UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

Previsão de Séries Financeiras utilizando Métodos de Clusterização e Máquinas de Vetor de Suporte

Lucas Faria e Souza Vilela

Itajubá, Setembro de 2016

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

Lucas Faria e Souza Vilela

Previsão de Séries Financeiras utilizando Métodos de Clusterização e Máquinas de Vetor de Suporte

Dissertação submetida ao Programa de Pós-Graduação em Ciência e Tecnologia da Computação como parte dos requisitos para obtenção do Título de Mestre em Ciência e Tecnologia da Computação

Área de Concentração: Matemática da Computação

Orientador: Prof. Dr. Otávio Augusto Salgado Carpinteiro
Coorientador: Prof. Dr. Carlos Alberto Murari Pinheiro

Setembro de 2016
Itajubá - MG

Resumo

Prever valores futuros de séries temporais financeiras é assunto de estudos e pesquisas há diversas décadas. Muitas propostas, utilizando modelos matemáticos lineares e não lineares, ou utilizando inteligência artificial, já foram formuladas, e os resultados vêm se aprimorando conforme os estudos avançam.

As séries temporais se caracterizam por possuírem diferentes contextos com o passar do tempo. Existem períodos de baixa e alta volatilidade, períodos com regime de expansão e de recessão, entre outros. Captar os contextos e tratá-los de forma distinta é desejável, visto que a relação entre os contextos é pequena ou nula.

Este trabalho propõe a aplicação de métodos de clusterização na série temporal, de forma a separar as informações da série em seus diversos contextos, chamados neste estudo de *clusters*. Os métodos *K-Means* e *C-Means* foram utilizados para este fim. Após este processo, uma SVM por *cluster* é treinada com as informações pertinentes apenas ao seu *cluster*. Desta forma, deseja-se inibir a influência de informações de contextos diferentes dentro de um *cluster*.

Utilizando uma série financeira de um fundo de ações de um banco brasileiro, os resultados mostraram-se positivos e superiores em precisão, quando comparados a uma arquitetura sem tratamento de contexto e a um estudo cujo tratamento de contexto se dá por meio de uma SOM. Tendo o tratamento de contexto se mostrado benéfico em estudos passados, esta proposta vem apresentar que ganhos ainda maiores podem ser obtidos quando a série é previamente processada de forma adequada, trabalhando assim no sentido de melhorar a exatidão do elemento previsor do sistema, a SVM.

Abstract

Financial time series prediction is subject of several studies and research over decades. Proposals using linear and nonlinear mathematical models and artificial intelligence have already been formulated and results are improving as research advance.

A key characteristic of time series is that they have different contexts. They have windows of high and low volatility, periods in that they are expanding and there are times of depression, and several others. Detect these contexts and work them separately is desirable, as there is no relationship between them.

This study proposes to split the time series into its several contexts by means of clustering methods, such as K-Means and C-Means. After that, SVM experts can be created, one for each cluster or context. By doing this, information that does not belong to a certain cluster or context will not affect learn acquired by its SVM.

To assess the model, it is used a financial time series from a stock market fund that belongs to a Brazilian bank. Prediction results are more precise when compared to an architecture that does not handle context and also when compared to a study which treats context using SOM. As past studies suggests that context handling improves prediction accuracy, this proposal shows that even better results can be obtained when time series is correctly processed, aiming to improve architecture's predictor node accuracy, which is SVM.

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

ANIFS	*Adaptive Neuro-Fuzzy Inference System*
ARIMA	*Autoregressive Integrated Moving Average*
BOVESPA	*Bolsa de Valores de São Paulo*
CEL	*Composição de Especialistas Locais*
COMIT	*Índice da bolsa de valores italiana*
CRB	*Carbon Copy*
DAX	*Deutscher Aktienindex*
DS	*Directional Symmetry*
FTSE	*Financial Times Stock Exchange*
HNM	*Modelo Neural Hierárquico*
HNM-V	*Modelo Neural Hierárquico com Volatilidade*
HSI	*Hong Kong Hang Seng Index*
IA	*Inteligência Artificial*
IBrX	*Índice Brasil*
INFS	*Integrated Nonlinear Feature Selection*
LIBSVM	*Biblioteca de software para Máquinas de Vetor de Suporte*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MLP	*Multi-Layer Perceptron*
MSE	*Mean Squared Error*
NASDAQ	*National Association of Securities Dealers Automated Quotations*
NIKKEI	*Índice da Tokyo Stock Exchange*
NMSE	*Normalized Mean Square Error*
RMSE	*Root Mean Square Error*
RS	*Rough Sets*
S&P	*Standard & Poors's*
SOM	*Self-Organizing Map*
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
TAIEX	*Taiwan Exchange Capitalization Weighted Stock Index*
WDS	*Weighted Directional Symmetry*

1 Introdução

1.1 Descrição do problema

A criação de modelos que descrevam o comportamento de uma série de preços de um recurso ou ativo é, historicamente, um dos primeiros aspectos já analisados em econometria, sendo a questão sobre o preço ser previsível uma das primeiras e mais duradouras perguntas deste ramo de pesquisa (CAMPBELL; LO; MACKINLAY, 2016).

Séries financeiras estão entre as mais ruidosas e mais difíceis de prever, sendo até considerada por alguns economistas como imprevisível e independente do passado (ABU-MOSTAFA; ATIYA, 1996). O preço de um ativo é a combinação de diversas informações vindas de fontes distintas, como comportamento do preço no passado, expectativa do ativo para o futuro, notícias e informações relevantes, rumores etc.

Além de possuir ruído, séries financeiras são não estacionárias. O ruído agrega informações errôneas ao modelo da série. Já a não estacionariedade torna a relação entre as variáveis de entrada (informações que geram o preço) e saída (o preço do ativo) dinâmica ao longo da série. Ambos os aspectos tornam mais desafiador o processo de previsão (CAO; TAY, 2001b)

Previsão de séries financeiras possui larga utilização no mercado de ações, servindo como indicador de operações para investidores. A decisão por comprar ou vender ações baseada em um sistema de previsão possui relação direta com seu resultado, sendo a precisão do sistema um aspecto fundamental para o sucesso do negócio, daí a importância de se desenvolver sistemas confiáveis.

Além da utilização no mercado de ações, a previsão de séries financeiras é também utilizada na compra e venda de contratos futuros, por exemplo, de *commodities*. A compra ou venda de um contrato a ser feito no futuro precisa ser precificada no momento do negócio, e a previsão, mais uma vez, desempenha papel importante neste processo.

A previsão de séries temporais não se restringe apenas às séries financeiras. Muitas

são as aplicações de previsão, como, por exemplo, previsão por demanda de carga de energia elétrica, previsões meteorológicas etc.

1.2 Propostas existentes

A história de construção de modelos que reproduzam o comportamento de séries temporais inicia-se com modelos lineares, como modelos autorregressivos e médias móveis, que não eram capazes de refletir fenômenos importantes de referidas séries, entre eles ciclos de mercado assimétricos, volatilidade do mercado de ações, e outros.

Devido às deficiências de tais modelos, os não lineares ganharam a atenção dos pesquisados no último quarto do século passado (TONG, 2002). Particularmente em séries financeiras, as não linearidades são caracterizadas pela presença de pelo menos dois regimes, de recessão e expansão, e de variáveis financeiras, como alta e baixa volatilidade(CLEMENTS; FRANSES; SWANSON, 2004). Apesar de adicionar complexidade ao modelo, e de muitos pesquisadores dizerem que tal complexidade extra não leva a ganhos significativos de precisão comparados aos modelos lineares, muitos são os trabalhos envolvendo não linearidade.

Paralelos aos sistemas matemáticos não lineares estão os sistemas construídos baseados em inteligência artificial, como redes neurais, algoritmos genéticos, enxame de partículas e máquinas de vetores de suporte (SVMs). Tais modelos, além de serem não lineares, possuem complexidade menor que os modelos matemáticos. Porém, necessitam de escolha de valor para parâmetros livres, sendo este um ponto crítico, visto que não existem métodos para obtenção de seus valores ótimos e, também, eles são dependentes do problema.

Muitos estudos preocupam-se em criar um sistema que contém apenas um elemento de IA que modele a série temporal e seja capaz de prever seus valores para novas entradas. Nestes, é muito comum o uso de redes neurais e SVMs. O elemento previsor é treinado com informações extraídas da série temporal, tendo seus parâmetros ajustados de forma a obter o melhor resultado de regressão. Esta técnica foi aplicada com sucesso em diversas séries diferentes, apresentando bons resultados finais.

Outras propostas, mais elaboradas, visam dar algum tipo de tratamento à informação existente na série, para posterior aplicação em um elemento previsor. A série apresenta, ao longo de sua existência, diferentes comportamentos para períodos distintos, sendo a relação entre eles nula ou praticamente nula. A primeira camada é responsável por separar

os dados da série nestes contextos, e é comum o uso de SOMs e algoritmos genéticos para este fim. Assim, a segunda camada, composta por um elemento previsor, normalmente uma rede neural ou SVM, é treinada levando em consideração também informações de contexto.

Os dados recebidos pela segunda camada podem conter a informação de contexto - ficando a cargo do elemento previsor tratá-lo de forma adequada -, ou podem estar separados por contexto - assim, cada subconjunto de dados pode ser tratado por um elemento exclusivo. Quando os dados estão separados por contexto, a informação de saída pode ser relativa apenas ao elemento cujo contexto seja pertencente aos novos dados, ou uma somatória ponderada da saída de todos os elementos previsores.

Mais uma vez, os parâmetros têm de ser ajustados, agora para ambas as camadas. Nos dois processos em camadas citados, os resultados obtidos são melhores quando comparados aos sistemas que não tratam os diferentes contextos da série.

1.3 Deficiências e necessidade de uma nova abordagem

Sendo a série financeira composta por contextos diferentes, podendo eles referenciar volatilidades distintas ou movimentos em direções opostas e, indo mais adiante, talvez sendo estes contextos os responsáveis pelas não linearidades presentes na série, é razoável considerar o tratamento de contexto como parte do sistema de previsão.

Também, deve-se ter em mente que, treinar e ajustar os parâmetros dos elementos responsáveis por modelar a série e prever seus valores futuros, é um processo não trivial e custoso.

Assim, é desejável que a obtenção de informações de contexto não adicione complexidade em excesso ao sistema. Além do mais, é interessante que as estas sejam, de certa forma, aparentes, para que possam ser analisadas e estudadas.

Considerando estes termos, as propostas existentes deixam lacunas a serem preenchidas. Primeiro, que o uso de SOMs e algoritmos genéticos não é trivial; é um processo custoso que envolve o ajuste de parâmetros livres. Segundo, que as informações de contexto obtidas por estas técnicas passam a ser parte inerente dos dados que serão utilizados como entrada do elemento previsor da segunda camada, tornando difícil a análise de contexto.

Na tentativa de aprimorar as deficiências existentes no processo de tratamento de série

e obtenção de contexto para posterior aplicação em um elemento previsor, este estudo propõe-se a aplicar métodos de clusterização como parte da primeira camada do sistema de previsão, sendo SVMs os elementos previsores da camada seguinte, uma para cada contexto extraído do processo de clusterização.

Este processo, que será descrito em detalhes no Capítulo 4, considera os padrões de entrada a serem aplicados no elemento previsor como sendo n-dimensionais, sendo n o número de informações da série pertencentes ao padrão. Os padrões são separados em *clusters* - o estudo comparou o resultado de dois processos: *K-Means* e *C-Means* -, e uma SVM por *cluster* é treinada apenas com os padrões pertencentes ao seu *cluster*. A pertinência de um novo padrão aos *clusters* é considerada na obtenção de valores futuros.

A obtenção dos *clusters* é um processo mais simples do que trabalhar com SOMs e algoritmos genéticos, além de ser relativamente rápido. Existe apenas um parâmetro livre, que é a quantidade de *clusters*.

Os resultados deste estudo são comparados à previsão feita por uma SVM simples, sem tratamento de contexto, e à previsão feita pelo modelo hierárquico, proposto por Leite (2010), composto por duas camadas: a primeira, para obtenção de contexto construída a partir de uma SOM, e a segunda, composta por uma SVM que é o elemento que modela a série e prevê valores futuros.

1.4 Estrutura da dissertação

A fundamentação teórica necessária para o desenvolvimento deste estudo está apresentada no Capítulo 2. Nele são abordados conceitos de séries temporais, contemplando suas características, informações intrínsecas e tratamentos dados às séries, bem como previsão de séries e medidas de erro. Também inclui elementos da teoria necessária sobre as clusterizações *K-Means* e *C-Means*, bem como sobre SVM e SVM de regressão.

O Capítulo 3 revisa os principais trabalhos existentes sobre previsões de séries financeiras, abordando as técnicas e modelos utilizados bem como os resultados alcançados.

Estão descritos no Capítulo 4 os detalhes sobre o modelo proposto neste estudo, desde o tratamento da informação até a sua construção, detalhando todos os seus componentes, inclusive ferramentas externas. Também descreve a série temporal utilizada - a mesma de Leite (2010) para fins de comparação. Por fim, esse capítulo descreve os resultados obtidos com a aplicação deste modelo e compara os resultados com modelos existentes.

Concluindo a dissertação, o último capítulo sumariza o problema e a proposta para resolvê-lo, bem como faz uma análise qualitativa dos resultados. Apresenta, também, propostas para novos trabalhos.

2 Fundamentação Teórica

2.1 Séries temporais

Definidas por Box, Jenkins e Reinsel (2008) como um conjunto de observações sequenciais de um evento retiradas ao longo do tempo, sendo que as observações podem obedecer a qualquer intervalo regular - horário, diário, semanal, mensal, anual etc (ANDERSON et al., 2016) -, as séries temporais são utilizadas em diversos campos da ciência, como economia, negócios, engenharia e ciências naturais e sociais. Alguns exemplos de séries temporais podem ser a série mensal de produtos fabricados em uma empresa, a série semanal de acidentes automobilísticos em uma rodovia, a sequência horária de observações de um processo químico ou o valor diário de um ativo da bolsa de valores.

Uma característica intrínseca das séries temporais consiste na dependência entre amostras ou observações adjacentes (BOX; JENKINS; REINSEL, 2008). O estudo de séries temporais passa por analisar essa dependência, podendo ser aplicado, dentre outras áreas, à previsão de valores futuros, objeto do presente trabalho.

Um forma de representar uma série temporal é através da plotagem de um gráfico, cujo eixo horizontal contenha informações temporais e o eixo vertical, informações de valor das observações, como exposto na Figura 1.

Anderson et al. (2016) citam que através desse gráfico é possível identificar padrões de comportamento da série, sendo alguns deles citados a seguir.

Horizontal: ocorre quando as observações flutuam entre uma média constante ao longo do tempo.
Tendência: se dá quando as observações se afastam gradualmente da média, durante um longo período de tempo, para cima (tendência de alta) ou para baixo (tendência de baixa).
Sazonal: padrão de comportamento repetido em sucessivos intervalos periódicos de tempo.

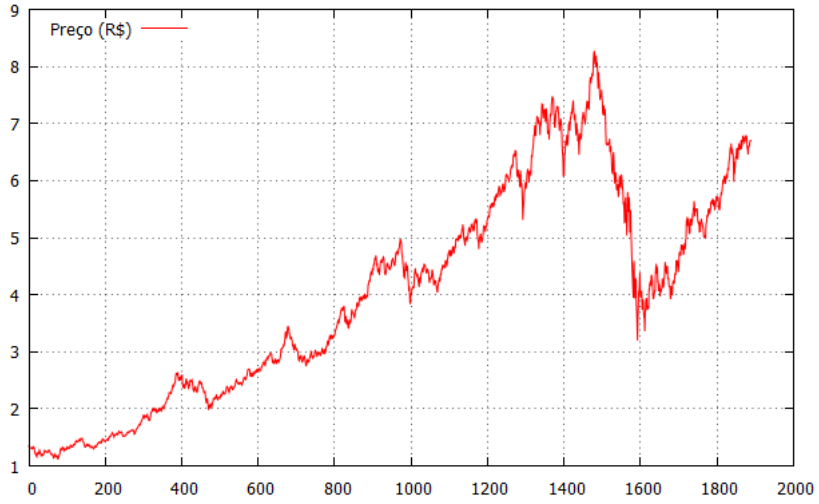

Figura 1: Série temporal de preços

2.1.1 Previsão de séries temporais

Previsão de valores futuros é utilizada nas mais diversas áreas com o propósito de ser uma ferramenta eficiente no auxílio a planejamento (MAKRIDAKIS; WHELLWRIGHT; HYNDMAN, 1998). Pode-se decidir por construir uma usina de energia elétrica, utilizando como um dos insumos a previsão de demanda por energia para os próximos dez anos. Outro exemplo é decidir pela compra de um ativo na bolsa de valores, baseado na previsão de seu valor para as próximas semanas ou meses.

Makridakis, Whellwright e Hyndman (1998) citam ainda que alguns eventos são mais fáceis de prever que outros, sendo a previsibilidade dependente de diversos fatores, os quais merecem destaque:

- grau de conhecimento dos fatores que influenciam o evento;
- quantidade de informação disponível;
- se previsões influenciam os eventos a serem previstos.

Na previsão por demanda de energia elétrica, estes três fatores são conhecidos. Tal demanda é dependente da temperatura e condições econômicas, e existem dados históricos suficientes de demanda e clima. Já na previsão de valor de ativos da bolsa de valores, tem-se apenas uma grande quantidade de informação - histórico de preços do ativo. Porém, fatores que influenciam este evento não são plenamente conhecidos, e previsões têm capacidade de afetar o valor diretamente, visto que pessoas podem ajustar o valor do negócio baseado em previsões.

Box, Jenkins e Reinsel (2008) definem previsão de séries temporais conforme a seguir. Considerando observações disponíveis na forma discreta e separadas por intervalos de tempo equidistantes, define-se como z_t a observação do evento em questão no período atual t, e as observações $z_{t-1}, z_{t-2}, z_{t-3}, ...$ nos períodos passados, sendo $t-1$ (imediatamente anterior a t), $t-2$ (imediatamente anterior a $t-1$) e assim sucessivamente. A previsão de séries temporais envolve a utilização de um ou mais destes períodos para determinação de valores futuros z_{t+l}, sendo $l = 1, 2, 3, ...$, obedecendo a mesma progressão descrita anteriormente.

Sendo, então, a previsão de séries temporais dependente de valores atuais e passados da série, quando propriamente identificados os padrões de comportamento da série neste período conhecido, é possível utilizá-los como guia para previsões futuras (ANDERSON et al., 2016).

Existem várias formas de construir um sistema de previsão de séries temporais, como aplicação de modelos matemáticos, modelos neurais ou de máquinas de aprendizado, sendo este último o método utilizado neste estudo.

2.1.2 Análise de erro e medidas de eficiência

Computar erro de previsão de séries temporais estabelece métricas de qualidade e possibilita a comparação dos diferentes modelos. Indo além, Armstrong e Collopy (1992) explicitam que medidas de erro têm papel importante na calibração e refinamento de previsores de séries temporais.

O erro de previsão é definido como

$$e_t = Y_t - F_t, \tag{2.1}$$

sendo Y_t a observação do evento no tempo t, e F_t a previsão feita de Y_t. O erro percentual é definido conforme a seguir:

$$p_t = \frac{e_t}{Y_t} 100. \tag{2.2}$$

Das definições anteriores, conforme organizado por Gooijer e Hyndman (2006), derivam-se diretamente, dentre outros, os mais básicos métodos de medida de erro, listados a seguir:

$$MSE = mean(e_t^2) = \frac{1}{n}\sum_{i=1}^{n}(Y_t - F_t)^2 \tag{2.3}$$

$$MAE = mean(|e_t|) = \frac{1}{n}\sum_{i=1}^{n}|Y_t - F_t| \tag{2.4}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_t - F_t)^2}, \tag{2.5}$$

sendo MSE (*Mean Squared Error*) o erro quadrático médio, MAE (*Mean Absolute Error*) o erro absoluto médio e RMSE (*Root Mean Square Error*) a raiz do erro quadrático médio.

Destes, um dos mais utilizados no passado, no meio acadêmico, é o RMSE, porém tanto ele quanto os outros citados possuem a desvantagem de não serem livres de escala, ou seja, seus valores são dependentes da magnitude dos dados de entrada. Atualmente, métodos absolutos ganharam a preferência (ARMSTRONG; COLLOPY, 1992). Dentre eles, um dos mais utilizados é o MAPE, cuja definição encontra-se a seguir.

$$MAPE = mean(|p_t|) = \frac{1}{n}\sum_{i=1}^{n}|\frac{e_t}{Y_t}100|, \tag{2.6}$$

sendo MAPE (*Mean Absolute Percentage Error*) o erro absoluto percentual médio.

Armstrong e Collopy (1992) citam duas desvantagens deste erro. A primeira é que ele é sensitivo a valores de observação iguais a zero - o que levaria a divisão por zero -, e a segunda é que a penalização para previsões acima do atual podem ser mais severas que as previsões abaixo do atual, visto que, para baixo, o MAPE é limitado em 100% - para séries que não suportem valores negativos - e para cima não há limite.

Neste estudo, a medida de erro utilizada será o MAPE. Vale ressaltar que a série (de retorno) utilizada não contempla valores zero e não é limitada a apenas valores positivos, sendo que as desvantagens citadas anteriormente não influenciarão o resultado. E também, este método foi o utilizado no estudo de Leite (2010), sendo de interesse utilizar os mesmos critérios de medida para fins de comparação.

Outra medida de eficiência adotada por Leite (2010) e também utilizada neste estudo é o cálculo do desvio padrão dos erros de previsão, que tem por objetivo medir a dispersão dos erros, sendo desejável valores menores, indicando a baixa incidência de valores de erro elevados.

2.1.3 Séries temporais financeiras

O estudo de séries temporais financeiras consiste em avaliar o valor de um ativo através do tempo. Descrito por Tsay (2005), o estudo de tais séries é considerado empírico, mas também envolve os conceitos de teoria financeira, que fornecem subsídios essenciais para a inferência de resultados.

Torna-o mais desafiador o fato de elementos de incerteza estarem sempre presentes, tanto na teoria financeira quanto na série temporal. A volatilidade, que mede o grau de variação de um ativo, é um exemplo de elemento de incerteza, não podendo ser diretamente observada em uma série de retorno. Assim, métodos estatísticos tornaram-se essenciais na análise de séries temporais financeiras.

As seções a seguir descrevem os conceitos e propriedades das séries temporais financeiras.

2.1.3.1 Série de retorno

Ao contrário das demais séries, em séries temporais financeiras é comum trabalhar com série de retorno ao invés da série de preços do ativo. Campbell, Lo e MacKinlay (2016) citam dois fatores favoráveis a esta abordagem. Primeiro, que a série de retorno é livre de escala, o que previne o perfil do investimento ser alterado pelo preço do ativo, e segundo, que ela contém propriedades estatísticas atrativas, como estacionariedade.

Existem diversas definições de série de retorno. As principais estão listadas a seguir.

Considere P_t o preço de um ativo no período t.

Retorno simples de um período O retorno simples computa a relação entre dois períodos subsequentes, conforme a seguir:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}. \tag{2.7}$$

Este retorno deriva do retorno simples bruto, calculado como a seguir:

$$1 + R_t = \frac{P_t}{P_{t-1}}. \tag{2.8}$$

Retorno composto contínuo Também chamado de *log return*, é definido com o logaritmo natural do retorno simples bruto, conforme a seguir:

$$r_t = ln(1 + R_t) = ln(\frac{P_t}{P_{t-1}}). \tag{2.9}$$

Este tipo de retorno possui propriedades estatísticas mais tratáveis, comparado a outras formas (CAMPBELL; LO; MACKINLAY, 2016), sendo o escolhido para aplicação neste estudo.

A Figura 1, apresentada anteriormente, mostra uma série de preços de um ativo, que contém características que impedem seu uso na análise de séries temporais, como tendências e escala.

Já a Figura 2 apresenta a série de retorno composto contínuo do mesmo ativo, resultado da aplicação direta da Equação 2.9 na série de preços, onde pode-se notar a ausência de tais características.

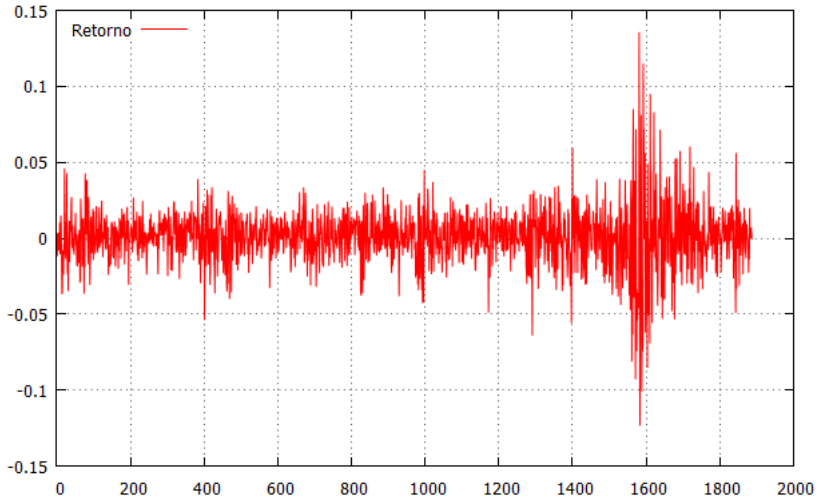

Figura 2: Série de retorno composto contínuo

2.1.3.2 Estacionariedade

É caracterizada como uma importante propriedade de uma série temporal e considerada a base de análise de séries temporais (TSAY, 2005). Uma série temporal é classificada como estritamente estacionária se a distribuição conjunta de um intervalo da série é invariante no tempo. Ou seja, uma série de retorno $\{r_t\}$ é estritamente estacionária se a distribuição conjunta de $(r_{t_1}, ..., r_{t_k})$ é idêntica a distribuição conjunta de $(r_{t_1+k}, ..., r_{t_k+k})$ para todo t, sendo k um número inteiro positivo arbitrário e $(t_1, ..., t_k)$ uma coleção de k

inteiros positivos.

Sendo esta uma propriedade dita forte e de difícil verificação empírica, um relaxamento é assumido de tal forma que ainda seja possível a aplicação de métodos de predição de valores futuros (TSAY, 2005). Assim, uma série é considerada fracamente estacionária se as seguintes condições forem satisfeitas:

- Média de $\{r_t\}$ invariante no tempo$E(r_t) = \mu$, sendo μ uma constante;
- Covariância de $\{r_t\}$ invariante no tempo$Cov(r_t, r_{t-l}) = \gamma_l$, sendo esta dependente apenas de l e tendo duas importantes propriedades: $(a)\gamma_0 = Var(r_t)$ e $(b)\gamma_{-l} = \gamma_l$.

Tsay (2005) cita que é comum assumir que a série de retorno de um ativo é fracamente estacionária, sendo possível provar empiricamente, através de dados históricos suficientes da série, divididos em amostras, e posterior verificação da consistência dos resultados obtidos através dessas amostras.

2.1.3.3 Função de autocorrelação

O coeficiente de correlação $\rho_{X,Y}$ entre duas variáveis aleatórias X e Y mede a dependência linear entre X e Y, sendo que $-1 < \rho_{X,Y} < 1$. X e Y são ditas descorrelacionadas se $\rho_{X,Y} = 0$, significando que variações em uma não afetam o valor da outra. Valores positivos maiores indicam maior relação direta entre as variáveis - variações positivas em uma provocam variações positivas na outra -, e negativos indicam relação inversa - variações positivas em uma provocam variações negativas na outra. A correlação é calculada conforme a seguir:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{E(X-\mu_x)^2 E(Y-\mu_y)^2}}. \tag{2.10}$$

Dada uma amostra $\{(x_t, y_t)\}_{t=1}^{T}$, é possível estimar a correlação através de

$$\hat{\rho}_{X,Y} = \frac{\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \overline{x})^2 \sum_{t=1}^{T}(y_t - \overline{y})^2}}, \tag{2.11}$$

sendo $\overline{x}$ e $\overline{y}$ as médias das amostras de X e Y, respectivamente.

Quando, considerando a série de retorno r_t, deseja-se obter a correlação entre r_t e seus valores passados r_{t-l}, generaliza-se o conceito de correlação para autocorrelação, chamado também de autocorrelação de *lag-l* de r_t e denotado por ρ_l. Sendo r_t fracamente estacionária e fazendo valer sua propriedade $Var(r_t) = Var(r_{t-l})$, pode-se definir a autocorrelação como

$$\rho_l = \frac{Cov(r_t, r_{t-l})}{\sqrt{Var(r_t)Var(r_{t-l})}} = \frac{Cov(r_t, r_{t-l})}{Var(r_t)} = \frac{\gamma_l}{\gamma_0}. \tag{2.12}$$

Novamente, dada uma amostra $\{(r_t)\}_{t=1}^{T}$, é possível estimar a autocorrelação da amostra *lag-l* através de

$$\hat{\rho}_l = \frac{\sum_{t=l+1}^{T}(r_t - \overline{r})(r_{t-l} - \overline{r})}{\sum_{t=1}^{T}(r_t - \overline{r})^2}. \tag{2.13}$$

2.1.3.4 Volatilidade

O preço de um ativo pode apresentar, ao longo do tempo, variações moderadas, ou seja, dentro de uma faixa esperada de valores, e também variações bruscas, quando os valores saltam quantidades maiores que o considerado normal para a série. A dimensão da variação dos preços é chamada de volatilidade, sendo a baixa volatilidade composta por variações moderadas e a alta volatilidade, por variações bruscas.

O desvio padrão de uma série temporal mede dispersão histórica dos valores em relação à média, podendo ser usado como referência para medida de volatilidade da série (YANG; CHAN; KING, 2002).

