CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS

UNIDADE ARAXÁ

ARTUR EMILIO ALVES NASCIMENTO

APLICAÇÃO DE REDES DE KOHONEN PARA DESAGRUPAMENTO DE DADOS PREFERENCIALMENTE AMOSTRADOS

ARAXÁ, MG

2022

ARTUR EMILIO ALVES NASCIMENTO

APLICAÇÃO DE REDES DE KOHONEN PARA DESAGRUPAMENTO DE DADOS PREFERENCIALMENTE AMOSTRADOS

Trabalho de Conclusão de Curso apresentado ao Centro Federal de Educação Tecnológica de Minas Gerais - Unidade Araxá como parte integrante dos pré-requisitos para obtenção do título de bacharel em Engenharia de Minas.

Orientador: Prof° Dr. Allan Erlikhman Medeiros Santos

Co-orientadora: Prof ª Me Silvânia Alves Braga de Castro

ARAXÁ– MG

2022

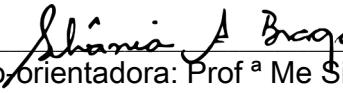
ARTUR EMILIO ALVES NASCIMENTO

APLICAÇÃO DE REDES DE KOHONEN PARA DESAGRUPAMENTO DE DADOS PREFERENCIALMENTE AMOSTRADOS

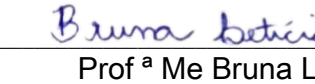
Trabalho de Conclusão de Curso apresentado ao Centro Federal de Educação Tecnológica de Minas Gerais - Unidade Araxá como parte integrante dos pré-requisitos para obtenção do título de bacharel em Engenharia de Minas.

Data de Defesa: Araxá, 10 de fevereiro de 2022.

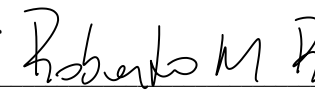
Orientador: Prof. Dr. Allan Erlikhman Medeiros Santos
Centro Federal de Educação Tecnológica de Minas Gerais - Unidade Araxá



Co-orientadora: Prof ª Me Silvânia Alves Braga de Castro
Centro Federal de Educação Tecnológica de Minas Gerais - Unidade Araxá



Prof ª Me Bruna Letícia dos Santos
Centro Federal de Educação Tecnológica de Minas Gerais - Unidade Araxá



Eng. Dr. Roberto Mentzingen Rolo
Universidade Federal do Rio Grande do Sul.

RESUMO

O processo de avaliação de recursos minerais é uma tarefa crucial que conduz a quantificação da mineralização e na qual, sob comprovação da viabilidade econômica, no todo ou em parte, pode tornar a lavra economicamente viável. Destaca-se que tal processo é feito, a partir dos conhecimentos de geologia e da topografia local, com base em dados de amostragens. Nesse sentido, a quantidade e qualidade dos dados amostrais são premissas fundamentais para a confiabilidade das estimativas desses recursos. Ao contrário disso, amostragens preferenciais em zonas de interesse, conduzem resultados enviesados. Em particular, diferentes fatores podem fazer com que áreas sejam preferencialmente amostradas, destacando as condições de acessibilidade, valores de atributos esperados e a própria estratégia de amostragem. Atualmente é sabido que a amostragem preferencial impacta a estatística da área levando a resultados incoerentes com a realidade. Posto isso, existem métodos para correção destes impactos amplamente utilizados na literatura, tais como o método das células móveis (*Cell Declustering*) e o método da poligonal (*Polygonal Declustering*). A presente pesquisa tem como objetivo principal a proposição de uma nova abordagem para as operações de desagrupamento de dados amostrais por meio da utilização da técnica de redes de Kohonen, conhecidas como *Self-Organizing Maps* (redes SOM). As redes de Kohonen são um tipo de rede neural artificial, utilizadas para classificação não-supervisionada. O princípio metodológico desta pesquisa parte da premissa de atribuir a cada amostra um peso para o cálculo da média, quanto maior for o adensamento amostral menor será o peso atribuído e quanto menor for o adensamento amostral maior será o peso, o somatório da multiplicação do peso com a amostra resultará na média desagrupada. A determinação dos pesos será feita pela classificação das redes de Kohonen. A pesquisa iniciará apresentando o potencial das redes de Kohonen para desagrupamento em dois bancos de dados conhecidos na literatura: *Walker Lake,* cuja variável de estudo é o elemento cobre e o banco de dados do Carvão que contém dados da espessura. Os resultados serão comparados com os métodos tradicionais de desagrupamento. A presente pesquisa não tem como objetivo substituir os métodos clássicos de desagrupamento, pelo contrário, apresentar uma nova abordagem para um problema de rotina na avaliação de reservas. Embora a matemática da técnica aplicada seja de fato complexa, os resultados podem ser promissores.

Palavras-chave: Redes SOM. Métodos de desagrupamento. Geoestatística.

ABSTRACT

The process of mineral resource evaluation is a crucial task that leads to the quantification of mineralization and which, upon proof of economic feasibility, in whole or in part, can make mining economically viable. It is noteworthy that such a process is done, based on knowledge of geology and local topography, and on sampling data. In this sense, the quantity and quality of sampling data are fundamental premises for the reliability of resource estimates. In contrast, preferential sampling in areas of interest leads to biased results. Different factors can cause areas to be preferentially sampled, highlighting accessibility conditions, expected attribute values, and the sampling strategy itself. It is currently known that preferential sampling impacts area statistics leading to results that are inconsistent with reality. That said, there are methods to correct these impacts widely used in the literature, such as the mobile cell method (Cell Declustering) and the polygonal method (Polygonal Declustering). The main objective of the present research is to propose a new approach to sample data disaggregation operations by using the Kohonen network technique, known as Self-Organizing Maps (SOM networks). Kohonen nets are a type of artificial neural network used for unsupervised classification. The methodological principle of this research starts from the premise of assigning each sample a weight to calculate the average, the greater the sample density the lower the weight assigned and the lower the sample density the higher the weight, the sum of the multiplication of the weight with the sample will result in the disaggregated average. The determination of the weights will be done by Kohonen's classification of the networks. The research will begin by presenting the potential of Kohonen's network for declustering methods in two databases known in the literature: Walker Lake, whose study variable is the element copper and the Coal database that contains thickness data. The results will be compared with traditional declustering methods The present research is not intended to replace classical declustering methods, but rather to present a new approach to a routine problem in reserve evaluation. Although the mathematics of the applied technique is indeed complex, the results can be promising.

Keywords: Self-organizing map. Declustering methods. Geostatistics.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O processo de amostragem corresponde a uma sequência de operações sistemáticas que visam representar, mediante a coleta de uma pequena porção denominada amostra, determinado universo. Diante disso, pode-se considerar que tal etapa pode ser considerada chave para o sucesso da etapa de Exploração Mineral, uma vez que amostras representativas fornecem subsídios para definir uma base de dados adequada que permita conhecer a continuidade geológica e a estimativa do potencial exploratório daquela área.

Nesse sentido, uma das principais dificuldades durante o processo de execução é cumprir o plano de amostragem de tal forma que se respeite a disposição espacial dos pontos a serem amostrados, segundo a estratégia previamente estabelecida no plano. Isso se deve ao fato de que devido à presença de cursos d'água, vegetação, falta de acesso à área, barreiras geográficas ou interesses em determinadas regiões resultam em um cenário em que a malha de amostragem torna-se irregular.

Como efeito, podem existir uma maior quantidade de pontos amostrados em determinadas regiões em prejuízo de outras gerando, por consequência, um agrupamento preferencial de pontos de amostragem. Em particular, no caso da mineração, tal fato é muito frequente uma vez que, o processo de amostragem visa zonas em regiões de altos teores que possam garantir (potencialmente) a viabilidade econômica do projeto.

Diante disso, é necessário ajustar o resultado da estatística de forma a evitar resultados não realísticos baseados em erros ou desvios na interpretação. Nesse sentido, os estudos em Geoestatística contribuem fornecendo métodos consagrados como o método de células móveis e o método da poliginal (*polygonal declustering*). Tais métodos permitem anular a influência que o adensamento das amostras provoca nas estatísticas aplicando ponderadores sobre o conjunto de dados agrupados.

Deste modo, o presente trabalho apresenta uma aplicação da inteligência artificial por meio da técnica de *Machine Learning*. Dentre todas as técnicas relacionadas a esse campo foi escolhido o modelo de rede neural artificial (RNA) baseado nas Redes SOM (*Self Organizing Map*)

também conhecidas como Redes de *Kohonen*, nome homônimo de seu criador, como ferramenta para identificar os adensamentos e por fim realizar o desagrupamento .

O presente trabalho apresenta uma avaliação da aplicação do desagrupamento utilizando as Redes SOM por meio da manipulação de dois bancos de dados *Walker Laker* que apresenta como variável o elemento cobre e um outro banco de dados contendo espessura de um depósito de Carvão conforme metodologia descrita ao longo dessa pesquisa. Dessa maneira, a partir dos resultados obtidos, poderão ser mensurados evidências iniciais da aplicabilidade (ou não) como uma ferramenta alternativa aos métodos clássicos a ser utilizada no desagrupamento de dados amostrais.

2. REVISÃO BIBLIOGRÁFICA

2.1 Amostragem

A amostragem torna-se necessária na medida em que não é possível ou não é conveniente acessar a totalidade de um universo amostral ou uma população. Assim, a partir desse, seleciona-se um determinado grupo, uma amostra, com o intuito de reduzir o tamanho do todo para se obter uma interpretação. Como tal, espera-se que a amostra reproduza com confiabilidade as características essenciais da população (MOON *et al*, 2006).

Conforme Yamamoto e Rocha (2001) destacam, no processo de avaliação de recursos minerais, a amostragem é a etapa inicial. Essa etapa é responsável por fornecer dados a serem utilizados no estudo da distribuição espacial dos teores associados a um elemento de interesse, além de possibilitar a inferência da continuidade geológica, geometria e a relação com as rochas encaixantes da mineralização.

Para tanto, o plano de amostragem deverá contemplar a área a ser amostrada e o método de amostragem que por sua vez, em particular, define a densidade amostral, isto é, a quantidade de pontos amostrais por unidade de área, o espaçamento dos pontos amostrais e a presença de suportes amostrais. Pinto e Deutsch (2017) destacam que o espaçamento de dados tem sido usado como uma medida da disponibilidade de dados e para avaliar os esquemas de amostragem.

Por outro lado, o suporte de dados diz respeito ao tamanho, forma e orientação das amostras, ou seja, volume n-dimensional no qual pode-se tomar medidas de determinada variável regionalizada. Suportes equivalentes são a base para a realização de estimativas em regiões não amostradas seja para fins de modelagem geológica por métodos geoestatísticos ou

para elaboração de estimativas de teores e tonelagens. Em contrapartida, suportes diferentes significam que as amostras possuem dispersões (desvios padrões) diferentes (Olea, 1991; PEREIRA, 2017).

Nesse contexto, são utilizadas grades ou malhas amostrais. As malhas amostrais são caracterizadas por um conjunto de pontos de amostragem, equidistantes entre si, sendo que as coletas são realizadas no entorno desses pontos. Quanto ao dimensionamento espacial da malha amostral, deverão ser considerados premissas de tal forma a reduzir as incertezas no modelo geológico, otimizar o inventário de recursos minerais, mas, sobretudo, objetivando a viabilidade econômica e o custo financeiro entre utilizar diferentes espaçamentos das malhas amostrais (GELAIN, 2016)

Nesse sentido, malhas amostrais dimensionadas com espaçamento menores entre as amostras conduzem a maior precisão, mas geram custos decorrentes de amostragens excessivas. Por outro lado, malhas mais espaçadas tendem a reduzir os custos com amostragens, com prejuízo na precisão podendo comprometer a confiabilidade do processo. Portanto, torna-se evidente a necessidade de otimizar a malha na busca pelo equilíbrio entre precisão e custo (RIVOIRARD, 2005).

Os tipos de malhas amostrais que, geralmente, são adotadas em abordagens geoestatísticas são malhas regulares, malhas irregulares ou a combinação de ambas. As malhas regulares permitem cobrir uniformemente a área disponível e por sua vez, estratificam a área de estudo favorecendo a obtenção de representatividade. Entretanto, tal configuração impõem uma distância mínima entre as amostras que poderá mascarar a correlação entre amostras à pequenas distâncias (ARIOLI; ANDRIOTTI, 2007).

Por outro lado, malhas irregulares apesar de abrangerem um grande universo de distâncias entre as amostras, não conseguem abranger de forma satisfatória a área em estudo. Como consequência, poderá privilegiar uma região em detrimento da outra e dessa forma induzir ao enviesamento dos pontos amostrados.

Segundo Yamamoto e Landin (2015) uma amostragem de determinada população pode ser classificada em três tipos gerais, descritas a seguir:

- Amostragem Aleatória Simples: as coordenadas geográficas dos pontos a serem amostrados são escolhidos de forma aleatória dentro de uma determinada região de estudo. Esse tipo de amostragem também pode ser conhecido como amostragem irregular.

- Amostragem Aleatória Estratificada: A amostragem é feita através de estratos, que são a subdivisão de uma região em células de dimensões fixas, a partir daí seleciona-se aleatoriamente uma coordenada geográfica para cada ponto amostral dentro de cada região.
- Amostragem Sistemática: As coordenadas geográficas dos pontos a serem amostrados estão definidos sobre os nós de uma malha regular. Nesse tipo de amostragem, ocorre uma disposição regular dos pontos de amostragem, o que justifica o fato desse tipo de amostragem ser conhecida também como amostragem regular.

Ao comparar os três métodos, a amostragem sistemática apresenta os melhores resultados pois permite cobrir uniformemente toda a área. Entretanto executar esse tipo de amostragem pode tornar impraticável uma vez que fatores e/ou situações externas podem impedir a coleta de amostras em determinados locais. Como alternativa, pode-se executar uma amostragem aleatória simples ou estratificada. (YAMAMOTO; LANDIM, 2015).

2.2 Agrupamento Preferencial

Como consequência, realizar amostragens aleatórias simples ou estratificadas podem ocasionar agrupamentos com maior quantidade de pontos em determinadas regiões em detrimento de outras, para estes casos, a amostragem é dita preferencial. Souza *et al* (2001) enumera três situações que podem ocasionar amostragens preferenciais em determinadas áreas:

- Condições de acessibilidade: Áreas com condições e/ou barreiras geográficas podem dificultar ou inviabilizar a coleta de amostras;
- Valores de atributos esperados: a amostragem é frequentemente adensada em áreas que são julgadas críticas (ou de interesse), por exemplo, com altos teores ou grande concentração de metais;
- Estratégia de amostragem: amostras agrupadas podem ter sido coletadas para caracterizar a variabilidade de curto alcance, para auxiliar na análise variográfica.

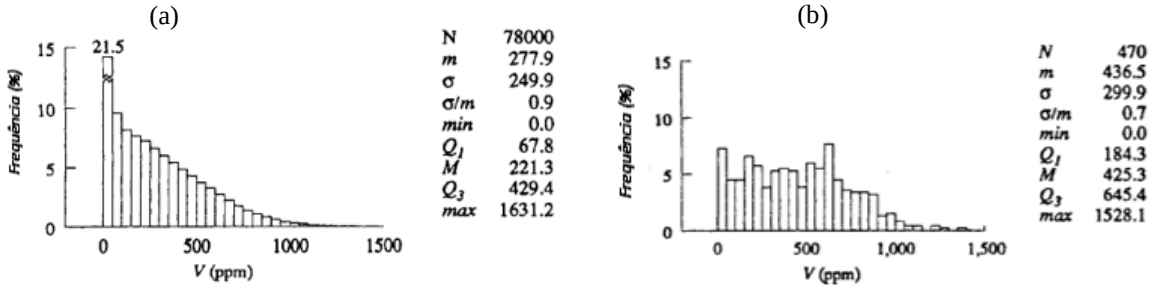
Em termos práticos, um agrupamento preferencial causa um falso enriquecimento do conjunto e influência na análise estatística do depósito em estudo. Como consequência, ocasionará uma distorção entre a estatística dos dados e os valores reais do depósito em estudo. (SOUZA, 2016).

Nesse sentido, é apresentado a seguir o conjunto de dados *Walker Lake* (Isaaks e Srivastava, 1989) como exemplo para mensurar o impacto ao assumir que medidas estatísticas descritivas de dados preferencialmente amostrados sejam representativas da distribuição real da população.

O conjunto de dados *Walker Lake* é uma base bidimensional derivada da topografia da área de mesmo nome localizada no estado de Nevada (EUA). Isaaks e Srivastava (1989), a partir de dados originais, criaram dois bancos de dados: um chamado de exaustivo composto por 78.000 dados (considerado como a distribuição real dos dados) e outro, correspondente as amostras com 470 dados preferencialmente amostrados. Três conjuntos de variáveis estão disponíveis nesses bancos de dados: *V*, *U* e *T* de forma que as duas primeiras são variáveis contínuas e a última categórica.

Nas Figuras 1(a) e 1(b) são apresentados, respectivamente, o histograma original do banco de dados *Walker Lake* (composto por 78.000 dados) e o histograma derivado desse com 470 amostras bem como a estatística descritiva respectiva a cada um. Para esses dois bancos de dados, a variável analisada, denominada *V*, representa o valor associado ao elemento cobre (Cu). Nesse sentido, ao verificar essas medidas descritivas de dados estatísticos é possível notar que o adensamento amostral do conjunto de 470 dados superestima o valor da média e subestima o valor da variância frente a população original.

Figura 1 – Histograma e estatística univariada para o banco de dados *Walker Lake* para a variável V composto por 78000 dados em (a) e 470 dados em (b).



Fonte - Isaaks e Srivastava (1989)

Tal distorção é justificada pelo fato de que amostras quando agrupadas em pequenas porções do espaço, não tem a mesma representatividade que as amostras dispostas em um espaço maior. Por consequência, a análise estatística básica se mostra simplista, ao atribuir o mesmo peso para cada amostra, e errônea, na medida em que não há uma preocupação com a distribuição espacial das amostras, se está interessado nos valores analisados de cada amostra (CORNETTI, 2003)

O exposto acima conduz a uma pergunta: Qual mecanismo poderia ser utilizado de forma a anular ou minimizar a influência que esses agrupamentos provocam sobre as estatísticas? Como resposta, aplicar de modo conveniente um conjunto de ponderadores fornecidos por diferentes métodos é uma solução para esse questionamento. (CORNETTI, 2003).

Nesse sentido, tal mecanismo é referente a técnica conhecida como desagrupamento (*declustering*). Segundo Cornetti (2003), a técnica de desagrupamento consiste em dar ponderadores p_i para amostras x_i de acordo com a irregularidade da amostragem de forma a minimizar a influência de amostras que estão próximas umas às outras. A seguir será apresentado uma breve revisão de três métodos que se propõem criar um conjunto de ponderadores: Polígonos de Influência, Vizinho mais próximo e Células Móveis.

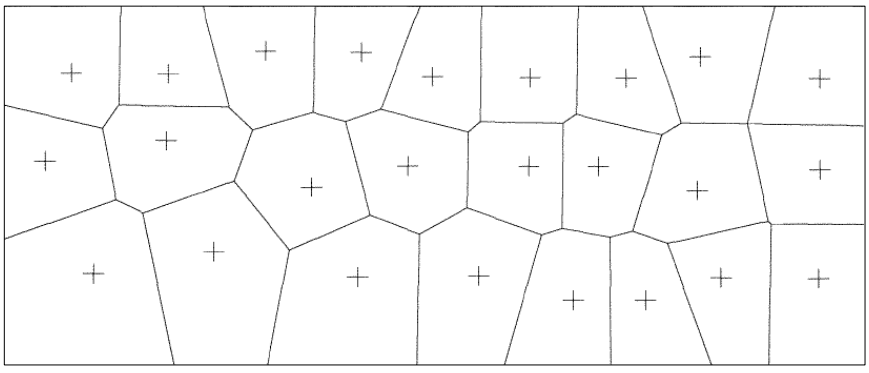
2.3 Métodos de desagrupamento

2.3.1 Polígonos de Influência

O método do polígono de influência atribui a cada amostra uma área (poligonal) ou volume (poliedro) que corresponde à sua influência na estimativa global. Essas áreas dos polígonos são utilizadas como pesos de desagrupamento.

Assim, amostras agrupadas tendem a possuir pesos menores correspondentes a polígonos de influência menores. A Figura 2 apresenta um esquema ilustrativo do referido método (CORNETTI, 2003).

Figura 2 - Exemplo de polígono de influência em dados não agrupados

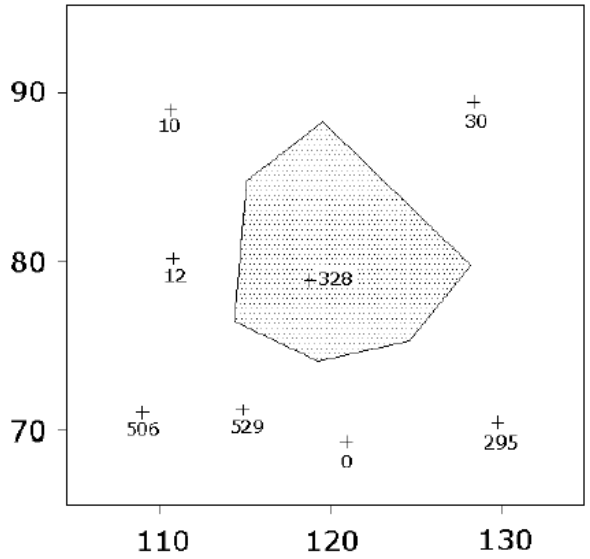


Fonte - CORNETTI (2003)

Além disso, o método se baseia na ideia de que ao redor de uma amostra existe uma população que possui relações com essa amostra e esta área está relacionada às distâncias entre o ponto em estudo e as amostras vizinhas mais próximas. A exemplo do que foi dito, na Figura

3 qualquer ponto localizado dentro da poligonal é mais próximo da amostra 328 do que qualquer amostra.

