Pesquisa Operacional para o Desenvolvimento
Vol. 10, n. 2, p. 93-107, 2018

ISSN 1984-3534

ANALISANDO O DESEMPENHO DA REMOÇÃO DE RUÍDOS DE UMA SÉRIE TEMPORAL DE VAZÃO DE AFLUENTES USANDO DBSCAN NA ABORDAGEM ANÁLISE ESPECTRAL SINGULAR[1]

Keila Mara Cassiano[a], Moisés Lima de Menezes[a]*

[a]Instituto de Matemática e Estatística, Departamento de Estatística,
Universidade Federal Fluminense - UFF, Niterói-RJ, Brasil

Recebido 24/05/2018, aceito 31/10/2018

RESUMO

O objetivo deste artigo é apresentar diferentes métodos para remoção de ruídos de séries temporais com o uso de Análise Espectral Singular (SSA - *Singular Spectrum Analysis*) e verificar o desempenho da Clusterização Baseada em Densidade em Aplicações com Ruído (DBSCAN) perante os demais. Para este propósito foram utilizadas quatro abordagens na fase de agrupamento do método SSA: análise de componentes principais (ACP), análise de agrupamentos integrada com ACP, análise gráfica dos vetores singulares e DBSCAN. Adicionalmente, testes estatísticos foram realizados a fim de se obterem evidências empíricas da existência de independência estatística e estacionariedade de segunda ordem na série temporal de ruídos removidos. Para ilustrar a aplicação dos métodos, considerou-se a série temporal de Vazão de Afluentes da Usina Hidrelétrica Governador Bento Munhoz, localizada na Bacia do Rio Paraná, Brasil.

Palavras-chave: Remoção de ruídos, Análise Espectral Singular, DBSCAN, Séries temporais.

ABSTRACT

The aim of this paper is to present different methods to remove noise from time series using the Singular Spectrum Analysis (SSA) and verify performance of Density Based Spatial Clustering of Applications with Noise (DBSCAN) before others. For this purpose, four approaches were used in the grouping step of the method SSA: Principal Component Analysis (PCA), Clustering Analysis integrated with PCA, Graphical Analysis of Singular Eigenvectors and DBSCAN. In addition, statistical tests were performed in order to empirically demonstrate statistical independence and second-order stationarity in the time series of noise removed. To illustrate the application of methods, we considered the time series of affluent flow of the Governor Bento Munhoz Hydroelectric Plant, located on the Paraná River Basin.

Keywords: Noise removal, Singular Spectrum Analysis, DBSCAN, Time series.

*Autor para correspondência. E-mail: moises_lima@msn.com
DOI: 10.4322/PODes.2018.007

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

1. Introdução

Uma das principais características do Sistema Elétrico Brasileiro (SEB) reside no fato da sua capacidade de geração ser predominantemente hidráulica. Não obstante, devido às incertezas nos regimes das *vazões naturais*, o SEB está submetido a um significativo risco hidrológico (Terry et al., 1986). Para mitigá-lo, o SEB conta com usinas termoelétricas que complementam a geração hidroelétrica no país. Adicionalmente, o sistema ainda dispõe de usinas hidroelétricas com reservatórios de grande capacidade de acumulação que permitem a regularização plurianual e, desta forma, protegem a geração de energia elétrica dos efeitos decorrentes de longos períodos secos. Para que estes recursos sejam utilizados de forma sustentável e contribuam à economicidade e à segurança do fornecimento de energia elétrica, a operação do parque termelétrico e das usinas hidroelétricas deve ser realizada de forma coordenada, a fim de se alcançar um equilíbrio entre o melhor uso da água e a minimização das despesas com combustíveis nas unidades termoelétricas. Tal coordenação é realizada por meio de uma cadeia de modelos de otimização e simulação que suportam os processos de tomada de decisão no planejamento e na programação da operação do Sistema Interligado Nacional (SIN) (Terry et al., 1986). Os resultados destes modelos são sensíveis às previsões das séries temporais de vazões nos aproveitamentos hidroelétricos. No entanto, estas séries temporais são caracterizadas por acentuada sazonalidade e incertezas (erros e imprecisões) na sua mensuração.

Uma forma factível de lidar com a sazonalidade e imprecisões (presentes nas séries de vazões) é o uso de Análise Espectral Singular (SSA). SSA é um método eficiente na extração e reconstrução de componentes periódicas e não periódicas de séries temporais com elevados níveis de ruído (Hassani, 2007). Portanto, útil no processamento de séries temporais de vazões.

Menezes et al. (2015) aplicam SSA para gerar previsão de consumo de energia elétrica e Teixeira Jr. et al. (2013) fazem uma combinação geométrica de métodos que incluem SSA. Em todos os casos, a abordagem SSA se mostrou mais eficiente que a modelagem sem o uso desta ferramenta.

Por meio do método SSA, uma matriz trajetória pode ser obtida a partir de uma série temporal original e ser expandida em termos da decomposição em valores singulares (Hassani, 2007). De acordo com Golyandina et al. (2001), cada componente desta expansão concentra uma parcela da energia contida na matriz trajetória. Dessa forma, um subconjunto de componentes concentra a maior parte da energia total com estrutura de dependência temporal e sazonalidade, enquanto que as componentes restantes concentram a parte da energia sem qualquer estrutura de dependência temporal ou informação (isto é, são constituídas apenas de ruído). Assim sendo, com o uso de algum método de seleção de componentes, é possível realizar a separação de tais componentes em dois grupos: um contendo as componentes que detêm a estrutura de dependência temporal e outro com as componentes que detêm apenas ruído. A soma das componentes que concentram a estrutura de dependência temporal gera uma versão aproximada e menos ruidosa da série temporal original. Isto é, por meio do método SSA e do método de seleção de componentes SSA, pode-se remover parte do ruído presente na série temporal original.

Nesta perspectiva, são apresentadas quatro diferentes abordagens para remoção de ruídos de séries de tempo com o uso do método SSA: análise de componentes principais (ACP), análise de agrupamentos integrada com ACP, análise gráfica dos vetores singulares e Clusterização Baseada em Densidade em Aplicações com Ruído (DBSCAN) (Tran et al., 2013). Os três primeiros métodos já mostraram eficácia quanto à remoção de ruídos quando comparados a outros métodos, como apresentados em Hassani (2007), Menezes et al. (2105) e Golyandina et al. (2001) enquanto o DBSCAN é alvo de estudos na implementação deste método em SSA. Adicionalmente, foram realizados testes estatísticos sob a série temporal de resíduos extraída de cada abordagem a fim de garantir, estatisticamente, a existência de independência e estacionariedade de segunda ordem (Hamilton, 1994), que são atributos das séries temporais de ruídos. Incialmente, são considerados processos estocásticos simulados para a avaliação do desempenho das quatro abordagens e, para a ilustração aplicada, considerou-se a

série temporal de Vazão da Usina Hidrelétrica Governador Bento Munhoz, localizada na Bacia do Rio Paraná, em Pinhão, estado do Paraná, Brasil.