Conforme descrito por Leite (2010), o desvio padrão da série como um todo, chamado desvio padrão total, deve ser comparado ao desvio padrão dos dias imediatamente anteriores ao dia analisado, sendo ambos calculados conforme a seguir:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{2.14}$$

Sendo o desvio padrão instantâneo maior que o desvio padrão total, o dia analisado é considerado como de alta volatilidade. Caso contrário, é considerado de baixa volatilidade. O mesmo estudo utiliza $n = 20$ para cálculo do desvio padrão instantâneo.

Ainda segundo Leite (2010), a informação de volatilidade é um conhecimento extra sobre a série temporal, podendo ser utilizado no momento da previsão. A Figura 3 ilustra a comparação do desvio padrão instantâneo com o desvio padrão total. Os pontos em que o primeiro ultrapassa o último são pontos de alta volatilidade. Os demais, de baixa volatilidade.

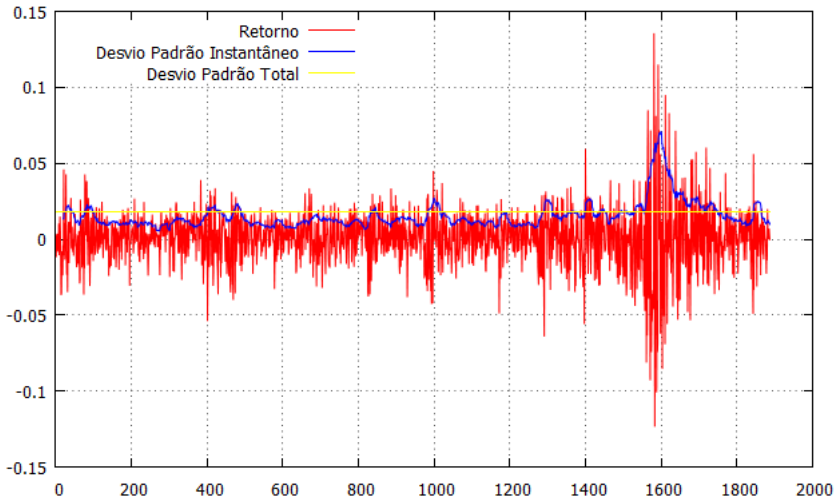

Figura 3: Desvio padrão aplicado à volatilidade

2.2 Clusterização

Agrupar e classificar objetos é um comportamento inerente aos seres vivos, que são capazes de perceber características intrínsecas de objetos e agrupá-los naturalmente para os mais diversos fins. Como exemplo, é possível perceber se animais são ferozes ou dóceis, ou se frutos são comestíveis ou venenosos (EVERITT et al., 2011).

Classificar também é parte importante da ciência, e sua utilização pode ser percebida há muitos séculos. Everitt et al. (2011) citam, como exemplo, a divisão de animais em dois grupos principais: um daqueles que possuem sangue vermelho e outro, dos que não possuem. Citam ainda a subdivisão desses grupos de acordo com a forma de nascimento de filhotes, podendo ser ovíparos, vivíparos etc.

Oliveira e Pedrycz (2007) introduzem o conceito de que um grupo desordenado de objetos, por vezes, possui estruturas que permitem a separação destes em classes ou subgrupos. Tais estruturas nem sempre são explícitas ou visíveis, sendo necessária a utilização de técnicas e algoritmos a fim de determiná-las.

Oliveira e Pedrycz (2007) dizem, ainda, que clusterização é um processo de aprendi-

zado não supervisionado, cujo propósito é decompor um conjunto de objetos em subgrupos ou *clusters*, baseado na similaridade entre eles. Em um processo ideal, os *clusters* representam de forma fiel as diferentes categorias de objetos existentes, sendo então a clusterização uma forma de reconstruir estruturas previamente desconhecidas, inerentes ao grupo de objetos. Everitt et al. (2011) ainda citam que clusterização é, essencialmente, descobrir grupos em um conjunto de dados.

O método de separação de objetos na clusterização procura agrupar, em um mesmo *cluster*, objetos que possuam maior similaridade possível, de forma que objetos pertencentes a *clusters* distintos possuam menor similaridade.

O processo de clusterização utiliza a função de distância - ou medida de proximidade -, que no fundo representa a medida de dissimilaridade, que é equivalente à medida de similaridade, para representar o grau de separação entre dois objetos (OLIVEIRA; PEDRYCZ, 2007). A comparação entre as separações dos diferentes objetos diz o quanto eles são similares entre si.

Everitt et al. (2011) atentam ao fato de que é possível haver um grupo homogêneo de objetos que não contém divisão natural em categorias. Neste caso, separar os objetos em grupos passa a ser referido como dissecação, e também é um processo válido e útil em diversas circunstâncias. Portanto, durante a clusterização, é importante ter cautela para que uma estrutura natural não seja imposta a um determinado grupo de objetos, pois nem sempre ela existirá.

Por exemplo, os dados representados na Figura 4

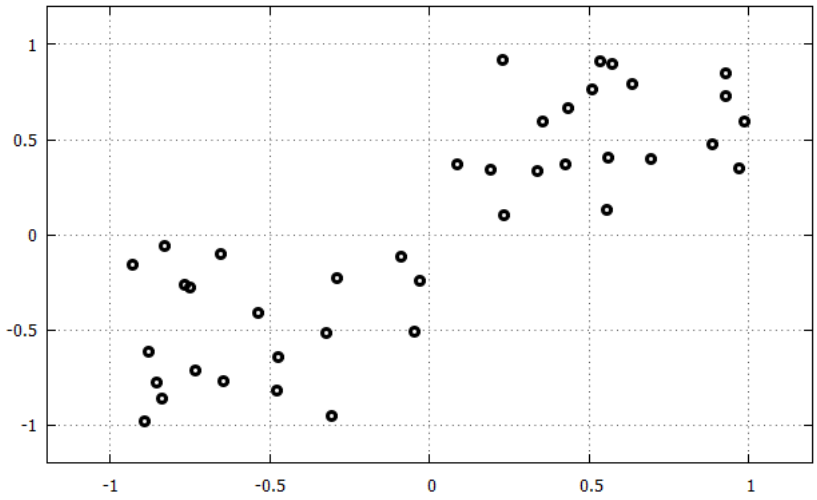

Figura 4: Dados aleatórios espalhados no espaço

podem ser agrupados conforme a Figura 5, sendo um *cluster* com os dados de coordenadas

positivas (vermelho) e outro de coordenadas negativas (verde).

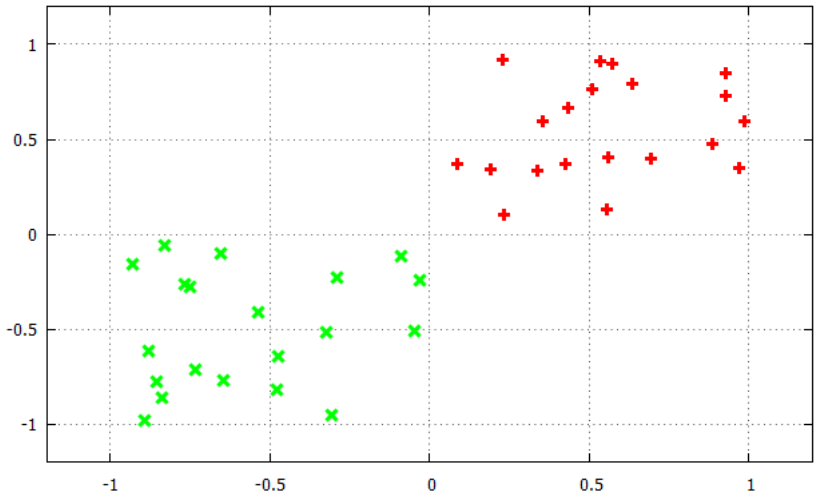

Figura 5: Dados separados em dois *clusters*

Oliveira e Pedrycz (2007) citam dois tipos distintos de métodos de clusterização:

Hierárquica Os objetos são divididos em diversos números de *clusters*, de forma que essas diferentes divisões formem uma estrutura hierárquica como uma árvore. A decisão sobre qual divisão utilizar é tomada a partir da análise de qual delas melhor representa o problema a ser resolvido.

Particionada ou não-hierárquica Os objetos são divididos em um número predefinido de *clusters*, baseado na medida de dissimilaridade entre eles, sendo tal número um parâmetro de entrada do processo. Everitt et al. (2011) classificam este método como de otimização, cujo propósito é minimizar ou maximizar um critério numérico. Este critério, também chamado de função objetivo, quantifica a qualidade da clusterização de um grupo de dados, externando propriedades que permitam tal visualização. A função objetivo é, por vezes, objeto de estudo com o propósito de refinar a medida de qualidade para problemas específicos (OLIVEIRA; PEDRYCZ, 2007).

A representação dos *clusters*, abordada por Oliveira e Pedrycz (2007), é feita por meio de protótipos $C_i, i = 1, ..., c$, sendo c a quantidade de *clusters*. Protótipos são enuplas de parâmetros contendo um centro c_i e, possivelmente, outros atributos como tamanho e forma do *cluster*.

De acordo com Everitt et al. (2011), a clusterização é aplicada para os mais diversos fins, entre eles:

- Pesquisa de mercado
- Astronomia
- Psiquiatria
- Classificações meteorológicas
- Arqueologia
- Bioinformática e genética

2.2.1 Medida de proximidade

Classificar o relacionamento entre dois objetos, considerando seus atributos, é papel fundamental no processo de clusterização. Dois objetos estão fortemente relacionados se seus valores de atributos são iguais, e igualmente afastados se os valores são diferentes. A medida de proximidade é a representação deste relacionamento entre dois objetos.

Everitt et al. (2011) classificam medida de proximidade entre dois grupos em proximidades diretas e proximidades indiretas, sendo a primeira extraída a partir da medida de similaridade - ou dissimilaridade - entre dois objetos em uma matriz n x n, e a segunda, mais aplicada em diversas áreas, derivada da matriz n x p de objetos e atributos, onde a distância entre dois objetos é calculada a partir dos valores de seus atributos, a fim de se formar uma matriz de dissimilaridade n x n.

Os valores dos atributos de um objeto podem ser categóricos, contínuos ou ambos, sendo que existem medidas de similaridade específicas para cada caso.

O caso mais simples para objetos, cujos valores de atributos são categóricos, ocorre quando estes são binários. Dados dois indivíduos i e j, é possível calcular o coeficiente de similaridade s_{ij}, que terá valor unitário se todas as variáveis dos indivíduos forem iguais, e valor zero se todas diferirem o máximo entre si.

Para o cálculo da similaridade, utiliza-se a contagem da comparação dos p atributos dos indivíduos, conforme Tabela 1,

		Indivíduo i	
	Valor	1	0
Indivíduo j	1	a	b
	0	c	d

Tabela 1: Contagem da comparação dos p atributos de dois indivíduos

sendo a o número de atributos presentes em dois indivíduos i e j, d o número de atributos ausentes nos dois indivíduos, e c e d o número de atributos presentes em um e não no outro. Ainda, $p = a + b + c + d$.

Everitt et al. (2011) descrevem que, tendo como origem estas contagens, diversas medidas de similaridade foram propostas, sendo algumas delas descritas na Tabela 2. Citam, ainda, que grande parte da razão de diferentes medidas existirem se dá pelas diferentes abordagens à contagem d, visto que o fato de dois indivíduos não possuírem certa característica não necessariamente significa que eles são semelhantes com relação a este aspecto. Por exemplo, quando os dois indivíduos não são do sexo masculino, podem ser considerados como semelhantes em relação ao aspecto sexo. Já dois que não possuem olhos castanhos, não podem ser considerados semelhantes em relação ao aspecto cor dos olhos.

Medida	**Fórmula**
S1: Matching coefficient	$s_{ij} = \frac{(a+d)}{(a+b+c+d)}$
S2: Jaccard coefficient	$s_{ij} = \frac{a}{(a+b+c)}$
S3: Rogers and Tanimoto	$s_{ij} = \frac{(a+d)}{(a+2(b+c)+d)}$
S4: Sneath and Sokal	$s_{ij} = \frac{a}{[a+2(b+c)]}$

Tabela 2: Medidas de similaridade para atributos categóricos binários

Já quando os atributos dos objetos são contínuos, Everitt et al. (2011) citam que a medida de proximidade utilizada normalmente é a medida de dissimilaridade, representada por δ_{ij}. Tal medida pode ser chamada de medida de distância quando

$$\delta_{ij} + \delta_{im} \geq \delta_{jm} \tag{2.15}$$

para todos indivíduos i, j e m. Sendo uma medida de distância, é então representada por d_{ij}, e as medidas de distâncias entre todos os pares de objetos do conjunto dão origem à matriz de distâncias $\mathbf{D}$ de tamanho n x n.

Dentre as diversas medidas de distância citadas por Everitt et al. (2011), destacam-se as citadas na Tabela 3. A distância D1 (Minkowski) é uma generalização de diversas medidas de distância, como D2 (Distância Euclidiana), que possui $r = 2$, e cuja medida pode ser interpretada como a distância física entre dois objetos em um espaço de dimensão p, e D3 (City Block), que possui $r = 1$ e cuja medida se refere ao somatório da decomposição da distância Euclidiana nos p eixos do espaço ao qual os objetos pertencem.

Medida	Fórmula
D1: Distância Minkowski	$d_{ij} = \left[\sum_{k=1}^{p} \lvert x_{ik} - x_{jk}\rvert^r\right]^{\frac{1}{r}} (r \geq 1)$
D2: Distância Euclidiana	$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}}$
D3: Distância City Block	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk}\rvert$

Tabela 3: Medidas de distância

2.2.2 Função objetivo

O conceito de qualidade de uma partição de *clusters* passa por verificar se características do agrupamento estão dentro de limites estabelecidos, ou se são comparativamente melhores que outros agrupamentos. Everitt et al. (2011) citam homogeneidade e separação como dois aspectos a serem empregados para este fim.

A medida de homogeneidade diz o quanto a estrutura de um grupo é coesa, ou seja, o quanto objetos pertencentes ao grupo estão relacionados entre si, considerando suas características, e a separação informa o grau de isolamento entre os grupos. É desejável que partições produzam grupos com estruturas coesas e com alto grau de isolamento uns dos outros (EVERITT et al., 2011).

A seguir, temos a descrição de duas formas, abordadas por Everitt et al. (2011), de utilizar como critério de otimização, também chamado de função objetivo, as medidas citadas anteriormente. Em ambas, é desejável minimizar a falta de homogeneidade dos grupos de uma partição, ou maximizar a separação entre eles, ou ainda um híbrido destes dois processos.

Critério de otimização a partir da matriz de dissimilaridade Considerando a matriz de dissimilaridade Δ contendo elementos δ_{ij}, cujos valores representam a dissimilaridade entre os objetos i e j, Everitt et al. (2011) citam três, dentre os diversos critérios já abordados, para quantificação da falta de homogeneidade de um grupo m.

$$h_1(m) = \sum_{l=1}^{n_m} \sum_{v=1, v\neq l}^{n_m} (\delta_{ml,mv})^r \tag{2.16}$$

$$h_2(m) = \max_{\substack{l=1,\ldots,n_m \\ v=1,\ldots,n_m \\ v\neq l}} [(\delta_{ml,mv})^r] \tag{2.17}$$

$$h_3(m) = \min_{v=1,\ldots,n_m} \left[\sum_{l=1}^{n_m} (\delta_{ml,mv})^r \right], \tag{2.18}$$

sendo $r \in \{1, 2\}$.

A Equação 2.16 é o somatório das dissimilaridades (quadráticas) entre todos os pares de elementos do grupo m. Já a Equação 2.17 apresenta apenas o maior valor de dissimilaridade entre dois elementos do grupo m. Por último, a Equação 2.18 mede a menor soma total da dissimilaridade (quadrática) entre um elemento e todos os outros do mesmo grupo m.

Para quantificar a separação entre grupos, Everitt et al. (2011) citam duas medidas, similares às medidas 2.16 e 2.17:

$$i_1(m) = \sum_{l=1}^{n_m} \sum_{k \neq m} \sum_{v=1}^{n_m} (\delta_{ml,kv})^r \tag{2.19}$$

$$i_2(m) = \min_{\substack{l=1,\ldots,n_m \\ k \neq m \\ v=1,\ldots,n_m}} \left[(\delta_{ml,kv})^r \right], \tag{2.20}$$

sendo $r \in \{1, 2\}$.

Assim como nas medidas de falta de homogeneidade, a Equação 2.19 retorna o somatório (quadrático) da separação entre os elementos de um grupo e todos os elementos dos outros grupos. Já a Equação 2.20 mede o menor valor de separação entre dois elementos existentes na equação anterior.

Seja qual for a medida escolhida, o critério de otimização passa por minimizar ou maximizar (dependendo da medida escolhida) a agregação da medida para todos os *clusters*. Considerando um agrupamento C de n elementos em c grupos, Everitt et al. (2011) citam que os critérios de otimização poderiam ser:

$$C_1(n, c) = \sum_{m=1}^{c} h(m) \tag{2.21}$$

$$C_2(n, c) = \max_{m=1,\ldots,c} [h(m)] \tag{2.22}$$

$$C_3(n, g) = \min_{m=1,\ldots,c} [h(m)], \tag{2.23}$$

quando levado em consideração a quantificação da falta de homogeneidade, sendo que 2.21 demonstra a falta de homogeneidade total do agrupamento, e 2.22 e 2.23 o pior e o melhor caso, respectivamente. Neste caso, o critério seria minimizar $C(n, c)$, visto que é desejada uma partição com maior homogeneidade possível.

Substituindo $h(m)$ por $i(m)$ nas equações citadas anteriormente, elas passariam a refletir o conceito de separação, sendo que 2.21 demonstraria a separação total do agrupamento, e 2.22 e 2.23 o melhor e o pior caso de separação, respectivamente. Para este cenário, o critério seria maximizar $C(n, c)$, buscando maior separação entre os grupos.

A fim de diminuir a ordem de grandeza do critério C_1, quando $h_1(m)$ é utilizado, Everitt et al. (2011) sugerem a utilização de

$$C_1^*(n, c) = \sum_{m=1}^{c} \frac{h_1(m)}{n_m}$$

em substituição à Equação 2.21.

Critério de otimização a partir dos atributos dos objetos Utilizando a matriz de objetos X de tamanho n x p, onde n é o número de objetos e p, o número de atributos de cada objeto, pode-se escrever a matriz de dispersão total T de dimensão p x p, conforme a seguir:

$$T = \sum_{m=1}^{c} \sum_{l=1}^{n_m} (x_{ml} - \overline{x})(x_{ml} - \overline{x})', \tag{2.24}$$

sendo x_{ml} o vetor de dimensão p de atributos do objeto l pertencente ao grupo m, e $\overline{x}$, o vetor de dimensão p que contém a média dos atributos de todos os objetos do conjunto. Everitt et al. (2011) citam que a matriz de dispersão total pode ser decomposta como a soma entre a matriz de dispersão intragrupo W e a matriz de dispersão intergrupo B:

$$T = W + B, \tag{2.25}$$

sendo tais matrizes definidas como

$$W = \sum_{m=1}^{c} \sum_{l=1}^{n_m} (x_{ml} - \overline{x}_m)(x_{ml} - \overline{x}_m)' \tag{2.26}$$

$$B = \sum_{m=1}^{c} n_m (\overline{x}_m - \overline{x})(\overline{x}_m - \overline{x})', \tag{2.27}$$

onde $\overline{x}_m$ é o vetor de dimensão p, que contém a média dos atributos dos objetos do grupo m.

Everitt et al. (2011) citam alguns critérios de otimização que podem ser utilizados a partir das matrizes 2.26 e 2.27, sendo destacado o de minimização do traço de W, ou seja, a soma dos elementos da diagonal principal de W, que pode ser calculada conforme a seguir:

$$E = \sum_{m=1}^{c} \sum_{l=1}^{n_m} (x_{ml} - \overline{x}_m)'(x_{ml} - \overline{x}_m). \tag{2.28}$$

Tal critério tem por objetivo minimizar a diferença total dos elementos de um grupo em relação à média dos elementos do grupo, e é equivalente a maximizar o traço de B, cujo objetivo é maximizar a diferença das médias dos grupos em relação à média de todos os objetos do conjunto. Minimizar o traço de W é equivalente a minimizar a soma da distâncias Euclidianas quadráticas entre os elementos de um *cluster* e seu centro, conforme a seguir:

$$E = \sum_{m=1}^{c} \sum_{l=1}^{n_m} d^2_{ml,m}. \tag{2.29}$$

Finalizando, pode-se definir função objetivo como sendo um critério matemático que quantifica a qualidade de uma clusterização, levando em conta os *clusters* e seus objetos, e servindo como função que deve ser minimizada, a fim de se obter soluções de clusterização ótimas (OLIVEIRA; PEDRYCZ, 2007).

2.2.3 K-Means

Descrito inicialmente por MacQueen (1967), o algoritmo de clusterização *K-Means* consiste em separar os n objetos pertencentes a um conjunto em k grupos. Os passos de sua execução clássica estão a seguir:

1. K grupos iniciais são formados contendo um único ponto, escolhido aleatoriamente do conjunto de objetos;

2. cada ponto do conjunto é adicionado ao grupo, cuja distância da média de seus pontos (centroide) até si seja a menor;

3. após a adição de um novo ponto a um grupo, os centroides são recalculados, considerando a nova pertinência do ponto.

Os passos 2 e 3 são repetidos até que nenhuma mudança nas pertinências ou nos centroides ocorram, sendo que cada objeto deve pertencer a exatamente um grupo (OLIVEIRA; PEDRYCZ, 2007). Assim, os mesmos citam que é definida uma matriz binária U de tamanho k x n, que representa a pertinência dos objetos aos *clusters*, cujos elementos são descritos individualmente como

$$u_{ij} \in \{0, 1\}, \tag{2.30}$$

sendo que $u_{ij} = 1$ indica que o objeto j pertence ao *cluster* i, e que $u_{ij} = 0$ indica que não pertence. Os centroides dos *clusters* são calculados conforme a seguir:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij} x_j}{\sum_{j=1}^{n} u_{ij}}. \tag{2.31}$$

Algoritmos que atribuem um objeto exclusivamente a um *cluster* também são classificados como sendo *hard clustering*, e satisfazem os três critérios a seguir, conforme descrito por Bezdek, Ehrlich e Full (1984). Dado um conjunto de objetos X clusterizados em k *clusters*, sendo eles C_1, C_2,..., C_k, com $2 \leq k \leq p$, sendo p o número de atributos dos objetos do conjunto, temos que:

$$C_i \neq \emptyset; \qquad 1 \leq i \leq k \tag{2.32}$$

$$C_i \cap C_j = \emptyset; \qquad i \neq j \tag{2.33}$$

$$\bigcup_{i=1}^{k} C_i = X. \tag{2.34}$$

O primeiro critério não permite que *clusters* sejam vazios. O segundo trata da pertinência de um objeto a um único *cluster*, condizente com a equação 2.30, e o terceiro mostra que a união de todos os *clusters* forma o conjunto original de objetos.

Conforme mostrado por Everitt et al. (2011), diversas variações do algoritmo original já foram apresentadas, sendo então chamados de *K-Means* os algoritmos que determinam a pertinência de objetos a um grupo tendo como base a distância do objeto até a média do grupo, e sendo esta recalculada a cada mudança. Ainda, buscar agrupamentos minimizando a distância dos objetos ao centro do *cluster*, quando a distância Euclidiana é utilizada, é equivalente a minimizar o traço de W, conforme equação 2.29, podendo a função objetivo para este algoritmo ser reescrita conforme a seguir:

$$\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij} d_{ij}^2. \tag{2.35}$$

Uma clusterização *K-Means* é dita ótima quando a soma da distâncias Euclidianas quadráticas entre os centroides dos *clusters* e seus objetos é mínima (OLIVEIRA; PEDRYCZ, 2007).

Everitt et al. (2011) também expõem a sensibilidade dos algoritmos *K-Means* aos grupos iniciais, citando que partições iniciais diferentes levam a agrupamentos com diferentes ótimos locais, e que diversas alternativas já foram desenvolvidas buscando determinar grupos iniciais que levassem a melhores agrupamentos. Oliveira e Pedrycz (2007) citam que é necessário executar o processo de clusterização por várias vezes, a fim de evitar ótimos locais.

Citado também por Everitt et al. (2011), os algoritmos *K-Means* encaixam-se na classificação de algoritmos *hill-climbing*, cujo comportamento busca encontrar o melhor agrupamento modificando a pertinência dos objetos nos *clusters*, e mantendo a nova configuração apenas se o critério de otimização for melhor que o da configuração anterior.

2.2.4 Fuzzy C-Means

A clusterização *Fuzzy* C-Means provê um relaxamento à regra descrita pela equação 2.30, permitindo que objetos tenham pertinência parcial a *clusters*. Tal relaxamento se dá ao se aceitar quaisquer valores entre os limites de não pertinência (0) e pertinência total (1), sendo então u_{ij} chamado de grau de pertinência e descrito conforme a seguir, de acordo com Oliveira e Pedrycz (2007):

$$u_{ij} \in [0, 1]. \tag{2.36}$$

Desta forma, o objeto j pode pertencer parcialmente ao *cluster* i, sendo o valor dado por u_{ij}. Oliveira e Pedrycz (2007) também citam que cada objeto tem associado a si um vetor, que indica seu grau de participação em cada um dos c *clusters* do agrupamento:

$$u_j = (u_{1j}, ..., u_{cj})^T. \tag{2.37}$$

É possível, também, definir a matriz de pertinência *fuzzy*, que engloba todos os objetos e todos o *clusters*, conforme a seguir:

$$U = (u_{ij}) = (u_1, ..., u_n). \tag{2.38}$$

O grau de pertinência descrito é proveniente do conceito de conjuntos *fuzzy*, descrito por Zadeh (1965).

Para que uma partição *fuzzy* seja válida, as seguintes restrições devem ser respeitadas:

$$\sum_{j=1}^{n} u_{ij} > 0, \qquad \forall i \in 1, ..., c \tag{2.39}$$

$$\sum_{j=1}^{n} u_{ij} = 1, \qquad \forall j \in 1, ..., n, \tag{2.40}$$

sendo que primeira restringe *clusters* vazios, condizente com a equação 2.32 de *hard clustering*, e a seguinte, que diz que a soma das pertinências de um elemento aos *clusters* é igual a unidade, garante que os pesos das pertinências são iguais e que todos os elementos são incluídos na partição de forma igualitária, e ainda, que esta restrição corresponde a uma normalização da pertinência dos dados (OLIVEIRA; PEDRYCZ, 2007).

O centro dos *clusters* de uma partição *Fuzzy C-Means* é calculado considerando todos os elementos do conjunto e o grau de pertinência de cada um ao *cluster* em questão, conforme a seguir:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \tag{2.41}$$

sendo m o fator *fuzificador*, que determina o grau de relaxamento dos limites dos *clusters*, ao passo que $m = 1$ leva à associação de objetos a *clusters* da mesma forma que o *hard clustering*, e $m > 1$ leva à associação relaxada, discutida anteriormente. Quanto maior o valor de m, mais relaxamento é aplicado às associações, sendo $m = 2$ um valor utilizado de forma geral (OLIVEIRA; PEDRYCZ, 2007), não existindo evidências que distinguam um valor ótimo de m (BEZDEK; EHRLICH; FULL, 1984).

O grau de pertinência dos objetos aos *clusters* considera a distância do objeto a todos os *clusters* da partição, sendo seu cálculo descrito a seguir:

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{ij}^2}{d_{il}^2}\right)^{\frac{1}{m-1}}}. \tag{2.42}$$

Diversas propostas já foram formuladas como critério de otimização de partições *fuzzy*, sendo a função a seguir uma das mais populares (BEZDEK; EHRLICH; FULL, 1984):

$$\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2. \tag{2.43}$$

A partir destas definições, Bezdek, Ehrlich e Full (1984) descrevem o algoritmo básico *Fuzzy C-Means*, conforme a seguir.

1. A quantidade de *clusters* c e o fator fuzificador m são escolhidos. Uma matriz de pertinência inicial $U^{(0)}$ é também escolhida (respeitando as restrições 2.39 e 2.40).
2. Os valores dos centroides $c_i^{(k)}$ são calculados, utilizando os valores de $U^{(k)}$, através da equação 2.41.
3. A matriz de pertinência $U^{(k+1)}$ é atualizada utilizando a equação 2.42, fazendo uso dos centroides calculados no passo 2.
4. A matriz de pertinência $U^{(k+1)}$ é comparada à matriz anterior $U^{(k)}$, e caso $||U^{(k+1)} - U^{(k)}|| < \in$, a iteração acaba. Caso contrário, volta-se ao passo 2.

Oliveira e Pedrycz (2007) citam que é possível escolher valores iniciais de centroides ao invés de uma matriz de pertinência inicial. E também, que pode ser utilizado como critério de parada um limite máximo τ de iterações, ou a comparação entre o centroide da iteração corrente e o da iteração anterior, sendo este critério atingido se a mudança entre os valores for menor que algum critério $\in$.

Duas vantagens citadas por Everitt et al. (2011), da clusterização *fuzzy* sobre as clusterizações do tipo *hard clustering*, são que a pertinência pode ser combinada com outras informações, e que se pode comparar os dois "melhores" *clusters* para um objeto, a fim de saber se ambos são igualmente bons, informação esta escondida em clusterizações *hard*.

2.3 Support Vector Machines

Support Vector Machines (Máquinas de Vetor de Suporte) pertencem à categoria de rede alimentada adiante, sendo uma máquina de aprendizado binária com propriedades interessantes, como capacidade de prover soluções ótimas. De forma simples, para um problema de classificação de padrões, pode ser definida como uma máquina capaz de construir um hiperplano de separação entre exemplos positivos e negativos, de forma que a separação entre eles seja maximizada (HAYKIN, 2009).

