

**UMA METODOLOGIA PARA AGRUPAMENTO DE AMOSTRAS DE ÁGUA DA
REGIÃO AMAZÔNICA**

**A METHODOLOGY FOR GROUPING WATER SAMPLES OF THE AMAZON
REGION**

Autor¹

Autor²

Resumo: A análise dos índices pluviométricos é de extrema importância pois estão relacionados ao desenvolvimento agrícola. Na agroindústria, por exemplo, condições climáticas que viabilizam o plantio de uma lavoura são previstas utilizando Redes Neurais Artificiais (RNAs) e oferecem taxa de acerto de 96% [3]. Predizer o ciclo de um fenômeno reduz incertezas e permite aplicar capital humano e econômico durante o período essencial. O objetivo deste trabalho é fazer um mapeamento dos índices pluviométricos da Região Amazônica e oferecer uma metodologia para identificar o melhor período de plantio. A coleta de dados e o agrupamento proposto denominado utiliza apenas 10 características da água como: o pH, a turbidez, a temperatura, dentre outros. Os dados foram coletados na Estação de Tratamento de Água (ETA) da cidade de Guajará-Mirim, captadora do Rio Madeira Mamoré localizando que banha os estados de Rondônia e Amazonas. Após o uso da abordagem verificou-se que as amostras de água podem ser agrupadas em três classes com SSE (Erro dos Quadrados das Distâncias Somados) de 50, 45% e 47, 61% menos atributos que outras metodologias.

Palavras-chave: Agrupamento. Agricultura de Precisão. Características Físico-químicas.

Abstract: The study of pluviometric indices are extremely important as they are related to agricultural development. In agro-industry, for example, climatic conditions that make planting viable are predicted using Artificial Neural Networks (ANNs) with an accuracy of 96% [3]. Predicting the cycle of a phenomenon reduces uncertainties and allows us to apply human and economic capital during the essential period. The objective of this work is to map the pluviometric indices of the North region and offer a methodology to identify the best planting period. The proposed cluster uses only 10 water features such as pH, turbidity and temperature. The data were collected at the Water Treatment Station (ETA) of the city of Guajará-Mirim, capturing the Madeira River Mamoré, located in the North region that bathes the states of Rondônia and Amazonas. After using the Water-Clu approach we found that water samples can be grouped into 3 classes with SSE of 50.45% and uses 47.61% less attributes than other approaches.

Keywords: Clustering. Precision Agriculture. Physical and chemical characteristics.

¹ Instituição
{email@email}

² Instituição
{email@email}

INTRODUÇÃO

A mineração de dados outorga vantagem competitiva à empresas quando estas utilizam a informação para tomar decisões estratégicas. Os dados viabilizam a aplicação de técnicas de Inteligência Artificial para a construção de modelos de classificação ou regressão. Na agroindústria, por exemplo, condições climáticas que inviabilizam o plantio de uma lavoura são previstas utilizando Redes Neurais Artificiais (RNAs): *Perceptrons de Múltiplas Camadas* (MLP) e oferecem taxa de acerto de 96% [3].

Predizer o ciclo de um fenômeno reduz incertezas e permite aplicar capital humano e econômico durante o período essencial. Por exemplo, as épocas de plantio de cana-de-açúcar são modeladas como Séries Temporais (ST) e algoritmos de Aprendizagem de Máquina (AM) procuram de padrões nas sequências (*motifs*) para identificar o período ideal de plantio [2].

O bater de asas de insetos vetores de doenças ou aqueles que devastam lavouras também são alvo da mineração de dados. Verificou-se um padrão no bater de asas dos insetos sendo possível identificar um animal pelo “som” emitido durante seu voo. A partir dos dados, interpretados como séries temporais, são construídos modelos para cada classe (espécie de animal) do conjunto. Os modelos atingem acurácia de 87, 33% para identificação automática de insetos como o *Psychodidae diptera* transmissor de *leishimaniose*, o *Anopheles gambiae* vetor da malária e o besouro *Cotinis mutabilis* que se alimenta de pétalas [1].

Um aspecto importante na agroindústria é a descoberta de regiões hidrográficas homogêneas [4]. Para isso o autor utiliza-se do algoritmo de agrupamento Árvore Geradora Mínima para analisar características físicas, químicas e toxicológicas da água com o objetivo de formar grupos semelhantes. A metodologia oferecida pelo autor proporciona um melhor conhecimento dos corpos d’água permitindo, por exemplo, reduzir a quantidade de pontos a serem analisados em programas de monitoramento. Semelhante ao trabalho de [4], a metodologia proposta neste trabalho é capaz de identificar grupos a partir da análise das propriedades da água, porém com uma quantidade menor de atributos.