O artigo está organizado em oito seções. Na Seção 2, tem-se uma breve apresentação do método SSA. Na Seção 3 são descritos os métodos para seleção das componentes, responsáveis pela separação das componentes SSA entre as categorias sinal e ruído. A metodologia utilizada está na Seção 4. Na Seção 5 são apresentados resultados simulados. Os dados relativos ao estudo de caso são descritos na Seção 6. Os principais resultados obtidos são apresentados na Seção 7. Por fim, as conclusões são apresentadas na Seção 8.

2. Análise Espectral Singular

SSA é um método de processamento de sinais que pode ser utilizado, dentre outras aplicações, na remoção de ruído de séries de tempo (Golyandina et al., 2001). A versão básica do método SSA pode ser dividida em duas etapas: decomposição e reconstrução.

2.1. Decomposição

A etapa da decomposição pode ser subdividida em duas fases: incorporação e decomposição em valores singulares (SVD - *Singular Value Decomposition*).

Seja $Y_T = [y_1, \dots, y_T]'$ uma *série temporal* de comprimento T. Por *incorporação*, entende-se como sendo um procedimento no qual uma série temporal Y_T é transportada a uma matriz $X = [X_1, \dots, X_K]_{L\times K}$, em que $X_k = [y_k, \dots, y_{k+L-1}]'$, para todo $k = 1, \dots, K$. A matriz X é conhecida como *matriz trajetória*, L $(2 \leq L \leq T)$ é um parâmetro estimado chamado *comprimento da janela da matriz trajetória* e $K = T - L + 1$ (Golyandina et al., 2001).

É possível decompor a matriz trajetória X como expresso em (1).

$$X = X_1 + X_2 + \cdots + X_L, \tag{1}$$

na qual $X_l = {\lambda_l}^{1/2} U_l V_l', l = 1, \dots, L$ e os conjuntos $\left\{{\lambda_l}^{1/2}\right\}_{l=1}^{L}$ e $\{U_l\}_{l=1}^{L}$ são, respectivamente, denominados por *espectro singular* e *vetores singulares* da matriz trajetória X. A coleção (λ_l, U_l, V_l) é conhecida como *autotripla na* SVD da matriz trajetória X. A contribuição de cada componente em (1) pode ser mensurada pela razão de valores singulares, dada por $(\lambda_l)^{1/2}/\sum_{l=1}^{L}(\lambda_l)^{1/2}$. Considere que d seja o *posto* (isto é, o número de autovalores não nulos) da matriz trajetória X. Segue que a identidade descrita em (1) pode ser reescrita tal como:

$$X = X_1 + X_2 + \cdots + X_d, \text{ onde } d \leq L. \tag{2}$$

Os detalhes sobre esta decomposição podem ser obtidos em Golyandina et al. (2001).

2.2. Reconstrução

A etapa de reconstrução pode ser subdividida em duas fases: *agrupamento* e *média diagonal*. A fase de *agrupamento* consiste no procedimento de agrupar algumas sequências de matrizes elementares resultantes da decomposição SVD em grupos *disjuntos* e, após isso, somá-las, gerando novas matrizes elementares (Elsner e Tsonis, 2010).

Considere a sequência $\{X_l\}_{l=1}^{d}$ de matrizes elementares na SVD, em (2). Agrupe-as em $m \leq d$ grupos *disjuntos* utilizando algum método, por exemplo, com o auxílio da *análise de componentes principais (ACP)* (Seção 3.1), ou *análise de agrupamentos integrada com ACP* (Seção 3.2), *análise gráfica de vetores singulares* (Seção 3.3) ou *DBSCAN* (Seção 3,4) e assuma que, após o agrupamento, o conjunto de índices gerado é dado por $\{I_1, \dots, I_m\}$, onde, para todo i, $I_i = \left\{I_{i1}, \dots, I_{ip_i}\right\}$ e p_i é a cardinalidade do grupo I_i. Note que $\{X_l\}_{l=1}^{d} = \bigcup_{i=1}^{m}\left\{X_{I_{ij}}\right\}_{j=1}^{p_i}$, em

que $m \leq d$. A matriz elementar X_{I_i} gerada a partir do grupo $\left\{X_{I_{ij}}\right\}_{j=1}^{p_i}$ é dada por $X_{I_i} = \sum_{j=1}^{p_i} X_{I_{ij}}$, de modo que a identidade em (2) pode ser reescrita como em (3).

$$X = X_{I_1} + X_{I_2} + \cdots + X_{I_m} \tag{3}$$

Para o caso particular da abordagem SSA neste artigo, $m = 2$ para os métodos que incluem ACP e DBSCAN e $m = 3$ nos métodos de Análise de Agrupamentos e Análise Gráfica dos Vetores Singulares.

Após a fase de agrupamento, as matrizes agrupadas em (3) retornam ao formato de série temporal via Média Diagonal a partir dos resultados obtidos em Golyandina et al. (2001). O número m de matrizes obtido na fase de agrupamento passa a ser o número de séries obtidas após a média diagonal, sendo uma delas classificada estatisticamente como ruído a ser removido.

2.3. Separabilidade

Por meio da separabilidade, é possível verificar estatisticamente se as duas componentes SSA estão bem separadas em termos de dependência linear. Por correlação ponderada entende-se como sendo a função que quantifica a dependência linear entre duas componentes SSA $Y_t^{(i)}$e $Y_t^{(j)}$. O estimador da correlação ponderada é definido em (4):

$$\rho_{ij}^{(w)} = \frac{\left(Y_t^{(i)}, Y_t^{(j)}\right)_w}{\left\|Y_t^{(i)}\right\|_w \left\|Y_t^{(j)}\right\|_w}, \tag{4}$$

Em que $\| \ \|$ é a norma euclidiana, $(\ )_w$ é o produto interno tal que: $\left\|Y_t^{(i)}\right\|_w = \sqrt{\left(Y_t^{(i)}, Y_t^{(i)}\right)_w}$ e $\left(Y_t^{(i)}, Y_t^{(j)}\right)_w = \sum_{k=1}^{T} w_k y_k^{(i)} y_k^{(j)}$; e $w_k = min\{k, L, T-k\}$.

Se o valor absoluto $\rho_{ij}^{(w)}$ 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 para geração de novas matrizes elementares na SVD (Golyandina et al., 2001).

