SINGULAR SPECTRUM ANALYSIS, ANÁLISE DE COMPONENTES PRINCIPAIS E CLUSTERIZAÇÃO BASEADA EM DENSIDADE NA REMOÇÃO DE RUÍDOS DE SÉRIES TEMPORAIS

Keila Mara Cassiano
Departamento de Estatística - Universidade Federal Fluminense
Rua Mário Santos Braga, S/N, Campus Valonguinho, Centro, Niterói - RJ, 24.220-900.
keilamath@hotmail.com

Moisés Lima de Menezes
Departamento de Estatística - Universidade Federal Fluminense
Rua Mário Santos Braga, S/N, Campus Valonguinho, Centro, Niterói - RJ, 24.220-900.
moises_lima@msn.com

Reinaldo Castro Souza
Departamento de Engenharia Elétrica - Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro – RJ, 22.451-900.
reinaldo@ele.puc-rio.br

José Francisco Moreira Pessanha
Instituto de Matemática e Estatística - Universidade do Estado do Rio de Janeiro
Rua São Francisco Xavier, 524, Maracanã, Rio de Janeiro – RJ, 20.550-013.
professorjfmp@hotmail.com

Resumo

O objetivo deste artigo é apresentar diferentes métodos para remoção de ruídos de séries temporais com o uso de Singular Spectrum Analysis (SSA) 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 obter 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 da Usina Hidrelétrica Governador Bento Munhoz, localizada na Bacia do Rio Paraná, Brasil.
Palavras Chave: Remoção de Ruídos, Singular Spectrum Analysis, Análise de Componentes Prinicipais, DBSCAN.

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 flow of the Governor Bento Munhoz Hydroelectric Plant, located on the Paraná River Basin.
Keywords: Remotion of Noise, Singular Spectrum Analysis, Principal Components Analysis, DBSCAN.

1. INTRODUÇÃO

Uma das principais características do Sistema Elétrico Brasileiro (SEB) reside no fato da sua capacidade de geração ser predominantemente hidráulica. Tal característica contribui substancialmente para que a matriz energética brasileira seja considerada uma das mais renováveis do mundo. Não obstante, devido às incertezas nos regimes das vazões naturais [12], o SEB está submetido a um significativo risco hidrológico. Para mitigá-lo, o SEB conta com usinas termoelétricas que complementam a geração hidroelétrica no país. Adicionalmente, o sistema ainda dispõe de 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. Desse modo, 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 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) [12]. 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 *Singular Spectrum Analysis* (SSA). SSA é um método eficiente na extração e reconstrução de componentes periódicas e não periódicas de séries temporais com elevados níveis de ruído [12]. Portanto, útil no processamento de séries temporais de vazões.

Por meio do método SSA, uma série temporal pode ser transformada, em uma matriz trajetória, uma matriz passível de ser expandida em termos da decomposição em valores singulares [8]. Cada componente desta expansão concentra uma parcela da energia [7] contida na matriz trajetória gerada a partir da série temporal. Dessa forma, um subconjunto de componentes concentra a maior parte da energia total com estrutura de dependência temporal e ruído; 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, pode-se 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) [13]. 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 [7] (que são atributos das séries temporais de ruídos). Para a ilustração das três abordagens, 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 seis 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. Os dados relativos ao

estudo de caso são descritos na seção 4. Os principais resultados obtidos são apresentados na seção 5. Por fim, as conclusões são apresentadas na seção 6.

2. SINGULAR SPECTRUM ANALYSIS

SSA é um método de processamento de sinais que pode ser utilizado, dentre outras aplicações, na remoção de ruído de séries de tempo [6, 8]. 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 $Y_T = [y_1, \dots, y_T] \in \mathbb{R}^T$ uma série temporal com cardinalidade igual a T e $F: \mathbb{R}^T \to \mathbb{R}^{L\times K}$ um mapa invertível. Por incorporação, entende-se como sendo um procedimento no qual uma série temporal $Y_T \in \mathbb{R}^T$ é transformada pelo mapa F em uma matriz $X = [X_1, \dots, X_T]_{L\times K} \in \mathbb{R}^{L\times K}$, onde $X_k = [y_K, \dots, y_{K+L-1}]^T \in \mathbb{R}^L$, para todo k [9]. Isto é, $Y_T \in \mathbb{R}^T \xrightarrow{F} X \in \mathbb{R}^{L\times K}$, onde $K = T - L + 1$. A matriz X é conhecida como matriz trajetória [8] e o parâmetro L, que assume algum valor inteiro no intervalo $2 \le L \le T$ é o tamanho da janela da matriz trajetória [6].

