

Luiz Paulo Fávero
Patrícia Belfiore
Fabiana Lopes da Silva
Betty Lilian Chan

análise de dados

MODELAGEM MULTIVARIADA
PARA TOMADA DE DECISÕES



Análise de Conglomerados

Todos somos iguais,
mas alguns são mais iguais que outros.

GEORGE ORWELL

AO FINAL DESTES CAPÍTULOS, VOCÊ SERÁ CAPAZ DE:

- Identificar situações de pesquisa que sejam apropriadas à utilização da técnica de análise de conglomerados.
- Explicar as principais medidas de similaridade ou distância utilizadas em análise de conglomerados.
- Entender as principais diferenças entre os procedimentos hierárquicos e não hierárquicos.
- Interpretar os resultados apresentados pela técnica.
- Determinar o número de agrupamentos mais adequado para solucionar o problema de pesquisa.
- Identificar elementos que compõem cada conglomerado.

6.1. APRESENTAÇÃO DO CAPÍTULO

A técnica de análise de conglomerados (*cluster analysis*), também conhecida como análise de agrupamentos, é uma técnica estatística de interdependência que permite agrupar casos ou variáveis em grupos homogêneos em função do grau de similaridades entre os indivíduos, a partir de variáveis predeterminadas.

A idéia principal da técnica é agrupar objetos com base em suas próprias características, buscando, assim, a estrutura “natural” desses objetos.

A análise de conglomerados pode ser aplicada em todas as áreas do conhecimento humano cujo objetivo seja segmentar as observações em grupos homogêneos internamente e heterogêneos entre si. Um exemplo usual é a aplicação desta técnica na área de marketing. Por exemplo, a fim de direcionar seus esforços de vendas, uma empresa pode segmentar sua base de clientes de acordo com os perfis dos consumidores, considerando aspectos econômicos, sociais e comportamentais e, assim, estabelecer estratégias mais adequadas a cada segmento.

Os três maiores objetivos deste capítulo são (1) introduzir os aspectos conceituais da análise de conglomerados hierárquicos e não hierárquicos, bem como as medidas de similaridade; (2) apresentar a aplicação da técnica; e (3) discutir os resultados obtidos.



6.2. UMA INTRODUÇÃO À ANÁLISE DE CONGLOMERADOS

A análise de conglomerados, ou *cluster analysis*, é uma técnica de interdependência que busca agrupar os elementos conforme sua estrutura “natural”.

É uma técnica que visa segregar elementos ou variáveis em grupos homogêneos internamente, heterogêneos entre si e mutuamente exclusivos, a partir de determinados parâmetros conforme uma medida de similaridade ou de distância. Neste sentido, tem por objetivo principal definir a estrutura dos dados de maneira a alocar as observações mais parecidas no mesmo grupo.

Esta análise busca identificar elementos semelhantes com base em suas características. Assim, a técnica identifica grupos de objetos, de modo que cada objeto é semelhante aos demais no agrupamento, buscando, assim, maximizar a homogeneidade dos objetos dentro dos grupos, e maximizar a heterogeneidade entre os demais grupos (HAIR, ANDERSON, TATHAM e BLACK, 2005).

Segundo Bailey (1975), a análise de conglomerados teve origem na psicologia, com Zubin (1938) e Tryon (1939), sendo este o primeiro a usar o termo *cluster analysis*, e na antropologia com Driver e Kroeber (1932). Mas somente após a década de 1950, com o desenvolvimento da computação, é que a técnica de fato ganhou destaque, tendo a maior contribuição dada por Sokal e Sneath (1963), em seu livro *Principles of Numerical Taxonomy*, que a utilizaram para a classificação biológica, visando identificar e agrupar organismos com características semelhantes. Após o agrupamento, tornava-se possível analisar as características de cada conglomerado e averiguar se eram ou não de espécies diferentes.

Depois desses estudos, houve um aumento significativo de publicações acerca da técnica em diversas áreas do conhecimento: economia, medicina, geografia, biologia, sociologia, psicologia, administração, entre outras. A justificativa para o crescimento, além do desenvolvimento de computadores para agilizar os cálculos, foi “a importância da classificação como método científico” (REIS, 2001).

A análise de conglomerado pode ser utilizada em diversas situações de pesquisa; por exemplo, quando:

- um gestor está interessado em identificar grupos de investimentos de acordo com perfis de risco;
- um diretor de marketing busca identificar segmentos homogêneos de consumidores, a fim de estabelecer programas de marketing específicos para cada público;
- um educador visa identificar grupos de alunos mais propensos à evasão escolar;
- uma seguradora busca identificar grupos de segurados de menor risco;
- um pesquisador busca segmentar empresas com base em indicadores financeiros (rentabilidade, liquidez, margem, entre outros).

A análise de conglomerados é uma importante técnica exploratória, uma vez que, ao estudar uma estrutura natural de grupos, possibilita avaliar a dimensionalidade dos dados, identificar *outliers* e levantar hipóteses relacionadas à estrutura (associações) dos objetos (JOHNSON e WICHERN, 2007).

Em análise de conglomerados, o agrupamento de variáveis se assemelha à análise fatorial, pois ambas as técnicas visam identificar grupos de variáveis relacionadas. Entretanto, a análise fatorial se mostra mais robusta para o agrupamento de variáveis em detrimento do agrupamento de observações, foco da análise de *cluster*.

A variável estatística de agrupamento pode ser definida como o conjunto de atributos ou características das observações que servirão de base para a determinação da similaridade entre elas. Cabe destacar que, na análise de conglomerados, diferentemente das demais técnicas multivariadas, a variável estatística de agrupamento é o conjunto de variáveis selecionadas pelo pesquisador, não sendo estimada empiricamente.

Cabe ressaltar que é uma técnica descritiva, sem base teórica e não-inferencial, isto é, embora apresente fortes propriedades matemáticas, não possui fundamentos estatísticos, não sendo apropriada para in-

ferências de características da população a partir de dados amostrais (HAIR, ANDERSON, TATHAM e BLACK, 2005).

Basicamente, a análise de conglomerados pode ser dividida nas seguintes etapas:

- análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização);
- seleção da medida de distância ou semelhança entre cada par de objetos;
- seleção do algoritmo de agrupamento: método hierárquico ou não-hierárquico;
- escolha da quantidade de agrupamentos formados;
- interpretação e validação dos agrupamentos.

Para fins didáticos, os tópicos a serem abordados no presente capítulo seguirão as etapas descritas anteriormente, a fim de facilitar para o leitor a visualização e o acompanhamento lógico da técnica.

Contribuições significativas sobre o assunto podem ser encontradas em Sokal e Sneath (1963), Sokal e Sneath (1973), Aldenderfer e Blashfield (1985), Bailey (1975), Frank e Green (1968), Milligan e Cooper (1987), Reis (2001), Mingoti (2005), Pestana e Gageiro (2005), Sharma (2006), Hair, Anderson, Tatham e Black (2005), Maroco (2007) e Johnson e Wichern (2007).

6.3. ANÁLISE DAS VARIÁVEIS E DOS OBJETOS A SEREM AGRUPADOS

6.3.1. Identificação de *Outliers* e Seleção de Variáveis

A seleção das variáveis deve ser feita com extremo cuidado, pois os grupos a serem formados refletirão a estrutura inerente das variáveis escolhidas, tendo-se em vista que serão utilizadas para determinar a medida de similaridade a qual corresponde ao critério de segregação dos grupos.

Cabe observar que a técnica não distingue se as variáveis são ou não relevantes para o estudo, ficando essa tarefa a cargo do pesquisador. Neste sentido, a inclusão de variáveis não representativas ou a presença de multicolinearidade podem distorcer os resultados do estudo. A questão da multicolinearidade interfere na ponderação das medidas de similaridade. Uma forma de mitigar seus efeitos é a adoção da medida de distância de Mahalanobis (D^2), a qual não apenas padroniza os dados, estabelecendo uma escala em termos de desvios padrão, como também soma a variância-covariância acumulada dentro dos grupos, ajustando as intercorrelações entre as variáveis, sendo, portanto, uma medida de distância comparável ao R^2 da regressão (HAIR, ANDERSON, TATHAM e BLACK, 2005).

Outro ponto importante a ressaltar é que esta técnica é altamente sensível à inclusão de variáveis com comportamentos atípicos, isto é, com a presença de *outliers*. Os *outliers* podem ser definidos como observações que fogem do padrão esperado em cada variável, ou seja, referem-se a observações com características muito destoantes dos demais membros da população, podendo prejudicar a qualidade dos resultados. Assim, antes de efetuar a análise de conglomerados, é recomendável verificar a existência de *outliers*, cabendo ao pesquisador decidir se devem continuar ou não na base de dados. No entanto, cabe esclarecer que é comum que indivíduos atípicos formem grupos isolados, o que, por vezes, é de interesse do próprio pesquisador esta constatação e, portanto, não necessariamente tais observações devem ser eliminadas da amostra.

6.3.2. Padronização de Variáveis

Um aspecto importante que deve ser considerado pelo pesquisador ao realizar uma análise de conglomerados é que a utilização de variáveis com medidas/escalas diferentes pode distorcer a estrutura do agrupamento.



A maior parte das medidas de distância sofre influência das diferentes escalas ou magnitudes das variáveis, sendo esse problema contornado com a padronização das variáveis.

Ou seja, a variável que apresentasse maior dispersão teria um peso mais elevado no cálculo das medidas de distância do que as demais. A padronização faz com que seja atribuído o mesmo peso para cada variável.

A forma mais utilizada para padronização dos dados consiste em transformar cada variável em escore padrão (Z scores), permitindo que seja eliminado o viés decorrente das diferenças de escalas. O método Z scores padroniza cada variável (x) de maneira a apresentar média zero e desvio padrão 1, sendo calculado da seguinte maneira:

$$Z = \frac{(x - \text{média})}{\text{desvio - padrão}} \quad (6.1)$$

Entretanto, deve-se destacar que a padronização de variáveis deve ser utilizada com cuidado, pois se existir alguma relação natural refletida nas escalas das variáveis, a padronização pode não ser adequada (HAIR, ANDERSON, TATHAM e BLACK, 2005).

Além do método de padronização Z scores, também são usuais:

- Método Range -1 a 1: faz com que a variável padronizada tenha amplitude 1.

$$\frac{x}{\text{amplitude}} \quad (6.2)$$

- Método Range 0 a 1: faz com que a variável apresente variação de 0 a 1.

$$\frac{x - \text{mínimo}}{\text{amplitude}} \quad (6.3)$$

- Método de máxima amplitude: confere à variável o valor máximo de 1.

$$\frac{x}{\text{máximo}} \quad (6.4)$$

- Método de média 1: como o próprio nome sugere, transforma a variável de maneira que apresente média 1.

$$\frac{x}{\text{média}} \quad (6.5)$$

- Método de desvio padrão 1: como o próprio nome sugere, transforma a variável de maneira que apresente desvio padrão 1.

$$\frac{x}{\text{desvio - padrão}} \quad (6.6)$$

6.4. MEDIDAS DE SIMILARIDADE OU DISTÂNCIA (DISSIMILARIDADE)

Após a seleção das variáveis e a verificação da necessidade ou não de padronizá-las, bem como avaliar a manutenção ou exclusão de *outliers*, a próxima etapa com a qual o pesquisador vai deparar relaciona-se à escolha da medida de similaridade que será utilizada no estudo.

O conceito de similaridade em análise de conglomerados é de vital importância, uma vez que a identificação de agrupamentos de sujeitos ou variáveis só é possível com a adoção de alguma medida de semelhança que permita a comparação objetiva entre os sujeitos.

Na análise de conglomerados, as observações são agrupadas segundo algum tipo de métrica de distância, e as variáveis são agrupadas conforme medidas de correlação ou associação.

De acordo com Tversky *apud* Reis (2001), “a análise teórica das relações de semelhança tem sido dominada pelos modelos geométricos. Esses modelos representam os objetos como pontos em um espaço de coordenadas, de forma que as dissimilaridades observadas entre objetos correspondam às distâncias métricas entre os respectivos pontos”.

Existem diversas maneiras de medir a similaridade ou a distância entre os objetos. O pacote estatístico SPSS fornece uma ampla gama de possibilidades.

Para ilustrar o conceito geométrico na concepção dos *clusters*, vamos nos valer de um pequeno exemplo: Imagine que um pesquisador pretende agrupar seis empresas do setor de comércio varejista publicadas pela Edição de Melhores e Maiores da Editora Abril do ano de 2007. Para tanto, utilizaremos as informações de vendas (em US\$ milhões) e o número de empregados, conforme apresentado na Tabela 6.1.

Tabela 6.1: Matriz de Dados de Seis Empresas Seleccionadas

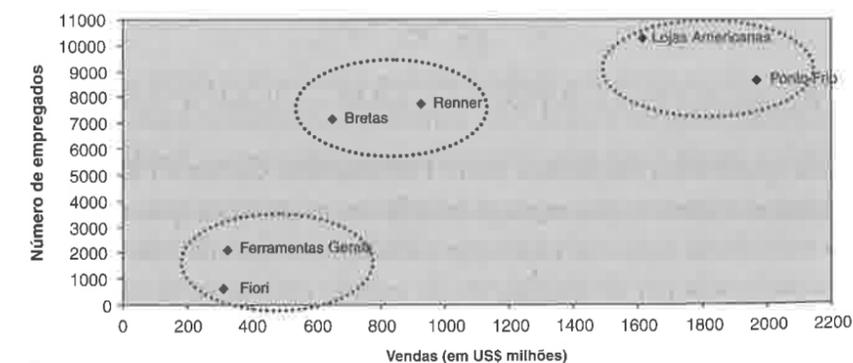
Empresas	Vendas (em US\$ milhões)	Número de empregados
Ferramentas Gerais (1)	327,5	2.150
Fiori (2)	312,2	661
Bretas Supermercados (3)	652,6	7.200
Renner (4)	929	7.764
Lojas Americanas (5)	1.613,5	10.281
Ponto Frio (6)	1.971	8.672

Seria possível separar a amostra de empresas em grupos similares em termos do porte, representado pelas variáveis faturamento e número de empregados?

