Rodrigo Arruda Torres

Aplicação de Métodos de Clusterização em um estudo sobre o Mercado Acionário Brasileiro

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Matemática do Departamento de Matemática da PUC-Rio.

Orientador: Prof. Hélio Côrtes Vieira Lopes

Rio de Janeiro
Março de 2013

Rodrigo Arruda Torres

Aplicação de Métodos de Clusterização em um estudo sobre o Mercado Acionário Brasileiro

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Matemática do Departamento de Matemática do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Hélio Côrtes Vieira Lopes
Orientador
Departamento de Informática – PUC-Rio

Prof. Cristiano Augusto Coelho Fernandes
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Eduardo Sany Laber
Departamento de Informática – PUC-Rio

Prof. Sinésio Pesco
Departamento de Matemática – PUC-Rio

Prof. José Eugênio Leal
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 22 de março de 2013

Rodrigo Arruda Torres

Graduou-se em Engenharia de Produção na Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) em dezembro de 2010 e cursou pós-graduação em Matemática, também na PUC-Rio (2012). É analista de empresas da GAP Asset Management, responsável pela cobertura de empresas de diversos setores da economia brasileira e internacional.

Ficha Catalográfica

Torres, Rodrigo Arruda

Aplicação de métodos de clusterização em um estudo sobre o mercado acionário brasileiro / Rodrigo Arruda Torres ; orientador: Hélio Côrtes Vieira Lopes. – 2013.
84 f. : il. (color.) ; 30 cm

Dissertação (mestrado)- Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Matemática, 2013.
Inclui bibliografia

1. Matemática – Teses. 2. Clusterização. 3. Clusters. 4. Métodos de validação de clusterização. 5. Finanças. I. Lopes, Hélio Côrtes Vieira. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Matemática. III. Título.

CDD: 510

Para meus pais, Luis Eduardo e Leila,
Pelo apoio e confiança.

Agradecimentos

Ao meu orientador Professor Hélio Lopes, pelo estímulo e parceria para a realização deste trabalho.

Ao CNPq e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos meus pais, pela educação, atenção e ajuda em todos os momentos.

A todos os familiares e amigos que, de alguma forma, me estimularam ou ajudaram.

Resumo

Torres, Rodrigo Arruda; Lopes, Hélio Cortês Vieira. **Aplicação de Métodos de Clusterização em um estudo sobre o Mercado Acionário Brasileiro**. Rio de Janeiro, 2013. 84p. Dissertação de Mestrado – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Evidências indicam que ações de empresas de um mesmo setor da economia apresentam retornos semelhantes ao longo do tempo, uma vez que estariam expostas a variáveis econômico-financeiras e técnico-operacionais semelhantes. Gestores de recursos, de maneira geral, utilizam esta evidência em suas avaliações diárias na busca pelos melhores investimentos. Entretanto, na grande maioria dos casos, não há um embasamento teórico e matemático que comprove essa relação entre as ações. O objetivo dessa dissertação é verificar se, para um grupo de ações classificadas como mais relevantes dentre as presentes na Bolsa de Valores brasileira, os preços diários de fechamento que se comportam analogamente correspondem a empresas de um mesmo setor econômico. Para testar tal hipótese, serão avaliados diferentes métodos de clusterização aplicados a matriz de dissimilaridade entre os dados estudados, que por sua vez será determinada a partir de diferentes técnicas não-paramétricas de cálculo de dependência entre dados. Os métodos testados serão comparados e o melhor escolhido através da aplicação de índices de validação de clusterizações.

Palavras-chave

Clusterização; *clusters*; métodos de validação de clusterização; finanças

Abstract

Torres, Rodrigo Arruda; Lopes, Hélio Cortês Vieira (Advisor). **Application of Clustering Methods in a study about the Brazilian Stock Market**. Rio de Janeiro, 2013. 84p. MSc. Dissertation – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Evidence indicates that shares of companies belonging to the same economic sector have similar returns over time, since they would be exposed to similar economic-financial and technical-operational variables. Portfolio managers, in general, use this evidence in their daily valuations in order to find the best investment alternatives. However, in most cases, there isn`t a theoretical and mathematical background proving this relationship between stocks exists. The objective of this dissertation is to determine whether, for a group of stocks classified as among the most important of the Brazilian stock market, the daily closing prices that behave similarly correspond to companies in the same economic sector. To test this hypothesis, various clustering methods were evaluated and applied to the dissimilarity matrix calculated for the analyzed data, which is determined using different non-parametric techniques for calculating the dependency between data. The models were compared and the best selected by applying clustering validation index.

Keywords

Clustering; *cluster*; cluster validation index; finance

Sumário

Lista de figuras

Lista de Tabelas

1.
Introdução

1.1.
Contexto da Pesquisa

Economia e Finanças são, em sua natureza, ciências quantitativas, cujas variáveis e suas respectivas quantificações caracterizam-se por elevado grau de incerteza. A mensuração de tais riscos é um primeiro exemplo da necessidade de utilização de métodos quantitativos para o estudo dessas ciências e ajuda a mostrar a diferença entre ambas e a Contabilidade pura e simples. A existência de variáveis e eventos econômicos e financeiros com alto grau de imprevisibilidade é um desafio ao uso de métodos quantitativos, porém não deve ser entendida como um empecilho. Tomemos como exemplo a Física, ciência na qual os modelos matemáticos são usados de maneira frequente. Muitos eventos físicos são caracterizados por alta imprevisibilidade, entretanto a análise quantitativa segue como alicerce importante em seus estudos.

Historicamente, a utilização de métodos quantitativos como ferramenta dentro do mercado financeiro esteve focada na avaliação de riscos de carteiras, especialmente aquelas mais complexas, que envolvem diversas classes de ativos. No entanto, a grande quantidade de informação disponível, a importância de análisá-las rapidamente e a necessidade de gerar retornos diferenciados levaram gestores a considerarem tais técnicas matemáticas também para a tomada de decisão e na seleção de estratégias de investimento. A especialização e a profissionalização dos gestores, a globalização dos mercados financeiros internacionais, bem como o desenvolvimento dos mesmos são outros fatores que explicam a necessidade de ferramentas que auxiliem os tomadores de decisão. Se, antigamente, especialmente em mercados menos maduros como o mercado acionário brasileiro, poucos eram os participantes relevantes e havia grande assimetria de informação, hoje empresas especializadas de todo o mundo acompanham têm acesso às informações em tempo real. As chances de arbitragem diminuíram dramaticamente, assim como o diferencial dos gestores locais que se

utilizavam apenas do julgamento próprio de poucas informações para suas decisões de investimento. Em outras palavras, a competição aumentou substancialmente e é, dentro deste contexto, que a análise quantitativa surge para auxiliar os gestores e, acima de tudo, diferenciá-los em relação a seus concorrentes.

Antes de prosseguir, é indispensável diferenciar o processo quantitativo de gestão daquele tido como tradicional. Enquanto este último é realizado por um gestor com base no seu julgamento em relação às informações a que tem acesso, o primeiro trata daqueles processos em que decisões são tomadas com base em saídas de modelos matemáticos que seguem regras fixas e pré-determinadas. Dentro da lista dos processos quantitativos, existem aqueles completamente automatizados, quando todos os procedimentos são computacionalmente determinados, desde a entrada dos dados, até a execução das ordens de compra e venda, passando pela geração de previsões e otimização de carteiras. Existem, claro, modelos híbridos, em que os métodos quantitativos são usados como mais uma fonte de informação no processo de tomada de decisão do gestor.

Dentre tais "modelos auxiliares", destacam-se aqueles que buscam confirmar ou quantificar uma percepção tida como senso comum dentro do mercado financeiro. Especialmente os participantes mais experientes entendem a forma como funcionam os diferentes mercados. Eles sabem como que alterações no preço de um certo ativo influenciam na precificação de um outro, ou como diferentes cenários econômicos interferem na valorização das diferentes classes de ativos. Além disso, entendem como os outros participantes agem diante dessas alterações de preços e cenários. Entretanto, usualmente, estes são aspectos subjetivos da avaliação que fazem, sem um embasamento matemático que os permita tomar decisões mais rápidas e seguras. Diversos exemplos podem ser dados para caracterizar este tipo de percepção dos participantes do mercado. Um gestor sabe que alterações no preço de uma certa ação tendem a influenciar significativamente (e, na maioria dos casos com correlação positiva) as ações de empresas que atuem no mesmo segmento econômico. Porém, esta é uma avaliação empírica, que deve ser testada e evidenciada estatisticamente.

1.2. Objetivo da Dissertação

O presente trabalho irá tratar da seguinte questão: é possível comprovar matematicamente que, para um certo grupo de ações listadas na Bolsa de Valores brasileira (Bovespa), os preços de fechamento que movem-se juntos são de ações que pertencem aos mesmos setores da economia? Para buscar tal resposta, foi considerada a técnica de análise de *clusters*. Como será descrito de forma mais precisa a seguir, os métodos de clusterização permitem encontrar, dentro de bases de dados, grupos com elementos que se assemelham, tomando como alicerce uma certa característica pré-determinada. Logo, a sua utilização na busca pela resposta à questão proposta baseia-se na idéia de que os preços diários de fechamento das cotações das ações diferenciarão os objetos da base de dados de tal maneira que serão (ou não, eventualmente) formados grupos cujos elementos serão ações de empresas dos mesmos segmentos econômicos.

1.3. Organização da Dissertação

O segundo capítulo desta dissertação irá descrever de forma detalhada as características gerais dos métodos de clusterização. Serão apresentadas as principais formas de classificar os mesmos, os tipos de *clusters* que podem ser formados, bem como as principais formas de clusterização. A definição matemática também será apresentada.

Os métodos de clusterização baseiam-se na premissa de que é possível encontrar semelhanças entre os dados e, desta forma, formar grupos de dados homogêneos dentro da base de dados original. Existem diversas maneiras de medir tal semelhança e, uma delas, é calcular a dissimilaridade entre os dados. O terceiro capítulo desta dissertação irá apresentar os principais conceitos associados a tal medição. Além disso, serão apresentadas neste capítulo três formas não-paramétricas de cálculo de dissimilaridades, bem como será descrito o coeficiente de Pearson, estatística comum na literatura para o mesmo cálculo.

O quarto capítulo, por sua vez, será dedicado caracterização de três métodos de clusterização que serão avaliados. São eles o método PAM, o AGNES e o DIANA.

No quinto capítulo, serão apresentadas as principais formas encontradas na literatura para a seleção apropriada do número *k* de *clusters* ou para a comparação de modelos na busca daquele que apresenta melhores resultados. Em outras palavras, serão descritas as chamadas estatísticas de validação de clusterizações.

O sexto capítulo consiste na apresentação dos resultados obtidos após a aplicação dos diferentes modelos testados. Foram, no total, analisados 9 modelos, obtidos através da combinação dos 3 métodos de clusterização com as 3 formas não-paramétricas de cálculo de correlação entre dados. Basicamente, serão avaliados os valores obtidos para diversos índices de validação de clusterizações aplicados a cada um dos modelos, de tal forma que será possível excluir aqueles inadequados para o estudo da base de dados escolhida e limitar a etapa seguinte de seleção, a qual consistirá uma avaliação precisa dos *clusters* efetivamente formados. Essa seção, vale destacar, contará ainda com uma breve apresentação do processo para seleção da base de dados, bem como será exposta a classificação setorial que a própria Bolsa de Valores brasileira apresenta para tais empresas, algo importante para se ter um resultado dito "teoricamente esperado".

O sétimo e último capítulo reúne as principais conclusões e apresenta sugestões para possíveis pesquisar futuras.

2.
Clusterização

Existem diversas formas de se conceituar um processo de clusterização. A mais simples consiste no agrupamento de dados similares em *clusters*. Em outras palavras, é uma análise através da qual formam-se grupos cujos dados contidos em cada deles um possuem maior semelhança com aqueles pertencentes ao mesmo grupo do que com qualquer outro. Essa técnica, atualmente aplicada a diversos ramos do conhecimento, é usada com dois propósitos básicos. O primeiro associado ao fato de que a busca por classes ou conjuntos conceitualmente significativos de dados, que compartilhem características similares, assume um importante papel na forma como as pessoas analisam e descrevem o mundo. O segundo diz respeito ao fato de que a clusterização permite uma abstração em relação aos dados individuais na busca por informações que somente são encontradas quando analisa-se toda a base.

A clusterização agrupa os dados com base apenas nas informações contidas neles próprios e que permitam descrevê-los, bem como permitam determinar a relação entre eles. É, portanto, chamada de "classificação não-supervisionada", em oposição aos métodos de "classificação supervisionada", que buscam classificar novos objetos com base também em informações contidas em dados anteriormente estudados. Ela está inserida em um cenário mais abrangente da análise de dados, associado ao processo de *machine learning*. Diante da necessidade de analisar e tratar conjuntos cada vez maiores e mais complexos, o conceito de *machine learning* surgiu com as bases técnicas para a extração de informações de dados brutos. Em suma, este processo envolve três etapas principais: (i) transformação dos dados em um formato adequado; (ii) limpeza dos mesmos para evitar presença ou *outliers*; (iii) inferência de conclusões a partir deste novo conjunto obtido.

Formalmente, o processo de clusterização pode ser definido da seguinte forma: considere $X \in R^{mxn}$, sendo X um conjunto de dados representados por m pontos x_i em R^n. A clusterização tem como objetivo dividir X em k grupos C_k de

tal forma que todos os dados pertencentes a um certo grupo C_j sejam mais "semelhantes" entre si do que com qualquer dado pertencente a um grupo C_q, tal que $q \neq j$.

Todos os processos de clusterização, por mais que apresentem suas particularidades, são estruturados de maneira semelhante. A primeira etapa consiste na seleção de uma certa característica dos dados que servirá de alicerce para o processo. É a característica segundo a qual os objetos serão divididos e que irá diferenciar os *clusters* formados. A segunda etapa consiste na escolha de um algoritmo apropriado dentre os vários possíveis, escolha esta que deverá considerar o tipo de base de dados analisada e a característica anteriormente escolhida. Para tal, é preciso identificar, em cada possível método, qual é a forma usada para a medição da proximidade entre os dados e o critério usado para a clusterização. Estas duas características permitem definir a adequação entre um certo método e a base de dados.

A terceira e a quarta etapas ocorrem após a implementação efetiva do algoritmo. A terceira consiste na validação dos resultados, isto é, a correção dos mesmos é verificada através de critérios e técnicas específicas. Ela é importante porque a clusterização define grupos que não são conhecidos *a priori*, de tal forma que a partição obtida deve ser avaliada para que a precisão do processo seja verificada. Por último, num quarto momento, é realizada a interpretação dos resultados, o que pode muitas vezes envolver a interface com outras experiências e técnicas matemáticas. O processo está esquematizado na figura 1 abaixo.

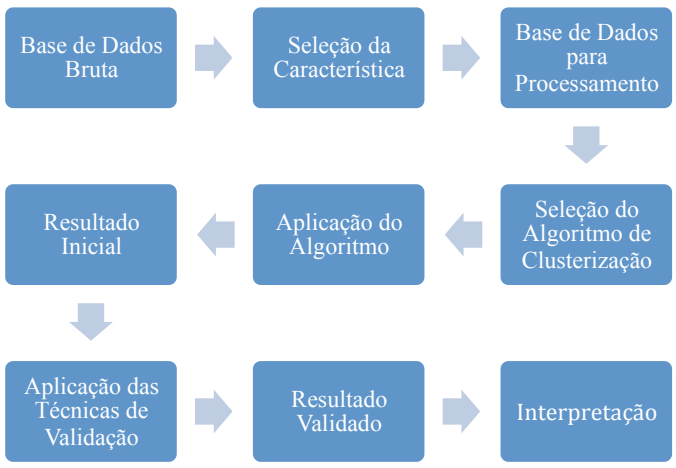

Figura 1 – Esquema das etapas de um processo de clusterização genérico.

São inúmeras as aplicações para os processo de clusterização. Por exemplo, eles podem ser usados para a compressão de informações quando a quantidade de dados disponível é muito grande. Assim, em vez de analisar a base de dados como um todo, pode-se estudar individualmente grupos formados a partir dela via clusterização. A geração e o teste de hipóteses em relação aos dados também são possíveis aplicações: o estudo dos dados agrupados permite inferir hipóteses de maneira simplificada. Destaca-se ainda a possibilidade de realizar previsões: padrões desconhecidos podem ser comparados com padrões sabidos, encontrados após a utilização desses métodos.

Especificamente no ramo das finanças, a técnica de clusterização pode assumir muitos e importantes papéis. São exemplos, a identificação de estruturas econômicas (ou padrões em processos econômicos), que permita uma melhora na capacidade preditiva; o estabelecimento de dependência entre variáveis econômicas, que permita uma melhor gestão de risco e escolha de estratégias de investimento, dentre outros.

Em suma, fica claro que a clusterização assume grande relevância na forma como dados são analisados atualmente. A literatura é rica em diferentes modelos, o que torna a classificação dos mesmos algo indispensável para a correta utilização. Portanto, na sequência deste capítulo, serão descritas as principais maneiras encontradas para a classificação dos processos de clusterização. Além disso, serão apresentados os tipos de *clusters* que podem ser formados, bem como as principais formas de clusterização.

2.1. Métodos de Clusterização

Existem, na literatura, diversos métodos de clusterização, os quais podem ser separados em diferentes grupos e classificados de acordo com uma série de possíveis características. Nesta seção, serão apresentadas os mais importantes.

Uma primeira forma de diferenciar é verificar se uma clusterização é particional ou hierárquica. Ela é particional quando os *clusters* formados não possuem interseção, ou seja, não existem os chamados *sub-clusters* e o processo de divisão dos dados em grupos ocorre somente uma vez. Em contrapartida, é dita

hierárquica quando existe esta interseção e diversas partições são obtidas ao longo do processo.

Os modelos particionais são comumente usados quando a quantidade de dados é muito grande e seria muito custoso determinar e armazenar todas as possibilidades, como seria necessários com um modelo hierárquico. De forma resumida, o problema de implementar uma clusterização particional pode ser definido da seguinte forma: dado um conjunto de n dados e a atributos, determine uma participação do conjunto inicial contendo k *clusters* de tal forma a maximizar a similaridade (ou minimizar a dissimilaridade) entre os dados pertencentes a um certo grupo.

Já os modelos hierárquicos produzem diversas partições da base de dados original e, dependendo da forma como realizam tal processo, podem ser classificados como aglomerativos ou divisivos. No primeiro caso, os *clusters* são formados a partir da união de *clusters* menores até que seja formado um único agrupamento contendo todos os dados. No segundo caso, a primeira partição contém a totalidade dos dados analisados e é sucessivamente dividida até que chegue-se a um total de n *clusters* formados por um único elemento cada. A Figura 2 a seguir esquematiza a diferença entre essas duas formas de métodos hierárquicos.

