REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034-7175

80 anos do IBGE

volume 76
número 241
janeiro / dezembro 2015

IBGE
Instituto Brasileiro de Geografia e Estatística

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 76 número 241 jan/dez 2015

ISSN 0034-7175

R. Bras. Estat., Rio de Janeiro, v. 76, n. 241, p. 1-115, jan./dez. 2015

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.

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

O número 241 da Revista Brasileira de Estatística contém cinco artigos com variada contribuição, em que são abordados temas envolvendo propostas de metodologia e aplicações diversas. São usados métodos Bayesianos, planejamento ótimo, métodos de análise de séries temporais, de análise de agrupamentos, confiabilidade e amostragem. O artigo de Gustavo da Silva Ferreira descreve uma aplicação da metodologia para obtenção do planejamento amostral ótimo no tempo em estudos de infestação do Aedes aegypti. O artigo de autoria de Pedro Guilherme Costa Ferreira, José Lisboa Gondin Junior e Daiane Marcolino de Mattos propõe o uso do programa X-13ARIMA-SEATS para fazer o ajuste sazonal de séries temporais da Sondagem da Indústria de Transformação (FGV/IBRE). O artigo de Marcos Garrido de Oliveira, Luís Pérez Zotes, Osvaldo Luiz Gonçalves Quelhas, Valdecy Pereira e Carlos Antônio da Silva Carvalho apresenta a comparação entre os métodos de agrupamentos automáticos K-means e Partitioning Around Medoids aplicados em dados simulados, sem e com outliers e sugere o diagrama de decisão para definição do número de grupos. O artigo de Laurinete do Nascimento Bacelar dos Reis Ximenes, Ronaldo M. Salles, Paulo Afonso Lopes da Silva propõe um indicador de resiliência que pesquisa uma infraestrutura confiável desde o início de um projeto, visando a reduzir os custos com manutenções corretivas. Por fim, o artigo de Juan José Orellana-Cáceres, Sergio Raúl Muñoz-Navarro e Alex Manuel Antequeda-Campos propõe um método para estabelecer o número de cupons em amostra Respondent-drivensampling aplicada a populações pequenas.

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

Aproveito para informar que o novo sítio da Revista Brasileira de Estatística está no ar no endereço http://rbes-seer.ibge.gov.br e convido todos a visitá-lo e a submeter seus trabalhos. Quero deixar agradecimentos especiais a José André de Moura Brito pela sua colaboração para a melhoria da revista.

Desejo a todos que tenham uma excelente leitura.

Lúcia Pereira Barroso
Editora Responsável

Planejamento Amostral Ótimo em Estudos de Infestação de Aedes aegypti

Gustavo da Silva Ferreira [1]

Resumo

O processo de seleção dos tempos de amostragem para obtenção de índices de infestação de mosquitos e demais vetores de doenças precisa levar em conta questões de praticidade, eficácia e economia de recursos. Este trabalho descreve uma aplicação da metodologia para obtenção do planejamento amostral ótimo no tempo em estudos de infestação do Aedes aegypti. Assumindo um Processo Gaussiano subjacente, desenvolve-se um critério de decisão baseado na maximização da utilidade esperada permitindo-se penalizar ou bonificar os candidatos a ponto amostral de acordo com o seu poder preditivo e com o risco de surto de infestações associado. O procedimento de inferência segue a abordagem Bayesiana e é aplicado aos dados de infestação resultantes do LIRAa no período de 2005 e 2009 no Rio de Janeiro. A metodologia mostrou-se flexível e com grande potencial de utilização para órgãos governamentais ou agentes de saúde que necessitem otimizar seus recursos materiais e financeiros.

Palavras-chave: amostragem; infestação; função utilidade; séries temporais; otimização; Aedes.

[1] Escola Nacional de Ciências Estatísticas – ENCE/IBGE (CEGRAD)

1. Introdução

A dengue é a arbovirose mais importante do mundo e, entre as doenças re-emergentes, é a que se constitui em problema mais grave de saúde pública (Tauil, 2002). A dengue é transmitida ao homem por algumas espécies do gênero *Aedes*, sendo o *Aedes aegypti* a mais importante. Em áreas urbanas, o *Aedes aegypti* costuma apresentar maior densidade populacional que o *Aedes albopictus*, outro vetor responsável pela transmissão da doença (Lima-Camara et al.,2006, Braks et al., 2003).

Assim, a dengue apresenta um ciclo epidemiológico urbano e seus principais elementos são o homem — o hospedeiro do vírus — e o *Aedes aegypti* — o vetor de transmissão (Rebêlo et al., 1999). Além do homem, não se conhece nenhum outro hospedeiro de importância significativa que atue como reservatório (Medronho, 1995). Os principais determinantes e implicações do controle da dengue atualmente no Brasil são discutidos em inúmeros artigos na literatura (Teixeira et al., 2009).

Em relação ao planejamento amostral para monitoramento do vetor, a periodicidade, os custos, a forma de coleta, as ferramentas utilizadas, as questões sociais e os aspectos metodológicos associados são de grande relevância e vários estudos podem ser encontrados na literatura tratando destes temas, como os estudos de Alves et al. (1991), Bracco e Fabbro (1995), Morato et al. (2005), Neto et al. (2006) e Lagrotta et al. (2008), entre outros.

No contexto da série histórica de infestação do mosquito *Aedes aegypti*, o processo de seleção dos tempos de amostragem para obtenção de índices de infestação precisa levar em conta questões de praticidade, eficácia e economia de recursos. Pela praticidade e reprodutibilidade demonstradas, os índices de infestação baseados em larvas acabam sendo os mais empregados como medidas de infestação e como indicadores de risco de transmissão de dengue (Gomes, 1998). Dentre os mais conhecidos destacam-se o índice Predial, que avalia o percentual de edifícios infestados, e o índice de Breteau, que avalia o percentual de recipientes positivos com larvas por domicílio. Na literatura há referência de que com um índice de infestação predial menor que 1% e um índice de Breteau abaixo de 5% não haveria transmissão de dengue. Entretanto, a relação entre índices baixos de infestação e a não-ocorrência de epidemias não é clara (Gomes, 1998 e Costa et al., 2012).

Além disso, nenhum dos índices é suficientemente capaz de medir com precisão a intensidade de infestação (Gomes, 1998). Estes fatos evidenciam a dificuldade e a complexidade de medidas de combate ao Aedes aegypti.

A fim de orientar os órgãos de saúde existentes no Brasil, a Secretaria de Vigilância em Saúde do Ministério da Saúde disponibilizou um documento contendo a metodologia para realização de um diagnóstico rápido conhecido como Levantamento de Índice Rápido de Infestação por Aedes aegypti – LIRAa (Brasil, Ministério da Saúde, 2013). Neste documento são apresentados os fundamentos que dão sustentação à metodologia que permite obter resultados dentro de uma segurança estatística aceitável, abordando critérios para a delimitação dos estratos, cuidados durante o planejamento das ações, desenho do plano amostral, formulários, dentre outras informações.

O LIRAa tem a vantagem de produzir de uma forma relativamente rápida os índices de infestação larvários Predial, definido como o percentual de edifícios pesquisados com a presença de larvas de Aedes aegypti, e Breteau, definido como o número de recipientes positivos em relação ao número de imóveis pesquisados. O LIRAa também pode substituir o levantamento tradicional, que, normalmente, apresenta o resultado somente após o fechamento de um ciclo bimestral de trabalho (Brasil, Ministério da Saúde, 2013).

Na constante busca pela otimização dos recursos para avaliação dos índices de infestação, evidencia-se a necessidade de planejar a periodicidade com que os agentes de saúde serão enviados a campo para obtenção do LIRAa. O conhecimento do comportamento da série de infestação do mosquito pode ser útil para definir esta periodicidade. Neste sentido, um bom planejamento deve ser realizado a fim de mobilizar os recursos disponíveis para os períodos de maior risco de surtos de infestação do mosquito e de epidemias de dengue.

Do ponto de vista da estatística, ao lidar-se com fenômenos que se distribuem de forma contínua no espaço ou no tempo, geralmente busca-se direcionar os esforços para obtenção de mapas e séries temporais preditas em locais ou instantes onde não se possui medição alguma. A utilização de Processos Gaussianos para descrever comportamentos de funções permite a quantificação da incerteza a respeito de variáveis que se distribuem no tempo ou espaço de forma contínua. Sob o enfoque bayesiano, estes processos podem ser utilizados para avaliar o comportamento de funções matemáticas desconhecidas (O'Hagan, 1978).

Assim, objetiva-se neste trabalho desenvolver um critério que auxilie no planejamento do momento ótimo no tempo para realização de novo trabalho de campo a fim de avaliar os índices de infestação na cidade do Rio de Janeiro. Para isso, serão utilizados Processos Gaussianos e métodos baseados na maximização de funções utilidades para obtenção do planejamento amostral ótimo levando-se em conta fatores associados ao poder de predição e ao risco de surtos de infestação do mosquito.

2. Materiais e Métodos

Os dados de infestação utilizados neste estudo referem-se aos índices de infestação predial produzidos pelos LIRAa realizados entre 02/01/2005 e 05/11/2009 na cidade do Rio de Janeiro, totalizando 18 levantamentos. Estes dados foram disponibilizados pela Secretaria Municipal de Saúde do município do Rio de Janeiro via Portal da Prefeitura da cidade do Rio de Janeiro (endereço eletrônico: http://www.rio.rj.gov.br, acessado em nov/2010).

No contexto da Geoestatística (Diggle e Ribeiro, 2007), podemos assumir que a infestação pelo mosquito na cidade do Rio de Janeiro pode ser modelada como um processo estocástico gaussiano $\{S(t): t \in D\}$, onde D representa o intervalo de tempo compreendido entre os anos de 2005 a 2009. Neste caso, considera-se que cada levantamento é uma observação da variável aleatória , Y_i , $i = 1,\ldots,18$, com distribuição condicional;

$$Y_i \mid S(t_i) \sim N(S(t_i)\,;\, \tau^2), \tag{1}$$

onde $S(t_i)$ representa o índice de infestação no tempo t_i e τ^2 representa a variabilidade associada aos erros de medição ou de efeitos de pequena escala. Em outras palavras, podemos considerar que Y é uma versão do processo estocástico S perturbada por um ruído aleatório. É importante salientar também que o período de realização do LIRAa compreende vários dias e que, para cada um dos 18 levantamentos, considerou-se o centro do período de coleta de dados como referência para o valor de t_i , referentes às datas 15/01/05, 13/04/05, 13/07/05, 12/10/05, 11/01/06, 12/07/06, 09/10/06, 14/01/07, 26/04/07, 05/07/07, 03/10/07, 04/06/08, 09/08/08, 16/10/08, 15/01/09, 18/03/09, 17/06/09 e 02/11/09.

Adicionalmente, vamos supor que o processo é estacionário com média e função de covariância dados por

$$E[S(t)] = \beta, \quad t \in (a,b) \text{ e}$$
$$Cov\,(S(t_1)\,;\,S(t_2)) = \sigma^2 \exp\{-\|t_1 - t_2\|/\phi\}, t_1, t_2 \in D, \tag{2}$$

onde ϕ é um parâmetro não-negativo que controla o alcance da correlação no tempo.

Após coletadas as observações Y_i , $i = 1,\ldots,18$, podemos predizer os valores do processo temporal em qualquer instante de tempo utilizando seus valores esperados e variâncias, isto é, $E[S(t) \mid \mathbf{y}]$ e $V[S(t) \mid \mathbf{y}]$, também conhecidos como *preditor* e variância de *krigagem*, respectivamente (Diggle e Ribeiro, 2007).

Para fenômenos que se distribuem de forma contínua no tempo podemos estar interessados em otimizar o seu planejamento amostral a fim de permitir que os novos instantes de coleta sejam aqueles que apresentem o maior potencial de ganho de informação a respeito do fenômeno. O planejamento de novos instantes d para amostragem é geralmente realizado via minimização ou maximização de funções que quantificam perdas ou utilidades de determinado ponto amostral.

Neste trabalho utilizou-se uma versão baseada no critério de maximização da utilidade esperada (Müller, 1999), definindo-se uma função utilidade $u(d, \theta, y_d)$, onde d representa o novo ponto do desenho amostral, $\theta = (\beta, \sigma^2, \phi, \tau^2)$ representa o vetor de parâmetros desconhecidos do modelo e y_d representa uma futura observação associada ao ponto amostral d. Desta forma, o ponto amostral candidato a instante ótimo será o valor de d, restrito ao intervalo de tempo fechado $[a, b]$, que maximiza a utilidade marginal

$$U(d) = \int u(d, \theta, y_d)\, p(y_d \mid \theta, \mathbf{y})\, p(\theta \mid \mathbf{y}) dy_d\, d\theta\,, \tag{3}$$

onde $p(y_d \mid \theta, \mathbf{y})$ representa a função densidade de probabilidade do modelo para y_d e $p(\theta \mid \mathbf{y})$ representa a distribuição a posteriori para θ.

A fim de incorporar pesos diferentes aos candidatos a ponto amostral de acordo com o seu poder preditivo e de acordo com o risco iminente de surto de infestações do mosquito, utilizou-se a seguinte função utilidade

$$u(d,\theta,y_d)=\int_a^b [V(S(t)\,|\,\theta,\mathbf{y})-V(S(t)\,|\,\theta,\mathbf{y},y_d)]dt+\exp\left\{-\alpha[\mathrm{P}(Y_d>k\,|\,\mathbf{y},y_d)-p_k]^2\right\}. \qquad (4)$$

Na função utilidade escolhida tem-se que o termo dentro da integral quantifica a redução da incerteza associada ao processo S após a escolha do instante d como novo ponto amostral, expressa por meio da diferença entre as variâncias. Por outro lado, o termo exponencial da expressão cumpre o papel de atribuir maior utilidade para os pontos amostrais que estejam associados aos instantes de tempo onde as probabilidades de eventos extremos (isto é, que ultrapassam um limite k são moderadas. No contexto de infestação pelo mosquito, este critério evita a escolha de instantes associados aos períodos de baixo risco, onde os gastos com o trabalho de campo para obtenção do LIRAa poderiam ser desnecessários, e de extremo risco, onde a realização da coleta poderia ser realizada demasiadamente tarde. De uma maneira geral, esta componente do critério permite que o modelo aponte para um instante no tempo que antecipe a ocorrência de um surto de infestação. O grau de moderação é definido pelo parâmetro p_k. Por fim, fixamos o parâmetro α a fim de ponderar o grau de influência desta componente de risco na função utilidade.

Para o procedimento de inferência utilizou-se uma versão discretizada desta função a partir de uma partição do período de estudo em M sub-intervalos, isto é,

$$u(d,\theta,y_d)\approx\sum_{j=1}^{M}\Delta\left[V(S(t_j)\,|\,\theta,\mathbf{y})-V(S(t_j)\,|\,\theta,\mathbf{y},y_d)\right]+\exp\left\{-\alpha[\mathrm{P}(Y_d>k\,|\,\mathbf{y},y_d)-p_k]^2\right\}, \qquad (5)$$

onde a variável aleatória $S(t_j),\ j=1,\cdots,M,$ representa o valor do processo S no centro do sub-intervalo $(t_{j-1},t_j]$ de comprimento Δ, e onde $a=t_0<t_1<\ldots<t_j<\ldots<t_M=b$.

Desta forma, o instante d com maior utilidade será aquele que minimizar a soma das variâncias preditivas de S calculadas para cada um dos M sub-intervalos $(t_{j-1}, t_j]$ e que, ao mesmo tempo, não estiver em períodos com índices de infestação extremos (muito altos ou muito baixos).

Como na prática o vetor de parâmetros do modelo é desconhecido, utiliza-se uma abordagem bayesiana e aproxima-se a utilidade marginal $U(d)$ a partir de simulações da distribuição a posteriori dos parâmetros do modelo e das respectivas distribuições preditivas.

O procedimento de inferência considera que o instante d é também um parâmetro a ser estimado no modelo (Müller, 1999). Para isto, define-se a seguinte função densidade de probabilidade *artificial*

$$h(d, \theta, y_d) \propto u(d, \theta, y_d)\, p(y_d \mid \theta, \mathbf{y})\, p(\theta \mid \mathbf{y}), \tag{6}$$

onde o termo *artificial* é utilizado para ressaltar que a quantidade d, apesar de não ser aleatória, é tratada no modelo como tal.

A partir de $h(d, \theta, y_d)$, podemos obter a distribuição de probabilidade marginal de d por meio de integração, isto é,

$$h(d) \propto \int u(d, \theta, y_d)\, p(y_d \mid \theta, \mathbf{y})\, p(\theta \mid \mathbf{y}) dy_d\, d\theta = U(d). \tag{7}$$

Assim, maximizar a utilidade esperada de d torna-se equivalente a encontrar a moda de $h(d)$. Em outras palavras, alteramos o problema de otimização para um problema de busca de moda de uma distribuição de probabilidade.

Adicionalmente foram especificadas distribuições a priori para todos os parâmetros desconhecidos do modelo. Para a média do processo utilizou-se $p(\beta) \propto 1$ e considerou-se prioris recíprocas para σ^2 e ϕ. Para $\tau^2_{rel} = \tau^2 / \sigma^2$ utilizou-se uma distribuição a priori Uniforme no intervalo $(0,1)$. Para a atualização de d no passo de simulação j, amostrou-se um valor proposto d_{prop} de uma distribuição normal truncada em $[a, b]$ e centrada em d_{j-1}. Para a realização da inferência foram utilizadas versões discretizadas aproximadas das distribuições dos parâmetros do modelo com o auxílio do pacote GeoR (www.r-project.gov).

3. Resultados

Para a análise dos dados realizou-se uma transformação logística na variável de infestação, alterando o seu suporte do intervalo $[0,1]$ para $\Re$, de forma a permitir o ajuste e as predições supondo um Processo Gaussiano subjacente. Foram utilizadas 5.000 amostras simuladas das distribuições a posteriori dos parâmetros do modelo para obtenção das distribuições preditivas e para o cálculo da função utilidade.

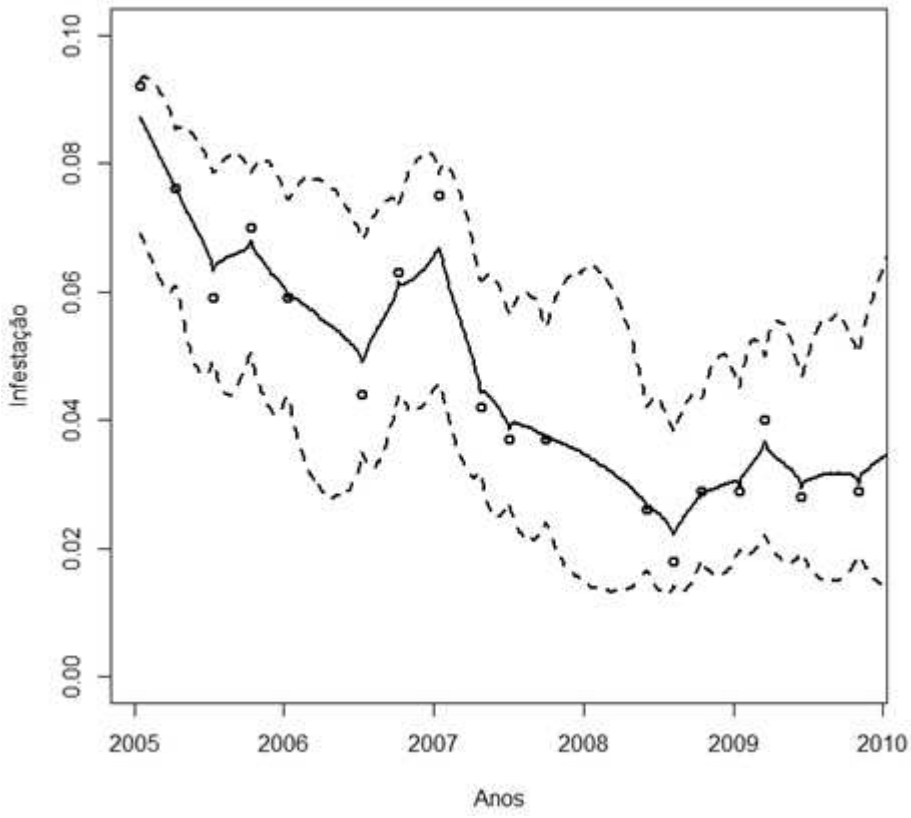
A Figura 1 apresenta os índices de infestação observados bem como os valores preditos destransformados para o período estudado e seus respectivos intervalos de 95% de credibilidade. Pode-se observar que os índices de infestação apresentaram uma queda no período 2005-2009, com um indicativo de estabilização abaixo dos 4% a partir de meados de 2009. A análise dos intervalos de 95% de credibilidade evidencia uma maior incerteza nos períodos entre a realização dos levantamentos LIRAa. No período que engloba o verão de 2007/2008, em especial, a incerteza associada aos índices de infestação se mostrou bastante elevada em virtude da não realização de levantamentos amostrais na cidade.

Para a avaliação da redução na variância preditiva da série temporal foi considerada uma partição de $M = 120$ sub-intervalos de tempo, referentes a uma janela de 240 dias existentes entre a realização do último LIRAa em nov/2009 e o final do mês de jun/2010. Assim, o intervalo $[a, b]$ da função utilidade $u(d, \theta, y_d)$ torna-se equivalente ao período de 03/nov/2009 à 30/jun/2010, onde cada sub-intervalo corresponde a um período de 48 horas (2 dias).

Posteriormente, esta partição foi utilizada para o cálculo da redução da variância preditiva na função utilidade avaliada para cada possível candidato a ponto amostral ótimo d.

O valor de k utilizado foi de 4%, baseado nos patamares de alerta atualmente empregados e recomendados pelo Ministério da Saúde no Brasil. A probabilidade máxima aceitável para este risco foi fixada em $p_k = 0{,}3$. A combinação destes parâmetros faz com que a função utilidade penalize os pontos amostrais associados a instantes de tempo onde a probabilidade de ocorrência de surtos de dengue seja muito elevada ou muito reduzida.

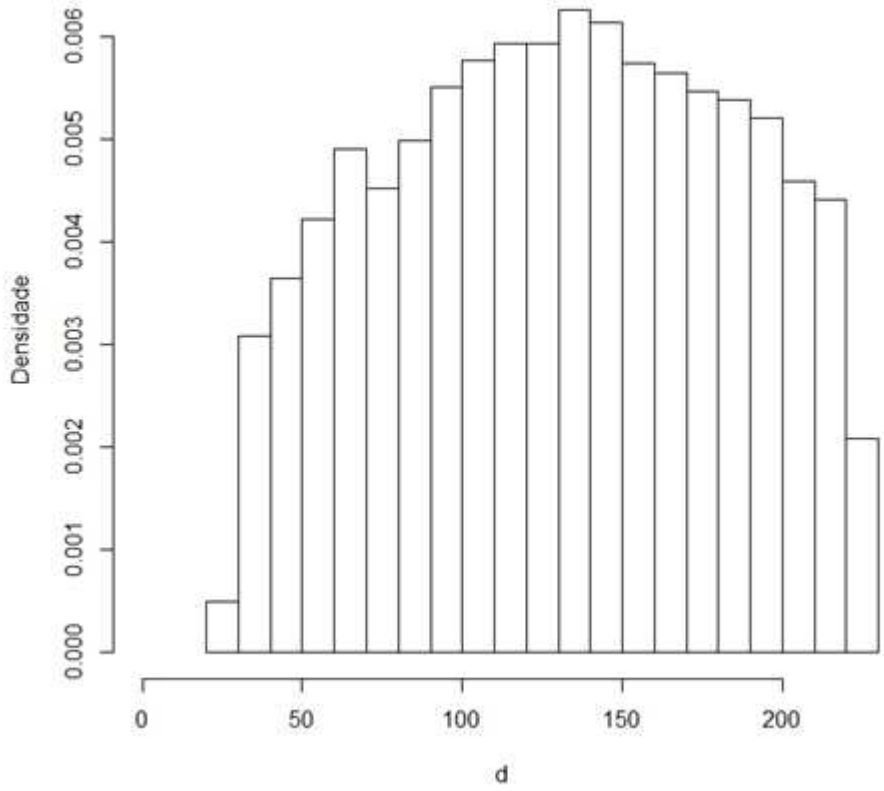
Figura 1: Valores observados (círculos), valores preditos (linha cheia) e respectivos IC 95% (linha tracejada) para a série temporal de infestação do *Aedes aegypti* no Rio de Janeiro.



A Figura 2 apresenta a distribuição artificial de *d* obtida a partir de 10.000 amostras simuladas considerando-se $\alpha = 100$. Pode-se observar que a moda da distribuição encontra-se no intervalo $[130, 140]$, o qual se refere ao período de 12/mar a 22/mar de 2010. Sendo assim, este resultado evidencia que o instante ótimo para realização de nova coleta ocorre em torno de 135 dias após a realização do LIRAa anterior.

Este resultado evidencia que a realização de um novo LIRAa com menos de 135 dias de diferença do último levantamento realizado seria demasiadamente custosa, já que o risco de surto de infestação ainda estaria sob controle. Por outro lado, a espera por mais de 135 dias incorporaria o risco de que um surto de infestação surgisse na cidade sem ser detectado.

Figura 2: Histograma da distribuição a posteriori artificial de d, onde o eixo das abscissas representa o número de dias após a realização do último LIRAa .



4. Discussão

O controle deste vetor é de grande importância para órgãos governamentais e inúmeros estudos recentes abordando aspectos relacionados ao tamanho populacional, taxa de sobrevivência e distribuição espaço-temporal da densidade de mosquitos têm sido realizados. No Brasil, destacam-se nesta área os trabalhos de Medronho (1995), Lagrotta et al. (2008), Souza-Santos e Carvalho (2000), Serpa et al. (2006), Urbinatti et al. (2007), Maciel-de-Freitas et al. (2008), Honório et al. (2009), Laporta et al. (2012) e Marteis et al. (2013), entre outros.

Tendo em vista que o padrão sazonal de infestação do Aedes aegypti ainda não está completamente determinado na literatura, a utilização dos processos gaussianos parece ser adequada para a realização de previsões de curto e médio prazo dos índices de infestação. Neste sentido, apesar de alguns estudos retratarem associações entre a série de infestação do *Aedes aegypti* e *Aedes albopictus* com fatores climáticos, estes também sugerem que a população destes mosquitos é afetada de forma diferenciada pela temperatura, umidade e precipitação, às vezes não apresentando sequer correlação significativa (Serpa et al., 2006, Urbinatti et al., 2007). Adicionalmente, a metodologia apresentada pode facilmente ser estendida de forma a contemplar covariáveis que eventualmente estejam correlacionadas com a série de infestação do *Aedes aegypti*.

Por outro lado, cabe destacar que para a utilização de processos gaussianos na modelagem de proporções ou taxas, como os índices de infestação prediais, é necessário ainda realizar transformações nos dados de forma a evitar problemas no processo de inferência e predição.

Em relação ao processo de escolha ótima de novos locais para amostragem, este se constitui de um tema bastante difundido na literatura da área de Estatística, além de possuir grande potencial de aplicação no planejamento amostral para controle de vetores. Por incorporar de forma natural a abordagem Bayesiana e a utilização de métodos de simulação intensivos, a metodologia empregada neste estudo torna-se bastante vantajosa no contexto de Processos Gaussianos. Estes processos, por sua vez, se constituem de modelos flexíveis que podem ser utilizados para modelar séries epidemiológicas irregularmente espaçadas no tempo.

O elevado custo computacional ao se realizar simulações das distribuições preditivas para obtenção das variâncias preditivas médias é o principal desafio desta metodologia, o que nos evidencia a necessidade de utilização de métodos aproximados de inferência. Neste sentido, a escolha do número de sub-intervalos M a partição do intervalo de predição estará bastante associada ao custo computacional envolvido.

Alguns estudos de simulação omitidos neste trabalho demonstraram que a decisão ótima é bastante sensível à escolha do valor de α, tendo em vista a escolha da função utilidade $u(d,\theta,y_d)$. Como alternativa, sugere-se a utilização de diferentes cenários a fim de reproduzir a análise sob a perspectiva de cenários pessimistas, realistas e otimistas em relação à disponibilidade de recursos. Assim, quanto maior o valor de α escolhido, maior será a utilidade alocada aos instantes de tempo onde é mais plausível a ocorrência de um surto de infestação.

Cabe salientar que o resultado obtido pode variar dependendo do patamar atual de infestação e que um novo planejamento amostral precisa ser realizado após cada realização de um levantamento do LIRAa. Apesar da redução dos custos de coleta, à medida que o intervalo entre a realização dos levantamentos cresce, maior se torna o risco de que um surto de infestação esteja ocorrendo sem ser detectado.

Dentre as vantagens desta abordagem, pode-se destacar que os parâmetros associados ao custo e ao risco de surtos de infestações permitem uma grande flexibilidade no planejamento de novos levantamentos, tornando esta metodologia uma satisfatória opção para atender as exigências de órgãos de saúde que necessitem otimizar seus recursos físicos e financeiros.

Por fim, cabe salientar que existem ainda outros aspectos, de ordem política, técnica e social, que também podem ser considerados na avaliação dos custos associados ao trabalho de controle de vetores e epidemias de dengue, conforme estudos realizados por Oliveira e Aragão (1995), Penna (2003) e Ferreira et al. (2009), entre outros. Entretanto, uma vez quantificados diferentes aspectos envolvidos nos custos, estes podem ser considerados na função utilidade descrita na seção de metodologia a fim de obter-se o planejamento amostral ótimo.