Para fins ilustrativos, estamos trabalhando com apenas duas variáveis, sendo cada observação indicada como um ponto em um espaço bidimensional. De forma análoga, hipoteticamente, poderíamos representar cada observação como um ponto em um espaço p -dimensional, sendo p o número de variáveis.

Os valores foram plotados no gráfico de dispersão mostrado a seguir, em que é possível visualizar três grupos, a saber: (a) o grupo 1, formado pelas empresas Ferramentas Gerais e Fiori; (b) o grupo 2, formado pelas empresas Bretas Supermercados e Renner; e (c) o grupo 3, formado pelas Lojas Americanas e Ponto Frio. Com base no Gráfico 6.1, é possível verificar como as empresas se relacionam em termos de similaridade quanto às variáveis de faturamento e número de empregados. Assim, o grupo 1 possui as empresas de pequeno porte, o grupo 2 as de médio porte e o grupo 3 as de maior porte.

Gráfico 6.1: Porte das Empresas segundo Vendas e Número de Empregados



O que acabamos de fazer foi alocar observações em grupos conforme o grau de similaridade. Geometricamente, o que fizemos foi representar os dados em um espaço bidimensional.

Para um grande número de variáveis, já não seria possível identificar visualmente os grupos, dada a limitação gráfica de espaço tridimensional, sendo necessário utilizar outros critérios de aglomeração, tais como medidas de distância e outras medidas de similaridade.

De modo geral, as medidas de similaridade ou distância podem ser classificadas em três tipos:

- medidas de distância;
- medidas correlacionais;
- medidas de associação.

A escolha das medidas de similaridade implica o conhecimento da natureza das variáveis (discreta, contínua, binária) e da escala de medida (nominal, ordinal, intervalar ou razão).

Além disso, tanto as medidas de distância quanto as medidas correlacionais requerem dados métricos, enquanto as medidas de associação são destinadas ao tratamento de dados não métricos (nominal ou ordinal).

Retomando o nosso exemplo das seis empresas, quando agrupamos Lojas Americanas e Ponto Frio, nós implicitamente utilizamos a distância entre os pontos (empresas) como uma medida de similaridade entre elas. Vale destacar que existe uma variedade de medidas de similaridade que serão abordadas no tópico seguinte. Dando continuidade ao exemplo, adotaremos a distância quadrática euclidiana como medida de similaridade.

Mas, antes de calcularmos as distâncias, vamos padronizar as variáveis para evitar o efeito das diferenças de escalas, uma vez que as vendas, por exemplo, podem ser representadas em milhões de dólares, milhares de dólares ou em dólares, magnitudes não comparáveis ao número de empregados. Neste caso, a variável que apresentar maior dispersão teria um peso mais elevado no cálculo das medidas de distância do que a outra. A padronização faz com que seja atribuído o mesmo peso para cada variável.

A tabela a seguir mostra as variáveis já padronizadas pelo método Z scores.

Tabela 6.2: Matriz de Dados Padronizados de Seis Empresas Seleccionadas

Empresas	Vendas (em US\$ milhões)	Número de empregados
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Portanto, a distância quadrática euclidiana entre Ferramentas Gerais e Fiori é dada por:

$$d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2$$

$$d_{12}^2 = (-0,931 - (-0,953))^2 + (-1,038 - (-1,427))^2 = 0,152$$

Sendo d_{12}^2 a distância quadrática euclidiana entre Ferramentas Gerais e Fiori (observações 1 e 2) e p o número de variáveis (vendas e número de empregados). Note, neste caso, que, como $p = 2$ (duas variáveis), a distância quadrática euclidiana nada mais é do que a tradicional distância de Pitágoras ao quadrado.

Repetindo o cálculo para cada par de elementos, de acordo com as variáveis de estudo, chega-se à seguinte matriz de similaridade (Tabela 6.3).



Tabela 6.3: Matriz de Similaridade pela Distância Quadrática Euclidiana para Seis Empresas Seleccionadas

	Ferramentas Gerais (1)	Fiori (2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Supermercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

A matriz de similaridade mostra as distâncias de cada par de empresas (distância quadrática euclidiana). Com base na matriz, percebe-se que a menor distância entre empresas são Ferramentas Gerais e Fiori, o que denotará o primeiro grupo a ser formado. Obviamente, a diagonal principal da matriz de similaridade possui apenas valores iguais a zero.

Cabe ressaltar que as medidas mais comuns de similaridade são as de distância, em especial, a distância euclidiana simples e a quadrática, as quais serão detalhadas no próximo tópico.

6.4.1 Medidas de Distância

As medidas de distância são consideradas medidas de dissimilaridade, pois, quanto maiores os valores, menor é a semelhança entre os objetos, e vice-versa. As principais medidas de distância tratadas na literatura quando do estudo de análise de conglomerados são:

- a) Distância Euclidiana: a distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j) para todas as p variáveis.

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \tag{6.7}$$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2} \tag{6.8}$$

Em que x_{ik} é o valor da variável k referente à observação i e x_{jk} representa a variável k para a observação j . Nesta abordagem, quanto menor a distância, mais similares serão as observações.

- b) Distância Quadrática Euclidiana: a distância entre duas observações (i e j) corresponde à soma dos quadrados das diferenças entre i e j para todas as p variáveis.

$$d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2 \tag{6.9}$$

O pacote estatístico SPSS utiliza como padrão de distância entre as observações a distância quadrática euclidiana. Além disso, a distância quadrática euclidiana é recomendada para os métodos de agrupamento centróide e Ward, a serem discutidos adiante (HAIR, ANDERSON, TATHAM e BLACK, 2005).

- c) Distância de Minkowski: A distância euclidiana é um caso particular de uma distância mais geral, chamada de Minkowski, dada pela seguinte expressão:

$$d_{ij} = \left(\sum_{k=1}^p (|x_{ik} - x_{jk}|)^n \right)^{1/n} \tag{6.10}$$



Em que d_{ij} é a distância de Minkowski entre as observações i e j , p é o número de variáveis, e $n=1, 2, \dots, \infty$.

Se aplicarmos $n=2$ na formulação de Minkowski, chegaremos à distância euclidiana. Entretanto, para $n=1$, temos uma nova distância, denominada Distância City-Block, ou também chamada de *Manhattan Distance*, apresentada na sequência. Outros valores de n resultam em outros tipos de distância, entretanto, elas não são usadas com frequência (SHARMA, 1996).

- d) Distância Absoluta, Bloco, City-Block ou Manhattan: representa a soma das diferenças absolutas entre os valores das p variáveis para os dois casos.

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (6.11)$$

- e) Mahalanobis: é a distancia estatística entre dois indivíduos i e j , considerando a matriz de covariância para o cálculo das distâncias.

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)} \quad (6.12)$$

Em que S é a estimativa amostral da matriz de variância-covariância Σ dentro dos agrupamentos.

- f) Chebychev: diferença absoluta máxima entre todas as p variáveis entre duas observações.

$$d_{ij} = \max_k |x_{ik} - x_{jk}| \quad (6.13)$$

Ressalta-se que voltaremos a estudar essas distâncias no Capítulo 9 (Escalonamento Multidimensional).

6.4.2 Medidas Correlacionais

De acordo com Hair, Anderson, Tatham e Black (2005), "as medidas correlacionadas representam similaridade pela correspondência de padrões ao longo das características (X variáveis)".

Nas ciências sociais, a correlação de Pearson, dada pela Fórmula (6.14), dentre as medidas correlacionais, é a mais utilizada (REIS, 2001).

$$r_{ij} = \frac{\sum_{k=1}^p (x_{1k} - \bar{x}_i)(x_{1j} - \bar{x}_j)}{\sqrt{\sum_{k=1}^p (x_{1k} - \bar{x}_i)^2 \sum_{k=1}^p (x_{1j} - \bar{x}_j)^2}} \quad (6.14)$$

Sendo:

- x_{ik} = valor da variável k para a observação i ;
- x_{ij} = valor da variável k para a observação j ;
- \bar{x}_i = representa a média de todas as variáveis para o indivíduo i ;
- \bar{x}_j = representa a média de todas as variáveis para o indivíduo j ;
- p = representa o número de variáveis.

O valor do coeficiente varia entre -1 e 1 , em que o zero significa que não há associação. Nessa abordagem, quanto maiores as correlações, mais próximas estarão as observações.

Cabe lembrar que a medida de similaridade mais utilizada em análise de conglomerado são as medidas de distância, tendo-se em vista que as medidas correlacionadas não focam a magnitude dos objetos, mas a correlação entre seus perfis.

6.4.3 Medidas de Associação

As medidas de associação são utilizadas para representar a similaridade quando tratamos de variáveis nominais, baseando-se em tabelas de contingência.

A presença ou ausência de determinada característica pode ser descrita matematicamente pela introdução de variáveis binárias, que assumem o valor 1 se a característica estiver presente e zero pela ausência.

Assim, considere dois indivíduos i e j , que são caracterizados por p variáveis nominais dicotômicas (binárias), em que 1 significa presença da característica e zero a ausência. Com base nessas informações é possível construir a seguinte tabela de contingência:

Tabela 6.4: Tabela de Contingência

		Indivíduo j		Total
		1	0	
Indivíduo i	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$p = a+b+c+d$

Em que a representa o número de características presentes em ambos os indivíduos, b representa o número de características presentes no indivíduo i e ausente no j , c representa o número de características ausente no i e presente no j , e, finalmente, d representa a ausência simultânea de características em i e j .

Os coeficientes de emparelhamento simples são os mais utilizados e são definidos, de acordo com Sharma (1996), Reis (2001) e Maroco (2007), como:

$$S_{ij} = \frac{a+d}{a+b+c+d} \quad \text{ou} \quad d_{ij} = \frac{b+c}{a+b+c+d} \quad (6.15)$$

Sendo que:

S_{ij} (medida de semelhança) é a relação entre o número de características presentes e ausentes simultaneamente para os dois indivíduos e o número total de características;

d_{ij} (medida de distância) representa o coeficiente entre o número de características presentes em um indivíduo e ausentes no outro e o número total de características.

Os coeficientes de Jaccard para semelhança e distância entre dois indivíduos i e j são definidos, respectivamente, como:

$$S_{ij} = \frac{a}{a+b+c} \quad \text{ou} \quad d_{ij} = \frac{b+c}{a+b+c} \quad (6.16)$$

Outros coeficientes de semelhança, a partir de variáveis binárias, são apresentados na tabela a seguir:

Tabela 6.5: Coeficientes de Similaridade

Coeficiente	Justificativa
$\frac{2(a+d)}{2(a+d)+b+c}$	Peso duplo para presença e ausência em ambos os indivíduos.
$\frac{(a+d)}{a+d+2(b+c)}$	Peso duplo às situações de discordantes, inclusão das ausências simultâneas.
$\frac{2a}{2a+b+c}$	Peso duplo às presenças simultâneas, exclusão das ausências simultâneas.
$\frac{a}{a+2(b+c)}$	Peso duplo às situações discordantes, exclusão das ausências simultâneas.
$\frac{a}{b+c}$	Quociente entre presenças simultâneas e situações discordantes, exclusão das ausências simultâneas.

Fonte: Reis (2001) e Johnson e Wichern (2007).

Voltaremos a discutir esses coeficientes de similaridade no Capítulo 9 (Escalonamento Multidimensional).

6.5. OS MÉTODOS DE AGRUPAMENTO EM ANÁLISE DE CONGLOMERADOS

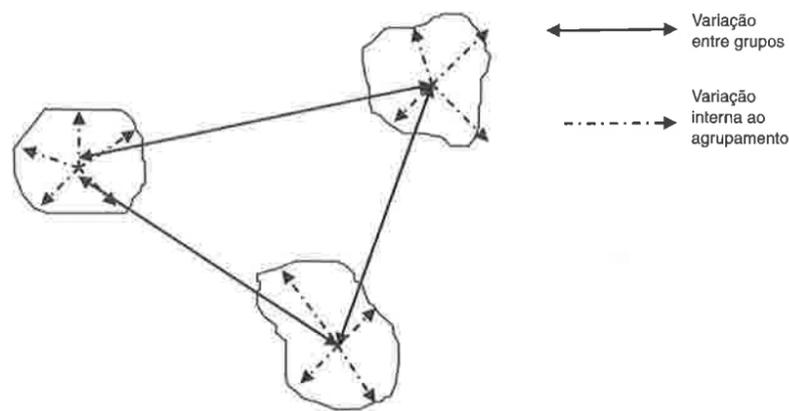
Uma vez selecionadas as variáveis do estudo e escolhida a medida de similaridade, é necessário determinar o algoritmo que fará o processo de agrupamento. Ou seja, a formação dos grupos decorre do critério de distância entre os vetores de dados e do método de agregação escolhido.

