1. Introdução

Uma das abordagens de sucesso mais recentes na área de previsão de séries temporais é a utilização da Análise Espectral Singular, SSA (do inglês, Singular Spectrum Analysis). A SSA é uma técnica relativamente nova e poderosa que diminui o ruído nas séries incorporando elementos de Análise Clássica de Séries Temporais, Estatística Multivariada, Geometria Multivariada, Sistemas Dinâmicos e Processamento de Sinais (HASSANI, 2007). Pouco explorada no Brasil, a metodologia SSA qualifica-se por ser uma abordagem completamente não paramétrica de análise de séries temporais, portanto não requer que as séries temporais sejam estacionárias e nem é necessário fazer suposições para as distribuições e parâmetros dos dados e ruídos. O objetivo do método é fazer Decomposição de Valor Singular da série em vários componentes aditivos, que podem tipicamente ser interpretados como componente de tendência, componentes oscilatórios e componentes de ruído. Em particular, para previsão, objetiva-se reconstruir a série separando os componentes de ruído, que serão desconsiderados.

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. Cada componente desta expansão concentra uma parcela da energia contida na matriz trajetória gerada a partir da série temporal (HAMILTON, 1994). Dessa forma, um subconjunto de componentes concentra a maior parte da energia total com estrutura de dependência temporal, 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 com menos ruído da série temporal original.

Tal análise básica do SSA é executada em quatro fases: 1) Decomposição da Série em uma Matriz *Hankel*, 2) Decomposição da Matriz *Hankel* via *Singular Value Decomposition* (SVD), 3) Agrupamento das Componentes da série (tendência, componentes harmônicas e ruído, com vistas à retirada do ruído) e 4) Reconstrução da série via Média Diagonal. Resumidamente, o procedimento de SSA objetiva expandir uma série temporal de interesse em termos de uma Soma Ponderada de Componentes Ortogonais (Auto-Vetores) onde cada componente pode ser identificada como tendência, componente periódica ou ruído. Em seguida, é realizado o Processo de Linearização dos Autovalores, e após isso, a série temporal original é reconstruída, com menos ruído, para modelagem e previsão.

1.1. Motivação

Até o presente momento, todos os trabalhos envolvendo SSA utilizaram um dos seguintes métodos na fase 3 do método SSA, a fase de separação das séries de ruído: Análise de Componentes Principais; Análise Gráfica dos Autovetores ou Clusterização Hierárquica.

O método de clusterização hierárquica até então é a última inovação na separação de ruído da abordagem SSA. Entretanto, há um consenso na literatura de que tal método de Clusterização é muito sensível a ruído, não deve ser usado em clusters com densidades variadas e não é eficiente na clusterização de séries temporais de diferentes tendências. Ao contrário, os métodos de Clusterização Baseados em Densidade são eficientes em separar o ruído dos dados e dedicados a trabalharem bem em dados de diferentes densidades (YIN et al., 2006). Além disso, ESTER et al. (1996) lembram que o principal problema com algoritmos de agrupamento hierárquico é a dificuldade de obter os parâmetros adequados para a condição de término. Em geral, é usada como condição de término a distância crítica D_{mim} entre todos os clusters, mas a escolha de um valor de D_{mim} não é fácil. O valor D_{mim} deve ser pequeno o suficiente para separar todos os clusters "naturais", e ao mesmo tempo suficientemente grande de tal modo que nenhum conjunto seja dividido em duas partes.

GUHA *et al.* (1998) também afirmam que há um problema de falta de robustez que importuna os métodos hierárquicos de clusterização: no caso de existir uma 'corrente densa' de objetos conectando dois clusters, ou seja, quando pontos de borda de dois clusters estão relativamente perto um do outro, o que acontece muito quando se tem dados densos, o método pode acabar por juntar os dois clusters.

