ANÁLISE ESPECTRAL SINGULAR COM CLUSTERIZAÇÃO BASEADA EM DENSIDADE NA MODELAGEM DE SÉRIES TEMPORAIS

Keila Mara Cassiano
Universidade Federal Fluminense – UFF
Rua Mário Santos Braga S/N, Campus Valonguinho, Centro - Niterói, Brasil.
keilamath@hotmail.com

José Francisco Moreira Pessanha
Universidade do Estado do Rio de Janeiro – UERJ
Av. Maracanã, Tijuca - Rio de Janeiro, Brasil.
professorjfmp@hotmail.com

RESUMO

Este trabalho propõe o uso de clusterização espacial baseada em densidade de aplicações com ruído (DBSCAN - density based spatial clustering of applications with noise) para separar os componentes de ruído das autotriplas na fase de agrupamento da Análise Espectral Singular (SSA – Sigular Spectrum Analysis) de séries temporais. DBSCAN é um método moderno e especialista em identificação de ruídos por regiões de menor densidade. Este trabalho mostra uma maior eficácia do DBSCAN perante os demais métodos já utilizado nesta fase da SSA. Este resultado é notório uma vez que a metodologia proposta permite considerável redução de ruídos e proporciona uma melhor eficácia na previsão. Modelos de Box-Jenkins são simulados e um modelo passeio aleatório com drift são avaliados com esta abordagem e uma série de velocidade do vento de uma estação anemométrica localizada no Nordeste do Brasil é utilizada como aplicação.

PALAVARAS CHAVE. Modelagem de Séries Temporais, Análise Espectral Singular, DBSCAN.
Área principal (EST)

ABSTRACT

This work proposes using DBSCAN (Density Based Spatial Clustering of Applications with Noise) to separate the noise components of eigentriples in the grouping stage of the Singular Spectrum Analysis (SSA) of Time Series. DBSCAN is a modern and expert method at identify noise through regions of lower density. This work shows better efficiency of DBSCAN over the others methods already used in this stage of SSA, because it allows considerable reduction of noise and provides better forecasting. The result is supported by experimental evaluations realized for simulated series of Box-Jenkins models and a random walk with drift model and the approach is applied at a real time series of speed wind.

KEYWORDS. Time Series Modeling, Singular Spectrum Analysis, DBSCAN.

Main area (EST)

1. INTRODUÇÃO

Singular Spectrum Analysis (SSA) é um poderoso método que vem sendo aplicado em diversas áreas da Matemática, Física, Economia, Matemática Financeira, Meteorologia, Oceanografia, Ciências Sociais e em recentes Análises de Séries Temporais (HASSANI, 2007). A abordagem incorpora elementos de análise clássica de séries temporais, estatística multivariada, geometria multivariada, sistemas dinâmicos e processamentos de sinais.

Quando aplicado a séries temporais o principal objetivo da SSA é fazer uma decomposição da série original em uma soma de componentes independentes, que podem ser interpretados como os componentes de tendência, os componentes oscilatórios e os componentes com estrutura de ruído. Uma vez identificadas os componentes de ruído, podem ser desconsiderados na reconstrução da série e obtém-se uma série menos ruidosa para modelagem e previsão. Para tal, é definido um algoritmo SSA de cinco etapas: 1) Decomposição da série original na matriz trajetória; 2) Decomposição Singular de Valor (SVD) da Matriz Trajetória; 3) Média Diagonal das Componentes da SVD 4) Agrupamento dos componentes de tendência, componentes os**c**ilatórios e a estrutura do ruído; e 5) Reconstrução da série menos ruidosa.

Até o presente, todos os trabalhos desenvolvidos e publicados em SSA tinham usado um dos seguintes métodos para a Fase 4 do SSA, a fase de separação de séries de ruído: Análise de Componentes Principais; Análise Heurística Visual Técnica do Comportamento dos Pares de Autovetores; ou Clusterização Hierárquica (que está implementado em RSSA, o pacote SSA no software R). O objetivo deste trabalho é o de propor um método de clusterização mais moderno e especialista em separação de ruído, o DBSCAN (Density Based Spatial Clustering of Applications with Noise).

Esta é uma proposta relevante, uma vez que o método de clusterização hierárquica foi a última inovação na separação de ruído na abordagem SSA. No entanto, este método é muito sensível ao ruído, não é capaz de separá-lo corretamente, não deve ser usado em conjuntos com diferentes densidades e não funciona bem no agrupamento de séries temporais de diferentes tendências (YIN et al., 2006). Se mesmo com essas desvantagens o método de clusterização hierárquica mostra bom desempenho em SSA é interessante investigar se um método mais recente pode superar o seu rendimento. O método hierárquico foi proposto em 1955 e desde então pelo menos 120 novos métodos de agrupamento foram publicados. É importante trazer para a abordagem SSA o progresso alcançado na área de clusterização nestes 65 anos. O método proposto, DBSCAN, foi lançado em ESTER et al. (1996), 41 anos depois do método hierárquico é especialista na identificação de ruídos, já foi muito explorado e revisado e todas as análises deste trabalho foram feitas com sua versão mais moderna e robusta, lançada em 2013 em TRAN et al. (2013).