Nessa etapa, podemos nos perguntar **como usaremos a matriz de similaridade para formar grupos?**

Basicamente, há dois métodos de agrupamento: o hierárquico e o não hierárquico. Podemos afirmar que os diversos métodos visam responder, de maneira diferente, às seguintes questões:

- “Distância entre indivíduos do mesmo grupo e distância entre indivíduos de grupos diferentes.
- Dispersão dos indivíduos dentro do grupo.
- Densidade dos indivíduos dentro e fora dos grupos” (REIS, 2001).

Ou seja, o algoritmo tenta maximizar a diferença entre agrupamentos, conforme mostra a figura a seguir:



Fonte: Hair, Anderson, Tatham e Black (2005).

Figura 6.1: Diagrama de agrupamento.

Na sequência, vamos discutir os métodos hierárquicos e não hierárquicos, contrapondo suas utilizações, vantagens e desvantagens de cada um.

6.5.1. Análise de Conglomerados Hierárquicos

Nas técnicas hierárquicas, distinguem-se dois tipos de procedimentos de agrupamento: os métodos aglomerativos e os divisivos.

No método aglomerativo, cada sujeito começa com seu próprio agrupamento e, a partir deste ponto, novos agrupamentos são realizados por similaridade, ou seja, no início cada indivíduo representa um grupo. Na etapa seguinte, os dois indivíduos mais similares (próximos) são agrupados primeiramente e, nas etapas subsequentes, vão se fundindo com os demais grupos de acordo com a proximidade. Assim, em cada etapa reduz-se o número de agrupamentos em uma unidade.

Uma característica importante dos procedimentos hierárquicos é que os resultados de um estágio anterior são sempre aninhados com os resultados de um estágio posterior, apresentando semelhança com a estrutura de uma árvore (HAIR, ANDERSON, TATHAM e BLACK, 2005).

Ao contrário do método aglomerativo, no método divisivo todas as observações começam em um grande grupo agregado, sendo separadas, primeiramente, as observações mais distantes, até que cada observação se torna um grupo isolado.

Segundo Johnson e Wichern *apud* Maroco (2007), o algoritmo do método aglomerativo se desenvolve nos seguintes passos:

1. “Começar com N clusters (um para cada sujeito ou variável) e calcular a matriz de distância (ou a matriz de proximidade) $D_{N \times N}$.”
2. Procurar na matriz D os pares de sujeitos (ou variáveis) i e j mais semelhantes (com menor $d_{i,j}$). Caso existam vários grupos com $d_{i,j}$ iguais, usar como primeiro agrupamento o que possuir o sujeito de menor valor alfanumérico.
3. Combinar os clusters i e j (os dois com menores $d_{i,j}$) para formar o cluster ij . Atualizar a matriz D , eliminando a linha e a coluna correspondentes ao cluster j e adicionando uma nova linha e coluna com as distâncias entre o novo cluster ij e os restantes clusters originais.
4. Repetir os passos 2 e 3 $N-1$ vezes, tomando nota dos clusters criados em cada um dos passos e das distâncias entre estes. Na última iteração do algoritmo, todos os sujeitos (ou variáveis) são agrupados em um único cluster”.

Após a formação do primeiro cluster, é preciso definir como a distância entre dois clusters será computada. Neste aspecto, há diversos métodos para a formação dos agrupamentos, sendo que o que os diferencia, principalmente, é a maneira como as distâncias são calculadas entre os grupos já formados e os que faltam ser agrupados.

Os métodos mais frequentes são:

- Menor Distância ou Ligação Individual (*Single Linkage* ou *Nearest Neighbor*).
- Maior Distância ou Ligação Completa (*Complete Linkage* ou *Furthest Neighbor*).
- Distância Média ou Ligação Média (*Average Linkage* ou *Between Groups*).
- Centróide (*Centroid*).
- Ward.

O método da **Ligação individual** ou **Menor Distância** (*Single Linkage* ou *Nearest Neighbor*) baseia-se na distância mínima entre dois grupos de elementos, buscando agrupar inicialmente os objetos separados pela menor distância. Neste método, o primeiro grupo é formado pelos dois elementos que possuem a menor distância entre eles, ou seja, será formado pelo vizinho mais próximo. Na próxima etapa, será agregado a este grupo o elemento que tiver a menor distância em relação a eles, sucessivamente, até que se chega a um único grupo formado por todos os elementos.

Dados dois grupos (i e j) e (k), a distância entre eles é representada pela distância mínima de qualquer ponto de um grupo até qualquer ponto do outro:

$$d_{(ij)k} = \min\{d_{ik}, d_{jk}\} \quad (6.17)$$

Segue esquematização da Ligação Individual (*Single Linkage*), em que a distância será dada por d_{24} .

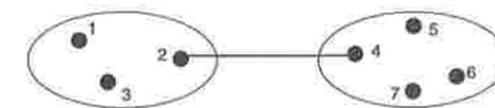


Figura 6.2: Esquematização do método da Ligação Individual.

O método da **Maior Distância** ou **Ligação Completa** (*Complete Linkage* ou *Furthest Neighbor*) baseia-se na distância máxima, ao contrário do método da ligação individual. Neste método, a distância entre dois grupos é definida como a distância máxima entre todos os pares de possibilidades de observações nos dois grupos. O método busca agrupar elementos cuja distância entre os mais afastados seja a menor.

Dados dois grupos (i e j) e (k), a distância entre eles é representada pela distância máxima de qualquer ponto de um grupo até qualquer ponto do outro:

$$d_{(ij)k} = \max\{d_{ik}, d_{jk}\} \quad (6.18)$$

Segue esquematização da Ligação Completa (*Complete Linkage*), em que a distância do agrupamento será d_{16} .

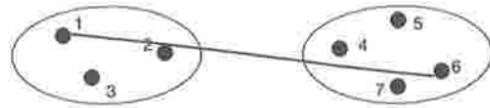


Figura 6.3: Esquematização do método da Ligação Completa.

Este método tende a formar grupos “mais compactos e composto de indivíduos muito semelhantes entre si. Embora os resultados deste método forneçam uma visão clara dos diferentes grupos encontrados, muitas vezes não apresentam um elevado grau de concordância com a estrutura inicial dos dados” (REIS, 2001).

O método da **Distância Média** ou **Ligação Média** (*Average Linkage* ou *Between Groups*) trata a distância entre dois grupos como sendo a distância média entre todos os pares de indivíduos dos dois grupos, buscando agrupar os agregados cuja distância média é a menor.

Como a técnica utiliza-se do valor médio, ao contrário da ligação individual e da ligação completa, há a vantagem de não se precisar de valores extremos e de se utilizarem todos os elementos do grupo, em vez de um único par de extremos.

Portanto, dado dois grupos (i e j) e (k), a distância entre eles é representada da seguinte maneira:

$$d_{(ij)k} = \text{média}\{d_{ik}, d_{jk}\} \quad (6.19)$$

Segue esquematização da Ligação Média (*Average Linkage* ou *Between Groups*), em que a distância do agrupamento será:

$$\frac{d_{14} + d_{15} + d_{16} + d_{17} + d_{24} + d_{25} + d_{26} + d_{27} + d_{34} + d_{35} + d_{36} + d_{37}}{7}$$

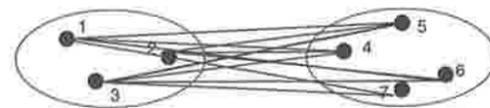


Figura 6.4: Esquematização do método da Ligação Média.

O método do **Centróide**, por sua vez, baseia-se na distância (geralmente euclidiana ou quadrática euclidiana) entre os centróides, priorizando a menor distância entre eles. Este método identifica os dois grupos separados pela menor distância entre os pontos mais próximos e os coloca no mesmo agrupamento.

Segundo Hair, Anderson, Tatham e Black (2005), “centróides são valores médios das observações sobre as variáveis na variável estatística de agrupamento. Neste método, toda vez que indivíduos são reunidos, um novo centróide é computado”. Este método também é mais robusto para observações atípicas.

Segue esquematização do método centróide:

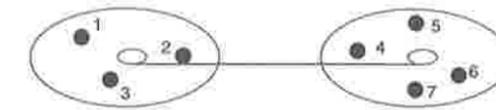


Figura 6.5: Esquematização do método Centróide.

Por fim, o método de Ward busca agrupar os agregados que apresentam menor soma dos quadrados entre os dois agrupamentos, calculada sobre todas as variáveis. Trata-se de um método que tende a proporcionar agregados com aproximadamente o mesmo número de observações (HAIR, ANDERSON, TATHAM e BLACK, 2005).

Segundo Reis (2001), o método de Ward pode ser resumido nas seguintes etapas:

- “Primeiramente, são calculadas as médias das variáveis para cada grupo.
- Em seguida, é calculado o quadrado da distância euclidiana entre estas médias e os valores das variáveis para cada indivíduo.
- Somam-se as distâncias para todos os indivíduos.
- Por último, pretende-se minimizar a variância dentro dos grupos. A função objetivo que se pretende minimizar é também chamada soma dos quadrados dos erros (em inglês, a sigla é ESS) ou soma dos quadrados dentro dos grupos (em inglês, a sigla é WSS)”.

Retomando o exemplo das seis empresas do comércio varejista, já tratado anteriormente, em que calculamos a matriz de similaridade com base na distância quadrática euclidiana, vamos agora proceder à formação dos grupos.

Consideremos a matriz de similaridade dada pela Tabela 6.3, anteriormente, e o método da **Ligação individual** ou **Menor Distância** (*Single Linkage* ou *Nearest Neighbor*). Neste método, como já abordado anteriormente, a distância entre dois objetos é representada pela distância mínima entre todas as possibilidades de pares de objetos em dois grupos.

Considerando cada empresa como se fosse um grupo, nós agrupamos as empresas Fiori e Ferramentas Gerais (1,2), por possuírem a menor distância entre pares de objetos. Portanto:

$$d_{(ij)k} = \min\{d_{ik}, d_{jk}\} = d_{12} = 0,152$$

Agora, após a formação do primeiro grupo (1,2), temos de recalculamos a matriz de distâncias utilizando o grupo já formado com as demais empresas.

Assim, a distância entre o grupo 1 (formado pelas empresas 1 e 2) e a empresa 3 (Bretas) é o mínimo das seguintes distâncias:

$$d_{(12)3}^2 = \min\{d_{13}^2, d_{23}^2\} = \{1,964; 3,163\} = 1,964$$

Da mesma forma, a distância entre o grupo (1,2) e as demais empresas 4, 5, 6 é dada por:

$$d_{(12)4}^2 = \min\{d_{14}^2, d_{24}^2\} = \{2,916; 4,248\} = 2,916$$

$$d_{(12)5}^2 = \min\{d_{15}^2, d_{25}^2\} = \{8,010; 9,898\} = 8,010$$

$$d_{(12)6}^2 = \min\{d_{13}^2, d_{23}^2\} = \{8,616; 10,200\} = 8,616$$

Substituindo a linha e a coluna correspondentes às empresas 1 e 2 e adicionando as novas distâncias calculadas para o novo grupo (1,2), obtemos a uma nova matriz, dada por:

Tabela 6.6: Matriz de Similaridade pela Distância Quadrática Euclidiana para Cinco Grupos

	(1,2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
(1,2)	0,000				
Bretas Supermercados (3)	1,964	0,000			
Renner (4)	2,916	0,183	0,000		
Lojas Americanas (5)	8,010	2,601	1,423	0,000	
Ponto Frio (6)	8,616	3,824	2,353	0,447	0,000

Com base na matriz anterior, percebe-se que as empresas mais semelhantes são a 3 e a 4 (Bretas e Renner), que formarão o próximo grupo com distância 0,183. Novamente, o procedimento para o cálculo da nova matriz é idêntico ao anterior. Agora, a distância entre o grupo (3,4) e as demais empresas é o mínimo das seguintes distâncias:

$$d_{(34)(12)}^2 = \min\{d_{3(12)}^2, d_{4(12)}^2\} = \{1,964; 2,916\} = 1,964$$

$$d_{(34)5}^2 = \min\{d_{35}^2, d_{45}^2\} = \{2,601; 1,423\} = 1,423$$

$$d_{(34)6}^2 = \min\{d_{36}^2, d_{46}^2\} = \{3,824; 2,353\} = 2,353$$

Substituindo a linha e a coluna correspondentes às empresas 3 e 4 e adicionando as novas distâncias calculadas para o novo grupo (3,4), obtemos uma nova matriz, dada por:

Tabela 6.7: Matriz de Similaridade pela Distância Quadrática Euclidiana para Quatro Grupos

	(1,2)	(3,4)	Lojas Americanas (5)	Ponto Frio (6)
(1,2)	0,000			
(3,4)	1,964	0,000		
Lojas Americanas (5)	8,010	1,423	0,000	
Ponto Frio (6)	8,616	2,353	0,447	0,000

No passo seguinte, a empresa 5 junta-se à empresa 6 em função da menor distância entre os pares (0,447), e as distâncias são:

$$d_{(56)(12)}^2 = \min\{d_{5(12)}^2, d_{6(12)}^2\} = \{8,010; 8,616\} = 8,010$$

$$d_{(56)(34)}^2 = \min\{d_{5(34)}^2, d_{6(34)}^2\} = \{1,423; 2,353\} = 1,423$$

Após esta etapa, temos a seguinte matriz de proximidades:

Tabela 6.8: Matriz de Similaridade pela Distância Quadrática Euclidiana para Três Grupos

	(1,2)	(3,4)	(5,6)
(1,2)	0,000		
(3,4)	1,964	0,000	
(5,6)	8,010	1,423	0,000

Agora, conforme se observa na Tabela 6.8, o grupo formado pelas empresas 3 e 4 se juntará ao grupo formado pelas empresas 5 e 6. Por fim, os dois grupos (1, 2) e (3, 4, 5, 6) se juntarão em um único grupo com distância de 1,9464.

