Pesquisa Operacional para o Desenvolvimento
Vol. 20, e20001, p. 1-19, 2026

ISSN 1984-3534

AVALIANDO O DESEMPENHO DO MÉTODO SSA - CROSS VALIDATION COM USO DE ANÁLISE DE CLUSTER NA MODELAGEM E PREVISÃO DE VELOCIDADE DO VENTO[1]

Moisés Lima de Menezes[a] *, **Guilherme Cruvello da Silveira Martins**[b]

[a] Departamento de Estatística,
Universidade Federal Fluminense - UFF, Niterói-RJ, Braisl

[b] Escola Nacional de Saúde Pública
Fundação Oswaldo Cruz - FIOCRUZ, Rio de Janeiro - RJ, Brasil

Recebido 19/05/2025, aceito 17/11/2025

RESUMO

A demanda por energia elétrica aumenta com o crescimento da população. Neste cenário, a energia eólica é uma possível fonte complementar à energia hidrelétrica instalada no Brasil e a previsão de velocidade do vento tem papel fundamental no planejamento de estratégias. Este trabalho propõe avaliar modelagens de séries temporais sem e com a filtragem Singular Spectrum Analysis (SSA) e a divisão cross Validation. A primeira técnica busca reduzir o ruído da série e a segunda busca reestimar a série de forma iterativa com a adição de uma nova observação a cada instante de tempo. Para validação, medidas de desempenho são utilizadas. Os resultados apontam melhor ajuste dos modelos após aplicação do filtro SSA e melhor desempenho na previsão quando este filtro é utilizado associado ao *Cross Validation*. Contudo, há situações em que o método não se mostra eficaz, levando a uma reflexão sobre o uso desse recurso.

Palavras-chave: Singular Spectrum Analysis, Análises de Cluster, Velocidade do Vento, Cross Validation.

ABSTRACT

Electricity demand increases with population growth. In this scenario, wind energy is a possible complementary source to hydroelectric energy installed in Brazil and wind speed forecasting plays a fundamental role in planning strategies. This work proposes to evaluate time series modeling under Singular Spectrum Analysis (SSA) filtering and Cross Validation division. The first technique seeks to reduce the noise of series and the second one seeks to re-estimate the series iteratively by adding a new observation at each instant of time. To validation, performance measures are used. The results indicate better fit of the models after application of SSA filter and better forecast performance when this filter carried out in association with Cross Validation. However, there are situations in which the method is not effective, leading to reflection on the use of this resource.

Keywords: Singular Spectrum Analysis, Cluster Analysis, Wind Speed, Cross Validation.

* Autor para correspondência. E-mail: moises_lima@id.uff.br
DOI: https://doi.org/10.4322/PODes.2026.001

[1] *Todos os autores assumem a responsabilidade pelo conteúdo do artigo.*

1. Introdução

Visando atender à crescente demanda de energia elétrica provocada pelo aumento da população principalmente nos centros urbanos e pelo avanço da tecnologia, as autoridades responsáveis pelo controle do uso de energia buscam por alternativas que possam ser complementares às usinas hidrelétricas. Atualmente, as usinas termelétricas são utilizadas nos períodos de secas e baixa geração de energia hidroelétrica. Essa ação gera um custo elevado na geração de energia. Diante deste cenário, a energia eólica surge como uma alternativa complementar de energia limpa e renovável (Tomasquim, 2012).

Apesar dos avanços tecnológicos sobre a energia eólica, existe uma preocupação com a geração em larga escala devido à dependência do vento, de própria velocidade do vento e de outros fatores climáticos. Desta forma, estudos de previsão de energia de origem eólica de qualidade são necessários uma vez que a mesma não pode ser acumulada ou armazenada para momentos de escassez.

Uma forma de planejamento da oferta e demanda por energia é feita com o conhecimento prévio destas variáveis, que pode ser obtido a partir do uso da análise de séries temporais (Cardoso, 2005). Diversos métodos podem ser utilizados para melhorar o desempenho destas modelagens, dentre eles, a filtragem da série antes de sua modelagem. *Singular Spectrum Analysis* (SSA) é uma técnica que pode filtrar uma série temporal a partir da remoção de uma componente ruidosa (Menezes et al., 2014). As análises *in-sample* e *out-of-sample* são formas de avaliação do desempenho do ajuste do modelo. Nesta técnica, busca-se avaliar o poder preditivo do modelo na amostra de teste (*out-of-sample*) a partir do modelo ajustado na amostra de treinamento (*in-sample*) com os dados históricos. *Time Series Cross Validation* (doravante chamada apenas de *Cross Validation*) é uma abordagem que pode melhorar a acurácia do modelo em alguns casos. A partir desta técnica, as modelagens *in-sample* e *out-of-sample* são atualizadas a cada instante de tempo com o novo dado permitindo uma análise de forma simultânea e interativa (Hyndman e Athanasopoulos, 2018).

Dalmaz (2007) utilizou redes neurais para a previsão da velocidade do vento em diferentes regiões de Santa Catarina para investigar o potencial de geração de energia eólica. Como métrica de aderência da previsão estatística foi utilizado o $RMSE$ (*root mean squared error*), concluindo que o município de Água Doce, na região Centro-Oeste de Santa Catarina, possui alto pontencial para viabilização da construção de um parque eólico. Cardoso (2005) propôs investigar um modelo para previsão da volatilidade de séries de demanda de energia elétrica para consumidores livres. Para isso, foi aplicado o modelo GARCH para dados de curto prazo. As análises apresentaram resultados satisfatórios. Hyndman et al. (2018) compararam estatísticas de aderências tradicionais com técnicas modernas de estimação do erro provenientes da área de *machine learning*. Entre elas, inclui-se o médodo *K-Fold Cross Validation*, utilizado primordialmente em problemas de classificação, e o método *Time Series Cross Validation*, que é uma versão adaptada do método anterior para dados que apresentam autocorrelação e problemas de estacionaridade da série. No estudo foi provado a eficiência de utilizar o *Time Series Cross Validation* tendo obtido menores erros de previsão ao serem aplicados em modelos de redes neurais e autoregressivos. Santos e Menezes (2021) utilizaram uma versão múltipla do SSA para filtrar 5 séries de precipitação pluviométrica simultaneamente. Na ocasião, verificou-se que, para mais de uma série, o método se mostra eficaz.

Com os objetivos de verificar qual o modelo mais acurado, qual a classe de modelo adequada para a série, se a metodologia *Cross Validation* melhora a qualidade da previsão e se a filtragem SSA com Clusterização Hierárquica produz ganho preditivo, neste artigo compara-se modelos da classe de Box & Jenkins com os modelos da classe de Holt-Winters, tanto em relação ao ajuste, quanto ao poder preditivo, na série de velocidade do vento com e sem filtragem SSA, usando as estratégias de dividir a amostra em treino/teste e uso do *Cross Validation* e com a clusterização hierárquica sendo utilizada na fase de agrupamento da abordagem SSA. O método *Time Series Cross Validation* é utilizado na validação da previsão, quando a amostra é dividida em amostra de treinamento e de teste. Para a seleção de modelos, o Critério de Informação Bayesiano (BIC) é

utilizado e para avaliar o desempenho preditivo dos modelos, as medidas Erro Médio Percentual Absoluto ($MAPE$) e a Raiz Quadrada do Erro Quadrático Médio ($RMSE$) são utilizadas. Para tanto, o software estatístico ***R version 4.5.0 (How About a Twenty-Six)*** é utilizado nas análises de dados, modelagens e previsões.

Este artigo está apresentado em nove seções. Na Seção 2 são apresentados os modelos de Holt-Winters, enquanto na Seção 3 estão os modelos de Box & Jenkins. Na Seção 4 são apresentadas as estatísticas de aderênica. Na Seção 5 o método *Cross Validation* é apresentado. Na Seção 6 é apresentada a técnica SSA, enquanto na Seção 7 é apresentada a Clusterização Hierárquica. Em seguida, na Seção 8, são apresentados os principais resultados após a aplicação das técnicas à série temporal de velocidade do vento e por fim, na Seção 9, são apresentadas as conclusões.

2. Modelos de Holt-Winters

Segundo Hyndman et al. (2002), o método de Holt-Winters, também conhecido como suavização exponencial tripla, é utilizado para séries que apresentam nível, tendência e sazonalidade. Para cada um das três componentes são definidos hiperparâmetros que decaem de forma exponencial com o passar do tempo. Este método pode ser definido como sendo uma média móvel ponderada exponencialmente, de modo que para observações mais recentes são dados pesos maiores. Este modelo também é dividido quanto a sazonalidade da série, podendo ser aditiva, se a mesma é constante em todo tempo, ou multiplicativa, se a mesma diminui ou aumenta com o passar do tempo. A equação de previsão do modelo multiplicativo é definida pela Equação (1).

$$\widehat{Z}_{t+n} = (L_t + nT_t)S_{t-s+n}, \tag{1}$$

em que L_t representa o nível da série, T_t a tendência, sazonalidade por S_t, t o instante de tempo, s a frequência e n o número de passos a frente.