3. Métodos de Remoção de Ruídos

No contexto de séries de tempo, os *métodos de remoção de ruído* consistem, basicamente, em gerar uma série temporal aproximada $[\tilde{y}_t]_{1\times T}$ que seja menos ruidosa que a série temporal original $[y_t]_{1\times T}$. Neste contexto, tem-se que qualquer série temporal pode ser decomposta em duas componentes: uma com estrutura de dependência temporal (*linear* ou *não linear*) e outra sem qualquer estrutura de dependência no tempo (que é conhecida como série temporal de ruídos). Tal decomposição é dada em (5).

$$[y_t]_{1\times T} = [\tilde{y}_t]_{1\times T} + [\varepsilon_t]_{1\times T}, \tag{5}$$

na qual $[y_t]_{1\times T}$ é a série temporal original, $[\tilde{y}_t]_{1\times T}$ é a série temporal aproximada e $[\varepsilon_t]_{1\times T}$ é o ruído.

Neste artigo, são utilizados quatro métodos com esta finalidade: análise de componentes principais (ACP), análise de agrupamentos integrada com ACP, análise gráfica dos vetores singulares e DBSCAN.

3.1. Análise de Componentes Principais

Cada valor singular λ_l resultante da SVD quantifica a energia da matriz trajetória X que está contida na matriz elementar X_l. Seja $X = \sum_{i=1}^{m} X_{I_i}$, por meio da *análise de componentes principais* (ACP), pretende-se determinar um valor ótimo N no conjunto de índices $\{1, \dots, m\}$, de tal forma que a série temporal $[y_t]_{1\times T} - \sum_{i=1}^{N} \left[y_t^{(i)}\right]_{1\times T}$ seja estatisticamente classificada como ruído (Manly, 2008).

3.2. Análise de Agrupamentos

Na literatura, podem ser encontrados diferentes métodos de *análise de agrupamentos* (ou *cluster analysis*) que consistem em *métodos de classificação não supervisionados* usados para encontrar uma estrutura natural de agrupamentos em objetos multidimensionais. De acordo com Aldenderfer e Blashfield (1984), a *análise de agrupamentos* visa a agrupar um conjunto com N objetos em K *cluster*s mutuamente *excludentes*, de tal forma que os objetos em um mesmo *cluster* apresentem *similaridades* entre si e *dissimilaridades* em relação aos objetos pertencentes aos outros *clusters*.

Os vetores singulares resultantes na SVD podem apresentar perfis semelhantes, de modo que podem ser agrupados por meio de *análise de agrupamentos*. Qualquer um dos métodos de análise de agrupamento pode ser utilizado na classificação dos vetores singulares na SVD. Neste artigo, foi o utilizado o *método de agrupamento hierárquico*, em virtude da sua simplicidade. Os métodos hierárquicos agrupam um conjunto de N objetos sequencialmente em 2, 3, 4 até $N - 1$ grupos, obtendo no final uma estrutura em árvore (Aldenderfer e Blashfield, 1984).

No estudo de caso, foi adotado o método de *agrupamento hierárquico de encadeamento simples* ou (*single-linkage*) para se dividir o conjunto de vetores obtidos pela expansão SVD em três grupos excludentes, sendo eles: tendência, harmônicos e ruídos. Os dois primeiros formam a série aproximada e o último é o ruído a ser removido. O algoritmo utilizado foi o *aglomerativo*, ou seja, o algoritmo inicia com N *cluster*s, cada um contendo apenas um vetor, e estes são sucessivamente fundidos dois a dois por meio de um procedimento interativo até que restem apenas dois *cluster*s. Em cada estágio do processo de aglomeração, o conjunto de N objetos é agrupado em um determinado número de grupos e a distância entre estes é calculada.

3.3. Análise Gráfica dos Vetores Singulares

A análise das coordenadas da série temporal na base definida pelos vetores singulares resultantes da SVD permite identificar as componentes de tendência e da sazonalidade da série. O problema geral aqui consiste em identificar e separar as componentes oscilatórias das componentes que fazem parte da tendência. De acordo com Golyandina et al. (2001), a análise gráfica de tais coordenadas aos pares permite identificar por meio visual as componentes harmônicas da série.

Considere um harmônico puro com frequência igual a ω, fase igual a δ, amplitude igual a ξ e período $\rho = \frac{1}{\omega}$ definido como um divisor do tamanho da janela L e K. Se o parâmetro ρ assume um valor inteiro, então ρ é classificado como *período do harmônico* (Morettin e Toloi, 2006). As coordenadas da série temporal em duas componentes ortogonais podem ser dispostas em um diagrama de dispersão de modo que, se a figura formada tem um formato de um polígono regular, o número de lados do polígono é o período das componentes harmônicas associadas.

Na ocasião da análise gráfica, os vetores singulares que apresentarem comportamento harmônico (seja na forma senoidal, seja na forma de polígonos em diagramas de dispersão) e de tendência (a partir dos vetores com comportamento mais suaves) se juntam para formar a série temporal aproximada, enquanto os demais são classificados como ruídos.

3.4. DBSCAN

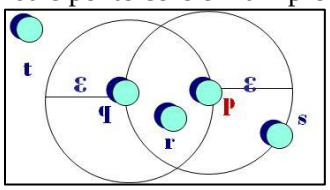
DBSCAN é o principal representante dos métodos de clusterização baseados em densidade e tem a qualificação de identificar clusters de formato arbitrário e separar eficientemente os ruídos dos dados. A versão revista e atualizada do DBSCAN, utilizada neste trabalho, foi apresentada por Tran et al. (2013) e tem um desempenho robusto para conjuntos de dados contendo estruturas densas com aglomerados conectados. Os resultados da clusterização não dependem da ordem em que os objetos são processados e a versão atualizada acabou com o problema de pertinência objeto na vizinhança de clusters densos e próximos. As definições a seguir caracterizam o método DBSCAN. Seja *D* uma base de dados de pontos, as seguintes definições são válidas:

Definição 1: (*Eps* - vizinhança de um ponto *p*) É a vizinhança de um objeto *p* com raio *Eps* dada por: $N_{Eps}(p) = \{q \in D | dist(p,q) < Eps\}$

Definição 2: (Ponto core) Se a vizinhança N_{Eps} de um objeto contém ao menos um número mínimo, $MinPts$, de objetos, então o objeto *p* é chamado interno ou **ponto core**.

Definição 3: (Ponto de borda) Se a vizinhança N_{Eps} de um objeto contém menos que *MinPts* mas contém algum ponto core, então o objeto é chamado de **ponto de borda**. Na Figura 1, para $MinPts = 4$, p é o único ponto core, enquanto q, r e s são pontos de borda.

Figura 1: Pontos de borda e ponto core em um processo de clusterização.

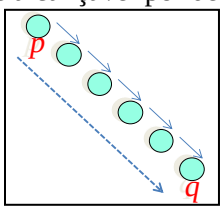


Fonte: Elaborada pelos autores.

Definição 4: (Alcance direto por densidade) Um objeto p é alcançável por densidade diretamente do objeto *q*, se *p* está na vizinhança Eps de *q*, e q é um core.