Tem-se como hipótese que apenas o uso de características da água como temperatura, pH e turbidez permite agrupar amostras e caracterizar sua etapa no ciclo das águas da região Amazônica. O objetivo deste trabalho é fortalecer arranjos produtivos locais baseados principalmente em atividades agrícolas por aumentar a produtividade do solo e incitar o uso eficiente dos recursos hídricos. Ao aplicar esta abordagem na Região Amazônica em dados coletados na Estação de Tratamento de Água (ETA) localizada em Rondônia, no Rio Mamoré

têm-se como resultado a identificação de 3 grupos, semelhante aos períodos chuvosos da Região Amazônica.

A Seção 2 apresenta a metodologia proposta, a Seção 3 demonstra os experimentos realizados e resultados obtidos, a Seção 4 as Conclusões e a Seção 5 os Agradecimentos.

2. PROPOSTA

Este trabalho propõe uma metodologia de agrupamento de amostras de água (Fig.1). Utilizam-se como características apenas três propriedades da água: o pH, a turbidez e a temperatura. Estas características físico-químicas da água sustentam esta metodologia para distinguir amostras e formar grupos de maneira não-supervisionada.

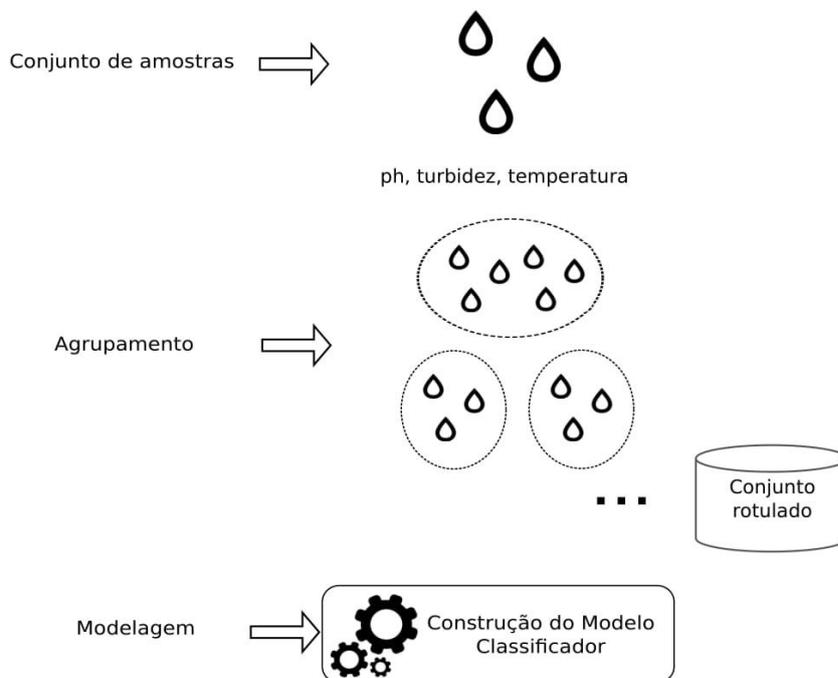


Figura 1. Metodologia de agrupamento.

2.1 AGRUPAMENTO

A metodologia utiliza o algoritmo de agrupamento SimpleKmeans para identificar os grupos. A partir dos dados normalizados e de um número de centroides definidos previamente, o algoritmo SimpleKmeans calcula a Distância Euclidiana Média no espaço de características para construir uma matriz de distâncias entre amostras e os centroides. O Algoritmo itera-se e os valores das distâncias são recalculados. A cada iteração o centroide

que está ‘mais perto’ da amostra a incorpora-a ao seu grupo (do inglês: *cluster*) [5]. O posicionamento dos centroides é quantificado em função da minimização do Erro dos Quadrados das Distâncias Somados (*SSE*), onde P é o ponto no espaço, m_i é um centroide do grupo C_i conforme Equação 1.

$$SSE = \sum_{i=1}^K \sum_{p \in C_i} (p - m_i)^2$$

Equação 1. Erro dos Quadrados das Distâncias Somados (*SSE*).