É capaz de resolver problemas lineares e não lineares de separação de padrões, como também problemas de regressão não linear, como será visto posteriormente.

2.3.1 Classificações de padrões linearmente separáveis

O propósito da separação linear de padrões, dado um conjunto de treinamento $\{(x_i, d_i)_{i=1}^n\}$, onde x_i é o padrão de entrada do exemplo i, e d_i é a resposta desejada para este padrão, é construir uma superfície de separação representada pela equação

$$w^{\mathrm{T}}x + b = 0, \tag{2.44}$$

onde x é um vetor de entrada, w, um vetor de peso ajustável perpendicular à superfície de separação, e b, um *bias*, de tal sorte que, para um dado vetor de entrada desconhecido u qualquer,

$$\begin{aligned} w^{\mathrm{T}}u + b \geq 0 \text{ então } &+1 \\ w^{\mathrm{T}}u + b < 0 \text{ então } &-1. \end{aligned} \tag{2.45}$$

A separação entre tal hiperplano e o exemplo mais próximo deve ser a máxima possível. Nestas condições, tem-se um hiperplano ótimo, representado por

$$w_o{}^{\mathrm{T}} x + b_o = 0. \tag{2.46}$$

O objetivo, então, é encontrar os valores de w_o e b_o para o hiperplano ótimo, dado um conjunto de treinamento, que satisfaçam as restrições a seguir:

$$\begin{aligned} w_o{}^{\mathrm{T}} x_i + b_o \geq +1 \text{ para } d_i = +1 \\ w_o{}^{\mathrm{T}} x_i + b_o \leq -1 \text{ para } d_i = -1, \end{aligned} \tag{2.47}$$

e que condicionam os padrões com valores desejados opostos em lados distintos do hiperplano de separação, espaçando-os de uma margem. As observações (x_i, d_i) que satisfaçam o sinal de igualdade de uma das equações de 2.47 são chamadas de vetores de suporte.

A Figura 6 mostra um exemplo de um hiperplano ótimo, em um espaço bidimensional.

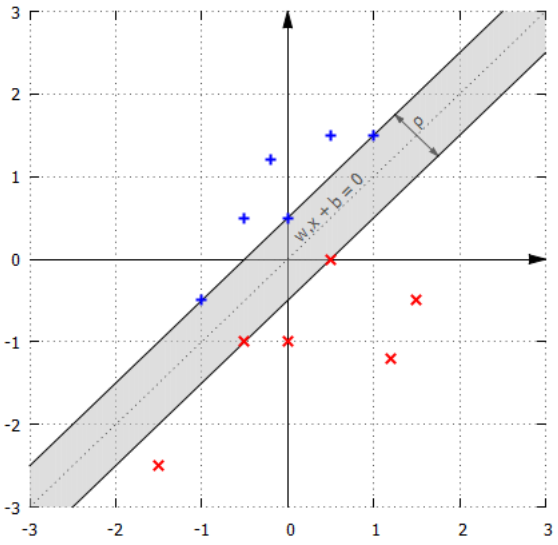

Figura 6: Exemplo de hiperplano de separação ótimo

Além de satisfazer as restrições mencionadas, deve-se também maximizar a margem de separação ρ entre as observações positivas e negativas, calculada pela diferença de dois vetores de suporte, um no lado positivo (x_+) e outro no lado negativo (x_-), projetada sobre o vetor unitário normal ao hiperplano, conforme a seguir:

$$\rho = (x_+ - x_-)^{\mathrm{T}} \left(\frac{w}{||w||} \right). \tag{2.48}$$

Aplicando as igualdades de 2.47 em $w^{\mathrm{T}} x_+$ e $w^{\mathrm{T}} x_-$ tem-se que

$$\rho = \frac{(1-b)-(-1-b)}{||w||} = \frac{2}{||w||}, \tag{2.49}$$

o que mostra que maximizar a margem de separação é equivalente a minimizar a magnitude do vetor w, que por sua vez, por conveniência matemática, é equivalente a minimizar $\frac{1}{2}w^{\mathrm{T}}w$.

Tem-se então um problema de otimização com restrição, que pode ser resolvido através dos multiplicadores de *Lagrange*, como descrito em Vapnik (1998).

Para tanto, constroi-se a função lagrangiana, levando-se em consideração a função a ser minimizada e as restrições. Neste ponto, pode-se simplificar as equações de 2.47 em uma única $d_i(w^{\mathrm{T}}x_i + b) \geq 1$.

$$L = \frac{1}{2}w^{\mathrm{T}}w - \sum_{i=1}^{N} \alpha_i[d_i(w^{\mathrm{T}}x_i + b) - 1], \tag{2.50}$$

sendo α_i os multiplicadores de *Lagrange*. Esta função deve ser minimizada em relação a w e b e maximizada em relação a $\alpha_i \geq 0$. Obtém-se as duas condições ótimas diferenciando a função lagrangiana em relação a w e b e igualando-as a zero, conforme a seguir:

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i d_i x_i \tag{2.51}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i d_i = 0. \tag{2.52}$$

A equação 2.51 mostra que w é dependente exclusivamente dos vetores de entrada x. Além do mais, para todas as restrições que não satisfaçam as igualdades de 2.47, o multiplicador α_i deve ser zero.

Substituindo as condições ótimas na função lagrangiana, tem-se agora a função a ser maximizada em relação a α_i:

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j d_i d_j {x_i}^{\mathrm{T}} x_j. \tag{2.53}$$

Encontrando os valores ótimos para os multiplicadores de *Lagrange* $\alpha_{o,i}$, é possível determinar w_o através de 2.51, e b_o através de um vetor de suporte positivo ou negativo

com uma das igualdades de 2.47.

É possível, então, reescrever as regras de decisão de 2.45 usando a equação 2.51 como a seguir:

$$\begin{aligned} &\sum_{i=1}^{N} \alpha_i d_i {x_i}^{\mathrm{T}} u + b \geq 0 \text{ então } +1 \\ &\sum_{i=1}^{N} \alpha_i d_i {x_i}^{\mathrm{T}} u + b < 0 \text{ então } -1. \end{aligned} \tag{2.54}$$

É pertinente notar que tanto a função a ser maximizada (Equação 2.53) quanto as regras de decisão (Equação 2.54) são dependentes apenas do produto escalar de dois vetores de entrada.

2.3.2 Classificações de padrões linearmente inseparáveis

Para os padrões que não são linearmente separáveis, a Máquina de Vetor de Suporte lança mão de um conceito sofisticado e simples para resolvê-los.

A Figura 7 mostra um padrão que não pode ser separado por um hiperplano em (x, y).

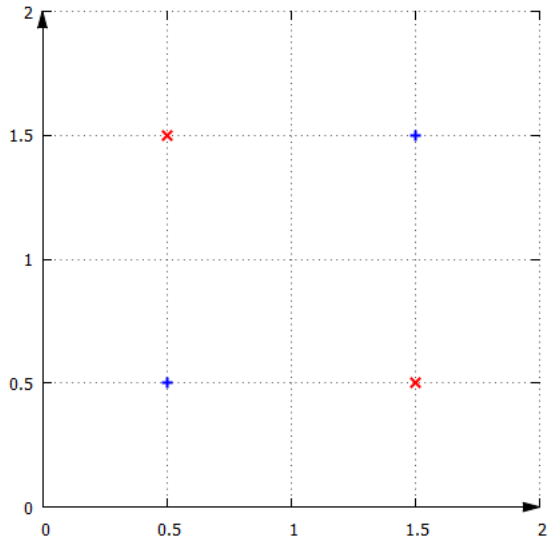

Figura 7: Padrão não separável linearmente

Porém, se os dados forem movidos para um hiperplano de alta dimensionalidade, mais conveniente para a solução do problema, tal hiperplano de separação passa a existir. A Figura 8 mostra os mesmos padrões movidos para um hiperplano de dimensionalidade

maior, contendo agora um hiperplano de separação em (x, y, z).

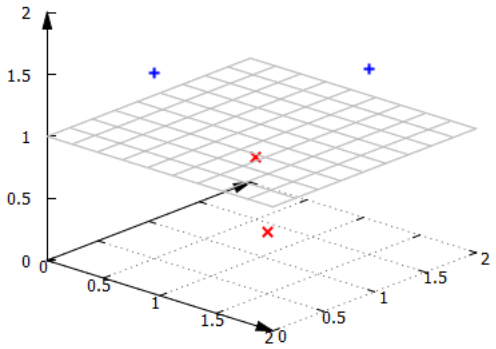

Figura 8: Hiperplano em alta dimensionalidade

Para tanto, é necessário transformar os padrões da dimensionalidade atual para um espaço de dimensionalidade maior. Esta transformação é representada por uma função não linear $\phi(x)$. Conforme visto anteriormente, tanto a função a ser maximizada (Equação 2.53) quanto as regras de decisão (Equação 2.54) são dependentes apenas do produto escalar de dois vetores de entrada. Para problemas não separáveis linearmente, ambas passam a ser dependentes do produto escalar dos vetores transformados, conforme a seguir:

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \phi^{\mathrm{T}}(x_i)\phi(x_j) \tag{2.55}$$

$$\begin{aligned} &\sum_{i=1}^{N} \alpha_i d_i \phi^{\mathrm{T}}(x_i)\phi(u) + b \geq 0 \text{ então } +1 \\ &\sum_{i=1}^{N} \alpha_i d_i \phi^{\mathrm{T}}(x_i)\phi(u) + b < 0 \text{ então } -1. \end{aligned} \tag{2.56}$$

Havendo então uma função

$$K(x_i, x_j) = \phi^{\mathrm{T}}(x_i)\phi(x_j) \tag{2.57}$$

que retorne o resultado do produto escalar de dois vetores transformados, não há necessidade de se saber a transformação em si. Tal função é chamada de *kernel*. A Tabela 4

apresenta alguns dos mais comuns *kernels* existentes.

Tipo	***kernel***
Função polinomial	$(x^{\mathrm{T}}x_i + 1)^p$
Função de base radial	$exp(-\frac{1}{2\sigma^2}\|\|x - x_i\|\|^2)$

Tabela 4: Exemplos de *kernel*

Vapnik (1999) cita que qualquer função que satisfaça o teorema de Mercer (MERCER, 1909) pode ser utilizada como função de *kernel*.

2.3.3 Regressão

De acordo com Haykin (2009), problemas de regressão não linear podem ser resolvidos minimizando o erro absoluto. Assim, a função de perda tem a seguinte definição

$$L(d, y) = |d - y|, \tag{2.58}$$

sendo d a resposta desejada e y a resposta produzida.

Uma extensão a esta função de perda, proposta por Vapnik (1998) e chamada função de perda insensível a ε, é utilizada para Máquinas de Vetor de Suporte para regressão, conforme a seguir:

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon, & \text{para } |d - y| \geq \varepsilon \\ 0, & \text{caso contrário} \end{cases}. \tag{2.59}$$

Esta função computa o erro apenas se este for superior a um limiar ε. A Figura 9 ilustra uma SVM de regressão com um tubo insensível a ε, e a Figura 10 ilustra a função de perda insensível a ε.

Portanto, para regressão, o objetivo é minimizar a função de risco

$$R(w, x, b) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} L_\varepsilon(d, y), \tag{2.60}$$

sujeita as seguintes restrições

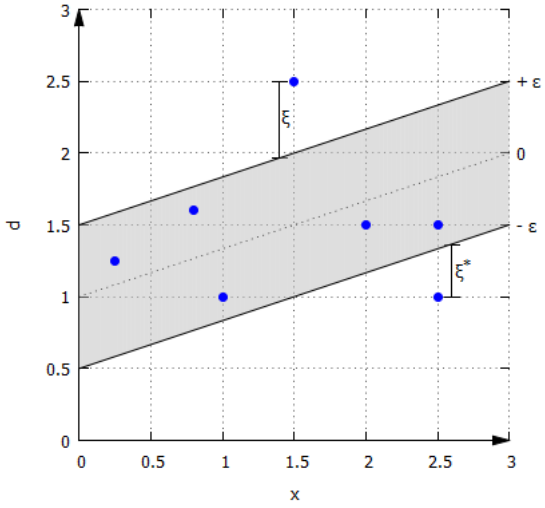

Figura 9: SVM de regressão - tubo insensível a ε

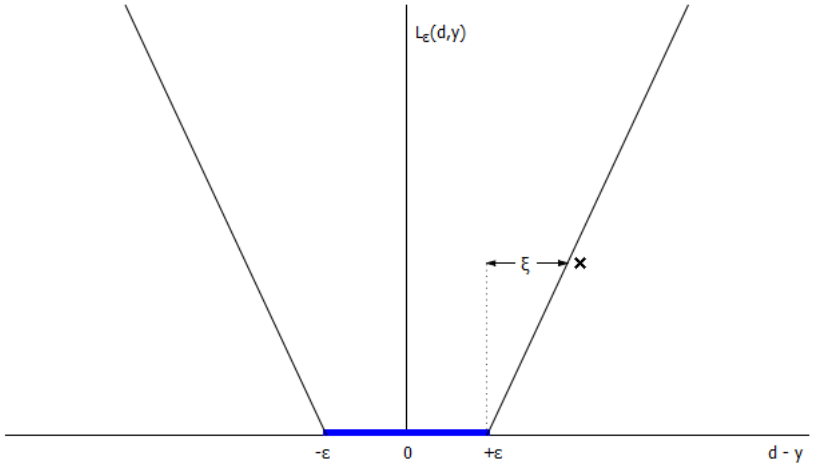

Figura 10: Função de perda insensível a ε

$$\begin{aligned} d_i - y_i &\leq \varepsilon + \xi_i \\ y_i - d_i &\leq \varepsilon + \xi_i^* \\ \xi_i &\geq 0 \\ \xi_i^* &\geq 0, \end{aligned} \tag{2.61}$$

sendo ξ_i e ξ_i^* variáveis soltas, que descrevem a função de erro insensível a ε da equação 2.59.

O primeiro termo da equação 2.60 é o termo de regularização e o segundo, o risco empírico, sendo a constante C, chamada de constante de regularização, a relação entre os

dois termos (CAO; TAY, 2001b). Valores maiores de C aumentam a importância do risco empírico em relação à regularização.

Utilizando novamente o método de *Lagrange* para otimização com restrição, obtém-se a função de decisão para regressão como (SMOLA; SCHOLKOPF, 2003):

$$f(x) = \sum_{i=1}^{n} (a_i - a_i^*) K(x, x_i) + b, \tag{2.62}$$

sendo a_i e a_i^* os multiplicadores de *Lagrange* que satisfazem as seguintes restrições:

$$\begin{aligned} &\sum_{i=1}^{n} (a_i - a_i^*) = 0 \\ &0 \leq a_i, a_i^* \leq C, \text{ sendo } i = 1, 2, ..., n. \end{aligned} \tag{2.63}$$

A quantidade de vetores de suporte utilizados na regressão, e consequentemente a suavidade da função de regressão, estão diretamente relacionadas ao valor de ε. Quanto menor a precisão requerida, menor o número de vetores de suporte necessários, atingindo-se isso com valores maiores de ε (SMOLA; SCHOLKOPF, 1998).

Ambos os parâmetros aqui citados, C e ε, são definidos pelo usuário de acordo com sua necessidade.

3 Trabalhos Existentes

A escolha e construção de um modelo, que trate de forma satisfatória uma série temporal com o propósito de prever valores futuros, é um processo não trivial e custoso. Modelos não lineares são amplamente utilizados quando a série é econômica ou financeira, seguindo, obviamente, a natureza não linear de tais séries. Porém, a escolha do modelo correto levanta diversas questões a serem consideradas. Perguntas sobre quais dados históricos da série considerar, qual modelo não linear utilizar, como estimar os parâmetros livres do modelo, como evitar ótimos locais, e diversos outros, são desafios a serem vencidos durante este processo.

Clements, Franses e Swanson (2004) tratam destas dificuldades, endereçando possíveis soluções para alguns dos problemas, consideradas o estado da arte corrente. O estudo termina concluindo que sistemas baseados em modelos não lineares ainda não tratam o problema de previsão de forma simples e confiável, mas que o futuro é promissor, citando que técnicas já existentes, quando aplicadas e testadas de forma cuidadosa, podem levar a avanços significativos no entendimento do problema.

As SVMs de regressão são amplamente utilizadas para construção de modelos não lineares a partir de séries temporais, com o propósito de prever valores futuros baseados no modelo construído. Os parágrafos seguintes descrevem alguns dos principais estudos nesta área, mostrando aplicações de sucesso deste modelo.

Cao e Tay (2001a) propuseram a aplicação de SVM em previsão de séries temporais financeiras, comparando o resultado com a aplicação de uma rede neural MLP. Tais técnicas foram empregadas em cinco contratos futuros, extraídos do *Chicago Mercantile Market*, utilizando como amostras o preço de fechamento do contrato em horizontes de aproximadamente três anos e meio, sendo a previsão aplicada em uma janela de cinco dias. A comparação entre elas foi feita utilizando os critérios NMSE, MAE, DS e WDS. Outro objetivo do estudo foi analisar o impacto dos parâmetros livres da SVM no resultado da previsão. O estudo apresenta que os resultados obtidos com a SVM são superiores em

qualidade à MLP, concluindo ser promissora a aplicação desta técnica para o fim proposto.

Um estudo mais detalhado dos efeitos dos diferentes valores dos parâmetros livres da SVM no resultado da previsão de séries financeiras foi abordado em Cao e Tay (2003), sendo então proposto um modelo com parâmetros adaptativos. Utilizando uma SVM com *kernel* de função Gaussiana, o estudo mostrou que os parâmetros C e δ^2 influenciam diretamente no resultado da previsão, mas que este é insensitivo ao parâmetro ϵ, que influencia o número de vetores de suporte. O modelo proposto - aplicado em cinco contratos futuros extraídos do *Chicago Mercantile Market* - utiliza parâmetros adaptativos em C e ϵ, sendo que estes dão mais peso para amostras mais próximas do período a se prever. Para todas as séries testadas, o modelo com parâmetros adaptativos mostrou-se mais preciso que o modelo utilizando uma SVM comum, bem como utilizando redes neurais dos tipos *back propagation* e função de base radial.

Também aprofundando o estudo nas características da SVM para regressão aplicada em previsão de séries temporais, Yang, Chan e King (2002) analisaram o efeito da margem da SVM de regressão no resultado da previsão, citando que margens pequenas levam ao *overfitting* e margens de valor elevado podem produzir excesso de generalização. Propuseram então uma margem adaptativa, calculada a partir do desvio padrão da série, cujo propósito é escolher o valor de margem utilizado a partir da volatilidade do ativo. Assim, margens pequenas podem ser utilizadas em períodos calmos - onde o ruído é menor -, e margens maiores, nos períodos de alta volatilidade. Utilizando o preço de fechamento do índice HSI da bolsa de Hong Kong, os autores mostraram que o uso de margem adaptativa aumenta consideravelmente a precisão de previsão comparado a sistemas com margem fixa.

Ainda com a atenção voltada à margem, Yang et al. (2004) estenderam o conceito de margem adaptativa para o tratamento de *outliers*. Utilizando uma função de perda insensitiva a ϵ, separada em margens superior e inferior a fim de prover características variáveis e assimétricas, os *outliers* são detectados quando a distância entre eles e a margem ultrapassa um dado limiar. Uma vez encontrados, a margem é adaptada a fim de diminuir seus efeitos sobre a SVM. Com dados de três fontes financeiras diferentes - índices NASDAQ, HSI e FTSE 100 -, o estudo mostra que diminuir os efeitos dos *outliers* no modelo de previsão construído a partir da SVM aumenta a precisão do sistema, quando comparado ao modelo que apenas considera a margem adaptativa.

O estudo apresentado por Huang, Nakamori e Wang (2005) diferencia-se dos demais por buscar a previsão da direção do movimento de séries temporais financeiras, ao invés

de prever os valores futuros propriamente, citando que operações com previsões precisas de movimentos são mais lucrativas que as feitas com previsões de preço que possuem um certo grau de erro. O estudo utilizou o valor semanal do índice japonês NIKKEI 255, associado a variáveis exógenas que possuem forte correlação com este, sendo estas o índice americano S&P500 e a taxa de câmbio entre dólar americano e iene japonês. Comparando os resultados com a análise discriminante linear e quadrática, e também com o modelo neural de Elman, o estudo mostra que os resultados com a SVM levam a taxa de acerto da direção do movimento a 73%, sendo superior a todos os outros modelos utilizados na comparação.

A teoria de conjuntos aproximados, cujo nome é oriundo do temo em inglês *Rough Sets* (RS), foi utilizada em Podsiadlo e Rybinski (2016) também com o propósito de prever a tendência do movimento de séries financeiras. A partir de um conjunto extenso de atributos de entrada, desde informações sobre a própria série, como preços de abertura e fechamento, até indicadores técnicos, o modelo foi construído utilizando RS para selecionar os atributos mais relevantes para sua construção e também como classificador para previsão da direção do movimento da série. O modelo foi comparado com uma SVM para classificação de padrões e com operações simuladas utilizando a técnica *Buy & Hold* (comprar um papel e não vendê-lo por um tempo longo). Os experimentos foram executados sobre três índices do mercado de ações, sendo eles S&P500, DAX e HSI. Em um dos índices, a precisão do modelo com RS foi similar à da SVM. Nos outros dois, sobressaiu-se. O mesmo aconteceu quando comparado com a técnica *Buy & Hold*.

Zhang, Zhang e Feng (2016) propõe o uso da série de Taylor em conjunto com o método ARIMA para previsão de séries financeiras, visando capturar componentes lineares e não lineares da série. O desempenho deste modelo foi comparado ao de um sistema híbrido, composto por uma SVM e pelo método ARIMA, sendo os experimentos executados em cima das séries de preços de diversas *commodities* negociadas no mercado americano, através da *Chicago Board of Trade*. Os resultados mostraram que o modelo utilizando a serie de Taylor apresentou resultados mais precisos que o modelo com SVM.

Su e Cheng (2016) utilizaram o sistema de inferência ANFIS para construir um modelo da séries temporais financeiras, tendo proposto um método para seleção de atributos relevantes, denominado INFS, responsável por construir o conjunto de dados de entrada do sistema ANFIS. As previsões feitas pelo sistema ANFIS são refinadas por um modelo de expectativas adaptativas. Os índices de mercado de ações TAIEX e HSI foram utilizados nos experimentos com o modelo proposto, que foi comparado com diversos outros modelos,

entre eles um composto por INFS + SVM e outro, por INFS + ANFIS, ambos não fazendo uso do modelo de expectativas adaptativas. Em todos os experimentos, o modelo proposto apresentou melhores resultados do que os modelos existentes.

As pesquisas listadas a seguir tratam da aplicação de técnicas que visam separar ou dar tratamento distinto a conhecimentos específicos da série. Todos têm em comum o fato de o sistema ser feito em estágios. O estágio inicial visa sempre tratar a série de forma a extrair conhecimentos inerentes e não visíveis. O estágio seguinte é a construção do modelo não linear utilizando sempre a informação tratada, vinda do estágio anterior. Como será visto, a aplicação de tais processos eleva consideravelmente a qualidade da previsão do segundo estágio.

Cao e Tay (2001b) apresentaram o uso de SVM *experts* para previsão de séries temporais financeiras. O modelo consiste em uma arquitetura de dois estágios, sendo o primeiro composto por SOMs, capazes de dividir a informação de entrada em regiões que contenham amostras similares, e o segundo contendo SVMs *experts*, conectadas a cada região. Utilizando séries financeiras de duas fontes distintas - séries provenientes da *Santa Fe Time Series Prediction Analysis Competition* e contratos futuros provenientes do *Chicago Mercantile Market* -, o estudo mostrou que o modelo proposto apresentou resultados superiores comparados à utilização de uma única SVM.

Já em Cao (2003), o uso de SVM *experts* foi extendido para séries temporais que não financeiras, sendo o modelo o mesmo utilizado em Cao e Tay (2001b). Com séries temporais de eventos variados - séries sobre manchas solares, séries provenientes da *Santa Fe Competition* e séries extraídas do concurso *The Great Energy Predictor Shootout - the First Data Analysis and Prediction* -, o estudo novamente demonstrou a superioridade do modelo de dois estágios sobre a SVM única, consolidando o conceito de separação do conhecimento como ferramenta de especialização de previsores SVM.

Seguindo o conceito de *experts*, Armano, Marchesi e Murru (2005) utilizaram algoritmos genéticos como responsáveis pela partição dos dados de entrada e redes neurais MLPs especializadas para previsão. Sendo a série considerada um processo multi estacionário, o propósito é idendificar e trabalhar com modelos locais ao invés de um único modelo global. Aproximadamente 9 anos de informações dos índices COMIT e S&P500 foram utilizados nos testes. Para comparação, os autores verificaram o resultado de operações simuladas utilizando a técnica *Buy & Hold* (comprar um papel e não vendê-lo por um tempo longo), e também o sistema proposto. Para conjuntos de dados grandes, o sistema proposto mostrou-se mais eficiente, trazendo resultados mais lucrativos para o investidor.

Melo (2003) utilizou a técnica de Composição de Especialistas Locais (CEL) para previsão de séries temporais. Tal técnica consiste em agrupar as informações de entrada em *clusters* disjuntos e utilizar várias técnicas de modelagem em cada *cluster*, de forma a construir modelos concorrentes em cada um. Para cada *cluster*, o modelo que apresenta o melhor resultado é escolhido como sendo o Modelo Especialista Local. As saídas dos Modelos Especialistas Locais são combinadas na formação do resultado final, ponderando-os pela distância dos novos dados ao centros dos *clusters*. Neste estudo, a clusterização ficou a cago de uma rede neural de *Kohonen*, e os modelos especialistas foram construídos através de redes neurais artificiais (RNA), análises de regressão múltiplas (ARM) e cópia carbono (CRB). Esta técnica foi aplicada em experimentos com duas séries temporais. A primeira foi extraída da *Santa Fe Time Series Prediction Analysis Competition*, e a segunda relativa às séries de preços diária e mensal do açucar, negociado na Câmara de Comércio de Nova York. A precisão de previsão da técnica CEL foi comparada com a precisão dos chamados especialistas globais. Estes especialistas são modelos criados com todos os dados da série, sem qualquer clusterização, utilizando isoladamente os métodos para modelagem aplicados no CEL. O modelo proposto CEL mostrou-se mais preciso que os especialistas globais.

Carpinteiro et al. (2012) propuseram um modelo hierárquico composto por uma SOM seguida de uma SVM, e compararam o resultado obtido com outros dois sistemas: uma rede neural MLP e uma SVM pura. Os dados utilizados pertencem a um fundo de ações do Banco do Brasil, sendo os mesmos deste estudo. Citaram, entre outras coisas, que o modelo hierárquico é capaz de reter grandes informações do passado pelo fato de segmentar a série em contextos menores, para posterior análise e combinação. Os resultados mostraram que o modelo hierárquico tem precisão ligeiramente superior à SVM pura, porém muito melhor que a MLP.

Panapakidis e Dagoumas (2016) trabalharam na previsão do preço de energia elétrica um dia a frente para o mercado italiano. Diversos modelos baseados em redes neurais artificiais (ANN) foram construídos, sendo um deles um modelo que clusteriza a série temporal e depois treina uma ANN por *cluster*. Este modelo mostrou-se superior aos modelos simples, compostos por apenas uma ANN. Outros modelos que utilizam não só a série temporal como entrada, mas também variáveis exógenas, mostraram-se ainda mais precisos.

Além dos trabalhos aqui citados, Sapankevych e Sankar (2009) produziram um estudo mostrando o estado corrente da aplicação de SVMs em previsão de séries temporais,

dizendo que este ramo de pesquisa é objeto de análise de diversos estudos nas mais variadas áreas de aplicação. De acordo este estudo, as pesquisas se concentram em duas áreas principais: séries financeiras e séries relacionadas a eletricidade. Porém, esta técnica também é aplicada em sistemas de controle, previsões do tempo e de caráter ambiental, entre outros. O estudo aborda sessenta e seis publicações, sendo que quase um terço corresponde à previsão de séries financeiras, explicitando a importância dada a este ramo no universo de pesquisa em séries temporais. Cita ainda também que o aspecto não linear do problema da previsão é o fator comum citado entre as publicações que torna o uso de SVMs o mais adequado para este fim, dizendo que esta técnica modela processos não lineares de forma mais satisfatória.

4 Metodologia

4.1 Modelo proposto

O modelo previsor desenvolvido neste estudo pode ser separado em duas partes. A primeira contempla os componentes responsáveis pelo treinamento e construção do sistema de previsão, a partir de uma série temporal e parametrização do usuário, e a segunda, o sistema de previsão em si.

4.1.1 Treinamento e construção do sistema de previsão

Esta seção do modelo, cujos componentes estão apresentados na Figura 11, é responsável pelo tratamento dos dados históricos da série (componentes em cinza), clusterização das informações de entrada (componentes em azul) e construção/treinamento de SVMs para previsão (componentes em verde). A seguir, a descrição de cada um.