Considere o operador normal e compacto $S \coloneqq XX^T$ [9]. Seja $\sigma(S)$ o espectro de S e $\{U_l U_l^T\}_{l=1}^L$ uma resolução de identidade sobre o *espaço de Hilbert* $(\mathbb{R}^L, \langle . \rangle)$ associada ao operador S, onde U_l é o autovetor associado ao autovalor $\lambda_l \in \sigma(S)$. Pode-se mostrar que S é um operador *semi-definido positivo* [9], de modo que $\lambda_l \ge 0$, para todo l. Seja $V_l \coloneqq X^T U_l/\sqrt{\lambda_l}$ e considere a ordenação parcial: $0 \ge \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_L$.

De acordo com [9], se $S \coloneqq XX^T$ e S é um operador normal, compacto e semi-definido positivo, então a matriz trajetória X pode ser expandida através da decomposição em valores singulares (SVD). Como esta condição suficiente [9] é verdadeira, segue que a matriz trajetória X pode ser expandida via SVD, em (1).

$$X = \sum_{\lambda_l \in \sigma(S)} (\lambda_l)^{\frac{1}{2}} U_l V_l^T = \sum_{l=1}^L X_l, \tag{1}$$

Na qual $X_l \coloneqq {\lambda_l}^{1/2} U_l \, V_l^T$ 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 de 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:

$$X = \sum_{l=1}^d X_l, \text{ onde } d \le L. \tag{2}$$

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.

Considere a sequência $\{X_l\}_{l=1}^d$ de matrizes elementares na SVD, em (2). Agrupe-as em $m \le d$ grupos disjuntos utilizando algum método [por exemplo, com o auxílio da análise de componentes principais [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 $\{X_l\}_{l=1}^d = \bigcup_{i=1}^m \left\{X_{I_{ij}}\right\}_{j=1}^{p_i}$,

onde $m \leq d$. 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).

$$X = \sum_{i=1}^{m} X_{I_i} \tag{3}$$

É importante salientar que o procedimento de agrupamento pode ser realizado também sob a sequência $\bigcup_{i=1}^{m}\left\{X_{I_{ij}}\right\}_{j=1}^{p_i}$ de igual forma ao realizado na sequência $\{X_l\}_{l=1}^{d}$ utilizando, inclusive, um método ou critério de agrupamento diferente do aplicado sob a sequência $\{X_l\}_{l=1}^{d}$ em (2). O objetivo do agrupamento é diminuir o número de componentes (ou matrizes elementares) na SVD da matriz trajetória X. A contribuição da componente X_{I_i} pode ser mensurada pela razão de valores singulares dada por $\sum_{j=1}^{p_i}\left(\lambda_{I_{ij}}\right)^{1/2} / \sum_{l=1}^{d}(\lambda_l)^{1/2}$. Tome o conjunto $\bigcup_{i=1}^{m}\left\{X_{I_{ij}}\right\}_{j=1}^{p_i}$ de matrizes elementares na SVD. Dado que F é um mapa invertível, segue que F^{-1} sobre a matriz trajetória X é tal que $F^{-1}(X) = [y_t]_{1\times T}$.

A componente $\left[y_t^{(i)}\right]_{1\times T}$ é conhecida como componente SSA da série temporal $[y_t]_{1\times T}$ gerada, por meio do mapa F^{-1}, a partir da matriz elementar X_{I_i}.