Figura 3 - Exemplo mostrando o polígono de influência de uma amostra



Fonte - Isaaks e Srivastava (1989)

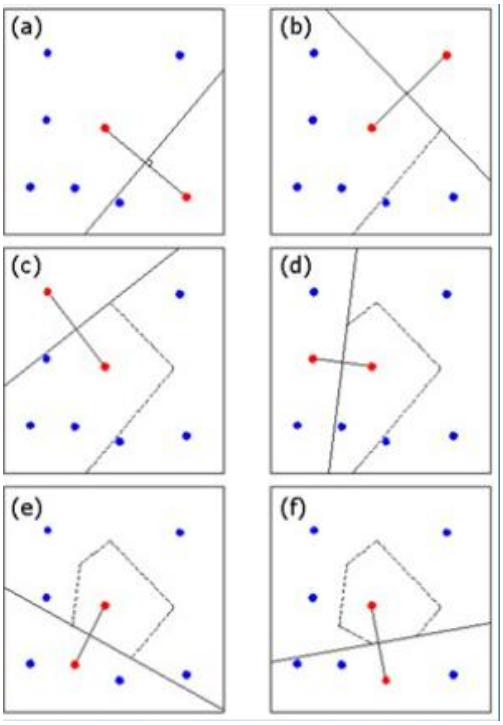
A média, pelo método dos polígonos, pode ser calculada conforme Equação 1:

$$m = \frac{1}{A} \sum_{\propto=1}^{n} \omega_{\propto} z(u_{\propto}) \quad (1)$$

onde: m é a média desagrupada; A é o somatório de todas as áreas do polígono; $w_{\propto}$ é área do polígono centrado em $u_{\propto}$; $z(u_{\propto})$ é o valor da variável resposta observado na amostra.

Os polígonos de influência são definidos pelo diagrama de Voronoi, que consiste em traçar uma bissetriz perpendicular ao segmento de reta que une as amostras próximas. Ao final do processo, a união dessas bissetrizes define a área de influência e consequentemente, formam os polígonos de cada amostra A Figura 4 exemplifica a técnica sequencial de construção para uma aplicação em 2D (ISAAKS; SRIVASTAVA, 1989).

Figura 4 Mecanismo sequencial de construção de um polígono de influência



Fonte – Adaptado de SOUZA *et al* (2001)

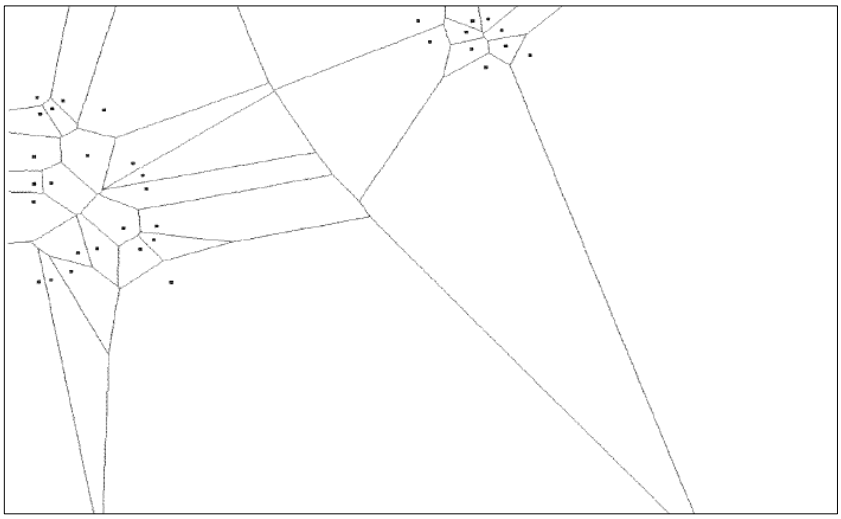
No entanto, essa técnica apresenta dois inconvenientes principais: a construção dos polígonos em torno de pontos que se encontram nas bordas da área amostrada e a presença de agrupamentos em que a distância entre os pontos pertencentes ao agrupamento e os dados que os rodeiam varia muito.

Com relação a deficiência do método, no que se diz respeito a definição do limite para casos em que as amostras se encontram nas extremidades, CORNETTI (2003) pondera que o usuário decida o valor da distância que a linha envoltória deva passar dos pontos da borda. Isaaks e Srivastava (1989) apresentam alternativas:

- Traçar um limite natural, como um contato geológico ou limite da jazida;
- Traçar um arco de influência da amostra na extremidade segundo um raio;
- Distância limite da metade da distância média entre as amostras.

Para o segundo inconveniente apresentado, para casos em que ocorre um agrupamento de dados (fato corriqueiro na amostragem em mineração), as áreas podem ficar de tamanhos diferentes para dados de um mesmo agrupamento. A Figura 5 ilustra o problema que o polígono de influência apresenta quando atua sobre dados que estão dispersos, observa-se que dados localizados em um mesmo agrupamento possuem ponderadores diferentes (CORNETTI, 2003).

Figura 5 - Exemplo de área de influência com agrupamentos distantes entre si



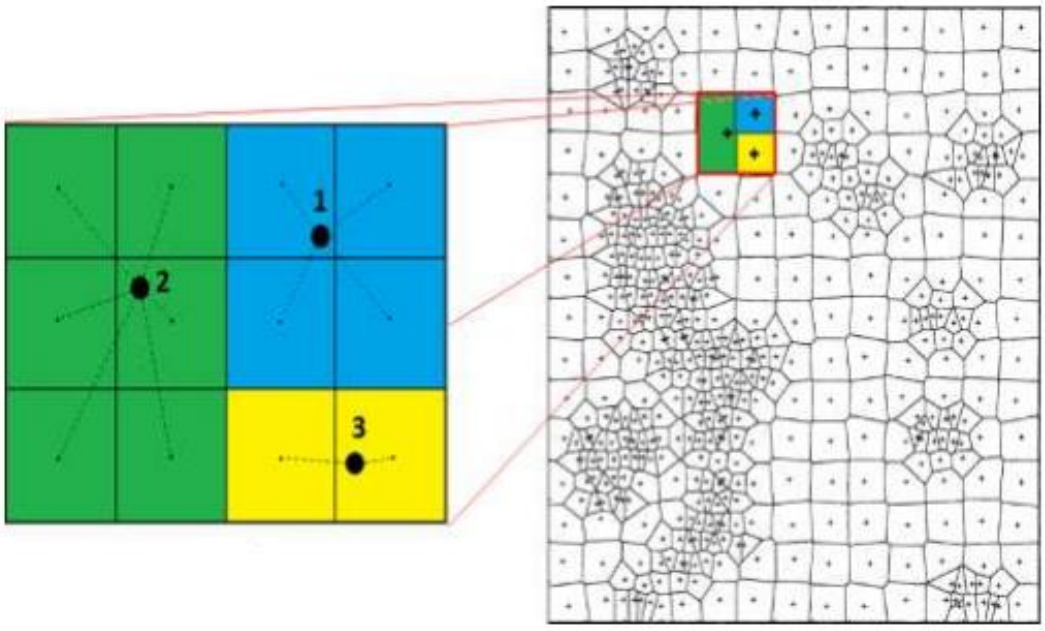
Fonte - CORNETTI, 2003

Em particular, no que se refere a definição de área de influência de uma determinada amostra, o método do vizinho mais próximo (*nearest neighbour*, NN) é uma ferramenta alternativa ao desagrupamento de dados.

De acordo com Sinclair e Blackwell (2002), o vizinho mais próximo é uma aproximação numérica do método da poligonal para a geração de *grids*, de tal forma que um valor de um nó é obtido pelo valor da amostra mais próxima a esse nó. Ramírez (2009) afirma que a metodologia parte do princípio ao assumir que a variável de estudo tem um valor constante dentro da região em torno do ponto amostrado. Assim, delimita-se essa área por retas perpendiculares traçadas a partir dos pontos médios das retas das distâncias entre as amostras. Por consequência, o valor no ponto estimado torna-se igual ao valor da amostra mais próxima a ele.

A Figura 6 ilustra o referido método onde a área de influência da amostra 1 corresponde aos quatro blocos azuis, ou seja, é a amostra mais próxima do centro dos blocos do que outras; já para a amostra 2 equivalem aos seis blocos verdes e para a amostra 3, aos dois blocos em amarelo. O processo é executado em todos os blocos dentro do domínio abrangendo assim, todo o conjunto de dados.

Figura 6 – Área de influência de amostras com destaque, a esquerda, pelo método do vizinho mais próximo, a direita, método da poligonal



Fonte: Rubio (2018)

2.3.2 Método de Células Móveis

O método de células móveis se baseia na divisão da área em estudo em regiões retangulares chamadas de células, de modo que uma ou mais amostras de um mesmo agrupamento pode(m) estar contida no interior da mesma célula. Desse modo, para cada amostra atribui-se um peso inversamente proporcional ao número total de amostras que estão contidas no interior da célula, conforme apresentado nas Equações 2 e 3 (CORNETTI, 2003; SOUZA, 2016).

$$\lambda_\alpha = \frac{1}{B \times n} \tag{2}$$

$$m = \sum_{\alpha=1}^{n} \lambda_\alpha \; z(u_\propto) \tag{3}$$

onde B é o número de células; n é o número de dados em cada célula; λ_α é o peso atribuído e $z(u_\propto)$ é o valor da variável resposta no ponto; m é a média desagrupada dos dados

De maneira geral, amostras agrupadas receberão pesos baixos, pois as células nas quais elas estão localizadas conterão diversas amostras. Ao contrário, pontos esparsos receberão pesos maiores por estarem sozinhas ou praticamente sozinhas na célula (ISAAKS; SRIVASTAVA ,1989; DEUTSCH; JOURNEL,1992).

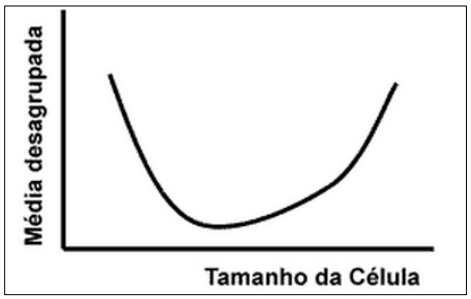
A eficiência do método dependerá da escolha adequada do tamanho da célula, uma vez que o peso do desagrupamento varia conforme o tamanho da mesma. Nesse sentido, ao escolher tamanhos de células pequenas, cada amostra cairá dentro de uma célula, $n = 1$, e, portanto, todas as amostras receberão o mesmo peso e a média desagrupada neste caso será igual a média dos dados originais. O mesmo verifica-se para casos em que a célula for demasiadamente grande, $B = 1$ em que todas as amostras localizam-se na mesma célula e, novamente recebem pesos iguais (COSTA; SOUZA)

COSTA e SOUZA destacam dois parâmetros chave do método: tamanho da célula e locação da célula no grid (origem e orientação). Por consequência, surge dois questionamentos: a) Qual a origem e orientação ideal? b) Qual o tamanho de célula ideal? No que se refere a origem, na prática testa-se várias células e utiliza-se como referência o valor médio; por outro lado, com relação a orientação, essa é a mesma do programa de amostragem.

Nesse mesmo sentido, se obtêm um tamanho ótimo de célula, testando diversos tamanhos. Conforme apresentado na Figura 7, deve-se iniciar com um tamanho pequeno de célula e ir aumentando até encontrar um valor ótimo, no qual o valor da média desagrupada volte a subir. (SOUZA *et al*, 2001)

O tamanho de célula ideal, é, portanto, aquele que produza a menor média, se as amostras estiverem concentradas nas regiões ricas. Dessa maneira, o método deve ser aplicado diversas vezes, em contraposição ao método dos polígonos de influência que é realizado uma vez, a fim de obter um tamanho ótimo de célula. (SOUZA *et al*, 2001).

Figura 7 - Variação do valor da média desagrupada com a alteração da dimensão da célula para amostragem preferencial em zonas de alto teor



Fonte – SOUZA *et al*, 2001

2.4 Aprendizado de máquina

2.4.1 Considerações iniciais

O aprendizado de máquinas é um campo de pesquisa da inteligência artificial que emprega técnicas e algoritmos computacionais para a criação de modelos computacionais. A característica principal é a capacidade de extrair conhecimentos e reconhecer padrões baseados em experiências acumuladas com exemplos e/ou problemas resolvidos (WEISS; KULIKOWSKI, 1991).

Na abordagem utilizando aprendizado de máquina, um modelo é construído através de um processo de treinamento onde a amostra na qual se deseja estudar é particionada em dois conjuntos. Um desses conjuntos, denominado “dados de treinamento” são utilizados para treinar o modelo enquanto o outro conjunto, denominado “dados de teste” servem para testar o modelo (BISHOP, 2006; AYACHE, 2021).

Nesse sentido, ao particionar as amostras, as variáveis que compõem os dados de cada conjunto assumirão duas representações. As variáveis preditoras (representadas por x_i) serão utilizadas como valores de entrada no modelo e como tal serão treinadas, testadas e responsáveis por gerar a saída desejada, denominada variável *target* que determina a resposta final (y_i) para cada amostra avaliada (BISHOP, 2006; AYACHE, 2021).

Durante o treinamento cada par (x_i, y_i) é submetido gradativamente ao algoritmo de aprendizado de máquina, de forma que parâmetros internos (ou pesos) são ajustados com o objetivo de aprender a associação entre o vetor de entrada x_i e a sua saída correspondente y_i (SARKAR, *et al*, 2017).

Depois que a máquina de aprendizado é treinada, o modelo deverá ser capaz de reconhecer dados inéditos que não foram utilizados no procedimento de ajuste. Tais dados, denominados conjunto de teste, definem a capacidade de generalização de uma hipótese, ou seja, a capacidade de classificação correta de conjunto de dados não vistos anteriormente (BISHOP, 2006).

Um algoritmo desenvolvido para uso de aprendizado de máquina pode executar tarefas de natureza preditiva ou descritivas. Enquanto a primeira busca encontrar um modelo ou uma hipótese que possa ser utilizada para prever um rótulo ou valor a partir dos demais conjuntos de dados a última busca explorar ou descrever um conjunto de dados que culminem na identificação de grupos de amostras (LANTZ, 2013).

Em especial, as tarefas de natureza preditiva podem ser classificadas em dois tipos de modelos: de regressão e os modelos de classificação. Ambos diferem na tipologia da variável

target: para o modelo preditivo, essa é composta por dados numéricos, ou seja, valores definidos fora do intervalo dos dados usados no treino do modelo, já no modelo de classificação a variável *target* é restrita aos valores predeterminados pelo conjunto de dados usados no seu treino (AYACHE, 2021).

2.4.2 Tipos de aprendizado

Nesta seção serão apresentadas as principais classificações dadas aos métodos de aprendizado de máquina quanto ao tipo de supervisão que recebem durante o treinamento: aprendizado supervisionado, não supervisionado, semi-supervisionado e aprendizado por esforço.

- Aprendizado supervisionado: neste tipo de aplicação os vetores de entrada x_i são rotulados, ou seja, se conhece previamente o valor correspondente y_i O objetivo do treinamento é criar uma aproximação da função que associa x_i a y_i baseado em um grande número de pares (x_i, y_i) (conjunto de treinamento). Após o treinamento este conhecimento é utilizado para prever o valor correspondente $\hat{y}$ de uma entrada desconhecida $\hat{x}$. De forma análoga, tem-se a figura de um "professor" ou "supervisor" na qual apresenta um problema e suas soluções e espera que o aluno consiga resolver problemas similares àquele (SARKAR, *et al*, 2017).

  O aprendizado supervisionado pode ser subdividido em duas categorias, dependendo do vetor de saídas desejadas y_i*:*

  i. classificação: nas tarefas de classificação a saída desejada pode assumir valores (reais ou inteiros) de um conjunto finito de categorias ou classes previamente estabelecidas (BISHOP, 2006).

  ii. regressão: quando a saída desejada é uma variável contínua e pode assumir qualquer valor real. Neste tipo de tarefa as variáveis de entrada ou atributos (elementos do vetor x_i) são denominadas de variáveis independentes, explicativas ou ainda preditoras, enquanto a saída que se deseja estimar é denominada de variável dependente (SARKAR, *et al* 2017).

- aprendizado não-supervisionado: neste tipo de treinamento os padrões de entrada x_i não são rotulados, ou seja, a saída desejada y_i não está disponível. Usando a analogia anterior, não há a presença de um "professor" ou

"supervisor", espera-se que o aluno deverá encontrar a solução com base no padrão de informações disponibilizadas a ele (SILVA, 2019).

- aprendizado semi supervisionado: neste cenário o algoritmo de aprendizado de máquina recebe uma grande quantidade de dados não rotulados e alguns exemplares rotulados. Segundo Sanches (2003) tal abordagem busca utilizar dados rotulados para se obter informações sobre o problema de maneira a utilizá-los para guiar o processo de aprendizado a partir dos exemplos não rotulados.
- aprendizado por reforço: de forma contrária as três abordagens anteriores nesse tipo de aprendizado, não está disponível para o algoritmo de aprendizado de máquina um conjunto de dados para treinamento. O aprendizado se dá pela interação com o ambiente que se deseja atuar por um determinado período com o objetivo de melhorar o desempenho de uma determinada tarefa. O algoritmo inicia com um conjunto de políticas e estratégias para interagir com o ambiente. Após observar o ambiente ele escolhe uma ação baseada nas políticas e estratégias previamente definidas. Neste momento o algoritmo recebe uma resposta na forma de punição ou recompensa. Assim ele pode atualizar as suas políticas e estratégias, se necessário, de forma iterativa até que ele aprenda o suficiente sobre o ambiente para receber as respostas desejadas (SARKAR *et al*, 2017).

2.4.3 Técnicas de aprendizado de máquina

Para cumprir o objetivo do aprendizado de máquinas é necessário extrair conhecimento a partir de um conjunto de amostras de dados. Para transformar os dados obtidos em informações uteis, diversas abordagens, isto é, métodos de aprendizado de máquina, podem ser aplicadas. A seguir são apresentados uma visão geral das técnicas derivadas de cada método.

1. Métodos de aprendizado de máquina baseado em estrutura de rede.

- Redes Neurais Artificiais (RNAs): Consistem em modelos inspirados no cérebro humano e constitui de várias camadas responsáveis pelo processamento da informação, de forma que cada camada é interligada e possui vários nós, uma representação análoga ao neurônio biológico. Cada "neurônio" possui um peso de acordo com a informação armazenada e a cada interação o peso do "neurônio" é recalculado até atingir a última camada neural cuja saída será a resposta do modelo (FALQUETO, 2007; AYACHE, 2021).

- Rede Bayesiana: Segundo Rezende (2005) as redes bayesianas são uma abordagem interpretativa e analítica baseado no aprendizado Bayeasino. Este tipo de aprendizado utiliza um modelo probabilístico baseado no conhecimento prévio do problema e na combinação de exemplos de treinamento para determinar a probabilidade final de uma dada hipótese. Nesse modelo os nós são os elementos principais e representam as variáveis aleatórias submetidas ao cálculo das probabilidades da rede.

2. Métodos de aprendizado de máquina baseados em análise estatística

- Regra de associação: A tarefa de associação permite estabelecer relações entre as ocorrências de um conjunto específico de itens com as ocorrências de um outro conjunto de itens. As afinidades/associações são expressas na forma de regras e expressas por combinações de itens que ocorrem com determinada frequência em uma base de dados qualquer (CARVALHO, 2001; VIERA JUNIOR, 2013).
- *Clustering*: Tal método agrupa objetos em *clusters* (agrupamentos) por meio de medidas de similaridade pré-definida. Dessa forma, agrupam-se em um mesmo *cluster* mediante a comparação de atributos similares que caracterizam os objetos. O objetivo é maximizar tanto a homogeneidade dos objetos pertencentes a um mesmo *cluster* enquanto maximiza-se a heterogeneidades entre objetos de *clusters* diferentes (SANCHES, 2003)
- *Ensemble Learning*: Métodos *ensemble* combinam diversos algoritmos de aprendizagem com a finalidade de obterem uma capacidade de desempenho preditivo superior aos de algoritmos individuais. Comumente são utilizados em aplicações de aprendizagem supervisionada e alguns casos podem ser utilizados em aprendizagem não supervisionada. Diversos estudos evidenciaram a superioridade desses métodos em diferentes aplicações para reconhecimento de padrões (ROKACH, 2010; KUNCHEVA; WHITAKER, 2003)
- Modelos ocultos de Markov (*Hidden Markov Models* - HMMs): são modelos estocásticos cuja distribuição de probabilidades geram uma observação, dependente de um estado pertencente a um processo de *Markov* não observado ou oculto. Nesse caso, o grande desafio de aplicação desse modelo é definir os parâmetros que devem permanecer observáveis e quais permanecerão ocultos (HERNÁNDEZ, 2013).
- Aprendizado indutivo: O objetivo dos algoritmos de aprendizagem indutivo é, a partir de um conjunto de instâncias de treinamento, cria um sistema que maximize a precisão alcançada nas classificações e, consequentemente, minimize o erro. Para tal são utilizados critérios para medir a qualidade do conceito induzido pelo algoritmo: precisão da

classificação, a transparência da descrição e a complexidade computacional. (NEVES, 2018; GONÇALVES, 2020).