Este trabalho traz assim uma combinação eficiente de duas ferramentas, método de clusterização baseado em densidade e SSA, sem precedentes na literatura de análise de séries temporais, trazendo o que há de mais moderno e revisado e até o momento em clusterização baseada em densidade. O desempenho da metodologia proposta é avaliado em séries sintéticas simuladas de modelos Box-Jenkins e de um processo passeio aleatório com drift e o método é também aplicado na previsão de uma série real de velocidade do vento. O trabalho está organizado da seguinte forma: na Seção 2 a metodologia SSA é apresentada e na Seção 3 disserta-se sobre clusterização baseada em densidade e o método DBSCAN. Na Seção 4 é relatada a metodologia proposta e as ferramentas computacionais utilizadas. Na Seção 5, resultados de simulações e aplicação do método a uma série de velocidade do vento são apresentados. A Seção 6 traz as principais conclusões.

2. SINGULAR SPECTRUM ANALYSIS (SSA)

Seja $Y_T = [y_1, \dots, y_T] \in \mathbb{R}^T$ uma *série temporal* com cardinalidade igual a T. Por *incorporação da série* Y_T entende-se como sendo um procedimento no qual a série temporal $Y_T \in \mathbb{R}^T$ é transformada por um mapa F em uma matriz $X = F(Y_T) = [X_1 \ X_2 \dots X_K] \in \mathbb{R}^{L\times K}$, onde X_k representa a k-ésima coluna de X, dada por $X_k = [y_k, \dots, y_{k+L-1}]^T \in \mathbb{R}^L$, para todo $k = 1, \dots, K$; onde $K = T - L + 1$. Assim definido, $F: \mathbb{R}^T \to \mathbb{R}^{L\times K}$, é um mapa invertível tal que

$$F(Y_t) = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_{K-1} & y_K \\ y_2 & y_3 & y_4 & \cdots & y_K & y_{K+1} \\ \vdots & & & \ddots & \vdots & \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_{T-1} & y_T \end{bmatrix}.$$

A matriz X é uma matriz Hankel conhecida como *matriz trajetória* (HASSANI, **2007**) e o parâmetro L, que assume algum valor inteiro no intervalo $2 \le L \le T$ é o *tamanho da janela* da matriz trajetória. A matriz trajetória X pode ser expandida via *decomposição em valores singulares (SVD)* por:

$$X = \sum_{i=1}^{L} (\lambda_i)^{\frac{1}{2}} U_i V_i' = \sum_{i=1}^{L} A_i, \qquad (1)$$

onde: $A_i = \lambda_i^{\frac{1}{2}} U_i V_i'$; $\{\lambda_i\}_{i=1}^{L}$ é o conjunto dos autovalores de $S = XX'$em ordem decrescente, $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_L$; $\{U_i\}_{i=1}^{L}$ é o sistema ortonormal dos autovetores singulares de S associados a estes autovalores; $V_i = X'U_i/\sqrt{\lambda_i}$.

A coleção (λ_i, U_i, V_i) é conhecida como autotripla na SVD da matriz trajetória X. A contribuição de cada componente $A_i$em (1) pode ser mensurada pela razão de valores singulares, dada por $(\lambda_i)^{1/2}/\sum_{i=1}^{L}(\lambda_i)^{1/2}$.

Considere que d seja o posto (isto é, o número de autovalores não nulos) da matriz S. Segue que a identidade descrita em (1) pode ser reescrita tal como:

$X = \sum_{i=1}^{d} A_i$, onde $d \le L$. (2)

A *etapa de reconstrução* pode ser subdividida em *média diagonal e classificação*. Para cada matriz $A_i$componente de X é aplicada a "Média Diagonal de A_i" que transforma cada matriz A_i em uma série $Y_T^{(i)} = \left\{y_t^{(i)}\right\}_{t=1,\dots,T}$. Considerando $L^* = \min(L, K)$ e $K^* = \max(L, K)$ e $a_{l,k}^{(i)}$ seja o elemento na linha l e coluna k na matriz A_i temos

$$y_t^{(i)} = \begin{cases} \frac{\sum_{l=1}^{t} a_{l,t-l+1}^{(i)}}{t}, & se \ 1 \le t < L^* \\ \frac{\sum_{l=1}^{L^*} a_{l,t-l+1}^{(i)}}{L^*}, & se \ L^* \le t < K^* \\ \frac{\sum_{l=t-K^*+1}^{T-K^*+1} a_{l,t-l+1}^{(i)}}{T-K^*+1}, & se \ K^* \le t \le T \end{cases} \qquad (3)$$

Em outras palavras, o sistema (3) diz que o t-ésimo termo $t = 1, \dots, T$ da i-ésima série $Y_T^{(i)}$ $i = 1, \dots, d$ é dada pela média da t-ésima diagonal da matriz A_i, uma relação inversa da incorporação que faz o operador F quando aplicado a uma série temporal. É possível provar que a série original Y_T é igual a soma das série s $Y^{(i)}$, ou seja, $Y_T = \sum_{i=1}^{d} Y_T^{(i)}$. A componente $Y_T^{(i)}$ é conhecida como *componente SSA* da série temporal Y_T gerada, por meio do mapa F^{-1}, a partir da matriz elementar A_i. De acordo com GOLYANDINA et al. (2001), as séries $Y^{(i)}$ podem ser classificadas em três categorias: *tendência*, *componentes harmônicas* e *ruído*. Esta classificação pode ser feita por Análise de Componentes Principais ou pela análise visual gráfica dos pares de coordenadas da série temporal na base definida pelos respectivos autovetores da SVD, ou por Clusterização. Quando realizada a *clusterização* objetiva-se agrupar as séries $Y_T^{(i)}$ $i = 1, \dots, d$ resultantes da Média Diagonal em grupos *disjuntos* que

possam ser identificados como grupos de séries componentes de tendencia, grupos de séries componentes oscilatórios e especialmente, grupos de séries com características de ruído.