Considere a matriz trajetória X e assuma que $L^* = \min(L, K)$ e que $K^* = \max(L, K)$. Considere que $x_{l,k}^{(i)}$ seja um elemento na linha l e coluna k na matriz X_{I_i}. O elemento $y_t^{(i)}$ da componente SSA $\left[y_t^{(i)}\right]_{1\times T}$ é calculado por meio da média diagonal, que é definida em (4), a partir da matriz elementar X_{I_i}.

$$y_t^{(i)} = \begin{cases} \frac{\sum_{l=1}^{t} x_{l,t-l+1}^{(i)}}{t}, & se\ \ 1 \leq t < L^* \\ \frac{\sum_{l=1}^{L^*} x_{l,t-l+1}^{(i)}}{L^*}, & se\ \ L^* \leq t < K^* \\ \frac{\sum_{l=t-K^*+1}^{T-K^*+1} x_{l,t-l+1}^{(i)}}{T-K^*+1}, & se\ \ K^* \leq t \leq T \end{cases} \tag{4}$$

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 valores singulares $\sum_{j=1}^{p_i}\left(\lambda_{I_{ij}}\right)^{1/2} / \sum_{l=1}^{d}(\lambda_l)^{1/2}$

De acordo com [6], as componentes SSA $\left[y_t^{(i)}\right]_{1\times T}$ podem ser classificadas em três categorias: tendência, componentes harmônicas (ciclo e sazonalidade) e ruído.

2.3. SEPARABILIDADE

Por correlação ponderada (w), entenda como sendo a função que quantifiva a dependência linear entre duas componentes SSA $Y_t^{(i)}$e $Y_t^{(j)}$, conforme definido por $\rho_{ij}^{(w)} = \frac{\left(Y_t^{(i)}, Y_t^{(j)}\right)_w}{\left\|Y_t^{(i)}\right\|_w \left\|Y_t^{(j)}\right\|_w}$, onde $\|\ \ \|$ 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 $w_k = min\{k, L, T-k\}$.

Por meio da separabilidade, pode-se verificar estatisticamente se as duas componentes SSA estão bem separadas, em termos de dependência linear. Se o valor absoluto w é pequeno [8], 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 [6].

3. MÉTODOS DE REMOÇÃO DE RUÍDO

No contexto de séries de tempo, os métodos de remoção de ruído consistem, basicamente, em se 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) [11], 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.

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 [1]. De acordo com [1], a análise de agrupamentos visa a agrupar um conjunto com *N* objetos em *K clusters* 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 [1], 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.

No estudo de caso, foi adotado o método de agrupamento hierárquico de encadeamento simples ou (*single-linkage*) [1] para se dividir o conjunto de vetores obtidos pela expansão SVD em três grupos excludentes. O algoritmo utilizado foi o aglomerativo [1], ou seja, o algoritmo inicia com *N clusters*, 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 *clusters*. 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 [1].

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 [6], 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 [10]. As coordenadas da série temporal em duas componentes ortogonais podem ser dispostas em um diagrama de dispersão [6].

3.4. DBSCAN

DBSCAN é o principal representante dos métodos de classificação 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 [13] e tem um desempenho robusto para conjuntos de dados contendo estruturas densas com aglomerados conectados. Os resultados da classificação 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.

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 *p* é chamado de **ponto de borda**.

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 $p1 = q$ 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.

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*.

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 [4].

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. 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 1.

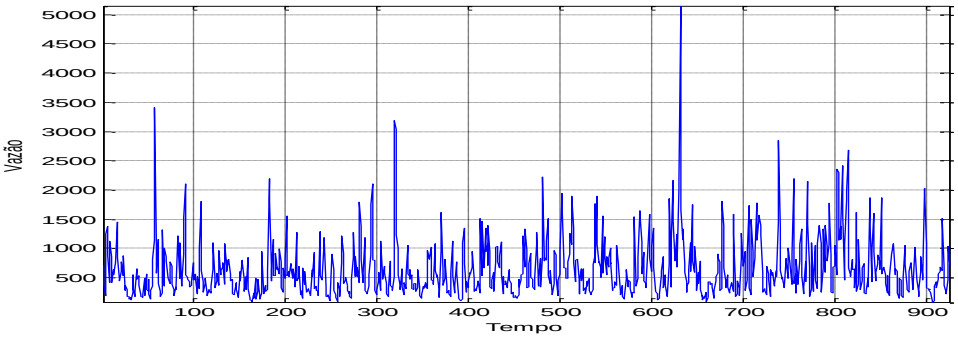

Figura 1 - Série temporal mensal de vazão (m^3/s) da UHE GB Munhoz.

5. 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 [2] e *Ljung-Box* [10].

5.1. ANÁLISE DA SÉRIE TEMPORAL 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* (tamanho 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 2, 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. Note que parte da energia (classificada estatisticamente como ruído) foi removida por ACP.

