UNIVERSIDADE FEDERAL FLUMINENSE

Antonio Augusto Corrêa Junior

ANÁLISE DE DISCREPÂNCIAS EM SÉRIES TEMPORAIS COM O USO DO DBSCAN E DA STGT

NITERÓI

2015

UNIVERSIDADE FEDERAL FLUMINENSE

Antonio Augusto Corrêa Junior

ANÁLISE DE DISCREPÂNCIAS EM SÉRIES TEMPORAIS COM O USO DO DBSCAN E DA STGT

Dissertação apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Mestre.
Área de Concentração: Inteligência Artificial.

Orientador:
Prof. Dr. José Viterbo Filho

Coorientador:
Prof. Dr. João Marcos Meirelles da Silva

NITERÓI
2015

ANÁLISE DE DISCREPÂNCIAS EM SÉRIES
TEMPORAIS COM O USO DO DBSCAN E DA STGT

ANTONIO AUGUSTO CORRÊA JUNIOR

Dissertação apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Mestre.
Área de Concentração: Inteligência Artificial.

Aprovada em Agosto de 2015.

BANCA EXAMINADORA

Prof. Dr. José Viterbo Filho – Orientador
UFF

Prof. Dr. João Marcos Meirelles da Silva – Coorientador
UFF

Prof. Dr. Luiz André Portes Paes Leme
UFF

Prof. Dr. Diego Barreto Haddad
CEFET-RJ

Niterói
2015

Para Camila, minha filha.

AGRADECIMENTOS

Nada mais natural em um trabalho tão repleto de referências do que iniciá-lo com uma importante citação. Assim, como escreveu Michael Cunningham em *Dias Exemplares* (2006), “Gratidão é a única resposta apropriada a *tudo* o que acontece.”. Por isso, agradeço a todos, indistintamente, que participaram da minha trajetória até este momento.

Além disso, faço um agradecimento especial à minha família pelo apoio incondicional ao longo do curso, ao Prof. Viterbo pelas frequentes palavras de incentivo e ao Prof. Torreão pela valiosa oportunidade de trabalhar com o algoritmo da STGT.

"É fácil, no mundo, viver de acordo com a opinião do mundo; é fácil, na solidão, viver de acordo consigo mesmo; mas o grande homem é aquele que, em meio à multidão, mantém com perfeita doçura a independência da solidão.”

Ralph Waldo Emerson, citado em *Uma Fração do Todo*

“(...) quando você é criança, para evitar que você siga a multidão você é atacado com a frase: 'Se todo mundo pulasse de uma ponte, você pularia?'; mas, quando você é adulto, e ser diferente é subitamente um crime, as pessoas parecem estar dizendo: 'Ei. Todo mundo está pulando de uma ponte. Por que você não?'”

Steve Toltz, *Uma Fração do Todo* (2011)

RESUMO

As principais estratégias adotadas atualmente em análise de investimentos no mercado de ações envolvem a análise fundamentalista, que procura avaliar a situação geral da economia, tendências setoriais e outros fatores econômico-financeiros de longo prazo, e a análise técnica, que faz uso de um conjunto de indicadores e de um outro conjunto de regras para reconhecimento de padrões de comportamento desses indicadores. Nesse último tipo de análise, que ainda pode ser subdividida em análise quantitativa e análise gráfica, os profissionais do mercado e os investidores costumam considerar que as circunstâncias históricas e o comportamento já observado nas séries temporais que representam os preços dos ativos apresentam uma boa chance de ocorrer novamente.

Por outro lado, é bastante comum encontrar elementos de um conjunto de dados que não obedecem ao comportamento geral apresentado pelo grupo como um todo. A sua ocorrência pode ser causada por erro na medição ou na execução de um processo e, por isso, costumam ser descartados. Muitos algoritmos de mineração de dados tendem a minimizar a influência dessas discrepâncias ou mesmo eliminá-las completamente. No entanto, aquilo que para alguém pode soar como um ruído, para outra pessoa pode soar como um sinal. Dessa forma, este estudo irá tratar a existência das discrepâncias no conjunto total de dados adotando um conceito comum em estatística, segundo o qual elas podem ser o resultado de diferentes funções de distribuição aplicadas a uma mesma população, ou seja, subpopulações distintas que coexistem em um mesmo ambiente ou contexto.

Assim sendo, este trabalho tem por objetivo propor um novo método para análise de séries temporais, visando priorizar a identificação e a interpretação de discrepâncias, baseado no uso de uma técnica de clusterização, através do algoritmo DBSCAN, e de análise tempo-frequência, através do algoritmo STGT. Espera-se que as discrepâncias que venham a ser identificadas possam sinalizar as mudanças de comportamento no mercado de forma mais clara e no momento oportuno, de forma a contribuir na redução da incerteza nas operações de compra e venda de ativos, auxiliando assim os investidores no processo de tomada de decisão. Esse método será validado em um estudo de caso com séries temporais formadas por dados do mercado de ações brasileiro que foram ampla e regularmente divulgados ao público.

Palavras-chave: Inteligência Artificial, Mineração de Dados, Séries Temporais, Clusterização, Análise Tempo-Frequência, Mercado de Ações, Bolsa de Valores.

ABSTRACT

The main strategies currently adopted in the analysis of investments in the stock market involves fundamental analysis, which seeks to assess the overall state of the economy, industry trends and other economic and financial factors of long-term, and technical analysis, which makes use of a set of indicators and another set of rules for recognizing behavior patterns in these indicators. In the latter type of analysis, which can be further subdivided into quantitative analysis and graphical analysis, market professionals and investors usually consider that the historical circumstances and the behavior already observed in the time series that represent asset prices have a good chance to reoccur.

On the other hand, it is quite common to find elements of a set of data that do not conform to the general behavior exhibited by the group as a whole. Their occurrence may be caused by error in the measurement or execution of a process and therefore often discarded. Many data mining algorithms tend to minimize the influence of these discrepancies or even eliminate them completely. However, what someone may sound like a noise, someone else may sound like a sign. Thus, this study will treat the existence of discrepancies in the total data set adopting a common concept in statistics, whereby it may be the result of two different distribution functions applied to the same population, that is, two distinct subpopulations coexisting in the same environment or context.

Therefore, this work aims to propose a new method for time series analysis for prioritizing the identification and interpretation of discrepancies based on the use of a clustering technique, through the DBSCAN algorithm, and time-frequency analysis, through the STGT algorithm. It is expected that the discrepancies which may be identified can signalize changes in market behavior in a more clearly and timely way, in order to contribute to the reduction of uncertainty in the purchase and sale of assets, thus helping investors in their decision-making process. This method will be validated in a case study with time series formed by the Brazilian stock market data that is being widely and regularly made public.

Keywords: Artificial Intelligence, Data Mining, Time Series, Clustering, Time-Frequency Analysis, Stock Market.

LISTA DE ILUSTRAÇÕES

2.1 Pontos conectados por densidade [Ester et al., 1996] 15
2.2 Representações de um sinal teórico e sua clusterização ..... 16 a 18
3.1: Representação do “dente de serra” por Série de Fourier [Williams e Spangler, 1981] ... 20
3.2: Duas DTF's de um mesmo sinal, utilizando o mesmo algoritmo [Boashash, 1992] 22
3.3: Representações de um sinal teórico e sua distribuição tempo-frequência ..... 25 a 26
4.1: Matriz gráfica para visualização das relações entre as variáveis da série temporal 31
4.2: Resultados preliminares da clusterização da série do IBOVESPA com o DBSCAN 33
4.3: Exemplo de protótipo de uma série temporal [Keogh e Pazzani, 1998] 38
4.4: Protótipo da série do IBOVESPA e suas representações no tempo 39
4.5: Representações do sinal de variação diária 42
4.6: Representações do sinal de máximo/mínimo–μ 44
4.7: Comparação entre DTF's dos sinais 46
4.8: 52 pontos de reversão da tendência de queda 49
4.9: 39 pontos de reversão da tendência de elevação 51
4.10: Análise de cluster com o uso do algoritmo SimpleKMeans 53
4.11: Exemplo construído com o auxílio do protótipo 55
5.1: Etapas da Mineração de Dados Indireta [Plastino, 2013] 56
5.2: Taxonomia dos Jogos [Kelly, 2003] 57
5.3: Aba Preprocess da Weka Explorer 59
5.4: Fluxo dos Processos ..... 62 a 63
6.1: Resultados da aplicação ao IBOVESPA 65
6.2: Aplicação ao IBOVESPA nos dias 248 e 266 ..... 67 a 68
6.3: Evolução na aplicação ao IBOVESPA entre os dias 351 e 486 ..... 70 a 73
6.4: Resultados da aplicação ao ativo PETR4 75
6.5: Evolução na aplicação ao ativo PETR4 entre os dias 549 e 651 ..... 77 a 80
6.6: Gráficos tipo *candlestick* com Bandas de Bollinger e MACD 84
6.7: Novas aplicações do modelo proposto ..... 87 a 89
6.8: Gráficos das aplicações no modelo tradicional ..... 92 a 93

LISTA DE TABELAS

4.1: Coeficiente de correlação de Pearson entre as variáveis da série do IBOVESPA 36
6.1: Análise do retorno com a aplicação ao IBOVESPA 69
6.2: Análise do retorno com a aplicação ao ativo PETR4 76
6.3: Análise do retorno com o modelo tradicional 85
6.4: Retorno das novas aplicações – modelo proposto 90
6.5: Retorno das novas aplicações – modelo tradicional ..... 93 a 94

SUMÁRIO

1 INTRODUÇÃO 01

1.1 Risco e Incerteza 05

1.2 Definição do Problema 05

1.3 Objetivo 06

1.4 Contribuição 06

1.5 Organização do Trabalho 07

2 MINERAÇÃO DE SÉRIES TEMPORAIS 08

2.1 Conceitos Básicos 08

2.2 Análise de Tendência 10

2.3 Taxonomia da Clusterização de Séries Temporais 11

2.4 Alguns Métodos de Clusterização 12

3 ANÁLISE TEMPO-FREQUÊNCIA DE SINAIS 20

3.1 Introdução à DTF – Distribuição Tempo-Frequência 21

3.2 Estudos Realizados com a Aplicação de DTF's 22

3.3 STGT – *Signal-Tuned Gabor Transform* 24

4 ANÁLISE EXPLORATÓRIA 28

4.1 Componentes do Processo de Clusterização 28

4.1.1 Escolha e Aplicação do Algoritmo 30

4.1.2 Redução da Dimensionalidade 32

4.1.3 Seleção das Medidas de Similaridade 34

4.1.4 Construção do Protótipo 37

4.2 Avaliação dos Resultados do Processo 40

4.3 Resultados da Aplicação da STGT 40

4.4 Reversão de Tendência 47

4.5 Avaliação das Técnicas Aplicadas 47

5 O MODELO PROPOSTO 56

5.1 Premissas na Extração dos Dados 57

5.2 Regras na Mineração da Informação 58

5.3 Critérios na Interpretação da Informação 60

5.4 Fluxo dos Processos 61

6 APLICAÇÕES DO MODELO 64

6.1 Aplicação ao IBOVESPA 64

6.2 Aplicação ao Ativo PETR4 74

6.3 Avaliação de Desempenho do Modelo Proposto 81

6.3.1 Critério de Avaliação 81

6.3.2 Outros Métodos de Análise Técnica 82

6.3.3 Outras Aplicações 86

6.3.4 Contribuição do Modelo Proposto 95

7 CONCLUSÃO 97

REFERÊNCIAS 99

1 – INTRODUÇÃO

Em março de 2000, enquanto seu livro intitulado *Irrational Exuberance* surgia nas prateleiras das livrarias nos Estados Unidos, o Professor Robert Shiller aproveitava o ano sabático da Universidade de Yale para se lançar em uma *tournée* por dez países divulgando o lançamento do livro. No prefácio àquela primeira edição, dentre muitos outros comentários, ele dizia parecer que os investidores se comportavam como se os indicadores do mercado de ações nunca poderiam vir a estar em níveis elevados demais, assim como nunca ficariam em níveis baixos por muito tempo. De certa forma essa lógica parecia ter alguma consistência pois, afinal, se milhões de pesquisadores e investidores estavam estudando os preços das ações e confirmando o seu valor aparente, por que perder tempo na tentativa de estabelecer preços mais razoáveis para os ativos? A resposta vinha a seguir:

> *But unknown to most investors is the troubling lack of credibility in the quality of research beeing done on the stock market, to say nothing of the clarity and accuracy with which it is communicated to the public. Some of this so-called research often seems no more rigorous than the reading of tea leaves. Arguments that the Dow is going 36,000 or 40,000 or 100,000 hardly inspire trust. Certainly* some *researchers are thinking more realistically about the market's prospects and reaching better-informed positions on its future, but these are not the names that grab the headlines and thus influence public attitudes.*
> *Instead the headlines reflect the news media's constant attention to trivial factoids and "celebrity" opinion about the market's price level. Driven as their authors are by competition for readers, listeners and viewers, media accounts tend to be superficial and thus to encourage basic misconceptions about the market. A conventional wisdom of sorts, stressing the seemingly eternal durability of stocks, has emerged from these media accounts. The public has learned to accept this conventional – but in my view shallow – wisdom. To be fair to the Wall Street professionals whose views appear in the media, it is difficult for them to correct the conventional wisdom because they are limited by the blurbs and sound bites afforded them. One would need to write books to straighten these things out.*

Naquele momento da história, ninguém sabia que março de 2000 representaria o pico do mercado de ações nos Estados Unidos. Daí em diante o mundo assistiria ao colapso da chamada "bolha" especulativa da internet. Em 2005, o Prof. Shiller lançou uma segunda edição de seu livro e nele, através de um gráfico que mostrava a evolução dos mercados de ações em dez países entre 2000 e 2001, apresentava a ocorrência de uma queda de mais da metade em um dos índices do mercado americano e de aproximadamente um terço no índice do mercado brasileiro. Para o período entre 2001 e 2002, o mesmo gráfico mostrava que o mercado brasileiro se recuperava quase que totalmente, enquanto o mercado americano permanecia em queda, atingindo quase um quarto do que havia sido o seu máximo. No prefácio da segunda edição, o Prof. Shiller acrescentou o seguinte comentário:

> *(This book) was, and is, about how errors of human judgement can infect even the smartest people, thanks to overconfidence, lack of attention to details, and excessive trust in the judgement of others, stemming from a failure to understand that others are not making independent judgements but are themselves following still others – the blind leading the blind.*

Com a crise dos títulos do mercado imobiliário nos Estados Unidos em 2008, o Professor Shiller lançou uma nova edição do seu livro em 2009, sempre enfatizando sua preocupação com os exageros incorridos pelos diversos agentes no mercado financeiro americano, o descaso do governo com esses excessos e as consequências sofridas pela população em geral. Na pauta das discussões, os mesmos três temas apontados anteriormente:

- excesso de autoconfiança: as pessoas tendem a fazer julgamentos em situações de incerteza através do uso de padrões já familiares, assumindo que os futuros padrões deverão ser semelhantes aos do passado, sem qualquer consideração sobre as razões ou as chances deles se repetirem;
- falta de atenção aos detalhes: a habilidade de focar a atenção em coisas importantes é uma das características que definem a inteligência humana;
- excesso de confiança no julgamento de terceiros: pesquisas na área de psicologia social demonstram que as pessoas estão sempre prontas para acreditar na opinião da maioria ou de autoridades, mesmo quando essas opiniões contrariam o senso comum ou a prática.

De fato, investidores e analistas de mercado sempre se concentraram em obter previsões acuradas sobre as tendências futuras dos preços dos ativos. Zarb e Kerekes [1970] argumentam que as principais estratégias adotadas em análise de investimentos envolvem a análise fundamentalista, que procura avaliar a situação geral da economia, tendências setoriais e outros fatores econômico-financeiros de longo prazo, e a análise técnica, que faz uso de um

conjunto de indicadores e de um outro conjunto de regras para reconhecimento de padrões de comportamento desses indicadores. Nesse último tipo de análise, que ainda pode ser subdividida em análise quantitativa e análise gráfica, os profissionais do mercado e os investidores costumam considerar que as circunstâncias históricas e os padrões de comportamento já observados para os ativos tendem a ocorrer repetidamente [Edwards et al., 2012; Murphy, 1999]. Assim sendo, esse método de análise confirma o primeiro problema apontado pelo Prof. Shiller, quando ele afirma que as pessoas tendem a usar padrões familiares e têm a expectativa de que eles se repitam ao longo do tempo.

No entanto, seria realmente necessário prever o comportamento futuro do preço de um ativo para realizar um investimento no presente? Ou bastaria para isso, conforme sugere o Prof. Shiller no seu segundo ponto, entender em maior profundidade e com riqueza de detalhes o atual estado em que se encontra o ativo em questão? A abordagem adotada por Lee e Jo [1999] em seu sistema especialista subdivide as mudanças nos mercados em Altas, Baixas, Posições Neutras, Confirmação de Tendência e Reversão de Tendência, sendo essa última a principal indicação para a ação dos investidores. Esse sistema usou como base as informações fornecidas por gráficos de barras (ou *candlestick charts*), que apresentam para cada dia de negociação os valores máximo, mínimo, de abertura e de fechamento dos ativos, indicando através de cores os movimentos diários de elevação e queda nesses valores.

Tsai e Quan [2014], que desenvolveram um método de processamento de imagem para analisar esse mesmo tipo de gráfico, argumentam que os aspectos visuais de um gráfico ou imagem são sinais que podem oferecer uma mensagem objetiva para os usuários, sejam eles analistas de investimentos ou meros investidores, sem requerer que eles possuam a experiência de um especialista em finanças. Em outras palavras, a interface gráfica parece ser a mais adequada para atingir um amplo espectro de usuários, evitando com isso a frequente disseminação de informações produzidas com base em dados de pouca relevância, que podem resultar em um comportamento indevido dos investidores conforme apontado pelo Prof. Shiller no seu terceiro e último ponto.

Ainda segundo o Prof. Shiller, a incessante troca de informações é uma característica fundamental da espécie humana. Na sociedade moderna, é bastante comum que haja uma rápida disseminação de informações sobre uma boa oportunidade de compra de um ativo ou alguma ameaça à nossa propriedade, pois esses assuntos lembram aqueles sobre os quais nossos ancestrais conversaram ao longo dos muitos séculos da existência da humanidade. Por outro lado, a comunicação não costuma fluir com a mesma facilidade quando se trata de temas mais abstratos, tais como cálculos financeiros ou estatísticas sobre

retorno de investimentos. A transmissão desse tipo de conhecimento costuma envolver muito esforço, ser pouco frequente e carregar muitas imperfeições.

As informações habitualmente divulgadas pelos profissionais envolvidos com o mercado de ações, sejam eles especialistas em finanças, analistas de investimento ou mesmo profissionais da imprensa, são carregadas de subjetividade e, acredita-se, de uma razoável dose de viés. Isso ocorre devido à própria natureza dos movimentos do mercado que, por apresentarem forte oscilação, são geralmente tratados como aleatórios. Assim, muitos erros, omissões e tentativas de manipulação se perdem em meio a mudanças bruscas de tendência e fortes volatilidades, ou são interpretados como discrepâncias e, com isso, são intencionalmente desprezados para não prejudicarem o resultado final da análise. Mas, afinal, o que viriam a ser essas discrepâncias?

É bastante comum encontrar elementos de um conjunto de dados que não obedecem ao comportamento geral apresentado pelo grupo como um todo. A sua ocorrência pode ser causada por erro na medição ou na execução de um processo e, por isso, costumam ser descartados. Muitos algoritmos de mineração de dados tendem a minimizar a influência dessas discrepâncias ou mesmo a eliminá-las completamente. No entanto, aquilo que para alguém pode soar como um ruído, para outra pessoa pode soar como um sinal. Dessa forma, este estudo irá tratar a existência das discrepâncias no conjunto total de dados adotando um conceito comum em estatística, segundo o qual elas podem ser o resultado de diferentes funções de distribuição aplicadas a uma mesma população, ou seja, subpopulações distintas que coexistem em um mesmo ambiente ou contexto.

De acordo com vários autores [Aghabozorgi et al., 2015; Keogh e Lin, 2004; Kosmelj e Batagelj, 1990], além de ser uma sub-rotina em algoritmos mais complexos de mineração de dados, a clusterização é a abordagem mais utilizada como técnica exploratória de séries temporais, sendo utilizada para a descoberta de regras de associação, na classificação e na detecção de anomalias ou discrepâncias. No entanto, as abordagens de clusterização usualmente adotadas privilegiam a comparação entre série temporais completas ou entre subsequências extraídas de uma mesma série, dando menor importância à identificação de pontos de ruído ou trechos de transição entre grupos de pontos da série, que podem caracterizar uma mudança de comportamento relevante.

Torna-se portanto necessário balancear a incerteza inerente aos investimentos em mercados de ações com informações mais objetivas e que retratem da melhor forma possível as reais características do momento presente vivido nesses mercados.

1.1 – Risco e Incerteza

De acordo com Kelly [2003], os jogos de chance são aqueles em que um único jogador enfrenta a natureza. Nesse caso, diferentemente do que acontece nos jogos de habilidade e nos de estratégia, o jogador em questão não é capaz de controlar completamente os resultados alcançados a partir de suas ações e as suas decisões estratégicas nem sempre levam a um resultado que pudesse ter sido antecipado.

Os jogos de chance podem ser subdivididos em dois outros tipos: jogos envolvendo risco, onde o jogador conhece a probabilidade de cada resposta da natureza às suas ações, e jogos envolvendo incerteza. Nesse último tipo, segundo Colman [1982; apud Kelly, 2003], não é possível atribuir uma probabilidade que contenha algum significado a quaisquer das respostas da natureza às ações do jogador.

Tendo como base o conceito exposto, este trabalho irá considerar as operações nos mercados financeiros denominados bolsas de valores como jogos de chance envolvendo incerteza, ficando fora do seu escopo qualquer avaliação de risco das operações.

1.2 – Definição do Problema

Resumindo os principais pontos anteriormente discutidos, é possível destacar como característica do problema a ser tratado o fato de os modelos atualmente adotados serem fortemente baseados na identificação de padrões históricos de comportamento da série temporal que, conforme está implícito na abordagem adotada pelas próprias técnicas utilizadas no processo de análise, supostamente apresentariam uma grande chance de se repetirem, sem que para isso exista qualquer justificativa plausível; não é de todo improvável que essas repetições possam vir a ser um resultado do uso de tais técnicas como mero “protocolo de comunicação” entre os agentes que atuam no mercado.

Além disso, como uma consequência dessa característica, esses modelos não são voltados para extrair da série temporal algum conhecimento mais profundo a respeito do momento presente, ou seja, do estado atual do processo representado pela série temporal. Esses mesmos modelos praticamente desprezam a ocorrência de discrepâncias no processo que está sendo analisado, tratando-as muitas vezes como pontos da série que, em vez de explicar um fenômeno, prejudicam o resultado final da análise.

Assim, mais do que compreender os fenômenos do passado ou tentar prever o comportamento no futuro, os estudos envolvendo séries temporais que caracterizam mercados financeiros deveriam demonstrar uma atenção especial para com a situação corrente nos mercados. É fundamental examinar cada ponto da série temporal no instante em que os dados

que o caracterizam sejam gerados e procurar interpretar corretamente as discrepâncias detectadas no processo. Só assim será possível produzir informações acuradas e divulgá-las a tempo, de tal forma que cada agente que atue no mercado possa tomar suas decisões e agir da maneira que lhe parecer mais conveniente.

1.3 - Objetivo

Este trabalho tem por objetivo propor um novo método para análise de séries temporais, visando priorizar a identificação e a interpretação de discrepâncias, baseado no uso de uma técnica de clusterização, através do algoritmo DBSCAN, e de análise tempo-frequência, através do algoritmo STGT. Espera-se que as discrepâncias que venham a ser identificadas possam sinalizar as mudanças de comportamento no mercado de forma mais clara e no momento oportuno, de forma a contribuir na redução da incerteza nas operações de compra e venda de ativos, auxiliando assim os investidores no processo de tomada de decisão.

Esse método será validado em um estudo de caso no qual as séries temporais serão formadas preferencialmente por dados do mercado de ações brasileiro, tendo em vista a menor dificuldade em associar os fenômenos identificados a um contexto econômico que possa ter sido vivenciado e, além disso, aos fatos noticiados e debatidos nos meios de comunicação de massa que atuam em uma região específica. Para isso, serão utilizados dados que estejam sendo ampla e regularmente divulgados ao público, procurando gerar conhecimento capaz de reduzir as incertezas nas decisões envolvendo operações na BM&FBOVESPA, buscando tornar mais claro o risco incorrido por agentes com menor poder de influência no mercado.

1.4 – Contribuição

Como principal contribuição deste trabalho para solucionar os aspectos mais relevantes do problema acima descrito é possível citar o desenvolvimento de um novo modelo para avaliação da incerteza nos investimentos em bolsas de valores, com características compatíveis com as de outros métodos de análise técnica amplamente utilizados. Apesar de estar voltado para profissionais do mercado, esse modelo também pode vir a ser utilizado pelos próprios investidores, necessitando para isso de pequenas adaptações para facilitar seu o uso e do treinamento apropriado para perfil dos usuários.

Além disso, os estudos de caso apresentados como demonstração da aplicabilidade do modelo podem servir de referência para os investidores no que diz respeito às reais expectativas de taxa de retorno e de prazo de aplicação envolvendo os investimentos no mercado de capitais brasileiro na atualidade.

1.5 – Organização do Trabalho

Este trabalho foi organizado de forma a refletir, da forma mais fiel possível, todas as etapas da investigação realizada ao longo da pesquisa que foi conduzida, buscando não prejudicar a compreensão dos conceitos envolvidos e dos resultados alcançados. Como este trabalho envolve temas de interesse de áreas de pesquisa bastante distintas, o capítulo descrevendo estudos anteriores foi suprimido e as referências aos trabalhos relacionados a esses temas foram incluídas nos respectivos capítulos e seções que os apresentam.

No Capítulo 2 será abordado o tema de Mineração de Séries Temporais, seus conceitos básicos e os especificamente aplicados a séries temporais, a taxonomia adotada na clusterização de séries e uma avaliação de dois diferentes algoritmos de clusterização baseados em forma, que resulta na escolha do DBSCAN como a alternativa mais adequada para aplicação neste estudo.

No Capítulo 3 será apresentada a técnica de Análise Tempo-Frequência de Sinais, juntamente com alguns importantes estudos nessa área e a definição da STGT – Signal Tuned Gabor Transform, na qual se baseia o algoritmo utilizado neste trabalho.

No Capítulo 4 serão apresentadas as etapas percorridas e os resultados alcançados na Análise Exploratória conduzida com a utilização do DBSCAN, da STGT e de outras técnicas adotadas, comparando séries temporais com diferentes características.

No Capítulo 5 será apresentado o Modelo Proposto para análise de séries temporais referentes a ativos negociados em bolsas de valores, suas carteiras e seus índices. Serão apresentadas as premissas, as regras e os critérios adotados em cada etapa do processo.

No Capítulo 6 as principais Aplicações do Modelo serão descritas, apresentando os estudos de caso conduzidos com as séries temporais produzidas com dados referentes ao índice IBOVESPA e ao ativo PETR4 – Petrobras PN. Além disso, será realizada uma avaliação de desempenho do modelo proposto, apresentando seu critério de avaliação, fazendo a comparação dos seus resultados com os de outros modelos que têm a mesma finalidade e evidenciando a sua contribuição.

Finalmente, no Capítulo 7 será apresentada a Conclusão do trabalho, incluindo algumas oportunidades para futuros estudos.

2 – MINERAÇÃO DE SÉRIES TEMPORAIS

De acordo com Fu [2011], a mineração em séries temporais pode ser compreendida como o objetivo de descobrir informações ocultas ou conhecimento inserido nos dados da série, sejam eles os próprios dados originais ou dados resultantes de algum tipo de transformação aplicada aos originais. Fu considera a busca por padrões a atividade mais comum envolvendo o tema e a clusterização a técnica mais comumente adotada.

Já Esling e Agon [2012], logo no início de seu trabalho, manifestam a opinião de que mineração de dados em séries temporais tem como propósito tentar extrair *da forma* dos dados todo o conhecimento que carregue significado. É interessante observar que, nesse caso em particular, o termo “forma” pode assumir um significado bem mais geral daquele normalmente usado para categorizar as medidas de similaridade.

Han e Kamber [2006] examinam os aspectos envolvendo a mineração de séries temporais dando foco à análise de tendência e a busca de similaridades. Nesse último tópico, os autores adotam o mesmo conceito apresentado por Aghabozorgi et al. [2015], que determina a busca de similaridades através da comparação de séries completas ou de subsequências de uma mesma série.

2.1 – Conceitos Básicos

Buscando a conceituação mais simples possível, a expressão mineração de dados se refere ao processo de extrair (ou minerar) conhecimento a partir de uma grande quantidade de dados. Segundo Han e Kamber [2006], essa expressão é, na verdade, um termo impróprio. Quando, por exemplo, se faz referência à mineração de ouro a partir de montanhas ou rochas, não nomeamos esse processo como mineração de montanhas ou mineração de rochas, mas sim mineração de ouro. Dessa forma, mineração de dados estaria mais apropriadamente nomeada

por “mineração de conhecimento a partir de dados”, o que, infelizmente, é uma expressão muito longa e que não parece ser tão eficaz quanto a primeira. De qualquer forma, o importante é que a palavra minerar induz a ideia de um processo no qual é necessário encontrar pequenas partes de material de maior valor em meio a uma grande quantidade de material bruto. Uma grande parte da comunidade científica também trata a mineração de dados como uma etapa essencial de um processo de KDD – *Knowledge Discovery from Data*. A descoberta de conhecimento é um processo que consiste dos seguintes passos:

- Limpeza de dados;
- Integração de dados;
- Seleção de dados;
- Transformação de dados;
- Mineração de dados;
- Avaliação de padrões;
- Apresentação do conhecimento.

Os quatro primeiros passos acima apresentados são formas de pré-processamento dos dados com o objetivo de prepará-los para a etapa de mineração. Após a mineração dos dados, dependendo das técnicas que estão sendo utilizadas, pode ser necessário avaliar se os padrões extraídos são realmente interessantes a ponto de representarem um conhecimento. Para isso é normalmente levado em consideração:

- se ele é facilmente compreendido pelas pessoas;
- se ele é válido para os dados utilizados com um certo grau de certeza;
- se ele é potencialmente útil;
- se ele é original ou fora do comum.

Os três últimos passos podem ser repetidos iterativamente, de forma a permitir que os padrões sejam avaliados adequadamente de acordo com os critérios acima apresentados.

Existem diferentes formas de classificação de um processo ou sistema de mineração de dados. Assim, é comum encontrar na literatura classificações de acordo com:

- o tipo de base de dados que está sendo minerada, seja pelo seu modelo de dados (relacional, transacional, etc.) ou pelo tipo dos dados (séries temporais, texto, etc.);
- o tipo de conhecimento minerado (previsões, regras de associação, etc.), seu grau de abstração e se representa um padrão regular ou uma irregularidade no padrão;
- o tipo de técnica utilizada (estatística, de aprendizado de máquina, de processamento científico, etc.), incluindo o grau de envolvimento do usuário;

- a sua área de aplicação (finanças, medicina, telecomunicações, varejo, etc.).

A classificação segundo o tipo de técnica utilizada, incluindo o grau de envolvimento do usuário, costuma caracterizar as duas principais abordagens adotadas em mineração de dados: direta, quando existe uma meta bem definida para orientar o processo, e indireta, quando não existe tal meta [Côrtes et al., 2002]. Caso seja adotada alguma técnica de aprendizado de máquina no processo, a abordagem direta pressupõe o uso de técnicas de aprendizado supervisionado, enquanto a abordagem indireta envolve o uso de técnicas de aprendizado não-supervisionado, como veremos adiante.

2.2 – Análise de Tendência

Segundo Han e Kamber [2006], uma base ou domínio de dados de séries temporais consiste de uma sequência de valores ou eventos obtidos através de repetidas medições ao longo do tempo, tipicamente medidas em intervalos iguais (ex.: horário, diário, semanal, etc.). Essas séries são bastante utilizadas em diversas aplicações na área econômico-financeira, tal como na análise do mercado de ações, em projeções econômicas, em previsões de vendas e em análise orçamentária, bem como de outras áreas, tal como controle de qualidade e de processos, observação de fenômenos naturais e tratamentos médicos.

De acordo com os mesmos autores, uma série temporal envolvendo a variável Y, que pode representar, por exemplo, o preço de fechamento diário de uma ação na bolsa de valores, pode ser vista como uma função Y = F (t). Em geral, existem dois objetivos na análise de uma série temporal: modelagem e previsão da série. Esse último está fora do escopo deste estudo por razões já apresentadas anteriormente. Já o primeiro, que nos permite obter maior entendimento sobre os mecanismos ou forças que geram os valores da série temporal, consiste em quatro componentes principais:

- tendência (T) ou movimentos de longo prazo: indica a direção geral na qual a série se move durante um longo intervalo de tempo; é importante ressaltar que a definição de “longo” é relativa, dependendo do intervalo de tempo usado nas medições da série temporal, e pode, no contexto deste estudo, envolver períodos que vão de dezenas de dias até alguns anos;
- movimentos cíclicos (C): se referem às oscilações de longo prazo em torno de uma curva de tendência, podendo ser periódicos ou não;
- movimentos sazonais (S): são sistemáticos e diretamente relacionados com o calendário, independentemente do intervalo de medição adotado;

- movimentos irregulares (I) ou aleatórios: caracterizam o movimento esporádico da série devido a eventos aleatórios ou de chance.

Assim, a modelagem da variável Y citada anteriormente costuma ser representada pelo produto ou pela soma das quatro componentes acima mencionadas (ex.: Y = T + C + S + I), sendo essa escolha tipicamente empírica. A componente T de uma série é habitualmente representada por uma curva de tendência, que costuma ser determinada por um conjunto típico de métodos, sendo a média móvel ponderada um dos mais utilizados. No entanto, esses métodos acabam por gerar a expectativa de que os valores obtidos no passado terão influência significativa nos valores a serem obtidos através das medições futuras. Idealmente, para uma função suave, poderíamos calcular a primeira e a segunda derivadas em relação ao tempo da variável Y = F (t) para conhecer a tangente e a sua taxa de variação em cada ponto dessa função. No entanto, mesmo com medições em intervalos de tempo tendendo a zero, as séries formadas por valores de ativos negociados em bolsas não possuem tangente definida.

2.3 – Taxonomia da Clusterização de Série Temporais

Analisando a literatura sobre o tema, é possível concluir que os processos de clusterização de séries temporais podem ser agrupados em três categorias [Aghabozorgi et al., 2015]:

- clusterização de séries temporais completas, que consiste na clusterização de um conjunto de séries temporais em relação à similaridade entre elas; nesse caso, considerando que a clusterização envolve a aplicação de técnicas a objetos, os objetos em questão são as séries temporais como um todo;
- clusterização de subsequências de uma série temporal, que envolve a clusterização de um conjunto de subsequências de uma única série temporal, extraídas a partir do uso de uma janela que se movimenta ao longo da série completa; nesse caso, os objetos clusterizados são segmentos de uma única série temporal mais longa;
- clusterização de pontos de uma série temporal, que trata da clusterização de pontos individuais de uma mesma série temporal baseada na proximidade entre eles; como o próprio nome indica, os objetos são os pontos da série; diferentemente da clusterização dos segmentos, nessa categoria nem todos os objetos precisam fazer parte de um cluster, ou seja, alguns objetos podem ser considerados ruído.

Surpreendentemente, Keogh e Lin [2004] argumentam que a clusterização de subsequências de uma série temporal não carrega qualquer significado (?!), demonstrando que os segmentos da série obtidos a partir de janelas móveis são forçados a obedecer certas

restrições, às quais os autores afirmam ser "*pathologically unlikely*" (patologicamente improvável) que qualquer conjunto de dados seja capaz de satisfazer. Por causa disso, os clusters extraídos da aplicação dos algoritmos a esses segmentos são, na sua essência, aleatórios. Considerando o domínio de dados a ser empregado, a caracterização da etapa inicial do estudo como exploratória e o fato de se acreditar que irregularidades no padrão ou discrepâncias (ruídos) possam vir a ser de grande relevância para o conhecimento do fenômeno que está sendo estudado, a estratégia adotada neste trabalho para o processo de clusterização foi a que utiliza os pontos da série temporal como objetos.