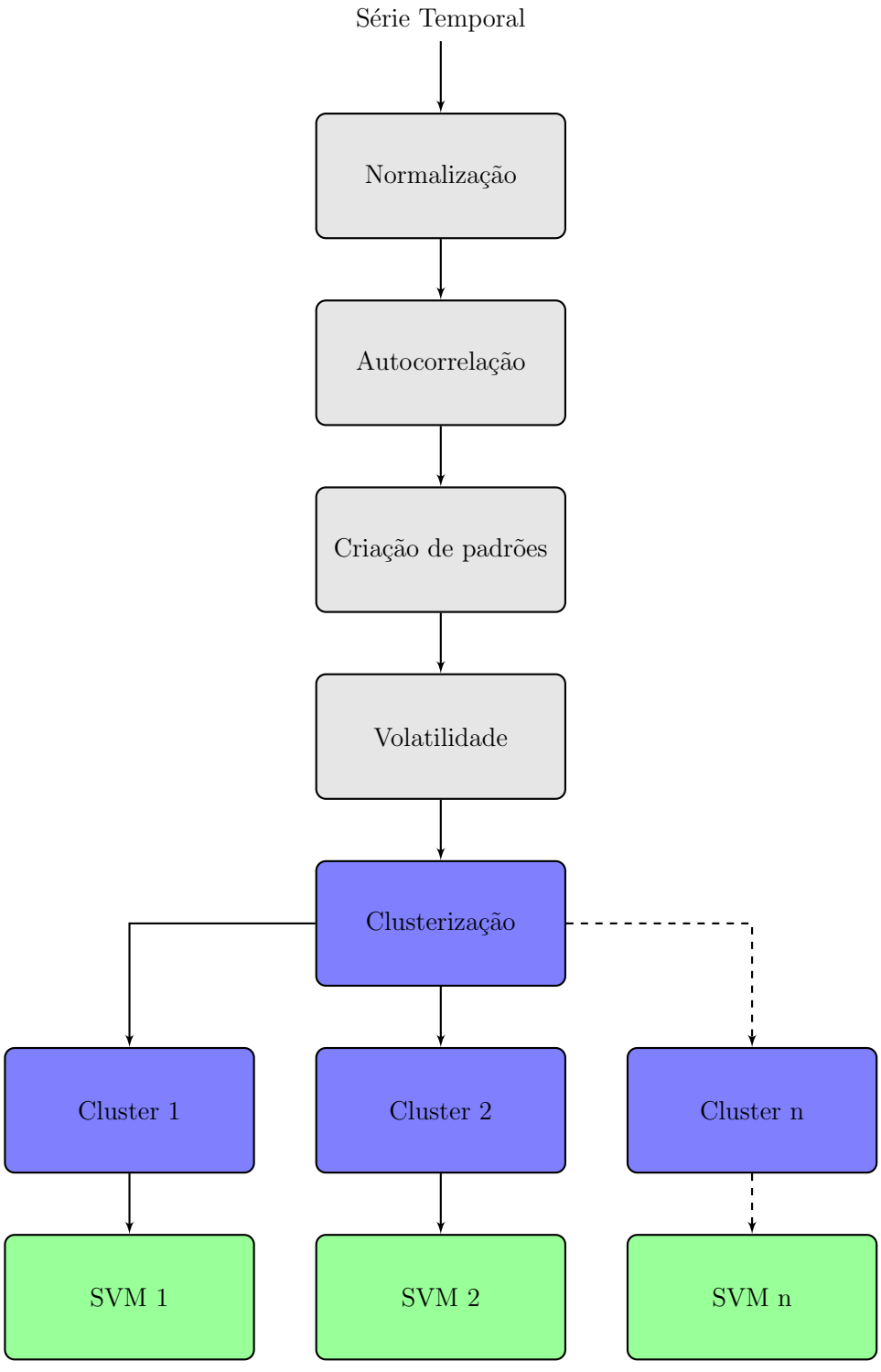

Figura 11: Passos para treinamento do modelo proposto

Normalização Recebe, como entrada, a série temporal pura e a transforma em uma série de retorno, com o propósito de remover escala e informações de tendência, e torná-la mais tratável para o uso em previsores, conforme descrito na Seção 2.1.3.1.

Autocorrelação A partir da série de retorno recebida, extrai as informações sobre quais períodos do passado são mais relevantes para previsão um passo a frente, utilizando o processo de autocorrelação e considerando significativos apenas valores superiores a 5%. Este componente fornece, então, uma sequência de *lags* significativamente relacionados a valores um passo a frente. Exemplificando, dizer que *lag*-1 e *lag*-3 são siginificativamente relacionados a um passo a frente é equivalente a ontem e três dias atrás estarem relacionados a hoje. Este processo está descrito na Seção 2.1.3.3.

Criação de padrões Tem por propósito criar uma tabela de decisão associando l atributos de condição - sendo estes os l valores passados, significativamente relevantes a valores um passo a frente, extraídos do processo de autocorrelação -, e um atributo de decisão - que é o valor um passo a frente. Por exemplo, sendo x o valor de *lag* mais distante de um passo a frente, uma tabela de decisão com três *lags* ($l = 3$), relativa a uma série de retorno contendo n observações, deverá conter $n - x$ linhas e $l + 1$ colunas, tendo cada linha o aspecto definido na Tabela 5. O valor *lag*-0 refere-se ao valor um passo a frente.

Atributos de condição			**Atributos de decisão**
lag-x	*lag*-b	*lag*-a	*lag*-0

Tabela 5: Exemplo de entrada de uma tabela de decisão

Volatilidade Componente responsável por adicionar informações de volatilidade à tabela de decisão recebida. Para tanto, uma coluna é inserida na seção de atributos de condição, passando então a tabela a ter o aspecto definido na Tabela 6. O valor de volatilidade v é calculado a partir da comparação do desvio padrão da série com o desvio padrão instantâneo, conforme descrito em 2.1.3.4, sendo o último sempre em relação ao atributo de decisão *lag*-0. Os valores possíveis são $vol = +1$, para períodos de alta volatilidade, e $vol = 0$, para períodos de baixa volatilidade. O propósito desta informação é inserir uma separação entre amostras com volatilidades diferentes, a fim de que o sistema previsor consiga distingui-las e tratá-las de maneiras diferentes. Quanto menor a separação entre os valores atribuídos às duas volatilidades possíveis, menor será a dissimilaridade entre padrões com volatilidades distintas, tendo o valor mencionado anteriormente se mostrado suficiente para a distinção de tais padrões.

Atributos de condição				**Atributos de decisão**
lag-x	*lag*-b	*lag*-a	*vol*	*lag*-0

Tabela 6: Exemplo de entrada de uma tabela de decisão com volatilidade

Clusterização Utilizando uma métrica de distância, realiza a clusterização da série temporal na forma dos atributos de condição da Tabela 6. Para tanto, considera os valores destes atributos como coordenadas de um ponto em R^{l+1}. Tem por finalidade agrupar padrões semelhantes em *clusters*, seccionando a Tabela 6 em c partes para uma clusterização com c *clusters* em *hard clustering*, ou aplicando pertinências parciais dos padrões desta tabela nos c *clusters* em *soft clustering*, estando o processo de clusterização descrito na Seção 2.2. Aliar os padrões da série com informações de volatilidade permite que a clusterização identifique e separe padrões com volatilidades distintas em *clusters* distintos, criando assim *clusters* conhecedores de determinados padrões para determinada volatilidade. O número de *clusters* deste processo é de escolha do usuário.

Previsores SVM Para cada *cluster* encontrado no passo anterior, uma SVM de regressão especialista é treinada para regressão de seus padrões. A ideia é ter SVMs de regressão com conhecimentos específicos de padrões semelhantes, de modo que padrões descorrelacionados não influenciem no conhecimento adquirido da SVM de regressão. Para atingir tal objetivo, os atributos da Tabela 6 de cada *cluster* são apresentados a cada SVM de regressão, sendo que os parâmetros C e ε são entradas de usuário. Assim, cada SVM será treinada com $l+1$ dados de entrada associados a um dado de saída, sendo a quantidade de dados de treinamento dependente do resultado do processo de clusterização. O conceito de SVM de regressão está descrito na Seção 2.3.3.

Os passos de Normalização até Clusterização referem-se à padronização, tratamento e mineração dos dados de entrada, a fim de torná-los mais adequados e precisos para a construção das SVMs previsoras, de forma a evitar que conhecimentos espúrios prejudiquem a assertividade das SVMs.

4.1.2 Sistema de previsão

O sistema de previsão nada mais é que aplicar dados desconhecidos, ou seja, que não fizeram parte do programa de treinamento, e computar a resposta. O novos dados devem obedecer aos mesmos critérios de tratamento e padronização aplicados aos dados históricos no treinamento. Dependendo da forma como a resposta final é construída, pode-se ter um previsor *hard* ou *soft*, sendo a resposta do primeiro proveniente apenas de uma SVM (como será visto, a SVM pertencente ao *cluster* mais pertinente ao novo dado), e a do segundo, por uma composição das respostas de cada SVM com pesos diferentes.

4.1.2.1 Sistema de previsão hard

Tendo como princípio que cada SVM é especializada em padrões previamente estabelecidos como semelhantes, computa como resultado final apenas o resultado da SVM cujos padrões mais se assemelham com o novo dado de entrada. O diagrama deste sistema está apresentado na Figura 12 e a descrição de cada componente encontra-se a seguir.

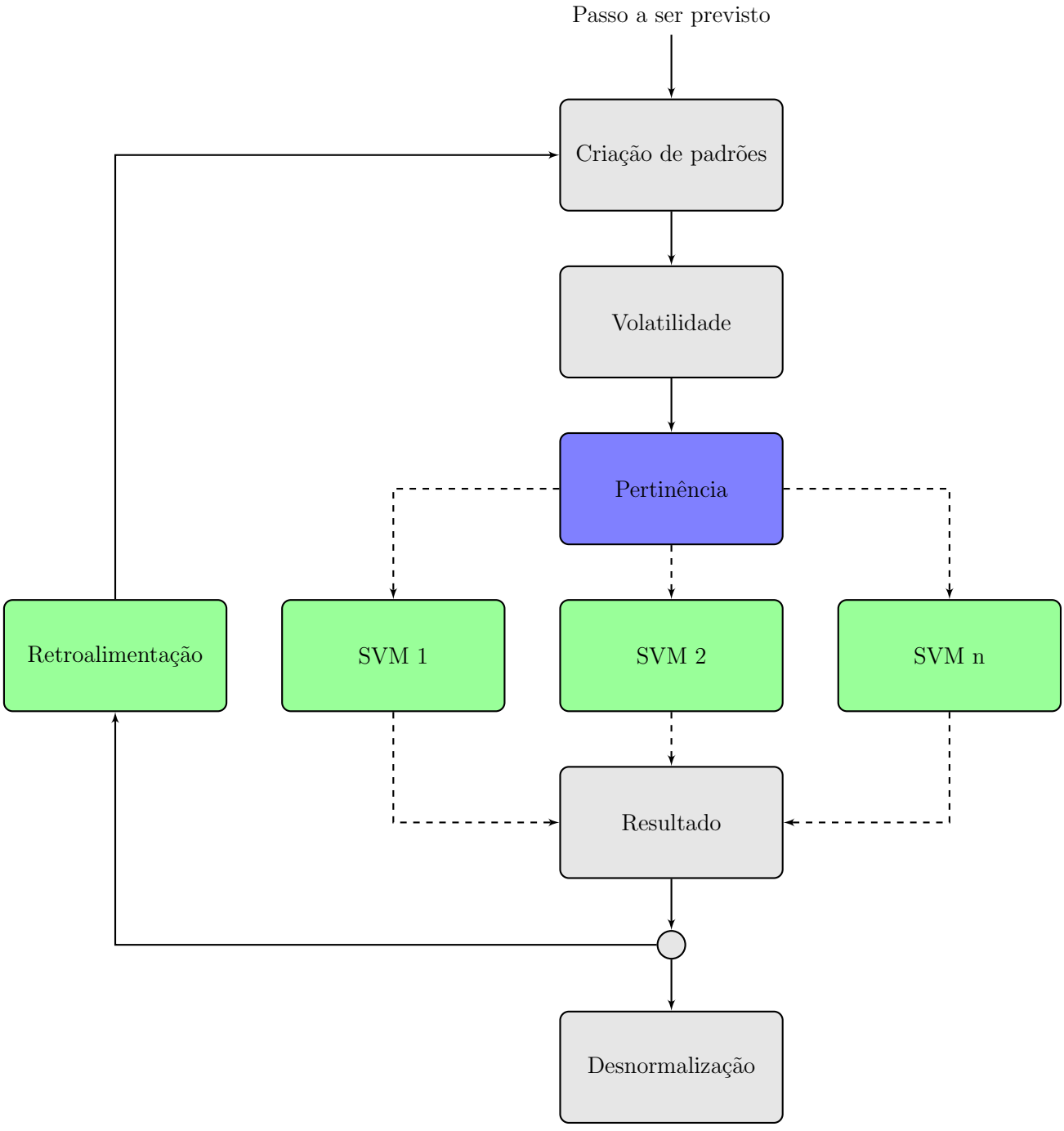

Figura 12: Sistema de previsão hard

Criação de padrões Tendo como entrada o passo a ser previsto, e sendo este não pertencente ao conjunto de treinamento, constroi-se uma tabela de decisão apenas com os atributos de condição a partir dos dados históricos, como também a partir de resultados previamente computados pelo sistema de previsão, no caso de previsão subsequente de vários passos. Tal tabela contém, assim como no treinamento, apenas os valores significativamente relacionados a valores um passo a frente, extraídos do processo de autocorrelação executado no treinamento, sendo seu aspecto definido pela Tabela 7, que exemplifica o caso de três valores significativamente relacionados a valores um passo a frente. Esta tabela sempre conterá uma única linha, relativa ao período que se deseja prever.

Atributos de condição		
lag-x	*lag*-b	*lag*-a

Tabela 7: Exemplo de uma tabela de decisão para previsão

Volatilidade Assim como no treinamento, uma coluna extra, com informações de volatilidade do período a ser previsto, é adicionada à tabela de decisão do passo anterior, tendo seu valor calculado da mesma forma que durante o treinamento, e utilizando tanto os dados históricos quanto os resultados previamente computados pelo sistema de previsão, no caso de previsão subsequente de vários passos. Seguindo o exemplo anterior, a tabela de decisão passa a ter o aspecto da Tabela 8.

Atributos de condição			
lag-x	*lag*-b	*lag*-a	*vol*

Tabela 8: Exemplo de uma tabela de decisão para previsão com volatilidade

Pertinência Utilizando a mesma métrica de distância do processo de clusterização durante o treinamento, define-se a qual *cluster* o novo dado de entrada pertence, sendo este o que possui a menor distância de seu centro até o novo dado. Importante ressaltar que, neste processo, o novo dado pertence a apenas um *cluster*.

Previsores SVM A partir desta informação de pertinência a um *cluster*, escolhe-se a SVM que receberá a incumbência de prever o valor futuro relativo ao novo dado da entrada, que é a SVM treinada para o *cluster* definido, sendo esta a única SVM a realizar a previsão para o novo dado.

Resultado O resultado da previsão é composto apenas do resultado extraído da SVM responsável pela previsão no passo anterior. Este resultado está normalizado, visto que as SVMs foram treinadas com dados nesta forma.

Retroalimentação Para previsões sequenciais de um passo a frente, valores previstos são utilizados para compor tabelas de decisão de novas previsões.

Desnormalização O resultado final da previsão corresponde ao dado previsto desnormalizado, com função inversa à utilizada no processo de normalização do treinamento. Com este resultado é possível calcular o erro de previsão, comparando-o com o valor real através de uma das métricas descritas na Seção 2.1.2.

4.1.2.2 Sistema de previsão soft

Conforme mencionado anteriormente, cada SVM pertencente ao sistema de previsão é especializada em determinados tipos de padrão. Porém, isto não significa que SVMs não possam contribuir entre si para formar um consenso sobre qual o valor previsto para um novo dado de entrada. O sistema de previsão *soft* procura extrair de cada SVM sua opinião sobre qual o valor previsto para um novo dado, e constroi a previsão final aplicando pesos diferentes a estas opiniões, de acordo com a importância da SVM em relação ao novo dado.

Por trabalhar com pertinência parcial de dados a *clusters*, este sistema pode ser utilizado apenas com clusterizações do tipo *soft*.

A Figura 13 apresenta o diagrama deste sistema, bem semelhante ao sistema de previsão *hard*. Os componentes Criação de padrões, Volatilidade e Retroalimentação possuem o mesmo comportamento que o sistema de previsão *hard*, portanto não serão descritos novamente. A seguir, o detalhamento do restante.

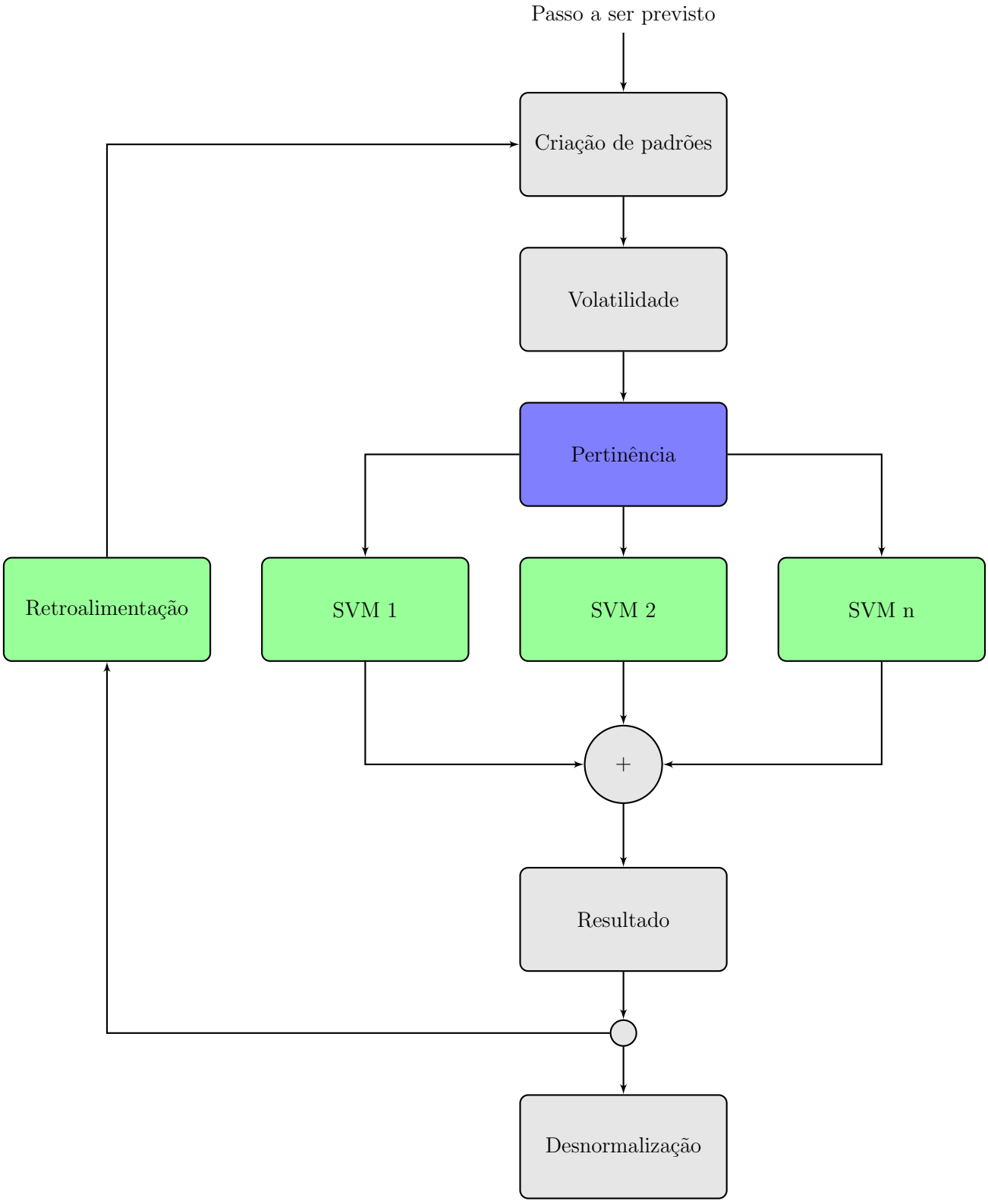

Figura 13: Sistema de previsão soft

Pertinência Calcula a pertinência parcial do novo dado em relação a todos os *clusters* do sistema de previsão, utilizando a mesma métrica de distância e a mesma forma de

cálculo de pertinência, aplicadas no processo de clusterização durante o treinamento. Desta forma, o novo dado de entrada pertence parcialmente a todos os *clusters*, alguns com maior intensidade e outros com menor, de acordo com o grau de pertinência a cada um, sendo que a pertinência total do novo dado aos *clusters* é sempre igual a unidade.

Previsores SVM Todas as SVMs pertencentes ao sistema de previsão recebem o novo dado de entrada, e calculam a saída relativa a ele a partir do conhecimento adquirido durante o treinamento.

Resultado As saídas de todas as SVMs acima são consideradas no cálculo da previsão final, sendo aplicado o peso da pertinência a cada *cluster* no resultado de sua respectiva SVM. Depois de aplicados os pesos, todos os dados são somados para compor o resultado final. Novamente, este resultado está normalizado, visto que as SVMs foram treinadas com dados nesta forma.

4.2 Ferramentas utilizadas

Os experimentos descritos neste estudo foram executados em um PC com processador Intel(R) Core(TM) i7 5500U, 8 GB de memória RAM e sistema operacional Windows 8.1 64 bits.

A implementação da SVM utilizada foi a LIBSVM (CHANG; LIN, 2011), elaborada por pesquisadores do Departamento de Ciência da Computação da Universidade Nacional de Taiwan, sendo desenvolvida continuamente desde o ano de 2000 e adotada com sucesso em diversos trabalhos em diferentes áreas, como visão computacional e processamento de linguagem natural, entre outros (CHANG; LIN, 2011). Possui soluções para classificação de padrões (C-SVC, ν-SVC) com suporte a múltiplas classes, estimação de distribuição (*one-class* SVM) e regressão (ε-SVR, ν-SVR).

Esta biblioteca possui implementações em C++ e Java, tendo extensões para outras linguagens como Python, Perl, PHP, C# e MATLAB. A versão utilizada neste estudo foi a Java 3.20, sendo o módulo de regressão ε-SVR o escolhido para os testes. Este módulo possibilita o ajuste dos parâmetros C e ε, bem como da função de *kernel*, essenciais para o desenvolvimento das atividades propostas.

Para os processos de clusterização *K-Means* e *Fuzzy C-Means* foi empregada a ferramenta MATLAB Release 2009a (THE MATHWORKS INC.,), amplamente utilizada por

pesquisadores para os mais diversos fins. Algumas funções foram modificadas de forma a padronizar a métrica de distância utilizada durante a clusterização.

Todo o restante do código, incluindo lógicas para normalização, padronização, volatilidade etc, foi desenvolvido na ferramenta MATLAB, sendo este código integrado à biblioteca LIBSVM citada anteriormente, quando necessário.

Por fim, as plotagens referentes à revisão da teoria e experimentos foram feitas utilizando a ferramenta *gnuplot* (WILLIAMS; KELLEY; many others, 2013).

4.3 Série temporal utilizada

Gerido pela BB Gestão de Recursos - Distribuidora de Títulos e Valores Mobiliários S.A. - BB DTVM, o Fundo BB Ações IBrX Indexado do Banco do Brasil S/A aplica seus recursos em cotas de fundo de investimento que apresenta uma carteira de ativos que reflete o comportamento da carteira teórica do IBrX - Índice Brasil.

O índice IBrX mede o retorno de uma carteira teórica composta pelas 100 ações mais negociadas (em número de negócios e volume financeiro medidos nos últimos doze meses) na BOVESPA, sendo estas ponderadas pelo seu respectivo número de ações disponíveis para negociação no mercado. A Figura 14 mostra a composição do índice por setor de atuação, em Dezembro de 2015.

A série temporal utilizada neste estudo é composta pelos valores diários da cota desse fundo e possui, pelo exposto anteriormente, todas as características de uma série temporal financeira, apresentando, por exemplo, tendências, escala, períodos com volatilidades distintas etc.

O período da série utilizado compreende os dias de operação do fundo no mercado de ações entre 1^o de Julho de 2002 e 31 de Dezembro de 2009, totalizando 1889 amostras. A Figura 15 apresenta o gráfico de valor das cotas deste período.

O período escolhido é exatamente o mesmo utilizado por Leite (2010) em seus estudos, com o propósito de efetuar comparações entre seus resultados e os deste estudo.

4.4 Estudo da série temporal

As seções descritas a seguir abordam, respectivamente, as técnicas utilizadas nos componentes Normalização, Autocorrelação, Criação de padrões e Volatilidade do modelo

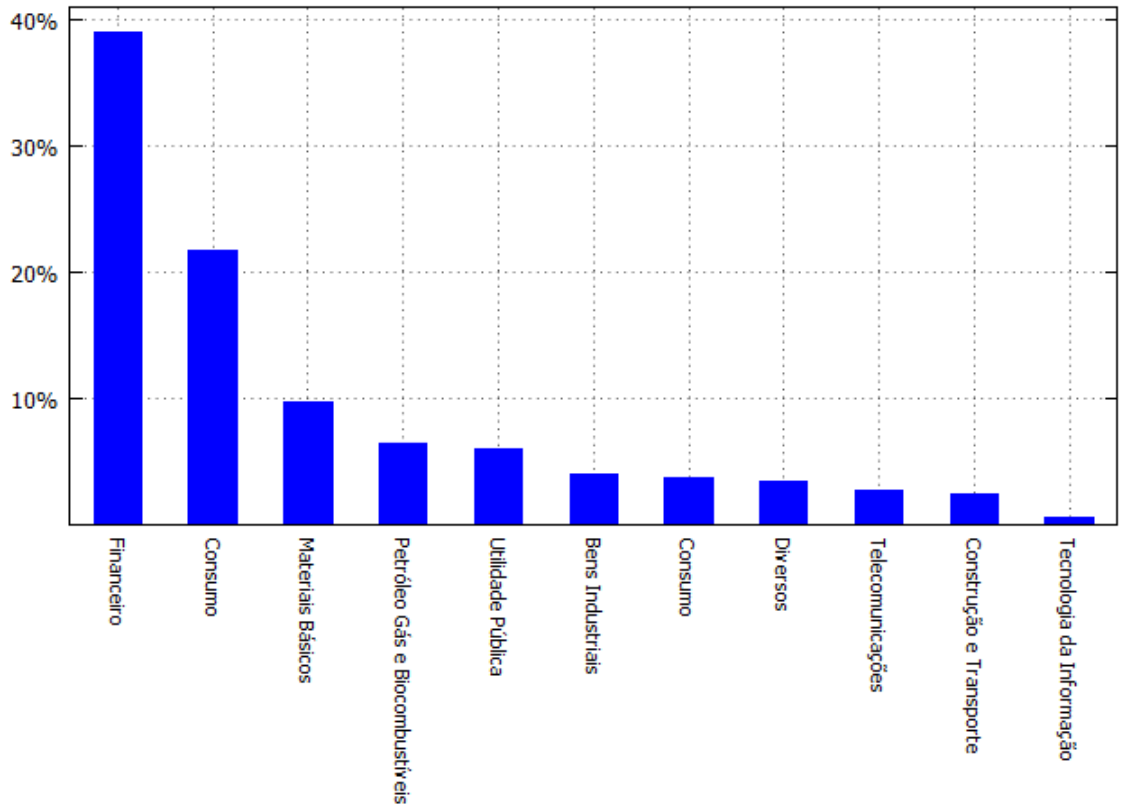

Figura 14: Composição do índice IBrX em Dezembro de 2015

Figura 15: Série de preços - Fundo BB Ações IBrX Indexado

proposto, apresentando os resultados obtidos após cada processamento.

4.4.1 Série de retorno

Conforme descrito na Seção 2.1.3.1, a série de retorno possui características que a tornam atrativas para análise e previsão como séries temporais que não estão presentes na série de preços, como ser livre de escala e poder ser considerada uma série fracamente estacionária. Isto a faz ser preferida, em relação à série de preços, nos estudos financeiros.

Dentre os tipos de retorno existentes, o utilizado neste estudo será o retorno composto contínuo, cujo cálculo é feito através da equação 2.9. O resultado da aplicação deste processo na série original, descrita na Seção 4.3, é uma nova série, contendo agora 1888 amostras relativas ao período de 2 de Julho de 2002 até 31 de Dezembro de 2009, e demonstrada na Figura 16. Apesar de carregar intrinsecamente as mesmas informações da série original, a série de retorno possui características completamente diferentes.

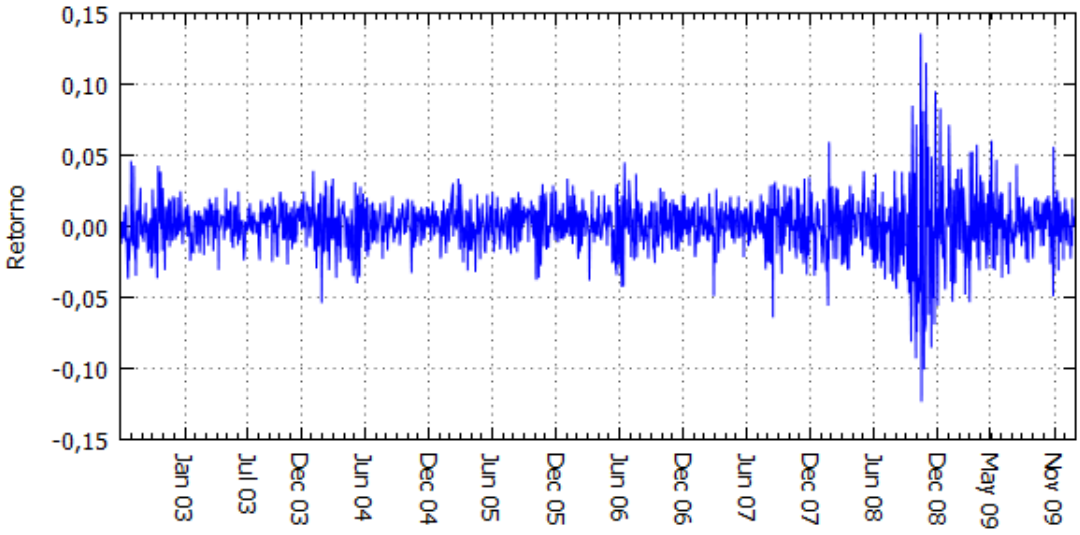

Figura 16: Série de retorno - Fundo BB Ações IBrX Indexado

É possível observar que, ao contrário da série original, a série de retorno não possui tendência; tem aspecto horizontal, demonstrando, nos pontos de magnitude elevada, períodos com variações abruptas de preço e, nos outros, períodos de calmaria, conforme Figura 17. A média da série de retorno é próxima a zero (0,00085), sendo da série de preços, 3,99471.