- *Naive Bayes*: Realiza a classificação com base nas probabilidades segundo o pressuposto de que todas as variáveis são condicionalmente independentes umas das outras. Para estimar os parâmetros (médias e variâncias das variáveis) necessários para a classificação, o classificador necessita de uma pequena quantidade de dados para treinamento. Além disso, tal modelo é capaz de lidar com dados reais e discretos e tem como objetivo encontrar a classe "ótima" de forma a maximizar a probabilidade futura do evento conforme uma equação estabelecida a priori (ELSALAMONY, 2014; OLIVEIRA *et al*., 2012).

3 Métodos de aprendizado de máquina baseado em evolução.

- Computação Evolutiva: Trata-se de um ramo da Inteligência Artificial que propõe um novo paradigma para solução de problemas baseado no evolucionismo darwiniano. A computação evolutiva está baseada em quatro processos fundamentais: reprodução, variação randômica, competição e seleção de indivíduos dentro da população. Sob esse viés, os algoritmos genéticos visam tratar as possíveis soluções do problema como "indivíduos" de uma "população", que irá "evoluir" a cada iteração ou "geração" (DARWIN, 1859; PASSOS, 2014).

2.4.4 Análise de agrupamentos

2.4.4.1 Considerações iniciais

A análise de agrupamento, ou *Cluster Analysis*, é um conjunto de técnicas de estatística multivariada que tem como objetivo identificar padrões e, a partir disso, formar grupos homogêneos A construção do grupo baseia-se no critério de que amostras e/ou observações pertencentes aquele grupo se pareçam mais entre si do que com os outros grupos formados, segundo uma medida de similaridade pré-estabelecida (BUSSAB *et al* 1990).

Mingoti (2005) salienta que em muitos casos apesar do número de grupos não ser conhecido *a priori* deve-se estimá-los via dados amostrais observados. Como consequência, a interpretação sobre o número e a estrutura dos grupos será realizada posteriormente. Por outro lado, a categorização do agrupamento é feita com base nas semelhanças ou distâncias (dissimilaridades) (JOHNSON; WICHERN, 2018).

Nesse sentido, os grupos são determinados de forma a obter uma certa homogeneidade interna e uma heterogeneidade entre agrupamentos distintos. Portanto, em uma classificação

bem-sucedida, graficamente, os objetos dentro dos agrupamentos aparecerão próximos na medida em que grupos diferentes serão exibidos a uma certa distância (HAIR *et al*., 2009).

Os algoritmos de agrupamento podem ser divididos em dois grupos: algoritmos hierárquicos e algoritmos particionais. Os algoritmos hierárquicos são classificados em métodos aglomerativos ou divisivos. No primeiro, os objetos mais semelhantes são agrupados, em um primeiro momento e, depois, de forma interativa, os subgrupos mais semelhantes até chegar ao topo da hierarquia. Por último, no método divisivo, o agrupamento no topo da hierarquia é iterativamente dividido até se chegar aos objetos estudados (SOUZA JUNIOR, 2018).

Os algoritmos de particionamento ou não-hierárquicos foram desenvolvidos com a finalidade de agrupar elementos em *K* grupos, de forma que *K* é uma quantidade de grupos definida a priori. Como consequência, nem todos os valores de *K* apresentam grupos satisfatórios de modo que se aplica o método repetidas vezes para diferentes valores de K e por fim, escolhem-se os resultados que apresentem melhor interpretação dos grupos ou que exibem uma melhor representação gráfica. O método mais conhecido é o método das K-médias (K-*means*) (DONI, 2004; BUSSAB, 1990).

2.4.4.2 Algoritmo das k-médias (K-means)

O algoritmo K-*means* também conhecido como K-médias é uma das técnicas de agrupamento particional mais utilizadas devido ao fato de possuir o maior número de variações e ter uma abordagem simplificada e fácil usabilidade em implementar em linguagens computacionais (MENDES, 2017).

A ideia do algoritmo K-*means* é fornecer uma classificação baseada em análises e comparações entre os valores dos dados. Dessa forma, o algoritmo de forma arbitrária fornece uma classificação automática sem a necessidade de intervenção e supervisão humana, não necessitando de uma nenhuma pré-classificação (MENDES 2017).

Segundo XAVIER (2012), o algoritmo de K-*means* pode ser descrito da seguinte forma:

1. Partição inicial dos objetos em *K* grupos definidos à priori;
2. Cálculo dos centróides para cada um dos *K* grupos e cálculo da distância euclidiana dos centroídes a cada indíviduo na base de dados;
3. Agrupar os indivíduos aos grupos cujos centroídes se encontram mais próximos e voltar ao passo anterior até que não ocorra uma variação significativa na

distância mínima de cada indivíduo da base de dados a cada um dos centroídes dos *K* grupos.

Em outras palavras, PRASS (2004) enfatiza que uma vez escolhidas as sementes iniciais, calcula-se a distância de cada elemento em relação às sementes, e o agrupamento do elemento se dá em função daquele que possuir a menor distância (mais similar) recalculado o centroíde. O processo é repetido iterativamente até que todos os elementos pertençam ao mesmo *cluster*.

Entretanto, o algoritmo de K-*means* apresenta alguns inconvenientes tais como:

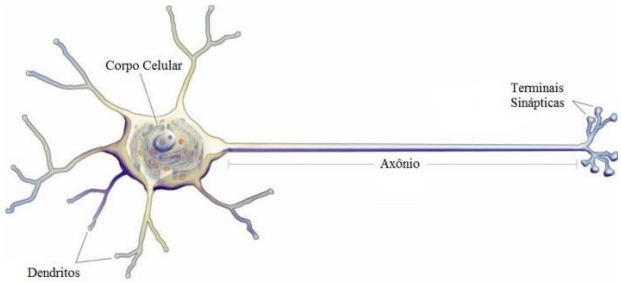
- Exige que as variáveis sejam numéricas ou binárias, para tal, frequentemente dados categorizados devem ser convertidos em valores numéricos (HUANG, 1997);
- É sensível a presença de *outliers*, ou seja, valores extremos podem modificar a distribuição dos dados (HAN; KAMBER, 2001).

2.4.5 Redes Neurais Artificiais

O modelo das redes neurais artificiais, conhecidas também como “redes neurais” foram desenvolvidas de forma análoga a estrutura física do cérebro humano. De modo similar a um computador, o cérebro é uma estrutura biológica altamente complexa que processa informações.

Nesse sentido, os neurônios, constituintes estruturais do cérebro, tem a capacidade de reconhecer padrões com um tempo de percepção e controle maior que os computadores mais velozes existentes atualmente. A Figura 8 apresenta os principais constituintes de um neurônio biológico.

Figura 8: Representação de um neurônio biológico



Fonte: LOHMANN, 2016

Pode-se definir uma rede neural artificial como um processador constituído de unidades paralelas de processamento simples com a capacidade intrínseca para armazenar conhecimento

empírico e disponibilizá-lo para uso. Pode-se correlacionar a rede neural artificial e o cérebro em dois aspectos:

- A rede adquire conhecimento por intermédio do seu ambiente através de um processo de aprendizagem;
- O conhecimento adquirido é armazenado pelas forças de conexão entre os neurônios denominadas pesos sinápticos (CINTRA, 2003).

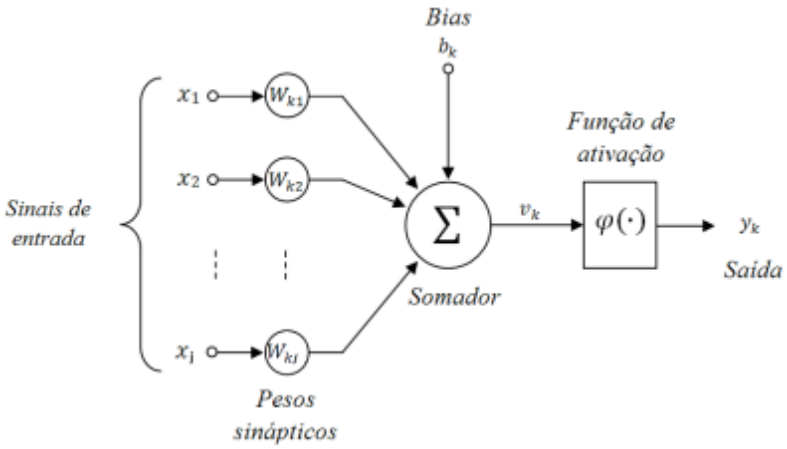
Dessa maneira, a rede neural, enquanto sistema de inteligência artificial, é capaz de adquirir, armazenar e aplicar o conhecimento para resolver problemas além de obter novos conhecimentos a partir de experiências (CINTRA, 2003).

Nesse contexto, sob a ótica computacional, pode-se dizer que em um neurônio pode ser considerado uma unidade fundamental processadora de informação. Nesse sentido, podemos correlacionar (a grosso modo) o dendrito a entrada, soma e ao processamento e o axônio, a saída.

Assim, HAYKIN (2001) propôs um modelo de neurônio artificial, como pode ser observado na Figura 9 e destaca três elementos básicos:

i. Um conjunto de ligações chamadas sinapses ou elos de conexão, cada uma caracterizada por um peso w_{kj}, onde o índice k corresponde ao número do neurônio e j ao sinal de entrada. O peso sináptico de um neurônio artificial pode estar em um intervalo que inclui valores negativos e positivos. Quanto maior o peso sináptico, maior será a contribuição da respectiva entrada para o somador;
ii. Um somador para adicionar os sinais de entrada, ponderados pelos respectivos pesos sinápticos configurando um combinador linear, cujo resultado é o valor u_k;
iii. Uma função de ativação para restringir a amplitude da saída de um neurônio, também referida como função restritiva já que limita o intervalo permissível de saída a valores normalizados entre 0 e 1 ou -1 e 1.

Figura 9 – Representação de um neurônio artificial.



Fonte –HAYKIN, 2001

Em termos matemáticos, um neurônio k pode ser descrito por meio da Equação 4 (HAYKIN, 2001),

$$u_k = \sum_{j=1}^{m} w_{kj}\, x_j \tag{4}$$

onde x_1, x_2, …, x_m são os sinais de entrada; w_{k1}, w_{k2},…, w_{km} são os pesos sinápticos do neurônio k; u_k é a saída do combinador linear devido aos sinais de entrada.

2.4.6 Redes SOM ou Redes de Kohonen

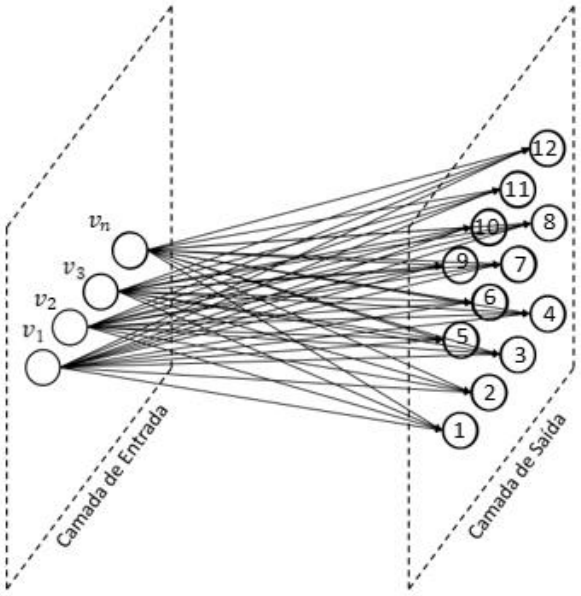
As redes SOM (*Self-Organizing Maps*, Mapas Auto Organizáveis) foram propostas por *Tuevo Kohonen* em 1982. Kohonen procurou estabelecer uma relação baseada na capacidade do cérebro humano em armazenar informações e seu modelo. Essa associação parte do princípio de que, tal como o cérebro humano é capaz de reconhecer padrões, o seu modelo poderia ser utilizado para auto-organizar padrões similares (TODT, 1998).

Os mapas auto-organizáveis podem ser definidos como sendo uma rede neural de aprendizado não supervisionado, onde a rede busca agrupar os dados de entrada baseado em suas similaridades formando classes ou agrupamentos denominados *clusters*. Além disso, pode ser considerada uma rede competitiva na medida em que os neurônios competem entre si até que sejam definidos os neurônios mais representativos de cada classe (SILVA,1998; TODT,1998).

A arquitetura do SOM consiste em duas camadas: entrada e saída, representada por uma grade pós-sináptica. Na Figura 10 pode-se observar que a camada de saída é formada por uma rede de neurônios interligados aos mais próximos, salienta-se que cada neurônio representa um

cluster. Já os neurônios que constituem a camada de entrada são conectados a todos os neurônios da grade pós-sináptica (BELUCO, 2013).

Figura 10 – Representação bidimensional da Rede de Kohonen com entrada v_n

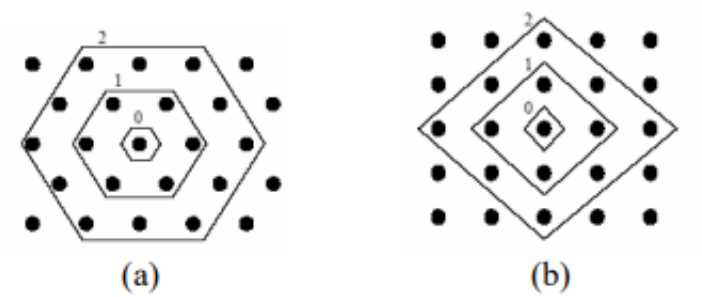


Fonte - Autor

Seja o conjunto de entrada contido no espaço $\Re^D$, $V = \{v_1, \dots, v_n\}, V \subseteq \Re^D$, de vetores $v_n = [v_{n1}, \dots, v_{nd}]^T \in \Re^D, n = 1, \dots, N$, onde cada vetor v_1 representa um dado no espaço D- dimensional, por meio de D atributos. Pode-se definir o SOM como um conjunto de neurônios $i, i = 1, \dots Q$, dispostos em um arranjo que define a vizinhança de cada neurônio (CAPPONI, 2019).

Considera-se que um neurônio é vizinho de outro no arranjo de acordo com a configuração adotada. Capponi (2019), enfatiza que o formato do arranjo tem influência direta na adaptação do SOM. Além disso, o autor afirma que o modelo hexagonal oferece resultados mais satisfatórios que o retangular. A respeito dessa vizinhança, existem diversas configurações de arranjo, como pode ser visto na Figura 11, (a) vizinhança retangular e em (b) um arranjo com vizinhança hexagonal.

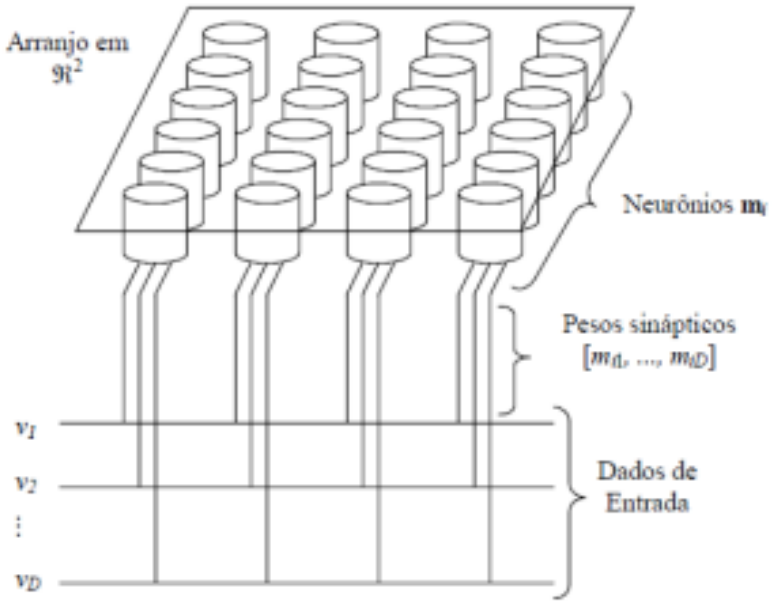
Figura 11 – Exemplos de configurações de arranjo para o SOM em (a) estrutura hexagonal e (b) estrutura retangular



Fonte - Vesanto (2000)

Cada neurônio i é representado por um vetor de pesos sinápticos $m_i = [m_{i1}, \dots, m_{iD}]^T \in \Re^D$ e, todos os neurônios são conectados ao sinal de entrada ou dados recebido. A Figura 12 apresenta os elementos básicos de uma mapa auto-organizável de Kohonen (ZUCHINI, 2003).

Figura 12 – Representação esquemática onde todos os neurônios estão conectados ao sinal de entrada ou dado recebido



Fonte -Adaptado de Zuchini (2003)

O treinamento das redes SOM consiste em submeter de maneira iterativa o conjunto de dados de entrada, ou seja, os vetores de entrada a um modelo de aprendizagem competitiva. Dessa maneira, de forma aleatória e repetitiva, um conjunto de dados representado por vetores no espaço $\Re^D$ é apresentado a uma rede composta por neurônios organizados segundo um arranjo específico, cada neurônio possui um correspondente vetor de pesos no $\Re^D$.

Como o conhecimento baseia-se nos pesos da rede, de forma análoga equivale aos pesos sinápticos, o objetivo é direcionar os pesos às entradas. Dessa forma, o neurônio vencedor é aquele cujo vetor de pesos é o mais próximo do vetor de entrada. Em outras palavras, distância menor significa maior semelhança. Consequentemente, este neurônio é o que produz a maior saída. Em especial, para esse neurônio é denominado BMU (*Best Matching Unit*) (Van Hulle, 2000; CAPPONI, 2019).

A determinação do neurônio mais similar (BMU) pode ser feita, por exemplo, por meio da distância euclidiana, conforme Equação 5. Seja $v_n \in V$ um dado de entrada tomado aleatoriamente ($n \in \{1, \dots, N\}$) e apresentado à rede. Uma vez que os neurônios do arranjo recebem a mesma entrada v_n , calcula-se a distância do vetor de pesos m_i de cada neurônio i ao vetor v_n ,

$$\mathrm{d}\,(m_i, v_n) = \|m_i - v_n\| = \sqrt{\sum_{j=1}^{D} \left|m_{ij} - v_{nj}\right|^2} \tag{5}$$

Uma vez calculado todas as distâncias, elege-se um neurônio BMU de índice c na forma da Equação 6:

$$c = \arg min_i \left\{\left\|m_{ij} - v_{nj}\right\|\right\} \tag{6}$$

Por fim, estabelecido o neurônio vencedor, os vizinhos a ele são adaptados. O próximo passo é então, a atualização de cada pesos de todos os neurônios na vizinhança física do neurônio vencedor. Dessa maneira, o aprendizado, isto é, o novo valor de peso sináptico do i-ésimo neurônio é dado no instante de tempo $(t + 1)$. A Equação 7 apresenta a equação de adaptação dada por:

$$m_i\,(t + 1) = m_i(t) + \propto (t) \times h_{ci}(t) \times [m_i(t) - v_n(t)\,] \tag{7}$$

Onde $t = 0,1,2$ ... é um número inteiro que representa a coordenada discreta de tempo e $\propto (t)$ representa a taxa de aprendizado. Pelo exposto na Equação 8 percebe-se que o grau de adaptação do neurônio BMU depende da função de vizinhança, h_{ci} e da taxa de aprendizado $\propto$. Para ocorrer a convergência do mapa, é necessário que $h_{ci}\,(t) \to 0$ quando $t \to \alpha$ ou seja, a função será ajustada reduzindo o grau de vizinhança relativo ao neurônio BMU (ZUCHINI, 2003, CAPPONI, 2019).

De forma resumida, o algoritmo para a determinação dos agrupamentos por meio das redes SOM, pode ser descrito como:

1. Determinação da estrutura da rede;
2. Inicialização de todos os pesos iniciais de forma aleatória;
3. Determinação de um ponto de dado aleatório a partir dos dados de treinamento;
4. Determinação do neurônio mais similar (BMU) ao ponto escolhido no passo 3 no mapa, sendo este o neurônio vencedor.
5. Ajuste dos pesos para os demais neurônios de acordo com os pesos do neurônio vencedor;
6. Os passos 2 a 5 são repetidos para t épocas, ou até se atinja um ponto de convergência predeterminado.

Apenas informações acerca das coordenadas no espaço de saída não são suficientes para visualização de mapas uma vez que informação de distância (dissimilaridade) entre os pesos dos neurônios são imperceptíveis. Para solucionar essa inconformidade, um método de visualização de um SOM treinado, denominado matriz de distancias unificadas, ou matriz U, foi proposta por A. Ultsch com o objetivo de permitir a detecção visual das relações topológicas, isto é, a distância, dos neurônios (MEDEIROS; COSTA, 2008).

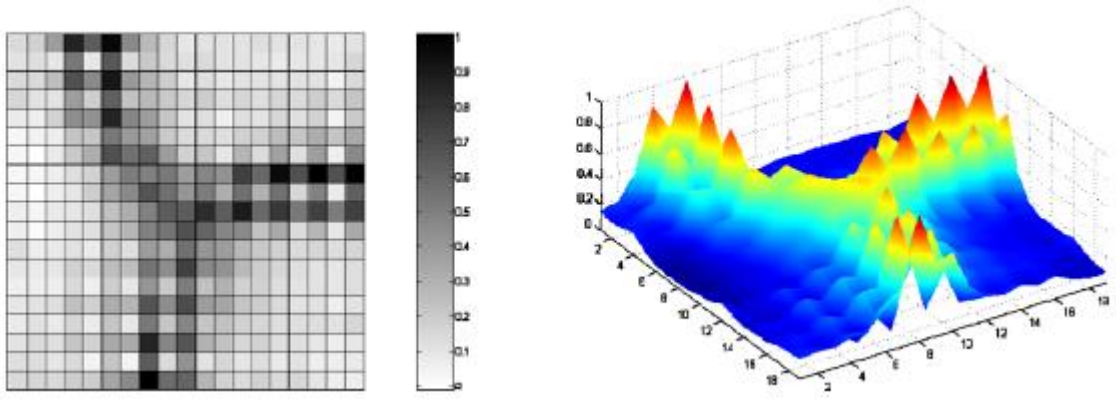
A ideia básica é calcular a distância entre os pesos sinápticos de neurônios adjacentes, tal qual é calculada durante o treinamento. O resultado é uma imagem $f(x, y)$onde as coordenadas (x, y) de cada pixel são derivadas das coordenadas dos neurônios no grid do mapa e a intensidade de cada pixel $f(x, y)$ na imagem corresponde à distância calculada. Uma imagem pode ser pensada como uma função tridimensional, onde um valor de pixel nas coordenadas (x, y) é representado por um ponto na coordenada z (MEDEIROS; COSTA, 2008).

