Research, Society and Development, v. 11, n. 14, e140111436235, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i14.36235

Identificação de padrões morfoestruturais utilizando Clustering Large Applications, um estudo de caso no Quadrilátero Ferrífero

Identification of morphostructural domains using Clustering Large Applications, a case study in Quadrilátero Ferrífero

Identificación de dominios morfoestructurales mediante Clustering Large Applications, un caso de estudio en Quadrilátero Ferrífero

Recebido: 07/10/2022 | Revisado: 14/10/2022 | Aceitado: 17/10/2022 | Publicado: 22/10/2022

Naim Khalil Ayache
ORCID: https://orcid.org/0000-0003-3834-6341
Centro Federal de Educação Tecnológica de Minas Gerais, Brasil
E-mail: naimayache98@gmail.com
Allan Erlikhman Medeiros Santos
ORCID: https://orcid.org/0000-0003-4302-3897
Centro Federal de Educação Tecnológica de Minas Gerais, Brasil
E-mail: allanerlikhman@cefetmg.br
Francisco de Castro Valente Neto
ORCID: https://orcid.org/0000-0002-9652-2588
Centro Federal de Educação Tecnológica de Minas Gerais, Brasil
E-mail: fcvn.araxa@cefetmg.br
Denise de Fátima Santos da Silva
ORCID: https://orcid.org/0000-0002-9695-2449
Universidade Federal de Minas Gerais, Brasil
E-mail: denisefss@yahoo.com.br

Resumo

Dentre as etapas de um projeto de mineração destaca-se a pesquisa mineral, com objetivos de identificar, estudar e avaliar os depósitos minerais. Nesta etapa específica ocorre a transformação dos recursos minerais inferidos, em indicados e por fim medidos, e caso seja viável sua explotação, em reservas minerais prováveis e/ou provadas. A descoberta destas reservas é marco impactante para o desenvolvimento industrial, tecnológico e econômico de uma sociedade. Este artigo tem como objetivo principal apresentar a utilização de uma técnica de machine learning para identificação de estruturas de particular interesse geológico, a partir de imagens de satélite. A técnica aplicada foi o Clustering Large Applications (CLARA) que é um algoritmo não-supervisionado para agrupamento de dados, com alta performance em banco de dados massivos. A área utilizada como estudo de caso foi o Quadrilátero Ferrífero, uma das maiores províncias minerais do planeta, localizada no estado de Minas Gerais, Brasil. Os resultados do modelo CLARA permitiram delinear todas as feições que formam o Quadrilátero Ferrífero. Neste contexto acredita-se que esta possa ser uma boa ferramenta para seleção de alvos exploratórios reduzindo incerteza e risco aos investidores. O que propicia não somente a atração de novas empresas para pesquisa mineral, além da ampliação das reservas dos recursos minerais brasileiros.

Palavras-chave: CLARA; Análise de agrupamento; Pesquisa mineral; Quadrilátero ferrífero.

Abstract

Among the stages of a mining project, mineral research stands out, with the objective of identifying, studying and evaluating mineral deposits. In this specific stage, the inferred mineral resources are transformed into indicated and finally measured, and if their exploitation is feasible, into probable and/or proven mineral reserves. The discovery of these reserves is an impacting milestone for the industrial, technological and economic development of a society. The main objective of this article is to present the use of a machine learning technique to identify structures of particular geological interest, from satellite images. The technique applied was the Clustering Large Applications (CLARA) which is an unsupervised algorithm for clustering data, with high performance in massive databases. The area used as a case study was the Quadrilátero Ferrífero, one of the largest mineral provinces on the planet, located in the state of Minas Gerais, Brazil. The results of the CLARA model allowed the delineation of all the features that form the Quadrilátero Ferrífero. In this context, it is believed that this can be a good tool for selecting exploratory targets, reducing uncertainty and risk to investors. This not only attracts new companies for mineral research, but also expands the reserves of Brazilian mineral resources.

Keywords: CLARA; Cluster analysis; Mineral search; Quadrilátero ferrífero.

Resumen
Entre las etapas de un proyecto minero se destaca la investigación minera, con el objetivo de identificar, estudiar y evaluar yacimientos minerales. En esta etapa específica, los recursos minerales inferidos se transforman en indicados y finalmente medidos, y si su explotación es factible, en reservas minerales probables y/o probadas. El descubrimiento de estas reservas es un hito impactante para el desarrollo industrial, tecnológico y económico de una sociedad. El objetivo principal de este artículo es presentar el uso de una técnica de aprendizaje automático para identificar estructuras de particular interés geológico, a partir de imágenes satelitales. La técnica aplicada fue el Clustering Large Applications (CLARA) que es un algoritmo no supervisado para el agrupamiento de datos, con alto rendimiento en bases de datos masivas. El área utilizada como caso de estudio fue el Quadrilátero Ferrífero, una de las mayores provincias mineras del planeta, ubicada en el estado de Minas Gerais, Brasil. Los resultados del modelo CLARA permitieron delimitar todos los rasgos que forman el Quadrilátero Ferrífero. En este contexto, se cree que esta puede ser una buena herramienta para seleccionar objetivos exploratorios, reduciendo la incertidumbre y el riesgo para los inversores. Esto no solo atrae nuevas empresas para la investigación minera, sino que también amplía las reservas de recursos minerales brasileños.
Palabras clave: CLARA; Análisis de conglomerados; Búsqueda de minerales; Quadrilátero ferrífero.

