UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Natália Maria Puggina Bianchesi

Estudo Comparativo entre Métodos de Agrupamento Clássicos e Redes Neurais Artificiais através de Planejamento de Experimento

Itajubá, Janeiro de 2020

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Natália Maria Puggina Bianchesi

Estudo Comparativo entre Métodos de Agrupamento Clássicos e Redes Neurais Artificiais através de Planejamento de Experimento

Dissertação submetida ao programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para obtenção do Título de Mestre em Ciências em Engenharia de Produção.

Área: Engenharia de Produção

Orientador: Prof. Dr. Pedro Paulo Balestrassi

Co-orientador

Janeiro de 2020
Itajubá

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Natália Maria Puggina Bianchesi

Estudo Comparativo entre Métodos de Agrupamento Clássicos e Redes Neurais Artificiais através de Planejamento de Experimento

Banca examinadora:

Itajubá
2020

DEDICATÓRIA

A Deus, aos meus pais Natálio e Lúcia, ao meu marido Pedro, aos meus familiares e aos meus amigos.

AGRADECIMENTOS

A Deus, por sempre estar ao meu lado nos momentos de felicidades e dificuldades.

Aos meus pais, Natálio e Lúcia, pelo incentivo, pelo apoio e por tudo o que já fizeram e ainda fazem por mim.

Ao meu marido, Pedro, pelo apoio incondicional, pelo companheirismo, pela paciência e amor em todos os momentos.

Sou muito grata ao professor Pedro Paulo Balestrassi pela confiança, pela orientação, pelos conselhos profissionais e, sobretudo, pela amizade. Obrigado pela oportunidade de fazer parte de um time que proporciona aprendizado constante, que se ajuda e que se dedica para alcançar bons resultados.

Aos meus amigos e colegas de mestrado: Marina, Estevão, Simone, José Giglio e outros, que contribuíram muito com ideias, sugestões e críticas.

Também agradeço aos outros professores e amigos do IEPG: Anderson, Turrioni, Carlos Mello, Gonzaga, José Henrique, e outros, por toda a ajuda.

Por fim, agradeço à UNIFEI e à CAPES, ao CNPq, à FAPEMIG e à FUPAI pelo apoio financeiro e estrutural à pesquisa brasileira viabilizando a realização deste trabalho e de muitos outros.

EPÍGRAFE

“Feliz aquele que transfere o que sabe e aprende o que ensina”.
Cora Coralina

RESUMO

A análise de *cluster* é uma técnica de mineração de dados multivariada amplamente usada em diversas áreas. Destina-se a agrupar automaticamente os *n* elementos da base de dados em *k clusters*, utilizando-se apenas das informações das variáveis de cada caso. No entanto, a precisão dos agrupamentos finais depende do método de *clustering* utilizado. Neste artigo, apresenta-se uma avaliação do desempenho dos principais métodos de análise de *cluster*: *Ward, K-means* e *Self-Organizing Maps.* Diferentemente de muitos estudos publicados na área, os conjuntos de dados foram gerados através de um Planejamento de Experimentos (DOE), de modo a simular diferentes estruturas de dados possíveis. Considerou-se o número de variáveis, número de *clusters*, tamanho da amostra, partição dos *clusters*, sobreposição dos *clusters*, e a presença de *outliers*, como os fatores do DOE. Os conjuntos de dados foram analisados por cada método de *clustering* e suas partições finais foram comparadas através do *Attribute Agreement Analysis.* Os resultados mostraram que o número de *clusters*, a sobreposição, e a interação entre o número de variáveis e o tamanho da amostra afetam significativamente todos os métodos estudados. Além disso, é possível afirmar que os métodos estudados não apresentam diferenças estatisticamente significativas, com um nível de significância de 5%, e não é possível classifica-los por desempenho.

Palavras-chave: Métodos de *Clustering*, K-*means*, *Self-Organizing Maps*, *Ward.*

ABSTRACT

Cluster analysis is a multivariate data mining technique that is widely used in several areas. It aims to group automatically the n elements of the database into k clusters, using only the information of the variables of each case. However, the accuracy of the final clusters depends on the clustering method used. In this paper, we present an evaluation of the performance of main methods for cluster analysis as Ward, K-means and Self-Organizing Maps. Differently from many studies published in the area, we generated the datasets using the Design of Experiment (DOE) technique, in order to achieve reliable conclusions about the methods through the generalization of the different possible data structures. We considered the number of variables and clusters, dataset size, sample size, cluster overlapping, and the presence of outliers, as the DOE factors. The datasets were analyzed by each clustering method and the clustering partitions were compared by the Attribute Agreement Analysis, providing invaluable information about the effects of the considered factors individually and about their interactions. The results showed that, the number of clusters, overlapping, and the interaction between sample size and the number of variables significantly affect all the studied methods. Moreover, it is possible to state that the methods have similar performances, with a significance level of 5%, and it is not possible to affirm that one outperforms the others.

Key words: Clustering methods, K-means, Self-Organizing Maps, Ward.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1 Contextualização

No decorrer dos últimos anos, verificou-se um crescimento significativo da quantidade de dados armazenados. Avanços nas tecnologias de armazenamento de dados, o aumento na velocidade e capacidade dos sistemas, o barateamento dos dispositivos de armazenamento e a melhoria dos sistemas gerenciadores de banco de dados e data *warehouse*, têm permitido transformar essa enorme quantidade de dados em grandes bases de dados (FAYYAD et al., 1996).

Então, surgiu a necessidade de se explorar esses dados para extrair informações e conhecimentos implícitos, a serem empregados na tomada de decisões (DONI, 2004). Para isso, utiliza-se técnicas de mineração de dados (MA; CHEN; CHEN, 2017), onde diversas ferramentas computacionais são aplicadas na busca de padrões de dados. Essas ferramentas empregam técnicas como indução, classificação, analise de *cluster*, regressão, redes neurais, redes bayesianas, algoritmos genéticos, entre outras. Neste estudo, serão abordados alguns métodos de análise de cluster.

A análise de *cluster*, também conhecida como análise de agrupamento, clusterização ou classificação não supervisionada, é uma das técnicas mais utilizadas no processo de mineração de dados para descoberta de agrupamentos e identificação de importantes distribuições e padrões para entendimento dos dados (HALDIKI, 2001).

A análise de agrupamento, é uma técnica que, somente a partir das informações das variáveis, tem por objetivo separar um conjunto de objetos em diferentes grupos, em que cada um deve conter objetos semelhantes segundo alguma função de distância estatística e, ao mesmo tempo, serem dissimilares dos objetos de outros *clusters*. Ou seja, o resultado obtido a partir da aplicação desse método é um conjunto de grupos com coesão interna e isolamento externo (EVERITT; LANDAU; LEESE, 2001).

O primeiro registro publicado sobre um método de clusterização foi oi apresentado por Sorensen (1948). Desde então, métodos de análise de agrupamento vêm sendo desenvolvidos devido à necessidade de análise da grande quantidade de dados coletados nas diversas áreas do conhecimento.

Everitt, Landau e Lesse (2001) apresentam aplicações de clusterização em diversas áreas, como marketing, medicina, educação e biologia. Exemplos de clusterização encontrados em artigos demonstram também essa diversidade de aplicação. Observa-se o uso de clusterização para identificar características de pessoas com tentativa

de suicídio (KIM et al., 2018); facilitar o diagnóstico e tratamento do câncer (Yu et al., 2015); identificar padrões residenciais e sociais de adultos sem-teto (LEE et al., 2016); e, também, em aplicações na área de engenharia de produção como, por exemplo, para planejamento de produção (MACCHIAROLI, RIEMMA, 1994; NACHTWEY; RIEDEL; MUELLER, 2009), e para análise de portfólios de produtos (HOCHDORFFER; LAULE; LANZA, 2017).

Em geral, sempre que é necessário classificar grandes quantidades de informações em um pequeno número de categorias, a análise de *clusters* pode ser útil (DONI, 2004).

Pesquisadores são frequentemente confrontados com a tarefa de classificar dados em estruturas significativas. No entanto, a precisão da partição final depende do método usado para agrupar os objetos, ou seja, a escolha inadequada do método pode comprometer os resultados obtidos. Há, portanto, uma crescente preocupação em fazer com que os métodos se adequem a determinadas situações e sejam também de menor complexidade.

Então, o objetivo deste trabalho é apresentar um estudo comparativo entre o desempenho dos principais métodos de agrupamento tradicionais e Redes Neurais Artificias (RNA). No entanto, diferentemente de outros estudos, este não considera uma abordagem usual de tentativa e erro para generalizar conjuntos de dados, o que poderia levar a conclusões restritas sem uma análise generalizada. Para isso, considerou-se a técnica de Design de Experimentos (DOE), que pode alcançar resultados mais confiáveis.

Usando a técnica DOE, é possível simular conjuntos de dados sintéticos e avaliar o desempenho de cada método, identificando os parâmetros que mais afetam seus resultados, e verificar a possibilidade de classificação do melhor método.

1.2 O problema de pesquisa

Para demonstrar a complexidade do problema de clusterização, uma definição formal é encontrada em Hruschka e Becken (2001). Dado um conjunto de p elementos $X = \{X_1, X_2, \dots, X_P\}$, o problema de agrupamento consiste na obtenção de um conjunto de k grupos, $G = \{G_1, G_2, \dots, G_k\}$, tal que os elementos contidos em um grupo G possuam uma maior similaridade entre si do que com os elementos de qualquer um dos demais grupos do conjunto G. O conjunto G é considerado um agrupamento com k grupos caso as seguintes condições sejam satisfeitas:

$$\bigcup_{i=1}^{k} G_i = X$$
$$G_i \neq \emptyset, para\ 1 \leq i \leq k \qquad (1.1)$$
$$G_i \cap G_j = \emptyset, para\ 1 \leq i, j \leq k\ e\ i \neq j$$

Enfatiza-se, por essas condições, que um elemento não pode pertencer a mais de um grupo e que cada grupo tem que ter ao menos um elemento.

A quantidade de grupos k pode ser conhecida ou não. Caso o valor de k seja desconhecido, o problema é conhecido como "problema de agrupamento automático" e a obtenção do valor de k faz parte do processo de solução do problema. Caso contrário, o k deve ser fornecido como parâmetro para a solução, e o problema é conhecido na literatura como "problema de *k*-agrupamento" (FASULO, 1999). Este último caso é o qual será abordado durante o desenvolvimento desse trabalho.

Em um *k*-agrupamento, o número total de diferentes formas de agrupamento de p elementos de um conjunto em *k* grupos, equivale à função $N\ (p, k)$:

$$N\ (p,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i} (k-i)^p \qquad (1.2)$$

Onde: *k* é o número de agrupamentos, *p* é o número total de objetos e *i* é o agrupamento em questão.

Considerando a equação acima, pode-se observar que o número de soluções possíveis para um problema de k-agrupamento tem um crescimento exponencial. Para combinar 10 elementos em 2 grupos e 100 elementos em 2 grupos, temos respectivamente 511 e 6,33825 x 10^{29} diferentes maneiras de combinar os elementos. Dessa maneira, nota-se a complexidade em encontrar a melhor solução de agrupamento dentro das possíveis soluções disponíveis e, consequentemente o método de clusterização mais indicado para encontrar tal solução.

De acordo com Bussab *et al.* (1990), a escolha de um método de clusterização exige o conhecimento de suas propriedades, aliado aos objetivos da pesquisa.

Portanto, como comparar de forma mais generalizada os métodos de clusterização e auxiliar na escolha do mais adequado para aplicação?

1.3 Relevância

A análise de *cluster* é explorada por diversos autores por meio de diversos métodos. Entretanto, ao se tratar da comparação dos métodos de agrupamento, encontra-se uma

lacuna na literatura. A maioria dos artigos estão limitados apenas em análises realizadas com base em bancos de dados específicos. Os poucos artigos que utilizam dados simulados apresentam limitação no processo de geração de dados, nas variáveis selecionadas e nos métodos de agrupamento utilizados.

Uma análise bibliométrica na base de dados *Web of Science* e *Scopus* pôde identificar tal deficiência. As pesquisas foram realizadas em 16 de fevereiro de 2020.

Na base de dados *Web of Science*, ao se pesquisar no título dos trabalhos as palavras *comparison, cluster* e *method*, foram encontrados 321 artigos de diversos temas.

Porém, ao substituir o termo *method* por *neural network* na busca, a quantidade de artigos encontrados reduziu-se para 17, sendo que a maior parte está relacionada com aplicações específicas de clusterização. Desta maneira, apenas três artigos têm como objetivo principal realizar a comparação dos métodos de maneira generalizada, visando encontrar o melhor método independentemente da aplicação. Esses artigos comparam redes neurais e agrupamento hierárquico (MANGIAMELI; CHEN; WEST, 1996), redes neurais e agrupamento não hierárquico (BALAKRISHNAN et al., 1994) e, apenas um artigo compara redes neurais, agrupamentos hierárquicos e não hierárquicos (WALLER et al., 1998).

Ainda, ao pesquisar termos similares como, por exemplo, substituir o termo *comparison* por *comparing*, apenas o artigo de Mingoti e Lima (2006) se destacou por abordar a comparação de redes neurais, *Fuzzy, k-means* e métodos hierárquicos.

Em geral, a pesquisa na *Web of Science*, permitiu identificar que a quantidade de publicações sobre esse tema é pequena, sendo que entre os anos de 2005 e 2008 ocorreram a maior parte das publicações. O número de citações também foi alto nesse mesmo período, e apresenta um crescimento a partir de 2017. A Figura 1.1 ilustra o comportamento descrito.

Em outra pesquisa, realizada na base de dados *Scopus*, com os termos *comparison*, *cluster* e *method*, encontrou-se 159 artigos diversos. E, em seguida, ao limitar a pesquisa inserindo o termo *neural network*, obteve-se apenas nove artigos, dos quais apenas o artigo de Waller *et al.* (1998) tem o objetivo de comparar os métodos de análise de *cluster*. Nenhum outro trabalho foi encontrado ao se pesquisar variações sobre os termos destacados.

Apesar dos trabalhos encontrados na literatura apresentarem resultados interessantes, nenhum avaliou o efeito do tamanho da amostra e o efeito do tamanho da amostra dos clusters serem iguais ou diferentes; os resultados não consideram interações

entre fatores; nenhum dos trabalhos comprovou os resultados por meio de resoluções de exemplos; e, principalmente, nenhum estudo utilizou Planejamento de Experimento (DOE) para simular os bancos de dados e avaliar o desempenho dos métodos. Tais trabalhos serão detalhados posteriormente.

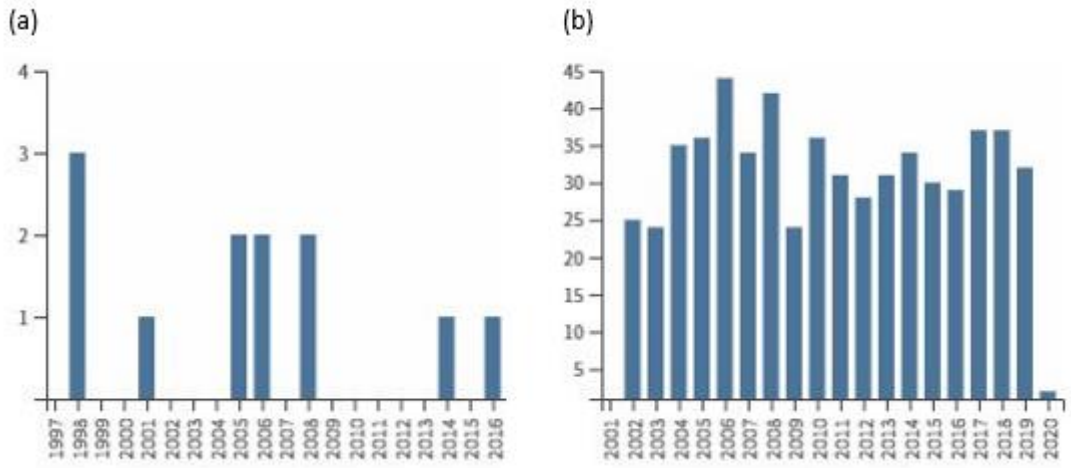

Figura 1.1- Publicações (a) e citações (b) na Web of Science.

1.4 Justificativa

O agrupamento de dados é um problema que pode estar presente em inúmeras situações de diversas naturezas. Por isso, diversos métodos e algoritmos de agrupamento foram desenvolvidos como, por exemplo, métodos hierárquicos e não hierárquicos, os quais possuem diversas ramificações devido a diferentes métodos de cálculo de similaridade. Com o surgimento e aprimoramento das redes neurais, mais um algoritmo mostrou-se capaz de solucionar problemas de agrupamento, o algoritmo de *Kohonen*, ou *Self Organizing Maps* (KOHONEN, 1997).

Portanto, para entender qual método é melhor aplicável em diferentes situações, pode-se compará-los através de testes estatísticos na resolução de diferentes estruturas de banco de dados. Poucos estudos foram desenvolvidos com esse objetivo (Mingoti e Lima (2006), enquanto que a busca pelo melhor método de clusterização é constantemente abordado.

1.5 Objetivos

O objetivo geral deste trabalho é o de comparar os métodos de análise de *cluster* clássicos, *Ward e K-means,* e Rede Neural Artificial (RNA), e analisar qual o método mais adequado para ser aplicado em diferentes situações. De modo a cumprir com o objetivo geral deste trabalho, têm-se como objetivos específicos:

1. Gerar dados simulados para diferentes cenários e estruturas de dados por meio do uso do DOE;
2. Realizar análise de agrupamento, comparar os resultados obtidos e extrair conclusões válidas sobre os efeitos dos fatores e suas interações;
3. Comprovar os resultados obtidos por meio da resolução de exemplos do livro "*Cluster Analysis*" (EVERITT; LANDAU; LEESE, 2001).

1.6 Estrutura do trabalho

O capítulo 2 contém uma fundamentação teórica sobre análise de *cluster*, métodos clássicos de agrupamento e RNA, a qual é aprofundada no método SOM (*Self-Organization Maps*). A fundamentação teórica também contém uma breve explicação sobre simulação de dados por DOE, visto que este será o processo utilizado para obtenção dos bancos de dados.

O capítulo 3 apresenta o método de pesquisa, abrangendo a classificação da pesquisa e o procedimento realizado, o qual inclui a geração dos dados, análise de *clusters*, análise estatísticas, interpretação e validação dos resultados.

O capítulo 4 contém discussões e conclusões sobre o trabalho e, também, sugestões para trabalhos futuros. Em seguida, são apresentados as Referências e os Apêndices, os quais contêm dados e informações necessárias à compreensão dos resultados obtidos.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados aspectos gerais da análise de agrupamento, trazendo algumas definições e classificações. Depois disso, será abordado, especificamente, os métodos clássicos de agrupamento e a metodologia de Redes Neurais aplicada em agrupamentos. Em seguida, é realizado um comparativo entre os métodos e destacado suas vantagens se desvantagens. Esta seção também apresentará trabalhos anteriores sobre comparação de métodos de agrupamento, e por fim, será abordada a metodologia de Planejamento de Experimentos para a geração de dados simulados.

2.1 Métodos de Agrupamento

A técnica multivariada de análise de agrupamento é uma maneira de se obter grupos homogêneos, por um esquema que possibilite reunir os indivíduos em um determinado número de grupos, de modo que exista grande homogeneidade dentro de cada grupo e heterogeneidade entre eles (JOHNSON e WICHERN, 1992; CRUZ e REGAZZI, 1994).

Porém, a busca pela melhor solução dentro do espaço de soluções viáveis, como exposto na Equação 1.2, é um problema complexo. Segundo Rodriguez (2009), a avaliação exaustiva de todas as configurações possíveis é computacionalmente inviável para problemas de médio ou grande porte, restringindo com isso o uso de métodos exatos. Por isso, segundo mesmo autor, métodos aproximados têm sidos propostos com frequência, os quais fornecem soluções sub-ótimas com significativa redução da complexidade na solução do problema.

Os algoritmos existentes para a solução de problemas de agrupamento podem ser classificados, de forma geral, em métodos hierárquicos e métodos não hierárquicos, também conhecido como particionamento, como descrito na Tabela 2.1 (FASULO, 1999).

Tabela 2.1- Métodos de *Clustering*. Fonte: Própria autora.

Hierarchical	***Non-Hierarchical***
Single Linkage	*Self-Organizing Maps (SOM)*
Complete Linkage	*Fuzzy clustering algorithms*
Centroid Linkage	*K-medoids*
Ward	*K-means*

Entre os métodos da Tabela 2.1, escolheu-se estudar um método hierárquico, um não hierárquico tradicional e o método SOM, que é uma abordagem de Rede Neural Artificial (RNA) mais recente para problemas de *cluster*.

O método hierárquico escolhido é o Ward. Entre os métodos hierárquicos de aglomeração, o método de Ward é o único algoritmo baseado em um critério clássico de soma de quadrados, grupos produtores que minimizam a dispersão dentro do grupo a cada fusão binária, e portanto, é visto como uma das melhores técnicas para medir distâncias entre clusters (MURTAGH; LEGENDRE, 2014). Mangiameli, Chen e West (1996) sugerem que o método *Ward* sempre deve ser empregado, pois apresenta melhores resultados entre os métodos hierárquicos. Mingoti e Lima (2006) acrescentam que o método Ward é mais estável e mais fácil de implementar.

O método tradicional não hierárquico que será usado é o *K-means*. Han e Kamber (2001) afirmam que os mais bem conhecidos e geralmente usados métodos de particionamento são o *k-means*, o *k-medoids*, e suas variações. O *k-means*, também conhecido como *k*-médias, é o mais popular (FUNG, 2001) e um dos algoritmos mais utilizados para análise de agrupamento devido à facilidade de implementação, simplicidade e eficiência (JAIN; MURTY; FLYNN, 1999).

Nas seções seguintes, os métodos *Ward*, *K-means* e SOM serão respectivamente explicados, apresentando seus conceitos e procedimentos.

2.1.1 Método Hierárquico

Os métodos hierárquicos são técnicas simples onde os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos (EVERITT; LANDAU; LEESE, 2001). Essa representação é apresentada através de um diagrama bidimensional em forma de árvore, chamado Dendograma, Figura 2.1. Pelo Dendograma é possível visualizar a partição dos dados e a formação dos agrupamentos em cada estágio onde ela ocorreu e com que grau de similaridade.

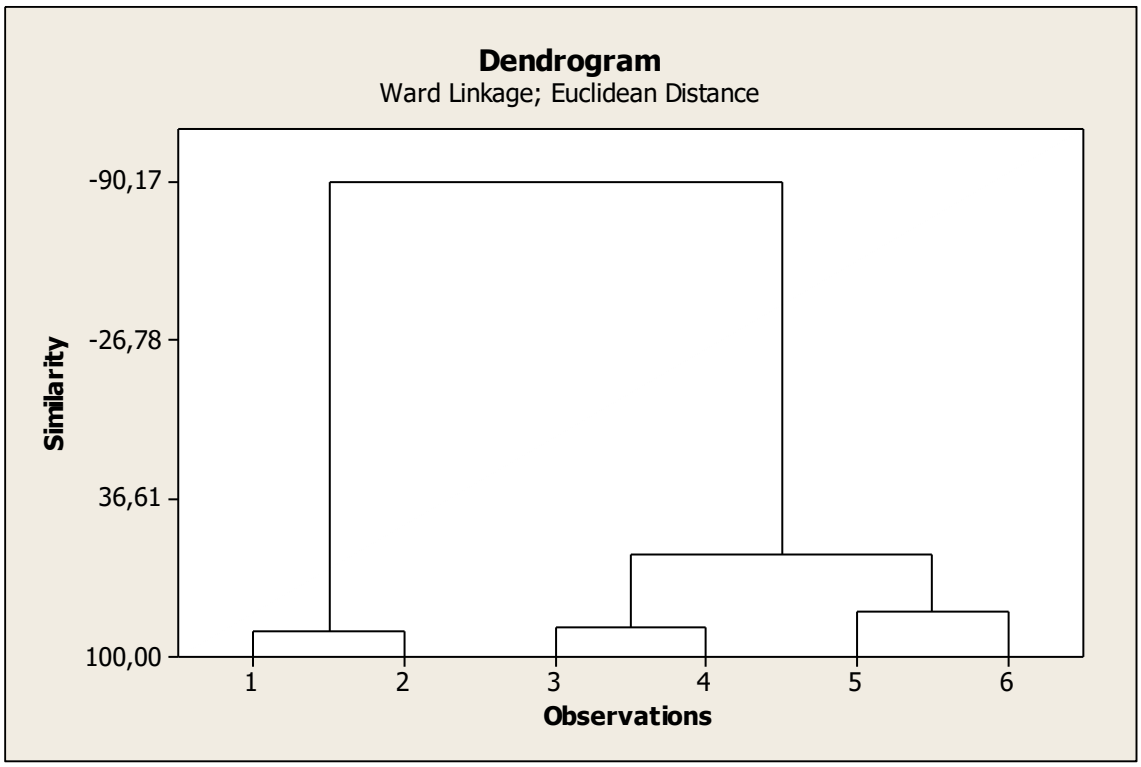

Figura 2.1- Exemplo de Dendograma. Fonte: Própria autora

As técnicas hierárquicas envolvem basicamente duas etapas: a primeira se refere à estimação de uma medida de similaridade entre os indivíduos, e a segunda refere-se à adoção de uma técnica de formação de grupos (SANTANA e MALINOVSKI, 2002).

Durante a primeira etapa, as métricas mais utilizadas para medir a similaridade são a distância Euclidiana, distância Mahalanobis e distância de Manhattan (NOVAES, 2002).

A distância Euclidiana é uma medida invariante a translações, porém assume covariâncias iguais entre as classes (COSTA, 1999). A distância de Manhattan é uma simplificação da distância Euclidiana e, segundo Kugler *et al.* (2003), é um método mais

eficiente para aplicações em tempo real devido sua simplicidade. Já a distância Mahalanobis tem características que visa suprir muitas das limitações da distância Euclidiana, porém pode ser bastante difícil determinar precisamente as matrizes de covariância, e o custo computacional cresce muito com o número de variáveis envolvidas (COSTA, 1999). Portanto, optou-se por utilizar a distância Euclidiana no desenvolvimento deste trabalho.