Referências bibliográficas

Alves MCGP, Gurgel SM, de Almeida MCRR (1991). Plano amostral para cálculo de densidade larvária de Aedes aegypti e Aedes albopictus no Estado de São Paulo, Brasil. Rev. Saúde Públ.; 25(4):251-256.

Bracco JE, Dal Fabbro ALD (1995). Amostragem por larva-única na vigilância de Aedes aegypti. Rev. Saúde Públ.; 29(2):144-146.

Braks MAH, Honório NA, Lourenço-De-Oliveira R, Juliano SA, Lounibos LP (2003). Convergent Habitat Segregation of Aedes aegypti and Aedes albopictus (Diptera: Culicidae) in Southeastern Brazil and Florida. J. Med. Entomol.; 40(6):785-794.

Brasil, Ministério da Saúde (2013). Levantamento rápido de Índices para Aedes aegypti — LIRAa para vigilância entomológica do Aedes aegypti no Brasil: metodologia para avaliação dos índices de Breteau e Predial e tipo de recipientes. Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância das Doenças Transmissíveis – Brasília - DF.

Costa F, Fattore G, Abril M (2012). Diversity of containers and buildings infested with Aedes aegypti in Puerto Iguazú, Argentina. Cad. Saúde Pública; 28(9):1802-1806.

Diggle P, Ribeiro PJ (2007). Model-based geostatistics. New York: Springer.

Ferreira ITRN, Veras MASM, Silva RA (2009). Participação da população no controle da dengue: uma análise da sensibilidade dos planos de saúde de municípios do Estado de São Paulo, Brasil. Cad. Saúde Pública; 25(12):2683-2694.

Gomes AC (1998). Medidas dos níveis de infestação urbana para Aedes (Stegomyia) aegypti e Aedes (Stegomyia) albopictus em programa de vigilância entomológica. Informe Epidemiológico do SUS – IESUS; VII(3), jul/set.

Honório NA, Castro MG, de Barros FSM, Magalhães MAFM, Sabroza PC (2009). The spatial distribution of Aedes aegypti and Aedes albopictus in a transition zone, Rio de Janeiro, Brazil. Cad. Saúde Pública; 25(6):1203-1214.

Lagrotta MTF, Silva WC, Souza-Santos R (2008). Identification of key areas for Aedes aegypti control through geoprocessing in Nova Iguaçu, Rio de Janeiro State, Brazil. Cad. Saúde Pública; 24(1):70-80.

Laporta GZ, Ribeiro MC, Ramos DG, Sallum MAM (2012). Spatial distribution of arboviral mosquito vectors (Diptera, Culicidae) in Vale do Ribeira in the South-eastern Brazilian Atlantic Forest. Cad. Saúde Pública; 28(2):229-238.

Lima-Camara TN, Honório NA, Lourenço-de-Oliveira R (2006). Frequência e distribuição espacial de Aedes aegypti e Aedes albopictus (Diptera, Culicidae) no Rio de Janeiro Brasil. Cad. Saúde Pública; 22(10):2079-2084.

Maciel-de-Freitas R, Eiras AE, Lourenço-de-Oliveira R (2008). Calculating the survival rate and estimated population density of gravid Aedes aegypti (Diptera, Culicidae) in Rio de Janeiro, Brazil. Cad. Saúde Pública; 24(12):2747-2754.

Marteis LS, Steffler LM, Araújo KCGM, Santos RLC (2013). Identificação e distribuição espacial de imóveis-chave de Aedes aegypti no bairro Porto Dantas, Aracaju, Sergipe, Brasil entre 2007 e 2008. Cad. Saúde Pública; 29(2):368-378.

Medronho RA (1995). Geoprocessamento e saúde: uma nova abordagem do espaço no processo saúde-doença. Rio de Janeiro: Núcleo de Estudos em Ciência e Tecnologia, Fundação Oswaldo Cruz.

Morato VCG, Teixeira MG, Gomes AC, Bergamaschi DP, Barreto ML (2005). Infestation of Aedes aegypti estimated by oviposition traps in Brazil. Rev. Saúde Públ; 39(4):553-558.

Müller P (1999). Simulation-based optimal design. Em Bayesian Statistics 6 (eds J.M. Bernardo, J.O. Berger, A.P. Dawid e A.F.M. Smith). Oxford University Press; 459-474.

Neto FC, Barbosa AAC, Cesarino MB, Favaro EA, Mondini A, Ferraz AA, Dibo MR, Vicentini ME (2006). Controle do dengue em uma área urbana do Brasil: avaliação do impacto do Programa Saúde da Família com relação ao programa tradicional de controle. Cad. Saúde Pública; 22(5):987:997.

O'Hagan A (1978). Curve fitting and optimal design for predictions. J R Stat Soc Series B; 40:1-42.

Oliveira SJ, Aragão MB (1995). Algumas sugestões para o controle dos mosquitos do gênero Aedes no Brasil. Cad. Saúde Pública; 11(4):629-630.

Penna MLF (2003). Um desafio para a saúde pública brasileira: o controle do dengue. Cad. Saúde Pública; 19(1):305-309.

Rebêlo JMM, Costa JML, Silva FS, Pereira YNO, Silva JMS (1999). Distribuição de Aedes aegypti e do dengue no Estado do Maranhão, Brasil. Cad. Saúde Pública; 15(3):477-486.

Serpa LLN, Costa KVRM, Voltolini JC, Kakitani I (2006). Variação sazonal de Aedes aegypti e Aedes albopictus no município de Potim, São Paulo. Rev. Saúde Públ.; 40(6):1101-1105.

Souza-Santos R, Carvalho MS (2000). Análise da distribuição espacial de larvas de Aedes aegypti na Ilha do Governador, Rio de Janeiro, Brasil. Cad. Saúde Pública; 16(1):31-42.

Tauil P.L. (2002). Aspectos críticos do controle do dengue no Brasil. Cad. Saúde Pública; 18(3):867-871.

Teixeira MG, Costa MGN, Barreto F, Barreto ML (2009). Dengue: twenty-five years since reemergence in Brazil. Cad. Saúde Pública; 25 Sup 1:S7-S18.

Urbinatti PR, Menezes RMT, Natal D (2007). Sazonalidade de Aedes albopictus em área protegida na cidade de São Paulo, Brasil. Rev. Saúde Públ.; 41(3):478-481.

Abstract

Sampling procedures in time to evaluate mosquitoes infestation and other disease vectors must take into account issues of practicality, efficiency and economy of resources. This paper describes an application of the methodology for obtaining optimal design time in *Aedes aegypti* infestation studies. Starting from an underlying Gaussian process assumption one can develops a decision criterion based on maximizing expected utility allowing penalize or reward the sampling points candidates according to its predictive power and its related infestation risk's outbreak. The inference procedure follows the Bayesian approach and is applied to LIRAa infestation's data for the years 2005 to 2009 in Rio de Janeiro. The methodology has proved to be very flexible with great potential for Health Agencies or Government's Agencies that need to optimize their material and financial resources.

Keywords: sampling; infestation; utility function; time series; optimization; *Aedes*

X-13ARIMA-SEATS: Uma proposta de ajuste sazonal para séries temporais de Business Tendency Survey

Pedro Guilherme Costa Ferreira [1]
José Lisboa Gondin Junior [2]
Daiane Marcolino de Mattos [3]

Resumo

Este artigo propõe o uso do programa X-13ARIMA-SEATS para fazer o ajuste sazonal das séries temporais de *business tendency survey* da Sondagem da Indústria de Transformação (FGV/IBRE). Por ser um índice composto por seis indicadores, discutiu-se o método de ajuste sazonal para séries temporais agregadas mais adequado (método direto e indireto). Como resultado, observou-se que, apesar de o método indireto ser mais eficaz em captar o padrão sazonal de cada série temporal que compõe o índice de confiança da indústria, o método direto apresentou fatores sazonais mais estáveis ao longo do tempo.

Palavras-chave: Tendência dos Negócios, X-13ARIMA-SEATS, Séries Temporais, R software, Sondagem da Indústria (FGV/IBRE).

[1] Núcleo de Métodos Estatísticos e Computacionais (FGV/IBRE), end. Rua Barão de Itambi, 60, Botafogo, RJ, Brasil. E-mail: pedro.guilherme@fgv.br
[2] E-mail: gondim@gmail.com
[3] Núcleo de Métodos Estatísticos e Computacionais (FGV/IBRE). E-mail: daiane.mattos@fgv.br

1. Introdução

Uma série temporal, segundo a decomposição clássica, pode ser decomposta em quatro componentes não observáveis: tendência, sazonalidade, ciclo e erro. A sazonalidade, principal objeto deste estudo, é causada por movimentos oscilatórios de mesma periodicidade que ocorrem em período intra-anual, como variações climáticas, férias, feriados, entre outros. A ocorrência desses eventos pode levar a conclusões inadequadas a respeito da série temporal em estudo. Por exemplo, a oferta de emprego costuma aumentar no final do ano devido às festas natalinas, isto é, há uma demanda maior por bens e serviços, elevando o nível de contratações de pessoas. Porém, como a maioria das vagas é temporária, geralmente, há diminuição no nível de pessoal ocupado no período seguinte. Para a análise econômica, o importante é detectar a diferença entre o que periodicamente ocorre e o que de fato ocorre de diferente naquele período específico, possibilitando observar a tendência e o ciclo da variável.

Para tal, precisa-se de uma ferramenta adequada que consiga remover essa componente (a sazonalidade). A remoção da sazonalidade de uma série temporal é chamada de ajuste sazonal ou dessazonalização. Diversos órgãos de estatística relevantes, como IBGE (2014), Eurostat (2014) e BLS (2014), aplicam métodos estatísticos que permitem um ajuste sazonal de qualidade confiável.

Na literatura, é possível encontrar diversas metodologias/softwares que permitem a remoção da sazonalidade de uma série temporal, como por exemplo, (i) *Dummies* Sazonais (Zellner (1979); Aguirre & Aguirre (1999)); (ii) *Holt-Winters* (Rasmussen, 2004); (iii) Modelos Estruturais (Harvey & Shepard (1993); Plosser (1979); Koopman, Harvey, Doornik & Shepard (2009)); (iv) Dainties (Fok, Franses, & Paap, 2005); (v) TRAMO-SEATS (Gómez & Maravall (1992, 1994, 1996, 1997, 2001); Gómez, Maravall & Peña (1999), European Comission Grant (2007)); e (vi) X11/X12/-ARIMA e X-13ARIMA-SEATS (Shiskin, Young & Musgrave (1967); Dagum (1980); Findley & Hood (1998); Findley, Monsell, Bell, Otto & Chen (1998); U.S. Bureau of the Census (2013)).

Apesar da variedade de opções, esse artigo está limitado a explicação da execução do último método (X-13ARIMA-SEATS), que é uma evolução do programa de ajuste sazonal X11, este desenvolvido pelo *U.S. Bureau of the Census* em 1965 (Shiskin, Young and Musgrave, 1967). Utilizava, basicamente, filtros de médias móveis[1] para estimar as componentes de tendência e sazonalidade, porém acarreta problemas com o início e o fim da série temporal pela baixa qualidade de filtros assimétricos, além de requerer um grande número de revisões. Em 1980, Estella Dagum do *Statistics Canada* desenvolveu o X11-ARIMA que permitia a extensão do início e do final das séries temporais através de um modelo ARIMA (Dagum, 1980). O método reduziu o número de revisões. Em meados de 1990, o X12-ARIMA foi implementado pelo *U.S. Bureau of the Census* possibilitando grandes melhorias ao X11-ARIMA (Findley & Hood, 1998). A principal foi a inserção da etapa de pré-ajuste em que a série temporal, antes de ser dessazonalizada, é filtrada por variáveis regressoras (regARIMA) que podem afetar o seu comportamento, como, por exemplo, a quantidade de dias úteis, feriados como páscoa e carnaval e também outliers. O X12-ARIMA era e ainda é utilizado em diversos órgãos internacionais além do *U.S. Bureau of the Census*, como, por exemplo, o IBGE no Brasil (2010), Eurostat na Europa (2009), *The Office for National Statistics* no Reino Unido (2007), *Statistics Canada* (2014).

Atualmente, já está disponível a nova versão do programa X12-ARIMA, o X-13ARIMA-SEATS, que une o X12-ARIMA com o TRAMO/SEATS. Com a finalidade de mostrar o uso do programa, o ajuste sazonal será aplicado às séries temporais de *Business Tendency Survey* da Sondagem da Indústria, divulgadas mensalmente pela FGV/IBRE (2014). Além disso, serão discutidos alguns pontos com relação à aplicação do ajuste sazonal via método direto e indireto, que são métodos usados em séries temporais que surgem a partir da agregação de outras séries temporais, e à estabilidade dos fatores sazonais.

---

[1] Existem outras metodologias que também utilizam filtros de medidas móveis, como, por exemplo, STL (Cleveland, Cleveland, McRae, & Terpenning 1990).

Dessa forma, além dessa introdução, este artigo está dividido em mais três seções. Na próxima seção, discute-se o software X-13ARIMA-SEATS e os métodos de dessazonalização para séries agregadas. Na seção 3, apresenta-se um estudo de caso aplicado à Sondagem da Indústria de Transformação (FGV/ IBRE), no qual é feito o ajuste sazonal, análises de adequação e, também, uma análise fora da amostra para a estabilidade dos fatores sazonais. Por fim, fazem-se as conclusões de relevo na seção 4.

2. X-13ARIMA-SEATS e Métodos de Ajuste Sazonal para séries agregadas

2.1 X-13ARIMA-SEATS

O X-13ARIMA-SEATS, criado em julho de 2012, é um programa de ajuste sazonal desenvolvido por *U. S. Census Bureau* com o apoio do *Bank of Spain*. Conforme observado na introdução, é uma versão aprimorada do X11 (Shiskin, Young and Musgrave, 1967). O programa é a junção dos *softwares* X12-ARIMA e TRAMO/SEATS com melhorias. As melhorias incluem uma variedade de novos diagnósticos que ajudam o usuário a detectar e corrigir inadequações no ajuste. O programa também inclui diversas ferramentas que superaram problemas de ajuste e permitiram um aumento na quantidade de séries temporais econômicas que podem ser ajustadas de maneira adequada (*U.S. Bureau of the Census*, 2013), além da possibilidade de realizar um pré-ajuste na série temporal, isto é, uma correção antes de ser feito, de fato, o ajuste sazonal.

Para realizar o ajuste sazonal, os seguintes passos[2] serão executados:

(i) visualização gráfica e inspeção visual da série temporal;

(ii) verificação da sazonalidade na série temporal;

(iii) aplicação do pré-ajuste e ajuste sazonal via X-13ARIMA-SEATS[3] ((Eurostat, Pre-treatament, 2014); Livsey, Pang & McElroy (2014));

(iv) diagnóstico do ajuse sazonal (Eurostat, 2014).

[2] Para maiores detalhes sobre esses passos ver: (Ferreira, Gondin Jr, & Mattos, 2015)

[3] O programa X-13ARIMA-SEATS pode realizar o ajuste de duas formas: através do X11 ((Shiskin, Young and Musgrave, 1967), (Findley, Monsell, Bell, Otto, & Chen, 1998)) ou através do SEATS. Neste artigo, optou-se pelo SEATS.

A análise gráfica de uma série temporal permite visualizar suas características para uma boa modelagem, por exemplo: seu padrão sazonal, quebras estruturais, possíveis *outliers*, se há necessidade (e possibilidade) de usar transformação logarítmica nos dados.

Também é sugerido testar estatisticamente se há indícios de sazonalidade na série temporal (Eurostat, 2009), uma vez que não é indicado, nem necessário, dessazonalizar uma série temporal sem sazonalidade. O X-13 oferece um diagnóstico de sazonalidade na série temporal que será explorado mais adiante.

O X-13, basicamente, funciona em duas etapas: pré-ajuste e ajuste sazonal. Na primeira, o software pode corrigir a série de efeitos determinísticos. É nesta etapa que o usuário pode especificar, por exemplo, *outliers* e efeitos do calendário (Páscoa, Carnaval, etc). Na segunda etapa, é feito o ajuste sazonal de fato. A execução do programa no modo automático pode trazer um ajuste sazonal de boa qualidade. O programa no modo automático verifica, entre outras coisas, se há necessidade de transformação logarítmica nos dados; se existem *outliers (additive, level shift e temporary change*); se há efeitos de calendário; e a ordem do modelo ARIMA. Essas verificações automáticas podem poupar o tempo do usuário e ajudá-lo na escolha de um bom modelo, principalmente na etapa do pré-ajuste. No entanto, este modelo precisa ser avaliado e o X-13ARIMA-SEATS fornece algumas ferramentas[4] para essa finalidade:

- *QS statistic*: É um diagnóstico para verificar a existência de sazonalidade em uma série temporal cuja hipótese nula é "não há evidências de sazonalidade". A tabela 1, a seguir, resume em quais séries temporais o programa calcula o teste de sazonalidade. Em um bom ajuste sazonal, o diagnóstico dado pela estatística QS, permitiria concluir indícios de sazonalidade somente na série original e não nas restantes.

[4] Há uma gama de recursos oferecidos pelo X-13ARIMA-SEATS que ainda não foram explorados nesse trabalho. Mais informações ver X-13ARIMA-SEATS Reference Manual Acessible HTML Output Version (U.S. Census Bureau, 2015).

- *Ljung-Box statistic*: O teste de Ljung and Box (1978) verifica a existência de autocorrelação em uma série temporal. O X-13 mostra o resultado desse teste aplicado aos resíduos do modelo SARIMA estimado na defasagem 24. Espera-se que os resíduos não sejam autocorrelacionados (hipótese nula).

Tabela 1 - Séries Temporais disponíveis para o diagnóstico dado pela estatística QS

Codificação	Significado
qsori	série original
qsorievadj	série original corrigida por *outliers*
qsrsd	resíduos do modelo SARIMA
qssadj	série com ajuste sazonal
qssadjevadj	série com ajuste sazonal corrigida por *outliers*
qsirr	componente irregular
qsirrevadj	componente irregular corrigida por outliers

- *Shapiro-Wilk statistic*: O teste de Shapiro & Wilk (1965) verifica se a distribuição de um conjunto de dados é normal. O X-13 mostra o resultado desse teste aplicado aos resíduos do modelo SARIMA estimado. Espera-se que os resíduos sigam distribuição normal (hipótese nula), no entanto esta não é uma propriedade obrigatória para considerar o ajuste sazonal adequado.
- Gráfico *SI ratio*: Útil para verificar se a decomposição das componentes da série temporal foi feita adequadamente. Espera-se que os fatores sazonais acompanhem o SI (componentes sazonal e irregular agregadas[5]), indicando que o SI não é dominado pela componente irregular.

Após a análise das ferramentas de diagnóstico, caso alguma não conformidade seja detectada no modelo automático, o modelo deve ser reajustado e rediagnosticado.

[5] Se for utilizada a decomposição aditiva (sem transformação log) então SI é a soma da componente sazonal e da componente irregular ($S+I$). Caso contrário, usa-se a multiplicação: $S \times I$.

2.2 Métodos de ajuste para séries agregadas

Comumente, uma série temporal pode surgir a partir da agregação de outras séries temporais. Nesses casos, é necessário discutir sobre a forma mais adequada de dessazonalização. Há dois métodos de ajuste sazonal para uma série agregada: (i) direto e (ii) indireto. Ambos são definidos a seguir.

> (i) *Método Direto*: o ajuste é feito apenas sobre a série já agregada, ou seja, primeiro ocorre a agregação[6] das séries temporais e, em seguida, a dessazonalização da série resultante.
>
> (ii) *Método Indireto*: o ajuste é feito, individualmente, em cada série temporal que compõe a série final de interesse e, em seguida, é feita a agregação das séries dessazonalizadas.

Os principais trabalhos sobre sazonalidade não apresentam evidência teórica ou empírica que indique uma preferência clara de um método em detrimento do outro em um contexto geral. Estudos empíricos indicam não haver grande diferença entre os dois métodos (Eurostat, Seasonal Adjustment, 2014), principalmente, quando todas as séries que compõem a série final, ou pelo menos as séries com maior importância (peso) na agregação, possuem sazonalidade identificável. Contudo, questões subjetivas devem ser consideradas em cada caso.

O método direto pode proporcionar maior transparência e precisão, por exemplo, no caso em que a série agregada é mais bem conhecida e compreendida do que suas componentes. Já o método indireto é indicado quando as séries temporais que compõem a série agregada têm padrões muito distintos, pois assim é possível uma atenção mais específica para cada série resultando em um pré-ajuste mais eficaz. Isso ocorre, uma vez que cada série temporal pode ser influenciada por efeitos determinísticos diferentes ou até mesmo apresentar um resultado diferente com relação ao mesmo evento determinístico, por exemplo, o impacto da crise de 2008 surge nas séries temporais (ou indicadores) da Sondagem da Indústria da Transformação em períodos distintos com impactos significativamente diferentes (ver seção 3). Assim, o método indireto pode

[6] Neste artigo a agregação é dada via método de padronização. Para mais informações ver European Comission (2002).

proporcionar um melhor diagnóstico, e consequentemente, uma melhor estimativa dos fatores sazonais.

3. Resultados: Um estudo de caso para a Sondagem da Indústria

A Sondagem da Indústria (FGV/IBRE, 2014) fornece indicações sobre o momento atual e as tendências de curto prazo do setor industrial brasileiro. Devido à sua importância no cenário econômico, por fornecer informações para a tomada de decisões empresariais e para a formulação de políticas econômicas pelo Governo, entre outras utilidades, é relevante uma análise de dados rigorosa em relação aos dados dessazonalizados divulgados pela instituição. A série histórica teve seu início em 1966, sendo divulgada trimestralmente até outubro de 2005 e mensalmente desde então.

O Índice de Confiança da Indústria (ICI) é o indicador-síntese da pesquisa. É calculado a partir da média ponderada pelo inverso da volatilidade das seis variáveis a seguir (figura 1):

(i) Nível de Demanda Global (NDG): Avaliação a respeito do nível atual de demanda por produtos da empresa. As respostas são desagregadas de acordo com a classificação da origem da demanda, entre interna (Nível de Demanda Interna – NDI) ou externa (Nível de Demanda Externa – NDE).

(ii) Nível de Estoques (NdE): Avaliação a respeito do nível atual dos estoques de produtos fabricados pela empresa.

(iii) Situação Atual dos Negócios (SAN): Ao responder este quesito, a empresa leva em consideração fatores de ordem microeconômica, tais como margens de lucro e faturamento, e fatores macroeconômicos que afetam o seu desempenho, como taxas de juros e câmbio. A variável é avaliada no momento de realização da pesquisa.

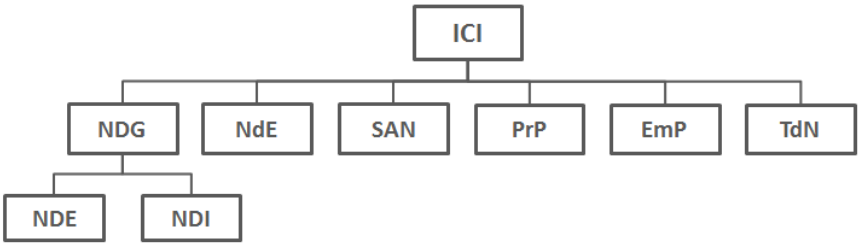
(iv) Produção Prevista (PrP): Perspectiva para a produção no trimestre seguinte.

(v) Emprego Previsto (EmP): Perspectiva em relação ao contingente de mão-de-obra empregado pela empresa no trimestre seguinte.

(vi) Tendências dos Negócio (TDN): Semelhants a SAN, porém a avalição é dada quanto às perspectivas para os próximos seis meses.

O valor do ICI em cada período permite avaliar o grau de aquecimento da atividade industrial: se o índice se encontra acima de 100, está acima da média histórica do período 1996-2005, refletindo, portanto, satisfação do setor industrial com o estado dos negócios e/ou otimismo em relação ao futuro. Analogamente, para valores abaixo dessa referência, configura-se uma situação de insatisfação/pessimismo (FGV/IBRE, 2014).

Figura 1 - Composição do Índice de Confiança da Indústria.



Fonte: (FGV/IBRE, 2014)

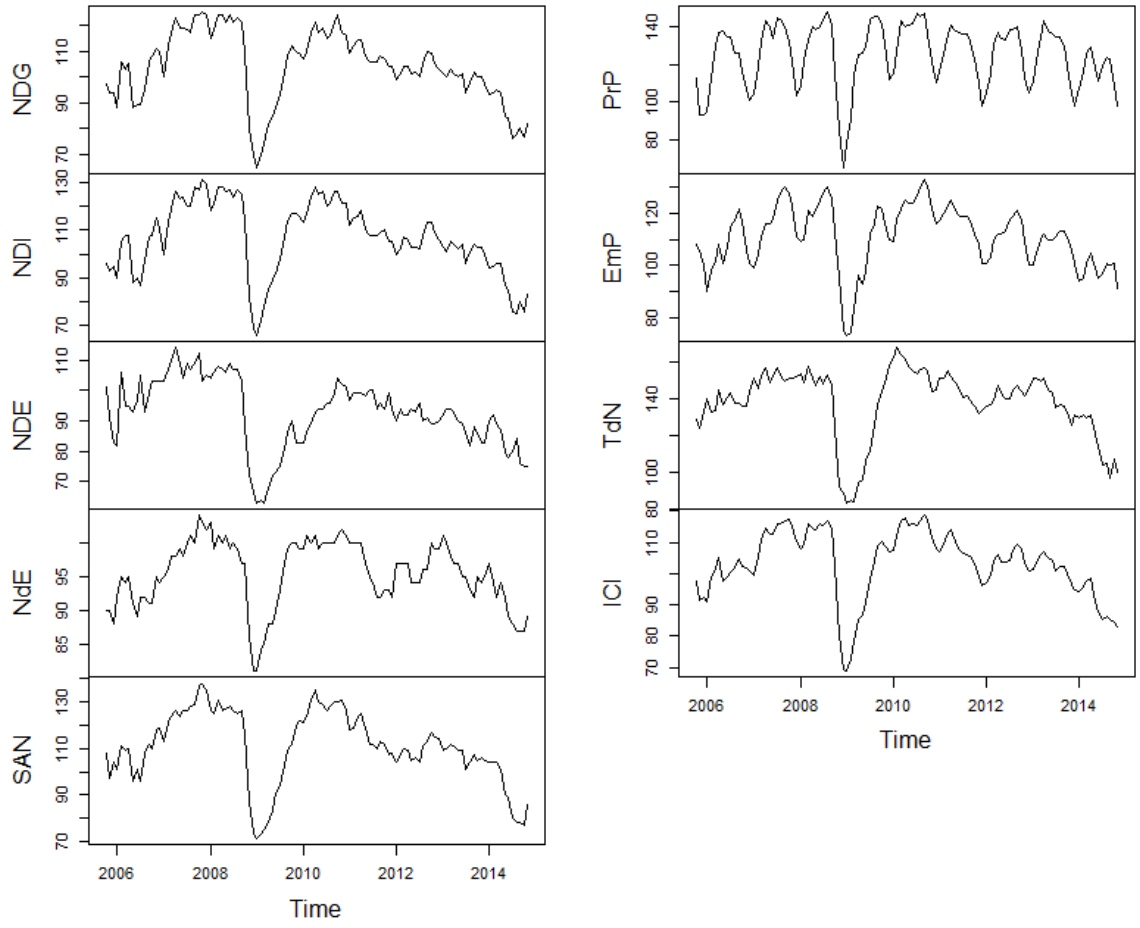
Para obter o ICI com ajuste sazonal, serão utilizados os métodos direto e indireto. No método direto, todas as séries que compõem o ICI serão agregadas e, por fim, a série resultante será ajustada. No método indireto, cada uma das séries que compõem o ICI será ajustada pelo método direto, com exceção da série NDG (método indireto), nesta as séries NDE e NDI serão ajustadas e agregadas para formar a série NDG com ajuste sazonal. Por fim, obtidas as séries com ajuste sazonal, estas serão agregadas e resultarão no ICI com ajuste sazonal.

Esse estudo de caso segue os passos delineados na seção 2.1, isto é: primeiro, será feita uma análise exploratória das séries temporais, pela qual se revelam indícios de existência de sazonalidade e outros comportamentos da série temporal; em seguida, há a seção de ajuste sazonal, em que é feito o teste de confirmação de sazonalidade e, consequentemente, a dessazonalização das séries temporais que compõem o ICI. Nessa seção são abordados os métodos direto e indireto, e é feita a comparação entre os dois. Por fim, perfaz-se a análise fora da amostra para estabilidade dos fatores sazonais, com o intuito de comparar a série ajustada final com a série obtida a partir dos fatores previstos.

3.1 Análise Exploratória dos Dados

Nesta seção, faz-se a análise exploratória dos dados com a finalidade de descrever o comportamento das séries temporais de interesse relatadas anteriormente. Na figura 2, com exceção da Produção Prevista (PrP), todas as séries que compõem o ICI sugerem uma quebra estrutural, em outras palavras, apresentam alteração no nível e flutuações irregulares ao longo do tempo. Também é notável que no período que antecede a crise econômica (final de 2008) há um crescimento na expectativa da atividade industrial, entretanto, ao longo do tempo, percebe-se um decrescimento em grande parte das séries.

Figura 2 - Séries históricas da Sondagem da Indústria – out/2005 a jun/2014.



Fonte: autores com base nos dados da (FGV/IBRE, 2014).

Além da inspeção visual sobre as séries temporais, aplicou-se também o teste de presença de raiz unitária *Augmented Dickey-Fuller,* ADF (Dickey & Fuller, 1979), visto a seguir na tabela 2. Nota-se a presença de raiz unitária no ICI e em grande parte de suas componentes. Embora o teste ADF não tenha detectado raiz unitária nas séries NDG e NDI, há evidências estatísticas sobre tendência determinística.