Tabela 6.9: Matriz de Similaridade pela Distância Quadrática Euclidiana para Dois Grupos

	(1,2)	(3,4,5,6)
(1,2)	0,000	
(3,4,5,6)	1,964	0,000

A Tabela 6.10 a seguir resume o processo de agrupamento desenvolvido.

Tabela 6.10: Processo de Agrupamento Hierárquico de Seis Empresas, segundo o Método *Single Linkage*

Etapa	Distância	Grupos	Número de Agrupamentos
1	0,152	(1,2), 3, 4, 5, 6	5
2	0,183	(1,2), (3,4), 5, 6	4
3	0,447	(1,2), (3,4), (5,6)	3
4	1,423	(1,2), (3,4,5,6)	2
5	1,964	(1,2,3,4,5,6)	1

Uma maneira de representar graficamente o processo de agrupamento hierárquico é por meio do dendrograma, que mostra em cada etapa o esquema de aglomeração e a distância entre os grupos, conforme a Figura 6.6.

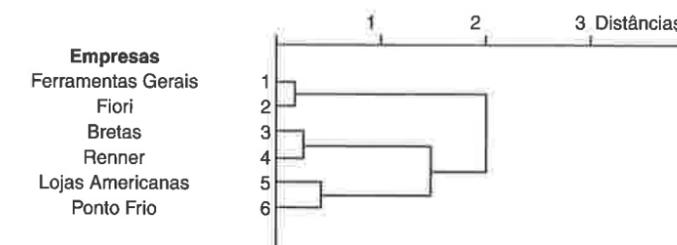


Figura 6.6: Dendrograma conforme critério *Single Linkage*.

Pelo dendrograma, é possível visualizar os elementos de cada *cluster*, conforme o corte que se queira fazer, imaginando-se uma linha reta vertical no gráfico. Para facilitar o entendimento, vamos imaginar uma linha reta que corta o gráfico, conforme mostrado na Figura 6.7. Por exemplo, nós podemos desenhar uma linha vertical tracejada que corta a distância 0,8, de acordo com a letra A. Neste ponto, identificamos três *clusters*, sendo um para cada ponto que intercepta o ramo do agrupamento (traços horizontais).

Agora, se cortássemos no ponto B, teríamos dois *clusters*, sendo o primeiro formado pelas empresas 1 e 2 e o segundo pelas empresas (3,4,5,6). Entretanto, o ideal é “cortar” antes que ocorram saltos muito grandes.

Com base no dendrograma, percebe-se que a formação de três grupos parece ser uma boa solução.

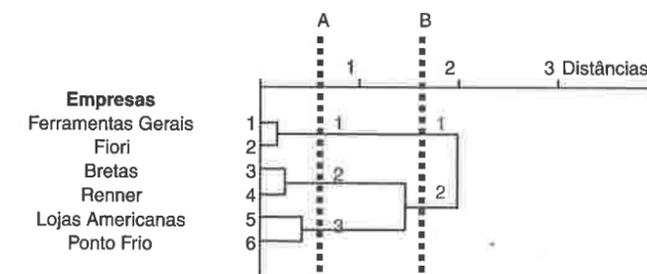


Figura 6.7: Dendrograma com linhas de corte para visualização de três grupos.

6.5.1.1. Análise de Conglomerados Hierárquicos: um Exemplo Prático

Imagine que um analista de mercado está interessado em segmentar um conjunto de empresas que atuam no setor de siderurgia e metalurgia, com base em informações financeiras publicadas. Para tanto, foram selecionadas 25 empresas que atuam neste segmento de acordo com os dados divulgados pela *Revista Exame – Melhores e Maiores do ano de 2005*, sendo escolhidos quatro indicadores para o estudo.

Para desenvolver o exemplo, abra o arquivo *Siderur_clusterteste.sav*. Este arquivo contém as seguintes variáveis:

- *Fatur*: Faturamento expresso em milhões de dólares;
- *Rent*: Rentabilidade do Patrimônio Líquido (%);
- *Endiv*: Endividamento geral da empresa (%);
- *Empreg*: Número de empregados.

No SPSS, selecione **Analyze** → **Classify** → **Hierarchical Cluster**, conforme Figura 6.8.

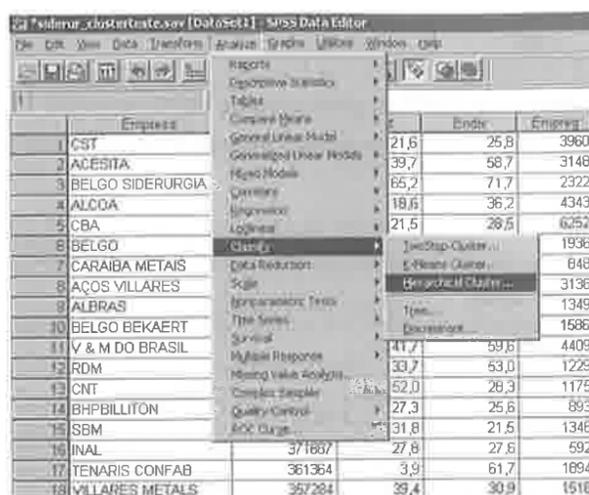


Figura 6.8: Análise de conglomerados – Hierarchical Cluster.

Na sequência, em **Variable(s)**, selecione as variáveis faturamento, rentabilidade, endividamento e número de empregados. Em **Label Cases**, coloque as empresas. Para agrupar as empresas, na opção **Cluster**, selecione **Cases**, mas se o interesse fosse agrupar variáveis a opção **Variables** seria selecionada, cabendo destacar que a análise com variáveis seria semelhante à análise fatorial, que será objeto de estudo do próximo capítulo. Em **Display**, selecione **Statistics** e **Plots**, conforme a Figura 6.9.

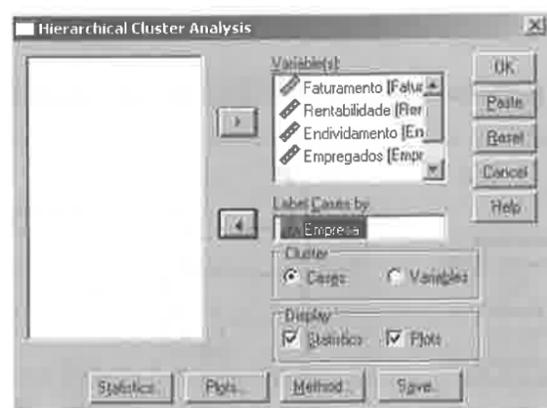


Figura 6.9: Hierarchical Cluster – seleção de variáveis.

Em seguida, clique em **Statistics...**, que será aberta a caixa de diálogo conforme a Figura 6.10, e selecione as opções **Agglomeration schedule** e **Proximity matrix**. O **Agglomeration Schedule** (esquema de aglomeração) mostra a ordem em que as empresas são agrupadas em cada estágio. Já a matriz de proximidade (**Proximity Matrix**) será dada em função da medida de proximidade a ser selecionada.

Além disso, se a opção **Cluster Membership** for selecionada, será gerado um quadro que mostra cada empresa e o *cluster* a que pertence. Nessa opção, o pesquisador pode escolher uma solução única (**Single solution**) para o número de *clusters* ou fornecer um intervalo de soluções (**Range of solutions**).

Neste exemplo, vamos imaginar que o ideal de agrupamentos estaria entre 2 e 5 grupos em função de estudos anteriores. Isto não é uma regra, mas o ideal é que o pesquisador faça a primeira análise de conglomerado sem marcar esta opção, para ter uma idéia do número de agrupamentos. Depois, clique em **Continue**.

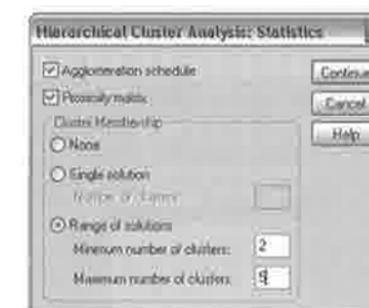


Figura 6.10: Hierarchical Cluster – Menu Statistics.

A próxima etapa é clicar em **Plots** e selecionar o **Dendrogram** e o **Icicle**, conforme Figura 6.11.

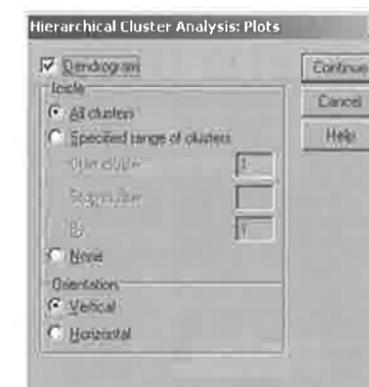


Figura 6.11: Hierarchical Cluster – Plots.

O Dendograma representa, graficamente, o esquema de aglomeração apresentado no **Agglomeration Schedule** em uma escala 0 a 25. Já o **Icicle Plot** mostra os elementos que compõem cada *cluster*.

Depois, clique no botão **Continue** e em seguida em **Method** para selecionar o tipo de método de agrupamento; neste caso, utilizaremos **Between-Groups Linkage**. Na mesma caixa de diálogo é possível escolher a medida de similaridade, e ela também permite a padronização das variáveis.

A medida de distância será a distância quadrática euclidiana. O SPSS adota como default essa medida, mas outras também poderão ser utilizadas, *vide* Figura 6.13.

Na parte teórica sobre análise de conglomerados, fizemos comentários sobre a necessidade de padronização das variáveis para se evitar a distorção nos agrupamentos em virtude de diferenças de escalas. No método hierárquico, é possível padronizar as variáveis na caixa de diálogo **Methods**, bastando escolher o tipo de padronização. Neste caso, escolhermos o Z scores (Figura 6.14). Depois, clique em **Continue**.



Figura 6.12: Hierarchical Cluster – método Between Groups.

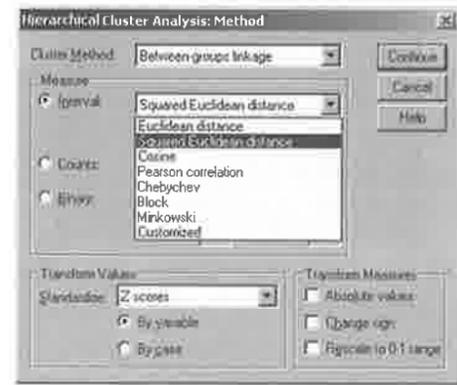


Figura 6.13: Hierarchical Cluster – medida de similaridade – distância quadrática euclidiana.

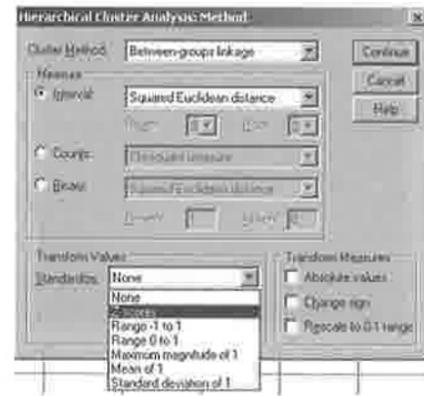


Figura 6.14: Hierarchical Cluster – padronização das variáveis pelo Z scores.

Na sequência, clique em **Save New Variables**, que permitirá que seja criada uma nova variável no banco de dados contendo o grupo a que cada elemento pertence (*vide* Figura 6.15). Entretanto, no primeiro momento, o pesquisador pode optar por não estabelecer um número determinado de agrupamentos ou intervalo de soluções, mas, em vez disso, fará uma primeira análise dos resultados para verificar qual seria uma solução possível dos agrupamentos naturais para serem salvos. Ou seja, o *cluster membership* poderia ser utilizado em uma segunda etapa. Entretanto, vamos supor que o ideal de agrupamentos estaria entre 2 e 5 grupos. Depois, clique em **Continue** e em **OK**.