Além das categorias acima apresentadas, os diferentes trabalhos de clusterização de série temporais podem ser agrupados de acordo com três diferentes abordagens passíveis de serem adotadas [Liao, 2005]:

- clusterização de dados brutos ou em seu estado natural, que trabalha diretamente com os dados originais da série, tendo apenas a preocupação de garantir que as medidas de similaridade ou distância usadas nas técnicas empregadas para dados estáticos sejam também adequadas, ou possam ser adaptadas, para uso em dados com variação temporal;
- clusterização baseada em características, que converte os dados da série em seu estado natural para um vetor de características da série com uma dimensionalidade inferior à da série original e, em seguida, aplica um algoritmo de clusterização convencional ao vetor obtido;
- clusterização baseada em modelo, que, analogamente à abordagem anterior, converte os dados da série em seu estado natural para um modelo com um conjunto de parâmetros e, em seguida, aplica um algoritmo de clusterização convencional ao modelo produzido.

2.4 – Alguns Métodos de Clusterização

Dentre as várias técnicas de aprendizado de máquina, a clusterização é um exemplo de método onde o aprendizado é não-supervisionado. Segundo Panait e Luke [2005], as abordagens de aprendizado de máquina diferem entre si pelo *feedback* que o papel de "crítico" do método fornece ao papel de "aprendiz": nos métodos supervisionados é fornecida a saída correta (ex.: classificação), nos métodos por recompensa é fornecida uma avaliação da qualidade da saída e nos métodos não-supervisionados não é fornecido qualquer *feedback*.

O objetivo do processo de clusterização é identificar uma estrutura em um conjunto de dados não rotulados, através da organização objetiva dos dados em grupos homogêneos, nos quais a similaridade entre os objetos dentro do grupo ao qual ele pertence é minimizada ao mesmo tempo em que a dissimilaridade entre objetos de diferentes grupos é maximizada. A clusterização é necessária quando não há nenhum rótulo disponível, independentemente do tipo de dado que está sendo analisado [Liao, 2005]. É por esse motivo que a clusterização é considerada uma forma de aprendizado por observação, ao contrário da classificação onde o aprendizado se dá através de exemplos.

Procurando expressar de uma forma mais simples, clusterização é o processo de agrupar um conjunto de objetos físicos ou abstratos em subconjuntos de objetos similares. Um cluster é uma coleção de objetos que são similares entre si e, ao mesmo tempo, diferentes dos objetos que formam um outro cluster. A análise de cluster é uma importante atividade humana. Já bem cedo, no início da infância, as crianças aprendem a diferenciar cães de gatos, ou animais de plantas, através do aprimoramento de esquemas de clusterização subconsciente. Através de processos automatizados de clusterização, é possível descobrir padrões de distribuição de objetos em uma ou várias dimensões e interessantes correlações entre atributos de uma mesma entidade de dados. A clusterização também é chamada de segmentação de dados em algumas aplicações, pois particiona grandes grupos de dados em subgrupos de acordo com a sua similaridade, assim como pode ser usada para detectar discrepâncias, que muitas vezes são mais significativas e interessantes do que os casos comuns, que costumam seguir um padrão de maior regularidade.

Existem vários algoritmos de clusterização disponíveis na literatura. De acordo com Han e Kamber [2006] a categorização dos métodos de clusterização é uma tarefa difícil, pois as categorias apresentam sobreposição quando avaliadas segundo diferentes critérios. Aghabozorgi et al. [2015] argumentam que todos os métodos que utilizam alguma medida de distância entre objetos como critério de similaridade (ou dissimilaridade) entre eles podem ser considerados como integrantes de um grande grupo de métodos baseados em forma. Assim, métodos de clusterização envolvendo particionamento e métodos baseados em densidade podem ser agrupados em uma categoria mais geral de métodos baseados em forma, sendo o primeiro subgrupo indicado para identificar clusters de formato esférico e o segundo para identificar clusters de formato irregular ou arbitrário. O SimpleKMeans é um exemplo do primeiro subgrupo e o DBSCAN pertence ao segundo, sendo ambos encontrados na ferramenta WEKA – Waikato Environment for Knowledge Analysis, versão 3.6.10.

O SimpleKMeans é considerado um exemplo pertencente à categoria de métodos de particionamento. Nesse caso, dado um conjunto de dados contendo *n* objetos, o algoritmo é capaz de formar *k* partições ou grupos, sendo $k<=n$, onde cada grupo contém pelo menos um objeto, cada objeto só pertence a um único grupo e cada grupo é representado por um elemento denominado centroide. Para esse algoritmo em particular, a posição do centróide é obtida através de uma função que calcula o valor médio das posições dos objetos que pertencem ao cluster.

Dado o número de partições a serem construídas, o método cria um primeiro particionamento e, em seguida, dá início a uma técnica iterativa de realocação dos centroides com o objetivo de melhorar o particionamento já obtido e, com isso, atingir a otimização global dos clusters. Um bom particionamento é aquele em que os objetos de um mesmo cluster estão próximos ou estão relacionados uns aos outros, enquanto objetos de clusters diferentes estão distantes ou dissociados uns dos outros. Devido à função adotada para estabelecer a posição do centroide, esse método é especialmente útil para encontrar clusters esféricos em conjuntos de dados de tamanhos pequenos a médios.

O DBSCAN – *Density Based Spatial Clustering of Applications with Noise* [Ester et al., 1996], como indica o seu nome, faz parte da categoria de métodos de clusterização por densidade. A ideia geral por trás desse tipo de método é dar prosseguimento ao crescimento de um cluster enquanto a sua densidade, ou número de objetos encontrados numa região ou vizinhança, for maior ou igual a um determinado valor mínimo esperado. Em outras palavras, para cada objeto em um determinado cluster, a vizinhança compreendida a uma distância menor ou igual a um certo raio contém no mínimo um certo número previamente definido de outros objetos.

Os métodos baseados em densidade são particularmente úteis para identificar clusters de formato irregular ou arbitrário [Ester et al., 1996], uma vez que os métodos de particionamento definem apenas clusters de formato circular ou esférico, considerando que o conjunto de dados pode estar distribuído em duas ou três dimensões respectivamente. Além disso, são muito eficazes para detectar discrepâncias no conjunto de dados, seja através da identificação de ruídos ou pela simples separação entre clusters, já que esses últimos são subespaços mais densos que são separados por subespaços mais rarefeitos. Apesar disso, conforme foi comentado por Aghabozorgi et al [2015], "*reviewing the literature it is noticeable that density-based clustering has not been used broadly for time-series data clustering because of its rather high complexity*". Mas, afinal, como os clusters são construídos através desse algoritmo?

A figura 2.1 a seguir [Ester et al., 1996], mostra que os pontos *r* e *s* estão conectados por densidade através do ponto *o*. O algoritmo DBSCAN inicialmente verifica a vizinhança de cada ponto do conjunto de dados; se a uma distância “epsilon” de um ponto for encontrado um número de pontos maior que “minPoints”, um novo cluster é formado com o ponto avaliado como centro do cluster. Em seguida, o algoritmo reúne os pequenos clusters que são alcançáveis por densidade, ou seja, os clusters cujos núcleos estão ligados por regiões que apresentam a densidade mínima requerida. É importante observar que nem todo ponto é um núcleo e, sendo assim, os pontos nas bordas de um cluster podem ser alcançados por densidade sem que eles próprios alcancem outros pontos por densidade; logo, esse conceito não apresenta simetria: o ponto *q* pode ser alcançado pelo ponto *p*, apesar do ponto *p* não ser alcançável pelo ponto *q*.

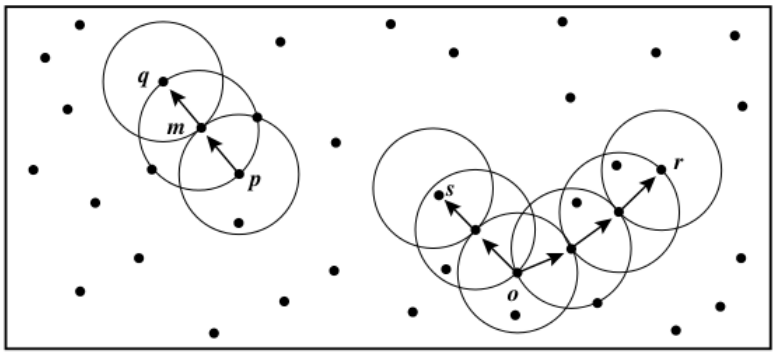
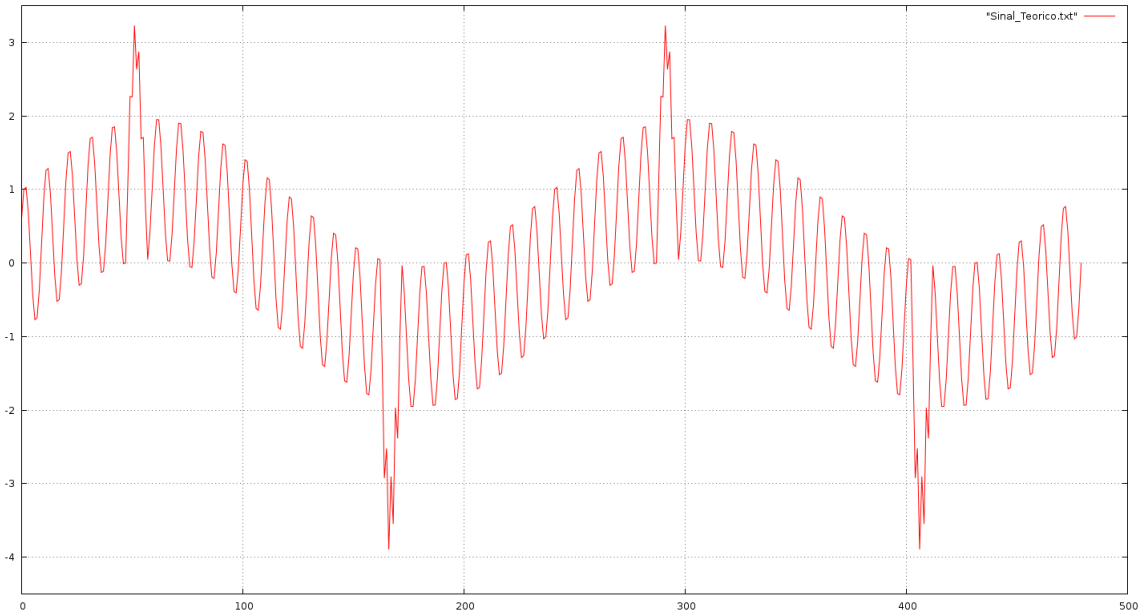
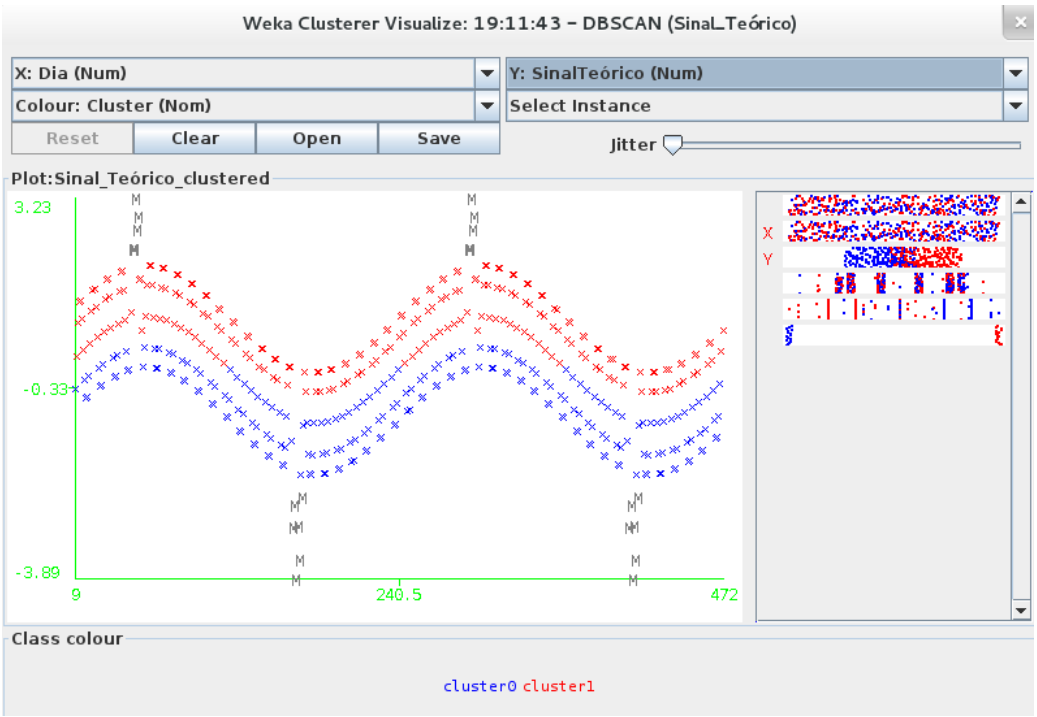


Figura 2.1: Pontos conectados por densidade [Ester et al., 1996].

Nas Figuras 2.2 (a), (b), (c), (d), (e) e (f) a seguir, é apresentada uma função teórica, construída com objetivo meramente didático, e os resultados da clusterização dos seus pontos através do DBSCAN e do SimpleKMeans.

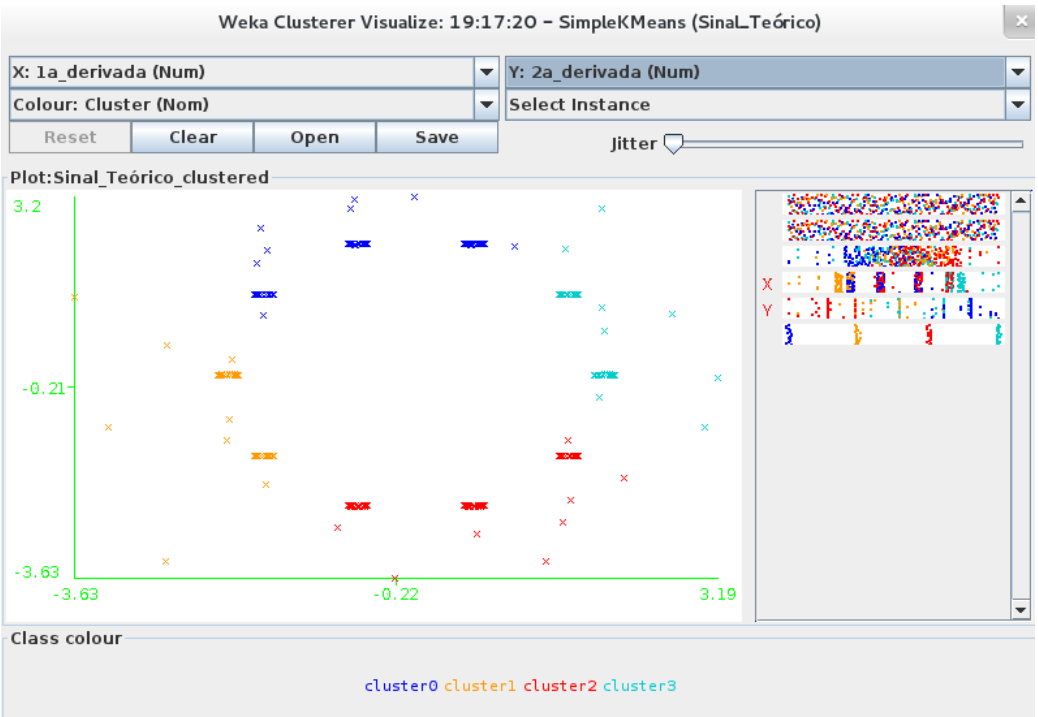


(a)

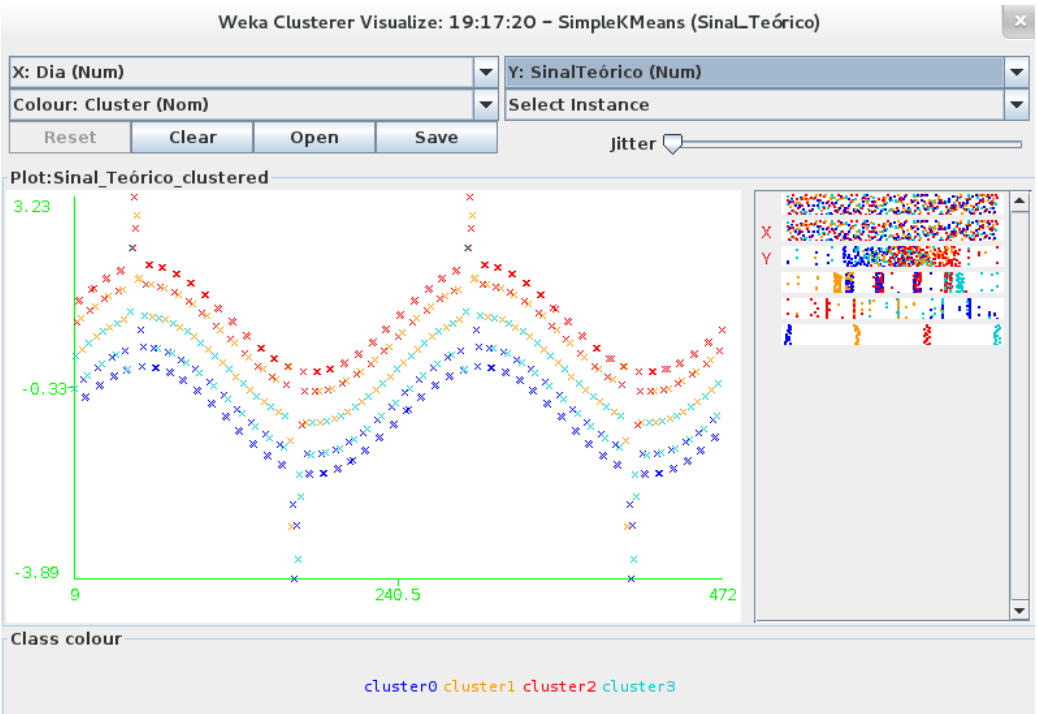


(b)

Figura 2.2: Representações de um sinal teórico e sua clusterização: (a) x(t); (b) x(t) pelo DBSCAN; (c) [x''(t)] vs [x'(t)] pelo SimpleKMeans; (d) x(t) com clusterização das estimativas das derivadas pelo SimpleKMeans; (e) [x''(t)] vs [x'(t)] pelo DBSCAN; (f) x(t) com clusterização das estimativas das derivadas pelo DBSCAN.

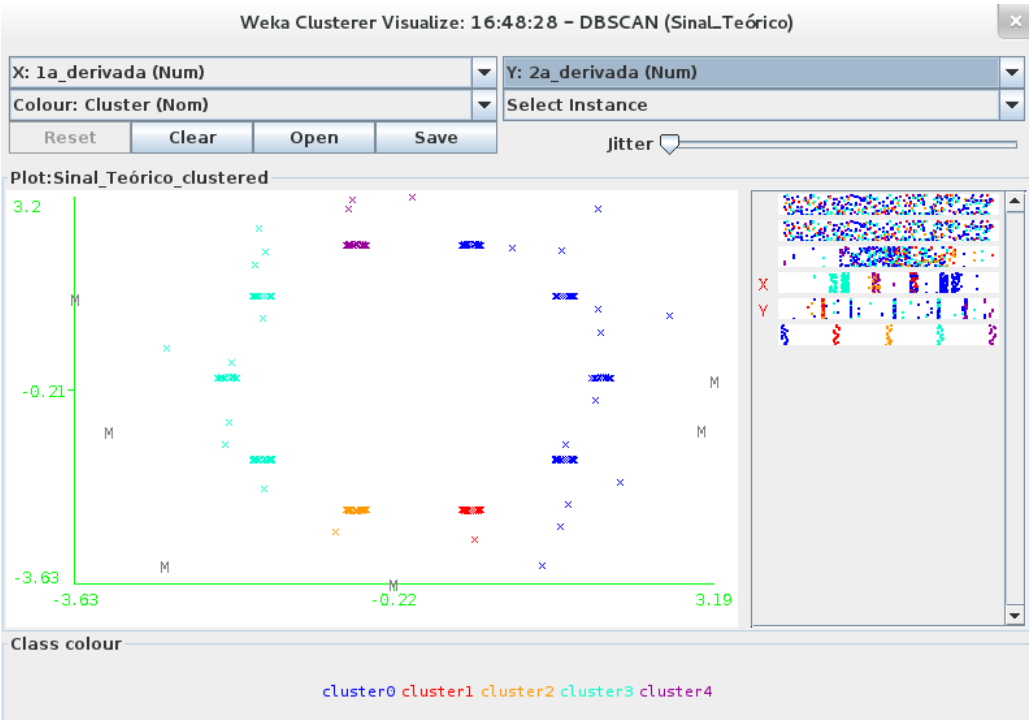


(c)

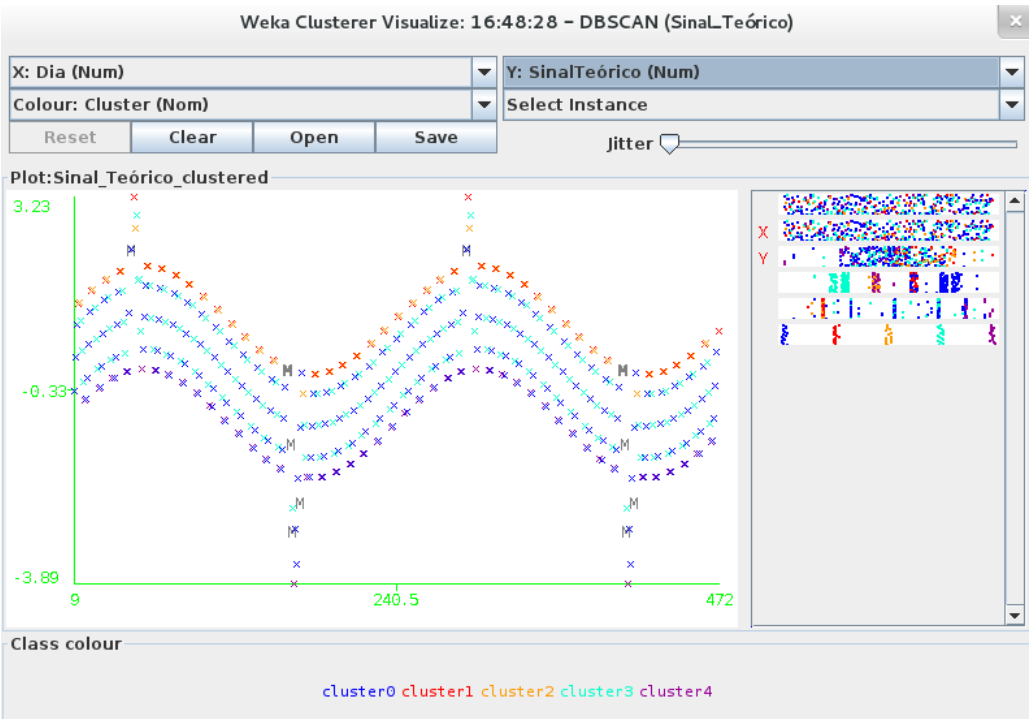


(d)

Figura 2.2 (continuação)



(e)



(f)

Figura 2.2 (continuação)

A função criada apresenta quatro variáveis: Dia, Sinal Teórico (um valor adimensional) e estimativas para a 1ª derivada e a 2ª derivada da função. Os resultados obtidos através dos processos de clusterização conduzidos nessa fase inicial do estudo foram fruto de uma abordagem empírica, adequada para a exploração do potencial das técnicas mas sem a preocupação com a fundamentação das decisões tomadas. Para a clusterização com o DBSCAN, foram ignoradas as variáveis 1ª derivada e 2ª derivada, adotado como parâmetro minPoints=4 e, após inúmeros testes, foi escolhido o valor 0.04225 para o parâmetro epsilon. Já para o caso do SimpleKMeans, foram ignoradas as variáveis Dia e Sinal Teórico, adotada a opção pela distância Euclidiana e, após a realização de algumas poucas tentativas, a optou-se pela busca de quatro clusters.

O DBSCAN foi capaz de identificar os pontos de discrepância que foram simulados na série, representando-os como ruídos assinalados com M (*Missing*). Além disso, foi capaz de representar claramente dois subconjuntos da série, formados pelos pontos localizados na parte superior e inferior do gráfico. Curiosamente, os pontos parecem formar cinco senóides e aquela localizada exatamente no meio foi subdividida entre os dois clusters produzidos. Já o SimpleKMeans, que não é capaz de identificar ruídos, representou as discrepâncias como pontos integrantes dos clusters identificados, produzidos em função exclusivamente dos valores das derivadas estimadas para esses pontos. Nesse caso também é possível observar um resultado intrigante, já que o conjunto de senóides formada pelos pontos da série parece representar uma espécie de gradiente de inclinação das curvas, que também surge como resultado da combinação dos valores estimados para as derivadas em cada ponto.

Analisando os resultados das duas abordagens adotadas, uma terceira alternativa foi avaliada: utilizando o DBSCAN, foram ignoradas as variáveis Dia e Sinal Teórico e, após algumas tentativas, foi escolhido o valor 0.127 para o parâmetro epsilon. O resultado apresentado nas Figuras 2.2 (e) e (f) parece representar de forma mais acurada a série, já que, apesar de haver uma certa simetria entre as senóides formadas pelos pontos, as discrepâncias não são simétricas e, portanto, têm influência na definição dos clusters e dos ruídos.

Assim, considerando os critérios previamente mencionados para a avaliação do processo de mineração de dados, o DBSCAN será a técnica mais adotada para dar continuidade ao estudo proposto, reservando o SimpleKMeans para conjuntos de dados com maior regularidade na forma. Esses dois métodos apresentaram facilidade de compreensão, resultados seguramente válidos para os dados utilizados, bom potencial de utilidade para os objetivos estabelecidos e uma certa originalidade nos resultados.

3 – ANÁLISE TEMPO-FREQUÊNCIA DE SINAIS

Uma conceituação já bastante popular estabelece que sinal é uma grandeza que contenha informação e que varie no tempo. Para sinais que sejam periódicos, uma das técnicas mais populares de análise é a utilização de Série de Fourier [Williams e Spangler, 1981], que é capaz de representar um sinal através de uma série de componentes senoidais de frequência harmônica. Conforme apresentado na Figura 3.1 abaixo [Williams e Spangler, 1981], os seis primeiros termos da série que representam o sinal "dente de serra" são: Y=f(t)=-sen(2πft)-1/2.sen(4πft)-1/3.sen(6πft)-1/4.sen(8πft)-1/5.sen(10πft)-1/6.sen(12πft).

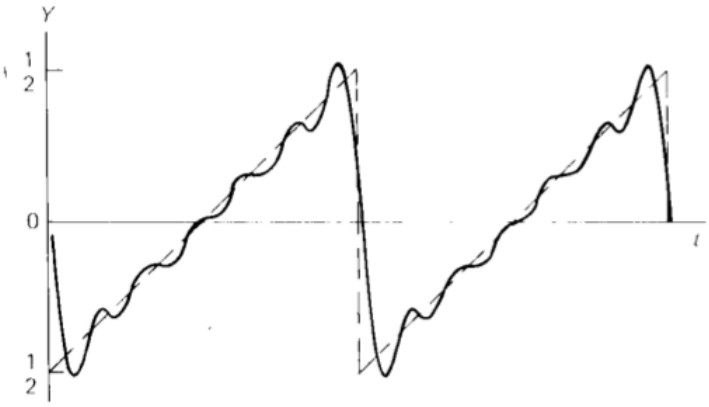


Fig. 3.1: Representação do "dente de serra" por Série de Fourier [Williams e Spangler, 1981].

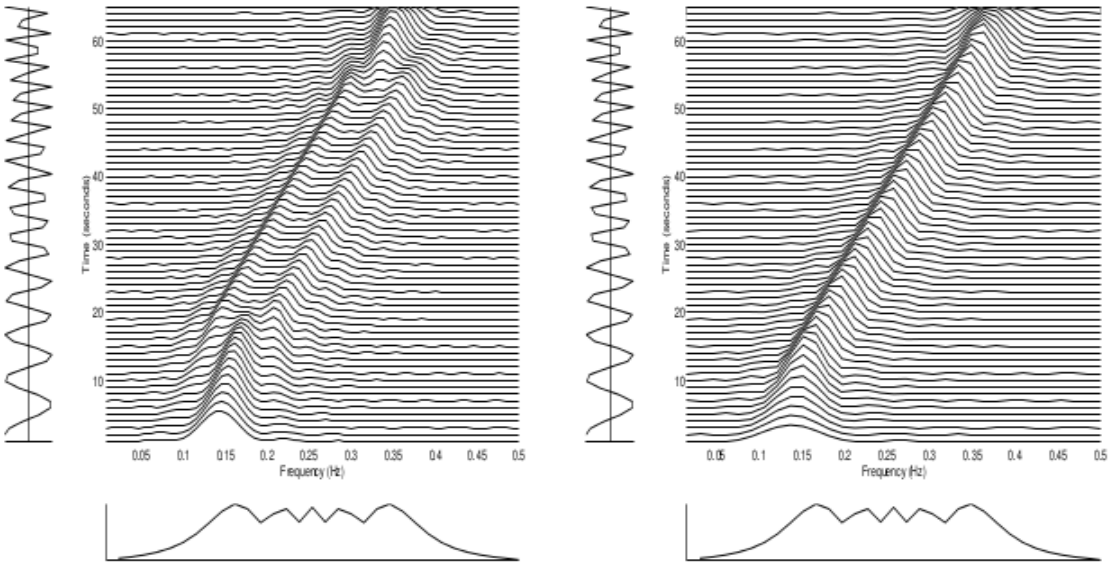
No entanto, em muitos casos práticos, como ocorre com os movimentos que representam a variação dos ativos em bolsas de valores, a frequência do sinal não é periódico e, portanto, a utilização de Séries de Fourier nesses casos é de pouca utilidade para a perfeita compreensão dos fenômenos.

De acordo com Boashash [1992], o processamento e a análise tempo-frequência de sinais compreende um conjunto de teorias e algoritmos utilizados para estudar sinais não estacionários ou, em outras palavras, sinais que carregam um conteúdo cuja frequência varia no tempo. Esse tipo de sinal é bem representado por uma distribuição tempo-frequência (ou DTF), que tem por objetivo apresentar como a energia do sinal estudado está distribuída pelo espaço bidimensional definido pelas dimensões tempo e frequência simultaneamente.

3.1 – Introdução à DTF – Distribuição Tempo-Frequência

As duas representações clássicas de um sinal são a representação no domínio do tempo s(t) e a representação no domínio da frequência S(f). Em ambos os casos, as variáveis tempo (t) e frequência (f) são consideradas mutuamente exclusivas. Consequentemente, cada uma dessas representações clássicas de um sinal é não localizada em relação à variável que foi excluída. Em outras palavras, a representação de frequência mostra essencialmente para cada valor da frequência a média das intensidades do sinal ao longo de todo o período de tempo e, por sua vez, a distribuição de tempo mostra essencialmente para cada instante de tempo a média das intensidades do sinal ao longo de toda a faixa de frequências. Na distribuição tempo-frequência as variáveis tempo (t) e frequência (f) estão juntas na mesma expressão $\rho(t,f)$, permitindo uma representação que seja localizada em relação a ambas as variáveis.

As Figuras 3.2 (a) e (b) mostram as três representações citadas (só em t, só em f e em ambas) para um mesmo sinal. Os dois gráficos tridimensionais em formato retangular mostram duas distribuições tempo-frequência do mesmo sinal que, como podemos observar, são bastante diferentes. Do lado esquerdo de cada imagem retangular, encontram-se as respectivas distribuições de cada sinal exclusivamente no domínio do tempo. Abaixo de cada imagem retangular, encontram-se as respectivas distribuições de cada sinal exclusivamente no domínio da frequência. Analisando as imagens, é possível constatar que o sinal tem representações idênticas quando apenas um dos domínios é utilizado, ou seja, excluindo a outra variável da representação. Isso ocorre porque as distribuições tempo-frequência envolvem um compromisso entre a resolução da representação no tempo e a resolução da representação na frequência. Esse problema pode ser denominado por otimização do tamanho de janela [Boashash, 1992], quando são utilizadas técnicas de janelamento de sinais para melhorar as características do sinal que está sendo analisado. Neste estudo, a avaliação da técnica de janelamento adotada no algoritmo utilizado na análise dos sinais foi essencialmente empírica, ou seja, foi feita com base na interpretação dos resultados obtidos.



(a) (b)

Figura 3.2: Duas DTF's de um mesmo sinal, utilizando o mesmo algoritmo [Boashash, 1992]: (a) com janela de tamanho 60% maior do que a ideal; (b) com janela do tamanho ideal.

3.2 – Estudos Realizados com a Aplicação de DTF's

Turhan-Sayan e Sayan [2001] avaliaram o potencial das DTF's, aplicando duas transformadas lineares (Transformada de Gabor e Transformada de Fourier de Tempo Curto) e duas quadráticas (*Wigner Distribution* e *Page Distribution*) ao longo de um período de dez anos de uma série temporal especial com base no índice da Bolsa de Valores de Istambul. Na estrutura do seu trabalho, os autores argumentam que uma série composta por preços de ações pode ser interpretada como uma combinação de tendência, vários ciclos e alguns ruídos. Além disso, é esperado que se encontre ciclos dominantes em diferentes frequências e que um deles venha a ser um ciclo econômico. Como o período é o inverso da frequência do ciclo (T=1/f), se o período varia de dois a quatro anos trata-se de um ciclo econômico curto e se o período chega a oito anos trata-se de um ciclo econômico longo.

Esses autores também concordam que uma série temporal pode ser analisada no domínio do tempo ou no domínio da frequência (ou do espectro), bastando para isso aplicar a transformada de Fourier para converter a representação do domínio do tempo para o domínio da frequência ou, no caso contrário, a transformada inversa de Fourier para converter a representação do domínio da frequência para o domínio do tempo. No entanto, eles reconhecem que, quando o conteúdo do espectro de um sinal varia com o tempo, a teoria convencional de Fourier não é suficiente para descrever de forma completa a contribuição de componentes de espectro arbitrariamente selecionados ao longo de certos períodos de tempo.

Assim sendo, concluem os autores, as técnicas para construção de representações tempo-frequência surgiram como uma solução viável para esse problema desafiador, já que são capazes de analisar um sinal nos domínios de tempo e frequência simultaneamente.

Aguiar-Conraria et al. [2008] aplicaram a transformada *wavelet* a várias séries temporais, com valores medidos em intervalos mensais ao longo de períodos entre os anos de 1920 e 2007, representando indicadores macroeconômicos europeus, tais como taxa de juros, índice de inflação, produção industrial, emissão de moeda e vários outros, com o objetivo de analisar a correlação entre os agregados monetários e a real atividade econômica da região. No artigo, apesar de reconhecerem a utilidade da análise de Fourier em vários contextos da literatura em economia, os autores argumentam que as séries temporais envolvendo dados econômicos são tipicamente não-estacionárias, além de complexas e repletas de ruídos. Para lidar com esse tipo de problema, foi introduzido nesses estudos o uso da transformada de Fourier de tempo curto, que não se mostrou muito eficiente devido ao fato da resolução da frequência ser a mesma para todos os valores de frequência das séries. Assim, como uma alternativa para melhorar o resultado da análise, o estudo propõe o uso da transformada *wavelet*. Esse tipo de técnica produz uma estimativa das características do espectro de uma série temporal como uma função do próprio tempo, revelando como os diferentes componentes periódicos da série temporal variam ao longo do tempo. Para atingir esse objetivo, a transformada *wavelet* expande a série temporal em versões defasadas e escalonadas de uma função – a chamada "*wavelet* mãe" – que possui tanto uma faixa de espectro quanto uma duração no tempo limitadas. De forma análoga, Aguiar-Conraria e Soares [2011] utilizam a análise *wavelet* para estudar a sincronização dos ciclos econômicos entre os países participantes da zona do euro. Baseados na transformada *wavelet*, os autores propõem uma métrica para medir e testar esses ciclos; questões envolvendo diferença de fase e janelamento são levantadas na utilização da transformada.

Em sua tese de mestrado, Nithin V George [2009] do *National Institute of Technology* – Rourkela, Índia, analisou vários ciclos econômicos através da aplicação da ST – *S Transform*, entre as quais destacamos as séries temporais mensais do índice DJIA – *Dow Jones Industrial Average* e do índice S&P 500 – *Standards and Poors 500*, ambos coletados no período entre 1950 e 2006, além de outras séries envolvendo taxas de desemprego nos Estados Unidos e preço do barril de petróleo no mercado mundial. O autor apresenta a transformada S como sendo uma versão híbrida entre a transformada de Fourier de tempo curto e a transformada *wavelet*. Em seguida, propõe uma versão aprimorada da transformada S, que oferece melhor resolução do que a sua versão original tanto no tempo quanto na

frequência. Essa melhoria foi alcançada a partir da introdução de um nova regra de escalonamento para a janela gaussiana utilizada na transformada S, fazendo com que o parâmetro de escalonamento, originalmente considerado constante, passasse a variar linearmente com a frequência do sinal. A análise experimental da nova versão proposta pelo autor mostrou que, para uma escolha adequada dos coeficientes angular e linear da função que estabelece o parâmetro, é importante levar em consideração o tipo e a natureza do sinal a ser analisado, já que o escalonamento da janela gaussiana tem uma forte influência na distribuição da energia do sinal no plano tempo-frequência.