1. Introdução

O sensoriamento remoto é a ciência de adquirir, processar e interpretar imagens e dados relacionados, adquiridos de aeronaves e satélites, que registram a interação entre matéria e energia eletromagnética (Sabins, 1997). Melhorias tecnológicas em sensoriamento remoto multiespectral proporcionam oportunidades para adquirir volumes cada vez maiores de informações, bem como fazer mais tipos de informações geológicas potencialmente deriváveis (Cloutis,1996).

No entanto, o aumento do volume de dados pode se tornar um problema em função da dificuldade de manipular e organizar essas informações de forma clara e rápida, otimizando a interpretação desses dados. Dessa forma, surge a necessidade da utilização de novas técnicas de análise estatística de dados com o objetivo de facilitar o processamento desses dados. Uma grande ferramenta utilizada para este tipo de problema são as técnicas de machine learning.

As técnicas de machine learning (ML), ou aprendizado de máquina, constam de uma abordagem empírica eficaz tanto para problemas de regressão quanto problemas de classificação (supervisionada ou não supervisionada) de sistemas não lineares. Esses sistemas podem ser massivamente multivariados envolvendo algumas ou literalmente milhares de variáveis. Em ML, um abrangente "conjunto de dados de treinamento" de exemplos é construído cobrindo o máximo possível do espaço de parâmetros do sistema, como enfatizado por Lary et al. (2016).

As técnicas de aprendizado de máquina são amplamente utilizadas em pesquisas em diversas áreas do conhecimento, especialmente naquelas com o objetivo de encontrar similaridade em grandes conjuntos de dados. Dentre alguns estudos utilizando técnicas de aprendizado de máquina e/ou técnicas de análise quantitativas, podemos citar o estudo de Alves et al. (2020) que realizaram a técnica de agrupamento por similaridade dos estados brasileiros com a finalidade de observar as medidas de combate a COVID-19 realizadas em cada um desse grupos pelo método não-hierárquico k-means considerando os coeficientes epidemiológicos como incidência, prevalência e letalidade. Reis et al. (2021) sistematizaram a epiderme foliar de 10 espécies florestais amazônicas, com ferramentas auxiliar a taxonomia, aplicando: o teste de Tukey, ANOVA, Análise de agrupamento e Análise de Componentes Principais. Posteriorente, Nascimento et al. (2022) compararam o desempenho de métodos de agrupamento formado por vários IDH dos 27 estados brasileiros. Ainda,

Lary et al. (2016) utilizaram duas técnicas distintas, uma supervisionada e outra não supervisionada, para estudar a eficiência que estas técnicas tem na resolução de problemas no campo do sensoriamento remoto. Seguindo essa mesma linha, o objetivo desta pesquisa foi estudar uma maneira de classificar imagens de satélite de forma a agrupar as regiões em grupos com maior similaridade. Deste jeito, facilitando a identificação de estruturas geomorfológicas e o seu comportamento estrutural superficial.

Para isso, foi utilizado a técnica CLARA (Clustering Large Applications) adaptado para interpretar imagens do satélite da missão Sentinel-2 da União Europeia. As imagens utilizadas compreendem a região do Quadrilátero Ferrífero (QF)

no estado de Minas Gerais, Brasil. A região do QF é conhecida mundialmente pela sua enorme potencialidade mineral e complexidade tectono-estrutural. A qualidade e o volume das reservas de minério de ferro, além da posição geográfica, logística e do desenvolvimento tecnológico presente no QF possibilitaram que a região tenha destaque na oferta de quase 250 Mt/a de produtos de minérios de ferro, representando aproximadamente 11% da produção mundial atual (USGS, 2017; Lima et al., 2020).