A Tabela 6.11 mostra a matriz de proximidade entre casos ou variáveis. Na **Matriz de Proximidade**, os valores representam a similaridade ou dissimilaridade entre cada par de observações. O exemplo

Tabela 6.11: Matriz de Proximidade pela Distância Quadrática Euclidiana

Case	Proximity Matrix																								
	1:CST	2:ACESITA	3:BELGO SIDERURGIA	4:ALCOA	5:CBA	6:BELGO	7:CARAÍBA METAIS	8:AÇOS VILLARES	9:ALBRAS	10:BELGO BEKAERT	11:V & M DO BRASIL	12:RDM	13:CNT	14:BHPBILLITON	15:SBM	16:INAL	17:TENARIS CONFAB	18:VILLARES METAIS	19:TERMOMECÂNICA	20:MM	21:ZAMPROGNA	22:AMSTED MAXION	23:MANGELS	24:VALESUL	25:ELUMA
1:CST	0,000	6,579	18,605	5,358	7,745	9,086	17,384	18,565	12,247	11,362	14,187	18,244	21,428	18,005	17,539	19,825	20,915	18,669	16,192	19,269	19,055	28,186	21,279	21,710	21,544
2:ACESITA	6,579	0,000	4,137	5,040	9,612	5,587	6,372	3,343	4,731	7,085	3,574	6,475	9,925	11,216	10,900	11,927	13,072	8,918	12,987	8,305	11,779	8,863	11,530	12,891	10,674
3:BELGO SIDERURGIA	18,605	4,137	0,000	17,001	22,867	13,568	11,996	2,999	10,638	14,283	7,772	10,500	10,657	18,412	17,437	18,578	24,886	13,041	24,259	11,759	19,369	8,825	21,069	20,431	18,314
4:ALCOA	5,358	5,040	17,001	0,000	2,011	3,757	9,579	10,283	5,565	5,425	4,781	8,223	12,755	8,560	8,031	9,835	7,988	8,589	5,675	8,568	8,411	14,051	7,789	10,020	8,414
5:CBA	7,745	9,612	22,867	2,011	0,000	9,642	19,479	15,026	13,150	11,727	6,823	16,264	19,377	16,358	14,664	18,260	15,990	15,030	11,239	14,701	16,000	18,428	15,642	18,863	16,631
6:BELGO	9,086	5,587	13,568	3,757	9,642	0,000	5,981	9,284	0,913	0,190	6,921	2,884	4,152	1,580	1,446	2,112	7,439	2,064	2,142	2,592	1,941	14,545	5,757	3,185	4,595
7:CARAÍBA METAIS	17,384	6,372	11,996	9,579	19,479	5,981	0,000	7,022	2,530	6,808	8,733	2,227	10,145	6,416	7,975	6,182	3,697	7,067	9,171	6,797	6,358	9,200	2,995	4,903	2,633
8:AÇOS VILLARES	18,565	3,343	2,999	10,283	15,026	9,284	7,022	0,000	6,425	10,090	1,094	0,775	3,691	1,622	1,999	12,448	13,848	7,677	15,353	6,225	12,192	10,506	3,681	2,140	2,454
9:ALBRAS	12,247	4,731	10,638	5,565	6,117	0,913	0,000	6,425	0,000	1,094	6,117	0,775	3,691	1,622	1,999	12,448	13,848	7,677	15,353	6,225	12,192	10,506	3,681	2,140	2,454
10:BELGO BEKAERT	14,187	3,574	7,772	4,781	8,223	6,921	2,227	2,530	6,117	8,052	0,000	5,661	8,511	10,508	9,263	11,299	11,072	6,769	10,544	5,555	10,156	3,225	8,749	11,360	8,100
11:V & M DO BRASIL	18,244	6,475	10,500	8,223	16,264	2,884	2,227	10,145	6,416	3,003	5,661	0,000	3,268	2,187	2,570	2,066	5,094	1,515	4,820	1,297	2,120	7,190	2,805	1,844	1,388
12:RDM	21,428	9,925	10,657	12,755	19,377	4,152	10,145	7,805	3,691	3,003	8,511	3,268	0,000	3,230	2,266	3,217	15,047	0,920	6,996	0,951	3,689	12,229	10,708	5,091	7,528
13:CNT	18,005	11,216	18,412	8,560	16,358	1,580	6,416	7,805	0,000	0,948	10,508	2,187	3,230	0,000	0,253	0,064	6,602	1,046	1,383	1,704	0,087	15,826	4,395	0,594	2,815
14:BHPBILLITON	17,539	10,800	17,437	8,031	14,664	1,446	7,975	11,399	1,995	0,740	9,263	2,570	2,266	0,253	0,000	0,460	8,222	0,546	1,416	1,103	0,359	15,151	5,694	1,386	3,776
15:SBM	19,825	11,927	18,578	9,835	18,260	2,112	6,182	12,448	1,771	1,360	11,299	2,066	3,217	0,064	0,000	0,718	1,154	1,908	1,503	1,806	0,155	15,925	4,413	0,440	2,709
16:INAL	20,915	13,072	24,886	7,988	15,990	7,439	3,697	13,848	5,446	8,464	11,072	5,094	15,047	6,602	8,222	6,718	0,000	8,981	5,503	8,951	5,696	13,206	0,472	4,181	1,551
17:TENARIS CONFAB	18,669	8,918	13,041	8,589	15,030	2,064	7,067	7,677	1,763	1,360	6,769	1,515	0,920	1,046	0,546	1,154	0,000	8,981	5,503	3,117	0,000	10,866	18,114	2,072	3,647
18:VILLARES METAIS	19,269	8,305	11,759	8,568	14,701	6,797	6,225	12,192	1,944	1,978	5,555	1,297	3,689	1,383	1,416	1,908	5,503	3,117	0,000	3,852	1,098	18,114	4,206	2,006	3,666
19:TERMOMECÂNICA	19,055	11,779	19,369	8,411	16,000	1,941	6,358	12,192	1,866	1,354	10,156	2,120	3,689	0,087	0,359	0,155	5,696	1,131	1,098	1,698	0,000	15,044	0,000	0,367	2,208
20:MM	28,186	8,863	8,825	14,051	16,428	14,545	9,200	1,871	10,506	15,419	3,225	7,190	12,229	15,826	15,151	15,925	13,206	10,866	18,114	8,859	15,044	0,000	10,269	0,000	1,230
21:ZAMPROGNA	21,279	11,530	21,069	7,789	15,642	5,757	2,995	10,790	3,681	6,394	8,749	2,805	10,708	4,395	5,684	4,413	0,472	5,776	4,206	5,634	3,568	10,269	0,000	2,383	0,346
22:AMSTED MAXION	21,710	12,891	20,431	10,020	18,863	3,185	4,903	12,633	2,140	2,649	11,360	1,844	5,091	0,594	1,386	0,440	4,181	2,072	2,006	2,563	0,367	14,827	2,383	0,000	1,230
23:MANGELS	21,544	10,674	18,314	8,414	16,631	4,595	2,633	9,240	2,454	4,783	8,100	1,388	7,528	2,815	3,776	2,709	1,551	3,647	3,666	3,536	2,208	9,353	0,346	0,000	
24:VALESUL	21,710	12,891	20,431	10,020	18,863	3,185	4,903	12,633	2,140	2,649	11,360	1,844	5,091	0,594	1,386	0,440	4,181	2,072	2,006	2,563	0,367	14,827	2,383	0,000	
25:ELUMA	21,544	10,674	18,314	8,414	16,631	4,595	2,633	9,240	2,454	4,783	8,100	1,388	7,528	2,815	3,776	2,709	1,551	3,647	3,666	3,536	2,208	9,353	0,346	0,000	

This is a dissimilarity matrix

mostra os coeficientes da distância quadrática euclidiana referente à combinação dos pares de empresas. Neste caso, a matriz de proximidade é uma matriz de dissimilaridade, uma vez que foi utilizada uma medida de distância (distância quadrática euclidiana) como medida de similaridade. Para medidas de distância, vale lembrar que, quanto menor a distância, mais próximos estão os elementos, ou seja, mais similares eles são.

Entretanto, a relação é inversa se uma medida de similaridade, como a correlação, for utilizada, pois, nesta situação, quanto maiores os valores, mais próximos estarão os elementos.

Conforme a Tabela 6.11, pode-se observar que os primeiros elementos a serem agrupados pela medida da distância quadrática euclidiana serão aqueles que apresentarem o menor coeficiente. No caso específico, trata-se das empresas Bhpbilliton (14) e Inal (16), com coeficiente de 0,064.

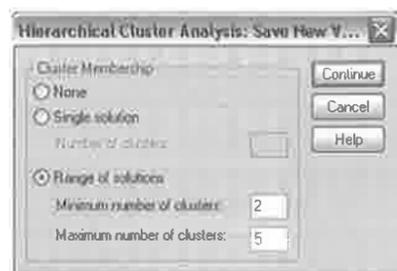


Figura 6.15: Hierarchical Cluster – salvar grupos.

O esquema de aglomeração (**Agglomeration Schedule**) mostrado na Tabela 6.12 indica a ordem de agregação (estágios) das empresas nos respectivos *clusters*, conforme o método utilizado. A coluna **Stage** mostra em que etapa cada elemento foi agrupado. As duas próximas colunas **Cluster Combined** indicam quais observações ou grupos foram unidos em cada estágio. A coluna **Coefficientes** mostra a distância entre dois grupos ou casos unidos em cada estágio, sendo que seu valor depende da medida de proximidade utilizada e do método usado na análise. Neste caso, utilizamos a distância quadrática euclidiana pelo método *Average Linkage (Between Groups)*, sendo que este procedimento busca agrupar os elementos de acordo com a menor distância média entre todos os pares de indivíduos dos dois grupos.

No nosso exemplo, no estágio 1, agrupou-se o registro 14 e o 16, que são, respectivamente, as empresas BHPBilliton e Inal, que apresentam a menor distância média na matriz de proximidade, sendo apresentado na coluna **Coefficients** o valor de 0,064. Nesse momento, tem-se 24 *clusters*. A distância é calculada considerando-se a padronização das variáveis pelo Z scores, da seguinte maneira:

$$d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2$$

$$d_{14,20}^2 = (-0,55459 - (-0,63818))^2 + (-0,26472 - (-0,22448))^2 + (-0,89966 - (-0,79869))^2 + (-0,90195 - (-1,11433))^2 = 0,064$$

No estágio 2, a empresa 18 (Villares Metals) foi agrupada à empresa 20 (CMM). O total de *clusters* no estágio 2 é de 23. No terceiro estágio, a empresa 21 (Zamproгна) é adicionada ao grupo formado por 14 e 16, e assim sucessivamente.

Os últimos a se juntarem são 11, 7, 2 e 3, sendo a última a empresa 1 (CST) no estágio 22.

As colunas **Stage Cluster First Appears** indicam justamente em qual estágio anterior cada elemento foi associado a outro. A coluna **Next Stage** indica quando o *cluster* formado naquele estágio vai se fundir com outro (por exemplo, o grupo formado no estágio 1 se fundirá com outro grupo no estágio 3).

Uma das formas de se determinar o número de agrupamentos é pela diferença dos coeficientes apresentados nesse esquema de aglomeração. Ou seja, um dos critérios de escolha do número de agrupamentos seria observar a maior diferença entre os coeficientes. O estágio anterior a esse salto indica o ponto de parada para novos agrupamentos. No caso específico, a maior diferença estaria entre o estágio 22 e 23, o que sugere a escolha de três *clusters*.

Tabela 6.12: Esquema de Aglomeração

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
	1	14		16	,064	
2	18	20	,116	0	0	9
3	14	21	,121	1	0	6
4	6	10	,190	0	0	12
5	23	25	,346	0	0	10
6	14	15	,357	3	0	7
7	14	24	,697	6	0	11
8	9	12	,775	0	0	15
9	13	18	,935	0	2	15
10	17	23	1,012	0	5	18
11	14	19	1,562	7	0	12
12	6	14	1,789	4	11	16
13	8	22	1,871	0	0	17
14	4	5	2,011	0	0	22
15	9	13	2,246	8	9	16
16	6	9	2,443	12	15	20
17	8	11	2,562	13	0	21
18	7	17	3,108	0	10	20
19	2	3	4,137	0	0	21
20	6	7	5,548	16	18	23
21	2	8	5,896	19	17	23
22	1	4	6,551	0	14	24
23	2	6	11,492	21	20	24
24	1	2	13,845	22	23	0

Na Tabela 6.13 (*Cluster Membership*), apresenta-se o número do *cluster* a que cada empresa pertence, de acordo com o número de grupos especificados. Neste caso, como escolhemos uma solução entre 2 a 5 grupos, cada coluna indica o número de grupos formados e as empresas que pertencem a cada grupo. Por exemplo, se selecionarmos a solução com três *clusters*, eles serão compostos das seguintes empresas: *Cluster 1* (1, 4 e 5), *Cluster 2* (2, 3, 8, 11 e 22), *Cluster 3* (6, 7, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25).

Na Tabela 6.14, é apresentado o **Icicle Plot**, que fornece a representação gráfica de quais empresas foram unidas em cada estágio de análise, mostrando os elementos que compõem cada *cluster*. Cada coluna branca entre as empresas representa a fronteira entre os grupos, e cada linha marcada com "X" indica os casos agrupados em cada *cluster*. Por exemplo, ao selecionar dois *clusters*, têm-se as empresas CST (1), Alcoa (4) e CBA (5) alocadas em um *cluster* e as demais empresas em outro.

Tabela 6.13: Cluster Membership

Case	Cluster Membership			
	5 Clusters	4 Clusters	3 Clusters	2 Clusters
1:CST	1	1	1	1
2:ACESITA	2	2	2	2
3:BELGO SIDERURGIA	2	2	2	2
4:ALCOA	3	3	1	1
5:CBA	3	3	1	1
6:BELGO	4	4	3	2
7:CARAÍBA METAIS	4	4	3	2
8:AÇOS VILLARES	5	2	2	2
9:ALBRAS	4	4	3	2
10:BELGO BEKAERT	4	4	3	2
11:V & M DO BRASIL	5	2	2	2
12:RDM	4	4	3	2
13:CNT	4	4	3	2
14:BHPBILLITON	4	4	3	2
15:SBM	4	4	3	2
16:INAL	4	4	3	2
17:TENARIS CONFAB	4	4	3	2
18:VILLARES METALS	4	4	3	2
19:TERMOMECÂNICA	4	4	3	2
20:CMM	4	4	3	2
21:ZAMPROGNA	4	4	3	2
22:AMSTED MAXION	5	2	2	2
23:MANGELS	4	4	3	2
24:VALESUL	4	4	3	2
25:ELUMA	4	4	3	2

A Figura 6.16 contém o dendrograma que representa graficamente o esquema de aglomeração apresentado no **Agglomeration Schedule**, mas os coeficientes (distâncias) apresentam-se em uma escala de 0 a 25. Pelo dendrograma, é possível visualizar os elementos de cada *cluster*, conforme o corte que se queira fazer, imaginando-se uma linha reta vertical no dendrograma. Com base na Figura 6.16 percebe-se a existência de três grupos.