Tabela 2 - Teste de raiz unitária Augmented Dickey-Fuller.

	Tipo de Equação	Lag	Estatística de teste τ	Valor crítico	Conclusão
ICI	constante	23	-2,0402	-2,88	Com raiz unitária + constante
NDG	constante + tendência	12	-3,8689	-3,43	Sem raiz unitária + tendência determinística
NDE	constante	12	-2,4525	-2,88	Com raiz unitária + constante
NDI	constante + tendência	12	-3,7371	-3,43	Sem raiz unitária + tendência determinística
NdE	constante	12	-3,8933	-2,88	Sem raiz unitária + constante
SAN	constante	18	-2,6898	-2,88	Com raiz unitária + constante
PrP	constante	12	-3,3701	-2,88	Sem raiz unitária + constante
EmP	sem constante e sem tendência	23	-0,744	-1,95	Com raiz unitária s/ constante e s/ tendência
TdN	constante	23	-2,8717	-2,88	Com raiz unitária + constante

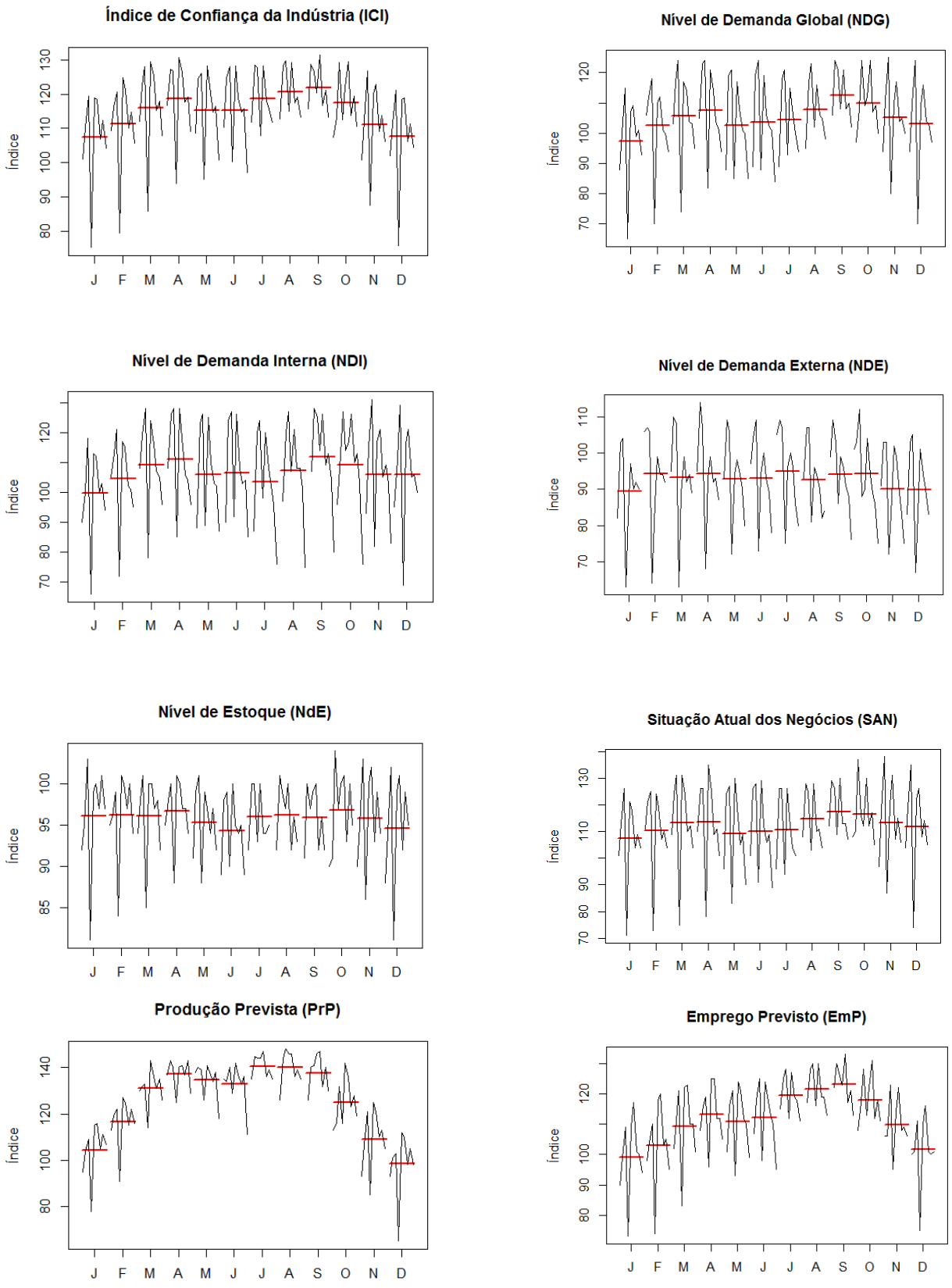
Fonte: autores.

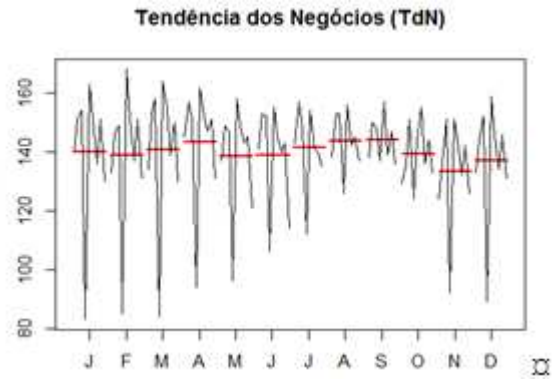
Como foram encontradas raízes unitárias em grande parte das séries, tomou-se a primeira diferença e recalculou-se o teste ADF, não tendo sido, então, encontrada raiz unitária. Para a série NDG e NDI, ambas com tendência determinística, o teste ADF foi feito sobre os resíduos obtidos a partir da série original subtraída da tendência encontrada. Para tanto foi usado um modelo de regressão linear simples cuja variável explicativa foi o tempo. Note, também, que a estatística de teste τ e o valor crítico para a série TdN são muito próximos, podendo não ser adequada a conclusão de que a série possua raiz unitária.

Em uma análise mais profunda do teste ADF, foi visto que nem todos os resíduos dos modelos ajustados possuem variância constante. Nesse caso o teste pode fornecer conclusões inadequadas. Dessa forma, aplicou-se também o teste de Phillips-Perron (Phillips & Perron, 1988). A conclusão do teste foi diferente para as séries NDG e NDI, nas quais foram encontradas raízes unitárias em vez de tendência determinística, como no teste ADF; Também houve diferença de conclusão para a série NdE, na qual o teste de Phillips-Perron encontrou raiz unitária, contrariamente ao teste ADF.

Ainda buscando descrever as séries temporais, na figura 3, pode-se analisar o nível de cada série histórica ao longo dos meses do ano. Este gráfico fornece indicativos de sazonalidade sobre as séries temporais. O Nível de Estoque (NdE), por exemplo, apresenta, relativamente, a mesma média (linha horizontal) em todos os meses, com os níveis de janeiro e dezembro apresentando maior variação. De forma oposta, a Produção Prevista (PrP) possui magnitudes de variações semelhantes em todos os meses, porém os meses de abril a setembro possuem médias maiores do que os meses restantes.

Figura 3 - Séries históricas da Sondagem da Indústria por mês – out/2005 a nov/2014.





Fonte: autores com base nos dados da (FGV/IBRE, 2014).

3.2 Ajuste Sazonal

Para efeito de comparações, serão executados os métodos direto e indireto. Este exemplo inicia-se pelo método direto, no qual o ICI sem ajuste sazonal foi obtido agregando-se as séries temporais que o compõem por meio do método de padronização, isto é, as séries componentes do ICI são padronizadas com média 0 e variância 1, em seguida é feita uma média aritmética simples entre elas e, por último, a série resultante é padronizada para ter média 100 e desvio-padrão 10. Posteriormente, realiza-se o método indireto, ou seja, após a finalização do ajuste sazonal em cada uma das séries componentes do ICI, estas são agregadas também pelo método de padronização. Para mais informações sobre o método, ver European Comission (2002).

3.2.1 Método Direto

O primeiro resultado que é sugerido observar é o teste de sazonalidade. Caso a conclusão do teste seja de não sazonalidade na série original, não é necessária a avaliação das outras saídas do programa e a série temporal permanecerá em sua forma observada. Na tabela 3, a seguir, o teste fornecido pelo X-13 pela estatística QS indica que foi encontrada sazonalidade na série original. Nota-se também que não foi encontrada sazonalidade na série ajustada, assim como nas séries dos resíduos do modelo e na componente irregular.

Tabela 3 - Diagnóstico QS para o ajuste sazonal de ICI

		QS	P-valor
Série Completa	*qsori*	18,2513	0,00011
	qsorievadj	17,8930	0,00013
	qsrsd	0	1
	qssadj	0	1
	qssadjevadj	0	1
	qsirr	0	1
	qsirrevadj	0	1
Últimos 8 anos	*qssori*	28,0794	0
	qssorievadj	36,6793	0
	qssrsd	0,0009	0,99951
	qsssadj	0	1
	qsssadjevadj	0	1
	qssirr	0	1
	qssirrevadj	0	1

Fonte: autores.

Na fase de pré-ajuste, foi modelado um ARIMA(0 1 2)(0 1 1)$_{12}$[7]. Três *outliers* foram detectados pelo programa. Todos são explicados pela crise econômica nos meses finais de 2008 e têm significância estatística (tabela 4). De acordo com os testes de Ljung-Box e de Shapiro Wilk, os resíduos não são autocorrelacionados e seguem distribuição normal.

[7] Para o ajuste sazonal utilizou-se o programa NMEC_AS (Programa de ajuste sazonal desenvolvido pela FGV\IBRE\NMEC que utiliza o X-13ARIMA-SEATS implementado no *software* R).

Tabela 4 - Pré-ajuste X-13ARIMA-SEAS - ICI

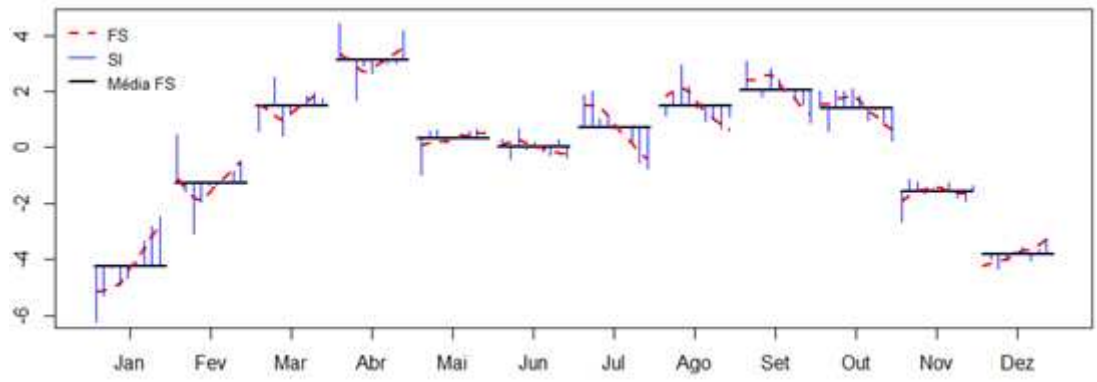
Coeficientes	Estimativas	Desvio-padrão	P-valor
LS2008.Oct	-9,6185	1,8699	$2,69e^{-07}$
LS2008.Nov	-16,5798	1,7812	$< 2e^{-16}$
LS2008.Dec	-7,5053	1,8565	$5,28e^{-05}$
MA Não Sazonal-1	-0,2315	0,0934	0,0132
MA Não Sazonal-2	-0,3656	0,0928	$8,16e^{-05}$
MA-Sazonal-1	0,6810	0,0809	$< 2e^{-16}$

Fonte: Autores

Ainda sobre o diagnóstico, a figura 4 mostra o comportamento dos fatores sazonais ao longo dos meses. Percebe-se que a média dos fatores sazonais (FS) possui níveis diferentes em cada mês, mas os fatores sazonais aparentam um comportamento estável, contribuindo para uma previsão menos errática. O SI (agregação dos fatores sazonais e da componente irregular) tende a acompanhar os fatores sazonais, indicando estabilidade da componente irregular.

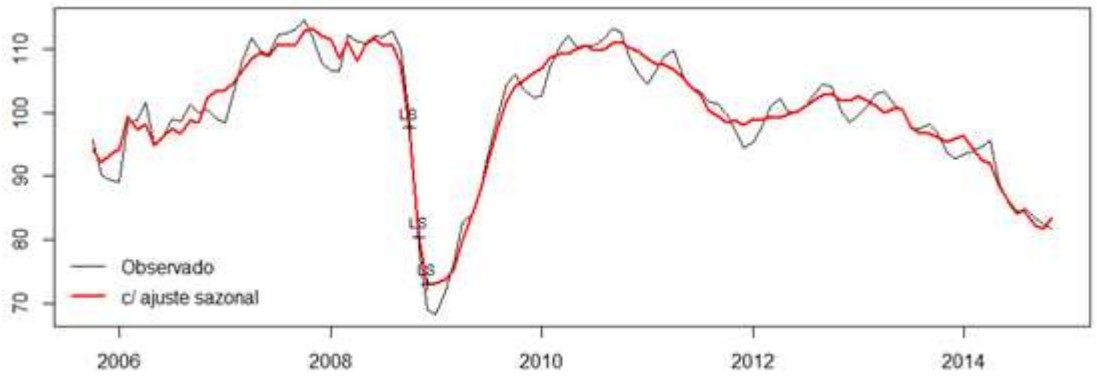
Dessa forma, considera-se o ajuste sazonal do ICI adequado. O gráfico do ICI com ajuste sazonal pode ser visto na figura 5, a seguir.

Figura 4 - Fatores Sazonais (FS) para o ICI out/2005 a nov/2014



Fonte: autores.

Figura 5 - Índice de Confiança da Indústria (ICI) com ajuste sazonal out/2005 a nov/2014 - método direto



Fonte: autores.

3.2.2 Método Indireto

A seguir, considera-se o ajuste[8] individual para cada uma das séries que compõem o ICI. O X-13ARIMA-SEATS não considerou que as séries NDE, NdE, SAN e TdN deveriam ser ajustadas por meio da análise do diagnóstico QS ao nível de significância de 5%, e assim a análise dos ajustes individuais será aplicada apenas às séries NDI, PrP e EmP, a começar pelo Nível de Demanda Interna.

a) Nível de Demanda Interna (NDI):

Na tabela 5, a seguir, tem-se o teste de sazonalidade para série NDI que indica sazonalidade na série original, mas não indica na série ajustada, nas séries dos resíduos do modelo e na componente irregular.

[8] Em nenhuma das séries foram inseridos efeitos de calendário devido à falta de justificativa teórica e/ou estatística.

Tabela 5 - Diagnóstico QS para o ajuste sazonal de NDI

		qs	p-val
Série Completa	*qsori*	18,88064	0,00008
	qsorievadj	40,14787	0
	qsrsd	0	1
	qssadj	0	1
	qssadjevadj	0	1
	qsirr	0	1
	qsirrevadj	0	1
Últimos 8 anos	*qssori*	17,29091	0,00018
	qssorievadj	34,70177	0
	qssrsd	0	1
	qsssadj	0	1
	qsssadjevadj	0	1
	qssirr	0	1
	qssirrevadj	0	1

Fonte: autores.

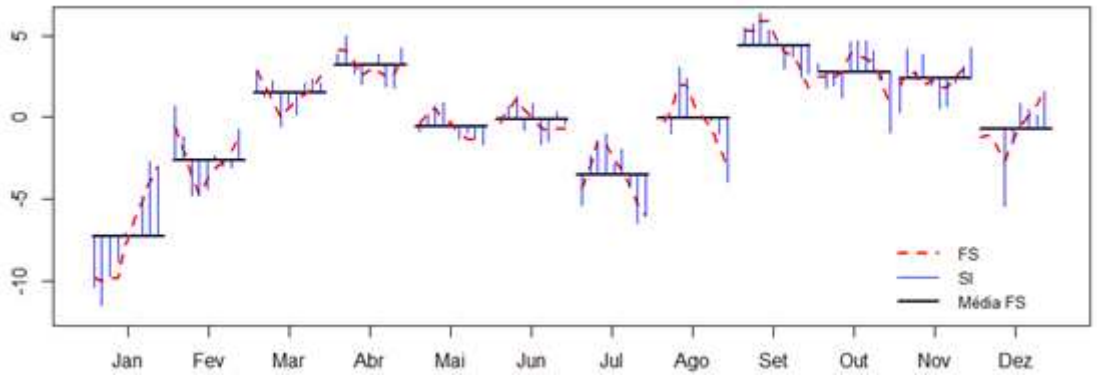
Na fase do pré-ajuste, foi modelado um ARIMA(0 1 2)(1 0 1)$_{12}$. Este segue as especificações vistas na tabela 6. Dois *outliers* foram detectados e mantidos, por melhorar a qualidade do ajuste. Com exceção do primeiro MA não sazonal, os parâmetros restantes do modelo são significativos, os resíduos não são autocorrelacionados de acordo com o teste de Ljung-Box e seguem distribuição normal (teste de Shapiro-Wilk). A seguir, na figura 6, tem-se o gráfico dos fatores sazonais (FS).

Tabela 6 - Pré-ajuste X-13 ARIMA-SEATS - NDI

Coeficientes	Estimativas	Desvio-padrão	P-valor
LS2006.May	-15,3835	3,4807	$9,88e^{-06}$
LS2008.Nov	-30,1277	3,2657	$< 2e^{-16}$
AR-Sazonal-1	0,9468	0,0263	$< 2e^{-16}$
MA Não Sazonal-1	-0,0504	0,0750	0,502
MA Não Sazonal-2	-0,4391	0,0757	$6,58e^{-09}$
MA Sazonal-1	0,6502	0,0816	$1,65e^{-15}$

Fonte: Autores.

Figura 6 - Fatores Sazonais (FS) para NDI – out/2005 a nov/2014

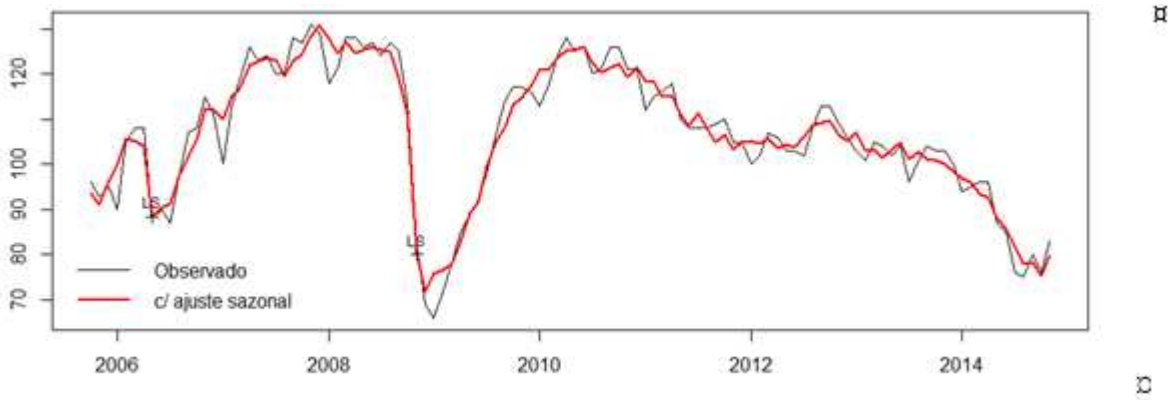


Fonte: autores

Pela figura 6, pode-se ver que os fatores sazonais acompanham bem os SI *ratios*, revelando indícios de não sazonalidade na componente irregular. Essa afirmativa é justificada pelo diagnóstico QS (tabela 5). Note também a mudança expressiva nos fatores sazonais em cada mês.

Com tais diagnósticos, considera-se esse ajuste sazonal para a série de Nível de Demanda Interna (NDI) adequado e o resultado gráfico pode ser visto na figura 7.

Figura 7- Nível de Demanda Interna (NDI) com ajuste sazonal.



Fonte: autores

b) Produção Prevista (PrP):

Na tabela 7, a seguir, tem-se o teste de sazonalidade para série PrP que indica sazonalidade na série original, mas não indica na série ajustada, nas séries dos resíduos do modelo e na componente irregular.

Tabela 7 - Diagnóstico QS para o ajuste sazonal de PrP

		QS	P-valor
Série Completa	*qsori*	78,6966	0
	qsorievadj	77,0188	0
	qsrsd	0,3960	0,8203
	qssadj	0	1
	qssadjevadj	0	1
	qsirr	0	1
	qsirrevadj	0	1
Últimos 8 anos	*qssori*	70,9508	0
	qssorievadj	72,5773	0
	qssrsd	0,2536	0,8809
	qsssadj	0	1
	qsssadjevadj	0	1
	qssirr	0	1
	qssirrevadj	0	1

Fonte: autores.

Na fase do pré-ajuste, foi modelado um ARIMA(0 1 0)(0 1 1)$_{12}$. Este segue as especificações vistas na tabela 8. Dois *outliers* significativos foram detectados pelo programa, e referem-se ao período da crise econômica de 2008. Todos os parâmetros do modelo são significativos, os resíduos não são autocorrelacionados de acordo com o teste de Ljung-Box e seguem distribuição normal (teste de Shapiro-Wilk).

Tabela 8 - Pré-ajuste X-13ARIMA-SEATS – PrP

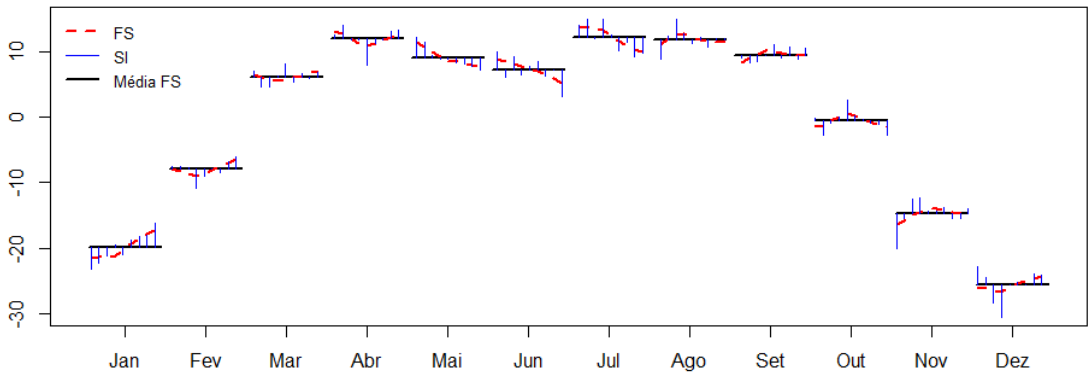
Coeficientes	Estimativas	Desvio-padrão	P-valor
LS2008.Oct	-16,0724	4,1023	$8{,}93e^{-05}$
LS2008.Nov	-17,0318	4,0627	$9{,}64e^{-07}$
MA-Sazonal-1	-0,3525	0,0782	$< 2e^{-16}$

Fonte: Autores

Note que, apesar de não haver evidências de raíz unitária na série temporal PrP (tabela 2), o modelo SARIMA estimado pelo programa faz uso da diferenciação. As diferenciações foram mantidas por tornar o modelo adequado.

O gráfico dos fatores sazonais (FS) e o gráfico de PrP com ajuste sazonal podem ser vistos, respectivamente, nas figuras 8 e 9, a seguir.

Figura 8 - Fatores Sazonais (FS) para PrP – out/2005 a nov/2014.



Fonte: Autores.

Figura 9: Produção Prevista (PrP) com ajuste sazonal out/2005 a nov/2014.

Fonte: Autores.

c) Emprego Previsto (EmP):

Na tabela 9, a seguir, tem-se o teste de sazonalidade para série EmP que indica sazonalidade na série original, mas não indica na série ajustada, nas séries dos resíduos do modelo e na componente irregular.

Tabela 9 - Diagnóstico QS para o ajuste sazonal de EmP

		qs	p-val
Série Completa	*qsori*	23,0404	0
	qsorievadj	23,0404	0
	qsrsd	0	1
	qssadj	0	1
	qssadjevadj	0	1
	qsirr	0	1
	qsirrevadj	0	1
Últimos 8 anos	*qssori*	26,1356	0
	qssorievadj	26,1356	0
	qssrsd	0	1
	qsssadj	0	1
	qsssadjevadj	0	1
	qssirr	0	1
	qssirrevadj	0	1

Fonte: autores.

Na fase do pré-ajuste, foi modelado um ARIMA(0 1 2)(1 1 0)$_{12}$. Este segue as especificações vistas na tabela 10. Nenhum *outlier* foi detectado. Como, ao final, o ajuste revelou-se de boa qualidade, não houve necessidade de inserir algum outro com referência à crise de 2008.

Tabela 10 - Pré-ajuste X-13ARIMA-SEATS - EmP

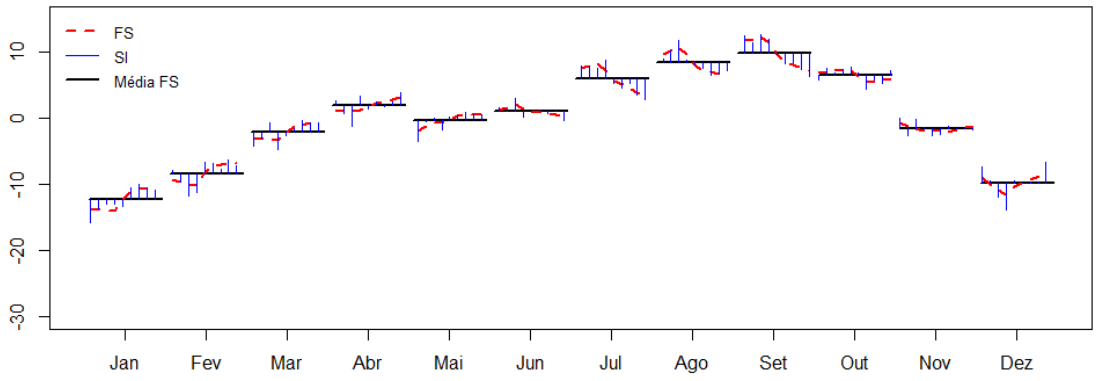
Coeficientes	Estimativas	Desvio-padrão	P-valor
AR Sazonal-1	-0,5553	0,0867	$1,49e^{-10}$
MA Não Sazonal-1	-0,0540	0,0902	0,5492
MA Não Sazonal-2	-0,3676	0,0970	0,0001

Fonte: autores.

Com exceção do primeiro AR não sazonal, todos os parâmetros do modelo são significativos, os resíduos não são autocorrelacionados de acordo com o teste de Ljung-Box e seguem distribuição normal (teste de Shapiro-Wilk). O gráfico dos fatores sazonais

(FS) é visto na figura 10, em que se pode perceber uma diferença considerável entre os fatores sazonais de cada mês.

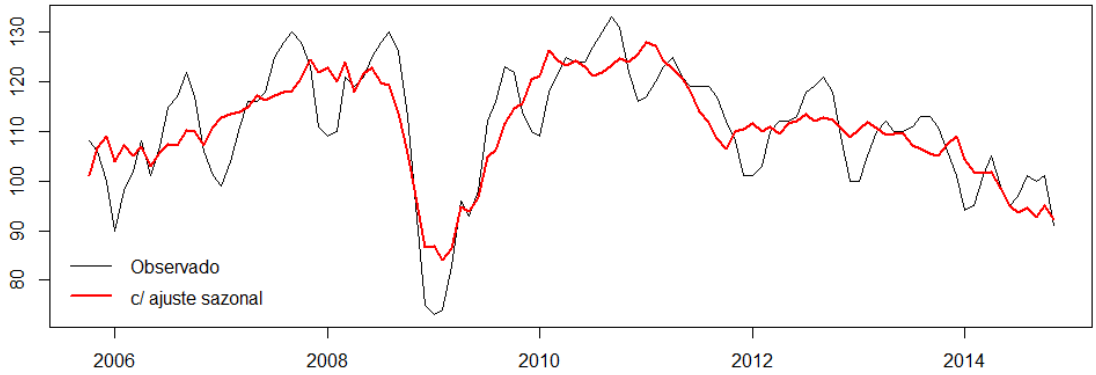
Figura 10 - Fatores Sazonais (FS) para EmP out/2005 a nov/2014.



Fonte: autores.

A partir desses diagnósticos, considera-se o ajuste sazonal para a série de Emprego Previsto (EmP) adequado e o resultado gráfico pode ser visto na figura 11.

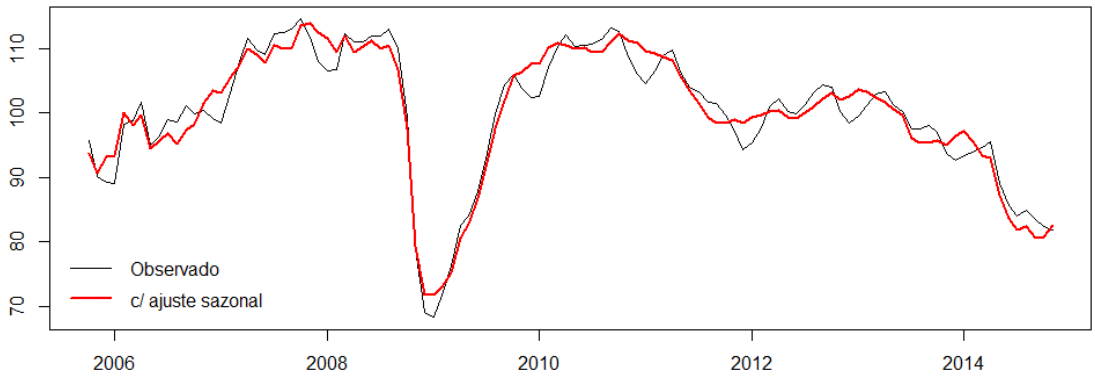
Figura 11 - Emprego Previsto (EmP) com ajuste sazonal out/2005 a nov/2014.