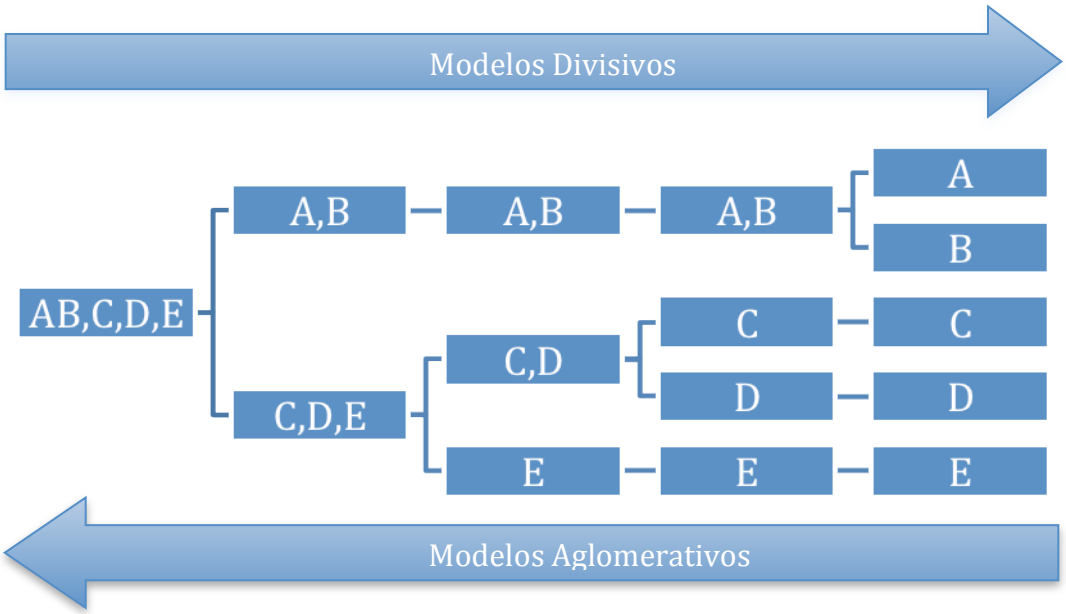

Figura 2 – Esquema sobre os modelos hierárquicos de clusterização.

Também é possível classificar os métodos de acordo com a relação de pertinência dos dados em relação aos *clusters* formados. Quando cada dado pertence somente a um grupo, diz-se que a clusterização é exclusiva. Já quando um dado pode pertencer, ao mesmo tempo, a mais de um *cluster*, diz-se que há sobreposição. Há ainda os casos em que cada dado é dito pertencente a todos os *clusters*, porém com um peso que pode variar entre 0 (quando certamente não pertence a certo grupo) e 1 (quando certamente pertence a um certo grupo). Claramente, esta é uma abordagem probabilística do problema de agrupamento, uma vez que os pesos representam a probabilidade com que certo dado pertence a um certo *cluster* (especialmente quando considera-se a premissa de que a soma de tais pesos para cada um dos dados deve ser igual a 1).

Uma clusterização pode ainda ser considerada completa ou parcial. No primeiro caso, todos os dados são alocados a pelo menos um grupo, enquanto no segundo, isso não é necessário. A motivação para a implementação de um processo parcial é a possibilidade de se considerar a existência de dados que não deveriam pertencer a nenhum *cluster* por serem ruídos ou *outliers*.

2.2. Tipos de *Clusters*

As características dos grupos formados a partir dos diferentes métodos também permitem classificá-los de diferentes formas. Os *clusters* são considerados "bem separados" quando a base de dados contém grupos naturais distantes uns dos outros, ou seja, a distância entre dois pontos quaisquer dentro de um mesmo grupo é menor do que a distância entre dois pontos quaisquer que pertençam a grupos distintos.

Uma outra classificação baseia-se na idéia de protótipos que representam os *clusters*. Neste caso, cada dado está mais próximo do protótipo que define seu grupo do que em relação ao protótipo de qualquer outro *cluster*, sendo que tal protótipo assume, normalmente, a figura do centróide.

Os *clusters* podem ainda ser definidos como um "componente de conexão" de uma estrutura de grafos. Nesse caso, eles são entendidos como um conjunto de dados conectados uns aos outros, porém sem conexão com os dados de fora. Por último, há a classificação com base no conceito de densidade: um *cluster* é uma região densa de dados cercada por uma região de densidade mais baixa.

2.3.
Principais Formas de Clusterização

2.3.1.
Modelos Paramétricos

O objetivo dos modelos paramétricos é sempre minimizar uma função-custo ou um critério de otimização que associe um custo a cada etapa do processo de atribuição de um *cluster* a um certo dado. Uma característica importante destes modelos é que usualmente assumem alguma premissa em relação a estrutura da base de dados analisada. Essa classe pode ser sub-classificada de duas formas: modelos reconstrutivos e modelos geradores.

Os modelos reconstrutivos assumem o objetivo geral dos modelos paramétricos de minimizar uma certa função-custo. O que irá diferenciá-los é a técnica usada para modelar tal função e otimizá-la.

Já os modelos geradores assumem uma postura probabilística, ao considerar que o vetor de entradas $x_{1,}x_2, \dots, x_m$ é formado por observações de um conjunto de K distribuições desconhecidas $E_1, E_2, \dots, E_K$. Suponha que a densidade de probabilidade de uma observação x_K em relação a E_i é dada pela função $f(x_K|\theta)$ para um certo conjunto desconhecido de parâmetros θ e que a probabilidade de que x_K pertença a E_i seja τ_K^i. Considerando que, normalmente, cada dado pertence a somente uma distribuição, podemos considerar válido que $\sum_{i=1}^{K} \tau_K^i = 1$. Logo, o objetivo deste tipo de modelo é encontrar os parâmetros θ e τ que maximizem a seguinte função de verossimilhança:

$$L(\theta, \tau) = \sum_{r=1}^{n} ln\left(\sum_{i=1}^{K} \tau_K^i f(x_K|\theta)\right) \quad (1)$$

O principal problema com este tipo de modelo é que, caso os parâmetros θ a serem estimados possam variar livremente para cada uma das distribuições, encontrar o ponto ótimo global será um processo demorado e que que exigirá grande capacidade computacional. Além disso, este tipo de modelo parte da premissa de que os dados apresentam alguma distribuição conhecida (usualmente, a distribuição Gaussiana), o que nem sempre é verdadeiro. Há ainda os casos em

que os dados não são numéricos, nos quais esses algoritmos não podem ser utilizados.

2.3.2. Modelos Não-paramétricos

Estes modelos baseiam-se na medição de semelhanças e diferenças entre os pontos presentes em um certo *cluster* que está sendo analisado a cada iteração, de tal maneira que podem determinar a união de dois *clusters* similares (modelos de aglomeração) ou a divisão de um grupo em dois ou mais (modelos de divisão). Uma importante e positiva característica desses modelos é que eles não exigem premissas sobre a distribuição que permeia a base de dados.

3.
Medição das Dissimilaridades

Para podermos aplicar diferentes métodos de clusterização a séries temporais financeiras e, além disso, comparar tais opções, é preciso processar os dados para que sirvam de entrada para os modelos. Tipicamente, exige-se a construção de uma certa estrutura, a qual pode apresentar duas formas principais. A primeira representa os dados como a média de p medidas ou atributos, estruturados em uma matriz $n\ x\ p$ em que as linhas correspondem aos dados e as colunas correspondem aos respectivos atributos. A segunda constrói uma matriz $n\ x\ p$ de índices de "proximidade" entre os dados.

Estes índices necessários a segunda estrutura podem, por sua vez, ser de dois tipos: de dissimilaridade, quando medem o quão distantes os dados estão uns dos outros, ou de similaridade, quando medem o quão próximos os dados estão uns dos outros. O cálculo deles é baseado na medição da correlação entre variáveis. Elas são ditas associadas entre si quando valores assumidos por uma afetam a distribuição da outra. Por outro lado, elas são ditas independentes quando alterações em uma não afetam a outra. Da mesma forma, uma correlação positiva existe quando um aumento em uma variável ocasiona aumento na outra e, vice-versa, uma correlação negativa existe quando um aumento em uma variável ocasiona redução na outra.

Neste trabalho, a avaliação dos dados será feita conforme a segunda estrutura mencionada, mais precisamente através de índices de dissimilaridades. Por isso, serão apresentadas, na sequência, as principais maneiras encontradas na literatura para o cálculo da correlação entre dados, informação esta que alicerça a medição da dissimilaridade. Na prática, uma vez conhecido o valor da correlação, a dissimilaridade é obtida a partir de uma simples transformação. As formas apresentadas serão o Coeficiente de Correlação de Pearson, o Coeficiente de Kendall, o Coeficiente de Spearman e o Coeficiente de Hoeffding.

Basicamente, os índices de dissimilaridades são números $d(i,j)$ não-negativos que assumem valores perto de 0 quando i e j são "próximos" entre si e aumentam na medida em que i e j ficam "diferentes" entre si.

Antes, entretanto, é importante apresentar aquela que é considerada a forma mais usual de calcular a distância entre dois pontos, a distância Euclidiana. Para dois pontos $I = (i_1, i_2, \dots, i_n)$ e $Q = (q_1, q_2, \dots, q_n)$, ela é dada por:

$$d(I,Q) = \sqrt{\sum_{j=1}^{n} (i_j - q_j)^2} \tag{2}$$

Lembrando que para uma função bivariada ser considerada uma função-distância, ela deve obedecer as seguintes relações:

$$\begin{gathered} d(i,j) \geq 0 \\ d(i,i) = 0 \\ d(i,j) = d(j,i) \\ d(i,j) \leq d(i,h) + d(h,j). \end{gathered} \tag{3}$$

É importante destacar que a distância Euclidiana está inserida dentro de um conceito mais amplo de medição de distância entre vetores. Ela faz parte do grupo de métricas expressas por Minkowski que assumem a seguinte estrutura geral:

$$d(I,Q) = \left(\sum_{j=1}^{n} |i_j - q_j|^r \right)^{\frac{1}{r}} \tag{4}$$

sendo que a distância Euclidiana é o subcaso em que $r = 2$.

Também são comumente utilizadas a distância de Manhattan e a distância "Sup", respectivamente os casos em que $r = 1$ e $r = \infty$ e expressas pelas seguintes equações:

$$d(I,Q) = \sum_{j=1}^{n} |i_j - q_j| \tag{5}$$

$$d(I,Q) = \max_{1 \le k \le n} |i_j - q_j| \tag{6}$$

3.1. Coeficiente de Correlação de Pearson (ρ_p)

O coeficiente de correlação de Pearson é uma medida de dependência linear entre duas variáveis aleatórias e pode assumir valores no intervalo [-1,1], de tal forma que quanto mais próximo de 1 ou -1 for seu valor, maior é grau de dependência linear (positiva ou negativa, respectivamente) entre as variáveis analisadas. Quando ele assume valor igual a 1, uma equação linear descreve a relação entre as variáveis X e Y, sendo que Y cresce na medida em que X cresce. Já quando assume valor igual a -1, o inverso pode ser interpretado: Y descresce na medida em que X cresce. Por sua vez, quando assume valor igual a zero, conclui-se que não existe relação linear entre as variáveis.

Considere duas variáveis aleatórias X e Y. O coeficiente é, então, calculado através da seguinte expressão:

$$\rho_p = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \tag{7}$$

tal que

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{8}$$

Já quando pretende-se calcular o coeficiente amostral, a fórmula (7) pode ser reescrita da seguinte maneira:

$$r_p = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{9}$$

Este coeficiente apresenta algumas importantes propriedades. Primeiro, ele não faz distinção entre a ordem de análise das variáveis, isto é, o valor para a correlação entre X e Y é o mesmo para a correlação entre Y e X. Segundo, a unidade de mensuração pode ser alterada sem que haja interferência no valor do coeficiente. Terceiro, este índice é adimensional, não possui unidade física. Desta forma, não pode ser interpretado em termos proporcionais (exemplo: coeficiente que assuma valor 0,5 não representa o dobro da correlação para um caso em que assuma valor de 0,25).

Uma outra importante característica é que, para certos casos, o coeficiente de Pearson pode ser escrito em termos da distância Euclidiana mencionada anteriormente. A equação (2) pode ser reescrita da seguinte maneira:

$$d(x,y) = \sqrt{\sum_j (x_j)^2 + \sum_j (y_j)^2 - 2\sum_j x_j y_j} \tag{10}$$

Quando os dados são padronizados, é possível afirmar que

$$\sum_j (x_j)^2 = \sum_j (y_j)^2 = n \tag{11}$$

de tal forma que o único termo não-constante é $\sum_j x_j y_j$. Assim, o coeficiente amostral pode ser calculado a partir da expressão

$$r_p = \frac{\frac{\sum_j x_j y_j}{n} - \bar{x}\bar{y}}{S_x S_y} \tag{12}$$

Ocorre que dados padronizados possuem média igual a zero e desvio-padrão igual a 1. Logo, a fórmula (12) pode novamente ser reeditada em termos da distância Euclidiana:

$$r_p = \frac{\sum_j x_j y_j}{n} = 1 - \frac{d^2(x,y)}{2n} \tag{13}$$

O coeficiente de correlação Pearson é bastante utilizado como medida de dependência entre variáveis, porém existem alguns problemas neste uso. A literatura mostra que os resultados podem ser bastante imprecisos quando as variáveis aleatórias não seguem distribuição Normal, caso no qual, acima de tudo, seu valor pode ser zero mesmo que haja dependência entre as variáveis. Além disso, ele não modela a estrutura de tal dependência, algo importante quando estudamos séries temporais financeiras, já que tal informação ajudaria a entender como dois mercados (ou, eventualmente, dois ativos) se relacionam. É também importante ressaltar que faz-se necessária uma análise dos dados a serem avaliados à procura de *outliers*, uma vez que o coeficiente de correlação de Pearson é bastante afetado por tal presença. Deve-se considerar ainda a necessidade de que as variáveis aleatórias analisadas sejam independentes.

3.2. Coeficiente de Kendall (ρ_τ)

Esta é uma primeira opção não-paramétrica de cálculo de dissimilaridade, alternativa ao Coeficiente de Pearson. Seja (X_1, Y_1) um vetor aleatório bivariado e seja (X_2, Y_2) uma cópia independente deste vetor. O coeficiente populacional de correlação Kendall pode ser definido como:

$$\rho_\tau = 2\mathcal{P}\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1 \quad (14)$$

Este evento $\{(Y_2 - Y_1)(X_2 - X_1) > 0\}$ ocorre se, e somente se, um dos seguintes eventos for verdadeiro: $\{X_2 > X_1 \; e \; Y_2 > Y_1\}$ ou $\{X_2 < X_1 \; e \; Y_2 < Y_1\}$. Como eles são mutualmente excludentes, temos que:

$$\mathcal{P}\{(Y_2 - Y_1)(X_2 - X_1) > 0\} = \mathcal{P}\{X_2 > X_1, Y_2 > Y_1\} + \mathcal{P}\{X_2 < X_1, Y_2 < Y_1\} \quad (15)$$

Desta maneira, o Coeficiente Kendall pode ser reescrito como:

$$\rho_\tau = 2\big\{\mathcal{P}\{X_2 > X_1, Y_2 > Y_1\} + \mathcal{P}\{X_2 < X_1, Y_2 < Y_1\}\big\} - 1 \quad (16)$$

Quando X e Y são independentes e tanto (X_1, X_2) quanto (Y_1, Y_2) são independentes e identicamente distribuídas (i.i.d.), valem as conclusões:

$$\begin{aligned} \mathcal{P}\{X_2 > X_1, Y_2 > Y_1\} &= \mathcal{P}\{X_2 > X_1\}\mathcal{P}\{Y_2 > Y_1\} = (1/2)(1/2) = 1/4 \\ \mathcal{P}\{X_2 < X_1, Y_2 < Y_1\} &= \mathcal{P}\{X_2 < X_1\}\mathcal{P}\{Y_2 < Y_1\} = (1/2)(1/2) = 1/4 \end{aligned} \tag{17}$$

as quais indicam que neste caso, $\rho_\tau = 2(1/4 + 1/4) - 1 = 0$.

Hollander e Wolfe [2] determinam o coeficiente Kendall amostral r_K da seguinte forma:

$$r_K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q\left((x_i, y_i), (x_j, y_j)\right) \tag{18}$$

para $1 \leq i < j \leq n$ e tal que

$$Q\big((a,b),(c,d)\big) = \begin{cases} 1, se\ (d-b)(c-a) > 0 \\ -1, se\ (d-b)(c-a) < 0 \end{cases} \tag{19}$$

3.3. Coeficiente de Spearman (ρ_s)

Para definir o coeficiente de Spearman, Kruskal [3], em sua demonstração, considera três observações hipotéticas da distribuição de interesse (X_1, Y_1), (X_2, Y_2) e (X_3, Y_3). Em seguida, ele explicita a probabilidade de concordância entre (X_1, Y_2) e a "observação cruzada" (X_2, Y_3), assim como a probabilidade de discordância entre (X_1, Y_2) e (X_2, Y_3), dadas pelas seguintes expressões:

$$\begin{aligned} i_c &= \mathcal{P}\{(X_1 - X_2)(Y_1 - Y_3) > 0\} \\ i_d &= \mathcal{P}\{(X_1 - X_2)(Y_1 - Y_3) < 0\} \end{aligned} \tag{20}$$

Na sequência, ele calcula a diferença entre i_c e i_d, que é o Coeficiente de Spearman, obtendo:

$$\rho_s = i_c - i_d = 2i_c - 1 \tag{21}$$

Como as variáveis consideradas são ordinalmente invariantes, é possível performar transformações nos componentes das variáveis aleatórias sem afetar os resultados acima obtidos. Logo, cada X_i pode ser substituídos por $X_i^* = F(X_i)$ e cada Y_i pode ser substituído por $Y_i^* = G(Y_i)$, tal que F e G são as distribuições cumulativas de probabilidades de X e Y, respectivamente. A partir desta nova relação, sabemos que X_i^* e Y_i^* possuem distribuições marginais de probabilidade uniformes entre 0 e 1. Logo, podemos escrever que:

$$\begin{aligned} i_c &= \mathcal{P}\{(X_1 - X_2)(Y_1 - Y_3) > 0\} = \mathcal{P}\{(X_1^* - X_2^*)(Y_1^* - Y_3^*) > 0\} \\ &= \mathbb{E}[\mathcal{P}\{(X_1^* - X_2^*)(Y_1^* - Y_3^*) > 0 | (X_1^*, Y_1^*)\}] \\ &= \mathbb{E}[X_1^* Y_1^* + (1 - X_1^*)(1 - Y_1^*)] \\ &= 2\mathbb{E}[X_1^* Y_1^*] \\ &= 2Cov(X_1^*, Y_1^*) + 1/2 \\ &= (2/12)\rho(X_1^*, Y_1^*) + 1/2 \end{aligned} \tag{22}$$

A partir desta fórmula encontrada, podemos concluir que i_c pode assumir valores entre $1/3$ e $2/3$ e que $2i_c - 1$ pode assumir valores entre $-1/3$ e $1/3$. Para que o coeficiente passe a variar entre -1 e 1, é comum multiplicar a fórmula (21) por 3, obtendo assim:

$$\rho_s = 3(i_c - i_d) = 6i_c - 3 = 3 - 6i_d \tag{23}$$

Para calcular o Coeficiente de Spearman amostral, basta considerar a seguinte expressão:

$$r_s = \frac{\sum_{i=1}^{n}\{[rank(X_i) - \overline{rank(X)}][rank(Y_i) - \overline{rank(Y)}]\}}{\sqrt{\sum_{i=1}^{n}[rank(X_i) - \overline{rank(X)}]^2 \sum_{i=1}^{n}[rank(Y_i) - \overline{rank(Y)}]^2}} \tag{24}$$

tal que $rank(W)$ denota o posto ou a "posição" de uma variável na sequência de dados analisados, ou seja, a observação com o menor valor possui $rank = 1$, a próxima com menor valor possui $rank = 2$ e assim sucessivamente até que a observação de maior valor assuma $rank = n$.