2. Metodologia

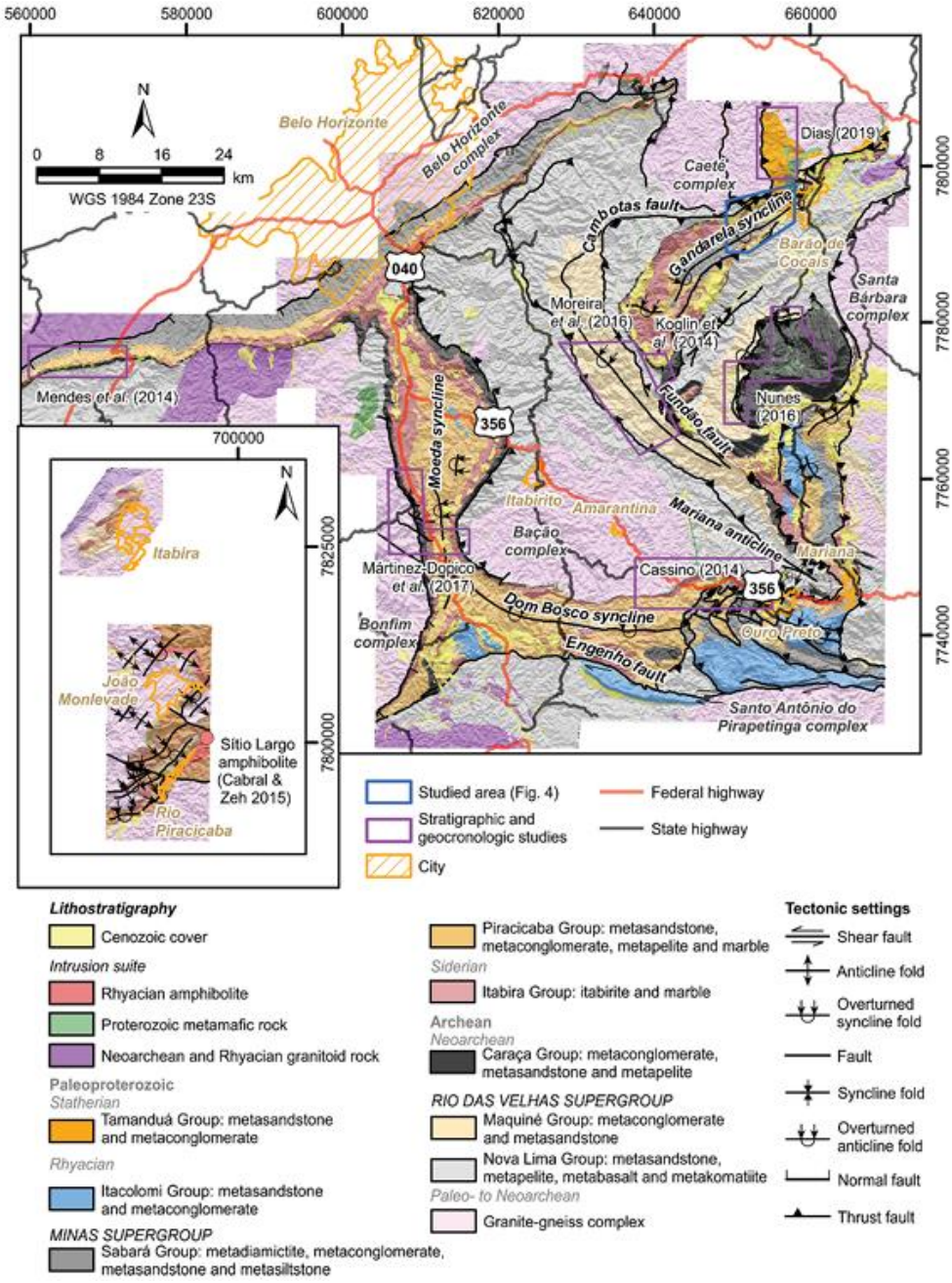
2.1 Área de estudo: Quadrilátero Ferrífero

O Quadrilátero Ferrífero (QF) localizado no sudeste do Brasil, ocupa 4% do estado de Minas Gerais, e apresenta como principal bioma a Mata Atlântica. Os depósitos de minério de ferro do Quadrilátero Ferrífero estão hospedados na Formação Cauê, parte do Supergrupo Minas. O Supergrupo Minas foi inicialmente depositado há aproximadamente 2,5 Ga atrás da borda do cráton do São Francisco (Dorr II, 1969).

O minério de ferro da região é do tipo conhecido como itabirítico. Sendo que o itabirito é uma variante metamorfoseada da formação de ferro bandado em que as bandas originais de quartzo e jaspe foram recristalizadas em grãos de quartzo distinguíveis macroscopicamente. De acordo com Cabral et al. (2012), os minerais de ferro (hematita e magnetita) estão tipicamente presentes em bandas finas. Esse material normalmente produz um minério de ferro de alto teor, já que impurezas como enxofre ou fosfato foram removidas durante os processos metamórficos.

A estruturação geológica da região pode ser vista na Figura 1, nela é possível ver os dois principais Supergrupos da região, os Supergrupos Minas e Rio das Velhas, e a litoestratigrafia do QF. O conhecimento da distribuição espacial desses grupos será importante nesta pesquisa pois espera-se que a utilização da CLARA consiga identificar essas morfologias na análise das imagens de satélite.

Figura 1. Mapa geológico do Quadrilátero Ferrífero.



Fonte: Dutra, Martins & Lana, (2019) baseado em Lobato et al. (2005).

2.2 Sentinel-2

Sentinel-2 é uma missão europeia para a captura de imagens multiespectrais de alta resolução e ampla faixa. A missão possui dois satélites gêmeos voando na mesma órbita, mas em fase de 180 °, que foram projetados para dar uma alta frequência de revisita de 5 dias no Equador. O Sentinel-2 carrega uma carga útil de instrumento óptico que irá amostrar 13 bandas espectrais: quatro bandas a 10m, seis bandas a 20m e três bandas a 60m de resolução espacial. A largura da faixa orbital será de 290 km (Engesat, 2015; ESA, 2021).

As imagens de satélite são compostas por bandas, estas que são capturadas pelo equipamento ótico instalado no satélite. Cada banda apresenta suas características em relação a um comprimento de onda especifico como pode ser visto na Tabela 1.