Fonte: autores.

Feito o ajuste nas séries que possuíam sazonalidade e agregando-as, obtém-se o Índice de Confiança da Indústria (ICI) dessazonalizado pelo método indireto. Foram cooptadas as séries com e sem ajuste sazonal. O gráfico com a série resultante é visto na figura 12.

Figura 12 - Índice de Confiança da Indústria (ICI) com ajuste sazonal out/2005 a nov/2014 – método indireto.

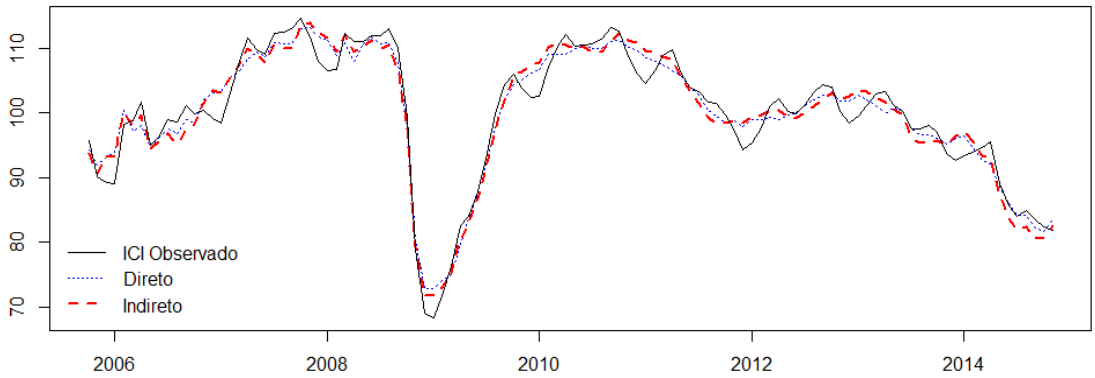


Fonte: autores.

Com a série do ICI ajustada sazonalmente, é possível concluir, por exemplo, que, no início de 2014, houve um leve decrescimento da expectativa enquanto que a série original indicava um crescimento.

A comparação entre as séries com ajuste sazonal pelos métodos direto e indireto encontra-se na figura 13. Nota-se a diferença sútil entre os métodos, havendo mudança de interpretação entre alguns meses em que o método direto indica crescimento e o indireto decrescimento, contudo em pequena escala.

Figura 13 - Comparação entre Índice de Confiança da Indústria (ICI) com ajuste sazonal direto e indireto – out/2005 a nov/2014.



Fonte: autores.

3.3 Análise de aderência fora da amostra para estabilidade dos fatores sazonais

Analisar a aderência fora da amostra permite verificar quais dos dois métodos propostos neste artigo fazem melhores previsões sobre os fatores sazonais estimados. Em resumo, são propostas as seguintes etapas:

i. ajuste sazonal da série temporal com dados até dezembro de 2010;
ii. previsão dos fatores sazonais para os doze meses seguintes;
iii. subtrair os fatores sazonais da série temporal observada para encontrar sua previsão com ajuste sazonal;
iv. armazenar os resultados e repetir as etapas *i*, *ii* e *iii* para os anos de 2011, 2012 e 2013.

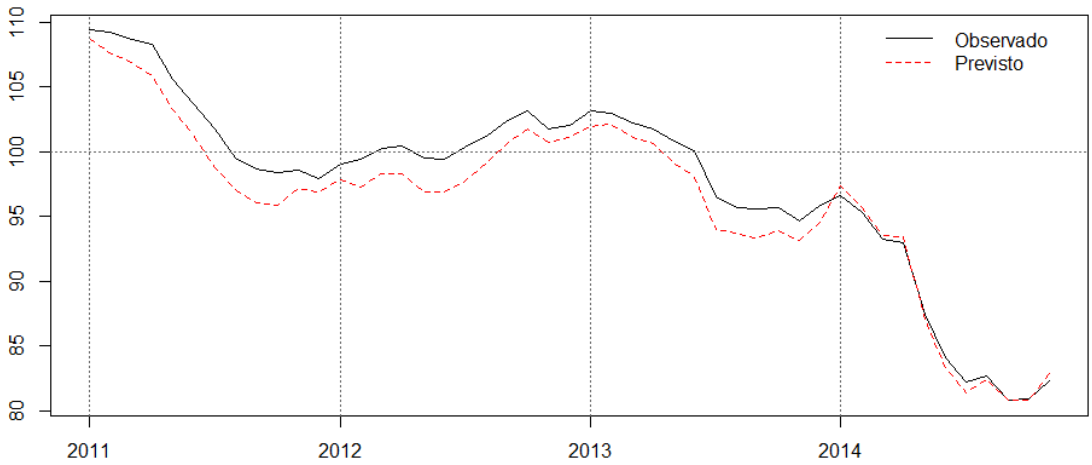
Ao concluir a execução das quatro etapas anteriores, têm-se quatro blocos de previsões para os anos 2011, 2012, 2013 e 2014. Essas previsões serão comparadas para os valores da série temporal ajustada com todos os dados disponíveis.

As figuras 14 e 15, a seguir, mostram o ICI com ajuste sazonal e as previsões dos quatro anos para os métodos direto e indireto, respectivamente. Quanto ao método direto, percebe-se que a série prevista para o ano de 2011 é relativamente diferente da série ajustada quando se utiliza todos os dados até novembro de 2014 (doravante série revisada). Há leves mudanças de sinal nos finais de 2011 e 2012, quando uma série indica crescimento e a outra não. Para o método indireto, também se notam algumas mudanças de sinal, mas no geral a trajetória das duas séries é a mesma. No entanto, os valores da série prevista estão em um nível inferior ao da série revisada. Importante ressaltar que nos dois métodos, a série prevista chegou a indicar otimismo (valores superiores a 100) enquanto a série revisada indicava pessimismo (valores inferiores a 100) e vice-versa, como se pode observar em meados de 2011 e 2012.

Figura 14 - Índice de Confiança da Indústria (ICI) com ajuste sazonal direto e previsões anuais – jan/2011 a nov/2014.

Fonte: autores.

Figura 15 - Índice de Confiança da Indústria (ICI) com ajuste sazonal indireto e previsões anuais – jan/2011 a nov/2014.



Fonte: autores.

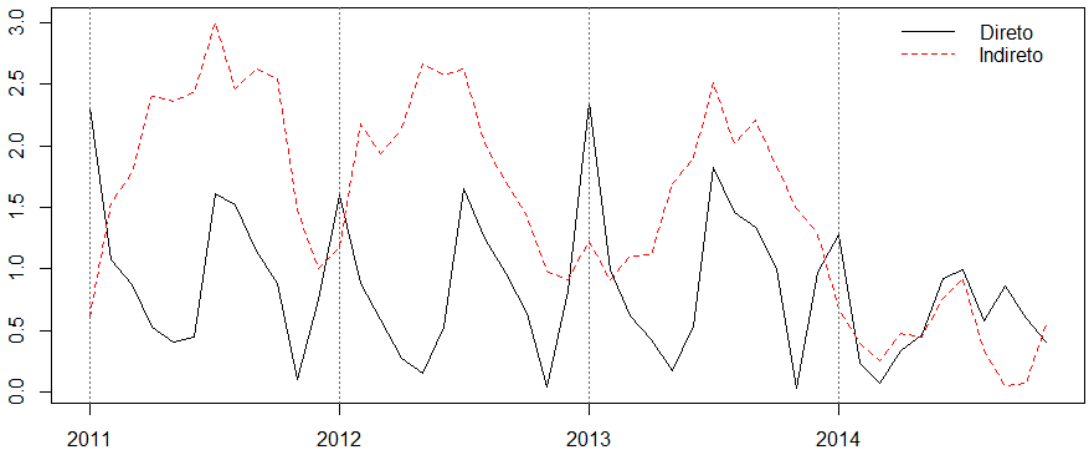
Para quantificar a diferença entre as duas séries obtidas, calculou-se o desvio absoluto para cada mês e o MAPE para cada ano e para os quatro anos juntos. A figura 16, a seguir, retrata que a diferença entre a série ajustada com todas as observações disponíveis e a prevista tende a ser maior para as observações mais antigas tanto no método direto quanto no indireto, sendo maior ainda para o indireto. Tem-se a mesma conclusão ao se observar os MAPE's na tabela 12.

Tabela 12 - MAPE entre as séries previstas e revisadas do Índice de Confiança da Indústria (ICI) para os métodos direto e indireto

Anos	Direto	Indireto
2011	0,93%	1,96%
2012	0,78%	1,85%
2013	0,98%	1,64%
2014	0,70%	0,51%
2011-2014	0,85%	1,51%

Fonte: autores.

Figura 16 - Desvio absoluto entre a previsão e a série revisada do Índice de Confiança da Indústria (ICI) pelos métodos direto e indireto jan/2011 a nov/2014.



Fonte: autores.

4. Considerações Finais

A partir dos resultados vistos nas duas seções anteriores, pode-se concluir, para as séries de dados da Sondagem da Indústria do IBRE/FGV, que não há grandes diferenças visuais entre os métodos direto e indireto. No entanto, em relação à estabilidade dos fatores sazonais, percebe-se que o método direto possui fatores mais estáveis do que o indireto, com um MAPE em torno de 1,51% no pior caso e 0,85% no melhor, embora os dois valores sejam razoáveis do ponto de vista estatístico. Notaram-se também poucas diferenças de sinal entre a série prevista e a série revisada para os dois métodos. Entre os anos de 2011 e 2012, o ICI revela duas interpretações diferentes: otimismo (acima de 100) e pessimismo (abaixo de 100) para as duas formas de revisão (figuras 14 e 15), embora não exista grande diferença em valor absoluto. Dessa forma, conclui-se que ambos os métodos oferecem resultados similares e são razoáveis para o ajuste sazonal das séries da Sondagem da Indústria, destacando-se o método direto por fornecer estimativas mais estáveis para os fatores sazonais.

Vale ressaltar, também, que as variáveis de calendário não foram utilizadas no modelo de pré-ajuste, mas sabe-se que estas podem ter justificativa econômica e influenciar, positivamente, na qualidade do ajuste. Como continuidade do trabalho, pretende-se acrescentá-las aos modelos e verificar se aprimoram a estabilidade dos fatores sazonais.

Referências bibliográficas

Aguirre, A., & Aguirre, L. A. (1999). Modelos estatísticos e econométricos para estudo da sazonalidade de preços: o caso do preço da carne de boi. *Nova Economia*, 9.

BLS. (2014). Acesso em 2014, disponível em Bureau of Labor Statistics: http://www.bls.gov/

Canada, S. (2014). *Statistics Canada Quality Guidelines*. Fonte: Statistics Canada: http://www.statcan.gc.ca/pub/12-539-x/2009001/seasonal-saisonnal-eng.htm

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedures Based on Loess. *Journal of Official Statistics*.

Dagum, E. B. (1980). The X11-ARIMA Seasonal Adjustment Method. Statistics Canada, Ottawa.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, 74.

European Comission. (2002). The Joint Harmonised EU Programme of Business and Consumer Surveys. *Economic and Financial Affairs*.

European Commission Grant. (2007). *Seasonal Adjustment Methods and Pratices*.

Eurostat. (2009). ESS Guidelines on *Seasonal Adjustment*. Methodologies and working papers, Eurostat - European Comission.

Eurostat. (2014). Acesso em 2014, disponível em Eurostat: http://ec.europa.eu/eurostat

Eurostat. (2014). *Pre-treatament*. Fonte: Seasonal Adjustment: http://www.sa-elearning.eu/pre-treatment-0

Eurostat. (2014). *Seasonal Adjustment*. Fonte: E-learning Courses: http://www.sa-elearning.eu/

Ferreira, P., Gondin Jr, J., & Mattos, D. M. (2015). Portal IBRE/FGV. Acesso em 2015, disponível em http://portalibre.fgv.br/main.jsp?lumPageId=402880811D8E34B9011D9CCBFDD1784C&contentId=8A7C82C54ADE6252014B4A982E0662F6

FGV/IBRE. (2014). *Sondagens e Índices de Confiança*. Fonte: Portal IBRE: http://portalibre.fgv.br/

Findley, & Hood. (1998). New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program. *Journal of Business and Economic Statistics*, 16.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., & Chen, B. C. (1998). New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program (with discussion). *Journal of Business and Economic Statistics*, 12.

Fok, D., Franses, P. H., & Paap, R. (2005). Performance of Seasonal Adjustment Procedures: Simulation and Empirical Results. *Econometric Institute Report*.

Gómez, V., & Maravall, A. (1992). Time Series Regression with ARIMA Noise and Missing Observations - Program TRAMO. *EUI Working Paper ECO*.

Gómez, V., & Maravall, A. (1994). Estimation, Prediction and Interpolation for Nonstationary Series with Kalman Filter. *Journal of the American Statistical Association,* 89.

Gómez, V., & Maravall, A. (1996). *Programs* TRAMO (Time series Regression with Arima noise, Missingo observations, and Outliers) and SEATS (Signal Extraction in Arima Time Series. Instructions for the User. *Working Paper*, Banco de España.

Gómez, V., & Maravall, A. (1997). *Programs TRAMO and SEATS; Instructions for the User. Working*.

Gómez, V., & Maravall, A. (2001). Automatic Modelling Methods for Univariate Series. *A Course in Time Series Analysis: J. Wiley and Sons*.

Gómez, V., Maravall, A., & Peña, D. (1999). Missing Observations in ARIMA Models. *Journal of Econometrics*, 88.

Harvey, A., & Shephard, N. (1993). Structural Time Series Models. In: *Handbook of Statistics* (Vol. 11). Elsevier Science Publishers B.V.

IBGE. (Novembro de 2010). *Nota Técnica - Aperfeiçoamento do Ajuste Sazonal*. IBGE.

IBGE. (2014). Fonte: Instituto Brasileiro de Geografia e Estatística: http://www.ibge.gov.br/home/

Koopman, S. J., Harvey, A., Doornik, J., & Shephard, N. (2009). Structuural Time Series Analyser, Modeler, and Predictor. *Timberlake Consultants*.

Livsey, J., Pang, O., & McElroy, T. (2014). Effect of Trading Day Regressors. *Research Report Series*.

Ljung, G. M., & Box, G. E. (1978). On a Measure of a Lack of Fit in Time Series Models. Biometrika.
ONS. (2007). *Guide to Seasonal Adjustment with* X-12-ARIMA.

Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. Biométrika.

Plosser, C. I. (1979). *A Time Series Analysis of Seasonality in Econometric Models, Seasonal Analysis of Economic Time Series*.

Rasmussen, R. (2004). On time series data and optimal parameters. *The International Journal of Management Science*.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*.

Shiskin, Young and Musgrave. (1967). The X-11 variant of the Census Method II seasonal adjustment program. Technical Paper No. 15, U.S. Department of Commerce, U. S. Census Bureau.

U.S. Bureau of the Census. (2013). X-13ARIMA-SEATS *Reference Manual Accessible HTML Output Version*.

Zellner, A. (1979). Front matter to "*Seasonal Analysis of Economic Time Series", Seasonal Analysis of Economic Time Series*.

Abstract

This article proposes the use of X-13ARIMA-SEATS to seasonally adjust the business tendency survey time series of the Brazil Business tendency surveys (manufacturing) (BBTS-M |FGV/IBRE). The BBTS-M is an aggregated time series composed of six indicators, and the article discussed the Direct and Indirect method to figure out the best way to perform the seasonal adjustment. As a result, we observed that although the indirect method had a better performance to capture seasonal patterns in each time series that composes the main indicator, the direct method showed up more stable seasonal factors during the time.

Keywords: Business Tendency Survey, X-13ARIMA-SEATS, Time Series, R software, Brazilian Industrial Tendency Survey (FGV|IBRE).

Comparação dos métodos de Análise Agrupamentos K-means e Partitioning Around Medoids (PAM)

Marcos Garrido de Oliveira [1]
Luis Pérez Zotes [2]
Osvaldo Luiz Gonçalves Quelhas [3]
Valdecy Pereira [4]
Carlos Antônio da Silva Carvalho [5]

Resumo

Este artigo propõe a comparação entre os métodos de agrupamentos automáticos *K-means* e *Partitioning Around Medoids (PAM)* aplicados em dados simulados, sem *outliers* e com *outliers* (valores atípicos). Para essa finalidade, foi sugerido o procedimento diagrama de decisão para definição do número de grupos, avaliação da significância das variáveis em estudo e a definição dos elementos que irão compor cada grupo. Considerando que ambos os métodos de Análise de Agrupamentos são ferramentas poderosas na segmentação das observações, a escolha dessas ferramentas, geralmente, está relacionada com a origem das bases de dados, custos de pesquisas e ao conhecimento do pesquisador sobre o tema. Pelos resultados expostos neste trabalho, percebemos que as técnicas *K-means e Partitioning Around Medoids (PAM)* apresentaram *performances* semelhantes com os dados da amostra sem *outliers.* Porém, quando analisadas com os valores atípicos, a técnica *Partitioning Around Medoids (PAM)* evidencia maior robustez e consistência se comparada com a técnica *K-means.*

Palavras-chave: Análise de Agrupamentos. Partitioning *Around Medoids*. *K-means*. Centróide. Medóide.

[1] UFF – Universidade Federal Fluminense, Escola de Engenharia, Rua Passo da Pátria, 156, São Domingos, Niterói, RJ, Brasil. CEP: 24210-240. E-mail: mgoliveirarj@gmail.com
[2] UFF – Universidade Federal Fluminense, Escola de Engenharia. E-mail: lpzotes@gmail.com
[3] UFF – Universidade Federal Fluminense, Escola de Engenharia. E-mail: quelhas@latec.uff.br
[4] UFF – Universidade Federal Fluminense, Escola de Engenharia. E-mail: valdecypereria@id.uff.br
[5] UFF – Universidade Federal Fluminense, Escola de Engenharia. E-mail: cascarvalho@uol.com.br

1. Introdução

Os modelos multivariados possuem o propósito de testar ou inferir hipóteses sobre um determinado fenômeno em estudo. Contudo, a correta utilização depende do conhecimento destas técnicas, assim com, de suas limitações. Os métodos de análise de agrupamentos, que fazem parte dos modelos multivariados, têm sido utilizados para diferentes fins tais como amostragem, segmentação de perfis de consumidores, biologia, indústria, transportes, meteorologia, etc. (BRITO; SEMAAN; BRITO, 2011).

Atualmente, com a grande geração de informações (*Big Data*), torna-se necessário tratamento dos dados primários ou secundários, a serem utilizados em pesquisas, considerando a importância e a necessidade da mensuração de determinados atributos para embasar e estabelecer fundamentos teóricos que norteiem a informação.

Muitas aplicações da análise de agrupamentos, devido à grande variação de técnicas e métodos, produzem soluções diversificadas para um mesmo conjunto de dados. Nesse sentido, com o objetivo de refinar as análises, sugere-se a combinação das técnicas de agrupamentos para obtenção de resultados mais consistentes e robustos.

As técnicas *K-means* e *Partitioning Around Medoids* são geralmente utilizadas como métodos de agrupamentos, devido à simplicidade e à alta *performance* quando aplicadas em grandes conjuntos de dados. Contudo, ambas as técnicas requerem que o usuário forneça, *a priori*, o número exato de *clusters*, que muitas vezes não são evidentes.

Este artigo tem como objetivo comparar as técnicas de agrupamentos *K-means* e *Partitioning Around Medoids* (PAM) aplicadas em dados simulados, com e sem *outliers* (valores atípicos), procedentes de distribuições Normal Gaussiana, simulados aleatoriamente. Para essa avaliação, é proposto um procedimento singular, de forma objetiva, que envolve o algoritmo hierárquico na definição do número de grupos pela análise gráfica do dendrograma e coeficiente de aglomeração, e os algoritmos não-hierárquicos na avaliação da significância das variáveis (*K-means*) e definição dos elementos para compor cada grupo *(K-means* e *Partitioning Around Medoids*).

Na primeira seção deste trabalho, apresenta-se o objetivo do artigo. Na segunda, a revisão da literatura acerca da análise multivariada de dados. Na terceira, é detalhada a metodologia da pesquisa. Na quarta, são apresentados os procedimentos e os resultados obtidos pelas técnicas mencionadas. Por fim, na quinta seção, são apresentadas as conclusões do estudo.

2. Revisão da Literatura

Das análises multivariadas de dados, a Análise de Agrupamentos apresenta uma ampla variedade de técnicas e soluções descritas na literatura. Referências importantes, como Anderberg (1973), Hartigan (1975), Jain e Dubes (1988), Kaufman e Rousseeuw (1990), Arabie et al. (1999) e Everitt et al. (2001) são bases de estudos, que complementam trabalhos recentes em algoritmos de agrupamentos. Dentre eles, podemos citar Marques (2005), Hair et al. (2005), Cargnelutti Filho et al. (2008), Ferreira (1996), Faceli et al. (2011), Corrar; Paulo; Dias Filho (2011), Brito, Semaan, Brito (2011), entre outros.

O problema de Agrupamento apresenta uma complexidade de ordem exponencial, tornando-se inviável enumerar todos os possíveis grupos e escolher a melhor configuração, à medida que aumenta a quantidade de elementos separados em grupos, Linden (2004) e Semaan et al. (2012) demonstram em seu trabalho que a quantidade de combinações possíveis de grupos é obtida pelo Número *Stirling* de segundo tipo, $NS\ (n,k)$, onde n representa o número de objetos e k a quantidade de grupos (Jr, 1968). Como exemplo, podemos considerar que, se quisermos separar 120 objetos em quatro grupos, existiriam $NS(120;4) = 7{,}3618627699099E + 70 \approx 10^{71}$ possíveis agrupamentos.

A Análise de Agrupamentos é uma das técnicas multivariadas, realizada com base em uma matriz de similaridade ou dissimilaridade das variáveis, tais como a medida de *distância Euclidiana* ou a *distância de Mahalanobis*, que considera o número de grupos, por exemplo. Essa técnica multivariada interliga as amostras por suas associações através de algoritmos predeterminados (FERREIRA, 1996; HAIR et al., 2005; CORRAR; PAULO; DIAS FILHO, 2011).

Na Análise de Agrupamentos, os algoritmos mais populares usados são classificados como hierárquicos e não-hierárquicos, por não serem excludentes, podem ser empregados separadamente ou como combinação de ambos os métodos. Hair et al. (2005) afirmam que a escolha do método não é uma questão simples, em função de centenas de soluções analíticas com diferentes algoritmos disponíveis e outros ainda em desenvolvimento. Entretanto, "o critério essencial de todos os algoritmos [...] é que eles tentam maximizar as diferenças entre agrupamentos relativamente à variação dentro dos mesmos" (HAIR et al., 2005).

O agrupamento hierárquico pode ser aglomerativo ou divisivo, conforme comentado por Faceli et al. (2011), Ferreira (1996), Hair et al. (2005) e Corrar; Paulo; Dias Filho (2011). Em ambos os procedimentos, é necessário especificar o método para estruturar e quantificar a distância entre os grupos a serem produzidos. Essa operação é efetuada pelos métodos de ligações mínima, máxima, média, Ward ou centróide, entre outros, que diferem na maneira por meio do qual a distância entre os grupos é computada.

Os cinco métodos de ligações aglomerativos acima são citados por Corrar; Paulo; Dias Filho (2011), como os mais populares usados nas formulações de agrupamentos. Esses métodos têm sido referenciados em estudos, também com nominações equivalentes, que são apresentados em *softwares* de soluções analíticas, descritas na tabela 1.

Tabela 1 - Nominações equivalentes dos métodos de ligação hierárquica

Ligação	Estudos	Software
Mínima	Single Linkage	Nearest neighbor
Máxima	Complete Linkage	Furthest neighbor
Média	Average Linkage	Between Groups
Ward	Ward	Ward
Centróide	Centroid	Centroid

Elaborado pelo Autor, 2014.

As abordagens, vantagens e limitações da utilização dos métodos de ligação descritos a seguir têm como referência os procedimentos apontados por Corrar; Paulo; Dias Filho (2011).

- Ligação mínima (*Single Linkage*) – é dada pela distância mínima entre os objetos dos dois grupos mais próximos. Uma limitação apresentada nesse método ocorre quando a composição dos objetos é pobremente estruturada, o que pode formar longas cadeias e, ocasionalmente, todos os indivíduos serem alocados nelas.
- Ligação máxima (*Complete Linkage*) – também conhecida como método do diâmetro, é similar a anterior, exceto pelo critério de agrupamento ter como base a distância máxima entre os objetos de dois grupos mais próximos. Esta técnica elimina o problema de encadeamento ou corrente prolongada, identificada na ligação mínima, o que é um diferencial para sua escolha.
- Ligação média (*Average Linkage*) – semelhante aos métodos anteriores, porém, o critério de agrupamento considera a distância de todos os elementos de um grupo em relação a todos de outro. Uma vantagem dessa técnica está na divisão que é baseada em todos os objetos, e não apenas em um par de dados extremos ou mais próximos, além de gerar grupos com menores variações internas.
- Método de *Ward* – baseia-se na perda de informações decorrente do agrupamento de objetos, que considera a soma do quadrado dos desvios de cada objeto em relação à média do conglomerado no qual foi alocado. Esse procedimento resulta em grupos com, aproximadamente, o mesmo número de observações.
- Método Centróide – este método utiliza a distância euclidiana ou euclidiana quadrada entre centróides dos grupos (valores médios das observações sobre as variáveis), que migram quando as fusões dos grupos ocorrem. A vantagem da utilização desse método é que o processamento dele não sofre impacto de valores atípicos (*outliers*).

É importante citar que a implementação da Análise de Agrupamentos em qualquer matriz de distância sempre irá gerar grupos. Para avaliar se o resultado produzido pelo algoritmo de agrupamento é representativo e se expressa uma estrutura do conjunto de dados, pode ser utilizada a correlação de *Pearson* entre a matriz de distância original dos dados e a matriz de distância cofenética do resultado do algoritmo hierárquico, denominado coeficiente de correlação cofenética, explicitada por Cargnelutti Filho et al. (2008).

Corrar; Paulo; Dias Filho (2011), argumentam que um "dendrograma" oriundo de um procedimento hierárquico é um resumo apropriado dos dados, validando o agrupamento gerado se o coeficiente de correlação cofenética for alto, próximo de 0,8. Nesse caso, identifica a existência de uma estrutura que diferencia os grupos da amostra. Caso contrário, o resultado deve ser visto como a descrição da saída do algoritmo de agrupamento. A distância cofenética entre duas observações agrupadas é definida como a dissemelhança entre os grupos em que as duas observações são primeiro combinadas em um único agrupamento (CARGNELUTTI FILHO et al., 2008).

As análises elementares e exploratórias de dados com os métodos gráficos auxiliam em geral, o entendimento da natureza da análise multivariada. As discussões de algumas técnicas gráficas adicionais, para agrupar objetos (itens ou variáveis), em conjunto com a utilização de algoritmos permitem encontrar, nos dados, uma composição natural de agrupamento, consideradas importantes técnicas exploratórias de dados (FERREIRA, 1996).

O procedimento de agrupamento não-hierárquico procura a partição de n objetos em k grupos especificados, ou seja, ele atribui objetos a cada grupo, uma vez que o número de grupos tenha sido definido pelo pesquisador. O método exige a determinação de critérios que produzam medidas sobre a qualidade da partição produzida (CORRAR; PAULO; DIAS FILHO, 2011). Nesse procedimento, duas técnicas são bem difundidas, a *K-means e Partitioning Around Medoids* (PAM) para selecionar os grupos.

A designação da técnica *K-means* deve-se ao fato do algoritmo representar cada um dos k *clusters* pela média (ou média ponderada) dos seus pontos, definido como centro de cada agrupamento. O número de *clusters* é definido inicialmente, mantendo-se o mesmo ao longo de todo o processo. Marques (2005) descreve que esse algoritmo faz uma partição inicial aleatória, atribuindo o elemento a um determinado agrupamento, considerando a distância de similaridade entre o mesmo e o centro de cada grupo. O processo inicia-se na procura pelos centróides, cujas coordenadas de cada um referem-se às coordenadas dos objetos no grupo e atribui cada objeto ao centróide mais próximo. Em seu trabalho, Ferreira (1996) apresenta o algoritmo da *K-means* de uma forma simplificada, dividindo em três passos:

- Particionar os itens em k grupos iniciais arbitrariamente;
- Percorrer os itens e calcular as distâncias de cada um deles em relação ao centróide (médias) dos grupos. Após esse processo, efetuar a realocação do item para o grupo em que ele apresentar mínima distância. Caso não seja o grupo ao qual este pertença, recalcular os centróides dos grupos que ganharam e perderam o item;
- Repetir o passo 2 até que nenhuma alteração seja feita.

Já a técnica *Partitioning Around Medoids* (PAM) utiliza os medóides (tipicamente o elemento mais central do grupo), em vez de centróide representativo da partição. Ao encontrar um conjunto de k medóides, k grupos *são* estabelecidos, associando cada observação ao medóide mais próximo. A finalidade é encontrar n objetos que minimizem a soma das dissimilaridades das observações do objeto mais próximo que os representa (MARQUES, 2005).