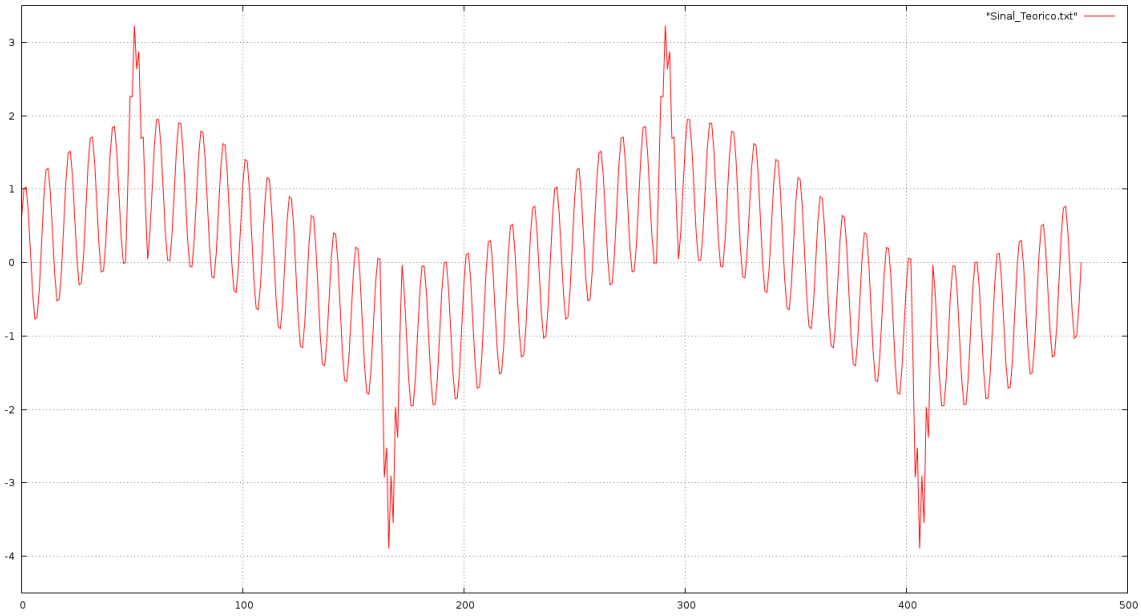
3.3 – STGT – *Signal-Tuned Gabor Transform*

A STGT [Torreão et al. 2013] é uma transformada que é capaz de gerar uma distribuição tempo-frequência de um sinal a ser analisado, desenvolvida com base em funções de Gabor e usando parâmetros fornecidos pela Transformada de Fourier aplicada ao mesmo sinal. A sua definição é dada pela seguinte equação:

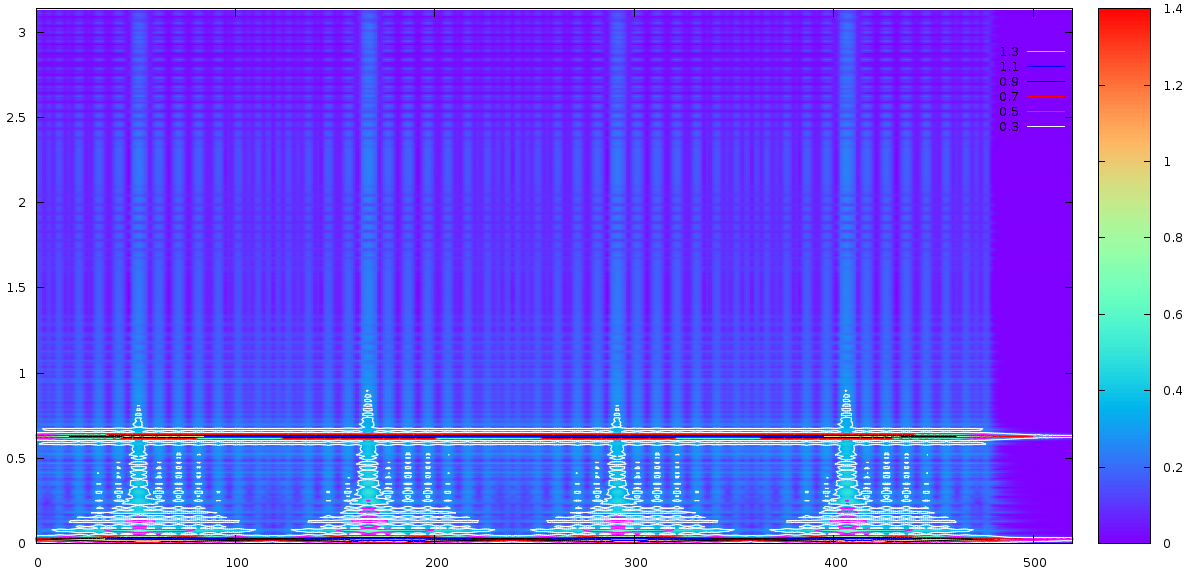
$$T_x(t,\omega) = \int_{-\infty}^{\infty} e^{-j[\omega\tau+\varphi(\omega)]} e^{-\frac{(\tau-t)^2}{2\sigma^2(\omega)}} x(\tau) d\tau.$$

No artigo publicado em 2013, foi apresentada uma análise experimental realizada com a STGT e as transformadas S e de Gabor, quando foram comparadas as distribuições tempo-frequência obtidas através da aplicação de cada uma delas a um determinado sinal. Após a avaliação dos resultados, a STGT demonstrou ser capaz de produzir uma distribuição tempo-frequência com uma melhor definição da localização das intensidades do sinal no tempo e na frequência simultaneamente, para todas as faixas de frequência e ao longo de todo o período de tempo. A melhor definição obtida com a STGT se deve exatamente à utilização dos parâmetros fornecidos pelo espectro de Fourier na aplicação da STGT, sendo a magnitude do espectro utilizada para definir a janela gaussiana $\sigma(\omega)$, que deixa de ser um parâmetro fixo como na transformada de Gabor convencional, e a frequência do espectro utilizada como ajuste de fase $\varphi(\omega)$ entre o sinal e o componente complexo da transformada.

O estudo experimental comparativo entre os diferentes tipos de transformada foi realizado através de uma versão discreta implementada por um programa codificado em C++, desenvolvida a partir da formulação da versão contínua da STGT [Torreão et al., 2013]. A Figura 3.3 a seguir, com as representações de um sinal teórico e sua distribuição tempo-frequência, bem como todas as demais distribuições tempo-frequência apresentadas adiante neste estudo foram obtidas através da utilização deste mesmo programa.

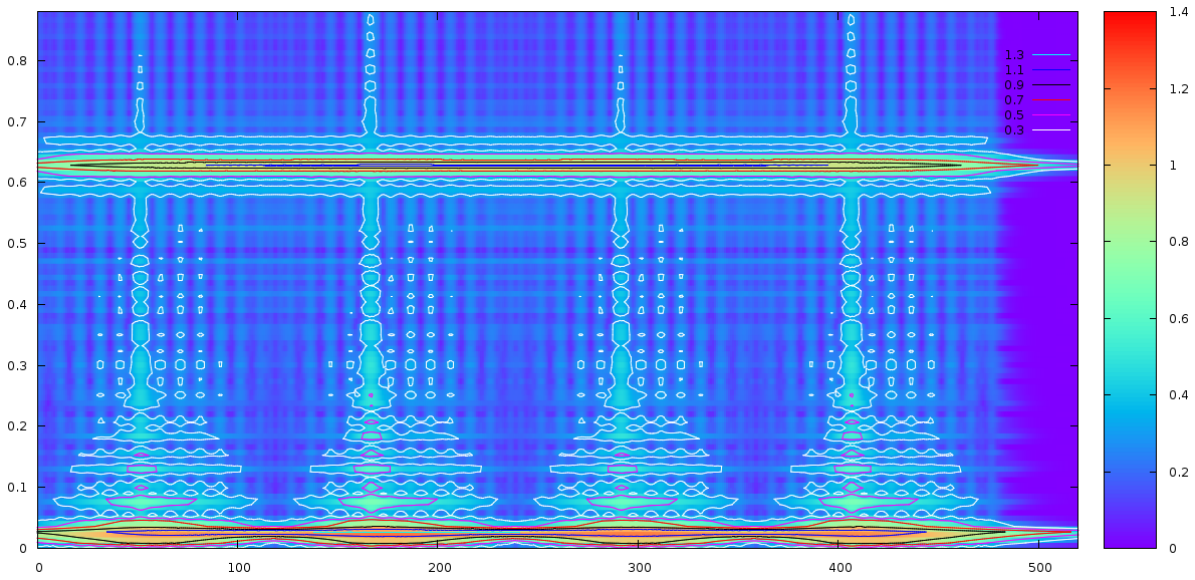


(a)

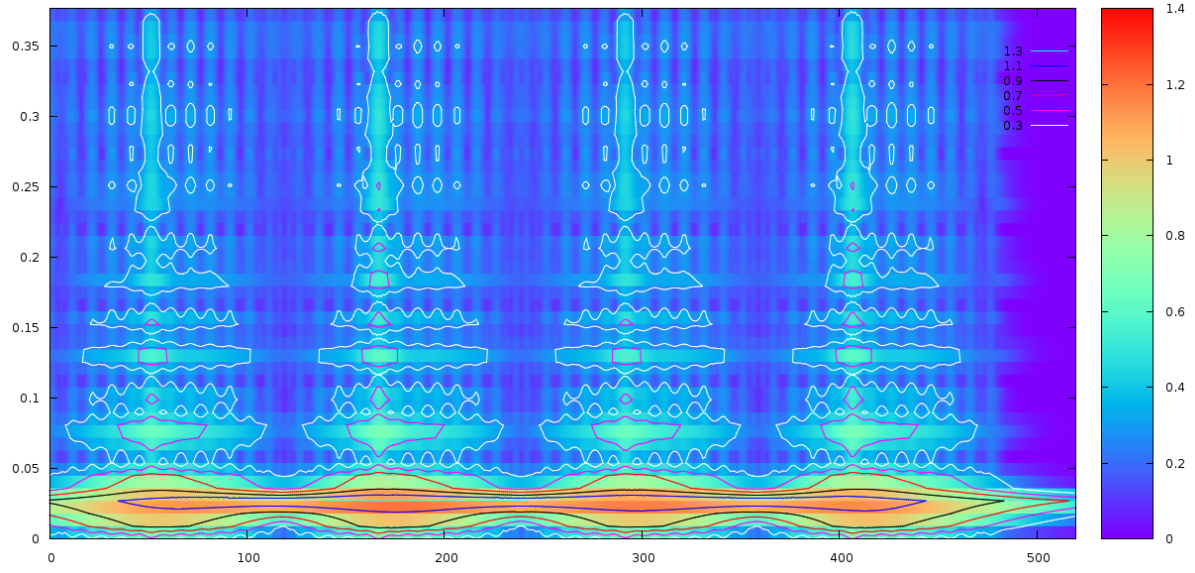


(b)

Figura 3.3: Representações de um sinal teórico e sua distribuição tempo-frequência: (a) x(t); (b) STGT com ω até 3.14; (c) STGT com ω até 0.88; (d) STGT com ω até 0.38.



(c)



(d)

Figura 3.3 (continuação)

O primeiro gráfico apresenta a mesma função criada para avaliação dos algoritmos de clusterização na seção 2.4. Assim, o gráfico da Figura 3.3 (a) mostra um Sinal Teórico (um valor adimensional) no eixo y em função do Dia no eixo x.

Os três outros gráficos apresentados logo em sequência mostram a distribuição tempo-frequência do sinal produzida com o algoritmo da STGT, que consiste em uma representação tridimensional que apresenta:

- no eixo x, a dimensão tempo, variando do dia 0 ao 520, com o objetivo de cobrir o período envolvendo as 480 instâncias e permitir a leitura da legenda no canto superior direito;
- no eixo y, a dimensão frequência, que pode apresentar qualquer variação de 0 a π, já que esse é o intervalo de frequências com o qual o algoritmo trabalha;
- para representar o eixo z, uma escala de cores, do lado direito do gráfico, e um conjunto de linhas de contorno (ou curvas de nível), situado no canto superior direito do gráfico, que significam a intensidade do sinal localizado simultaneamente no tempo e na frequência; as regiões com maior intensidade, indicando uma discrepância na distribuição, serão tratados como “aspectos”.

Os valores máximos da frequência no eixo y foram alterados nas Figuras 3.3 (b), (c) e (d) para permitir uma melhor visualização dos componentes de frequência constante do sinal e das quatro discrepâncias. Assim, é possível notar através da análise dos gráficos:

- nas Figuras 3.3 (b) e (c), um componente com formato totalmente uniforme, de frequência constante ao longo do tempo com ω=0.628, referente à senoide com período igual a dez dias que compõe o sinal;
- nas Figuras 3.3 (c) e (d), um componente com formato uniforme na intensidade 1.1, de frequência variando entre ω=0.02 e ω=0.03 ao longo do tempo, referente à senoide com período igual a 240 dias que compõe o sinal;
- nas Figuras 3.3 (b), (c) e (d), quatro momentos ao longo do tempo com formato bastante irregular, com intensidade 0.3 em frequências até ω=0.9 e intensidade de 0.5 a 0.9 em frequências variando de zero a ω=0.3, referente às discrepâncias.

Pela avaliação realizada através da função que representa o sinal teórico, a STGT apresentou um grande potencial para identificar discrepâncias em uma série temporal, evidenciando características estruturais da série que são de difícil observação quando analisadas exclusivamente no domínio temporal.

4 – ANÁLISE EXPLORATÓRIA

De acordo com vários autores [Aghabozorgi et al., 2015; Keogh e Lin, 2004; Košmelj e Batagelj, 1990], além de ser uma sub-rotina em algoritmos mais complexos de mineração de dados, a clusterização é a abordagem mais utilizada como técnica exploratória de séries temporais, sendo utilizada para a descoberta de regras de associação, na classificação e na detecção de anomalias ou discrepâncias.

Assim, após dar início à etapa empírica deste estudo através de um processo de clusterização com o uso do DBSCAN e avaliar os resultados alcançados, foi incorporado o uso da STGT como técnica de análise tempo-frequência. Mais adiante, a identificação de pontos de reversão de tendência completou esta etapa.

4.1 – Componentes do Processo de Clusterização

Nesta seção serão apresentados os componentes metodológicos envolvidos no processo de clusterização de uma série do IBOVESPA, composta por 492 registros diários referentes ao período de 16/01/2012 a 13/01/2014. É importante ressaltar que, como o processo de análise foi conduzido essencialmente de forma empírica, algumas decisões tomadas ao longo do processo considerando um determinado critério resultaram na escolha de um outro componente, sem que esse último tivesse sido formalmente analisado (ex.: a escolha de um determinado algoritmo acarretou na adoção de uma medida de similaridade). A ordenação da apresentação dos componentes está de acordo com a sequência de decisões tomadas durante o processo de análise dos dados.

Considerando os trabalhos existentes sobre o tema, Aghabozorgi et al. [2015] identificaram os quatro principais componentes metodológicos envolvidos no processo de clusterização de séries temporais que, apesar de terem sido identificados com base na análise

de trabalhos envolvendo a clusterização de séries completas, também podem ser considerados para as demais abordagens:

- redução de dimensionalidade (ou método de representação) da série temporal;
- seleção das medidas de similaridade ou distância;
- construção e avaliação do protótipo da série temporal;
- aplicação do(s) algoritmo(s) e avaliação dos resultados.

Segundo os autores, os processos de clusterização de séries temporais que são geralmente executados costumam envolver alguns ou todos esses quatro componentes. Tendo como referência o objetivo inicialmente estabelecido, utilizaremos informações diárias sobre as operações realizadas na Bolsa de Valores de São Paulo a partir do ano de 2012, que podem ser facilmente obtidas por qualquer investidor através do *site* www.infomoney.com.br . Esse *site* fornece uma planilha contendo os seguintes dados:

- Data;
- Histórico (em pontos para os índices e em R$ para os ativos);
- Fechamento (em pontos para os índices e em R$ para os ativos);
- Variação Dia (%);
- Abertura (em pontos para os índices e em R$ para os ativos);
- Mínimo (em pontos para os índices e em R$ para os ativos);
- Médio (em pontos para os índices e em R$ para os ativos);
- Máximo (em pontos para os índices e em R$ para os ativos);
- Volume (em R$);
- Negócios (em número de transações).

A esses dados foram acrescentadas as seguintes relações:

- Máximo/Mínimo (%), que representa o tamanho da banda de variação do índice ou do preço do ativo entre a abertura e o fechamento; é importante destacar que, diferentemente da variação no dia, esse valor é sempre positivo;
- Volume/Negócios (em R$), que representa o valor médio das transações realizadas ao longo do dia, seja para toda a carteira que compõe um índice ou para um ativo isoladamente.

A seguir, são apresentadas as alternativas adotadas para cada um dos componentes do processo de clusterização envolvendo a série temporal deste estudo.

4.1.1 – Escolha e Aplicação do Algoritmo

O primeiro passo do processo consistiu em utilizar ferramenta WEKA, acessar a aplicação Weka Explorer, e importar os dados da série com o objetivo de analisá-los através da matriz apresentada na aba Visualize. Antes disso, algumas pequenas alterações foram realizadas:

- a coluna Data foi transformada em um sequencial inteiro, pois a aplicação não suporta os tradicionais formatos de data;
- a coluna Histórico foi eliminada por ser igual ao Fechamento para toda a série;
- a coluna Médio foi desconsiderada, por não retratar um valor do índice em um instante de tempo específico.

A matriz gráfica que representa as relações entre as dez variáveis que formam a série está representada na Figura 4.1 a seguir. Considerando que a série envolve dez variáveis, a figura apresenta uma matriz com dez linhas e dez colunas, na qual cada variável é relacionada com ela mesma, apresentando o formato da reta y=x, e as outras nove variáveis, apresentando formatos originalmente desconhecidos. A avaliação da matriz constatou a diversidade de formatos obtidos através das relações entre as variáveis da série. Esse fato levou à utilização do DBSCAN como algoritmo de clusterização, tendo em vista que os métodos de clusterização baseados em densidade são particularmente úteis para identificar clusters de formato irregular ou arbitrário e detectar ruídos que não pertencem a nenhum cluster [Ester et al., 1996]. Conforme mencionado anteriormente, essa decisão teve impacto na escolha da primeira medida de similaridade usada na análise, como veremos mais adiante.

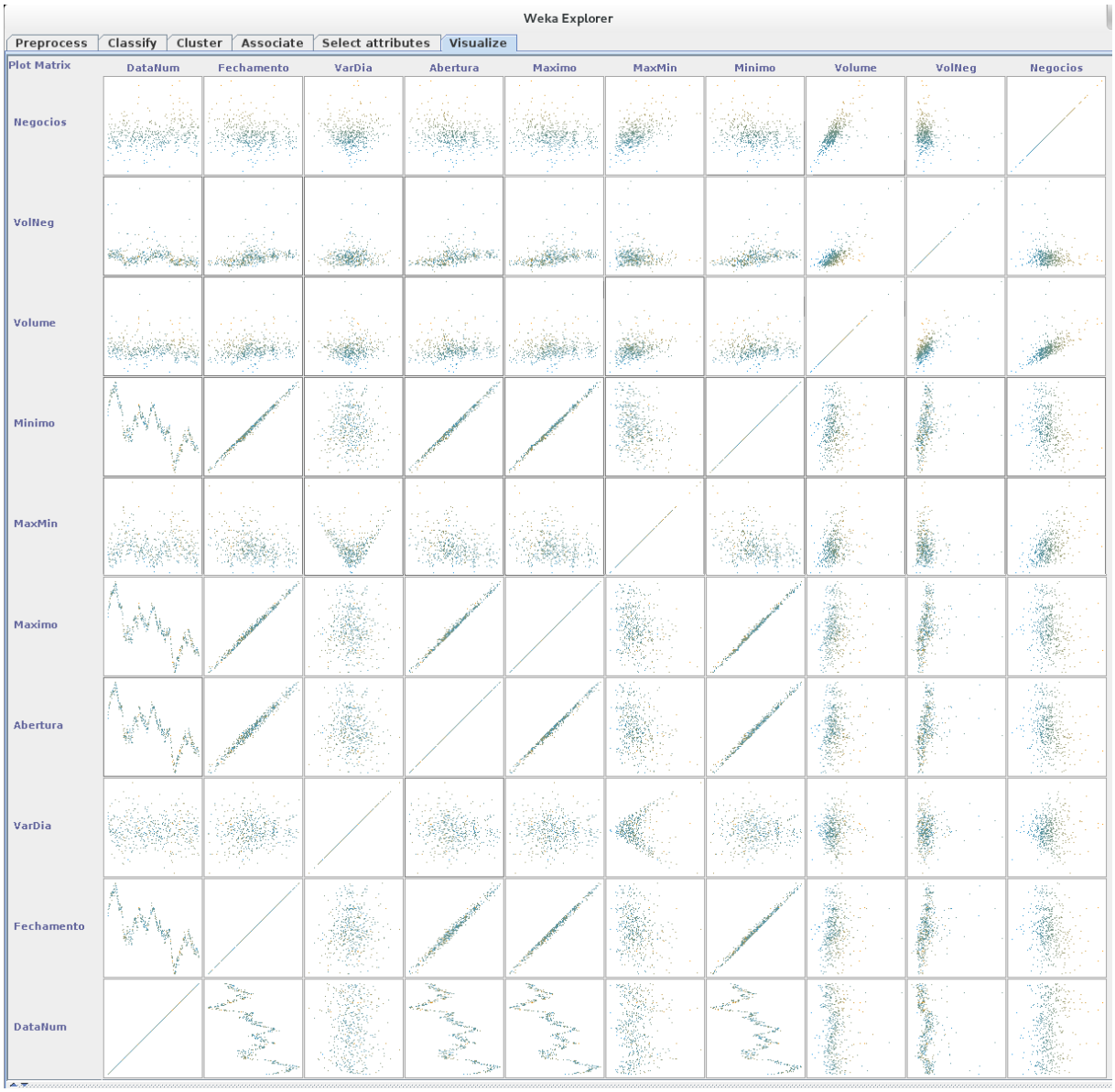


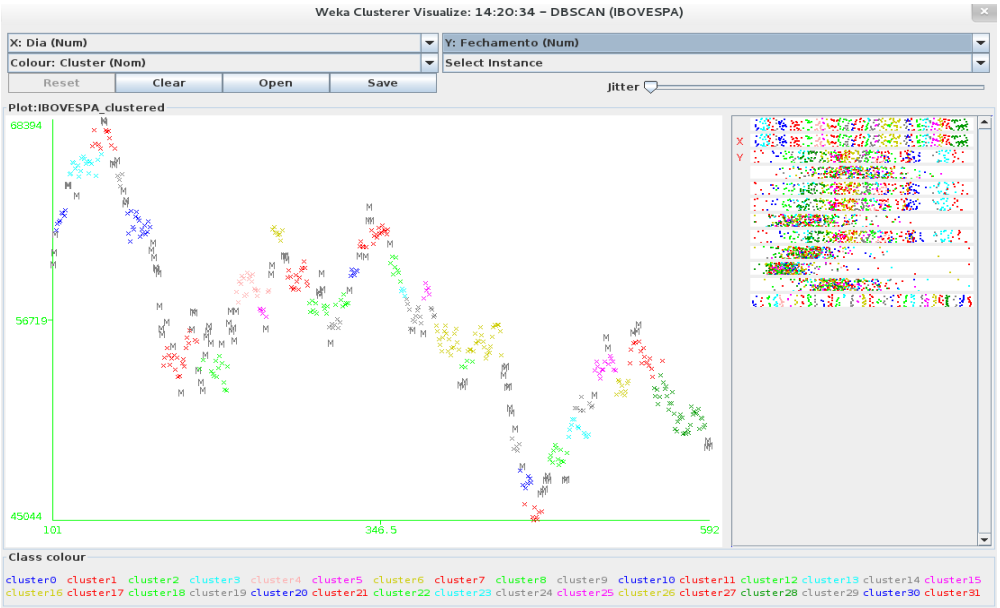
Figura 4.1: Matriz gráfica para visualização das relações entre as variáveis da série temporal.

4.1.2 – Redução da Dimensionalidade

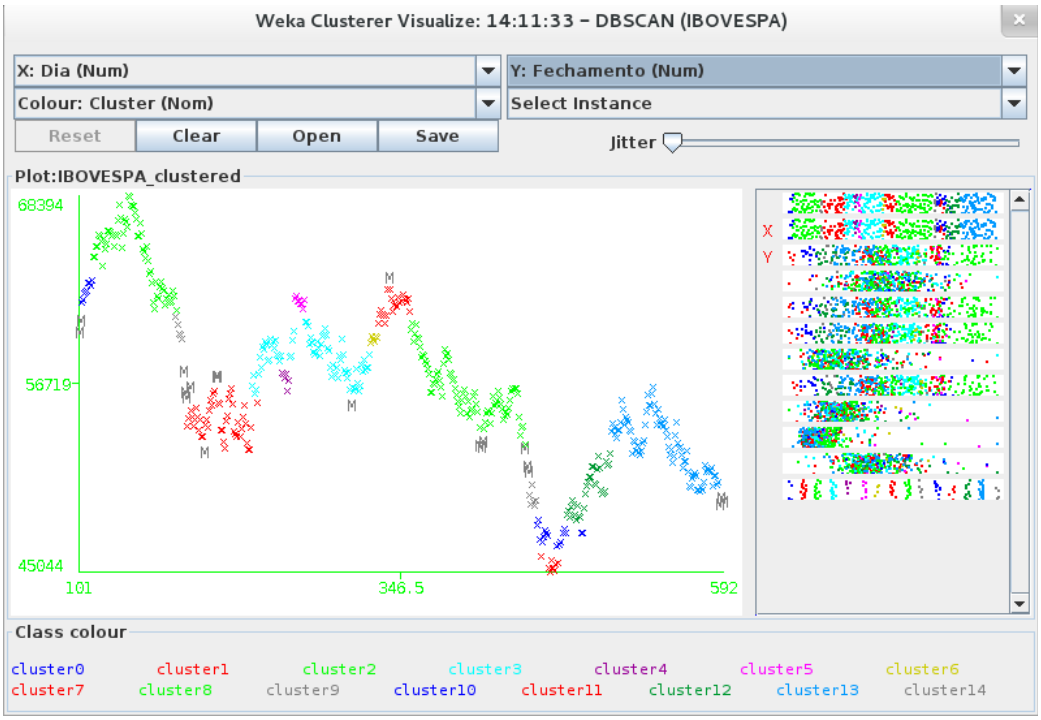
O DBSCAN é um algoritmo que funciona com a escolha de dois parâmetros: o raio em torno de um ponto no espaço usado pelo algoritmo para a busca de outros pontos da série, denominado epsilon, e o número de mínimo de pontos que deve ser encontrado em uma região do espaço definida pelo raio (epsilon), denominado minPoints. A aparente simplicidade do conceito envolvido encobre a dificuldade que os usuários, de uma forma geral, costumam encontrar para obter resultados satisfatórios com a aplicação do algoritmo.

Devido a essa complexidade inerente à utilização do algoritmo DBSCAN, principalmente em conjuntos de dados compreendendo várias dimensões, foram adotados duas medidas simplificadoras apresentadas pelos próprios criadores do algoritmo na seção do artigo que discute a determinação dos parâmetros a serem utilizados [Ester et al., 1996]: primeiramente, na definição do parâmetro minPoints, "*we eliminate the parameter MinPts by setting it to 4 for all databases (for 2-dimensional data)*" e, adicionalmente, no ajuste do parâmetro epsilon inicialmente arbitrado, "*the user can estimate the percentage of noise*" para avaliar os resultados obtidos. Assim, após várias tentativas de clusterização envolvendo mais de duas dimensões, todas elas apresentando um percentual inaceitável de pontos classificados como ruído, o estudo passou a considerar processos de clusterização envolvendo apenas duas dimensões, mantendo sempre o parâmetro minPoints=4 e ajustando o parâmetro epsilon de forma a obter menos de 20% (vinte por cento) do total de pontos da série classificado como ruído.

É importante ressaltar que a redução de dimensionalidade acima descrita foi uma consequência natural da escolha do algoritmo utilizado. Os dois resultados preliminares mais significativos do processo de clusterização executado estão representados a seguir através das Figuras 4.2 (a) e (b).



(a)



(b)

Figura 4.2: Resultados preliminares da clusterização da série do IBOVESPA com o DBSCAN: (a) epsilon=0.020 – 93 pontos como ruído; (b) epsilon=0.030 – 20 pontos como ruído.

Košmelj e Batagelj [1990] introduziram uma abordagem para a clusterização de séries que variam no tempo baseada na ideia de incorporar a dimensão temporal no processo de definição da dissimilaridade. Usando a mesma ideia na clusterização dos pontos da série do IBOVESPA, na Figura 4.2 (a) é apresentado o resultado envolvendo as dimensões Dia e Fechamento com o menor valor do parâmetro epsilon que ainda é capaz de atender à condição de um máximo de 20% de pontos de ruído, que são representados no gráfico pelas várias letras M (*Missing*). Em seguida, foram analisados os resultados com valores do parâmetro variando de 0.021 a 0.031. A Figura 4.2 (b) apresenta o resultado com o maior valor do parâmetro epsilon que ainda apresenta o final da série temporal analisada formada por três pontos como ruído, ou seja, é o resultado com o menor número de clusters que ainda representa o final da série com as mesmas características da alternativa na qual o percentual de pontos classificado como ruído atingiu o limite da tolerância estabelecida.

A representação com menor número de clusters, que também apresenta menor número de pontos como ruído, torna as discrepâncias mais significativas no contexto da série. Por discrepância podemos denominar os próprios ruídos e as mudanças de um cluster para outro. A análise e a interpretação dessas discrepâncias será abordada a seguir.

4.1.3 – Seleção das Medidas de Similaridade

A escolha do DBSCAN como técnica exploratória da série temporal por questões envolvendo a distribuição dos pontos da série no espaço e, consequentemente, o formato das regiões a serem investigadas, resultou na adoção da primeira medida de similaridade utilizada na análise da série.

Mas, enfim, seriam os três pontos encontrados ao final da série temporal, representada pela clusterização com o DBSCAN, suficientes para caracterizar o seu estado naquele momento? Seria esse fenômeno algo característico do que estaria por vir? Como podemos observar pelo gráfico da Figura 4.2 (a), ruídos podem ser encontrados em máximos e mínimos locais, mas não necessariamente em todos eles, e em algumas transições entre clusters nos trechos da série que apresentam forte tendência de elevação ou de queda. Certamente carregam informação relevante, porém não o suficiente para produzir conclusões de muita profundidade. Com isso, foi decidido aplicar uma outra medida de similaridade ainda nessa fase exploratória inicial.

Alguns estudos envolvendo clusterização de séries temporais com muitas variáveis assumem não existir correlação entre as variáveis da série [Liao, 2005]. Não é o caso neste estudo em particular pois, como podemos observar na matriz gráfica apresentada

anteriormente na Figura 3.1, algumas variáveis apresentam forte correlação linear entre elas (ex.: entre o Fechamento e o Mínimo, entre o Fechamento e o Máximo, etc.), já que algumas células da matriz apresentam uma distribuição de pontos que se aproxima da reta y=x. Dessa forma, a escolha de uma nova medida de similaridade recaiu sobre o coeficiente de correlação de Pearson, também denominada de coeficiente de correlação produto-momento [Rodgers e Nicewander, 1988], representado pela seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

O coeficiente de Pearson é uma função estatística encontrada na grande maioria dos *softwares* disponíveis no mercado, desde planilhas eletrônicas até sofisticados ambientes de programação como o MATLAB. Os resultados mais significativos obtidos através da aplicação do coeficiente de correlação de Pearson à série temporal com dados do IBOVESPA é apresentada a seguir na Tabela 4.1.

Esse coeficiente mede o grau de correlação entre variáveis de escala intervalar, assumindo valores entre –1 e +1:

- $\rho=+1$ indica correlação perfeita positiva;
- $\rho=0$ indica que x e y não variam linearmente uma com a outra;
- $\rho=-1$ indica correlação perfeita negativa.

Na prática, a interpretação do valor do coeficiente de correlação de Pearson deve se orientar pelas seguintes faixas de valores, sejam eles positivos ou negativos:

- $\rho>70\%$ indica forte correlação;
- $30\%<\rho<70\%$ indica correlação moderada;
- $\rho<30\%$ indica fraca correlação.

É importante enfatizar que a adoção do coeficiente de Pearson como medida de similaridade nesta etapa exploratória do estudo foi determinante para a sua boa evolução, mas esse coeficiente não possui grande relevância para o modelo de análise proposto, as aplicações do modelo no estudo de caso desenvolvido e a conclusão final deste trabalho. A correta aplicação do coeficiente de Pearson exige a análise da significância da correlação entre as variáveis e só costuma ser aceito como parte de uma distribuição normal com amostras envolvendo pelo menos trinta elementos.

Tabela 4.1: Coeficiente de correlação de Pearson entre as variáveis da série do IBOVESPA.

GERAL

	Fecham	VarD (%)	Abertura	Máximo	M/m (%)	mínimo	Volume	Vol/Neg
VarD (%)	0,085							
Abertura	0,989	-0,061						
Máximo	0,996	0,014	0,996					
M/m (%)	-0,256	-0,039	-0,251	-0,217				
mínimo	0,996	0,016	0,996	0,997	-0,293			
Volume	0,129	0,063	0,119	0,136	0,264	0,112		
Vol/Neg	0,339	0,052	0,331	0,334	-0,107	0,336	0,680	
Negócios	-0,184	0,033	-0,189	-0,169	0,489	-0,204	0,649	-0,098

Cluster 2 – do dia 35 ao dia 42

	Fecham	VarD (%)	Abertura	Máximo	M/m (%)	mínimo	Volume	Vol/Neg
VarD(%)	0,151							
Abertura	0,661	-0,642						
Máximo	0,963	-0,073	0,803					
M/m(%)	0,505	0,870	-0,270	0,319				
mínimo	0,756	-0,502	0,968	0,885	-0,158			
Volume	0,592	0,726	-0,088	0,456	0,791	0,087		
Vol/Neg	-0,090	0,130	-0,166	-0,186	-0,069	-0,158	0,212	
Negócios	0,630	0,559	0,067	0,570	0,758	0,221	0,803	-0,408

Cluster 1 – do dia 121 ao dia 128

	Fecham	VarD (%)	Abertura	Máximo	M/m (%)	mínimo	Volume	Vol/Neg
VarD(%)	0,576							
Abertura	0,330	-0,581						
Máximo	0,767	-0,052	0,826					
M/m(%)	-0,478	-0,745	0,384	-0,065				
mínimo	0,837	0,097	0,724	0,980	-0,260			
Volume	-0,472	-0,622	0,244	-0,021	0,573	-0,133		
Vol/Neg	-0,340	-0,544	0,288	0,052	0,606	-0,069	0,878	
Negócios	-0,408	-0,357	0,002	-0,145	0,110	-0,162	0,537	0,070

Cluster7 – do dia 242 ao dia 249

	Fecham	VarD (%)	Abertura	Máximo	M/m (%)	mínimo	Volume	Vol/Neg
VarD(%)	0,408							
Abertura	0,335	-0,724						
Máximo	0,844	0,090	0,548					
M/m(%)	0,217	0,764	-0,623	0,124				
mínimo	0,706	-0,256	0,800	0,897	-0,328			
Volume	-0,433	-0,068	-0,261	-0,261	0,208	-0,341		
Vol/Neg	-0,348	-0,326	0,066	-0,360	-0,301	-0,209	0,691	
Negócios	-0,094	0,329	-0,404	0,143	0,615	-0,137	0,313	-0,470

Cluster11 – do dia 357 ao dia 364

	Fecham	VarD (%)	Abertura	Máximo	M/m (%)	mínimo	Volume	Vol/Neg
VarD(%)	0,381							
Abertura	0,748	-0,329						
Máximo	0,871	-0,056	0,931					
M/m(%)	-0,607	-0,821	-0,029	-0,199				
mínimo	0,977	0,339	0,755	0,888	-0,627			
Volume	0,049	-0,616	0,489	0,264	0,372	0,039		
Vol/Neg	0,215	-0,066	0,270	0,212	-0,103	0,221	0,647	
Negócios	-0,025	-0,705	0,476	0,232	0,498	-0,045	0,963	0,421

Final da Série – do dia 485 ao dia 492

	Fecham	VarD (%)	Abertura	Máximo	M/m (%)	mínimo	Volume	Vol/Neg
VarD(%)	0,448							
Abertura	0,494	-0,556						
Máximo	0,566	-0,414	0,932					
M/m(%)	-0,517	-0,860	0,360	0,346				
mínimo	0,922	0,170	0,691	0,768	-0,336			
Volume	0,235	0,335	-0,109	-0,089	-0,119	-0,005		
Vol/Neg	0,155	0,600	-0,445	-0,454	-0,601	-0,043	0,720	
Negócios	0,140	-0,380	0,504	0,549	0,659	0,101	0,340	-0,403

Primeiramente, foi desenvolvida uma matriz que permite correlacionar cada uma das nove variáveis envolvidas com as outras oito demais variáveis (a variável que representa a dimensão tempo foi excluída dessa análise), ao longo das 492 instâncias da série, resultando em um total de 36 valores de correlação, representados na primeira matriz da Tabela 4.1, com título “Geral”. Esse resultado permitiu confirmar a forte correlação linear positiva entre algumas variáveis, como entre o Fechamento e o Mínimo, entre o Fechamento e o Máximo, etc., e determinar o nível de correlação entre as demais variáveis. Os casos extremos, como de forte correlação positiva ou negativa e de correlação fraca, são importantes para comparação com os resultados do próximo passo na análise.