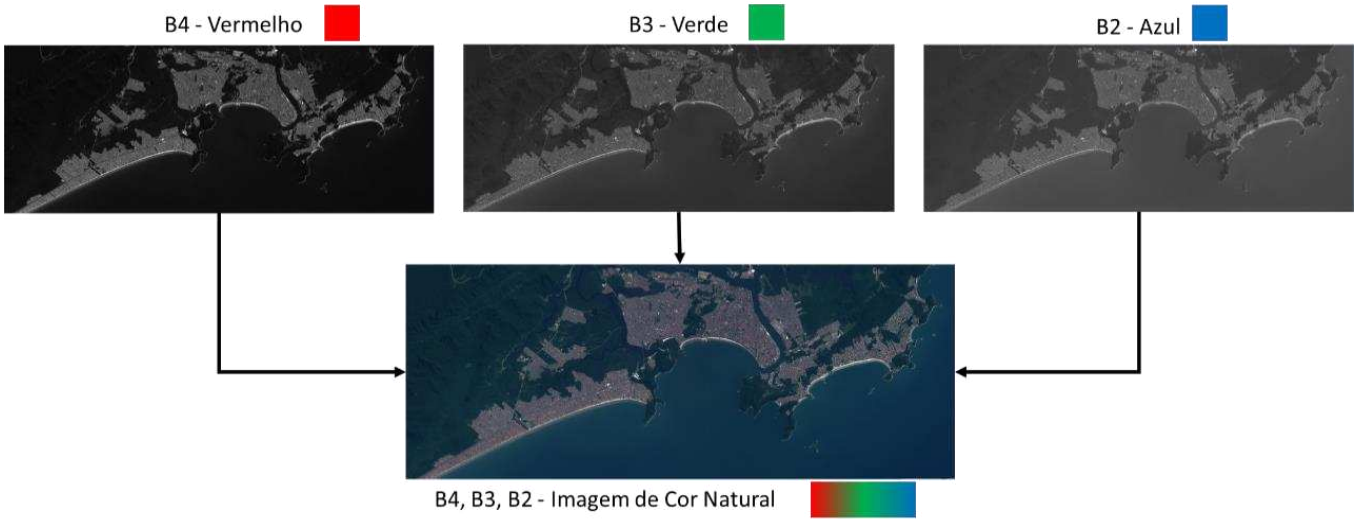
Tabela 1. Descrição das bandas da missão Sentinel-2.

Nome	Descrição	Resolução	Comprimento de onda
B1	Aerossol costeiro	60 metros	443.9nm (S2A) / 442.3nm (S2B)
B2	Azul	10 metros	496.6nm (S2A) / 492.1nm (S2B)
B3	Verde	10 metros	560nm (S2A) / 559nm (S2B)
B4	Vermelho	10 metros	664.5nm (S2A) / 665nm (S2B)
B5	Vermelho 1	20 metros	703.9nm (S2A) / 703.8nm (S2B)
B6	Vermelho 2	20 metros	740.2nm (S2A) / 739.1nm (S2B)
B7	Vermelho 3	20 metros	782.5nm (S2A) / 779.7nm (S2B)
B8	NIR	10 metros	835.1nm (S2A) / 833nm (S2B)
B8A	Vermelho 4	20 metros	864.8nm (S2A) / 864nm (S2B)
B9	Vapor d'água	60 metros	945nm (S2A) / 943.2nm (S2B)
B10	Cirrus	60 metros	1373.5nm (S2A) / 1376.9nm (S2B)
B11	SWIR 1	20 metros	1613.7nm (S2A) / 1610.4nm (S2B)
B12	SWIR 2	20 metros	2202.4nm (S2A) / 2185.7nm (S2B)

Fonte: Engesat (2015); The European Space Agency (2021).

Com a seleção das bandas é possível juntá-las para formar diferentes imagens. Diferentes combinações podem gerar imagens com diferentes finalidades. Um exemplo comum de combinação de bandas pode ser visto na Figura 2. A partir dela é possível ver as combinações das bandas 4, 3 e 2 do Sentinel-2 para formar uma imagem de cor natural da região litoral do Estado de São Paulo, na altura da cidade de Santos.

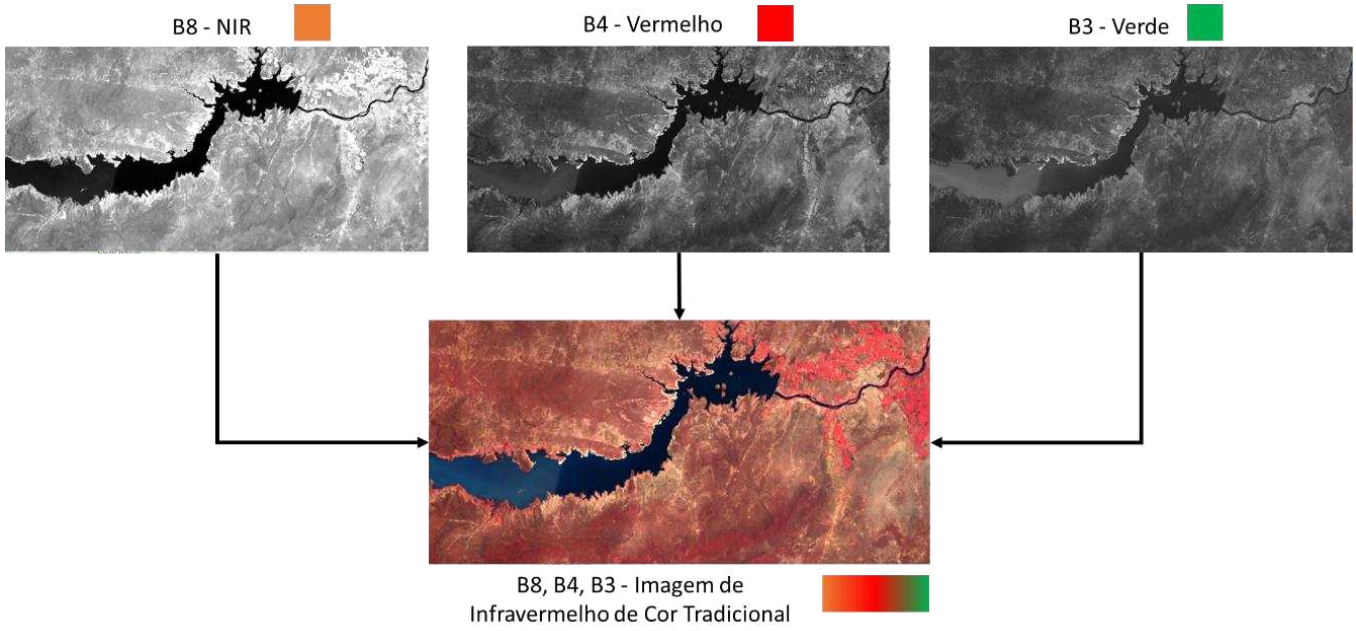
Figura 2. Representação da junção de bandas do Sentinel-2 para formação de imagem de cor natural do litoral do Estado de São Paulo.