Assim, em uma matriz-U tridimensional, o resultado é uma representação da superfície topográfica em 3D de forma que a topografia revela a configuração dos neurônios obtido pelo treinamento: vales correspondem a regiões onde neurônios são similares, isto é, em áreas de alta densidade de dados; cadeias montanhas, por outro lado, refletem dissimilaridades entre neurônios vizinhos e estão associados a valores altos visto que são regiões onde existe pouco ou nenhum dado. Em outras palavras, a cadeias de montanhas são as fronteiras dos *clusters* (ULTSCH, 2004).

A Figura 13 apresenta um exemplo de uma projeção de uma matriz-U tridimensional, obtida de um determinado mapa treinado, à direita, pode-se observar a interpretação

tridimensional correspondente a imagem a esquerda, onde é possível notar regiões de alta densidade (vales) e os picos representando as regiões de fronteiras. A barra de intensidade cinza indica o nível de proximidade entre os neurônios: quanto mais escuro, maior a distância.

Figura 13 – Esquema de uma projeção bridimensioanal de uma matriz U (esquerda) e sua respectiva interpretação espacial (direita)



Fonte – Medeiros e Costa (2018)

3. METODOLOGIA

Para o desenvolvimento desta pesquisa, algumas etapas foram feitas com o objetivo de adequar os dados originais ao modelo proposto e paralelamente, esses foram tratados utilizando as técnicas convencionais de desagrupamento para serem utilizadas como dados de validação. Estes dados serão a base para determinar a eficiência do modelo proposto em comparação com os resultados tradicionalmente utilizados.

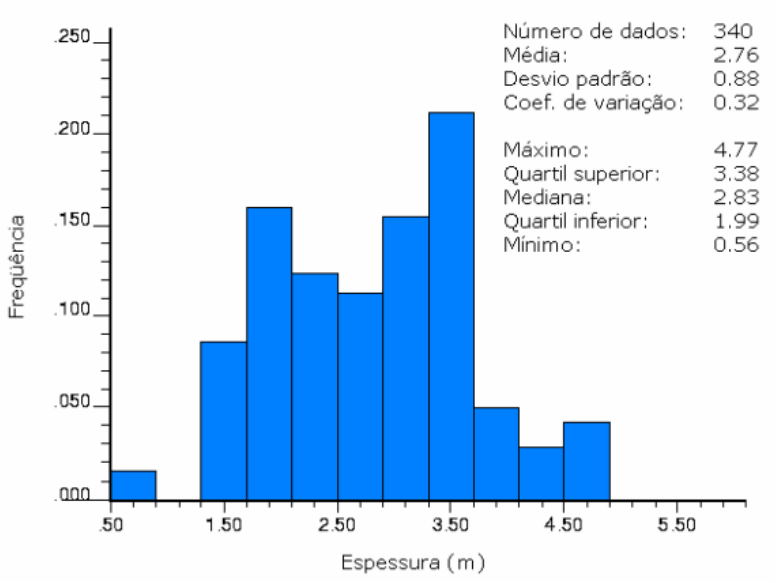
Para dar início nesta etapa, um apanhado geral da metodologia será discutido no tópico a seguir deste trabalho

3.1 Metodologia Geral

Primeiramente, foi feita a seleção dos bancos de dados para testar o modelo proposto. Por causa do critério didático e comparativo dessa pesquisa, foram escolhidos os dados de *Walker Lake* (Isaaks e Srivastava,1989) e do Carvão (Souza, 2002). A justificativa para o primeiro é em relação a sua utilização em diversos estudos e pesquisas na área de geoestatística, facilitando a comparação com os resultados de outros autores. Já para o segundo, o principal critério está associado à heterogeneidade na distribuição espacial dos dados. Este aspecto foi fundamental para forçar as capacidades de interpretação do modelo.

O banco de dados *Walker Lake*, conforme descrito no tópico 2.2, possui 470 locais amostrados e foi utilizado a variável cobre como atributo de interesse. Por outro lado, o banco de dados do carvão consiste em 340 furos de sondagem com dados de espessura de camada de carvão. Segundo SOUZA (2002), o depósito de carvão está inserido na bacia Carbonífera Sul-catarinense, que se localiza no flanco sudeste do Estado, estendendo desde o sul de Araranguá além de Lauro Muller, com aproximadamente 100 km de comprimento e uma largura média de 20 km. A Figura 14 apresenta o histograma para a variável espessura de carvão.

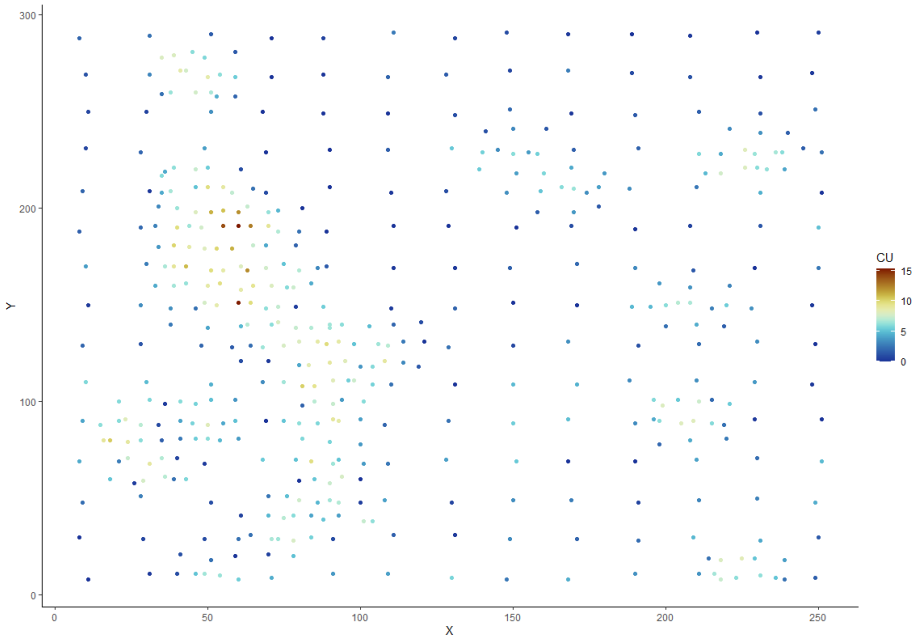
Figura 14 – Histograma dos dados desagrupados para a variável espessura de carvão



Fonte – SOUZA (2002)

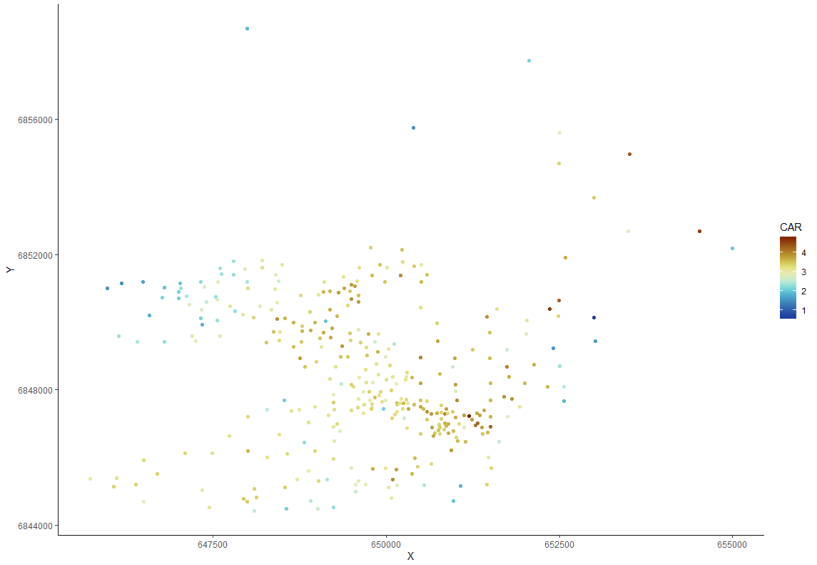
As Figuras 15 e 16 apresentam a malha de dados do *Walker Lake* e do Carvão respectivamente.

Figura 15 – Mapa de localização das amostras disponíveis no *Walker Lake*



Fonte – R CORE TEAM, (2016).

Figura 16 - Mapa de localização das amostras disponíveis de espessura do carvão



Fonte - R CORE TEAM, (2016).

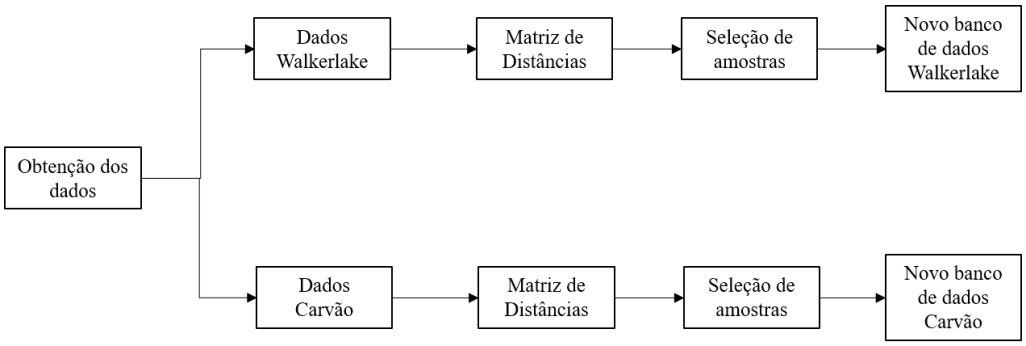
Estes dados foram trabalhados para adequar as suas informações com o funcionamento do modelo desenvolvido. Primeiramente, as amostras foram submetidas a uma matriz de distâncias utilizando as variáveis espaciais *x* e *y*. Para tal, foi utilizado a distância de Manhattan cuja métrica é tal que a distância entre dois pontos é a soma das diferenças absolutas de suas coordenadas $\{(x_{i1},\ x_{j1}), (x_{i2},\ x_{j2}), \dots, (x_{in},\ x_{jn})\}$ conforme Equação 8.

$$d(i,j) = \sqrt{(|x_{i1}\ - x_{j1}| + |x_{i2}\ - x_{j2}| + \cdots + |x_{in}\ - x_{jn}|)} \tag{8}$$

De acordo com os dados da matriz de distância foi feito uma seleção das 5 menores distâncias obtidas para cada amostra do banco de dados. Quanto a seleção desse número de distâncias, justifica-se que a partir de 5 seleções, o ganho de informação era pequeno, daí optou-se por selecionar tal quantidade para reduzir o processamento de dados.

Uma vez selecionada, foi possível gerar um novo banco de dados com 5 variáveis que representassem, em ordem crescente, as 5 menores distâncias para cada amostra. É importante salientar que a distância da amostra com ela mesmo, obtida pela matriz de distância, foi ignorada pois essa continha valores de distância igual a 0. Essas novas variáveis foram utilizadas nas Redes de Kohonen (Kohonen, *et al.*, 1995). A Figura 17 apresenta a sequência de etapas desenvolvidas que serão apresentadas nesse tópico.

Figura 17 - Tratamento inicial dos dados



A seguir são apresentados as Tabelas 1 e 2 que contém os cabeçalhos dos bancos de dados normalizados extraídos do pacote estatístico R (R CORE TEAM, 2016) após o tratamento de dados de acordo com a metodologia apresentada no Tópico 3.2.

Tabela 1 – Valores associados as variáveis relacionadas ao banco de dado *Walker Lake*

	SOMAA	**SOMBB**	**SOMCC**	**SOMDD**	**SOMEE**
1	0.9543716	0.9546646	0.9087716	0.6711022	0.7248217
2	0.8393683	0.8941034	0.6798847	0.6664143	0.6405884
3	0.6808057	0.7403124	0.5553416	0.385402	0.3813862
4	0.5760663	0.5055157	0.318949	0.2382755	0.2372276
5	0.2264755	0.4308991	0.26042	0.1954903	0.1757644
6	0.6737917	0.791569	0.5874241	0.3945471	0.4266184

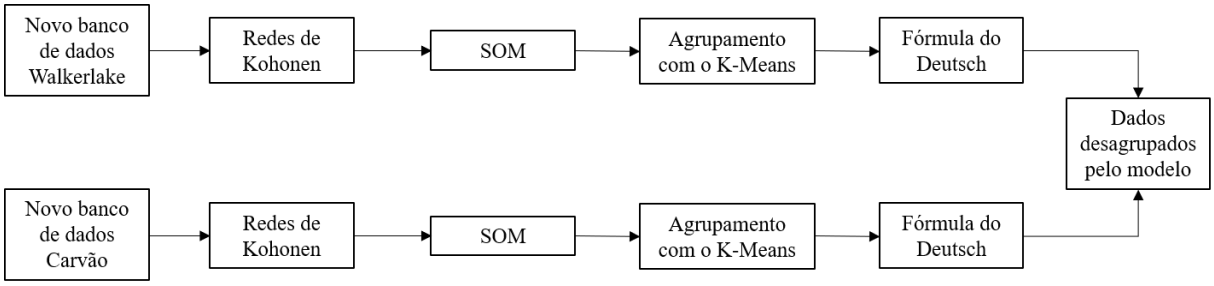
Tabela 2: Valores associados as variáveis relacionadas ao banco de dado do Carvão

	SOMAA	**SOMBB**	**SOMCC**	**SOMDD**	**SOMEE**
1	0.234732	0.279195	0.355770	0.340777	0.315663
2	0.552165	0.557000	0.470585	0.526665	0.489605
3	0.234732	0.247030	0.196322	0.361343	0.341726
4	0.286748	0.260975	0.244700	0.291677	0.262468
5	0.110150	0.142316	0.114073	0.136095	0.126772
6	0.010131	0.031455	0.022042	0.014651	0.014974

Após gerar o Mapa SOM, os neurônios passaram pelo K-*means* (Macqueen, 1967) para serem agrupados em conjuntos maiores. A partir daí, esses grupos foram submetidos a fórmula desenvolvida por Deutsch (1989) para obtenção dos pesos para o desagrupamento.

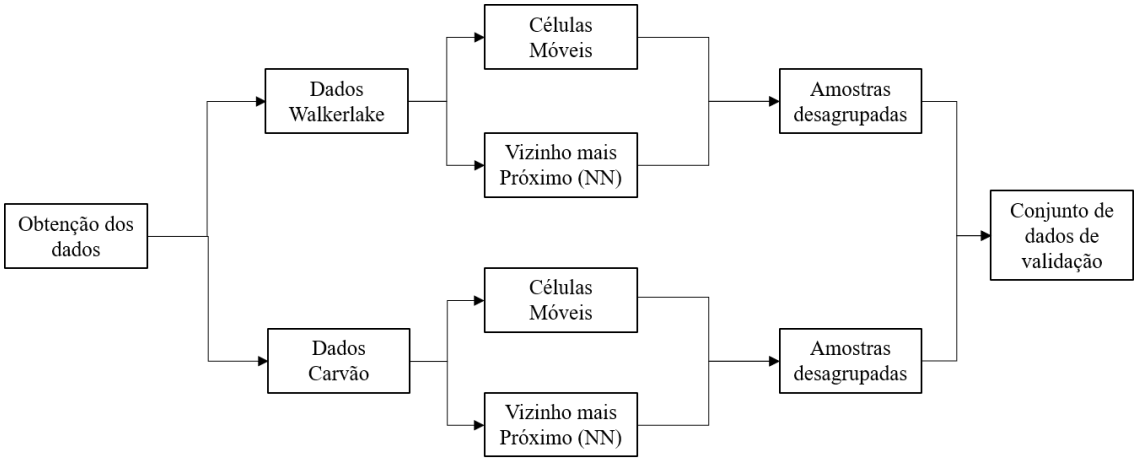
Essa descrição feita é a base para o funcionamento do modelo proposto nesta pesquisa. Na Figura 18 é possível ver o diagrama que representa essas primeiras etapas do desenvolvimento.

Figura 18 - Diagrama do Modelo



Por fim, foi necessário validar os dados desagrupados pelo modelo. Para esta determinação, esses resultados foram comparados com os resultados obtidos pelos métodos tradicionais de desagrupamento: Células Móveis (Deutsch, 1989) e vizinho mais próximo (Cover e Hart, 1967) De acordo com esta comparação foi possível extrair as métricas para determinar a qualidade da estimativa, conforme Figura 19.

Figura 19 - Diagrama dos dados de validação



Todo o desenvolvimento deste trabalho foi realizado utilizando a linguagem de programação *freeware* R (R CORE TEAM, 2016), uma linguagem de programação estruturada que tem como objetivo a manipulação, análise e visualização de dados.

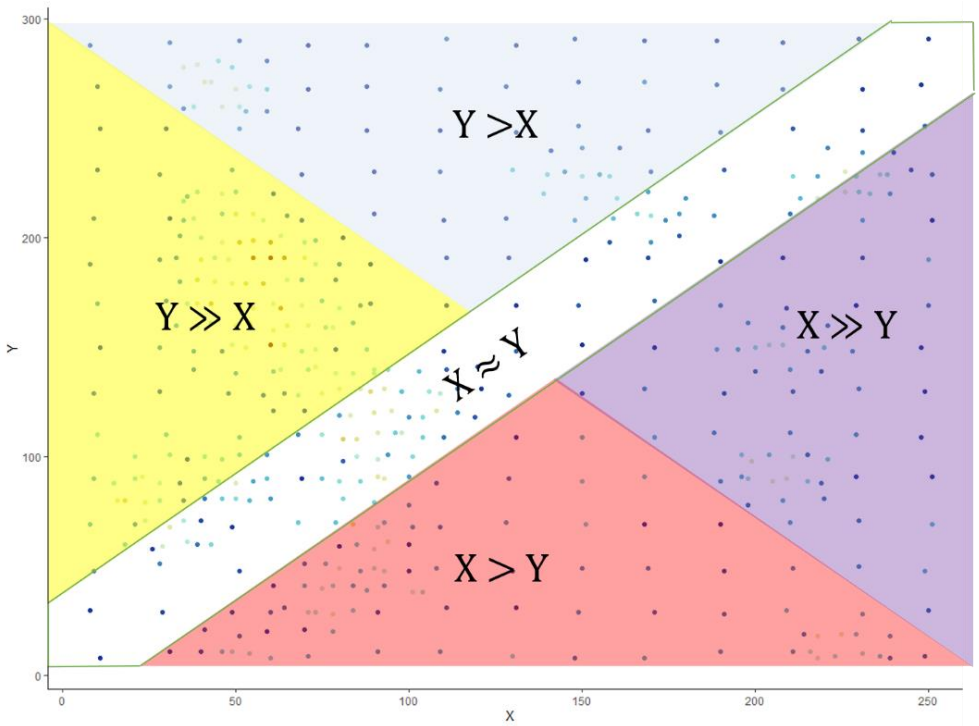
Todas estas etapas serão abordadas de forma mais profunda nos demais tópicos deste trabalho. Seguindo a ordem apresentada nesse tópico, a primeira etapa será o tratamento dos dados.

3.2 Tratamento dos dados

Ao ler o Tópico 3.1 uma pergunta pode ter surgido “Porque não utilizar as variáveis X e Y no modelo ao invés de utilizar a matriz de distâncias para criar um novo banco de dados?” Realmente, seria mais fácil utilizar as variáveis originais pois reduziria o pré-processamento das amostras. No entanto, há uma característica muito importante no algoritmo das Redes de Kohonen que dificulta a obtenção de bons resultados ao utilizar as coordenadas como variáveis de entrada.

Qualquer técnica de inteligência artificial busca encontrar padrões dentro das variáveis de entrada para gerar uma resposta de saída coerente. No entanto, quando se trata de dados espaciais, esses padrões podem causar um viés “espacial” que não condiz com a realidade dos dados. Este efeito pode ser visto na Figura 20 na qual representa a tendência que esses padrões podem causar em dados espaciais, formando 5 grupos distintos.

Figura 20 - Representação dos padrões das variáveis nos dados do *Walker Lake*.



Mesmo possuindo uma coerência em função das variáveis X e Y, esse não consegue representar bem o agrupamento de algumas amostras, misturando em um mesmo grupo dados agrupados e não agrupados, o principal foco desta pesquisa.

Ao utilizar a matriz de distâncias, é possível “traduzir” os dados espaciais em variáveis escalares, facilitando uma interpretação do modelo e gerando dados de saída mais coerentes.

Para o desenvolvimento da matriz de distâncias, o pacote *Rgeos* (*Interface to Geometry Engine - Open Source*) foi usado pois permite criar diversas aplicações e técnicas para dados espaciais. O pacote *Rgeos* foi desenvolvido por Bivand e Rundel (2021).