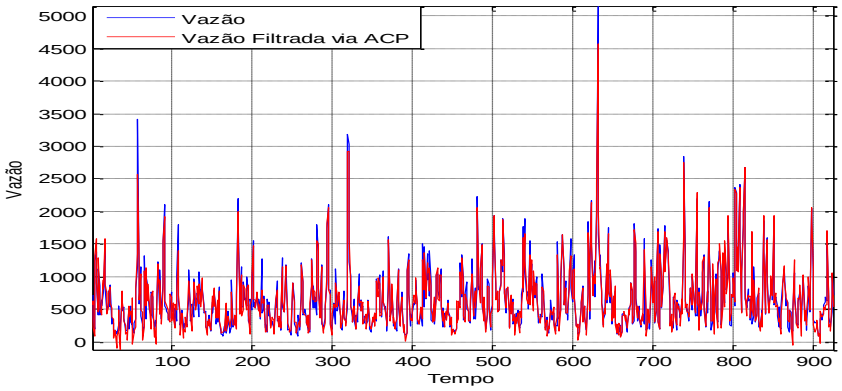

Figura 2 - Série temporal de vazão da UHE GB Munhoz e sua versão aproximada através da ACP.

5.2. ANÁLISE DA SÉRIE TEMPORAL 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* 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 3 mostra a série original e a série filtrada via análise de agrupamento com ACP.

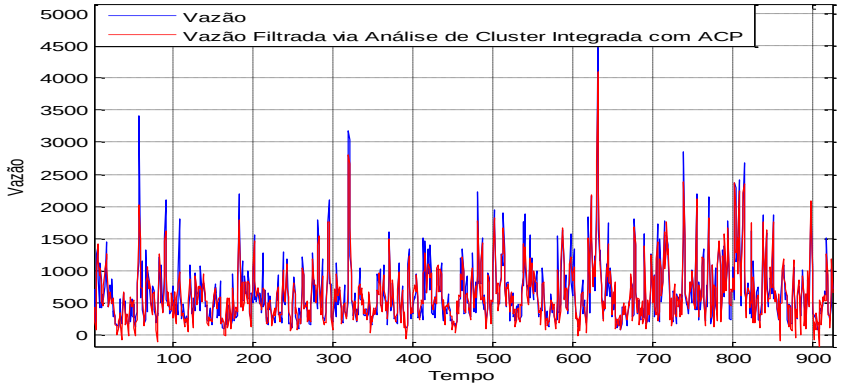

Figura 3 - Série temporal de vazão da UHE GB Munhoz e sua versão filtrada através da análise de agrupamento integrada com ACP.

Na Figura 3, pode-se visualizar que parte da energia (classificada estatisticamente como ruído) foi removida por meio da análise de agrupamento integrada com ACP.

5.3. ANÁLISE DA SÉRIE TEMPORAL 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 4, 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* [5] e no MATLAB.

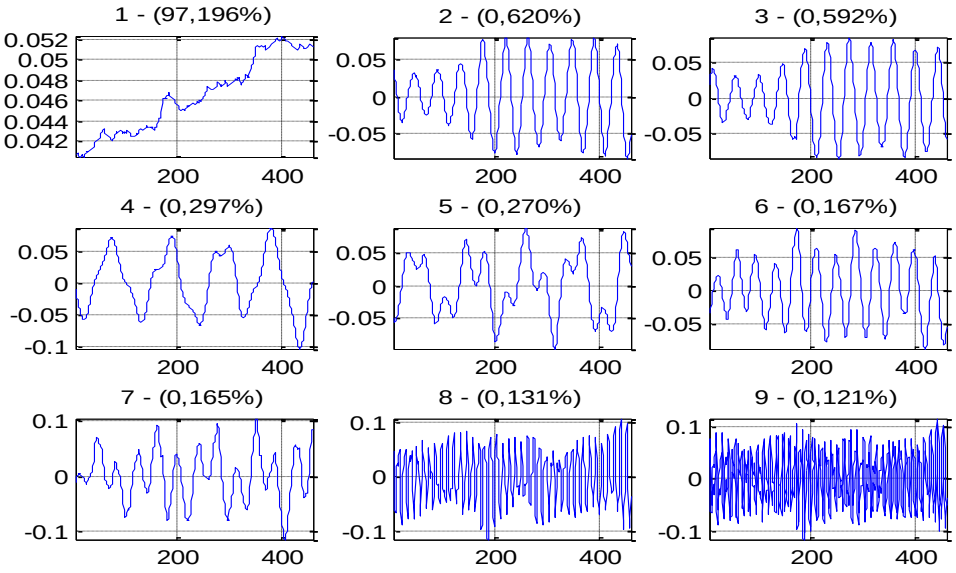

Figura 4 – Os nove primeiros vetores singulares na SVD da matriz trajetória da Série temporal de vazão da UHE GB Munhoz.