Fonte: Autores (2022).

Outra combinação de bandas tradicional é a Infravermelho de Cor Tradicional (B8, B4 e B3). Essa combinação enfatiza a diferença entre regiões com mais vegetação. Na Figura 3 é possível ver um exemplo na região do Rio São Francisco à leste de Sobradinho no Estado de Minas Gerais, a vegetação aparece em vermelho, com vegetação mais densas apresentando cores de tonalidades mais vibrantes. Por acentuar bem essa diferença na vegetação, essa combinação de bandas foi escolhida para compor o banco de dados usado na CLARA.

Figura 3. Imagem de Infravermelho de Cor Tradicional da região de Sobradinho MG.



Fonte: Autores (2022).

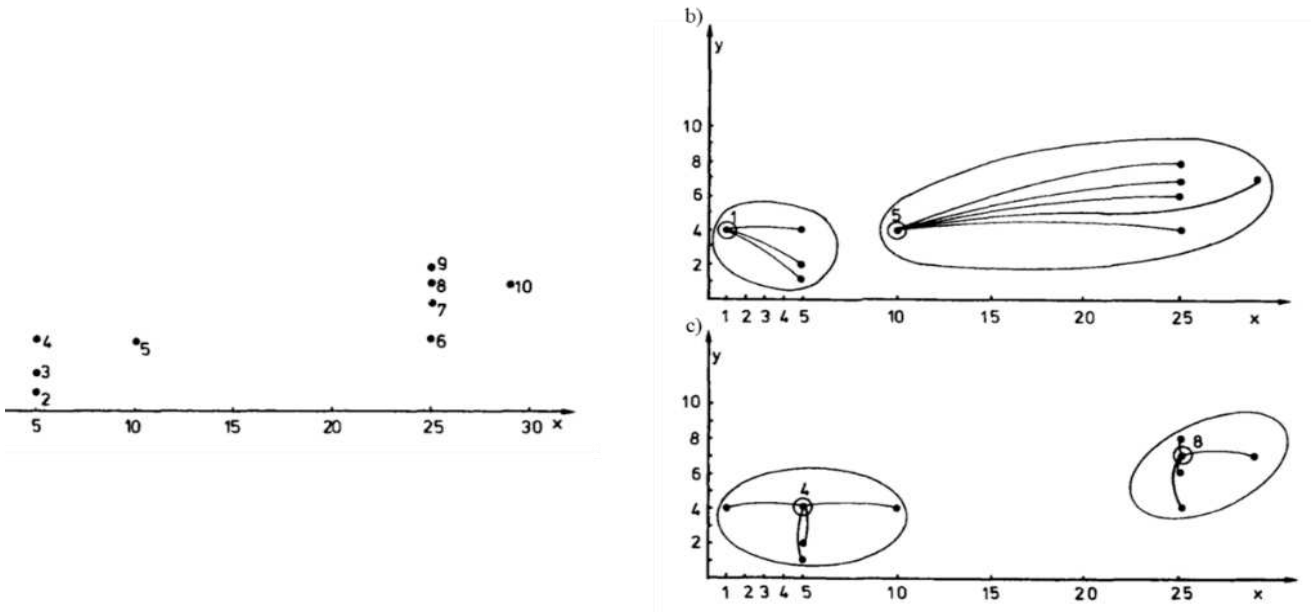
2.3 Clustering Large Applications (CLARA)

Para entender corretamente o que é o programa CLARA, é preciso visitar alguns conceitos mais básicos sobre o agrupamento de dados e como esta técnica funciona. Primordialmente, a utilização de técnicas de agrupamento (clustering) tem como objetivo particionar um conjunto de objetos ou amostras em k grupos, juntando esses indivíduos clusters que apresentam um alto grau de similaridade, enquanto objetos pertencentes a clusters diferentes são tão diferentes quanto possível (Kaufman et al., 1990b).