Para o caso dos dados do *Walker Lake*, uma matriz 470x470 foi criada e a partir das amostras geradas para cada amostra original foram retiradas as 5 menores distâncias. O mesmo procedimento foi feito para os dados do Carvão, no entanto, por ter menos amostras, a matriz criada tinhas as dimensões de 340x340.

Uma informação importante a ser dita é que as matrizes de distâncias sempre geram matrizes quadráticas, ou seja, possuem o mesmo número de linhas para o mesmo número de colunas sendo que, este número é igual ao número de amostras no banco de dados. Outra informação importante a ser comentada é que, essencialmente, a seleção das 5 menores distâncias foi feita a partir da segunda menor distância, isso é feito pois em todas as amostras, a menor distância sempre será 0, pois todos os dados são comparados com as suas próprias coordenadas. Sendo assim, não haveria ganho nenhum em adicionar a primeira menor distância como variável, sendo essa eliminada da seleção.

A segunda etapa feita no tratamento foi a normalização das variáveis. Este processo envolve a padronização das escalas das variáveis em uma mesma métrica. Isso facilita o entendimento e interpretação dos algoritmos, melhorando os resultados do modelamento.

Para isso a Equação 9 foi usada para transformar cada amostra de cada variável em uma escala de 0 a 1, sendo que 0 é a menor distância do conjunto e 1 é a maior. Sendo que x_i° representa a amostra normalizada, x_i é a amostra original e V_n a variável analisada.

$$x_i^{\circ} = \left\{ \left(x_i - \min_n(V_n) \right) \Big/ \left(\max_n(V_n) - \min_n(V_n) \right) \right\} \tag{9}$$

Após esta etapa, os dois novos conjuntos de dados foram criados e para analisar os resultados obtidos, duas matrizes de correlação foram desenvolvidas. Estes resultados serão apresentados mais à frente. Após esta etapa completa, foi possível avançar para a implementação do modelo.

3.3 Implementação do modelo

Para esta etapa, a metodologia foi dividida em 3 passos diferentes para facilitar o entendimento. A primeira parte abordará a implementação das Redes de Kohonen para criação

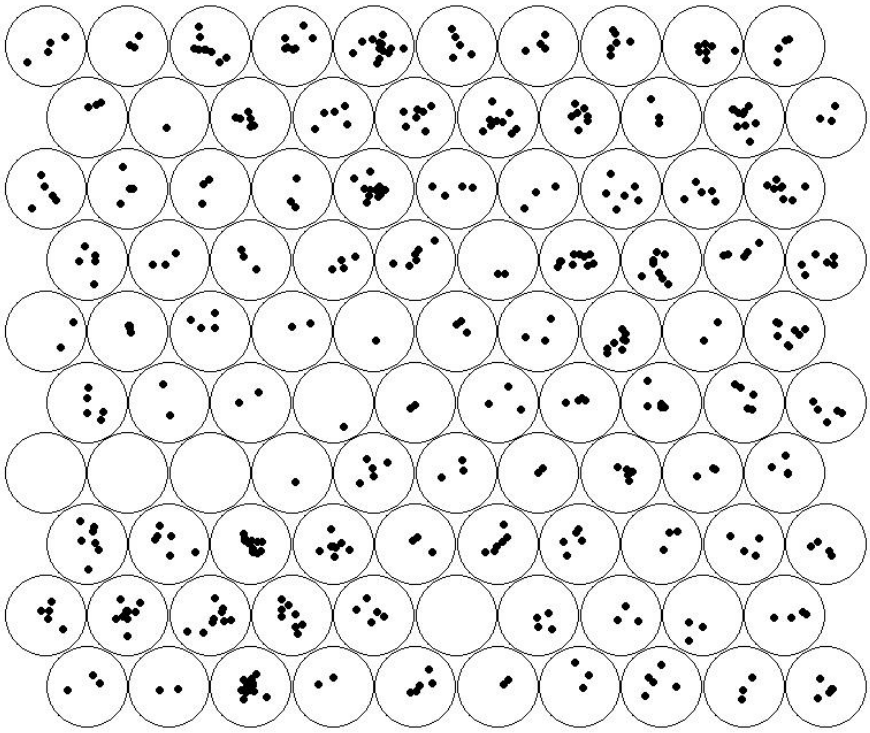
do Mapa SOM. Na segunda parte, a utilização do K-*means* será descrita e por fim, na terceira etapa será demonstrada a visualização dos resultados obtidos.

3.3.1 Rede de Kohonen

Para implementar esta etapa, o pacote Kohonen foi utilizado, desenvolvido por Wehrens e Kruisselbrink (2019), este possui a capacidade implementar várias formas de Mapas Auto-organizáveis (SOMs). A base utilizada para desenvolvimento vem dos estudos desenvolvidos por Kohonen *et al*., (1995).

O grid criado é do tamanho de 10x10 com neurônios de topologia circular. Essa escolha foi feita partir da premissa de otimizar o tempo de processamento com o aumento da segregação das amostras em cada neurônio. Com o grid dessas dimensões, 100 neurônios foram criados como pode ser visto na Figura 21.

Figura 21 - Grid de neurônios



Fonte - R CORE TEAM, (2016).

Com o grid criado, foi possível gerar o SOM. Foram realizadas 1000 interações no processo de criação, além de utilizar o método de Manhattan como fator das distâncias entre as camadas dos dados.

Esses processos foram feitos para os dois bancos de dados transformados e após esta etapa, os resultados obtidos para cada neurônio foram utilizados como dados de entrada para o K-*means*.

3.3.2 Agrupamento dos neurônios das redes SOM

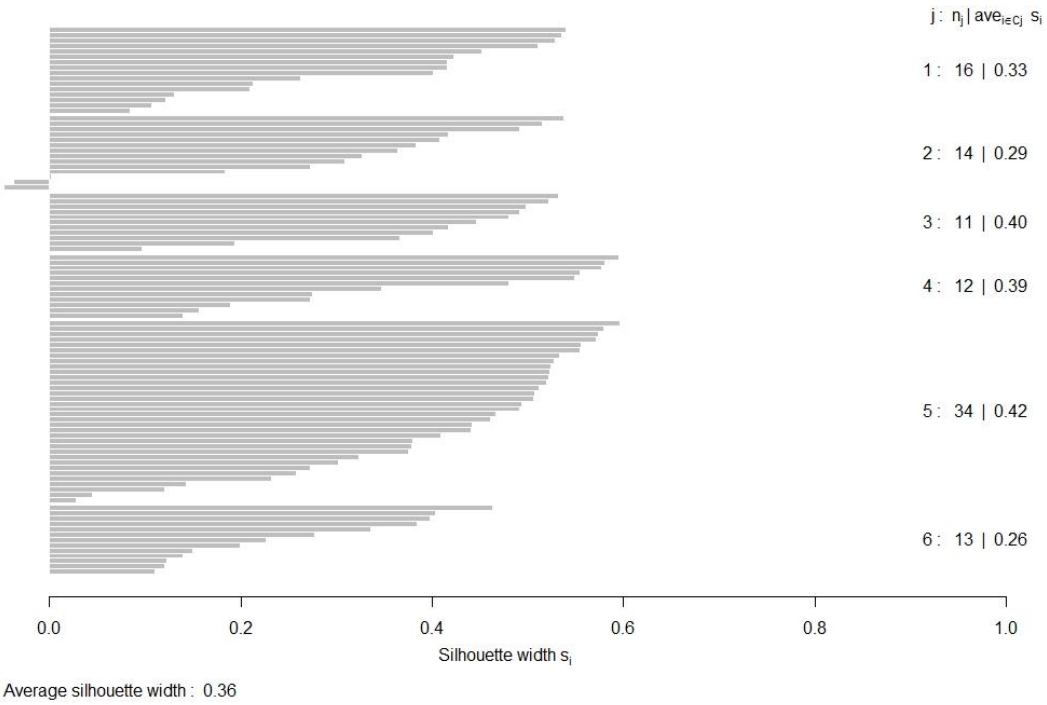
Com as amostras classificadas em diferentes neurônios, foi possível agrupá-las e determinar os conjuntos com maior semelhança. Este procedimento pode ser feito manualmente de acordo com uma validação visual do resultado deste agrupamento de neurônios. No entanto, para ampliar as capacidades do modelo, e retirar a subjetividade desta seleção manual, o K-*means* foi utilizado para realizar esta tarefa.

Os dados de entrada foram os neurônios criados pelo SOM e as informações de saída foram os clusters criados pelo K-*means* para as amostras dos dois bancos de dados. O algoritmo usado no K-*means* foi o Hartigan-Wong (1979) (Hartigan e Wong, 1979). O algoritmo Hartigan-Wong geralmente faz um trabalho melhor se comparado com outras técnicas (Forgy (1965); Lloyd (1957); Macqueen (1967) e por isso este método foi escolhido para o modelamento.

Além disso, outros parâmetros usados no K-*means* foram o número de interações e número de inícios aleatórios. Para o primeiro, o valor selecionado foi de 50 sendo o suficiente para processar os dados de entrada sem extrapolar no tempo de processamento. Já para o segundo, o valor selecionado foi igual a 1000 para reduzir ao máximo a variação dos resultados de saída obtidos. Porém, ainda existe um terceiro parâmetro de entrada que deve ser determinado, o número de grupos (*clusters*).

Para determinar o número de clusters otimizado para os dados de entrada, é necessário que o K-*means* seja feito variando o número de grupos em um intervalo, neste caso entre 2 a 50 grupos. Cada K-*means* realizado foi então submetido a uma análise por diagrama de silhueta (Rousseeuw, 1987) e de acordo com a médias das dissimilaridades obtidas, a mais próxima de 1 foi selecionada como número ótimo de grupos. Na Figura 22 é possível ver um exemplo de um diagrama de silhueta com a dissimilaridade de cada grupo e a média de todos os grupos. Neste caso, 6 grupos foram utilizados.

Figura 22 - Diagrama de silhueta.



Fonte - R CORE TEAM, (2016).

Após a obtenção do melhor grupo otimizado, o K-*means* foi realizado com as variáveis otimizadas. O resultado obtido determinou quais neurônios, e por consequência as amostras, seriam agrupadas em cada grupo, gerando o agrupamento final.

Com os grupos determinados, estes foram submetidos às equações desenvolvidas por Deutsch (1989) para as Células Móveis. Com a aplicação dessas equações foi possível determinar o peso de cada amostra e a média desagrupada. As Equações 10 e 11 representam os cálculos usados para determinação dos pesos e das médias desagrupadas respectivamente, de acordo com Deutsch (1989).

$$w_i = \frac{1}{n_l . l_o} \tag{10}$$

$$\bar{Z} = \sum_{i=1}^{n} w_i . z_i \tag{11}$$

Sendo que, w_i é o peso das amostras do grupo, n_l é o número de grupos, l_o o número de amostras no grupo, $\bar{Z}$ a média desagrupada e z_i uma amostra i do grupo de dados. Para o cálculo do desvio padrão dos dados desagrupados, a Equação 12 pode ser utilizada.

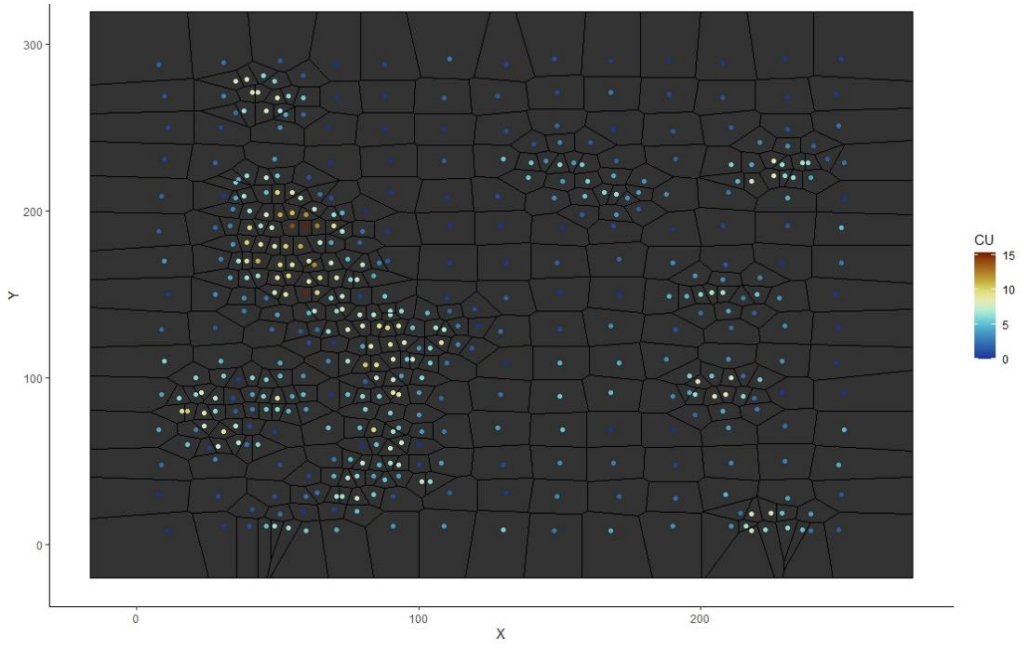
$$S = \sqrt{\sum_{i=1}^{n} w_i (z_i - \bar{Z})^2} \tag{12}$$

Com os dados desagrupados, é possível dar continuidade para a visualização dos dados. Para esta etapa foi usado a técnica de Polígonos de Voronoi (Voronoy, 1904) como será descrito no Tópico 3.3.3.

3.3.3 Visualização dos resultados do modelo

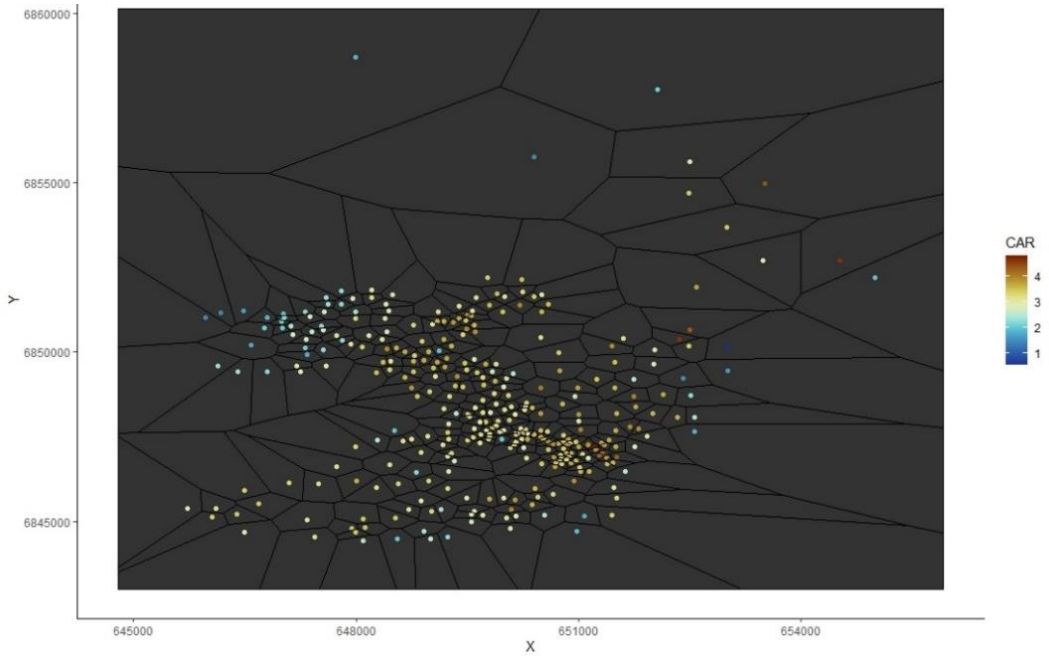
Como dito anteriormente para a visualização dos resultados foi utilizada a técnica de Polígonos de Voronoi (Voronoy, 1904). A escolha desse método teve como principal justificativa a fácil interpretação especial dos resultados do desagrupamento além de facilitar a interpretação visual dos resultados do agrupamento dos neurônios. Nas Figuras 23 e 24 é possível visualizar o resultado desta técnica nos dados do *Walker Lake* e do Carvão sem a aplicação dos grupos do K-*means*.

Figura 23 - Polígono de Voronoi para o Walkerlake.



Fonte – R CORE TEAM, (2016).

Figura 24 - Polígono de Voronoi para o Carvão.



Fonte - R CORE TEAM, (2016).

Após esta etapa, foi possível dar prosseguimento para o último estágio desta pesquisa, a validação dos modelos. Este será abordado no Tópico 3.4.

3.3.4 Comparação com os métodos clássicos

Na Tabela 3 é possível ver as médias obtidas para cada método tradicional em função do conjunto de dados estudado (*Walker Lake* ou Carvão). Essas informações são a base para determinar a eficiência do modelo utilizando o SOM e descobrir se a metodologia aqui proposta realmente consegue desagrupar os dados de forma eficiente.

Tabela 3 - Médias desagrupadas pelos métodos tradicionais.

	RESULTADOS DAS MÉDIAS DESAGRUPADAS	
	Células Móveis	Vizinho mais Próximo
Walker Lake	291,90 ppm	277,52 ppm
Carvão	2,76 metros	2,62 metros

Com essas informações, foi possível determinar a eficiência da metodologia de desagrupamento desenvolvida nesta pesquisa. Desta forma é possível dar seguimento e apresentar os resultados obtidos.

4. RESULTADOS E DISCUSSÕES

Para apresentação dos resultados deste trabalho, a mesma divisão usada na metodologia será utilizada para facilitar a interpretação e discussão das informações obtidas em cada etapa. Desta forma, a primeira parte que será abordada é o tratamento dos dados vindos do *Walker Lake* e do Carvão.

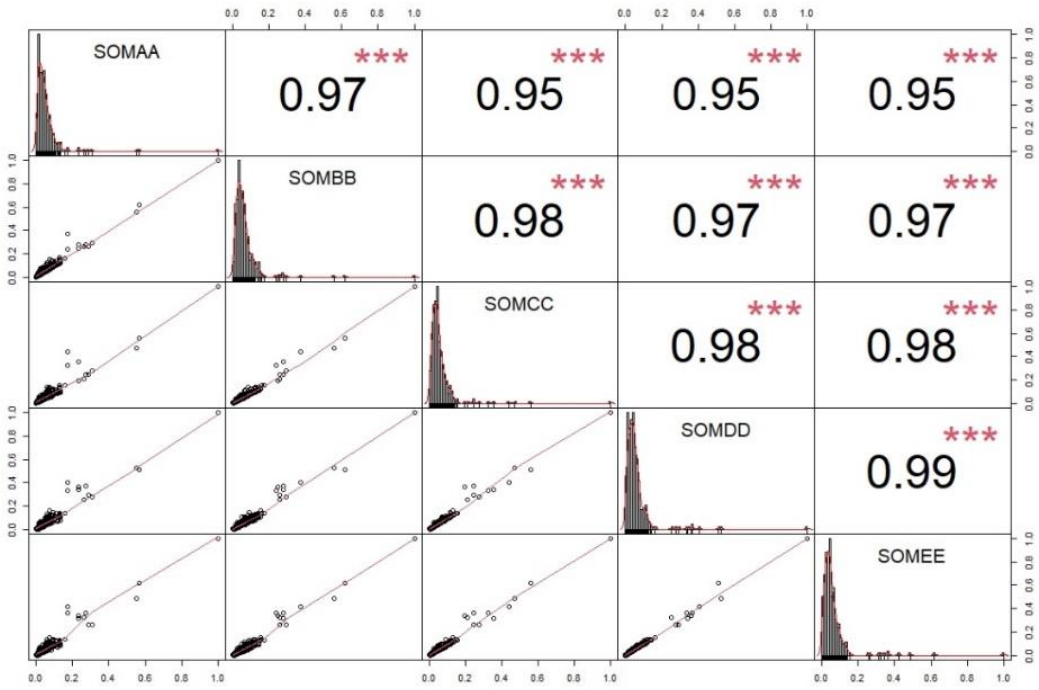
4.1 Resultados do tratamento de dados

Após a utilização da matriz de distâncias, a seleção das 5 menores distâncias e a normalização dessas 5 novas variáveis, como explicado no Tópico 3.2, obteve-se os dois novos bancos de dados que foram utilizados na implementação do modelo.

Para comprovar o efeito positivo após a inserção das novas variáveis, primeiramente no banco de dados do Carvão, essas 5 variáveis foram submetidas a uma matriz de correlação para determinar a correlação existente de cada variável entre si mediante a verificação por pares. Assim, uma vez identificada a associação entre as variáveis as mesmas foram nomeadas como SOMAA, SOMBB, SOMCC, SOMDD e SOMEE para os dois novo bancos de dados. Respectivamente estas variáveis representam da menor distância até a quinta menor distância da matriz de distâncias.

Na Figura 25 é possível ver a matriz de correlação, utilizando o método de Pearson. É notável o efeito positivo que o tratamento de dados fez com o resultado do banco de dados do Carvão, obtendo-se elevados valores de correlação entre as variáveis.