Em seguida, a matriz originalmente criada para trabalhar com a série completa foi alterada de forma a obter a correlação entre as mesmas variáveis ao longo de uma sequência de apenas sete dias, permitindo que a matriz percorresse toda a série como uma “janela” que se move ao longo da dimensão temporal. O principal objetivo era o de analisar as correlações entre as mesmas variáveis nos intervalos de tempo próximos aos máximos e mínimos locais da série. Os resultados mais importantes para o estudo são apresentados nas cinco últimas matrizes da Tabela 4.1, identificadas com o número do cluster apresentado na figura 4.2 (b) e o intervalo de tempo em que os resultados foram obtidos.

É importante notar que a correlação entre a Variação Diária e a relação Máximo/Mínimo, quando computada para toda a série, apresentou resultado próximo de zero. No entanto, em intervalos de sete dias próximos aos trechos que representam máximos locais, essas mesmas variáveis apresentaram forte correlação positiva, enquanto em intervalos de sete dias próximo aos mínimos locais, elas apresentaram forte correlação negativa. Curiosamente, os sete últimos pontos da série, com quatro pontos pertencentes ao Cluster 13 e outros três pontos de ruído, também apresentaram forte correlação negativa entre as variáveis mencionadas. É importante deixar claro que em outros trechos compreendendo máximos e mínimos locais ao longo da série a avaliação dessas mesmas variáveis também apresentou correlação forte, ainda que esse resultado não tenha sido verificado em todos os trechos envolvendo máximo ou mínimo local. Além disso, como será explicado mais adiante, também foram identificados alguns trechos de forte tendência de elevação e de forte tendência de queda do IBOVESPA nos quais as variáveis analisadas também apresentaram forte correlação positiva e forte correlação negativa respectivamente.

4.1.4 – Construção do Protótipo

De forma semelhante à abordagem adotada por Keogh e Pazzani [1998], conforme ilustrada

na figura 4.3 a seguir e demonstrada pelos autores a partir de aplicações com dados médicos, dados de telemetria espacial e dados sintéticos, foi desenvolvido um protótipo da série temporal formada pelos dados do IBOVESPA. Em seu artigo, Keogh e Pazzani representaram as séries temporais graficamente através de segmentos de reta, que indicam a variação da série, associados a “blocos”, que indicam os pesos de cada segmento de reta.

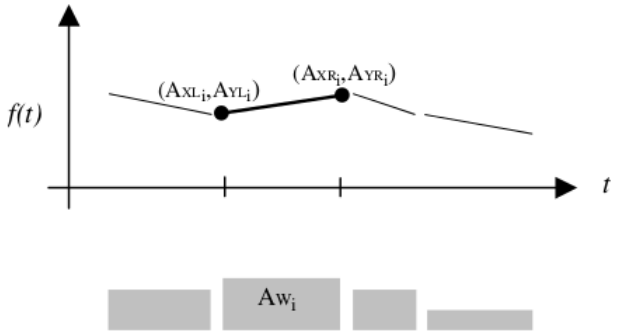


Figura 4.3: Exemplo de protótipo de uma série temporal [Keogh e Pazzani, 1998].

Aghabozorgi et al. [2015] argumentam que, especialmente nos processos de clusterização envolvendo algoritmos de particionamento, a qualidade dos clusters obtidos é altamente dependente da qualidade do protótipo desenvolvido para a série. Assim, ainda que os métodos de particionamento sejam apenas uma das possíveis alternativas para extração de conhecimento da série temporal em estudo, o desenvolvimento do protótipo foi realizado com o objetivo de contribuir na validação dos resultados preliminares obtidos, auxiliar na compreensão da evolução da série temporal e fundamentar as decisões futuras envolvendo o desenvolvimento da sua análise.

Tendo como base os resultados da clusterização com o uso do DBSCAN e os indícios gerados com a aplicação do coeficiente de correlação linear de Pearson, a opção natural pareceu ser a da representação conjunta da Variação Diária (em %) com a relação entre Máximo/Mínimo (em %). Com isso, considerando que a primeira instância da série temporal representaria o estado inicial da série ou a referência de variação zero, foi produzido o gráfico apresentado na Figura 4.4 (a), cuja legenda se encontra no canto superior direito.

Separando as ocorrências diárias da relação Máximo/Mínimo dos correspondentes valores da Variação Diária, é possível produzir um gráfico de barras conforme apresentado na Figura 4.4 (b), que se assemelha à representação de um sinal estritamente positivo. Essa representação carrega consigo alguns significados interessantes para a análise desse tipo de série temporal.

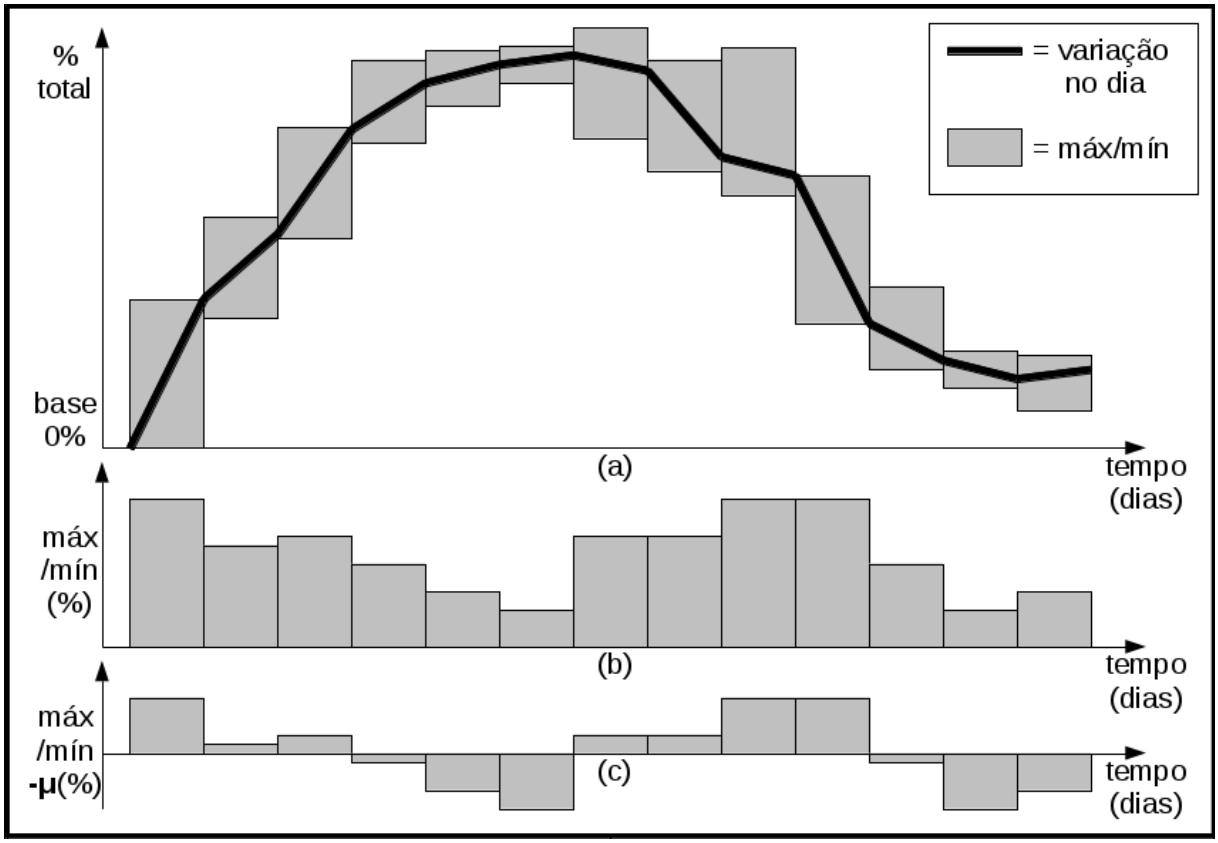


Figura 4.4: Protótipo da série do IBOVESPA e suas representações no tempo: (a) variação diária e relação máximo/mínimo; (b) relação máximo/mínimo; (c) máximo/mínimo–μ.

Quando se procura um determinado item para comprar em algumas lojas de uma região da cidade, em alguns *stands* de venda em um mercado ou mesmo em algumas barracas de uma feira livre, espera-se encontrar variações de preço entre as diferentes ofertas do mesmo item. As variações podem ocorrer entre os diferentes locais em que se busca o item, mas também podem ocorrer ao longo do tempo, mesmo que o intervalo da pesquisa de preços seja de um único dia. Se forem considerados apenas aqueles itens que se costuma comprar diariamente, ainda assim é esperado encontrar variações de preço e, além disso, cada item costuma apresentar uma variação esperada, como se essa fosse uma característica sua em partiular. Nos dias em que a variação entre os preços ofertados para um item excede em muito aquilo que costuma ser encontrado nos mercados, os compradores costumam estranhar o fato e reclamar. Isso não quer dizer que compras desse item não sejam realizadas, mas que certamente há algo de diferente no cenário. Mesmo quando a variação é muito aquém da esperada, os compradores costumam estranhar o ocorrido, muito embora nesse caso não demonstrem uma reação muito clara sore o fato.

Foi com base nas reflexões acima comentadas que uma segunda representação da relação Máximo/Mínimo foi adotada, dessa vez subtraindo a média μ dos valores dessa relação, calculada para toda a série temporal, do valor da relação em cada instância da série. O resultado dessa nova representação é apresentado na Figura 4.4 (c), que se assemelha à representação de um novo sinal, que dessa vez assume valores positivos e negativos. Com isso, a série temporal do IBOVESPA passa a ser representada pelo protótipo composto pelos três gráficos apresentados na Figura 4.4.

4.2 – Avaliação dos Resultados do Processo

A identificação de fenômenos associados à série temporal através da segunda medida de similaridade adotada, o coeficiente de correlação linear de Pearson, desencadeou a busca por novas técnicas de mineração de dados em séries temporais, de forma a permitir, também na clusterização de pontos, a detecção de características análogas às que Aghabozorgi et al. [2015] denominaram como "*finding similar time-series in change (structural similarity)*" na clusterização de séries temporais completas.

As mudanças que são alvo de estudos a respeito de séries temporais costumam se repetir ao longo do tempo, mas não necessariamente com o mesmo aspecto ou o mesmo formato. Na verdade, as mudanças são as mesmas: uma tendência de elevação que se torna um máximo local e, em seguida, se transforma em uma tendência de queda; uma tendência de queda que se torna um mínimo local e, mais tarde, passa a apresentar tendência de elevação; dentre outras mais. No entanto, essas mudanças semelhantes não ocorrem porque em algum trecho da série os formatos dos gráficos apresentam semelhança. Em outras palavras, máximos são semelhantes entre si, assim como mínimos ou pontos de inflexão; quanto a isso não resta dúvida. A grande questão é o que leva à formação de um máximo, um mínimo ou um ponto de inflexão.

Tendo como motivação as questões acima apresentadas, o estudo passou a também fazer uso de técnicas com a capacidade de extrair informações que pudessem caracterizar a própria formação da série temporal, ou seja, que definissem a sua estrutura.

4.3 – Resultados da Aplicação da STGT

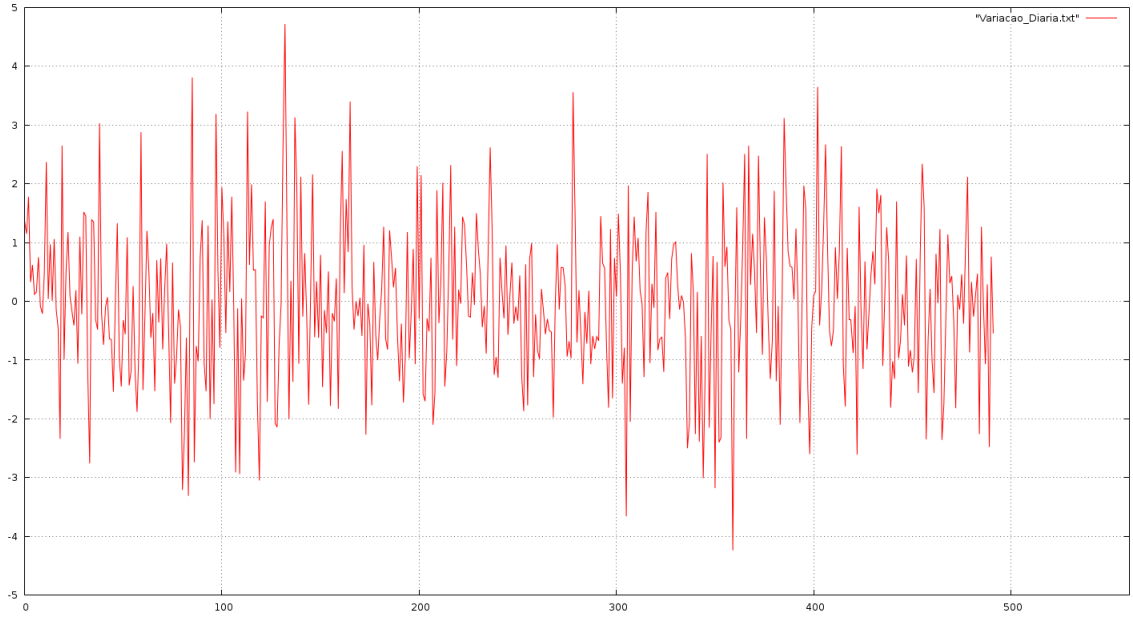
Antes da interpretação dos gráficos produzidos, é importante reforçar o padrão adotado na elaboração e apresentação dos mesmos, principalmente no caso da distribuição tempo-frequência, que voltará a ser apresentada nos próximos capítulos. Todos os gráficos foram elaborados com a utilização do *software* GNUPLOT *Version 4.6 patchlevel 1*.

Em cada par de gráficos, o primeiro apresenta no eixo x a dimensão tempo e no eixo y a amplitude do sinal. Assim, para o sinal da Variação Diária o eixo da amplitude varia de –5 a 5% e para a relação Máximo/Mínimo–μ o eixo da amplitude varia de –2 a 4%. Em ambos os casos, o eixo do tempo varia de 0 a 550, de forma a apresentar as 492 instâncias da série.

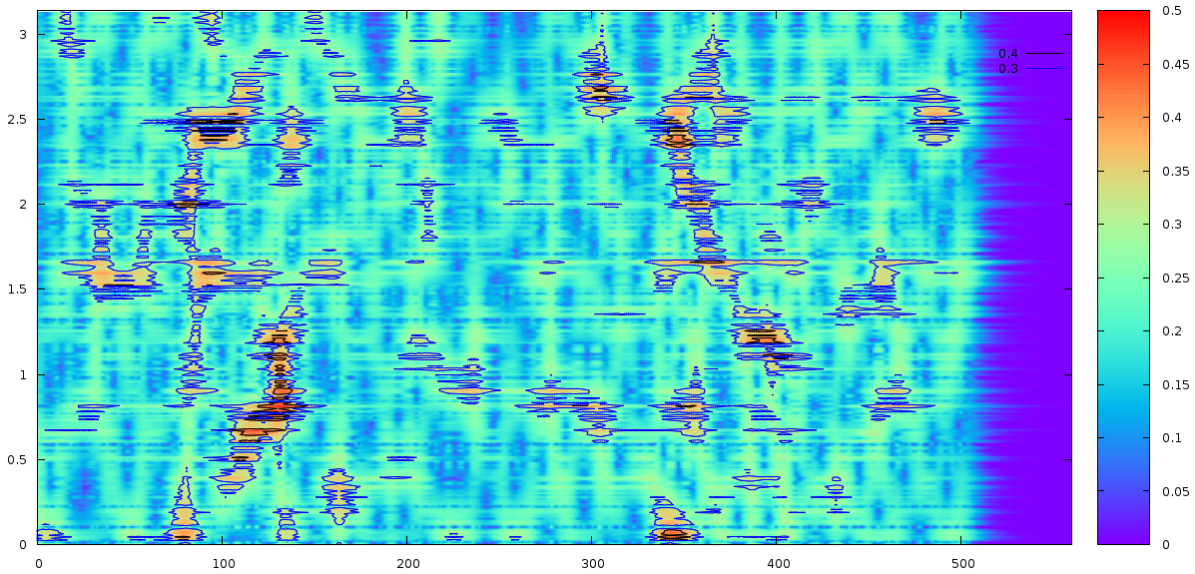
O segundo gráfico, que mostra a distribuição tempo-frequência de cada sinal, é uma representação tridimensional que apresenta:

- no eixo x, a dimensão tempo, variando do dia 0 ao 550, com o objetivo de cobrir o período envolvendo as 492 instâncias e permitir a leitura da legenda no canto superior direito;
- no eixo y, a dimensão frequência, que pode apresentar qualquer variação de 0 a π, já que esse é o intervalo de frequências com o qual o algoritmo trabalha;
- para representar o eixo z, uma escala de cores, do lado direito do gráfico, e um conjunto de linhas de contorno (ou curvas de nível), situado no canto superior direito do gráfico, que significam a intensidade do sinal localizado simultaneamente no tempo e na frequência; as regiões com maior intensidade, indicando uma discrepância na distribuição, serão tratados como “aspectos”.

As Figuras 4.5 (a) e (b) a seguir apresentam, respectivamente, o gráfico que representa o sinal formado pela variável Variação Diária exclusivamente no domínio do tempo e o gráfico que representa a distribuição tempo-frequência desse mesmo sinal. Observando a Figura 5.3 (b) é possível notar áreas de intensidade mais elevada distribuídas por quase todo o plano tempo-frequência definido pelos eixos x e y. A intensidade máxima obtida, como podemos observar pela escala de cores do lado direito do gráfico, é de 0.5 e as curvas de contorno com intensidade 0.3 e 0.4 são as que representam discrepância do padrão mais geralmente encontrado. Na tentativa de buscar alguma associação entre os resultados obtidos na clusterização da série com o uso do DBSCAN, conforme representado anteriormente na Figura 4.2 (b), e a distribuição tempo-frequência do sinal apresentado a seguir, não foi encontrada nenhuma indicação clara de que a frequência da Variação Diária possa indicar algum fenômeno simultâneo ao que se observa no Fechamento do IBOVESPA.



(a)



(b)

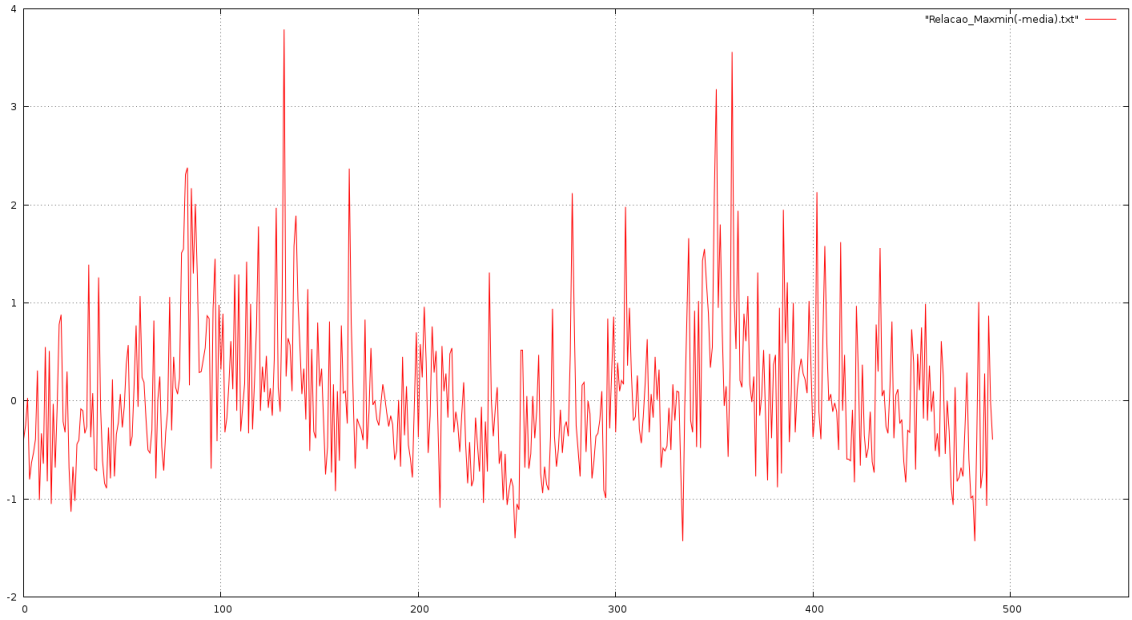
Figura 4.5: Representações do sinal de variação diária: (a) no tempo; (b) DTF.

De forma análoga à adotada para os gráficos anteriormente apresentados, as Figuras 4.6 (a) e (b) a seguir apresentam, respectivamente, o gráfico que representa o sinal formado pela relação Máximo/Mínimo–μ exclusivamente no domínio do tempo, conforme apresentado anteriormente na Figura 4.4 (c) da subseção 4.2.4, e o gráfico que representa a distribuição tempo-frequência desse mesmo sinal.

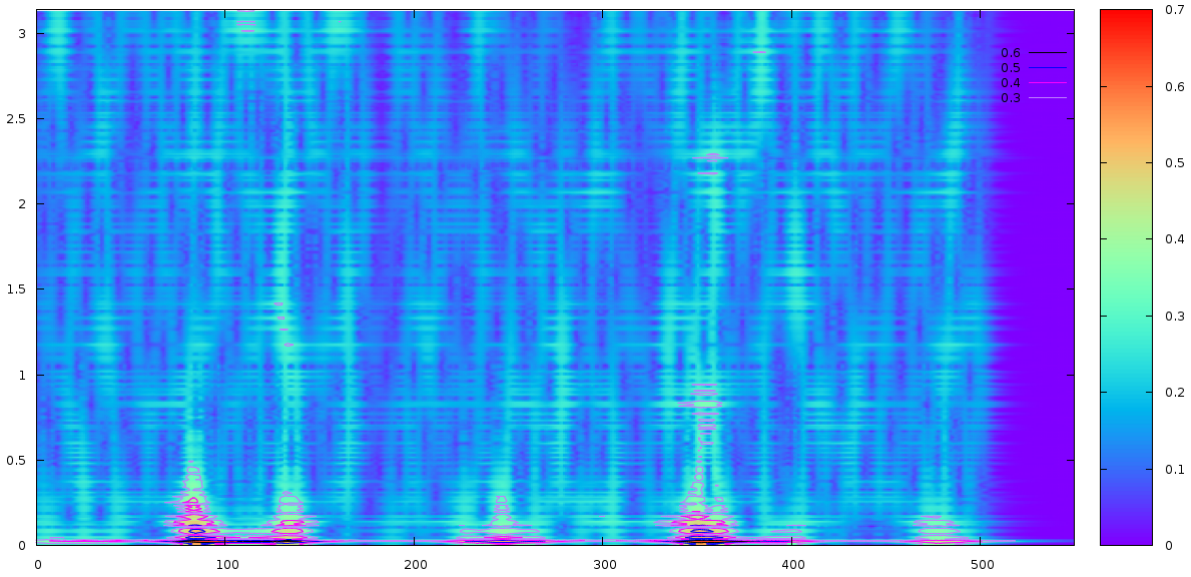
Analisando a Figura 4.6 (b) e comparando-a com a 4.5 (b) avaliada anteriormente, algumas diferenças significativas chamam a atenção. Primeiramente, a intensidade máxima obtida, como podemos observar pela escala de cores do lado direito do gráfico, é de 0.7 e as curvas de contorno com intensidade variando entre 0.3 e 0.6 são as que representam discrepância do padrão mais geralmente encontrado. Em uma verificação bastante simples, bastando para isso alterar um simples parâmetro do *software* GNUPLOT, foi observado que mesmo a intensidade 0.2 não está homogeneamente distribuída por todo o plano tempo-frequência do gráfico.

Em segundo lugar, o que talvez seja a diferença mais importante, somente nas frequências inferiores a 0.5 foram observados aspectos que representam as discrepâncias do padrão mais geralmente encontrado. Isso indica menor presença de trechos do sinal cujas frequências equivalem a períodos inferiores a um dia, como resultado de menos mudanças bruscas nesse sinal, e, além disso, como as frequências dos aspectos são mais baixas, o período do sinal ao longo do qual eles ocorrem é mais longo, representando uma maior estabilidade desse sinal se comparado àquele anteriormente analisado.

Finalmente, tendo em vista a expectativa de poder estabelecer alguma associação entre os resultados da clusterização da série com o uso do DBSCAN, que apresentavam discrepâncias do Fechamento ao longo do tempo, e os resultados da análise tempo-frequência dos sinais selecionados com o uso do coeficiente de correlação linear de Pearson, que apontavam discrepâncias para a correlação entre a Variação Diária e a relação Máximo/Mínimo, foi realizada uma comparação entre os resultados obtidos através da aplicação do DBSCAN e da STGT.



(a)



(b)

Figura 4.6: Representações do sinal de máximo/mínimo–μ: (a) no tempo; (b) DTF.

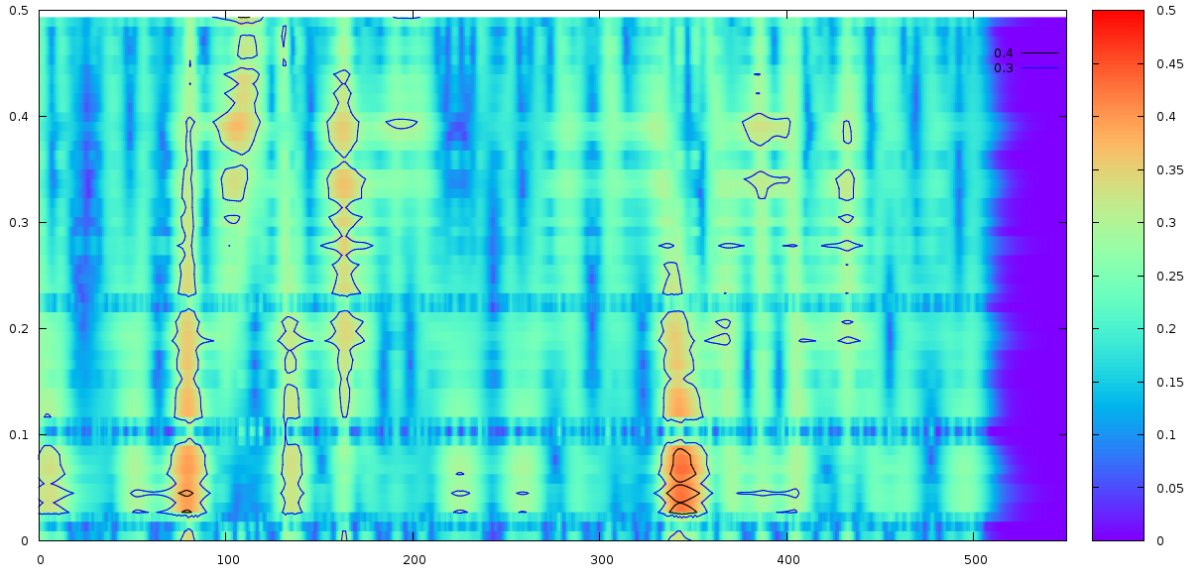
O primeiro passo foi alterar as escalas de frequência das duas distribuições já apresentadas e discutidas, fazendo com que apresentassem valores de 0 a 0.5 no máximo. Isso é feito através de simples parâmetros do *software* GNUPLOT. O resultado encontra-se representado pelas Figuras 4.7 (a) e (b) a seguir. Em seguida, as duas representações foram comparadas entre si e comparadas com o resultado da clusterização apresentado anteriormente na Figura 4.2 (b). O processo envolve uma precisão bastante razoável, já que a ferramenta WEKA permite a identificação de cada ponto da série no exato dia em que ele foi gerado e o GNUPLOT permite a determinação do tempo em milésimos da unidade adotada (nesse caso, dia). Assim, a qualidade da localização no tempo depende fundamentalmente da capacidade da STGT em localizar o sinal nessa dimensão.

Como pode ser observado, os aspectos encontrados em ambos os gráficos sugerem uma certa coerência entre as informações transmitidas pelos sinais, mas não é possível considerá-los semelhantes. É possível notar, mesmo sem observar o resultado da clusterização com o DBSCAN, que a representação da Figura 4.7 (b) apresenta aspectos com formato mais uniforme e de intensidade mais elevada, caracterizando discrepâncias mais claramente delimitadas na frequência, mas também com uma boa localização no tempo.

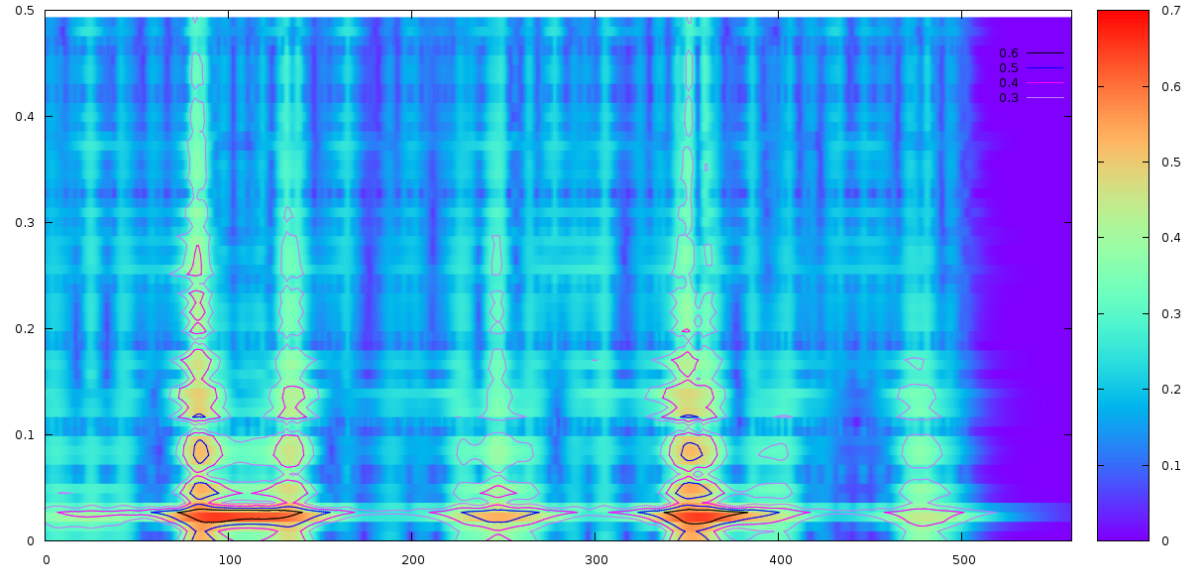
Quando comparamos as Figuras 4.7 (b) com a Figura 4.2 (b), é possível identificar claramente as seguintes associações:

- o 1º aspecto, com intensidade 0.4 entre os dias 10 e 50, corresponde ao cluster 2 (verde claro) que apresenta o máximo global da série;
- o 2º aspecto, com intensidade 0.6 entre os dias 70 e 140, corresponde ao cluster 1 (vermelho), que apresenta vários mínimos locais;
- o 3º aspecto, com intensidade 0.5 entre os dias 230 e 270, corresponde ao cluster 7 (vermelho), que apresenta um máximo local;
- o 4º aspecto, com intensidade 0.6 entre os dias 340 e 380, corresponde aos clusters 9 (cinza), 10 (azul escuro) e 11 (vermelho), sendo esse último o que apresenta o mínimo global da série;
- o 5º e último aspecto, com intensidade 0.4 entre os dias 460 e 500, corresponde ao final da série, incluindo os três pontos de ruído.

Já a Figura 4.7 (a) apresenta fracas indicações do 2º e 4º aspectos já relatados. É importante mencionar que essa figura também apresenta um outro aspecto, entre os dias 155 e 175, que pode corresponder à rápida elevação entre os clusters 4 (roxo) e 5 (rosa).



(a)



(b)

Figura 4.7: Comparação entre DTF's dos sinais: (a) variação diária; (b) máximo/mínimo–μ.

As observações realizadas indicam que os fenômenos analisados através do gráfico da distribuição tempo-frequência do sinal Máximo/Mínimo–μ, produzido com o uso da STGT, podem corresponder a discrepâncias observadas na clusterização nas dimensões tempo e Fechamento, elaborada com o uso do DBSCAN.

Parece importante o fato do último aspecto observado ter se estendido até o dia 500, já que a série analisada termina no dia 492. Isso pode indicar uma quantificação da precisão envolvendo o uso da transformada, como também pode ser uma indicação da maneira como as imagens geradas devem ser analisadas. Em outras palavras, os aspectos analisados no gráfico talvez não devam ser medidos do início até o final das linhas de contorno que representam os limites de uma faixa de intensidade. Essa dúvida só poderá ser esclarecida se as mudanças nos gráficos produzidos pela STGT e naqueles elaborados através do uso do DBSCAN sejam analisados na medida em que os fenômenos forem sendo gerados.

Dando continuidade ao estudo, o DBSCAN e a STGT passarão a ser utilizados sistematicamente para o acompanhamento do crescimento natural da série e, além disso, o surgimento de cada um dos aspectos já encontrados deverá ser “simulado”, como se os respectivos pontos da série estivessem se incorporando a ela no momento da análise, de forma a confirmar a ocorrência simultânea dos fenômenos. Antes disso porém, serão apresentados os resultados de uma abordagem de análise de reversão de tendência em série temporais e, em seguida, os resultados dessa abordagem serão avaliados para verificar a sua aplicabilidade na análise das variações de preço em ações da BOVESPA.

4.4 – Reversão de Tendência

Considerando o exposto anteriormente na seção 2.2, este estudo tem como proposta inicial adotar o coeficiente angular da reta de regressão linear determinada de forma móvel, calculada para cada ponto da série usando o seu valor e os quatro valores que o antecedem, e a taxa de variação desse coeficiente angular, calculada de forma análoga, para obter uma estimativa da primeira e da segunda derivadas em relação ao tempo da variável $Y = F(t)$.

A decisão de adotar cinco pontos para cálculo do coeficiente angular foi tomada com base em avaliações empíricas: a ideia era adotar o menor número de pontos possível, de forma a manter a abordagem coerente com o conceito de derivada como o ângulo a reta que passa por dois pontos quando a distância entre eles tende a zero, mas também estabelecendo um compromisso de não permitir que a brusca variação entre apenas dois pontos pudesse causar distorções na análise. A opção de adotar cinco pontos para o cálculo do coeficiente angular também apresenta a vantagem de representar a variação média ao longo de um

intervalo de uma semana (cinco dias úteis), já que esse unidade intervalar tem significado prático para muitos analistas de mercado e investidores.