Dessa forma, para atualizar a equação de previsão se faz necessário obter estimações do nível, da tendência e dos fatores sazonais visto anteriormente. Esses valores podem ser obtidos pelas Equações (2), (3) e (4).

$$L_t = \alpha\frac{Z_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + T_{t-1}), \tag{2}$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}, \tag{3}$$

$$S_t = \gamma\frac{Z_t}{L_t} + (1-\gamma)S_{t-s}, \tag{4}$$

em que α é a constante de amortecimento do nível, β é a constante de amortecimento da tendência, γ é a constante de amortecimento dos fatores sazonais.

3. Modelos de Box & Jenkins

O modelo geral proposto por Box e Jenkins (1970) assume a realização $(z_1, z_2, \ldots, z_T)$ do processo estocástico $\{Z_t\}_{t\in I}$, em que, comumente em Séries Temporais, o conjunto de índices I pode ser considerado como o conjunto dos números naturais ou inteiros. Hamilton (1994) apresenta uma introdução aos processos estocásticos e as suas caracterísicas, tais como estacionariedade. Na ocasião, a Equação (5) considera os erros estocásticos ε_t e os valores observados z_t em tempos defasados.

$$z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}. \tag{5}$$

Este modelo pode ser representado em termos de dois polinômios obtidos via operador de defasagem B definido por $B^d Z_t = Z_{t-d}$ (Hamilton, 1994). Desse modo, a Equação (5) pode ser descrita como em (6).

$$(1 - \phi_1 B^1 - \cdots - \phi_p B^p) z_t = (1 - \theta_1 B^1 - \cdots - \theta_q B^q) \varepsilon_t, \tag{6}$$

sendo $\phi(B) = 1 - \phi_1 B^1 - \cdots - \phi_p B^p$ e $\theta(B) = 1 - \theta_1 B^1 - \cdots - \theta_q B^q$ os respectivos polinônimos autorregressivo e de média móvel. Assim, a Equação (6) pode ser reescrita por (7).

$$\phi(B) z_t = \theta(B) \varepsilon_t, \tag{7}$$

A Equação (7) representa o modelo Autorregressivo e de Média Móveis, $ARMA(p, q)$. Quando $q = 0$, o polinômio de médias móveis tem grau zero, sendo identicamente igual a 1. Assim a Equação (7) é descrita por $\phi(B) z_t = \varepsilon_t$, sendo este o modelo Autorregressivo de ordem p - $AR(p)$. De forma análoga, o modelo $z_t = \theta(B) \varepsilon_t$ é o modelo de Média Móveis de ordem q - $MA(q)$.

Quando a série temporal é uma realização de um processo estocástico não estacionário, sucessivas diferenças são feitas até que se obtenha a estacionariedade (Hamilton, 1994). O operador de diferenças Δ é definido por $\Delta z_t = z_t - z_{t-1}$. Considerando que uma série não estacionária necessite de d diferenças para se tornar estacionária, o modelo para estas séries será o modelo Atorregressivo Integrado e de Média Móveis $ARIMA(p, d, q)$, como descrito em (8).

$$\phi(B)(1 - B)^d z_t = \theta(B) \varepsilon_t, \tag{8}$$

de modo que é possível verificar que $1 - B = \Delta$.

Na presença de sazonalidade, o modelo $SARIMA(P, D, Q) \times (p, d, q)_S$, em (9), é utilizado para descrever a série.

$$\phi(B) \Phi(B^S)(1 - B)^d (1 - B^S)^D z_t = \theta(B) \Theta(B^S) \varepsilon_t, \tag{9}$$

em que S é o período sazonal, $\Phi(B^S)$ é o polinômio autorregressivo sazonal, $\Theta(B^S)$ é o polinômio de médias móveis sazonal e D é o número de diferenças sazonais necessárias para que a série se torne sazonalmente estacionária (Morettin e Toloi, 2018).

4. Medidas de Desempenho Preditivo e Critério de Seleção

Os modelos de Holt-Winters e de Box & Jenkins são ajustados na amostra de treino tanto na série original, quanto na série filtrada SSA. Os desempenhos preditivos dos modelos são avaliados nas amostras de testes. Neste trabalho são utilizadas três medidas, sendo um critério de seleção de modelos e duas medidas de desempenho, anotadas a seguir:

$$MAPE = \frac{\sum_{t=1}^{T} \left| \frac{Z_t - \hat{Z}_t}{Z_t} \right|}{T} \times 100, \qquad RMSE = \sqrt{\frac{\sum_{t=1}^{T} (Z_t - \widehat{Z}_t)^2}{T}},$$

$MAPE$ (*Mean Absolute Percentage Error*) e $RMSE$ (*Root Mean Squared Error*) são medidas de desempenho que avaliam o desvio do modelo perante a série original. Quanto menor, melhor.

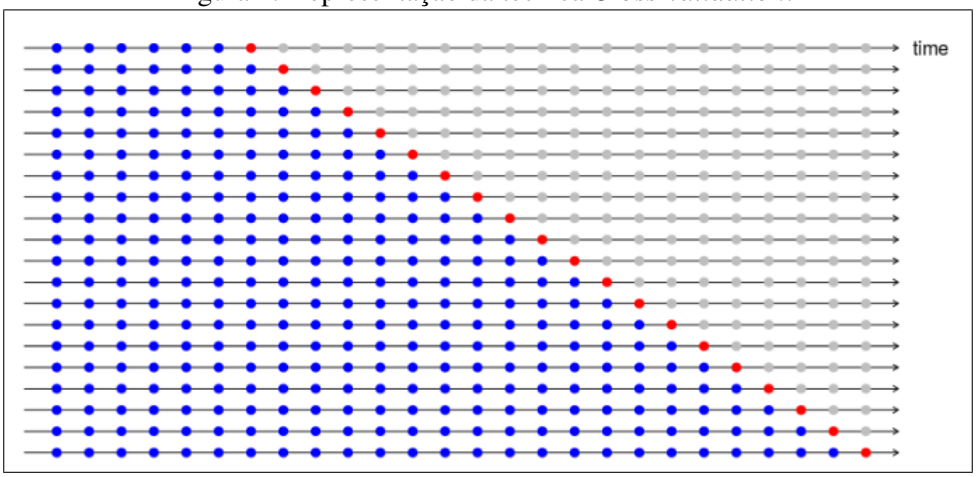
$$BIC = -2 \log L_p + [(p+1)+1] \log T.$$

O critério de seleção de modelo BIC (*Bayesian Information Criterion*) penaliza modelos com muitos parâmetros, utilizando o princípio da parcimônia. Dessa forma, quanto menor o valor do BIC, melhor é o modelo.

5. Time Series Cross Validation

De acordo com Hyndman et al. (2018), *Time Series Cross Validation*, ou simplesmente *Cross Validation*, é uma forma de mensurar a performance do modelo mais apropriadamente do que o método de *hold-out (in-sample e out-of-sample)*, em que se divide a série temporal em apenas dois subgrupos denominados amostra de treino e amostra de teste. Uma vantagem desse novo cenário é a redução do problema conhecido como *Overfitting*, isto é, quando o erro da modelagem medido pela amostra de teste é baixo ou aceitável para colocar o modelo em produção, no entanto, para dados futuros o mesmo modelo apresenta erros superiores aos medidos anteriormente inviabilizando a continuação do processo de previsão. A metodologia alternativa proposta pelo atual estudo é dado pelo método *Cross Validation* conforme apresentado na Figura 1.

Figura 1: Representação da técnica *Cross Validation*.



Fonte: `https://www.sciencedirect.com/science/article/abs/pii/S0167947317302384?via%3Dihub.`

Na Figura 1, os pontos azuis representam a amostra de treinamento (dentro da amostra) e os pontos vermelhos são a amostra de teste (fora da amostra) usados para estimar o erro de previsão um passo à frente. A cada iteração, o modelo acrescenta uma nova observação na amostra de treinamento que, por sinal, é a amostra de teste (ponto vermelho) obtida no período anterior, reestima os parâmetros e recalcula o erro para o próximo passo à frente. Este processo pode ser refeito inúmeras vezes definido pelo usuário e no final calcula-se a média simples dos erros de previsão de um passo à frente. Com isto, espera-se diminuir a variância das previsões, uma vez que foram estimadas inúmeras previsões um passo à frente. Enquanto o método de *hold-out* estima a previsão uma única vez e calcula o erro.

Note que, apesar de a figura representar apenas um passo à frente, nada impede que usuário possa estimar mais de um passo à frente pelo método do *Cross Validation*.

