6. Metodologia

A proposta dessa tese é 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*) na terceira fase do método SSA, para análise de séries temporais. Para investigar a eficiência desta proposta, a abordagem foi aplicada exaustivamente a modelos sintéticos de séries temporais.

Os experimentos foram realizados seguindo esta metodologia: para uma dada série $\{y_t: 1 \leq t \leq N\}$, aplicou-se o procedimento SSA até a fase 3. Ou seja, foi realizada a primeira fase de SSA, fazendo a incorporação de Y_t em uma matriz trajetória X; na segunda fase SSA fez-se a SVD de X; obtendo-se assim a decomposição de X em d matrizes. As fases seguintes de SSA são as fases de agrupamento e média diagonal.

O padrão teórico é que o agrupamento seja feito nas matrizes, mas na prática, sem perda de generalidade, para alguns métodos de agrupamentos é mais conveniente primeiramente fazer a média diagonal das d matrizes resultantes da SVD, e obter d séries componentes da série temporal, e depois fazer agrupamento destas d séries temporais. Ou seja, pode-se apropriadamente, inverter a ordem dos procedimentos na fase de reconstrução SSA. Em outras palavras, pode-se fazer o agrupamento baseando-se não nas matrizes nem nas séries resultantes das matrizes da expressão (24) por média diagonal, mas nos d autovalores ou d autovetores. Isto porque cada uma das d matrizes está biuvunicamente relacionada a cada um dos d autovalores e a cada um dos d autovetores, e fazendo a média diagonal de cada uma destas d matrizes, também teremos d séries temporais componentes componentes da série original biunivocamente relacionada às d matrizes.

Na terceira fase, pela média diagonal, obteve-se uma série temporal a partir de cada matriz componente de X. Tinha-se então, para a série $\{y_t: 1 \leq t \leq T\}$, d séries de tamanho T a serem agrupadas na fase 4. O próximo objetivo foi então realizar a

identificação das componentes de ruído entre estas d séries. Nesta fase, procedeu-se para 4 abordagens distintas:

Abordagem 1: Identificação padrão por Análise Visual Gráfica dos Vetores Singulares;

Abordagem 2: Identificação por Análise de Componentes Principais (ACP)

Abordagem 3: Identificação por Clusterização Hierárquica (HC)

Abordagem 4: Identificação por Clusterização realizada por DBSCAN

Pra cada uma das abordagens é feita a reconstrução da nova série temporal Realiza-se a modelagem e previsão de cada uma das séries. Também é feita a modelagem e previsão da série sem abordagem SSA, que vamos chamar de **Abordagem 5**, para analisar a eficiência do uso de SSA na previsão de séries temporais. As abordagens 1,2,3 e 5 são clássicas e conhecidas e a abordagem 4 é a inovação proposta nesta tese.

Uma vez modelada a série original $\{y_t: 1 \leq t \leq N\}$ depois das 5 abordagens empregadas de filtragem, Previsão usando SSA+Análise dos Autovetores, Previsão usando SSA+ ACP; Previsão usando SSA+HC, Previsão usando SSA+DBSCAN e Previsão Sem SSA; observou-se a qualidade das previsões pelos valores de MAPE (Mean Absolute Percentage Erro) e RMSE (Root-Mean-Square Error) de cada abordagem, calculados com referência à série original y_t como em (14) e (12).

6.1. Processos Geradores de Dados Simulados

A metodologia foi experimentada em séries sintéticas simuladas de modelos descritos nas Seções 2.4 e 2.5. As séries y_t foram obtidas considerando 4 processos estacionários 4 processos não estacionários.

a) Processsos geradores de dados estacionários
   1. $y_t = \varepsilon_t,\ \varepsilon_0$=0
   2. AR(1); $\emptyset_1 = 0.4$
   3. MA(2); $\theta_1 = -0.3$ e $\theta_2 = 0.8$
   4. ARMA (1,2) ; $\emptyset_1 = 0.4, \theta_1 = -0.3$ e $\theta_2 = 0.8$

b) Processos Geradores de Dados não estacionários

5. Passeio Aleatório com drift $\mu = 0.1$ representando crescimento suave do componente de tendência determinística.
6. Passeio Aleatório com drift $\mu = 0.6$ representando crescimento brusco do componente de tendência determinística.
7. ARIMA (0,1,1) ; $\theta_1 = 0.4$
8. ARIMA (1,1,2); $\emptyset_1 = 0.4, \theta_1 = -0.3$ e $\theta_2 = 0.8$.

Os erros ε_t são ruído branco, possuindo média 0 (quando existente) e variância constante (quando finita). Neste trabalho foi considerada distribuição normal padrão para ε_t. A escolha destes modelos e parâmetros foi motivada pelos trabalhos de CASSIANO (2003) e ESQUÍVEL (2012). Os processos geradores considerados aqui são simples, mas são suficientes para oferecer uma noção razoável do comportamento das abordagens propostas.

As simulações destes processos foram realizadas através da função `arima.sim` do software R, que gera séries da classe dos modelos ARIMA. Para o passeio aleatório foi usada a função `cumsum`. Para cada tipo de série especificado acima, foram simuladas 100 séries de tamanho 500, para cada série é feito o procedimento proposto nas 5 abordagens, é feita a modelagem e calculadas as médias dos 100 valores de MAPE e do RMSE.

6.2. Ferramentas Computacionais

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

O Caterpillar foi desenvolvido por um grupo de pesquisadores chamado GistaT, da Universidade de São Petersburgo na Rússia especialmente para análises em SSA. Informações sobre o software e distintas versões para análise são disponíveis na página oficial do Gistat http://www.gistatgroup.com.