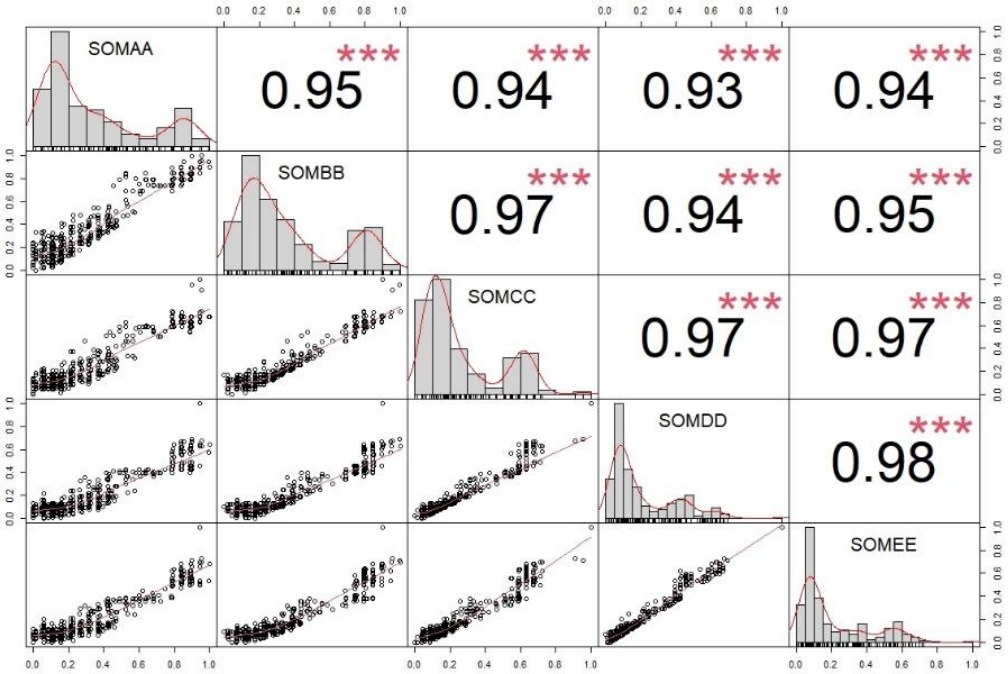
Figura 25 Matriz de correlação dos dados do Carvão.



Fonte - R CORE TEAM, (2016).

O mesmo pode ser visto na Figura 26, o mesmo comportamento do tratamento pode ser visto para os novos dados do *Walker Lake*.

Figura 26: - Matriz de correlação dos dados do *Walker Lake*.



Fonte - R CORE TEAM, (2016).

4.2 Resultados da implementação do Modelo do *Walker Lake*

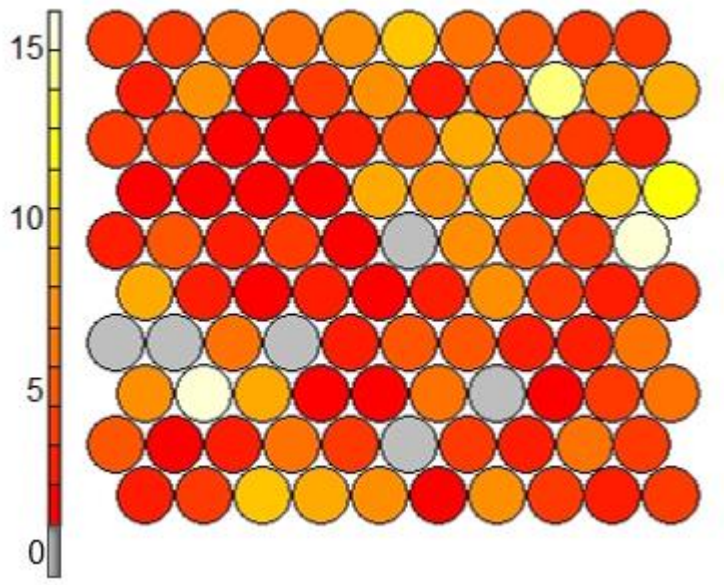
Para facilitar o entendimento dos resultados do modelo do *Walker Lake* e do Carvão, cada um será abordado separadamente e no final, os dados obtidos serão comparados e validados para determinar a eficiência da metodologia para desagrupar dados amostrais.

Primeiramente será abordado os resultados do *Walker Lake*, as primeiras discussões abordarão o SOM e suas características resultantes.

4.2.1 SOM do *Walker Lake*

Os primeiros resultados obtidos foram os diagramas de neurônios resultantes das Redes de Kohonen. Na Figura 27 é possível ver o número de amostras selecionadas para cada neurônio do *grid* criado. O ideal dessa etapa é que não haja muitos neurônios vazios e nem neurônios sobrecarregados, o equilíbrio é necessário para que não se tenha um excesso de processamento desnecessário ou uma baixa taxa de processamento respectivamente as condições citadas anteriormente. O resultado obtido foi satisfatório demonstrando a eficiência na seleção do grid de neurônios desenvolvido.

Figura 27 - Contagem de amostras nos neurônios do grid do *Walker Lake*.

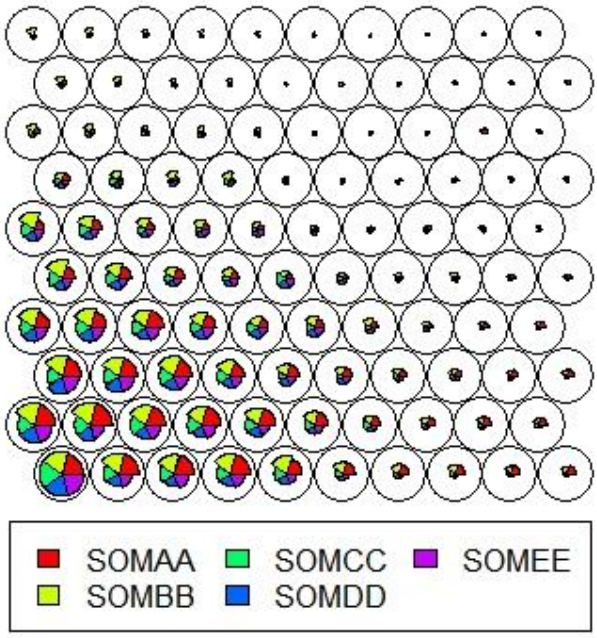


Fonte - R CORE TEAM, (2016).

O próximo diagrama está relacionado com a seleção das variáveis para cada neurônio. Cada neurônio possui uma regra de classificação em função das variáveis de entrada, de acordo com essa classificação as amostras são selecionadas para cada neurônio. A representação das variáveis em cada neurônio é representada por um gráfico de pizza, cada fatia representa uma

variável e quanto maior for o raio da fatia de pizza, maior será o intervalo de valores aceitáveis nesse neurônio conforme pode ser observado na Figura 28.

Figura 28 - Relação das variáveis nos neurônios do *Walker Lake*

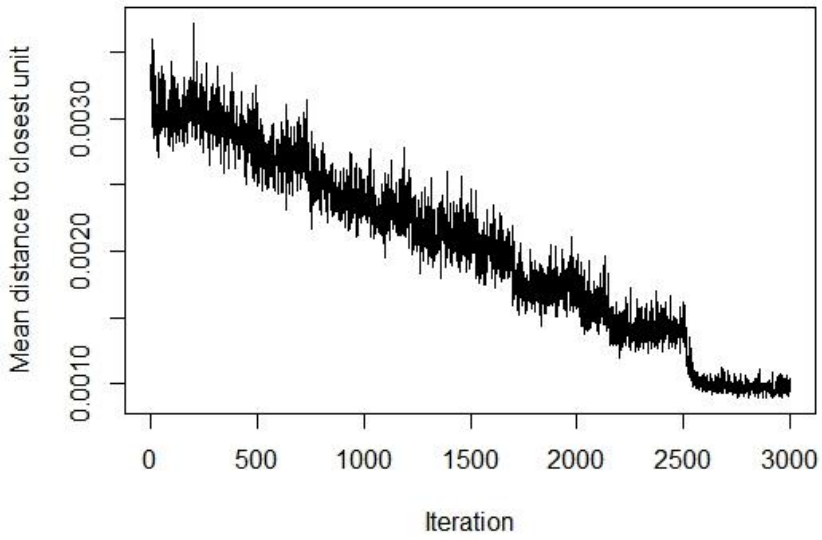


Fonte - R CORE TEAM, (2016).

Por fim, o último diagrama gerado é o gráfico do treinamento do modelo do SOM, esse representa a distância média entre as amostras de entrada dentro da rede. Quanto menor for essa distância, melhor é a qualidade do modelamento. O número de interações entre as amostras determina o quanto este processo de “aproximação” deve ser feito, quando este processo se estabiliza, recomendasse que o desenvolvimento da rede tenha fim pois não há mais ganhos de aprendizado neste ponto.

Como pode ser visto na Figura 29 o resultado obtido para o treinamento foi satisfatório em 3000 interações, alcançando todas as metas de desempenho necessárias para a criação do modelo SOM.

Figura 29: Gráfico do treinamento do SOM para o *Walker Lake*



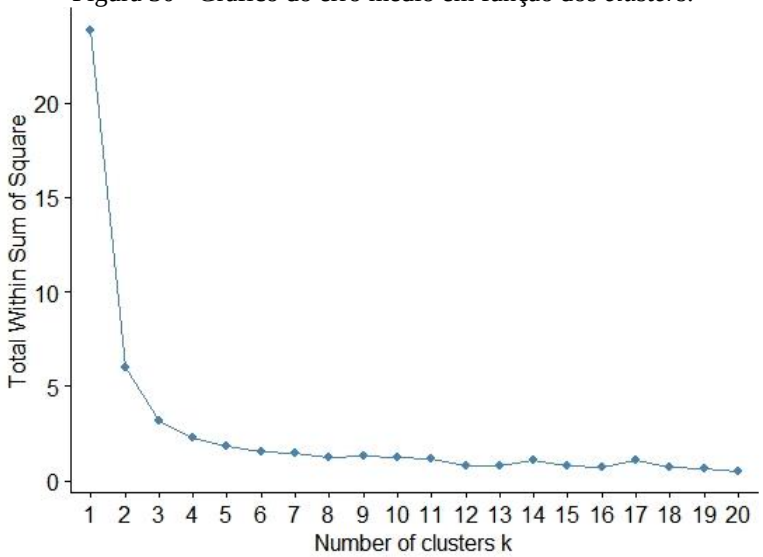
Fonte - R CORE TEAM, (2016).

Após esta etapa, os resultados foram passados para o processamento do K-*means*. Como explicado no Tópico 3.3.2, está procedimento teve como objetivo principal agrupar os neurônios em grupos de acordo com a dissimilaridade entre eles.

4.2.2 Resultado do K-*means* para o *Walker Lake*

Após a obtenção dos neurônios do SOM, este foram submetidos ao agrupamento do K-*means*. A seleção do número de grupos ótimo pode ser vista na Figura 30, a curva representa o erro médio do K-*means* em função do número de clusters, quando a diminuição desse erro se estabiliza, o número de grupos é escolhido.

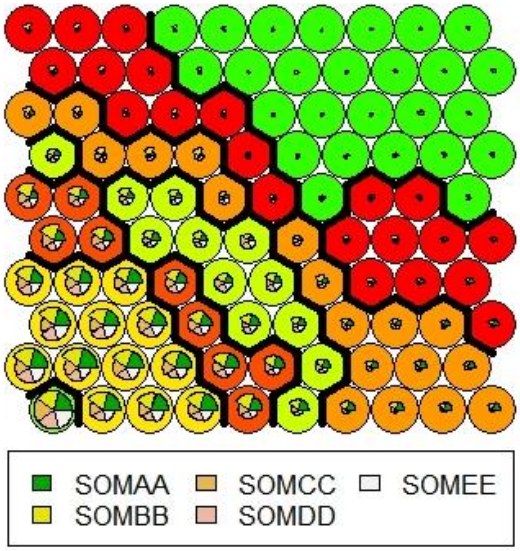
Figura 30 - Gráfico do erro médio em função dos *clusters.*



Fonte - R CORE TEAM, (2016).

A otimização de parâmetros selecionou 7 grupos para segregar de forma eficiente os neurônios como pode ser visto na Figura 31. Nela é possível ver em diferentes cores a seleção feita pelo K-*means* e as fronteiras entre grupos dos neurônios. Uma característica importante para se notar é que os grupos de neurônios não necessariamente estão conectados diretamente, podendo ser separados por outros neurônios.

Figura 31 - Seleção dos grupos nos neurônios do *Walker Lake*.



Fonte - R CORE TEAM, (2016).

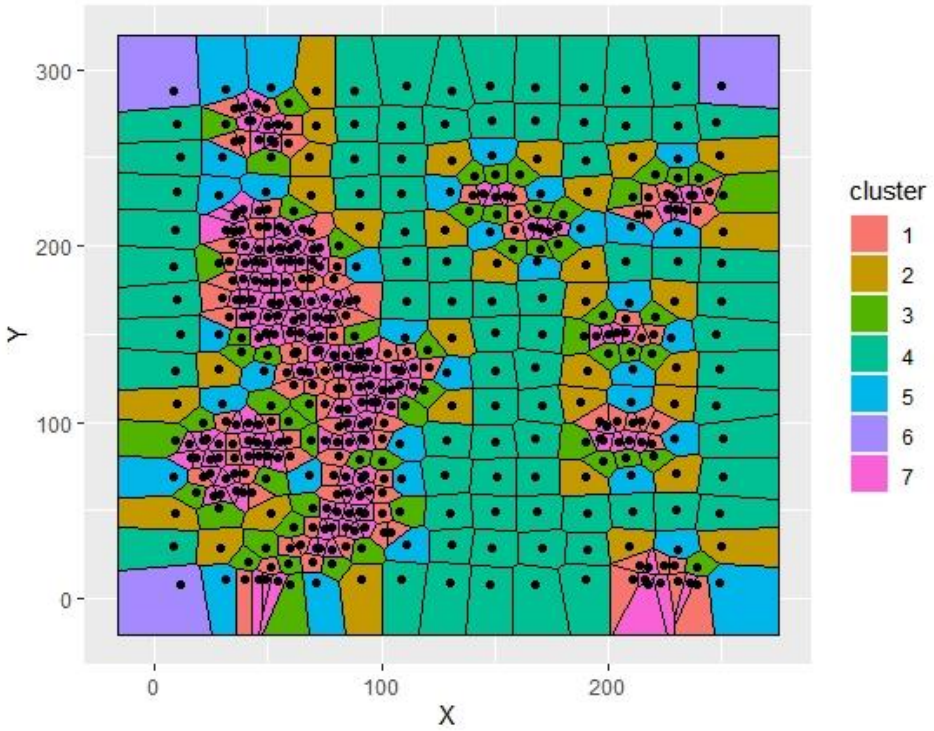
Com a seleção dos grupos feita, é possível prosseguir para a visualização dos resultados. Como dito anteriormente no Tópico 3.0 dessa pesquisa, essa etapa foi feita utilizando os polígonos de Voronoi.

4.2.3 Resultados da visualização do modelo *Walker Lake*

Esse tópico da pesquisa é de grande importância para a estrutura geral do modelo: além de facilitar a observação dos resultados obtidos, também possui um caráter de validação visual, ou seja, proporciona coerência mediante o que foi observado. Essa validação deve ser feita para garantir que a seleção das amostras nos grupos tenha respeitado a estruturação original dos dados, ou seja, que representem a espacialidade dos dados agrupados e não agrupados mesmo após a transformação feita nos dados originais.

Os resultados obtidos foram computados e representados na Figura 32. O resultado obtido foi satisfatório pois alcançou os objetivos explicados anteriormente em relação a espacialidade dos dados. Ao mesmo tempo em que conseguiu agrupar bem as diferentes nuances do agrupamento.

Figura 32: Representação dos grupos do modelo nos dados do *Walker Lake* com polígonos de Voronoi.



Fonte - R CORE TEAM, (2016).

Após a apresentação desses dados, serão abordados os resultados obtidos com a modelagem do Carvão. Ao final dessa apresentação, que passará pelos mesmos tópicos que o *Walker Lake* passou, os dois modelos se encontrarão para a última parte dos resultados, a validação dos modelos.

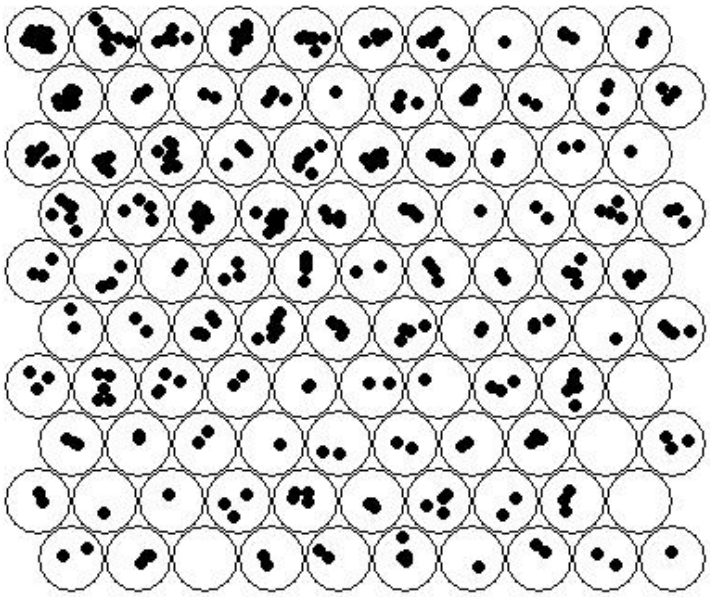
4.3 Resultado da implementação do Modelo do Carvão

O mesmo desenvolvimento descrito no Tópico 4.2 foi utilizado para apresentar os resultados do modelamento dos dados do Carvão. Este procedimento foi feito para facilitar a leitura e entendimento dos resultados obtidos em cada etapa.

4.3.1 SOM do Carvão

Os primeiros resultados obtidos, como anteriormente, foram os diagramas de neurônios resultantes das Redes de Kohonen. Na Figura 33 é possível ver o número de amostras selecionadas para cada neurônio do *grid* criado. Porém desta vez, o gráfico de contagem é apresentado de forma diferente, neste caso as amostras em cada neurônio são representadas visualmente. O resultado obtido foi satisfatório demonstrando a eficiência na seleção do *grid* de neurônios desenvolvido.

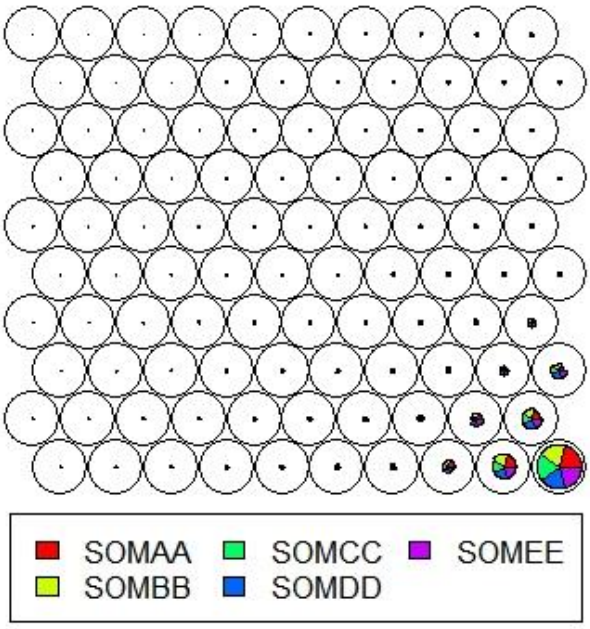
Figura 33: Contagem de amostras no grid do SOM Carvão



Fonte - R CORE TEAM, (2016).

O próximo diagrama está relacionado com a seleção das variáveis para cada neurônio. Como explicado anteriormente, as variáveis em cada neurônio são representadas por um gráfico de pizza, cada fatia representa uma variável e quanto maior for o raio da fatia de pizza, maior será o intervalo de valores aceitáveis nesse neurônio, conforme Figura 34.

Figura 34 Relação das variáveis do Carvão em cada neurônio

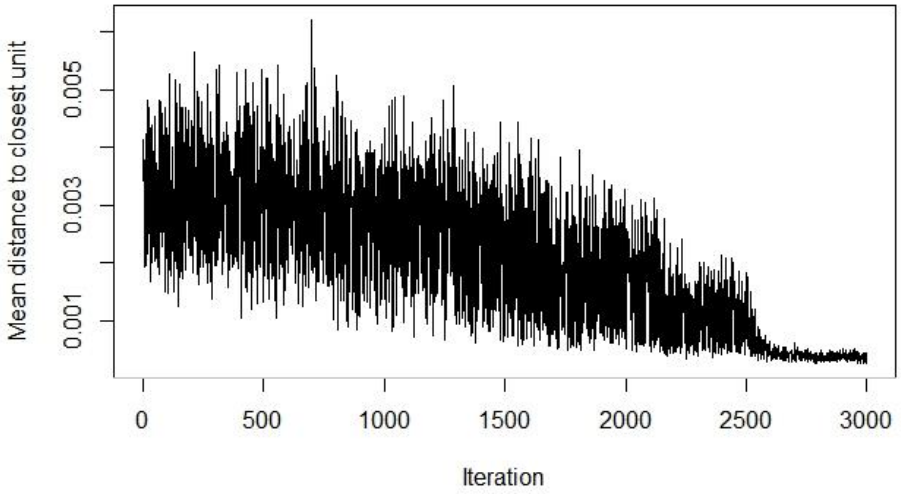


Fonte - R CORE TEAM, (2016).

Por fim, o último diagrama gerado é o gráfico do treinamento do modelo do SOM. Como pode ser visto na Figura 35 o resultado obtido para o treinamento foi satisfatório em 3000

interações, alcançando todas as metas de desempenho necessárias para a criação do modelo SOM. Um detalhe importante nos resultados obtidos aqui é que, as oscilações no treinamento visto na Figura 29 são maiores se comparados com o mesmo diagrama do SOM do *Walker Lake*. Isso já era esperado pela heterogeneidade na distribuição espacial dos dados do Carvão, o que tornaria mais "difícil" o treinamento das redes. Porém, mesmo com essa dificuldade, o resultado obtido foi satisfatório.

Figura 35 - Diagrama do treinamento do Carvão.