As análises feitas como o comportamento dos gráficos dos vetores singulares levam a decomposição da série em três componentes: tendência, harmônica e ruído. Com efeito, segue que 3 componentes SSA foram geradas e aquela classificada como ruído (via estatísticas BDS e de *Ljung-Box*), removida.

Na Figura 5, tem-se a série temporal de Vazão da UHE GB Munhoz e a sua versão filtrada através da análise gráfica dos vetores singulares.

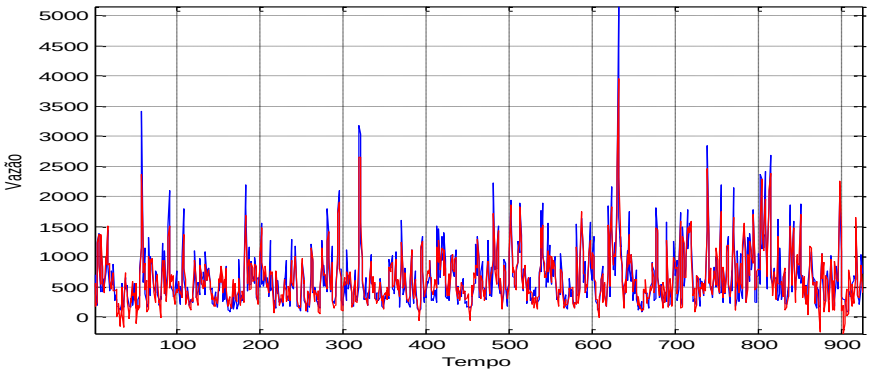

Figura 5 - Série temporal de vazão e sua aproximação através da análise gráfica de vetores singulares.

5.4. ANÁLISE DA SÉRIE TEMPORAL VIA DBSCAN

No processo de clusterização na fase de agrupamento utilizando o DBSCAN, o software utilizado (R), 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 Figura 6 apresenta a sobreposição da série original com a série aproximada via DBSCAN.

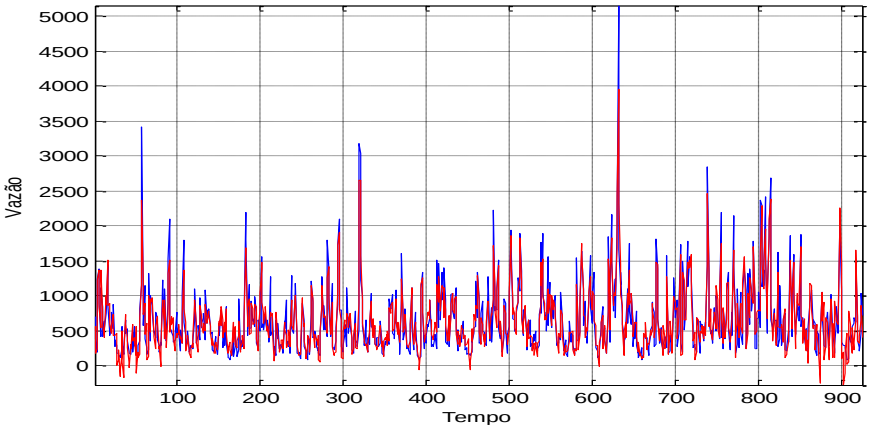

Figura 6 - Série temporal de vazão e sua aproximação através da abordagem DBSCAN.

Os resultados dos testes BDS e Ljung-Box aplicados às séries de ruído dos quatro métodos estudados estão na Tabela 1. Em seguida, na tabela 2 estão os resultados do teste de *Dickey-Fuller* [3] sobre as séries temporais de resíduos das quatro abordagens e de correlação ponderada $\rho_{1,2}^{(\omega)}$ entre a série temporal de resíduos oriundas das quatro abordagens (1) e a série temporal de vazão da UHE GB Munhoz filtrada (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.