Se ainda com estas características, o método de clusterização hierárquica tem demonstrado boa eficiência, quando comparado aos métodos da Análise de Componentes Principais e Análise Gráfica dos Autovetores, torna-se interessante investigar se um método de clusterização mais moderno pode superar seu desempenho. NG e HAN (1994) afirmam que os clusters produzidos por métodos não hierárquicos são de qualidade superior aos produzidos por métodos hierárquicos, e que, por isso, o desenvolvimento de métodos não hierárquicos tem sido um dos principais focos de pesquisa de análise de clusters, havendo muitos métodos particionais e baseados em densidade, grade ou em inteligência artificial descritos na literatura. O método hierárquico foi proposto em 1955 e desde então foram desenvolvidos pelo menos 125 novos algoritmos de Clusterização. É importante a investigação de novos modelos para a abordagem SSA, trazendo pra SSA a evolução alcançada na área de Clusterização nestes anos.

O estudo desta tese iniciou-se então com o intuito de utilizar um método de clusterização não hierárquico 'especializado' na separação de ruído, que determina automaticamente o número natural de clusters, robusto ao problema de fronteira de clusters adjacentes e que não tenha sido ainda aplicado a estudos da área de séries temporais nem de SSA.

A presente pesquisa também foi motivada especificamente pelo interesse de propor métodos de previsão inéditos ou melhorar métodos de previsão já conhecidos, que possam ser usados para previsão de séries relacionadas à indústria elétrica (séries de demanda de energia, séries de potência, séries de consumo, séries de vazão, séries velocidade do vento e outras).

1.2. Objetivo do Trabalho

A presente tese tem por objetivo propor o uso de um método de Clusterização Baseada em Densidade eficiente para separação do ruído, o DBSCAN (*Density Based Spatial Clustering of Applications with Noise*), a ser utilizado na terceira fase do método SSA, para análise de séries temporais.

Tem-se também por propósito aplicar tal combinação de metodologia na previsão de séries temporais de velocidade do vento.

1.3. Relevância do Tema

Sob o ponto de vista metodológico, esta tese traz uma contribuição diferenciada que é a combinação inédita do uso de um Método de Clusterização Baseado em Densidade Revisado com a análise de séries temporais por SSA, que pode ser usada na modelagem e previsão de qualquer série temporal. A análise de séries temporais por SSA é uma abordagem relativamente recente que está sendo muito explorada e investigada na atualidade.

Sob o ponto de vista da aplicação do método, esta tese traz tal metodologia inédita para previsão de velocidade do vento. A produção de energia elétrica de fonte eólica representa importante avanço no que se refere ao uso dos recursos naturais renováveis e vem sendo explorada com mais intensidade desde 2001 no Brasil. A ABEEólica, Associação Brasileira de Energia Eólica, avalia que o Brasil tem um potencial de 300 GW de potencial eólico, com a contratação de no mínimo 2 GW por ano até 2020. Com este crescente potencial eólico brasileiro e uso cada vez mais acentuado deste tipo de energia, a previsão de velocidade do vento também se torna necessária e, dentro deste contexto, ferramentas mais precisas de previsão tornam-se cada vez mais necessárias.

1.4. Organização da Tese

Esta tese está organizada em 10 capítulos. No segundo capítulo disserta-se sobre Análise de Séries Temporais e no terceiro capítulo é apresentada a técnica *Singular Spectrum Analysis*, SSA. O quarto capítulo traz aspectos sobre Clusterização de Dados. O método DBSCAN é definido e discutido no Capítulo 5. No capítulo 6 é relatada qual será a metodologia adotada e são listadas as ferramentas computacionais utilizadas. O capítulo 7 apresenta resultados da aplicação da metodologia proposta SSA+Métodos de Clusterização às séries sintéticas. No oitavo capítulo são apresentados os resultados de aplicação a uma série de velocidade do vento. Finalmente, as conclusões são apresentadas no Capítulo 9. As referências bibliográficas citadas são discriminadas no Capítulo 10.