A distância Euclidiana é a distância geométrica no espaço dimensional. Para calcular a distância euclidiana entre cada par de objetos *i* e *j*, objeto *i*, *i* = 1,2,...,*n*, objeto *j*, *j* = 1,2,...,*n*, caracterizado pelos atributos $x_j = (x_{j1} \quad ..., \quad x_{jp})$, sendo *p*, *p* = 1,2,...,*k*, o número de características, utiliza-se (MONTEGOMERY, 2005):

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + \left(x_{ip} - x_{jp}\right)^2}, \qquad i \neq j \tag{2.1}$$

Na segunda etapa, a técnica de formação de grupos que será utilizada, como mencionado anteriormente, é o método *Ward.* O método de *Ward* é uma técnica aglomerativa, que procura por partições que minimizem a perda associada a cada agrupamento (Ward, 1963). Essa perda é quantificada pela diferença entre a soma dos erros quadráticos de cada padrão e a média da partição em que está contido. A soma dos erros quadráticos para cada agrupamento é definida como:

$$ESS_k = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{2.2}$$

Onde: *k* é o agrupamento em questão, *n* é o número total de objetos do agrupamento *k* e x_i é o *i*-ésimo objeto do agrupamento *k*.

O procedimento geral do método aglomerativo pode ser descrito em poucos passos (EVERITT; LANDAU; LEESE, 2001):

- Início: considera-se cada elemento como um único grupo;
- Na primeira etapa, cada elemento é comparado entre si através de uma medida de similaridade, as quais são armazenadas em uma matriz de similaridade;
- Na segunda etapa, os grupos formados são comparados entre si através da técnica de Ward e, em seguida, os grupos mais similares se unem;
- Este procedimento se repete diversas vezes até que todos os elementos estejam agrupados conforme o número de grupos desejado;

- Apenas dois grupos podem ser unidos em cada estágio e eles não podem ser separados posteriormente.

2.1.2 Método Não Hierárquico

Os métodos de particionamento, ou não hierárquicos, buscam encontrar a melhor partição dos *n* elementos em um número *k* de grupos escolhido a priori. A ideia central da maioria dos métodos por particionamento é escolher uma partição inicial dos elementos e, em seguida, alterar os membros dos grupos para obter-se a melhor partição (ANDERBERG, M.R., 1973).

O algoritmo *k-means*, além da definição do número de grupos *k*, exige a definição inicial do centro de cada grupo k_1, k_2,..., k_p no espaço. O centro do grupo é chamado de centróide, que é o ponto médio do grupo. Sua definição pode ser realizada automaticamente pelo algoritmo ou ser indicada pelo usuário.

Portanto, a técnica de formação de grupo utilizada é em relação ao centroide. A distância dos centroides é dada pela fórmula:

$$D(C_1, C_2) = d(\mu_1, \mu_2) \tag{2.3}$$

Onde: μ_1 e μ_2 são respectivamente os centróides dos agrupamentos C_1 e C_2 e $d(\mu_1, \mu_2)$ é a distância entre eles.

A medida de distância utilizada da Equação 2.3, é a distância Euclidiana, pois é a mais recomendada por apresentar resultados mais precisos (KAUFFMAN, 1990).

Os passos básicos do algoritmo *k-means* é ser descrito por (RODRIGUEZ, 2009):

- Selecionar *n* elementos para serem os centroides iniciais dos *k* grupos;
- Cada elemento é associado a um grupo, esse elemento é alocado no grupo em que a dissimilaridade entre ele e o centroide do grupo for menor que a dissimilaridade entre este e os demais centroides;
- Os centroides dos grupos são recalculados, redefinindo cada um, em função dos atributos de todos os elementos pertencentes ao grupo;
- Retorna ao passo 2 até que os centros dos grupos se estabilizem.

2.1.3 Redes Neurais Artificiais

Uma rede neural artificial (RNA) é uma estrutura computacional modelada com os princípios dos processos biológicos (ADYA; COLLOPY, 1998). Esse método foi inspirado na estrutura de funcionamento da rede de neurônios do cérebro pelo fato da

facilidade e eficácia com que o cérebro realiza tarefas difíceis e complexas e principalmente pela sua capacidade de aprender.

O primeiro modelo de redes neurais foi proposto no trabalho de McCulloch e Pitts (1943), no qual descrevem o cálculo lógico das redes neurais que unificava o estudo de neurofisiologia e de lógica matemática. Em seguida, Hebb (1949) propôs um ajuste no modelo adicionando pesos aos valores de entrada. Adiante várias melhorias foram agregadas ao modelo, mas apenas nos anos 80 o RNA começou a ser mais utilizado.

O objetivo da RNA não é replicar a operação do sistema biológico, mas fazer o uso do que é conhecido sobre a funcionalidade das redes biológicas para resolver problemas complexos (BASHEER; HAJMEER, 2000).

Segundo Haykin (2001), uma rede neural é um processador maciçamente paralelamente distribuído, constituído de unidades de processamento simples, os neurônios, que se unem por meio de conexões sinápticas e têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:

- O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem;
- Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Uma simplificação do modelo de neurônio artificial apresentado por Haykin (2001) está representada na Figura 2.2.

Este modelo é composto por três elementos básicos:

- Um conjunto de sinapses, cada uma delas caracterizada por um peso característico;
- Um combinador linear para somar os sinais de entrada, ponderados pela respectiva sinapse do neurônio;
- Uma função de ativação para limitar a amplitude de saída do neurônio. A função de ativação limita a faixa de amplitude permitida do sinal de saída a um valor finito.

Pode-se descrever matematicamente um neurônio *k* com as equações 2.4 e 2.5:

$$u_k = \sum_{j=0}^{m} w_{kj} x_j \tag{2.4}$$

$$y_k = \varphi(u_k + b_k) \tag{2.5}$$

Onde: x_1, x_2,..., x_m são sinais de entrada, w_{k1}, w_{k2},..., w_{km} são os pesos sinápticos do neurônio k, u_k é a saída do combinador linear, b_ké o bias (desvio), $\varphi(.)$ é a função de ativação e y_k é a função de saída do neurônio (HAYKIN, 1998).

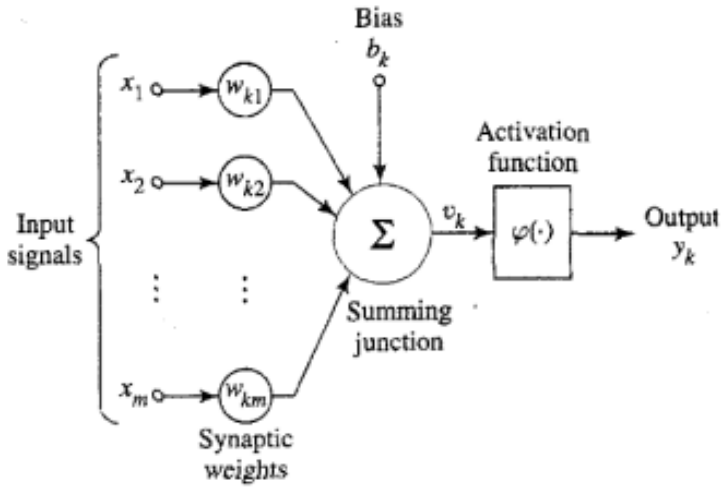

Figura 2.2- Modelo matemático de um neurônio. Fonte: Haykin (2001)

O neurônio recebe sinais de entrada (*x*), que podem ser informações de outros neurônios ou estímulos externos, através de conexões simuladas por pesos (*w*). O efeito de um sinal é determinado pela multiplicação do valor do sinal pelo peso da conexão correspondente $(x \times w)$. Em seguida, é efetuada a soma dos valores $(x \times w)$ de todas as conexões, resultando em um valor u_k. Esse valor é somado a um elemento polarizador b_k (bias), que tem o efeito de aumentar ou diminuir o argumento da função de ativação. O valor resultante é processado por uma função de ativação, que produz um sinal de saída (*y*).

Combinados diversos neurônios, forma-se uma rede neural artificial. Uma representação simplificada de uma rede neural pode ser vista na Figura 2.3, em que os nós são os neurônios e as ligações fazem a função das sinapses.

Segundo Ferneda (2006), as redes neurais se diferem pela sua arquitetura e topologia, e pela forma como os pesos associados às conexões são ajustados durante o processo de aprendizado.

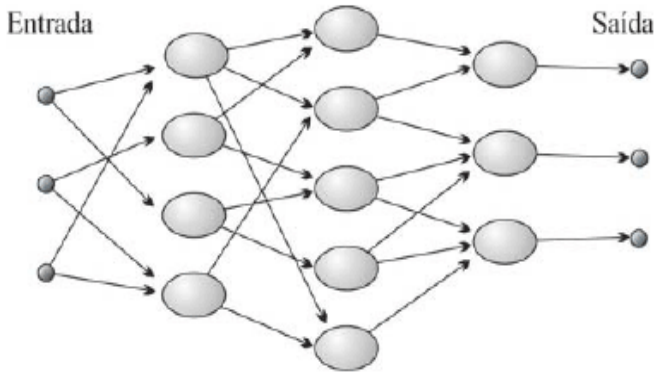

Figura 2.3 - Representação simplificada de uma rede neural artificial. Fonte: Ferneda (2009)

Arquitetura refere-se à disposição dos neurônios, um em relação ao outro. A arquitetura de uma rede neural restringe o tipo de problema no qual a rede poderá ser utilizada, e é definida pelo número de camadas (camada de entrada, camadas de processamento e camada de saída) e pelo tipo de conexão entre os nós (redes alimentadas adiante ou redes recorrentes) (HAYKIN, 2001).

A topologia da rede refere-se às diferentes composições estruturais possíveis com diferentes quantidades de neurônios nas camadas de entrada, intermediária e de saída.

Por processo de aprendizagem entende-se processo pelo qual os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação pelo ambiente no qual a rede está inserida (HAYKIN, 2001). Existem dois fatores a serem considerados no processo de aprendizagem: o paradigma de aprendizagem e o algoritmo de aprendizagem. O paradigma de aprendizagem refere-se ao modo pelo qual uma rede neural se relaciona com o seu ambiente, ou seja, a maneira como a rede aprende através de exemplos provenientes de casos conhecidos. O algoritmo de aprendizagem é determinado pela maneira pela qual a modificação dos parâmetros ocorre.

Existem dois paradigmas de aprendizado: aprendizado supervisionado e aprendizado não supervisionado. No primeiro, a rede recebe conjuntos de exemplos de padrões de entrada e seus correspondentes padrões de saída, tais exemplos podem ser modelos existentes ou dados históricos; a resposta fornecida pela rede é comparada com a saída esperada e o erro verificado é informado à rede para que o algoritmo seja ajustado visando uma melhor aproximação nas respostas futuras. Opondo-se ao supervisionado, o aprendizado não supervisionado, não utiliza conjuntos previamente conhecidos, ou seja, apenas os padrões de entradas são inseridos na rede, a qual tenta progressivamente codificar características dos dados para classifica-los automaticamente.

O algoritmo de aprendizagem são regras bem definidas e é determinado pelo modo como os pesos das conexões são ajustados (BRAGA; CARVALHO; LUDEMIR, 2000). As principais regras de aprendizagem são: aprendizagem por correção de erro, aprendizagem baseada em memória, aprendizagem hebbiana, aprendizagem competitiva e aprendizagem de Boltzmann.

Para problemas de agrupamento, o algoritmo que tem sido amplamente utilizado em pesquisas recentes é o chamado auto-organizado, o qual pode ser também utilizado para projetar e visualizar objetos (KOHONEN, 1997).

2.1.3.1 Aprendizado Auto-organizado (SOM)

O algoritmo de aprendizado auto-organizado também é chamado de Mapa Auto-Organizado de *Kohonen* (*Self-Organizing Maps*, ou *SOM*), desenvolvido pelo finlandês Teuvo Kohonen no começo dos anos 80.

O mapa auto-organizado de Kohonen constitui uma classe de redes neurais artificiais baseadas em aprendizado competitivo e utiliza o treinamento não-supervisionado, em que a rede busca agrupar os dados de entrada baseando-se apenas em suas similaridades.

Um mapa de *Kohonen* é um arranjo de neurônios, geralmente restrito a um espaço uni ou bidimensional, que procura a preservação topológica. A sua topologia possui duas camadas, na qual todas as unidades de entrada encontram-se conectadas a todas as unidades de saída através de conexões sinápticas, Figura 2.4. A quantidade de elementos de entrada depende do banco de dados a ser utilizado e cada unidade de saída representa um *cluster*, o que limita a quantidade de *clusters* ao número de saídas.

Durante o treinamento, a rede determina a unidade de saída que melhor responde ao vetor de entrada; o vetor de pesos é ajustado de acordo com o algoritmo de treinamento. A principal característica desse algoritmo, que se difere dos demais em redes neurais, é a utilização de regras de aprendizado competitivo.

Rodriguez (2009) descreve o aprendizado competitivo como o fato de após os neurônios receberem os padrões de entrada, cada um calcula seu nível de ativação multiplicando o seu vetor de pesos pelo vetor de entrada e, em seguida, apenas o neurônio com maior nível de ativação (neurônio vencedor) terá atividade diferente de zero na saída da rede, ou seja, o padrão de entrada que estiver sendo apresentado à rede provocará a ativação de apenas um neurônio da rede neural.

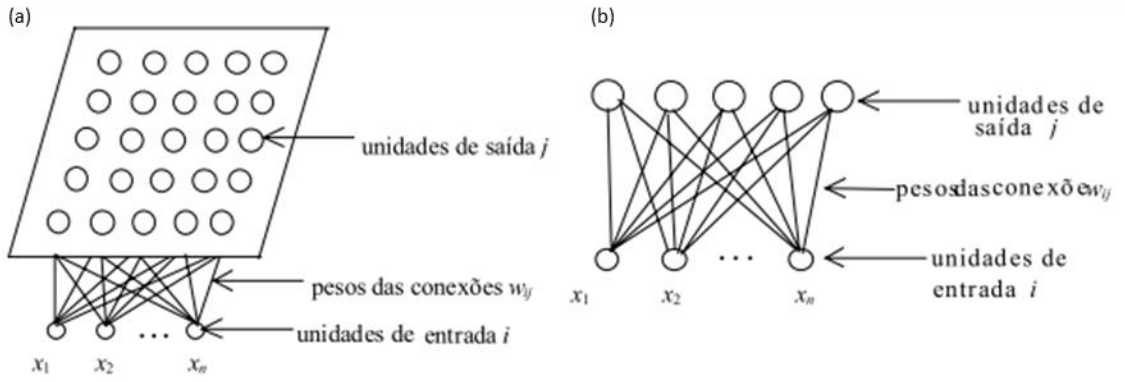

Figura 2.4 -(a) Arquitetura bidimensional. (B) Arquitetura unidimensional. Fonte: Haykin (2008).

Viana (2004) sugere a decomposição do algoritmo de auto-organização em quatro etapas: inicialização do mapa; processo competitivo; o processo cooperativo; adaptação sináptica.

1. Inicialização do mapa

A inicialização consiste na atribuição de um vetor de pesos às conexões entre neurônios das camadas de entrada e saída. A escolha do vetor de pesos pode ser feita atribuindo pequenos valores aleatórios; desta forma, nenhuma ordem prévia é imposta ao mapa.

Assim, após um vetor padrão de entrada ser apresentado à rede, cada neurônio recebe este padrão e calcula o seu nível de ativação. O nível de ativação é representado pelo valor de proximidade entre o vetor de entrada *x* e cada neurônio de saída *j*, e é medido através da distância euclidiana d_j, dado por:

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2 \qquad (2.6)$$

Onde: $x_i(t)$ é a entrada ao neurônio i, no instante de tempo t; $w_{ij}(t)$ é o peso entre o neurônio de entrada *i* e o de saída *j*, no instante *t*.

2. Processo competitivo

No processo competitivo, os neurônios competem entre si através dos níveis de ativação, sendo que apenas um neurônio será o vencedor. O neurônio vencedor é aquele cujo vetor de pesos tiver a menor distância euclidiana com o padrão de entrada, ou seja, tiver um d_j de valor mínimo.

3. Processo cooperativo

O processo cooperativo implica na influência que o neurônio vencedor exerce no estado dos neurônios vizinhos. O neurônio vencedor determina a localização espacial de

uma vizinha topológica de neurônios excitados, fornecendo assim a base para cooperação entre neurônios vizinhos.

O neurônio vencedor tende a excitar mais os neurônios em sua vizinhança imediata, do que aqueles neurônios que estão mais distantes. Assim, a vizinhança topológica ao redor do neurônio vencedor decai suavemente com a distância lateral.

Por exemplo, seja $h_{j,i}$ a vizinhança topológica centrada no neurônio vencedor *i* e circundada por um conjunto de neurônios excitados cooperativos, dos quais um neurônio típico é denotado por *j*. Seja, $d_{j,i}$ a distância lateral entre o neurônio vendedor *i* e o neurônio excitado *j*. Então, pode-se assumir que a vizinhança topológica $h_{j,i}$ é uma função unimodal da distância lateral $d_{j,i}$, tal que satisfaça a dois requerimentos (Haykin, 1999):

- A vizinhança topológica $h_{j,i}$ é simétrica e máxima ao redor do neurônio vencedor *v* definido por $d_{j,i} = 0$;
- A amplitude da vizinhança topológica $h_{j,i}$ decresce monotonicamente com o aumento da distância lateral $d_{j,i}$, decaindo para zero quando $d_{j,i} \to \infty$; condição necessária para convergência. Portanto, se $d_{j,i} = 0$, $h_{j,i}$ obtém valor máximo; e se $d_{j,i} \to \infty$, $h_{j,i} = 0$.

Outra característica importante desse algoritmo é que a largura da vizinhança decresce com o tempo *t*. Portanto, uma escolha para $h_{j,i}$, por exemplo, é a função Gaussiana variante com o tempo, definido na Equação (2.7).

$$h_{j,i}(t) = exp\left(\frac{d_{j,i}^2}{2\sigma^2(t)}\right) \quad , \qquad t = 0,1,2, \ldots \tag{2.7}$$

Onde: $\sigma(t)$ é a "largura efetiva" da vizinhança topológica, conforme Figura 2.5 e definido pela Equação (2.8).

$$\sigma(t) = \sigma_0 exp\left(-\frac{t}{\tau_1}\right) \tag{2.8}$$

Onde: σ_0 é o valor inicial de σ, e τ_1é uma constante de tempo.

Assim, à medida que o tempo *t* (número de iterações) aumenta, a largura $\sigma(t)$ diminui a uma taxa exponencial, e a vizinhança topológica é reduzida.

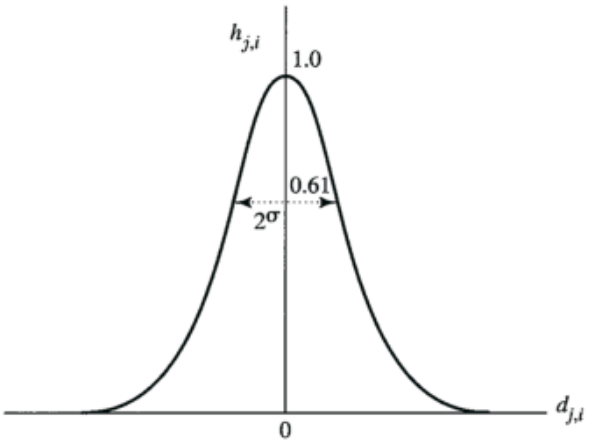

Figura 2.5 - Função de vizinhança Gaussiana. Fonte: Haykin (1999)

4. Adaptação sináptica

Uma vez determinado o neurônio vencedor e seus vizinhos, seus vetores de peso (*w*) são atualizados de modo a se aproximarem espacialmente do padrão de entrada (*x*), porém essa aproximação vai diminuindo à medida que o neurônio correspondente fica mais distante do vencedor (*i*). Os vetores de pesos dos neurônios são modificados de acordo com a Eq. 2.9, para todo neurônio dentro da vizinhança topológica do vencedor *i.*

$$w_j(t+1) = w_j(t) + \eta(t)h_{j,i}(t)(x - w_i(t)) \tag{2.9}$$

O termo η corresponde à taxa de aprendizagem do algoritmo e também tem seu valor alterado a cada iteração, segundo:

$$\eta(t) = \eta(0)exp\left(-\frac{t}{\tau_2}\right), \qquad t = 0,1,2,\ldots \tag{2.10}$$

Onde: $\eta(0)$ é o valor inicial de η que decresce ao longo das iterações *t*, e τ_2 é outra constante de tempo.

Assumindo uma inicialização aleatória, são necessárias duas fases de Adaptação Sináptica: auto-organização e convergência.

Auto-organização:

É nessa fase que ocorre o ordenamento topográfico, ou seja, à medida que a rede vai treinando vetores de neurônios vizinhos, os neurônios sensibilizados com mesmas características começam a se aproximar uns dos outros. Assim, elementos semelhantes vão sendo posicionados próximos entre si, formando um gradiente de características.

A fase de ordenação pode durar 1000 ou mais interações, e deve-se escolher cuidadosamente o parâmetro de aprendizado e da função de vizinhança.

Haykin (1999) sugere que o parâmetro $\eta(t)$ deve iniciar com um valor próximo a 0,1, decrescendo gradualmente, mas permanecendo acima de 0,01. Estes valores são atingidos através dos seguintes parâmetros para a Equação 2.10:

$$\eta(0) = 0.1$$

$$\tau_2 = 1000$$

A função vizinhança $h_{j,i}(t)$deve incluir inicialmente quase todos os neurônios da rede centrados no neurônio vencedor i e, então encolher lentamente com o tempo. Assumindo o uso de um mapa bidimensional, pode-se, então, ajustar o tamanho inicial σ_0 igual ao "raio" do mapa.

Convergência:

Nessa fase ocorre o refinamento do mapa, ou seja, a taxa de aprendizagem η deve permanecer pequena, mas não caindo a zero, caso contrário, a rede poderá ficar presa em um estado metaestável. A função de vizinhança deve englobar apenas o próprio neurônio. Dessa maneira, o mapa converge para uma disposição estável e os agrupamentos não sofrerão mais alterações.

Desse modo, os passos 2 a 4 do algoritmo são repetidos a cada nova apresentação de um elemento ao mapa, ou seja, a cada iteração. Cada elemento apresentado gera um estímulo que é transmitido aos neurônios e aquele que reagir mais fortemente aos estímulos, fica com o elemento.

Além disso, esse neurônio reforça suas ligações com os vizinhos próximos, sensibilizando-os um pouco mais às características do elemento, ou seja, a vizinhança será sensibilizada a um perfil similar, porém cada neurônio reagirá mais intensamente a um elemento um pouco diferente.

2.2 *Ward versus* K-*means versus* SOM

Berkhin (2002) aponta como vantagens dos métodos de agrupamento hierárquicos a facilidade em lidar com qualquer medida de similaridade utilizada e a sua consequente aplicabilidade a qualquer tipo de atributo. Outra vantagem é que a quantidade de grupos *k* a serem formados não é um parâmetro, isto ocorre porque o resultado final é só um grupo contendo todos os elementos. Dessa maneira, é possível visualizar o agrupamento dos elementos para diferentes quantidades de *k*. Por outro lado, traz como desvantagem o fato de que os agrupamentos não podem ser corrigidos, ou seja, os elementos de um

determinado agrupamento permanecerão nesse agrupamento até o final da execução do método.

Os métodos particionais são extremamente mais rápidos que os métodos hierárquicos (FUNG, 2001), pois não é necessário calcular e armazenar, durante o processo, a matriz de similaridade. Outra vantagem dos métodos particionais em relação ao método hierárquico é a possibilidade de um elemento poder mudar de agrupamento com a evolução do algoritmo e a possibilidade de se operar com bases de dados maiores.

No entanto, a principal desvantagem está relacionada com o fato do número de agrupamento ter que ser escolhido a priori (KAINULAINEN, 2002). Se escolhido erroneamente o número de agrupamentos *k*, o método irá impor uma estrutura aos dados, no lugar de buscar a estrutura inerente a estes. Portanto, Bussab, Miazaki e Andrade (1990) sugerem a aplicação do método diversas vezes para diferentes valores de *k*, escolhendo os resultados que apresentem melhor interpretação dos grupos.

Assim como os métodos particionário, o SOM exige uma definição prévia do número de agrupamento, nesse caso também denominado como dimensão do mapa topológico. Para a maioria dos problemas, determinar os números corretos pode exigir certa quantidade de tentativa e erro.

Entretanto, destacam-se três principais vantagens no método SOM, apontadas por LEE *et al.*, (2003): I) são relativamente fáceis de projetar; II) fornecem respostas rápidas e III) possuem capacidade de modelar dados dinâmicos, não lineares e com ruídos.

Portanto, nota-se que cada método apresenta suas vantagens e desvantagem, e a busca por entender qual método é melhor aplicável para cada situação impulsionou pesquisadores a realizar estudos comparativos entre os diversos métodos de agrupamento. Esses estudos serão apresentados na próxima seção.

2.3 Pesquisas anteriores comparativas de métodos de agrupamento

Como mencionado no Capítulo 1, realizou-se uma pesquisa nas principais bases de dados do portal periódico CAPES com o objetivo de analisar a forma com que o tema da comparação entre métodos de agrupamento vem sendo tratado na literatura.

Dentre os artigos que tem como objetivo comparar métodos de agrupamento com redes neurais através de dados simulados, destacaram-se o trabalho de Mangiameli, Chen e West (1996) com 177 citações; Mingoti e Lima (2006) com 113 citações; Balakrishnan *et al.* (1994) com 60 citações; Waller *et al.* (1998) com 40 citações.