Conforme descrito por Brito; Semaan; Brito (2011) para esse procedimento, primeiro ocorre a seleção aleatória de k objetos, considerando os medóides iniciais em cada um dos grupos. Após definidos os medóides iniciais, cada um dos $(n-k)$ objetos restantes é alocado ao grupo cujo medóide está mais próximo. Essa estratégia efetua vários experimentos de troca de medóides por não medóides e reavalia a qualidade dos novos agrupamentos resultantes, até que se obtenha o menor valor da função objetivo (Equação 1), da soma das dissimilaridades.

$$fo = mininizar \sum_{i=1}^{k} \sum_{\forall O_j \in med_i} d_{ij} \qquad (1)$$

Sendo: $O_j \in$ ao conjunto $\boldsymbol{X}$ (formado pelos n objetos), k quantidade de medóides e $M = \{med_1, med_2, \ldots, med_k\}$ objetos medóides que definem os grupos. Essa técnica proposta por Kaufman e Rousseeuw (1990) é considerada mais poderosa e robusta se comparada à *K-means*.

Como observado por Park et al. (2009) Kumar et al. (2009), comparações entre técnicas de análise de agrupamentos são, por vezes, propostas com o intuito de aperfeiçoar as escolhas de modelos, que produzam grupos cada vez mais consistentes e homogêneos para as mais diversas finalidades: amostragem, segmentação de perfis de consumidores, entre outros. Hu et al. (2004), por exemplo, utilizam simultaneamente várias técnicas de agrupamentos e combinam seus resultados para tomar uma decisão final.

Semaan et al. (2013) propõem um método baseado em combinação de soluções que considera a técnica de uma matriz de coassociação, utilizada para identificar do número ideal de grupos em problemas de agrupamento automático. Em seu trabalho, foi utilizado o índice silhueta, que combina as características como coesão e separação.

3. Metodologia

3.1 População e Amostra da Pesquisa

Assumindo que os dados que formam a população para avaliação e desenvolvimento do estudo são derivados de uma distribuição Normal (Equação 2), também conhecida como distribuição de Gauss ou Gaussiana, com função de densidade de probabilidade $f(x)$, média μ e variância o^2, cada grupo das amostras (dois vetores com trinta observações) foi gerado considerando as médias e desvios padrão (tabela 2).

$$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\left(-\frac{x-\mu}{2\sigma^2}\right)^2}, -\infty < x < \infty, \sigma > 0 \qquad (2)$$

Tabela 2 - Vetor de médias e desvio padrão da amostra

Estimadores	Grupo 01	Grupo 02	Grupo 03	Grupo 04
Vetor de Médias	(14; 15)	(15; 6)	(10; 15)	(7; 4)
Desvio Padrão	1,414	1,414	1,414	1,000

Elaborado pelo Autor

3.1.1 Seleção Amostral

A fim de avaliar o desempenho dos métodos propostos, 120 dados foram simulados em planilhas eletrônicas pela função geração de números aleatórios em quatro (4) grupos distintos.

3.1.2 Delineamento Amostral

As observações foram delineadas como pontos no espaço p-dimensional, cujas coordenadas dadas por vetores formam a matriz de dados composta de p respostas das n observações ou unidades experimentais de ordem $n \times p$. Se considerarmos os pontos $P = (x_1, x_2)$ *e* $Q = (y_1, y_2)$ no plano cartesiano, a distância D definida por $d(P, Q)$ pode ser descrita pelo teorema de Pitágoras conforme a Equação de Distância 3:

$$d(P,Q) = D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}\ , \tag{3}$$

A distância euclidiana (ou distância métrica) é a distância entre dois pontos, provada pela aplicação recursiva do teorema de Pitágoras. Dado que o ponto P e Q possuam p coordenadas descritas como $P = (x_1, x_2, \dots, x_p)$ *e* $Q = (y_1, y_2, \dots, y_p)$ com distância D generalizada é representada pela Equação de Distância em um Espaço Métrico (4):

$$D = \sqrt[2]{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_p - y_p)^2} = \sqrt[2]{\Sigma_{i=1}^{p} (x_i - y_i)^2}\ , \tag{4}$$

A aplicação da Equação (4) como distância, no espaço euclidiano, torna-se um espaço métrico.

Com a obtenção de dados aleatórios descritos no item 3.1.1, foram geradas matrizes com $p = 2$ variáveis ou características, tomadas em cada unidade da amostra. A representação desses dados é demonstrada adiante com a notação x_{jk} **,** que indica um valor particular da *j-ésima* unidade amostral, com $j = 1, 2, \dots, n$, e da *k-ésima* variável mensurada com $k = 1, 2, \dots, p$. Essas medidas de $n = 120$ unidades amostrais e $p = 2$ variáveis podem ser representadas conforme o arranjo da tabela 3.

Tabela 3 - Representação de dados através da notação X_{jk} para indicar um valor particular da *k*-ésima variável mensurada na *j*-ésima unidade amostral.

Unidades Amostrais	Variáveis	
	1	*p*
1	X_{11}	X_{1p}
2	X_{21}	X_{2p}
.	.	.
.	.	.
.	.	.
j	X_{j1}	X_{jp}
.	.	.
.	.	.
.	.	.
n	X_{n1}	X_{np}

Elaborado pelo Autor

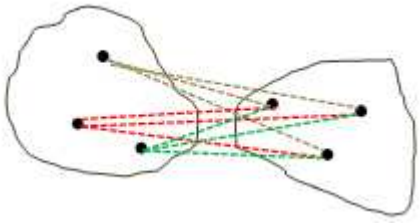
É importante destacar que a distância Euclidiana pode ser insatisfatória em muitas situações estatísticas, pois para cada coordenada é atribuído o mesmo peso no cálculo da distância. Para esse caso, Ferreira (1996) sugere o uso de uma medida de distância que considera as diferenças de variação, a presença de correlação e a magnitude dos dados, como a distância de Mahalanobis, citado por Hair et al. (2005, p.382) e ressaltado por Corrar; Paulo; Dias Filho (2011, p.339) "não apenas desenvolve um processo de padronização nos dados, utilizando a escala em termos de desvio-padrão, mas também soma a variância-covariância total do grupo, com ajustes das intercorrelações entre as variáveis".

3.2 Aplicação da Análise de Agrupamentos

A Análise de Agrupamentos é uma tarefa não supervisionada, ou seja, explora ou descreve um conjunto de dados, pois não há uma classificação predefinida para os grupos de elementos (objetos) que serão gerados dessa amostra. A validação de um resultado de agrupamento deve, de forma objetiva, determinar se a solução é representativa para o conjunto de dados analisados e, para essa finalidade, foi utilizado o coeficiente de correlação cofenética. A determinação da quantidade de grupos em cada Análise de Agrupamentos teve como base a visualização gráfica através do dendrograma, considerando todas as variáveis e o coeficiente de aglomeração do procedimento hierárquico.

O método selecionado para o procedimento hierárquico aglomerativo foi o de ligação média ou encadeamento médio (*Average Linkage*), por ter apresentado o maior coeficiente de correlação cofenética, comparado com os demais métodos delineado no item 3.1.2. Além disso, as poucas restrições e o não enviesamento dos resultados foram fatores de ponderação na escolha desse método para análise da amostra gerada.

Figura 1 - Procedimento hierárquico de agrupamento (Ligação Média).



Fonte: Adaptado de Ferreira (1996, p.298).

Como a escolha da solução final da Análise de Agrupamentos requer o conhecimento tácito sobre o tema pelo pesquisador, essa é considerada, ainda, por muitos, como uma técnica subjetiva. Mesmo que métodos mais elaborados tenham sido desenvolvidos para auxiliar na avaliação das soluções de agrupamentos, Hair et al. (2005), descrevem que "ainda cabe ao pesquisador tomar a decisão final quanto ao número de agrupamentos a ser aceito como solução final". Contudo, o coeficiente de aglomeração também é considerado por Hair et al. (2005), como uma das regras de parada.

Para avaliação da significância das variáveis foi utilizada a ANOVA *(Analysis of Variance)* proveniente do procedimento de agrupamento não-hierárquico *(K-means),* com o número de agrupamentos determinados pelo coeficiente de aglomeração do procedimento hierárquico com todas as variáveis. Por fim, para definição de quais elementos da amostra devem compor cada grupo para comparação, foram utilizadas as técnicas *K-means e Partitioning Around Medoids (PAM).*

Os dados da amostra foram trabalhados em planilha Excel (elaboração de tabelas). No *software Statistical Package for the Social Sciences – SPSS* (versão 18.0.0) foi obtida a análise de variância (ANOVA) e alocação dos elementos (objeto) em cada *cluster*, proveniente do procedimento não-hierárquico *K-means*. Pelo *software* R (versão 3.1.2), foram implementadas as correlações cofenéticas e dendrogramas dos métodos hierárquicos, além da alocação dos elementos (objeto) em cada *cluster* para descrição dos

agrupamentos originados pelo procedimento não-hierárquico *Partitioning Around Medoids* (PAM).

O *software* SPSS (IBM) é reconhecido no ambiente corporativo pelas suas diversas aplicabilidades, nas análises estatísticas e gerenciamento de dados. Já o R *(software* livre) é uma solução viável bastante utilizada no meio acadêmico que fornece uma ampla variedade de medidas estatísticas e técnicas gráficas.

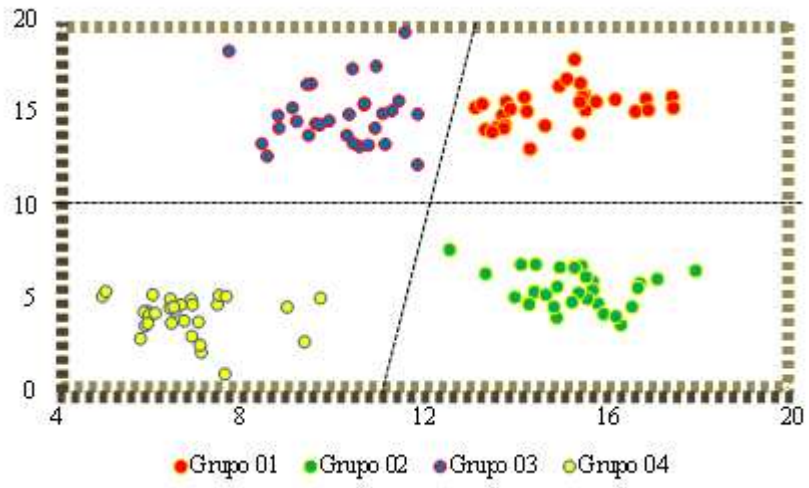
4. Resultados

Nesta seção, são apresentados, graficamente, os dados da amostra, o diagrama de decisão, a definição do método de ligação e, por fim, as composições e análises de agrupamentos com e sem *outliers.*

4.1 Representação Gráfica da Amostra e o Diagrama de decisão

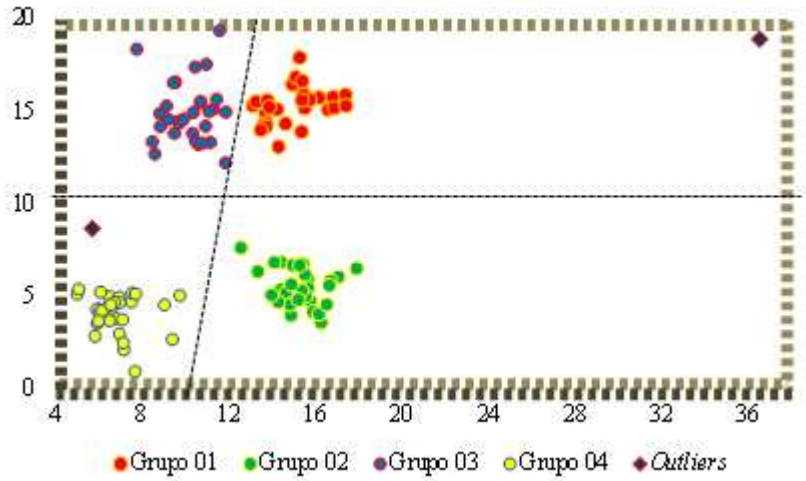
Os dados da amostra sem *outliers*, com quatro (04) grupos distintos de trinta (30) observações cada, são representados na figura 2.

Figura 2 - Gráfico de dispersão dos dados amostrais sem *outiliers*



Já a figura 3 representa a distribuição dos mesmos dados amostrais, porém, substituindo duas (2) observações ou quatro (4) pontos/coordenadas por dados destoantes *(outliers).*

Figura 3 - Gráfico de dispersão dos dados amostrais com *outliers*.

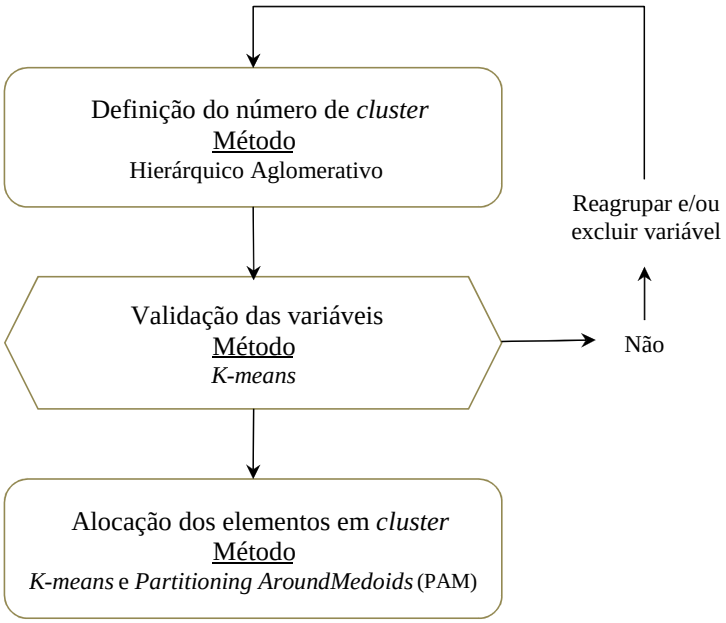


Com as duas amostras definidas, iniciamos o procedimento de comparação entre as técnicas *K-means* e PAM.

O processo de decisão em Análise de Agrupamentos segue a utilização dos procedimentos descritos na figura 04:

- Algoritmo hierárquico aglomerativo para definição do número de grupos pela análise gráfica do dendrograma e coeficiente de aglomeração;
- Algoritmo não-hierárquico *K-means* na avaliação da significância das variáveis (ANOVA); e
- Algoritmo não-hierárquico *Partitioning Around Medoids* (PAM) para definição dos elementos pertencentes a cada grupo.

Figura 4 - Procedimento do diagrama de decisão



4.2 Definição do Método de Ligação

A definição do método utilizado no procedimento Hierárquico Aglomerativo ocorreu através da análise do coeficiente de correlação cofenética. Nessa avaliação, o método de ligação média *(Average Linkage)* apresentou resultado superior aos demais. Esse método não utiliza valores extremos e a partição considera todos os elementos da amostra (tabela 4).

Tabela 4 - Correlação cofenética dos métodos hierárquicos

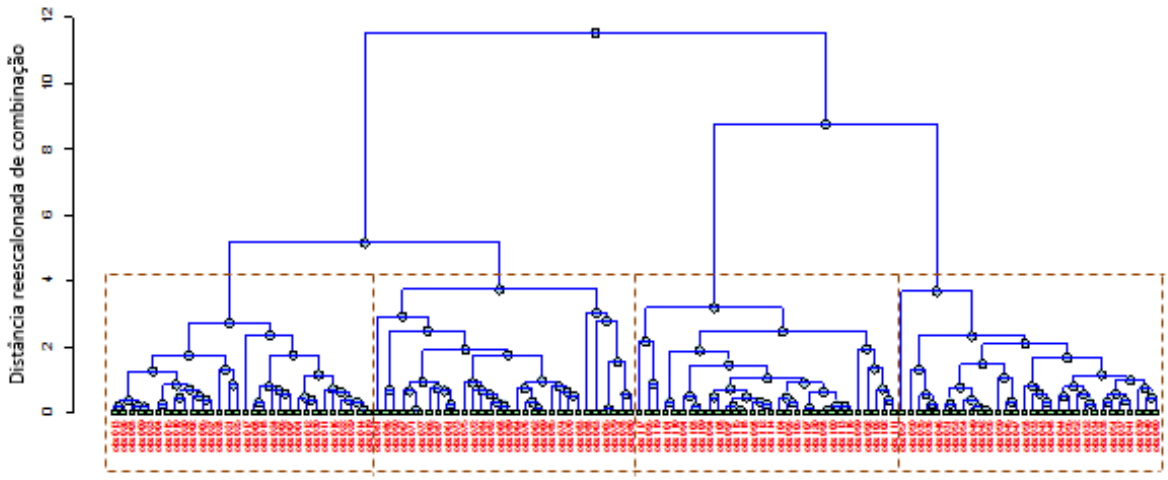
Métodos	*Avarege*	*Mcquitty*	*Complete*	*Single*	*Centroid*	*Ward*
Correlação	0,911	0,903	0,902	0,900	0,892	0,891

Fonte: Dados da amostra

4.3 Definição do Número de *Clusters* – Variáveis sem *Outliers*

Iniciamos a análise gráfica do dendrograma gerado pelo algoritmo hierárquico, com todas as variáveis. Nessa visão, o resultado gráfico (figura 5) com quatro grupos representou aglomerados mais homogêneos entre si.

Figura 5 - Dendrograma do procedimento hierárquico (método a*verage linkage*)



Fonte: Elaborado a partir dos dados da amostra

Na avaliação pela ANOVA *(Analysis of Variance),* proveniente do procedimento não-hierárquico *(K-means)* com 4 *clusters* e considerando o nível de significância $\alpha = 0{,}05$, as duas variáveis simuladas mantiveram-se como significativas conforme tabela 5. Caso alguma variável fosse considerada como não significativa, neste momento, essa seria reagrupada, considerando conceitos teóricos das variáveis ou excluída do modelo e retornaria ao procedimento do primeiro estágio do diagrama de decisão (método hierárquico), para gerar novo dendrograma e nova tabela de coeficiente de aglomeração.

Tabela 5 - Análise de variância da amostra – ANOVA

Variáveis	Agrupamento		Erro		Valor F	Nível de Significância
	Quadrático Médio	Graus de Liberdade	Quadrático Médio	Graus de Liberdade		
Var 1	494,294	3	1,312	116	376,802	0,000
Var 2	1075,542	3	1,500	116	716,866	0,000

Fonte: Elaborado a partir dos dados da amostra

Para corroborar a definição do número de grupos com as variáveis validadas pela ANOVA, foi analisado o coeficiente de aglomeração resultante do procedimento hierárquico aglomerativo (tabela 6) onde as maiores variações, sublinhadas em destaques, foram definidas como possíveis números de *clusters* (regra de parada). As duas grandes variações percentuais desses coeficientes ocorreram nas passagens de 2 para 3 agregados (69,8%) e de 3 para 4 agregados (38,2%), o que gerou dois prováveis resultados para análises com 3 ou 4 grupos. Porém, após o valor observado para 4 agregados, percebeu-se uma forte queda de variação de 4 para 5 grupos (1,3%), indicativo para o número de *clusters* a ser estabelecido. Além disso, o dendrograma (figura 5), gerado pelo algoritmo hierárquico, corrobora a quantidade de 4 *clusters*.

Tabela 6 - Coeficiente de aglomeração da análise hierárquica

Número de Agrupamentos	Coeficiente de Aglomeração	Variação
9	2,787	4,4%
8	2,910	4,0%
7	3,027	5,5%
6	3,195	15,1%
5	3,679	1,3%
4	**3,727**	**38,2%**
3	**5,150**	**69,8%**
2	8,745	31,9%
1	11,535	-

Fonte: Dados da amostra

4.4 Composição e Análise e Agrupamentos sem *Outliers*

Utilizando as variáveis sem *outliers*, validadas na Análise de Variância (tabela 5) e definido o número de *clusters*, foram aplicados os algoritmos *K-means* e *Partitioning Around Medoids* (PAM) com o perfil de quatro (4) agrupamentos (tabela 7) para comparação entre os dois algoritmos em relação aos dados amostrais.

Tabela 7 - *Clusters* das observações da amostra com as variáveis validadas e sem *outliers*

Observação	Grupo	Observação	Grupo	Observação	Grupo	Observação	Grupo
Obs. 001	1	Obs. 031	2	Obs. 061	3	Obs. 091	4
Obs. 002	1	Obs. 032	2	Obs. 062	3	Obs. 092	4
Obs. 003	1	Obs. 033	2	Obs. 063	3	Obs. 093	4
Obs. 004	1	Obs. 034	2	Obs. 064	3	Obs. 094	4
Obs. 005	1	Obs. 035	2	Obs. 065	3	Obs. 095	4
Obs. 006	1	Obs. 036	2	Obs. 066	3	Obs. 096	4
Obs. 007	1	Obs. 037	2	Obs. 067	3	Obs. 097	4
Obs. 008	1	Obs. 038	2	Obs. 068	3	Obs. 098	4
Obs. 009	1	Obs. 039	2	Obs. 069	3	Obs. 099	4
Obs. 010	1	Obs. 040	2	Obs. 070	3	Obs. 100	4
Obs. 011	1	Obs. 041	2	Obs. 071	3	Obs. 101	4
Obs. 012	1	Obs. 042	2	Obs. 072	3	Obs. 102	4
Obs. 013	1	Obs. 043	2	Obs. 073	3	Obs. 103	4
Obs. 014	1	Obs. 044	2	Obs. 074	3	Obs. 104	4
Obs. 015	1	Obs. 045	2	Obs. 075	3	Obs. 105	4
Obs. 016	1	Obs. 046	2	Obs. 076	3	Obs. 106	4
Obs. 017	1	Obs. 047	2	Obs. 077	3	Obs. 107	4
Obs. 018	1	Obs. 048	2	Obs. 078	3	Obs. 108	4
Obs. 019	1	Obs. 049	2	Obs. 079	3	Obs. 109	4
Obs. 020	1	Obs. 050	2	Obs. 080	3	Obs. 110	4
Obs. 021	1	Obs. 051	2	Obs. 081	3	Obs. 111	4
Obs. 022	1	Obs. 052	2	Obs. 082	3	Obs. 112	4
Obs. 023	1	Obs. 053	2	Obs. 083	3	Obs. 113	4
Obs. 024	1	Obs. 054	2	Obs. 084	3	Obs. 114	4
Obs. 025	1	Obs. 055	2	Obs. 085	3	Obs. 115	4
Obs. 026	1	Obs. 056	2	Obs. 086	3	Obs. 116	4
Obs. 027	1	Obs. 057	2	Obs. 087	3	Obs. 117	4
Obs. 028	1	Obs. 058	2	Obs. 088	3	Obs. 118	4
Obs. 029	1	Obs. 059	2	Obs. 089	3	Obs. 119	4
Obs. 030	1	Obs. 060	2	Obs. 090	3	Obs. 120	4

Fonte: Dados da amostra

Ambas as técnicas, *K-means* e PAM, alocaram os mesmos elementos em cada grupo, conforme os dados da amostra. Além disso, pela tabela 8, observam-se as proximidades entre *os k-center* ou centróides (*K-means*) e os medóides (PAM).

Tabela 8 - Definição de Centróides e Medóides para 4 agrupamentos sem *outliers*

Clusters	**Centróides**	**Variável 1**	**Variável 2**	**Medóides**	**Variável 1**	**Variável 2**
1	Cent. 1	14,91	15,34	Obs. 29	15,36	15,65
2	Cent. 2	15,28	5,59	Obs. 51	15,34	5,35
3	Cent. 3	10,15	14,99	Obs. 76	9,91	14,65
4	Cent. 4	6,81	4,11	Obs. 97	6,56	4,38

Fonte: Dados da amostra

Na tabela 9, o vetor da soma dos quadrados $SQ_{IntCluster\,(i)}$ de cada grupo, oriundo da *K-means*, revela a variância interna em cada grupo. A soma dos quadrados $SQ_{ExtCluster}$ representa a variância entre grupos. E esse valor dividido pela soma dos quadrados total SQ_{Total} define o quanto o agrupamento formado pelo algoritmo *K-means* com 4 grupos pode ser considerado aderente aos agrupamentos originais.

Tabela 9 - Variâncias dos 4 agrupamentos formados pelo algoritmo *K-means* sem *outliers*

Clusters	$SQ_{IntCluster\,(i)}$	$SQ_{ExtCluster}$	SQ_{Total}	**% de aderência**
1	76,16	4.708,73	5.034,93	93,5
2	67,02			
3	114,83			
4	68,19			

Fonte: Dados da amostra

Empregando o percentual de aderência, assim como o coeficiente de aglomeração descrito pelo procedimento hierárquico, o incremento percentual pode ser considerado como regra de parada na utilização do algoritmo *K-means* para o número de grupos. Conforme a tabela 10, nas passagens de 2 para 3 *clusters* e de 3 para 4 *clusters,* o percentual de incremento é de 22,1% e 6,8%, respectivamente, o que representa um o ganho ao aumentar em uma unidade o número de grupos. Após, para as demais quantidades de *clusters,* ocorre uma redução acentuada desse incremento, o que representa um pequeno ganho a cada acréscimo de grupo.

Tabela 10 - Variâncias dos 4 agrupamentos formados pelo algoritmo *K-means* sem *outliers*

Clusters	$SQ_{ExtCluster}$	SQ_{Total}	% de aderência	Incremento
2	3.256,81	5.034,93	64,7	-
3	4.367,68		86,7	22,1
4	4.708,73		**93,5**	**6,8**
5	4.775,01		94,8	1,3
6	4.808,85		95,5	0,7
7	4.839,18		96,1	0,6

Fonte: Dados da amostra

No resultado do algoritmo PAM, por default, foi considerado um conjunto de observações como medóides iniciais. As observações da tabela 11 representam a estrutura de cada grupo gerado. Esse conjunto de medóides formou os 4 grupos construídos através da atribuição de cada observação para o medóide mais próximo. Nessa tabela destacam-se, concomitantemente, o tamanho dos *clusters*, os medóides em cada grupo, as dissimilaridades, o diâmetro e a separação dos agrupamentos.

Tabela 11 - Descrição dos 4 agrupamentos sem *outliers* formados pelo algoritmo PAM

Tamanho de Agrupamentos	Medóides	Dissimilaridade Máxima	Dissimilaridade Média	Diâmetro	Separação
30	Obs. 29	2,74	1,52	4,93	1,30
30	Obs. 51	3,64	1,32	5,51	3,83
30	Obs. 76	5,05	1,66	7,38	1,30
30	Obs. 97	3,60	1,24	5,18	3,83

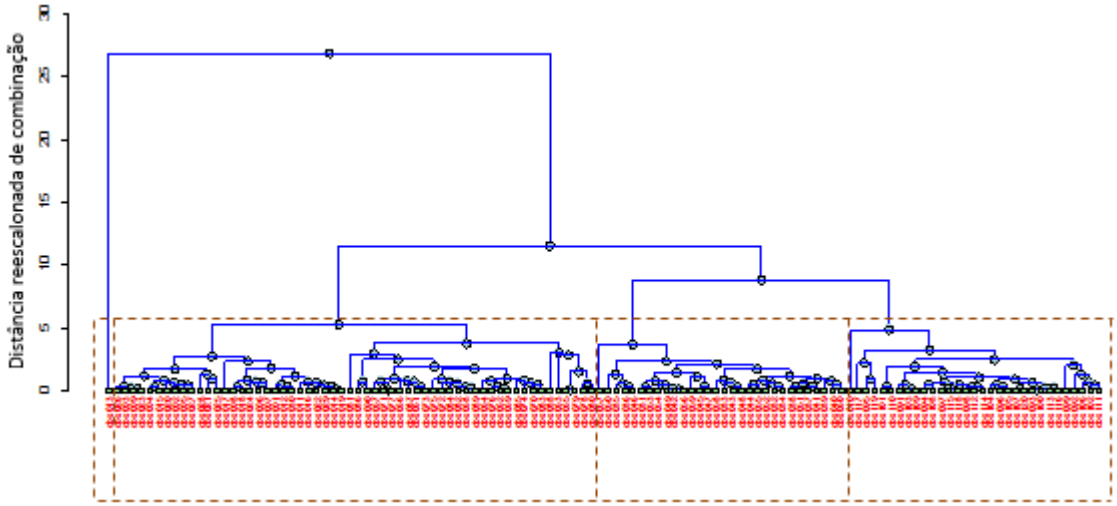
Fonte: Dados da amostra

Apesar de podermos especificar os medóides centrais, quando isso não ocorre, o algoritmo procura primeiro por um conjunto inicial de medóides ótimos (essa é chamada a fase de construção). Em seguida, ele encontra um mínimo local para a função objetivo, ou seja, uma solução de tal modo que não exista nenhuma opção de medóides que irá diminuir essa função (essa é chamada a fase de troca).

4.5 Definição do Número de *Clusters* – Variáveis com 4 *Outliers*

Para a definição do número de grupos ou *clusters*, foi utilizada a análise gráfica do dendrograma gerado pelo algoritmo hierárquico, com todas as variáveis. Nessa visão, o resultado gráfico (figura 6), também com 4 grupos, representou aglomerados mais homogêneos entre si.

Figura 6 - Dendrograma do procedimento hierárquico (método average linkage).



Fonte: Elaborado a partir dos dados da amostra

Na avaliação pela a ANOVA (*Analysis of Variance*), proveniente do procedimento não-hierárquico *(K-means)* com 4 *clusters*, ao nível de significância de $\alpha = 0{,}05$, as duas variáveis simuladas mantiveram-se como significativas conforme tabela 12.