Assim, esta será a série que alimentará o restante do modelo previsor, ocorrendo o processo inverso (conversão de retorno para preço) apenas ao final, quando se desejar saber o valor real de uma previsão.

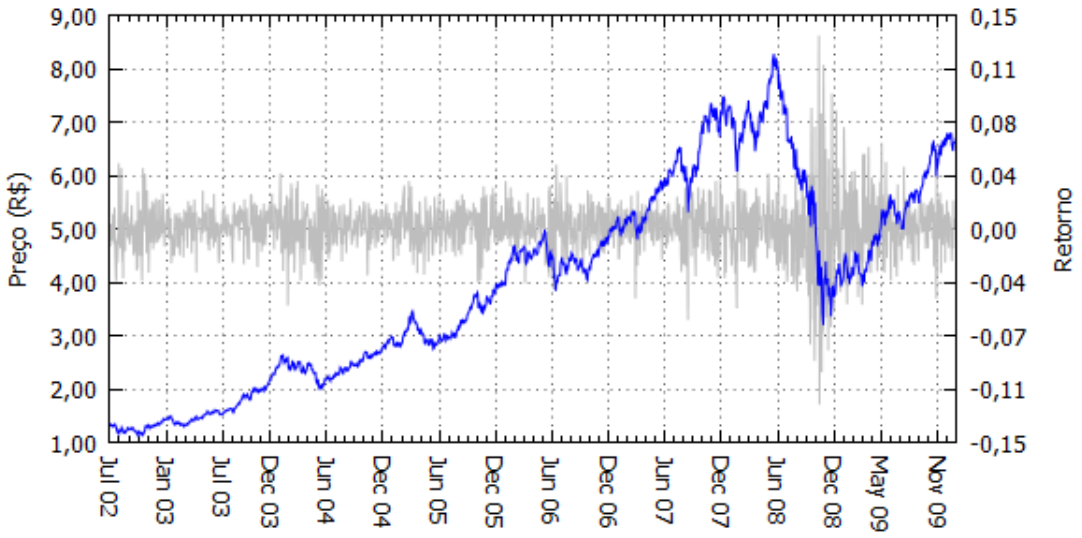

Figura 17: Série de preços comparada à série de retorno - Fundo BB Ações IBrX Indexado

4.4.2 Autocorrelação

Como mencionado na Seção 2.1.3.3, nem todos os períodos do passado influenciam significativamente em valores futuros. Tendo isso em mente, é possível selecionar os períodos mais apropriados para se trabalhar com previsões futuras, o que traz diversas vantagens para um sistema de previsão, como: diminuição da complexidade, já que o sistema trabalhará com um número menor de variáveis; eliminação de dados espúrios, visto que períodos passados, com baixa relevância na alteração de comportamentos futuros, pouco ou nada têm a contribuir com o processo de previsão; e também estabelecimento de limite de análise do passado.

Para atingir este objetivo, emprega-se a função de autocorrelação na série de retorno, conforme Equação 2.13. O resultado desta aplicação, na série de retorno da Figura 16, para uma janela de vinte períodos passados, está plotado no correlograma da Figura 18.

Valores positivos de autocorrelação indicam que a tendência do período futuro é variar no mesmo sentido do período passado, e valores negativos indicam que os períodos variam em sentidos opostos. Quanto maior a magnitude, maior a relação entre os períodos. Já o valor zero indica a completa falta de correlação entre os períodos, sendo que variações em um não afetam o outro.

Os valores relevantes são aqueles cuja magnitude é superior a 5%. Assim, para a série de retorno em análise, os valores de *lag* a serem empregados como padrões de entrada para previsão de valores futuros são 1, 2, 3, 17 e 18.

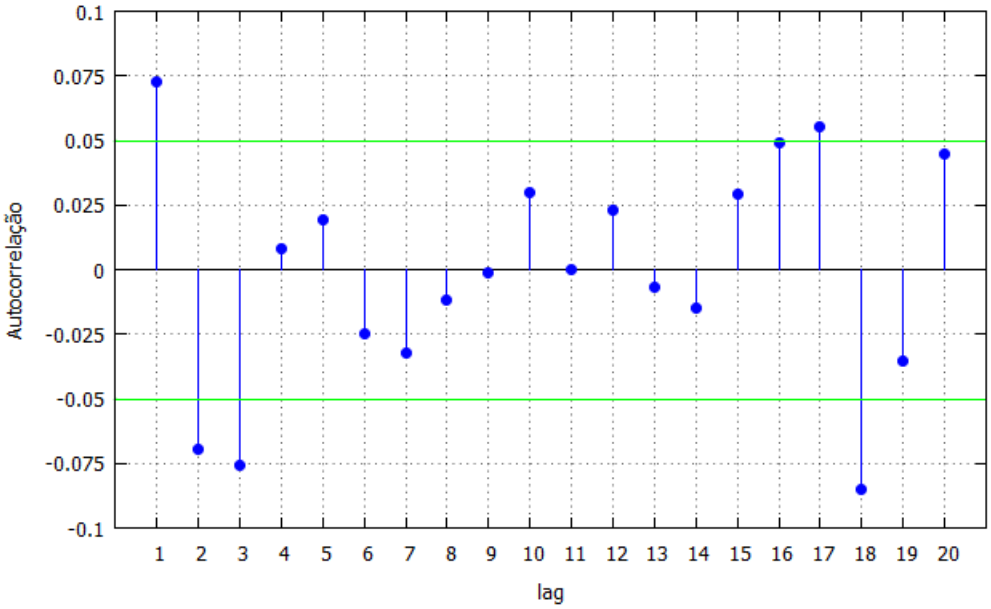

Figura 18: Correlograma de 20 dias da série de retorno - Fundo BB Ações IBrX Indexado

4.4.3 Criação de padrões

Os dados da série temporal devem ser dispostos de maneira que possam ser utilizados por mecanismos posteriores do sistema de previsão. Para tanto, a série de retorno deve ser transformada em um tabela de decisão, levando em conta a associação de valores passados, extraídos do processo de autocorrelação, com um valor futuro, conforme descrito na Seção 4.1.

Para a série em estudo, a tabela de decisão deve ter o formato da Tabela 9.

Atributos de condição					**Atributos de decisão**
lag-18	*lag*-17	*lag*-3	*lag*-2	*lag*-1	*lag*-0

Tabela 9: Padrão da tabela de decisão para série de retorno IBrX

A tabela conterá 1870 linhas correspondentes aos 1870 valores futuros (atributos de decisão), que podem ser associados individualmente aos cinco valores passados (atributos de condição).

4.4.4 Volatilidade

Descrita na Seção 2.1.3.4, a volatilidade vem agregar conhecimento aos padrões pré-estabelecidos da série temporal, permitindo que o sistema de previsão possa discernir entre períodos de volatilidade distinta e aplicar assim regressões diferentes para cada caso.

Para tanto, uma nova coluna é adicionada à Tabela 9, contendo um valor numérico que indica se o período *lag*-0 da entrada da tabela de decisão é de baixa ou alta volatilidade, sendo este valor 0 ou +1, respectivamente, calculado pela Equação 2.14.

A tabela de decisão passará a ter o formato da Tabela 10, porém agora com 1868 linhas, correspondentes aos 1868 valores futuros que podem ter calculados seus valores de volatilidade, bem como ser associados individualmente aos cinco valores passados extraídos do processo de autocorrelação.

Atributos de condição						**Atributos de decisão**
lag-18	*lag*-17	*lag*-3	*lag*-2	*lag*-1	vol	*lag*-0

Tabela 10: Padrão da tabela de decisão com volatilidade para série de retorno IBrX

A Figura 19 mostra em amarelo o desvio padrão instantâneo e em vermelho, o desvio padrão da série de retorno como um todo. Os pontos em que a linha amarela ultrapassa o limiar vermelho são considerados de alta volatilidade. Todos os outros são de baixa volatilidade.

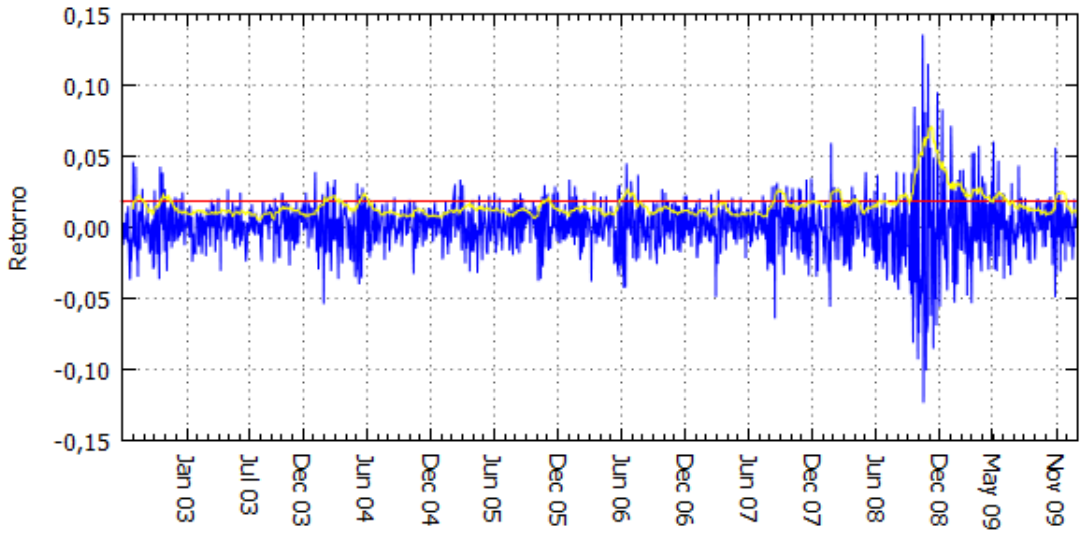

Figura 19: Análise de volatilidade para série de retorno - Fundo BB Ações IBrX Indexado

5 Experimentos e Resultados

Todos os experimentos realizados tiveram como referência a série temporal descrita na Seção 4.3, devidamente tratada conforme descrições da Seção 4.4.

Os testes do modelo foram realizados em dez janelas distintas da série temporal, da mesma maneira que realizados pelo estudo de Leite (2010). Desta forma, torna-se possível realizar comparações entre os modelos, de forma a elencar pontos positivos e negativos do atual estudo, tendo como base um sistema cuja capacidade de previsão se mostrou bastante eficiente.

Cada uma das dez janelas de previsão para teste são compostas por vinte dias consecutivos de operação do fundo, sendo o sistema retroalimentado, a fim de utilizar previsões já realizadas como entrada para previsões a frente. O sistema realiza, então, previsões vinte passos a frente. As janelas possuem características por vezes distintas, como direção do movimento (ascendente, horizontal ou descendente) e volatilidade (possuindo períodos de baixa ou alta volatilidade, ou uma mistura destes), sendo também espaçadas no tempo.

A tabela 11 descreve as dez janelas, citando algumas de suas características.

Experimento	**Período**		**Dias de volatilidade**		**Desvio Padrão**
	Início	**Fim**	**Baixa**	**Alta**	
1	24-06-2003	21-07-2003	20	0	0,005809
2	22-11-2004	17-12-2004	20	0	0,007526
3	23-08-2005	20-09-2005	20	0	0,007031
4	28-05-2007	25-06-2007	20	0	0,009718
5	11-09-2009	08-10-2009	20	0	0,009881
6	28-01-2005	28-02-2005	20	0	0,012627
7	19-05-2006	16-06-2006	5	15	0,026984
8	18-01-2008	18-02-2008	2	18	0,025356
9	03-09-2008	30-09-2008	1	19	0,049167
10	14-10-2009	11-11-2009	11	9	0,024033

Tabela 11: Janelas de teste

Os experimentos de 1 a 5 referem-se a períodos com menor variação dos preços, com-

parados aos períodos de 6 a 10. Os períodos de 7 a 10 possuem dias que são caracterizados por serem de alta volatilidade, em função do desvio padrão instantâneo elevado.

As informações de cada janela foram devidamente removidas do processo de treinamento, com o propósito de não haver influência dos resultados originais no processo de previsão, sendo possível desta forma medir a capacidade de generalização do sistema. Assim sendo, para cada teste, vinte dias pertencentes à janela de teste em questão foram removidos da tabela de decisão descrita na Seção 4.4.4, deixando-a então com 1848 linhas e exatamente com o mesmo aspecto já descrito.

Nas clusterizações realizadas nos experimentos, os dados apresentados para o algoritmo foram os 1848 atributos de condição dos padrões de treinamento, estando eles no formato da Tabela 10, previamente discutida. Para o treinamento das SVMs, os dados apresentados foram os padrões completos, conforme Tabela 10, pertinentes aos *clusters* de cada SVM. Para previsão, os dados são apresentados à SVM no formato dos atributos de condição desta mesma tabela, e o resultado generalizado pela SVM é fornecido no formato do atributo de decisão, também da mesma tabela.

O processo de escolha do número de *clusters* foi empírico, sendo apresentados neste estudo os melhores resultados obtidos. O mesmo vale para os valores dos parâmetros C e ε das SVMs de regressão. Vale ressaltar que todas as SVMs de um mesmo experimento compartilham os mesmos valores de parâmetros. Para cada um dos três parâmetros livres do sistema, uma faixa de valores foi varrida a fim de se obter o melhor para o problema em questão. Para o número de clusters, a faixa considerada foi de 2 até 18 clusters. Já o parâmetro C foi de 550 a 10.050, em passos de 500, e o parâmetro ε, de $0,0001$ a $0,1001$ em passos de $0,01$.

Para medida de eficiência e qualidade do sistema, foi adotada a medida de erro MAPE, descrita na Seção 2.1.2, como também o cálculo do desvio padrão dos erros de previsão. Tal escolha tem por objetivo, dentre outros, facilitar a comparação com os resultados obtidos por Leite (2010), visto que estas foram as medidas adotadas em seus experimentos.

As seções seguintes descrevem os resultados alcançados usando os diferentes métodos de clusterização e as diferentes variações aplicadas ao sistema de previsão.

5.1 Cópia Carbono

Com o propósito de servir como referência, os experimentos descritos neste estudo foram executados utilizando como método de previsão o sistema mais simples possível, chamado de Cópia Carbono (*Carbon Copy* - CRB). Neste método, o valor de previsão de um dado período é igual ao valor do período imediatamente anterior, conforme equação a seguir.

$$\widehat{Y}_{t+1} = Y_t \tag{5.1}$$

Sendo assim, para um dada janela de previsão de vinte períodos, o primeiro valor previsto será igual ao valor do período imediatamente anterior, sendo este um valor real. O segundo valor previsto será igual ao valor do período imediatamente anterior, sendo este um valor previsto, descrito anteriormente. Este processo continua até o último período a ser previsto. Desta forma, todos os períodos previstos terão o mesmo valor, que é igual ao valor do último período antes da janela de previsão, conforme a seguir.

$$\widehat{Y}_{t+20} = \widehat{Y}_{t+19} = ... = \widehat{Y}_{t+2} = \widehat{Y}_{t+1} = Y_t \tag{5.2}$$

Além dos erros MAPE, o indicador estatístico U de *Theil*, foi também utilizado como meio de comparação de desempenho entre os sistemas propostos neste estudo e o método de referência CRB. Este indicador é calculado através da razão entre os valores do erro RMSE do sistema cujo desempenho queira-se avaliar e do método de referência CRB (MAKRIDAKIS; WHELLWRIGHT; HYNDMAN, 1998).

O valor do indicador U de *Theil* é interpretado conforme a seguir.

- Valor igual a um indica que ambos os métodos possuem desempenho igual.
- Valores menores que um indicam que o método em avaliação possui desempenho superior ao método de referência.
- Valores maiores que um indicam que o método em avaliação possui desempenho inferior ao método de referência.

Sendo então desejáveis valores pequenos do indicador, visto que quanto menor seu valor, melhor é o desempenho do sistema em avaliação em relação ao método de referência.

Os resultados da aplicação deste método nos experimentos serão apresentados nas seções seguintes, juntamente com os resultados dos sistemas que utilizam métodos de clusterização, a fim de se fazer comparações entre os modelos.

5.2 SVM pura

Para que seja possível fazer comparações entre o modelo aqui proposto e um sistema convencional e conhecido, os experimentos também foram executados aplicando os dados em uma SVM pura. Tal sistema consiste em uma única SVM, treinada com os dados no formato da Tabela 7.

A SVM pura é treinada com todos os dados da série, exceto os pertencentes à janela de previsão do experimento em questão. Processo algum de clusterização é aplicado e tampouco informações de volatilidade são inseridas.

Os parâmetros utilizados na SVM pura estão apresentados na Tabela 12, sendo estes os que produziram os melhores resultados de previsão.

SVM pura	
$\boldsymbol{C}$	$\boldsymbol{\varepsilon}$
550	0,0061

Tabela 12: Parâmetros para experimento com SVM pura

Sendo o propósito deste estudo apresentar uma nova proposta de previsão de séries temporais envolvendo aplicação de técnicas de clusterização, os resultados relativos à SVM pura serão apresentados nas seções seguintes, que envolvem a aplicação de tais técnicas no processo de previsão, como forma de comparação, a fim de elaborar conclusões sobre a proposta aqui em questão.

5.3 K-Means

Sendo *K-Means* um algoritmo de clusterização *hard*, o sistema de previsão utilizado no experimento descrito nesta seção é o sistema de previsão *hard*, descrito na Seção 4.1.2.1.

Os parâmetros que apresentaram o melhor resultado de previsão, quando utilizado o algoritmo de clusterização *K-Means*, estão descritos na Tabela 13.

Tendo cada padrão informação sobre sua volatilidade, fica a critério do algoritmo de clusterização a quantidade de *clusters* de baixa e alta volatilidade que serão criados, visto

Número de *clusters*	**SVM**	
	C	ε
12	550	0,0601

Tabela 13: Parâmetros para experimento com clusterização *K-Means*

que as informações de volatilidade são separadas por um valor cuja magnitude faz com que, em um dado *cluster*, apenas padrões de uma volatilidade estejam presentes.

Quando utilizada a clusterização *K-Means*, a quantidade de *clusters* de baixa e alta volatilidades foi igual em todos os experimentos, conforme Tabela 14. É possível notar que o processo de clusterização deste algoritmo produz sempre resultados consistentes, não importando as pequenas variações de padrões de cada experimento.

Experimento	**Quantidade de *clusters***	
	Baixa	**Alta**
1	5	7
2	5	7
3	5	7
4	5	7
5	5	7
6	5	7
7	5	7
8	5	7
9	5	7
10	5	7

Tabela 14: Quantidade de *clusters* de baixa e alta volatilidade para clusterização *K-Means*

Conforme já mencionado, o experimento aqui descrito foi realizado utilizando o Sistema Previsor *Hard*, o que indica que cada previsão é realizada exclusivamente por uma única SVM, sendo esta a que pertence ao *cluster* mais próximo do novo padrão de entrada relativo ao período futuro que se deseja prever.

A Tabela 15 indica a quantidade de *clusters* de baixa e alta volatilidade utilizada na previsão de cada experimento. Se a SVM de um dado cluster for responsável pela previsão de ao menos um dia no experimento, o *cluster* é considerado como utilizado no experimento.

Nesta tabela, é possível notar que todos os experimentos utilizaram apenas SVMs de baixa volatilidade, o que inclui os experimentos que englobam períodos de alta volatilidade. Vale ressaltar que a decisão sobre qual SVM utilizar cabe exclusivamente ao sistema de previsão.

Experimento	**Quantidade de *clusters* utilizados**	
	Baixa	**Alta**
1	5	0
2	5	0
3	4	0
4	5	0
5	3	0
6	3	0
7	4	0
8	4	0
9	4	0
10	4	0

Tabela 15: Quantidade de *clusters* de baixa e alta volatilidade utilizada na previsão de cada experimento para clusterização *K-Means*

Tal resultado pode ser interpretado como um excesso de generalização do sistema, tornando-o incapaz de produzir valores com variação de amplitude significativa (ou seja, a ponto de elevar o desvio padrão instantâneo acima do desvio padrão da série), fazendo-o refém das informações de baixa volatilidade.

É importante também notar que as previsões foram realizadas por várias SVMs em cada experimento, mostrando que o conhecimento sobre a série temporal foi espalhado entre as diversas SVMs.

A Tabela 16, a seguir, apresenta o erro MAPE de cada experimento, comparando-os com o erro MAPE da SVM pura, descrita na Seção 5.2, e com o erro MAPE do método CRB, descrito na Seção 5.1.

Experimento	**CRB**	**SVM pura**	***K-Means***
1	2.7418	1,0203	0,5046
2	4.3019	2,3018	0,6772
3	4.3986	2,9234	1,0332
4	3.2120	1,2918	0,9571
5	4.0398	2,5090	0,9775
6	7.4515	5,8195	1,3543
7	5.7648	8,1299	3,4139
8	4.1066	2,6267	1,9366
9	7.1954	9,1502	6,9022
10	2.2059	2,6163	1,9587
Erro médio	4.5418	3,8389	1,9715
Desvio padrão	1.7635	2,8463	1,9064

Tabela 16: MAPE dos experimentos utilizando clusterização *K-Means*

A análise numérica dos resultados mostra que o erro médio, resultante da previsão utilizando a clusterização *K-Means*, é $48,65\%$ menor comparado à previsão com a SVM pura. É possível também observar que os erros de todos os experimentos foram menores quando utilizado o sistema de previsão com clusterização *K-Means*. Outro fator importante é o desvio padrão dos erros - $33,02\%$ menor quando utilizado o sistema clusterizado.

A comparação entre o sistema com clusterização *K-Means* e o método de referência CRB mostra que o primeiro obteve erro $56,59\%$ menor. O desvio padrão apresentado pelo método de referência CRB foi menor que o apresentado pelo sistema com clusterização *K-Means*, porém tal informação é irrelevante, dado o alto erro MAPE apresentado pelo método CRB.

O indicador U de *Theil* para o sistema com clusterização *K-Means* foi calculado através do erro RMSE, apresentado na Tabela 17. Seu valor foi de $0,4928$, calculado através da razão entre os erros RMSE médios do sistema com clusterização *K-Means* e do método CRB. Tal valor indica que o modelo proposto é substancialmente mais preciso que o modelo de referência.

Experimento	**CRB**	***K-Means***
1	0.0539	0.0094
2	0.1373	0.0243
3	0.1915	0.0407
4	0.2227	0.0662
5	0.2792	0.0701
6	0.2785	0.0573
7	0.3139	0.1763
8	0.3280	0.1649
9	0.4401	0.4232
10	0.1688	0.1571
Erro médio	0.2414	0.1190

Tabela 17: RMSE dos experimentos utilizando clusterização *K-Means*

As Figuras de 20 a 29 ilustram as previsões destes experimentos quando utilizada a clusterização *K-Means*, comparando-as com o valor real da série, e também com a previsão utilizando apenas a SVM pura, descrita na Seção 5.2.

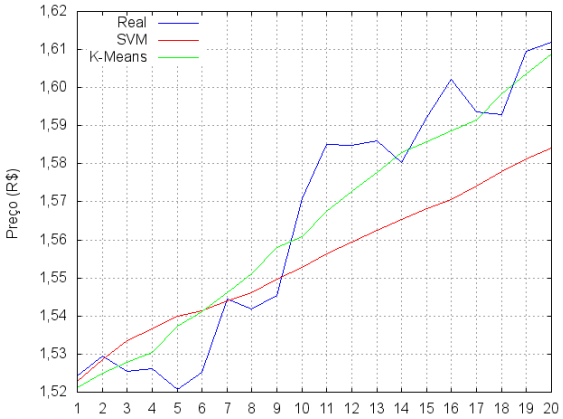

Figura 20: Experimento 1 - *K-Means*

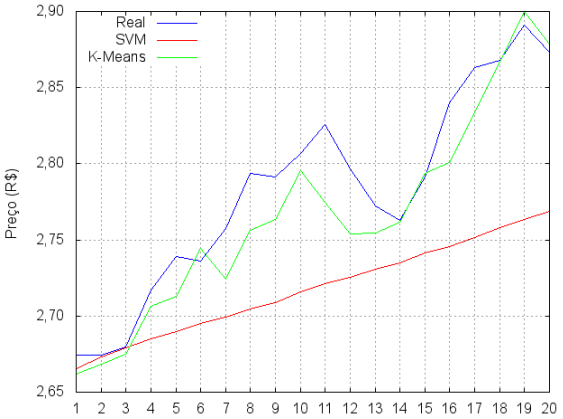

Figura 21: Experimento 2 - *K-Means*

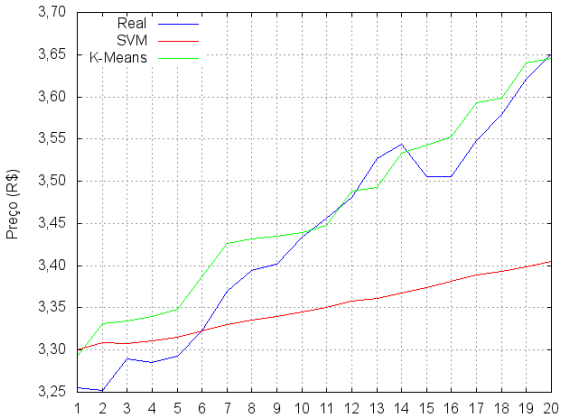

Figura 22: Experimento 3 - *K-Means*

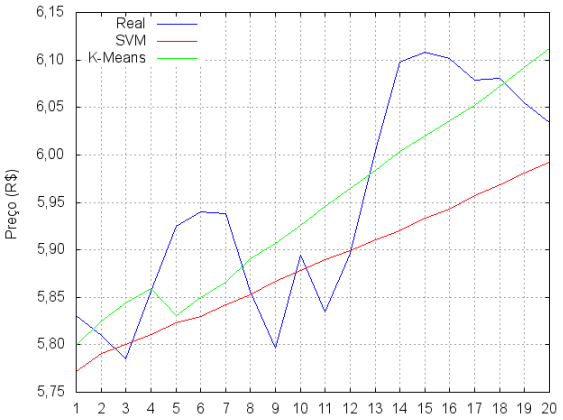

Figura 23: Experimento 4 - *K-Means*

Figura 24: Experimento 5 - *K-Means*

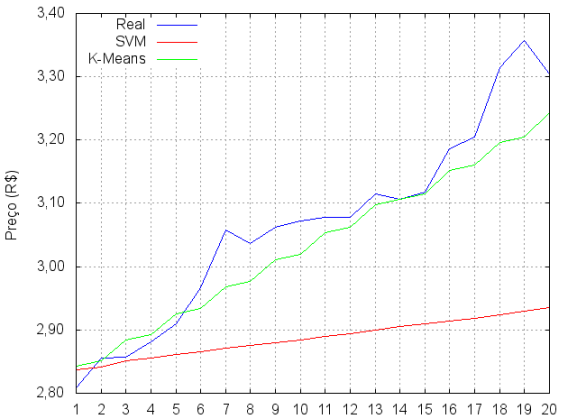

Figura 25: Experimento 6 - *K-Means*

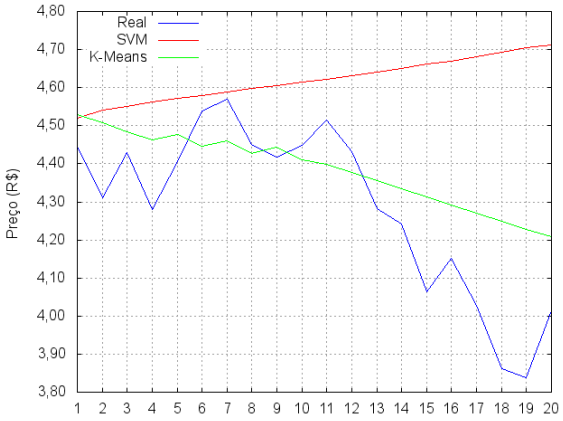

Figura 26: Experimento 7 - *K-Means*

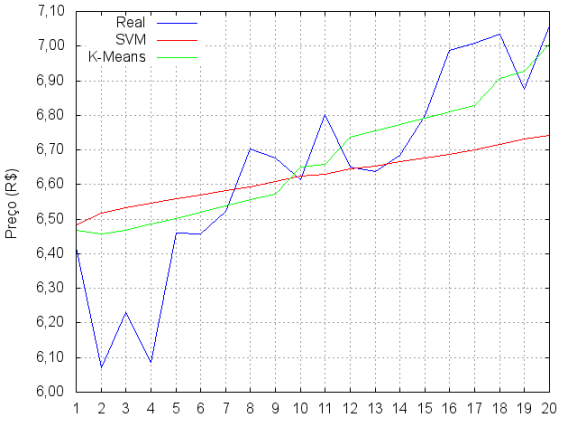

Figura 27: Experimento 8 - *K-Means*

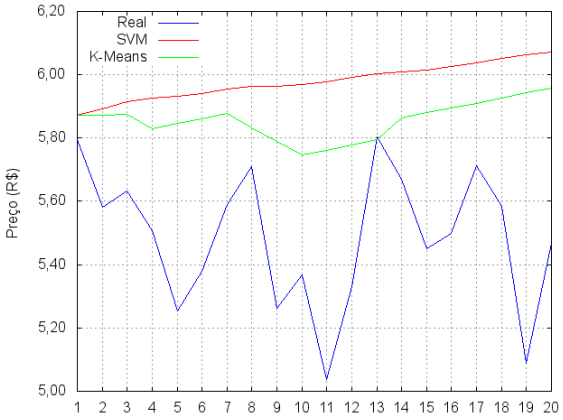

Figura 28: Experimento 9 - *K-Means*

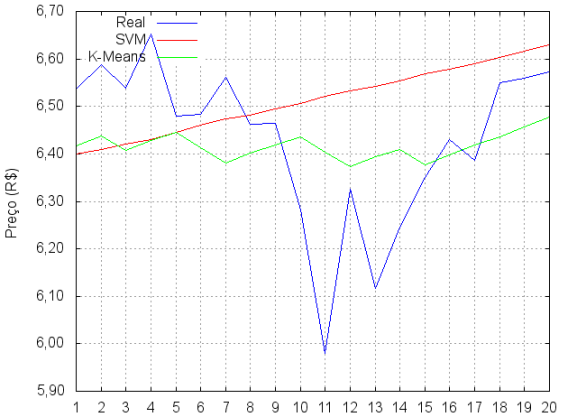

Figura 29: Experimento 10 - *K-Means*

Analisando o resultado graficamente, pode-se observar que o sistema com clusterização *K-Means* acertou a tendência dos preços em nove dos dez experimentos, contra sete tendências corretas do sistema com a SVM pura. É possível notar também que, em alguns dos experimentos, a série prevista tende a acompanhar os movimentos da série real, fato que não ocorre em nenhum dos experimentos com a SVM pura, cujo resultado praticamente não apresenta variações de movimento.