A análise dessas pesquisas foi feita individualmente e, em seguida, foram criados quadros comparativos para classifica-las quanto ao método de agrupamento utilizado, os fatores analisados e os níveis dos fatores utilizados.

Em Mangiameli *et al.* (1996) o método de redes neurais para agrupamento, conhecido como SOM (*Self-Organizing Method*), foi comparado com vários métodos hierárquicos aglomerativo, incluindo os métodos de centróide, simples, completo, médio e *Ward*. Os dados foram gerados em uma distribuição normal sem correlação entre as variáveis e considerando o número de *cluster* k = 2, 3, 4, 5; número de variáveis p = 4, 6, 8; e três diferentes graus de dispersão *intracluster* chamados de alto, médio e baixo. O grau de dispersão determina a taxa de sobreposição dos *clusters*. A adição de variáveis irrelevantes e *outliers* também foram investigadas. Um total de 252 conjuntos de dados foi gerado, contendo 50 observações em cada cluster. Para uma baixa dispersão *intracluster*, mostrou-se que todos os métodos tiveram uma boa taxa correta de alocação (90%), exceto para o método hierárquico simples (76,9%). Para grau médio de dispersão, o SOM ainda apresenta uma boa taxa correta de alocação (98%), seguido do método *Ward* (86,2%). Os demais métodos clássicos apresentaram uma queda na taxa correta de alocação para menos de 45%. Para níveis altos de dispersão *intracluster*, a porcentagem de classificação correta de SOM foi 82,5% maior que o método de *Ward* (50,4%), que foi o melhor entre os métodos hierárquicos. Os métodos de ligação simples, centróide e médio apresentaram desempenho muito ruim nas condições de alta e média dispersão intracluster. Quando *outliers* e variáveis irrelevantes foram adicionadas aos dados, a taxa correta de alocação do SOM diminuiu para cerca de 80% e foi semelhante ao método de *Ward*. Os outros métodos hierárquicos foram novamente muito afetados, apresentando taxas corretas de alocação inferiores a 40%. Nenhum resultado mostrou o efeito do número de variáveis na precisão dos métodos de agrupamento. Em geral, os resultados mostraram que a taxa correta de alocação diminui conforme o número de *clusters* e o grau de dispersão *intracluster* aumentam. Além disso, em todas as situações testadas, o algoritmo SOM apresenta resultado superior aos métodos hierárquicos.

Em Mingoti e Lima (2006) o SOM foi comparado com métodos hierárquicos (ligações simples, completas, centróide, média e Ward), método *k-means* e *Fuzzy*. Os dados foram gerados em uma distribuição normal multivariada, considerando o número de *cluster* de mesma dimensão k = 2, 3, 4, 5, 10; número de variáveis p = 4, 6, 8, 10, 20. O número total de observações geradas para cada população foi definido como n = 500 e número de observações geradas para cada *cluster* foi igual a n / k. Diferentes graus de

correlação (0,25; 0,5; 0,75; 1) entre as variáveis *p* foram investigados. Também foram introduzidos diferentes graus de sobreposição *intracluster* (40%, 60%) e *outliers* (10%, 20%, 40%). Os resultados mostram que todos os métodos obtiveram bons resultados de taxa correta de alocação para todos os valores de *p* e *k* (iguais ou superiores a 99%), com exceção do SOM, que foi afetado pela quantidade de *clusters* e variáveis. SOM apresentou melhor resultado para $p = 4$ (94,99%) e $k = 2$ (99,9%) e piores resultados para $p = 20$ (74,98%) e $k = 10$ (76,43%). A introdução de correlação das variáveis não afetou o desempenho dos métodos. Ao introduzir sobreposição *intracluster*, o desempenho diminuiu para todos os métodos, exceto para o *fuzzy* que apresentou uma taxa correta de alocação próxima a 90% em 40% de sobreposição e uma taxa de 88% em 60% de sobreposição. Nos demais a taxa correta de alocação caiu para cerca de 80% em 40% de sobreposição e para 66% em 60% de sobreposição. O SOM foi o método que apresentou pior resultado em relação à sobreposição. Para 40% de sobreposição a taxa correta de alocação caiu para 75% e para 60% sobreposição a taxa ficou em torno de 50%. Quando os *outliers* foram introduzidos, o desempenho de todos os métodos diminuiu. Para 10% de *outliers* as taxas de alocação foram de 95% para todos os métodos, exceto para *k-means* (89,82%) e SOM (50,51%). Resultados semelhantes foram encontrados para 20% dos outliers. Para 40% de outliers a taxa correta de alocação de *Fuzzy* foi menor do que a ligação simples (88,91% e 98,10 respectivamente) e o SOM teve a menor taxa correta de alocação (50%). Todos os outros métodos apresentaram taxa correta de alocação acima de 80%. Em geral, conclui-se que a análise *cluster* foi mais afetada pela sobreposição que pelos *outliers*, e que o método SOM foi o que sofreu maiores variações de desempenho.

Em Balakrishnan *et al.* (1994) as redes neurais (*self organization maps – SOM*) foram comparadas com o método não hierárquico *k*-médias (*k-means*) utilizando um procedimento de simulação de dados. Os dados foram simulados de acordo com uma distribuição normal sem correlação entre as variáveis e considerando três fatores: números de clusters de mesma dimensão $k = 2, 3, 4, 5$; número de variáveis $p = 4, 6, 8$; e perturbação na matriz de distância (erro), medido em três níveis: livre, baixo e alto. Gerou-se um total de 108 conjuntos de dados, contendo 50 observações por *cluster*. Mostrou-se que, em geral, o SOM não teve um bom desempenho. Considerando o fator erro, o melhor e o pior desempenho foram observados para a estrutura livre de erros (89,34%) e pela estrutura com nível alto de erros (86,44%), respectivamente. Para o fator número de *clusters*, a melhor taxa correta de alocação foi observada em $k = 2$ (97,04%) e pior em $k = 5$ (74,82%). Para o fator número de variáveis, o melhor resultado foi em $p =$

8 (88,78%) e pior em $p = 6$ (86,22%). A taxa correta de alocação média global foi 98,77% para *k*-médias e 87,79% para SOM. Considerando os três fatores (erro, número de clusters e número de variáveis) a taxa correta de alocação variou de 100% a 96,22% para *k*-médias e de 97,04% para 74,82% para o SOM.

Em Waller *et al.* (1998), o SOM foi também comparado com métodos hierárquicos e não hierárquicos. Os dados foram gerados em uma distribuição normal multivariada, considerando o número de *cluster* $k = 2, 3, 4, 5, 6$; número de variáveis $p = 4, 6, 8, 10, 20$; correlação intracluster (zero, baixa, moderada, alta); grau de não sobreposição *(nonoverlapping)* dos clusters $v = 0,4; 0,6; 0,8$, ou seja, quanto menor o índice de *nonoverlapping*, maior a sobreposição entre os *clusters*. Também foi avaliado o nível de compactação dos *clusters*, ou seja, se os *clusters* tinham coeficientes de variação iguais ou diferentes entre si. Foram gerados 480 dados com cinco réplicas, e o tamanho da amostra em cada *cluster* foi determinada por uma distribuição uniforme entre 10 e 50. Os resultados mostram que a correlação *intracluster* e níveis de compactação de cluster tiveram pouca influência nas precisões de agrupamento. De modo geral, os métodos SOM, *k-means* e *Ward* apresentaram uma taxa correta de alocação alta, aproximadamente 87%, e produziram resultados equivalentes. Os demais métodos hierárquicos apresentaram taxa correta de alocação em torno de 60%. Nenhum dos métodos apresentou bom desempenho para *nonoverlapping* $v = 0,4$, porém todos os métodos tiveram bom desempenho para $v = 0,8$. Para o efeito de número de *clusters*, nenhum os métodos funcionou bem em conjunto de dados com 5 ou 6 *clusters*, no entanto, todos os métodos executaram bem em conjuntos de dados com 2 clusters. Em relação ao número de variáveis, todos os métodos apresenta melhor desempenho em dados com maior número de variáveis. O método SOM foi o que apresentou menor variação, apresentando melhor resultado para $k = 2$ (100%), $p = 8$ (97%) e $v = 0,8$ (99%) e pior resultado para $k = 6$ (71%), $p = 2$ (64%) e $v = 0,4$ (67%). Os autores ainda analisam a interação dos efeitos de maneira indireta, através da análise de efeitos da variância nas taxas de porcentagem de erro de classificação. A análise mostra que pouquíssimas são as interações que respondem por até 1% da variação na precisão da classificação.

Os trabalhos descritos acima estão resumidos nos quadros abaixo, de acordo com o método de agrupamento utilizado, Quadro 2.1, e os fatores avaliados, Quadro 2.2.

Quadro 2.1 - Referências- Métodos de agrupamento. Fonte: Própria autora.

Referência	**Métodos de agrupamento**		
	Hierárquico	Particionário	RNA
Balakrishnan et al. (1994)		*	*
Mangiameli, Chen e West (1996)	*		*
Waller et al. (1998)	*	*	*
Mingoti e Lima (2006)	*	*	*

Quadro 2.2 - Referências - Fatores avaliados. Fonte: Própria autora.

Referência	**Fatores avaliados**						
	Número variáveis	Número *clusters*	Correlação *intraclusters*	Sobreposição dos *clusters*	*Outliers*	Erro	Coeficiente de variação
Balakrishnan et al. (1994)	*	*				*	
Mangiameli, Chen e West (1996)	*	*		*	*		
Waller et al. (1998)	*	*	*	*			*
Mingoti e Lima (2006)	*	*	*	*	*		

Os Quadros 2.3 a 2.7 apresentam os níveis dos fatores analisados nas pesquisas anteriores. Destaca-se que os fatores número de c*lusters* e número de variáveis estão presentes em todos os trabalhos. Os níveis de trabalho são similares para todos os fatores, exceto para os *outliers*, que no estudo de Mangiameli *et al.* (1996) teve classificação como "ausente/presente" enquanto que no estudo de Mingoti e Lima (2006) foram nivelados em porcentagens.

Quadro 2.3 - Referências - Número de Clusters. Fonte: Própria autora.

Número de *clusters*	**2**	**3**	**4**	**5**	**6**	**10**
Balakrishnan et al. (1994)						
Mangiameli, Chen e West (1996)						
Waller et al. (1998)						
Mingoti e Lima (2006)						

Quadro 2.4 - Referências - Número de Variáveis. Fonte: Própria autora.

Número de variáveis	**4**	**6**	**8**	**10**	**20**
Balakrishnan et al. (1994)					
Mangiameli, Chen e West (1996)					
Waller et al. (1998)					
Mingoti e Lima (2006)					

Quadro 2.5 - Referências - Grau de sobreposição dos *clusters*. Fonte: Própria autora.

Grau de sobreposição dos *clusters*	**0,2**	**0,4**	**0,6**
Balakrishnan et al. (1994)			
Mangiameli, Chen e West (1996)			
Waller et al. (1998)			
Mingoti e Lima (2006)			

Quadro 2.6 - Referências - Correlação *intracluster*. Fonte: Própria autora.

Correlação *intracluster*	**0**	**0,25**	**0,5**	**0,75**	**1**
Balakrishnan et al. (1994)					
Mangiameli, Chen e West (1996)					
Waller et al. (1998)					
Mingoti e Lima (2006)					

Quadro 2.7 - Referências - *Outliers*. Fonte: Própria autora.

Correlação *intracluster*	**10%**	**20%**	**40%**	**SIM**	**NÃO**
Balakrishnan et al. (1994)					
Mangiameli, Chen e West (1996)					
Waller et al. (1998)					
Mingoti e Lima (2006)					

Em geral, os trabalhos mostram que os desempenhos dos métodos de agrupamento pioram a medida que o número de *clusters*, número de variáveis e grau de sobreposição aumentam. Os resultados mostram também que a correlação *intracluster* (WALLER *et al.*, 1998; MINGOTI; LIMA, 2006), o coeficiente de variação (WALLER *et al.*, 1998), e o erro aleatório na matriz de distância (BALAKRISHNAN *et al.*, 1994) tiveram muito pouca influência nas precisões de agrupamento.

Além disso, os resultados mostram opiniões contraditórias a respeito dos *outliers,* uma vez que para Mangiameli *et al.* (1996) os *outliers* apresentaram pouca influência no desempenho dos métodos, sendo o método *Ward* e o SOM os menos afetados, e Mingoti, e Lima (2006) sugerem que a presença de *outliers* afeta bastante o desempenho do agrupamento, principalmente do método SOM.

Outra divergência nas pesquisas é em relação ao desempenho do método SOM. Para Mangiameli *et al.* (1996), Balakrishnan *et al.* (1994) e Waller *et al.* (1998) o SOM é o método de agrupamento que apresenta melhor desempenho. Porém, Mingoti, S. A e Lima, J.O. (2006) discordam e defendem a ideia de que o SOM é o método que é mais afetado pelas variações de estrutura de dados e o que apresenta pior desempenho.

No entanto, como mencionado anteriormente, (I) nenhum trabalho avaliou o efeito do tamanho da amostra; (II) não foi avaliado o efeito do tamanho da amostra dos clusters serem iguais ou diferentes, (III) nenhum estudo utilizou Planejamento de Experimento (DOE) para simular os bancos de dados e avaliar o desempenho dos métodos; (IV) os resultados não consideram efeitos diretos de interações entre fatores; (V) nenhum dos trabalhos comprovou os resultados por meio de exemplos de agrupamento de dados.

2.4 Projeto e Análise de Experimentos (DOE)

Montgomery (2005) define Planejamento de Experimentos, que vem do inglês *Design of Experiments* (DOE), como um processo onde se planejam os experimentos para que dados apropriados sejam coletados e depois analisados por métodos estatísticos, resultando em conclusões válidas e objetivas.

DOE é definido também como um teste, ou uma série de testes, em que um conjunto de variáveis de entrada ou fatores (*x*) são alterados pelo experimentador de maneira controlada (*c*) a fim de observar e identificar como as respostas (*y*) desse sistema são afetadas devido as alterações (STAICULESCU *et al.*, 2005), Figura 2.6. Assim, é possível compreender quais fatores são significativos e como eles interagem um com o outro.

Para Gomes (2010), as técnicas do DOE têm encontrado uma ampla aplicação nas mais variadas áreas do conhecimento, mostrando-se como um conjunto de ferramentas de grande importância para o desenvolvimento de produtos e processos.

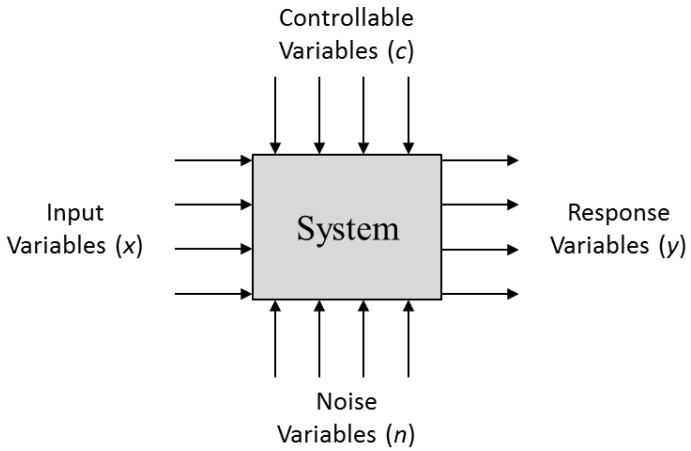

Figura 2.6 - Modelo experimental de um sistema genérico. Fonte: STAICULESCU *et al.* (2005).

O DOE é uma técnica comumente usada em processos para encontrar a solução ótima e robusta (LEE *et al.*, 2007; DASCALEUSCU *et al.*, 2008). No entanto, a técnica DOE pode ser usada para outros fins, por exemplo, pode ser aplicada em problemas de simulação. Nesse caso, aumenta a transparência do comportamento do modelo de

simulação e a eficácia dos relatórios dos resultados da simulação (LORSCHEID; HEINE; MEYER, 2012). Além disso, permite controlar os fatores que serão utilizados na simulação e apresentar resultados melhores e mais rápidos do que a simulação de tentativa e erro. Portanto, o DOE é uma parte útil e necessária da análise de simulação (LEE *et al.*, 2007).

Neste trabalho, o DOE é usado para simular conjuntos de dados sintéticos por meio da combinação de diferentes fatores descritos na próxima seção. Neste contexto, a Figura 2.6 pode ser representada analogamente pela Figura 2.7.

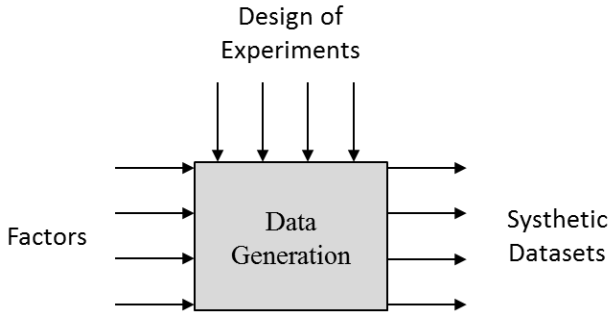

Figura 2.7- Modelo experimental de um sistema de simulação. Fonte: Própria autora.

Antes de iniciar qualquer experimentação, é importante estabelecer o planejamento dos testes. Para uma melhor condução do Planejamento de Experimentos, Montgomery (2005) sugere que o DOE seja dividido nas seguintes etapas:

1. Definição do problema;
2. Escolha dos fatores e definição dos níveis de trabalho;
3. Seleção das variáveis de resposta;
4. Execução dos experimentos;
5. Análise estatística dos dados;
6. Conclusões e recomendações.

Portanto, é fundamental que o experimentador tenha um bom grau de conhecimento a respeito do fenômeno ou modelo que se pretende estudar, como os dados serão coletados, sendo também necessário um conhecimento básico sobre as ferramentas de análise estatísticas utilizadas. Sendo assim, qualquer problema experimental necessita ser sustentado por dois elementos: o projeto dos experimentos e a análise estatística dos dados (GOMES, 2010).

3 MÉTODO DE PESQUISA

A partir dos conceitos apresentados no capítulo anterior, este capítulo tem o objetivo de reafirmar o problema de pesquisa para, em seguida, classificar a pesquisa, o método experimental adotado e demonstrar a sequência das etapas executadas.

3.1 Problema de pesquisa

Este trabalho consiste na simulação de bancos de dados, com seis parâmetros em diferentes níveis, os quais devem ser analisados por diferentes métodos de agrupamento (*Ward, K-means,* SOM), a fim de avaliar o desempenho dos métodos através de uma taxa de alocação correta dos agrupamentos e, assim, identificar os parâmetros que mais impactam nos desempenhos dos mesmos, Figura 3.1.

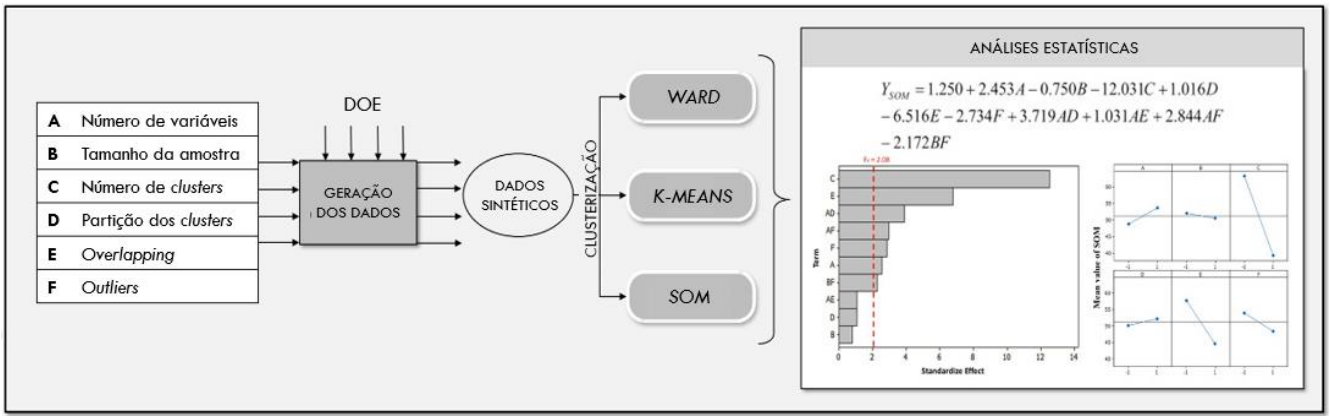

Figura 3.1 – O Problema de Pesquisa. Fonte: Própria autora.

3.2 Classificação da pesquisa

A classificação da pesquisa baseou-se em Miguel *et al.* (2014), que sugerem a classificação da pesquisa científica segundo os critérios: natureza, objetivos, forma de abordar o problema e métodos utilizados, Figura 3.2.

Portanto, essa pesquisa será classificada da seguinte maneira:

- Quanto à natureza: Segundo Appolinário (2006), a pesquisa básica visa produzir novos conhecimentos e teorias, enquanto que a pesquisa aplicada tem como foco a solução de problemas reais com a utilização dos resultados obtidos. Nesse sentido, este trabalho se classifica como uma pesquisa básica, pois busca desenvolver conhecimento através de dados sintéticos que possam eventualmente ser utilizados por outros pesquisadores e empresas na resolução de problemas;

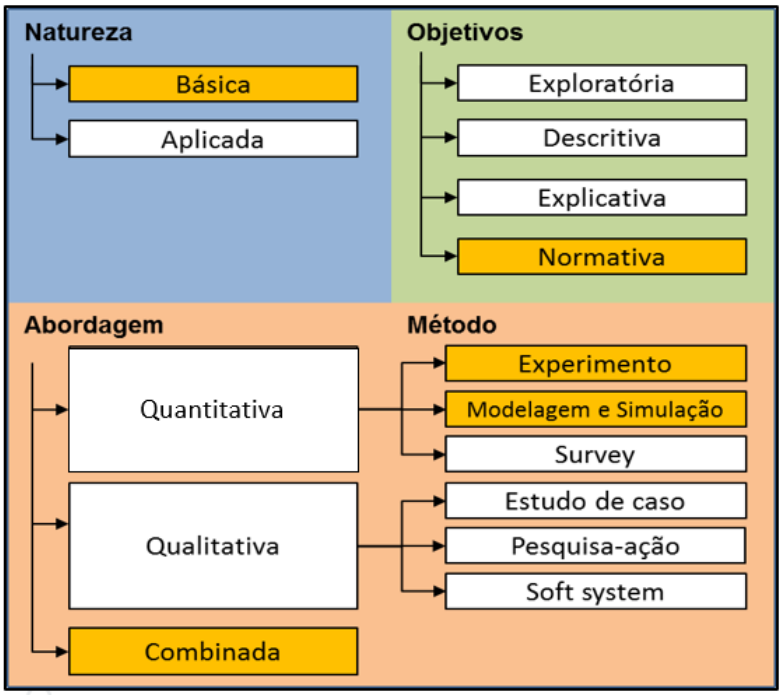

Figura 3.2 - Classificação da pesquisa. Fonte: adaptado de Miguel et al. (2014)

- Quanto ao objetivo: Em relação aos seus objetivos, esta pesquisa se caracteriza como normativa, pois procura desenvolver novas maneiras, estratégias e ações para resolver determinado problema. A pesquisa normativa está primariamente interessada no desenvolvimento de políticas, estratégias e ações para aperfeiçoar os resultados disponíveis na literatura existente, para encontrar uma solução ótima para novas definições de problemas ou para comparar várias estratégias relativas a um problema específico (BERTRAND e FRANSOO, 2002);

- Quanto à forma de abordar o problema: Quanto à abordagem, trata-se de pesquisa combinada do experimental e modelagem e simulação, pois será empregado primeiramente o método de experimentação durante a elaboração do banco de dados e, em seguida, modelagem e simulação, na aplicação de métodos de agrupamento e na avaliação de seus desempenhos.

- Quanto aos métodos: Na pesquisa experimental, o pesquisador analisa o problema, formula hipóteses e manipula variáveis de entrada selecionadas, de forma a estabelecer uma relação entre tais variáveis e os resultados sob investigação (KÖCHE, 2013). Já o método de modelagem e simulação deve ser utilizado quando se deseja prever o efeito de mudanças no sistema ou avaliar seu desempenho ou comportamento (BERTRAND, J. W. M., FRANSOO, J. C., 2012).

3.2.1 Experimentação

Segundo Bryman (1989), a pesquisa experimental adquiriu considerável importância devida à facilidade que o pesquisador que emprega os projetos experimentais encontra para estabelecer relações de causa e efeito, fazendo com que o experimento seja considerado um modelo de delineamento de pesquisa.

Portanto, é importante reafirmar que a principal característica da pesquisa experimental é que o pesquisador tem como objetivo demonstrar, usando técnicas de análise estatística, as relações causais entre a variável independente (fator ou parâmetros) e a variável dependente (repostas ou efeitos).

Assim, entre as técnicas de experimentação, foi empregado o DOE, definido no item 2.5 como o processo de planejamento dos experimentos para que dados apropriados sejam gerados e depois analisados por métodos estatísticos, o que resulta em conclusões válidas e objetivas (MONTGOMERY, 2005).

3.2.2 Modelagem e Simulação

Segundo Chung (2004), a modelagem e simulação é o processo de criar e experimentar um sistema físico, conjunto de componentes, que se interage e que recebe entradas e oferece resultados para algum propósito.

Para Pereira (2000), sistema físico é um modelo computacional de um sistema real, em que se pode visualizar esse sistema, implementar mudanças e responder a testes do tipo "o que aconteceria se" *(what-if)*, minimizando custos e tempo. Desse modo, o objetivo da simulação é estudar o comportamento de um sistema, sem que seja necessário modifica-lo ou mesmo construí-lo fisicamente.

Os sistemas podem ser classificados como discretos, contínuos, ou combinação de ambos. Nesse trabalho será abordado o sistema contínuo, pois as variáveis utilizadas mudam continuamente no tempo.