Vale destacar que, conforme citado anteriormente, para utilizar os coeficientes de correlação apresentados como indicadores de dissimilaridade, é preciso transformá-los de tal forma que variáveis com alta correlação positiva tenham coeficientes de dissimilaridade próximos a zero e variáveis com alta correlação negativa tenham coeficientes próximos a 1. A transformação necessária é dada pela seguinte expressão:

$$d(x,y) = \frac{1-\rho(x,y)}{2} \tag{25}$$

3.4. Coeficiente de Hoeffding ($\mathcal{D}$)

Em seu trabalho, Hoeffding [4] considerou o problema de testar a independência de duas variáveis aleatórias X e Y com base em uma amostra aleatória de tamanho n e na função $f(x,y)$, que é considerada contínua. Se $F(x,y)$ é uma função de distribuição bivariada, então podemos definir:

$$\begin{aligned} D(x,y) &= F(x,y) - f(x,\infty)F(\infty,y) \\ \Delta = \Delta(F) &= \int D^2(x,y)\,dF(x,y) \end{aligned} \tag{26}$$

Nesse mesmo estudo, Hoeffding mostra que $0 \leq \Delta \leq 1/30$. Ele define ainda três variáveis auxiliares:

$$\begin{aligned} C(u) &= \begin{cases} 1, se\ u \geq 0 \\ 0, se\ u < 0 \end{cases} \\ \psi(x_1,x_2,x_3) &= C(x_1-x_2) - C(x_1-x_3) \\ \phi(x_1,y_1;\ldots;x_5,y_5) &= \frac{1}{4\psi(x_1,x_2,x_3)\psi(x_1,x_4,x_5)\psi(y_1,y,y_3)\psi(y_1,y_4,y_5)} \end{aligned} \tag{27}$$

Além disso, mostra que Δ pode ser escrito da seguinte maneira:

$$\Delta = \int \ldots \int \phi(x_1,y_1;\ldots;x_5,y_5)\,dF(x_1,y_1)\ldots dF(x_1,y_5) \tag{28}$$

Seja $(X_1, Y_1), \dots, (X_n, Y_n)$ uma amostra aleatória de uma população com função distribuição de probabilidade dada por $F(x, y)$, com $n \geq 5$. O coeficiente de Hoeffding é definido como:

$$D = D_n = \frac{1}{n(n-1)\dots(n-4)} \sum \phi(X_{\alpha 1}, Y_{\alpha 1}; \dots; X_{\alpha 5}, Y_{\alpha 5}) \tag{29}$$

tal que $\alpha_i = 1, \dots, n\,;\ \alpha_i \neq \alpha_j\ se\ i \neq j; i, j = 1, \dots, 5$. Das equações (27) e (28) é possível obter a seguinte expressão:

$$D = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)} \tag{30}$$

tal que

$$\begin{aligned} A &= \sum_{\alpha=1}^{n} a_\alpha (a_\alpha - 1)\, b_\alpha (b_\alpha - 1) \\ B &= \sum_{\alpha=1}^{n} (a_\alpha - 1)\,(b_\alpha - 1) c_\alpha \\ C &= \sum_{\alpha=1}^{n} c_\alpha\,(c_\alpha - 1) \\ a_\alpha &= \sum_{\beta=1}^{n} C\big(X_\alpha - X_\beta\big) - 1 \\ b_\alpha &= \sum_{\beta=1}^{n} C\big(Y_\alpha - Y_\beta\big) - 1 \\ c_\alpha &= \sum_{\beta=1}^{n} C\big(X_\alpha - X_\beta\big) C\big(Y_\alpha - Y_\beta\big) - 1 \end{aligned} \tag{31}$$

Logo, para computar o coeficiente de Hoeffding para uma certa amostra, é preciso calcular os valores de $a_\alpha, b_\alpha, c_\alpha$ para cada integrante da amostra e, então, calcular os valores de A, B, C necessários para finalizar o cálculo da equação (30).

Este coeficiente já foi bastante estudado por outros pesquisadores após ter sido desenvolvido por Hoeffding. Hollander e Wolfe [2], por exemplo, mostraram

que apesar de a estatística D não ser sensível a todas as alternativas de independência entre as variáveis X e Y, existem funções F tais que X e Y são dependentes e D é consistente. Além disso, eles mostraram que D é um indicador robusto em relação a uma grande variedade de alternativas de independência entre as variáveis, ou seja, quanto maior o valor da estatística D, mais dependentes são as variáveis.

4.
Métodos de Clusterização

Existem, na literatura, diversos métodos de clusterização, os quais foram apresentados em uma série de trabalhos, com destaque para Kaufman e Rousseeuw [1]. Na sequência, serão apresentados três diferentes algoritmos, que serão posteriormente aplicados à base de dados analisada e comparados, em busca daquele que apresenta melhor resultado.

4.1.
Método PAM (*Partitioning Around Medoids*)

O algoritmo PAM, também conhecido como *K-Medoid,* é um modelo particional que pode ser aplicado tanto quando o *input* é o conjunto de atributos dos dados que compõem a base, quanto em casos em que se trabalha com a matriz de dissimilaridades. Ele busca encontrar objetos, chamados de *medoids*, que possuam uma localização central dentro de seus respectivos *clusters*. O objetivo do algoritmo, em suma, é minimizar a dissimilaridade média entre todos os objetos presentes na base de dados e o seu respectivo *medoid* mais próximo.

Este método funciona em duas fases, as quais podem ser descritas em diferentes etapas. Considere o conjunto S formado pelos objetos definidos como *medoids* e o conjunto U formado pelos objetos não definidos como tal. Considere ainda d_p como sendo a dissimilaridade entre um objeto p e o objeto mais próximo que pertença a S e e_p como sendo a dissimilaridade entre p e o segundo objeto mais próximo que pertença a S. Na primeira fase, S é inicializado adicionando a ele o objeto cuja soma das dissimilaridades em relação a todos os demais seja mínima.

Em seguida, considere o objeto $i \in U$ como sendo um candidato a entrar no conjuto S. Para todo objeto $j \in U - \{i\}$, calcule d_j. Se $d_j > d(i,j)$, então j irá contribuir para a decisão de selecionar i. Calcule, então, o total de ganho ao se acrescentar i a S através da seguintes equação:

$$g_i = \sum_{j \in U} C_{ji} = \sum_{j \in U} max\{d_j - d(i,j); 0\} \tag{32}$$

O objeto i deve ser selecionado de tal forma que maximize g_i. Esta primeira etapa deve ser repetida até que k dados sejam escolhidos para compor S.

Na segunda fase, busca-se melhorar a qualidade da clusterização ao trocar objetos entre S e U. Para tal, deve-se considerar todos os pares $(i,h) \in S \times U$ e calcular o efeito t_{ih} sobre a soma das dissimilaridades entre objetos e o seus respectivos *medoids* ao trocar i por h, ou seja, ao transferir i de S para U e h de U para S.

Considere w_{jih}, o qual representa a contribuição de cada objeto $j \in U - \{h\}$ para o cálculo de t_{ih} e pode assumir os seguintes valores:

$$\begin{aligned} &se\ d(i,j) > d_j, w_{jih} = min\{d(j,h) - d_j; 0\} \\ &se\ d(i,j) = d_j, w_{jih} = min\{d(j,h); e_j\} - d_j \end{aligned} \tag{33}$$

Assim, pode-se calcular $t_{ih} = \sum\{w_{jih} | j \in U\}$. Na sequência, deve-se selecionar um par $(i,h) \in S \times U$ que minimize t_{ih}. Se $t_{ih} < 0$, então a troca é realizada, d_p e e_p são atualizados para todos os objetos p e retorna-se ao início do algoritmo. Já se $min\{t_{ih}\} > 0$, o algoritmo atinge seu ponto de parada.

Kaufman e Rousseeuw [1] explicitaram em seus estudos que, em comparação com outro método particional importante, o *K-Means*, o PAM é mais robusto quando há presença de *outliers*. Outra característica importante é que ele busca formar *clusters* "esféricos", de tal forma que não é um bom método na busca de *clusters* em formato mais alongado.

4.2. AGNES (*Agglomerative Nesting*)

O AGNES é um modelo hierárquico de clusterização e, como tal, apresenta três importantes vantagens em relação a um modelo particional. Primeiro, não requer que seja pré-definido um número-alvo de *clusters*. Segundo, não é preciso assumir nenhuma premissa em relação a distribuição da base de dados. Terceiro,

exige apenas a existência de uma matriz de dissimilaridade entre os possíveis pares de objetos.

Este método promove a união de *clusters* dois a dois ao longo do processo, até que reste apenas um. Em suma, a primeira iteração do algoritmo une dois objetos da base de dados que estejam mais próximos entre si formando um primeiro *cluster*. Nas demais iterações, promove-se a união entre dois grupos já formados que estejam mais próximos entre si. Fica claro, nesse processo, a necessidade de se definir uma forma de calcular a dissimilaridade entre *clusters*, a qual deverá ser baseada nos conceitos, já descritos, de dissimilaridade entre objetos. Antes, entretanto, de apresentar as principais formas de realizar este cálculo – explicitadas especialmente no trabalho de Kaufman e Rousseeuw [1] – destacaremos o passo-a-passo detalhado para a implementação do AGNES.

O algoritmo tem início com um conjunto de n *clusters*, tal que n é o tamanho da base de dados. Determina-se, então a matriz $D_{n\times n} = d(i,j)$, composta pelas dissimilaridades entre os n pontos. Dentro desta matriz deve-se encontrar o menor valor $d(w,r)$ e promover a união entre os *clusters* w e r (a ser chamado wr). Na sequência, calcula-se a dissimilaridade $d(wr,q)$ entre wr e todos os demais *clusters* $q \neq wr$, cálculo este que poderá assumir uma das formas que serão apresentada a seguir. No próximo passo, forma-se uma nova matriz $D^2_{(n-1)\times(n-1)}$, na qual são excluídas as linhas e colunas referentes a w e a r e são acrescentadas uma nova linha e uma nova coluna com valores de $d(wr,q)$. Esse processo, desde a etapa em que se determina o menor valor para $d(w,r)$, deve ser repetido $(n-1)$ vezes, até que todos os objetos estejam formando um único *cluster*.

Kaufman e Rousseeuw [1] apresentaram oito regras para o cálculo de dissimilaridade entre *clusters*, dentre quais quatro têm uso mais comum. São elas:

• Ligação Média: a dissimilaridade $d(R,Q)$ entre dois *clusters* R e Q pode ser calculada como a média de todas as dissimilaridades $d(i,j)$ tal que $i \in R$ e $j \in Q$.

$$d(R,Q) = \frac{1}{|R||Q|}\sum_{\substack{i\in R\\ j\in Q}} d(i,j) \tag{34}$$

• Ligação Simples: neste caso, a dissimilaridade entre os *clusters* é definida como o valor mínimo da dissimilaridade entre todos os pares de pontos $i \in R$ e $j \in Q$. Logo, podemos escrever que:

$$d(R,Q) = \min_{\substack{i \in R \\ j \in Q}} d(i,j) \tag{35}$$

Esta regra apresenta melhores resultados quando busca-se encontrar *clusters* com formato mais alongado. Entretanto, em outros casos, os resultados encontrados não são robustos. Quando dois grupos claramente distintos aproximam-se um do outro, mesmo que através de somente um ponto, este método irá uni-los, quando na verdade eles deveriam ficar separados.

• Ligação Completa: apresenta-se como o método exatamente oposto ao segundo apresentado. A dissimilaridade entre dois *clusters* é calculada como o maior valor da dissimilaridade entre dois pontos quaisquer dos grupos analisados.

$$d(R,Q) = \max_{\substack{i \in R \\ j \in Q}} d(i,j) \tag{36}$$

Esta regra tende a formar *clusters* muito compactos, isto é, resulta num número grande de grupos com diâmetro pequeno.

• Método de Ward: nesse caso, a dissimilaridade entre dois *clusters* baseia-se na distância euclidiana entre seus centróides, multiplicada por um certo fator:

$$d^2(R,Q) = \frac{2|R||Q|}{|R| + |Q|} \|\bar{x}(R) - \bar{x}(Q)\|^2 \tag{37}$$

As figuras 3 a 6 abaixo esquematizam as quatro formas descritas para determinação da dissimilaridade entre *clusters*.

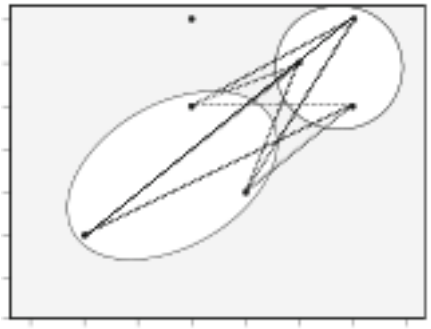

Figura 3 – Ligação Média.

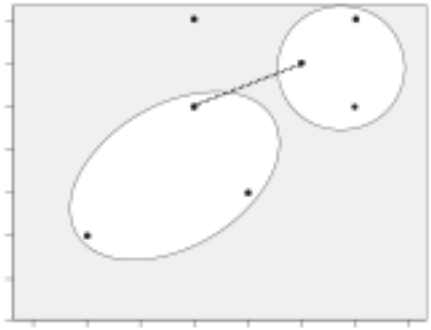

Figura 4 – Ligação Simples.

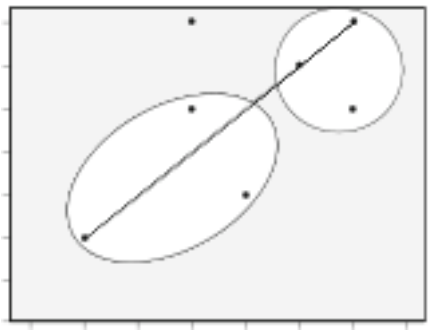

Figura 5 – Ligação Completa

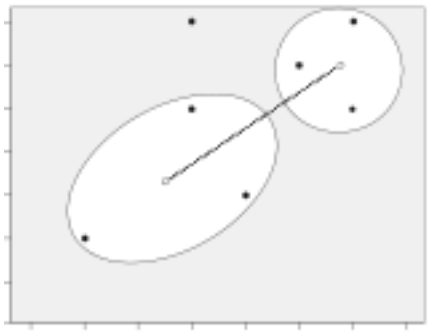

Figura 6 – Método de Ward

É importante ressaltar que Kaufman e Rousseeuw [1] demonstraram uma forma eficiente de performar este método AGNES, que utiliza apenas a matriz de

dissimilaridades resultante da união entre dois *clusters*, sem a necessidade da matriz de dissimilaridades original associada aos dados em si. A equação de Lance-Williams, como também é conhecida, é a seguinte:

$$d(R,Q) = \alpha_A d(A,Q) + \alpha_B d(B,Q) + \beta d(A,B) + \gamma|d(A,Q) - d(B,Q)| \qquad (38)$$

Para comparar estas quatro (ou até as oito) formas de cálculo da dissimilaridade entre *clusters*, Kaufman e Rousseeuw [1] descreveram três condições. A primeira delas é que a dissimilaridade entre *clusters* sendo unidos deve ser monotônica, algo necessário para que a clusterização possa ser apresentada graficamente. Isso ocorre quando

$$\alpha_A \geq 0, \alpha_B \geq 0, \gamma = 0 \; e \; (\alpha_A + \alpha_B + \beta) \geq 1 \qquad (39)$$

Essa condição é satisfeita por todos os quatro métodos acima apresentados.

A segunda condição exige que a dissimilaridade seja inequívoca, isto é, não pode haver contradições que façam com que duas escolhas determinem dissimilaridades diferentes para o mesmo par de *clusters*. Novamente, os quatro métodos mencionados satisfazem tal condição. Por último, a dissimilaridade entre *clusters* deve ser estatisticamente consistente. Logo, na medida em que a amostra aumenta, espera-se que a dissimilaridade entre grupos torne-se significativa. No caso da regra da Ligação Simples, a dissimilaridade tende a zero neste cenário, enquanto para a regra da Ligação Completa, este valor tende ao infinito quando trabalha-se com mais pontos, assim como ocorre para a regra de Ward. Conclui-se, portanto, que somente a regra da Ligação Média satisfaz as três condições propostas.

4.3. DIANA (*Divise Clustering*)

Este método trabalha de maneira oposta aos algoritmos aglomerativos, no sentido em que, a cada etapa, promove a divisão de um *cluster* em dois outros menores até que todos possuam somente um único elemento. Inicialmente, considere a existência de somente um grupo contendo todos os n objetos. A cada etapa, o *cluster* com maior diâmetro (medida esta calculada como sendo a maior

dissimilaridade entre objetos internos ao *cluster*) é dividido em dois, processo este que é repetido $(n-1)$ vezes.

De maneira mais precisa, o algoritmo DIANA pode ser descrito através das seguintes etapas. Primeiro, selecione o *cluster* C_k com o maior diâmetro. Considere, então, o cálculo das seguintes estatísticas:

$$\begin{aligned} D^k &= \{d(i,j) | i,j \in C_k\} \\ D_i^k &= \left\{d(i,j) \middle| j \in C_k \backslash \{i\}\right\} \end{aligned} \tag{40}$$

Inicie um novo grupo $C_s = \{s\}$, tal que $s = \text{argmax}_i\left(\overline{D_i^k}\right)$. Considere, então

$$\begin{aligned} D_i^{ks} &= \{d(i,j) | i \in C_k, j \in C_s\} \\ D_i^{kk} &= \{d(i,j) | i,j \in C_k, j \notin C_s\} \end{aligned} \tag{41}$$

e, ainda, calcule o valor:

$$h = \underset{i \in C_k \backslash C_s}{\text{argmax}}\left(\overline{D_i^{kk}} - \overline{D_i^{ks}}\right) \tag{42}$$

Se $\overline{D_i^{kk}} - \overline{D_i^{ks}} > 0$, então h é, na média, mais próximo do grupo recém formado C_s do que de C_k e, portanto, deve ser movido para C_s. As etapas descritas a partir das equações apresentadas em (41), inclusive, devem ser repetidas até que $\overline{D_i^{kk}} - \overline{D_i^{ks}} < 0$. Caso haja algum grupo com mais de um objeto, então todo o algoritmo, desde sua primeira etapa, deve ser refeito.