Definição 5: (Alcance por densidade) Um objeto *p* é alcançável por densidade do objeto *q* com respeito a *Eps* e *MinPts* em um conjunto *D*, se existe uma cadeia de objetos $\{p_1, \ldots, p_n\}$, tais que $p_1 = q$ e $p_n = p$ e $p_i + 1$ é alcançável por densidade diretamente de p_i com respeito a Eps e *MinPts*, para $1 \leq i \leq n$, $p_i \in D$. Há, portanto, um fechamento transitivo do alcance por densidade. A Figura 2 ilustra um objeto alcançável por densidade de outro objeto.

Figura 2: O objeto q é alcançável por densidade pelo objeto p.



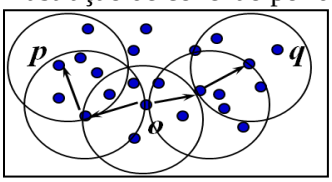
Fonte: Elaborada pelos autores.

Na Figura 1, q é diretamente alcançável por densidade a partir de p mas p não é diretamente alcançável pela densidade de q porque q não é um ponto core.

Definição 6: (Conexão por densidade) Um objeto *p* é conectado por densidade ao objeto *q* com respeito a *Eps* e *MinPts* em um conjunto de objetos *D*, se existe um objeto *r* em *D* tal que ambos

p e q são alcançáveis por densidade do objeto r com respeito a *Eps* e *MinPts*. A Figura 3 ilustra a conexão por densidade: um ponto p é conectado por densidade a um ponto q em relação a Eps e $MinPts$ se houver um ponto O tal que tanto p quanto q sejam alcançáveis por densidade de O em relação a Eps e $MinPts$.

Figura 3: Ilustração de conexão por densidade.



Fonte: Elaborada pelos autores.

Definição 7: (Cluster DBSCAN) Um cluster com respeito a e *MinPts* é um conjunto não vazio e satisfazendo as seguintes condições:
(Maximilidade) $\forall\, p, q$: se $p \in C$ $(Cluster)$ e q é alcançável por densidade de com respeito a *Eps* e *MinPts*. Então $q \in C$.
(Conectividade) $\forall\, p, q \in C$, p é conectado por densidade a q com respeito a *Eps* e *MinPts.* Em outras palavras, um cluster DBSCAN o conjunto de pontos conectados por densidade que é maximal com respeito a alcançabilidade por densidade. E um cluster DBSCAN é inequivocamente determinado por qualquer de seus centros (Tran et al., 2013).

Definição 8: (Ruído): Sejam $C_1, C_2, \ldots, C_k$, clusters do conjunto de dados *D* com respeito a *Eps* e *MinPts.* Se um ponto p não pertence a nenhum destes k clusters, ele é um ruído. Em outras palavras ruídos são pontos que não são diretamente alcançados por algum ponto core.

O método DBSCAN encontra clusters verificando a vizinhança *Eps* de cada ponto na base de dados, começando por um objeto arbitrário. Se a vizinhança *Eps* de um ponto *p* contém mais do que *MinPts*, um novo cluster com *p* como um centro é criado. O método DBSCAN, então, iterativamente coleta objetos alcançáveis por densidade diretamente destes centros, que pode envolver a união de alguns clusters alcançáveis por densidade. O processo termina quando nenhum novo ponto pode será adicionado a qualquer cluster. Para o algoritmo DBSCAN assim definido, quaisquer dois pontos core que são pertos suficientes com distância menor ou igual a *Eps* são colocados no mesmo cluster. Qualquer ponto de borda que está perto de um ponto core é colocado no mesmo cluster do ponto core. Pontos de ruído, ou seja, pontos que não são diretamente atingíveis por algum ponto core são descartados.

4. Metodologia

A metodologia a ser considerada no processo de execução deste artigo consiste em remover a parte ruidosa de uma série temporal original de Vazão de Afluentes usando uma decomposição em sinal e ruído via SSA de modo que na sua fase de agrupamento sejam aplicadas as quatro metodologias descritas e, após este processo, sejam testadas as partes ruidosas removidas via teste BDS (Brock et al., 1996) e *Ljung-Box* (Ljung e Box, 1978) para verificar a independência dos dados e teste de Dickey-Fuller (Dickey e Fuller, 1979) para a estacionariedade e a correlação ponderada, utilizada para avaliar se a decomposição está adequada, ou seja, se não há ruído no sinal nem sinal no ruído.

5. Experimentos Computacionais

Com o objetivo de avaliar a eficácia do método DBSCAN para séries de distintas características em relação aos demais métodos de agrupamento em SSA, foram simulados processo estacionários e não estacionários via simulação de Monte Carlo utilizando o MATLAB

(2010) com erros normalmente distribuídos com média zero e variância igual a 1. Os processos foram filtrados via SSA com os quatro métodos, modelados e o RMSE (*Root Mean Square Error*) foi obtido para cada caso e para o processo original. Os processos simulados foram:

Y_1: $y_t = \varepsilon_t$; $\varepsilon_0 = 0$;
Y_2: $AR(1)$; $\phi_1 = 0{,}4$;
Y_3: $MA(2)$; $\theta_1 = -0{,}3$ e $\theta_2 = 0{,}8$;
Y_4: $ARMA\ (1{,}2)$; $\phi_1 = 0{,}4$, $\theta_1 = -0{,}3$ $e\ \theta_2 = 0{,}8$;
Y_5: $Passeio\ Aleatório\ com\ drift$ $\mu = 0{,}1$;
Y_6: $Passeio\ Aleatório\ com\ drift$ $\mu = 0{,}6$;
Y_7: $ARIMA\ (0{,}1{,}1)$; $\theta_1 = 0{,}4$;
Y_8: $ARIMA\ (1{,}1{,}2)$; $\phi_1 = 0{,}4$, $\theta_1 = -0{,}3$ e $\theta_2 = 0{,}8$.

Os primeiros 4 processos gerados $Y_i, i = 1, \ldots, 4$ são processos estacionários e os últimos $Y_i, i = 5, \ldots, 8$, são processos não estacionários. Os resultados obtidos com o teste de *Dickey-Fuler* e a correlação ponderada estão na Tabela 1.

Tabela 1: Teste de *Dickey-Fuller* (ADF) e correlação ponderada para os processos simulados.