Outra classificação encontrada na literatura é sobre os modelos serem determinísticos ou estocásticos. O modelo estocástico é o que será utilizado, pois utiliza variáveis probabilísticas.

Bertrand e Fransoo (2012) sugerem também a classificação da pesquisa como axiomática ou empírica, e descritiva ou normativa. Nesse caso, será utilizada uma pesquisa axiomática normativa. Segundo os mesmos autores, a pesquisa axiomática normativa produz conhecimento sobre o comportamento de certas variáveis no modelo, com base em suposições sobre o comportamento de outras variáveis.

Miltrof (1974) sugere que a abordagem operacional da pesquisa de modelagem e simulação consiste em quatro fases:

- Conceitualização;
- Modelagem;
- Solução pelo modelo;
- Implementação.

3.3 Procedimento

O procedimento da pesquisa baseou-se na combinação da metodologia de modelagem e simulação introduzida por Mitroff *et al.* (1974), item 3.2.2, e na metodologia de experimentação proposta por Montgomery (2005), item 2.5, gerando um procedimento de pesquisa combinado, Figura 3.3.

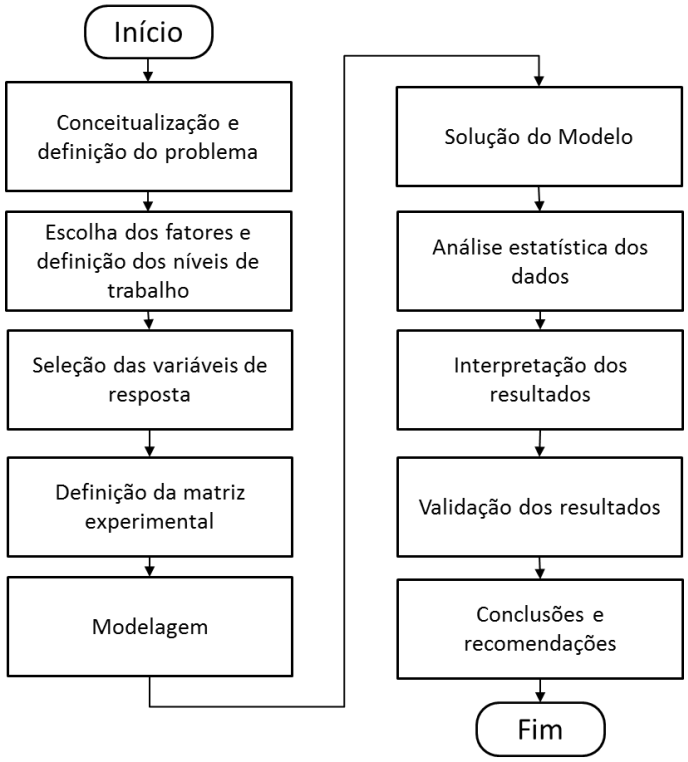

Figura 3.3 – Procedimento da Pesquisa. Fonte: Própria autora.

Portanto, para a condução desta pesquisa foram realizadas as etapas a seguir.

3.3.1 Conceitualização e definição do problema

Nesta fase inicial é importante definir o escopo do problema e os objetivos da pesquisa. Como ressaltado no item 3.1, o escopo do problema é determinar qual método

de agrupamento de dados oferece maior taxa de alocação correta, gerando assim melhor agrupamento.

Os objetivos específicos dessa pesquisa, assim como mencionado no item 1.5, são: gerar um banco de dados simulando diversas situações de amostras; realizar a análise de agrupamento dos dados gerados utilizando o método hierárquico *Ward*, o método particionário *K-means*, e Redes Neurais de *Kohonen* (SOM); comparar os resultados de agrupamento obtido e extrair conclusões válidas sobre os efeitos dos fatores e suas interações; comprovar os resultados obtidos através da resolução de exercícios do livro "*Cluster Analysis*" (EVERITT; LANDAU; LEESE, 2001).

3.3.2 Escolha dos fatores e níveis de trabalho

Nessa etapa, os fatores a serem estudados foram selecionados através da análise bibliográfica, Seção 2.4, e também se buscou assimilá-los aos fatores presentes nos exercícios clássicos do livro "*Cluster Analysis*" (EVERITT, B. S.; LANDAU, S.; LEESE, M., 2001), de modo a permitir uma comparação final entre os exemplos clássicos e o resultado. Fatores contidos em outros trabalhos que não apresentaram resultados relevantes não foram inseridos nessa pesquisa. Portanto, para este trabalho, serão utilizados os fatores e níveis a seguir:

- *Tamanho da Amostra:*

Esse fator indica a quantidade de observações presente em cada banco de dados. Foram gerados dados contínuos aleatórios N (μ;1) contendo 200 e 400 amostras, de modo que o mínimo de observações em cada *clusters* fosse no mínimo 20.

- *Número de Variáveis*

Esse fator indica a quantidade de variáveis medidas em cada elemento observado. Os níveis utilizados foram: variáveis de 4 e 6 dimensões. Visto que a literatura sugere que a porcentagem de classificação correta aumenta com o número de variáveis (Milligan e Cooper, 1980), esses níveis foram escolhidos por serem mencionados em todos os trabalhos e mais plausíveis de simulação dos dados.

- *Número de Clusters*

O número de *clusters* indica a quantidade de grupos finais em que os elementos serão separados. Os dados foram gerados considerando situações em que a problemática

requer o agrupamento de dados em 2 ou 4 *clusters*, visto que são valores considerados em todos os trabalhos analisados.

- *Partição dos Clusters*

Esse fator representa a distribuição dos dados em cada cluster. Os *clusters* podem ser classificados como homogêneos, ou seja, os tamanhos das amostras dos *clusters* são iguais; ou os clusters podem ser heterogêneos, apresentando grupos com partições diferentes. Visto que os níveis de agrupamento são k = 2 e 4, a partição heterogêna dos *clusters* heterogêneos foram gerados conforme distribuição abaixo, Quadro 3.1.

Tais valores foram baseados em trabalhos publicados por outros autores que utilizaram esse fator no estudo de agrupamento, como Pereira (1993), que comparou o desempenho de dez métodos hierárquicos em diferentes estruturas de dados; e Rodrigues (2009), que estudou algumas técnicas de agrupamento aplicadas a dados de expressão gênica.

Quadro 3.1 - Partição dos *Clusters.* Fonte: Própria autora.

Grupos homogêneos				**Grupos heterogêneos**			
Para k =2		Para k =4		Para k =2		Para k =4	
Grupos	Amostras	Grupos	Amostras	Grupos	Amostras	Grupos	Amostras
G1	50%	G1	25%	G1	30%	G1	10%
G2	50%	G2	25%	G2	70%	G2	20%
		G3	25%			G3	30%
		G4	25%			G4	40%

- *Overlapping*

Este fator tem por finalidade criar situações em que os grupos apresentam regiões de sobreposição (*overlapping*). Milligan e Cooper (1980) sugerem que não é adequado considerar este aspecto, uma vez que os métodos hierárquicos de agrupamento não foram desenvolvidos para detectar estruturas com grupos sobrepostos ("*overlapping*"). Porém, considerando o fato de outros métodos de agrupamento serem avaliados também, decidiu-se por inserir esse fator nos seguintes níveis: baixo e alto. Para criar essas situações, as amostras foram geradas utilizando os conceitos de *"effect size".*

Effect size, ou tamanho do efeito (TDE), é um conceito relacionado à significância de testes estatísticos. O TDE refere-se à diferença entre duas amostras (ROSENTHAL, 1994). Para calcula o TDE usa-se normalmente a medida de "*d de Cohen*", Equação 3.1, (COHEN, 1988).

$$d = \frac{x_1 - x_2}{s_p} \quad , onde \tag{3.1}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{3.2}$$

Cohen (1998) classificou o TDE como pequeno; médio; grande. TDE pequeno (*d*=0,2) significa que a diferença entre amostras é difícil de ser vista a olho nu. Um TDE médio (*d*=0,5) é provavelmente suficientemente maior para ser percebido a olho nu; enquanto que um TDE grande (*d*=0,8) é evidente a olho nu.

Neste trabalho serão utilizados os níveis *d*=0,8 e *d*=0,2 para gerar dados com TDE grande, o que indica pequena sobreposição, e TDE pequeno, grande sobreposição.

Segundo Cohen (1998), para *d*=0,2, Figura 3.4(a), 58% do grupo 2 estará acima da média do grupo 1; 92% dos dois grupos se sobreporão e há uma chance de 56% de que um elemento escolhido aleatoriamente do grupo 2 seja maior que um elemento escolhido aletoriamente do grupo 1.

E para *d*=0,8, Figura 3.4(b), 79% do grupo 2 estará acima da média do grupo 1; 69% dos dois grupos estarão sobrepostos e há uma chance de 71% de se escolher aleatoriamente um elemento do grupo 2 maior que um elemento do grupo 1 também escolhido aleatoriamente.

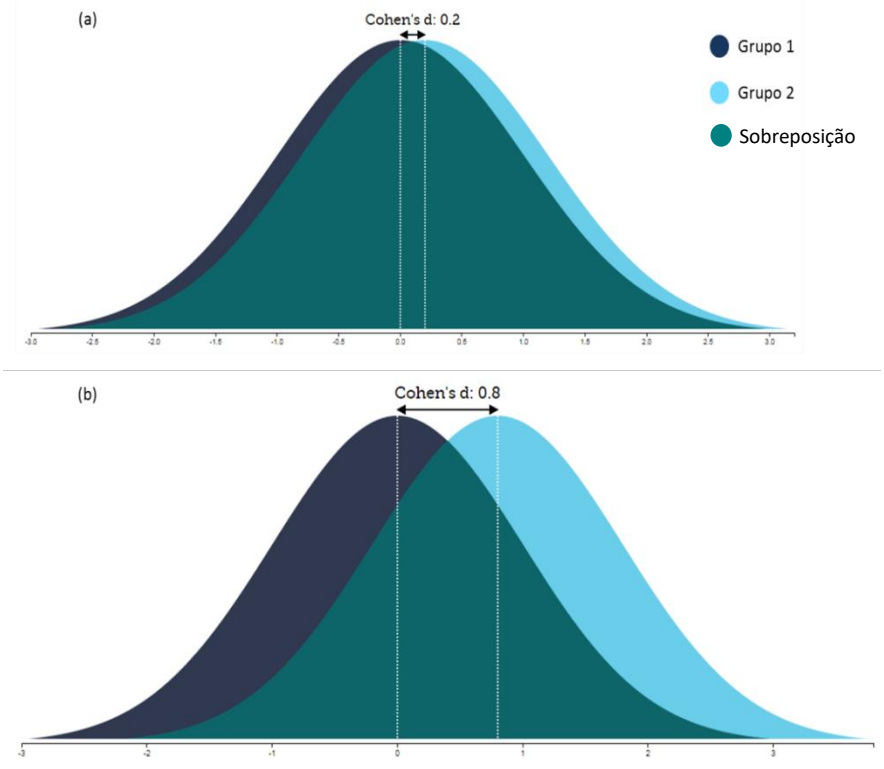

Figura 3.4 - (a) Cohen's d=0,2; (b) Cohen’s d=0,8. Fonte: Própria autora.

- *Outliers*

A inserção de *outliers* nas amostras tem como objetivo simular um erro discrepante da mensuração de dados e verificar como os métodos de agrupamento reagem diante desse fator, ou seja, se *outliers* interferem no resultado do agrupamento. Para isso, foi introduzida uma "contaminação" sobre 20% das amostras aleatórias geradas, em todas as variáveis. Essa contaminação consiste em uma distribuição normal com um desvio padrão cinco vezes maior que o das observações originais N(μ;5).

3.3.3 Seleção das variáveis de respostas

Nessa etapa determinam-se as variáveis de resposta utilizadas no DOE. Nesse caso, será utilizada a "Taxa Correta de Alocação" (*Attribute Agreement Analysis*), a qual é utilizada para comparar os agrupamentos de cada elemento obtidos pelos métodos de *cluster* com os respectivos agrupamentos simulados. Ela é calculada pela proporção de observações alocadas corretamente aos grupos originalmente simulados, tendo como valor de referência 100%.

3.3.4 Definição da matriz experimental

Para selecionar ou construir uma matriz experimental, devem ser considerados o número de fatores e níveis escolhidos anteriormente, Tabela 3.1.

Tabela 3.1 - Parâmetros do Experimento. Fonte: Própria autora.

A	B	C	D	E	F
Número de Variáveis	Tamanho da Amostra	Número de *clusters*	Partição dos *clusters*	*Overlapping*	*Outliers*
4	200	2	Igual	d=0,8	0%
6	400	4	Diferente	d=0,2	20%

Considerando os seis fatores da Tabela 3.1, construiu-se a matriz de DOE, Tabela 3.2, usando o software estatístico MINITAB®. A matriz de delineamento utilizada foi o experimento fatorial fracionado 2^k com resolução IV e duas repetições, resultando em 32 experimentos.

O experimento fatorial fracionário é uma classe do DOE amplamente utilizada em experimentos envolvendo vários fatores, onde é necessário estudar a significância dos fatores e o efeito conjunto dos mesmos em uma resposta (MONTGOMERY, 2005), e

quando o recurso disponível para o experimento é escasso (GIESBRECHT; GUMPERTZ, 2004).

Tabela 3.2 - Matriz experimental fatorial fracionado. Fonte: Própria autora.

	A	**B**	**C**	**D**	**E**	**F**
Ordem	Número de Variáveis	Tamanho da Amostra	Número de *clusters*	Partição dos *clusters*	*Overlapping*	*Outliers*
1	4	200	2	Igual	0.8	0%
2	6	200	2	Igual	0.2	0%
3	4	400	2	Igual	0.2	20%
4	6	400	2	Igual	0.8	20%
5	4	200	4	Igual	0.2	20%
6	6	200	4	Igual	0.8	20%
7	4	400	4	Igual	0.8	0%
8	6	400	4	Igual	0.2	0%
9	4	200	2	Diferente	0.8	20%
10	6	200	2	Diferente	0.2	20%
11	4	400	2	Diferente	0.2	0%
12	6	400	2	Diferente	0.8	0%
13	4	200	4	Diferente	0.2	0%
14	6	200	4	Diferente	0.8	0%
15	4	400	4	Diferente	0.8	20%
16	6	400	4	Diferente	0.2	20%
17	4	200	2	Igual	0.8	0%
18	6	200	2	Igual	0.2	0%
19	4	400	2	Igual	0.2	20%
20	6	400	2	Igual	0.8	20%
21	4	200	4	Igual	0.2	20%
22	6	200	4	Igual	0.8	20%
23	4	400	4	Igual	0.8	0%
24	6	400	4	Igual	0.2	0%
25	4	200	2	Diferente	0.8	20%
26	6	200	2	Diferente	0.2	20%
27	4	400	2	Diferente	0.2	0%
28	6	400	2	Diferente	0.8	0%
29	4	200	4	Diferente	0.2	0%
30	6	200	4	Diferente	0.8	0%
31	4	400	4	Diferente	0.8	20%
32	6	400	4	Diferente	0.2	20%

Uma característica particular do planejamento fatorial fracionário é o fato de não apresentar um arranjo experimental completo, apresentando, portanto, um confundimento entre os principais efeitos e as interações. A intensidade do confundimento é chamada de

resolução (BOX; HUNTER; HUNTER, 1978). Quanto maior a resolução, menor é a confusão.

Neste trabalho, a matriz gerada possui resolução IV, ou seja, os principais fatores são confundidos com interações de 3ª ordem, que são geralmente fracas, portanto, desprezíveis. Para compor esse arranjo, calculou-se um fatorial completo 2^{k-2}. Seja *k* o número de fator igual a 6, então o arranjo terá 24 experimentos completos. Assim, os fatores *p, z, k, n* compõem um fatorial completo, e assumiram-se $ov = p \times z \times k$ e $ot = z \times k \times n$.

Neste caso, não foi necessário randomizar os experimentos porque os mesmos são usados apenas como simulação de dados. A matriz de design é um guia para indicar a combinação de fatores para gerar os conjuntos de dados que serão usados para análise de *cluster*.

3.3.5 Modelagem

Nessa etapa, o modelo conceitual é convertido em modelo computadorizado. Os modelos são bancos de dados gerados a partir da matriz de experimentos. Para cada teste (*run*) foi gerado um banco de dados utilizando o *software* MINITAB®. Por exemplo, o 'banco de dados 1' é composto por 4 variáveis, 200 amostras, 2 *clusters* com 100 amostras cada, baixa sobreposição e sem *outliers*. No total, gerou-se 32 conjuntos de dados, Apêndice A, os quais foram analisados pelos métodos de agrupamento, que será descrito na seção a seguir.

3.3.6 Solução dos Modelos

Com os modelos obtidos, foi possível realizar a análise de agrupamento. É importante explicitar que cada modelo representa um banco de dado e que todos os 32 modelos foram submetidos à análise de todos os métodos de agrupamentos: *Ward, K-means, SOM*. Como exemplo, a seguir será descrito o procedimento utilizado para o agrupamento 'banco de dados 1'. Para a solução dos modelos, serão utilizados dois *softwares:* MINITAB, para métodos clássicos de agrupamento; STATISTICA, para o método SOM.

3.3.6.1 Ward

Primeiramente, gerou-se o Apêndice B, em que são mostradas as etapas de agrupamento *Ward*, sendo que em cada etapa, dois *clusters* são unidos. O quadro mostra quais *clusters* foram unidos; a distância entre eles; o nível se similaridade correspondente;

o número de identificação do novo *cluster*; o número de observações no novo *cluster* e o número do *cluster*. O agrupamento continua até que haja apenas um *cluster*.

Idealmente, os *clusters* devem ter um nível de similaridade relativamente alto e um nível de distância relativamente baixo. As etapas de agrupamento mostram que o nível de similaridade teve uma queda abrupta entre o passo 2 (-242,510) e o passo 1 (-776,677); e também aumentou ligeiramente o nível de distância entre o passo 2 (27,2675) para o passo 1 (69,7930), o que indica que dois *clusters* são razoavelmente suficientes para a partição final.

Ao realizar a análise com a partição final de dois *clusters*, obteve-se a Figura 3.5. Essa tabela resume cada *cluster* pelo número de observações; a soma de quadrados dentro do *cluster*; a distância média da observação ao centróide do *cluster* e a distância máxima de observações até o centróide do *cluster*.

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	74	233,279	1,66582	3,36422
Cluster2	126	426,085	1,72101	3,26758

Figura 3.5 - *Ward* – Análise de cluster. Fonte: Própria autora.

Nota-se que no *cluster* 1 contém 74 elementos, enquanto que no *cluster* 2 contém 126. O *cluster* 1 é o grupo mais compacto, ou seja, com menor variabilidade; enquanto que o *cluster* 2 apresenta maior variabilidade, ou seja, seus elementos são menos similares entre si.

Outro resultado obtido está na Figura 3.6. O centroide de cada variável dentro dos *clusters* é apresentado e seu valor indica a influência das variáveis em cada cluster. O *cluster* 1 tem maior influência da variável *x2*, enquanto o *cluster* 2 é influenciado pela variável *x1*.

Variable	Cluster1	Cluster2	Grand centroid
x1	-1,35313	0,212178	-0,366985
x2	-0,73173	-0,160666	-0,371958
x3	-0,95849	0,009005	-0,348968
x4	-0,84949	-0,137310	-0,400819

Figura 3.6 - *Ward*- Centroide. Fonte: Própria autora.

Por fim, a Figura 3.7 apresenta a distância entre os centróides de cada cluster.

	Cluster1	Cluster2
Cluster1	0,00000	2,05415
Cluster2	2,05415	0,00000

Figura 3.7 - *Ward*- Distância entre centroides. Fonte: Própria autora.

3.3.6.2 K-means

Inicialmente, determinou-se o número de *clusters* = 2. Nota-se na Figura 3.8, que no *cluster* 1 existem 83 elementos, enquanto que no *cluster* 2 contém 117. Nota-se também que o *cluster* 2 apresenta menor variabilidade.

Final Partition Number of clusters: 2				
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	83	266,439	1,686	3,460
Cluster2	117	386,600	1,697	3,268

Figura 3.8 - *K-means* – Análise de cluster. Fonte: Própria autora.

O centróide de cada variável dentro dos *clusters* é apresentado na Figura 3.9. Esse resultado mostra que o *cluster* 1 é influenciado pela variável *x2* e o *cluster* 2 pela *variável x1*. A distância entre os centróides gerais de cada *cluster* está descrito na Figura 3.10.

Variable	Cluster1	Cluster2	Grand centroid
x1	-1,2563	0,2639	-0,3670
x2	-0,7333	-0,1157	-0,3720
x3	-0,9421	0,0718	-0,3490
x4	-0,7980	-0,1190	-0,4008

Figura 3.9 - *K-means* – Centroides. Fonte: Própria autora.

	Cluster1	Cluster2
Cluster1	0,0000	2,0449
Cluster2	2,0449	0,0000

Figura 3.10 - *K-means* – Distância entre centroides. Fonte: Própria autora.

3.3.6.3 SOM

Primeiramente, definiu-se o método de amostragem (Figura 3.11). Para esse problema foi selecionada a opção de amostragem aleatória, ou seja, o STATISTICA atribuirá aleatoriamente casos aos subconjuntos de Treinamento, Teste e Validação, com base em porcentagens especificadas de 60% dos dados para Treinamento de rede; 20%

para Teste e 20% para Validação, com a soma total da porcentagem para não mais que 100.

É importante mencionar que a que a soma das porcentagens de amostra pode ser menor que 100. Esse pode ser o caso se o número de casos de dados presentes no conjunto de dados for grande. O treinamento de redes neurais em grandes conjuntos de dados pode consumir muito tempo, e a omissão aleatória de casos de um grande conjunto de dados pode ajudar a reduzir o tempo de computação e também produzir bons modelos desde que seja incluada porcentagem de casos de dados para a análise (STATISTICA, 2005).

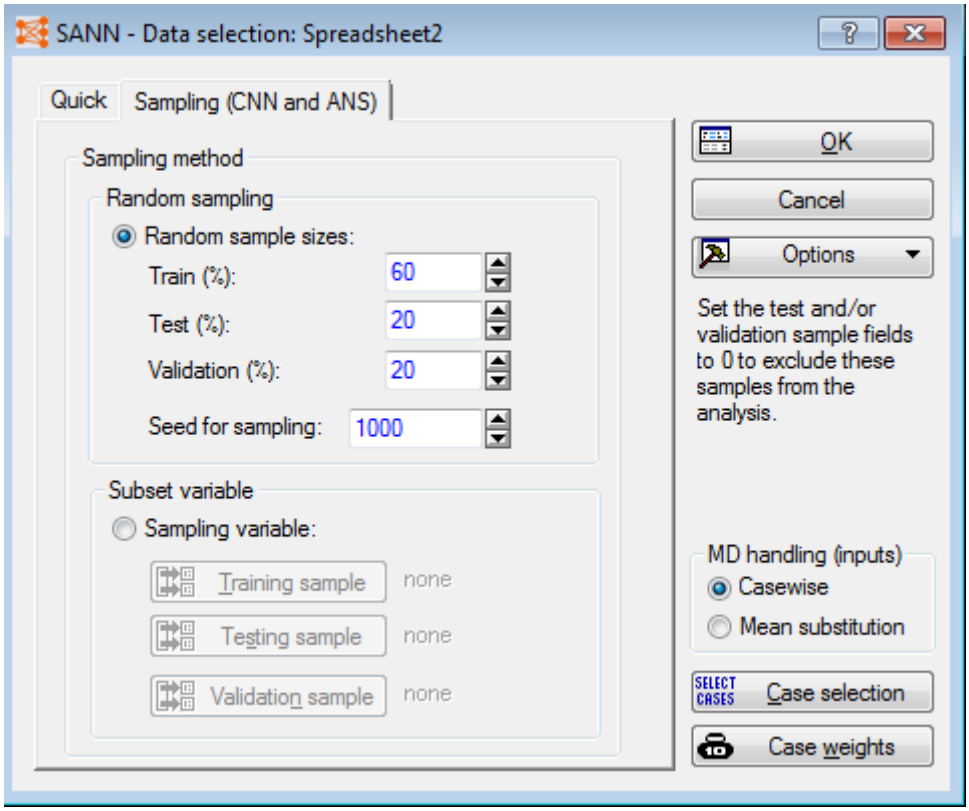

Figura 3.11 - Redes Neurais -*Sampling*. Fonte: Própria autora

A opção *"Seed for sampling"* refere-se a um número positivo inteiro que é utilizado como um gerador de números aleatórios que produz as amostras aleatórias dos dados. O mesmo número sempre obterá a mesma amostra. Para alterar as amostras, é preciso alterar o valor inserido.

A próxima etapa, Figura 3.12, é definir as dimensões do mapa topológico, que é padronizado pelo *software* como uma estrutura retangular. Essa estrutura é a camada de saída e indica a quantidade de agrupamento final, que neste caso é de 2 *clusters.*

Na aba *"Kohonen Training",* Figura 3.13, é definido o treinamento, a randomização da rede e condições de parada do treinamento. O treinamento é caraterizado pelo ciclo de treinamento e pela vizinhança.