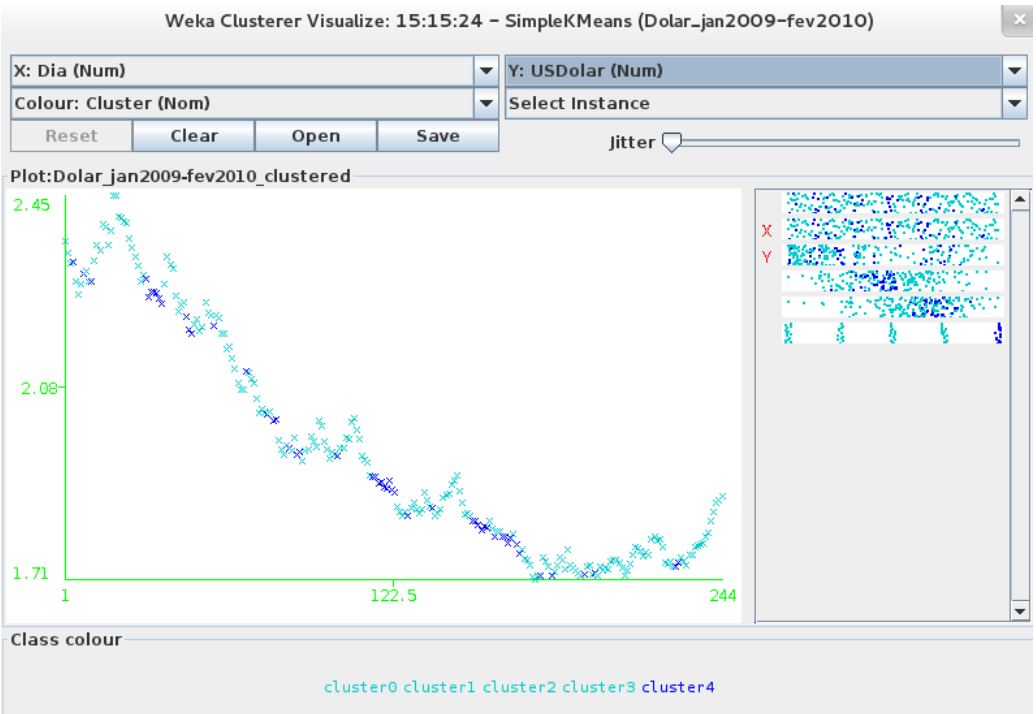
Definido o conceito, foi aplicada a técnica de clusterização pelo SimpleKMeans, conforme apresentada no Capítulo 3, a duas séries que representam um ano de evolução do valor do dólar comercial no Brasil. A primeira série cobre o período de fevereiro de 2009 a janeiro de 2010, quando o país se recuperava da crise financeira de 2008 e o dólar apresentava uma clara tendência de queda. A segunda série cobre o período de maio de 2014 a abril de 2015, quando as eleições presidenciais e os primeiros ajustes na política fiscal feitas pelo governo reeleito trouxeram maior incerteza ao cenário econômico e, com isso, provocaram uma clara tendência de elevação na cotação do dólar.

O objetivo era buscar um grupo de pontos da série que representasse uma das duas seguintes situações:

- pontos com primeira derivada negativa e segunda derivada positiva, indicando um potencial de reversão de tendência de queda;
- pontos com primeira derivada positiva e segunda derivada negativa, indicando um potencial de reversão de tendência de elevação;

Primeiramente foi avaliada a série correspondente ao período de fevereiro de 2009 a janeiro de 2010. O melhor resultado do processo de clusterização está representado a seguir nas Figuras 4.8 (a) e (b).

(a)



(b)

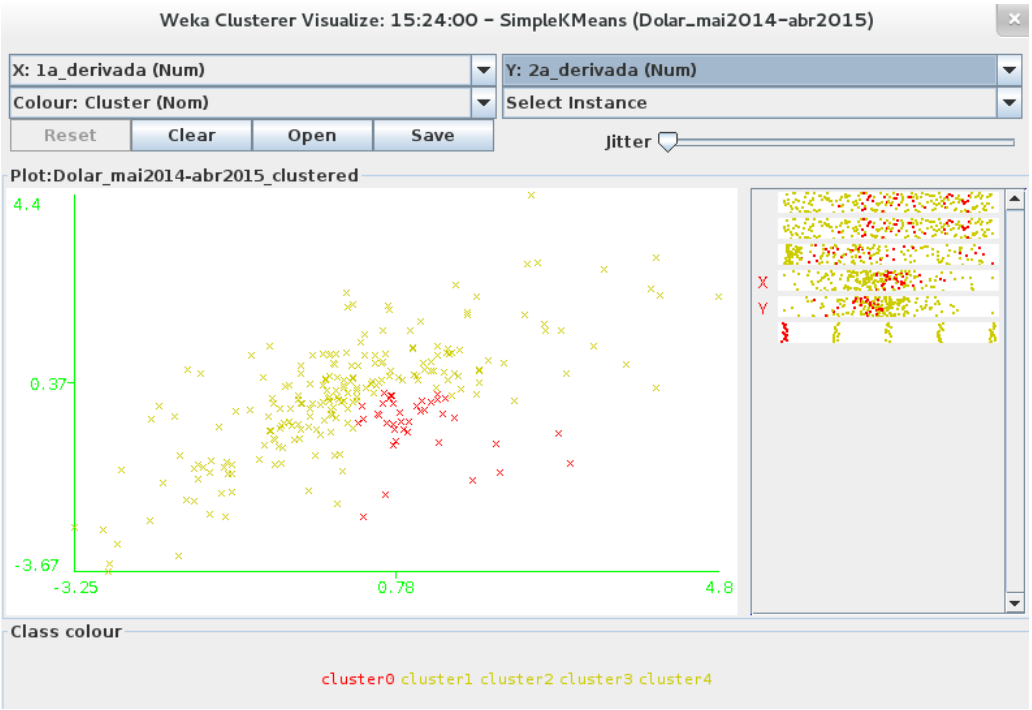
Figura 4.8: 52 pontos de reversão da tendência de queda: (a) [x''(t)] vs [x'(t)]; (b) x(t).

A clusterização foi realizada ignorando os atributos Dia, que representa a dimensão do tempo, e USDolar, que representa o valor da moeda, com a intenção de aplicar o algoritmo para identificar clusters de formato circular na região do plano formado pela primeira derivada representada no eixo x e pela a segunda derivada representada no eixo y. Foram utilizados diferentes números de clusters como parâmetro e avaliadas as duas medidas de distância disponíveis para esse algoritmo: a Euclidiana e a *Manhattan*. O melhor resultado, apresentado anteriormente, foi obtido com a busca de cinco clusters e a utilização da distância *Manhattan*. Como é possível observar, foram encontrados 52 pontos de reversão de tendência de queda (representados na cor azul escura do cluster4) ao longo de um total de 244 pontos que formam a série. Ainda que existam alguns poucos pontos de reversão de tendência de elevação, eles não chegam a formar um cluster específico, como se pode notar pelo formato da nuvem de pontos na Figura 4.8 (a).

De forma análoga, foi avaliada a série correspondente ao período de maio de 2014 a abril de 2015. O melhor resultado do segundo processo de clusterização está representado a seguir nas Figuras 4.9 (a) e (b).

Essa nova clusterização foi executada seguindo os mesmos passos do processo realizado para a primeira série temporal, adotando o mesmo parâmetro de número de clusters e a mesma medida de distância. Dessa vez, como se pode observar, foram encontrados 39 pontos de reversão de tendência de elevação (representados na cor vermelha do cluster0) ao longo de um total de 248 pontos que formam a segunda série. Também nesse caso, mesmo que existam alguns poucos pontos de reversão de tendência de queda, eles não chegam a formar um cluster específico, como também é fácil perceber pelo formato da nuvem de pontos na Figura 4.9 (a).

Parece natural que, em uma série onde prevalece uma clara tendência, seja ela de queda ou de elevação, sejam encontrados vários pontos de continuação de tendência e alguns pontos onde agentes que atuam no mercado entrem em ação com o objetivo de reverter a tendência apresentada. Por outro lado, não faz muito sentido buscar pontos que revertam uma tendência que o gráfico não aponta. No entanto, é importante destacar que, em ambos os casos, as tentativas de reversão nem sempre surtem um efeito muito marcante. É fácil perceber que há casos em que, momentaneamente, a tendência é revertida. Porém, em alguns pontos de reversão, ocorre somente uma leve desaceleração na evolução da série e, em alguns outros, até mesmo quando localizados em sequência, curiosamente nenhuma mudança facilmente perceptível é alcançada.



(a)

(b)

Figura 4.9: 39 pontos de reversão da tendência de elevação: (a) [x''(t)] vs [x'(t)]; (b) x(t).

4.5 – Avaliação das Técnicas Aplicadas

Para demonstrar a utilidade do protótipo apresentado na Figura 4.4 (c) da subseção 4.2.4, foram realizados dois processos de clusterização utilizando SimpleKMeans de duas séries com 600 instâncias, referente a um ativo que envolve pouca especulação no mercado.

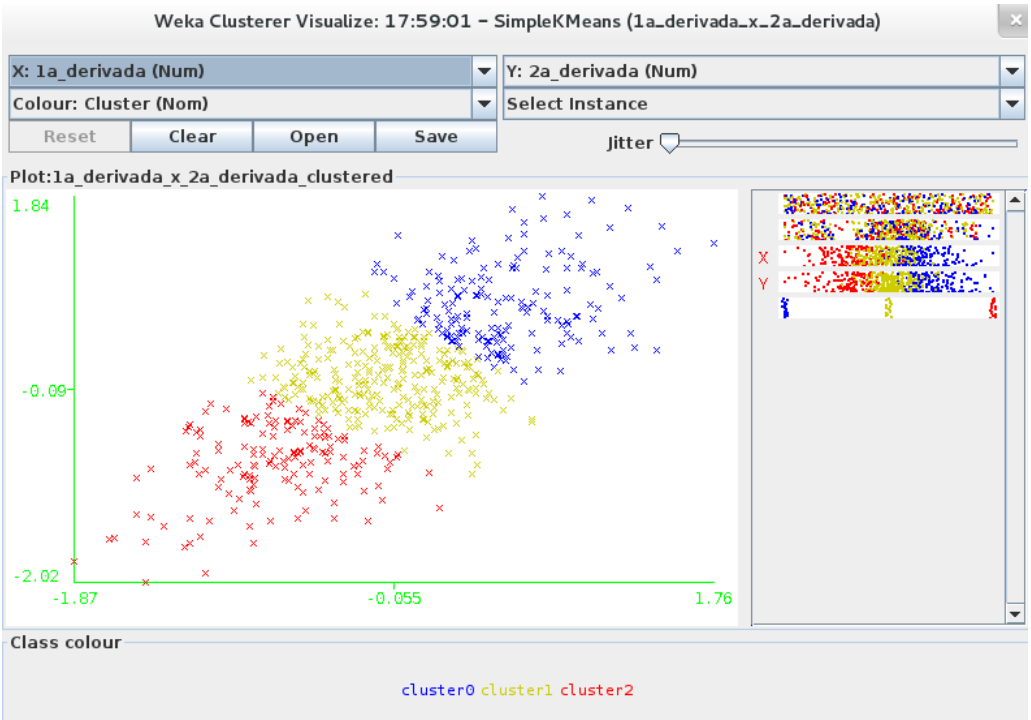
Na Figura 4.10 (a) a seguir é apresentado o gráfico que relaciona uma estimativa para a primeira derivada da evolução do valor do ativo (no eixo x) e a estimativa para a sua segunda derivada (no eixo y). A primeira é calculada através do coeficiente angular de uma regressão linear realizada sobre a série de valores do ativo para cada ponto da série, considerando esse último e os quatro pontos anteriores, e a segunda é, de forma análoga, o coeficiente angular da regressão linear realizada sobre os valores da estimativa da primeira derivada. A utilização de cinco pontos da série no cálculo das estimativas segue os mesmos critérios apresentados na seção 4.4. Dessa forma, considerando que CORR é o coeficiente de correlação linear de Pearson e STDEV é o desvio padrão, as estimativas das derivadas são obtidas através de:

- 1ª derivada=CORR(valor,tempo)*[STDEV(valor)/STDEV(tempo)];
- 2ª derivada=CORR(1a_derivada,tempo)*[STDEV(1a_derivada)/STDEV(tempo)].

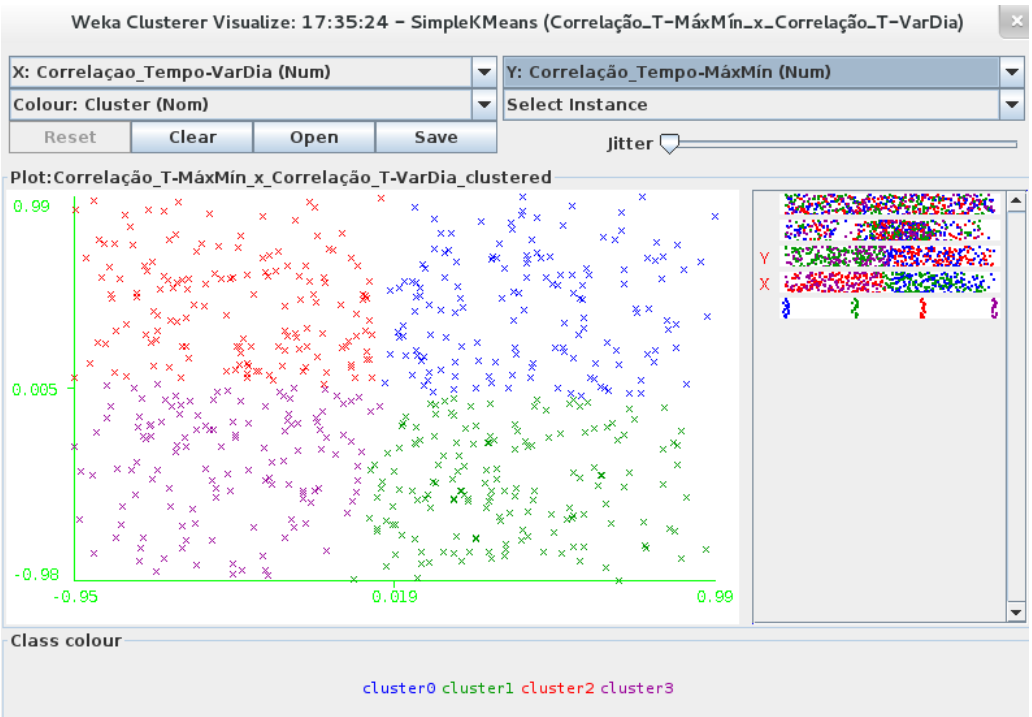
Já na Figura 4.10 (b) é apresentado o gráfico que relaciona a correlação entre a Variação Diária e o tempo (no eixo x) e a correlação entre a relação Máximo/Mínimo e o tempo (no eixo y). Ambas são simplesmente a aplicação do coeficiente de correlação linear de Pearson entre o tempo e a outra variável mencionada para cada ponto da série, considerando esse último e os quatro pontos anteriores, mantendo assim os mesmos critérios de medição intervalar adotados anteriormente. A ideia por trás da análise da relação entre essas duas medidas é a de que, considerando que o tempo sempre cresce, seria possível comparar não os valores assumidos pelas variáveis, mas as tendências com que elas evoluem. Assim, considerando que CORR é o coeficiente de correlação linear de Pearson, as referidas medidas são obtidas através de:

- Correlação_Tempo-VarDia=CORR(Variação Diária, tempo);
- Correlação_Tempo-MáxMín=CORR(relação Máximo/Mínimo, tempo).

Em ambos os casos foi utilizada a distância Euclidiana; o melhor resultado para o primeiro foi obtido com a busca por três clusters e para o segundo com a busca por quatro clusters. Pode ser facilmente observado nas Figuras 4.10 (a) e (b) que os dois gráficos apresentam diferenças bastante significativas na distribuição dos pontos pelo plano.



(a)



(b)

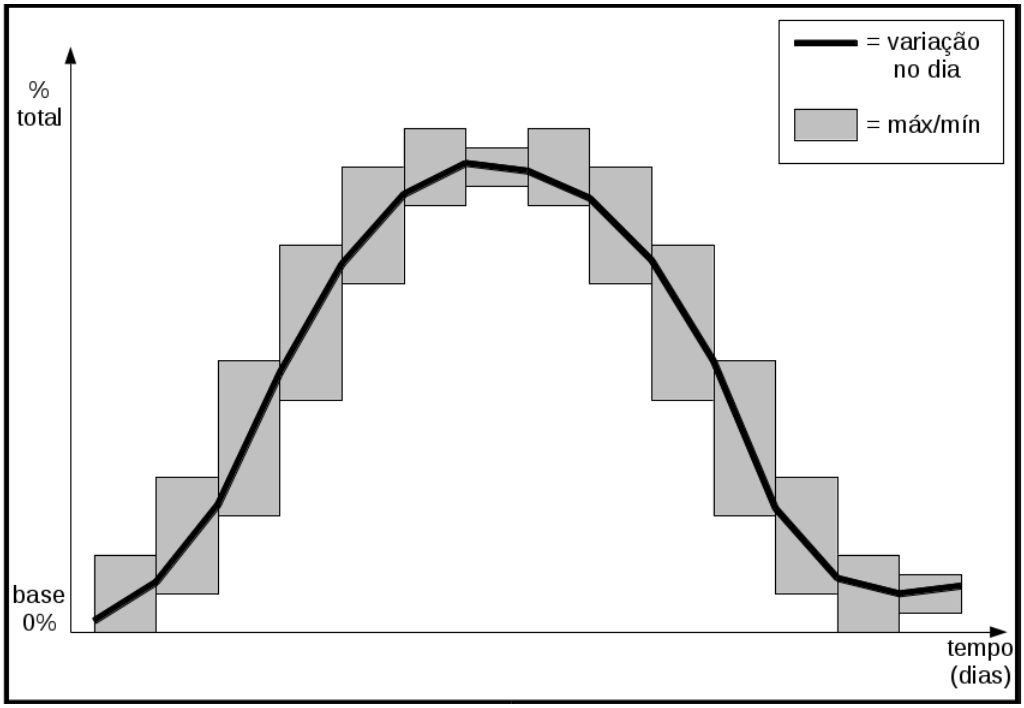
Figura 4.10: Análise de cluster com o uso do algoritmo SimpleKMeans: (a) das estimativas das derivadas; (b) das correlações do tempo com a variação diária e com máximo/mínimo.

Observando a distribuição dos pontos no primeiro gráfico, é possível notar que a Figura 4.10 (a) apresenta uma “nuvem” de pontos que parece estar distribuída em torno de uma reta imaginária y=x. Também é possível notar que o processo de clusterização não foi capaz de gerar clusters de reversão de tendência, ou seja, um grupo de pontos em que a primeira derivada é positiva e a segunda é negativa (canto inferior direito, abaixo da nuvem) ou um grupo de pontos em que a primeira derivada é negativa e a segunda é positiva (canto superior esquerdo, acima da nuvem), indicando respectivamente que uma tendência de alta ou de baixa está desacelerando. Existem alguns poucos pontos de reversão de tendência nas bordas dos três clusters mas, como são poucos e estão dispersos, não formam um cluster, o que retrata a dificuldade para realizar uma previsão acurada sobre a evolução do valor de um ativo do tipo que é negociado em bolsas de valores.

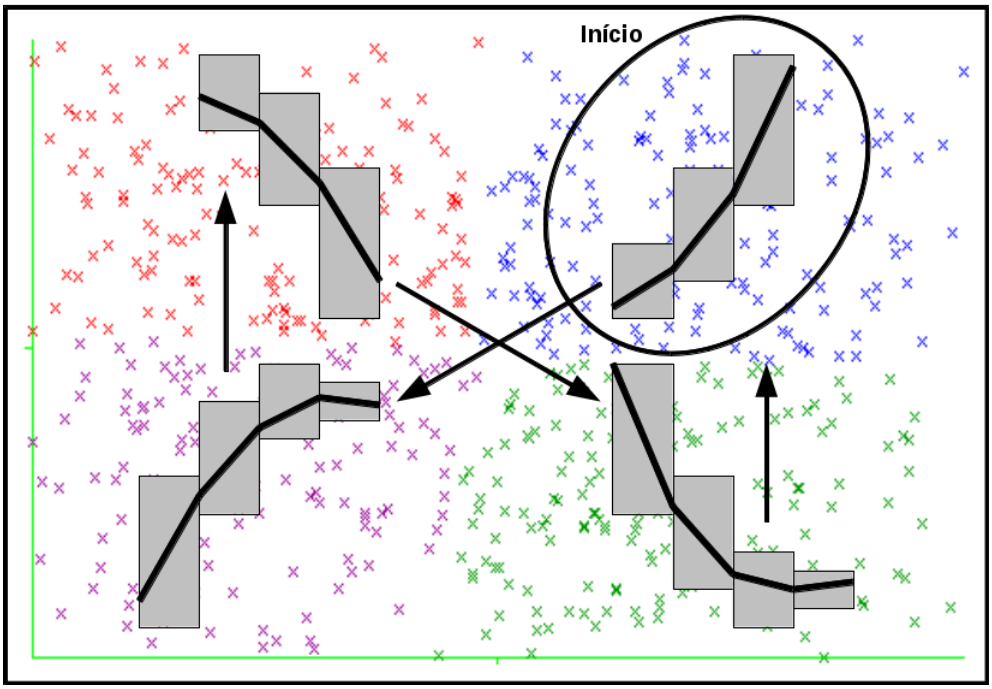
Já no segundo gráfico, é possível notar que a distribuição dos pontos que representam a relação entre as medidas de correlação é bastante homogênea e, com isso, identificamos quatro quadrantes que representam estados significativos da evolução do valor do ativo, sendo particularmente importantes para este estudo aqueles em que verificamos tendência de redução da relação Máximo/Mínimo com reversão de tendência na Variação Diária: de aumento na Variação Diária no cluster1 (na cor verde) e de redução na Variação Diária no cluster3 (na cor roxa). Para melhor ilustrar os quatro estados mencionados, o protótipo desenvolvido e apresentado na Figuras 4.4 (a), (b) e (c) da subseção 4.2.4 será utilizado juntamente com o gráfico da Figura 4.10 (b).

As Figuras 4.11 (a) e (b) a seguir apresentam os resultados da combinação do protótipo com a última clusterização realizada. Primeiramente, foi desenvolvido e apresentado na Figura 4.11 (a) um exemplo utilizando a simbologia do protótipo, com objetivo meramente didático, que representa um ciclo completo envolvendo as etapas de tendência de elevação, reversão de tendência de elevação, tendência de queda e reversão de tendência de queda de um ativo. Em seguida, as partes do exemplo de protótipo correspondentes a essas quatro etapas do ciclo foram desmembradas e colocadas sobrepostas aos quadrantes a que elas correspondem no gráfico da clusterização, conforme apresentado na Figura 4.11 (b). Essa representação fornece mais uma indicação de que a variação do tamanho das barras que representam a relação Máximo/Mínimo pode ser determinante na análise de tendência de uma série temporal envolvendo valores de ativos.

Considerando os resultados obtidos, o modelo proposto será desenvolvido com o uso do DBSCAN para a clusterização do Fechamento no tempo e da STGT para a análise tempo-frequência do sinal formado por Máximo/Mínimo–μ.



(a)



(b)

Figura 4.11: Exemplo construído com o auxílio do protótipo: (a) da evolução dos valores no tempo; (b) do significado dos clusters.

5 – O MODELO PROPOSTO

O modelo apresentado a seguir foi desenvolvido com o objetivo de analisar o comportamento de ativos negociados em bolsas de valores, bem como suas combinações em carteiras que, por sua vez, podem representar índices de mercado. Na verdade é fortemente recomendável que todo processo de análise tenha início com um agregado de ativos, já que os resultados assim obtidos costumam representar a percepção dos investidores a respeito de um mercado ou de um setor da economia. O conjunto de técnicas empregadas compõe uma abordagem denominada Mineração de Dados Indireta, conforme representada na Figura 5.1 a seguir. Nessa abordagem, a etapa de interpretação das informações mineradas é indispensável para o sucesso do processo de geração de conhecimento.

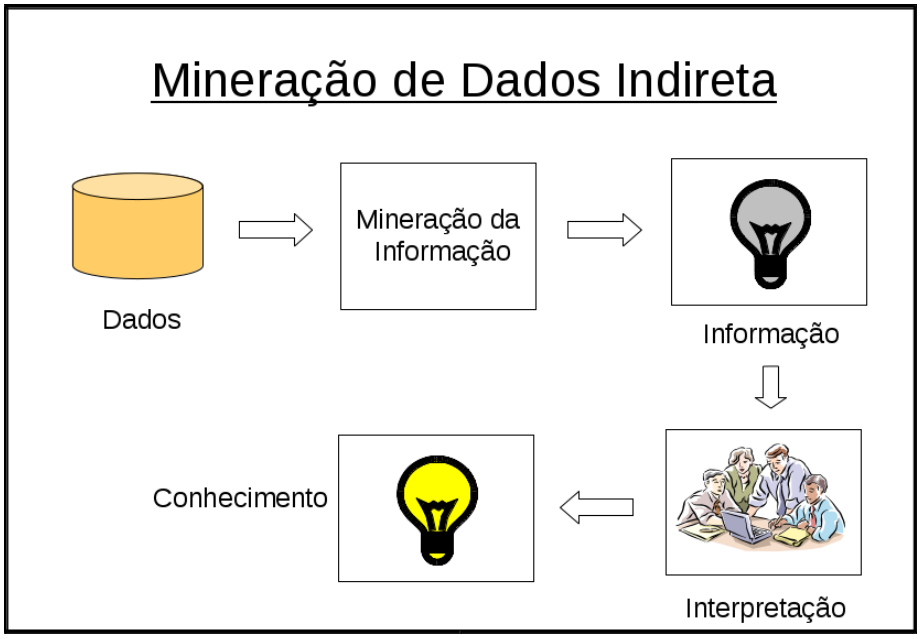


Figura 5.1: Etapas da Mineração de Dados Indireta [Plastino, 2013].

Como mencionado no Capitulo 1 – INTRODUÇÃO e repetido aqui apenas para enfatizar esta importante característica do modelo proposto, este trabalho considera as operações nos mercados financeiros denominados bolsas de valores como jogos de chance envolvendo incerteza, conforme apresentado na Figura 5.2 a seguir [Kelly, 2003].

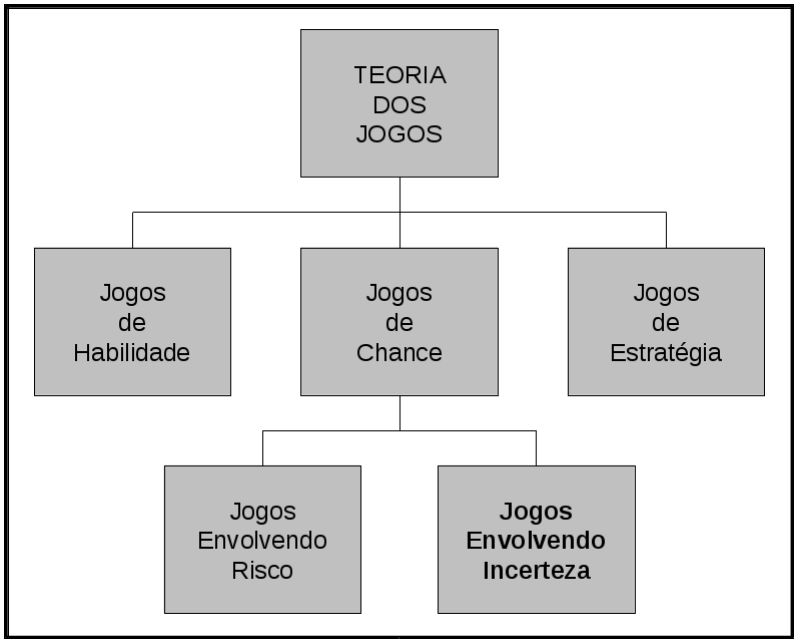


Figura 5.2: Taxonomia dos Jogos [Kelly, 2003].

De acordo com Kelly [2003], os jogos de chance são aqueles em que um único jogador enfrenta a natureza. Nesse caso, diferentemente do que acontece nos jogos de habilidade e nos de estratégia, o jogador em questão não é capaz de controlar completamente os resultados alcançados a partir de suas ações e as suas decisões estratégicas nem sempre levam a um resultado que pudesse ter sido antecipado.

Os jogos de chance podem ser subdivididos em dois outros tipos: jogos envolvendo risco, onde o jogador conhece a probabilidade de cada resposta da natureza às suas ações, e jogos envolvendo incerteza. Nesse último tipo, segundo Colman [1982; apud Kelly, 2003], não é possível atribuir uma probabilidade que contenha algum significado a quaisquer das respostas da natureza às ações do jogador.

5.1 – Premissas na Extração dos Dados

A primeira recomendação na escolha dos ativos a serem analisados é selecionar somente aqueles que apresentem volume de negociação significativo, usando como parâmetro uma movimentação mínima de R$100 milhões por dia. No caso de empresas que apresentem mais

de um tipo de ativo (ex.: ON, PN, etc.), é recomendável trabalhar exclusivamente com os papéis do tipo PN, pois esse tipo costuma, no médio prazo, atrair investidores preocupados com a distribuição de dividendos, que é um importante fator ligado à rentabilidade do ativo.

O conjunto mínimo de instâncias recomendado para dar início à análise é de 250 registros, o que compreende o período aproximado de um ano. Com o crescimento da série, é conveniente trabalhar com um conjunto de 500 a 750 registros, sempre procurando excluir períodos de comportamento atípico do mercado (ex.: eleições presidenciais de 2014).

Após calcular a relação MÁXIMO/MÍNIMO diária e obter a média dessa variável para toda a série, é importante avaliar a incerteza decorrente do tamanho da banda de oscilação diária do valor do ativo. Em geral, ativos com grandes bandas de oscilação, superiores a 3,0% como apresentado pela PETR4 – Petrobras PN, tendem a atrair investidores com comportamento mais especulativo. Entre 2,0% e 3,0% estão a maior parte dos ativos de primeira linha do mercado que, quando adequadamente combinados, podem resultar em uma carteira mais balanceada, com banda inferior a 2,0%, conforme ocorre com a própria carteira do índice IBOVESPA.

A seguir é apresentado um resumo dos critérios acima descritos:

- #instâncias mínimo >= 250; 700 >= #instâncias ideal >= 500
- VOLUME (do ativo no dia) >= R$ 100 milhões
- Análise da incerteza decorrente da banda de oscilação diária do valor do ativo:
    - Baixa: MÉDIA [MÁXIMO (ativo) / MÍNIMO (ativo)] < 2,0%
    - Média: 2,0% <= MÉDIA [MÁXIMO (ativo) / MÍNIMO (ativo)] <= 3,0%
    - Alta: MÉDIA [MÁXIMO (ativo) / MÍNIMO (ativo)] > 3,0%

5.2 – Regras na Mineração da Informação

Após selecionados os ativos e obtidas suas correspondentes séries temporais, a próxima etapa do processo compreende a mineração de informações através do uso do DBSCAN na ferramenta Weka Explorer e da STGT com o seu algoritmo implementado via programa específico. No DBSCAN é analisada a série dos valores diários do FECHAMENTO e na STGT é analisada a série dos valores diários da relação MÁXIMO/MÍNIMO subtraída da média dessa nova variável para toda a série, calculada e analisada na etapa anterior.

Antes de executar os dois métodos acima mencionados, é importante realizar uma outra avaliação de incerteza, levando em consideração os valores máximo e mínimo do FECHAMENTO na série, sua média e seu desvio padrão e, de forma qualitativa, o gráfico que

representa a função densidade de probabilidade dessa variável, utilizando para isso as informações encontradas na aba Preprocess da Weka Explorer, conforme apresentada na figura 5.2 a seguir. Idealmente, se a função densidade apresenta um formato que possa ser aproximado por uma curva normal e se o valor atual do ativo se encontra a, no máximo, um desvio padrão abaixo da sua média, parece ser válida uma análise mais detalhada do momento atual com intenção de investir nesse ativo. Caso contrário, a menos que existam importantes valores de referência do ativo, tal como máximos e mínimos anuais ao longo de mais de uma década ou valores registrados em importantes momentos da economia, passa a ser recomendável aguardar o crescimento da série para retomar o processo.

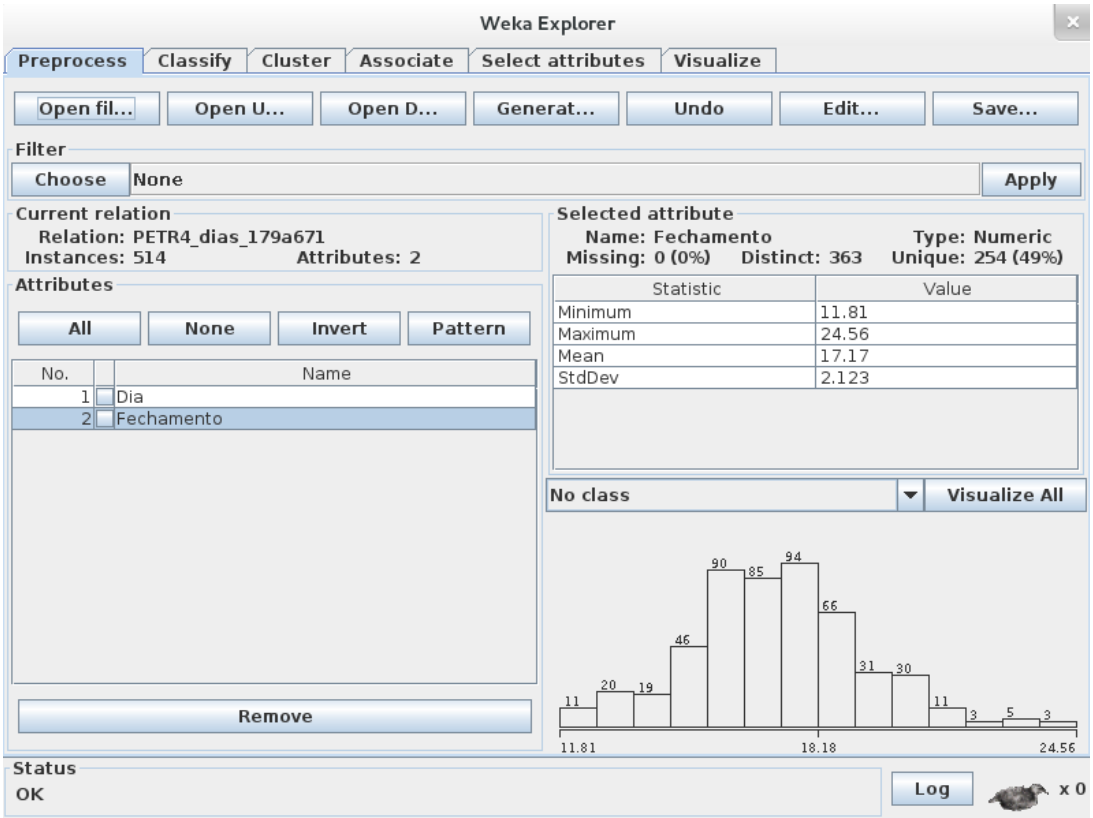


Figura 5.3: Aba Preprocess da Weka Explorer.

A seguir é apresentado um resumo dos critérios acima descritos:

- Na aba Preprocess da Weka Explorer, análise da incerteza pela função densidade:
  - Baixa: Maximum / Minimum > 200%
  - Média: 150% <= Maximum / Minimum >= 200%
  - Alta: Maximum / Minimum < 150%

- Na mesma aba, com o mesmo objetivo:
  - Baixa: FECHAMENTO < Mean – StdDev
  - Média: Mean – StdDev <= FECHAMENTO <= Mean + StdDev
  - Alta: FECHAMENTO > Mean + StdDev
- Na aba Cluster, usando o DBSCAN, para 5 a 8 clusters com 250 instâncias:
  - 0.020 <= epsilon <= 0.040; minPoints = 4
- Na STGT, usando Input Signal 8, para obter o espectro de Fourier e a STGT:
  - SINAL(i) = MÁXIMO/MÍNIMO(i) – MÉDIA[MÁXIMO/MÍNIMO]

5.3 – Critérios na Interpretação da Informação

A utilização do DBSCAN fornece gráficos da evolução do valor (eixo y) de FECHAMENTO no tempo (eixo x), destacando em diferentes cores os clusters formados por trechos que atendem aos parâmetros de densidade mínima estabelecidos na etapa anterior e através da letra “M” (*Missing*) os pontos considerados como ruído. O arquivo gerado pela STGT é analisado pela ferramenta gráfica Gnuplot, de modo a produzir gráficos tridimensionais da distribuição das frequências (eixo y) da relação MÁXIMO/MÍNIMO no tempo (eixo x) com as correspondentes intensidades (3º eixo) representadas pela curvas de nível (legenda no canto superior direito) e sua escala de cores (barra vertical).

É importante proceder a análise dos pares de gráficos em conjunto, de forma a buscar indicações de mudança de comportamento de ambos os indicadores, ainda que as alterações em um deles não sejam necessariamente refletidas no outro no mesmo instante. Por outro lado, quando as alterações ocorrem simultaneamente nos dois gráficos, a interpretação do fenômeno se torna menos duvidosa. Vale ressaltar que, para observar o surgimento de pontos de ruído e a sua transformação em pontos pertencentes a clusters, a periodicidade de execução das duas técnicas deve ser inferior a quatro dias, já que esse foi um dos parâmetros adotados na etapa anterior para a identificação dos clusters (minPoints = 4). A qualquer momento ao longo do processo de interpretação das informações, pode ser útil realizar simulações através da redução do parâmetro “epsilon” do DBSCAN. Dessa forma é possível avaliar a homogeneidade de um cluster, ou seja, se em uma análise mais detalhada ele pode ser decomposto em outros clusters menores e eventuais ruídos.