Processo Estocástico	Teste de *Dickey-Fuller* (ADF) e Correlação ponderada	Método				
		SSA + Análise Gráfica	SSA + ACP	SSA + Agrupamento com ACP	SSA + DBSCAN	Original
Y_1	ADF	-12,546	-11,200	-13,025	**-13,004**	-12,853
	Correlação	0,00087	0,00070	0,00084	0,00060	0,00051
Y_2	ADF	-12,032	-11,998	-12,033	**-13,202**	-12,224
	Correlação	0,00045	0,00091	0,00078	**0,00044**	0,00028
Y_3	ADF	-13,255	-12,288	-11,336	**-13,247**	-12,032
	Correlação	0,00088	0,00094	0,00023	**0,00072**	0,00035
Y_4	ADF	-12,333	-12,655	-13,360	**-12,008**	-11,540
	Correlação	0,00020	0,00039	0,00088	**0,00024**	0,00081
Y_5	ADF	-12,004	-12,000	-13,002	**-12,254**	-11,335
	Correlação	0,00029	0,00052	0,00087	**0,00033**	0,00080
Y_6	ADF	-13,335	-13,278	-13,697	**-11,298**	-12,672
	Correlação	0,00077	0,00072	0,00058	**0,00068**	0,00065
Y_7	ADF	-12,999	-12,668	-13,770	**-13,669**	-13,885
	Correlação	0,00022	0,00037	0,00031	**0,00054**	0,00099
Y_8	ADF	-11,379	-13,950	-12,987	**-13,089**	-13,879
	Correlação	0,00039	0,00055	0,00074	**0,00025**	0,00022

Fonte: Elaborada pelos autores.

Os resultados apresentados na Tabela 1 mostram que as séries de ruídos removidas possuem características de estacionariedade e que o agrupamento usando os quatro métodos obteve uma boa separabilidade, indicando que não há parte de sinal na série de ruído nem ruído na série filtrada obtida. Estes resultados mostram que os métodos são adequados nos diversos processos propostos. A estatística RMSE encontra-se na Tabela 2.

Tabela 2: RMSE (*In Sample*) para os processos simulados.

Processo Estocástico	Método				
	SSA + Análise Gráfica	SSA + ACP	SSA + Agrupamento com ACP	SSA + DBSCAN	Original
Y_1	0,954	0,654	0,988	**0,555**	1,025
Y_2	0,741	0,511	0,635	**0,312**	0,958
Y_3	0,601	0,900	0,801	**0,247**	0,994
Y_4	0,825	0,555	0,562	**0,451**	1,048
Y_5	3,001	2,781	2,199	**2,101**	4,803
Y_6	2,478	3,000	2,589	**1,549**	3,758
Y_7	3,999	3,709	3,747	**2,864**	4,281
Y_8	2,574	2,989	2,951	**1,921**	3,494

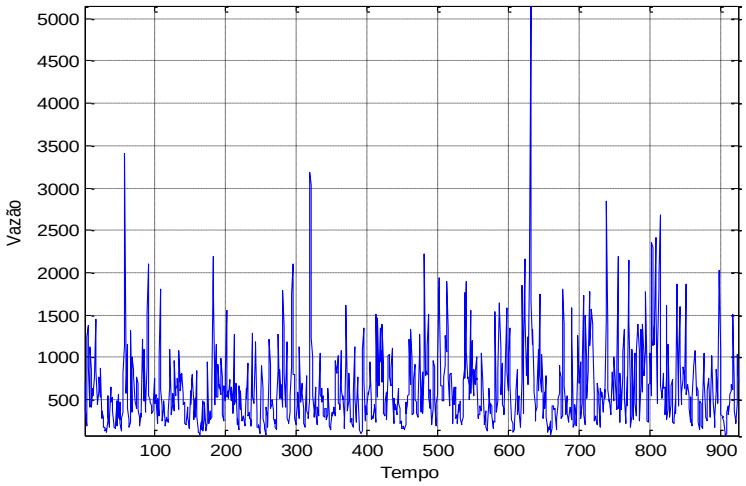
Fonte: Elaborada pelos autores.

Os resultados apresentados na Tabela 2 mostram que, para os processos simulados, a remoção da parte ruidosa foi eficaz com os métodos abordados uma vez que minimizam os erros de previsão no ajuste em relação ao processo sem tal remoção. Também pode-se perceber que o DBSCAN tem desempenho melhor que os demais métodos em SSA sendo, portanto, uma ferramenta que pode ser utilizada nesta abordagem.

6. Estudo de Caso

Localizada na Bacia do Rio Paraná, a Usina Hidrelétrica Governador Bento Munhoz da Rocha Neto (UHE GB Munhoz) foi construída ao longo do Rio Iguaçu, no município de Pinhão, e fica a 5 km da jusante da foz do Rio Areia e a 240 km de Curitiba. A série temporal mensal das *médias diárias* de vazão da UHE GB Munhoz tem cardinalidade igual a 924 meses (de janeiro de 1931 a dezembro de 2007) e está apresentada na Figura 4.

Figura 4: Série Temporal mensal de Vazão (m^3/s) da UHE GB Munhoz.



Fonte: Operador nacional do Sistema Elétrico (ONS).

7. Resultados e Discussões

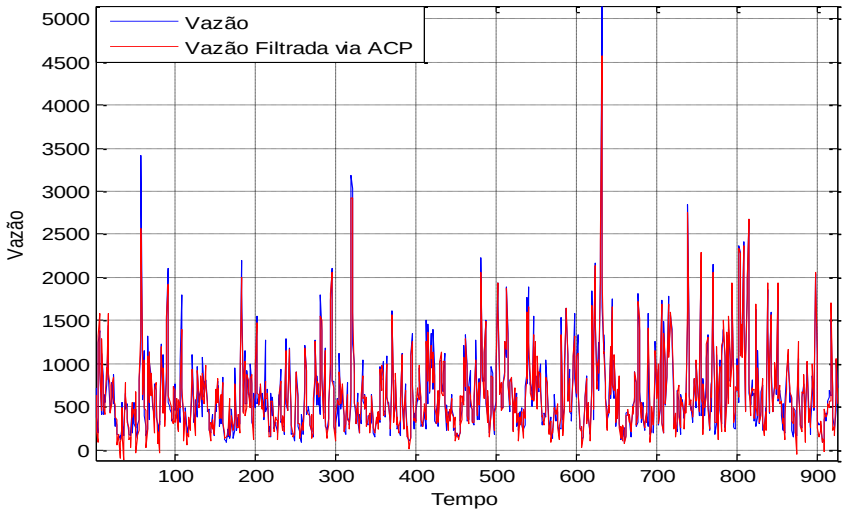
Os métodos de análise de componentes principais (ACP), de análise de agrupamento integrada com ACP, de análise gráfica de vetores singulares e DBSCAN foram utilizados na fase de agrupamento do método SSA com a finalidade de se remover as matrizes elementares

que geram componentes SSA estatisticamente classificadas com ruídos pelos testes BDS (Brock et al., 1996) e *Ljung-Box* (Ljung e Box, 1978).

7.1. Remoção de Ruídos via Análise de Componentes Principais

A análise de componentes principais (ACP) sobre SVD foi implementada no *software* MATLAB. O valor ótimo para o parâmetro *L* (comprimento da janela na fase de decomposição do método SSA) foi igual a 362 (ou seja, o número de vetores singulares na SVD foi igual a 362). O valor ótimo de truncamento *N* foi igual a 201 (ou seja, dos 362 vetores singulares na SVD, foram mantidos na SVD somente os 201 primeiros vetores singulares; enquanto que os outros foram classificados como ruído e removidos). O método utilizado para obtenção dos valores ótimos dos parâmetros *L* e *N* foi o método de tentativa e erro. Na Figura 5 têm-se os gráficos sobrepostos da série temporal de vazão da UHE GB Munhoz (original) e sua aproximação (ou versão filtrada) através do método de ACP na SVD.