6. Filtragem SSA

SSA é uma técnica não paramétrica útil para filtrar dados de séries temporais que permite a sua decomposição em sinal e ruído. SSA incorpora elementos de análise clássica de séries temporais, estatística multivariada, geometria multivariada, sistemas dinâmicos e processamentos de sinais (Elsner e Tsonis, 1996). SSA tem sido aplicada com sucesso em diversas áreas como matemática, física, economia, matemática financeira, meteorologia, oceanografia e ciências sociais (Golyandina et al., 2001).

De acordo com Hassani et al. (2012), a versão básica do método SSA pode ser dividida em duas etapas: decomposição e reconstrução.

6.1. Decomposição

A etapa de decomposição pode ser subdividida em: Incorporação e decomposição em valores singulares (SVD – *Singular Value Decomposition*). Considere $(z_1, \ldots, z_T)$ as observações de uma série temporal $\{Z_t\}_{t\in I}$ e considere L tal que $2 \leq L \leq T$ de modo que L é um parâmetro a ser estimado e é chamado de comprimento da janela (Golyandina et al., 2001). Entende-se por Incorporação o procedimento no qual uma série temporal Z_T é levada a uma matriz $\mathbf{X}$ chamada "Matriz Trajetória" dada por (10).

$$\mathbf{X} = \begin{bmatrix} z_1 & z_2 & z_3 & \cdots & z_k \\ z_2 & z_3 & z_4 & \cdots & z_{k+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_L & z_{L+1} & z_{L+2} & \cdots & z_T \end{bmatrix} \tag{10}$$

A matriz $\mathbf{X}$ é uma matriz Hankel, ou seja, os elementos de $x_{i,j}$ tal que $i + j =$ constante são iguais.

Considere $\mathbf{S} = \mathbf{X}\mathbf{X}'$. Os autovalores de $\mathbf{S}$ dispostos em ordem de significância como seguem $\lambda_1 \geq \cdots \geq \lambda_L \geq 0$ são obtidos e os respectivos autovetores $U_1, \ldots, U_L$ são encontrados. Considere $V' = (X'U_L)/\sqrt{\lambda}$, como $\mathbf{S}$ é positiva semi-definida, então a matriz trajetória $\mathbf{X}$ pode ser expressa pela decomposição em valores singulares (SVD) apresentada em (11):

$$\mathbf{X} = \mathbf{E}_1 + \mathbf{E}_2 + \cdots + \mathbf{E}_L, \tag{11}$$

em que $E_l = \sqrt{\lambda} U_l V_l'$, para todo $l = 1, \ldots, L$. A coleção $(\sqrt{\lambda_l}, U_l, V_l)$ é conhecida como autotripla da expansão SVD de $\mathbf{X}$. Os elementos da autotripla são definidos respectivamente por: valor singular, vetor singular à esquerda e vetor singular à direita de $\mathbf{X}$ (Menezes et al., 2014). A contribuição de cada componente em (11) pode ser mensurada por $\lambda_l / \sum_{l=1}^{L} \lambda_l$, em que λ_l são autovalores.

6.2. Reconstrução

A etapa de reconstrução está subdividida em duas partes: agrupamento e média diagonal. A etapa de agrupamento consiste em agrupar algumas sequências de matrizes elementares resultantes da decomposição SVD em grupos disjuntos e, após isso, realizar a soma delas, gerando novas matrizes elementares.

Considere a sequência $\sum_{l=1}^{L} \mathbf{E}_l$ de matrizes elementares da expansão SVD. Agrupe as mesmas em m grupos disjuntos utilizando algum método, como por exemplo análise gráfica de vetores singulares, análise de componentes principais ou clusterização hierárquica. Assuma que o conjunto de índices gerado é dado por $\{I_1, \ldots, I_m\}$, de modo que a expansão (11) pode ser reescrita como em (12), sendo $\mathbf{X}_{I_i}$ arbitrária tal que $\mathbf{X}_{I_i} = \sum_{j=1}^{p_i} \mathbf{X}_{I_i j}$.

$$\mathbf{X} = \sum_{l=1}^{L} \mathbf{E}_l = \sum_{i=1}^{m} \mathbf{X}_{I_i} \tag{12}$$

O objetivo do agrupamento é diminuir o número de componentes na expansão da matriz trajetória $\mathbf{X}$. A contribuição de cada componente é mensurada pela razão (13).

$$\frac{\sum_{j=1}^{p_i} \lambda_{I_{i_j}}}{\sum_{l=1}^{L} \lambda_l}. \tag{13}$$

Considere a matriz trajetória X e assuma que $L^* = \min(L, K)$ e $K^* = \max(L, K)$. Considere $x_{l,k}^{(i)}$ um elemento na linha l e coluna k na matriz X_{I_i}. O elemento $y_t^{(i)}$ da componente $\left[y_t^{(i)}\right]_{1\times T}$ da série temporal $[y_t]_{1\times T}$ é calculado por meio da *média diagonal* da matriz elementar X_{I_i} definida em (14), a partir da matriz elementar X_{I_i}.

$$y_t^{(i)} = \begin{cases} \frac{\sum_{l=1}^{t} x_{l,t-l+1}^{(i)}}{t}, & \text{se} \quad 1 \leq t < L^* \\ \frac{\sum_{l=1}^{L^*} x_{l,t-l+1}^{(i)}}{L^*}, & \text{se} \quad L^* \leq t < K^* \\ \frac{\sum_{l=t-K^*+1}^{T-K^*+1} x_{l,t-l+1}^{(i)}}{T-K^*+1}, & \text{se} \quad K^* \leq t \leq T \end{cases} \tag{14}$$

Cada componente $\left[y_t^{(i)}\right]_{1\times T}$ concentra parte da energia da série temporal original $[y_t]_{1\times T}$ que pode ser mensurada pela razão de autovalores $\sum_{j=1}^{p_i} \lambda_{I_{ij}} / \sum_{l=1}^{d} \lambda_l$. De acordo com Hassani et al. (2012), podemos classificar as componentes SSA $\left[y_t^{(i)}\right]_{1\times T}$ de uma série temporal arbitrária $[y_t]_{1\times T}$ em três categorais: *tendência*, *componentes harmônicas* (ciclo e sazonalidade) e *ruído* (Golyandina et al., 2001).

Um dos principais conceitos estudados em SSA é a propriedade de separabilidade (Hassani et al., 2012). Tal propriedade caracteriza quão bem separados estão as diferentes componentes umas das outras. Uma boa medida de separabilidade é a Correlação Ponderada. Por correlação ponderada *weighted correlation* ou w-correlação entende-se como uma função que quantifica a dependência linear entre duas componentes SSA $Y_T^{(1)}$ e $Y_T^{(2)}$ definida em (15).

$$\rho_{ij}^{(w)} = \frac{\left(Y_T^{(i)}, Y_T^{(j)}\right)_w}{||Y_T^{(i)}||_w ||Y_T^{(j)}||_w}. \tag{15}$$

em que $||Y_T^{(i)}||_w = \sqrt{\left(Y_T^{(i)}, Y_T^{(i)}\right)_w}$; $||Y_T^{(j)}||_w = \sqrt{\left(Y_T^{(j)}, Y_T^{(j)}\right)_w}$;
$\left(Y_T^{(i)}, Y_T^{(j)}\right)_w = \sum_{k=1}^{T} w_k y_k^{(i)} y_k^{(j)}$ e $w_k = \min\{k, L, T-k\}$.

Através da separabilidade, pode-se verificar se duas componentes SSA estão bem separadas. Segundo Hassani et al. (2012), se o valor absoluto da w-correlação é pequeno, então as componentes SSA correspondentes são classificadas como w-ortogonais (ou quase w-ortogonais). Caso contrário, são ditas mal separadas. Salienta-se que comumente utiliza-se a correlação ponderada na fase de agrupamento SSA (Golyandina et al., 2001).

7. Clusterização Hierárquica

A clusterização hierárquica é uma técnica que permite agrupar diferentes observações de um banco de dados em subgrupos de acordo com um conjunto de variávies. Essa técnica é bastante explorada em cenários de *Churn*, isto é, cenários de cancelamentos de serviços entre cliente e empresa (Andrade, 2004).

Existem diversas técnicas de clusterização tais como Clusterização Hierárquica, *K-Means* e *Dynamic Time Warping*. Todas essas técnicas baseiam-se numa medida de similariedade ou dissimilariedade entre as observações, tendo como as principais medidas: distância Euclediana e Mahalanobis.

Em séries temporais, o banco de dados é composto de apenas duas informações: o tempo e a variável de estudo. Dessa forma, não existe nenhum conjunto de variáveis que possam ser utilizadas para clusterização com o intuito de separar as séries temporais em grupos distintos. No entanto, quando utilizada junto à filtragem SSA, é possível separar as componentes de tendência, harmônica e ruidosa.