O objetivo fundamental da interpretação é identificar o momento mais adequado para investir em ativos ou carteiras e, de forma análoga, o momento recomendável para realizar o ganho (ou mesmo a perda, se for para minimizá-la) de um investimento realizado.

Para isso usamos como fundamentação dois fenômenos: mudanças na densidade da série de valores de fechamento e aumentos na intensidade das frequências da série de valores da relação máximo/mínimo. Em outras palavras, buscamos sinais da alteração de um padrão de comportamento da série de valores de fechamento, tais como a formação de ruídos ou a formação de novos clusters, e o surgimento de intensidades mais elevadas (de 0.3 a 0.5 ou maior) nas frequências mais baixas (inferiores a 0.3) da série de valores da relação máximo/mínimo. Vale ressaltar que as alterações observadas em carteiras de ativos costumam ser mais sutis do que aquelas observadas isoladamente através da análise dos ativos que entram na sua composição.

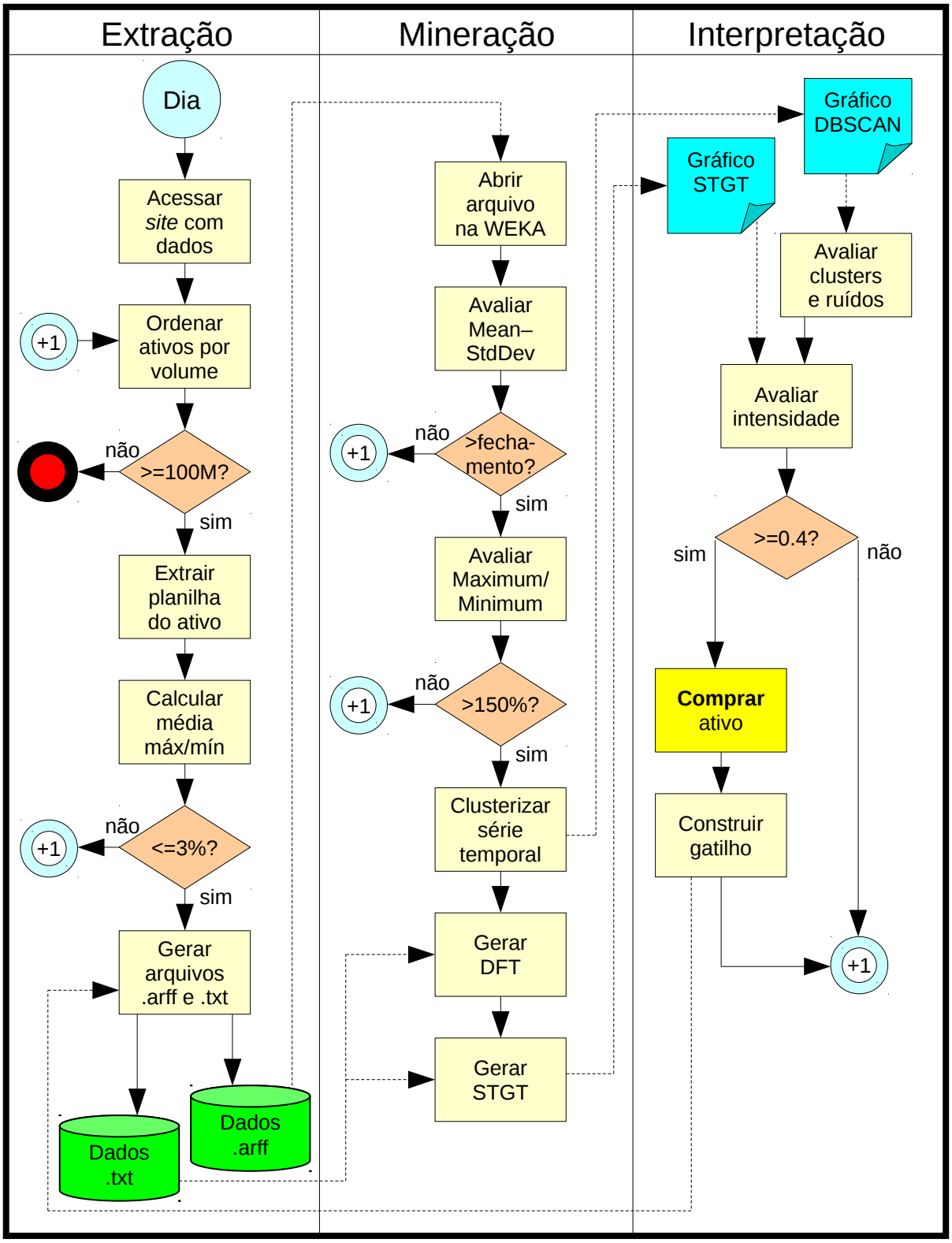
Quando for identificado um momento com potencial para compra de um ativo, deve ser criada uma curva de retorno do investimento realizado no DBSCAN, que servirá de "gatilho" para a venda do ativo caso a STGT não sinalize o momento adequado para essa venda. Com isso, caso o valor de fechamento em um momento futuro seja inferior ao valor da compra corrigido pelo retorno esperado, é recomendável agir para realizar o ganho já obtido.

A seguir é apresentado um resumo dos critérios acima descritos:

- Análise da incerteza na compra, analisando na STGT:
  - Alta: INTENSIDADE [MÁXIMO / MÍNIMO] >= 0,3
  - Média: INTENSIDADE [MÁXIMO / MÍNIMO] >= 0,4
  - Baixa: INTENSIDADE [MÁXIMO / MÍNIMO] >= 0,5
- Tendo efetuado a compra, adicionar ao DBSCAN:
  - GATILHO=FECHAMENTO(compra)*EXP(1+taxa;#dias); taxa=0,1% a.d.
- Análise da incerteza na venda, analisando na STGT:
  - Baixa: INTENSIDADE [MÁXIMO / MÍNIMO] >= 0,3
  - Média: INTENSIDADE [MÁXIMO / MÍNIMO] >= 0,4
  - Alta: INTENSIDADE [MÁXIMO / MÍNIMO] >= 0,5
- Não sendo sinalizada a venda na STGT, analisar no DBSCAN:
  - FECHAMENTO (hoje) < GATILHO

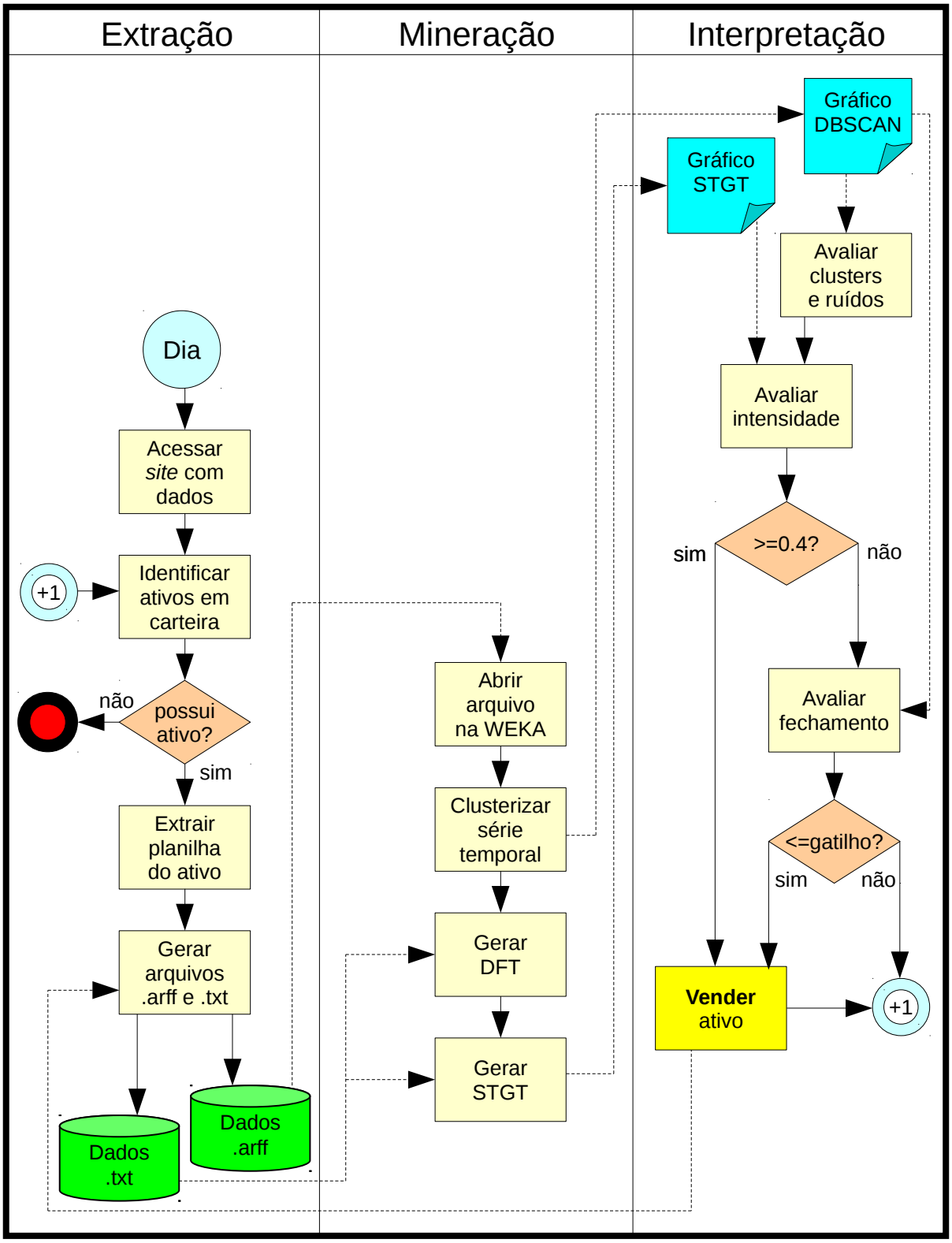
5.4 – Fluxo dos Processos

As Figuras 5.4 (a) e (b) apresentadas a seguir ilustram, respectivamente, o fluxo dos processos Formação da Carteira e Gestão da Carteira no padrão *BPM – Business Process Model*. Os objetos utilizados nos diagramas estão de acordo com a *BPM-Notation*, mas sofreram pequenas adaptações aos objetos disponíveis no editor de texto LibreOffice.



(a)

Figura 5.4: Fluxo dos Processos: (a) Formação da Carteira; (b) Gestão da Carteira.



(b)

Figura 5.4 (continuação)

6 – APLICAÇÕES DO MODELO

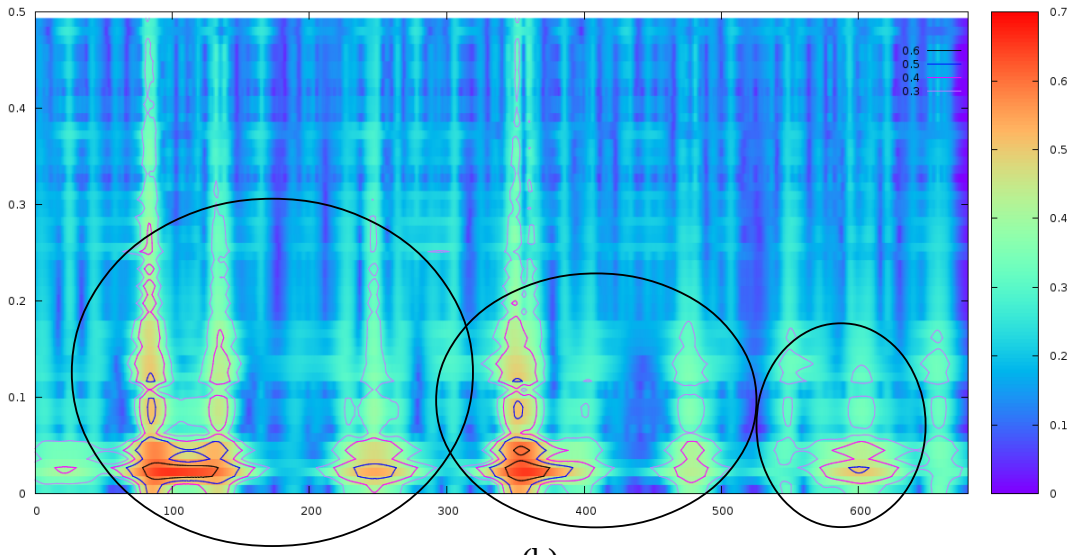
A seguir são apresentados os resultados de duas importantes aplicações conjuntas do DBSCAN com a STGT e os seus correspondentes gráficos, que representam os movimentos referentes ao IBOVESPA e ao ativo PETR4 – Petrobras PN.

6.1 – Aplicação ao IBOVESPA

A primeira série temporal utilizada envolve o período transcorrido do primeiro dia útil após a semana do Carnaval de 2012 (27/02/2012, ou dia 27) até o último dia útil antes do primeiro turno das eleições de 2014 (03/10/2014, ou dia 671). A clusterização foi realizada com minPoints=4 e epsilon=0.028 a 0.030 como parâmetros.

O resultado final da aplicação ao IBOVESPA está representado a seguir através das Figuras 6.1 (a) e (b). A série da carteira IBOVESPA apresenta três ciclos de investimento, com a identificação assinalada tanto no DBSCAN quanto na STGT. No segundo ciclo, foi criada uma curva que representa um retorno esperado para o investimento com uma taxa de retorno de 0,1% ao dia. Essa curva também poderia ser traçada para o primeiro e o terceiro ciclos, mas tornaria mais difícil a visualização dos fenômenos nos gráficos, pois atravessariam as regiões ocupadas por vários pontos da própria série temporal. Esses fatos envolvendo o primeiro e o terceiro ciclos podem ser interpretados como a postura do investidor em relação ao risco envolvido nessas operações mas, conforme mencionado anteriormente, qualquer análise envolvendo risco está fora do escopo deste trabalho.

(a)



(b)

Figura 6.1: Resultados da aplicação ao IBOVESPA: (a) pelo DBSCAN; (b) pela STGT.

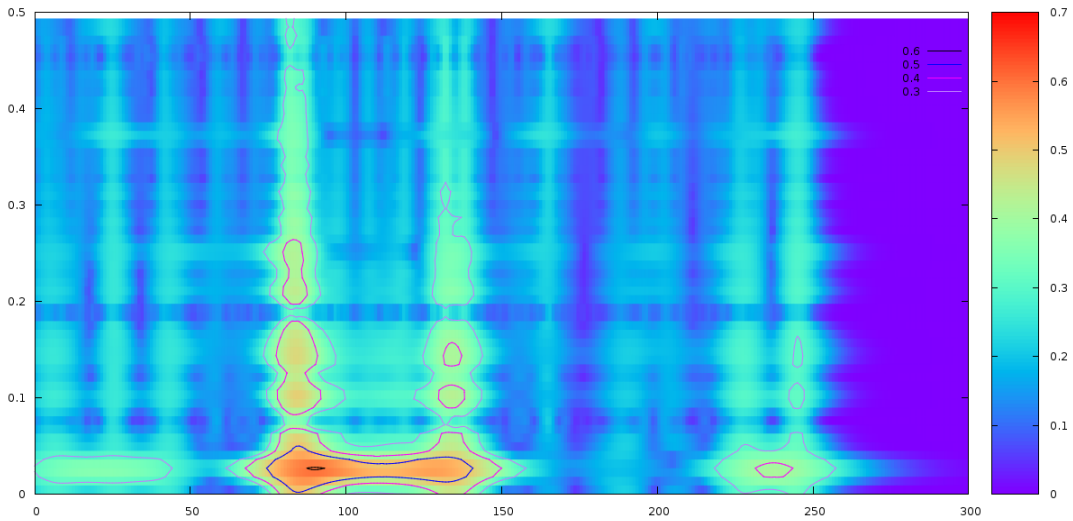
Este estudo evoluiu com a aplicação sistemática das duas técnicas selecionadas a uma base inicialmente composta de 631 registros, buscando realizar dois papéis: acompanhar o crescimento da série em intervalos de um a três dias, observando eventuais fenômenos que surgissem, e simular o surgimento dos fenômenos do passado que já estavam retratados pelas técnicas. Com isso, foi necessário caminhar em dois sentidos ao longo da linha do tempo, ou seja, para frente ao observar os novos fenômenos e para trás ao simular aqueles que já haviam ocorrido. No DBSCAN, esses fenômenos serão tratados por clusters, ruídos, máximos e mínimos. Na STGT, eles serão tratados como aspectos, destacando suas frequências e as correspondentes intensidades no tempo. Tendo concluído todas as observações e simulações planejadas, os resultados serão apresentados em ordem cronológica, como se todos os fenômenos tivessem sido observados pela primeira vez no exato momento em que ocorreram.

Com um número muito pequeno de registros na série, tal como algo inferior a 50, já podemos observar alguns fenômenos interessantes, mas os resultados obtidos através de ambas as técnicas são extremamente sensíveis à entrada de um único novo elemento. No DBSCAN a ocorrência de ruídos se torna muito frequente e na STGT o formato e a intensidade máxima dos aspectos sofrem fortes alterações. Quando a série se aproxima de 100 registros, os resultados passam a apresentar mais estabilidade, considerando os valores do sinal em cada instante de tempo variando em torno de 2/3 da amplitude média.

Sendo assim, no dia 85 o DBSCAN apresentou um conjunto de quatro pontos de ruído e na STGT surgiu um aspecto de intensidade 0.5 nas frequências até 0.2, indicando um provável fim da tendência de queda do IBOVESPA. Em seguida, desse dia até o dia 132, o aspecto se estende horizontalmente ao longo do tempo, com intensidades 0.4 e 0.3 nas frequências inferiores a 0.1. No DBSCAN podemos notar que esse período apresentou grande volatilidade no índice. Sete dias depois, no dia 139 surge na STGT a intensidade 0.5 no mesmo aspecto e o DBSCAN passa a apontar uma leve tendência de alta.

Depois disso, a série cresce sem qualquer alteração em ambos os modelos até o dia 235, quando surge na STGT um pequeno aspecto de intensidade 0.3 em frequências inferiores a 0.1, ao mesmo tempo em que o DBSCAN apresenta um pequeno cluster no final da série. No dia 266 o aspecto cresce para 0.4 e o seu tamanho aumenta, enquanto o cluster começa a apresentar uma curva suave, indicando uma provável mudança na tendência do índice. Nos 18 dias subsequentes o DBSCAN apresenta uma clara tendência de queda do índice e a intensidade máxima do aspecto na STGT chega a 0.5, com um acentuado aumento no seu tamanho. A evolução dos fenômenos mencionados pode ser observada nos gráficos relativos aos dias 248 e 266, que encontram-se representados na Figura 6.3 (a), (b), (c) e (d).

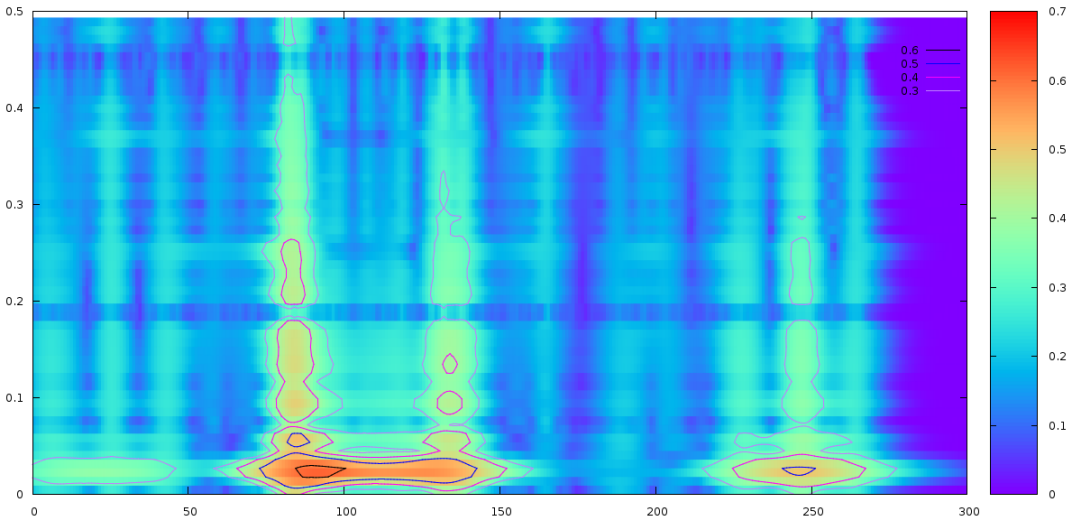
(a)



(b)

Figura 6.2: Aplicação ao IBOVESPA nos dias 248 e 266: (a) dia 248 pelo DBSCAN; (b) dia 248 pela STGT; (c) dia 266 pelo DBSCAN, (d) dia 266 pela STGT.

(c)



(d)

Figura 6.2 (continuação)

O trecho da série acima analisado parece representar um ciclo de investimento entre os dias 85 (aos 56.590 pontos) e 266 (aos 58.951 pontos), onde o potencial de ganho máximo com suporte das técnicas de análise adotadas ocorreu entre os dias 130 (aos 52.607 pontos) e 250 (aos 61.692 pontos), resultando em um retorno de 17,3% em 80 dias, que equivale a 0,13% ao dia.

Após a observação desses fenômenos, quase 90 dias transcorreram sem qualquer alteração em ambas as técnicas de análise. No DBSCAN, apesar de uma significativa volatilidade, a tendência de queda do IBOVESPA se mantém. No dia 347 ele passa a apresentar uma sequência de ruídos, com o índice atingindo um mínimo global. No mesmo dia, a STGT mostra o surgimento de um novo aspecto de intensidade 0.3 nas frequências inferiores a 0.2. Do dia 351 ao 359 a intensidade cresce para 0.4 e 0.5, juntamente com o tamanho do aspecto, enquanto o DBSCAN apresenta a formação de pequenos clusters, ainda em tendência de queda, mas já indicando uma desaceleração no processo. No dia 360 (aos 45.044 pontos), o índice renova o mínimo global e passa a apresentar uma tendência de alta.

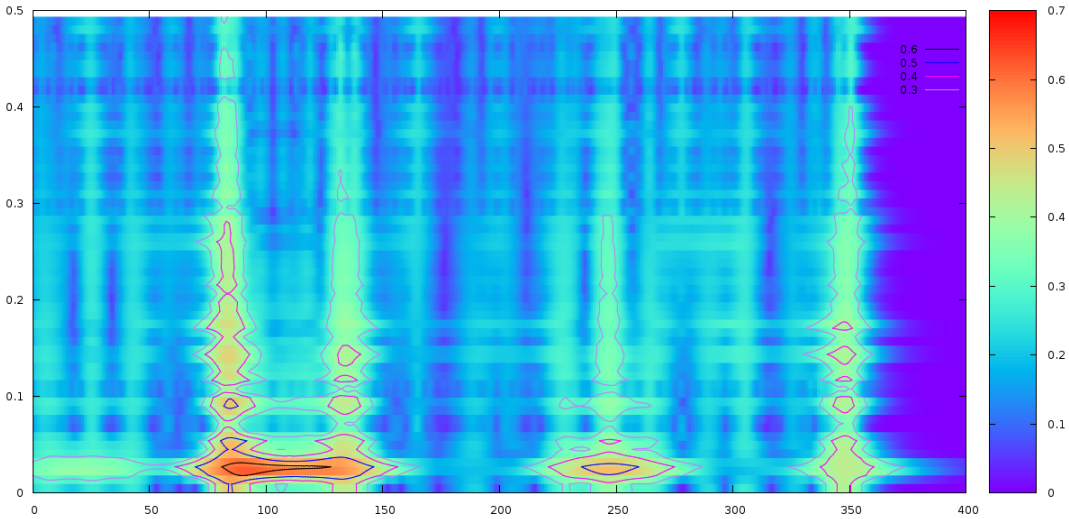
Após atingir dois máximos locais, o IBOVESPA dá início a uma fase com clara tendência de queda, sem qualquer sinal nas técnicas de análise. No entanto, entre os dia 479 e 486 a STGT mostra o surgimento e o crescimento de um novo aspecto nas frequências inferiores a 0.2, iniciando com intensidade 0.3 e chegando a 0.4 no final desse período. A tabela 8.1 a seguir apresenta a análise de retorno compreendendo esse segundo ciclo analisado.

Tabela 6.1: Análise do retorno com a aplicação ao IBOVESPA.

	Compra no dia 351 a 48.214 pontos	**Compra no dia 360 a 45.044 pontos**
Venda no dia 479 a 51.185 pontos	6,2% em 128 dias (0,05% a.d.)	13,6% em 119 dias (0,11% a.d.)
Venda no dia 486 a 50.973 pontos	5,7% em 135 dias (0,04% a.d.)	13,2% em 126 dias (0,10% a.d.)

A evolução dos fenômenos relacionados ao crescimento da série do IBOVESPA envolvendo o segundo ciclo de investimento pode ser bem observada a partir dos gráficos relativos ao período compreendido entre os dias 351 e 486, que encontram-se representados na Figura 6.3 (a), (b), (c), (d), (e), (f), (g) e (h) apresentada logo a seguir.

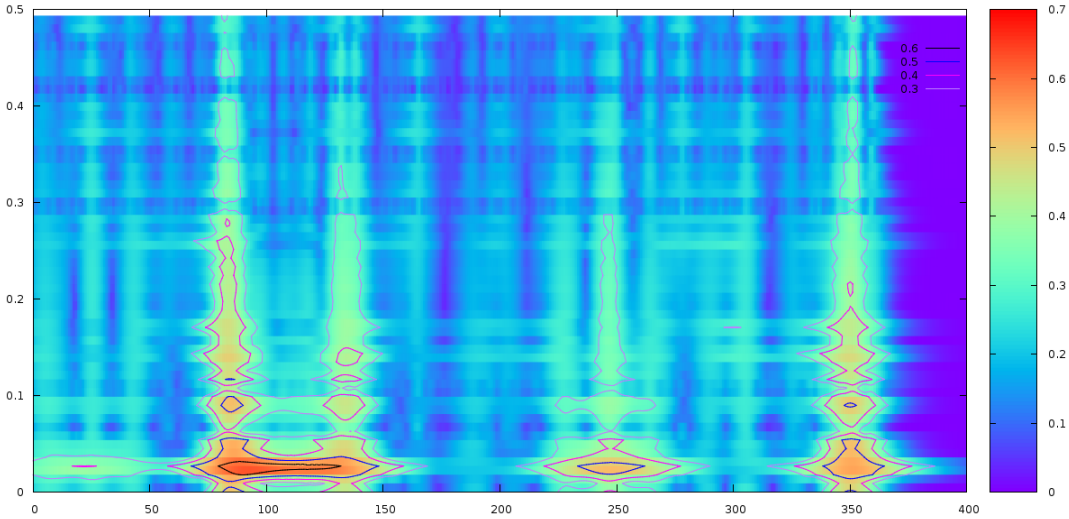
(a)



(b)

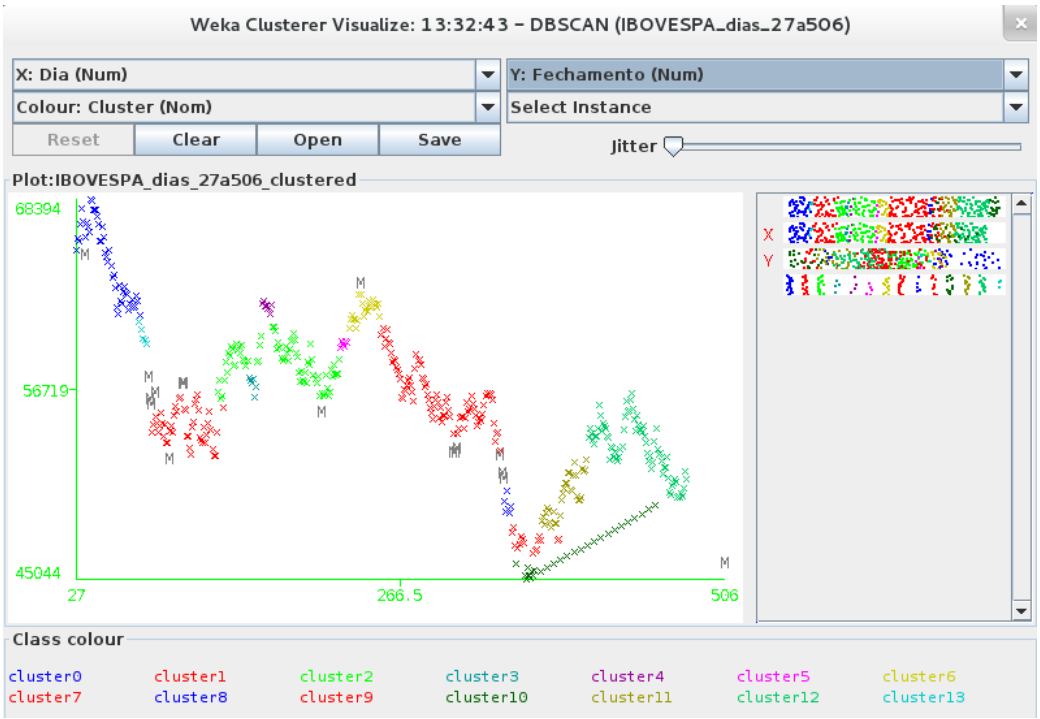
Figura 6.3: Evolução na aplicação ao IBOVESPA entre os dias 351 e 486: (a) dia 351 no DBSCAN; (b) dia 351 na STGT; (c) dia 360 no DBSCAN, (d) dia 360 na STGT; (e) dia 479 no DBSCAN; (f) dia 479 na STGT; (g) dia 486 no DBSCAN, (h) dia 486 na STGT.

(c)

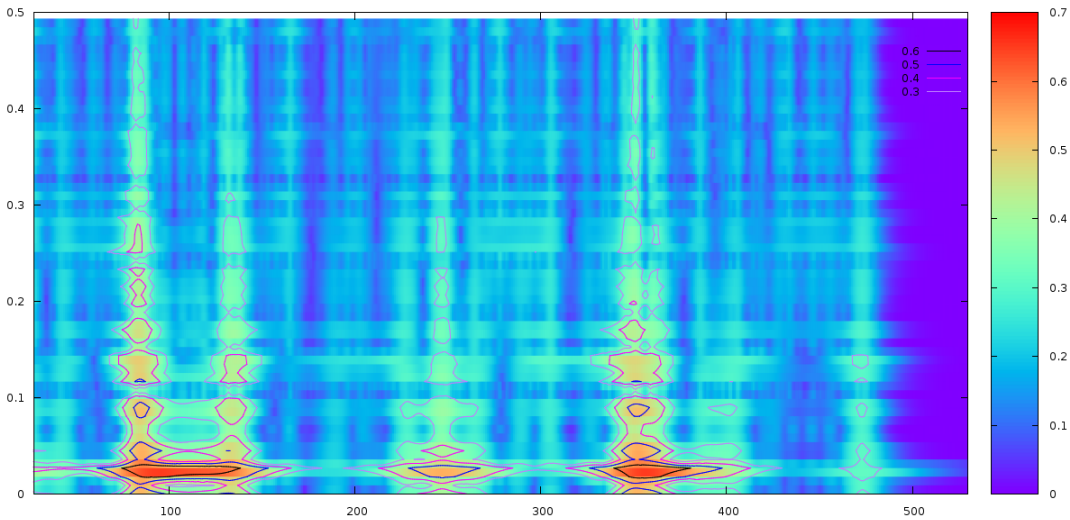


(d)

Figura 6.3 (continuação)

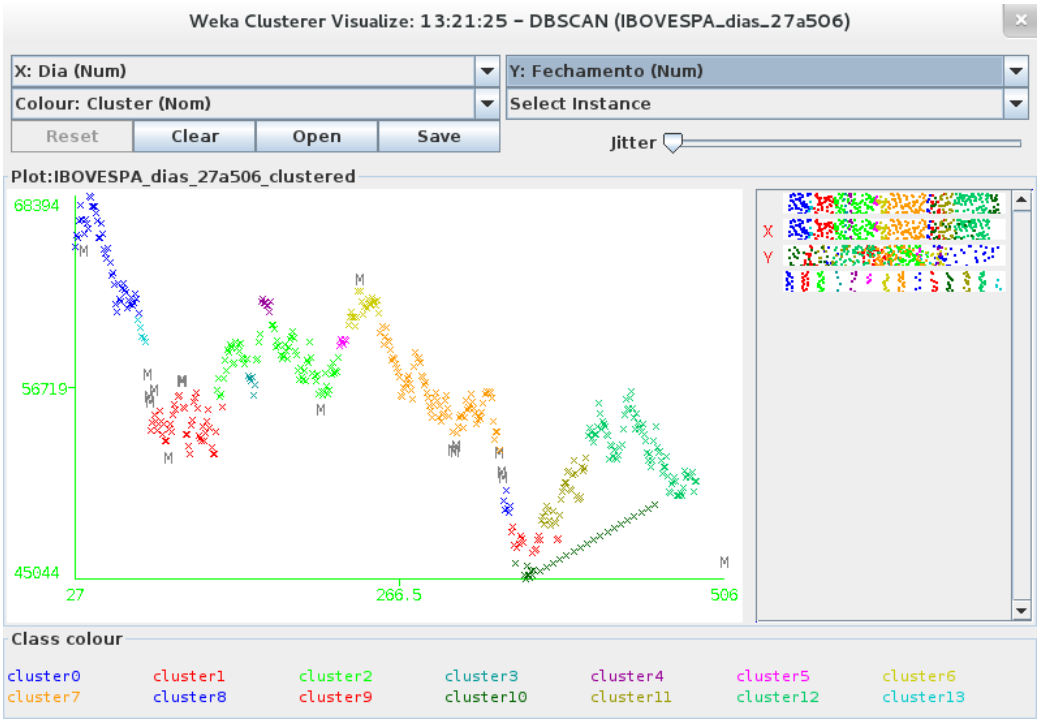


(e)

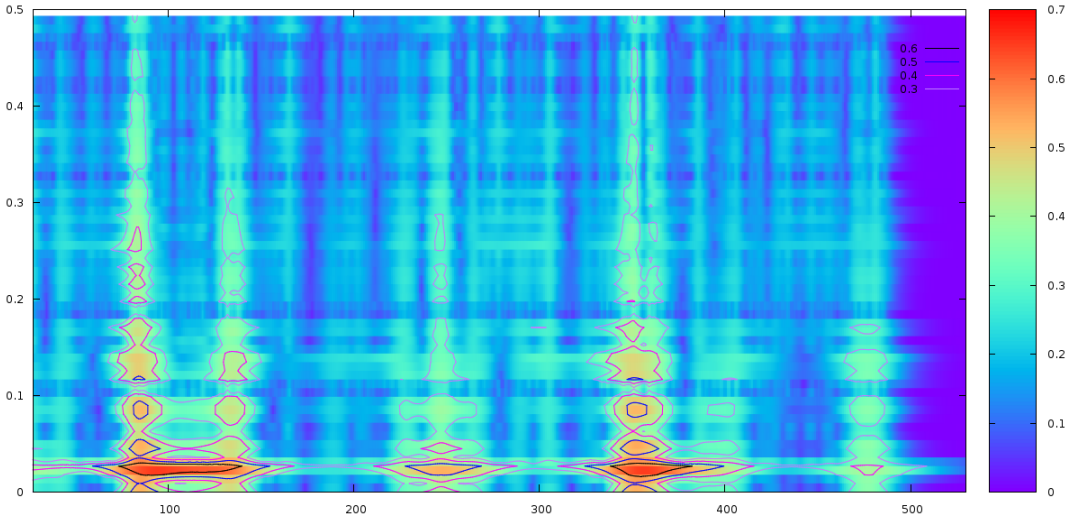


(f)

Figura 6.3 (continuação)



(g)



(h)

Figura 6.3 (continuação)

Passados mais 50 dias, o IBOVESPA renova o mínimo global no dia 533 (aos 44.965 pontos), sem qualquer indicação nas duas técnicas de análise adotadas. No dia 564 (aos 52.980 pontos), já demonstrando tendência de alta, o DBSCAN apresenta o primeiro ponto do cluster14 (cinza) logo após o cluster13 (azul claro), conforme apresentado através do terceiro ciclo de investimento assinalado na Figura 6.1 (a), e a STGT mostra um novo aspecto de intensidade 0.3 nas frequências baixas, analogamente conforme apresentado através do terceiro ciclo de investimento assinalado na Figura 6.1 (b). Mantendo a tendência de alta, apesar de uma considerável volatilidade, a série evolui apresentando a formação e a combinação de novos clusters. Do dia 601 ao 633, a STGT mostra o surgimento de um novo aspecto nas frequências baixas e o seu crescimento de uma intensidade 0.3 até 0.5, com um correspondente aumento de tamanho.