Figura 5: Série Temporal de Vazão da UHE GB Munhoz e sua Versão Aproximada através da ACP.



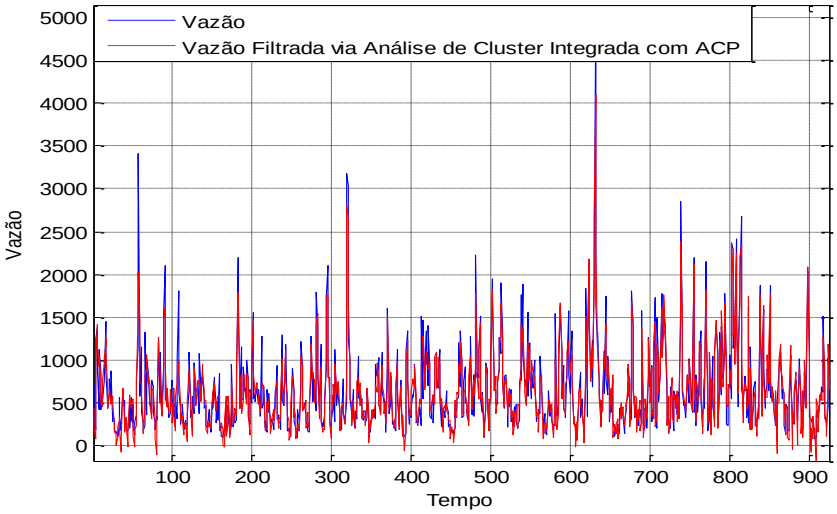
Fonte: Elaborada pelos autores.

É possível perceber, na Figura 5, que parte da energia (classificada estatisticamente como ruído) foi removida por ACP, mas a série aproximada resultante manteve o processo gerador da série original.

7.2. Remoção de Ruídos via Análise de Agrupamentos Integrada com ACP

A ACP foi implementada no *software* MATLAB. Os valores ótimos para os parâmetros *L* e *N* foram, respectivamente, iguais a 362 e 201 (os mesmos valores da Seção 5.1). Após a ACP, foram tomados os vetores singulares remanescentes na SVD e agrupados, por meio do método do agrupamento *hierárquico* (referenciado na Seção 3.3), em 3 *clusters* (*grupos*). Em cada *cluster* foi gerada uma componente, portanto, a série temporal da vazão da UHE GB Munhoz foi decomposta em 3 componentes. Os testes estatísticos BDS e *Ljung-Box* (apresentados na Tabela 1) foram aplicados às componentes SSA e verificou-se que a componente SSA 3 (oriunda do *cluster* 3) possui propriedades estatísticas de ruído. A análise de agrupamento foi implementada no *software* R, com o uso do pacote *Rssa*, e os testes estatísticos, no *software EViews*. A componente ruidosa foi removida e combinação das duas componentes restantes deu origem à série aproximada. A Figura 6 mostra a série original e a série filtrada via análise de agrupamento com ACP.

Figura 6: Série Temporal de Vazão da UHE GB Munhoz e sua Versão Filtrada através da Análise de Agrupamento Integrada com ACP.



Fonte: Elaborada pelos autores.

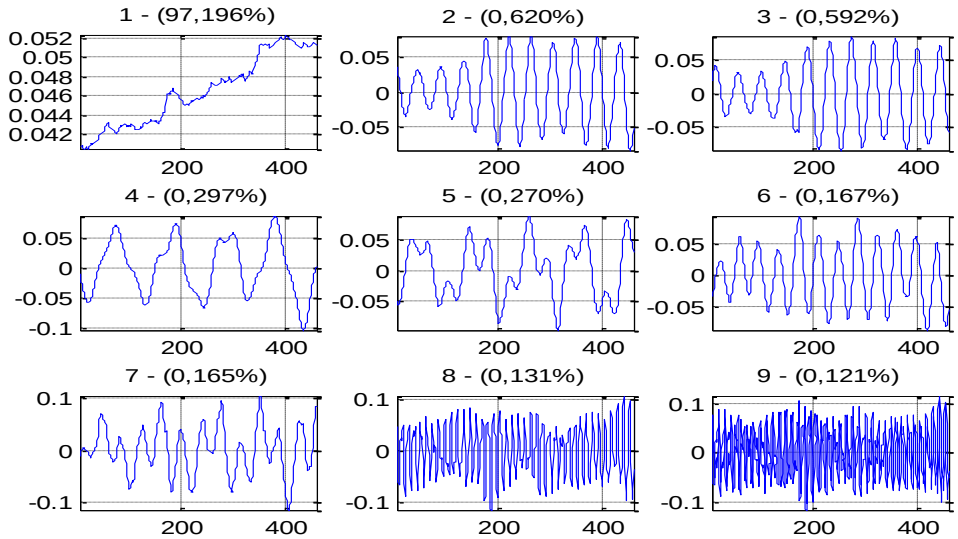
Na Figura 6 é possível visualizar que parte da energia (classificada estatisticamente como ruído) foi removida por meio da análise de agrupamento integrada com ACP.

7.3. Remoção de Ruídos via Análise Gráfica dos Vetores Singulares

Na abordagem de análise gráfica de vetores singulares na SVD, foi utilizado o valor para *L* igual a 462. Na Figura 7 têm-se os 9 primeiros (e principais) vetores singulares na SVD da matriz trajetória da série temporal de vazão da UHE GB Munhoz. Na primeira linha, da direita para a esquerda, têm-se os vetores singulares 1, 2 e 3. E assim sucessivamente. Esta abordagem foi implementada no *software Caterpillar SSA* (2013) e no MATLAB (2010).

É possível perceber, nas análises gráficas, que o vetor 1 pertence à componente de tendência, os vetores de 2 a 7 pertencem à componente harmônica e que os vetores 8 e 9 pertencem à componente ruidosa.

Figura 7: Os nove primeiros vetores singulares na SVD da matriz trajetória da série temporal de vazão da UHE GB Munhoz.