Conclui-se então que o excesso de informação, além de gerar erros de previsão superiores, generaliza em demasiado os resultados, de tal sorte que os valores previstos seguem sempre movimentos suaves. Porém, esta generalização excessiva também se mostra presente em alguns dos resultados do sistema com clusterização *K-Means*, principalmente nos experimentos que possuem períodos de alta volatilidade.

Conforme já demonstrado anteriormente, apesar de o sistema possuir informações de baixa e alta volatilidade, apenas as informações de baixa volatilidade foram utilizadas.

Isso explica a falta de precisão das previsões dos experimentos de alta volatilidade (experimentos de 6 a 10). O erro médio dos experimentos de baixa volatilidade foi de $0,8299\%$, enquanto o de alta volatilidade foi de $3,1131\%$.

Apesar do erro de alta volatilidade ser muito superior ao de baixa volatilidade, pode-se considerar o resultado satisfatório, visto que foi ainda inferior ao erro do sistema com a SVM pura. Além disso, previsão de períodos de alta volatilidade tende a ser mais desafiadora que a de baixa devido à constante variação brusca do movimento. Ainda, pode-se dizer que as informações de baixa volatilidade carregam consigo informações relevantes sobre os períodos de alta volatilidade.

Analisando os resultados por completo, pode-se dizer que o sistema clusterizado é mais preciso, com erros menores e mais próximos à média - fato comprovado pelo desvio padrão do erro inferior produzido por este sistema. Este dado pode ser classificado como de suma importância, visto que sistemas que produzem resultados com erros previsíveis são mais confiáveis, comparados aos que produzem, ora resultados excelentes, ora resultados péssimos.

Em termos gerais, os resultados descritos anteriormente mostram que o excesso de informação pode levar a resultados imprecisos, visto que informações pouco ou nada relevantes ao problema a ser resolvido tendem a contribuir negativamente ao processo de a sua resolução. A aplicação de clusterização das informações, e posterior treinamento por *cluster*, caminham no sentido de separar os conhecimentos em grupos especializados, de forma que grupos que não possuam conhecimento sobre um determinado problema não atuem de forma negativa no processo de busca por uma solução.

As Figuras de 30 a 39 mostram o resultado da previsão um período a frente dentro de cada janela dos dez experimentos de todos os *clusters* - linhas em cinza -, e as marcações em verde correspondem ao valor previsto relativo ao *cluster* escolhido para realizar a previsão pelo sistema, sendo o critério de escolha baseado na menor distância entre o padrão a ser previsto e o centro do *cluster*.

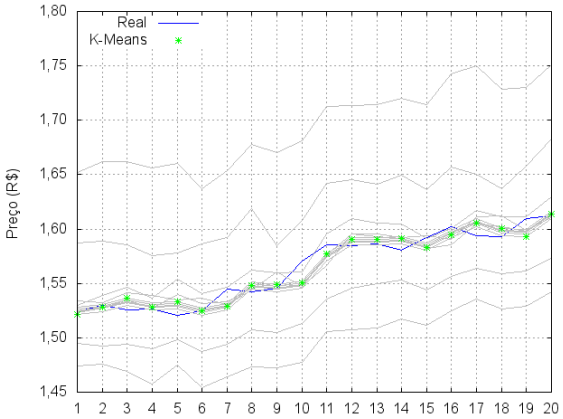

Figura 30: Experimento 1 - Análise de *clusters K-Means*

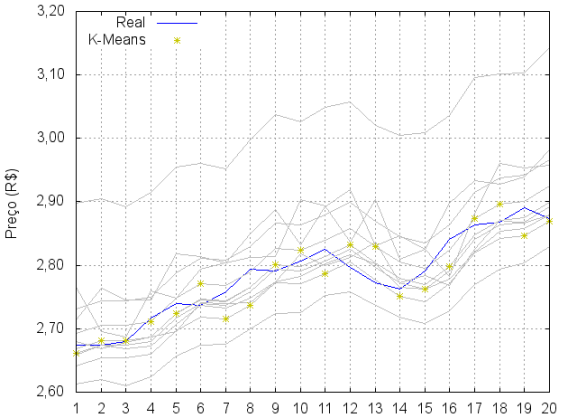

Figura 31: Experimento 2 - Análise de *clusters K-Means*

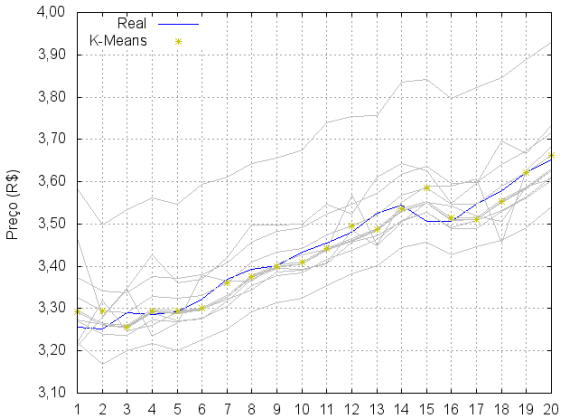

Figura 32: Experimento 3 - Análise de *clusters K-Means*

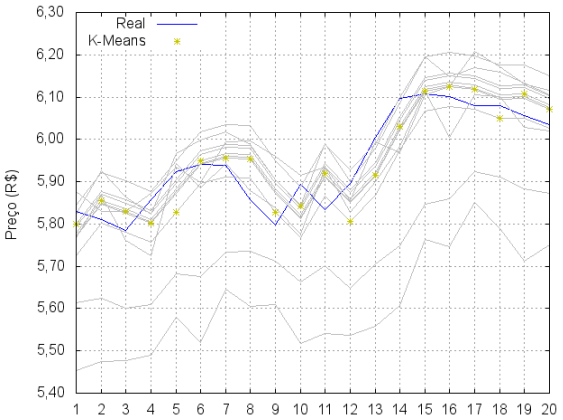

Figura 33: Experimento 4 - Análise de *clusters K-Means*

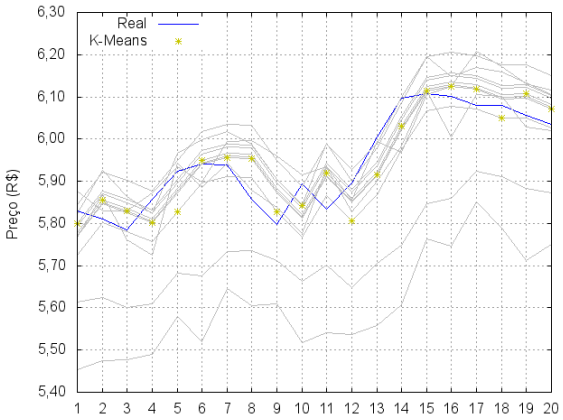

Figura 34: Experimento 5 - Análise de *clusters K-Means*

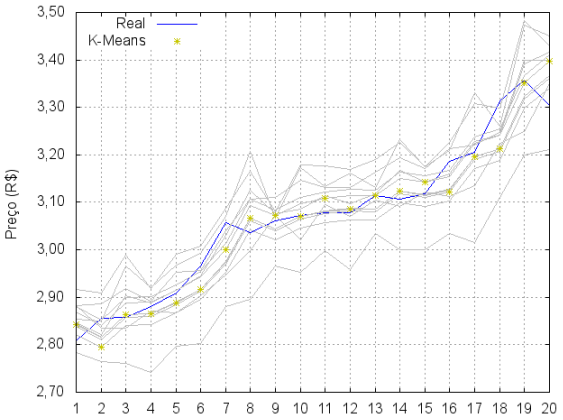

Figura 35: Experimento 6 - Análise de *clusters K-Means*

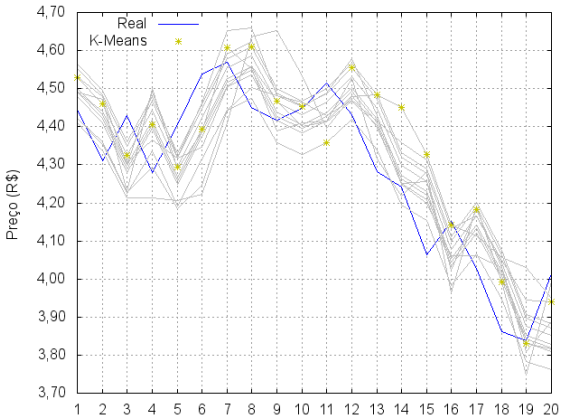

Figura 36: Experimento 7 - Análise de *clusters K-Means*

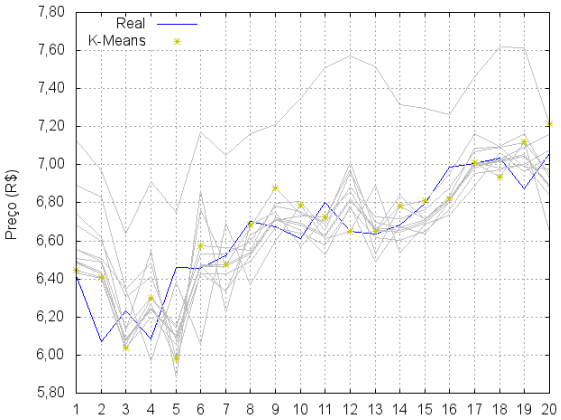

Figura 37: Experimento 8 - Análise de *clusters K-Means*

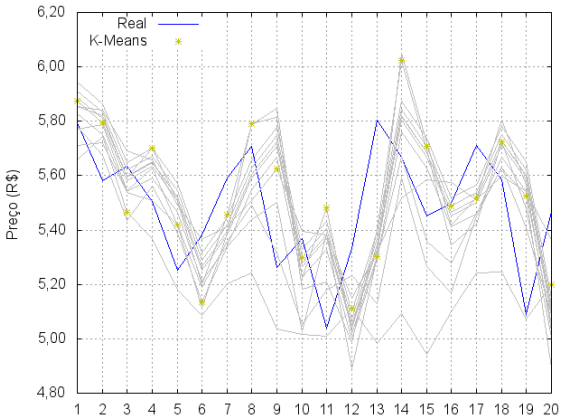

Figura 38: Experimento 9 - Análise de *clusters K-Means*

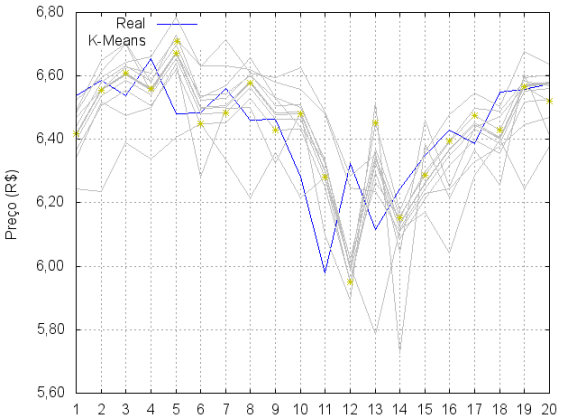

Figura 39: Experimento 10 - Análise de *clusters K-Means*

É possível observar que, em todos os experimentos, diversos *clusters* são capazes de fazer previsões razoáveis da série, porém outros, devido à natureza diferente dos padrões, não possuem tal habilidade. Ainda referenciando esta análise, a Tabela 18 mostra as escolhas feitas pelo sistema de previsão que utilizaram o *cluster* responsável pelo melhor resultado.

Mesmo sendo baixa a taxa de acerto, a análise visual do gráfico mostra que, para todos os cenários, *clusters* distintos são capazes de produzir previsões com baixo erro, e grande parte das escolhas concentra-se em *clusters* pertencentes a este grupo.

Experimento	Escolhas corretas	MAPE
1	1	0,5010
2	3	0,8982
3	3	0,6405
4	1	0,8459
5	2	0,7779
6	2	1,0382
7	2	2,6927
8	4	2,0582
9	2	4,1728
10	2	1,9300
Taxa de acerto / Erro médio	11%	1,5555

Tabela 18: Acerto na escolha de *cluster* - *K-Means*

5.4 C-Means

Para a clusterização *C-Means*, além do sistema de previsão *hard*, descrito na Seção 4.1.2.1, foi utilizado também o sistema de previsão *soft*, apresentado na Seção 4.1.2.2, visto que tal clusterização trabalha com pertinências parciais, podendo estas serem aplicadas no último sistema de previsão.

Para esta clusterização, os parâmetros descritos na Tabela 19 foram os que apresentaram os resultados mais satisfatórios.

Número de *clusters*	SVM	
	C	ε
14	550	0,0601

Tabela 19: Parâmetros para experimento com clusterização *C-Means*

Assim como na clusterização *K-Means*, a quantidade de *clusters* de baixa e alta volatilidade é definida pelo próprio algoritmo de clusterização. Os padrões de *clusters* de baixa e alta volatilidade utilizados variaram conforme o experimento, produzindo em cada caso padrões distintos, conforme apresentado na Tabela 20.

Quando utilizado o sistema de previsão *hard*, onde cada previsão é realizada exclusivamente por uma única SVM - a que pertence ao *cluster* com maior pertinência do padrão que se deseja prever -, apenas um conjunto de *clusters* foi utilizado pelo sistema em cada experimento, assim como também observado no sistema com clusterização *K-Means*.

Para cada experimento, a quantidade de *clusters* de baixa e alta volatilidade utilizada

Experimento	**Quantidade de *clusters***	
	Baixa	**Alta**
1	4	10
2	6	8
3	2	12
4	5	9
5	2	12
6	6	8
7	11	3
8	1	13
9	9	5
10	7	7

Tabela 20: Quantidade de *clusters* de baixa e alta volatilidade para clusterização *C-Means*

na previsão está apresentada na Tabela 21. Novamente, todos os experimentos utilizaram apenas SVMs de baixa volatilidade, sendo esta decisão de responsabilidade do sistema previsor. Portanto, o mesmo excesso de generalização visto no sistema com clusterização *K-Means* pode ser observado no sistema com clusterização *C-Means*.

Experimento	**Quantidade de *clusters* utilizados**	
	Baixa	**Alta**
1	3	0
2	6	0
3	2	0
4	5	0
5	2	0
6	5	0
7	8	0
8	1	0
9	7	0
10	7	0

Tabela 21: Quantidade de *clusters* de baixa e alta volatilidade utilizada na previsão de cada experimento para clusterização *C-Means*

Novamente, pode-se observar a divisão do conhecimento entre as SVMs, devido às previsões serem realizadas por várias delas na maioria dos experimentos.

Já quando utilizado o sistema de previsão soft, todos os *clusters* atuam no processo de previsão, sendo que o peso da previsão de cada *cluster* e dado pela pertinência do novo padrão de entrada a cada um. Esta pertinência varia a cada previsão, visto que os dados apresentados são diferentes para cada período.

O erro MAPE de cada experimento utilizando os sistemas *hard* e *soft*, comparado ao

erro MAPE utilizando o sistema com clusterização *K-Means*, está apresentado na Tabela 22.

Experimento	***K-Means***	***C-Means Hard***	***C-Means Soft***
1	0,5046	0,5879	0,6583
2	0,6772	0,8142	1,1165
3	1,0332	1,8059	2,5969
4	0,9571	1,0168	1,0368
5	0,9775	1,0829	1,5694
6	1,3543	4,2342	4,2619
7	3,4139	3,7544	4,4896
8	1,9366	2,2313	2,2312
9	6,9022	4,1613	5,1193
10	1,9587	1,6965	2,1330
Erro médio	1,9715	2,1385	2,5213
Desvio padrão	1,9064	1,4119	1,5793

Tabela 22: MAPE dos experimentos utilizando clusterização *C-Means*

De forma geral, ambos os sistemas utilizando a clusterização *C-Means* apresentaram precisão inferior ao sistema com clusterização *K-Means*. O sistema *hard* obteve resultados melhores em dois experimentos e o *soft*, em apenas um. Porém ambos os sistemas *C-Means* apresentaram desvio padrão do erro inferior ao *K-Means*, indicando que apesar do nível de erro ser maior, eles são ligeiramente mais constantes.

Analisando estes números, pode-se concluir que o sistema com clusterização *K-Means* apresenta resultados melhores que os sistemas com clusterização *C-Means*, dado que o ganho obtido no desvio padrão pela clusterização *C-Means* pouco contribui para a estabilidade das previsões.

A comparação entre os sistemas *hard* e *soft* mostra que o sistema *soft* adiciona erro às previsões quando passa a considerar os *clusters* mais distantes como responsáveis também do resultado final. Porém, se comparado o resultado numérico do sistema *soft* com a SVM pura, conclui-se que a utilização do conhecimento de toda a série, aplicando pesos conforme semelhança dos padrões, produz resultados melhores do que considerá-los todos como pertinentes, visto que, no primeiro caso a parcela de contribuição dos padrões menos semelhantes é menor e, consequentemente, é também o erro introduzido.

Traduzindo em números, o erro, quando utilizada a clusterização *K-Means*, é $7,81\%$ menor que quando utilizada a clusterização *C-Means* com o sistema *hard*, e $21,81\%$ com o sistema *soft*.

Quando comparados ambos os sitemas *C-Means* com o método de referência CRB

(cujos erros estão descritos na Tabela 16), é possível constatar que o erro MAPE médio do sistema *hard* é $52,92\%$ menor que o do método de referência, e que o erro do sistema *soft*, $44,48\%$ menor. Assim como no sistema com clusterização *K-Means*, os erros dos sistemas com clusterização *C-Means* apresentaram precisão muito superior ao método de referência.

A Tabela 23 apresenta os erros RMSE utilizados no cálculo do índice U de *Theil* de ambos os sistemas *C-Means*. O cálculo do índice também envolve o RMSE do método de referência CRB, apresentado na Tabela 17. O sistema *C-Means hard* apresentou o valor do índice de $0,5106\%$, enquanto o sistema *C-Means soft*, $0,5996\%$. Tais resultados, semelhantes ao apresentado pelo sistema com clusterização *K-Means*, demonstram a superioridade em desempenho destes sistemas em relação ao sistema de referência.

Experimento	**C-Means hard**	***C-Means soft***
1	0.01128	0.0123
2	0.0303	0.0417
3	0.0745	0.1092
4	0.0704	0.0730
5	0.0783	0.1167
6	0.1732	0.1693
7	0.1933	0.2359
8	0.1914	0.1914
9	0.2788	0.3355
10	0.1307	0.1624
Erro médio	0.1232	0.1447

Tabela 23: RMSE dos experimentos utilizando clusterização *C-Means*

As Figuras de 40 a 49 ilustram as previsões destes experimentos utilizando a clusterização *C-Means* para os sistemas *hard* e *soft*, comparando-as com o valor real da série e também com a previsão utilizando a clusterização *K-Means*.

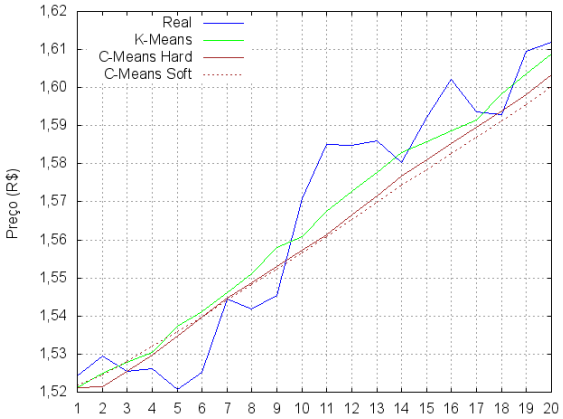

Figura 40: Experimento 1 - *C-Means*

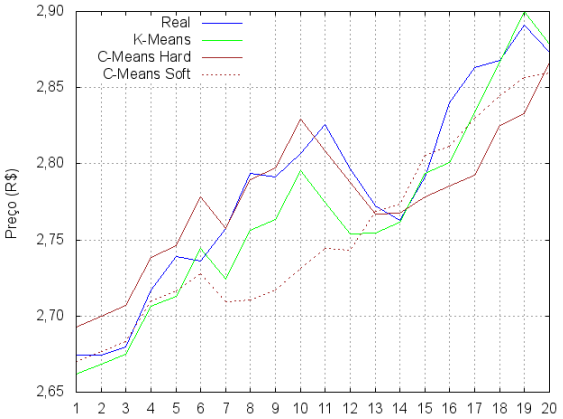

Figura 41: Experimento 2 - *C-Means*

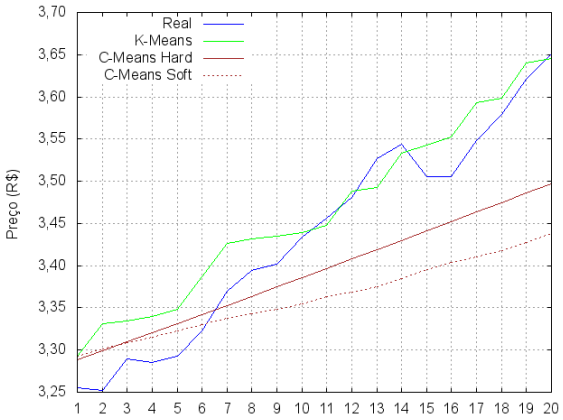

Figura 42: Experimento 3 - *C-Means*

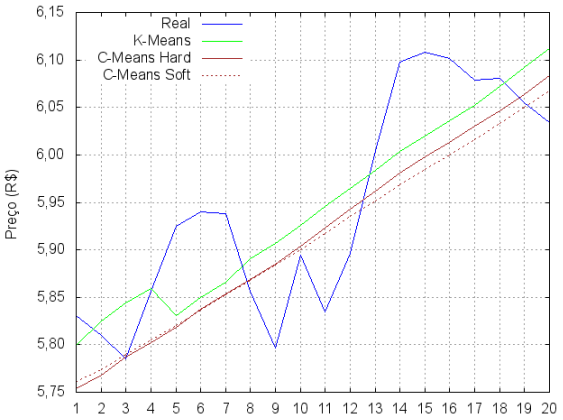

Figura 43: Experimento 4 - *C-Means*

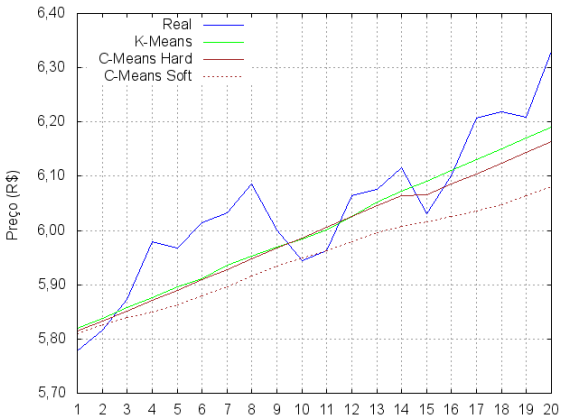

Figura 44: Experimento 5 - *C-Means*

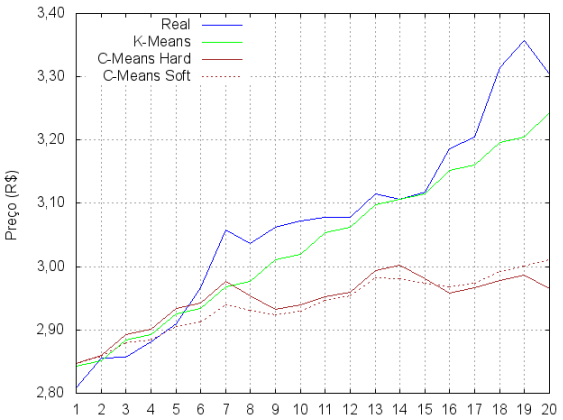

Figura 45: Experimento 6 - *C-Means*

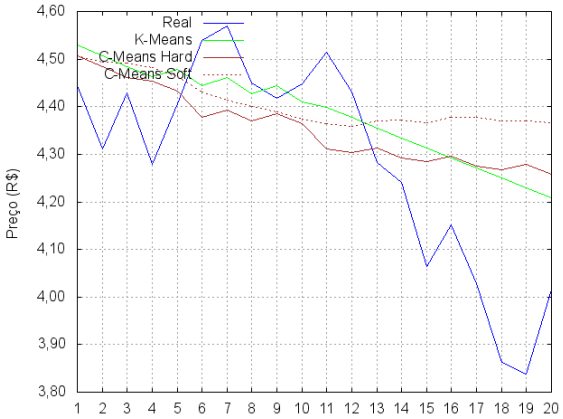

Figura 46: Experimento 7 - *C-Means*

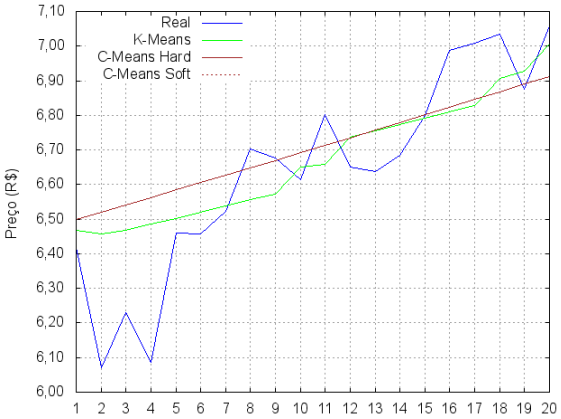

Figura 47: Experimento 8 - *C-Means*

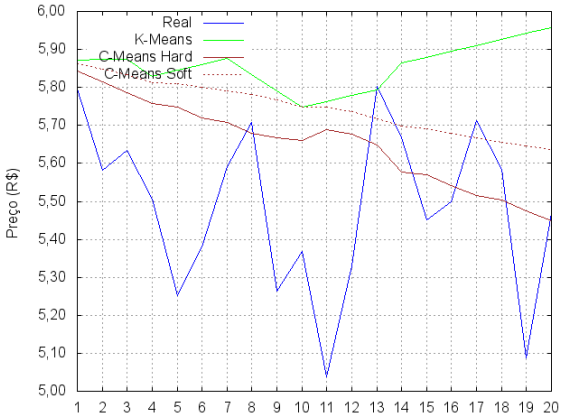

Figura 48: Experimento 9 - *C-Means*

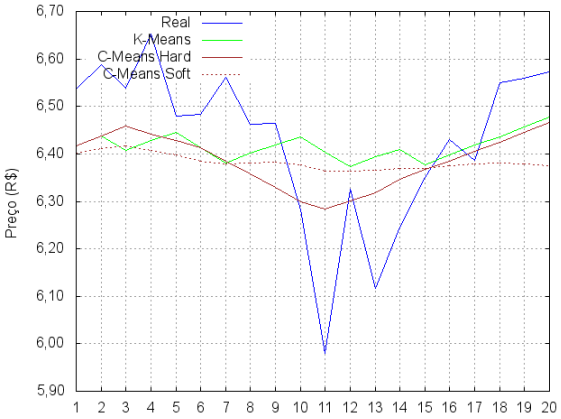

Figura 49: Experimento 10 - *C-Means*

Os gráficos mostram que ambos os sistemas com clusterização *C-Means* acertaram a tendência de todos os experimentos, porém em alguns casos, como nos experimentos 3 e 6 , o resultado encontrado pela clusterização *K-Means* foi mais contundente.

Assim como na clusterização *K-Means*, o excesso de generalização pode ser visto nos resultados de ambos os sistemas com *C-Means*, mostrando que eles são, por muitas vezes, *flats*, não tentando acompanhar o movimento da série ao longo dos dias de previsão.

Também, mais uma vez, apenas *clusters* de baixa volatilidade foram utilizados na previsão, levando a um erro baixo para os experimentos de baixa volatilidade - média de $1,0615\%$ no sistema *hard* e $1,3956\%$ no sistema *soft* -, e consideravelmente mais alto para alta volatilidade - média de $3,2155\%$ no sistema *hard* e $3,6470\%$ no sistema *soft*.

Apesar de o erro ser mais elevado nos experimentos de alta volatilidade, as informações de baixa volatilidade se mostraram capazes de generalizar os resultados de alta volatilidade

com certa precisão, corroborando com os resultados obtidos na clusterização *K-Means*.

Assim, em vista dos resultados obtidos serem muito semelhantes aos da clusterização *K-Means*, pode-se generalizar as conclusões dos sistemas em separado para os sistemas clusterizados. Ambos os sistemas, com *K-Means* e *C-Means*, levaram a resultados mais precisos a partir do processo de separar o conhecimento em grupos específicos.