A clusterização hierárquica aplicada neste artigo é realizada por agrupamento aglomerativo, de modo que cada item é considerado como um grupo individual no primeiro momento e recursivamente vão fundindo a outros grupos de menor distância até que se obtenha a clusterização final. A outra abordagem considera um único grupo no primeiro momento para, recursivamente, ocorrerem as divisões até chegar a clusterização desejada. A relação entre os grupos é calculada pela distância Euclidiana dada em (16):

$$d(X_i, X_j) = \left[\sum_{l=1}^{p}(x_{il} - x_{jl})^2\right]^{\frac{1}{2}} \tag{16}$$

em que X representa a variável, x as respectivas observações da variável e l representa a dimensão dos dados de 1 a p.

8. Resultados

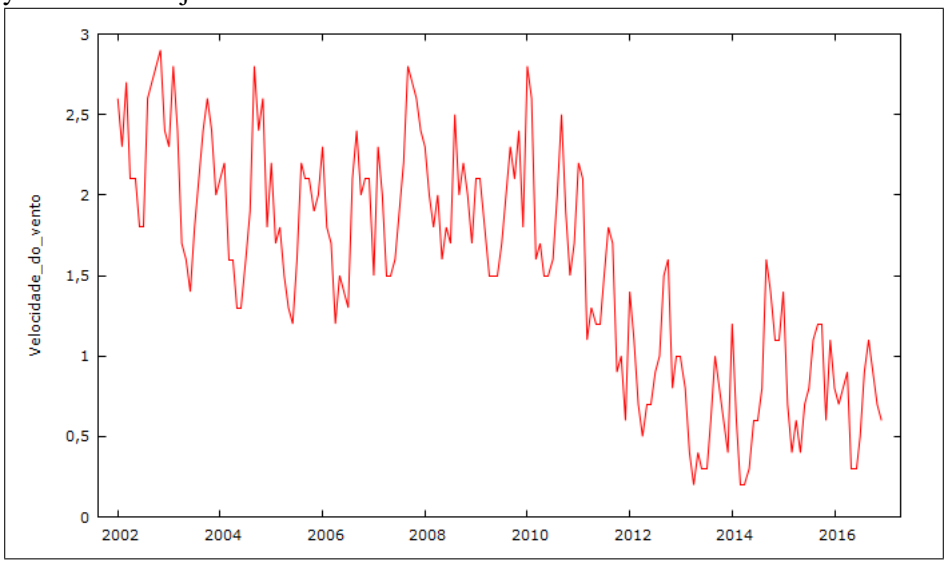
A fim de avaliar a capacidade preditiva dos modelos sem e com a aplicação das metodologias descritas, uma série temporal de média mensal de velocidade do vento (em m/s) com dados coletados a partir de estações anemométricas localizadas no município de Campos dos Goytacazes, no Estado do Rio de Janeiro, foi ajustada e seus resultados comparados via medidas de desempenho preditivo. Os dados variam de janeiro de 2002 a dezembro de 2016 (180 observações) e estão disponíveis na página do Instituto Nacional de Meteorologia (INMET) `[https://portal.inmet.gov.br/]`. A título de entendimento, o ajuste de modelo na amostra de treino também é conhecido como ajuste dentro da amostra ou ajuste *in sample* e a predição feita na amostra de teste é conhecida como predição fora da amostra ou predição *out-of-sample*.

A Figura 2 apresenta o comportamento da série original. Os últimos 12 meses da série são utilizados como amostra de teste dos modelos, enquanto as anos de 2002 a 2015 representam a amostra de treinamento.

A Figura 3 apresenta a componente sazonal da série. A autocorrelação temporal apresenta um comportamento não estacionário, como pode ser visto na Figura 4. Tal comportamento também pode ser observado na Figura 2. Isso implica na necessidade de se fazer pelo menos uma diferença na série antes do ajuste do modelo.

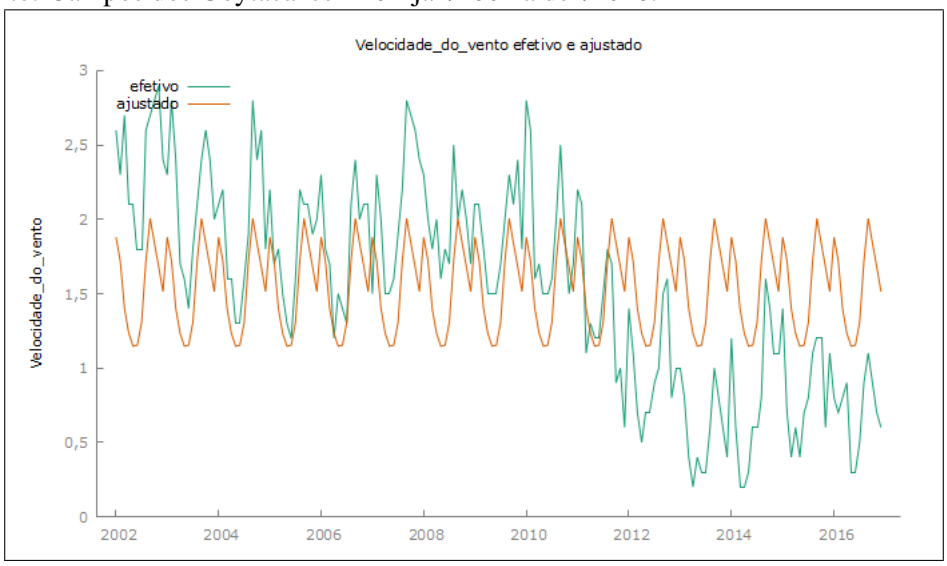
Os resultados estão divididos em quatro etapas, sendo elas: modelagens treino/teste sem *Cross Validation* e sem SSA; modelagens treino/teste com *Cross Validation* e sem SSA; modelagens treino/teste sem *Cross Validation* e com SSA e modelagens treino/teste com *Cross Validation* e com SSA. Em todos os casos, são feitas modelagens de Holt-Winters e de Box & Jenkins. Todas as filtragens SSA são feitas a partir da Clusterização Hierárquica e, em todos os casos, são

Figura 2: Série temporal de médias mensais de velocidade do vento. Campos dos Goytacazes - RJ - jan/2002 a dez/2016.



Fonte: `https://portal.inmet.gov.br/`

Figura 3: Componente Sazonal da Série temporal de médias mensais de velocidade do vento. Campos dos Goytacazes - RJ - jan/2002 a dez/2016.



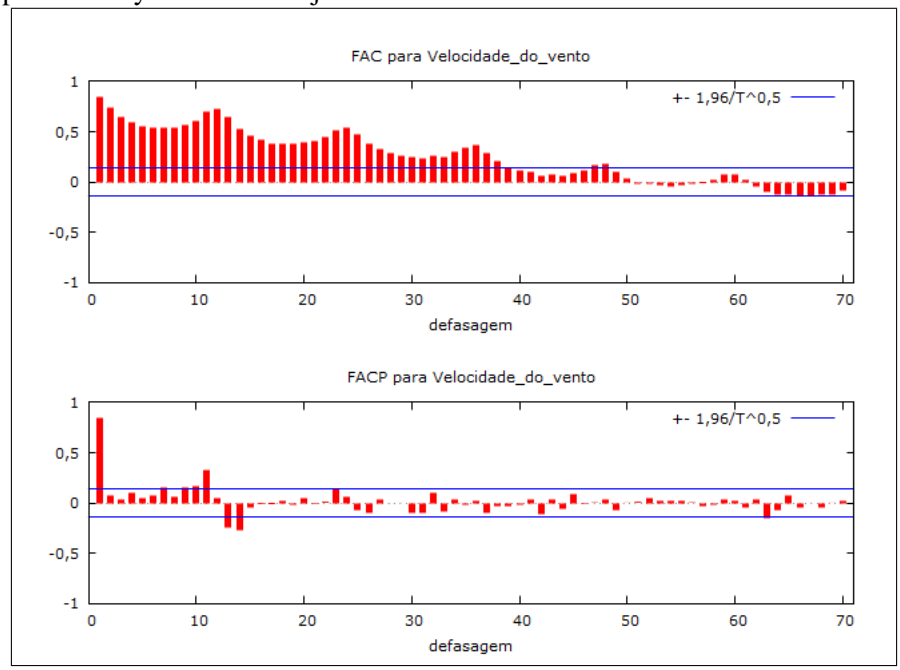
Fonte: Elaboração própria.

utilizados os modelos mais adequados em cada classe de acordo com as análises dos resíduos e são aplicadas as medidas de desempenho preditivo a fim de obter o modelo mais acurado.