5.
Métodos de Validação de Clusterizações

Um dos principais problemas associados à implementação dos métodos de clusterização é a avaliação e a comparação dos resultados obtidos. Em métodos de classificação supervisionados, por exemplo, é possível computar um erro de classificação, uma vez que sabe-se *a priori* o tipo de resultado esperado. Da mesma forma, em problemas de regressão, pode-se comparar o valor obtido com aquele previsto em teoria, calculando-se assim o erro cometido pelo modelo. Os métodos de clusterização, entretanto, como citado anteriormente, são formas não-supervisionadas de classificação e, assim, não existem estruturas esperadas *a priori*, muito menos um resultado teórico que possa ser comparado com aquele efetivamente obtido. Na grande maioria das aplicações, não é possível sequer determinar o número correto de *clusters* que deverão ser gerados.

Diante deste cenário, torna-se indispensável a criação de uma ferramenta que permita, ao menos, a comparação entre diferentes métodos de clusterização e diferentes valores para o número *k* de *clusters* a serem formados. Este processo é chamado na literatura de "validação de clusterizações" e pode assumir três diferentes abordagens: critérios externos, internos ou relativos. Os critérios externos avaliam este problema com base numa estrutura pré-definida que reflita um conhecimento anterior sobre a base de dados ou uma eventual intuição de qual deveria ser o resultado esperado. Já os critérios internos consideram apenas medidas associadas a própria base de dados, sem considerar aspectos exteriores. Por último, os critérios relativos comparam estruturas geradas por diferentes métodos de clusterização aplicados à mesma base de dados.

Uma outra forma de entender a questão de validação de clusterizações é visualizá-la como o problema de determinar o valor-ótimo para o número de *clusters*. Neste caso, os índices sugeridos na literatura podem ser divididos em dois grupos. O primeiro considera que, se a estatística de teste associada ao número de grupos não exibir uma tendência de aumento ou decréscimo na medida em que este número de *clusters* aumenta, então a melhor solução será o valor

máximo ou mínimo dependendo de como a estatística for calculada. Já o segundo trata dos casos em que esta tendência de aumento ou decréscimo é observada, de tal forma que a melhor solução é encontrada quando ocorre uma abrupta alteração local na estatística calculada, observada através da presença de um "joelho" no gráfico formado (exemplos serão apresentados na sequência).

De maneira geral, os índices de validação são definidos com base em dois conceitos importantes. O primeiro, de compactação, mede o quão próximos estão os elementos pertencentes a cada um dos *clusters*. Espera-se sempre que tais membros de um mesmo grupo estejam o mais próximo possível entre si e tal medição é usualmente feita através de uma medida de variância, que deve ser mínima para que haja maior qualidade da clusterização. O segundo conceito, de separabilidade, indica o quão distintos os *clusters* são entre si. Normalmente, mede-se tal característica a partir dos métodos apresentados anteriormente na descrição do algoritmo AGNES.

Na sequência desta seção do trabalho, serão apresentados diversos índices de validação de clusterização. Destes, os principais serão posteriormente utilizados para comparar e validar os métodos de clusterização propostos quando aplicados à base de dados estudada.

5.1. Critérios Externos

O conceito básico por trás dos critérios externos é testar se a base de dados é aleatoriamente estruturada ou não. Usualmente, testa-se tal hipótese através de técnicas de simulação de Monte Carlo, especialmente nos casos em que o tamanho da base é grande e, consequentemente, é o esforço computacional necessário para o processamento dos dados. O objetivo da aplicação desta técnica é determinar a função densidade de probabilidade da estatística de teste alvo e, assim, construir intervalos de confiança que permitam ou não rejeitar a hipótese nula, a qual afirma que os objetos são aleatoriamente estruturados.

A utilização dos critérios externos pode assumir duas formas diferentes. Na primeira, compara-se o resultado da estrutura C encontrada com uma partição independente P da base de dados construída com base na intuição de como deveria ser a estrutura correta. Na segunda, compara-se a matriz de proximidade resultante da clusterização com a tal partição independente P. Os principais

índices que compõem este grupo de critérios externos estão associados à primeira forma descrita no parágrafo anterior.

Considere $X = \{x_1, \dots, x_n\}$ um conjunto de pontos da base de dados, $C = \{c_1, \dots, c_k\}$ a estrutura de *clusters* formada, $P = \{p_1, \dots, p_m\}$ a partição esperada da base de dados e $Y = \{x_a, x_b | x_a \neq x_b; x_a, x_b \in X\}$ o conjunto de todos os pares de pontos com tamanho $M = \frac{n(n-1)}{2}$. Cada par de pontos (x_a, x_b) irá pertencer a um certo subconjunto de Y dependendo da sua situação:

Y^{ss} → pontos que pertencem ao mesmo *cluster* e partição, com tamanho n_{ss};

$$(x_a, x_b) \in Y^{ss} \Leftrightarrow \exists\, c_q \in C, p_w \in P: x_a, x_b \in c_q;\ x_a, x_b \in p_w.$$

Y^{sd} → pontos que pertencem ao mesmo *cluster*, mas a partições diferentes, com tamanho n_{sd};

$$(x_a, x_b) \in Y^{sd} \Leftrightarrow \exists\, c_q \in C, p_{w1} \neq p_{w2}: x_a, x_b \in c_q;\ x_a \in p_{w1}, x_b \in p_{w2}$$

Y^{ds} → pontos que pertencem a *clusters* diferentes, porém à mesma partição, com tamanho n_{ds};

$$(x_a, x_b) \in Y^{ds} \Leftrightarrow \exists\, c_{q1} \neq c_{q2}, p_w \in P: x_a \in c_{q1}, x_b \in c_{q2};\ x_a, x_b \in p_w$$

Y^{dd} → pontos que pertencem a *clusters* e partições diferentes, com tamanho n_{dd};

$$(x_a, x_b) \in Y^{dd} \Leftrightarrow \exists\, c_{q1} \neq c_{q2}, p_{w1} \neq p_{w2}:\ x_a \in c_{q1}, x_b \in c_{q2}, x_a \in p_{w1}, x_b \in p_{w2}$$

Os valores de n_{ss}, n_{sd}, n_{ds} e n_{dd} podem ser vistos como o número de verdadeiros positivos, falsos positivos, falsos negativos e verdadeiros negativos, respectivamente. Alguns índices de validação externos utilizam-se exatamente destes valores para medir a similaridade entre C e P. O primeiro deles é a medida F ($F \in [0,1]$), calculada a partir da seguinte expressão:

$$F = \frac{2 \times \left(\frac{n_{dd}}{n_{dd} + n_{sd}}\right) \times \left(\frac{n_{ss}}{n_{ss} + n_{ds}}\right)}{\left(\frac{n_{dd}}{n_{dd} + n_{sd}}\right) + \left(\frac{n_{ss}}{n_{ss} + n_{ds}}\right)} \tag{43}$$

Um outro índice, chamado Estatística Rand ($R \in [0,1]$) mede o percentual de pares corretamente designados em relação ao total de pontos:

$$R = \frac{n_{ss} + n_{dd}}{n_{ss} + n_{sd} + n_{ds} + n_{dd}} \tag{44}$$

O Coeficiente de Jaccard ($J \in [0,1]$) é simplesmente uma variação da Estatística Rand, ao considerar que o valor de n_{dd} pode ser derivado a partir dos demais e, portanto, a expressão pode ser reescrita da seguinte maneira:

$$J = \frac{n_{ss}}{n_{ss} + n_{sd} + n_{ds}} \tag{45}$$

Podemos destacar ainda o Índice de Folkes e Mallows ($FM \in \left[0, \sqrt{2}\right]$):

$$FM = \frac{n_{ss}}{\sqrt{(n_{ss} + n_{sd})(n_{ss} + n_{ds})}} \tag{46}$$

Todos estes índices apresentados até o momento atuam de tal forma que, quanto maior o seu valor, mais parecido C é de P.

Existem ainda, dentre os critérios externos, índices que comparam matrizes formadas a partir de diferentes esquemas de clusterização, cujos elementos (i, j) podem assumir valor igual 1 se x_i e x_j pertencem ao mesmo *cluster* ou valor 0 caso contrário. Considere agora que P representa esta matriz para a partição intuitiva da base de dados, com média μ_p e variância σ_p^2 e que C, analogamente, representa esta matriz para a estrutura resultante da clusterização, também possuindo média μ_c e variância σ_c^2. Podemos, então, apresentar a Estatística de Hubert (Γ) e seu valor normalizado, dados pelas seguintes expressões:

$$\Gamma = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij} P_{ij}}{n_{ss} + n_{sd} + n_{ds} + n_{dd}}$$

$$\bar{\Gamma} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (C_{ij} - \mu_c)(P_{ij} - \mu_p)}{(n_{ss} + n_{sd} + n_{ds} + n_{dd})\sigma_c \sigma_p} \tag{47}$$

5.2.
Critérios Internos

Como citado anteriormente, os critérios internos são aqueles que buscam medir a qualidade da clusterização com base apenas nas informações contidas na própria base de dados. Existem diversos índices dentro desta lista, dentre os quais os principais serão destacados a seguir.

O primeiro deles, comumente utilizado para avaliação de métodos hierárquicos, chama-se $CPCC$ (do inglês, *Cophenetic Correlation Coefficient*) e é usado para comparar a clusterização resultante com uma certa matriz C tal que cada um de seus elementos C_{ij} representa o nível de proximidade com que dois objetos i e j aparecem juntos em um mesmo *cluster* pela primeira vez no processo hierárquico. Quanto mais próximo de 1 for o valor para este índice, melhor pode ser considerada a clusterização. Considere P como sendo a matriz de proximidade dos dados, $M = \frac{n(n-1)}{2}$ e μ_c e μ_p as médias das matrizes C e P. Assim, o coeficiente $CPCC$ pode ser calculado a partir da seguinte expressão:

$$CPCC = \frac{\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_{ij}P_{ij} - \mu_p\mu_c}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P_{ij}^2 - \mu_p^2\right)\left(\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_{ij}^2 - \mu_c^2\right)}} \tag{48}$$

tal que

$$\mu_p = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P_{ij}$$

$$\mu_c = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_{ij} \tag{49}$$

Um outro índice interno é conhecido como Estatística *Gap*. Ele pode ser aplicado a qualquer método de clusterização, tanto hierárquico quanto particional, e avalia mudanças na dispersão interna do *cluster*. É dado pela seguinte fórmula:

$$Gap_n(k) = E_n^*(\log W_k) - \log W_k \tag{50}$$

tal que E_n^* é o valor esperado para uma amostra de tamanho n da distribuição de referência e

$$W_k = \sum_{w=1}^{k} \frac{\sum_{i=1}^{n_w-1} \sum_{j=i+1}^{n_w} d(i,j)}{2n_w} \tag{51}$$

sendo n_w o número de elementos no *cluster* C_w. O número ótimo de grupos k^* será aquele que maximizar o valor de $Gap_n(k)$.

Existe ainda o *Bayesian Information Criteria* (*BIC*), cuja expressão é:

$$BIC = -L(\theta) + v \ln n \tag{52}$$

tal que $L(\theta)$ é o log da função de verossimilhança associada a cada método, v é o número de parâmetros livres do modelo Gaussiano e n é o tamanho da base de dados. Quanto menor o índice *BIC*, melhor o modelo.

Em seu trabalho, Kaufman e Rousseeuw [1] apresentaram um outro índice de validação baseado no conceito da "largura da silhueta" de uma clusterização. Para qualquer objeto i pertencente à base de dados e para cluster C_k a que foi designado, a largura da silhueta de i é dada por:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{53}$$

tal que $a(i)$ representa a dissimilaridade média do objeto i em relação a todos os demais objetos de C_k e $b(i)$ representa a dissimilaridade média de i em relação a todos os objeto pertencentes ao cluster mais próximo de C_k, denominado C_h. Estes valores, por sua vez, são calculados a partir das seguintes expressões:

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq i} d(i,j) \tag{54}$$

$$b(i) = \min_{C_h \neq C_k} \frac{1}{|C_k|} \sum_{j \in C_h} d(i,j)$$

A estatística $s(i)$ pode assumir valores entre -1 e 1. Quando se aproxima de 1, isso significa que a dissimilaridade "interna" $a(i)$ é muito menor do que a menor dissimilaridade externa $b(i)$. Logo, o objeto i é considerado bem classificado. Por outro lado, quando a estatística assume valor próximo a -1, pode-se dizer que a classificação é imprecisa. Já quando $s(i)$ fica próximo de zero, a conclusão não é direta, uma vez que i está situado a distâncias equivalentes de seu próprio *cluster* e do mais próximo. Para avaliar os modelos, calcula-se a largura média da silhueta para diferentes valores de k.

Calisnki e Harabasz [5], por sua vez, definiram em seus estudos um índice que, posteriormente, foi generalizado para medidas de dissimilaridades por Hennig e Liao [6] da seguinte maneira:

$$CH = \frac{B(k)(n-k)}{W(k)(k-1)} \tag{55}$$

tal que

$$W(k) = \sum_{h=1}^{k} \frac{1}{|C_H|} \sum_{i,j \in C_H} d(i,j)^2$$

$$B(k) = \frac{1}{n} \sum_{i,j=1}^{n} d(i,j)^2 - W(k) \tag{56}$$

O melhor valor de k é aquele que maximiza a estatística CH. Uma importante característica deste índice é seu comportamento. A estatística $W(k)$ inicia com um valor comparativamente grande e diminui à medida em que se aproxima da solução ótima, já que a compactação do *cluster* aumenta. Quando a solução ótima é ultrapassada, um aumento nesta compactação e, consequentemente, um decréscimo no valor de $W(k)$ pode ser notado. Por outro

lado, $B(k)$ se comporta de maneira inversa, aumentando enquanto aproxima-se do valor ótimo de k e mostrando uma redução quando ultrapassa-se tal ponto ótimo.

Outro índice foi proposto por Davies e Bouldin [7]. O objetivo deste é identificar grupos de *clusters* que sejam compactos e bem-separados e a melhor solução é aquela que minimiza seu valor. A sua definição é a seguinte:

$$DB(K) = \frac{1}{K}\sum_{k=1}^{K} R_k \tag{57}$$

tal que:

$$\begin{gathered} R_k = \max_{j=1,\dots,K, j\neq k}\left(\frac{S_k + S_j}{d_{k,j}}\right), k \in [1,\dots,K] \\ S_k = \frac{1}{\sum_{i=1}^{N} w_{k,i}}\sum_{i=1}^{N} w_{k,i}\|x_i - \bar{x}_k\| \\ d_{k,j} = \left\|\bar{x}_k - \bar{x}_j\right\| \end{gathered} \tag{58}$$

Destacam-se ainda, outros quatro índices internos. O primeiro, chamado de Coeficiente Gama de Goodman e Kruskal (*g2*), foi apresentado por Gordon [8] como uma comparação entre todas as dissimilaridades intra-*cluster* e todas as dissimilaridades inter-*clusters*. Tal comparação é considerada concordante se a dissimilaridade intra é estritamente menor do que a dissimilaridade inter. Da mesma forma, ela é considerada discordante se o primeiro valor é estritamente maior do que o segundo. Matematicamente, $g2(k)$ é calculado da seguinte forma:

$$g2(k) = \frac{S_+ - S_-}{S_+ + S_-} \tag{59}$$

tal que S_+ e S_- denotam o número de comparações concordantes e discordantes, respectivamente. O valor-ótimo de k é aquela que maximiza $g2(k)$.

Gordon [8] definiu também o Coeficiente *g3* da seguinte maneira:

$$g3(k) = \frac{D(k) - D_{min}}{D_{max} - D_{min}} \tag{60}$$

tal que $D(k)$ representa a soma de todas as dissimilaridades intra-*cluster* de uma partição de k *clusters*. Se a partição possui r dissimilaridades intra-*cluster*, então D_{min} é definido como a soma das r menores e D_{max} como a soma das r maiores. Ao contrário dos demais casos, k é escolhido de tal forma que minimize $g3(k)$.

O terceiro foi proposto por Maulik e Bandyopadhyay [9], é chamado de Índice $I(k)$ e pode ser definido da seguinte forma:

$$I(k) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^p \tag{61}$$

tal que:

$$E_K = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{k,i} \|x_i - \bar{x}_k\|$$
$$D_K = \max_{p,q=1,\ldots,k, p \neq q} \left\|\bar{x}_p - \bar{x}_q\right\| \tag{62}$$

O primeiro termo da expressão de $I(k)$ normaliza o índice pela número total de *clusters*. Já o segundo determina a soma do erro quadrático da base completa em relação ao erro intra-*cluster* para uma certa clusterização. O terceiro considera a máxima diferença observada entre dois dos *clusters* formados. O parâmetro p controla certas diferenças entre os métodos analisados e, usualmente, assume valor igual a 2, por recomendação dos próprios desenvolvedores do índice.

O quarto é o índice Dunn, cujo objetivo principal é identificar *clusters* compactos e bem separados. Matematicamente, é definido da seguinte forma:

$$Dunn_{nc} \min_{i=1,\ldots,nc} \left\{ \min_{j=i+1,\ldots,nc} \left(\frac{d(c_i, c_j)}{\max_{k=1,\ldots,nc} diam(c_k)} \right) \right\} \tag{63}$$

tal que $d(c_i, c_j)$ é a função de dissimilaridade entre dois *clusters* e $diam(c_k)$ é o diâmetro de um certo *cluster*. Quanto maior o valor desta estatística, maior a evidência da existência de clusters compactos e bem separados.

5.3. Critérios Relativos

Os critérios externos e internos acima destacados baseiam-se em testes estatísticos para a validação dos métodos de clusterização e, portanto, exigem um elevado poder computacional. Os critérios relativos surgem, portanto, como uma alternativa, ao não exigirem a performance de testes estatísticos. A ideia fundamental por trás destes critérios é escolher uma clusterização, dentre um conjunto de métodos previamente selecionados. A utilização destes critérios pode ser considerada mais restrita do que dos demais e, portanto, o aprofundamento de seu estudo foge do escopo deste trabalho.