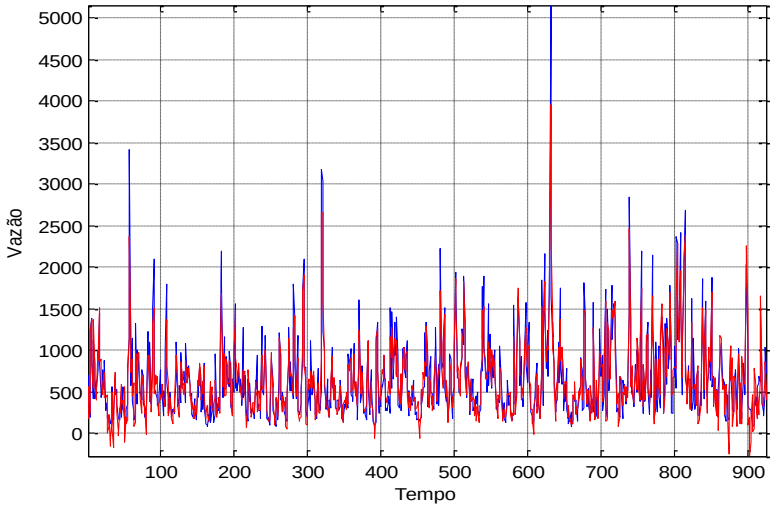


Fonte: Elaborada pelos autores.

Esta análise gráfica foi feita para todos os 462 vetores singulares na SVD e, após esta análise, as três componentes foram obtidas. Com isso, foi obtida a decomposição da série em três componentes: tendência, harmônica e ruído. Esta última, classificada como ruído (via estatísticas BDS e de *Ljung-Box*), foi removida.

Na Figura 8, tem-se a série temporal de vazão da UHE GB Munhoz e a sua versão filtrada através da análise gráfica dos vetores singulares.

Figura 8: Série temporal de vazão e sua aproximação através da análise gráfica de vetores singulares.



Fonte: Elaborada pelos autores.

Percebe-se, na Figura 8, que a parte ruidosa foi removida e que a série aproximada após a remoção do ruído manteve o comportamento da série original.

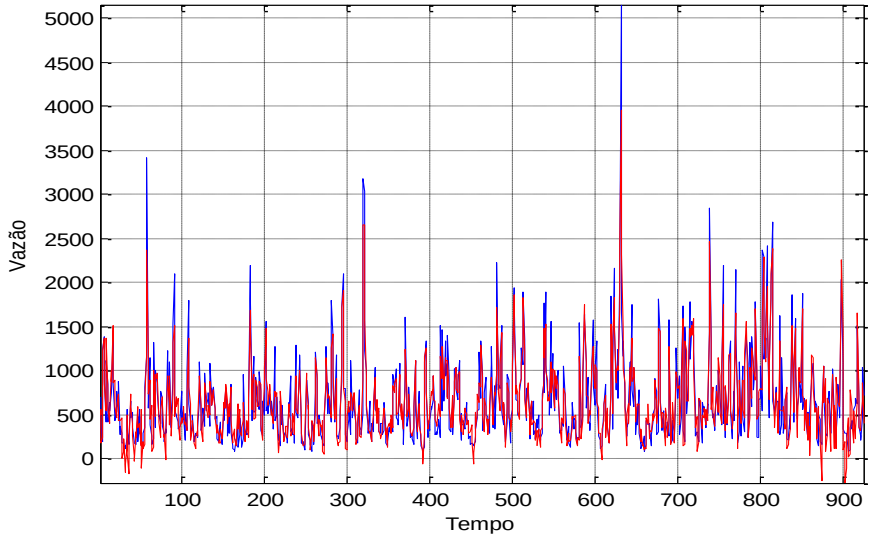
7.4. Remoção de Ruídos via DBSCAN

No processo de clusterização na fase de agrupamento SSA utilizando o DBSCAN, o software utilizado (R, a partir do pacote "*dbscan*" proposto por Hahsler et al. (2018) forneceu duas séries, sendo uma de sinal e outra de ruídos. A série de ruídos que foi obtida pelo conjunto dos pontos não atingíveis por algum ponto core foi removida e a série aproximada contendo os demais pontos foi obtida. A série de ruídos foi removida e a série remanescente foi comparada à série original. A Figura 9 apresenta a sobreposição da série original com a série aproximada via DBSCAN.

Assim como nos métodos anteriores, a Figura 9 apresenta a série aproximada após a remoção da parte ruidosa (em vermelho) que acompanha o desenvolvimento da série original (azul).

Os resultados dos testes BDS e Ljung-Box aplicados às séries de ruído dos quatro métodos estudados estão na Tabela 3. Em seguida, na Tabela 4, estão os resultados do teste de *Dickey-Fuller* (Dickey e Fuller, 1979) sobre as séries temporais de ruídos das quatro abordagens e de correlação ponderada $\rho_{1,2}^{(\omega)}$ entre as séries temporais de ruído oriundas das quatro abordagens (1) e a série temporal de vazão da UHE GB Munhoz aproximada (2). Espera-se que esta correlação seja a menor possível para se ter certeza de que não há sinal sendo excluído juntamente com o ruído e que não haverá ruído na série aproximada.

Figura 9: Série Temporal de Vazão e sua Aproximação através da abordagem DBSCAN.



Fonte: Elaborada pelos autores.

Tabela 3: Testes BDS e de *Ljung-Box*.

Método	***Teste BDS***				***Teste de Ljung-Box***				
	Dim.	*Estatística BDS*	*Estatística Z*	*Prob.*	*Lag*	*FAC*	*FACP*	*Estat.Q*	*Prob.*
ACP	2	-0,0000029	-0,31945	0,7949	1	0,0038	0,0038	1,3043	0,253
	3	-0,0000042	-0,20315	0,8390	2	-0,0941	-0,0985	5,9841	0,104
	4	-0,0000013	-0,03995	0,9681	3	-0,0925	-0,0922	8,1544	0,095
	5	-0,0000062	-0,12257	0,9024	4	-0,0812	-0,0865	6,2540	0,134
	6	-0,0000022	-0,03313	0,9736	5	0,0027	0,0028	2,0687	0,352
Análise de Agrupamento integrada com ACP	2	-0,001258	-0,54459	0,5860	1	-0,024	-0,024	0,5517	0,458
	3	0,000671	0,18361	0,8543	2	-0,052	-0,053	3,0656	0,216
	4	0,002181	0,50365	0,6145	3	-0,004	-0,007	3,0813	0,379
	5	0,004454	0,99176	0,3213	4	0,005	0,002	3,1046	0,540
	6	0,005500	1,27606	0,2019	5	0,048	0,047	5,2170	0,390
Análise Gráfica dos Vetores Singulares	2	-0,0000412	-0,510289	0,5849	1	0,0037	0,0029	1,3210	0,228
	3	-0,0000917	-0,410241	0,6580	2	0,0908	0,0921	5,2365	0,204
	4	-0,0000566	-0,144583	0,9025	3	-0,0360	-0,0326	4,1524	0,423
	5	-0,0000244	-0,045879	0,9287	4	-0,0021	-0,0015	3,2860	0,564
	6	-0,0000302	-0,482100	0,6024	5	0,0031	0,0033	2,5897	0,583
DBSCAN	2	-0,0004938	-0,913443	0,7865	1	0,0023	0,0150	5.4983	0,694
	3	-0,0001328	-0,670130	0,9710	2	0,0075	0,0098	3,9870	0,574
	4	-0,0002310	-0,193585	0,9467	3	-0,0034	-0,0045	6,9420	0,882
	5	-0,0007392	-0,127877	0,9861	4	-0,0012	-0,0077	4,4902	0,589
	6	-0,0002081	-0,123099	0,8603	5	0,0039	0,0069	7,6293	0,821

Fonte: Elaborada pelos autores.