***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Average Linkage (Between Groups)

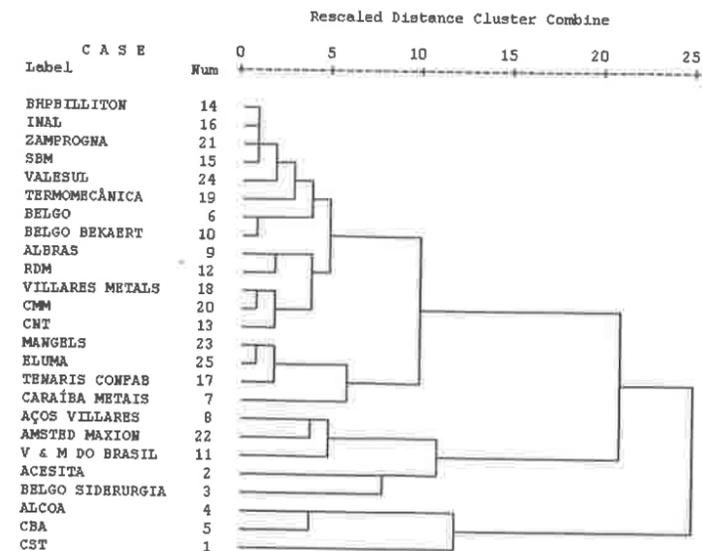


Figura 6.16: Dendrograma.

Tabela 6.14: Gráfico Icicle

Vertical Icicle	Case
1	1:CST
2	2:ACESITA
3	3:BELGO SIDERURGIA
4	4:ALCOA
5	5:CBA
6	6:BELGO
7	7:CARAÍBA METAIS
8	8:AÇOS VILLARES
9	9:ALBRAS
10	10:BELGO BEKAERT
11	11:V & M DO BRASIL
12	12:RDM
13	13:CNT
14	14:BHPBILLITON
15	15:SBM
16	16:INAL
17	17:TENARIS CONFAB
18	18:VILLARES METALS
19	19:TERMOMECÂNICA
20	20:CMM
21	21:ZAMPROGNA
22	22:AMSTED MAXION
23	23:MANGELS
24	24:VALESUL
25	25:ELUMA

6.5.2. Análise de Conglomerados não Hierárquicos

Como vimos anteriormente, no método hierárquico, o algoritmo estabelece uma relação de hierarquia entre os sujeitos e os grupos. Este fato não ocorre no método não hierárquico, pois, uma vez especificado o número de agrupamentos, o processo é dinâmico e interativo, tendo como objetivo identificar a melhor solução.

Os procedimentos não hierárquicos são utilizados para agrupar indivíduos (e não variáveis) cujo número inicial de *clusters* é definido pelo pesquisador.

Mingoti (2005) explica que os procedimentos não hierárquicos são métodos que têm como objetivo encontrar diretamente uma partição de n elementos em K grupos (*clusters*), de modo que a partição satisfaça dois requisitos básicos: ‘coesão’ interna (ou semelhança interna) e isolamento (ou separação) dos *clusters* formados.

Os métodos não hierárquicos não requerem o cálculo e armazenamento de uma nova matriz de distâncias a cada etapa do processo, o que reduz o tempo computacional e possibilita sua aplicação em grandes bases de dados.

Conforme Hair, Anderson, Tatham e Black (2005), há basicamente três abordagens para designar as observações individuais de determinado agrupamento:

- Referência sequencial: começa pela seleção de uma semente de agrupamento e inclui todos os objetos dentro de uma distância pré-especificada. Após a inclusão de todos os objetos, surge uma segunda semente e, dentro de uma distância pré-especificada, todos os objetos são incluídos de forma sucessiva.
- Referência paralela: são escolhidas diversas sementes de agrupamento e, dentro da distância pré-especificada, todos os objetos são agrupados de forma simultânea à semente mais próxima.
- Otimização: é semelhante às duas abordagens anteriores, porém, neste caso, é permitida a redesignação de objetos.

A probabilidade de acontecerem classificações erradas nos agrupamentos é menor nos métodos não hierárquicos, mas, em contrapartida, há a dificuldade de se estabelecer o número de *clusters* de partida. Uma alternativa consiste na realização do método hierárquico como técnica exploratória para, posteriormente, propiciar a utilização do número indicado de *clusters* na técnica não hierárquica (MAROCO, 2007). Dos métodos não hierárquicos, o mais popular é o *K-means*, também chamado de *K-médias*.

O método não hierárquico *K-Means* pode ser usado para o agrupamento de grandes conjuntos de observações. Este método produz apenas uma solução para o número de conglomerados predefinido, que deve ser especificado pelo analista, enquanto o método hierárquico fornece uma série de soluções correspondentes a diferentes números de agrupamentos.

De acordo com Anderberg *apud* Gouvêa e La Plata (2006), “o algoritmo usado para a determinação da alocação das observações em cada conglomerado no método *K-Means* é denominado *nearest centroid sorting*. O critério de distância para formação dos grupos no método *K-Means* é a distância euclidiana”.

No método *K-means*, o processo é composto de três passos, de acordo com Maroco (2007):

- “Partição inicial dos indivíduos em K *clusters* definidos pelo analista.
- Cálculo dos centróides para cada um dos K *clusters* (no SPSS, as primeiras n observações são usadas como centróides dos K *clusters* no primeiro passo da rotina; ou o analista pode definir qual o valor dos centróides a usar) e cálculo da distância euclidiana dos centróides a cada sujeito na base de dados.



- Agrupar os sujeitos aos *clusters* cujos centróides se encontram mais próximos, e voltar ao passo 2 até que não ocorra variação significativa na distância mínima de cada sujeito da base de dados a cada um dos centróides do K *clusters* (ou até que o número máximo de interações ou o critério de convergência – definido pelo analista – seja alcançado)”.

De acordo com Gouvêa e La Plata (2006), “com o método *K-Means*, persegue-se o objetivo de minimização da variância interna aos grupos e maximização da variância entre os grupos”.

Em *K-means*, o pesquisador pode fornecer informações sobre os centróides, ou pontos de sementes iniciais de agrupamento, que serão a base para a alocação dos indivíduos. Esses centróides podem ser obtidos por meio do *Hierarchical Cluster*.

Mas, caso os centróides iniciais sejam desconhecidos, sua estimação é feita de forma aleatória, com base nas observações. No SPSS, as primeiras n observações são utilizadas como centróides iniciais dos K *clusters* resultantes da primeira etapa do algoritmo, formando conglomerados temporários.

O principal problema enfrentado pelos métodos não hierárquicos é justamente a seleção das sementes de agrupamento. A opção de referência sequencial pode gerar resultados distintos, em função de uma alteração na ordem das observações no banco de dados, uma vez que este método se baseia em um conjunto de pontos iniciais para calcular a semente. Ou seja, a alteração da ordem dos dados no banco de dados pode afetar os resultados iniciais e, principalmente, os finais. Mas mesmo a semente aleatória poderá produzir diferentes resultados para cada conjunto de pontos (HAIR, ANDERSON, TATHAM e BLACK, 2005).

O procedimento *K-means* também é frequentemente utilizado com o intuito de se verificar se as variáveis adotadas no procedimento hierárquico são estatisticamente significantes para a formação daqueles K *clusters* obtidos. Como veremos mais adiante, essa verificação será elaborada por meio da análise de variância (ANOVA), cuja finalidade não é verificar se os *clusters* são ou não diferentes, mas identificar qual ou quais das variáveis permitem a separação desses *clusters*.

6.5.2.1. Análise de Conglomerados não Hierárquicos *K-means*: um Exemplo Prático

Para o procedimento *K-means*, vamos utilizar os mesmos dados do exemplo hierárquico, apesar de este método ser mais utilizado para grandes amostras. Ao se efetuar o *K-means*, como já abordado nos aspectos teóricos, o pesquisador deve indicar qual será o valor de K a ser inserido no SPSS. O pesquisador pode fazer uso da técnica hierárquica para obter um indicativo do número de *clusters* a ser adotado no *K-means*. O *K-means* fornece apenas uma solução de agrupamento com base no que for especificado pelo pesquisador.

Assim, abra o arquivo *Siderur_clusterteste.sav*.

Vale destacar que, diferentemente do *Hierarchical Cluster*, que possui a opção internamente de padronizar as variáveis, no *K-means* as variáveis a serem consideradas já deverão estar padronizadas. Desta forma, é necessário efetuar a transformação antes de iniciar a técnica. Clique em **Analyze** → **Descriptives Statistics** → **Descriptives...**, conforme a Figura 6.17.

Selecione todas as variáveis e coloque-as em **Variable(s)**, marque a opção **Save standardized values as variables** e clique em **OK** para salvar as variáveis no banco de dados, conforme a Figura 6.18.

Feito isso, selecione **Analyze** → **Classify** → **K-means Cluster...**, conforme a Figura 6.19.

Conforme a Figura 6.20, selecione as variáveis padronizadas criadas anteriormente (*zFatur*, *zRent*, *zEndiv*, *zEmpreg*) e coloque-as em **Variables:**. Em **Label Cases by:**, especifique a variável *Empresa*. Em **Number of Clusters:**, digite 3, que corresponde ao número de *clusters* indicado pelo método hierárquico, conforme abordado anteriormente.

Case	Mean	Std. Deviation	Minimum	Maximum
1	21,5	28,5	6252	3960
2	39,7	53,0	1229	3148
3	65,2	71,7	2322	3148
4	18,6	36,2	4343	3148
5	21,5	28,5	6252	3960
6	22,5	72,1	848	3148
7	51,6	74,2	3136	3148
8	30,0	43,1	1349	3148
9	30,9	22,5	1686	3148
10	41,7	59,6	4409	3148
11	33,7	53,0	1229	3148
12	52,0	28,3	1175	3148
13	27,3	25,6	893	3148
14	31,8	21,5	1348	3148
15	27,8	27,6	592	3148
16	3,9	61,7	1894	3148
17	39,4	30,9	1518	3148
18	17,1	18,9	2089	3148

Figura 6.17: Padronização de variáveis.



Figura 6.18: Caixa de diálogo Descriptives.

Case	Mean	Std. Deviation	Minimum	Maximum
1	21,6	26,8	3960	3960
2	39,7	58,7	3148	3148
3	65,2	71,7	2322	3148
4	18,6	36,2	4343	3148
5	21,5	28,5	6252	3960
6	22,5	72,1	848	3148
7	51,6	74,2	3136	3148
8	30,0	43,1	1349	3148
9	30,9	22,5	1686	3148
10	41,7	59,6	4409	3148
11	33,7	53,0	1229	3148
12	52,0	28,3	1175	3148
13	27,3	25,6	893	3148
14	31,8	21,5	1348	3148
15	27,8	27,6	592	3148
16	3,9	61,7	1894	3148
17	39,4	30,9	1518	3148
18	17,1	18,9	2089	3148
19	17,1	18,9	2089	2089

Figura 6.19: Análise de conglomerados – K-means.

O **Method** permite que se escolha entre **Iterate and classify** ou **Classify only**, sendo que, no primeiro caso, o procedimento se encarrega de estimar os centróides interativamente a cada nova observação designada e de classificar os sujeitos. Na opção **Classify only**, os centróides não são atualizados e é utilizado quando se buscam atribuir casos adicionais nos *clusters* já criados.

Além disso, ainda por meio do quadro apresentado na Figura 6.20, tem-se a opção do **Cluster Centers**, que possibilita incluir os valores dos centróides iniciais. A opção **Read initial** permite que o usuário decida quais valores utilizar como centróides iniciais. O botão **File** serve para indicar o caminho em que se



Figura 6.20: K-means – seleção de variáveis.

encontra o arquivo que contém os valores dos centróides. A opção **Write final** guarda os centróides dos conglomerados finais em um arquivo externo. Tal arquivo pode ser utilizado posteriormente para classificar novos casos. Ou seja, por meio do *K-means*, o pesquisador pode fornecer o centróide, ou ponto de semente inicial de agrupamento, que será a base para a alocação dos indivíduos. Estes centróides podem ser obtidos a partir do *Hierarchical Cluster*.

Mas, caso os centróides iniciais sejam desconhecidos, sua estimação é feita aleatoriamente, com base em todas as observações, de acordo com Hair, Anderson, Tatham e Black (2005). No SPSS, as primeiras *K* observações são utilizadas como centróides iniciais dos *K clusters* na primeira etapa do algoritmo, formando conglomerados temporários.

A Figura 6.21 mostra a caixa de diálogo **Iterate**, que permite controlar alguns detalhes relacionados ao processo de iteração utilizado para calcular os centróides finais, determinando quando a solução torna-se estável e, neste caso, cessando o algoritmo. No critério de convergência, pode-se inserir outro valor diferente do padrão do SPSS, que é zero. O valor introduzido representa a proporção da distância mínima existente entre os centróides iniciais dos conglomerados, devendo estar entre 0 e 1.