3. CLUSTERIZAÇÃO BASEADA EM DENSIDADE E DBSCAN

Cada vez mais, grandes quantidades de aplicações requerem o gerenciamento de dados espaciais. Devido a isso, Knowledge Discovery in Databases (KDD) tem se tornado uma ferramenta cada vez mais importante. KDD pode ser definida como um processo não- trivial de identificar padrões em análise compreensível de dados. Data Mining é um passo no processo de KDD que consiste na aplicação de análise de dados e algoritmos de descoberta que, sob as limitações de eficiência computacional aceitáveis, produzem uma numeração particular de padrões em relação aos dados (FAYYAD et al., 1996). Por sua vez, Clusterização é uma das técnicas mais importantes em Data Mining que tem por objetivo particionar um conjunto de objetos em grupos tais que os objetos dentro de um conjunto têm padrões mais semelhantes entre si do que os padrões em diferentes clusters.

Os estudos em clusterização começaram em Antropologia, a partir de obras de DRIVER & KROEBER (1932) e em Psicologia por ZUBIN (1938) e TRYON (1939). No entanto, o primeiro trabalho publicado referindo-se a um "método de clusterização"' foi o trabalho SORENSEN (1948), onde o autor definiu o método hierárquico aglomerativo de ligação completa. Diferentes técnicas de clusterização foram desenvolvidos, mas ficaram restritas muito tempo a um pequeno grupo de pesquisadores, devido à complexidade matemática. O desenvolvimento da tecnologia computacional promoveu a disseminação da técnica entre os diferentes ramos do conhecimento e o desenvolvimento de novos métodos. Centenas de algoritmos de clusterização têm sido propostos na literatura, em muitos campos científicos diferentes, e eles divergem sobre a escolha da função objetivo, os modelos probabilísticos generativos e a heurística. Hás três categorias gerais de classificação de um método de clusterização: particional, hierárquica e baseada em densidade.

A maioria dos métodos Particionais e Hierárquicos clusteriza objetos baseando-se na distância entre eles. Por isso tais métodos podem encontrar dificuldades para descobrir clusters de formas arbitrárias. Nos métodos de Clusterização baseados em Densidade, clusters são definidos como regiões densas, separadas por regiões menos densas que representam ruídos. As regiões densas podem ter uma forma arbitrária e os pontos dentro de uma região podem também estar distribuídos arbitrariamente e, por isso, os métodos baseados em densidade são *experts* para filtrar ruídos e descobrir clusters com forma arbitrária, tais como elíptica, cilíndrica, espiralada, etc. até os completamente cercados por outro "cluster" (HAN e KAMBER, 2001).

Para entender a ideia dos métodos baseados em densidade, ESTER et al. (1996) atentam que ao observar um conjuntos de objetos tais como os da Figura 1 a seguir, pode-se, facilmente, e de forma não ambígua, detectar clusters circulares no conjunto 1, clusters de formatos arbitrários no conjunto 2 e clusters de objetos e óbvios ruídos não pertencentes a qualquer dos clusters no conjunto 3. Temos este reconhecimento automático porque sem saber nosso cérebro reconhece visualmente, neste caso bidimensional, que dentro de cada cluster tem-se uma densidade de objetos típica que é consideravelmente maior do que fora dos clusters. Além disso, a densidade de áreas de ruído é menor do que a densidade em qualquer dos clusters. Involuntariamente, o cérebro humano entende a formação dos clusters usando o conceito de grupos formados por densidade para reconhecer os clusters e ruídos nos exemplos mostrados na Figura 1. Um método baseado em densidade clusteriza objetos baseado nesta noção de densidade e capta esta este comportamento, observados de forma visualmente óbvia nos exemplos acima, em conjuntos de dados de maiores dimensões onde a compreensão visual humana não é capaz de trabalhar.

Figura 1. Conjunto de objetos com clusters visualmente não globulares (conjunto 2) e presença de ruídos (conjunto 3).

O DBSCAN, abreviação de Density Based Spatial Clustering of Application with Noise (Clusterização Espacial Baseada em Densidade de Aplicações com Ruído), 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. O método DBSCAN é muito explorado e tem diversas variações, pelo menos 19 algoritmos melhorados DBSCAN foram propostos, nos quais os investigadores tentam melhorar a complexidade técnica de DBSCAN, o tempo de execução e seu desempenho em densidades variadas. 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.

Definição 1: (Eps-vizinhança de um ponto p) A vizinhança de um objeto p com raio Eps é chamada de Eps-vizinhança de p e é dada por: $N_{Eps}(p) = \{ q \in D | dist(p,q) < Eps$.