6.
Estudo de Caso

Após a caracterização de diferentes tipos de clusterização existentes e da descrição dos principais métodos de validação, esta seção da dissertação irá apresentar um estudo de caso para um conjunto de ações listadas na Bolsa de Valores de São Paulo (Bovespa). Conforme será detalhado a seguir, este conjunto foi formado a partir dos principais índices já existentes e comumente utilizados pelo mercado, como o Ibovespa e o IbrX. Entretanto, o escopo de empresas estudadas foi alterado para considerar algumas que não compõem estes índices, porém têm representatividade dentro da economia real nacional, e para excluir empresas cujo histórico de cotações não seja relevante.

O objetivo desta seção é verificar, com a utilização de técnicas de clusterização, se ações cujos preços de fechamento apresentam comportamento semelhante estão associadas a empresas que participam do mesmo setor da economia. Esta é uma percepção comum dentre os agentes de mercado, o que garante a importância de sua comprovação. A Bolsa de Valores de São Paulo estabelece uma classificação setorial oficial que, aqui, consideramos uma classificação teórica. Esta abertura será apresentada e descrita na sequência, para que seja possível compará-la com os resultados obtidos.

Os modelos analisados foram determinados a partir da combinação dos três diferentes métodos de clusterização (PAM, AGNES e DIANA) com as três formas não-paramétricas de cálculo de dissimilaridade entre dados apresentadas (Spearman, Kendall e Hoeffding). O Coeficiente de Pearson, apesar de descrito, não será analisado porque há forte evidência de que séries temporais financeiras não possuem distribuição Normal de probabilidade e, conforme mencionado, este aspecto representa um empecilho à sua utilização. Para a avaliação e comparação dos modelos, também nesta etapa de apresentação dos resultados, serão expostos os valores obtidos para diversos índices de validação.

6.1.
Séries Temporais Financeiras

Os mercados financeiros são sistemas complexos e estudos mostram que há fortes evidências contrárias a tese de que retornos de ativos financeiros exibem distribuição Normal Multivariada. As séries de retornos financeiros apresentam certas regularidades estatísticas, conhecidas como "fatos estilizados" que as diferenciam de outros tipos de séries e, acima de tudo, as caracterizam. O primeiro deles diz respeito a auto-correlação dos retornos: diferentes estudos ao longo do tempo mostraram que séries de retornos semanais e mensais de ações apresentam fraca correlação negativa, enquanto retornos diários, semanais e mensais de índices de ações são positivamente correlacionados. Por outro lado, a auto-correlação quando se considera os retornos absolutos ou quadráticos é positiva e significativa, com decaimento lento.

O segundo fato estilizado afirma que as séries são não-estacionárias. Elas apresentam agrupamentos de volatilidade, característica esta evidenciada pelos modelos ARCH e GARCH comumente utilizados para sua modelagem, os quais garantem que o desvio-padrão dos retornos não é constante ao longo do tempo. Além disso, as distribuições de probabilidade de séries financeiras são aproximadamente simétricas e apresentam alta curtose ou leptocurtose (caudas gordas comparadas as da distribuição Normal de probabilidade).

Outros fatos estilizados estão associados a volatilidade deste tipo de série. Volatilidade, cabe destacar, é o desvio-padrão da variação em valor de um instrumento financeiro e, desta forma, é considerada uma forma de mensuração de risco. A distribuição de probabilidade da volatilidade é log-Normal e sua função de auto-correlação exibe dependência de longo-prazo. Há presença de heterogeneidade e as séries são não-lineares.

Na sequência, as ações estudadas foram *clusterizadas* considerando os preços diários de fechamento (representados por S_t). Para comparar os diferentes valores, é preciso normalizar os dados, algo que foi feito ao se considerar o log-retorno das ações (sendo o retorno representado por R_t):

$$R_{t+1} = \log\frac{S_{t+1}}{S_t} = \log\frac{S_t + \Delta_t}{S_t} = \log\left(1 + \frac{\Delta_t}{S_t}\right) \tag{64}$$

Utilizando a aproximação pela Série de Taylor, podemos escrever que:

$$R_{t+1} \cong \frac{\Delta_t}{S_t} = \frac{S_{t+1} - S_t}{S_t} \tag{65}$$

6.2. Descrição das Empresas Estudadas

O Índice Ibovespa, da Bolsa de Valores de São Paulo, é considerado o principal índice representativo do desempenho médio das cotações do mercado de ações brasileiro. Criado em 1968, é composto atualmente por 69 ações, as quais representam as principais empresas listadas no país. Sua determinação considera certos critérios, especialmente de negociabilidade e liquidez, que limitam o número de empresas consideradas. Como o objetivo deste trabalho é analisar ações de firmas que representem a economia como um todo, é importante considerar um escopo maior. Por esta razão, decidiu-se tomar como base inicial as empresas que compõem um outro importante índice da bolsa nacional, o IbrX (Índice Brasil), que mede o retorno de uma carteira teórica de 100 ações dentre as mais negociadas na Bovespa. É, portanto, mais abrangente do que o Ibovespa e surge como uma opção especialmente porque, atualmente, todas as ações que compõem o Ibovespa estão presentes no IbrX.

Para formar o grupo final estudado, entretanto, foi preciso fazer alguns ajustes. Conforme mencionado, mesmo que as medições sejam feitas com as ações, este trabalho envolve uma análise das empresas em si, uma vez que consideramos sua representatividade dentro da economia real. Assim, estudar ao mesmo tempo ações ordinárias e preferenciais de uma mesma companhia representaria um efeito de "dupla contagem". Da mesma forma, considerar ações que representem *holdings* as quais, na prática, têm seus desempenhos associados exclusivamente ao ativo do qual são donas, traria o mesmo efeito. Portanto, foram excluídas da análise as ações ordinárias de empresas com listagem dupla, considerando para esta escolha o fato de que tal classe usualmente é menos líquida e representativa do desempenho das empresas do que a classe preferencial. Também foram excluídas ações de *holdings* que possuam somente um investimento significativo em carteira, já representado por alguma outra ação. Saem assim, do escopo de análise, os seguintes papéis: AMBV3 (ações ordinárias

da Ambev), ELET3 (ações ordinárias da Eletrobrás), PETR3 (ações ordinárias da Petrobras), USIM3 (ordinárias da Usiminas), VALE3 (idem para Vale), BBDC3 (idem, Bradesco), OIBR3 (idem, Oi), GOAU4 (holding com ações da Gerdau), BRAP4 (holding com ações da Vale), UGPA3 (holding da Ultrapar) e Itaúsa (holding com ações do Itaú).

Uma outra alteração foi desconsiderar empresas cujos históricos de cotação sejam mais recentes do que o primeiro dia de negociação do ano de 2009. Desta maneira, foram considerados dados desde o dia 2 de janeiro de 2009 até 31 de outubro de 2012. Esta data limite foi escolhida com base em dois principais motivos. O primeiro foi eliminar o período mais intenso da última crise econômica mundial, ocorrido no segundo semestre de 2008. Evidências empíricas mostram que a movimentação das ações nestes momentos tende a estar mais associada a questões como liquidez e base de acionistas (a eventual falência de um fundo de investimentos, por exemplo, pode levar a vendas excessivas de uma certa ação que esteja na carteira do mesmo) do que questões associadas às características setoriais que pretende-se estudar neste trabalho. O segundo motivo foi manter uma quantidade mínima de observações que permita uma avaliação mais precisa dos resultados. Diante desta limitação, foram excluídas da análise empresas que tenham passado por processos de mudança na listagem na Bovespa, como a unificação de suas classes de ações, e também empresas cujas ofertas iniciais tenham sido realizadas após o início de 2009. Portanto, saem do escopo de análise as seguintes ações: AEDU3 (Anhanguera), ALLL3 (ALL), BRSR3 (Banrisul), BRPR3 (BR Properties), CCXC3 (CCX), CTIP3 (Cetip), CIEL3 (Cielo), ECOR3 (Ecorodovias), FIBR3 (Fíbria), HRTP3 (HRT), MGLU3 (Magazine Luiza), MPLU3 (Multiplus), OSXB3 (OSX), QGEP3 (Queiroz Galvão) e SANB11 (Santander).

À lista restante, decidiu-se pelo acréscimo de algumas ações que, apesar de não estarem presentes nos índices considerados, estão associadas a empresas importantes para a economia nacional. Um total de 11 novos ativos foram considerados, os quais representam as seguintes empresas: Equatorial Energia (EQTL3), Estácio (ESTC3), Kroton (KRTO11), WEG (WEGE3) Santos Brasil (STBP11), Magnesita (MAGG3), Helbor (HBOR3), Lopes Brasil (LPSB3), Lojas Marisa (AMAR3), Guararapes (GUAR3) e M Dias Branco (MDIA3).

Considerando todos estes ajustes, foram analisados 86 ativos no total. Cabe destacar que o número de empresas que são relevantes na economia nacional, é maior do que o número de companhias que foram consideradas no estudo. Entretanto, muitas delas não satisfazem o critério do histórico mínimo de cotação citado anteriormente, e, portanto, não puderam ser incluídas.

6.2. Classificação Setorial Oficial da Bovespa

A própria Bovespa disponibiliza uma forma de classificação de empresas de acordo com o setor econômico a que pertencem, estabelecendo ainda os possíveis subsetores e segmentos. Esta classificação está disposta abaixo e todas as 86 empresas consideradas estão devidamente associadas a sua classificação oficial. Este critério, que considera a existência de 10 setores e 46 subsetores, será usado em comparações com os resultados obtidos após a aplicação da clusterização.

Setor Econômico	Subsetor	Segmento	Empresas
1. Petróleo, Gás e Biocombustíveis	1.1. Petróleo, Gás e Biocombustíveis	1.1.1. Exploração e/ou Refino	OGX; Petrobras
		1.1.2 Equipamentos e Serviços	
2. Materiais Básicos	2.1. Mineração	2.1.1. Minerais Metálicos	MMX; Vale
		2.1.2. Minerais Não Metálicos	
	2.2. Siderurgia e Metalurgia	2.2.1. Siderurgia	Gerdau; CSN; Usiminas
		2.2.2. Artefatos de Ferro e Aço	
		2.2.3. Artefatos de Cobre	
	2.3. Químicos	2.3.1. Petroquímicos	Braskem

		2.3.2. Fertilizantes e Defensivos	
		2.3.3. Químicos Diversos	
	2.4. Madeira e Papel	2.4.1. Madeira	Duratex
		2.4.2. Papel e Celulose	Klabin; Suzano
	2.5. Embalagens	2.5.1. Embalagens	
	2.6. Materiais Diversos	2.6.1. Materiais Diversos	Magnesita
3. Bens Industriais	3.1. Material de Transporte	3.1.1. Material Aeronáutico e de Defesa	Embraer
		3.1.2. Material Ferroviário	
		3.1.3. Material Rodoviário	Iochpe; Marcopolo; Randon
	3.2. Equipamentos Elétricos	3.2.1. Equipamentos Elétricos	
	3.3. Máquinas e Equipamentos	3.3.1. Motores, Compressores e Outros	WEG
		3.3.2. Máquinas e Equipamentos Industriais	
		3.3.3. Máquinas e Equipamentos Construção e Agrícolas	
		3.3.4. Máquinas e Equipamentos	

		Hospitalares	
		3.3.5. Armas e Munições	
	3.4. Serviços	3.4.1. Serviços Diversos	
	3.5. Comércio	3.4.2. Material de Transporte	
		3.4.3. Máquinas e Equipamentos	
4. Construção e Transporte	4.1. Construção e Engenharia	4.1.1. Materiais de Construção	
		4.1.2. Construção Civil	Brookfield; Cyrela; Even; Eztec; Gafisa; MRV; PDG; Rossi; Tecnisa; Helbor
		4.1.3. Construção Pesada	
		4.1.4. Engenharia Consultiva	
		4.1.5. Serviços Diversos	
		4.1.6. Intermediação Imobiliária	BR Brokers; Lopes Brasil
		4.1.7. Comércio de Material de Construção	
	4.2. Transporte	4.2.1. Transporte Aéreo	GOL
		4.2.2. Transporte Ferroviário	

		4.2.3. Transporte Hidroviário	
		4.2.4. Transporte Rodoviário	
		4.2.5. Exploração de Rodovias	CCR; OHL
		4.2.6. Serviços de Apoio e Armazenagem	LLX; Santos Brasil
5. Consumo Não Cíclico	5.1. Agropecuária	5.1.1. Agricultura	V-Agro
	5.2. Alimentos Processados	5.2.1. Açúcar e Álcool	Cosan
		5.2.2. Café	
		5.2.3. Grãos e Derivados	
		5.2.4. Carnes e Derivados	Brasil Foods; JBS; Marfrig
		5.2.5. Laticínios	
		5.2.6. Alimentos Diversos	M Dias Branco
	5.3. Bebidas	5.3.1. Cervejas e Refrigerantes	Ambev
	5.4. Fumo	5.4.1. Cigarros e Fumo	Souza Cruz
	5.5. Produtos de Uso Pessoal e de Limpeza	5.5.1. Produtos de Uso Pessoal	Natura
		5.5.2. Produtos de Limpeza	
	5.6. Saúde	5.6.1. Medicamentos e Outros Produtos	
		5.6.2. Serviços Médico-	Amil; Dasa; Odontoprev

		Hospitalares, Análises e Diagnósticos	
	5.7. Diversos	5.7.1. Produtos Diversos	Hypermarcas
	5.8. Comércio e Distribuição	5.8.1. Alimentos	Pão de Açúcar
		5.8.2. Medicamentos	RaiaDrogasil
6. Consumo Cíclico	6.1. Tecidos, Vestuário e Calçados	6.1.1. Fios e Tecidos	
		6.1.2. Vestuário	Hering
		6.1.3. Calçados	
		6.1.4. Acessórios	
	6.2. Utilidades Domésticas	6.2.1. Eletrodomésticos	
		6.2.2. Móveis	
		6.2.3. Utensílios Domésticos	
	6.3. Automóveis e Motocicletas	6.3.1. Automóveis e Motocicletas	
	6.4. Mídia	6.4.1. Produção e Difusão de Filmes e Programas	
		6.4.2. Jornais, Livros e Revistas	
		6.4.3. Publicidade e Propaganda	
	6.5. Hotéis e Restaurantes	6.5.1. Hotelaria	
		6.5.2. Restaurantes e Similares	
	6.6. Lazer	6.6.1. Bicicletas	
		6.6.2. Brinquedos e Jogos	

		6.6.3. Parques de Diversão	
		6.6.4. Produção de Eventos e Shows	
		6.6.5. Atividades Esportivas	
	6.7. Diversos	6.7.1. Serviços Educacionais	Estácio; Kroton
		6.7.2. Aluguel de Carros	Localiza
		6.7.3. Programas de Fidelização	
	6.8. Comércio	6.8.1. Tecidos, Vestuário e Calçados	Lojas Renner; Lojas Marisa; Guararapes
		6.8.2. Eletrodomésticos	
		6.8.3. Produtos Diversos	B2W; Lojas Americanas
7. Tecnologia da Informação	7.1. Computadores e Equipamentos	7.1.1. Computadores e Equipamentos	
	7.2. Programas e Serviços	7.2.1. Programas e Serviços	Totvs
8. Telecomunicações	8.1. Telefonia Fixa	8.1.1. Telefonia Fixa	Oi; Telefônica
	8.2. Telefonia Móvel	8.2.1. Telefonia Móvel	Tim
9. Utilidade Pública	9.1. Energia Elétrica	9.1.1. Energia Elétrica	AES Tietê; Cemig; Cesp; Copel; CPFL; Eletrobras; Eletropaulo;

			EdB; Light; MPX; Tractebel; CTEEP; Equatorial
	9.2. Água e Saneamento	9.2.1. Água e Saneamento	Copasa; Sabesp
	9.3. Gás	9.3.1. Gás	
10. Financeiros e Outros	10.1. Intermediários Financeiros	10.1.1. Bancos	Bradesco; Banco do Brasil; Itaú
		10.1.2. Sociedades Crédito e Financiamento	
		10.1.3. Sociedades Arrendamento Mercantil	
	10.2. Securitizadoras de Recebíveis	10.2.1. Securitizadoras de Recebíveis	
	10.3. Serviços Financeiros Diversos	10.3.1. Gestão de Recursos e Investimentos	
		10.3.2. Serviços Financeiros Diversos	BMFBovespa
	10.4. Previdência e Seguros	10.4.1. Seguradoras	Porto Seguro; Sul América
		10.4.2. Corretoras de Seguros	
	10.5. Exploração de Imóveis	10.5.1. Exploração de Imóveis	BRMalls; Iguatemi; Multiplan
	10.6. Holdings Diversificadas	10.6.1. Holdings Diversificadas	

	10.7. Serviços Diversos	10.7.1. Serviços Diversos	
	10.8. Outros	10.8.1. Outros	
	10.9. Fundos	10.9.1. Fundos Imobiliários	
		10.9.2. Fundos de Ações	
		10.9.3. Fundos de Direitos Creditórios	
		10.9.4 Fundos de Incentivo Setorial	
		10.9.5. Outros Títulos	

Tabela 1 – Classificação setorial fornecida pela BMFBovespa das empresas do mercado acionário brasileiro.

6.3. Apresentação de Resultados

No total, foram testados nove diferentes métodos de clusterização, uma vez que foram consideradas todas as combinações possíveis entre os três métodos propostos (PAM, AGNES e DIANA) e as três formas de cálculo de correlação entre os dados (Spearman, Kendall e Hoeffding). Para cada um destes métodos, foram calculados seis diferentes índices de validação (ASW – silhueta média, $g2$, $g3$, $Gamma$, $Dunn$ e CH) para valores de k variando entre três e treze.

Nesta seção serão apresentados os valores obtidos para os diferentes índices de validação considerados. O objetivo é verificar qual dos métodos, se algum, é capaz de modelar satisfatoriamente a base de dados estudada. Conforme anteriormante destacado, para verificar esta adequabilidade, deve-se procurar por mudanças abruptas nos valores dos índices de validação na medida em que alteramos o valor k do número de *clusters*. Dependendo do índice, esta mudança deve representar a presença de um ponto de máximo ou mínimo local.

Uma vez selecionado(s) o(s) método(s) eficaz(es), haverá uma análise mais precisa para verificar se as ações classificadas como pertencentes a um mesmo *cluster* por tal(is) método(s) de fato estão associadas a um mesmo setor econômico. Em outras palavras, será realizado um segundo processo de validação considerando uma abordagem mais qualitativa do(s) método(s) selecionado(s), a fim de avaliar a relação com a realidade do mercado.

6.3.1. Método PAM

Os gráficos abaixo mostram os índices de validação calculados quando utilizamos o método PAM junto com cada uma das formas de determinação da correlação entre os dados. De maneira resumida, somente o caso Spearman apresentou resultados satisfatórios, com a indicação de um valor-ótimo para $k = 10$ e, desta forma, será melhor analisado em seção posterior.