A técnica básica utilizada para criar o programa CLARA é o Partitioning Around Medoids, mais conhecido como PAM. Como o nome diz, o algoritmo usado no programa PAM é baseado na busca por k objetos representativos entre as amostras do conjunto de dados. Esses objetos devem representar vários aspectos da estrutura dos dados. No algoritmo PAM, os objetos representativos são os chamados medoides dos clusters (Kaufman & Rousseeuw, 1987). Depois de encontrar um conjunto de k objetos representativos, os k clusters são construídos atribuindo cada objeto do conjunto de dados para o objeto representativo mais próximo. Os k medoides escolhidos aleatoriamente entre as amostras do banco de dados são submetidas ao programa DAISY (Kaufman et al., 1990a,) para determinar a dissimilaridade de todas as amostras em relação aos medoides.

Um exemplo descrito por Kaufman et al., (1990b) pode ser visto na Figura 4ª. A partir dela é possível ver um conjunto de dados com duas dimensões criados pelos autores. Esse banco de dados é submetido ao programa PAM e dois pares distintos de medoides são escolhidos, os pares 1 e 5, Figura 4b, e os pares 4 e 8, Figura 4c. Com esses medoides, os autores calcularam usando o programa DAISY, a dissimilaridade desses pontos com todas as demais amostras. Utilizando as menores dissimilaridades calculadas, a dissimilaridade média foi obtida e comparada entre os dois pares. A média mínima para o par 1 e 5 foi de 9.37 enquanto que a do par 4 e 8 foi de 2.30. Sendo assim, a segunda seleção de medoides alcançou um melhor resultado em agrupar as amostras em clusters mais semelhantes (Kaufman et al., 1990b).

Figura 4. a) Conjunto de dados propostos por Kaufman, Leonard & Peter 1990b; **b)** Agrupamento do par de medoides 1 e 5 e **c)** Agrupamento do par de medoides 4 e 8.



Fonte: Adaptado de Kaufman et al., (1990b).

Por mais eficaz que seja o agrupamento de dados resultante do PAM, este possuí uma certa dificuldade no tratamento de conjunto de dados muito grandes. Desta forma Kaufman et al., (1990c) desenvolveram o CLARA justamente para adaptar as capacidades de clustering do PAM à conjuntos mais extensos.

O agrupamento de um conjunto de objetos com CLARA é realizado em duas etapas. Primeiro, uma amostra é retirada do conjunto de objetos e agrupada em k subconjuntos (medoides) mesma forma que o PAM. Então, cada objeto não pertencente à amostra é atribuído ao mais próximo dos k objetos representativos.

Isso produz um agrupamento de todo o conjunto de dados. Uma medida da qualidade deste agrupamento é obtida calculando a distância media entre cada objeto do conjunto de dados e seu objeto representativo. Depois que cinco amostras foram retiradas e agrupadas, é selecionado aquele para no qual a menor distância média foi obtida.

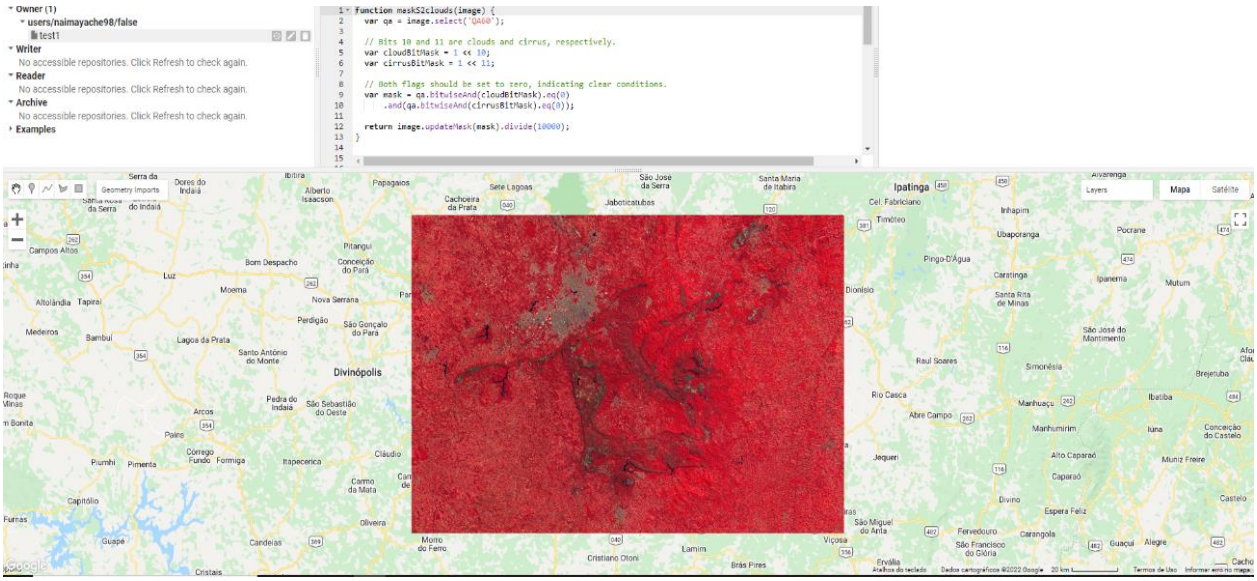
Essa técnica reduz drasticamente o processamento necessário para o processamento de extensos conjuntos de dados. Como imagens de satélite possuem milhares de amostras, esta técnica se encaixa de forma satisfatória para analisar este tipo de problema.