Tabela 1 - Testes BDS e de *Ljung-Box*.

Método	***Teste BDS***				***Teste de Ljung-Box***				
	Dimensão	*Estatística BDS*	*Estatística Z*	*Prob.*	*Lag*	*FAC*	*FACP*	*Estatística 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

A partir dos resultados apresentados na Tabela 1, pode-se concluir que as séries temporais de ruídos oriundas das quatro abordagens são estacionárias de segunda ordem.

Tabela 2 - 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	$\rho_{1,2}^{(\omega)} = 0{,}00009$

Os resultados apresentados na Tabela 2 mostram a não significância das correlações ponderadas. Pode-se observar que na abordagem DBSCAN, a correlação quase nula mostra que este método é eficiente na remoção de ruídos de séries temporais, sobretudo em SSA. Sobre o teste de Dickey-Fuller, os resultados mostram que a hipótese nula de raiz unitária é rejeitada em todos métodos e em todos os níveis.

6. CONCLUSÕES

Neste artigo, foi proposto o uso do método SSA para a filtragem de séries temporais com quatro abordagens na fase de agrupamento SSA: ACP; ACP integrada com Análise de Agrupamento; Análise Gráfica de Vetores Singulares e DBSCAN. Para a ilustração dos métodos, foi utilizada a série temporal de vazão da UHE GB de Bento Munhoz.

Para verificar a qualidade da filtragem, foram aplicados testes estatísticos de *Ljung-Box*, BDS, correlação ponderada e de *Dickey-Fuller* nas componentes ruidosas obtidas de cada aplicação em que a séries é decomposta em duas partes: sinal e ruído. Nas Tabelas 1 e 2 foram apresentados os resultados destes testes para cada tipo de agrupamento. Como mostrados nestas tabelas, pode-se perceber que todas as metodologias aplicadas cumprem bem a função de separação de componentes.

Pode-se perceber no item de correlação ponderada na Tabela 2 que o método de agrupamento DBSCAN obteve uma correlação mais próxima de zero, indicando ser esta mais eficiente que as demais para uso em SSA na filtragem de séries temporais.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Aldefender, M.S., Blashfield, R.K. (1984). *Cluster Analysis*. Sage Publications. California. Páginas 35-45.

[2] Brock, W. A., Dechert, W., Scheinkman, J., and LeBaron, B. (1996). A test for Independence based on the correlation dimension. *Econometric reviews*, 15 (3), pp. 197-235.

[3] Dickey, D, and Fuller, W. A. (1979) Distribution of the estimates for Autorregressive Time Series with a Unit Root. Journal of the American Statistical Association, 7, pp. 427 – 431.

[4] Ester, M., Kriegel, H. P., Snader, J., and Xu, X. (1996). Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD-96 Proceedings. 226-231.

[5] GISTATGROUP (2010). Caterpillar SSA. Petersburg University. Department of Mathematics. Russia. Site: http://www.gistatgroup.com/cat/index.html.

[6] Golyandina, N., Nekrutkin, V., Zhihgljavsky, A. (2001). *Analysis of time series structure: SSA and reletade techniques.* Chapman&Hall/CRC. New York, USA.
[7] Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.
[8] Hassani, H. (2007). Singular Spectrum Analysis: Methodology and Comparison. *Journal of Data Science* 5, 239-257.
[9] Kubrusly, C. S. (2001). *Elements of Operator Theory*. Birkhäuser, Boston.
[10] Morettin, P.A., Toloi, L.M.C (2006). *Análise Séries Temporais, 2ª Ed. ABE. Projeto Fisher.* Editora: Edgard Blucher.
[11] Manly, B. J. F. (2008). *Métodos Estatísticos Multivariados: Uma introdução.* 3ª. Ed. Editora Bookman.
[12] Terry, L.A., Pereira, M.V.F., Silva, L.F. Sales, P.R., Araripe N. T.A. (1986). Coordinating the Energy Generation of the Brazilian System. Interfaces Special lssue.
[13] Tran, T. N., Drab, K. and Daszykowski, M. (2013),.Revised DBSCAN algorithm to cluster data with dense adjacent clusters". *Chemometrics and Intelligent Laboratory Systems*, Vol.120. 92-96, 2013.