A primeira e mais forte evidência desta conclusão é o comportamento da estatística ASW. Apesar ser possível observar um crescimento de seu valor para $k \geq 12$, é clara a existência de um "joelho" quando $k = 10$. Para os casos Kendall e Hoeffding, o valor deste índice é crescente na medida em que o valor de k aumenta, impossibilitando uma conclusão em relação a adequação dos métodos de clusterização.

Quando observamos o comportamento das estatísticas $g2$, $g3$ e $Gamma$ também para o caso Spearman, os resultados evidenciam um ponto ótimo quando $k = 10$, porém de forma menos clara. Em outras palavras, apesar de ser possível visualizar um ponto de inflexão quando $k = 10$, seus valores voltam a crescer – no caso dos índices $g2$ e $Gamma$ – ou diminuir – no caso do índice $g3$ – de maneira mais pronunciada. Da mesma forma que no caso da estatística ASW, ao analisarmos o comportamento destas outras três estatísticas quando implementados os métodos Kendall e Hoeffding, não é possível aceitar a hipótese de que os métodos são eficientes.

As estatísticas $Dunn$ e CH, por sua vez, apresentam comportamento semelhante para os três casos estudados. Em todos, os resultados são inconclusivos no que tange o objetivo de selecionar os melhores métodos.

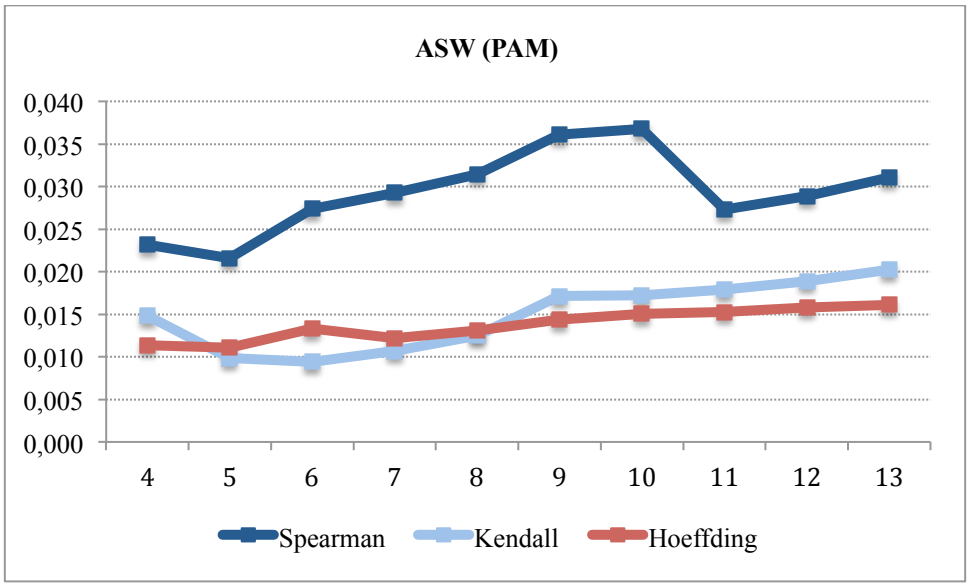

Figura 7 – Valores para o índice ASW para o método PAM aplicado junto aos métodos Spearman, Kendall e Hoeffding.

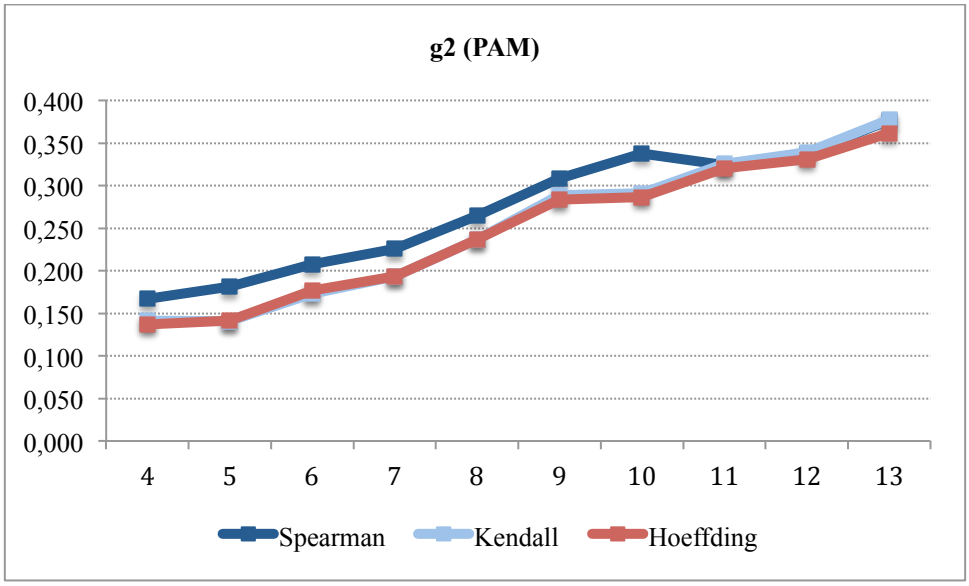

Figura 8 – Valores para o índice *g2* para o método PAM aplicado junto aos métodos Spearman, Kendall e Hoeffding.

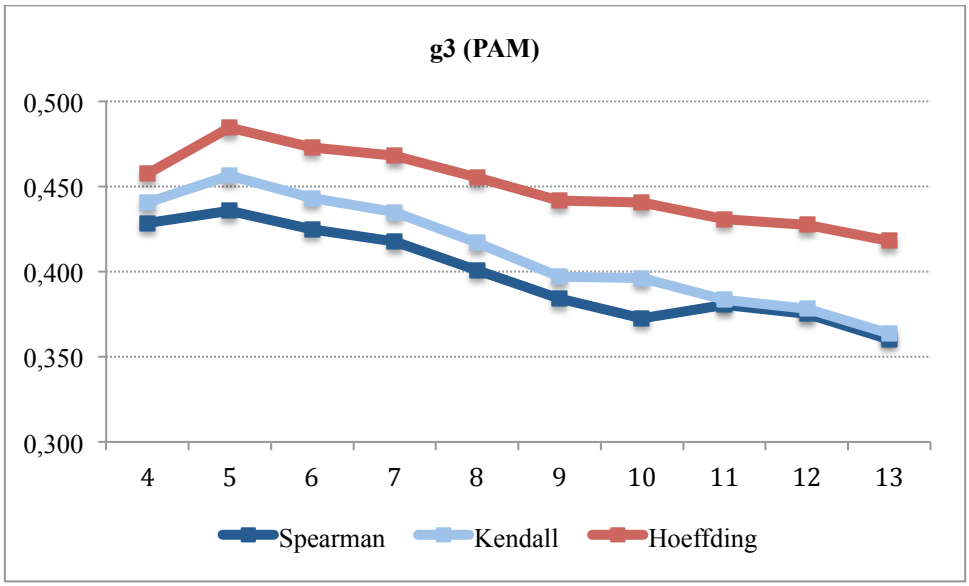

Figura 9 – Valores para o índice *g2* para o método PAM aplicado junto aos métodos Spearman, Kendall e Hoeffding.

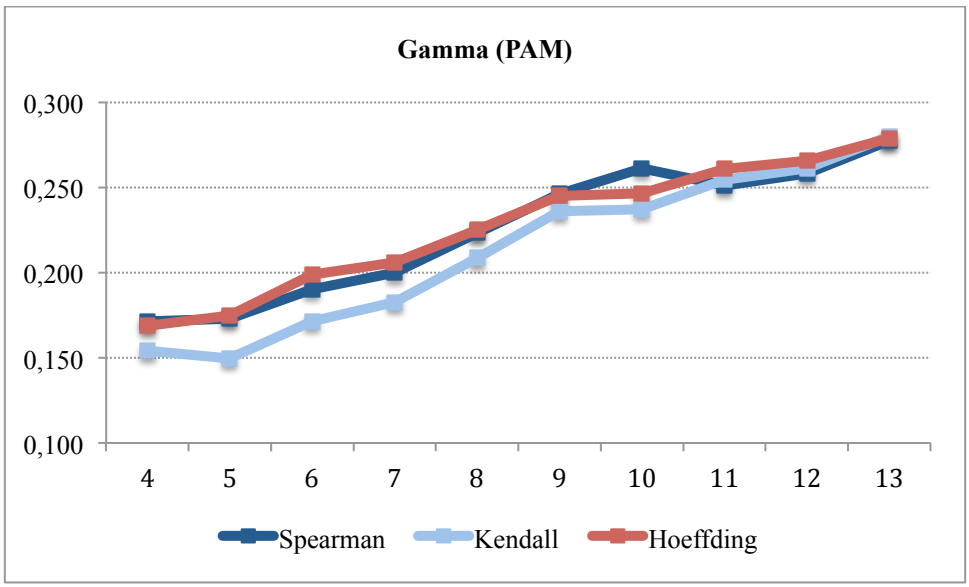

Figura 10 – Valores para o índice Gamma para o método PAM aplicado junto aos métodos Spearman, Kendall e Hoeffding.

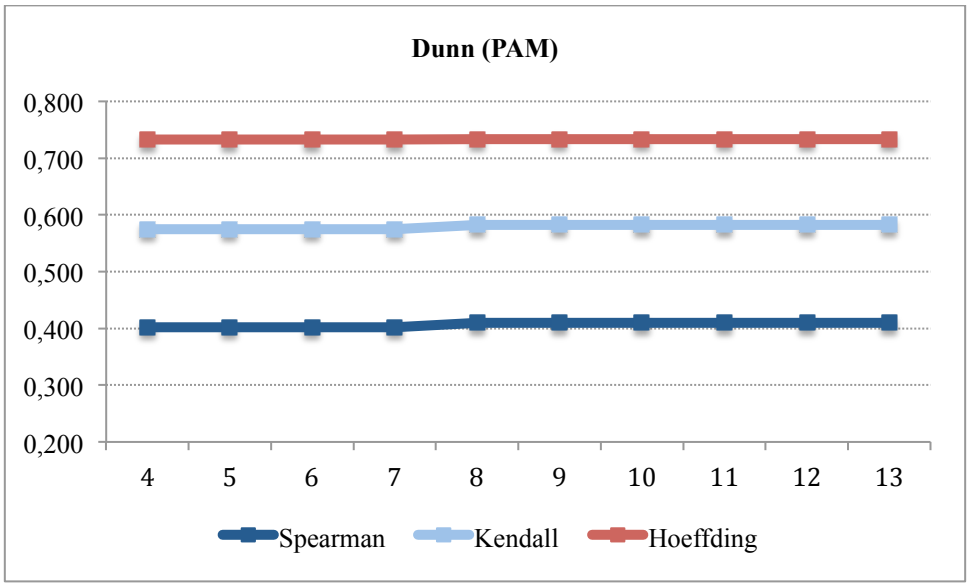

Figura 11 – Valores para o índice Dunn para o método PAM aplicado junto aos métodos Spearman, Kendall e Hoeffding.

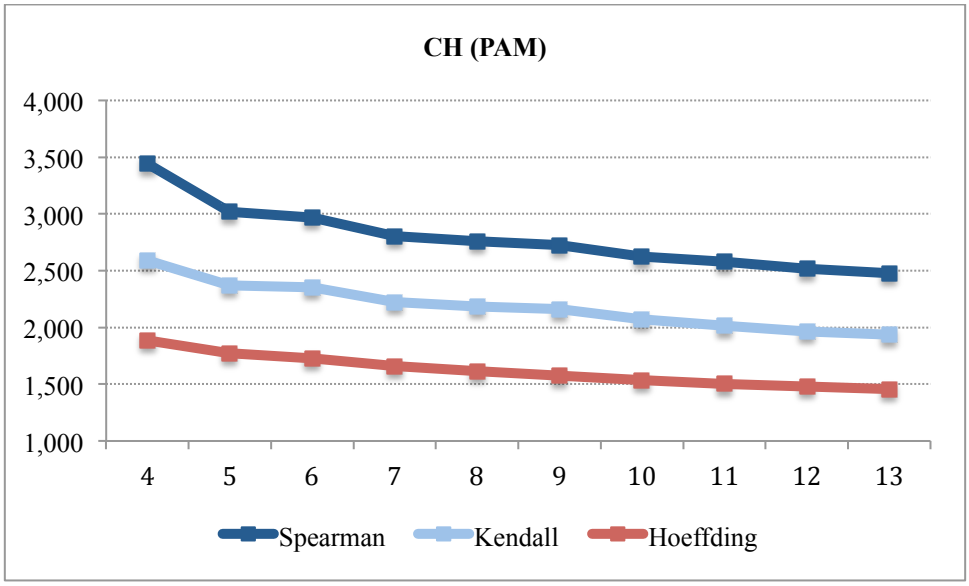

Figura 12 – Valores para o índice CH para o método PAM aplicado junto aos métodos Spearman, Kendall e Hoeffding.

6.3.2.
Método AGNES

Os índices de validação apresentados a seguir mostram comportamentos distintos entre si. Todos os resultados encontrados quando utilizou-se o método Hoeffding indicaram pouca ou nenhuma aderência do método, com os índices se comportando de forma contrária ao que deveria ser esperado. Desta forma, esta combinação pôde ser desconsiderada.

Por outro lado, para os casos em que foram usados os métodos Spearman e Kendall, os índices $g2$ e $g3$ apresentaram pontos claros de máximo e mínimo, respectivamente, para $k = 10$, indicando bom comportamento dos métodos. Já os demais índices computados (ASW, $Gamma$, $Dunn$ e CH) tiveram movimentação anômala, não permitindo qualquer conclusão quanto a eficácia dos métodos testados.

Com base, portanto, nos resultados acima destacados, pode-se concluir que as combinações do método AGNES com os métodos Spearman e Kendall deverão ser melhor estudados na seção seguinte desta dissertação.

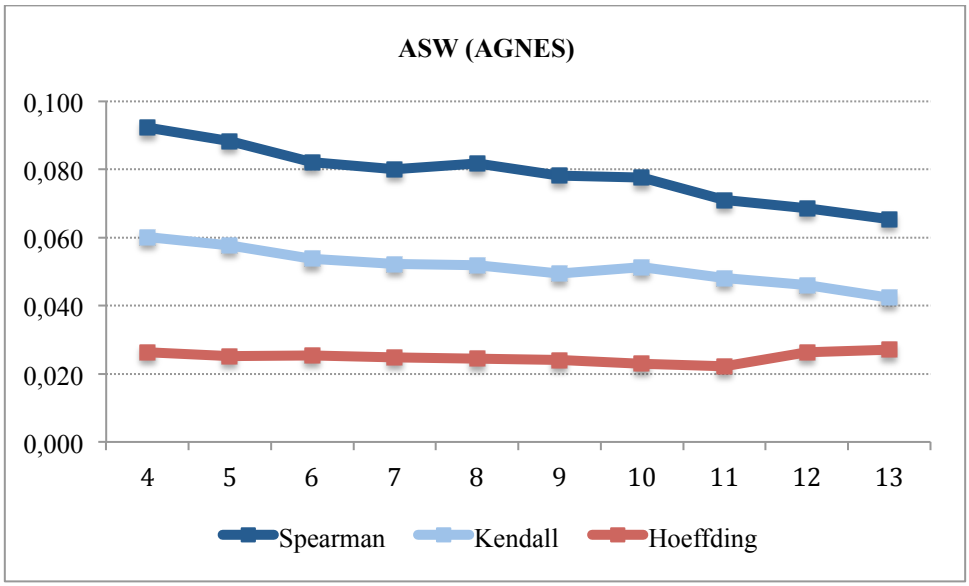

Figura 13 – Valores para o índice ASW para o método AGNES aplicado junto aos métodos Spearman, Kendall e Hoeffding.

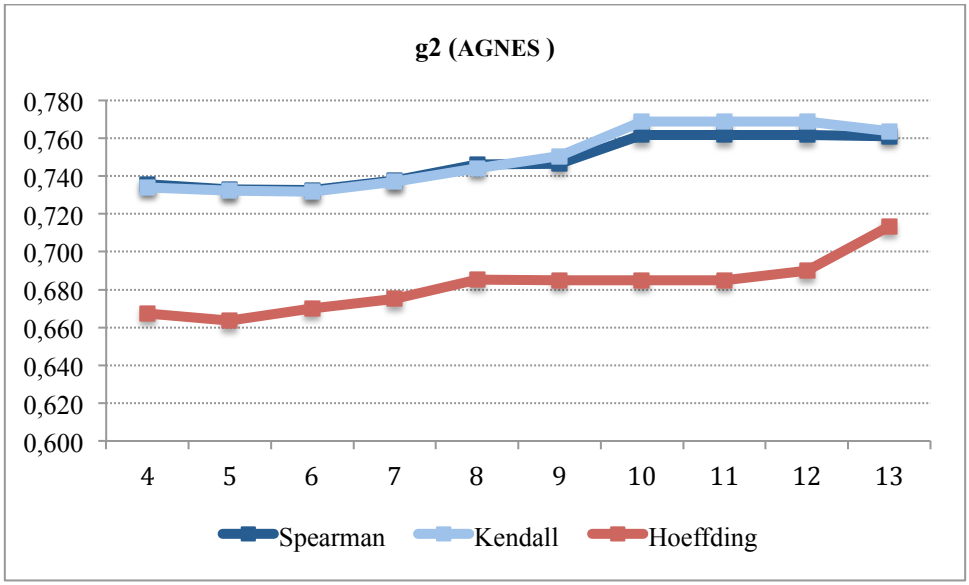

Figura 14 – Valores para o índice *g2* para o método AGNES aplicado junto aos métodos Spearman, Kendall e Hoeffding.

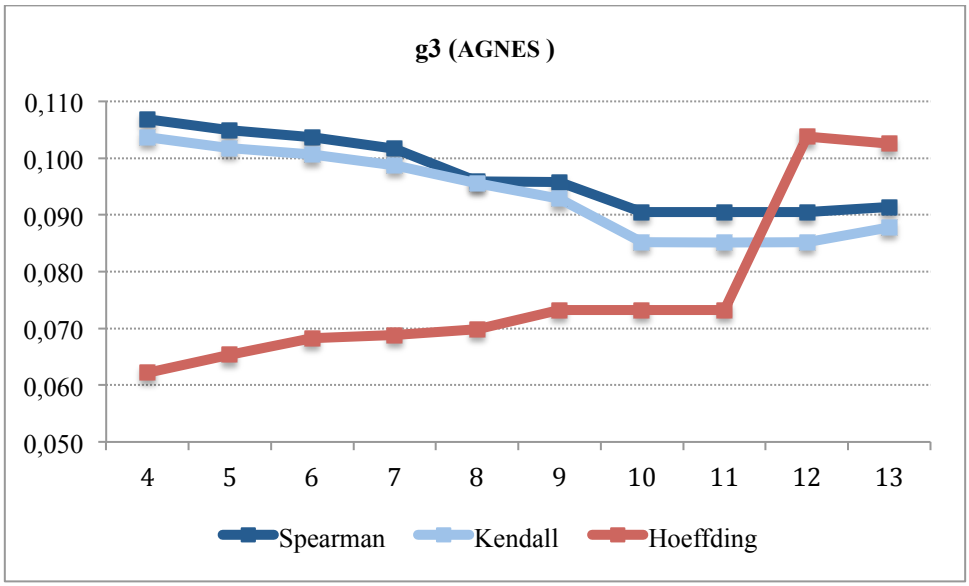

Figura 15 – Valores para o índice *g3* para o método AGNES aplicado junto aos métodos Spearman, Kendall e Hoeffding.