A análise *cluster* a *cluster* para os sistemas *hard* e *soft* estão representadas nas Figuras de 50 a 59.

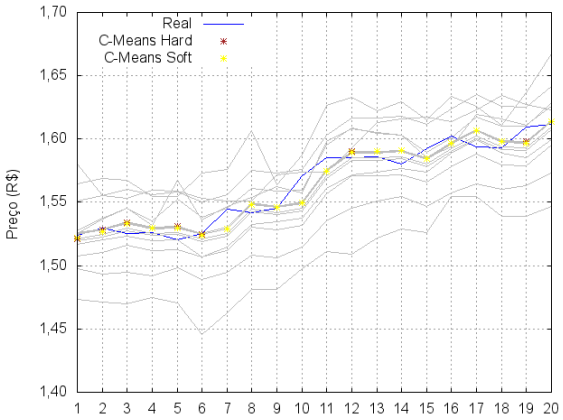

Figura 50: Experimento 1 - Análise de *clusters C-Means*

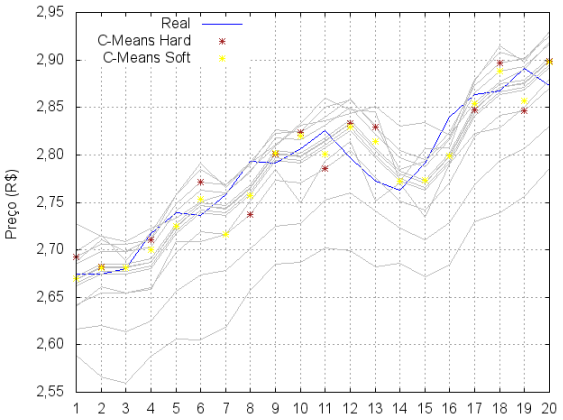

Figura 51: Experimento 2 - Análise de *clusters C-Means*

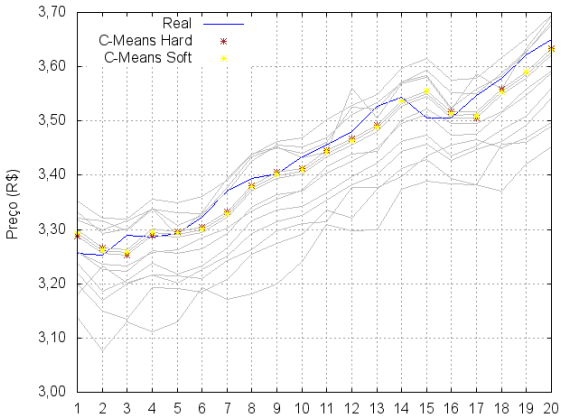

Figura 52: Experimento 3 - Análise de *clusters C-Means*

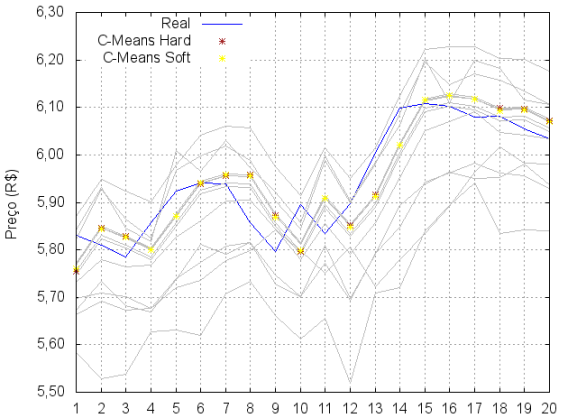

Figura 53: Experimento 4 - Análise de *clusters C-Means*

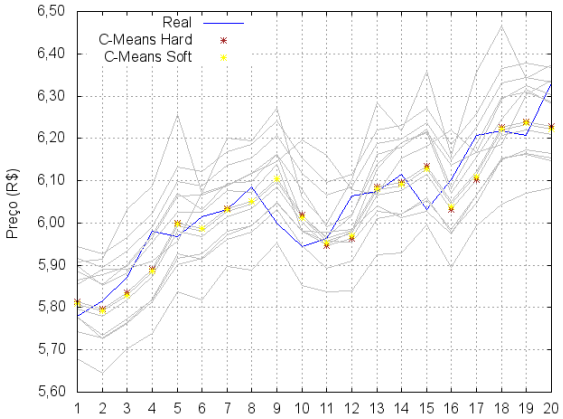

Figura 54: Experimento 5 - Análise de *clusters C-Means*

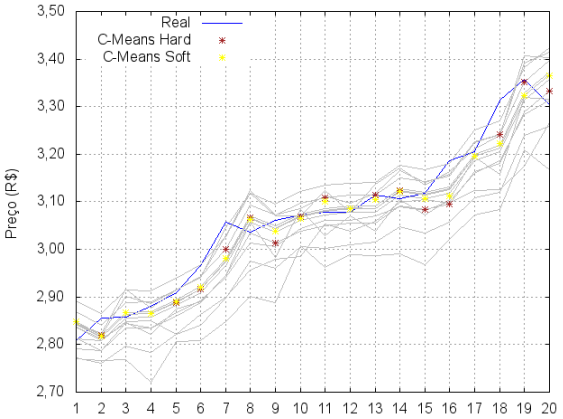

Figura 55: Experimento 6 - Análise de *clusters C-Means*

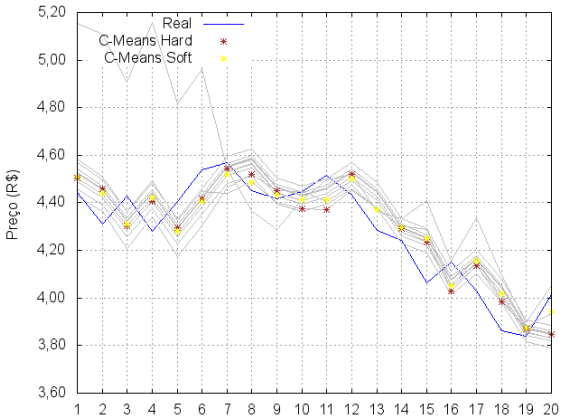

Figura 56: Experimento 7 - Análise de *clusters C-Means*

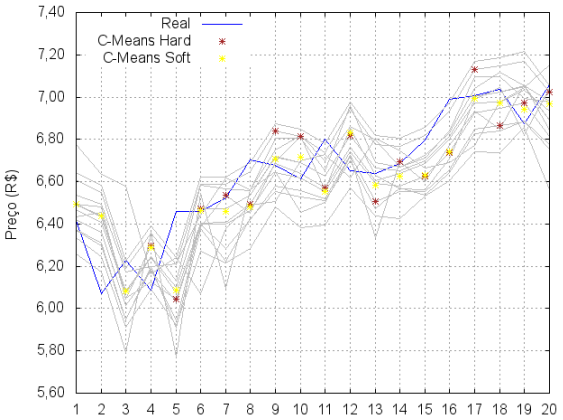

Figura 57: Experimento 8 - Análise de *clusters C-Means*

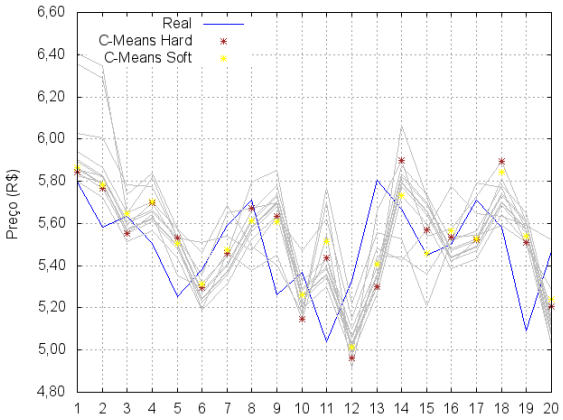

Figura 58: Experimento 9 - Análise de *clusters C-Means*

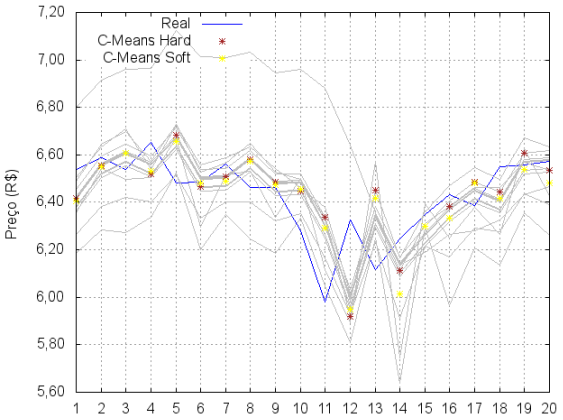

Figura 59: Experimento 10 - Análise de *clusters C-Means*

Assim como visto anteriormente, diversos *clusters* são capazes de realizar previsões aceitáveis da série, sendo a escolha de qual *cluster* prover o resultado (ou prover o principal do resultado, no caso da sistema *soft*) nem sempre ser a melhor.

Em diversos experimentos é possível notar que o resultado provido pelo sistema *soft* é muito semelhante à escolha do sistema *hard*, mostrando alta pertinência do padrão a ser previsto a um único *cluster*. Em outros, onde o espaçamento entre os resultados é maior, há uma divisão entre resultados melhores para cada sistema.

A Tabela 24 mostra, para o sistema C-Means *hard*, a quantidade de escolhas corretas por experimento. Para o sistema C-Means soft, mostra a o número de vezes que sua composição de preço foi melhor que os resultados de todas as SVMs individualmente. A taxa de acerto e o erro médio foi semelhante entre os dois sistemas, corroborando com a análise gráfica anterior.

Experimento	***C-Means Hard***		***C-Means Soft***	
	Escolhas corretas	**MAPE**	**Escolhas corretas**	**MAPE**
1	1	0,4665	1	0,4674
2	1	0,9363	1	0,7484
3	5	0,6081	2	0,6278
4	0	0,8541	1	0,8359
5	3	0,8414	1	0,8170
6	2	0,9890	0	1,0289
7	0	2,3388	1	2,1718
8	3	2,4400	1	2,1275
9	1	4,1261	2	3,6249
10	0	2,0229	2	2,0604
Taxa de acerto / Erro médio	8%	1,5623	6%	1,4510

Tabela 24: Acerto na escolha de *cluster* - *C-Means*

5.5 Explorando a alta volatilidade

Como visto nas Seções 5.3 e 5.4, a execução dos sistemas previsores descritos neste estudo, em sua forma natural, leva à utilização de apenas *clusters* de baixa volatilidade. Apesar destes *clusters* terem se mostrado capazes de generalizar, com razoável precisão, previsões de alta volatilidade, os números e os gráficos mostram que as maiores responsáveis pelo valor do erro médio e do desvio padrão são estas previsões.

Levando isso em consideração, uma nova abordagem é aqui proposta visando melhorar

os resultados de alta volatilidade, mesmo que ao custo de perda de precisão com baixa volatilidade.

O novo modelo proposto considera apenas os dados de alta volatilidade, descartando as informações de baixa volatilidade. O oposto foi realizado anteriormente, de forma natural, pelas características do sistema. Neste novo modelo, a escolha dos dados de alta volatilidade é uma imposição.

No treinamento e construção do sistema de previsão, descrito na Seção 4.1.1, as únicas modificações estão na criação de padrões e na volatilidade. A primeira foi alterada para inserir na tabela de decisão apenas atributos de decisão de alta volatilidade, ignorando os de baixa volatilidade; os atributos de condição de baixa volatilidade continuam fazendo parte da tabela, quando existentes como condição de um atributo de decisão de alta volatilidade. Já a informação de volatilidade foi completamente removida, visto que o sistema trabalhará com apenas uma volatilidade; deste modo, a tabela de decisão passa a ter uma coluna a menos, tendo o aspecto da Tabela 10.

A tabela de decisão deste sistema tem apenas 422 linhas, contra 1868 linhas do sistema original. Ou seja, o sistema utiliza apenas 22% das informações disponíveis.

Para esta nova abordagem, a quantidade de *clusters* e os parâmetros das SVMs estão descritos na Tabela 25.

Clusterização	**Número de *clusters***	**SVM**	
		C	ε
K-Means	11	550	0,0601
C-Means	13	550	0,0601

Tabela 25: Parâmetros para experimento com alta volatilidade

A Tabela 26 apresenta os erros de cada experimento para ambas as clusterizações, comparando-os com o resultado obtido pelo sistema convencional quando utilizada a clusterização *K-Means*.

Para ambos os casos, é possível notar uma ligeira deterioração do resultado em quase todos os experimentos de baixa volatilidade. Enquanto o sistema convencional com clusterização *K-Means* apresentou erro médio de $0,8299\%$ nestes experimentos, os sistemas de alta volatilidade com *K-Means* e *C-Means* apresentaram $0,9903\%$ e $1,0119\%$, respectivamente.

Em contrapartida, houve uma melhora significativa nos experimentos de alta volatilidade, sendo $3,1131\%$ para o sistema convencional com clusterização *K-Means*, e $2,7197\%$ e

Experimento	*K-Means*	*K-Means HighV*	*C-Means HighV*
1	0,5046	0,9301	0,9979
2	0,6772	1,0661	0,8290
3	1,0332	1,1647	1,0876
4	0,9571	0,9550	0,8242
5	0,9775	0,8356	1,3206
6	1,3543	2,2751	1,6690
7	3,4139	3,1603	2,1395
8	1,9366	2,9599	3,1074
9	6,9022	3,2326	3,0539
10	1,9587	1,9705	1,7730
Erro médio	1,9715	1,8550	1,6802
Desvio padrão	1,9064	0,9897	0,8529

Tabela 26: MAPE dos experimentos utilizando alta volatilidade

$2,3486\%$ para os sistemas de alta volatilidade com *K-Means* e *C-Means*, respectivamente.

Pode-se constatar também que os resultados ficaram mais homogêneos pela diminuição substancial do desvio padrão do erro. No primeiro sistema, a diferença entre o menor e maior erro é de $6,3976$, enquanto nos dois seguintes as diferenças são de $2,3970$ e $2,2784$.

Dados os erros semelhantes obtidos entre os sistemas de alta volatilidade e dois sistemas *hard* de baixa volatilidade, por comparação indireta, pode-se concluir que ambos os modelos de alta volatilidade apresentaram precisão superior ao modelo de referência CRB.

Já as Figuras de 60 a 69 mostram as previsões da Tabela 26, comparando-as com a série real.

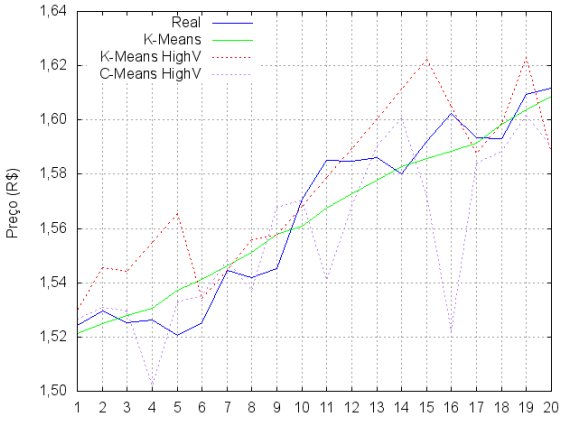

Figura 60: Experimento 1 - Alta volatilidade

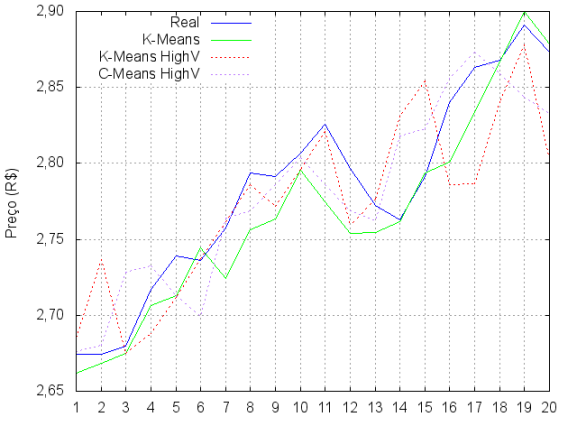

Figura 61: Experimento 2 - Alta volatilidade

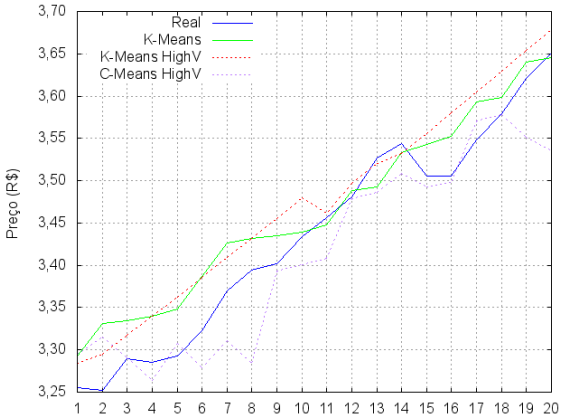

Figura 62: Experimento 3 - Alta volatilidade

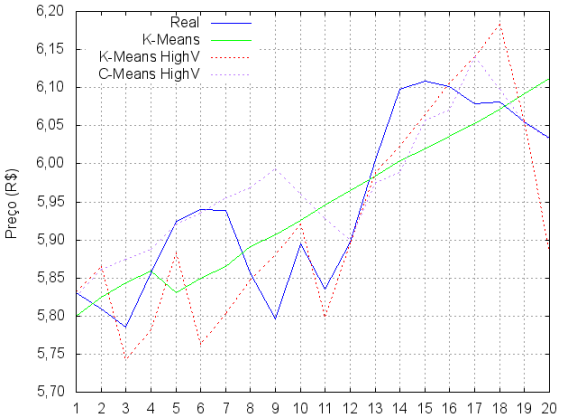

Figura 63: Experimento 4 - Alta volatilidade

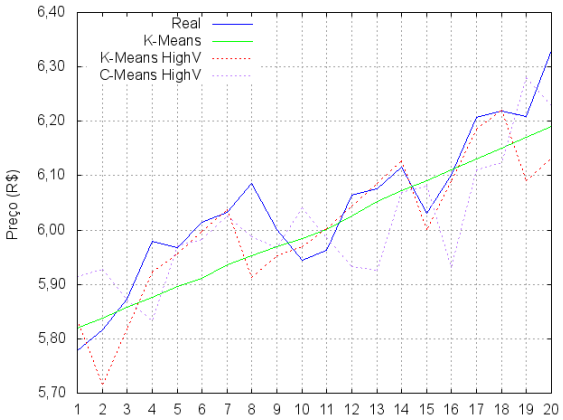

Figura 64: Experimento 5 - Alta volatilidade

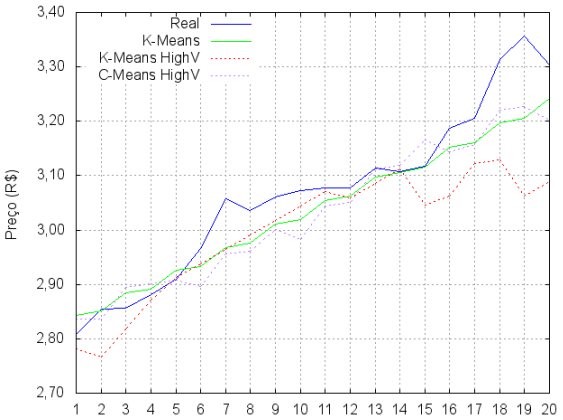

Figura 65: Experimento 6 - Alta volatilidade

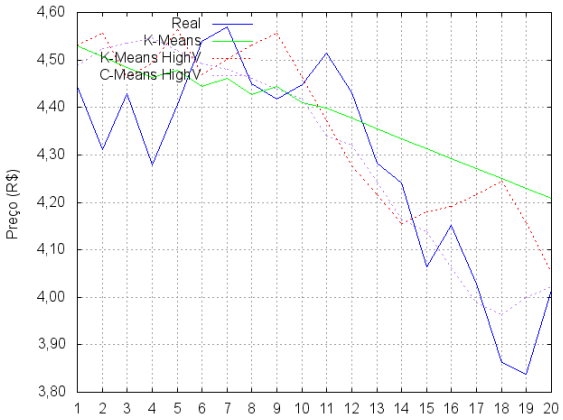

Figura 66: Experimento 7 - Alta volatilidade

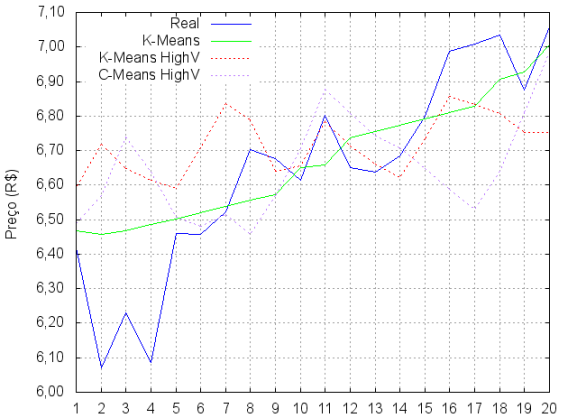

Figura 67: Experimento 8 - Alta volatilidade

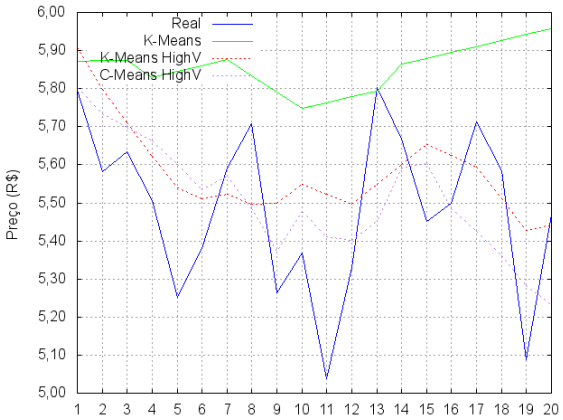

Figura 68: Experimento 9 - Alta volatilidade

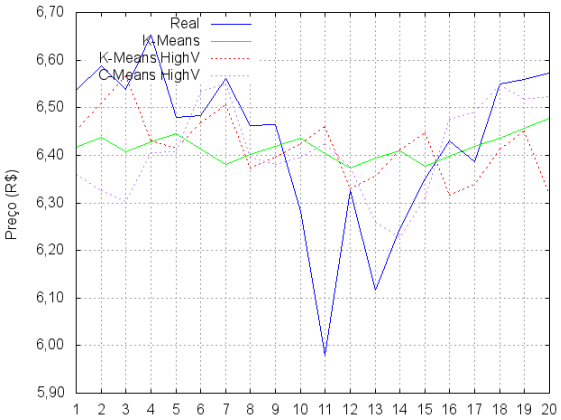

Figura 69: Experimento 10 - Alta volatilidade

É notável a diferença de comportamento da série prevista do sistema convencional, comparada aos sistemas de alta volatilidade. Enquanto a primeira é *flat* na maioria das vezes, nos outros dois sistemas o resultado busca acompanhar o comportamento real da série, deixando de lado o excesso de generalização visto anteriormente.

Os gráficos mostram que as previsões dos experimentos de baixa volatilidade são muito boas, embora piores que as obtidas com o sistema K-Means que utilizarou apenas os dados de baixa volatilidade, e que as previsões dos experimentos de alta volatilidade melhoraram significativamente, comparadas a este mesmo sistema. Para ambos os sistemas de alta volatilidade, houve acerto de tendência em todos os experimentos.

Os gráficos e os resultados numéricos demonstram que as informações de alta volatilidade são capazes de generalizar padrões de baixa volatilidade com boa precisão, com qualidade muito próxima ao sistema convencional, sendo o ganho maior, o apresentado nos experimentos de alta volatilidade.

As Figuras de 70 a 79 apresentam a análise *cluster* a *cluster* para o sistema de alta volatilidade *K-Means*, e as Figuras de 80 a 89, para *C-Means*.