Têm-se como objetivo formar o menor número de grupos possíveis que expressem a realidade dos dados. Para tal, utiliza-se o método heurístico *Elbow* para relacionar o *SSE* e o número de centroides utilizados. Ao analisar o gráfico oriundo do método *Elbow* deve-se procurar o *Elbow Effect*, ou seja, uma queda abrupta do *SSE* seguida de uma estabilização [5].

3. EXPERIMENTO

Nesta seção é apresentada a configuração e o resultado da aplicação da metodologia de agrupamento em dados oriundos da análise de água. Verifica-se que a técnica é eficiente e viável por alcançar formar 3 grupos de dados utilizando como métrica o método *Elbow Effect* com *SSE* de 50, 45%.

3.1 DATASET

A partir de parceria realizada entre o Instituto Federal de Rondônia (IFRO) e a Companhia de Águas e Esgotos do Estado de Rondônia (CAERD) foram compartilhados relatórios técnicos de diferentes etapas do processo de tratamento de água. A primeira etapa do tratamento consiste na aferição da temperatura da água e do índice de turbidez ao entrar na Estação de Tratamento de Água (ETA). A partir desta amostra realiza-se a análise química e sabe-se a quantidade necessária de compostos químicos para tratar a água. Coletou-se outras amostras de água até a sua entrada na rede de distribuição para as residências.

Ao todo se utilizam 10 características em diferentes estágios do processo de tratamento da água como: temperatura máxima e mínima, pH máximo e mínimo medidos antes e depois da aplicação do sulfato para floculação e a turbidez máxima e mínima coletada no início e término do processo.

Foram utilizadas ao todo 366 amostras normalizadas coletadas em Guajará-Mirim/RO de 1 de janeiro de 2016 até 31 de dezembro de 2016. Cada amostra de água representa a média diária de doze amostras coletadas de duas em duas horas durante o ano (24 horas por dia).

Tabela 1. Experimento de adicionar um centroide para computo do SSE.

Número de Centroides	SSE	Taxa de Redução do SSE
2	75,53	–
3	50,45	–33,21%
4	42,6	–15,56%
5	38,62	–9,34%
6	36,09	–6,55%
7	33,74	–6,51%
8	31,2	–7,53%
9	28,51	–8,62%
10	26,83	–5,89%

3.2 CONFIGURAÇÃO E RESULTADOS

A metodologia de agrupamento tem-se como métrica de aferição da qualidade dos grupos formados a taxa obtida ao incrementar um centroide relacionando-o com a redução do SSE. Utilizando os dados coletados da ETA em 2016 verifica-se na Tabela 1 que a partir da utilização de 3 centroides a taxa de redução do SSE estabiliza-se em média a 8,57% ao acrescentar um novo centroide.

Uma forma de verificar a tendência de incrementar o número de centroides e o SSE não variar é ao analisar o Gráfico 1 em busca do *Elbow Effect* (Efeito Cotovelo) confirmado ao utilizar 3 centroides. A partir destes resultados infere-se que as amostras de águas podem ser agrupadas em 3 grupos.

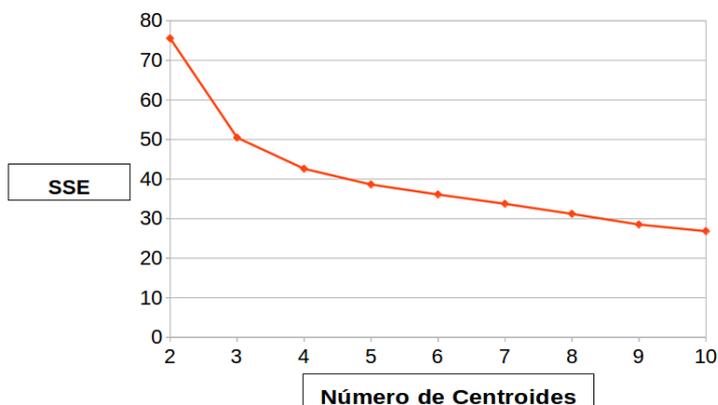


Gráfico 1. Identificação visual do *Elbow Effect* ao relacionar número de centroides e a redução do SSE.

A ferramenta de processamento utilizada (Weka 3.6.11) oferece a opção de visualização de atribuição das amostras aos respectivos clusters apresentada na Figura 3. No eixo das ordenadas verifica-se os grupos formados e no eixo das ordenadas a representação das amostras considerando a distância atribuída ao seu centroide.