2.4 Considerações específicas da metodologia

Para a obtenção dos dados, tratamento e implementação das técnicas foram utilizados dois softwares distintos. O primeiro foi o Google Earth Engine (Gorelick et al., 2017). O Earth Engine consiste em um catálogo de dados pronto para análise co-localizado com uma computação intrinsecamente paralela de alto desempenho serviço. Ele é acessado e controlado por meio de uma interface de programação de aplicativos (API) acessível pela Internet e um ambiente de desenvolvimento interativo (IDE) baseado na Web associado que permite prototipagem e visualização de resultados. No banco de dados dessa aplicação é possível encontrar as informações de diversos satélites que diariamente coletam imagens da superfície da Terra.

Um desses satélites é justamente o Sentinel-2 utilizado para o desenvolvimento dessa pesquisa. Um script foi criado para extrair as imagens com as bandas corretas como descrito anteriormente. A Figura 5 apresenta a interface do programa.

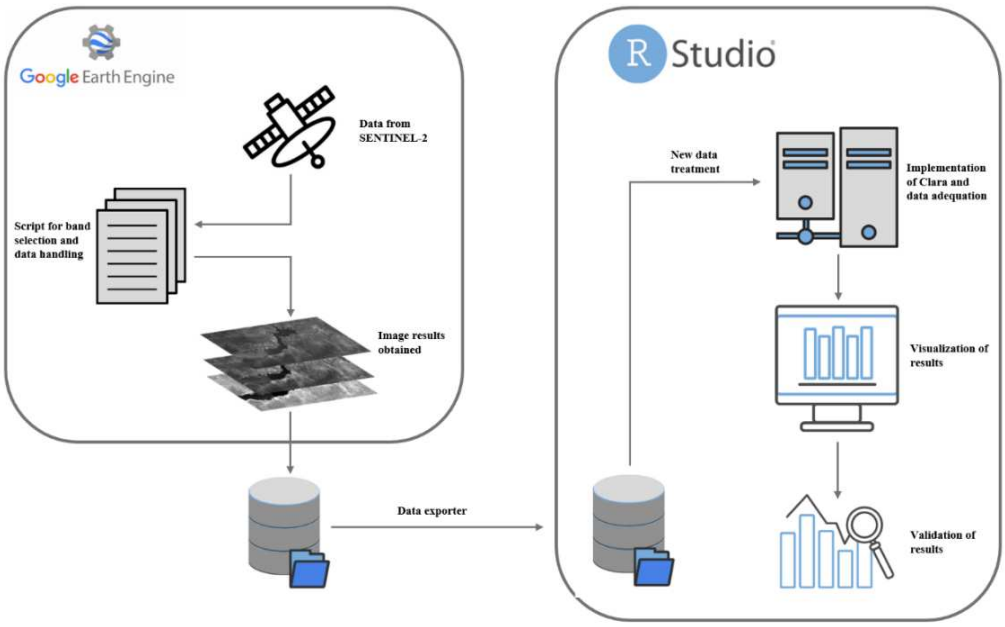
Figura 5. Interface do Google Earth Engine.



Fonte: Autores (2022).

A partir dos dados coletados foi possível prosseguir para a próxima aplicação. Utilizando a linguagem de programação livre R (R Core Team, 2016), uma linguagem de programação estruturada que tem como objetivo a manipulação, análise e visualização de dados. Este software é muito utilizado dentro das áreas estatísticas e analíticas de dados para o desenvolvimento de modelos estatísticos e análise de dados (Ayache, 2021). A linguagem R foi usada para criar os scripts necessários para implementar o Clara e suas visualizações, além de criar os tratamentos necessários para a adequação dos dados de entrada vindos da exportação do Google Earth Engine. O caminho feito pelos dados até a obtenção dos resultados finais pode ser visto na Figura 6.

Figura 6. Diagrama metodológico do desenvolvimento da pesquisa.



Fonte: Autores (2022).

A partir da figura anterior é possível perceber que os dados do Sentinel-2 são importados para o script criado que seleciona as informações necessárias. Após está etapa, as imagens com as bandas são extraídas do Google Earth Engine e importadas para o ambiente do RStudio. Ao entrar no novo software, um novo tratamento das informações é feito com o objetivo de transformar os dados num formato de dataframe, esse formato consiste de uma matriz com n dimensões estruturadas para a interpretação correta do programa. Essa dataframe possuí 3 dimensões, a primeira representa o número de pixels no comprimento da imagem, o segundo representa o número de pixels na largura da imagem e por fim, a terceira representa as 3 diferentes bandas RGB da imagem.

Por fim, esses dados foram submetidos aos scripts desenvolvidos para implementar a técnica Clara. Foi feito primeiramente uma avaliação de otimização de número de grupos, esta etapa tem como objetivo otimizar os resultados do agrupamento diminuindo as diferenças entre os dados dentro dos grupos. Essa validação foi feita utilizando um Diagrama de Silhuetas. Após a obtenção do número ótimo de grupos, o algoritmo foi novamente executado para a obtenção dos resultados finais. Estes resultados foram organizados e apresentados no Tópico 3 desta pesquisa.