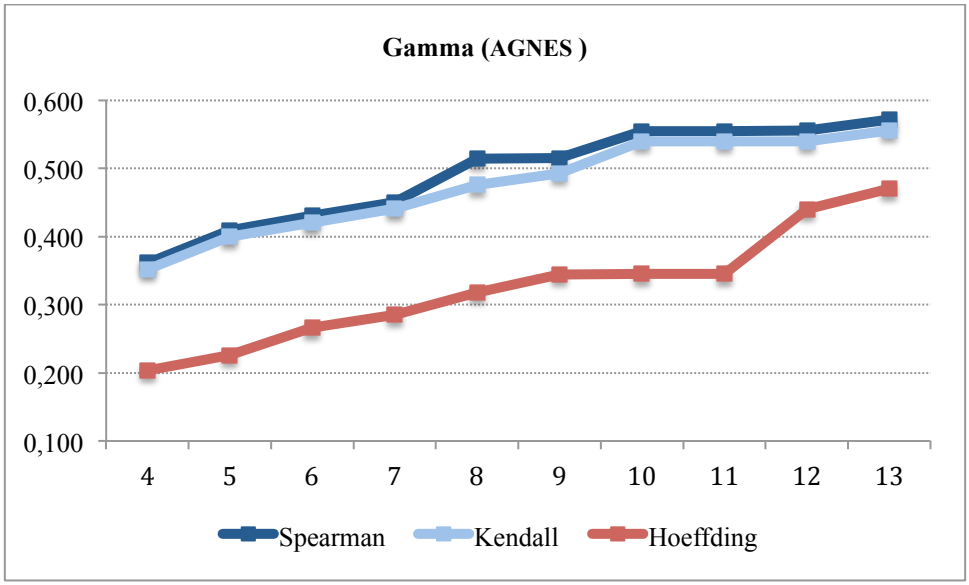

Figura 16 – Valores para o índice Gamma para o método AGNES aplicado junto aos métodos Spearman, Kendall e Hoeffding.

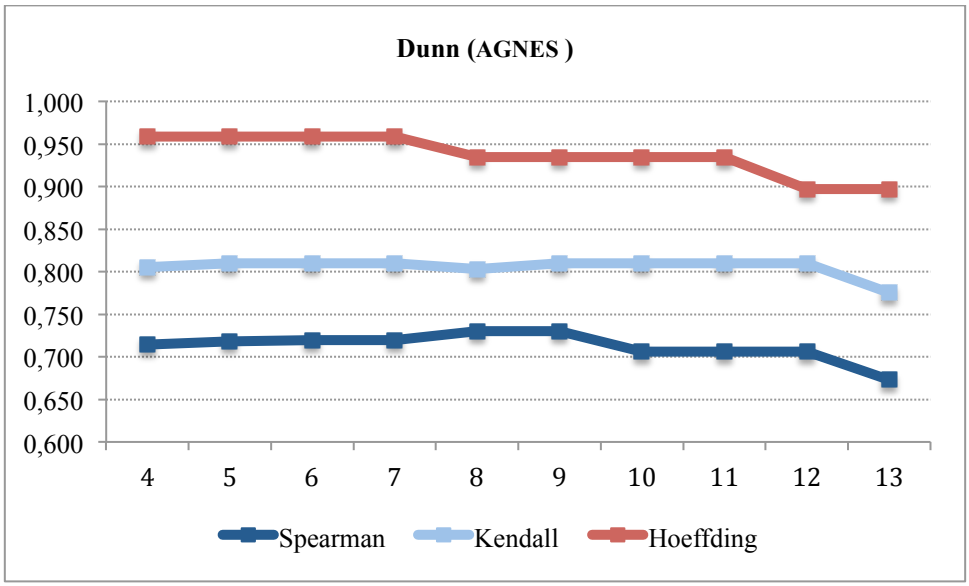

Figura 17 – Valores para o índice Dunn para o método AGNES aplicado junto aos métodos Spearman, Kendall e Hoeffding.

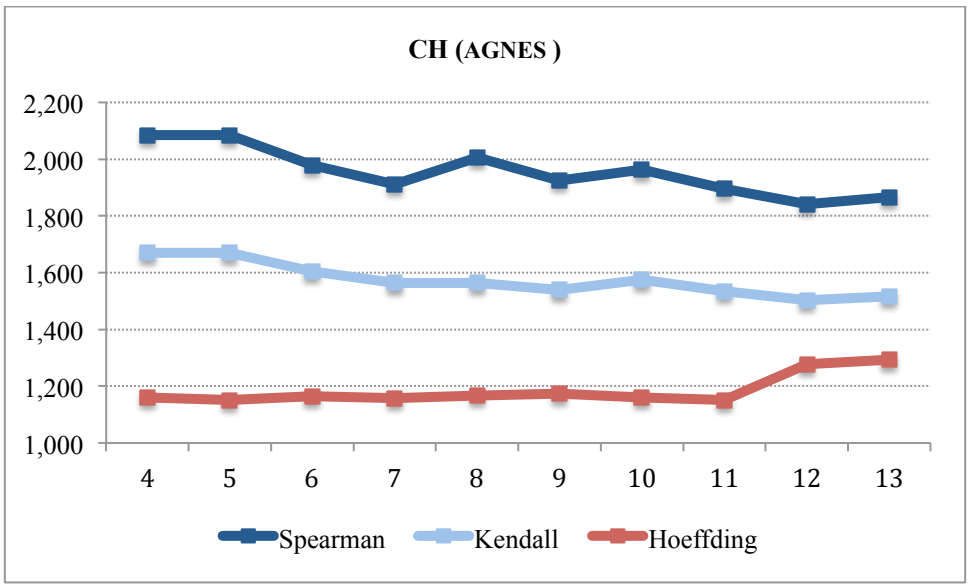

Figura 18 – Valores para o índice CH para o método AGNES aplicado junto aos métodos Spearman, Kendall e Hoeffding.

6.3.3.
Método DIANA

Este foi o método que pior aderiu à base de dados e apresentou resultados mais discrepantes em relação do que seria, *a priori*, esperado. Todos os índices de validação, para as três diferentes formas de cálculo de correlação, indicaram resultados que não permitem determinar um valor ótimo para o número k de *clusters*, isto é, apresentaram comportamento crescente sem a presença de máximos e mínimos locais bem definidos.

Diante deste fato, foram excluídos todas as opções baseadas no método DIANA testados.

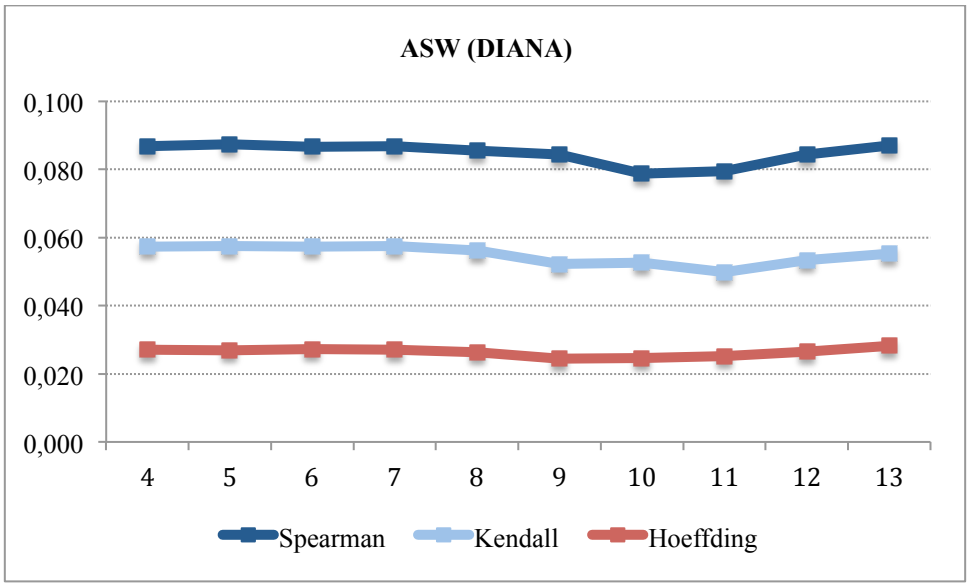

Figura 19 – Valores para o índice ASW para o método DIANA aplicado junto aos métodos Spearman, Kendall e Hoeffding.

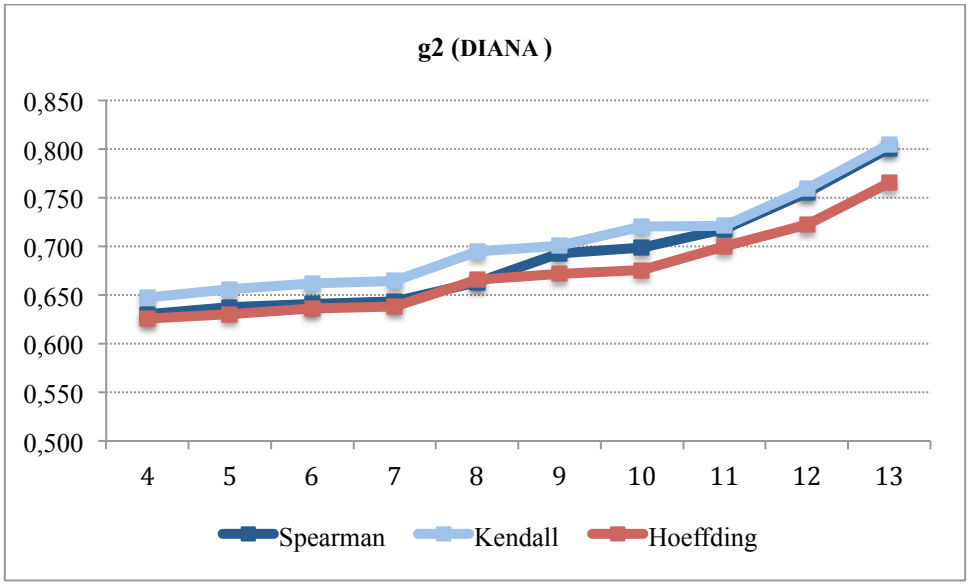

Figura 20 – Valores para o índice *g2* para o método DIANA aplicado junto aos métodos Spearman, Kendall e Hoeffding.

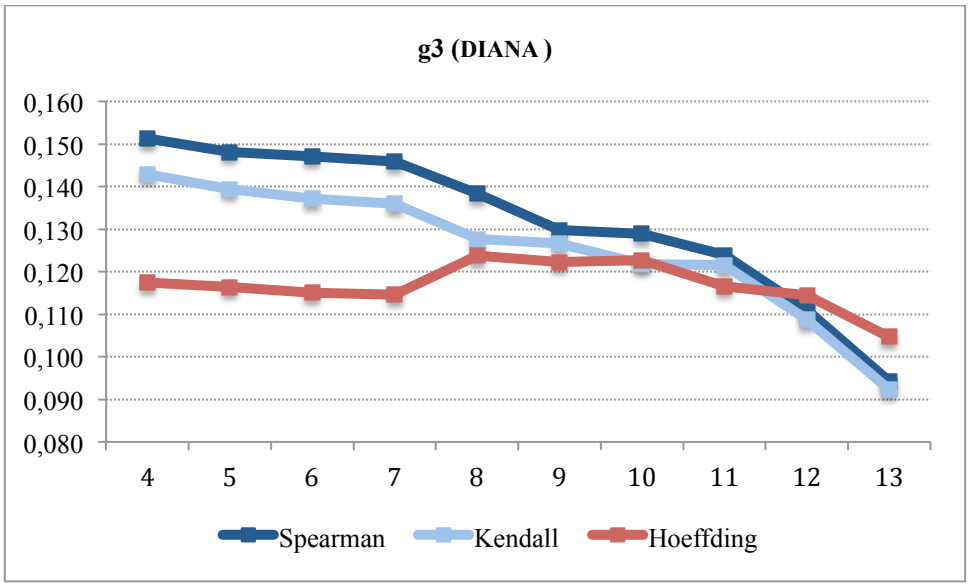

Figura 21 – Valores para o índice *g3* para o método DIANA aplicado junto aos métodos Spearman, Kendall e Hoeffding.

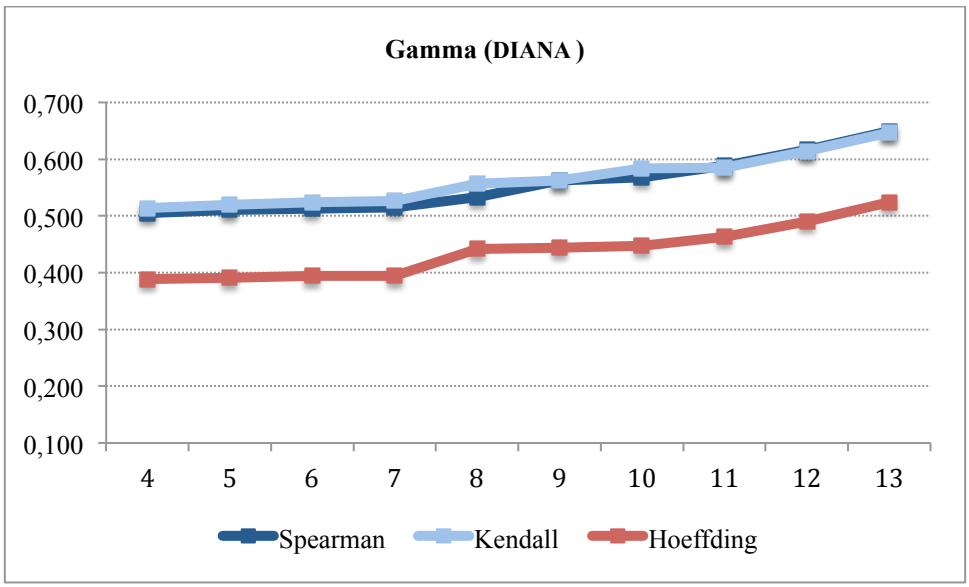

Figura 22 – Valores para o índice Gamma para o método DIANA aplicado junto aos métodos Spearman, Kendall e Hoeffding.

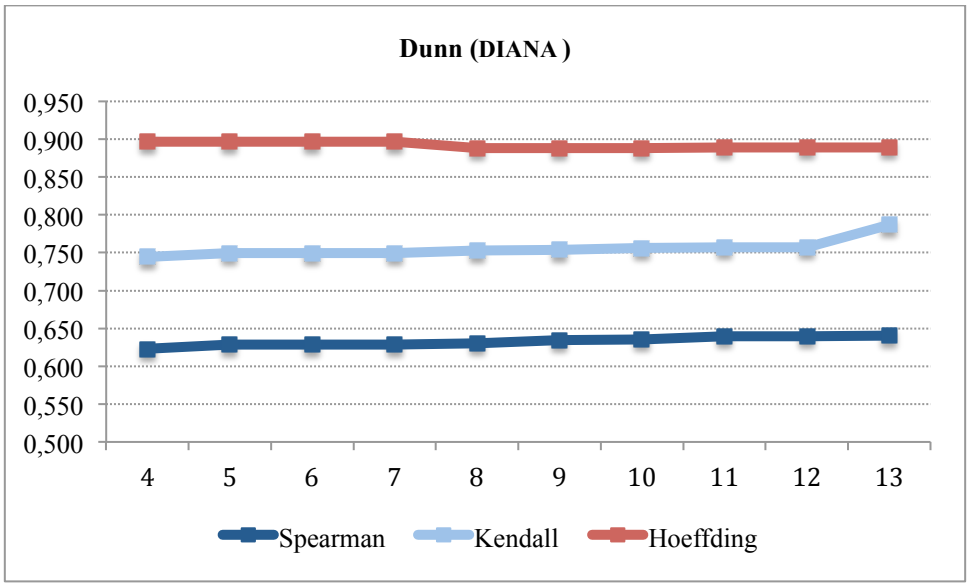

Figura 23 – Valores para o índice Dunn para o método DIANA aplicado junto aos métodos Spearman, Kendall e Hoeffding.

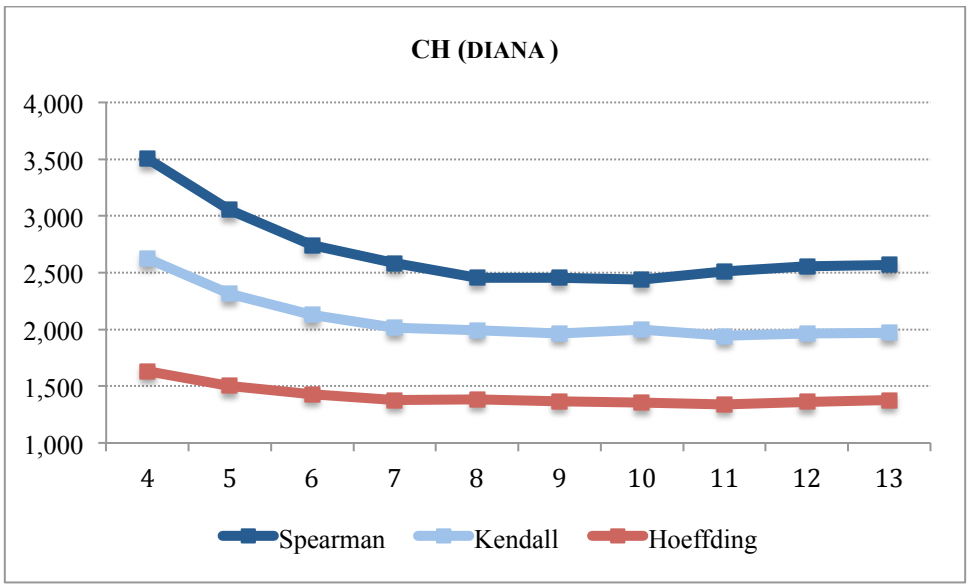

Figura 24 – Valores para o índice CH para o método DIANA aplicado junto aos métodos Spearman, Kendall e Hoeffding.

6.4.
Avaliação dos Métodos

De acordo com os resultados apresentados anteriormente, pode-se afirmar que apenas três dos nove diferentes métodos testados estariam aptos a corretamente modelar a base de dados estudada. São eles o método PAM quando aplicado junto ao método Spearman e o método AGNES quando aplicado junto com métodos Spearman e Kendall. Cabe, portanto, dar continuidade a análise, ao verificar se os *clusters* efetivamente formados por tais métodos condizem com as características reais das ações analisadas, isto é, verificar se existe sentido na combinação de objetos presentes em cada um dos *clusters.* Mais precisamente, dado que o presente trabalho busca comprovar uma hipótese associada aos setores econômicos a que pertencem as empresas associadas às ações, este "sentido na combinação" envolve a garantia de que os agrupamentos formados satisfazem tal hipótese.