8.1. Análise dos Resultados sem a Filtragem SSA

Como a amplituade das variações sazonais da série são constantes ao longo do tempo, o modelo sazonal aditivo é o mais adequado dentro da classe dos modelos de amortecimento exponencial de Holt-Winters. Já na classe dos modelos de Box & Jenkins, alguns procedimentos são necesários, uma vez que já se verificou perante a FAC e a FACP que a série não apresenta estacionariedade. Tal condição também é verificada pelo resultado obtido no Teste Aumentado de Dickey-Fuller. Nota-se também no correlograma da Figura 4, que há picos nos *lags* múltiplos de

Figura 4: Correlograma da Série temporal de médias mensais de velocidade do vento. Campos dos Goytacazes - RJ - jan/2002 a dez/2016.



Fonte: Elaboração própria.

12, indicando não estacionariedade sazonal. Diante deste cenário, a primeira diferença foi feita e o resultado obtido foi testado via Teste Aumentado de Dickey -Fuller. O teste rejeitou a hipótese nula de existência de raiz unitária para a parte não sazonal. No entanto, é possível perceber na Figura 5 que há um descresimento lento nos *lags* múltiplos de 12, indicando a necessidade de se fazer diferenças sazonais. Após a diferença sazonal, um novo correlograma foi obtido e neste, finalmente, foi apresentado o comportamento estacionário, resultado este confirmado pelo Teste Aumentado de Dickey-Fuller. A Figura 6 mostra este comportamento, que serviu de base para a escolha dos modelos.

Analisando o comportamento do Correlograma da Figura 6, pode-se perceber um decaimento rápido na FACP e uma significância no *lag* 1, sugerindo um modelo $MA(1)$ e significâncias no *lag* 12 tanto na FAC, quanto na FACP, podendo indicar um modelo $ARMA(1,1)$ sazonal. Considerando que há diferenças simples e sazonal, então um possível modelo da classe de Box & Jenkins a ser avaliado será $SARIMA(0,1,1) \times (1,1,1)_{12}$. Contudo, ainda é possível considerar que na FACP haja uma significância no *lag* 1, assim como há na FAC. Isso implicaria em um modelo $ARMA(1,1)$, levando a outro possível modelo: $SARIMA(1,1,1) \times (1,1,1)_{12}$. A Tabela 1 apresenta os valores do Critério de Informação Bayesiano (BIC) obtidos na amostra de treino para cada modelo.

Tabela 1: Critério de Informação Bayesiano - seleção do modelo de Box & Jenkins.

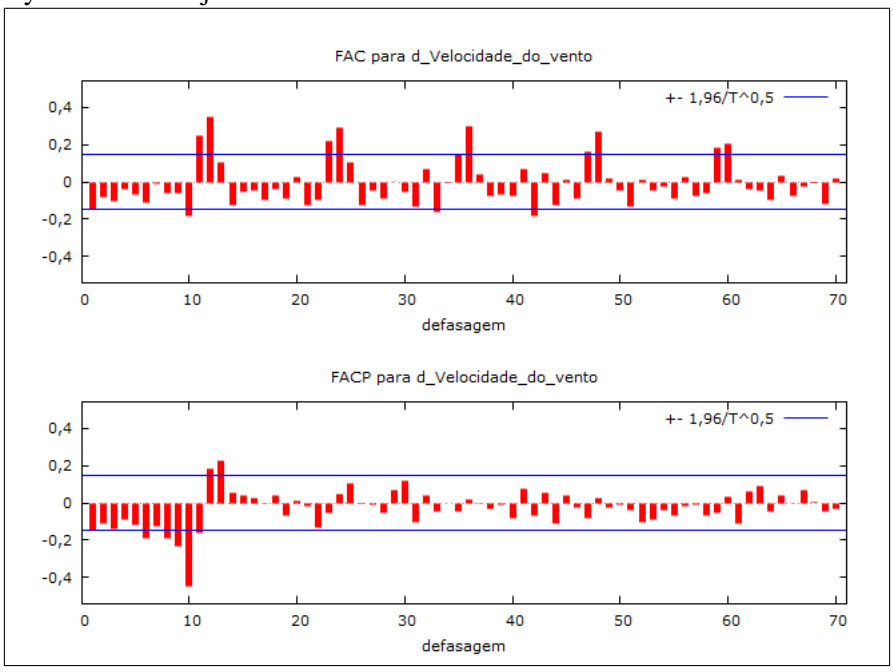
MODELO	BIC
$SARIMA(0,1,1) \times (1,1,1)_{12}$	74,13
$SARIMA(1,1,1) \times (1,1,1)_{12}$	79,25

Fonte: Elaboração própria.

Conforme pode ser observado na Tabela 1, o modelo $SARIMA(0,1,1) \times (1,1,1)_{12}$ obteve o menor critério BIC. Portanto, este modelo será considerado para as abordagens a serem verificadas neste artigo.

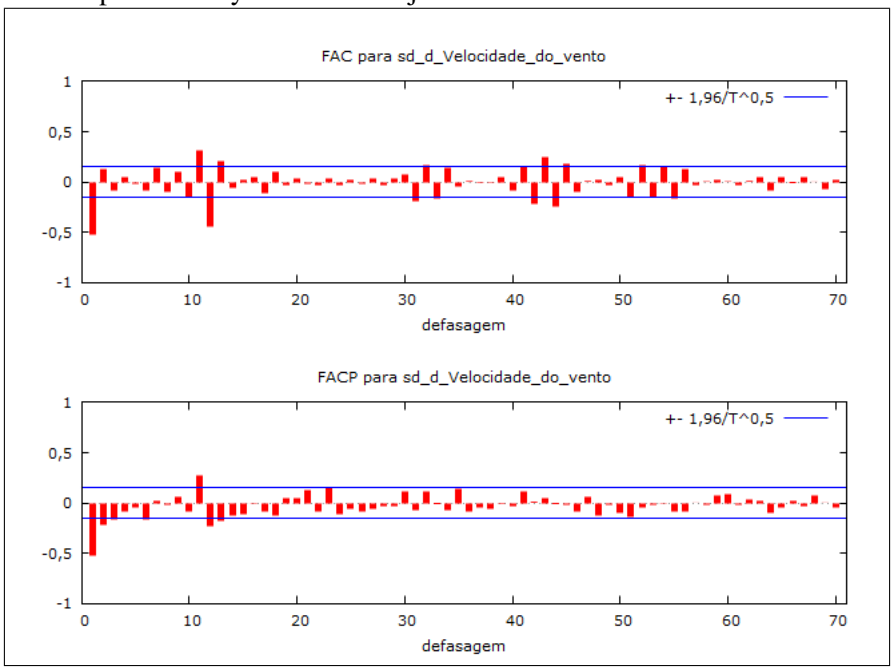
Após definir os modelos de Holt-Winters e de Box & Jenkins, o desempenho preditivo dos

Figura 5: Correlograma da Série temporal após a primeira diferença simples. Campos dos Goytacazes - RJ - jan/2002 a dez/2016.



Fonte: Elaboração própria.

Figura 6: Correlograma da Série temporal após uma diferença simples e uma diferença sazonal. Campos dos Goytacazes - RJ - jan/2002 a dez/2016.



Fonte: Elaboração própria.

modelos é avaliado a partir das medidas $MAPE$ e $RMSE$. Os resultados estão na Tabela 2.

No procedimento utilizando amostras treino e teste, o poder preditivo sempre é avaliado na amostra teste. Os resultados na Tabela 2 mostram melhor desempenho do modelo de Holt-Winters, uma vez que apresenta menores valores das medidas avaliadas. Com isso, considera-se que o modelo de Holt-Winters com sazonalidade aditiva seja o mais adequado a esta série sem a filtragem

Tabela 2: Medidas de desempenho (*out-of-sample*) - Série sem filtro SSA.

	$MAPE$	$RMSE$
Holt-Winters	0,362	0,328
Box & Jenkis	0,625	0,366

Fonte: Elaboração própria.

SSA e sem o *Cross Validation*.

Na segunda etapa, o processo *Cross Validation* é aplicado à série sem a filtragem SSA. Neste processo, a cada iteração, um novo elemento de amostra é incluído. A Tabela 3 apresenta as medidas de desempenho fora da amostra quando o processo *Cross Validation* é utilizado.

Tabela 3: Medidas de desempenho (*out-of-sample*) - Série sem filtro SSA com *Cross validation*.

	$MAPE$	$RMSE$
Holt-Winters	0,372	0,324
Box & Jenkins	0,348	0,301

Fonte: Elaboração própria.

A maior contribuição do processo *Cross Validation* está na análise fora da amostra, uma vez que os elementos de amostra são inseridos a cada iteração na amostra de teste. A partir daí, percebe-se na Tabela 3, que o efeito desse processo afeta nas medidas de desempenho avaliadas, que apresentam melhores resultados para a classe de modelos de Box & Jenkins, ao contrário do que acontece no caso sem esse processo.