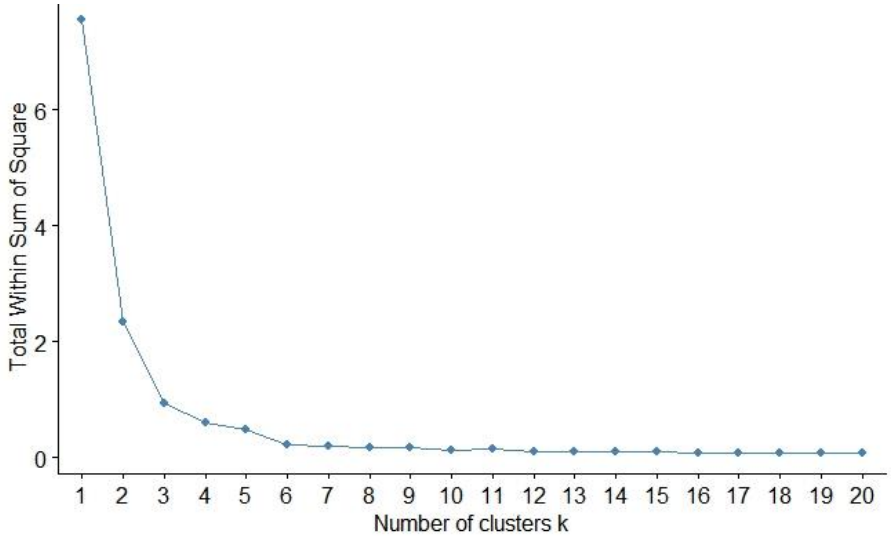
Fonte - R CORE TEAM, (2016).

Após esta etapa, os resultados foram passados para o processamento do K-*means*. Como explicado no Tópico 3.3.2, está procedimento teve como objetivo principal agrupar os neurônios em grupos de acordo com a dissimilaridade entre eles, como feito para o SOM do *Walker Lake*.

4.3.2 Resultado do K-*means* para o Carvão

Após a obtenção dos neurônios do SOM, este foram submetidos ao agrupamento do K-*means*. A seleção do número de grupos ótimo pode ser vista na Figura 36, a curva representa o erro médio do K-*means* em função do número de *clusters*, quando a diminuição desse erro se estabiliza, o número de grupos é escolhido.

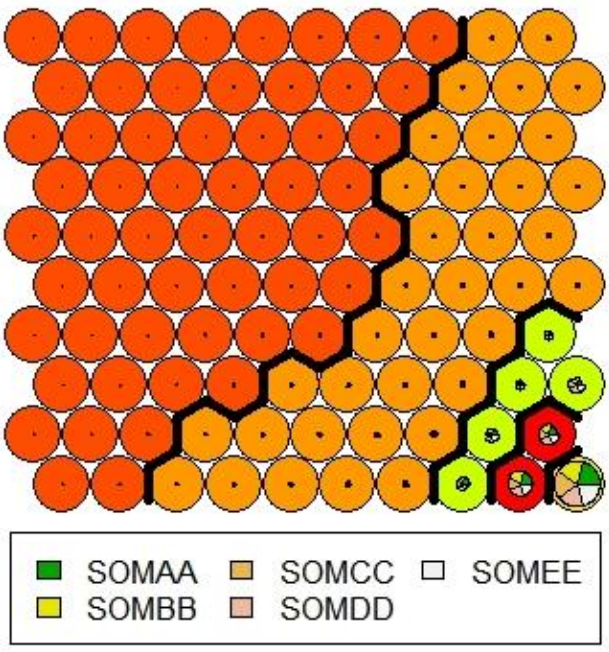
Figura 36 Gráfico do erro médio em função dos clusters.



Fonte - R CORE TEAM, (2016).

A otimização de parâmetros selecionou 5 grupos para segregar de forma eficiente os neurônios como pode ser visto na Figura 37. Nela é possível ver em diferentes cores a seleção feita pelo K-*means* e as fronteiras entre grupos dos neurônios.

Figura 37: Seleção dos grupos nos neurônios do Carvão.



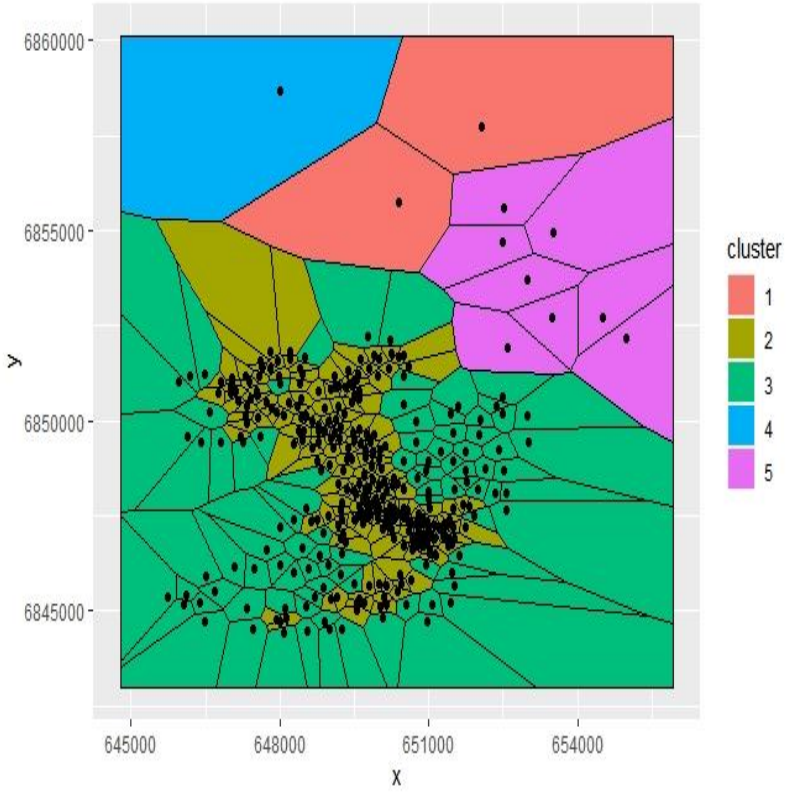
Fonte - R CORE TEAM, (2016).

Com a seleção dos grupos feita, é possível prosseguir para a visualização dos resultados. Como dito anteriormente no Tópico 3.0 dessa pesquisa, essa etapa foi feita utilizando os polígonos de *Voronoi*.

4.3.3 Resultado da visualização do modelo do Carvão

Os resultados obtidos foram computados e representados na Figura 38. O resultado obtido foi satisfatório pois alcançou os objetivos em relação a espacialidade dos dados. Ao mesmo tempo que conseguiu agrupar bem as diferentes nuances do agrupamento.

Figura 38: Representação dos grupos do modelo nos dados do Carvão com polígonos de Voronoi.



Fonte - R CORE TEAM, (2016).

Com todos os resultados dos modelos em mão, foi possível partir para a última etapa da pesquisa, a validação dos resultados pela comparação com os métodos tradicionais de desagrupamento.

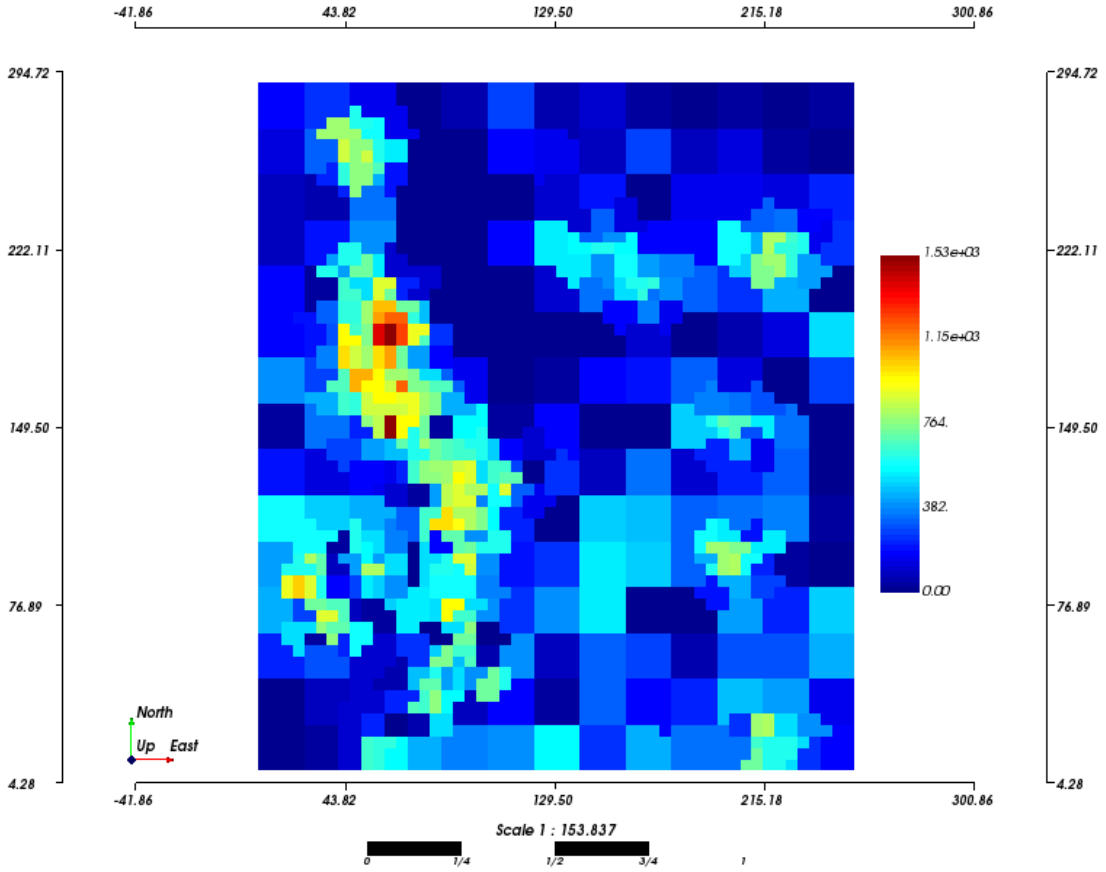
4.4 Validação dos resultados dos modelos

Como descrito no Tópico 3.4, a validação dos resultados foi feita a partir da comparação do desagrupamento feito utilizando Células Móveis e Vizinho mais próximo. A média de cada técnica foi calculada e comparada com a média obtida pelo método desenvolvido nesta pesquisa como descrito no Tópico 3.3.2.

4.4.1 Métodos Tradicionais

As primeiras médias desagrupadas apresentadas neste Tópico será do Vizinho mais próximo (NN) para os dados do *Walker Lake* e do Carvão. Utilizando esta técnica, além do resultado numérico obtido, também se obteve resultados espaciais em relação a estimativa de teor no *grid*. Na Figura 39 está representado o NN para o *Walker Lake*. Nela é possível verificar a estimativa do NN de forma que regiões em azul representam teores menores de cobre e a regiões em vermelho, teores maiores.

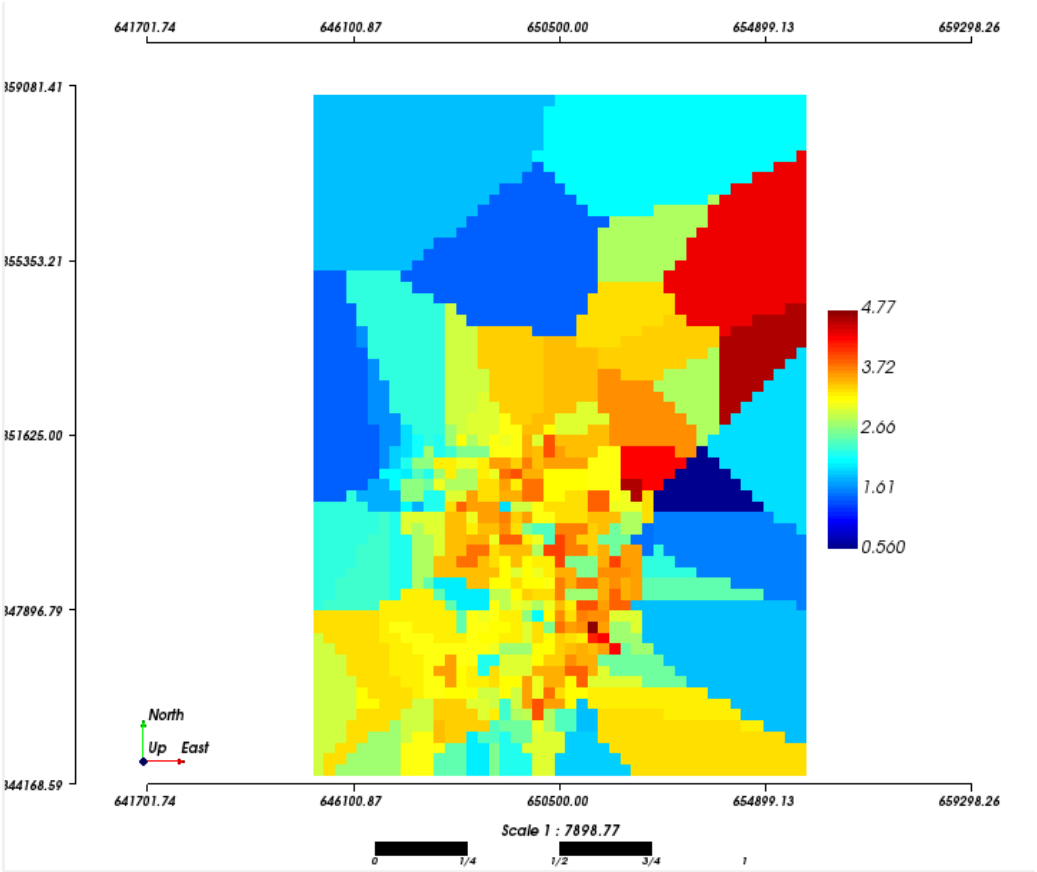
Figura 39 - NN para o *Walker Lake*.



Fonte - SGeMS, (2011).

Da mesma forma, a Figura 40 apresenta o NN do Carvão e tal qual foi descrito anteriormente, a representação das cores na escala permanece a mesma, porém para esse caso, representa-se dados da espessura de carvão na região.

Figura 40 - NN para o Carvão



Fonte R CORE TEAM, (2016).

A seguir são apresentadas as médias desagrupadas para os dois métodos: NN e Células Móveis. Considerado o NN, para o *Walker Lake* e para o Carvão as médias foram 277,52 ppm de cobre e 2,62m de espessura. Já no Método de Células Móveis, para o *Walker Lake* considerando um tamanho de célula de 20x20, a média desagrupada foi de 291,90 ppm; por outro lado, para o carvão, considerando uma célula de 4400x4400, a média desagrupada foi de 2,76 metros de espessura. As Figuras 41 e 42 apresentam os gráficos referentes ao método de Células Móveis que mostram a relação tamanho de célula *versus* média desagrupada.

Figura 41 – Método de Células Móveis para o *Walker Lake*

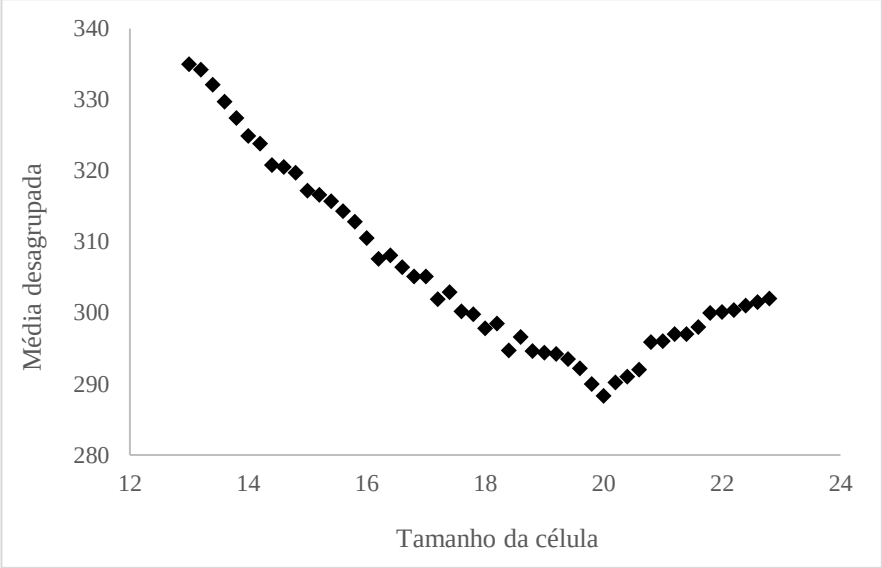
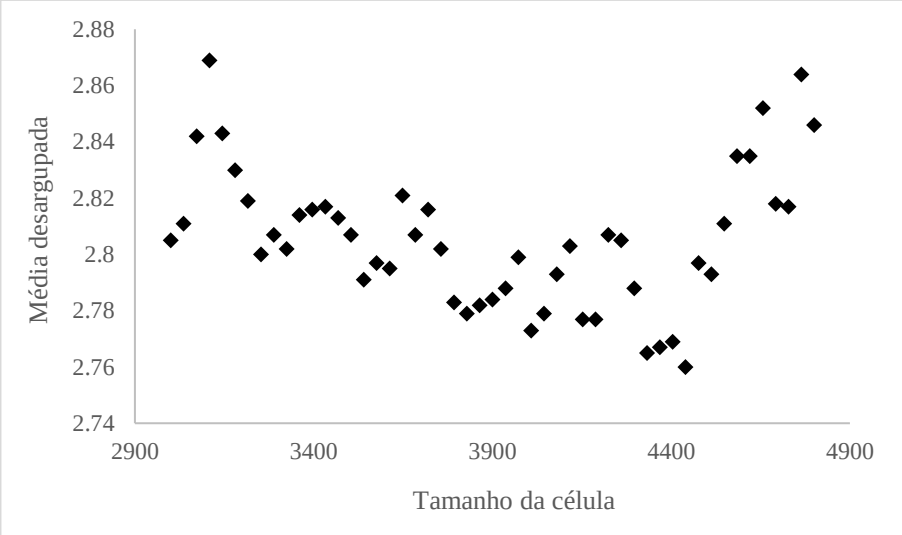


Figura 42 – Método de Células Móveis para o Carvão



4.4.2 Comparação das médias

Como explicado no Tópico 3.3.2, as médias do *Walker Lake* e do Carvão para esta nova metodologia proposta foi feita de acordo com as fórmulas desenvolvidas por Deutsch, (1989) para as Células Moveis. Utilizando as Equações 2 e 3 foi possível determinar os pesos de cada amostra e a média desagrupada.

Para validar as capacidades de estimativa do modelo proposto, foi necessário comparar os resultados das médias desagrupadas pelos métodos tradicionais (Células Móveis e Vizinho mais próximo) com as médias desagrupadas obtidas pelo modelo desenvolvido. Além dessas, para facilitar a discussão dos resultados, as médias originais dos bancos de dados do *Walker Lake* e do Carvão também foram disponibilizados na Tabela 4.

A primeira informação que se pode extrair desses resultados é que todas as médias das 3 técnicas conseguiram desagrupar as amostras dos dois bancos de dados. Isso é visto pela diminuição das médias calculadas se comparadas com as médias originais.

Tabela 4 Médias finais obtidas.

	RESULTADOS DAS MÉDIAS DESAGRUPADAS		
	Células Móveis	Vizinho mais Próximo	Modelo SOM
Walkerlake	291,90 ppm	277,52 ppm	293,92 ppm
Carvão	2,76 metros	2,62 metros	2,65 metros
	RESULTADOS DAS MÉDIAS ORIGINAIS (Amostras)		
Walkerlake	435,30 ppm		
Carvão	3,16 metros		

A segunda observação que pode ser feita é a relação mais próxima das médias entre as Células Móveis e o Modelo SOM se comparado com o NN. Isso ocorreu devido ao fato de que essas duas técnicas compartilham a mesma metodologia para os cálculos dos pesos das amostras e da média desagrupada. Além disso, há uma tendência geral em que o NN tende a inferiorizar demais a média desagrupada por causa da influência que a área no entorno das amostras com menor teor tem neste cálculo.

Com estes resultados é possível chegar em um veredito em relação à eficiência do modelo de desagrupamento desenvolvido. As duas médias obtidas para o Modelo SOM demonstram as capacidades de desagrupamento da técnica, obtendo um resultado satisfatório para esta tarefa. Como dito anteriormente, as médias obtidas se aproximam mais dos resultados obtidos pelas Células Móveis.

O grande diferencial da técnica aqui criada, em relação as duas metodologias tradicionais, é a redução da subjetividade na seleção de parâmetros do modelo. Como o Modelo SOM se baseia exclusivamente em técnicas estatísticas para fazer essa seleção de parâmetros, os erros associados a inexperiência dos usuários são muito reduzidos, como ocorre nos outros métodos. Além disso, ao utilizar técnicas robustas de estimativa, o Modelo SOM gera respostas mais confiáveis e de fácil manipulação e melhoramento.

4 CONCLUSÕES

O processo de amostragem, no âmbito da Exploração Mineral, é uma etapa fundamental para o sucesso do empreendimento mineiro na medida em que quantifica o teor, volume e extensão do corpo mineralizado. Nesse sentido, uma das grandes dificuldades é o desenvolvimento de uma malha com dados homogeneamente distribuídos na área de estudo. Como existem interesses específicos em regiões com maior incidência de minérios há um adensamento do número de furos nessas regiões causando uma superestimativa do teor da reserva. Técnicas de desagrupamento conseguem mitigar esse efeito causado pela hetereogeneidade da distribuição dos dados no espaço e da mesma forma, essa pesquisa utilizou técnicas de aprendizado de máquina para alcançar esse mesmo objetivo.