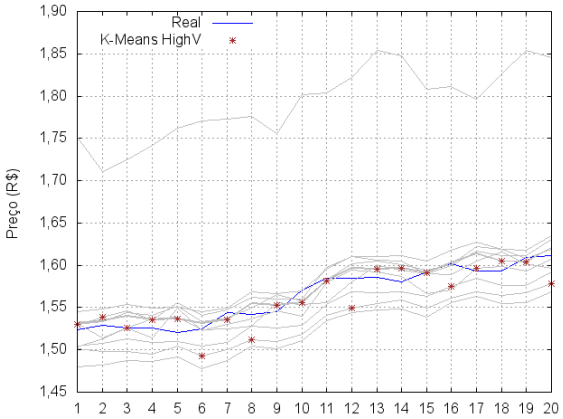

Figura 70: Experimento 1 - Análise de *clusters K-Means* de alta volatilidade

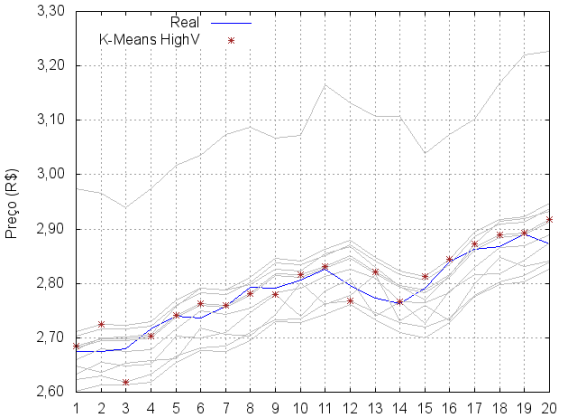

Figura 71: Experimento 2 - Análise de *clusters K-Means* de alta volatilidade

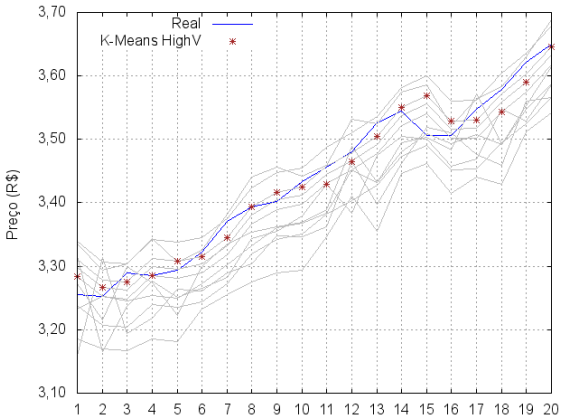

Figura 72: Experimento 3 - Análise de *clusters K-Means* de alta volatilidade

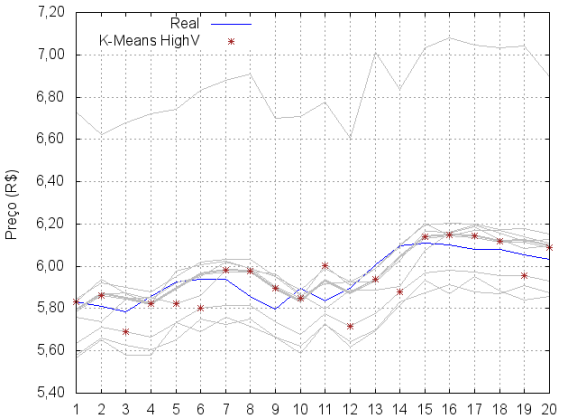

Figura 73: Experimento 4 - Análise de *clusters K-Means* de alta volatilidade

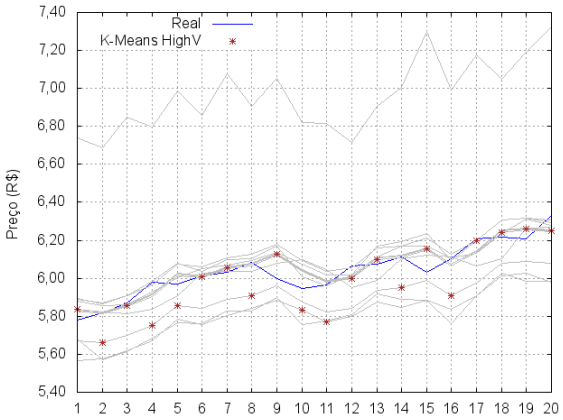

Figura 74: Experimento 5 - Análise de *clusters K-Means* de alta volatilidade

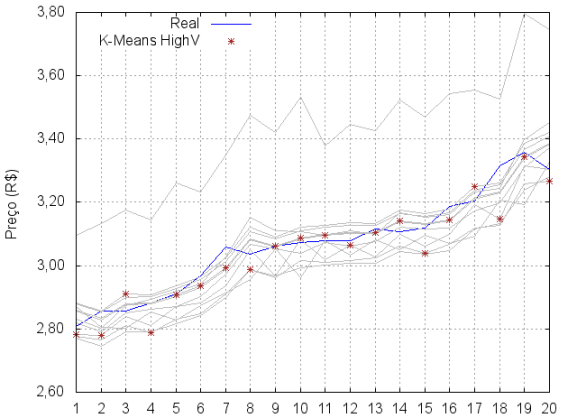

Figura 75: Experimento 6 - Análise de *clusters K-Means* de alta volatilidade

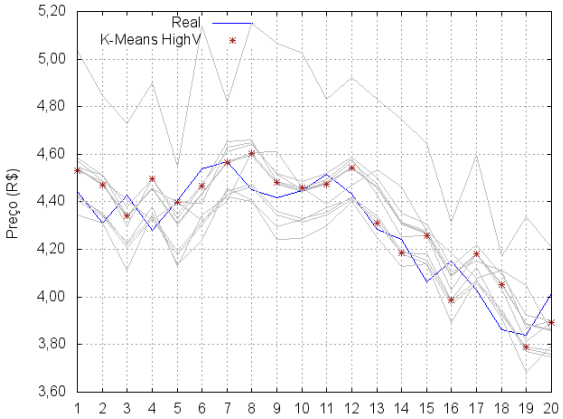

Figura 76: Experimento 7 - Análise de *clusters K-Means* de alta volatilidade

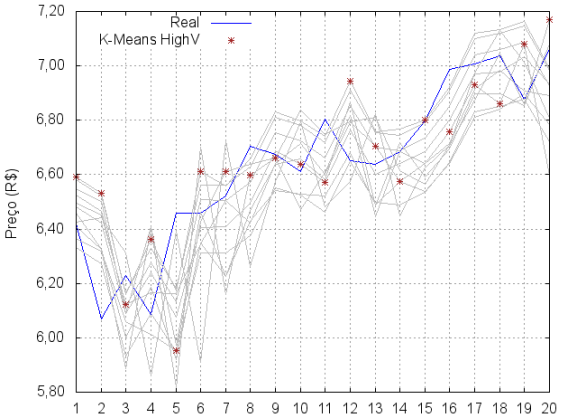

Figura 77: Experimento 8 - Análise de *clusters K-Means* de alta volatilidade

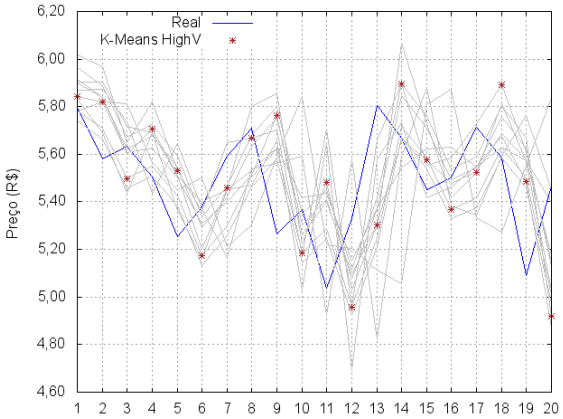

Figura 78: Experimento 9 - Análise de *clusters K-Means* de alta volatilidade

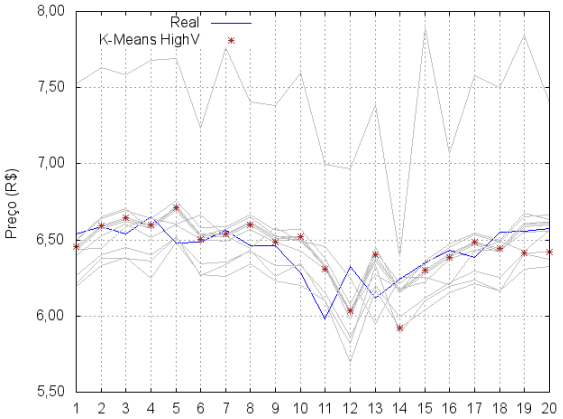

Figura 79: Experimento 10 - Análise de *clusters K-Means* de alta volatilidade

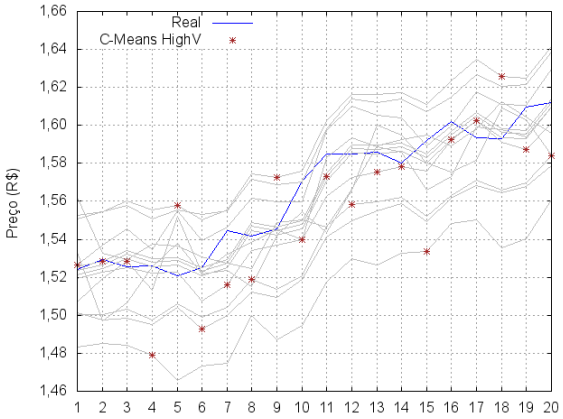

Figura 80: Experimento 1 - Análise de *clusters C-Means* de alta volatilidade

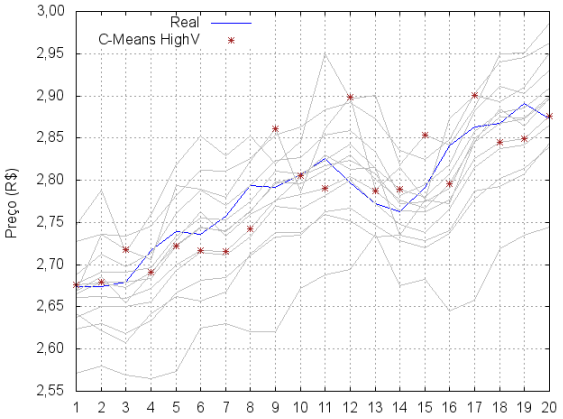

Figura 81: Experimento 2 - Análise de *clusters C-Means* de alta volatilidade

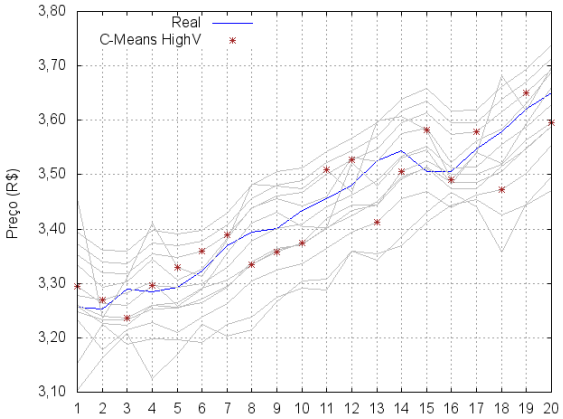

Figura 82: Experimento 3 - Análise de *clusters C-Means* de alta volatilidade

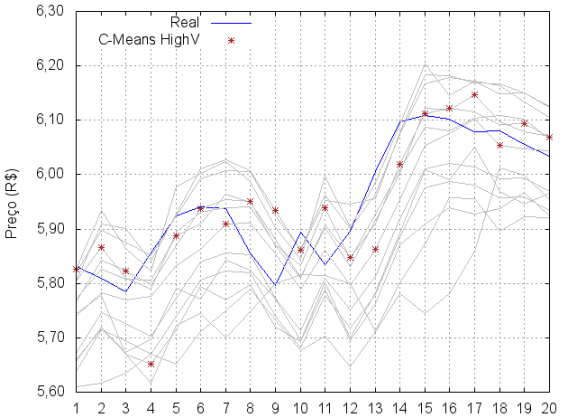

Figura 83: Experimento 4 - Análise de *clusters C-Means* de alta volatilidade

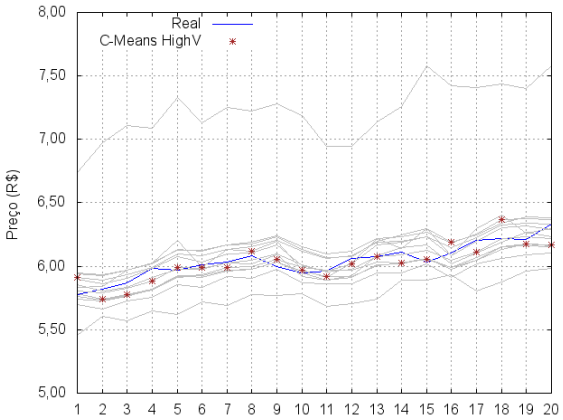

Figura 84: Experimento 5 - Análise de *clusters C-Means* de alta volatilidade

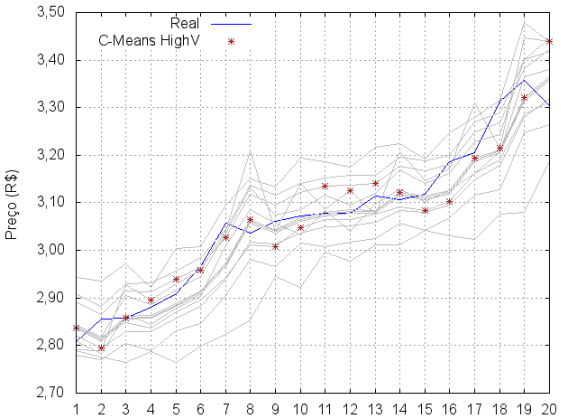

Figura 85: Experimento 6 - Análise de *clusters C-Means* de alta volatilidade

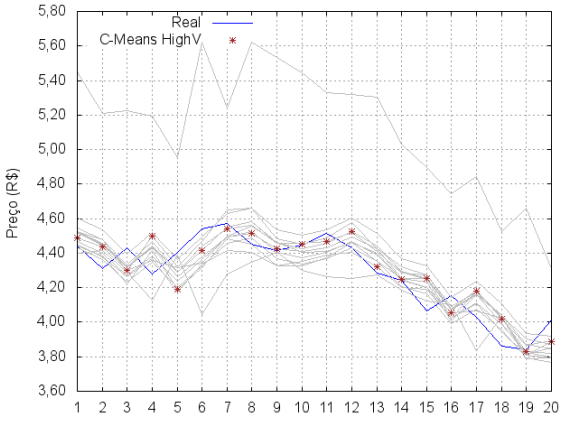

Figura 86: Experimento 7 - Análise de *clusters C-Means* de alta volatilidade

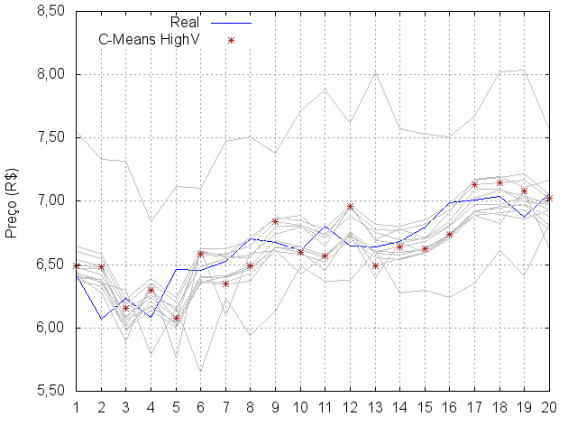

Figura 87: Experimento 8 - Análise de *clusters C-Means* de alta volatilidade

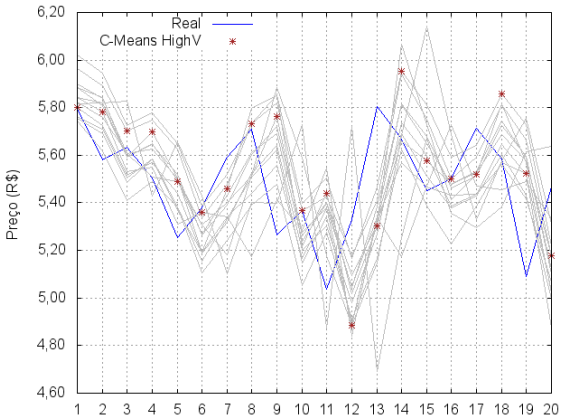

Figura 88: Experimento 9 - Análise de *clusters C-Means* de alta volatilidade

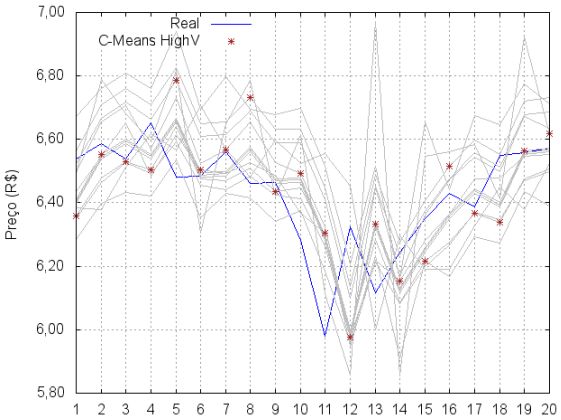

Figura 89: Experimento 10 - Análise de *clusters C-Means* de alta volatilidade

Assim como nas análises anteriores, para ambos os casos, vários *clusters* são capazes de prover previsões aceitáveis da série, sendo a escolha por proximidade ou pertinência nem sempre a melhor opção. Tal fato pode também ser visualizado na Tabela 27.

Experimento	***K-Means HighV***		***C-Means HighV***	
	Escolhas corretas	**MAPE**	**Escolhas corretas**	**MAPE**
1	2	0,8930	2	1,4169
2	2	0,7046	2	1,1871
3	4	0,5427	1	1,3653
4	3	1,4335	3	1,0193
5	2	1,6212	1	1,1090
6	4	1,4169	2	1,3156
7	1	2,3340	2	2,2183
8	2	2,6145	0	2,6292
9	0	4,8015	4	3,9996
10	2	2,1796	3	2,1201
Taxa de acerto / Erro médio	11%	1,8542	10%	1,8380

Tabela 27: Acerto na escolha de *cluster* - Alta volatilidade

5.6 Comparação com modelos existentes

Conforme já mencionado anteriormente, Leite (2010) apresenta um estudo de previsão de séries temporais utilizando uma arquitetura hierárquica composta por um mapa auto-organizável (SOM) e uma SVM. A finalidade de se utilizar esta arquitetura é trabalhar

os dados de entrada, de forma a extrair suas características estatísticas mais importantes através da SOM, para posterior transferência para a SVM.

Dois modelos foram propostos por Leite (2010): o primeiro chamado HNM, composto por uma SOM seguida de uma SVM, e o segundo, chamado HNM-V, que possui a mesma arquitetura do modelo HNM, porém adiciona informação de volatilidade aos dados.

A Tabela 28 apresenta a comparação dos resultados obtidos por Leite (2010) com alguns dos resultados descritos no presente estudo.

Experimento	**HNM**	**HNM-V**	***K-Means***	***C-Means Hard***	***C-Means HighV***
1	4,1581	1,2434	0,5046	0,5879	0,9979
2	0,9902	3,2745	0,6772	0,8142	0,8290
3	5,8239	1,6491	1,0332	1,8059	1,0876
4	1,9167	1,4496	0,9571	1,0168	0,8242
5	2,5711	1,8358	0,9775	1,0829	1,3206
6	7,1402	4,0869	1,3543	4,2342	1,6690
7	5,0196	5,3138	3,4139	3,7544	2,1395
8	2,1515	3,7272	1,9366	2,2313	3,1074
9	6,2069	5,4878	6,9022	4,1613	3,0539
10	1,9342	4,9700	1,9587	1,6965	1,7730
Erro médio	3,7912	3,3038	1,9715	2,1385	1,6802
Desvio padrão	2,1554	1,6652	1,9064	1,4119	0,8529

Tabela 28: Comparação com resultados alcançados por Leite (2010)

Os resultados obtidos pelos sistemas que aplicam clusterização são superiores a ambos os modelos hierárquicos estabelecidos. Se comparados o melhor resultado entre os modelos hierárquicos — HNM-V — e o melhor entres os modelos clusterizados — *C-Means HighV* —, pode-se constatar que o segundo apresenta erro menor em todos os experimentos, sendo que, no geral, apresenta erro 49% menor e desvio padrão 48%, também menor.

Os sistemas utilizados na comparação possuem características próprias, que não são observadas nos modelos hierárquicos citados. Durante o treinamento dos elementos previsores, apenas as informações do *cluster* de uma dada SVM são utilizadas, o que significa que informações não relacionadas aos padrões deste *cluster* não influenciam no conhecimento da SVM. No processo de previsão, novamente, apenas o conhecimento relativo aos padrões mais semelhantes ao novo padrão de entrada é utilizado, sendo conhecimentos de outros padrões descorrelacionados não considerados.

Apesar da semelhança entre os modelos, visto que o modelo hierárquico constroi a

informação a ser repassada para a SVM através da clusterização dos dados de entrada por contexto utilizando da camada SOM, o sistema apresentado neste estudo, por excluir do processo de treinamento e previsão informações não pertinentes ao padrão a ser previsto, não é susceptível a estas informações, tornando-o consequentemente mais preciso.

6 Conclusão

6.1 Discussão dos resultados e considerações finais

A previsão de séries financeiras baseada em modelos construídos a partir da própria série temporal é objeto de estudo de diversos pesquisadores, tendo sido muitos os trabalhos desenvolvidos neste ramo. Com o propósito de auxiliar a tomada de decisão de investidores e de operadores de mercado, um sistema de previsão deve ser preciso e confiável. Além disso, é desejável que sua construção seja simples ou, pelo menos, que não seja complexa ao ponto de torná-lo inviável. Neste quesito, o tempo e os recursos dispensados para obtenção do modelo são primordiais.

As séries financeiras são, de longe, das mais difíceis de modelar e prever. Não só a oferta e a procura ditam o comportamento do preço de um ativo, mas também fatos como rumores sobre negócios por trás dos preços, condições econômicas locais e globais, notícias e informações de momento, fatores externos que influenciem, direta ou indiretamente, o ramo de negócios que gera o preço, entre muitos outros. Desta forma, o preço é composto por uma somatória de itens complexos, que nem sempre são visíveis ou acessíveis ao investidor.

Tudo isso faz com que uma série financeira tenha um comportamento errático quando analisado sem as ferramentas corretas. Neste ponto, ter um sistema capaz de extrair as informações importantes e necessárias de uma série temporal, que seja capaz de aprender sobre seu comportamento com base nessas informações, e abstrair todas as informações inerentes à série de forma a tornar a processo decisório do investidor mais simples e direto, é fundamental.

Um número expressivo de estudos foi desenvolvido com este propósito, começando com modelos matemáticos lineares e, posteriormente, modelos matemáticos não lineares. É particularmente interessante trabalhar com modelos não lineares, visto que séries financeiras são caracterizadas, entre outras coisas, por possuírem não linearidades, como

os períodos de volatilidade distinta e os movimentos em direções opostas.

Em um passado não muito recente, sistemas baseados em inteligência artificial passaram a ser utilizados como modelo para as séries financeiras. Redes neurais artificiais, por suas características atrativas, como ser auto adaptativa, ser capaz de captar relações funcionais da série e poder generalizar a partir do conhecimento adquirido, tornaram-se alternativas para a solução do problema de previsão. Apesar dos bons resultados apresentados, as redes neurais possuem deficiências, como o alto número de parâmetros livres - sendo que não existem métodos para obtenção de seus valores ótimos -, e os resultados inconstantes - em virtude dos ótimos locais.

Máquinas de vetor de suporte (SVMs) começaram a ser empregadas para este fim, tornando-se logo uma alternativa promissora. Uma das grandes vantagens da SVM é a capacidade de encontrar sempre soluções ótimas, fugindo da instabilidade observada nas redes neurais artificiais. Vários estudos compararam a capacidade de previsão de modelos baseados em SVM com modelos baseados em redes neurais, tendo os primeiros se mostrado superiores.

A partir destes estudos, os sistemas evoluíram de forma a tratar a série antes de apresentá-la ao elemento previsor. O tratamento consiste em separar a série em seus diversos contextos, ou adicionar informações sobre os contextos nos padrões apresentados ao elemento previsor adiante - redes neurais ou SVMs. Neste primeiro estágio de tratamento, é comum a presença de SOMs e de algoritmos genéticos. Estes sistemas apresentaram resultados mais precisos quando comparados aos que não tratam os contextos, mostrando ser este um caminho produtivo na obtenção de melhores resultados.

A complexidade é fator a ser considerado na montagem destes sistemas construídos em estágios. Muitos são os parâmetros livres do elemento previsor, e que já demandam certo esforço em sua construção. Neste contexto, a inserção de SOMs, ou de outros métodos que precisam ser treinados e aprimorados, mostra-se ser um processo complexo extra adicionado ao sistema. Apesar dos resultados positivos obtidos pelos estudos, a construção e treinamento de uma SOM, passando pelo ajuste de seus parâmetros livres, é um processo que demanda tempo e recursos.

Além disso, é interessante tratar os contextos da série de forma palpável, de maneira que possam ser analisados individualmente em uma etapa posterior, fato este não possível quando ocorre transformação dos padrões de entrada da série.

Assim, este estudo propôs utilizar métodos de clusterização como processamento pré-

vio da série temporal. O propósito é separar a série em seus diversos contextos, e depois construir modelos específicos, um para cada contexto. Novos padrões de entrada podem ser previstos apenas pelo contexto mais relacionado a si, ou ainda podem ser uma composição ponderada de todos os contextos.

Não menos importante, não há transformação dos padrões de entrada. Eles são simplesmente separados de acordo com a distância entre si, podendo então haver análise individual dos diversos *clusters* formados a partir da série, o que possibilita uma visão mais profunda do processo de previsão executado pelo sistema.

Além do mais, clusterizar um grupo de dados é um processo mais simples e rápido que construir e treinar uma SOM. A clusterização possui apenas um parâmetro livre: a quantidade de *clusters*. Não havendo método para a sua determinação, a escolha deste número foi feita baseada no melhor resultado.

Os resultados deste estudo foram comparados com os apresentados pelos modelos hierárquicos HNM e HNM-V, sendo ambos compostos de uma SOM, para extração de contexto, seguido de SVMs, sendo que o segundo trata volatilidades distintas de forma separada. Para todos os experimentos, o resultado do modelo clusterizado mostrou ser mais preciso e confiável, tendo inclusive acertado todas as tendências em alguns dos casos, fato que não ocorreu em nenhum dos modelos hierárquicos.

Quando comparado a um modelo simples, sem tratamento de contexto, ou seja, apenas uma SVM, a diferença de precisão foi ainda maior.

Ao comparar o resultado dos três modelos (simples, hierárquico e clusterizado), é possível concluir que o tratamento de contexto agrega precisão e confiabilidade ao sistema previsor, fato já demonstrado em outros estudos. Já os métodos de clusterização mostraram-se superiores, no que tange a separação da série em contextos, em dois aspectos: além de a precisão ter aumentado significativamente quando utilizada a clusterização, sua complexidade é menor, o que exige menores tempo e recursos computacionais.

6.2 Sugestões para novos trabalhos

Apesar dos resultados do estudo estarem de acordo com o esperado - a segmentação da série de acordo com contextos intrínsecos permite melhor absorção de conhecimento e de generalização por parte da SVM -, e serem positivos - a precisão apresentada foi superior do que modelos preestabelecidos -, há ainda espaço para melhorias.

A análise *cluster* a *cluster* feita nos experimentos aponta dois pontos que podem ser atacados. O primeiro é que diversos *clusters* são capazes de fazer previsões razoáveis da série - é possível ver que, na previsão um passo a frente, vários deles acompanham o movimento da série. O segundo é que a escolha do *cluster* responsável pela previsão nem sempre é ótima - em muitos casos, *clusters* com resultados mais próximos do real não foram os escolhidos para realizar a previsão.

Ambos os pontos passam por um mesmo aspecto que pode ser estudado, que é a métrica de distância entre os padrões. A distância utilizada no estudo foi a distância euclidiana quadrática que, apesar de se mostrar suficiente para levar a resultados aceitáveis, não capta fielmente a diferença entre os padrões formados a partir da série.

Um estudo mais aprofundado sobre como correlacionar estes padrões, de forma a tornar a comparação entre eles mais fiel aos seus comportamentos, poderá levar a melhor identificação e separação de contextos, o que deverá aumentar ainda mais a precisão das previsões.

Além disso, aspectos mais genéricos como tratamento dos parâmetros livres, tanto os da SVM quanto a quantidade ideal de *clusters*, também merecem atenção, visto que a utilização de métodos que propiciem a obtenção de seus valores ótimos tornaria o processo de construção do sistema mais simples e rápido.

Referências

ABU-MOSTAFA, Y. S.; ATIYA, A. F. Introduction to financial forecasting. *Applied Intelligence*, v. 6, n. 3, p. 205–213, 1996.

ANDERSON, D. R. et al. An introduction to management science: Quantitative approaches to decision making. *International Journal of Forecasting*, v. 8, n. 1, p. 69–80, 2016.

ARMANO, G.; MARCHESI, M.; MURRU, A. A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences*, v. 170, n. 1, p. 3–33, 2005. Computational Intelligence in Economics and Finance.

ARMSTRONG, J.; COLLOPY, F. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, v. 8, n. 1, p. 69–80, 1992.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, v. 10, n. 2–3, p. 191–203, 1984.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis, Forecasting and Control*. Hoboken, New Jersey: John Wiley & Sons, 2008.

CAMPBELL, J. Y.; LO, A. W.; MACKINLAY, A. C. *The Econometrics of Financial Markets*. [S.l.]: Cengage Learning, 2016.

CAO, L. Support vector machines experts for time series forecasting. *Neurocomputing*, v. 51, p. 321–339, 2003.

CAO, L.; TAY, F. E. H. Application of support vector machines in financial time series forecasting. *Omega*, v. 29, n. 4, p. 309–317, 2001.

CAO, L.; TAY, F. E. H. Improved financial time series forecasting by combining support vector machines with self-organizing feature map. *Intelligent Data Analysis*, IOS Press, v. 5, n. 4, p. 339–354, Sep 2001.

CAO, L.; TAY, F. E. H. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on Neural Networks*, v. 14, n. 6, p. 1506–1518, Nov 2003.

CARPINTEIRO, O. A. S. et al. Forecasting models for prediction in time series. *Artificial Intelligence Review*, v. 38, n. 2, p. 163–171, 2012.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, n. 3, p. 27:1–27:27, 2011. Library available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

CLEMENTS, M. P.; FRANSES, P. H.; SWANSON, N. R. Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, v. 20, n. 2, p. 169–183, 2004.

EVERITT, B. S. et al. *Cluster Analysis*. [S.l.]: John Wiley & Sons, 2011.

GOOIJER, J. G. D.; HYNDMAN, R. J. 25 years of time series forecasting. *International Journal of Forecasting*, v. 22, n. 3, p. 443–473, 2006.

HAYKIN, S. *Neural Networks and Learning Machines*. [S.l.]: Prentice Hall, 2009.

HUANG, W.; NAKAMORI, Y.; WANG, S. Y. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, v. 32, n. 10, p. 2513–2522, 2005.

LEITE, J. P. R. R. *Aplicação de Modelos Neurais na Previsão de Séries Temporais*. Dissertação (Mestrado) — Universidade Federal de Itajubá, 2010.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.: s.n.], 1967.

MAKRIDAKIS, S.; WHELLWRIGHT, S. C.; HYNDMAN, R. J. *Forecasting: Methods and Appplications*. [S.l.]: John Wiley & Sons, 1998.

MELO, B. d.; MILIONI, A. Z.; JUNIOR, C. L. N. Daily and monthly sugar price forecasting using the mixture of local expert models. *Pesquisa Operacional*, v. 27, p. 235–246, 2007.

MELO, B. de. *Previsão de Séries Temporais Usando Modelos de Composição de Especialistas Locais*. Dissertação (Mestrado) — Instituto Tecnológico de Aeronáutica, 2003.

MERCER, J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, The Royal Society, v. 209, n. 441–458, p. 415–446, 1909.

MUKHERJEE, S.; OSUNA, E.; GIROSI, F. Nonlinear prediction of chaotic time series using support vector machines. In: *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*. [S.l.: s.n.], 1997. p. 511–520.

MULLER, K.-R. et al. Predicting time series with support vector machines. In: *Artificial Neural Networks - ICANN97*. [S.l.]: Springer Berlin Heidelberg, 1997, (Lecture Notes in Computer Science, v. 1327). p. 999–1004.

MULLER, K.-R. et al. Advances in kernel methods. In: . [S.l.]: MIT Press, 1999. cap. Using Support Vector Machines for Time Series Prediction, p. 243–253.

OLIVEIRA, J. V.; PEDRYCZ, W. *Advances in Fuzzy Clustering and its Applications*. [S.l.]: John Wiley & Sons, 2007.

PANAPAKIDIS, I. P.; DAGOUMAS, A. S. Day-ahead electricity price forecasting via the application of artificial neural network based models. *Applied Energy*, v. 172, p. 132 – 151, 2016.

PODSIADLO, M.; RYBINSKI, H. Financial time series forecasting using rough sets with time-weighted rule voting. *Expert Systems with Applications*, v. 66, p. 219 – 233, 2016.

SAPANKEVYCH, N. I.; SANKAR, R. Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, v. 4, n. 2, p. 24–38, May 2009.

SMOLA, A. J.; SCHOLKOPF, B. *A Tutorial on Support Vector Regression*. [S.l.], 1998.

SMOLA, A. J.; SCHOLKOPF, B. A tutorial on support vector regression. *Regression. Statistics and Computing*, v. 14, n. 3, p. 199–222, 2003.

SU, C.-H.; CHENG, C.-H. A hybrid fuzzy time series model based on {ANFIS} and integrated nonlinear feature selection method for forecasting stock. *Neurocomputing*, v. 205, p. 264 – 273, 2016.

THE MATHWORKS INC. "MATLAB 2009a". Overview available at `http://www.mathworks.com/products/matlab/`.

TONG, H. Nonlinear time series analysis since 1990: Some personal reflections. *Acta Mathematicae Applicatae Sinica*, v. 18, n. 2, p. 177–184, 2002.

TSAY, R. S. *Analysis of Financial Time Series*. [S.l.]: John Wiley & Sons, 2005.

VAPNIK, V. N. *Statistical Learning Theory*. [S.l.]: John Wiley & Sons, 1998.

VAPNIK, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, v. 10, n. 5, p. 988–999, 1999.

WILLIAMS, T.; KELLEY, C.; many others. *Gnuplot 4.6: an interactive plotting program*. April 2013. `http://gnuplot.sourceforge.net/`.

YANG, H.; CHAN, L.; KING, I. Support vector machine regression for volatile stock market prediction. In: *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*. [S.l.: s.n.], 2002. p. 391–396.

YANG, H. et al. Outliers treatment in support vector regression for financial time series prediction. In: *Proceedings of the International Conference on Neural Information Processing (ICONIP)*. [S.l.: s.n.], 2004. p. 1260–1265.

ZADEH, L. A. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338–353, 1965.

ZHANG, G.; ZHANG, X.; FENG, H. Forecasting financial time series using a methodology based on autoregressive integrated moving average and taylor expansion. *Expert Systems*, v. 33, n. 5, p. 501–516, 2016.

ZHANG, G. P.; KLINE, D. M. Quarterly time-series forecasting with neural networks. *IEEE Transactions on Neural Networks*, v. 18, n. 6, p. 1800–1814, 2007.