8.2. Análise dos Resultados com a Filtragem SSA

A abordagem SSA é feita no R com o pacote Rssa. Neste pacote, as escolhas dos autovetores no processo de reconstrução da série é feito por clusterização hieráquica, de modo que a decomposição da série original em componentes harmônica, tendência e ruidosa é feita a partir de escolha de parâmetros adequados. No caso deste artigo, o comprimento de janela utilizado foi $L = T/2 = 180/2 = 90$ com base nas observações feitas em Golyandina et al. (2001). As Figuras 7 e 8 apresentam os comportamentos das componentes harmônica, tendência e ruidosa no processo de reconstrução da série.

A fim de verificar se as componentes estão bem separadas, a correlação ponderada foi aplicada entre elas. O resultado das correlações está na Tabela 4.

Tabela 4: Correlação Ponderada entre as componentes no processo SSA.

	Tendência	Harmônica	Ruidosa
Tendência	1,00	-0,04	0,05
Harmônica	-0,04	1,00	0,07
Ruidosa	0,05	0,07	1,00

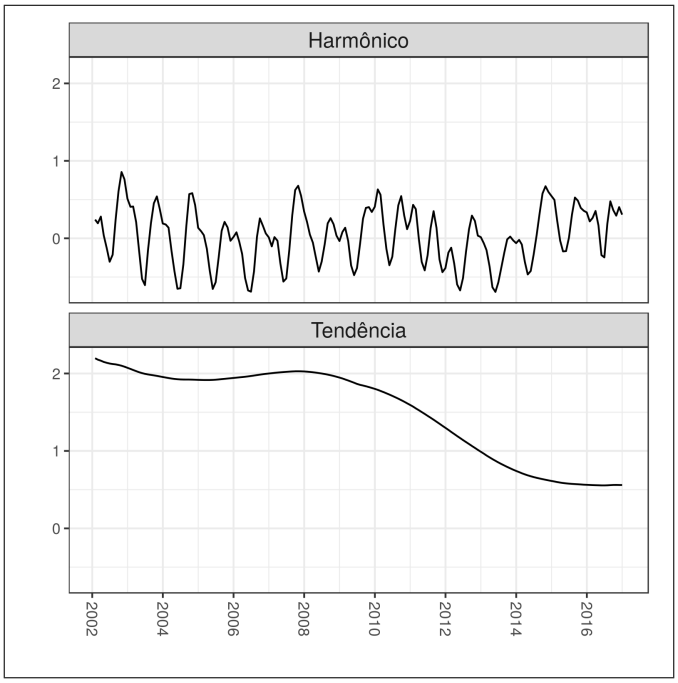
Fonte: Elaboração própria.

Os resultados apresentados na Tabela 4 indicam boa separabilidade, dado que as correlações ponderadas são valores muito baixos entre componentes distintas.

Por fim, a componente ruidosa é removida e a série é reconstruída com as componentes harmônica e tendência. A Figura 9 apresenta a comparação entre a série sem o filtro SSA e a série filtrada SSA.

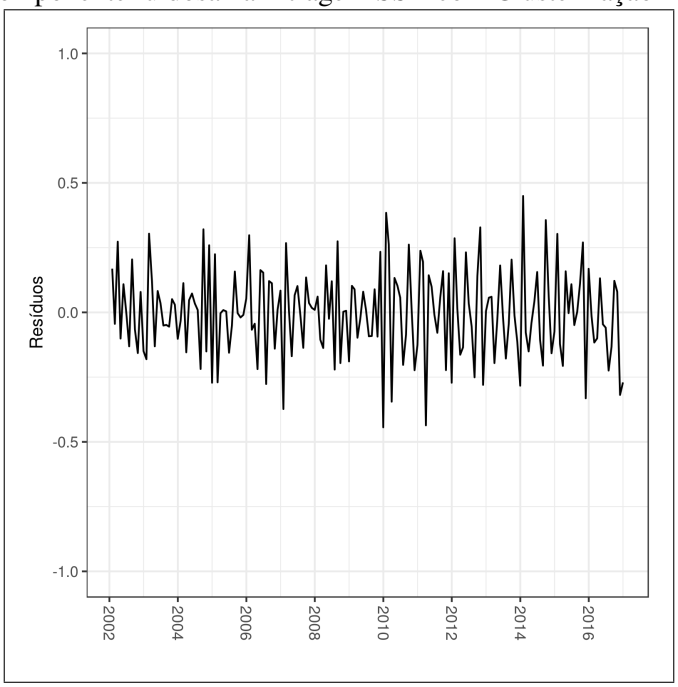
A série filtrada apresentada na Figura 9 está suavizada via SSA. Ela será utilizada para as modelagens a seguir sem e com o procedimento *Cross Validation*.

Figura 7: Componentes harmônica e tendência na filtragem SSA com Clusterização Hierárquica.



Fonte: Elaboração Própria.

Figura 8: Componente ruidosa na filtragem SSA com Clusterização Hierárquica.

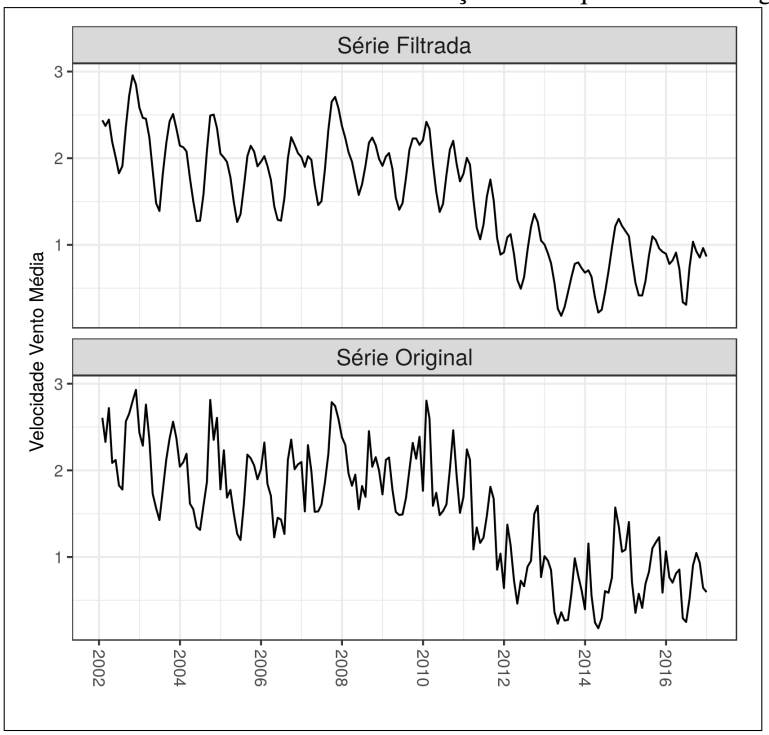


Fonte: Elaboração Própria.

Na ocasião dos ajustes dos modelos de Holt-Winters e de Box & Jenkins, o modelo de suavização exponencial com sazonalidade aditiva foi o mais adequado, enquanto na classe de modelos de Box & Jenkins, o modelo $SARIMA(0,1,2) \times (1,1,1)_{12}$, apresentou menor BIC.

As Figuras 10 e 11 apresentam a série original juntamente com as séries ajustadas via modelos ARIMA de Box & Jenkins e Suavização Exponencial de Holt-Winters e a série filtrada

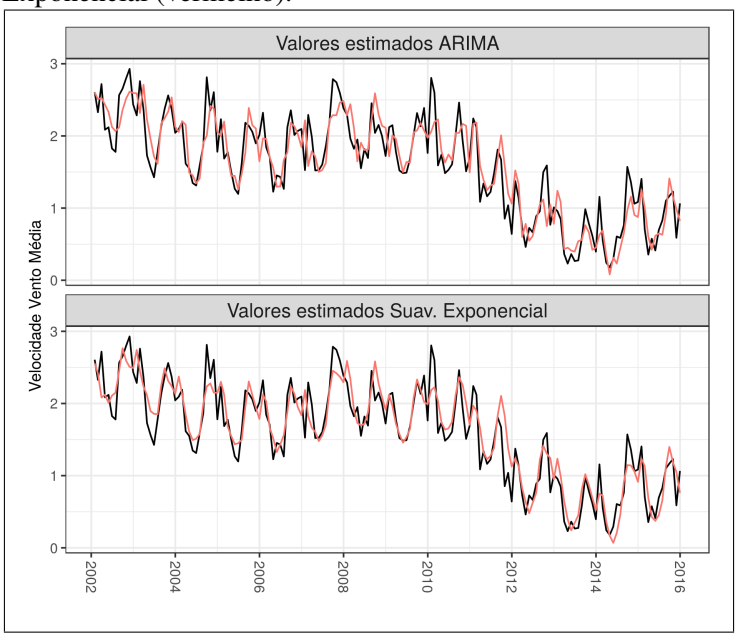
Figura 9: Série filtrada SSA com Clusterização hierárquica e série origial.