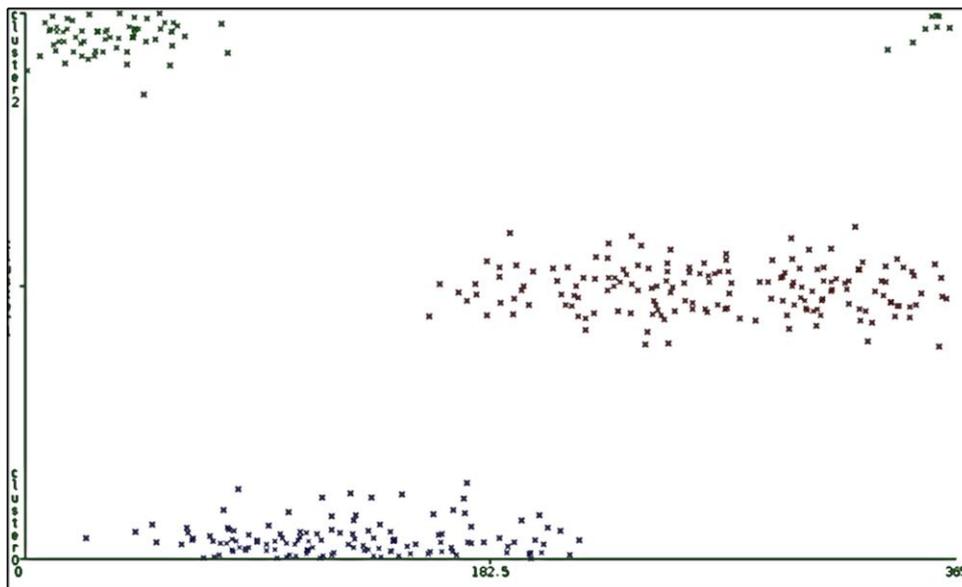


Figura 3. Atribuição das amostras aos clusters.

Ao analisar a Figura 3 têm-se amostras posicionadas no canto superior esquerdo representando àquelas coletadas no começo do ano de 2016. Nota-se no quadrante inferior amostras coletadas no 2º trimestre de 2016. E por fim, no quadrante superior direito outras amostras que apresentem características físico-químicas do final do ano de 2016 (período chuvoso na Região Amazônica).

CONCLUSÃO

Verifica-se a partir dos experimentos realizados que a abordagem pode ser utilizada para inferir grupos e posteriormente rótulos para as amostras, alvo de trabalho futuro para produção de um classificador automático de amostras de água. Comparando com o trabalho relacionado [4] que se vale de outras características de amostras de água, esta abordagem reduz a quantidade de características utilizadas em aproximadamente 47,61% e obter resultados semelhantes.

Realizando uma análise temporal das amostras, têm-se como trabalho futuro a investigação das amostras coletadas no período de estiagem na Região Amazônica representadas no centro da Figura 3 serem em maior número.

AGRADECIMENTOS

Agradecemos ao Instituto Federal de Rondônia (IFRO), ao Instituto Federal do Acre (IFAC) e a Companhia de Águas e Esgotos do Estado de Rondônia (CAERD).

REFERENCIAS

- [1] D. F. Silva, V. De Souza, G. E. Batista, E. Keogh, and D. P. Ellis. **Applying machine learning and audio analysis techniques to insect recognition in intelligent traps**. Machine Learning and Applications (ICMLA), 2013, volume 1, p. 99–104. IEEE, 2013.
- [2] D. Y. Chino, R. R. Goncalves, L. A. Romani, C. Traina, and A. J. Traina. **Discovering frequent patterns on agrometeorological data with triemotif**. ICEIS, p. 91–107. Springer, 2014.
- [3] E. R. Coutinho, R. M. Silva, and A. R. S. Delgado. **Using computational intelligence technique for the meteorological data prediction**. Revista Brasileira de Meteorologia, p. 24–36, 2016.
- [4] L. Bertholdo, L. C. Júnior, G. de Aragão Umbuzeiro, and C. G. da Silva. **Mineração de dados de qualidade de água para agrupamento de pontos de amostragem usados no monitoramento de recursos hídricos**. WCAMA-CSBC, p. 1036–1046, 2013
- [5] T. M. Kodinariya and P. R. Makwana. **Review on determining number of cluster in k-means cluste-ring**. International Journal, p. 90–95, 2013.