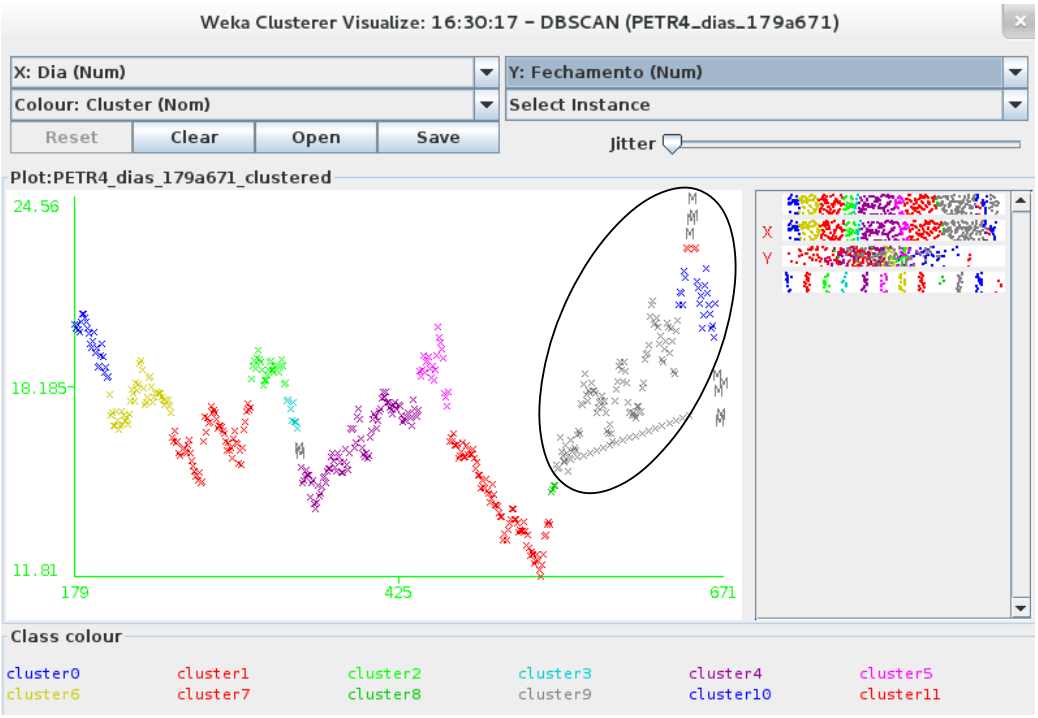
O trecho da série acima analisado parece representar um ciclo de investimento entre os dias 551 (aos 51.127 pontos) e 640 (aos 58.992 pontos), onde o potencial de ganho máximo com suporte das técnicas de análise adotadas ocorreu exatamente entre esses mesmos dias, resultando em um retorno de 15,4% em 89 dias, que equivale a 0,16% ao dia.

Ainda que os aspectos mais antigos na STGT sofressem alguma mudança de forma com o surgimento dos aspectos mais recentes, em decorrência de características inerentes à própria técnica, a série histórica se mostrava bem preservada até o surgimento de um aspecto muito pequeno no dia 660 (18/09/14), faltando pouco mais de duas semanas para o primeiro turno das eleições de 2014. O que pode ser observado desse momento em diante na STGT foi o crescimento progressivo do tamanho e da intensidade do aspecto, chegando a atingir intensidade de 0.7, juntamente com uma grande volatilidade do IBOVESPA, representada no DBSCAN pelo surgimento de vários pequenos clusters e muitos pontos de ruído.

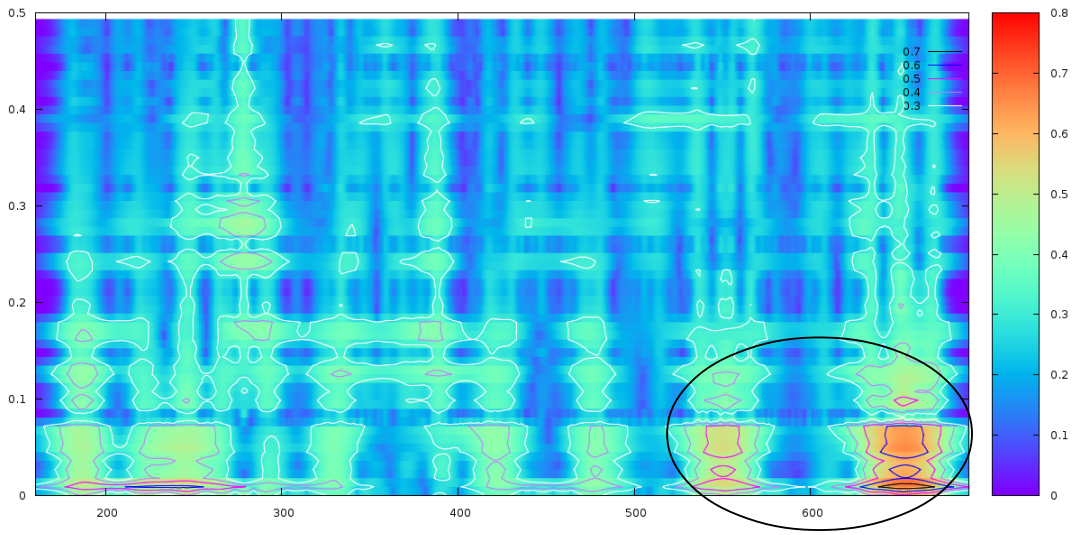
6.2 – Aplicação ao Ativo PETR4

A segunda série temporal utilizada, cuja análise só passou a ser realizada após a constatação de fenômenos relevantes na série do IBOVESPA, teve início vários meses após a primeira e envolve o período transcorrido de 04/10/2012 (ou dia 179) até o último dia útil antes do primeiro turno das eleições de 2014 (03/10/2014, ou dia 671). A clusterização foi realizada com minPoints=4 e epsilon=0.040 como parâmetros.

A fraca sinalização do início do terceiro ciclo de investimento do IBOVESPA foi o principal motivador para essa aplicação. O resultado final da aplicação das técnicas de análise à série do ativo PETR4 está representado a seguir através das Figuras 6.4 (a) e (b).



(a)



(b)

Figura 6.4: Resultados da aplicação ao ativo PETR4: (a) pelo DBSCAN; (b) pela STGT.

A série do ativo PETR4 apresenta um único ciclo de investimento, conforme assinalado nas figuras, identificado pela sua curva de retorno no DBSCAN, criada com a adoção de uma taxa de 0,1% ao dia, e os correspondentes trechos de aumento de intensidade na STGT. O ativo apresentou um potencial total de ganho de 108,0% em 115 dias, equivalente a 0,64% ao dia, entre o mínimo global do dia 533 (a R$11,81) e o máximo global do dia 648 (a R$24,56) tendo apresentado uma forte influência no terceiro ciclo de investimento identificado na análise da série do IBOVESPA. Dois aspectos qualitativos muito importantes devem ser destacados como resultado da análise dessa série:

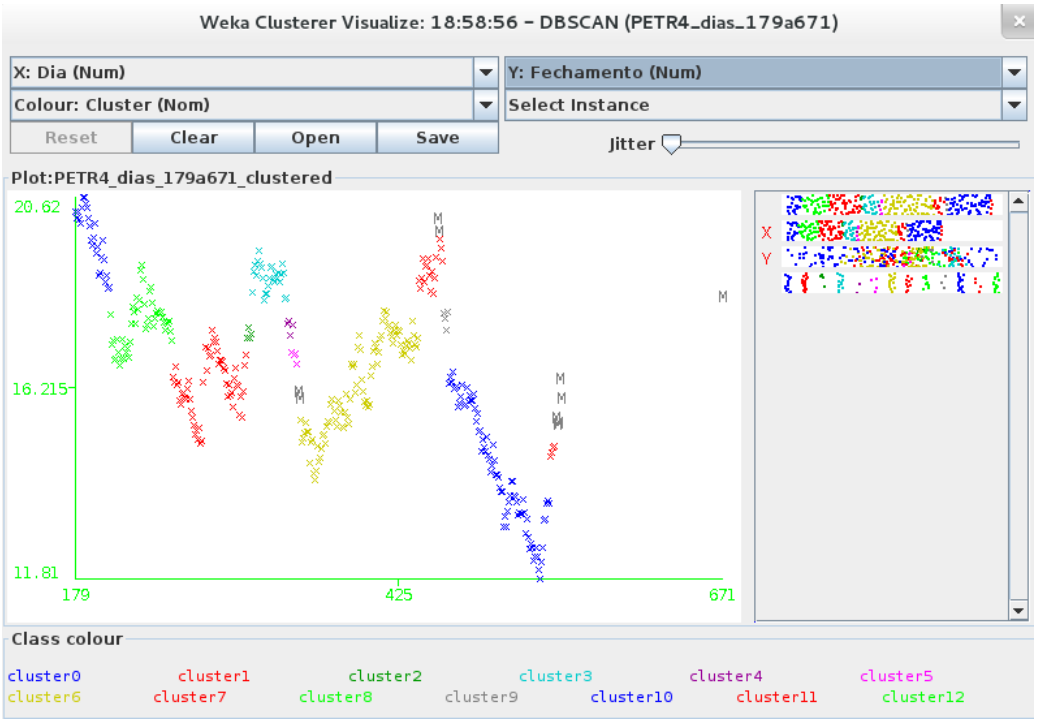
- a constatação de que a sinalização do início de um ciclo de investimento pode não coincidir com o exato momento em que é observada a reversão da tendência da série temporal;
- de forma análoga, a constatação de que a sinalização do término de um ciclo de investimento pode não estar associado a um cluster contendo um máximo local, mas também a uma sequência de ruídos.

Independentemente desses novos fenômenos observados, o uso conjunto do DBSCAN e da STGT demonstrou ser de grande ajuda nas decisões envolvendo investimento no ativo avaliado, permitindo um retorno maior do que com o IBOVESPA, além de apresentar uma sinalização mais clara do início e do final do ciclo identificado. A tabela 6.2 a seguir apresenta a análise de retorno compreendendo esse ciclo analisado.

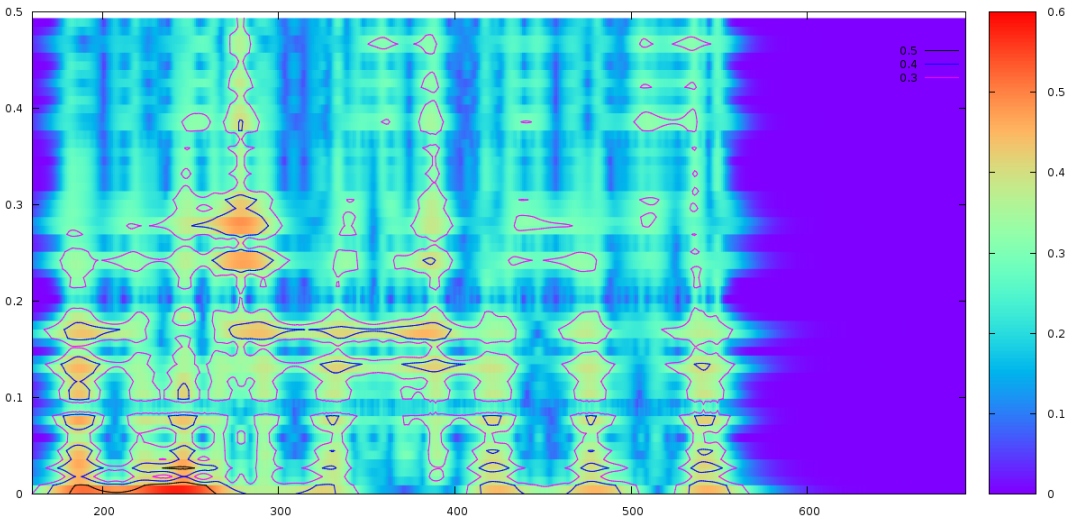
Tabela 6.2: Análise do retorno com a aplicação ao ativo PETR4.

	Compra no dia 549 a R$15,99	**Compra no dia 557 a R$15,96**
Venda no dia 644 a R$22,84	42,8% em 95 dias (0,38% a.d.)	43,1% em 87 dias (0,41% a.d.)
Venda no dia 651 a R$22,82	42,7% em 102 dias (0,35% a.d.)	43,0% em 94 dias (0,38% a.d.)

A evolução dos fenômenos relacionados ao crescimento da série do ativo PETR4 envolvendo o ciclo de investimento identificado pode ser bem observada a partir dos gráficos relativos ao período compreendido entre os dias 549 e 651, que encontram-se representados na Figura 6.5 (a), (b), (c), (d), (e), (f), (g) e (h) apresentada logo a seguir.

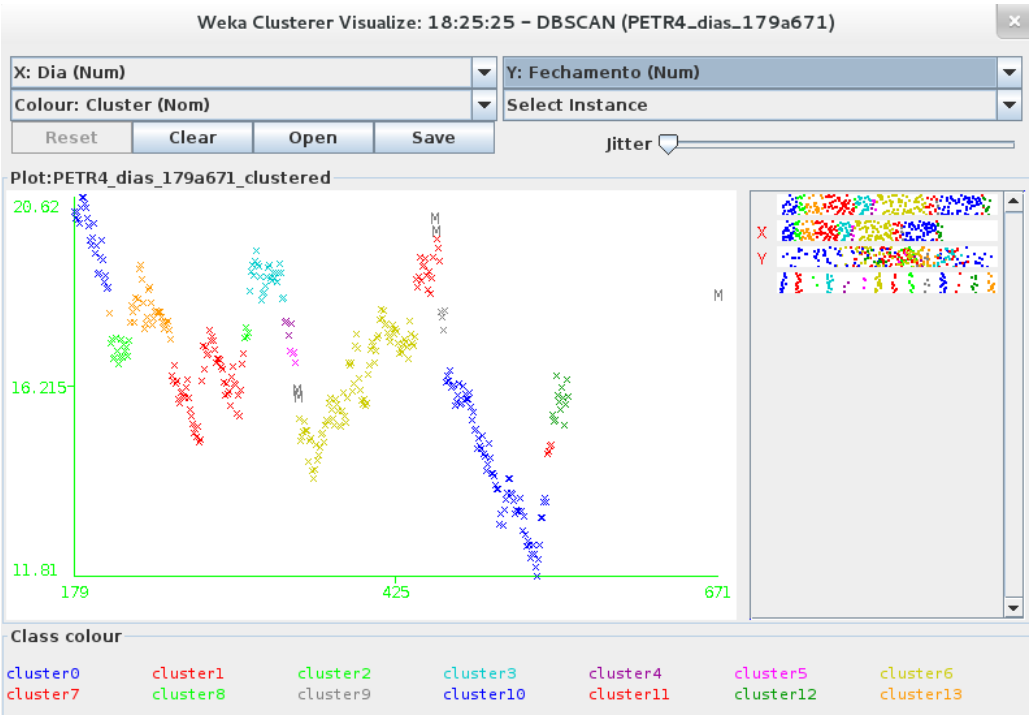


(a)

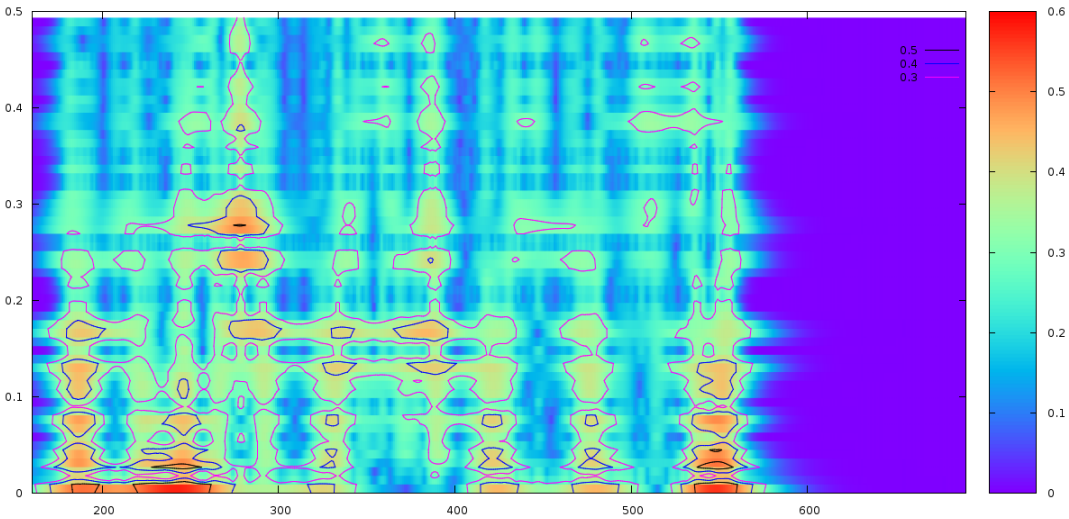


(b)

Figura 6.5: Evolução na aplicação ao ativo PETR4 entre os dias 549 e 651: (a) dia 549 no DBSCAN; (b) dia 549 na STGT; (c) dia 557 no DBSCAN, (d) dia 557 na STGT; (e) dia 644 no DBSCAN; (f) dia 644 na STGT; (g) dia 651 no DBSCAN, (h) dia 651 na STGT.



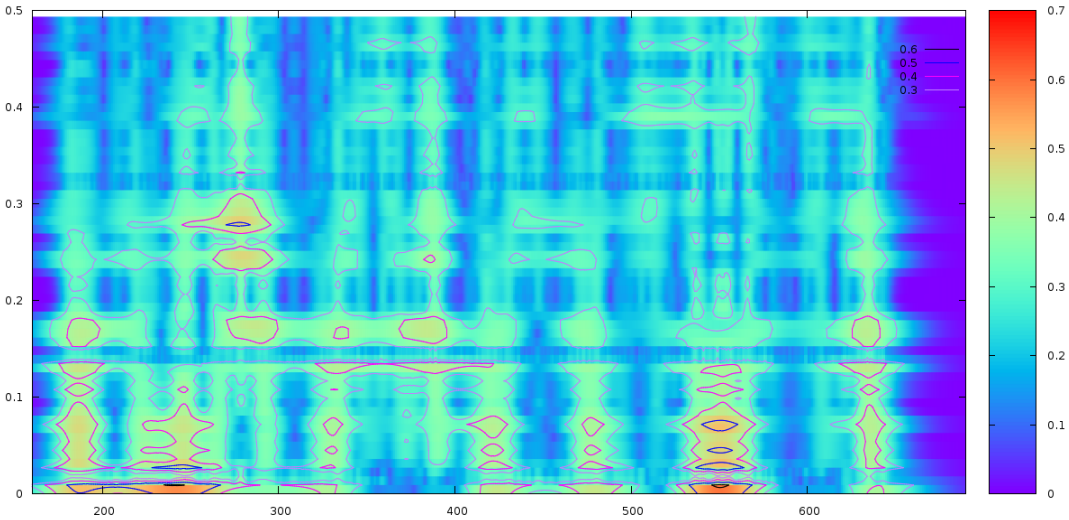
(c)



(d)

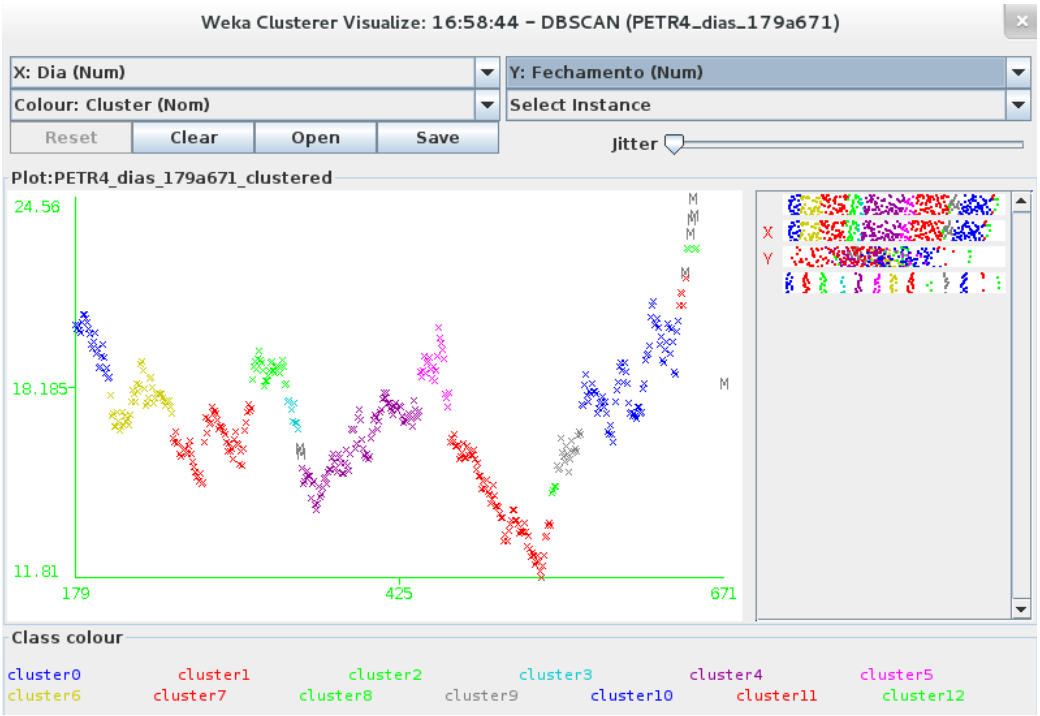
Figura 6.5 (continuação)

(e)

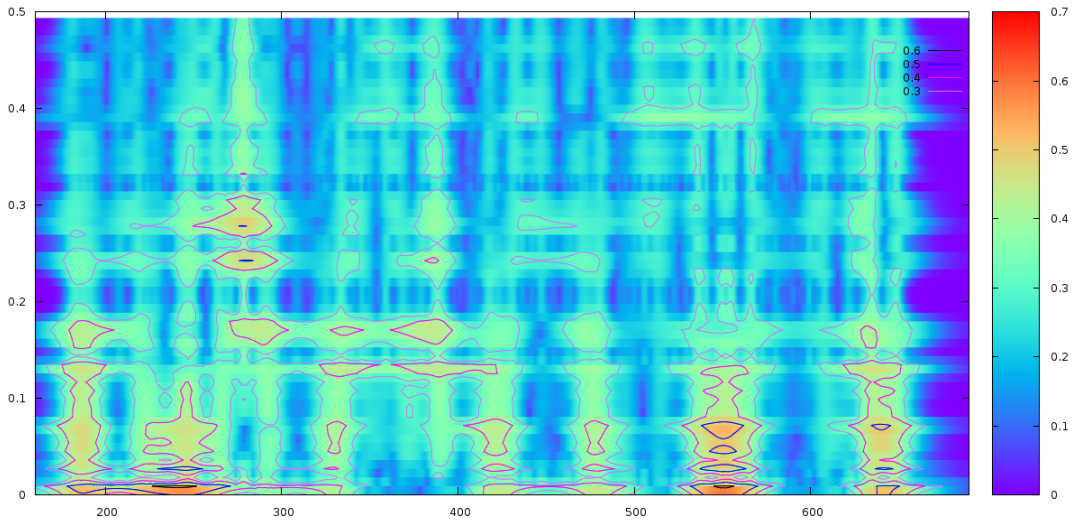


(f)

Figura 6.5 (continuação)



(g)



(h)

Figura 6.5 (continuação)

6.3 – Avaliação de Desempenho do Modelo Proposto

Conforme mencionado anteriormente, segundo Colman [1982; apud Kelly, 2003] em jogos de chance envolvendo incerteza não é possível atribuir uma probabilidade que contenha algum significado a quaisquer das respostas da natureza às ações do jogador. Então como deve agir um jogador ao tomar decisões nessas circunstâncias?

Nas próximas subseções será apresentado um critério de avaliação do modelo proposto, seu desempenho será comparado com outros métodos tradicionalmente usados em análise técnica de investimentos, incluindo novas aplicações a três outras séries temporais, e a sua contribuição será destacada.

6.3.1 – Critério de Avaliação

Ainda segundo Kelly [2003], existem três princípios que podem ser aplicados na tomada de decisão em situações envolvendo incerteza. O primeiro recomenda que o jogador escolha a alternativa que apresenta o maior ganho estimado; essa abordagem é considerada de extremo otimismo. O segundo recomenda que o jogador evite a alternativa que apresenta a maior perda estimada; essa abordagem é considerada de extremo pessimismo. O terceiro e último adota uma abordagem intermediária entre a otimista e a pessimista, recomendando que o jogador evite a alternativa que possa lhe causar o maior arrependimento. Em outras palavras, se há uma oportunidade ou sinal de ganho potencial com um ativo ele deve evitar não comprá-lo e, por outro lado, se há uma ameaça ou sinal de perda potencial com um outro ativo ele deve evitar não vendê-lo.

Assim, em meio a inúmeras dúvidas a respeito de qual a melhor decisão a ser tomada, o indivíduo que avalia investimentos procura combinar uma certa dose de otimismo com um comportamento cauteloso, tomando como base as informações disponíveis no mercado e os métodos de análise com os quais tem familiaridade. Muitas vezes, a combinação de dois ou mais métodos pode trazer o balanceamento adequado entre o otimismo e a cautela, de forma a obter um retorno satisfatório do investimento realizado, sem que necessariamente o ganho seja aquele apresentado pelo potencial máximo do ativo na circunstância e, eventualmente, podendo trazer algumas perdas em determinados momentos ao longo do período em que o investimento é realizado.

Conforme apresentado anteriormente nas seções envolvendo as aplicações do modelo proposto ao IBOVESPA e ao ativo PETR4, o critério a ser adotado para avaliar o desempenho do modelo será a taxa de retorno total obtida ao longo do período completo do investimento, a sua taxa equivalente ao dia e a duração total do período. Esses valores devem

ser comparados aos de outros investimentos disponíveis no mercado, como os da caderneta de poupança e os de fundos de investimento em renda fixa ou renda variável, bem como com os valores obtidos em investimentos nos mesmos ativos quando as decisões forem todas através da utilização de métodos convencionais de análise técnica. O objetivo nesse último caso não é necessariamente obter uma taxa de retorno superior a todo e qualquer método, mas identificar de forma clara a aplicabilidade do método proposto a um perfil de investidor, que pode ser diferente daqueles que adotam, ou são forçados a adotar, os métodos de análise técnica tradicionalmente usados no mercado de ações.

6.3.2 – Outros Métodos de Análise Técnica

Conforme mencionado anteriormente, as aplicações do modelo proposto ao IBOVESPA e ao ativo PETR4 deveriam ter seus resultados comparados aos obtidos através da aplicação de modelos tradicionais de análise técnica. Para isso, foram escolhidos os métodos denominados Bandas de Bollinger e *MACD – Moving Average Convergence Divergence*.

As Bandas de Bollinger são formadas por uma média móvel simples do valor de fechamento que forma a chamada banda média (*Bollinger Mid*) associada a uma banda superior (*Bollinger High*) e outra inferior (*Bollinger Low*). Os parâmetros utilizados são:

- Banda Média = Média Móvel Simples de 20 períodos;
- Banda Superior = Banda Média + (2 * desvio padrão do fechamento);
- Banda Inferior = Banda Média – (2 * desvio padrão do fechamento).

Nas Bandas de Bollinger os momentos de agir são determinados pelos pontos em que o valor do ativo atravessa as bandas inferior e superior. Portanto, quando o preços rompem a banda superior é um sinal de que o preço do ativo está mais alto do que deveria e poderá inverter o seu movimento; é o momento de vender. Ao contrário, quando o preço rompe a banda inferior é um sinal de que o preço está baixo demais e poderá apresentar uma reversão da tendência; é o momento de comprar.

Os profissionais do mercado de ações orientam que o uso de um único método de análise isoladamente é uma forma muita arriscada de operar nesse mercado e, dessa forma, outro método foi selecionado dentre os vários adotados no mercado para permitir a utilização conjunta de dois métodos de análise técnica.

A *MACD – Moving Average Convergence Divergence* é formado por duas linhas, sendo a primeira linha a diferença entre duas médias móveis exponenciais (MME) do valor de fechamento calculadas para dois diferentes períodos (*MACD*) e a segunda linha uma nova

média móvel exponencial dos valores da primeira linha calculada para um terceiro período (*Signal*), juntamente com um histograma que representa a diferença entre os valores das duas linhas para cada ponto no tempo (*Divergence*). Os parâmetros utilizados são:

- Linha *MACD* = MME(fechamento) 12 períodos – MME(fechamento) 26 períodos;
- Linha *Signal* = MME(linha *MACD*) 9 períodos;
- Histograma *Divergence* = Linha *MACD* – Linha *Signal*.

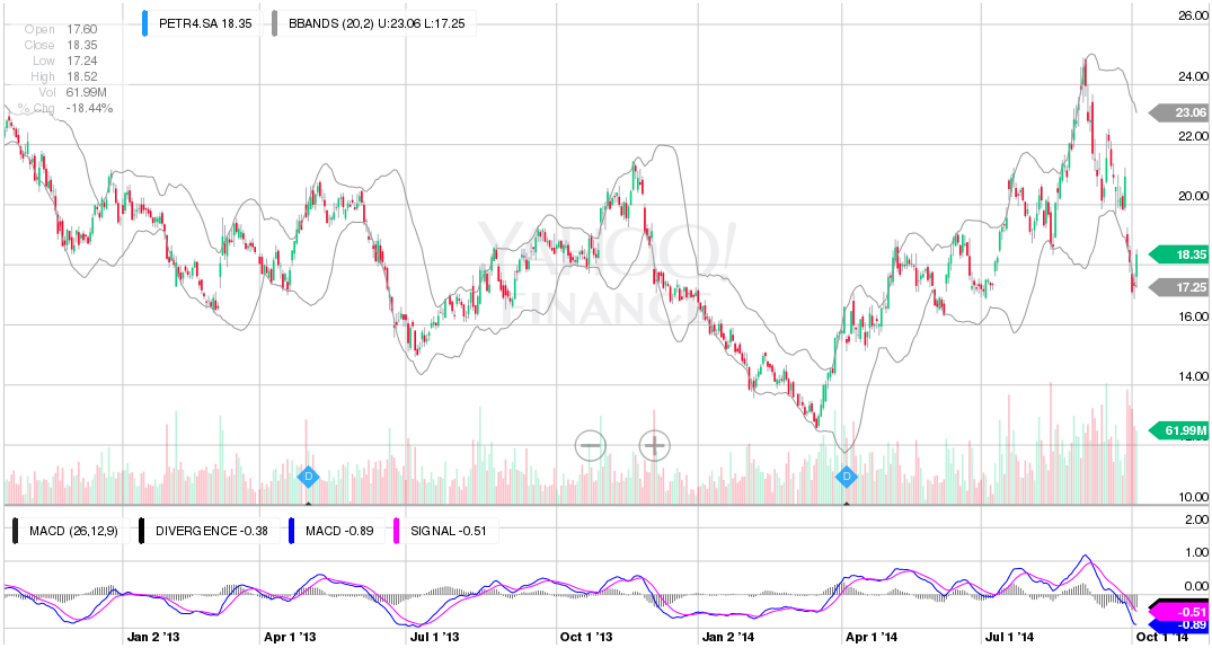
Na *MACD* os momentos de agir são determinados pelos pontos em que as linhas *MACD* e *Signal* se cruzam. Dessa forma, quando as linhas se cruzam e o histograma *Divergence* passa a apresentar valores positivos é um sinal de tendência de elevação do preço do ativo; é o momento de comprar. Já quando as linhas se cruzam e o histograma *Divergence* passa a apresentar valores negativos é um sinal de tendência de queda do preço do ativo; é o momento de vender.

Os métodos selecionados representam os mais tradicionais modelos de análise técnica adotados pelos profissionais que atuam no mercado de ações. No entanto é importante ressaltar que, quando utilizadas em conjunto, as duas técnicas selecionadas podem apresentar sinais conflitantes simultaneamente, indicando ações contraditórias ao investidor. Nesses casos, cabe exclusivamente a ele tomar a decisão quanto ao que fazer e assumir a responsabilidade pela ação executada.

As Figuras 6.6 (a) e (b) a seguir apresentam as séries do IBOVESPA e do ativo PETR4 analisadas anteriormente, dessa vez modeladas pelos métodos de análise técnica que acabaram de ser descritos. Logo a seguir, nas Tabela 6.3 (a) e (b) , são apresentados os resultados de cada ciclo de compra e venda indicado pelo modelo tradicional para cada uma das séries, juntamente com os valores agregados que permitem a comparação com aqueles obtidos através da aplicação do modelo proposto e apresentados nas Tabelas 6.1 e 6.2 deste mesmo Capítulo.

Pela comparação entre os resultados, é possível constatar melhores desempenhos para as aplicações envolvendo o modelo tradicional, tanto para a taxa de retorno total obtida ao longo do período completo do investimento quanto para a sua taxa equivalente ao dia e a duração total do período.

(a)



(b)

Figura 6.6: Gráficos[1] tipo *candlestick* com Bandas de Bollinger e MACD: (a) IBOVESPA de 27/02/12 a 03/10/14; (b) PETR4 de 10/10/12 a 03/10/14.

1 Sinais (–) e (+) nas figuras são ícones da ferramenta gráfica do *YAHOO! Finance*.

Tabela 6.3: Análise do retorno com o modelo tradicional: (a) IBOVESPA; (b) ativo PETR4.

IBOVESPA	Data	Pontos	Técnica	Dias	%total	%diário
Comprar	10/07/13	45.483	MACD			
Vender	12/08/13	50.299	BB	33	10,59%	0,31%
Comprar	03/09/13	51.626	MACD			
Vender	06/09/13	53.749	BB	3	4,11%	1,35%
Comprar	15/10/13	54.981	MACD			
Vender	16/10/13	55.973	BB	1	1,80%	1,80%
Comprar	08/11/13	52.249	BB			
Vender	11/11/13	52.624	MACD	3	0,72%	0,24%
Comprar	03/12/13	50.394	BB			
Vender	10/01/14	49.696	MACD	38	-1,39%	-0,04%
5 ciclos				78	15,84%	0,19%

(a)

PETR4	Data	Pontos	Técnica	Dias	%total	%diário
Comprar	19/03/14	13,34	MACD			
Vender	27/03/14	15,57	BB	8	16,72%	1,95%
Comprar	01/04/14	15,81	MACD			
Vender	02/04/14	16,66	BB	1	5,38%	5,38%
Comprar	03/04/14	15,40	MACD			
Vender	02/05/14	17,60	BB	29	14,29%	0,46%
Comprar	09/05/14	17,64	MACD			
Vender	19/05/14	17,94	MACD	10	1,70%	0,17%
Comprar	30/05/14	16,65	BB			
Vender	03/06/14	16,90	MACD	4	1,50%	0,37%
Comprar	05/06/14	16,32	BB			
Vender	18/07/14	20,52	BB	43	25,74%	0,53%
Comprar	23/07/14	20,26	MACD			
Vender	31/07/14	19,10	MACD	8	-5,73%	-0,73%
Comprar	22/08/14	20,92	MACD			
Vender	25/08/14	22,04	BB	3	5,35%	1,75%
Comprar	26/08/14	21,84	MACD			
Vender	27/08/14	22,84	BB	1	4,58%	4,58%
Comprar	28/08/14	22,80	MACD			
Vender	02/09/14	24,56	BB	5	7,72%	1,50%
10 ciclos				60	77,24%	0,96%

(b)

Com o objetivo de realizar uma comparação mais cuidadosa entre os desempenhos dos dois tipos de modelo e, com isso, investigar as possíveis causas da obtenção de resultados eventual ou sistematicamente melhores através do emprego de um deles, foram realizadas três outras aplicações envolvendo os ativos BRFS3 (Brasil Foods ON – Bovespa), ITUB4 (Itaú Unibanco PN – Bovespa) e FB (Facebook – Nasdaq), apresentados a seguir.