Fonte: Elaboração Própria.

via SSA juntamente com as séries ajustadas via modelos ARIMA de Box & Jenkins e Suavização Exponencial de Holt-Winters respectivamente.

Figura 10: Série original (preto) juntamente com as séries ajustadas via ARIMA e Suavização Exponencial (vermelho).

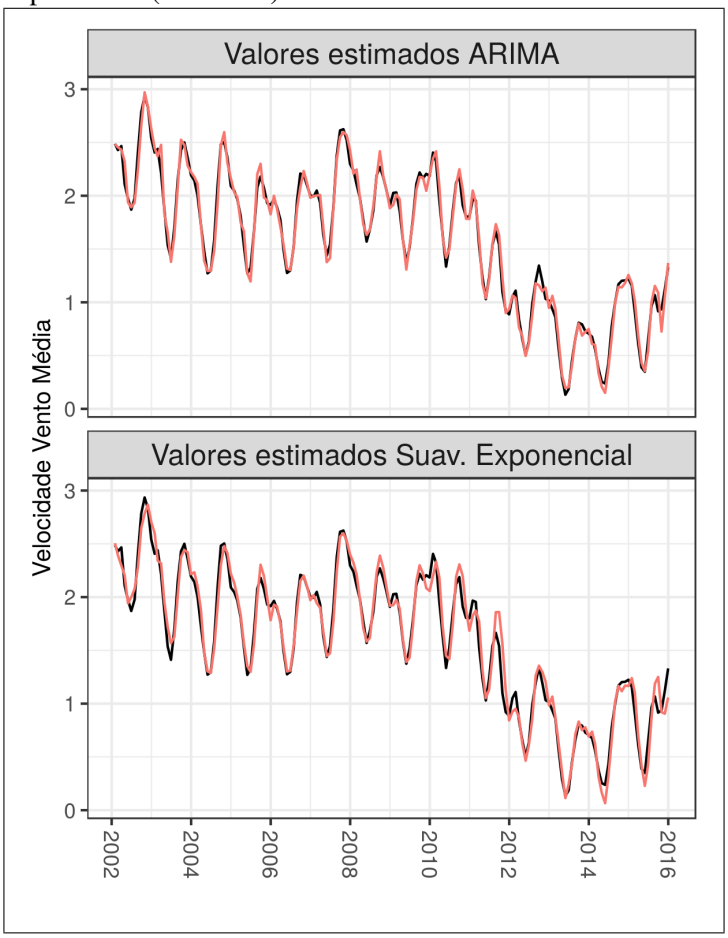


Fonte: Elaboração Própria.

Observando a Figura 11, é possível perceber que os modelos aderem melhor à série filtrada do que o que pode ser verificado na Figura 10 com a série sem o filtro SSA.

As medidas de desempenho são obtidas na verificação do melhor modelo fora da amostra

Figura 11: Série filtrada SSA (preto) juntamente com as séries ajustadas via ARIMA e Suavização Exponencial (vermelho).



Fonte: Elaboração Própria.

(amostra de teste) em relação a capacidade preditiva. Os resultados dessas medidas nos casos aplicados à série filtrada SSA estão na Tabela 5.

Tabela 5: Medidas de desempenho (*out-of-sample*) - Série filtrada via SSA.

	$MAPE$	$RMSE$
Holt-Winters	0,724	0,469
Box & Jenkins	0,618	0,397

Fonte: Elaboração própria.

Na Tabela 5, é possível perceber que o modelo de Box & Jenkins possui melhores resultados, uma vez que apresenta menores valores das medidas avaliadas.

Na implementação *Cross Validation* com filtragem SSA, a cada iteração uma nova observação é implementada e a série é filtrada.

Para concluir as análises, a acurácia da previsão é verificada fora da amostra utilizando *Cross validation* na série filtrasa SSA. A Tabela 6 apresenta os resultados das medidas de desempenho sobre os modelos de suavização exponencial de Holt-Winsters e de Box & Jenkins.

Conforme apresentado na Tabela 6, o modelo de Holt-Winters minimiza as medidas $MAPE$ e $RMSE$ e, com isso, apresenta o melhor desempenho fora da amostra para o modelo com o filtro SSA e *Cross Validation*.

A Tabela 7 apresenta comparações das medidas de desempenho fora da amostra entre os modelos da classe de amortecimento exponencial de Holt-Winters e de Box & Jenkins utilizando

Tabela 6: Medidas de desempenho (*out-of-sample*) - Série filtrada via SSA com *Cross Validation*.

	$MAPE$	$RMSE$
Holt-Winters	0,226	0,168
Box & Jenkins	0,247	0,173

Fonte: Elaboração própria.

Cross Validation sem e com a filtragem SSA.

Tabela 7: Medidas de desempenho (*out-of-sample*) - *Cross Validation*.

	$MAPE$	$RMSE$
Holt-Winters	0,372	0,324
Holt-Winters com SSA	0,226	0,168
Box & Jenkins	0,348	0,301
Box & Jenkins com SSA	0,247	0,173

Fonte: Elaboração própria.

A Tabela 7 resume os resultados das medidas de desempenho fora da amostra com *Cross Valiadation* para as séries sem e com a filtragem SSA. Os resultados corroboram que este filtro melhora a qualidade do ajuste. Contudo, neste caso, o melhor desempenho para a série original é apresentado pelo modelo de Box & Jenkins, minimizando as medidas $MAPE$ e $RMSE$. Enquanto para a série filtrada, o modelo de amortecimento exponencial de Holt-Winters apresenta melhor desempenho.

A Tabela 8 apresenta o resumo geral com todos os resultados fora da amostra considerando todas as situações testadas com os modelos propostos.

Tabela 8: Medidas de desempenho - Resumo geral.

	(*out-of-sample*)	
	$MAPE$	$RMSE$
Holt-Winters	0,362	0,328
Holt-Winters - *Cross Validation*	0,372	0,324
Holt-Winters - SSA	0,724	0,618
Holt-Winters - SSA - *Cross Validation*	**0,226**	**0,168**
Box & Jenkins	0,625	0,366
Box & Jenkins - *Cross Validation*	0,348	0,301
Box & Jenkins - SSA	0,618	0,397
Box & Jenkins - SSA - *Cross Validation*	0,247	0,173

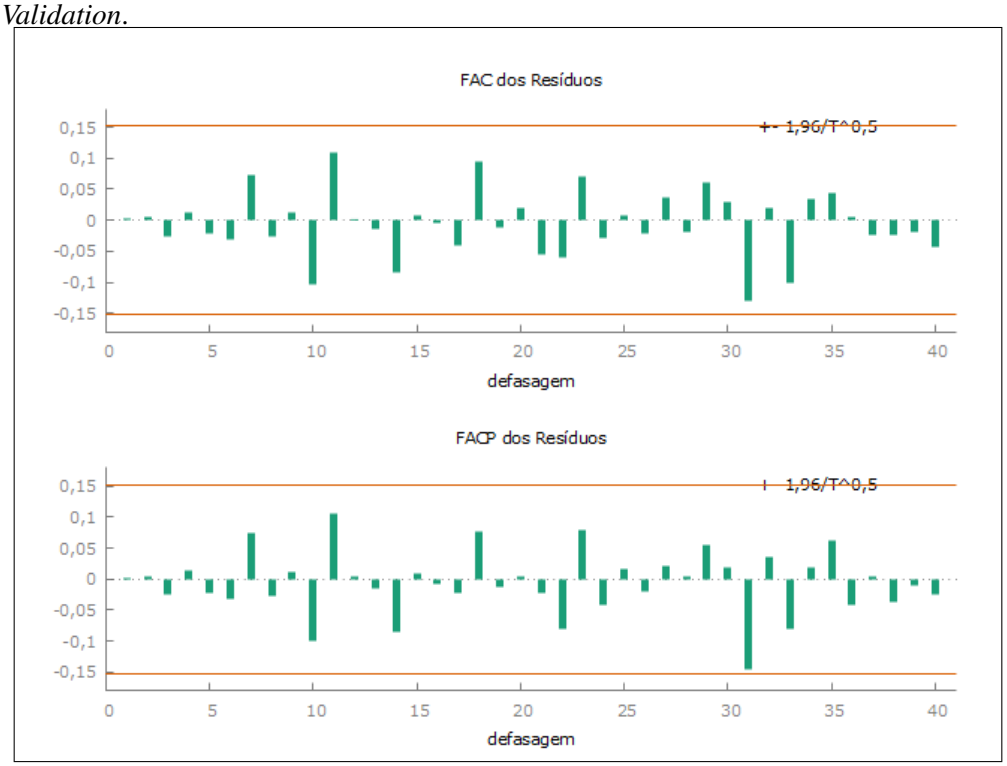
Fonte: Elaboração própria.