Tabela 12 - Análise de variância da amostra – A NOVA– variáveis com *outliers*

Variáveis	Agrupamento		Erro		Valor F	Nível de Significância
	Quadrático Médio	Graus de Liberdade	Quadrático Médio	Graus de Liberdade		
Var 1	594,741	3	4,218	116	141,001	0,000
Var 2	1078,277	3	1,677	116	642,821	0,000

Fonte: Elaborado a partir dos dados da amostra

Assim como ocorreu no conjunto de dados sem *outlires*, as variáveis mantiveram-se válidas pela ANOVA. Na análise do coeficiente de aglomeração resultante do procedimento hierárquico aglomerativo (tabela 13), foram identificadas duas grandes variações, sublinhadas e em destaques, nas passagens de 1 para 2 agregados (133,2%) e de 3 para 4 agregados (68,0%). Na primeira passagem, torna-se evidente a identificação do primeiro *outlier*; já na segunda, a queda de variação de 4 para 5 grupos (8,2%) é o indicativo para 4 *clusters* a ser estabelecido, corroborado pelo dendrograma (figura 6).

Tabela 13 - Coeficiente de aglomeração da análise hierárquica

Número de Agrupamentos	Coeficiente de Aglomeração	Variação
9	3,030	2,2%
8	3,197	15,1%
7	3,678	1,3%
6	3,726	29,7%
5	4,833	8,2%
4	**5,230**	**68,0%**
3	8,789	31,0%
2	**11,510**	**133,2%**
1	26,840	-

Fonte: Dados da amostra

4.6 Composição e Análise de Agrupamentos – Com 4 *Outliers*

Definido o número de grupos (4), foram aplicados os algoritmos *K-means* e *Partitioning Around Medoids* (PAM) para comparação entre eles e em relação aos dados amostrais (tabela 14).

Tabela 14 - ***Clusters*** **das observações da amostra com** ***outliers***

(continua)

Observação	Simulação	PAM	*K-means*	Observação	Simulação	PAM	*K-means*
Obs. 001	1	1	1	Obs. 061	3	3	**1**
Obs. 002	1	1	1	Obs. 062	3	3	**1**
Obs. 003	1	1	1	Obs. 063	3	3	**1**
Obs. 004	1	1	1	Obs. 064	3	3	**1**
Obs. 005	1	1	1	Obs. 065	3	3	**1**
Obs. 006	1	1	1	Obs. 066	3	3	**1**
Obs. 007	1	1	1	Obs. 067	3	3	**1**
Obs. 008	1	1	1	Obs. 068	3	3	**1**
Obs. 009	1	1	1	Obs. 069	3	3	**1**
Obs. 010	1	**4**	**4**	Obs. 070	3	3	**1**
Obs. 011	1	1	1	Obs. 071	3	3	**1**
Obs. 012	1	1	1	Obs. 072	3	3	**1**
Obs. 013	1	1	**3**	Obs. 073	3	3	**1**
Obs. 014	1	1	1	Obs. 074	3	3	**1**
Obs. 015	1	1	1	Obs. 075	3	3	**1**
Obs. 016	1	1	1	Obs. 076	3	3	**1**
Obs. 017	1	1	1	Obs. 077	3	3	**1**
Obs. 018	1	1	1	Obs. 078	3	3	**1**
Obs. 019	1	1	1	Obs. 079	3	3	**1**
Obs. 020	1	1	1	Obs. 080	3	3	**1**
Obs. 021	1	1	1	Obs. 081	3	3	**1**
Obs. 022	1	1	1	Obs. 082	3	3	**1**
Obs. 023	1	1	1	Obs. 083	3	3	**1**
Obs. 024	1	1	1	Obs. 084	3	3	**1**
Obs. 025	1	1	1	Obs. 085	3	3	**1**
Obs. 026	1	1	1	Obs. 086	3	3	**1**
Obs. 027	1	1	1	Obs. 087	3	3	**1**
Obs. 028	1	1	1	Obs. 088	3	3	**1**
Obs. 029	1	1	1	Obs. 089	3	3	**1**
Obs. 030	1	1	1	Obs. 090	3	3	**1**
Obs. 031	2	2	2	Obs. 091	4	4	4
Obs. 032	2	2	2	Obs. 092	4	4	4
Obs. 033	2	2	2	Obs. 093	4	4	4
Obs. 034	2	2	2	Obs. 094	4	4	4
Obs. 035	2	2	2	Obs. 095	4	4	4
Obs. 036	2	2	2	Obs. 096	4	4	4
Obs. 037	2	2	2	Obs. 097	4	4	4
Obs. 038	2	2	2	Obs. 098	4	4	4
Obs. 039	2	2	2	Obs. 099	4	4	4
Obs. 040	2	2	2	Obs. 100	4	4	4
Obs. 041	2	2	2	Obs. 101	4	4	4
Obs. 042	2	2	2	Obs. 102	4	4	4

Tabela 14 - *Clusters* das observações da amostra com *outliers*

(conclusão)

Observação	Simulação	PAM	*K-means*	Observação	Simulação	PAM	*K-means*
Obs. 043	2	2	2	Obs. 103	4	4	4
Obs. 044	2	2	2	Obs. 104	4	4	4
Obs. 045	2	2	2	Obs. 105	4	4	4
Obs. 046	2	2	2	Obs. 106	4	4	4
Obs. 047	2	2	2	Obs. 107	4	4	4
Obs. 048	2	2	2	Obs. 108	4	4	4
Obs. 049	2	2	2	Obs. 109	4	4	4
Obs. 050	2	2	2	Obs. 110	4	4	4
Obs. 051	2	2	2	Obs. 111	4	4	4
Obs. 052	2	2	2	Obs. 112	4	4	4
Obs. 053	2	2	2	Obs. 113	4	4	4
Obs. 054	2	2	2	Obs. 114	4	4	4
Obs. 055	2	2	2	Obs. 115	4	4	4
Obs. 056	2	2	2	Obs. 116	4	4	4
Obs. 057	2	2	2	Obs. 117	4	4	4
Obs. 058	2	2	2	Obs. 118	4	4	4
Obs. 059	2	2	2	Obs. 119	4	4	4
Obs. 060	2	2	2	Obs. 120	4	4	4

Fonte: Dados da amostra

Com a inclusão de duas coordenadas destoantes, a técnica *K-means* gerou *clusters* desbalanceados e não homogêneos em comparação com os dados amostrais, formando as seguintes composições: 58 observações no 1º *cluster*, 30 observações no 2º *cluster,* uma (1) observação no 3º *cluster* e 31 observações no 4º *cluster*. Já a técnica PAM manteve a distribuição dos elementos balanceada e grupos homogêneos, apresentando apenas uma observação em grupo diferente do arranjo inicial. Essa altercação, também, é ressaltada com o distanciamento entres os *k-center* ou centróides e os medóides (tabela 15)

Tabela 15 - Definição de Centróides e Medóides para 4 agrupamentos com *outliers*

Clusters	Centróides	Variável 1	Variável 2	Medóides	Variável 1	Variável 2
1	Cent. 1	12,49	15,17	Obs. 29	15,36	15,65
2	Cent. 2	15,28	5,59	Obs. 51	15,34	5,35
3	Cent. 3	36,50	18,95	Obs. 76	9,91	14,65
4	Cent. 4	6,77	4,25	Obs. 97	6,56	4,38

Fonte: Dados da amostra

Pelo descrito na tabela 16, observa-se que o vetor da soma dos quadrados $SQ_{IntCluster\ (i)}$ de cada grupo procedente da *K-means* apresenta, principalmente, grande variância interna para grupo 1 que absorveu diversas observações, comparado com os dados sem *outliers*. O *cluster* 3, por alocar apenas uma observação, não possui variância. A soma dos quadrados $SQ_{ExtCluster}$, que representa a variância entre grupos, exibiu aumento de 6,6% (309,42).

Em relação à análise com os dados sem *outliers*, a $SQ_{ExtCluster}$ passou de 4.708,63 para 5.019,05. Esse aumento impactou o percentual de aderência ($SQ_{ExtCluster}$ / SQ_{Total}) reduzindo em 5,5%. Porém, mesmo com essa configuração, o modelo formado pelo algoritmo *K-means* com 4 grupos pode ser considerado aderente ao agrupamento possivelmente original (tabela 16).

Tabela 16 - Variâncias dos 4 agrupamentos formados pelo algoritmo *K-means* com *outliers*

Clusters	$SQ_{IntCluster\ (i)}$	$SQ_{ExtCluster}$	SQ_{Total}	% de aderência
1	527,11	5.019,05	5.702,92	88,0
2	67,02			
3	0,00			
4	89,74			

Fonte: Dados da amostra

Conforme a tabela 17, na passagem de 3 para 4 *clusters,* o percentual de incremento chegou a 21,9%, o que representa o melhor ganho ao aumentar em uma unidade o número de grupos. Após, para as quantidades superiores a *4 clusters*, observam-se variações negativas.

Tabela 17 - Variâncias dos 4 agrupamentos formados pelo algoritmo *K-means* com *outliers*

Clusters	$SQ_{ExtCluster}$	SQ_{Total}	% de aderência	Incremento
2	3.305,98	5.702,92	58,0	-
3	3.768,94		66,1	8,1
4	5.019,05		**88,0**	**21,9**
5	4.925,27		86,4	-1,6
6	4.983,41		87,4	-1,0
7	4.948,82		86,7	-0,6

Fonte: Dados da amostra

No resultado do algoritmo PAM, a estrutura de cada grupo gerado (tabela 18), em relação à descrição do conjunto de dados anterior, revela as principais alterações que ocorreram no *cluster* 1, como aumento da dissimilaridade máxima e do diâmetro. Já o quarto *cluster* registrou alteração moderada no diâmetro, produto da entrada e saída de elemento entre o primeiro e terceiro grupo.

Tabela 18 - Descrição dos 4 agrupamentos formados pelo algoritmo PAM – com *outliers*

Tamanho de Agrupamentos	Medóides	Dissimilaridade Máxima	Dissimilaridade Média	Diâmetro	Separação
29	Obs. 29	21,40	2,18	23,69	1,30
30	Obs. 51	3,64	1,32	5,51	3,83
30	Obs. 76	5,05	1,66	7,38	1,30
31	Obs. 97	4,40	1,34	7,99	3,83

Fonte: Dados da amostra

5. Conclusões

Este trabalho teve como objetivo avaliar as técnicas de agrupamentos *K-means* e *Partitioning Around Medoids* (PAM), aplicadas em dados simulados aleatoriamente sem e com *outliers* (valores atípicos). O procedimento sugerido, diagrama de decisão, mostrou-se eficaz para definição do número de grupos, avaliação da significância das variáveis e definição dos elementos para compor cada grupo, além de identificar possíveis *outliers*.

Pelos resultados apresentados neste artigo, as técnicas *K-means e Partitioning Around Medoids* (PAM) apresentaram os mesmos desempenhos com os dados da amostra sem *outliers*. Porém, quando analisada com os valores atípicos, a técnica *Partitioning Around Medoids* (PAM) mostrou-se mais robusta, consistente e pouco impactada pelos valores destoantes. Já, na técnica *K-means*, os valores atípicos afetaram, fortemente, os centróides, que são formados pelas médias ou médias ponderadas dos pontos, e provocaram o desbalanceamento dos grupos.

Considerando as duas técnicas de Análise de Agrupamentos, estudos corroboram que ambas são ferramentas poderosas na segmentação das observações em grupos. A escolha para utilização dessas técnicas, geralmente, está relacionada com a origem e custo da base de dados em estudo e com o conhecimento do pesquisador sobre o tema. A correta utilização dessas ferramentas é fator preponderante para delineamento dos dados e obtenção de resultados consistentes nas pesquisas que as empregam.

Em um campo vasto de pesquisas, diversas variantes podem ser aplicadas sobre as análises propostas neste artigo. A quantidade de pontos contaminados para a avaliação entre as duas técnicas pode ser considerada uma das limitações deste trabalho, a ser testada em futuros estudos, assim como, a utilização de critérios complementares ao dendrograma para a determinação da quantidade de grupos descrito por Semaan et al. (2012); *Bisecting K-means*.

Referências bibliográficas

ANDERBERG, M. R. Cluster analysis for applications, volume 19, Academic Press, New York, 1973

ARABIE, P.; HUBERT, L. J., DE SOETE, G. Clustering and Classification, World Scientific, 1999.

BRITO, W. M.; SEMAAN, G. S.; BRITO, J. A. Um Algoritmo Genético para o Problema dos K-Médoides. 10th Brazilian Congress on Computational Intelligence (CBIC'2011), Fortaleza, Ceará Brasil, 2011.

CARGNELUTTI FILHO, A.; RIBEIRO, N. D.; CITTON, R.; REIS, P.; SOUZA, J. R.; JOST, E. Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão. Ciência Rural, Santa Maria, volume 38, nº 8, páginas 2138 a 2145, 2008.

CORRAR, J. L.; PAULO, E.; DIAS FILHO, J. M. (Coordenadores). Análise Multivariada. 1ª Edição – 3ª reimpressão – São Paulo – Editora Atlas, 2011.

EVERITT, B. S.; LANDAU, S.; LEESE, M. Cluster Analysis, Hodder Arnold Publication, 2001.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A., C., P., L. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina – Rio de Janeiro – Livros Técnicos e Científicos Editora Ltda., 2011.

FERREIRA, D. F. Análise Multivariada, Universidade Federal de Larvas, Minas Gerais, 1996.

HAIR, J. F. Jr.; ANDERSON, I.; TATHAM, R. E. II.; BLACK, W. C. Análise Multivariada de Dados, 5ª Edição – Porto Alegre – Editora Bookman, 2005.

HARTIGAN, J. A. Clustering Algorithms (Probability& Mathematical Statistics), Editora John Wiley & Sons Inc, 1975.

HU, X.; ILLHOY, Y. Cluster ensemble and its applications in gene expression analysis. ACM International Conference Proceeding Series; Proceedings of the second conference on Asia-Pacific bioinformatics – volume 29, páginas 297 a 302, Nova Zelândia, 2004.

JAIN, A. K.; DUBES, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.

Jr, H. S. Cardinality of finite topologies. Journal of Combinatorial Theory, nº 5, volume 1, páginas 82 a 86, 1968.

KAUFMAN, L; ROUSSEEUW, P. J. Finding Groups in Data: An Introduction to Cluster Analysis. Editora Wiley-Interscience, 1990.

KUMAR, V.; STEINBACH, M.; TAN, P. N. Introdução ao Data Mining - Mineração de Dados. Editora Moderna, 2009.

LINDEN, R. Técnicas de Agrupamento. Revista de Sistemas de Informação da FSMA, nº 4, páginas 18 a 36, Macaé – Rio de Janeiro, 2009.

MARQUES, M. A. M. Caracterização da Contribuição dos Consumidores BT para as Perdas na Rede. Porto, Dissertação (Mestrado em Gestão de Informação) – Universidade do Porto – Faculdade de Engenharia – FEUP, 2005.

PARK, H. S.; LEE, J. S.; JUN, C. H. A K-means-like Algorithm for K-medoids Clustering and Its Performance. Department of Industrial and Management Engineering, POSTECH. San 31 Hyoja-dong, Pohang 790-784, S. Korea, 2009.

SEMAAN, G. S.; BRITO, J. A. M.; OCHI, L. S. Um Método Hierárquico par Determinação de Número Ideal de Grupos. Revista Brasileira de Estatística, volume 73, nº 236, páginas 81 a 113, 2012.

SEMAAN, G. S.; TORRES, C. C.; BRITO, J. A. M.; OCHI, L. S. Um Método Baseado em Combinação de Soluções com Coassociação para o Problema de Agrupamento Automático. Revista Brasileira de Estatística, volume 74, nº 239, páginas 43 a 68, 2013.

Abstract

This article proposes a comparison of performance between automatic clusters methods, *K-means and Partitioning Around Medoids* (PAM), applied in simulated data with and without *outliers*. For this analysis, it's suggested the decision diagram procedure to define the number of groups, assessing the significance of the variables and the definition of the elements that will compose each group. Whereas both cluster analysis techniques are powerful tools in segmenting the observations, the choice of these tools usually are related to the origin of databases, research costs and knowledge of the researcher on the subject. In the results presented in this paper, the techniques *K-means and Partitioning Around Medoids* had similar performances with the sample data without outliers. However, when analyzed with the outliers, the *Medoids Partitioning Around* technique proves to be more robust and consistent compared with the *K-means* technique.

Keywords: Clusters Analysis. Partitioning Around Medoids. K-means. K-Centers. Medoids.

Resiliência em Redes de Computadores, baseada na Teoria da Confiabilidade

Laurinete do Nascimento Bacelar dos Reis Ximenes [1]
Ronaldo M. Salles [2]
Paulo Afonso Lopes da Silva [3]

Resumo

O crescente número de aplicações executadas na rede mundial de computadores tem incentivado o estudo da resiliência em redes de computadores. Pesquisadores e especialistas buscam a padronização de métricas para a correta leitura do nível de resiliência nas redes. O presente artigo propõe um indicador de resiliência que pesquisa uma infraestrutura confiável desde o início do projeto, com base na teoria da confiabilidade, cujo objetivo é aumentar a da rede, não apenas dos componentes como também do sistema propriamente dito, por redundância, visando reduzir os custos com manutenções corretivas. Testou-se o indicador proposto em *backbones* de topologias reais, com resultados melhores que os atuais, e os gerentes de redes, ao utilizarem a presente metodologia no planejamento de seus sistemas, também terão uma ferramenta para facilitar recuperações em eventuais incidentes de falhas. A utilização de *links* com maior confiabilidade proporciona uma estrutura otimizada para enfrentar as ameaças à rede, além de promover o aumento da qualidade dos serviços fornecidos pela rede e o consequente cumprimento dos acordos de nível de serviço.

[1] Instituto Militar de Engenharia (IME) – End. Praça General Tibúrcio, 80 CEP: 22290-270 – Rio de Janeiro, RJ, Brasil. E-mail: lauranbr_79@yahoo.com.br
[2] Instituto Militar de Engenharia (IME). E-mail: salles@ime.eb.br
[3] Instituto Militar de Engenharia (IME). E-mail: pauloafonsolopes@uol.com.br

1. Introdução

O objetivo deste artigo é apresentar uma metodologia para determinar o grau de resiliência de uma rede, utilizando-se da teoria da confiabilidade (GROSH, 2012; LOCKS, 1995; NACHLAS, 2005), que acrescentará às questões topológicas o enfoque probabilístico necessário para se compreender o comportamento das falhas que podem ocorrer não apenas em uma operação normal, mas também em face de diversas ameaças, como desastres naturais e conflitos, resultando um projeto de rede com maior confiabilidade que a atual.

A internet tem se tornado essencial em todos os aspectos da vida moderna, e interrupções na operação das redes podem acarretar graves consequências (STERBENZ et al., 2010). O dia a dia dos cidadãos de todos os países está fortemente relacionado a operações com base na rede mundial de computadores. Entretanto, o aumento de tal dependência e a sofisticação dos serviços torna a internet mais vulnerável a problemas. O contínuo crescimento da confiança nesses serviços produz dois efeitos principais: em primeiro lugar, o aumento dessa confiança aumenta a gravidade das interrupções e, em segundo lugar, as redes tornam-se mais atrativas aos ataques cibernéticos (STERBENZ et al., 2010). Por causa desses fatos, surge a preocupação com a proteção do funcionamento do sistema de redes de computadores, diante das ameaças que podem lhes sobrevir. Esse tema tem sido objeto das recentes pesquisas nas áreas de sistemas tolerantes a falhas e de resiliência.

Para se aumentar a confiabilidade de uma rede é preciso investir-se nos seus componentes desde o projeto, isto é, o ganho em confiabilidade em qualquer sistema, seja ele uma rede de comunicações ou uma linha de produção, depende do projeto de confiabilidade das partes que o compõem.

A presente metodologia proporciona a mensuração do grau de resiliência de uma rede por meio da comparação de diferentes características de falhas dos seus componentes, resultando na possibilidade de se optar pelo melhor dos meios de transmissão disponíveis no mercado, com diferentes probabilidades de falhas. Cabe ressaltar que este artigo não somente propõe uma medida que possibilita comparar topologias, como também um método de simulação para obterem-se os perfis de falhas, com imediata aplicação no planejamento de estruturas de rede.

Desenvolveu-se a metodologia com o foco na eficácia do cálculo da métrica da resiliência, ao considerar três aspectos: o número de nós remanescentes no maior componente conexo, LCC-*Largest Connected component* (BEYGELZIMER, 2005), a quantidade de componentes conexas após a ocorrência das falhas aleatórias, representadas pela retirada das arestas do grafo da rede e, finalmente, o número total de nós na configuração inicial para normalizar esse Indicador de Resiliência (IR). Com isso, obtém-se uma medição da resiliência para representar diferenças na robustez da rede após alterações da topologia original.

Este artigo está organizado da seguinte maneira: na seção 2, revisão dos principais artigos relacionados. A seção 3 apresenta a metodologia, definindo o Indicador de Resiliência, suas etapas de cálculo e implementação. A seção 4 detalha os testes realizados e apresenta os resultados dos experimentos nos *backbones* das topologias reais estudadas. A seção 5 apresenta as considerações finais e as conclusões.

2. Revisão de Literatura

2.1 Confiabilidade

A Confiabilidade, de acordo com (NBR, 1994), é um termo para medida de desempenho, e significa a capacidade de um item em desempenhar uma função requerida, sob condições especificadas, durante um dado intervalo de tempo.

Entre os autores que podem ser considerados referências em confiabilidade, citar Kapur e Lamberson (1977), Ireson, Coombs e Moss (1988), Lewis (1995), O´Connor e Kleiner (2012), Bazovsky (2004, Elsayed (2012) e Tobias (2011), a partir dos quais se pode escrever um resumo dos conceitos básicos da confiabilidade para facilitar, conforme a seguir.

A teoria da confiabilidade nada mais é que aplicar a teoria da probabilidade na modelagem e na previsão da probabilidade de falhas, baseando-se, principalmente, nos conceitos de função densidade de probabilidade $f(t)$ e função de distribuição $F(t)$, onde t é o tempo até a ocorrência da falha.

A probabilidade do tempo T da ocorrência da falha pertencer ao intervalo $(t, t + \Delta t)$ é:

$$Pr(t \leq T \leq t + \Delta t) = f(t)\,\Delta t = F(t + \Delta t) - F(t) \qquad (1)$$

onde $F(t)$, e $f(t)$ são, respectivamente, as funções distribuição e de densidade de probabilidades desse tempo.

Como a função de distribuição é a integral da função densidade de probabilidade,

$$F(t) = \int_0^t f(x)dx \tag{2}$$

ou seja, essa função equivale à probabilidade de falha até o tempo t.

Por essa razão, pode-se escrever que

$$F(t) = \Pr\ (0 \leq T \leq t) = \int_0^t f(x)dx \tag{3}$$

Define-se a probabilidade de não haver falha até o tempo t, $R(t)$, como a probabilidade de que o tempo até falhar é maior que t , isto é, $T > t$, da seguinte maneira:

$$R(t) = \Pr(T > t) = 1 - F(t) = Pr \int_t^\infty f(x)dx \tag{4}$$

na qual $R(T)$ denomina-se função confiabilidade.

Entretanto, ao se estudar dados de falhas, $f(t)$ não é muito útil, e então deduz-se outra função, denominada taxa de falhas $h(t) = \frac{f(t)}{R(t)}$, taxa instantânea de falhas no tempo t, sabendo-se que se sobreviveu até o tempo t.

De acordo com a (NBR, 1994), a medida da confiabilidade é representada por uma função para descrever uma variável aleatória. Seja X essa variável aleatória, que modele o período da vida útil do componente, período de ocorrência de somente falhas aleatórias (LEWIS, 1995). Como a função confiabilidade, $R(t)$, é

$$R(t) = P(X > t) = 1 - F(t) \tag{5}$$

e que, na vida útil, a taxa de falhas é constante e modela-se pela distribuição exponencial com parâmetro $\lambda, \lambda > 0$, então a função de confiabilidade é dada por:

$$R(t) = P(X > t) = e^{-\lambda t} \quad (6)$$

O exposto resume, matematicamente, a maior parte dos conceitos que se deve saber a respeito da teoria da confiabilidade.

Conceito integrado à confiabilidade no contexto das redes de comunicações é o de resiliência, definida como a capacidade de uma rede fornecer e manter um nível adequado de serviço, mesmo na presença de falhas ou ameaças à operação normal (STERBENZ et al., 2010). Essa característica é importante, principalmente no caso de empresas provedoras de serviço de grande abrangência, sobre as quais existe uma demanda de tráfego de nível elevado que, em caso de falhas e sobrecarga, exige a manutenção do fluxo de informações, com a proteção não apenas da conectividade, mas também dos níveis de *QoS (Quality of Service-Qualidade do Serviço)*, além de um imediato retorno ao estado de normalidade (MENTH, 2009).

As estratégias de resiliência aplicadas a uma rede visam minimizar os impactos das falhas, garantindo restaurar os níveis das camadas IP (Internet Protocol), ou seja, a restauração da conectividade entre os links afetados.

O surgimento de novos problemas, causado pelo desenvolvimento da tecnologia, torna necessário obter-se um método de mensuração da resiliência de uma rede de computadores que atenda aos desejados parâmetros de confiabilidade e de disponibilidade em uma quantidade considerável de cenários de falhas.

2.2 Estado da Arte

Entre as vantagens do estudo da confiabilidade de sistemas com enfoque na resiliência, tem-se a redução dos custos com os programas de manutenção, a melhoria na disponibilidade (PHAM, 2006), o aumento da vida útil dos componentes da rede e o decréscimo nas redundâncias, tanto de equipamentos quanto de enlaces, porque, como define (OMER, 2009), um sistema resiliente é o que tem probabilidades de falhas, consequências e tempo de recuperação reduzidos.

Os poucos artigos referentes ao tema resiliência em redes de computadores e tolerância a falhas, atêm-se, principalmente, na questão topológica da rede (SHOOMAN,2002).

Em (MARKOPOULOU, 2004), foi constatado que as falhas ou recuperação em um *link* IP acarretam alterações na topologia da camada de rede. Quando essas mudanças ocorrem, os roteadores nas duas pontas do *link* notificam o restante da rede via IS-IS (MARKOPOULOU, 2004), comprovando-se que a maioria das falhas simultâneas acontece devido a problemas em roteadores, e como um roteador tem vários *links*, em consequência todos os componentes são afetados ao mesmo tempo, buscando retornarem ao estado *UP'* no mesmo instante.

MARKOPOULOU (2004) indicou que 20% das falhas devem-se a manutenções programadas e, das corretivas, 30% das falhas classificam-se como falhas compartilhadas, a metade das quais afetou *links* com um roteador em comum, enquanto a outra metade afetou *links* que compartilhavam uma estrutura ótica, indicando falhas nessa camada. Além disso, concluiu que 16% e 11% das falhas não planejadas podem ser atribuídas a problemas relacionados a roteadores e à estrutura ótica, respectivamente.

Portanto, comprova-se a necessidade de se investir e manter, primeiramente, as camadas IP e ótica com o intuito de prevenção das falhas.

Segundo (LEE, 2005), a estrutura topológica da rede, diante dos ataques cibernéticos, influencia, diretamente, sua robustez. Esse artigo demonstra, por meio de métricas sugeridas para avaliação da resiliência e técnicas para modelar os ataques, utilizando teoria dos grafos, que a robustez da estrutura da internet não é melhor do que a de uma topologia de uma rede aleatória.

Segundo (LEE, 2005), os estudos comprovaram que um colapso total da internet, causado por ataques aos seus nós seria impraticável devido à alta conectividade da rede, porém ataques bem direcionados, aqueles que visam a remoção dos nós mais importantes, poderiam debilitar a rede, o que, em termos práticos, equivaleria a um colapso total.

MENTH (2009) elaborou um processo algorítmico de avaliação da indisponibilidade e de sobrecarga da rede a partir de uma descrição probabilística que envolvesse as situações de falhas em geral, *hotspots* locais e roteamento entre domínios. Para esse problema, (MENTH, 2009) utilizou duas abordagens: na primeira, somente cenários com probabilidade mínima de ocorrência, nos quais se avaliaram as potenciais causas de sobrecarga para otimizar a velocidade computacional, e estabeleceram-se limites inferiores e superiores para os resultados obtidos. Na segunda abordagem, foram propostos algoritmos que reutilizavam resultados intermediários em seus diferentes cálculos. MENTH (2009) também verificou que a análise por meio de gráficos, resumindo as estatísticas calculadas, promove a praticidade da análise da resiliência da rede.

Este artigo trata da resiliência sob o aspecto de estratégias com base em um método de avaliação do perfil de falhas dos componentes. Além disso, considera a influência da topologia da rede, de comprovada capacidade em assegurar a conectividade entre os nós, possibilitando a ação dos protocolos de roteamento, e garantindo-se os níveis de QoS contratados, aspectos não contemplados nas pesquisas já realizadas.

3. Metodologia Proposta

A presente metodologia proporciona a mensuração do grau de resiliência de uma rede, comparando-se diferentes perfis de falhas dos seus componentes, permitindo optar-se entre os meios de transmissão disponíveis no mercado, cada um deles com diferentes probabilidades de falhas. Cabe ressaltar que este artigo não somente propõe uma medida para comparar topologias, como também um método de simulação para obterem-se os perfis de falhas para uso nos planejamentos de estruturas da rede.