A partir dos resultados apresentados na Tabela 3, pode-se concluir que as séries temporais de ruídos oriundas das quatro abordagens são independentes. Este resultado pode ser observado a partir do teste BDS, cuja hipótese nula de independência dos dados não é rejeitada ao nível de 5% de significância uma vez que o valor-*p*, em todos os casos e em todas as dimensões, é maior que 0,05. O mesmo tipo de análise pode ser feito em relação ao teste de Ljung-Box para as quatro abordagens.

Tabela 4: Testes de *Dickey-Fuller* e de correlação ponderada.

Método	***Teste de Dickey Fuller***			***Correlação Ponderada***
	Estatística ADF	*Nível*	*Valores Críticos*	
ACP	-13,12375	1% 5% 10%	-3,4402 -2,8651 -2,5687	$\rho_{1,2}^{(\omega)} = 0{,}00123$
Análise de agrupamento integrada com ACP	-12,96724	1% 5% 10%	-3,4402 -2,8651 -2,5687	$\rho_{1,2}^{(\omega)} = 0{,}00051$
Análise gráfica dos vetores	-13,22529	1% 5% 10%	-3,4402 -2,8651 -2,5687	$\rho_{1,2}^{(\omega)} = 0{,}00031$
DBSCAN	-12,89376	1% 5% 10%	-3,4402 -2,8651 -2,5687	$\rho_{1,2}^{(\omega)} = 0{,}00009$

Fonte: Elaborada pelos autores.

Os resultados apresentados na Tabela 4 mostram a não significância das correlações ponderadas entre as séries de ruído e as séries aproximadas remanescentes, uma vez que os valores das correlações ponderadas são bastante baixos indicando uma boa separabilidade das componentes. Pode-se observar que, na abordagem DBSCAN, a correlação quase nula mostra que este método é mais eficiente na remoção de ruídos de séries temporais em SSA que os demais métodos aplicados. Sobre o teste de *Dickey-Fuller*, os resultados mostram que a hipótese nula de raiz unitária é rejeitada em todos os métodos e em todos os níveis, indicando que estas séries de ruído são estacionárias.

8. Conclusões

Neste artigo, foi proposto o uso do método SSA para a filtragem de séries temporais removendo as partes ruidosas. Para tanto, quatro abordagens foram propostas: ACP; ACP integrada com Análise de Agrupamento; Análise Gráfica de Vetores Singulares e DBSCAN. As quatro metodologias foram usadas em oito séries simuladas e para a ilustração dos métodos foi utilizada a série temporal de vazão de afluentes da UHE GB Munhoz.

Para verificar a adequação do processo nas séries simuladas, a estatística RMSE foi utilizada nos ajustes. Para testar a independência dos dados e a estacionariedade das séries de ruídos removidas nas quatro abordagens, foram utilizados os testes de *Ljung-Box* e BDS. Para testar a estacionariedade das séries removidas, o teste de *Dickey-Fuller* foi aplicado e a correlação ponderada foi utilizada para verificar a separabilidade no processo de remoção das séries de ruídos.

Os resultados apresentados dão conta que o processo de filtragem de séries temporais à partir de remoção de ruídos sob a abordagem SSA melhora o desempenho do modelo e que os quatro métodos utilizados na fase de agrupamento são eficazes. Pode-se também perceber que as séries removidas usando os quatro métodos de agrupamento tem comportamento de ruído uma vez que os testes utilizados mostram que elas são independentes e estacionárias. A abordagem DBSCAN, por sua vez, superou as demais métodos no quesito correlação ponderada, apresentando uma melhor separabilidade dos dados quando aplicada. Este fato corrobora que o DBSCAN promove uma melhora na remoção dos ruídos por promover uma separabilidade mais adequada na decomposição da série. Portanto, o método DBSCAN é adequado para realizar o agrupamento da abordagem SSA para a remoção de ruído da série de vazão de afluentes da Usina GB Munhoz.

Referências

Aldenderfer, M. e Blashfield, R. *Cluster Analysis*. California: Sage Publications, 1984.

Brock, W., Scheinkman, J., Dechert, W., e LeBaron, B. A test for independence based on the correlation dimension. *Econometric Reviews*, v. 15, n. 3, p. 197-235, 1996.

Caterpillar-SSA. Version 3.4. Standard M Edition. [S.l.]: Gistat Group, 2013.

Dickey, D. e Fuller, W. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, v. 7, n. 366a, p. 427-431, 1979.

Elsner, J. e Tsonis, A. *Singular Spectrum Analysis. A New Tool in Time Series Analysis*. New York and London: Plenum Press, 2010.

Golyandina, N., Nekrutkin, V. e Zhigljavsky, A. *Analysis of Time Series Structure:* SSA and Related Techniques. New York: Chapman & Hall/CRC, 2001. 300p.

Hahsler, M., Piekenbrock, M., Arya, S. e Mount, D. Density based clustering of applications with noise (DBSCAN) and related algorithms. R package version 1.1-2, 2018.

Hamilton, J. *Time Series Analysis*. New Jersey: Princeton University Press, 1994.

Hassani, H. Singular spectrum analysis: Methodology and comparison. *Journal of Data Science*, v. 5, n. 2, p. 239-257, 2007.

Ljung, G. e Box, G. On a measure of lack of fit in time series models. *Biometrika*, v. 65, n. 2, p. 297-303, 1978.

Manly, B. *Métodos Estatísticos Multivariados: Uma introdução.* 3ª. Ed. São Paulo: Bookman, 2008.

MATLAB. Version R2010b. [S.l.]: MathWorks, 2010.

Menezes, M., Souza, R. e Pessanha. J. Electricity consumption forecasting using singular spectrum analysis. *DYNA*, v. 82, n. 190, p. 138-146, 2015.

Morettin, P. e Toloi, C. *Análise de Séries Temporais, 2ª Ed.* São Paulo: Edgard Blucher, 2006.

Teixeira Jr., L., Menezes, M., Cassiano, K., Pessanha, J. e Souza, R. Residential electricity consumption forecasting using a geometric combination approach. *International Journal of Energy and Statistics*, v. 1, n. 2, p. 113-125, 2013.

Terry, L., Pereira, M., Araripe Neto, T., Silva, L., e Sales, P. Coordinating the energy generation of the Brazilian national hydrothermal electrical generating system. *Interfaces*, v. 16, n. 1, p. 16-38, 1986.

Tran, T., Drab, K. e Daszykowski, M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, v.120, p. 92-96, 2013.