Diante de todos os cenários apresentados, aquele que minimiza as medidas de desempenho fora da amostra é o modelo de Holt-Winters com sazonalidade aditiva para a série filtrada SSA e com *Cross Validation*. O modelo a ser considerado para previsão é aquele com melhor desempenho fora da amostra, uma vez que esta abordagem avalia a capacidade de o modelo prever.

A Figura 12 apresenta o correlograma dos resíduos para este modelo. Este correlograma atesta que os resíduos são não correlacionados. Além disso, os testes de Ljung-Box e Shapiro-Wilk confirmam a independência e normalidade dos resíduos, certificando a adequação do modelo para previsão.

Verificada a adequação do modelo, previsões para 24 meses foram geradas a partir do modelo *Holt-Winters - SSA - Cross Validation*. A Figura 13 apresenta o ajuste do modelo e as previsões de 24 meses. Para estes ajustes, são considerados intervalos de 95% de confiança.

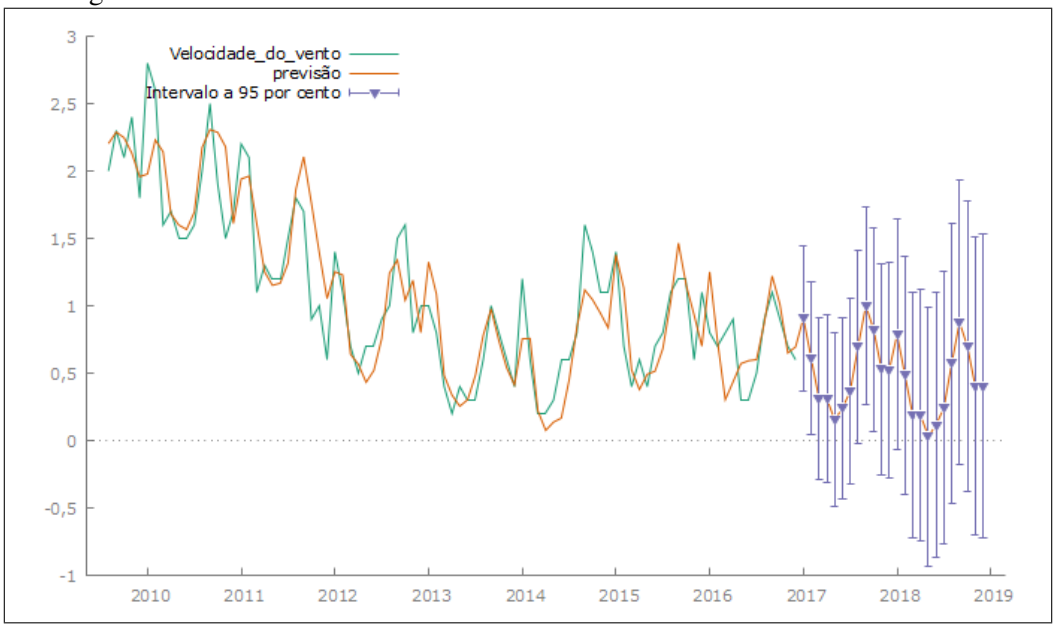
Figura 12: Correlograma dos resíduos do modelo Holt-Winters - SSA - *Cross Validation*.



Fonte: Elaboração própria.

Pode-se observar que as previsões acompanhas a tendência e a zasonalidade da série.

Figura 13: Previsão 24 meses. Modelo Holt-Winters - SSA - *Cross Validation*.



Fonte: Elaboração própria.

9. Conclusões

A fim de se obter uma fonte alternativa de energia para os momentos em que as fontes hidrelétricas não são suficientes para atender às crescentes demandas, modelos de previsão de uma série temporal de velocidade do vento foram ajustados com o objetivo de se verificar a possibilidade de instalação de usinas para geração de energia eólica.

Para isso, modelos de Suavização Exponencial de Holt-Winters e de Box & Jenkins foram usados considerando amostras de treinamento e de teste. Além disso, duas metodologias foram testadas nos ajustes: Singular Spectrum Analysis (SSA) com Clusterização Hierárquica e *Time Series Cross Validation*. Na ocasião, os modelos foram ajustados sem e com a filtragem SSA e sem e com o processo *Cross validation*.

De acordo com as medidas de desempenho, o modelo de Box & Jenkins se mostra mais eficiente na maioria dos cenários fora da amostra. Tanto na série original com *Cross Validation*, quanto na série filtrada SSA sem o uso do procedimento *Cross Validation*. Contudo, ao se aplicar o procedimento *Cross Validation* na série filtrada SSA, o modelo amortecimento exponencial de Holt-Winters se mostra mais eficiente preditivamente.

Porém, o modelo de Holt-Winters obteve melhor resultado global considerando a abordagem SSA-*Cross Validation*, o que nos leva a crer que para este caso particular seria o métodos mais adequado para a modelagem e previsão.

Mas, é preciso ter cautela ao se pensar em utilizar esta técnica devido a dependência temporal. A inclusão de dados a cada iteração pode gerar informações desnecessárias que podem quebrar a ordem temporal. Isso pode causar probelmas nos resultados como o desempenho do modelo de Box & Jenkins piorar ao se utilizar a técnica *Cross Validation* na série fora da amostra quando filtrada via SSA ou provocar maior valor na medida de desempenho, diminuindo a sua capacidade preditiva.

Concluis-se, então, que os modelos de Holt-Winters são mais adequados que os modelos de Box & Jenkins para previsão neste caso desta série temporal de velocidade do vento e que o filtro SSA com uso da Clusterização Hierárqueica melhora a capacidade preditiva do modelo. Também pode-se concluir que o uso da técnica *Cross Validation*, melhora a acurácia das previsões no caso estudado sem ou com o uso da filtragem SSA.

Para estudos futuros, sugere-se testar e comparar tecnicas como *Walk-forward Validation* com o *Cross Validation* para verificar suas eficácia em outras séries temporais.

Agradecimentos

Os autores agradecem à Sociedade Brasileira de Pesquisa Operacional (SOBRAPO).

Referências

Andrade, L. P. *Procedimento Interativo de Agrupamento de Dados*. . Dissertação de Mestrado. Programa de Mestrado de Engenharia Civil, Universidade Federal do Rio de Janeiro, 2004.

Box, G. E. P. e Jenkins, G. M. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1970.

Cardoso, M. M. *Simulação de Modelos GARCH para Séries Temporais Univariadas de Demanda de Energia Elétrica para Consumidores Livres em Regime de Curto Prazo*. 173p. Dissertação de Mestrado. Programa de Mestrado em Engenharia de Produção, Universidade Federal de Itajubá, 2005.

Dalmaz, A. *Estudo do Potencial Eólico e Previsão de Ventos para Geração de Eletricidade*

em Santa Catarina. 193p. Dissertação de Mestrado. Programa de Mestrado em Engenharia de Produção, Universidade Federal de Santa Catarina, 2007.

Elsner, J. B. e Tsonis, A. *Singular Spectrum Analysis. A New Tool in Time Series Analysis*. New York: Plenum Press, 1996.

Golyandina, N., Nekrutkin, V., e Zhigljavsky, A. *Analysis of Time Series Structure: SSA and Related Techniques*. New York: Chapman & Hall / CRC, 2001.

Hamilton, J. D. *Time Series Analysis*. Princeton, NJ: Princeton University Press, 1994.

Hassani, H., Heravi, S., e Zhigljavsky, A. Forecasting UK industrial production with multivariate singular spectrum analysis. In: *Annals of 2012 International Conference on the Singular Spectrum Analysis and its Application*. Beijing, 2012.

Hyndman, R. J. e Athanasopoulos, G. *Forecasting: principles and practice, 2nd edition*. Melbourne, Australia: OTexts, 2018.

Hyndman, R. J., Bergmeir, C., e Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, v. 120, 2018.

Hyndman, R. J., Coehlher, A. B., Ord, J. K., Snyder, R. D., e Grose, S. A. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, v. 18, n. 3, 2002.

Menezes, M. L., Cassiano, K. M., Souza, R. M., Jr, L. A. T., Pessanha, J. F., e Souza, R. C. Modelagem e previsão de demanda de energia com filtragem ssa. *Revista da Estatística da UFOP*, v. 3, n. 2, 2014.

Morettin, P. A. e Toloi, C. M. C. *Análise de Series Temporais*. São Paulo, SP: Editora Blucher, 2018.

Santos, J. M. A. e Menezes, M. L. Análise e previsão de precipitação pluviométrica sob a abordagem MSSA. *Pesquisa Operacional para o Desenvolvimento*, v. 14, 2021.

Tomasquim, M. T. Perspectiva e planejamento do setor energético no brasil. *Revista de Estudos Avançados da USP*, v. 26, n. 74, 2012.