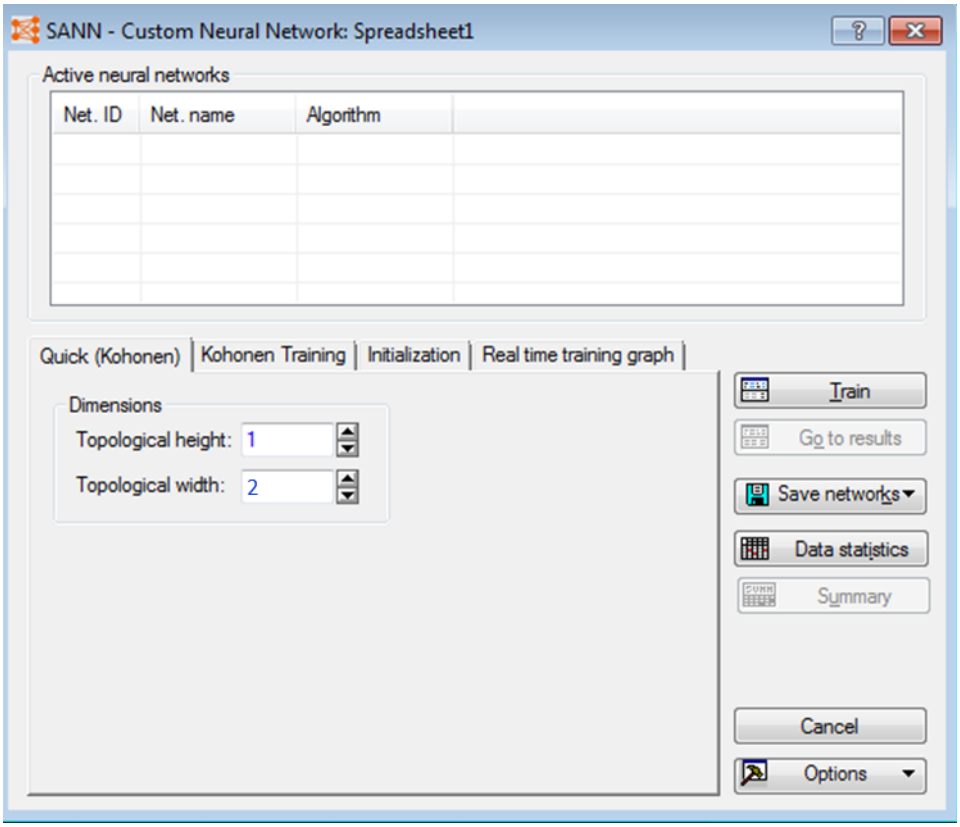

Figura 3.12 - Redes Neurais- *Kohonen*. Fonte: Própria autora

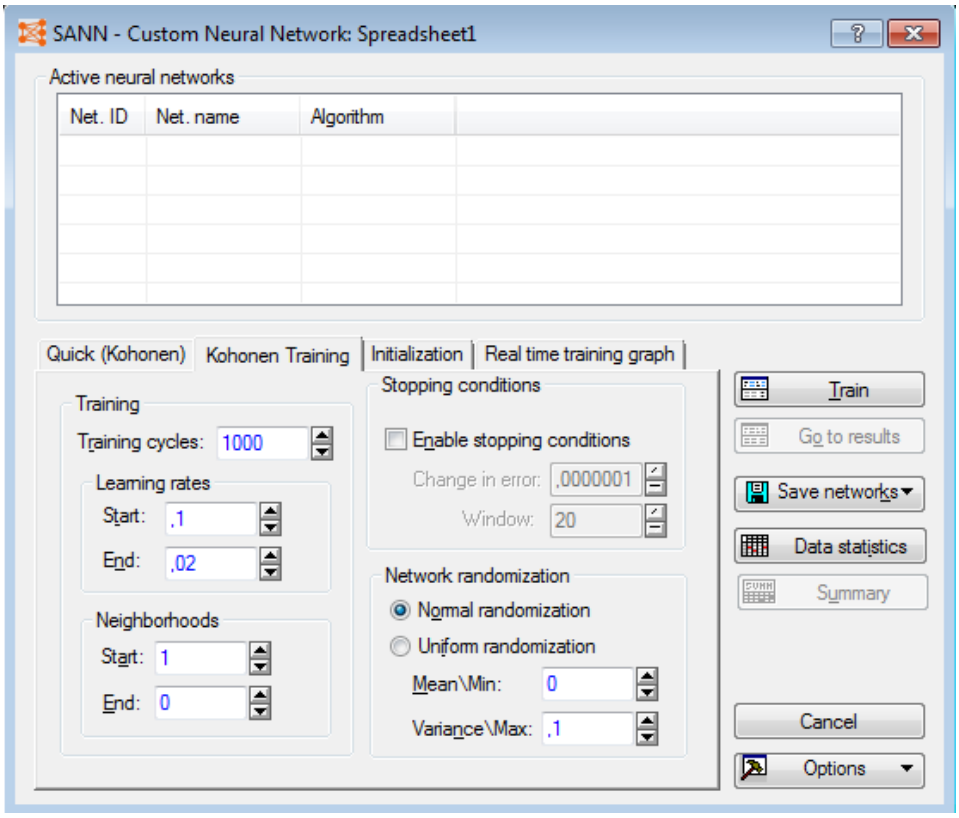

Figura 3.13 - Redes Neurais- *Kohonen Training*. Fonte: Própria autora

Na opção *"Training cycles"* foi inserido o valor de 1000 ciclos, conforme item 2.2.3.1. De acordo com o mesmo item, inseriu-se a taxa de aprendizagem de Kohonen inicial em 0.1 e final em 0.02.

A vizinhança é definida como o "raio" de uma vizinhança quadrada centrada no neurônio vencedora. Por exemplo, um tamanho de vizinhança de 2 especifica um quadrado de 5 x 5 (STATISTICA, 2005). Nesse caso, para uma topologia 1 x 2 o tamanho da vizinhança foi definido em 1. Se o neurônio vencedor estiver posicionado próximo ou na borda do mapa topológico, a vizinhança é cortada na borda. A vizinhança é dimensionada linearmente desde o valor inicial até o valor final fornecido.

Na opção *"Network randomization"* é especificada como os pesos devem ser inicializados no início do treinamento. Nesse caso, foi selecionada a aleatorização normal. As configurações padrão de *"Mean\Min"* e *"Variance\Max"* foram mantidas, pois é recomendável definir a *"Mean\Min"* como zero e a *"Variance\Max"* não superior a 0,1. Isso ajudará a rede a crescer gradualmente de seu modo linear (valores pequenos de peso) para o modo não linear (grandes valores de peso) para modelar os dados durante o processo de treinamento (STATISTICA, 2005).

As condições de parada antecipada, *"stopping conditions",* não foram selecionadas, pois o *software* indica o uso da parada antecipada apenas quando nenhuma amostra de teste foi selecionada.

Porém, caso seja necessário ativar as condições de parada, será necessário definir os itens *"Change in error"* e "*Window".* Quando as condições de parada são aplicadas, o treinamento da rede terminará se a melhoria média do erro da rede sobre um número especificado de ciclos de treinamento *("Window")* for menor que o valor de mudança no erro ("*Change in error").*

Os resultados do agrupamento estão na Figura 3.14, a qual mostra a frequência dos *clusters*, ou seja, a quantidade de elemento em cada grupo.As frequências indicam os centros de clusters no mapa topológico. Frequências iguais a zero são geralmente consideradas como uma indicação de que o aprendizado não foi bem-sucedido, pois a rede não está usando todos os recursos disponíveis (STATISTICA, 2005).

	Win Frequencies (2) (Spreadsheet1)
	0
0	94,000
1	106,000

Figura 3.14 - Redes Neurais- Resultados. Fonte: Própria autora

3.3.7 Análises estatísticas dos dados

Nessa etapa é realizada a modelagem matemática das respostas, análise de resíduos, adequação e ajuste dos modelos.

Assim que todos os bancos de dados foram analisados pelos três métodos, os resultados dos agrupamentos foram submetidos ao *Attribute Agreement Analysis*, a fim de verificar o grau de conformidade dos agrupamentos com os grupos originais. As respostas estão descritas na Tabela 3.3.

Tabela 3.3 - Matriz experimental - Respostas. Fonte: Própria autora.

StdOrder	Respostas		
	Ward	***K-means***	***SOM***
1	72,00	74,50	76,00
2	54,50	51,00	53,00
3	50,25	50,50	50,50
4	50,25	73,50	72,00
5	23,50	27,00	28,50
6	44,50	44,50	48,00
7	56,50	59,25	50,50
8	36,50	34,50	34,50
9	68,50	70,50	54,50
10	62,50	65,50	69,50
11	50,25	50,50	57,25
12	79,75	82,25	79,25
13	29,00	36,50	32,50
14	61,50	56,00	52,00
15	46,50	50,00	38,50
16	35,50	38,50	37,25
17	72,50	84,50	75,50
18	50,00	53,00	50,50
19	50,00	50,50	50,50
20	50,50	79,50	52,75
21	26,50	22,50	24,00
22	44,50	47,50	45,50
23	56,75	59,50	56,50
24	33,00	32,50	35,50
25	63,50	67,50	73,00
26	62,00	68,00	68,50
27	50,50	50,00	51,75
28	79,00	80,25	78,00
29	33,00	28,00	35,00
30	62,00	54,50	46,00
31	45,50	47,50	26,25
32	35,00	37,50	37,00

A partir dos dados experimentais apresentados pela Tabela 3.3, é possível verificar qual o método que apresenta melhor desempenho e estabelecer relações matemáticas entre as respostas analisadas e os parâmetros de entrada.

Tendo em vista os resultados, realizou-se primeiramente o teste *paired-t* e *power and sample size* para verificar se as diferenças entre os métodos de agrupamento eram estatisticamente significativas, a um nível de confiança de 95%.

A hipótese nula é a diferença entre as amostras igual a zero, e a hipótese alternativa é a diferença entre as amostras maiores que zero. Comparando *Ward* com SOM, obteve-se *p-value* de 0,923 com poder de 0,05. Comparando *K-means* e SOM, o *p-value* obtido foi de 0,05 com poder de 0,50. E ao se comparar *Ward* e *K-means* o *p-value* foi de 0,033 com poder de 0,58.

Apesar dos valores de *p-value* indicarem que os métodos *Ward* e SOM apresentam resultados similares entre si (p-value >0,05) e divergentes do método *K-means* (*p-value* <0,05), os resultados do *power and sample size* indicam que o poder dos testes são baixos para detectar tal diferença entre os desempenhos dos métodos. Portanto, a um nível de significância de 5%, não é possível afirmar que um método é melhor que outro.

Em seguida, foi possível estabelecer relações matemáticas entre as respostas e os parâmetros de entrada usando o Método dos Mínimos Quadrados Ordinários (OLS) e Análise de Variância no Minitab®. Os coeficientes estimados em OLS estão indicados na Tabela 3.4. Os coeficientes são codificados pelas letras A, B, C, D, E e F. Toda a análise seguinte considerou um nível de significância de 5%.

Tabela 3.4 - Coeficiente estimado. Fonte: Própria autora.

Coeficientes	*Ward*	*K-means*	*SOM*
Constante	**51,117**	**53,976**	**51,250**
A	**1,445**	**2,180**	**2,453**
B	**-0,758**	0,789	-0,750
C	**-9,258**	**-11,742**	**-12,031**
D	**2,883**	**1,211**	1,016
E	**-8,492**	**-10,477**	**-6,516**
F	**-3,680**	**-1,445**	**-2,734**
AB	**-1,867**	0,367	0,328
AC	**0,758**	**-1,227**	0,297
AD	**4,211**	**2,945**	**3,719**
AE	**2,055**	**1,883**	1,031
AF	**-0,789**	**2,102**	**2,844**
BD	-0,492	**-1,414**	-0,860
BF	**-1,242**	0,117	**-2,172**
ABD	**0,773**	-0,430	0,719
ABF	**-1,414**	-0,836	-1,469

Coeficiente em **negrito** indicam os termos significativos.

Após, realizou-se a Análise de Variância (ANOVA). A análise também foi realizada para cada resposta, *Ward*, *K-means* e SOM individualmente, conforme Figuras (3.15) – (3.17) respectivamente. Os resultados do ANOVA indicam que todos os modelos tiveram bons ajustes, indicando grande confiabilidade e preditividade.

Em seguida, realizou-se a análise dos resíduos. Os resíduos são definidos como a diferença entre o valor previsto do modelo e o valor experimental observado para cada condição e sua análise é importante para garantir que os modelos matemáticos desenvolvidos representam bem as respostas de interessa (GOMES, 2010).

S = 1,63279 PRESS = 170,625
R-Sq = 99,38% R-Sq(pred) = 97,52% R-Sq(adj) = 98,80%

Analysis of Variance for Ward (coded units)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	6	5834,82	5834,82	972,47	364,77	0,000
A	1	66,85	66,85	66,85	25,07	0,000
B	1	18,38	18,38	18,38	6,89	0,018
C	1	2742,63	2742,63	2742,63	1028,74	0,000
D	1	265,94	265,94	265,94	99,75	0,000
E	1	2307,75	2307,75	2307,75	865,62	0,000
F	1	433,28	433,28	433,28	162,52	0,000
2-Way Interactions	7	909,51	909,51	129,93	48,74	0,000
A*B	1	111,56	111,56	111,56	41,85	0,000
A*C	1	18,38	18,38	18,38	6,89	0,018
A*D	1	567,42	567,42	567,42	212,84	0,000
A*E	1	135,10	135,10	135,10	50,67	0,000
A*F	1	19,92	19,92	19,92	7,47	0,015
B*D	1	7,75	7,75	7,75	2,91	0,107
B*F	1	49,38	49,38	49,38	18,52	0,001
3-Way Interactions	2	83,13	83,13	41,56	15,59	0,000
A*B*D	1	19,14	19,14	19,14	7,18	0,016
A*B*F	1	63,99	63,99	63,99	24,00	0,000
Residual Error	16	42,66	42,66	2,67		
Pure Error	16	42,66	42,66	2,67		
Total	31	6870,12				

Figura 3.15 - ANOVA - WARD. Fonte: Própria autora.

S = 2,92918 PRESS = 549,125
R-Sq = 98,48% R-Sq(pred) = 93,92% R-Sq(adj) = 97,05%

Analysis of Variance for Kmeans (coded units)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	6	8210,12	8210,12	1368,35	159,48	0,000
A	1	152,03	152,03	152,03	17,72	0,001
B	1	19,92	19,92	19,92	2,32	0,147
C	1	4412,13	4412,13	4412,13	514,23	0,000
D	1	46,92	46,92	46,92	5,47	0,033
E	1	3512,27	3512,27	3512,27	409,35	0,000
F	1	66,85	66,85	66,85	7,79	0,013
2-Way Interactions	7	649,25	649,25	92,75	10,81	0,000
A*B	1	4,31	4,31	4,31	0,50	0,488
A*C	1	48,14	48,14	48,14	5,61	0,031
A*D	1	277,60	277,60	277,60	32,35	0,000
A*E	1	113,44	113,44	113,44	13,22	0,002
A*F	1	141,33	141,33	141,33	16,47	0,001
B*D	1	63,99	63,99	63,99	7,46	0,015
B*F	1	0,44	0,44	0,44	0,05	0,824
3-Way Interactions	2	28,27	28,27	14,13	1,65	0,224
A*B*D	1	5,91	5,91	5,91	0,69	0,419
A*B*F	1	22,36	22,36	22,36	2,61	0,126
Residual Error	16	137,28	137,28	8,58		
Pure Error	16	137,28	137,28	8,58		
Total	31	9024,92				

Figura 3.16 - ANOVA - K-*MEANS*. Fonte: Própria autora

S = 5,61249 PRESS = 2016
R-Sq = 93,68% R-Sq(pred) = 74,73% R-Sq(adj) = 87,76%

Analysis of Variance for SOM (coded units)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	6	6473,37	6473,37	1078,90	34,25	0,000
A	1	192,57	192,57	192,57	6,11	0,025
B	1	18,00	18,00	18,00	0,57	0,461
C	1	4632,03	4632,03	4632,03	147,05	0,000
D	1	33,01	33,01	33,01	1,05	0,321
E	1	1358,51	1358,51	1358,51	43,13	0,000
F	1	239,26	239,26	239,26	7,60	0,014
2-Way Interactions	7	916,19	916,19	130,88	4,16	0,009
A*B	1	3,45	3,45	3,45	0,11	0,745
A*C	1	2,82	2,82	2,82	0,09	0,769
A*D	1	442,53	442,53	442,53	14,05	0,002
A*E	1	34,03	34,03	34,03	1,08	0,314
A*F	1	258,78	258,78	258,78	8,22	0,011
B*D	1	23,63	23,63	23,63	0,75	0,399
B*F	1	150,95	150,95	150,95	4,79	0,044
3-Way Interactions	2	85,56	85,56	42,78	1,36	0,285
A*B*D	1	16,53	16,53	16,53	0,52	0,479
A*B*F	1	69,03	69,03	69,03	2,19	0,158
Residual Error	16	504,00	504,00	31,50		
Pure Error	16	504,00	504,00	31,50		
Total	31	7979,12				

Figura 3.17 - ANOVA - SOM. Fonte: Própria autora.

Segundo Montgomery (2005), os resíduos devem ser normais, aleatórios e não correlacionados. Dessa forma, os resíduos referentes aos modelos foram analisados e o resultado dessa análise está descrito na Tabela 3.5.

Tabela 3.35 - Análise de resíduos. Fonte: Própria autora.

Análise dos resíduos		**Respostas**		
		Ward	***K-means***	**SOM**
Teste de normalidade	AD	1,519	0,394	1,202
	P-value	**<0,005**	0,355	**<0,005**
Análise de correlação	Pearson	-0,170	-0,023	-0,188
	p-value	0,352	0,901	0,304
Análise de aleatoriedade *(p-value)*	*Clustering*	0,509	0,078	**0,037**
	Mixtures	0,491	0,922	0,963
	Trends	0,194	0,500	0,194
	Oscillation	0,806	0,500	0,806

A Tabela 3.5 indica que apenas os resíduos do modelo *k-means* são normais, pois é o único que apresenta coeficientes *Anderson-Darling* (AD) menor que 1 e *p-value* maior que 5% de significância. Se os resíduos não seguirem uma distribuição normal, os intervalos de confiança e os valores-p podem ser inexatos. A análise de correlação apresentou todos os coeficientes de *Pearson* próximos a 0 e *p-value* >0,05, o que indica que os resíduos de todos os modelos não são correlacionados. Finalmente, o teste de aleatoriedade apresentou ausência de causas especiais nos resíduos dos modelos *Ward* e *K-means*, mas no modelo SOM foi detectada uma causa especial de *clustering* (*p-value* <0,05), o que indica que os resíduos mostram um padrão quando exibidos em ordem temporal e, portanto, é necessário verificar o pressuposto deles não serem independentes um do outro.

Portanto, a fim de satisfazer os pressupostos dos resíduos e garantir que o modelo ajuste bem os dados, os dados dos modelos *Ward* e SOM passaram por uma verificação de adequação dos modelos, através da redução do modelo, ou seja, eliminando os termos e/ou interações não significantes, mantendo bons resultados de R^2 *(adj.)* e R^2 *(pred.)*. Assim, a Tabela 3.6 indica os novos ajustes dos modelos.

Para obtenção dos modelos reduzidos, foram retiradas as interações BD e ABD no modelo *Ward*, e no modelo SOM foram retirados os termos AB, AC, BD e todas as interações triplas.

Tabela 3.6 - Resultados para os modelos ajustados. Fonte: Própria autora.

Respostas		***Ward***		***K-means***		**SOM**	
Modelo		Completo	Reduzido	Completo	Reduzido	Completo	Reduzido
	R^2 (adj.) (%)	98,8	98,26	97,05	97,05	87,76	88,54
	R^2 (pred.) (%)	97,52	96,80	93,92	93,92	74,73	81,97
	S	1,63279	1,96569	2,92918	2,92918	5,61249	5,43122
Teste de normalidade	AD	1,519	0,336	0,394	0,394	1,202	0,576
	P-value	**<0,005**	**0,486**	0,355	0,355	**<0,005**	**0,123**
Análise de correlação	Pearson	-0,170	-0,133	-0,023	-0,023	-0,188	-0,169
	P-value	0,352	0,468	0,901	0,901	0,304	0,354
Análise de aleatoriedade *(P-value)*	*Clustering*	0,509	0,360	0,078	0,078	**0,037**	**0,640**
	Mixtures	0,491	0,640	0,922	0,922	0,963	0,360
	Trends	0,194	0,194	0,500	0,500	0,194	0,667
	Oscillation	0,806	0,806	0,500	0,500	0,806	0,333

Assim, os modelos apresentam os formatos descritos pelas Eqs. (3.3) – (3.5).

$$W = 51{,}117 + 1{,}445A - 0{,}758B - 9{,}258C + 2{,}883D - 8{,}492E - 3{,}680F - 1{,}867AB + 0{,}758AC + 4{,}211AD + 2{,}055AE - 0{,}789AF - 1{,}242BF - 1{,}414ABF \quad (3.3)$$

$$K = 53{,}977 + 2{,}180A + 0{,}790B - 11{,}742C + 1{,}211D - 10{,}477E - 1{,}445F + 0{,}367AB - 1{,}227AC + 2{,}9453AD + 1{,}883AE + 2{,}102AF - 1{,}414BD + 0{,}117BF - 0{,}430ABD - 0{,}836ABF \quad (3.4)$$

$$S = 51{,}250 + 2{,}453A - 0{,}750B - 12{,}031C + 1{,}016D - 6{,}516E - 2{,}734F + 3{,}719AD + 1{,}031AE + 2{,}844AF - 2{,}172BF \quad (3.5)$$

Portanto, nesta etapa, os conceitos estatísticos foram aplicados nas soluções do modelo para analisar os resíduos e a significância do modelo, a fim de ajustá-los para obter a melhor modelagem matemática das repostas. Através dessa modelagem matemática será possível, na próxima etapa, descrever o comportamento das variáveis, a relação entre elas e estimar os efeitos produzidos nas respostas observadas.

3.3.8 Interpretação dos resultados

A partir do desenvolvimento dos modelos matemáticos finais, torna-se possível analisar a maneira como as repostas se comportam devido às alterações nos parâmetros de entrada (GOMES, 2010). Portanto, nessa etapa devem-se extrair as conclusões práticas

dos resultados, descrevê-lo em gráficos e questionar se as respostas satisfazem as questões experimentais.

Sendo assim, a seguir será discutida a influência dos fatores, número de variáveis; tamanho da amostra; número de *clusters*; partição dos *clusters; overlapping; outliers*, nas análises de agrupamento *Ward, K-means* e SOM, através da análise dos modelos matemáticos desenvolvidos.

Ao analisar as Eqs. (3.3), (3.4) e (3.5), conclui-se que para todos os métodos os fatores mais significativos são o número de *clusters* (C) e a sobreposição (E), e também a interação entre o número da variável (A) e a partição dos *clusters* (D). Além disso, existem outros fatores e interações que também são significativos ao nível de confiança de 95%, mas em uma menor intensidade. Outro resultado interessante é que o tamanho da amostra (B) não é significativo para os métodos *K-means* e SOM, e a partição dos *clusters* (D) não é significativo para o SOM. Para melhor visualização dos fatores significantes, gerou-se o gráfico de Pareto para cada método, Figuras 3.18 a 3.20.

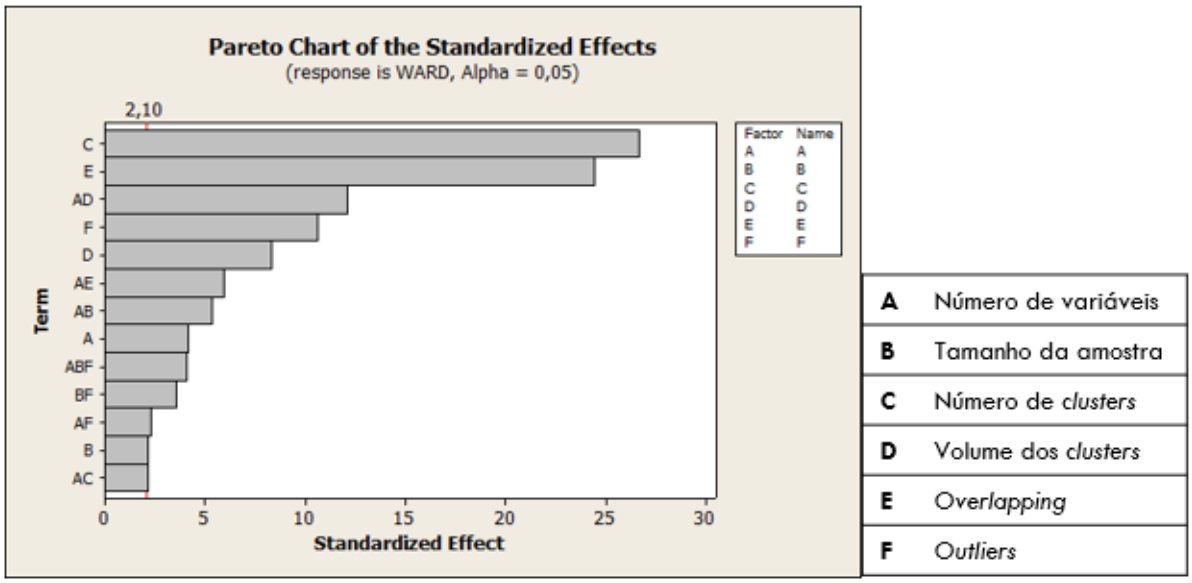

Figura 3.18 - Gráfico de Pareto para método *Ward*. Fonte: Própria autora.