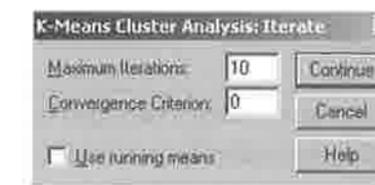


Figura 6.21: K-means – caixa de diálogo Iterate.

Na sequência, clicando no botão **Save...**, será aberta a caixa de diálogo mostrada na Figura 6.22. Esta opção permite criar uma nova variável, como o *cluster* a que cada elemento pertence.

Pode-se também criar uma variável que contém a distância euclidiana de cada observação ao centróide do respectivo conglomerado. Ressalta-se que a distância padrão utilizada pelo SPSS para o procedimento *K-means* é a distância euclidiana, e não a distância quadrática euclidiana. É muito comum o pesquisador notar que, após a aplicação do procedimento hierárquico com a utilização da distância quadrática euclidiana e a alocação das observações em cada um dos *clusters* obtidos, o procedimento *K-means* apresenta, como *output*, alocações diferentes daquelas obtidas anteriormente. Para que isso não ocorra, é preciso que seja utilizada, no SPSS, a distância euclidiana no procedimento hierárquico.

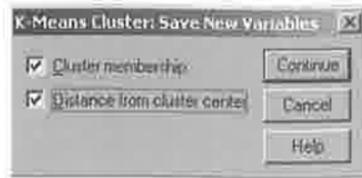


Figura 6.22: K-means – caixa de diálogo Save New Variables.

Clique em **Continue**.

O próximo passo consiste em clicar em **Options...** para selecionar as opções **Initial cluster centers**, **ANOVA table** e **Cluster information for each case**. Clique em **Continue**, conforme a Figura 6.23.

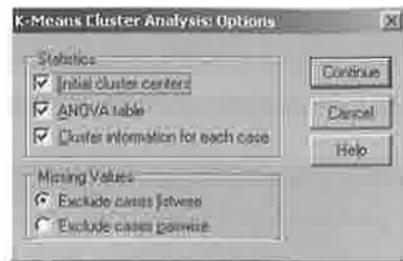


Figura 6.23: K-means – caixa de diálogo ANOVA.

Finalmente, clique em **OK** para que sejam gerados os *outputs* do SPSS. A Tabela 6.15 apresenta os valores iniciais do centróide de cada *cluster*, representando a média de cada variável dentro de cada *cluster* inicial.

Tabela 6.15: Centróides Iniciais

	Cluster		
	1	2	3
Zscore: Faturamento	-,91682	,81204	1,59547
Zscore: Rentabilidade	-,63141	-,67936	2,46637
Zscore: Endividamento	-,35632	-,75516	1,40555
Zscore: Empregados	-1,08469	2,87922	,10632

A Tabela 6.16 mostra o *Iteration History*, que indica a variação do centro dos *clusters* em cada passo da iteração. No caso, utilizou-se a atribuição dos indivíduos em três *clusters*.

Tabela 6.16: Processo de Iteração

Iteration	Change in Cluster Centers		
	1	2	3
1	,793	1,260	1,454
2	,000	,658	,299
3 ^a	,000	,000	,000

^a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 3. The minimum distance between initial centers is 4,343.

A Tabela 6.17 apresenta a composição de cada um dos três *clusters* selecionados, em que se mostra a qual grupo pertence cada elemento e a distância de cada observação ao centro do respectivo *cluster*.

Tabela 6.17: Composição dos *Clusters*

Cluster Membership			
Case Number	Empresa	Cluster	Distance
1	CST	2	1,640
2	ACESITA	3	1,457
3	BELGO SIDERURGIA	3	1,699
4	ALCOA	2	,881
5	CBA	2	1,254
6	BELGO	1	1,154
7	CARAÍBA METAIS	1	1,964
8	AÇOS VILLARES	3	,403
9	ALBRAS	1	,802
10	BELGO BEKAERT	1	1,072
11	V & M DO BRASIL	3	1,197
12	RDM	1	,885
13	CNT	1	1,840
14	BHPBILLITON	1	,711
15	SBM	1	,900
16	INAL	1	,774
17	TENARIS CONFAB	1	2,111
18	VILLARES METALS	1	,968
19	TERMOMECÂNICA	1	1,340
20	CMM	1	1,090
21	ZAMPROGNA	1	,659
22	AMSTED MAXION	3	1,642
23	MANGELS	1	1,586
24	VALESUL	1	,793
25	ELUMA	1	1,151

A Tabela 6.18 oferece a distância final entre os centróides dos *clusters*. Os valores na tabela indicam a média de cada variável em cada um dos *clusters* finais.

Tabela 6.18: Distância Final entre os Centróides dos *Clusters*

	Final Cluster Centers		
	1	2	3
Zscore: Faturamento	-,44698	1,57672	,57371
Zscore: Rentabilidade	-,25442	-,74587	1,31254
Zscore: Endividamento	-,27816	-,66978	1,34761
Zscore: Empregados	-,57460	1,89118	,81895

Esta tabela permite identificar e avaliar os *clusters*. O *cluster* final reflete as características de cada caso agrupado e, no nosso exemplo, percebe-se que no *Cluster* 1 todas as variáveis apresentam sinal negativo, ao contrário do *Cluster* 3, em que as variáveis apresentam valores com sinal positivo. Esse fato permite que sejam feitas interpretações dos agrupamentos formados, possibilitando a descrição dos perfis encontrados em cada grupo.

A Tabela 6.19 mostra as distâncias euclidianas entre os centróides dos *clusters*.

Tabela 6.19: Distância entre os Centróides

Cluster	Distances between Final Cluster Centers		
	1	2	3
1		3,251	2,843
2	3,251		3,235
3	2,843	3,235	

A Tabela 6.20 refere-se à análise de variância ANOVA (*analysis of variance*). Entretanto, a finalidade do quadro ANOVA não é verificar se os *clusters* são ou não diferentes, mas identificar qual ou quais das variáveis permitem a separação dos *clusters*. A nota de rodapé do *output* do SPSS chama a atenção para este fato. Em outras palavras, a tabela ANOVA indica quais variáveis mais contribuíram para a solução dos *clusters*.

Por meio da ANOVA, pode-se afirmar que, se uma variável conseguir distinguir bem os agrupamentos, é de se esperar que sua variabilidade entre os grupos (*Cluster Mean Square*) seja elevada. Do mesmo modo, é de se esperar que a variabilidade interna seja mínima (*Error Means Square*). Portanto, as variáveis que mais discriminam os grupos são aquelas com maior valor da estatística F.

Com base nos valores de sig. F para cada variável, pode-se afirmar que as quatro variáveis utilizadas no estudo são significantes para a formação dos três *clusters*, ao nível de significância de 5%. A variável referente ao número de empregados ($F = 50,338$) foi a que mais discriminou as empresas de cada *cluster*.

Tabela 6.20: Análise de Variância ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore: Faturamento	6,250	2	,523	22	11,957	,000
Zscore: Rentabilidade	5,692	2	,573	22	9,925	,001
Zscore: Endividamento	5,871	2	,557	22	10,536	,001
Zscore: Empregados	9,848	2	,196	22	50,338	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

A Tabela 6.21 mostra os números de empresas em cada *cluster* selecionado.

Tabela 6.21: Número de Casos em Cada Cluster

Number of Cases in each Cluster		
Cluster	1	17,000
	2	3,000
	3	5,000
Valid		25,000
Missing		,000

6.5.3. Qual o Melhor Método?

Nesta etapa, surgem os seguintes questionamentos: **qual a técnica que deve ser escolhida (hierárquica ou não hierárquica)?** Após a escolha da técnica, **qual o método mais adequado (ligação completa, Ward, centróide etc., se for hierárquico)?** A resposta a essas perguntas dependerá do problema de pesquisa.

Sharma (1996) discute brevemente as propriedades, vantagens e desvantagens de cada método e procedimentos que poderão facilitar a escolha do método mais adequado.

O **método hierárquico** não requer que o pesquisador conheça, *a priori*, o número de *clusters* para iniciar a partição, o que é uma vantagem sobre os métodos não hierárquicos. Mas uma desvantagem é que os objetos, uma vez agrupados, permanecerão no mesmo *cluster*. Uma alternativa seria utilizar o método hierárquico como exploratório e os resultados sugeridos dos agrupamentos serem inseridos no método não hierárquico para melhorar a solução final, tornando os métodos complementares, e não competitivos.

O **método não hierárquico** requer o conhecimento de um número sugerido de agrupamentos e necessita que uma semente inicial (centróide) seja selecionada. O principal problema está justamente na seleção dessa semente inicial e no estabelecimento desse número indicativo de *clusters*, já que o método não

hierárquico é muito sensível à definição da semente inicial e do número de *clusters* proposto pelo pesquisador. Alguns estudos indicam que o desempenho do método *K-means* é bastante inadequado quando a semente inicial aleatória é utilizada sem a aplicação prévia de um método hierárquico. Assim, recomenda-se que se utilize o número de *clusters* indicado *a priori* pelo método hierárquico.

6.6. QUANTOS AGRUPAMENTOS DEVEM SER SELECIONADOS?

Dados os agrupamentos formados, o próximo passo é avaliar e determinar o número de *clusters* presentes nos dados. Em análise de conglomerados, não existe de fato um procedimento padrão e altamente objetivo para estabelecer o número adequado de grupos.

O método hierárquico resulta em diversas soluções de agrupamento, sendo atribuição do pesquisador o julgamento do número mais adequado de *clusters*. Embora haja formas de norteá-lo nesta tarefa, como, por exemplo, por meio da utilização do coeficiente de aglomeração como parâmetro, ainda assim trata-se de uma decisão de natureza subjetiva.

No exemplo das seis empresas do comércio varejista abordado anteriormente, percebemos que, entre as etapas 3 e 4, houve um salto na distância entre os agrupamentos, sugerindo a existência de três *clusters*. Além disso, a representação gráfica com base no dendrograma possibilita uma boa visualização da distância entre os agrupamentos.

O *software* SAS fornece quatro estatísticas para auxiliar na determinação do número de *clusters*. As estatísticas disponíveis no SAS são:

- *Root-mean-square standard deviation* (RMSSTD) do número de *clusters*: verifica a homogeneidade do novo agrupamento.
- *Semipartial R-square* (SPR): verifica a homogeneidade da fusão dos agrupamentos.
- *R-square* (RS): identifica a heterogeneidade dos agrupamentos.
- *Distance between two clusters*: assim como o SPR, identifica a homogeneidade dos agrupamentos.

6.7. INTERPRETAÇÃO E VALIDAÇÃO DOS AGRUPAMENTOS

Finalmente, após a especificação da quantidade de *clusters*, o pesquisador fará o exame e a interpretação dos agrupamentos formados.

Sobre a validação das soluções de agrupamento no método hierárquico, uma alternativa consiste no emprego de diferentes medidas de similaridade, buscando avaliar a consistência dos resultados. No método não hierárquico, pode ser elaborado um comparativo entre a utilização de sementes aleatórias, como resultados obtidos, com o uso de sementes especificadas. Assim, caso haja consistência nos resultados, pode-se afirmar, com maior segurança, sobre a validade da análise.

Vale destacar que, após a especificação do número de *clusters*, o pesquisador deve ficar atento ao número de observações que compõem cada *cluster*, ou seja, verificar se a dimensão de cada grupo é muito distinta. Por exemplo, se for encontrado um agrupamento com apenas um elemento ou composto de uma quantidade bem pequena de observações, se comparado aos demais, pode haver indícios de ocorrência de observações atípicas e, neste caso, caberá ao pesquisador proceder ou não à eliminação dessas observações e refazer a análise.



6.8. RELAÇÃO COM OUTRAS TÉCNICAS

Em análise de conglomerados, o agrupamento de variáveis assemelha-se à análise fatorial, pois ambas as técnicas visam identificar grupos de variáveis relacionadas. Entretanto, a análise fatorial mostra-se mais adequada para o agrupamento de variáveis, uma vez que oferece ao pesquisador testes não gerados quando da aplicação da análise de conglomerados, cujo principal objetivo, conforme discutido ao longo do capítulo, é o de agrupamento de observações.

Em função das variáveis disponíveis e do objetivo de pesquisa, um pesquisador poderá decidir qual a técnica mais apropriada para resolver seu problema de investigação. Se o objetivo for identificar o grupo ao qual cada indivíduo pertence, ele poderá se valer da análise de conglomerados ou da análise discriminante. Entretanto, se a resposta desejada consistir em saber quais variáveis discriminam determinados grupos, a análise discriminante, que é uma técnica confirmatória de dependência, é mais apropriada. Desta forma, podemos afirmar que a análise discriminante pode ser utilizada como uma técnica confirmatória da análise de conglomerados, ou seja, se o pesquisador possuir uma base de dados e não apresentar sequer uma idéia de como as observações se comportam em termos de similaridade, pode inicialmente aplicar uma análise de conglomerados para estratificar a amostra em grupos homogêneos e, a partir daí, elaborar um modelo de análise discriminante para confirmar os agrupamentos obtidos ou até para classificar novas observações não pertencentes à amostra inicialmente trabalhada, já que o grupo a que pertence cada observação (*output* da análise de conglomerados) servirá de variável dependente no modelo confirmatório (*input* da análise discriminante).