Definição 2: (Ponto core) Se a vizinhança N_{Eps} de um objeto p 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_ε 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 $\{p_1,..., p_n\}$, 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 em 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 $\boldsymbol{C}$ com respeito a $\boldsymbol{Eps}$ *e MinPts* éum conjunbto não vazio de $\boldsymbol{D}$ satisfazendo as seguintes condições.
(**Maximilidade**)$\forall\, \boldsymbol{p}, \boldsymbol{q}$: se $\boldsymbol{p \in C}$, e $\boldsymbol{q}$ é alcançável por densidade de $\boldsymbol{p}$ com respeito a $\boldsymbol{Eps}$ *e MinPts*. Então $\boldsymbol{q \in C}$.

(Conectividade) $\forall \boldsymbol{p}, \boldsymbol{q} \boldsymbol{\epsilon} \boldsymbol{C}$, $\boldsymbol{p}$ é conectado por densidade a $\boldsymbol{q}$ com respeito a $\boldsymbol{Eps}$ *e MinPts.*

Em outras palavras, um cluster DBSCAN é 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 (ESTER *et al.*, 1998).

Definição 8 (Ruído): Sejam $\boldsymbol{C_1}$, $\boldsymbol{C_2}$, ..., $\boldsymbol{C_k}$, clusters do conjunto de dados D com respeito a $\boldsymbol{Eps}$ *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. Isso qualifica o método em especialista em classificar ruídos que em outros métodos de clusterização, como o k-means, hierárquico, CLARANS, seriam colocados obrigatoriamente em algum cluster, não necessariamente só formados por ruídos. É o que nós precisamos na fase 4 da Análise Espectral Singular de Séries Temporais é isso: distinguir propriamente as componentes de ruído das componentes de tendência e das componentes oscilatórias.

A Figura 2 a seguir mostra o desempenho superior do DBSCAN em identificar formatos arbitrários de clusters, quando comparado com o método SOM e o método k-means em avaliação de desempenho feita por MUNTAZ & DURAISWAMY (2010).

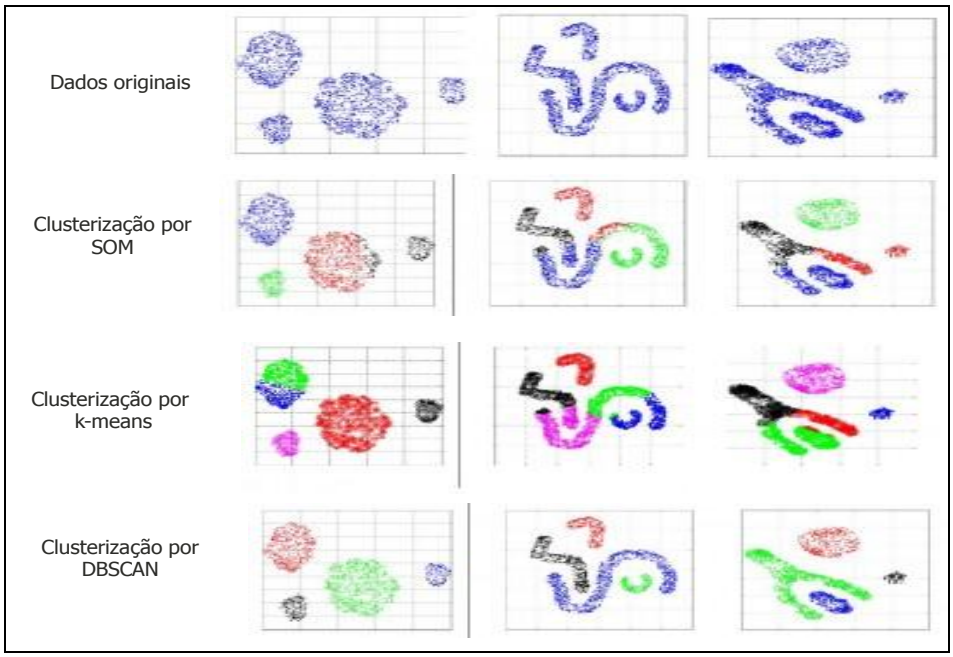

Figura 2: Desempenho de Diferentes Métodos de Clusterização para Dados Espaciais
Fonte: MUNTAZ & DURAISWAMY (2010).

O algoritmo DBSCAN tem uma complexidade computacional da ordem de O(n^2), mas se um índice espacial é usado, como uma *R-tree*, o método DBSCAN alcança melhor desempenho, obtendo a complexidade computacional de O(n log n) (SHEIKHOLESLAMI et al., 1998). ESTER et al. (1996) escrevem que a noção de clusters e o algoritmo DBSCAN se aplicam para espaços *Euclidianos* de duas e três dimensões, como para qualquer espaço métrico característico de alta dimensão, sendo um dos mais eficientes algoritmos em bases de dados grandes. Eles salientam, ainda, que a abordagem trabalha com qualquer função de distância, de maneira que uma função apropriada pode ser escolhida para alguma dada aplicação. A forma da vizinhança é determinada pela escolha da função de distância.