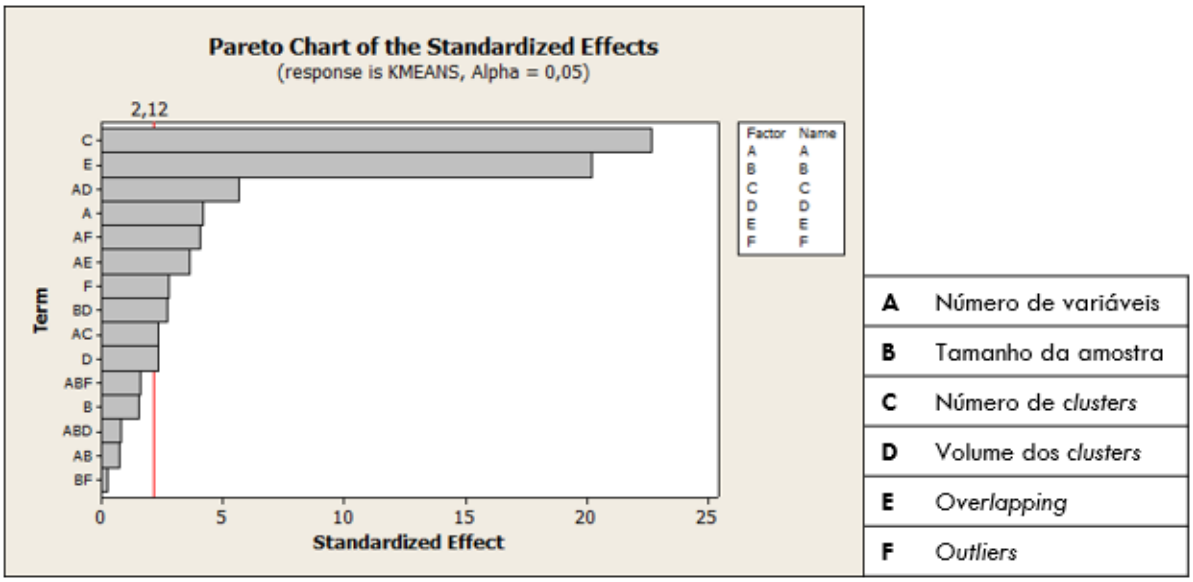

Figura 3.19 - Gráfico de Pareto para método *K-means*. Fonte: Própria autora.

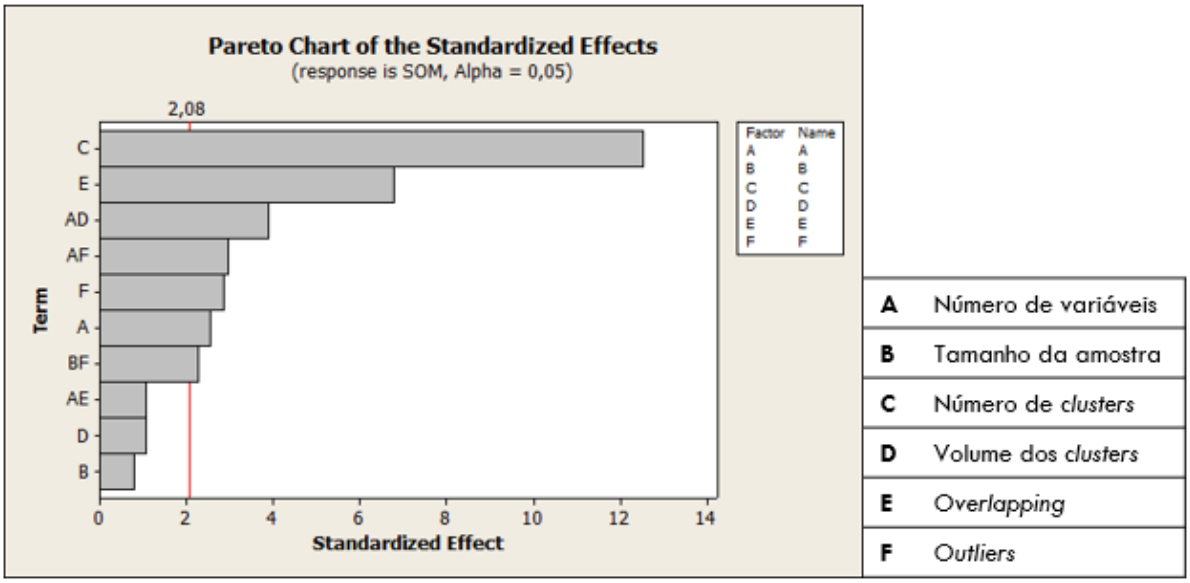

Figura 3.20 - Gráfico de Pareto para método SOM. Fonte: Própria autora.

Neste momento, é necessário entender como cada fator significativo influencia cada modelo. Então, foi desenvolvida uma análise dos principais efeitos e os gráficos de interação para todas as respostas, como mostrado nas Figuras 3.21 a 3.23.

Inferindo das Figuras 3.21, 3.22 e 3.23, o aumento do número de variáveis (A) implica em um aumento no desempenho para todos os métodos, e o método *Ward* é o menos sensível a esse fator. Seu desempenho diminuiu de 49,67% para 52,56%, enquanto o *K-means* apresentou um aumento de 51,80% para 56,16% e SOM de 48,80% para 53,70%.

O tamanho da amostra (B) não pode ser considerado um fator significativo para os métodos *K-means* e SOM.

O aumento do número de *clusters* (C) faz com que o desempenho de todos os métodos de *clustering* piore. Para o número de *clusters* k = 2, *Ward* apresentou desempenho de 60,38%, *K-means* de 65,72%, e SOM de 63,28%. Para k = 4 *Ward* teve desempenho de 41,86%, *K-means* de 42,23% e SOM de 39,22%.

A partição dos *clusters* (D) é significativo apenas para *Ward* e *K-means*, que apresentam melhores resultados em situações que os clusters têm diferentes tamanhos de amostra: 54% para *Ward* e 55,19% para *K-means*.

Clusters sem sobreposição (E) apresenta melhores resultados que *cluster* com sobreposição. Para sobreposição de *ov* = 20%, *Ward*, *K- médias* e SOM apresentaram desempenho igual a 59,61%, 64,45% e 57,76%, respectivamente. Para *ov* = 80%, o desempenho do *Ward* diminuiu para 42,63%, *K-means* para 43,5% e SOM para 44,74%.

A presença de *outliers* (F) piora o desempenho de todos os métodos, sendo o *K-means* o mais estável. O desempenho do *Ward* reduziu de 59,80 para 47,4%, *K-means* de 55,42% para 52,53%, e SOM de 53,98% para 48,52%.

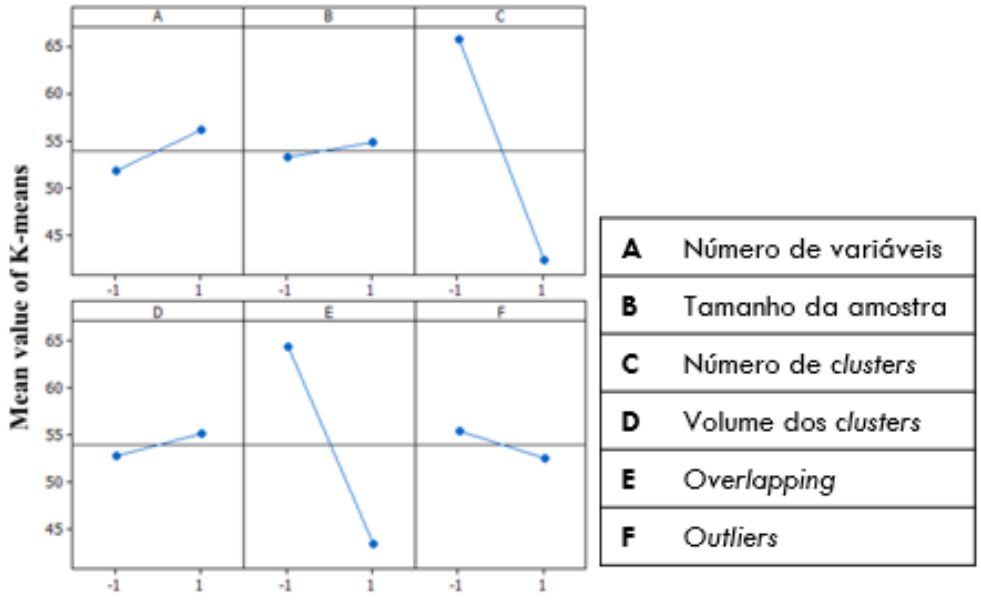

Figura 3.21- Efeitos Principais para *Ward*. Fonte: Própria autora.

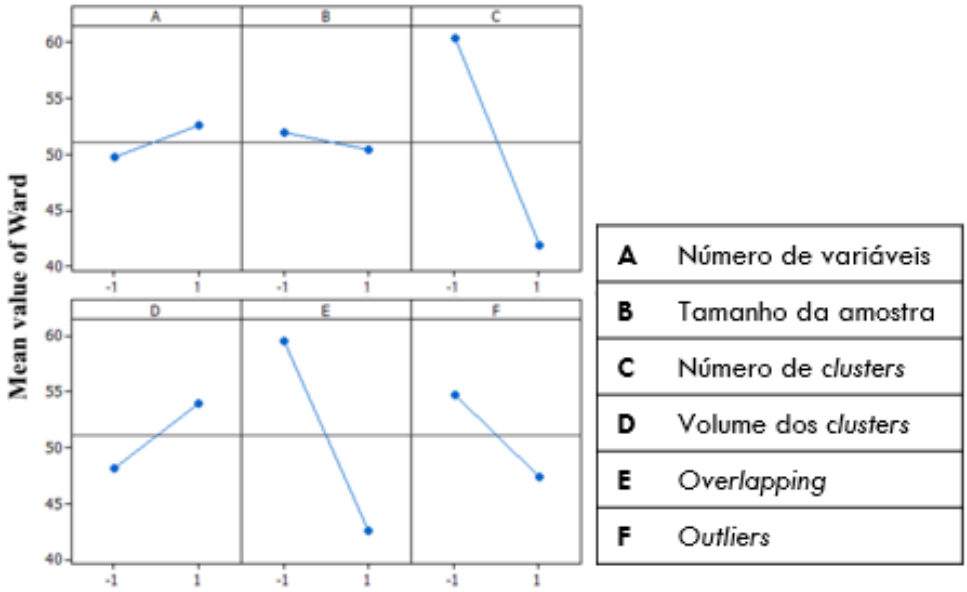

Figura 3.22- Efeitos Principais para *K-means.* Fonte: Própria autora.

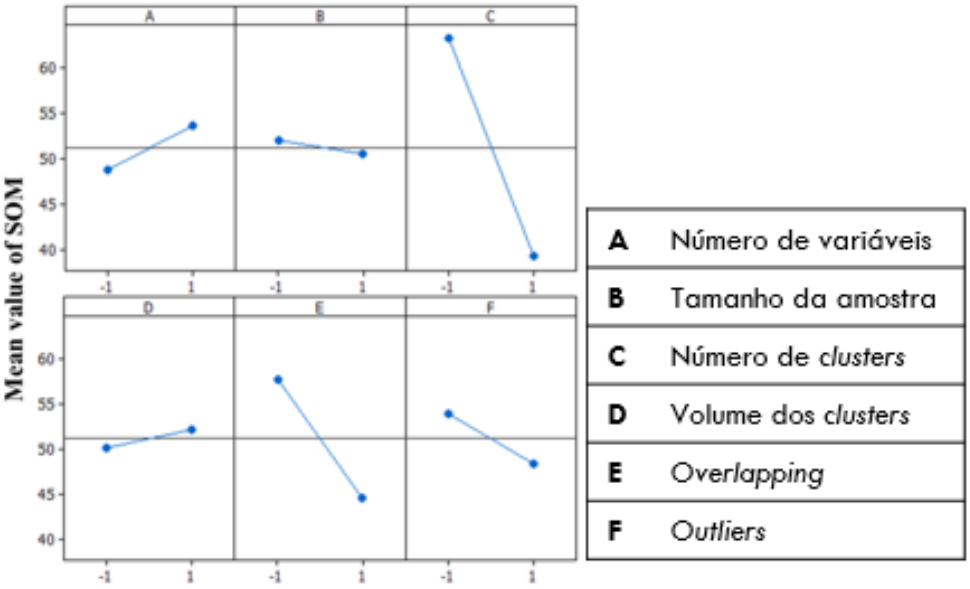

Figura 3.23 - Efeitos Principais para SOM. Fonte: Própria autora.

Além da influência dos fatores, a influência da interação entre o número de variáveis (A) e a partição dos *clusters* (D) também é significativo. Como mostrado na Figura 3.24, o melhor desempenho para todos os métodos ocorre quando o número de variáveis é alto e a partição dos *clusters* é diferente. Para os métodos *K-means* e SOM, o efeito de A é maior quando se tem partições diferentes de *clusters*, então os menores (50,06% e 46,09%) e maiores (60,31% e 58,43%) valores das performances ocorrem quando A é igual a 4 e 6, respectivamente, neste contexto. Para o método *Ward*, a situação

é diferente. O pior desempenho do *Ward* é de 45,63% e ocorre com pequeno número de variável e partição de *clusters* iguais, enquanto seu melhor desempenho é de 59,66%.

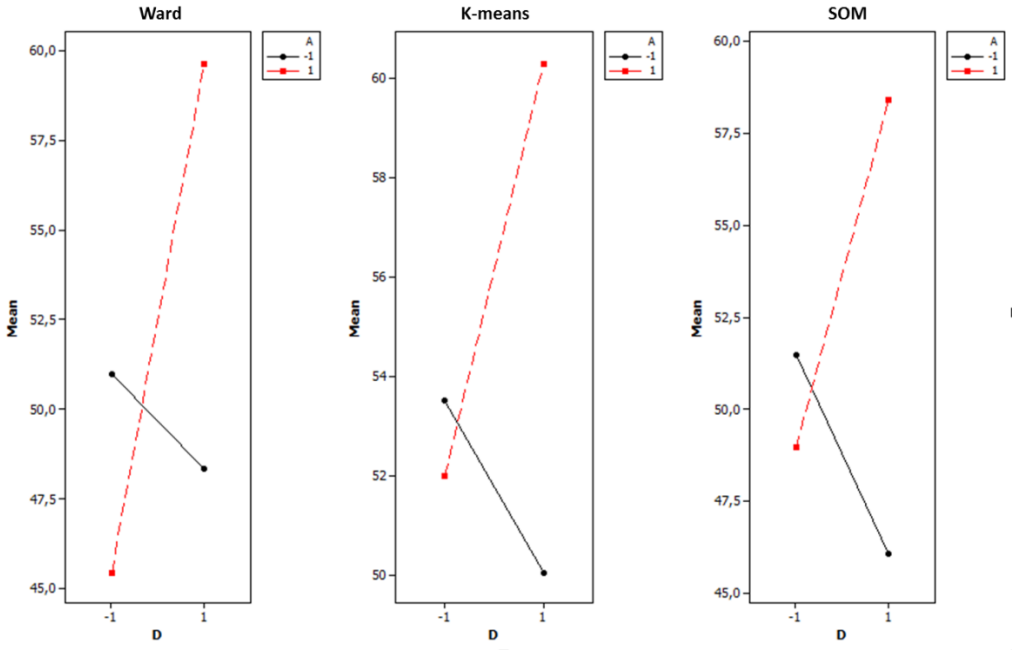

Figura 3.24 - Interação: número de variáveis e tamanha da amostra. Fonte: Própria autora.

3.3.9 Validação dos resultados

Todo resultado experimental deve ser comprovado com um experimento de confirmação. Portanto, essa etapa é fundamental para validar os resultados obtidos na etapa anterior. Para isso, utilizou-se de exemplos clássicos do livro "*Cluster Analysis*" (EVERITT; LANDAU; LEESE, 2001), os quais contêm características similares aos bancos de dados gerados.

3.3.9.1 Exemplo 1

O primeiro exemplo é chamado de "*body measurements*" (medidas do corpo) e apresenta condições favoráveis a todos os métodos. O exemplo contém medidas (em polegadas) de três variáveis (peito, cintura e quadril) para 20 pessoas, Figura 3.25; e o objetivo é dividi-los em dois grupos: homens e mulheres. A partição dos clusters não são necessariamente iguais, e as amostras não apresentam *overlapping* significativos, Figura 3.26 (d_1=1,13; d_2=1,04; d_3=0,87), e nem *outliers*. Apesar de o número de variáveis ser pequeno, os fatores mais significativos em ambos os métodos, número de *clusters* e *overlapping*, estão nos níveis de melhor desempenho, portanto, espera-se que os métodos de agrupamento apresentem bons resultados.

Subject	Chest	Waist	Hips
1	34	30	32
2	37	32	37
3	38	30	36
4	36	33	39
5	38	29	33
6	43	32	38
7	40	33	42
8	38	30	40
9	40	30	37
10	41	32	39
11	36	24	35
12	36	25	37
13	34	24	37
14	33	22	34
15	36	26	38
16	37	26	37
17	34	25	38
18	36	26	37
19	38	28	40
20	35	23	35

Figura 3.25 - Dados "*Body measurements*". Fonte: Everitt; Landau e Lleese (2001)

Fonte: EVERITT, B. S.; LANDAU, S.; LEESE, M. (2001)

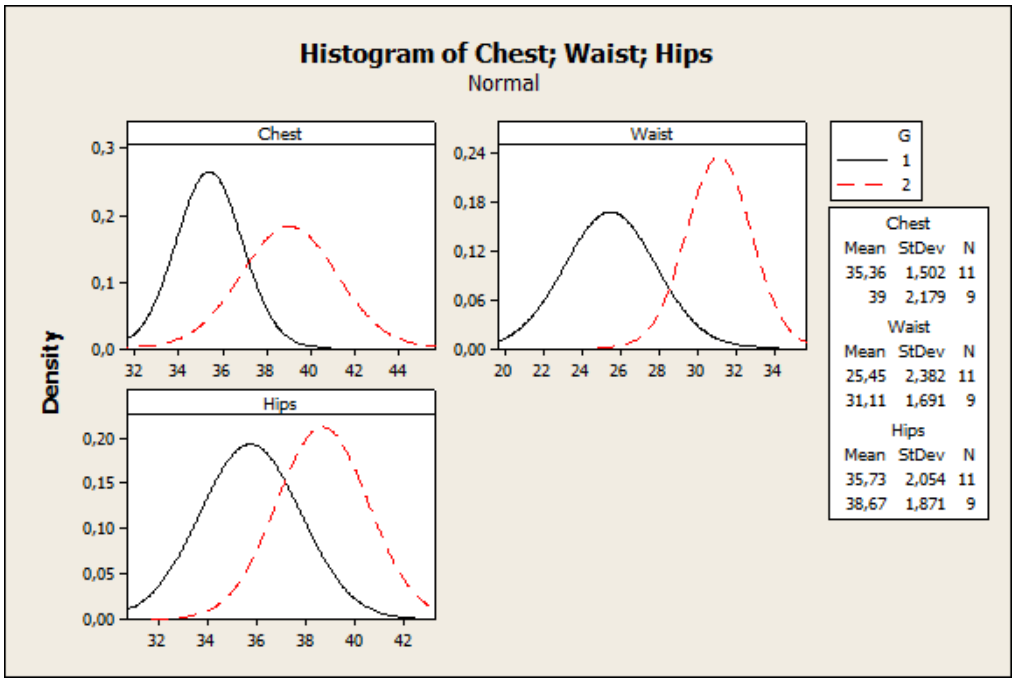

Figura 3.26 - Histograma "*Body measurements*". Fonte: Própria autora.

Ao resolver o problema a cima através dos métodos *Ward, K-means* e SOM, obtêm-se 100% de alocação correta para todos os métodos, contendo 11 elementos no grupo 1 e 9 elementos no grupo 2.

A Figura 3.27 mostra os resultados das etapas de agrupamento para o método *Ward*, sendo que a partição ideal é realmente 2 *cluster*, uma vez que o nível de similaridade teve uma queda abrupta entre os passos 2 e 3; e também ocorreu um aumento do nível de distância entre esses mesmos passos.

Cluster Analysis of Observations: Chest; Waist; Hips

Euclidean Distance, Ward Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	19	93,463	1,0000	16	18	16	2
2	18	91,658	1,2761	15	16	15	3
3	17	91,206	1,3452	12	15	12	4
4	16	90,755	1,4142	11	20	11	2
5	15	90,755	1,4142	13	17	13	2
6	14	86,926	2,0000	8	19	8	2
7	13	85,382	2,2361	6	10	6	2
8	12	85,382	2,2361	3	9	3	2
9	11	83,987	2,4495	2	4	2	2
10	10	76,100	3,6560	11	14	11	3
11	9	72,265	4,2426	1	5	1	2
12	8	68,196	4,8651	6	7	6	3
13	7	66,315	5,1528	12	13	12	6
14	6	62,591	5,7225	2	3	2	4
15	5	56,741	6,6174	2	8	2	6
16	4	39,225	9,2968	11	12	11	9
17	3	34,572	10,0085	2	6	2	9
18	2	1,674	15,0409	1	11	1	11
19	1	-175,693	42,1730	1	2	1	20

Final Partition
Number of clusters: 2

Figura 3.27 - Resultado Ward- "*Body measurements"*. Fonte: Própria autora.

Portanto, ao realizar o agrupamento utilizando a partição final de 2 *clusters*, através no método *Ward*, obteve-se o Dendograma, Figura 3.28, que mostra o agrupamento hierárquicos dos dados.

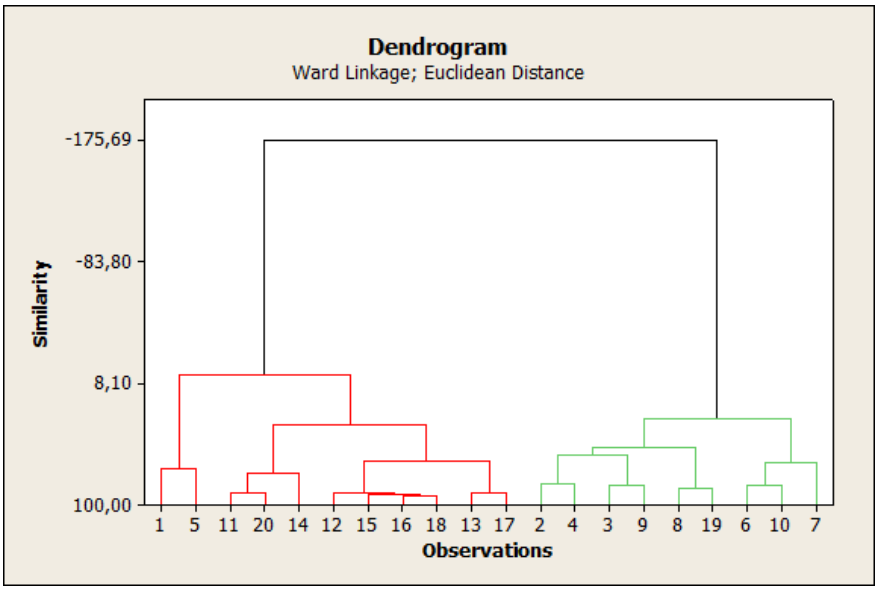

Figura 3.28 - Dendograma- "*Body measurements*". Fonte: Própria autora.

Em seguir, tanto para o método *Ward* quanto para o *K-means*, obteve-se o resultado da Figura 3.29. Nota-se que o *cluster* 1 contém 11 observações, e o *cluster* 2 contém 9 observações. O *cluster* 2 é o mais compacto, pois apresenta menor variabilidade, enquanto que o *clusters* 1 apresentam maior variabilidade.

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	11	121,455	2,97499	6,03434
Cluster2	9	88,889	3,05168	4,15145

Cluster Centroids

Variable	Cluster1	Cluster2	Grand centroid
Chest	35,3636	39,0000	37,00
Waist	25,4545	31,1111	28,00
Hips	35,7273	38,6667	37,05

Distances Between Cluster Centroids

	Cluster1	Cluster2
Cluster1	0,00000	7,33893
Cluster2	7,33893	0,00000

Figura 3.29 - Resultado Ward e K-means- "Body measurements". Fonte: Própria autora.

Ainda na Figura 3.29, pode-se observar o centroide de cada variável dentro dos *clusters*, o que indica que o *cluster* 1 é caracterizado por medida menores, principalmente das variáveis "*waist*", enquanto que o *cluster* 2 apresenta elementos com medidas maiores das variáveis.

A disposição dos dados do espaço e os elementos agrupados podem ser visualmente representados pelo "*3D Scatterplot*", Figura 3.30, em que também se pode observar a ausência de *outliers* e sobreposição de grupos.

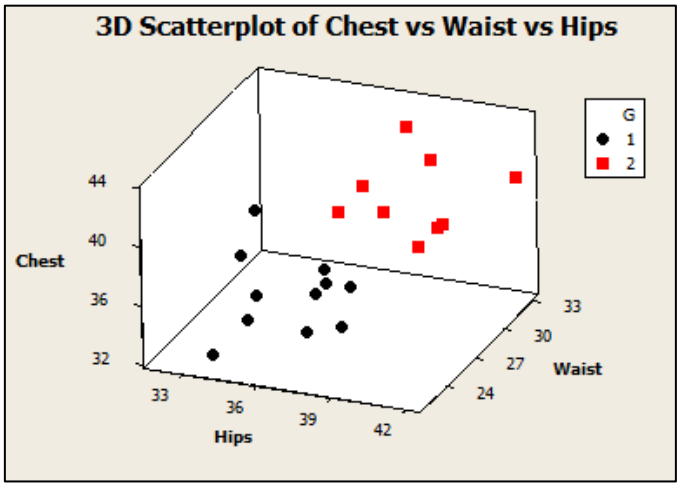

Figura 3.30 - 3D *Scatterplot*- "*Body measurements*". Fonte: Própria autora.

No método SOM, o resultado obtido está representado na Figura 3.31, que mostra a ativação de cada elemento e sua localização no mapa topológico, sendo que os elementos com localização (1,1) pertencem ao agrupamento 1, enquanto que os elementos (2,1) pertencem ao agrupamento 2.

name	Neuron location	Neuron ID	Activation	Var1	Var2	Var3
1	(2, 1)	2	0,591203	34,00000	30,00000	32,00000
2	(1, 1)	1	0,270235	37,00000	32,00000	37,00000
3	(1, 1)	1	0,309993	38,00000	30,00000	36,00000
4	(1, 1)	1	0,360827	36,00000	33,00000	39,00000
5	(2, 1)	2	0,538652	38,00000	29,00000	33,00000
6	(1, 1)	1	0,367817	43,00000	32,00000	38,00000
7	(1, 1)	1	0,391835	40,00000	33,00000	42,00000
8	(1, 1)	1	0,265822	38,00000	30,00000	40,00000
9	(1, 1)	1	0,214346	40,00000	30,00000	37,00000
10	(1, 1)	1	0,180897	41,00000	32,00000	39,00000
11	(2, 1)	2	0,158070	36,00000	24,00000	35,00000
12	(2, 1)	2	0,159543	36,00000	25,00000	37,00000
13	(2, 1)	2	0,189984	34,00000	24,00000	37,00000
14	(2, 1)	2	0,387684	33,00000	22,00000	34,00000
15	(2, 1)	2	0,259450	36,00000	26,00000	38,00000
16	(2, 1)	2	0,248517	37,00000	26,00000	37,00000
17	(2, 1)	2	0,247842	34,00000	25,00000	38,00000
18	(2, 1)	2	0,179383	36,00000	26,00000	37,00000
19	(1, 1)	1	0,400479	38,00000	28,00000	40,00000
20	(2, 1)	2	0,204297	35,00000	23,00000	35,00000

Figura 3.31 - Resultado SOM - "*Body measurements*". Fonte: Própria autora.