Ao observarmos detalhadamente os resultados associados aos dois casos de aplicação do método AGNES, foi possível rapidamente excluí-los da análise. Apesar de os índices de validação apontarem para um valor ótimo de $k = 10$, quando aplicam-se os métodos para tal valor, encontram-se grupos que não apresentam relação sequer próxima da realidade do mercado. Verifica-se a criação de um grande *cluster* contendo 70 das 86 empresas estudadas, enquanto os demais 9 conjuntos contêm, no máximo, 3 elementos, sendo que, em 5 deles, há a presença de apenas um objeto. É claro que, na grande maioria dos casos, empresas do mesmo setor foram consideradas num mesmo cluster, porém não houve diferenciação entre os setores, comportamento evidenciado por este grande agrupamento contendo quase a totalidade das ações. Em suma, as empresas foram agrupadas da seguinte forma para ambos os casos:

- Cluster 1 → GETI4 RADL3, GUAR3.

- Cluster 2 → TRPL4, AMBV4, CRUZ3, NATU3, BTOW3, LAME4, LREN3, BISA3, CYRE3, GFSA3, MRVE3, PDGR3, RSID3, EVEN3, BVMF3, BBDC4, ITUB4, BBAS3, CSNA3, USIM5, VALE5, PETR4, LLXL3, GOLL4, JBSS3, OGXP3, DTEX3, RENT3, SUZB5, CSAN3, MAGG3, TCSA3, PCAR4, HYPE3,

VAGR3, AMAR3, MYPK3, POMO4, RAPT4, BRFS3, EMBR3, MFRG3, LPSB3, HBOR3, CMIG4, CPLE6, ELET6, CPFE3, ENBR3, CESP6, OIBR4, SBSP3, LIGT3, CCRO3, OHLB3, MPXE3, PSSA3, BRML3, MULT3, IGTA3, TIMP3, CSMG3, MDIA3, WEGE3, AMIL3, DASA3, ODPV3, ESTC3, KRTO11, STBP11.

- Cluster 3 → TBLE3.
- Cluster 4 → EZTC3, BBRK3, EQTL3.
- Cluster 5 → GGBR4, ELPL4, TOTS3.
- Cluster 6 → MMXM3.
- Cluster 7 → BRKM5, HGTX3.
- Cluster 8 → KLBN4.
- Cluster 9 → SULA11.
- Cluster 10 → VIVT4.

Diante deste resultado fortemente díspare, optou-se por focar a análise somente na terceira opção selecionada na etapa anterior: método PAM aplicado em conjunto ao método Spearman.

Neste caso, os resultados obtidos são bastante mais condizentes com uma expectativa inicial baseada numa percepção qualitativa das empresas e ações estudadas, apesar de certas disparidades serem observadas. A figura 25, apresentada a seguir, resume os *clusters* que foram formados, sendo que as cores utilizadas correspondem, cada uma, a um dos diferentes setores explicitados na classificação oficial da Bovespa.

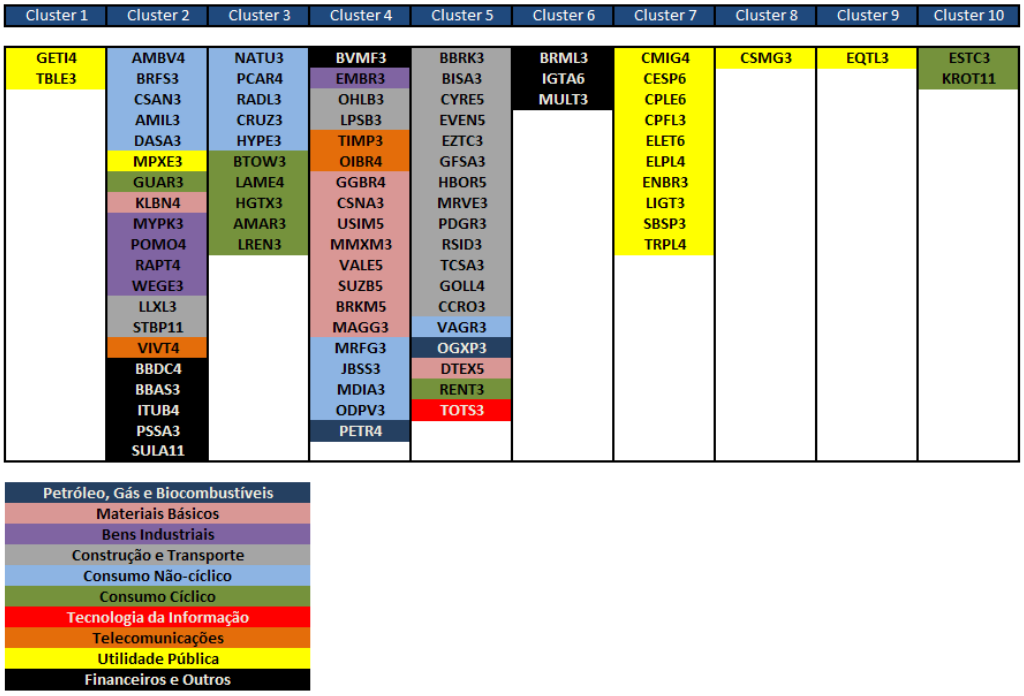

Figura 25 – Resumo do resultado da clusterização aplicando o método PAM junto ao método Spearman à base de dados analisada.

Algumas conclusões importantes podem ser consideradas ao observamos o agrupamento obtido:

• De maneira geral, empresas pertencentes ao mesmo segmento, não necessariamente ao mesmo setor econômico foram corretamente alocadas em um mesmo *cluster*. Por exemplo, são claros os seguintes agrupamentos: BRFS3 e CSAN3; BBDC4, BBAS3 e ITUB4; PSSA3 e SULA11; AMIL3 e DASA3; MYPK3, POMO4, RAPT4 e WEGE3; LLXL3 e STBP11; BTOW3, LAME4, HGTX3, HYPE3, AMAR3 e LREN3; NATU3, PCAR4, RADL3 e CRUZ3; TIMP3 e OIBR4; GGBR4, CSNA3, USIM5, MMXM3, VALE5, SUZB5, BRKM5 e MAGG3; MRFG3, JBSS3 e MDIA3; BBRK3, BISA3, CYRE3, EVEN3, EZTC3, GFSA3, HBOR3, MRVE3, PDGR3, RSID3 e TCSA3; CCRO3 e GOLL4, BRML3, IGTA3, e MULT3; CMIG4, CESP6, CPLE6, CPFE3, ELET6, ELPL4, ENBR3, LIGT3, SBSP3 e TRPL4; ESTC3 e KROT11.

• Segmentos que, de acordo com a classificação oficial, deveriam ser unidos por estarem associados a um mesmo setor econômico não necessariamente encontram-se desta forma. Por exemplo, o segmento de “Bancos” e o segmento de

"Seguradoras" não estão no mesmo *cluster* que o segmento de "Exploração de Imóveis". Ou então, os segmentos de "Serviços Educacionais" e "Comércio de Tecidos, Vestuário e Calçados" também não estão unidos, apesar de fazerem parte do setor de "Consumo Cíclico". É importante ressaltar, entretanto, que tais discrepâncias fazem sentido quando observamos mais atentamente as diferenças entre tais segmentos. A classificação que une Bancos e Seguradoras aos Shoppings, por exemplo, de fato não faz sentido empírico, logo seria de se esperar tal separação. O mesmo pode ser dito do segundo exemplo, apesar de as similaridades serem um pouco maiores neste caso.

• Por outro lado, os *clusters* formados são muito heterogêneos, isto é, diferentes segmentos econômicos que, na prática, não deveriam apresentar forte correlação, foram classificados de maneira análoga. Este fato é bastante evidenciado quando observam-se os grupos 2 a 5, os mais numerosos. Por exemplo, no segundo existem tanto empresas do setor financeiro (bancos e seguradoras), quanto dos setores de bens industriais, saúde e transporte. Mais importante, cabe destacar que não há evidências claras de que outras características poderiam estar determinando essa união entre setores diferentes. Por exemplo, poderia-se esperar que empresas ditas estáveis, de setores como Telecomunicações, Shoppings e Elétricas, estariam agrupadas entre si e separadas de outros como Bens Industriais e Commodities, ditos mais cíclicos. Entretanto, claramente este não é o caso.

• Verifica-se que 12 empresas podem ser consideradas erradamente classificadas, uma vez que estão alocadas em um certo *cluster* sem que haja outras empresas do mesmo segmento, ou ao menos, do mesmo setor. São elas: MPXE3, GUAR3, KLBN4 e VIVT4 no segundo grupo; BVMF3, EMBR3 e PETR4 no quarto; VAGR3, OGXP3, DTEX5, RENT3 e TOTS3 no quinto.

• O setor "Utilidade Pública", cujos segmentos envolvem o de "Energia Elétrica" e "Água e Saneamento", chama atenção, uma vez que cinco dos dez *clusters* possuem empresas deste setor. Entretanto, é possível afirmar que ele foi bastante bem representado. Dez das quinze empresas foram classificadas em um mesmo *cluster* sem que houvesse nenhuma empresa de outro setor presente. Duas empresas – Tractebel (TBLE3) e AES Tietê (GETI4) formaram um grupo

separado que pode, facilmente, ser entendido: são as duas únicas empresas que atuam somente no segmento de geração de energia (sem atividades de distribuição ou transmissão) e que são controladas por entidades privadas, ao contrário da Cesp (CESP6), que apesar de também uma geradora pura, é controlada pelo Estado de São Paulo. A MPXE3, ao longo de grande parte de sua existência na bolsa brasileira, foi uma empresa com ativos não apenas de geração de energia, mas também de mineração de carvão. Desta forma, pode-se dizer que um comportamento diferenciado em relação às demais empresas do mesmo setor não é surpreendente. As demais, Equatorial (EQTL3) e Copasa (CSMG3) são *outliers* no resultado, pois deveriam estar presentes no sétimo *cluster*.

Apesar de mais satisfatório, pode-se considerar que o resultado obtido após a aplicação do melhor método selecionado encontra-se ainda distante do que poderia ser esperado, quer seja comparando-se com evidências empíricas, quer seja comparando com a classificação oficial da Bovespa. Isso porque, conforme acima destacado, mesmo que as empresas tenham sido agrupadas de maneira relativamente precisa, houve uma união bastante heterogênea entre os grupos formados.

7.
Conclusão

A presente dissertação buscou avaliar uma evidência, comumente verificada nos mercados acionários, de que ações cujos preços de fechamento variam de forma semelhante pertencem a um mesmo setor econômico. Mais precisamente, esta premissa foi avaliada para um conjunto de ações negociadas na Bolsa de Valores de São Paulo. Para isso, foi considerada a técnica de clusterização, a qual tem como objetivo central encontrar em bases de dados grupos tais que os elementos pertencentes a cada um são mais semelhantes entre si do que em relação a dados classificados em um grupo distinto. Além disso, para avaliar esta similaridade entre os objetos, foram consideradas três diferentes maneiras de cálculo de correlação, as quais foram posteriormente empregadas da determinação da chamada matriz de dissimilaridade. São elas: Spearman, Kendall e Hoeffding.

No total, foram estudados nove diferentes métodos, uma vez que cada uma das três técnicas de cálculo de correlação foi empregada junto a três distintos métodos de clusterização, nomeadamente PAM, AGNES e DIANA. Diante desta grande quantidade de possibilidades, foi usada uma técnica de avaliação que consiste na verificação de diferentes índices de validação de clusterizações. Tais índices, não apenas indicam a qualidade dos métodos, como também permitem determinar o melhor valor para o número de *clusters* a ser formado.

A Bolsa de Valores disponibiliza uma classificação setorial teórica de todas as empresas listadas. Portanto, foi possível comparar os resultados obtidos a partir do emprego dos métodos com um resultado teoricamente esperado.

Após este processo, somente um dos métodos apresentou resultado minimamente satisfatório: o método PAM empregado junto ao método Spearman. Neste caso, foram encontrados dez grupos, mesmo número que seria esperado de acordo com a classificação oficial da Bolsa. Entretanto, algumas diferenças importantes foram encontradas.

A primeira delas mostra que empresas de um mesmo segmento econômico, e não necessariamente do mesmo setor, foram classificadas em grupos iguais.

Alguns exemplos desse comportamento ajudam a concluir que tal discrepância deveria de fato se encontrada. Em outras palavras, a classificação setorial teórica não contempla certas diferenças entre as empresas que são bastante claras quando olhamos uma classificação mais precisa, segmento a segmento. As empresas donas de shoppings centers, por exemplo, que formaram um *cluster* único bastante bem definido, deveriam estar agrupadas junto com Bancos e Seguradoras se levada em consideração a classificação oficial.

A segunda diz respeito a heterogeneidade dos *clusters* formados. Apesar de, como supracitado, o agrupamento ter respeitado semelhanças mais precisas do que somente o setor econômico a que pertencem as empresas, houve uma clara junção de segmentos que, *a priori*, não deveriam ter comportamento análogo. Este fato fica claro ao observarmos quatro dos dez *clusters* formados, os quais concentram a grande maioria das empresas.

Um terceiro aspecto que deve ser citado em relação aos resultados obtidos é que 12 das 86 empresas avaliadas podem ser considerada incorretamente classificadas se, claro, observarmos apenas a questão setorial estudada nesta tese. Chega-se a essa conclusão uma vez que tais empresas foram alocadas a um certo grupo sem que nele houvesse pelo menos uma outra empresa do mesmo setor.

Em suma, apesar de os resultados encontrados a partir da aplicação do método PAM de clusterização junto ao método Spearman de cálculo de correlação apresentarem uma boa aderência ao resultado teoricamente esperado e, até, a certas variações que poderiam ser esperadas dadas as imperfeições de tal classificação, pode-se afirmar que haveria espaço para um resultado ainda mais próximo do observado no mercado real. Algumas hipóteses podem ser levantadas para justificar tal comportamento.

A primeira delas diz respeito ao tamanho do mercado acionário brasileiro. Se comparado a outros mercados mais desenvolvidos, tal como o norte-americano, o brasileiro é bastante pequeno, com um número reduzido de empresas listadas. Tomemos como exemplo os principais índices destes dois mercados. Enquanto nos Estados Unidos destaca-se o S&P500, composto pelas 500 principais companhias, no Brasil o Ibovespa comportam um pouco mais de 60 empresas e o IbrX, 100. Desta forma, a quantidade de ações representantes de cada um dos setores e segmentos da economia é pequena, criando um reduzido espaço amostral para avaliação.

A segunda está, na verdade, associada a primeira. Além de ainda pequeno, o mercado acionário brasileiro é composto por muitas empresas que recentemente fizeram sua estréia. Desta forma, são poucas as empresas com um histórico de dias de negociação mínimo para uma avaliação mais precisa. O conjunto estudado neste trabalho, por exemplo, contou com 86 empresas, apesar de o índice IbrX usado como base ser composto por 100. Foi preciso alterar, retirando e acrescentando certas empresas, para formar uma base de dados mínima. Vale considerar ainda que questões de liquidez impedem a utilização de certas ações neste tipo de estudo, isto é, apesar de presentes na Bolsa há bastante tempo, certas ações não são negociadas diariamente, tornando suas cotações mais expostas especulações e menos representativas do valor real das empresas.

Como proposta de sequencia do estudo realizado nesta dissertação, é possível considerar outros métodos de clusterização para avaliação da base de dados. Além disso, pode-se realizar a aplicação dos métodos de clusterização considerando-se outras estatísticas obtidas a partir da base de dados que não o retorno diário das ações, tais quais a média, variância, volatilidade dentre outras. Outras possibilidades seriam validar os métodos utilizando também critérios externos de validação ou realizar o mesmo estudo de maneira mais restrita, com um número menor de empresas, porém com um histórico de negociação maior.

Referências Bibliográficas

[1] KAUFMAN, L.; ROUSSEUW, P.. **Finding Groups in Data: An Introduction to Cluster Analysis**. Jonh Wiley ans Sons Inc, 1990.

[2] HOLLANDER, M.; WOLFE, D.. **Nonparametric Statistical Methods**. Jonh Wiley ans Sons Inc, 1999.

[3] KRUSKAL, W.. **Ordinal Measures of Association**. Journal of The American Statistical Association, 284:814-861, 1958.

[4] HOEFFDING, W.. **A non-parametric test of independence**. The Annals of Mathematical Statistics, 19:546-557, 1948.

[5] CALINSKI, T.; HARABASZ, J.. **A Dendrite Method For Cluster Analysis**. Communications in Statistics, 3:1-27, 1974.

[6] HENNIG, C.; LIAO, T.. **Comparing Latent Class and Dissimilarity Based Clustering for Mixed Type of Variables With Application do Social Stratification**. Departamento de Ciência Estatística, UCL. http://www.ucl.ac.uk/Stats/research/reports/psfiles/rr308.pdf.

[7] DAVIES, D. L.; BOULDIN, D. W.. **A Cluster Separation Measure**. IEEE T, 1:224-227, 1979.

[8] GORDON, A. **Classification**. Chapman and Hall, 2ª edição, 1999.

[9] MAULIK, U.; BANDYOPADHYAY, S.. **Performance Evaluation of Some Clustering Algorithms and Validity Indices**. IEEE T, 24:1650-1654, 2002.

[10] YEUNG, K.; HAYNOR, D.; RUZZO, W.. **Validating clustering for gene expression data**. Bioinformatics, 17:309-318, 2001.

[11] MILLIGAN, G. W.; COOPER, M. C.. **An Examination of Procedures for Determining the Number of Clusters in a Data set**. Psychometrika, 50:159-179, 1985.

[12] TIBSHIRANI, R.; WALTHER, G.; HASTIE, T.. **Estimating the Number of Clusters in a Dataset via the Gap Statistic**. Journal of the Royal Statistical Society, Series B, 63:411-423, 2001.

[13] MILLIGAN, G. W.; SOON, S. C.; SOKOL, L. M.. **The Effect of Cluster Size, Dimensionality and the Number of Clusters on Recovery of True Cluster Structure**. IEEE T, 5:40-47, 1983.

[14] MUSETTI, A.. **Clustering methods for financial time series**. Swiss Federal Institute of Technology, 2012.

[15] JAIN, A. K.; MURTY, M. N.; FLYNN, P. J.. **Data Clustering: A review**. ACM Computer, 31:264-323, 1999.

[16] BEZDEK, J. C.; PAL, N. R.. **Some new indexes of cluster validity**. IEEE T, 28:301-315, 1998.