Em ESTER et al. (1996) os autores propõem um método para estimar os parâmetros iniciais Eps e MinPts. Eles recomendam que para um dado k (recomendado ser igual a 4, sem perda de generalidade) seja definida uma função Distk, definida no conjunto de dados D para os números reais, que mapeia cada ponto $p \in D$ do conjunto de dados à distancia de p ao seu k-ésimo vizinho mais próximo. Quando classificados os pontos em ordem crescente dos seus valores de Distk ordenados {Distk}, deve-se traçar o Gráfico de {1,...n} vs {Distk} e procurar pelo 'joelho' da curva resultante, ou seja, o ponto em que há uma mudança nítida na tendência do gráfico. O valor de Distk neste ponto é o valor estimado para Eps; o valor de k é o valor estimado para MinPts. O estudo da distribuição de distk supondo um percentil da distk para separar ruído dos dados também pode ser usado.

Para um conjunto de dados com clusters bem definidos, com fronteiras distantes, o método DBSCAN definido por ESTER et al. (1996) trabalha bem. Mas no caso de existir uma cadeia densa de objetos conectando dois clusters, ou seja, quando pontos de borda de dois clusters estão relativamente muito perto um do outro, GUHA *et al.* (1998) lembram que o DBSCAN original sofre do problema de falta de robustez que também importuna os métodos hierárquicos de clusterização que utilizam todos os objetos: pode acabar por juntar os dois clusters ou atribuir pontos de bordas a cluster errados e crescer os clusters de forma errada perto da borda. Além disso, nestes casos, a clusterização final dependerá da ordem em que os objetos foram processados na fase de extensão do algoritmo.

O que acontece neste caso é que haverá pelo menos um ponto de borda compartilhando cadeias de pontos core alcançáveis por densidade originadas de dois clusters diferentes, ou seja, nestes casos os pontos de borda podem ser alcançados no algoritmo por diferentes caminhos, por diferentes pontos core na Eps-vizinhança deste ponto de borda e assim um ponto de borda pode ser alcançado via por densidade por cadeias de pontos core diferentes, potencialmente originárias de clusters diferentes. Não podendo ser assignado para os dois clusters, este ponto de borda será alcançado pela primeira cadeia visitada pelo algoritmo; e o ponto de borda será assignado ao cluster descoberto primeiro, pna expansão do algoritmo. Uma vez que o primeiro objeto core de cada cluster é qualquer objeto que cumpre a propriedade de core objeto, a ordem de descoberta dos clusters vai interferir totalmente no resultado final da clusterização, mostrando a fragilidade do método. Além disso, a fim de identificar clusters corretamente, objetos dos dados na área de contato poderão ser reconhecidos como pontos de borda. Daí como uma regra, quanto mais áreas de contato no espaço de dados, mais objetos de borda seriam detectados e uma vizinhança mais ampla poderia ser construída equivocadamente. Ou seja o DBSCAN na presença de clusters densos e adjacentes pode produzir uma clusterização sensível a ordem de busca do algoritmo e equivocada: juntando dois clusters e/ou atribuindo pontos de bordas a cluster errados e/ou crescendo os clusters de forma errada perto da borda.

Em TRAN et al.(2013) os autores revisaram o conceito do método DBSCAN e ajustaram o algoritmo para alcançar uma performance robusta contra este problema. O algoritmo de TRAN et al. (2013) preserva todas as características e vantagens do DBSCAN

original de ESTER et al. (1996) e estende a aplicabilidade do método para muitos tipos de dados, ao superar o problema de amostras de fronteira pertencentes a grupos adjacentes.

Neste trabalho, o DBSCAN será utilizado para clusterizar dados muito densos, d séries de tamanho N, para tal esta robustez se faz necessária e por isso essa nova versão do DVSCAN Revisado é a versão utilizada para os resultados.

A cadeia de objetos alcançáveis por densidade $\{p_1,..., p_n\}$ ou pode estar na forma $\{p_{core\ 1},..., p_{core\ n-1,}\ p_{core\ n}\}$ com todos os pontos core ou $\{p_{core\ 1},..., p_{core\ n-1,}\ p_{border}\}$com todos os core objetos exceto o ultimo objeto sendo um ponto de borda. O objeto de borda, portanto, não contribui para o mecanismo da expansão da cadeia de alcance por densidade, o passo essencial do DBSCAN. Por esta razão, a versão revisada e melhorada do DBSCAN, proposta por TRAN et al. (2013) tem por objetivo desconectar o ultimo ponto de borda da cadeia de alcance por densidade.

Isto é alcançado através do conceito adicional de objetos core alcançáveis por densidade no curso da Clusterização, que são cadeias de objetos $\{p_1,..., p_n\}$ onde p_i é um ponto core, para todo i. Daí na nova versão do algoritmo DBSCAN, utilizada neste trabalho, o passo de expansão é revisado a fim de usar cadeias de core alcançáveis por densidade, somente com pontos core, ao invés da cadeia de pontos alcançáveis por densidade. Uma vez que elas contêm um número similar de pontos core, a abordagem antiga e revisada identificam os mesmos objetos core. Entretanto, pontos de borda (originalmente designados durante o passo de expansão) permanecem temporariamente não classificados até todos os pontos core do cluster serem identificados. Somente após a detecção de todos os grupos, quando todas cadeias alcançáveis por densidade que alcançam tal objeto core já são conhecidas, no ultimo passo do DBSCAN revisado, é que cada objeto de borda passa a ser designado a sua melhor cadeia alcançável por densidade, que é aquela cadeia alcançável por densidade mais próxima, e então este ponto de borda é assignado ao cluster ao qual essa cadeia de objetos core alcançáveis por densidade mais próxima pertence. Esta versão revisada do DBSCAN fornece uma atribuição robusta de um ponto de borda para o cluster esperado que independe da ordem em que os clusters são descobertos.