3.3.9.2 Exemplo 2

Outro exemplo do livro "*Cluster Analysis*" (EVERITT; LANDAU; LEESE, 2001) para confirmação do experimento é chamado de "*Chemical Pottery*" e está descrito a seguir. A Tabela 3.7 mostra a composição química da cerâmica, em termos de nove óxidos, conforme determinado por espectrofotometria de absorção atômica, de 46 exemplos de cerâmica Romano-Britânica (Tubb *et al*., 1980).

Tabela 3.7 - Resultado da análise química da cerâmica. Fonte: Tubb *et al*. (1980).

Sample number	Chemical component								
	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	MnO	BaO
1	18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015
2	16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018
3	18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014
4	16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019
5	17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019
6	18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017
7	16.5	7.05	1.81	1.73	0.33	3.20	0.95	0.066	0.019
8	18.0	7.42	2.06	1.00	0.28	3.37	0.96	0.072	0.017
9	15.8	7.15	1.62	0.71	0.38	3.25	0.93	0.062	0.017
10	14.6	6.87	1.67	0.76	0.33	3.06	0.91	0.055	0.012
11	13.7	5.83	1.50	0.66	0.13	2.25	0.75	0.034	0.012
12	14.6	6.76	1.63	1.48	0.20	3.02	0.87	0.055	0.016
13	14.8	7.07	1.62	1.44	0.24	3.03	0.86	0.080	0.016
14	17.1	7.79	1.99	0.83	0.46	3.13	0.93	0.090	0.020
15	16.8	7.86	1.86	0.84	0.46	2.93	0.94	0.094	0.020
16	15.8	7.65	1.94	0.81	0.83	3.33	0.96	0.112	0.019
17	18.6	7.85	2.33	0.87	0.38	3.17	0.98	0.081	0.018
18	16.9	7.87	1.83	1.31	0.53	3.09	0.95	0.092	0.023
19	18.9	7.58	2.05	0.83	0.13	3.29	0.98	0.072	0.015
20	18.0	7.50	1.94	0.69	0.12	3.14	0.93	0.035	0.017
21	17.8	7.28	1.92	0.81	0.18	3.15	0.90	0.067	0.017
22	14.4	7.00	4.30	0.15	0.51	4.25	0.79	0.160	0.019
23	13.8	7.08	3.43	0.12	0.17	4.14	0.77	0.144	0.020
24	14.6	7.09	3.88	0.13	0.20	4.36	0.81	0.124	0.019
25	11.5	6.37	5.64	0.16	0.14	3.89	0.69	0.087	0.009
26	13.8	7.06	5.34	0.20	0.20	4.31	0.71	0.101	0.021
27	10.9	6.26	3.47	0.17	0.22	3.40	0.66	0.109	0.010
28	10.1	4.26	4.26	0.20	0.18	3.32	0.59	0.149	0.017
29	11.6	5.78	5.91	0.18	0.16	3.70	0.65	0.082	0.015
30	11.1	5.49	4.52	0.29	0.30	4.03	0.63	0.080	0.016
31	13.4	6.92	7.23	0.28	0.20	4.54	0.69	0.163	0.017
32	12.4	6.13	5.69	0.22	0.54	4.65	0.70	0.159	0.015
33	13.1	6.64	5.51	0.31	0.24	4.89	0.72	0.094	0.017
34	11.6	5.39	3.77	0.29	0.06	4.51	0.56	0.110	0.015
35	11.8	5.44	3.94	0.30	0.04	4.64	0.59	0.085	0.013
36	18.3	1.28	0.67	0.03	0.03	1.96	0.65	0.001	0.014
37	15.8	2.39	0.63	0.01	0.04	1.94	1.29	0.001	0.014
38	18.0	1.50	0.67	0.01	0.06	2.11	0.92	0.001	0.016
39	18.0	1.88	0.68	0.01	0.04	2.00	1.11	0.006	0.022
40	20.8	1.51	0.72	0.07	0.10	2.37	1.26	0.002	0.016
41	17.7	1.12	0.56	0.06	0.06	2.06	0.79	0.001	0.013
42	18.3	1.14	0.67	0.06	0.05	2.11	0.89	0.006	0.019
43	16.7	0.92	0.53	0.01	0.05	1.76	0.91	0.004	0.013
44	14.8	2.74	0.67	0.03	0.05	2.15	1.34	0.003	0.015
45	19.1	1.64	0.60	0.10	0.03	1.75	1.04	0.007	0.018

O exemplo consiste em nove variáveis; 45 amostras; pequena sobreposição de dados presente em apenas algumas variáveis, Figura 3.32; sem presença de *outliers* significativos; e agrupamento em 3 *clusters* com partições diferentes. Assim como no exemplo anterior, apenas um fator não está no nível favorável, dessa vez é o número de *clusters*. Portanto, espera-se que este exemplo também apresente bons resultados para todos os métodos.

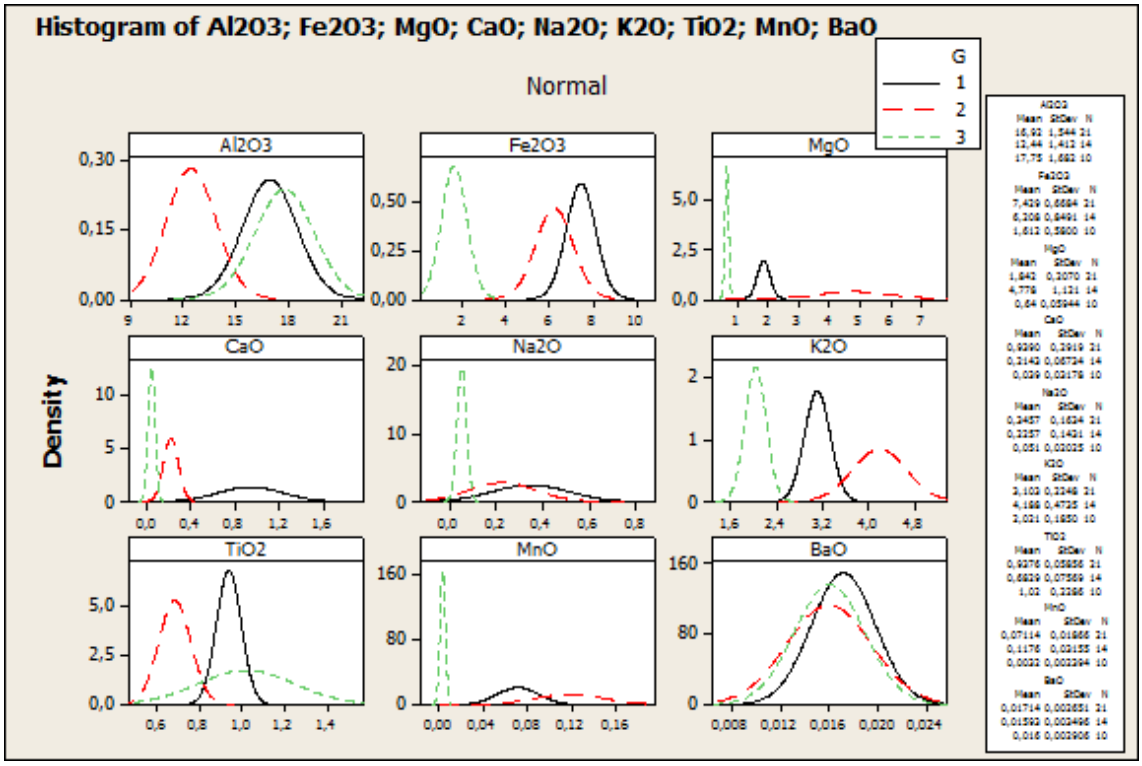

Figura 3.32 - Histograma *"Chemical Pottery"*. Fonte: Própria autora

Ao resolver o exemplo, os resultados obtidos pelos métodos SOM, *Ward* e *K-means* e foram iguais, com 100% de alocação correta. A alocação dos dados em cada *cluster* pode ser observada no Dendograma obtido pelo método *Ward*, Figura 3.33, e no resultado do SOM, Apêndice C, em que o *cluster* 1 recebeu 20 elementos, o *cluster* 2 obteve 15, e o *cluster* 3 ficou com 10 elementos.

No Apêndice C é possível verificar a ativação de cada elemento durante o método SOM e sua localização no mapa topológico, sendo que os elementos com localização (1,1) pertencem ao agrupamento 1, enquanto que os elementos (2,1) pertencem ao agrupamento 2 e os elementos (3,1) estão alocados no agrupamento 3.

Além disso, nos resultados do SOM e dos métodos *Ward* e *K-means (*Figura 3.34), também nota-se que o *clusters* 2 representa o grupo de cerâmica que, em geral, contém menor quantidade de Al2O3, e maior quantidade de K2O e MnO; enquanto que o *cluster* 3 apresenta elementos com alta composição de Al2O3 e TiO2, e baixa composição dos demais elementos; e, por fim, o *cluster* 1 contém elementos caracterizados por, principalmente, alta composição de Fe2O3 e CaO.

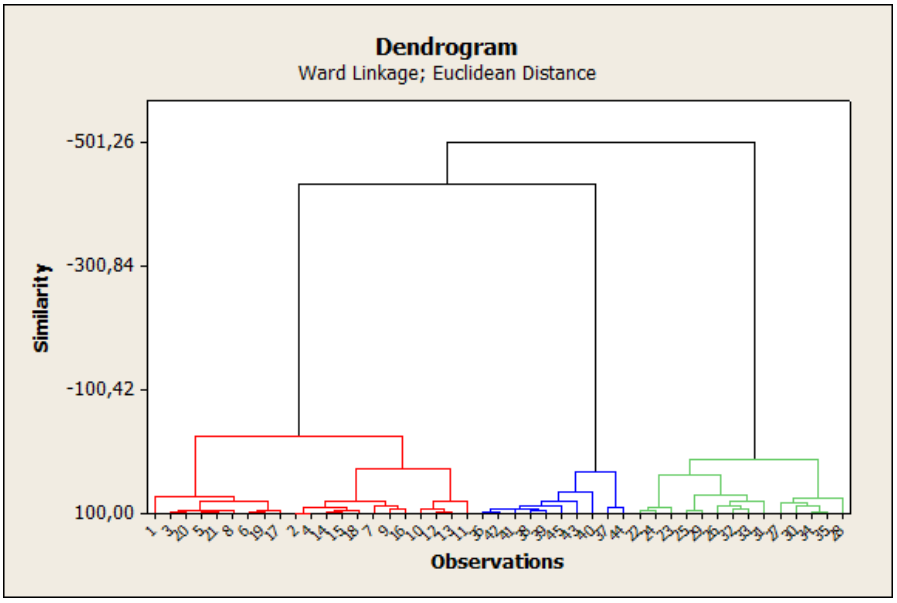

Figura 3.33 - Dendograma "*Chemical Pottery*". Fonte: Própria autora

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	20	46,190	1,316	2,662
Cluster2	15	70,311	2,045	3,787
Cluster3	10	29,315	1,378	3,177

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centroid
Al2O3	17,0800	12,5200	17,7500	15,7089
Fe2O3	7,5085	6,1827	1,6120	5,7562
MgO	1,8595	4,5593	0,6400	2,4884
CaO	0,9530	0,2440	0,0390	0,5136
Na2O	0,3565	0,2193	0,0510	0,2429
K2O	3,1455	4,0587	2,0210	3,2000
TiO2	0,9470	0,6873	1,0200	0,8767
MnO	0,0730	0,1121	0,0032	0,0705
BaO	0,0174	0,0157	0,0160	0,0165

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0,0000	5,5915	6,2376
Cluster2	5,5915	0,0000	8,2432
Cluster3	6,2376	8,2432	0,0000

Figura 3.34 - Resultado Ward e K-means- "*Chemical Pottery*". Fonte: Própria autora.

3.3.10 Conclusões e recomendações

Essa última etapa sintetiza os resultados, identifica as limitações práticas e teóricas encontradas, indica recomendações para futuros trabalhos e ressalta as conclusões gerais obtidas. Esta etapa é importante porque demostra que o estudo desenvolvido é um processo de aprendizado contínuo (MONTGOMERY, 2005), e os resultados obtidos servirão como base para futuros estudos e aplicações práticas de agrupamento.

3.4 Considerações Finais

Este capítulo teve dois objetivos principais: apresentar o método experimental utilizado para que os objetivos definidos para o presente trabalho pudessem ser alcançados; e descrever a aplicação do método seguindo as etapas definidas através do fluxograma apresentado. Considerando que o objeto de estudo foi considerado como um problema de modelagem de dados e análise de experimentos, o método experimental foi combinado de modo a mesclar fases do método de modelagem e simulação com o método de experimentação, e assim foi desenvolvido um procedimento experimental focado neste trabalho, o que contribuiu para que resultados importantes fossem obtidos.

4 CONCLUSÕES

A análise de *cluster* é amplamente usada em várias áreas para resolver problemas reais importantes, e a precisão da solução final depende do método de agrupamento usado. Motivado por isso, foi desenvolvido um estudo comparativo entre os métodos de *clustering Ward, K-means* e SOM a fim de avaliar o desempenho de cada método. A análise baseou-se em um conjunto de dados sintéticos, criados por DOE, cujos fatores foram: número de variáveis, número de clusters, partição dos *clusters*, tamanho da amostra, sobreposição de *clusters* e presença de *outliers*. Para resolvê-los pelo método *Ward e K-means*, utilizou-se o software Minitab® e, para o método SOM, o software Statistica®. As partições dos *clusters* foram comparadas pelo método "*Attribute Agreement Analysis*" e analisadas por técnicas estatísticas.

Os resultados apresentados neste trabalho mostram que os desempenhos dos métodos de *clustering* não podem ser comparados entre si, pois o poder do teste de hipótese é baixo para detectar a diferença entre os desempenhos dos métodos. Portanto, a um nível de significância de 5%, não é possível afirmar qual o mlehor método. No entanto, é possível afirmar que todos os métodos são significativamente afetados pelo

número de *clusters*, pela sobreposição e pela interação entre a partição dos *clusters* e o número de variáveis.

Em relação ao número de clusters, para $k = 2$ todos os métodos apresentaram bons resultados e para $k = 4$ todos os métodos tiveram uma queda abrupta em seus desempenhos. O mesmo é encontrado em Mingoti e Lima (2006), Balakrishnan *et al.* (1994), e Waller *et al.* (1998). Para situações com baixo nível de sobreposição, todos os métodos tiveram bons desempenhos e, para um alto nível de sobreposição, os desempenhos foram piores. Isso corrobora com Mangiameli, Chen e West (1996), que sugere que todos os métodos apresentam melhor desempenho para amostras sem sobreposição e com Waller *et al.* (1998), que propõe que o SOM é o método que apresenta o melhor resultado para altos níveis de sobreposição. Quanto ao número de variáveis, o melhor resultado para todos os métodos foi encontrado em $p = 6$. Esta análise corrobora com Balakrishnan *et al.* (1994) e Mingoti e Lima (2006), que sugerem uma melhora no desempenho de todos os métodos com o aumento do número de variáveis. O tamanho da amostra não é significativo para ambos os métodos, e partição dos *clusters* apresentou-se pouco significativo. Quando *outliers* são introduzidos, os desempenhos de todos os métodos diminuem, como mostrado em Mangiameli, Chen e West (1996) e em Mingoti e Lima (2006). Mangiameli, Chen e West (1996) acrescentam que o desempenho dos métodos SOM e *Ward* são similares na presença de o*utliers*, e Mingoti e Lima (2006) enfatizam que os resultados do método *K-means* são superiores à SOM. A interação entre o número de variáveis e o tamanho da amostra de *cluster* também foi significativa, e os métodos *K-means* e SOM são muito afetados pelo número de variáveis quando a partição dos *clusters* são diferentes.

Esse trabalho difere de outros principalmente ao que diz respeito ao método usado para generalizar e simular os conjuntos de dados. Utilizou-se a técnica DOE, que permite uma combinação de níveis de fatores, e resultam em um arranjo de experimentos com estruturas controladas. No entanto, outros trabalhos apresentaram uma abordagem usual de tentativa e erro, o que pode levar a conclusões restritas.

Existe outra diferença entre os trabalhos mencionados. Alguns deles consideram apenas a interação quando sobreposição e *outliers* são introduzidos nos dados (MANGIAMELI; CHEN; WEST, 1996; MINGOTI; LIMA, 2006; WALLER *et al.*, 1998). Nenhum demostra a interação entre todos os fatores como foi realizado neste trabalho. A partição dos *clusters* é um fator que também não foi estudado nos artigos anteriores, bem como o tamanho das amostras. Todos os artigos especificaram o tamanho

da amostra, mas nenhum o utilizou como um fator controlado. Mesmo os resultados mostrando que o tamanho da amostra não é significativo, é um resultado importante, pois infere que os métodos de agrupamento estudados não têm seu desempenho afetado pelo tamanho da amostra.

Outra diferença é a aplicação dos métodos *Ward* e *K-means* que foram implementados usando o software Minitab® e a implementação do SOM que usou o software Statistica® com os parâmetros de treinamento cuidadosamente definidos.

Outra possível razão para conclusões diferente, é o fato de ter sido utilizado nesse trabalho o teste estatístico '*paired-t'* e '*power and sample size'* para comparar os resultados dos métodos. Caso contrário, se um teste estatístico não for aplicado, pode-se rejeitar a hipótese nula e afirmar que um método supera o outro, quando na verdade eles têm desempenhos semelhantes com 95% de nível de confiança, ou então, o experimento não tem poder para detectar diferenças no desempenho dos métodos. Nesse contexto, alguns autores acreditam que o SOM é o método de agrupamento que apresenta o melhor desempenho (MANGIAMELI; CHEN; WEST, 1996; WALLER *et al.*, 1998), enquanto outros indicam que o melhor método são os métodos tradicionais hierárquicos ou não hierárquicos (BALAKRISHNAN *et al.*, 1994; MINGOTI; LIMA, 2006).

Muitos outros estudos ainda podem ser realizados como exemplo: 1) Comparação da Rede Neural Artificial com outros métodos estatísticos usando delineamento de experimentos, 2) Comparação dos métodos de agrupamento usando outras métricas de similaridade diferente da Distância Euclidiana, 3) Geração do conjunto de dados por uma distribuição diferente da normal, e 4) Estudo de caso real comprovando os resultados obtidos neste trabalho.

Por fim, este estudo revelou que, independentemente do método escolhido, pode-se obter resultados satisfatórios, desde que o pesquisador conheça as principais características do conjunto de dados e aplique o método corretamente.

APÊNDICE A – Banco de dados

Algoritmo para geração dos bancos de dados. Exemplo: Banco de dados 1

Tabela 4 – Parâmetros do banco de dados 1. Fonte: Própria autora.

Número de variáveis	4
Tamanho da amostra	200
Número de *clusters*	2
Volume dos *clusters*	Igual
Overlapping	d=0,8
Outliers	0

Passo 1. Geram-se quatro amostras aleatórias (*x1g1; x2g1, x3g1, x4g1)* de tamanho 100 por uma distribuição normal N (0; 1);

Passo 2. Geram-se quatro amostras aleatórias (*x1g2; x2g2, x3g2, x4g2)* de tamanho 100 por uma distribuição normal N (0,8; 1);

Passo 3. Agrupam-se as amostras formando quatro amostras finais (*X1, X2, X3, X4*) de modo que *X1=x1g1+x1g2; X2=x2g1+x2g2; X3=x3g1+x3g2; X4=x4g1+x4g2.*

Nota: Para bancos de dados com outliers, deve ser geradas amostras adicionais nos passos 1 e 2 por uma distribuição normal N (0; 5) e N (0,8; 5).

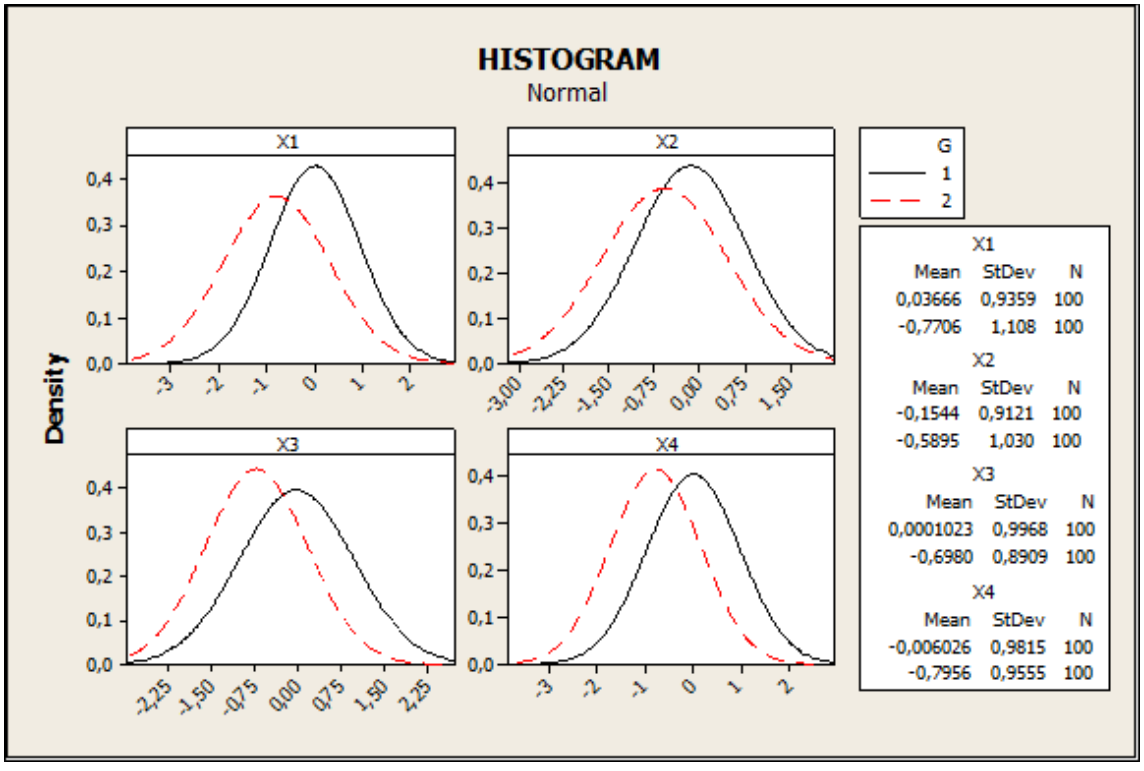

Figura 4 – Histograma do banco de dados 1

APÊNDICE B – *Ward*: Resultado 1

Cluster Analysis of Observations: x1; x2; x3; x4

Euclidean Distance, Ward Linkage
Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	199	98,141	0,1480	133	167	133	2
2	198	97,299	0,2150	136	177	136	2
3	197	97,270	0,2173	41	87	41	2
4	196	96,722	0,2609	32	89	32	2
5	195	96,322	0,2928	153	190	153	2
6	194	96,137	0,3076	96	163	96	2
7	193	95,915	0,3252	36	61	36	2
8	192	95,581	0,3518	74	128	74	2
9	191	95,300	0,3742	100	123	100	2
10	190	95,197	0,3823	113	154	113	2
11	189	95,187	0,3832	112	143	112	2
12	188	95,151	0,3860	10	77	10	2
13	187	95,061	0,3932	58	126	58	2
14	186	95,057	0,3936	28	101	28	2
15	185	94,987	0,3991	25	161	25	2
16	184	94,637	0,4270	78	80	78	2
17	183	94,544	0,4344	26	187	26	2
18	182	94,439	0,4427	97	137	97	2
19	181	94,331	0,4513	74	111	74	3
20	180	94,203	0,4615	16	141	16	2
21	179	94,099	0,4698	60	173	60	2
22	178	93,884	0,4869	76	134	76	2
23	177	93,747	0,4978	117	133	117	3
24	176	93,733	0,4989	12	99	12	2
25	175	93,733	0,4990	15	17	15	2
26	174	93,717	0,5002	43	118	43	2
27	173	93,699	0,5016	198	200	198	2
28	172	93,668	0,5041	19	65	19	2
29	171	93,319	0,5318	98	197	98	2
30	170	93,255	0,5370	38	149	38	2
31	169	93,232	0,5388	8	50	8	2
32	168	93,186	0,5425	120	139	120	2
33	167	92,935	0,5625	7	40	7	2
34	166	92,598	0,5893	2	194	2	2
35	165	92,553	0,5929	66	142	66	2
36	164	92,545	0,5935	27	138	27	2
37	163	92,457	0,6005	67	191	67	2
38	162	92,391	0,6057	14	82	14	2
39	161	92,248	0,6172	54	140	54	2
40	160	92,205	0,6206	30	104	30	2
41	159	92,094	0,6294	116	196	116	2
42	158	91,986	0,6380	55	180	55	2
43	157	91,641	0,6654	10	62	10	3
44	156	91,531	0,6742	25	117	25	5
45	155	91,473	0,6788	2	170	2	3
46	154	91,404	0,6843	122	165	122	2
47	153	91,381	0,6862	35	45	35	2
48	152	91,306	0,6921	44	83	44	2
49	151	91,261	0,6957	6	135	6	2
50	150	90,966	0,7192	150	168	150	2
51	149	90,925	0,7225	37	39	37	2
52	148	90,796	0,7327	47	90	47	2
53	147	90,700	0,7404	145	182	145	2
54	146	90,668	0,7429	58	88	58	3
55	145	90,589	0,7492	13	31	13	2
56	144	90,541	0,7531	28	41	28	4
57	143	90,435	0,7614	1	147	1	2
58	142	90,403	0,7641	22	36	22	3
59	141	90,355	0,7679	108	178	108	2
60	140	90,195	0,7806	113	124	113	3
61	139	90,185	0,7814	162	188	162	2
62	138	90,128	0,7859	20	64	20	2
63	137	89,924	0,8021	19	166	19	3