Inicia-se a pesquisa da métrica proposta pelo cálculo das confiabilidades das arestas, por meio da estimação do limite inferior, com uma confiança de $(1-\alpha)100\%$, da Confiabilidade, R_i (IRESON, 1988):

$$R_i \frac{y}{y + (n - y + 1) \times F_{\alpha,2(n-y+1),2y}} \tag{7}$$

Na equação (7), encontrada em (IRESON, 1988) como referência à obra (KAPUR et al., 1982), quando $y = 0$, tem-se o caso de ensaio de sucesso, e a equação se resume a $R_i = (1 - \alpha)^{\frac{1}{n}}$, (BACELAR, 2014).

Pela dificuldade na obtenção dos dados relativos às características necessárias para o desenvolvimento do presente artigo, utiliza-se o Método de Monte Carlo (MMC) (CHWIF, 2006), para a geração de dígitos pseudoaleatórios correspondentes à variável aleatória $\mathrm{Y} \sim \mathrm{Binomial}\,(\mathrm{n}, \mathrm{p})$, que representa o número de componentes que sobreviveram durante o experimento, normalmente executado como um dos processos de fabricação dos produtos, onde n é o número de retiradas em cada tentativa, e p é a probabilidade de sucesso, isto é, a probabilidade do componente ser aprovado.

Devem ser gerados tantos números quantos forem os *links* (número de arestas do grafo subjacente à topologia real). A primeira probabilidade de sucesso a ser usada para avaliar-se seu impacto no planejamento da rede, é a probabilidade no ponto máximo da variância da distribuição Binomial, $\mathrm{p} = 1/2$.

Determina-se o valor de p pela a primeira derivada de $\mathrm{VAR(Y)} = \mathrm{n} \times \mathrm{p} \times (1 - \mathrm{p})$ com respeito a p, obtendo-se $\mathrm{d(VAR(Y))/dp} = \mathrm{n} - 2\mathrm{np}$, e igualar-se essa equação a zero. Daí que, $1 - 2p = 0$, resultando em $\mathrm{p} = 0.5$, valor para o pior caso, já que, nesse ponto, a variação de Y com relação à sua média, assume o valor máximo, comprovado pela derivada segunda, $\frac{d^2(\mathrm{VAR(Y)})}{dp^2} = -2\mathrm{n}$, o que caracteriza o ponto de máximo. Aplicam-se esses valores à equação (7), obtendo-se o limite inferior de confiança de 95% para os R_i.

A partir desse ponto, retiram-se os enlaces que apresentarem valores de R_i menores do que o requisito mínimo de confiabilidade (RMC). Uma vez que os menores valores de confiabilidade representam as maiores probabilidades de falhas, é razoável pensar que, em uma operação real da rede, esses serão os primeiros enlaces a serem desconectados.

A presente metodologia considera os seguintes parâmetros: o número de nós sobreviventes no maior componente conexo (LCC) (BEYGELZIMER, 2005), a quantidade de componentes conexas após a ocorrência das falhas aleatórias, representadas pela retirada das arestas do grafo da rede e, por último, o número total de nós da configuração inicial.

O número de nós no maior componente conexo e a quantidade de componentes conexas têm uma relação inversa, ou seja, quando o número de nós no maior componente conexo diminui, a quantidade de componentes conexas aumenta.

Na configuração inicial, antes das falhas ocorrerem, o grafo G é conexo e, portanto, tem apenas uma componente, após as falhas, o grafo se divide em subgrafos, correspondentes às porções contíguas de sua representação geométrica (SZWARCFITER, 1988).

À medida em que a rede vai sendo fragmentada, indicando perda de potencial de transmissão da informação, essa razão tende a diminuir.

Define-se o proposto Indicador de Resiliência (IR), pela equação:

$$IR = \frac{n^{o}\ de\ nós\ no\ lcc}{n^{o}\ de\ componentes\ conexas\ x\ n^{o}\ de\ nós\ no\ grafo} \qquad (8)$$

O resultado da métrica proposta serve para auxiliar o gerente de rede na tomada de decisão entre topologias diferentes e, não apenas como um valor absoluto a ser interpretado de maneira isolada.

O algoritmo para o cálculo do (IR), foi implementado com o software SciLab 5.5.0 para Windows, software livre, compatível com o MATLAB, disponível em http://www.scilab.org/.

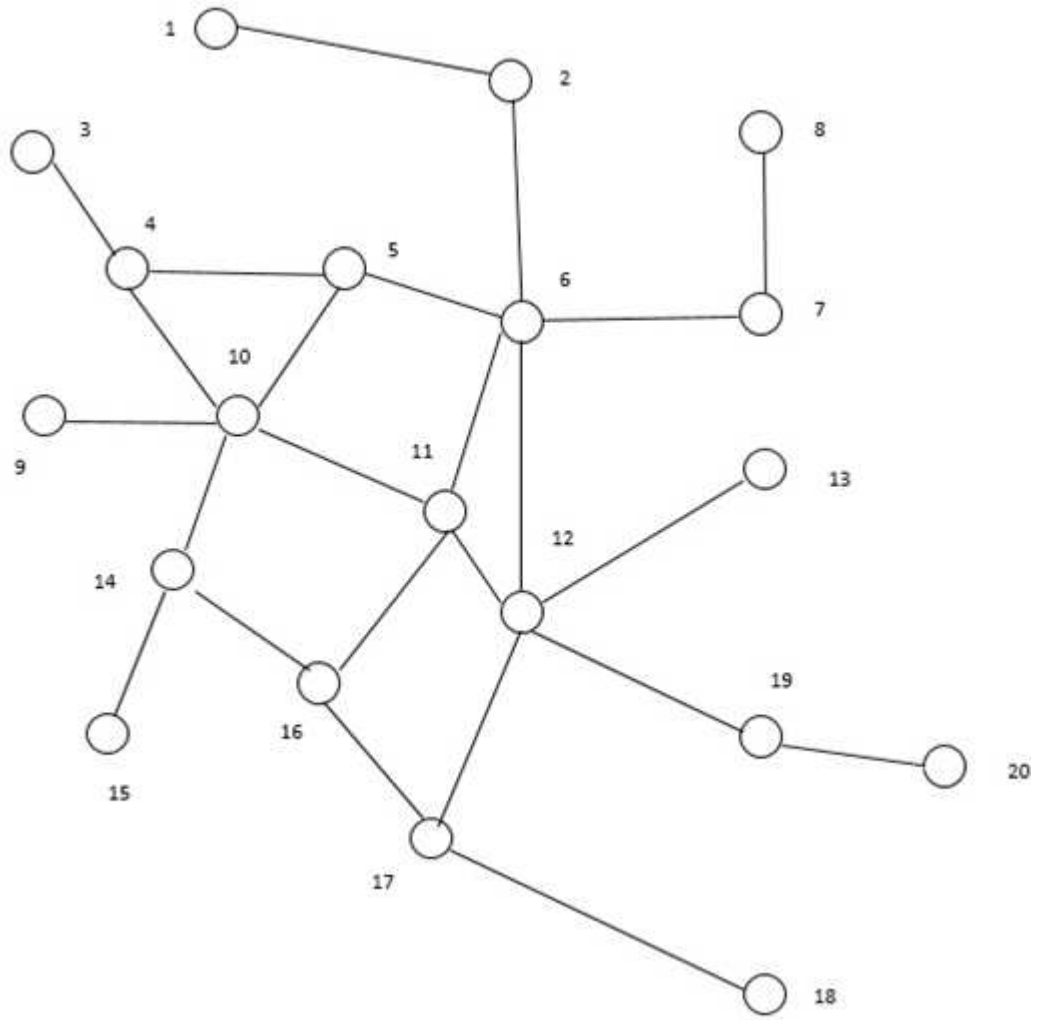
Para a obtenção do IR, desenvolveu-se um algoritmo que lê a matriz de confiabilidade associada ao grafo da topologia, dada pelos R_i, o número de nós do grafo, o número de arestas de uma árvore *spanning tree* associada ao grafo, o número de arestas da configuração inicial, o RMC e o número de iterações decorrentes da escolha desse parâmetro. Após essa etapa, o algoritmo ordena os valores de confiabilidade e efetua a retirada das arestas com valores menores que o RMC. Em seguida, procede-se com a verificação da conectividade do grafo subjacente, a contagem do número de componentes conexas, o número de nós no maior componente conexo (LCC), e finalmente, a obtenção do Indicador de Resiliência Proposto (IR).

Portanto, a métrica proposta indica a robustez da rede frente às prováveis falhas nos *links*, tendo evidenciada ser mais precisa ao detectar as alterações na topologia, quando comparada às medidas de proporção de nós no LCC, AISPL e Diâmetro do LCC, encontradas em (BEYGELZIMER, 2005), conforme provado na Seção 4. Resultados.

3.1 Exemplo de aplicação da metodologia

Seja a rede representada por uma topologia composta de 20 nós e 24 enlaces, conforme a figura (1):

Figura 1 - Exemplo de Topologia de Rede



O objetivo é avaliar quanto uma rede pode suportar de redução na quantidade de seus componentes ao ocorrerem as falhas aleatórias ou provenientes de ataques maliciosos e, ainda assim permanecer com um nível de serviço especificado.

O primeiro passo na obtenção do Indicador proposto é simular valores para uma variável aleatória $Y \sim Binomial\,(n,p)$, (Tabela 1), onde n é o número de retiradas de arestas em cada teste, e y é a probabilidade de sucesso. Essa etapa é necessária para se obterem os valores de y, o número de componentes que sobreviveram após o experimento, e do limite inferior de confiança de $(1-\propto)\,100\%$, da Confiabilidade, R_i. Após calculados os valores de y, determinam-se os valores da distribuição F de Snedecor, com nível de significância de $\propto = 0{,}5$ e $2(n-y+1)$ e $2y$, graus de liberdade (Tabela 1).

Após o cálculo dos R_i, estabelece-se o Requisito Mínimo de Confiabilidade (RMC), para os quais foram supostos os valores de 0,18, 0,22, 0,25 e 0,30, em todas as topologias (Tabela 2). Numa situação real, o RMC deverá ser determinado tomando-se por base os valores de confiabilidade reais para cada meio de transmissão de dados, que podem ser obtidos com os fornecedores.

Tendo-se realizado essa etapa, retiram-se as arestas do grafo cuja confiabilidade seja inferior ao RMC; no exemplo em questão, para o RMC de 0,18, foram retiradas as arestas 4-5, 4-10, 5-6, 5-10, 6-11, 9-10, 12-13 e 16-17. Calcula-se então, o número de nós remanescentes no maior componente conexo (LCC) e a quantidade de componentes conexas, possibilitando-se a obtenção do valor do Indicador de Resiliência proposto (IR), cujos valores encontram-se na Tabela (2).

Tabela 1 - Cálculo dos valores de y (número de componentes que sobreviveram após o experimento – e estimação dos R_i para cada aresta

Aresta	y	2y	2(n-y + 1)	F	Ri
1-2	6	12	10	2,75	0,30
2-6	6	12	10	2,75	0,30
3-4	8	16	6	2,74	0,49
4-5	3	6	16	3,92	0,09
4-10	4	8	14	44	0,15
5-6	4	8	14	44	0,15
5-10	4	8	14	44	0,15
6-7	7	14	8	2,70	0,39
6-11	4	8	14	44	0,15
6-12	5	10	12	2,91	0,22
7-8	8	16	6	2,74	0,49
9-10	4	8	14	44	0,15
10-11	7	14	8	2,70	0,39
10-14	5	10	12	2,91	0,22
11-12	5	10	12	2,91	0,22
11-16	8	16	6	2,74	0,49
12-13	4	8	14	44	0,15
12-17	6	12	10	2,75	0,30
12-19	5	10	12	2,91	0,22
14-15	7	14	8	2,70	0,39
14-16	6	12	10	2,75	0,30
16-17	4	8	14	44	0,15
17-18	5	10	12	2,91	0,22
19-20	6	12	10	2,75	0,30

Fonte: Autores

Tabela 2 - Valores do Indicador de Resiliência proposto (IR) de acordo com o Requisito Mínimo de Confiabilidade (RMC) proposto

RMC	IR
0,18	0,15
0,18	0,15
0,25	0,05
0,30	0,05

Fonte: Autores

Há uma a relação inversa entre os dois indicadores, ou seja, à medida em que a exigência por maiores valores de confiabilidade dos *links* aumenta, a resiliência diminui, apesar de ainda ser capaz de operar durante um intervalo de tempo, uma vez que a estimação indica um patamar mínimo de confiabilidade para a rede. Tal resultado indica que de acordo com o tipo de *link*, ou seja, o meio de transmissão dos dados, que pode ser via fibra ótica, rádio, cabo de par trançado etc., a resiliência irá variar, e isso condiz com a realidade, porque os meios de transmissão têm diferentes níveis de confiabilidade. A decisão pela melhor escolha compete ao gerente de redes no momento do planejamento da nova topologia física.

Com a finalidade de validar o Indicador proposto, a metodologia foi aplicada em cinco topologias, com diferentes números de nós, enlaces e redundâncias. Além disso, foram utilizadas métricas propostas na literatura para avaliação da robustez da rede frente às possíveis alterações, apenas em uma das topologias o IR não superou tais métricas.

Estudaram-se as topologias de rede reais REANNZ (Karen), RedClara, RedIRIS, GÉANT e RNP para verificar a influência da estrutura topológica da rede e da confiabilidade dos componentes no desempenho do sistema.

A rede REANNZ (*Research and Education Advanced Network New Zealand*), mais conhecida como KAREN, é uma rede de alta capacidade, sem fins lucrativos, que promove a pesquisa, educação e inovação na Nova Zelândia. Apresenta, em sua topologia atual, 22 nós e 26 enlaces.

A segunda topologia a ser estudada é a da RedCLARA, rede de Cooperação Latino Americana de Redes Avançadas, sem fins lucrativos, visando também, a promoção da pesquisa, da inovação e da educação. Sua configuração consiste de 19 nós e 21 enlaces.

A RedIRIS é a rede acadêmica e de investigação espanhola, proporcionando serviços avançados de comunicações à comunidade científica e universitária nacional. Sua topologia conta com 45 nós e 59 enlaces.

A premiada rede GÉANT, rede Pan-Europeia de pesquisa e educação, considerada a mais avançada e bem conectada do mundo, oferece uma capacidade de até 2Tbps, além de interconectar diversos países. A GÉANT interconecta as redes de Educação e Pesquisa Nacionais da Europa (NRENs) (*Europe's National Research and Education Networks*). Sua topologia tem 41 nós e 59 enlaces.

A RNP é responsável pelo funcionamento da rede Ipê, uma infraestrutura de rede Internet dedicada à comunidade brasileira de ensino superior e pesquisa, que interconecta universidades e seus hospitais, institutos de pesquisa e instituições culturais. Atualmente, a rede Ipê tem uma topologia composta por 28 nós e 38 enlaces.

4. Resultados

Os resultados obtidos para o Indicador de Resiliência para as cinco topologias reais estão na Tabela 3.

Tabela 3 - Valores do Indicador de Resiliência (IR) para quatro valores do (RMC) das cinco topologias

RMC	REANNZ (KAREN)	RedCLARA	RedIRIS	GÉANT	RNP
0,18	0,129	0,114	0,107	0,082	0,131
0,22	0,129	0,114	0,107	0,082	0,131
0,25	0,045	0,053	0,022	0,024	0,036
0,30	0,045	0,053	0,022	0,024	0,036

Fonte: Autores

Observando-se os resultados da Tabela 3, inicialmente poder-se-ia concluir que, de todas as topologias, a RNP seria a mais resiliente, porque inicia os testes com um valor de IR = 0,131, isto é, o máximo entre as demais. Contudo, há uma diminuição de 72,52% (Tabela 4) no seu valor de IR que, em relação às outras topologias, foi uma das maiores reduções, quando se aumenta o RMC para 0,25. Enquanto que a rede KAREN sofre redução de 65% no IR, a RedCLARA de 54%, as redes GÉANT e RedIRIS têm seu valor de IR decrescido em 70% e 79%, respectivamente. Conclui-se que a RedCLARA é a mais resiliente.

Tabela 4 - Valores percentuais de redução do Indicador de Resiliência (IR) das cinco topologias

REANNZ (KAREN)	RedCLARA	RedIRIS	GÉANT	RNP
65%	54%	79%	70%	72,52%

Fonte: Autores

4.1 Análise Comparativa entre as métricas

Com o objetivo de validar o Indicador de Resiliência proposto, utilizaram-se, neste artigo, as métricas *Largest Connected Component* (LCC), Diâmetro do LCC e Average *Inverse Shortest Path Length* (AISPL), estudadas em (BEYGELZIMER, 2005), para medir-se a robustez de uma topologia após alterações provocadas por ataques cibernéticos ou falhas aleatórias.

O maior componente conexo é o maior subgrafo que resulta da desconexão da topologia após a ocorrência de falhas ou ataques. A proporção de nós no LCC é uma medida utilizada em (BEYGELZIMER, 2005) para se avaliar a robustez de uma rede, ou ainda, um indicador da manutenção de sua disponibilidade máxima, após a ocorrência de falhas ou ataques. Entretanto, essa métrica não indica quantas componentes se formaram após a desconexão; por esse motivo, não deveria ser utilizada em separado de outras métricas, fato que será comprovado na análise comparativa com o Indicador proposto.

O Diâmetro do LCC (BEYGELZIMER, 2005) representa o maior comprimento de menor caminho do maior componente conexo.

O *Average Inverse Shortest Path Length* (AISPL) (BEYGELZIMER, 2005) é o comprimento médio inverso dos menores caminhos.

A redução acentuada no Indicador proposto, ao aumentar-se o RMC, não foi observada na medida proporção de nós no LCC (%LCC), conforme mostra a Tabela 5.

Tabela 5 - Valores do Indicador de Resiliência proposto (IR) e (%LCC) das cinco topologias

Topologias	KAREN		RedCLARA		RedIRIS		GÉANT		RNP	
RMC	IR	%LCC	IR	%LCC	IR	%LCC	IR	%LCC	IR	%LCC
0,18	0,129	0,77	0,114	0,68	0,107	0,53	0,082	0,66	0,131	0,79
0,22	0,129	0,77	0,114	0,68	0,107	0,53	0,082	0,66	0,131	0,79
0,25	0,045	0,45	0,053	0,63	0,022	0,51	0,024	0,34	0,036	0,5
0,30	0,045	0,45	0,053	0,63	0,022	0,51	0,024	0,34	0,036	0,5

Fonte: Autores

O Indicador proposto teve uma redução de 65% na rede KAREN, Tabela 5, 58.5% maior do que a observada na métrica proporção de nós no LCC; na RedCLARA, o Indicador proposto sofreu uma redução de 54%, tendo sido 635% maior do que a apresentada pelo % LCC; para a rede RNP, o Indicador proposto apontou uma redução de 72,52%, 98% maior do que a do % LCC.

Fazendo a mesma análise para as topologias RedIRIS e a GÉANT (Tabela 6), observou-se o mesmo comportamento entre o Indicador proposto e o %LCC, ou seja, na RedIRIS houve uma redução de 79% no IR, ao passo que o %LCC reduziu em 3,8%; a rede GÉANT teve uma redução de 70% no IR, enquanto que o %LCC foi reduzido de 48,5%.

Tabela 6 - Valores percentuais de redução do Indicador de Resiliência proposto (IR) e (% LCC) das cinco topologias

KAREN		RedCLARA		RedIRIS		GÉANT		RNP	
IR	%LCC	IR	%LCC	IR	%LCC	IR	%LCC	IR	%LCC
65%	41%	54%	7,35%	79%	3,8%	70%	48,5%	72,52%	36,7%

Fonte: Autores

Finalizando a validação do Indicador proposto, encontram-se na Tabela 7 os resultados das métricas: AISPL e Diâmetro do LCC (D).

A rede KAREN percebeu queda de 65% no AISPL e de 60% no Diâmetro do LCC, quando a exigência por confiabilidade foi aumentada, enquanto que o IR decresceu 65%; na RedClara a redução do AISPL foi de 66%, do Diâmetro foi de 42,85%, e do IR foi de 54%; na RedIRIS o AISPL teve redução de 76%, o Diâmetro de 46,15%, e o IR de 79%; na rede GÉANT o AISPL reduziu em 53%, o Diâmetro em 22.22%, e o IR em 70%. Finalmente, na rede RNP o AISPL percebeu redução de 63%, o Diâmetro de 22.22% e o IR de 72,52%.

Tabela 7 - Valores do Indicador de Resiliência proposto (IR), AISPL e Diâmetro do LCC (D) das cinco topologias

Topologias	KAREN		RedCLARA		RedIRIS		GÉANT		RNP	
RMC	AISPL	D	AISPL	D	AISPL	D	AISPL	D	AISPL	D
0,18	0,20	10	0,19	7	0,17	13	0,15	9	0,22	9
0,25	0,07	4	0,064	4	0,04	7	0,07	7	0,08	7

Fonte: Autores

O Indicador proposto (IR) foi mais preciso em identificar as alterações das topologias, exceto para a RedClara, porque quanto mais se aumentou o Requisito Mínimo de Confiabilidade (RMC), mais acentuada foi a redução do IR, provando que as demais métricas não foram capazes de perceber tanto quanto o IR as alterações nas topologias. Porque, em sua constituição, o (IR) considera a quantidade de componentes conexas após a desconexão do grafo, portanto quanto maior for o Requisito Mínimo de Confiabilidade (RMC) dos *links,* mais subgrafos surgirão após a desconexão.

5. Considerações Finais e Conclusões

O tema Resiliência em Redes de Computadores é um dos assuntos de relevância no campo de sistemas tolerantes a falhas, sistemas críticos e aplicações em tempo real, em crescente expansão na atualidade. A necessidade de se otimizarem os investimentos em equipamentos e *links* com maior confiabilidade que compõem as redes motivou o presente estudo, resultando em uma nova metodologia como parte do desenvolvimento de um

projeto de rede capaz de sobreviver às mais diversas ameaças, além da obtenção de um indicador de resiliência capaz de identificar alterações ocorridas em diversas topologias.

Esse Indicador de Resiliência demonstrou ser eficiente na mensuração da robustez após alterações efetuadas na topologia, ao retirarem-se os enlaces mais propensos a falhar, isto é, com menor confiabilidade.

Para a validação da presente metodologia, utilizaram-se cinco topologias reais de apoio às pesquisas científicas, das quais foram retirados os enlaces que apresentaram menor confiabilidade, mediante um Requisito Mínimo de Confiabilidade (RMC), seguindo-se os cálculos do Indicador de Resiliência (IR), evidenciando-se a sua consistência. Além disso, para avaliação da robustez da rede frente às possíveis alterações, foram utilizadas métricas propostas na literatura, e somente em uma das topologias o (IR) não superou tais métricas.

Existe uma grande dificuldade de obtenção de dados relativos às falhas dos enlaces, por isso neste artigo utilizou-se da técnica de simulação dos dados. As falhas ocorrem de diversas formas, de acordo com a tecnologia de transmissão, além de outros fatores, que somente os administradores das redes poderiam citar, devido à sua experiência nos projetos e implantações de redes, como por exemplo, o evento de roubo de enlaces, um tema recorrente.

Minha proposta seria a de juntarem-se os profissionais de pesquisa e de operações, para pensarem em um projeto que melhor pudesse abarcar as possibilidades de eventos geradores de falhas, e daí proporem um modelo de probabilidades mais real, uma vez que nos estudos atuais, têm-se focado em um ou poucos eventos de falhas, o que não retrata a realidade das ocorrências nas instituições.

Referências bibliográficas

BACELAR, L. [E-mail] 30 mai. 2014, Rio de Janeiro [para] KAPUR et al., USA. 1f. Re: Ireson Handbook of Reliability Engineering and Management.

BAZOVSKY, I., Reliability Theory and Practice, Dover Publications, 2004, 304 p.

BEYGELZIMER, A., GRINSTEIN, G., LINSKER, R. and RISH, I. Improving Network Robustness by Edge Modification. *Physica A: Statistical Mechanics and its Applications*, 357(3):593-612, 2005.

CHWIF, L. e MEDINA, A. C. Modelagem e Simulação de Eventos Discretos. 2006.

ELSAYED, E., Reliability Engineering, 2nd ed., John Wiley and Sons, 2012, 792 p.

GROSH , D. L. Primer of Reliability Theory Wiley (1989) IEEE, págs. 156-162. IEEE, 2009.

IRESON, W. G., COOMBS, C. F. e MOSS, R. Y. Handbook of Reliability Engineering and Management. McGraw-Hill New York, 1988.

KAPUR, K.C. and LAMBERSON, L.R. Reliability in Engineering Design Wiley 1977.

LEE, H. e KIM, J. Attack Resiliency of Network Topologies. *Em Parallel and Distributed Computing*: *Applications and Technologies*, págs. 638-641. Springer, 2005.

LEWIS, E. E. Introduction to Reliability Engineering, Wiley; 1995.

LOCKS , M. O. Reliability, Maintainability, and Availability Assessment, 2nd ed., Amer Society for Quality, 1995.

MARKOPOULOU, A., IANNACCONE, G., BHATTACHARYYA, S., CHUAH, C.-N. e DIOT, C. Characterization of Failures in an IP Backbone. Em INFOCOM 2004. *Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 4, págs. 2307-2317. IEEE, 2004.

MENTH, M., DUELLI, M., MARTIN, R. e MILBRANDT, J. Resilience Analysis ofPacket-Switched Communication Networks. *IEEE/ACM Transactions on Networking (TON),* 17(6):1950-1963, 2009.

NACHLAS , J. A. Reliability Engineering: Probabilistic Models and Maintenance Methods CRC Press 2005.

NBR 5462 Confiabilidade e Mantenabilidade-Terminologia. Rio de Janeiro: Associação Brasileira de Normas Técnicas, 1994.

O´CONNOR, P.P. e KLEINER, A., Practical Reliability Engineering, 5th ed., John Wiley and Sons, 2012, 512p.

OMER, M., NILCHIANI, R. e MOSTASHARI, A. Measuring the Resilience of the Global Internet Infrastructure System. Em Systems Conference, 2009 3rd Annual IEEE, págs. 156-162. IEEE, 2009.

Pham, H. Reliability Modeling, Analysis and Optimization, World Scientific Publishing: Singapore 2006.

SHOOMAN, M.L. Reliability of computer Systems and Networks: Fault Tolerance, Analysis, and Design. John Wiley & Sons, 2002

STERBENZ, J. P., HUTCHISON, D., ÇETINKAYA, E. K., JABBAR, A., ROHRER, J. P., SCHÖLLER, M. and SMITH, P. Resilience and Survivability in Communication Networks: Strategies, Principles, and Survey of Disciplines. *Computer Networks*. 54(8): 1245-1265, 2010.

SZWARCFITER, J. L. Grafos e Algoritmos Computacionais, volume 2. Campus, 1988.

TOBIAS, P. A., Applied Reliability, 3rd Edition, Chapman and Hall/CRC, 2011, 600 p.

Abstract

The increasing number of applications running on the internet has motivated the study of resilience in computer networks. Researchers and experts aims the standardization of metrics for correct measurement of the resilience level in networks, and this article proposes an indicator of resilience pursuing a good infrastructure from the beginning of the project. This project will be made viable using reliability theory, whose objective is to increase the reliability of network components, reducing costs of maintenance and redundancies, resulting in savings for the project. The proposed indicator was tested in backbone topologies, indicating good results, which allows network managers to use this methodology in the proper planning, in addition to the recovery of any fault incidents. The use of more trusted links will provide a structure better prepared to oppose the threats, and to foment the gain of the quality of services provided by the network, and to comply with Service Level Agreements.

Estimación de la prevalencia por método RDS en poblaciones de tamaño pequeño

Juan José Orellana-Cáceres [1]
Sergio Raúl Muñoz-Navarro [2]
Alex Manuel Antequeda-Campos [3]

Resumen

Introducción: Respondent-driven sampling (RDS) es un método de muestreo para poblaciones "ocultas". Utiliza cadenas referenciales desde sujetos "semillas", usando un sistema de "cupones" para contactar nuevos sujetos elegibles. El objetivo de este trabajo es proponer un método para establecer el número de semillas/cupones en muestreo RDS en poblaciones pequeñas (~1000) para la estimación de prevalencia. Simulaciones computacionales de una población oculta de tamaño pequeño y validación del método. Se obtuvo la distribución muestral del tamaño de muestra para diferentes combinaciones de número de semillas y cupones. La utilización de un número bajo de semillas y cupones muestra ser inadecuado en poblaciones de tamaño pequeño. El método propuesto es válido, obteniéndose estimaciones eficientes de prevalencia. Se debe estudiar la robustez del método ante desviaciones a los supuestos de RDS. Se entrega a la comunidad científica, recomendaciones para establecer el número de semillas/cupones para muestreos RDS para la estimación de prevalencia en poblaciones pequena.

[1]Académico Departamento de Salud Pública y miembro del Centro de Excelencia CIGES de la Universidad de La Frontera y Estudiante Programa de Doctorado en Salud Pública de la Universidad de Chile. E-mail: juan.orellana@ufrontera.cl.
[2] Académico Departamento de Salud Pública y miembro del Centro de Excelencia CIGES de la Universidad de La Frontera
[3] Ingeniero Civil Matemático, Departamento de Matemática y Estadística, Facultad de Ingeniería y Ciencias, Universidad de La Frontera.

1. Introducción

Respondent-driven sampling (RDS) permite estimar promedios y proporciones em poblaciones de difícil acceso donde los métodos de muestro tradicionales son inviables o presentan sesgo de selección. Es um método que combina el método de "Bola de nieve" (al hacer que los individuos recluten a otros desde su red de contactos en la población) y un modelo que proporciona estimaciones insesgadas de los parámetros, corrigiendo así los problema de sobre muestreo y representatividad de la muestra obtenida por Bola de nieve. RDS fue propuesto por Heckathorn en la década de 90', ha demostrado ser especialmente útil para la vigilancia del comportamiento del VIH (Virus de la Inmunodeficiencia Humana) siendo frecuentemente empleado en estudios auspiciados por el Centro para el Control y Prevención de Enfermedades, EEUU (Heckathorn et al., 2002; Schonlau & Liebau, 2012; Zhang et al., 2012).