A análise de conglomerados também pode fornecer as medidas de similaridade que irão compor a base de dados de um escalonamento multidimensional, por exemplo, que terá por objetivo a elaboração de mapas de posicionamento relativo de cada observação em estudo.

Se outro pesquisador estiver trabalhando com variáveis nominais e também possuir a intenção de classificar as observações de sua amostra em grupos internamente homogêneos e heterogêneos entre si, a fim de tentar investigar associações e possíveis classificações, as análises de correspondência (Anacor) e de homogeneidade (Homals) podem ser mais indicadas.

6.9. CONSIDERAÇÕES FINAIS

A análise de conglomerados permite ao pesquisador agrupar casos e até variáveis em grupos homogêneos, em função das similaridades entre as observações, a partir de variáveis previamente selecionadas. O objetivo da técnica é agrupar objetos com base em suas próprias características, buscando assim a estrutura “natural” dos dados.

A seleção das variáveis é o que define a estratificação obtida e, portanto, qualquer afirmação de que duas observações são similares para um conjunto de variáveis não consideradas na análise de conglomerados é puramente especulativa e sem fundamento. Por ser a análise de conglomerados uma técnica exploratória, o simples fato de identificarmos que alguns objetos pertencem ao mesmo agrupamento não caracteriza que sejam semelhantes em todos os aspectos. Os agrupamentos formados refletem a estrutura inerente das variáveis escolhidas. Cabe destacar que, na análise de conglomerados, diferentemente das demais técnicas multivariadas, a variável estatística de agrupamento é o conjunto de variáveis selecionadas pelo pesquisador, não sendo estimada empiricamente.

Segundo Hair, Anderson, Tatham e Black (2005), por ser uma técnica descritiva, sem base teórica e não inferencial, embora apresente fortes propriedades matemáticas, a análise de conglomerados não possui fundamentos estatísticos, não sendo apropriada para inferências de características de população a partir de dados amostrais.

6.10. EXERCÍCIOS – APLICAÇÃO DE BANCOS DE DADOS

1. Com base no arquivo *Siderur_clusterteste.sav* utilizado no exemplo prático ao longo deste capítulo (agrupamento hierárquico e *K-means*), elabore a análise de conglomerados hierárquicos, porém desconsiderando as três maiores empresas do setor. Compare os resultados.
2. O arquivo *500MM.sav* fornece informações financeiras das 50 maiores empresas do *ranking* da *Revista Exame Melhores e Maiores de 2005*. As variáveis que constam da base são:
 - Vendas (em US\$ mil);
 - Rentabilidade (%) do PL (Patrimônio Líquido);
 - Índice de Liquidez Geral (em %);
 - Índice de Endividamento Geral (%).

Com base nas informações fornecidas, aplique a técnica de análise de conglomerados hierárquicos, utilizando a distância quadrática euclidiana e o método *nearest neighbor*. Antes, padronize as variáveis pelo método Z scores. Analise todas as saídas do processamento (*hierarchical cluster* e *K-means*).

3. Refaça a análise de conglomerados com os dados do exercício anterior (*500MM.sav*), porém agora excluindo as duas maiores empresas, em função do faturamento. Analise todos os *outputs* da técnica gerados pelo *software*. Quais variáveis foram significativas para formar os *clusters*?
4. A tabela a seguir apresenta o resultado de um questionário aplicado em uma amostra de 10 alunos para saber o grau de satisfação destes sobre o curso de administração. Os quesitos avaliados foram: conhecimento adquirido, avaliação do corpo docente e mercado de trabalho. As notas são atribuídas em uma escala quantitativa de 0 a 10 para cada questão.

Aluno	Conhecimento	Docentes	Mercado
A	9	8	9
B	9	7	6
C	8	7	8
D	2	3	2
E	6	5	4
F	6	7	8
G	4	2	3
H	3	2	4
I	6	4	5
J	3	1	3

Com base nesta base de dados e por meio da aplicação da técnica de análise de conglomerados hierárquicos com a distância quadrática euclidiana com a utilização do método da ligação individual (*Single Linkage*), identifique o número de conglomerados e interprete os *outputs* gerados.

5. A base de dados a seguir apresenta características de 25 empresas varejistas (itens no sortimento, número de lojas com mais de 1.000 m² e faturamento mensal em R\$). Por meio da elaboração da análise de conglomerados, interprete todas as saídas do processamento (*hierarchical cluster* e *K-means*), após padronização das variáveis pelo método Z scores, utilizando a distância quadrática euclidiana e o método *Between Groups*.



Empresa	Itens no Sortimento	Número de Lojas com mais de 1.000 m ²	Faturamento Mensal (R\$)
A	2500	3	250000
B	2700	3	240000
C	4000	4	310000
D	3700	3	390000
E	9850	5	540000
F	17000	7	740000
G	25000	8	850000
H	3600	3	290000
I	4500	4	350000
J	6900	4	450000
K	12800	8	650000
L	32000	16	980000
M	1000	1	120000
N	1200	2	190000
O	1450	3	190000
P	14500	6	695000
Q	8500	5	490000
R	9800	3	570000
S	72000	18	1250000
T	4500	3	290000
U	35000	10	1000000
V	17000	8	820000
X	7000	4	410000
Y	79000	24	1950000
Z	7000	3	390000

6. Sejam as variáveis relativas a faturamento bruto em 2005 (R\$), área de vendas (m²) e número de funcionários de 30 grupos supermercadistas do Brasil:

Empresa	Faturamento bruto em 2005 (R\$)	Área de vendas (m ²)	Número de funcionários
COMPANHIA BRASILEIRA DE DISTRIBUIÇÃO	16.168.968.046	1.206.254	62.803
CARREFOUR COM E IND. LTDA.	12.546.232.768	1.013.247	48.072
WAL-MART BRASIL LTDA.	11.731.759.991	1.170.021	50.112
CIA ZAFFARI COMÉRCIO E INDÚSTRIA	1.410.185.010	113.075	7.727
GBARBOSA COMERCIAL LTDA.	1.227.023.563	66.707	6.730
DMA DISTRIBUIDORA S/A – EPA	1.212.276.038	95.998	7.000
IRMÃOS BRETAS, FILHOS E CIA. LTDA.	1.144.635.735	101.615	6.655
COOP – COOPERATIVA DE CONSUMO	1.076.619.724	62.990	3.682
A ANGELONI & CIA. LTDA.	992.810.531	64.812	6.065
PREZUNIC COMERCIAL LTDA.	781.709.800	34.175	12.000
Y. YAMADA S.A.	767.725.801	64.627	4.668
CONDOR SUPER CENTER LTDA.	765.895.534	62.057	3.820
SONDA SUPERMERCADOS EXP E IMP. LTDA.	752.195.264	42.171	3.700
LIDER SUPERMERCADOS E MAGAZINE LTDA.	733.881.395	38.804	5.564
EMPRESA BAIANA DE ALIMENTOS S.A.	496.123.154	64.713	3.660
SAVEGNAGO SUPERMERCADOS LTDA.	424.125.485	23.347	1.694

(Continua)

Empresa	Faturamento bruto em 2005 (R\$)	Área de vendas (m ²)	Número de funcionários
SUPERMERCADO GIMENES LTDA.	423.298.693	34.899	2.219
CARVALHO & FERNANDES LTDA.	405.905.643	35.914	2.873
SUPERMERCADO BAHAMAS LTDA.	397.979.384	24.413	1.993
D'AVÓ SUPERMERCADOS LTDA.	394.149.650	34.000	1.669
GIASSI & CIA. LTDA.	386.710.656	28.076	1.838
SUPERMERCADO NORDESTÃO LTDA.	347.220.472	14.701	1.367
REALMAR DISTRIBUIDORA LTDA.	278.755.563	20.585	2.085
SUPERMERCADO MODELO LTDA.	271.168.108	26.854	2.194
AM/PM COMESTÍVEIS LTDA.	269.821.182	24.684	2.710
SUPERMERCADOS IRMÃOS LOPES LTDA.	248.585.812	18.443	1.281
NAZARÉ COMERCIAL DE ALIMENTOS E MAGAZINES LTDA.	231.183.502	14.000	1.295
FORMOSA SUPERMERCADOS E MAGAZINE LTDA.	223.706.399	12.000	1.388
PETROBRAS DISTRIBUIDORA S/A	218.337.838	48.988	3.700
COMERCIAL DELTA PONTO CERTO LTDA.	205.183.214	16.899	1.604

Fonte: Abras – Associação Brasileira de Supermercados.

Pede-se:

- Aplique a técnica de análise de conglomerados hierárquicos, utilizando a distância quadrática euclidiana e o método *Nearest Neighbor*. Antes, padronize as variáveis pelo método Z scores. Analise todas as saídas do processamento (*hierarchical cluster* e *K-means*) e ofereça recomendações quanto ao agrupamento das empresas, justificando sua decisão.
- Refaça o item (a), porém agora com o método *Between Groups* e distância quadrática euclidiana.
- Refaça o item (a), porém agora com método *Furthest Neighbor* e a correlação de Pearson (antes, padronize as variáveis pelo método de média 1).
- Refaça o item (a), porém agora sem os três maiores *players*, identificando os elementos de cada *cluster* no gráfico Icycle, se a solução escolhida for a de quatro conglomerados.
- Para os resultados do item (d), a partir do esquema de aglomeração, descreva passo a passo as agregações dos elementos nos conglomerados até o estágio 12 (inclusive).
- Refaça o item (a), porém agora sem os três maiores *players*, identificando os elementos de cada *cluster* no gráfico Icycle, se a solução escolhida for a de seis conglomerados.
- Para os resultados do item (f), identifique os elementos de cada *cluster* no dendrograma, se a solução escolhida for a de três conglomerados.

6.11. RESUMO

A análise de conglomerados, de agrupamentos ou de *clusters*, é uma técnica multivariada exploratória que busca agrupar os elementos conforme a estrutura “natural” dos dados. Ou seja, é uma técnica que visa segregar elementos ou variáveis em grupos homogêneos internamente, heterogêneos entre si e mutuamente exclusivos, a partir de determinados parâmetros, conforme uma medida de distância ou similaridade. Neste sentido, tem por objetivo principal definir a estrutura dos dados de maneira a alocar as observações mais parecidas no mesmo grupo.

Trata-se de uma técnica multivariada que não requer a elaboração de premissas para sua validade, visto que consiste apenas em medidas de similaridade ou distâncias. Entretanto, é importante destacar que ela

não diferencia as variáveis relevantes das irrelevantes, o que significa que a inclusão de variáveis irrelevantes tende a distorcer os resultados.

Há basicamente dois tipos de procedimentos para formação de agrupamentos: hierárquicos e não hierárquicos.

Os procedimentos hierárquicos são utilizados por pesquisadores quando não há possibilidade de se saber, *a priori*, quantos *clusters* são formados por determinada amostra de observações, por meio de alguns critérios (variáveis) preestabelecidos. Assim, este procedimento é exploratório, já que muitas soluções são obtidas e o pesquisador consegue ter apenas indícios de quantos *clusters* são formados com a utilização de alguma medida de similaridade e segundo algum método de agrupamento.

Em relação às medidas de similaridade, as quais representam a base do processo de agrupamento, três métodos merecem destaque: medidas correlacionais, medidas de distância e medidas de associação, sendo que as duas primeiras requerem dados métricos, enquanto a terceira é destinada ao tratamento de dados não métricos (nominais ou ordinais).

Os métodos de agrupamento podem ser aglomerativos ou divisivos. No método aglomerativo, cada objeto começa com seu próprio agrupamento e novos agrupamentos são realizados por similaridade. No método divisivo, por outro lado, todas as observações começam em um grande agregado, sendo separadas as observações mais distantes de forma sucessiva. Entre os métodos aglomerativos, destacam-se: Ligação individual (*Single Linkage* ou *Nearest Neighbor*), Ligação completa (*Complete Linkage* ou *Furthest Neighbor*), Ligação média (*Average Linkage* ou *Between Groups*), Centróide e Ward.

Os procedimentos de agrupamentos não hierárquicos (entre eles, o *K-means*), diferentemente dos métodos hierárquicos, necessitam de informações iniciais a respeito do número de *clusters* e auxiliam na determinação de quais variáveis são significantes para a discriminação dos grupos. A informação do número de *clusters* pode ser proveniente dos procedimentos hierárquicos de forma exploratória, razão pela qual frequentemente os procedimentos não hierárquicos são utilizados após a aplicação da análise de conglomerados hierárquicos.

6.12. QUESTÕES COMPLEMENTARES

- a) Defina análise de conglomerados.
- b) Qual é o objetivo da técnica de análise de conglomerados?
- c) Indique três situações de pesquisa nas quais poderia ser utilizada a técnica de análise de conglomerados?
- d) Quais são as principais medidas de similaridade ou de distância utilizadas em análise de conglomerados?
- e) Discorra brevemente sobre os principais métodos hierárquicos utilizados.
- f) Quais as vantagens e desvantagens do procedimento hierárquico *versus* o procedimento não hierárquico? Em quais circunstâncias um procedimento é preferível a outro?
- g) Explique os principais pontos relevantes que devem ser observados pelo pesquisador na escolha dos métodos hierárquicos.
- h) Os métodos hierárquicos e não hierárquicos podem ser considerados complementares? Justifique sua resposta.
- i) Como o pesquisador pode identificar e decidir o número de grupos formados pela técnica de análise de conglomerados?
- j) Em quais circunstâncias devem-se padronizar as variáveis antes de utilizá-las na análise de conglomerados? Qual o seu impacto?