4. DESCRIÇÃO DA ANÁLISE E FERRAMENTAS COMPUTACIONAIS

Os experimentos foram realizados seguindo a seguinte metodologia: para uma dada série $\{Y_t: 1 \leq t \leq N\}$ e $L = N/2$, aplica-se o procedimento SSA até a fase 4. Ou seja, faz a incorporação de Y_t em uma matriz trajetória X; faz a SVD de X; pela média diagonal e obtém-se uma série temporal a partir de cada matriz componente de X. Tem-se então $d = L$ séries de tamanho T a serem agrupadas na fase 4.

O próximo objetivo é então realizar a identificação das componentes de ruído entre estas d séries. Aqui que a proposta trabalha. Nesta fase, procede-se para o agrupamento padrão por Análise Gráfica dos Vetores Singulares, pelo método de clusterização hierárquica e pelo DBSCAN. Daí, é feita a reconstrução da nova série temporal para cada uma das abordagens. Realiza-se a modelagem e previsão de cada uma das séries reconstruídas. Também é feita a modelagem e previsão das séries puras, sem abordagem SSA, para comparação dos resultados.

Todas as modelagens e previsões deste trabalho são realizadas no software *FPW-Pro for Windows*. Para cada série de entrada, este software encontra entre modelos tradicionais, de suavização exponencial e SARIMA do enfoque Box-Jenkins, o modelo que melhor se adequa a série fornecida e identifica tal modelo automaticamente.

Após modelar cada série depois das quatro abordagens empregadas: Sem SSA, SSA+Análise dos Autovetores, SSA+Hierárquico, SSA+DBSCAN, compara-se a qualidade da previsão destas abordagens para da série Y_t, pelo MAPE calculado com referência à série original.

A metodologia foi experimentada em séries sintéticas simuladas de modelos Box-Jenkins e séries simuladas de passeio aleatório. As séries nomedas do tipo Y1 são séries simuladas segundo um modelo AR(1) ; as séries do tipo Y2 são séries também simuladas que seguem um modelo MA(1). As séries Y3 são séries simuladas de modelos ARMA(1,1). As séries Y4 são de séries de um modelo ARIMA (0,1,1); as séries Y5 são séries de modelos ARIMA(1,1,2) e as séries Y6 são séries simuladas de um processo passeio aleatório com drift. Sendo assim, as séries do tipo Y1, Y2 e Y3 são amostras de processos estacionários e as séries Y4, Y5 e Y6 são amostras de processos não estacionários. Todos os modelos são definidos com pertubações gaussianas com média 0 e variância 1. Para cada tipo de séries foram simuladas 100 séries de tamanho 500, feito o procedimento proposto nas 4 abordagens e calculado a média do MAPE – Mean Absolute Percentage Error, para cada tipo de série, em cada abordagem.

A aplicação a dados reais é feita pra uma série de dados de média mensal de velocidade do vento, em m/s, medida de 10 em 10 minutos à 50m de altura na estação anemométrica de Petrolina no período de janeiro de 1996 até dezembro de 2012, N=192 observações. Os dados são provenientes do Instituto Nacional de Meteorologia – INMET. Além do MAPE, foram usadas as medidas de qualidade de ajuste RMSE (Root-Mean Square Error) e MAE (Mean Absolute Error).

Para obtenção dos resultados foram utilizados os seguintes softwares para análise e programação: R, MatLab, Cartepillar (GOLYANDINA & OSIPOV, 2007) e FPW (Forecast Pro for Windows).

5. RESULTADOS

A Tabela 1 a seguir traz o resultado da média do MAPE (in sample) de 100 séries de cada um dos tipos Y1, Y2, Y3. A partir dos valores apresentados observa-se que a abordagem SSA+DBSCAN garante para todos os tipos de modelo observados, previsões de melhor qualidade, ao apresentar os menores valores de MAPE. Além disso, a previsão pura da série sem utilizar SSA é, para todos os casos, a pior dentre as abordagens usadas para previsão de séries temporais.

Tabela 1: Média do MAPE da previsão de séries de Modelos AR(1); Modelos MA(1) e modelos ARMA(1,1), ARIMA (0,1,1), ARIMA(1,1,2) e passeio aleatório com drift,.