6.3.3 – Outras Aplicações

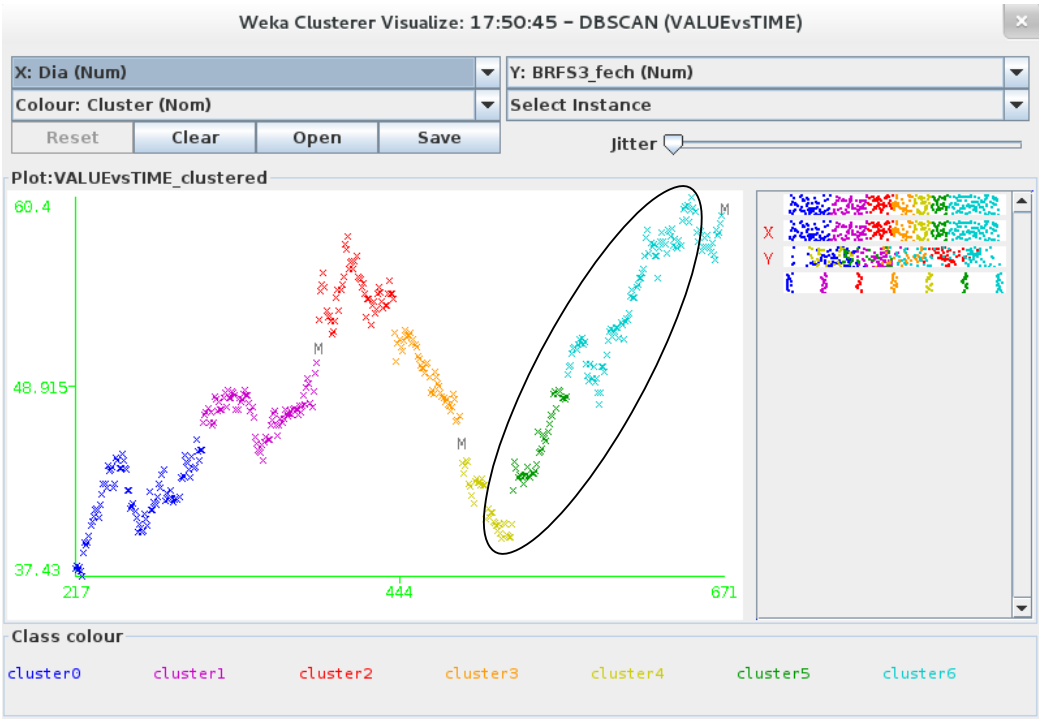
As duas primeiras séries temporais utilizadas nesta nova etapa do trabalho, referentes aos ativos BRFS3 (Brasil Foods ON – Bovespa) e ITUB4 (Itaú Unibanco PN – Bovespa), envolvem um período semelhante ao das séries do IBOVESPA e do ativo PETR4 já estudadas, sendo apenas um pouco menores: vão do primeiro dia útil de dezembro de 2012 (03/12/2012, ou dia 217) até o último dia útil antes do primeiro turno das eleições de 2014 (03/10/2014, ou dia 671). Utilizando esse período, é possível manter a influência dos mesmos fenômenos que produziram as séries anteriormente analisadas nessas novas séries, sem introduzir um contexto completamente diferente no estudo. A clusterização foi realizada com minPoints=4 e espsilon=0.04 como parâmetros.

Já com a terceira série, referente ao ativo FB (Facebook – Nasdaq), não houve a mesma preocupação, por se tratar de um ativo negociado em um mercado com características bem diferentes do mercado brasileiro. Na verdade, o objetivo com essa última série foi exatamente este: analisar o desempenho dos modelos quando aplicados a um ativo que sofre influências diferentes das que são sentidas no mercado que já estava fazendo parte do estudo, procurando manter um cenário de relativa estabilidade econômica e com bastante atratividade para com os investidores. Assim, o período dessa última série foi de 07 de fevereiro de 2014 a 10 de agosto de 2015. A clusterização também foi realizada com minPoints=4 e espsilon=0.04 como parâmetros.

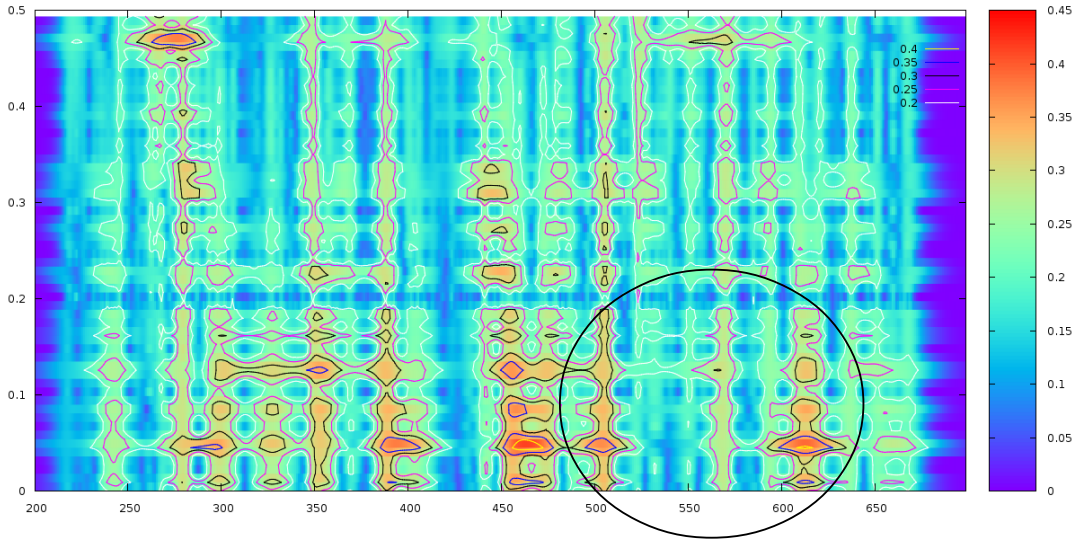
As políticas de compra e venda sinalizadas pelo modelo proposto apresentaram as seguintes características, que foram assinaladas com curvas ovais no DBSCAN e na STGT:

- para BRFS3, um único ciclo, envolvendo um único momento de compra e um único momento de venda;
- para ITUB4, um único ciclo, envolvendo um único momento de compra e quatro momentos de venda (1, 2, 3 e final);
- Para FB, um único ciclo, envolvendo um único momento de compra e dois momentos de venda (parcial e final).

O resultado final da aplicação às novas séries temporais está representado a seguir através das Figuras 6.7 (a), (b), (c), (d), (e) e (f). Logo a seguir, nas Tabela 6.4 (a), (b) e (c) , são apresentados os resultados de cada ciclo de compra e venda indicado pelo modelo proposto para cada uma das séries.

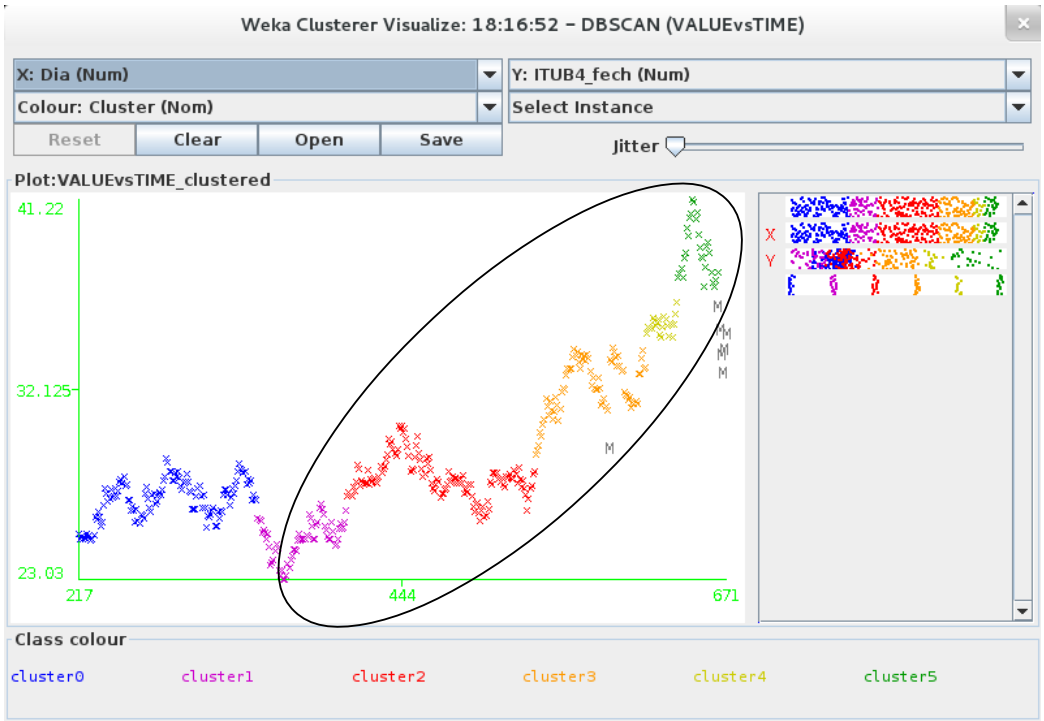


(a)

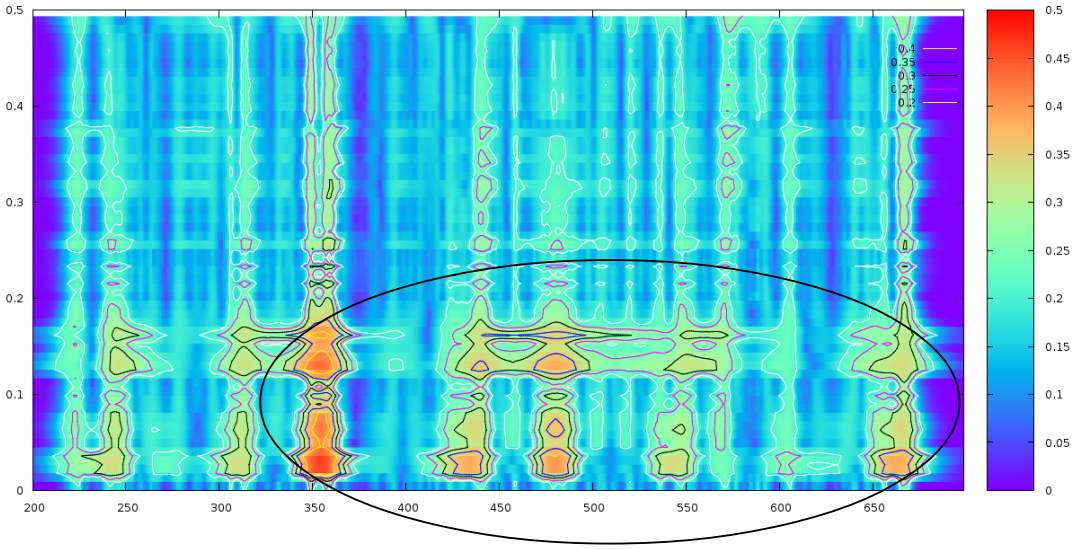


(b)

Figura 6.7: Novas aplicações do modelo proposto: (a) DBSCAN de BRFS3; (b) STGT de BRFS3; (c) DBSCAN de ITUB4; (d) STGT de ITUB4; (e) DBSCAN de FB; (f) STGT de FB.

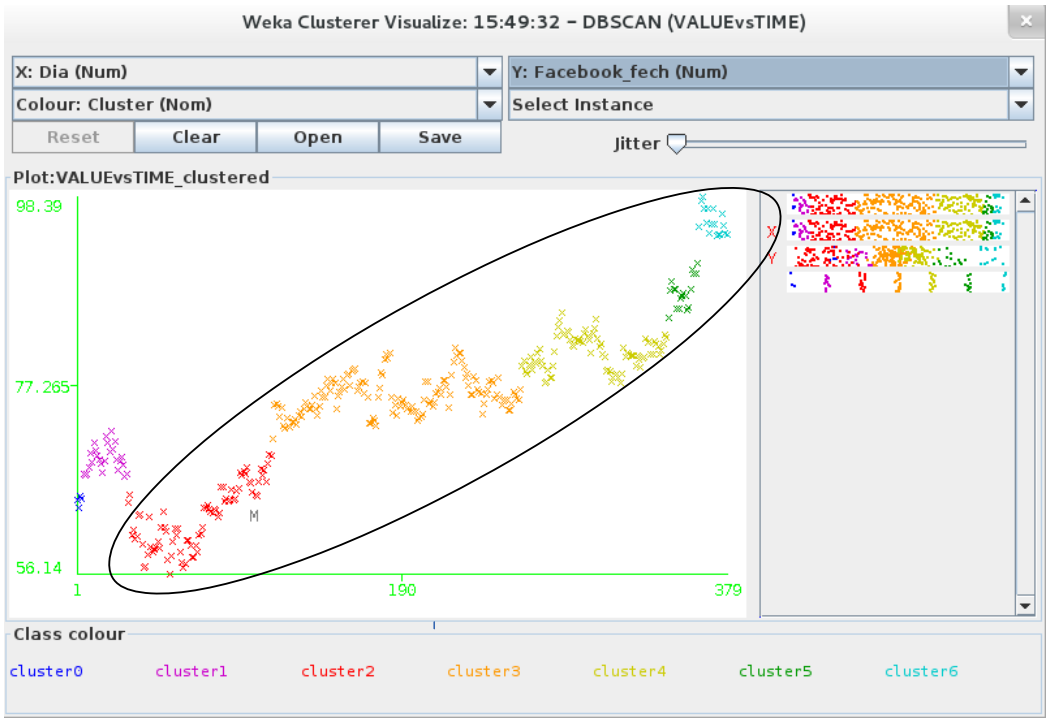


(c)

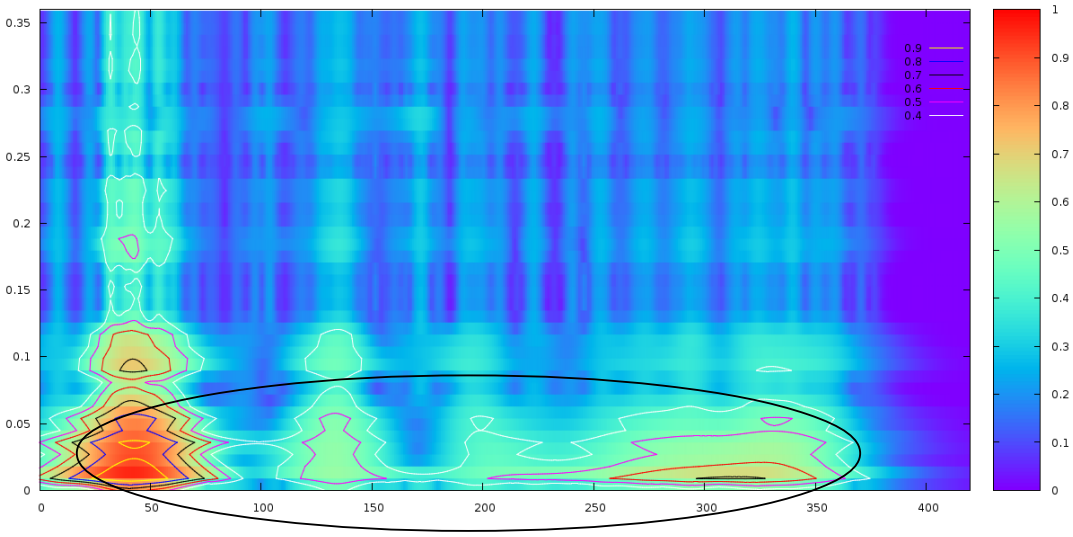


(d)

Figura 6.7 (continuação)



(e)



(f)

Figura 6.7 (continuação)

Tabela 6.4: Retorno das novas aplicações– modelo proposto: (a) BRFS3; (b) ITUB4; (c) FB.

Compra média	dia	520,0	@	42,07	
Venda média	dia	640,0	@	58,07	
Retorno médio	em	120,0 dias		38,03 %	0,27 %ad

(a)

Compra média	dia	365,0	@	24,34	
Venda média (1)	dia	430,0	@	28,43	
Retorno médio	em	65,0 dias		16,82 %	0,24 %ad
Compra média	dia	365,0	@	24,34	
Venda média (2)	dia	480,0	@	27,54	
Retorno médio	em	115,0 dias		13,14 %	0,11 %ad
Compra média	dia	365,0	@	24,34	
Venda média (3)	dia	570,0	@	33,63	
Retorno médio	em	205,0 dias		38,18 %	0,16 %ad
Compra média	dia	365,0	@	24,34	
Venda média (final)	dia	657,5	@	38,32	
Retorno médio	em	292,5 dias		57,45 %	0,16 %ad

(b)

Compra média	dia	42,5	@	59,68	
Venda média (parcial)	dia	135,0	@	74,7	
Retorno médio	em	92,5 dias		25,19 %	0,24 %ad
Compra média	dia	42,5	@	59,68	
Venda média (final)	dia	325,0	@	83,90	
Retorno médio	em	282,5 dias		40,57 %	0,12 %ad

(c)

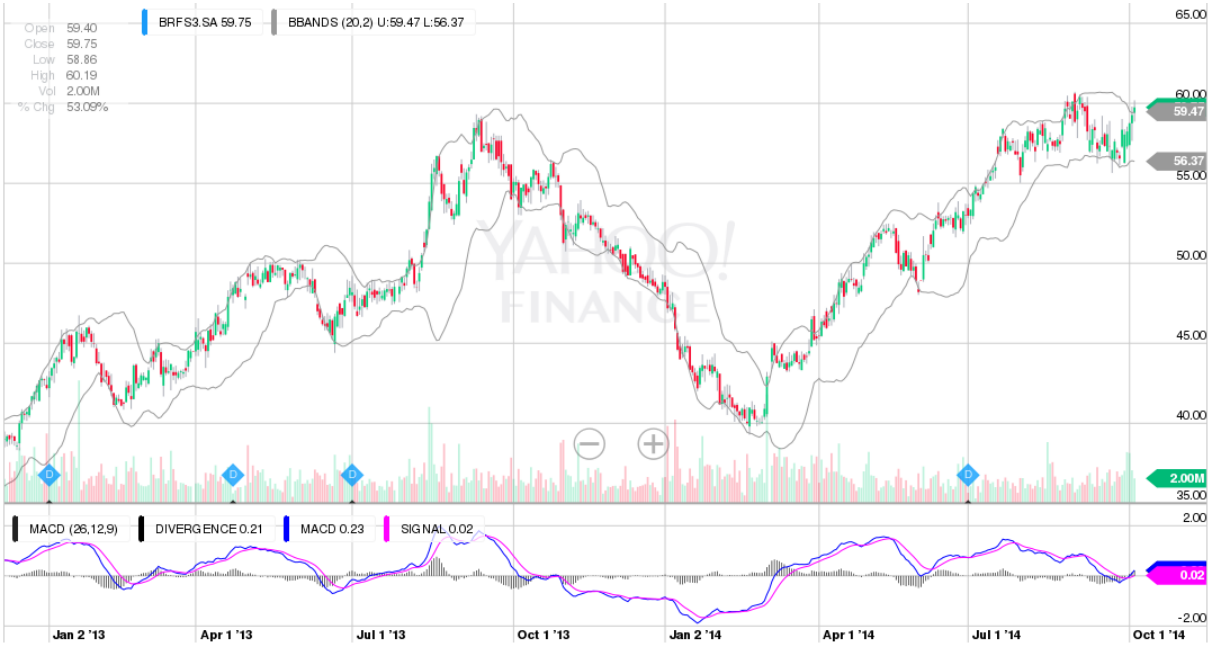
De maneira análoga à que foi realizada para as séries temporais do IBOVESPA e do ativo PETR4, foi aplicado um modelo tradicional de análise técnica, envolvendo os métodos de Bandas de Bollinger e *MACD – Moving Average Convergence Divergence*, às novas séries dos ativos BRFS3, ITUB4 e FB.

Conforme já mencionado, os profissionais do mercado de ações orientam que o uso de um único método de análise isoladamente é uma forma muita arriscada de operar nesse mercado e, dessa forma, os dois métodos de análise técnica foram utilizados de forma conjunta. Mais uma vez, exatamente por serem utilizadas em conjunto, as duas técnicas selecionadas apresentaram sinais conflitantes simultaneamente, indicando ações contraditórias ao investidor. Nesses casos, foi necessário tomar uma decisão pessoal quanto ao que fazer e assumir a responsabilidade pela ação executada.

As Figuras 6.8 (a), (b) e (c) a seguir apresentam as séries dos ativos BRFS3, ITUB4 e FB analisadas anteriormente, dessa vez modeladas pelos métodos de análise técnica tradicionais. Logo a seguir, nas Tabela 6.5 (a), (b) e (c), são apresentados os resultados de cada ciclo de compra e venda indicado pelo modelo tradicional para cada uma das séries, juntamente com os valores agregados que permitem a comparação com aqueles obtidos através da aplicação do modelo proposto e apresentados na Tabelas 6.4 deste mesmo Capítulo.

Pela comparação entre os resultados, é possível constatar que o modelo proposto foi capaz de produzir um retorno total maior para o ativo BRFS3, embora a duração total do ciclo tenha sido mais do que o dobro daquele obtido através do modelo tradicional, o que também resultou em uma taxa equivalente ao dia menor do que aquela apresentada pelo modelo tradicional. Para os ativos ITUB4 e FB, o modelo tradicional confirmou os resultados anteriormente obtidos para o IBOVESPA e o ativo PETR4, obtendo melhores desempenhos tanto para a taxa de retorno total obtida ao longo do período completo do investimento quanto para a sua taxa equivalente ao dia e a duração total do período.

Os resultados obtidos com a comparação entre os dois modelos, apesar de parecerem prejudiciais à proposta desenvolvida neste trabalho, são bastante coerentes e, de certa forma, já eram esperados. Esses resultados têm como principal razão o fato do modelo proposto ser o único capaz de identificar ciclos de investimento mais longos, que é uma funcionalidade difícil de ser encontrada em outros métodos de análise técnica. Na próxima subseção, os resultados obtidos serão comentados com mais detalhes e as contribuições do modelo proposto serão destacadas.



(a)

(b)

Figura 6.8: Gráficos[2] das aplicações no modelo tradicional: (a) BRFS3; (b) ITUB4; (c) FB.

2 Sinais (–) e (+) nas figuras são ícones da ferramenta gráfica do *YAHOO! Finance*.

(c)

Figura 6.8 (continuação)[3]

Tabela 6.5: Retorno das novas aplicações– modelo tradicional: (a) BRFS3; (b) ITUB4; (c) FB.

BRFS3	Data	R$	Técnica	Dias	%total	%diário
Comprar	25/04/13	48,48	MACD			
Vender	16/05/13	52,00	MACD	21	7,26%	0,33%
Comprar	30/05/13	48,20	BB			
Vender	04/07/13	54,77	BB	35	13,63%	0,37%
Comprar	01/10/13	58,75	MACD			
Vender	03/10/13	59,75	BB	2	1,70%	0,85%
3 ciclos				58	22,59%	0,35%

(a)

3 Sinais (–) e (+) nas figuras são ícones da ferramenta gráfica do *YAHOO! Finance*.

Tabela 6.5 (continuação)

ITUB4	Data	R$	Técnica	Dias	%total	%diário
Comprar	12/07/13	22,63	MACD			
Vender	16/08/13	24,13	MACD	35	6,63%	0,18%
Comprar	04/09/13	24,03	MACD			
Vender	10/09/13	25,66	BB	6	6,78%	1,10%
Comprar	11/09/13	25,54	MACD			
Vender	18/09/13	26,86	BB	7	5,17%	0,72%
Comprar	20/01/14	21,60	BB			
Vender	06/02/14	26,32	BB	17	21,85%	1,17%
Comprar	14/03/14	24,21	BB			
Vender	25/03/14	26,82	BB	11	10,78%	0,94%
Comprar	22/04/14	26,55	MACD			
Vender	12/05/14	31,04	MACD	20	16,91%	0,78%
Comprar	30/05/14	26,56	BB			
Vender	25/06/14	29,38	MACD	26	10,62%	0,39%
Comprar	11/07/14	30,14	MACD			
Vender	31/07/14	31,22	BB	20	3,58%	0,18%
Comprar	20/08/14	34,26	MACD			
Vender	26/08/14	35,15	BB	6	2,60%	0,43%
Comprar	28/08/14	35,87	MACD			
Vender	08/09/14	35,36	MACD	11	-1,42%	-0,13%
Comprar	29/09/14	31,87	BB			
Vender	03/10/14	31,69	MACD	4	-0,56%	-0,14%
11 ciclos				67	82,94%	0,91%

(b)

FB	Data	R$	Técnica	Dias	%total	%diário
Comprar	07/02/14	64,32	MACD			
Vender	28/02/14	68,46	MACD	21	6,44%	0,30%
Comprar	24/03/14	64,10	BB			
Vender	27/05/14	63,48	BB	64	-0,97%	-0,02%
Comprar	08/07/15	62,76	BB			
Vender	24/07/15	74,98	BB	16	19,47%	1,12%
Comprar	10/10/15	72,91	BB			
Vender	30/10/15	74,11	MACD	20	1,65%	0,08%
Comprar	25/11/14	75,63	MACD			
Vender	26/11/14	77,62	BB	1	2,63%	2,63%
Comprar	01/12/14	75,10	MACD			
Vender	22/12/14	81,45	BB	21	8,46%	0,39%
Comprar	15/01/15	74,05	BB			
Vender	20/02/15	79,90	BB	36	7,90%	0,21%
Comprar	30/04/15	78,77	BB			
Vender	03/06/15	82,44	BB	34	4,66%	0,13%
Comprar	16/06/15	81,06	MACD			
Vender	22/06/15	84,74	BB	6	4,54%	0,74%
Comprar	13/07/15	90,10	MACD			
Vender	17/07/15	94,97	BB	4	5,41%	1,32%
10 ciclos				80	60,18%	0,59%

(c)

6.3.4 – Contribuição do Modelo Proposto

Antes de destacar as contribuições do modelo proposto, torna-se importante apresentar e comentar as possíveis razões que levaram o modelo tradicional a apresentar um desempenho superior ao do modelo proposto na maioria das aplicações realizadas. Dessa forma, os seguintes fatos devem ser levados em consideração:

- o modelo tradicional, que é capaz de identificar ciclos muito curtos (por vezes de um único dia), só foi aplicado após a identificação do início e do término dos ciclos longos pelo modelo proposto; é natural que, identificando pequenos períodos de rápida elevação de preço dentro de um período maior com tendência de crescimento e, além disso, desprezando os pequenos períodos de queda ao longo do ciclo maior, o modelo tradicional apresente um resultado melhor do que o obtido pelo modelo proposto;
- conforme mencionado anteriormente, os profissionais do mercado sugerem que mais de um método seja utilizado para avaliar a evolução do preço de um ativo e, consequentemente, nada impede que o modelo proposto seja aplicado em conjunto com algum método tradicional de análise técnica voltado para a identificação de ciclos curtos de investimento, buscando assim uma sinergia entre os métodos;
- a utilização conjunta de um método capaz de identificar ciclos longos com outro voltado para a identificação de ciclos curtos eliminaria as indicações conflitantes que costumam surgir ao se utilizar dois métodos igualmente voltados para a busca de ciclos de investimento curtos, como no caso do uso conjunto das Bandas de Bollinger com o *MACD – Moving Average Convergence Divergence* ou em muitas outras combinações de métodos tradicionais;
- os métodos de análise tradicional funcionam como um protocolo de comunicação entre os agentes que atuam no mercado; muitos investidores, ou até mesmo empresas, baseiam seu comportamento nas informações fornecidas por esses métodos, mesmo quando eles apontam sinais pouco relevantes (ex.: valor do ativo maior que a banda superior de Bollinger), por saber que grande parte do mercado age da mesma forma;
- o DBSCAN pode vir a ser utilizado em ciclos curtos, com indicações de compra ou venda em períodos não inferiores a quatro dias, ainda que a STGT seja mantida como um método de identificação de ciclos longos, compreendendo a sinalização de momentos de compra ou venda separados por dezenas de dias;

- tanto o DBSCAN quanto a STGT podem vir a ser utilizados em ciclos curtos, com indicações de compra ou venda a cada hora, ou mesmo a cada intervalo de 15 minutos, sendo necessário para isso avaliar o funcionamento do modelo proposto com dados produzidos nas unidades temporais apropriadas.

Finalmente, considerando os resultados obtidos em todas as aplicações realizadas, é possível destacar as seguintes contribuições do modelo proposto:

- o atendimento às expectativas de investidores pessoa física e de administradoras de fundos de investimento, que buscam uma boa rentabilidade em aplicações financeiras de longo prazo, através da identificação de ciclos de investimento longos, envolvendo dezenas de dias entre o momento da compra e o da venda;
- a obtenção dos melhores níveis de desempenho em termos da taxa de retorno total obtida no período completo do investimento para aqueles ativos que apresentam pouca oscilação de preço durante o ciclo de investimento, como pode ser observado na aplicação de ambos os modelos ao ativo BRFS3;
- a orientação aos investidores quanto ao melhor período para a aplicação de outros métodos de análise técnica, sejam eles tradicionais ou inovadores, capazes de se valer de períodos curtos de valorização dos ativos e de desprezar os de desvalorização, de forma a maximizar simultaneamente a taxa de retorno total do investimento e a sua taxa equivalente ao dia, como uma consequência da sinergia entre os diferentes métodos;
- o alerta da formação de fenômenos chamados de “bolhas especulativas”, como pode ser observado na aplicação dos modelo proposto ao ativo PETR4, quando um ativo passa a apresentar uma valorização bem acima do consenso entre os analistas sobre a expectativa do seu valor máximo de mercado; esse fenômeno é traduzido por um aumento na intensidade das frequências baixas na STGT e a formação de vários pontos de ruído ou de vários pequenos clusters no DBSCAN quando o ativo apresenta forte alta de preço.

7 – CONCLUSÃO

As principais conclusões extraídas do estudo foram as seguintes:

- O surgimento, aumento de tamanho e intensificação dos aspectos nas baixas frequências na STGT apresentam grande importância na interpretação da evolução do IBOVESPA e do ativo PETR4 no DBSCAN e, consequentemente, nas decisões sobre compra e venda de posições, mas não estão necessariamente associadas a uma inversão na tendência na evolução dos índices ou dos ativos. Quando analisados em conjunto, sugerem a existência de importantes ciclos de investimentos, representados no modelo pelo resultado agregado das ações independentes de diversos agentes.
- Os resultados da aplicação das duas técnicas adotadas, o DBSCAN e a STGT, demandam um acompanhamento do processo em constante sincronismo com o crescimento da série, já que a entrada de poucos novos elementos podem alterar o resultado de parte significativa do histórico apresentado no modelo. Não foi possível identificar o potencial para a aplicação de outras técnicas, mas alguns testes demostraram que pode vir a ser bastante útil desenvolver novos indicadores para acompanhamento do processo, assim como passar a acompanhar todos os ativos relevantes na composição dos índices ou das carteiras analisadas.
- Se a existência desses ciclos for aceita como fato, é possível estimar um retorno para esses conjuntos ou agregados de operações, avaliando dessa forma o desempenho das aplicações do modelo. A observação continuada do processo é fundamental para confirmar a ocorrência de tais ciclos e a capacidade do modelo em identificá-los, definindo as condições que permitem a sua aplicação e os fatores que influenciam nos seus resultados.

Dentre as oportunidades para estudos futuros é possível destacar:

- A avaliação da aplicabilidade do modelo proposto para analisar longos ciclos econômicos, sejam eles no mercado de capitais ou em outras atividades de serviços financeiros. As características do sinal desenvolvido para ser analisado através da STGT, conforme apresentado no protótipo construído para a série do IBOVESPA (Figura 4.4), sugerem que os dados obtidos com medição diária podem vir a ser agregados em unidades temporais mais longas (semanas, meses, trimestres, etc.), de forma a produzir resultados tão úteis para a análise dos fenômenos macroeconômicos de longo prazo quanto aqueles apresentados e discutidos neste trabalho para os ciclos de investimento envolvendo dezenas de dias.
- De forma análoga, a avaliação da aplicabilidade do modelo proposto para analisar ciclos curtos de investimento, especificamente no mercado de ações. Como mencionado por Gomber et al. [2011], uma nova modalidade de negociação denominada *high-frequency trading* vem recebendo cada vez mais atenção do público envolvido com estratégias voltadas para grandes volumes de negócios, principalmente em mercados nos quais tecnologias sofisticadas são adotadas em todas as etapas da cadeia de valor do processo. A característica de autossimilaridade das curvas de evolução dos ativos, apresentadas no DBSCAN, sugere a existência de potencial para adoção do modelo proposto em intervalos de tempo bem mais curtos, como hora ou mesmo alguns poucos minutos.

REFERÊNCIAS:

Aghabozorgi, S. et al., *Time-series clustering – A decade review.* Information Systems, http://dx.doi.org/10.1016/j.is.2015.04.007 (2015).

Aguiar-Conraria, L.; Azevedo, N.; Soares, M.J., *Using wavelets to decompose the time-frequency effects of monetary policy*. Physica A, Elsevier (2008).

Aguiar-Conraria, L.; Magalhães C. P., e Soares, M. J., *Cycles in Politics: Wavelet Analysis of Political Time-Series*. The American Journal of Political Science (2011).

B. Boashash (ed.), *Time Frequency Signal Analysis and Processing – A Comprehensive Reference*. Elsevier, Amsterdam (2003).

Cortês, S. C.; Porcaro; R. M., Lifschitz, S., *Mineração de Dados – Funcionalidades, Técnicas e Abordagens*. PUC-RioInf.MCC10/02 (2002).

Edwards, R. D.; Magee, J.; and Bassetti, W. H. C., *Technical Analysis of Stock Trends*. 10th Ed., CRC Press, Boca Raton, FL (2012).

Esling, P.; Agon, C., *Time-series data mining*. ACM Computing Surveys, 45.1:12 (2012).

Ester, M.; Kriegel, H. P.; J. Sander, J., *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), vol. 1996, no. 6, pp. 226–231 (1996).

Fu, T., *A review on time series data mining*. Engineering Applications of Artificial Intelligence 24.1:164-181 (2011).

George, N. V, *S Transform: Time Frequency Analysis & Filtering*. Department of Electronics and Communication Engineering , National Institute of Technology , Rourkela, India (2009).

Han, J.; Kamber, M., *Data Mining: Concepts and Techniques*. 2nd Ed., Elsevier (2003).

Gomber, P., Arndt, B., Lutat, M., & Uhle, T., *High-frequency trading*. Available at SSRN 1858626 (2011).

Higham, D. J., *An Introduction to Financial Option Valuation – Mathematics, Stochastics and Computation*. Cambridge University Press (2004).

Lee, K. H. and Jo, G. S., *Expert system for predicting stock market timing using a candlestick chart*. Expert Syst. Appl. 16, 4, 357-364 (1999).

Liao, T. W., *Clustering of time series data – a survey.* Pattern Recognition Society, 38 (11), 1857–1874 (2005).

Kelly, A., *Decision Making Using Game Theory*. Cambridge University Press (2003).

Keogh, E.; Lin, J., *Clustering of time-series subsequences is meaningless: implications for previous and future research*. Knowl. Inf. Syst. 8 (2) 154–177 (2005).

Keogh, E.; Pazzani, M., *An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback*. Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining, pp. 239–241 (1998).

Košmelj, K.; Batagelj, V., *Cross-sectional approach for clustering time varying data*. Journal of Classification 7, 99–109 (1990).

Murphy, J. J., *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Apps*. 1st Ed., New York Institute of Finance, New York, NY (1999).

Meyer, P., *Probabilidade – Aplicações à Estatística*. 2a edição, São Paulo: LTC (2000).

Panait, L.; Luke, S. Cooperative Multi-agent Learning: The State of the Art. *Autonomous Agents and Multi-Agent Systems Journal*, Volume 11, Issue 13, Pages 387-434 (2005).

Plastino, A., *Mineração de Dados – Notas da Aula 02 – Introdução* (2013).

Rodgers, J. L.; Nicewander, W. A., *Thirteen ways to look at the correlation coefficient*. The American Statistician 42 (1) 59–66 (1988).

Russel,S.;Norvig,P., *Artificial Intelligence: a modern approach*. 2nd Ed, Prentice Hall (2003).

Segaran, T., *Programming Collective Intelligence*. O'Reilly (2007).

Shiller, R. J., *Irrational Exuberance*. 2nd Ed., Princeton University Press (2009).

Torreão, J. R. A.; Victer, S. M. C.; Fernandes, J. L., *A Signal-Tuned Gabor Transform with Aplication to EEG Analysis*. International Journal of Modern Physics C, Vol. 24, No. 4, 1350017 (24 pages), World Scientific Publishing Company (2013).

Tsai, C.-F. and Quan, Z.-Y., *Stock prediction by searching for similarities in candlestick charts*. ACM Trans. Manage. Inf. Syst. 5, 2, Article 9 (2014).

Turhan-Sayan, G.; Sayan, S., *Use of Time-Frequency Representations in the Analysis of Stock Market Data*. Middle East Technical University, Ankara, Turkey (2001).

William, D.; Spangler, J., *Physics for Science and Engineering*. Nostrand, New York (1981).

Zarb, F. G. and Kerekes, G. T., *The Stock Market Handbook: Reference Manual for the Securities Industry*. 1st Ed., Dow Jones-Irwin, Homewood, IL (1970).