Utilizando as matrizes de distância foi possível criar dois bancos de dados derivados das coordenadas dos pontos das amostras dos bancos de dados do *Walker Lake* e do Carvão. Essa transformação foi necessária para "traduzir" os dados espaciais originais em informações escalares utilizando as menores distâncias entre cada amostra original. A partir daí, as Redes de Kohonen foram "alimentadas" com essas novas informações e o resultado submetido ao agrupamento K-*means* gerando os grupos necessários para o cálculo dos pesos do desagrupamento.

Com os resultados dos pesos foi possível calcular as médias desagrupadas segundo a metodologia descrita nesta pesquisa e compará-las com as médias das técnicas tradicionais de desagrupamento. A média obtida pela aplicação das redes SOM foi próxima do valor encontrado para as médias tradicionais: para o banco de dados *Walker Lake*, o valor encontrado foi de 293,92 ppm diante dos valores de 291,9 ppm utilizando o método de Células Móveis e 277,52 ppm utilizando o NN. Da mesma forma, para o banco de dados do Carvão, o valor obtido aplicando as redes SOM foi de 2,65 m enquanto que o método de Celulas Móveis foi de 2,76m e o NN foi de 2,62m.

Portanto, pelo exposto acima, a aplicação das redes SOM no desagrupamento de dados se mostrou satisfatória além do fato de que reduz a subjetividade na seleção de parâmetros, minimiza erros associados a inexperiência gerando respostas mais assertivas e confiáveis. No entanto, quanto a aplicação desse tipo de algoritmo em inteligência artificial, ele gera incertezas quanto ao seu resultado na medida em que não se consegue explicar de forma clara parâmetros específicos, nesse caso a metodologia aplicada pode ser considerada como *black box* e mostra uma limitação quanto as redes SOM.

Nesse sentido, a afirmação de George E. P. Box "*essentially, all models are wrong, but some are useful*" pode ser contextualizada para essa pesquisa acadêmica. Na medida em que modelos se aproximam da realidade eles são utilizados para fazerem previsões de eventos que não ocorreram, daí o caráter utilitarista desses modelos: mesmo que simplistas, eles têm a capacidade de permitir avaliar o espectro de possibilidades, mesmo que algumas delas estejam erradas elas podem capilarizar *insights* e fornecer novas perspectivas.

Além disso, dado o caráter de análise inicial dessa pesquisa, é importante salientar que por mais que essa pesquisa obteve bons resultados, como proposta para trabalhos futuros, é interessante replicar a metodologia em outros bancos de dados e adicionalmente, que contenham dados mais heterogêneos para validar e consequentemente, comprovar eficiência da técnica bem como sua aplicabilidade. Além disso, considerar a análise de deriva como ferramenta de verificação dos teores nos eixos X e Y; avaliar os coeficientes de variação dos resultados obtidos pelo método; considerar a inclusão além das 5 menores distâncias bem como analisar o desempenho computacional inerente a essa inclusão, isto é, a otimização dos resultados frente a esse incremento.

5 REFERÊNCIAS

ARIOLI, E. E.; ANDRIOTTI, J. L. S. Representatividade da amostragem na prospecção geoquímica. *In*: LICHT, O. A. B.; MELLO, C. S. B. de; SILVA, C. R. da (Ed.). **Prospecção Geoquímica**. Rio de Janeiro: CPRM; SBGq, 2007 Disponível em: <https://rigeo.cprm.gov.br/handle/doc/487>. Acesso em: 20 jan.2022.

AYACHE, N. K. **Avaliação das condições de estabilidade de taludes de mina por meio de árvores de decisão**.2021. Trabalho de Conclusão de Curso (graduação) – Centro Federal de Educação Tecnológica de Minas Gerais, CEFET-MG: Araxá, 2021.

BELUCO, A. **Modelo híbrido SOM-ANN/BP para previsão de índices da NYSE através de redes neurais artificiais**. 2013. Dissertação (Mestrado em Programa de Pós-Graduação em Administração) – Universidade Federal do Rio Grande do Sul, Porto Alegre: UFRGS, 2013.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. Singapore: Springer,2006.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and regression trees. **Wadsworth International**: California, USA, 1984.

BUSSAB, W. de O.; MIAZAKI, E. S; ANDRADE, D. F. de. **Introdução à Análise de Agrupamentos**. 9º Simpósio Nacional de Probabilidade e Estatística., 9., São Paulo: ABE, 1990.

CAPPONI, L. **Planejamento estocástico de curto prazo incorporando a incerteza da estimativa no controle de teores**. 2019. Tese (doutorado) – Universidade Federal do Rio Grande do Sul, Escola de Engenharia. Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e de Materiais, Porto Alegre: UFRGS, 2013.

CARVALHO, L.A.V. **Datamining: a mineração de dados no Markenting, Medicina, Economia, Engenharia e Adminstração**. São Paulo: Érica, 2001.

CINTRA, E. C. **Aplicação de redes neurais no controle de teores de cobre e ouro do depósito de Chapada (GO)**. 2003. Tese (doutorado) – Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas, 2003. Disponível em: <http://hdl.handle.net/11449/103035>. Acesso em: 12 jan.2022.

COSTA, J. F.C. L.; SOUZA, L. E. **Métodos de Desagrupamento**: Método de Células Móveis. Rio Grande do Sul. (Notas de aula).

CORNETTI, M. A. **O impacto do uso de ponderadores de dados agrupados na geoestatistica**. 2003. Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Geociencias, Campinas, SP: UNICAMP, 2003 Disponível em: <http://www.repositorio.unicamp.br/handle/REPOSIP/287474>. Acesso em: 10 jan. 2022.

COVER, T. HART, P. (1967) Nearest Neighbor Pattern Classification. **IEEE Transactions on Information Theory**, IT-11, pp. 21-27, 1967. Disponível em: <http://dx.doi.org/10.1109/TIT.1967.1053964>. Acesso em: 12 jan.2022.

DARWIN, C. **The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life**. 6. ed. London: John Murray, 1872. Disponível em: < http://darwin-online.org.uk/converted/pdf/1861_OriginNY_F382.pdf >. Acesso em: 28 dez. 2021

DEUTSCH, C. V.; JOURNEL, A. G. **Geostatistical software library and user's guide** - GSLIB. Oxford University Press, 1992.

DONI, M. V. **Análise de *cluster*:** métodos hierárquicos e de particionamento. 2004. Trabalho de Conclusão de Curso (graduação) – Mackenzie: São Paulo, 2004. Disponível em: < http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF>. Acesso em: 28 dez. 2021

ELSALAMONY, H. A. Bank direct marketing analysis of data mining techniques. **International Journal of Computer Applications, Foundation of Computer Science (FCS)**, v. 85, n. 7, p. 12–22, 2014.

ESTEVÃO JÚNIOR, J. **Avaliação de incerteza associada à Krigagem de variáveis indicadoras e à modelagem de *grade shell* em depósito de ouro**. Dissertação (Mestrado em Recursos Minerais e Meio Ambiente) – Universidade de São Paulo, Instituto de Geociências, São Paulo, 2019. Disponível em: <http:// doi:10.11606/D.44.2020.tde-13022020-151348>. Acesso em: 10 jan. 2022.

FALQUETO, D. **Rede neural artificial para reconhecimento de horários de arme/desarme no sistema SIGMA**. 2007. Trabalho de Conclusão de Curso (graduação em Ciência da Computação) – Universidade do Vale do Itajaí: São José, SC, 2007 Disponível em: <http://siaibib01.univali.br/pdf/Daniel%20Falqueto.pdf>. Acesso em: 10 jan. 2022.

FERREIRA, T. C. de O. **Análise de incertezas do modelo de teores associado aos investimentos de pesquisa de longo prazo**. Dissertação (Mestrado em Recursos Minerais e Meio Ambiente) – Universidade de São Paulo, Instituto de Geociências, São Paulo, 2016. Disponível em: <http://doi:10.11606/D.44.2016.tde-17062016-095800>. Acesso em: 10 jan. 2022.

FRANCISCO, C.A.C. **REDE DE KOHONEN**: Uma ferramenta no estudo das relações tróficas entre espécies de peixes. Dissertação (Mestrado) – Universidade Federal do Paraná, Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Curitiba, 2004. Disponível em: <https://acervodigital.ufpr.br/bitstream/handle/1884/34757/R%20-%20D%20%20CLAUDIA%20APARECIDA%20CAVALHEIRO%20FRANCISCO.pdf?sequence=1&isAllowed=y>. Acesso em: 03 jan. 2022.

GELAIN, E. **Arranjo amostral para mapeamento de atributos do solo**. 2016. Tese (Doutorado em Agronomia) – Universidade Federal da Grande Dourados, Faculdade de Ciências Agrárias, Dourados, MS, 2016. Disponível em:< http://repositorio.ufgd.edu.br/jspui/handle/prefix/457>. Acesso em: 03 jan. 2022.

GONÇALVES, S V. **Aprendizado Baseado em Instâncias Subsidiado por Conjuntos Locais**. 2020. Dissertação (Mestrado em Ciência da Computação) – Centro Universitário Campo Limpo Paulista, Campo Limpo Paulista, SP, 2020. Disponível em:<

https://www.cc.faccamp.br/Dissertacoes/SandroVieiraGoncalves.pdf>Acesso em: 03 jan. 2022.

HAIR, J. F. ANDERSON, R. E. TATHAM, R. L. BLACK, W. C. **Análise Multivariada deDados**, 6.ed. Porto Alegre: Bookman, 2009.593 p.

HAN, Jiawei; KAMBER, Micheline. ***Data Mining:*** *concepts and techniques*. San Diego: Academic Press, 2001.

HAYKIN, S. **Redes neurais**: princípios e prática. 2.ed. Porto Alegre: Bookman, 2001. 900 p.,

HERNÁNDEZ, L. C.H. **Modelos de Markov com estados ocultos na modelagem de séries de vazões anuais**. 2013. Dissertação (Mestrado em Tecnologia Ambiental e Recursos Hidrícos) Universidade de Brasilia, Departamento de Engenharia Civil e Ambiental, Brasília. Disponível em: <https://repositorio.unb.br/bitstream/10482/17997/1/2013_LuisCarlosHernandezHernandez.pdf> Acesso em: 13 jan. 2022

HUANG, Z. Clustering Large Data Sets with Mixed Numeric and Categorical Values. *In*: **Proceedings of the 1st Pacific-Asia Conference on knowledge discovery and datamining (PAKDD'97)**. Singapore: PAKDD'97,1997. p. 21–34.

ISAAKS, E. H.; SRIVASTAVA, R. M. **An introduction to applied geoestatistics**. New York: Oxford University Press, 1989.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6. ed. Pearson,2018. 808 p.

KUNCHEVA, L.; WHITAKER, Ch. Measures of Diversity in Classifier Ensembles, **Machine Learning**, v. 51, 2003 p. 181-207

LANTZ B. **Machine learning with R**. Birmingham: Packt Publishing, 2013. p. 375.

LOHMANN, E. S. **Técnica de Redes Neurais Artificiais aplicada ao Reconhecimento de Imagens para a interação com um Jogo Computacional**. 2016. Trabalho de Conclusão de Curso (graduação) – Universidade de Santa Cruz do Sul, Departamento de Computação, Santa Cruz do Sul, 2016. Disponível em: <https://repositorio.unisc.br/jspui/bitstream/11624/2127/1/Eduardo%20Lohmann.pdf>. Acesso em: 20. dez 2021.

MEDEIROS, C., COSTA, J.A. F. Uma Comparação Empírica de Métodos de Redução de Dimensionalidade Aplicados a Visualização de Dados. **Learning & Non Linear Models**, vol. 6, n. 2, p. 81-110. 2008. Disponível em:<http://www.deti.ufc.br/~lnlm/papers/vol6-no2-art1.pdf>. Acesso em: 22 dez. 2021.

MENDES, J. C. **Agrupamento de dados e suas aplicações**.2017. Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Maranhão. São Luís, 2017. Disponível em: <https://monografias.ufma.br/jspui/bitstream/123456789/3570/1/JAKELSON-MENDES.pdf>. Acesso em: 21 dez.2021.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Ed. UFMG. 2005.

MOON, C. J.; WHATELEY, M. K. G; EVANS, A.M. **Introduction to mineral exploration**. 2 ed. New Delhi: Blackwell Publishing, 2006.

NEVES, S. A. das**. Técnicas de aprendizado de máquina aplicadas a classificação da qualidade de pavimentos asfálticos utilizando *smartphones***. 2018. Trabalho de Conclusão de Curso (graduação em Engenharia de Computação) – Universidade Federal de Ouro Preto, Instituto de Ciências Exatas e Aplicadas. João Monlevade, 2018. Disponível em: < https://www.monografias.ufop.br/bitstream/35400000/799/1/MONOGRAFIA_T%C3%A9cnicasAprendizadoM%C3%A1quina.pdf >. Acesso em: 21 dez.2021.

OLEA, R. A. **Geostatistical Glossary and Multilingual Dictionary.** New York: Oxford University Press, 1991. 177p.

OLIVEIRA, V. L. M. **Analytical customer relationship management in retailing supported by data mining techniques**. 2012.Thesis (Doctor of Industrial Engineering and Management) – Universidade do Porto, Faculdade de Engenharia. Porto,2012. Disponível em: < https://repositorio-aberto.up.pt/bitstream/10216/69283/1/000155270.pdf> Acesso em: 10 dez. 2021.

PASSOS, U. R. C. **Computação evolutiva e aprendizado de máquina aplicados ao apoio do diagnóstico da cardiopatia isquêmica**. 2014. Dissertação (Mestrado). – Universidade Cândido Mendes. Campos dos Goytacazes, 2014. Disponível em: <https://mpoic.ucam-campos.br/wp-content/uploads/2014/11/Ubiratan_Roberte_Cardoso_Passos.pdf> Acesso em: 08 dez. 2021.

PRASS, F. S. **Estudo comparativo entre algoritmos de análise de agrupamentos em *data mining***. Dissertação (Mestrado). – Universidade Federal de Santa Catarina, Santa Catarina, 2004. Disponível em: <http://repositorio.ufsc.br/xmlui/handle/123456789/87267> Acesso em: 08 dez. 2021.

PEREIRA, P. E. C. **Estimativa de recursos minerais e otimização de cava aplicados a um estudo de caso de uma mina de calcário**. Dissertação (Mestrado em Modelagem e Otimização) – Universidade Federal de Goiás, Catalão, 2017. Disponível em: <http://repositorio.bc.ufg.br/tede/handle/tede/7140>. Acesso em: 18 dez. 2021.

PINTO F. A. C.; DEUTSCH, C. V. Calculation of high resolution data spacing models. In Deutsch, J. L., **Geostatistics Lessons**. 2017. Disponível em: < http://www.geostatisticslessons.com/lessons/dataspacing> Acesso em 12 dez.2021.

RAMÍREZ, J. E. G. **Variabilidade espacial do parâmetro geomêcanico RQD no depósito mineral Animas-Peru.** Dissertação (Mestrado). – Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009. Disponível em: < https://www.maxwell.vrac.pucrio.br/colecao.php?strSecao=resultado&nrSeq=14482@2 >. Acesso em: 19 dez 2021.

R CORE TEAM (2016). **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri-SP: Malone, 2005.

RIVOIRARD, J. Concepts and methods of geostatistics. In: Bilodeau M., Meyer F.; Schmitt M. (eds.), **Space, Structure and Randomness**: Contributions in Honor of Georges Matheron in the Field of Geostatistics, Random Sets and Mathematical Morphology. New York: Springer, 2005.

ROKACH, L. Ensemble-based Classifiers. **Artificial Intelligence Review**, v. 33, n. 1–2, p. 1-39, 2010.

RUBIO, R. J. H. Otimização de parâmetros de krigagem baseada na minimização do erro absoluto e o erro quadrático. 2018. Tese (Doutorado em Engenharia). – Escola de Engenharia. Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e de Materiais, Porto Alegre: UFRGS, Porto Alegre, RS:2018. Disponível em; < https://lume.ufrgs.br/handle/10183/198144>. Acesso em: 03 jan. 2022.

SANCHES, M. K. **Aprendizado de máquina semi-supervisionado**: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados. 2003. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, 2003. Disponível em: <doi:10.11606/D.55.2003.tde-12102003-140536>. Acesso em: 10 dez.2021.

SARKAR, D.; BALI, R.; SHARMA, T. **Practical Machine Learning with Python**: A Problem-Solver's Guide to Building Real-World Intelligent Systems. California: Ed. Berkely, USA: Apress, 2017. Disponível em: < https://doi.org/10.1007/978-1-4842-3207-1_1. >. Acesso em: 10 dez.2021.

SILVA, L. C. e. **Aprendizado de máquina com treinamento continuado aplicado à previsão de demanda de curto prazo**: o caso do Restaurante Universitário da Universidade Federal de

Uberlândia. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Uberlândia, Uberlândia, 2019. Disponível em: < http://dx.doi.org/10.14393/ufu.di.2019.2001.> Acesso em: 10 dez.2021.

SILVA, L. N. de C. **Análise e síntese de estratégias de aprendizado para redes neurais artificiais**. 1998. Dissertação (Mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP, 1998. Disponível em: <http://www.repositorio.unicamp.br/handle/REPOSIP/259070>. Acesso em: 21 dez 2021.

SOUZA, L. E. **Estimativa de incertezas e sua aplicação na classificação de recursos minerais**. 2002. Dissertação (Mestrado em Engenharia) – Universidade Federal do Rio Grande do Sul, Escola de Engenharia. Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e de Materiais, Porto Alegre: UFRGS, 2002. Disponível em: < https://lume.ufrgs.br/handle/10183/77812 >. Acesso em: 21 dez 2021.

SOUZA, L. E., WEISS, A. L., COSTA, J.F.C.L., KOPPE, J. C. Impacto do agrupamento preferencial de amostras na inferência estatística: aplicações em mineração. **Rem**, Revista da Escola de Minas, v.54, p.257 -266, 2001 Disponível em: < https://doi.org/10.1590/S0370-44672001000400005. >. Acesso em: 01 dez 2021.

SOUZA, R. A. **Análise da influência da incerteza geológica no planejamento de lavra**. 2016. Dissertação (Mestrado em Engenharia Mineral) – Universidade Federal de Ouro Preto Escola de Minas, Ouro Preto, 2016. Disponível em: < http://www.repositorio.ufop.br/handle/123456789/7265>. Acesso em: 01 dez 2021.

SOUZA JUNIOR, C. T. de; **Análise de agrupamento**: o problema da identificação de línguas em textos por meio de bi-gramas.2018. Dissertação (Mestrado em Modelagem Computacional e Tecnologia Industrial) – SENAI CIMATEC, Salvador, 2018. Disponível em: <http://repositoriosenaiba.fieb.org.br/handle/fieb/891>. Acesso em: 21 dez 2021.

Stanford University. 2011. **Stanford Geostatistical Modeling Software (SGeMS)**. Disponível em: <http://sgems.sourceforge.net/.>. Acesso em: 21 dez 2021.

ULTSCH, A.; SIEMON, H.P. Kohonen's Self-Organizing Feature Maps for Exploratory Data Analysis. *In*: **Proceedings of International Neural Networks Conference (INNC)**, p. 305 - 308, 1990.

VAN HULLE, M. M. **Faithful Representation and Topographic Maps**: From Distortion to Information-Based Self-Organization. John Wiley e Sons, 2000.

VESANTO, J.; HIMBERG, J.; ALHONIEMI, E.; PARHANKANGAS, J. "**SOM Toolbox for Matlab 5**": report A57, April 2000. Libella Oy: Finland: SOM Toolbox Team, Helsinki University of Technology, 2000b. 59 p. Disponível em:< http://www.cis.hut.fi/projects/somtoolbox/download/.> Acesso em: 11 dez 2021.

VIERA JUNIOR, V. **Classificação de dados usando técnicas de Datamining e aprendizagem de máquina**. Trabalho de Conclusão de Curso (graduação) –Universidade Federal do Maranhão, São Luís, 2013. Disponível em:< https://monografias.ufma.br/jspui/bitstream/123456789/3254/1/ValdecyJunior.pdf> Acesso em: 11 dez 2021.

TODT, Viviane. **Avaliação do Desempenho de Classificadores Neurais para Aplicações em Sensoriamento Remoto**.1998. Dissertação (mestrado). – UFRGS, Porto Alegre: UFRGS, 1998. Disponível em: < https://lume.ufrgs.br/handle/10183/2062 > Acesso em: 01 dez 2021.

WEISS, S. M.; KULIKOWSKI C.A., **Computer Systems that Learn**: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems, Morgan Kaufmann Publishing, San Mateo, 1991.

XAVIER, Vinicius Layter. **Resolução do problema de agrupamento segundo o critério de minimização da soma de distancias**. 2012.Dissertação (mestrado). – Universidade Federal do Rio de Janeiro, Rio de Janeiro; 2012. Disponível em: < http://objdig.ufrj.br/60/teses/coppe_m/ViniciusLayterXavier.pdf> Acesso em: 11 dez 2021.

YAMAMOTO, J. K. **Avaliação e classificação de reservas minerais**. São Paulo: EdUSP, 2001.

YAMAMOTO, J.K.; LANDIN, P.M.B. **Geoestatística**: conceitos e aplicações. São Paulo: Oficina de Textos: Online Book, 2015.

ZUCHINI M. H. **Aplicação de mapas auto-organizáveis em mineração de dados e recuperação de informação**. 2003. Dissertação (Mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP, 2003. Disponível em: < https://www.dca.fee.unicamp.br/~vonzuben/theses/zuchini_mest/zuchini_mest.pdf> Acesso em: 11 dez 2021.