Método usado para tratar as séries antes da modelagem e previsão	Média do MAPE (in sample) N=500; 100 replicações					
	Y1	**Y2**	**Y3**	**Y4**	**Y5**	**Y6**
Nenhum	1.357	1.772	3.678	4.587	3.845	2.646
SSA+ Análise Autovetores	1.173	1.921	1.859	2.613	0.5846	2.169
SSA+Hierárquico	1.089	1.847	1.893	1.999	0.9852	2.151
SSA+DBSCAN	0.899	1.795	1.101	0.614	0.5209	0.3633

Quando aplicada a metodologia a série real de velocidade do vento, verificou-se também melhor eficiência da abordagem proposta SSA combinada com DBSCAN, conforme medidas de qualidade de ajuste listadas na Tabela 2 a seguir. O modelo ajustado para a série de velocidade do vento é um modelo ARIMA(1,1,2)*(1,1,2) com transformação logarítimica, onde a(1)=-0.458; b(1)=-0.076; b(2)= -0.773; A(12)=0.982; B(12)=0.010; B(24)=0.786.

A Figura 3 exibe o comportamento da série original, dos valores ajustados da combinação SSA+DBSCAN e do modelo ajustado sem considerar a abordagem SSA, que apresentou o pior rendimento observado. Como pode ser visto no gráfico, sem usar a abordagem SSA a suavização da série fica muito destoante da série original, ocasionando altos índices de erros in sample. Este resultado confirma o sucesso da abordagem SSA na análise de modelagem e previsão de séries temporais.

Tabela 2: Medidas de qualidade do ajuste para os modelos de previsão propostos aplicados a série de velocidade do vento de Petrolina.

Medida de Erro de Previsão	Método usado para tratar as séries antes da previsão			
	Nenhum	SSA+Análise dos Autovetores	SSA+Hierárquico	SSA+ DBSCAN
MAPE	10.04%	6.78%	6.93%	4.98%
RMSE	0.775	0.453	0.398	0.151
MAE	0.708	0.595	0.632	0.336

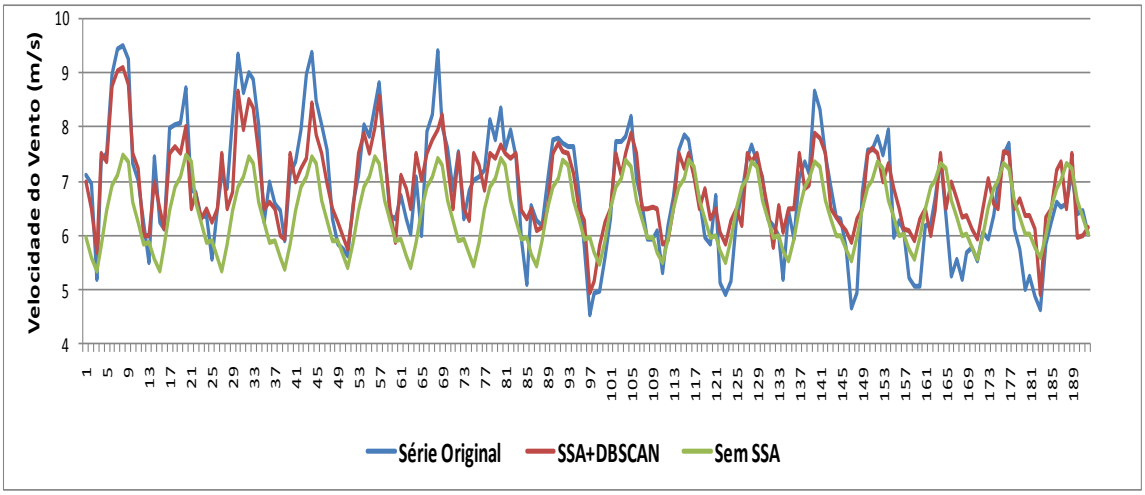

Figura 3: Série Original e Valores Ajustados da cobinação SSA+DBSCAN, Valores Ajustados para Modelagem sem SSA.

6. CONCLUSÕES

Os resultados obtidos apontam para o sucesso da proposta, a combinação do método de clusterização DBSCAN com SSA para filtragem e posterior modelagem e previsão de séries temporais. A abordagem que usa a combinação SSA+DBSCAN rendeu, em todos os casos simulados, estacionários e não estacionários, e na série real de velocidade do vento, previsões de melhor qualidade quando comparadas com aquelas obtidas pelos outros métodos tradicionais, considerando como critério o MAPE, o RMSE e o MAE. Tal resultado de previsão sugere afirmar que o DBSCAN quando aplicado a SSA está separando melhor o que são de fato as componentes de ruído e fornecendo uma série menos ruidosa para previsão. Importante notar também que em todos os casos, a previsão sem utilizar SSA é a de menor eficiência em todas as avaliações, confirmando o sucesso do uso da Análise Singular Espectral na Análise de Séries Temporais antes da modelagem e previsão.

Uma idéia para trabalhos futuros seria usar o DBSCAN com Distância de Mahalanobis, ao invés da distância euclidiana. A Distância de Mahalanobis leva em consideração a matriz de covariâncias calculada com todos os objetos. A distância de Mahalanobis é amplamente utilizada em análise de clusters, mas no caso específico desta

abordagem, em que os objetos a serem clusterizados são séries temporais, considerar a matriz de covariâncias pode ser mais interesante ainda, capturando esta informação de covariância entre as séries e deixar isso influenciar na formação dos clusters que apresentem a máxima verossimilhança.