Con este método, el proceso de muestreo se inicia identificando un número determinado de sujetos accesibles, llamados "semillas", quienes pertenecen a la población objetivo. Estas semillas participan en el muestreo reclutando personas que cumplen con los criterios de elegibilidad para la población objetivo, desde su red de contacto. Este proceso se realiza mediante un sistema de cupones, cuya cantidad define el número máximo de contactos que puede reclutar desde el total de su red de contactos elegibles. Los sujetos contactados que aceptan participar, reclutan nuevos sujetos elegibles desde su propia red de contacto, usando el mismo sistema de cupones. Cada uno de estos procesos genera el llamado nivel referencial u "ola"; de este modo las semillas corresponden a la ola 0 y sus contactos corresponden a la ola 1, y así sucesivamente hasta el final del muestreo. Cada contacto es llamado "nodo" los que pueden ser representados gráficamente en un árbol de relaciones.

RDS crea un modelo de expansión basado en cadenas referenciales. Cada contacto debe informar el tamaño de su red, llamado "grado". El muestreo termina cuando se alcanza el tamaño de muestra predeterminado, no se encuentran más contactos, o se ha alcanzado la estabilidad en las estimaciones con una precisión razonable.

El método exige que cada persona tenga una doble participación, primero como sujeto que aporta datos al estudio y segundo como reclutador de nuevos elegibles. En cada forma de participación, el otorgamiento de un incentivo es clave para lograr la minimización del rechazo a participar (Heckathorn et al., 2002). El reclutamiento debe ser conducido por cada entrevistado y no por el entrevistador (Respondent-driven sampling). El papel del entrevistador es verificar que los entrevistados que posean el cupón de invitación, sean elegibles, que no estén repetidos por doble invitación o estar simulando diferentes identidades (Heckathorn et al., 2001), entregar los incentivos correspondientes y gestionar la aplicación de los cuestionarios del estudio a cada participante.

La determinación del tamaño de muestra, cuyo diseño es complejo, es otro aspecto a considerar en el diseño de un estudio con método RDS. Una alternativa propuesta en la literatura es ajustar por efecto de diseño no menor a 2, al cálculo obtenido asumiendo muestreo aleatorio simple (Wejnert et al., 2012).

Para más detalles del funcionamiento de RDS y de sus propiedades estadísticas sobre el control de sesgo, definición del tamaño de muestra y estimación de parámetros, se recomiendan los papers: Heckathorn et al. (2001), Heckathorn et al, (2002) Wejnert (2009), y Schonlau & Liebau, (2012).

Los softwares Stata, SPSS y R entre otros, han implementado los principales métodos estadísticos para el análisis de datos RDS. El software Netdraw (Harvard:Technologies & Borgatti, 2002), permite representar gráficamente las redes de contactos de la muestra RDS obtenida. RDSAT 7.1 es un software libre diseñado exclusivamente para el análisis de datos RDS (Volz et al., 2012).

En adición a los fundamentos teóricos de RDS, se han implementado estrategias analíticas mediante simulación computacional para estudiar el desempeño de los estimadores propuestos. Los resultados obtenidos confirman que, cuando los supuestos del método se cumplen, los estimadores son eficientes asintóticamente insesgados (Salganik & Heckathorn, 2004). Se ha demostrado, también por simulaciones computacionales, que el método de estimación es muy sensible a la validez del tamaño de la red que reportan los entrevistados. La preferencia por los dígitos 0 y 5 genera un considerable sesgo en la estimación de la prevalencia (Mills et al., 2014). En ambos estudios el desempeño de RDS fue evaluado en poblaciones relativamente grandes (10.000 sujetos).

No se encontraron estudios que evalúen el desempeño del método RDS, en poblaciones pequeñas como puede ocurrir cuando se acota la población geográficamente y a condiciones técnicas específicas; como por ejemplo, la población de micro y pequeñas empresas de una región determinada, donde se desea estimar la proporción de empresas sin iniciación formal de actividades. Se prevé que en estas situaciones la muestra deberá cubrir una gran proporción de la población, por lo cual la elección del número de semillas y cupones es un elemento clave y aun no discutido.

2. Objetivo

En el contexto de una población oculta de tamaño pequeño, el objetivo de este artículo es proponer un método para la selección del número de semillas y cupones en función del tamaño de muestra inicial que se desea alcanzar. Adicionalmente se mostrará el desempeño de RDS y de los métodos estadísticos disponibles para la estimación de una proporción.

3. Metodo

Simulación computacional de una población oculta de tamaño pequeño y de muestreo RDS sobre dicha población. Adicionalmente se simula una nueva población de tamaño pequeño para la evaluación del método de muestreo y estimación de prevalencia.

En una primera etapa se simula una población oculta de tamaño pequeño, con una prevalencia predeterminada de una característica de interés, asociada al tamaño de la red o grado de cada sujeto. De dicha población, se toman repetidas muestras irrestrictas RDS con las que se construye una distribución de frecuencias de tamaño de muestra máximo alcanzado en cada repetición, según el número de semillas y cupones.

A partir de la distribución de muestreo del tamaño de muestra alcanzado, se establecen criterios para la determinación del número de semillas y de cupones para lograr el tamaño de muestra especificado en el diseño del estudio.

En una segunda etapa, se ejecutan repetidas simulaciones de muestreo RDS en una población adicional sobre las cuales se validará la recomendación sobre número de semillas y cupones, y se verificarán los resultados teóricos del método en relación a las propiedades de los estimadores.

La simulación se inicia definiendo el número de semillas y cupones para alcanzar el tamaño de muestra predeterminado.

El algoritmo computacional parte identificando todos los pares de semillas que son incapaces de generar árboles referenciales. Dichos pares forman un vector que contiene todos los nodos terminales con su respectivo contacto referencial, y todos los nodos terminales que compartan un mismo referente. Esta restricción se traduce en RDS a que, no serían elegibles como semillas, pares sujetos donde uno de ellos solo conoce a su par, o pares de sujetos que solo conocen a un mismo y único referente.

La simulación prosigue seleccionando en forma aleatoria y sin reemplazo un número predefinido de semillas elegibles. Se selecciona al azar una de dichas semillas y uno de sus contactos directos, luego escoge la segunda semilla y le asigna al azar uno de sus contactos directos elegibles (no usados por la otra semilla), así hasta la última semilla. El proceso anterior se repite idénticamente hasta que cada semilla alcance un máximo c' de contactos ($c' \leq c$), donde c es el número de cupones especificado) impidiendo repeticiones con los contactos de otra semilla ya utilizado. Todos los contactos asignados en la etapa anterior, pasan a ser las semillas en una siguiente etapa, que repite el algoritmo de las semillas iniciales. Este proceso se repite hasta alcanzar el tamaño de muestra. En cada muestra RDS se registra para cada sujeto, el tamaño de su red de contacto, que se le asignó en la construcción de la población, la condición en estudio y sus contactos directos seleccionados según el número de cupones.

Finalmente, usando los comandos rds_network, rds y svy asociados al análisis RDS en Stata 13 ("StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP," n.d.), se obtienen el tamaño de muestra alcanzado, la prevalencia estimada con su correspondiente intervalo de confianza.

Para el estudio de la distribución muestral del tamaño de muestra máximo alcanzado, la simulación usa una subrutina para 200 muestreos RDS, sin restricción en el tamaño de muestra, y para todas las combinaciones entre número de semillas (5, 6, 9, 12, 15, 20, 30 y 35) y de cupones (3, 4, 5, 6 y 7). En cada simulación se obtiene la distribución de frecuencia del tamaño de muestra y se registran los percentiles 5, 10 y 15 de tamaño de muestra alcanzado como indicadores de probabilidad de alcanzar un tamaño de muestra específico (1-percentil).

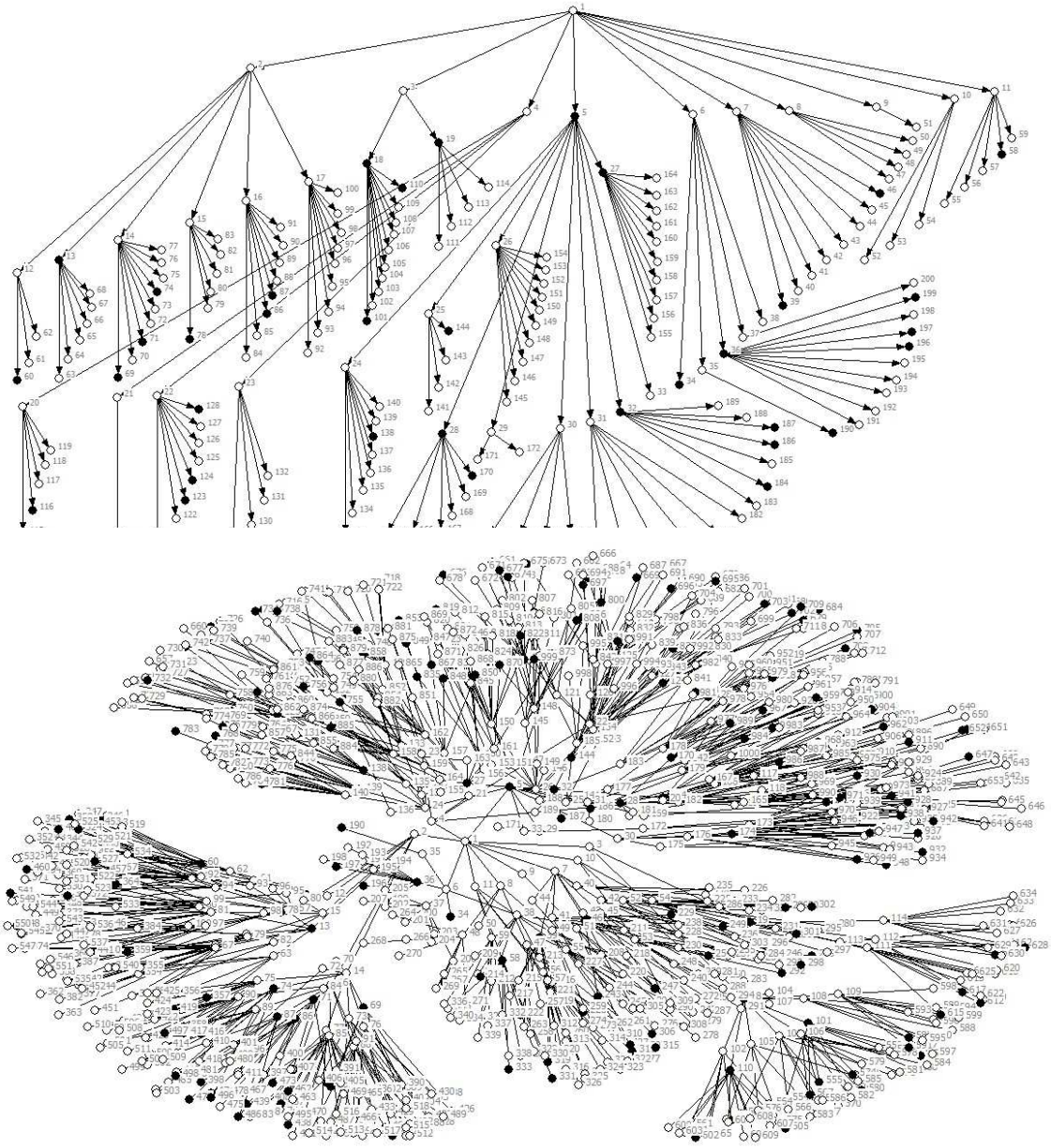
La validación del método propuesto para la determinación del número de semillas y cupones en el contexto de una población pequeña, se ensaya sobre una nueva población oculta con la misma estructura de la población anterior, pero con diferente esquema de aleatorización. Se obtienen 200 muestras RDS para distintos tamaños de muestra según tres efectos de diseño (1, 2 y 3) y dos niveles de precisión relativa en las estimaciones (20 y 30%). Como resultados, se obtiene el tamaño de muestra promedio alcanzado en las 200 muestras RDS, la proporción de muestras que logró el tamaño calculado en el diseño, el promedio de prevalencia estimada, la proporción de intervalos de confianza que contienen el valor del parámetro, promedio de precisión observada y promedio de sesgo absoluto y relativo en la estimación de la prevalencia.

4. Resultados

Usando códigos en Stata 13, se simuló una población oculta de tamaño 1.000 individuos, representada por un árbol de relaciones (Figura 1). El grado de cada persona, sigue una distribución uniforme con rango 0 – 10 (media 5 y varianza 10). Se fija una prevalencia poblacional de 20,8%. La población se estratifica en dos grupos según el grado de los sujetos, el estrato "populares" queda compuesto con sujetos cuyo grado es igual o superior a 5, y el estrato "impopulares" definido por su complemento. Se asume que la prevalencia poblacional en el estrato "populares" es de 31,2% y un 10,4% en el estrato "impopular". El link correspondiente, contiene los códigos Stata que generan la población y el archivo con los datos poblacionales para NetDraw.

En este link se presentan los códigos para la selección de las muestras RDS desde el árbol poblacional. Este archivo contiene las rutinas de selección de semillas, de muestreo RDS, y de análisis de datos.

Figura 1 - **Primeros 200 nodos del total de 1000 que componen el árbol poblacional. Más abajo el árbol referencial completo.**



● Con la condición en estudio
○ Sin la condición en estudio

Distribución muestral para el tamaño de muestra máximo alcanzado.

Los resultados del estudio para determinar el tamaño de muestra en muestreo RDS en poblaciones pequeñas, se presentan en la Tabla 1.

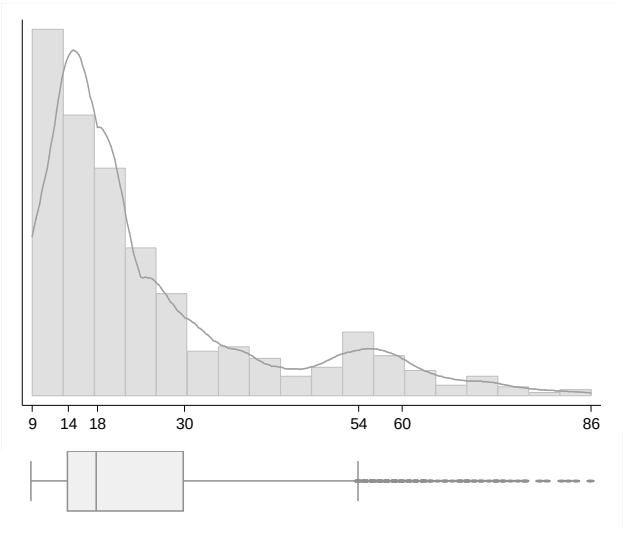
Tabla 1 - Percentiles 5, 10 y 15 de la distribución del tamaño de muestra máximo alcanzado en muestreos RDS para la combinación de semillas y cupones.

Percentiles p5, p10, p15		N° de cupones				
		3c	4c	5c	6c	7c
Semillas	5s	36, 44, 48	62, 88, 209	308, 416, 452	573, 626, 633	746, 756, 758
	6s	41, 53, 62	115, 241, 293	428, 460, 473	632, 651, 655	755, 761, 767
	9s	81, 102, 133	296, 314, 337	460, 500, 518	651, 661, 670	763, 773, 777
	12s	118, 151, 192	358, 378, 399	512, 537, 548	663, 674, 682	770, 784, 793
	15s	199, 227, 241	376, 405, 417	551, 564, 570	686, 693, 696	787, 797, 803
	20s	250, 267, 273	419, 439, 454	583, 591, 596	708, 717, 719	812, 815, 825
	30s	327, 341, 346	401, 502, 508	624, 637, 640	743, 751, 756	845, 851, 856
	35s	361,369, 374	508, 515, 520	646, 656,661	766, 770, 774	860, 865, 869
	1s y todos los cupones					999

La tabla anterior orienta en la selección del número de semillas y cupones necesarios para realizar el muestreo RDS en una población pequeña. Así por ejemplo, si se requiere realizar un muestreo RDS para un tamaño de muestra 182 (producto del cálculo de tamaño de muestra para muestreo aleatorio simple, con una prevalencia esperada de 10%, precisión relativa de 30%, efecto de diseño de $2\,y$ un nivel de confianza de 95%); se recomienda usar 15 semillas y 3 cupones pues se espera una probabilidad de alcanzar el tamaño de muestra $[P(n > p5)]$ de 95%; o bien 6 semillas y 4 cupones pues se espera alcanzar el tamaño de muestra con probabilidad de 90% $[P(n > p10 = 90\%)]$.

La utilización de un número bajo de semillas y cupones muestra ser inadecuado en poblaciones de tamaño pequeño ya que el tamaño de la muestra alcanzado tiene una alta probabilidad quedar por debajo del determinado en el diseño del estudio. Así por ejemplo, en la población en estudio, el 75% de las muestras RDS con tres semillas y dos cupones no superan los 30 sujetos; solo en uno de los muestreos RDS, el árbol muestral queda compuesto por 86 sujetos (Figura 2).

Figura 2 - Distribución de frecuencia, gráfico kernel-density y de caja, del tamaño de muestra en 1000 muestreos RDS desde la población de referencia con tres semillas y dos cupones.



Desempeño del método en la estimación de la prevalencia en una población pequena.

A continuación se presenta el resultado de la estimación de la prevalencia poblacional al seleccionar 500 muestras RDS desde una nueva población oculta con similar estructura de la población anterior, pero con distinta esquema de aleatorización. Se ensayarán 12 situaciones de muestreo RDS, según, prevalencia esperada $Po = 10\%\ y\ 30\%$, precisión relativa $pr = 30y20\%$, efecto diseño $ED = 1,2\ y\ 3\ y$ con un nivel de confianza en la estimación de 95%. La Tabla 2 y Figura 3, presenta los resultados.

Tabla 2 - Estimación proporción $(P = 0,203)$. Cálculo tamaño de muestra (n) para una prevalencia esperada $Po = 0,1\ y\ 0,3$, precisión relativa $pr = 20\ y\ 30\%$ y efecto de diseño $ED = 1,2\ y\ 3$. Simulación de 500 muestreos RDS usando un número de semillas y cupones según la Tabla 1, reportando: media de tamaño muestra alcanzado\proporción de muestras sobre el valor del diseño, promedio de prevalencia, proporción de intervalos de confianza que contienen el valor del parámetro, promedio precisión observada y promedio sesgo absoluto\relativo.

Po: Prev.esp. pr: Prec.relat. pa: Po·pr Δ	ED	n	RDS	N°semillas y cupones escogido	ni prom (DE)	% alcance muestra	pi prom (%)	% casos P∈ IC95%	Promedio	
									Precisión observada ⊥ (Exceso / Déficit)□	Sesgo absol.Ψ (relativo %)ϕ
Po=0,3	1	91	1	6s y 4c	90 (5,6)	97,0	20,0	95,8	8,2 (+0,8)	0,0 (+0,2)
pr=30%	2	182	2	15s y 3c	180 (8,1)	93,8	20,4	96,8	5,9 (+3,1)	0,4 (+1,8)
pa=9pp	3	273	3	9s y 4c	269 (21)	95,4	20,0	96,4	4,8 (+4,2)	0,0 (- 0,2)
Po=0,3	1	184	4	15s y 3c	183 (7,4)	95,4	20,4	97,4	5,8 (+0,2)	0,4 (+2,1)
pr=20%	2	368	5	15s y 4c	367 (7,4)	98,6	20,1	98,6	4,1 (+1,9)	0,1 (+0,7)
pa=6pp	3	552	6	5s y 6c	550 (27,1)	99,6	20,4	100,0	3,4 (+2,6)	0,4 (+1,8)
Po=0,1	1	278	7	9s y 4c	275 (15,8)	96,0	20,0	98,6	4,7 (-1,7)	0,0 (0,0)
pr=30%	2	556	8	15s y 5c	556 (3,2)	98,0	20,2	100,0	3,3 (-0,3)	0,2 (+1,2)
pa=3pp	3	834	9	30s y 7c	834 (2,2)	98,6	20,2	100,0	2,7 (+0,3)	0,2 (+1,1)
Po=0,1	1	464	10	9s y 5c	462 (19,5)	98,4	20,1	99,4	3,7 (-1,7)	0,1 (+0,7)
pr=20%	2	928	11	35s y 7c	882 (13)	0,0	20,2	100,0	2,6 (-0,6)	0,2 (+0,9)
pa=2pp	3	1392+	12	1s y 10c	999(0)	0,0	20,3	100,0	-	-

+ Equivalente al método de bola de nieve para cubrir toda la población, donde basta solo una semilla y todos los cupones.

Δ Prevalencia esperada (Po), precisión relativa (pr) y precisión absoluta (pa) en puntos porcentuales (pp).

⊥ Distancia media entre los extremos del IC (pp).

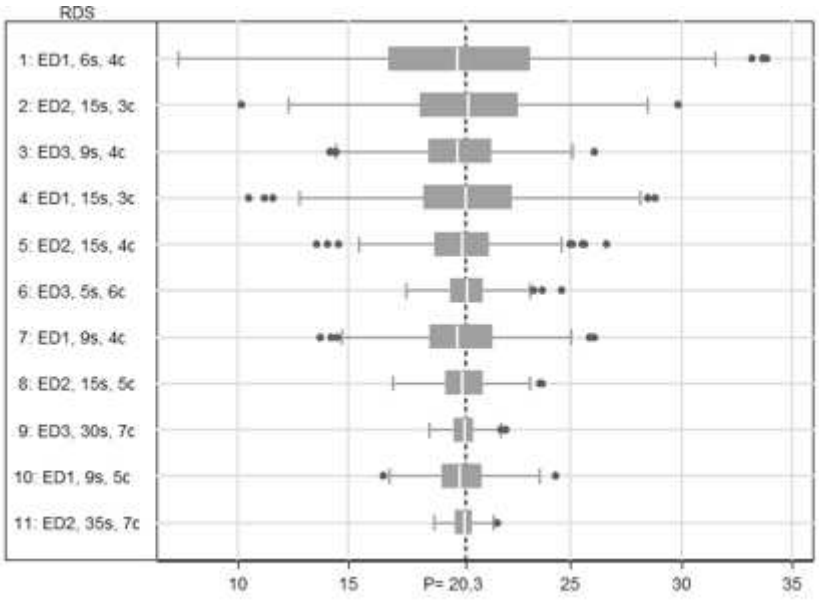
□ Precisión deseada en el diseño, menos la precisión observada (pp)

Ψ Desviación absoluta de la estimación puntual respecto del parámetro poblacional $\left(\hat{P} - P\right) 100 \quad pp$

ϕ Desviación relativa de la estimación puntual respecto del parámetro poblacional $\hat{P} - P)/P \cdot 100 \quad \%$

+ Equivalente al método de bola de nieve para cubrir toda la población, donde basta solo una semilla y todos los cupones.

Figura 3 - Distribución de muestreo en 500 réplicas RDS sobre una población relacional



Solo en el muestreo que se perseguía un tamaño de muestra muy cercano a la población 928 no se consiguió alcanzar el tamaño de muestra requerido, pero las muestras alcanzaron un promedio muy cercano a dicho valor 882. En el resto de los muestreos, se alcanzó un porcentaje de cumplimiento del tamaño de muestra que varió entre 93,8 y 99,6% (promedio 98,5%). En todos los muestreos RDS, las prevalencias estimadas fueron insesgadas con sesgos absolutos y relativos despreciables.

Al usar tamaños de muestra 91 y 184, se alcanzó el nivel de precisión (9 y 6 pp) y confianza (95%) estipulada en el diseño muestral, sin necesidad de usar un efecto de diseño mayor que 1. Cuando la presión absoluta es de $3\ y\ 2$ puntos porcentuales es preciso usar un efecto de diseño de $3y$ más para satisfacer los requerimientos del diseño muestral.

Un interesante resultado que se deriva de uno de los supuestos del método[1] es que, independiente de la semilla, cuando se utilizan todos los contactos de los sujetos y sin límites de tamaño de muestra, el resultado invariable es cubrir a toda la población de referencia. Esta situación es equivalente al método de bola de nieve que pretenda construir un marco muestral para diseños de muestreo tradicionales.

5. Discusión Y Conclusiones

El método propuesta para establecer el número de semillas y cupones, se comportó adecuadamente, sin grandes brechas entre el tamaño de muestra del diseño y el alcanzado en los muestreos RDS. Esta situación debiera ser estudiada ante variaciones de los supuestos del método, como por ejemplo el redondeo del tamaño de la red de cada persona, selección no aleatoria desde la red de cada sujeto o la presencia de valores perdidos.

El método de estimación mostró ser robusto a la estratificación de la población según el tamaño de la red, con distinto valor de prevalencia en cada estrato.

---

[1] Cada entrevistado puede ser alcanzado por cualquier outro através de uma serie de vínculos em red, es decir, la red forma um solo componente

La asignación del “grado” mediante la distribución uniforme, puede ser discutida por algunos autores aduciendo que normalmente sigue distribuciones asimétricas positivas, sin embargo creemos importante evaluar la situación de encontrar personas con igual probabilidad de tener redes de contacto pequeñas, medianas y grandes.

Es muy importante una buena selección de las semillas, dado que esto es determinante para conseguir la muestra deseada. El análisis de datos RDS elimina automáticamente toda semilla que no pueda referenciar al menos un contacto y se espera que dicha estrategia también se aplique en muestreos RDS reales.

Si bien se ha simulado una población de sujetos conectados en un único árbol como lo asume el método, es muy probable de existan casos de poblaciones con más de un árbol relacional, como es el caso de población de bandas de delincuentes o poblaciones ocultas estratificadas, por ejemplo en “tribus urbanas”. En estos casos, necesariamente se deben seleccionar semillas de cada estrato y aplicar el procedimiento de muestreo RDS anterior en cada uno.

Referências bibliográficas

Harvard:Technologies, & Borgatti, S. P. (2002). NetDraw: Graph Analytic, Visualization Software.

Heckathorn, D., Broadhead, R., & Sergeyev, B. (2001). A methodology for reducing respondent duplication and impersonation in samples of hidden populations. *Journal of Drug Issues*.

Heckathorn, D., Semaan, S., Broadhead, R., & Hughes, J. (2002). Extensions of Respondent-Driven Sampling : A New Approach to the Study of Injection Drug Users Aged 18 – 25. *AIDS and Behavior*, 6(1), 18–25. doi:10.1023/A:1014528612685

Mills, H. L., Johnson, S., Hickman, M., Jones, N. S., & Colijn, C. (2014). Errors in reported degrees and respondent driven sampling: Implications for bias. *Drug and Alcohol Dependence*, 142, 120–126. doi:10.1016/j.drugalcdep.2014.06.015

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193–239.

Schonlau, M., & Liebau, E. (2012). Respondent-driven sampling. *The Stata Journal*, 12(1), 72–93.

StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP. (n.d.). Retrieved from www.stata.com

Volz, E., Wejnert, C., Cameron, C., Spiller, M., Barash, V., Degani, I., & Heckathorn, D. . (2012). Respondent-Driven Sampling Analysis Tool (RDSAT). Ithaca, NY: Cornell University.

Wejnert, C. (2009). An empirical test of Respondent-Driven Sampling: Point estimates, variance, deree measures, and out-of-equilibrium data. *Sociological Methodology*, 39(1), 73–116. doi:10.1111/j.1467-9531.2009.01216.x.AN

Wejnert, C., Pham, H., Krishna, N., Le, B., & DiNenno, E. (2012). Estimating design effect and calculating sample size for Respondent-driven sampling studies of injection drug users in the United States. AIDS and Behavior, 16(4), 797–806. doi:10.1007/s10461-012-0147-8

Zhang, L., Ding, X., Lu, R., Feng, L., Li, X., Xiao, Y., … Qian, H.-Z. (2012). Predictors of HIV and syphilis among men who have sex with men in a Chinese metropolitan city: comparison of risks among students and non-students. PloS One, 7(5), e 37211. doi:10.1371/journal.pone.0037211

Abstract

Respondent-driven sampling (RDS) is a sampling procedure for "hidden" populations. Use referencial chains from subject "seeds", using a system of "coupons" to contact new eligible subjects. The objective of this work it to propose a method to set the number of seeds/coupons on RDS in small population (~1000) to estimate prevalence. Computational simulations of small hidden population and validation of the method. The sampling distribution of the sample size for different combinations of number of seeds and coupons was obtained. The use of a small number of seeds and coupons shown to be inadequate in small populations. The proposed method is valid, obtaining efficient prevalence estimates. It should study the robustness of the method to deviations from the assumptions of RDS. It is delivered to the scientific community the recommendations to set the number of seeds coupons for RDS samples to estimate prevalence in small populations.

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.seer.ibge.gov.br

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 2142-8762 (José André de Moura Filho – Editor–Executivo)

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais enviados para publicação devem obedecer às normas seguintes:

1. Devem ser submetidos originais processados pelo editor de texto *Word for Windows*;

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. 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;

4. 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;

5. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;

6. A citação de referências no texto e a listagem final de referências devem ser feitas de acordo com as normas da ABNT;

7. 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;

8. 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.

9. 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;

10. 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 Word) para o envio da versão corrigida; e

11. Como parte do processo de submissão, os autores são obrigados a verificar a conformidade da submissão em relação a todos os itens listados a seguir. As submissões que não estiverem de acordo com as normas serão devolvidas aos autores.

ISSN 0034-7175