64	136	89,905	0,8036	51	136	51	3
65	135	89,853	0,8078	49	199	49	2
66	134	89,729	0,8177	20	35	20	4
67	133	89,625	0,8259	112	158	112	3
68	132	89,520	0,8344	9	32	9	3
69	131	89,343	0,8484	63	179	63	2
70	130	89,158	0,8632	5	68	5	2
71	129	89,120	0,8662	107	172	107	2
72	128	89,113	0,8668	23	48	23	2
73	127	89,082	0,8692	42	91	42	2
74	126	89,049	0,8718	2	153	2	5
75	125	88,899	0,8838	55	76	55	4
76	124	88,826	0,8896	19	79	19	4
77	123	88,716	0,8983	37	160	37	3
78	122	88,694	0,9001	120	189	120	3
79	121	88,632	0,9050	102	175	102	2
80	120	88,625	0,9056	159	186	159	2
81	119	88,355	0,9271	3	121	3	2
82	118	88,311	0,9306	57	106	57	2
83	117	88,167	0,9420	69	71	69	2
84	116	87,945	0,9597	12	195	12	3
85	115	87,838	0,9682	18	75	18	2
86	114	87,783	0,9726	34	60	34	3
87	113	87,719	0,9777	96	144	96	3
88	112	87,256	1,0146	38	130	38	3
89	111	87,174	1,0211	1	103	1	3
90	110	87,118	1,0256	171	198	171	3
91	109	86,923	1,0411	145	184	145	3
92	108	86,687	1,0598	53	95	53	2
93	107	86,511	1,0738	30	185	30	3
94	106	86,451	1,0786	78	169	78	3
95	105	86,399	1,0828	7	46	7	3
96	104	86,304	1,0903	6	29	6	3
97	103	86,279	1,0923	73	74	73	4
98	102	86,277	1,0925	108	115	108	3
99	101	86,188	1,0996	9	15	9	5
100	100	86,085	1,1078	25	183	25	6
101	99	85,540	1,1512	159	164	159	3
102	98	85,490	1,1552	25	43	25	8
103	97	85,428	1,1601	67	156	67	3
104	96	85,314	1,1692	4	85	4	2
105	95	85,151	1,1821	6	86	6	4
106	94	85,051	1,1901	12	84	12	4
107	93	85,031	1,1917	33	181	33	2
108	92	84,794	1,2105	66	105	66	3
109	91	84,724	1,2161	122	155	122	3
110	90	84,608	1,2254	54	151	54	3
111	89	84,370	1,2443	26	55	26	6
112	88	84,288	1,2508	58	97	58	5
113	87	83,864	1,2846	14	44	14	4
114	86	83,639	1,3025	59	94	59	2
115	85	83,309	1,3288	30	150	30	5
116	84	83,164	1,3403	8	27	8	4
117	83	82,962	1,3564	9	92	9	6
118	82	82,682	1,3787	102	176	102	3
119	81	82,651	1,3812	98	129	98	3
120	80	82,564	1,3881	127	131	127	2
121	79	81,285	1,4899	4	70	4	3
122	78	81,178	1,4984	5	100	5	4
123	77	81,125	1,5027	157	159	157	4
124	76	81,123	1,5028	10	34	10	6
125	75	81,101	1,5045	13	30	13	7
126	74	80,964	1,5155	107	192	107	3
127	73	80,791	1,5292	11	125	11	2
128	72	80,678	1,5382	1	3	1	5
129	71	80,367	1,5630	109	146	109	2
130	70	80,133	1,5816	112	148	112	4
131	69	80,103	1,5840	66	113	66	6
132	68	80,064	1,5871	51	119	51	4
133	67	79,379	1,6417	21	47	21	3
134	66	78,946	1,6761	93	110	93	2
135	65	78,729	1,6934	162	171	162	5
136	64	78,215	1,7343	5	23	5	6
137	63	78,155	1,7391	102	114	102	4
138	62	77,880	1,7610	72	152	72	2
139	61	77,280	1,8088	108	174	108	4
140	60	76,571	1,8652	20	52	20	5

141	59	76,551	1,8668	18	59	18	4
142	58	76,002	1,9105	19	33	19	6
143	57	75,876	1,9206	53	93	53	4
144	56	75,354	1,9620	73	96	73	7
145	55	75,059	1,9856	10	37	10	9
146	54	74,796	2,0065	5	28	5	10
147	53	74,433	2,0354	67	78	67	6
148	52	73,772	2,0880	49	81	49	3
149	51	73,690	2,0945	22	49	22	6
150	50	73,178	2,1353	16	98	16	5
151	49	72,591	2,1821	2	69	2	7
152	48	72,468	2,1918	6	38	6	7
153	47	72,074	2,2232	63	72	63	4
154	46	71,948	2,2333	12	14	12	8
155	45	71,470	2,2713	18	56	18	5
156	44	69,476	2,4301	107	108	107	7
157	43	68,811	2,4830	116	120	116	5
158	42	68,150	2,5356	2	112	2	11
159	41	67,561	2,5825	42	54	42	5
160	40	67,154	2,6149	73	145	73	10
161	39	66,317	2,6816	8	19	8	10
162	38	65,951	2,7107	51	122	51	7
163	37	65,515	2,7454	11	127	11	4
164	36	64,619	2,8167	9	21	9	9
165	35	63,329	2,9194	132	162	132	6
166	34	62,365	2,9962	12	24	12	9
167	33	62,061	3,0204	7	63	7	7
168	32	60,199	3,1686	109	193	109	3
169	31	59,133	3,2534	25	58	25	13
170	30	50,097	3,9728	6	26	6	13
171	29	48,476	4,1019	16	132	16	11
172	28	47,608	4,1710	5	57	5	12
173	27	43,672	4,4844	4	10	4	12
174	26	41,428	4,6630	11	116	11	9
175	25	40,602	4,7287	2	66	2	17
176	24	37,506	4,9752	67	157	67	10
177	23	37,463	4,9786	42	53	42	9
178	22	35,843	5,1076	8	12	8	19
179	21	35,821	5,1093	13	25	13	20
180	20	27,762	5,7509	7	18	7	12
181	19	26,826	5,8254	102	107	102	11
182	18	17,147	6,5960	9	11	9	18
183	17	16,138	6,6763	51	73	51	17
184	16	15,444	6,7316	102	109	102	14
185	15	6,434	7,4489	5	8	5	31
186	14	-0,113	7,9701	4	6	4	25
187	13	-2,666	8,1733	51	67	51	27
188	12	-8,695	8,6533	20	42	20	14
189	11	-12,169	8,9299	1	16	1	16
190	10	-35,914	10,8202	5	22	5	37
191	9	-59,876	12,7278	1	2	1	33
192	8	-62,303	12,9210	7	9	7	30
193	7	-83,314	14,5938	4	13	4	45
194	6	-84,823	14,7139	51	102	51	41
195	5	-164,422	21,0509	1	51	1	74
196	4	-181,183	22,3852	5	20	5	51
197	3	-222,441	25,6698	5	7	5	81
198	2	-242,510	27,2675	4	5	4	126
199	1	-776,677	69,7930	1	4	1	200

Final Partition
Number of clusters: 2

APÊNDICE C – Resultado SOM *"Chemical Pottery"*

Neurônio	ID	Ativação	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
(1, 1)	1	0,3633	18,8000	9,5200	2,0000	0,7900	0,4000	3,2000	1,0100	0,0770	0,0150
(1, 1)	1	0,1793	16,9000	7,3300	1,6500	0,8400	0,4000	3,0500	0,9900	0,0670	0,0180
(1, 1)	1	0,2774	18,2000	7,6400	1,8200	0,7700	0,4000	3,0700	0,9800	0,0870	0,0140
(1, 1)	1	0,2668	16,9000	7,2900	1,5600	0,7600	0,4000	3,0500	1,0000	0,0630	0,0190
(1, 1)	1	0,2667	17,8000	7,2400	1,8300	0,9200	0,4300	3,1200	0,9300	0,0610	0,0190
(1, 1)	1	0,2460	18,8000	7,4500	2,0600	0,8700	0,2500	3,2600	0,9800	0,0720	0,0170
(1, 1)	1	0,6019	16,5000	7,0500	1,8100	1,7300	0,3300	3,2000	0,9500	0,0660	0,0190
(1, 1)	1	0,1924	18,0000	7,4200	2,0600	1,0000	0,2800	3,3700	0,9600	0,0720	0,0170
(1, 1)	1	0,2020	15,8000	7,1500	1,6200	0,7100	0,3800	3,2500	0,9300	0,0620	0,0170
(1, 1)	1	0,4231	14,6000	6,8700	1,6700	0,7600	0,3300	3,0600	0,9100	0,0550	0,0120
(1, 1)	1	0,7114	13,7000	5,8300	1,5000	0,6600	0,1300	2,2500	0,7500	0,0340	0,0120
(1, 1)	1	0,4988	14,6000	6,7600	1,6300	1,4800	0,2000	3,0200	0,8700	0,0550	0,0160
(1, 1)	1	0,4407	14,8000	7,0700	1,6200	1,4400	0,2400	3,0300	0,8600	0,0800	0,0160
(1, 1)	1	0,3472	17,1000	7,7900	1,9900	0,8300	0,4600	3,1300	0,9300	0,0900	0,0200
(1, 1)	1	0,3559	16,8000	7,8600	1,8600	0,8400	0,4600	2,9300	0,9400	0,0940	0,0200
(1, 1)	1	0,7066	15,8000	7,6500	1,9400	0,8100	0,8300	3,3300	0,9600	0,1120	0,0190
(1, 1)	1	0,2546	18,6000	7,8500	2,3300	0,8700	0,3800	3,1700	0,9800	0,0810	0,0180
(1, 1)	1	0,6429	16,9000	7,8700	1,8300	1,3100	0,5300	3,0900	0,9500	0,0920	0,0230
(1, 1)	1	0,3623	18,9000	7,5800	2,0500	0,8300	0,1300	3,2900	0,9800	0,0720	0,0150
(1, 1)	1	0,4100	18,0000	7,5000	1,9400	0,6900	0,1200	3,1400	0,9300	0,0350	0,0170
(1, 1)	1	0,2450	17,8000	7,2800	1,9200	0,8100	0,1800	3,1500	0,9000	0,0670	0,0170
(2, 1)	2	0,6352	14,4000	7,0000	4,3000	0,1500	0,5100	4,2500	0,7900	0,1600	0,0190
(2, 1)	2	0,5551	13,8000	7,0800	3,4300	0,1200	0,1700	4,1400	0,7700	0,1440	0,0200
(2, 1)	2	0,4961	14,6000	7,0900	3,8800	0,1300	0,2000	4,3600	0,8100	0,1240	0,0190
(2, 1)	2	0,5384	11,5000	6,3700	5,6400	0,1600	0,1400	3,8900	0,6900	0,0870	0,0090
(2, 1)	2	0,5051	13,8000	7,0600	5,3400	0,2000	0,2000	4,3100	0,7100	0,1010	0,0210
(2, 1)	2	0,5429	10,9000	6,2600	3,4700	0,1700	0,2200	3,4000	0,6600	0,1090	0,0100
(2, 1)	2	0,4851	10,1000	4,2600	4,2600	0,2000	0,1800	3,3200	0,5900	0,1490	0,0170
(2, 1)	2	0,2872	11,6000	5,7800	5,9100	0,1800	0,1600	3,7000	0,6500	0,0820	0,0150
(2, 1)	2	0,2807	11,1000	5,4900	4,5200	0,2900	0,3000	4,0300	0,6300	0,0800	0,0160
(2, 1)	2	0,5060	13,4000	6,9200	7,2300	0,2800	0,2000	4,5400	0,6900	0,1630	0,0170
(2, 1)	2	0,5197	12,4000	6,1300	5,6900	0,2200	0,5400	4,6500	0,7000	0,1590	0,0150
(2, 1)	2	0,3436	13,1000	6,6400	5,5100	0,3100	0,2400	4,8900	0,7200	0,0940	0,0170
(2, 1)	2	0,3509	11,6000	5,3900	3,7700	0,2900	0,0600	4,5100	0,5600	0,1100	0,0150
(2, 1)	2	0,4325	11,8000	5,4400	3,9400	0,3000	0,0400	4,6400	0,5900	0,0850	0,0130
(3, 1)	3	0,4774	18,3000	1,2800	0,6700	0,0300	0,0300	1,9600	0,6500	0,0010	0,0140
(3, 1)	3	0,4734	15,8000	2,3900	0,6300	0,0100	0,0400	1,9400	1,2900	0,0010	0,0140
(3, 1)	3	0,1052	18,0000	1,5000	0,6700	0,0100	0,0600	2,1100	0,9200	0,0010	0,0160
(3, 1)	3	0,4717	18,0000	1,8800	0,6800	0,0100	0,0400	2,0000	1,1100	0,0060	0,0220
(3, 1)	3	0,4512	20,8000	1,5100	0,7200	0,0700	0,1000	2,3700	1,2600	0,0020	0,0160
(3, 1)	3	0,3656	17,7000	1,1200	0,5600	0,0600	0,0600	2,0600	0,7900	0,0010	0,0130
(3, 1)	3	0,2610	18,3000	1,1400	0,6700	0,0600	0,0500	2,1100	0,8900	0,0060	0,0190
(3, 1)	3	0,3155	16,7000	0,9200	0,5300	0,0100	0,0500	1,7600	0,9100	0,0040	0,0130
(3, 1)	3	0,5588	14,8000	2,7400	0,6700	0,0300	0,0500	2,1500	1,3400	0,0030	0,0150
(3, 1)	3	0,2112	19,1000	1,6400	0,6000	0,1000	0,0300	1,7500	1,0400	0,0070	0,0180

REFERÊNCIAS

ADYA, M.; COLLOPY, F.L. How effective are neural networks at forecasting and prediction? A review and evaluation. **Journal of forecasting**, v.17, n.5-6, p.481-495, 1998.

ANDERBERG, M.R. **Cluster Analysis for Application**. New York: Academic Press, Inc., 1973.

APPOLINÁRIO, F. **Metodologia da ciência: filosofia e prática da pesquisa.** São Paulo: Pioneira Thomson Learning, 2006.

BALAKRISHNAN, P.V.; COOPER, M. C.; JACOB, V. S.; LEWIS, P. A. A study of the classification of neural networks using unsupervised learning: A comparison whit *K*-means clustering. **Psychometrika**, v.59, n.4, p.509-525, 1994.

BASHEER, I.A. AND HAJMEER, M. Artificial Neural Networks: Fundamentals, Computing, Design, and Application. **Journal of Microbiological Methods**, v. 43, p.3-31, 2000.

BRYMAN, A. **Research methods and organization studies.** London: Unwin Hyman, London, 1989. 283 p.

BERKHIN, P. **Survey of clustering data mining techniques**. Accrue Software, 2002.

BERTRAND, J. W. M., FRANSOO, J. C. Modelling and Simulation: operations management research methodologies using quantitative modeling. **International Journal of Operations & Production Management**, v.22, p.241-264, 2002.

BOX, G.E.P.; HUNTER, W.G.; HUNTER, J.S. **Statistics for Experimenters**, John Wiley & Sons, 1ed., 1978.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990.

CHUNG, C. A. **Simulation Modeling Handbook: a practical approach**. Washington, D. C: CRC Press, 2004.

COHEN, J. **Statistical power Analysis for the behavioral sciences.** 2ed. Hillsdale: Lawrence Erlbaum Associates, 1988.

COSTA, J.A.F. **Classificação Automática e Análise de Dados por Redes Neurais Auto-Organizáveis**, Tese de Doutorado, Faculdade de Engenharia Elétrica e de Computação (FEEC/Unicamp), 1999.

CRUZ, C. D.; REGAZZI, A. J. **Modelos biométricos aplicados ao melhoramento genético.** 2 ed. Viçosa: UFV, p.390, 1994.

DASCALESCU, L.; MEDLES, K.; DAS, K.; YOUNES, M.; CALIAP, L.; MIHALCIOIU, A., Using Design of Experiments and Virtual Instrumentation to Evaluate the Tribocharging of Pulverulent Materials in Compressed-Air Devices**, IEEE Transactions on Industry Applications,** v. 44, n. 1, p. 3–8, Jan/Feb. 2008.

DONI, M.V., **Análise de cluster: métodos hierárquicos e de particionamento.** São Paulo: Universidade Presbiteriana Mackenzie, 2004.

EVERITT, B. S.; LANDAU, S.; LEESE, M. **Cluster -analysis.** Arnold Publishers, 2001.

FASULO, D. **An analysis of Recent Work on Clustering Algorithms**, Technical Report 01-03-02, Department of Computer Science & Engineering, University of Washington, Washington, 1999.

FAYYAD, U. M., PIATETSKY, S. G., SMYTH, P., UTHURUSAMY, R., "Advances in Knowledge Discovery and Data Mining", **AAAIPress**, The Mit Press, 1996.

FERNEDA, E. **Redes Neurais e sua aplicação em sistemas de recuperação de informação**. Ci. Inf., Brasília, v. 35, n. 1, p. 25-30, 2006.

FUNG, G., **A Comprehensive Overview of Basic Clustering Algorithms**, 2002.

GERCHMAN, M. **Problemas de otimização na engenharia de produção e transporte.** Dissertação no Programa de Pós Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

GIESBRECHT, G.F, GUMPERTZ, M. **Planning, Construction, and Statistical Analysis of Comparative Experiments,** New Jersey: John Wiley & Sons, Inc., 2004.

GOMES, J.H.F. **Análise e otimização da soldagem de revestimento de chapas de aço ABNT 1020 com utilização de arama tubular inoxidável austenítico**. Dissertação no Programa de Pós Graduação em Engenharia de Produção, Universidade Federal de Itajubá, Itajubá, 2010.

GORDON, A. D. A review of hierarchical classification. **Journal of Royal Statistical Society**, v.150, n.2, p.119-137, 1987.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. Porto Alegre: Bookman, 2001.

HAN, J.; KAMBER, M., **Data Mining: Concepts and Techniques**, 2nd ed., San Francisco, USA: Morgan Kaufmann Publishers, 2006.

HEBB, D.O. **The Organization of Behavior.** John Wiley, New York, 1949.

HOCHDORFFER, J., LAULE, C., LANZA, G., "Product variety management using data-mining methods — Reducing planning complexity by applying clustering analysis on product portfolios," in IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2017.

HRUSHKA, E. R., EBECKAEN, N. F. F. **A Genetic algorithm for cluster analysis**. IEEE Transactions on Evolutionary Computation , 2001.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, New York, v. 31, n. 3, p. 265-323, Sept., 1999.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis.** New jersey, USA: Englewood Cliffs, p. 642, 1992.

KAUFMAN, L.; ROUSSEEUW, P.J. **Finding groups in data. An Introduction**. John Wiley & Sons,1990.

KERLINGER, Thomas C.; TAYLOR, James R. **Marketing research: an applied approach**. Tóquio:, McGraw-Hill Kogakusha, 1979.

KIM, H.; KIM B.; KIM, S.; PARK, C.; KIM, E.; AHN Y. Classification of attempted suicide by cluster analysis: A study of 888 suicide attempters presenting to the emergency department. **Journal of Affective Disorders**, v. 235, p.184-190, 2018.

KÖCHE, J. C. **Fundamentos de metodologia científica: teoria da ciência e iniciação à pesquisa.** 26. ed. Petrópolis, RJ: Vozes, 2013.

KOHONEN, T. **Self-Organizing Maps.** Springer-Verlag, Berlin, 1995.

KUGLER, M. & JÚNIOR, J. T. & LOPES, H. S. **Desenvolvimento de uma Rede Neural LVQ em Linguagem VHDL para Aplicações em Tempo-Real.** Proceedings of the VI Brazilian Conference on Neural Networks - VI Congresso Brasileiro de Redes Neurais, p. 103–108, 2003.

LEE, C. T.; GUZMAN, D.; PONATH, C.; TIEU, L.; RILEY, E.; KUSHEL, M. Residential patterns in older homeless adults: Results of a cluster analysis. **Social Science and Medicine**, v.153, p. 131-140, 2016.

LEE, J.; CHANG, J.; KANG, D.; KIM, S.; HONG, J., Tooth Shape Optimizationfor Cogging Torque Reduction of Transverse Flux Rotary Motor Using Design of Experiment and Response Surface Methodology, **IEEE Transactions on Magnetics,** v. 43, n. 4, p. 1817–1820, Apr. 2007.

LORSCHEID, I.; HEINE, B.; MEYER, M., Opening the black box of simulations: increased transparency and effective communication through the systematic design of experiments, **Computational and Mathematical Organization Theory**, v.18, p.22–62, 2012.

MA, X.; CHEN, S.; CHEN, F., ‘‘Multivariate space-time modeling of crash frequencies by injury severity levels,’’ **Anal. Methods Accident Res.,** vol. 15, pp. 29–40, Sep. 2017.

MACCHIAROLI, R., RIEMMA, S., “Clustering Methods for Production Planning and Scheduling in a Flexible Manufacturing System, in **IEEE International Conference on Robotics & Automation**, San Diego, CA, USA, 1994.

MANGIAMELI, P.; CHEN, S.K.; WEST, D. A comparison of SOM neural network and hierarchical clustering methods. **European Journal of Operation Research**, v.93, n.2, p.402-417, 1996.

MCCULLOCH, W.S., PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v.5, n.1, p. 115–133, 1943.

MIGUEL, P. A. C.; FLEURY, A.; MELLO, C. H. P.; NAKANO, D. N.; TURRIONI, J. B.; LEE HO, L.; MORABITO, R.; MARTINS, R. A.; PUREZA, V. **Metodologia de pesquisa em engenharia de produção e gestão de operações.** 2ª. Ed. Rio de Janeiro: Elsevier, 2014.

MILLIGAN, G.W.; COOPER, M.C. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. **Psychometrika**, v.50, n.1, p.123-127, 1980.

MITROFF I. I., BETZ F., PONDY L. R., SAGASTI F. On managing science in the system age: two schemas for the study of science as a whole system phenomenon. **Interfaces,** v.4, n.3, p.46-58, 1974.

MONGOTI, S. S.; LIMA, J. O. Comparing SOM neural network with Fuzzy c-means, *K*-means and traditional hierarchical clustering algorithms. **European Journal of Operation Research**, v.174, p.1742-1759, 2006.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 6 ed. New York: John Wiley, 2005.

MURTAGH, F., LEGENDRE, P., “Ward’s hierarchical agglomerative clustering method: Which algorithms implement ward’s criterion**. Journal of Classification**, v.31, p.274-295, 2014.

NACHTWEY, R. RIEDEL, E. MUELLER, “Cluster analysis as a method for the planning of production systems,” in **IEEE International Conference on Computers & Industrial Engineering**, Troyes, France, 2009.

PEREIRA, J.R.G. **Um estudo sobre alguns métodos hierárquicos para análise de agrupamento.** Dissertação de Mestrado. Instituo de Matemática, Estatística e Ciência da Computação (UNICAMP), 1993.

PEREIRA, I. C. **Proposta de sistematização da simulação para fabricação em lotes**. Dissertação no Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Itajubá, Itajubá, 2000.

RODRIGUEZ, F. S. **Métodos de agrupamento na análise de dados de expressão gênica**. Dissertação de Pós-Graduação em Estatística, Universidade Federal de São Carlos, São Carlos, 2009.

ROSENTHAL, R. Parametric measures of effect size. Em H. Cooper e L. V. Hedges (Eds.). **The handbook of research synthesis**. New York: Russell Sage, p. 231-244, 1994.

SANTANA, C. M.; MALINOVSKI, J. R. Uso da análise multivariada no estudo de fatores humanos em operadores de motosserra, **Cerne**, v.8, n.2, p.101-107, 2002.

STATSOFT. **Statistica**, 2005.

SIEGMUND, K. D.; LAIRD, P. W.; LAIRDOFFRINGA, I.A. A comparison of cluster analysis methods using DNA methykation data. **Bioinformatics**, v.20, n.12, p.1896-1904, 2004.

SORENSON, T. A Method of Establishing Groups of Equal Amplitudes in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. **Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter**, v.5, p.1-34, 1948.

STAICULESCU, D.; BUSHYAGER, N.; OBATOYINBO, A.; MARTIN, L. J.; TENTZERIS, M. M., Design and Optimization of 3-D Compact Stripline and Microstrip Bluetooth/WLAN Balun Architectures Using the Design of Experiments Technique, **IEEE Transactions on Antennas and Propagation**, v.53, n.5, p. 1805–1812, May 2005.

YU, Z., CHEN, H., YOU, J., LIU, J., WONG, HAN, G., LI, L., “Adaptive Fuzzy Consensus Clustering Framework for Clustering Analysis of Cancer Data,” **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v.12, n. 4, Jul/Aug 2015.

WALLER, N. G.; KAISER, H. A.; ILLIAN, J. B.; MANRY, M. A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning ad three hierarchical cluster analysis algorithms. **Psychometrika**, v.63, n.1, p.5-22, 1998.

WARD, J.H. Hierarchical grouping to optimize an objective function. **Journal of America Statistical Association**, v.58, n.1, p.236-244, 1963.

WEIGEND, A.S., RUMELHART, D.E., HUBERMAN, B.A. **Generalization by weigth-elimination with application to forecasting**. Proceedings of the 1990 conference on Advances in neural information processing system 3, Denver, Colorado, USA, p. 875-882, 1990.