Outra idéia seria utilizar DBSCAN usando distâncias adaptativas tal como foi usada com sucesso por Cavalcanti Junior (2006) para o método de Clusteriação fuzzy c-Means e por Souza *et al.*, 2003 no método de clusterização k-means. Nesta abordagem deseja-se clusterizar séries temporais, sabe-se que uma série pode ter muita similaridade a outra em um período e nem tanta similaridade em outro, as distâncias adaptativas podem ajustar bem essa característica.

A metodologia proposta também pode ser mais exaustivamente testada em modelos sintéticos mais complexos e combinada com previsão SSA ao invés da previsão por modelos clássicos como foi feito neste trabalho . A previsão SSA é uma abordagem recente que vem ganhando espaço e popularidade ao apresentar melhores resultados do que os métodos de modelagem e previsão clássicos de séries temporais, como pode ser visto em Hassani (2007) ; Hassani et al. (2009), Esquivel (2012), Hassani et al. (2013)-b e Hassani et al (2013)-b.

7. REFERÊNCIAS

Cavalcanti Júnior, N. L. (2006). ***Clusterização baseada em Algoritmos Fuzzy.*** Dissertação de Mestrado em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco.

Driver, H. E. & Kroeber, A. L. (1932) Quantitative Expressions of Cultural Relationships. Berkeley: University of California Press.

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

Ester, M., Kriegel, H.P, Sander, J, Wimmer, M. & Xu, X. (1998) Incremental Clustering for Mining in Data Warehousing Environment. Proceedings of the 24th VLDB Conference, New York, USA.

Esquievel, R.M. (2012). Análise Espectral Singular: Modelagem de séries Temporais através de etstudos comparativos usando diferentes estratégias de previsão. Dissertação de Mestrado em Modelagem Computacional e Tecnologia Industrial do SENAI-CIMATEC. Salvador-BA.

Fayyad U. M.,.J., Piatetsky-Shapiro G., Smyth P. (1996)"From Data Mining to Knowledge Discovery: An Overview*"*, in: "Advances in Knowledge Discovery and Data Mining", AAI Press, Menlo Park, 1996, pp. 1 - 34.

Golyandina, N. (2010). Singular Spectrum Analysis: Methodology and Comparison. *Journal of Data Science*. 5, 239-257.

Golyandina, N., Nekrutkin, V. & Zhihgljavsky, A. (2001). *Analysis of time series structure: SSA and related techniques.* Chapman & Hall/CRC. New York, USA.

Golyandina, N., Osipov, E. (2007). Caterpillar, SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference*. Vol 137, 8, 2642-2653.

Guha,S., Rastogi, R. & Shim, K. (1998). *CURE: An efficient clustering algorithm for large databases*. In Proc. ACM SIGMOD. Int. Conf. Management of Data, pp. 73–84.

Han, J., & Kamber, M. (2001). *Cluster Analysis*. In: Morgan Publishers (eds.), *Data Mining: Concepts and Techniques*, 1 ed., chapter 8, NewYork, USA, Academic Press.

Han, J. & Ng, R.T. (1994) Efficient and Effective Clustering Methods for Spatial data Mining. Proceedings of the 20th VLDB Conference Santiago, Chile. pp. 144-155.

Hassani, H. (2007). Singular Spectrum Analysis: Methodology and Comparison**.** *Jornal of Data Science* 5, 239-257.

Hassani, H.; HERAVIC, S.; ZHIGLJAVSKYA (2009) Forecasting European Industrial Prodution with SSA. International of Forecasting 25,103-118.

Hassani, H., Heravi, S., Brown, G. & Ayoubkhani, D. (2013a) Forecasting before, during, and after recession with singular spectrum analysis. *Journal of Applied Statistics*, DOI: 10.1080/02664763.2013.810193. Vol. 40 (10), pp. 2290 - 2302.

Hassani, H., Soofi, S. & Zhigljavsky, A. (2013b). Predicting Inflation Dynamics with Singular Spectrum Analysis. *Journal of the Royal Statistical Society*. Vol. 176 (3), pp. 743 – 760. DOI: 10.1111/j.1467-985X.2012.01061.x.

Muntaz, K & Duraiswamy, K. (2010) An Analysis on Density Based Clustering of Multi Dimensional Spatial Data. *Indian Journal of Computer Science and Engineering*, Vol 1(1):8-12.

Sheikholeslami, G., Chatterjee, S. & Zhang, A. (1998). WaveCluster: A multiresolution clustering approach for very large spatial databases**.** In Proc. 24th VLDB Conf., pp. 428–439.

Sorensen T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content / Kongelige Danske Videnskabernes Selskab. Biol. krifter. Bd V. n.o 4.pg 1-34.

Souza, R. M. C. R.. (2003). *Métodos de cluster para intervalos usando algoritmos do tipo nuvens dinâmicas*. Tese de Doutorado, Centro de Informática-Universidade Federal de Pernambuco.

Tran, T. N., Drab, K., Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, Vol.120. pp. 92-96.

Tryon, R. (1939). *Cluster Analysis*. New York: McGraw-Hill.

Yin, J., Zhou, D., & Xie, **Q-Q.** (2006) A Clustering Algorithm for Time Series Data. Proceedings of the Seventh International Conference on Parallel and distributed Computing of IEEE.

Zubin, J. A. (1938). "A technique for measuring likemindedness". *Journal of Abnormal and Social Psychology*, 33, pp.508-516.