3. Resultados e Discussão

3.1 Imagens do Sentinel-2

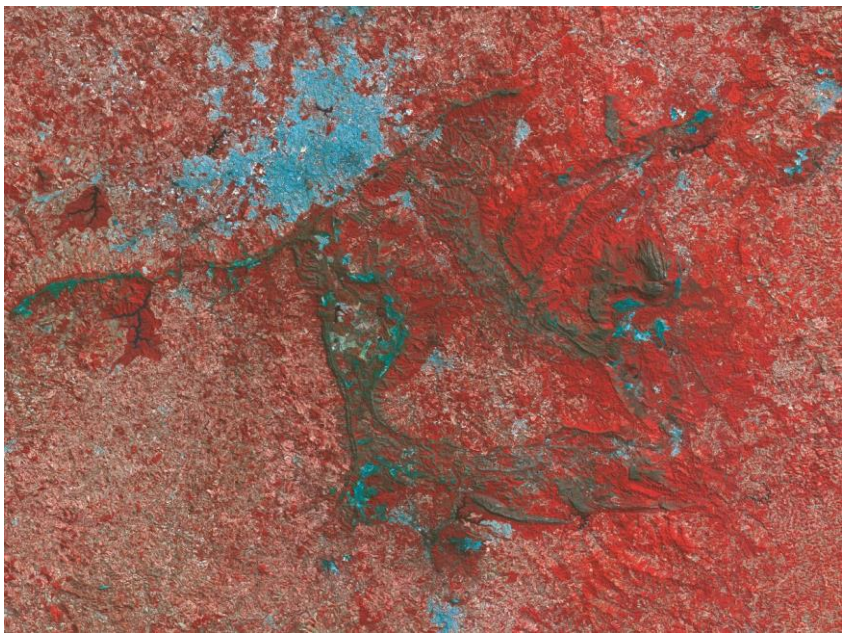
Os primeiros resultados obtidos foram as imagens de satélite retiradas do Sentinel-2 como explicado anteriormente. Na Figura 7 é possível ver a imagem da região do Quadrilátero Ferrífero com as cores naturais retiradas do satélite.

Figura 7. Quadrilátero Ferrífero com bandas naturais.

Fonte: Autores (2022).

Porém, para evidenciar melhor as estruturas geológicas da região, os conjuntos de bandas do Infravermelho de Cor Tradicional (B8, B4 e B3) foi usada. O resultado pode ser visto na Figura 8 e com essa modificação é possível ver como as estruturas geológicas ficam mais evidentes da vegetação, o que tornou mais fácil o agrupamento desses dados nas próximas etapas. Outro ponto importante dessa modificação está relacionado com o destaque dado às zonas urbanas. Essa mudança também possibilitou diferenciar esses pontos dos demais, aumentando a distinção entre os grupos.

Figura 8. Quadrilátero Ferrífero do Infravermelho de Cor Tradicional (B8, B4 e B3).



Fonte: Autores (2022).

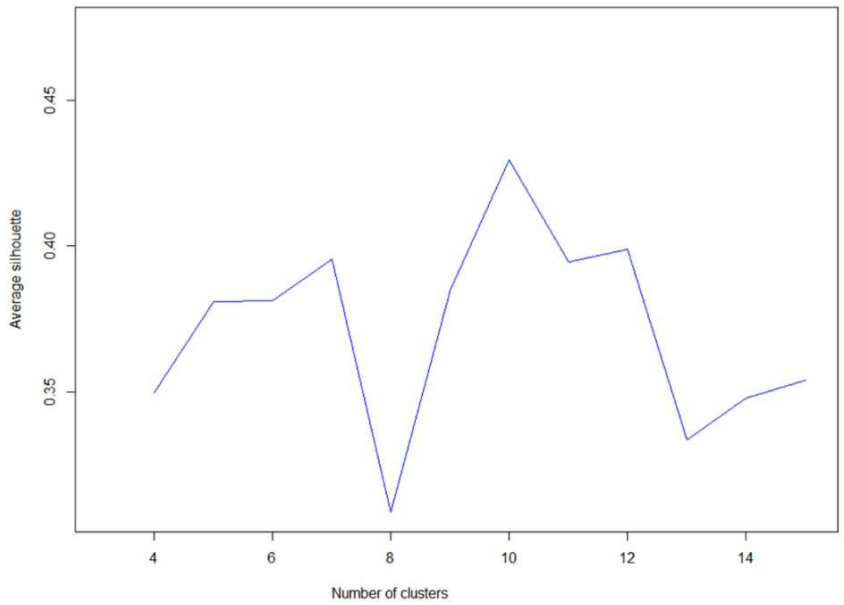
Deve-se destacar que as imagens foram exportadas em alta resolução (3840 x 2879). O objetivo dessa decisão foi garantir que os algoritmos tivessem uma grande variedade de informações melhorando o ajuste fino do seu agrupamento. Com os dados obtidos foi possível importar essas informações para o RStudio (RStudio Team, 2020). Como explicado anteriormente, a imagem foi transformada em uma dataframe para a otimização de parâmetros do modelo.

3.2 Otimização do modelo

Para chegar no número ótimo de grupos para a utilização na CLARA, o algoritmo foi executado variando o número de clusters. Esse número variou de 4 à 15 grupos e a média da silhueta dos grupos em cada repetição foi guardada. Além dessa variação, o número de amostras também foi testado utilizando 4 configurações distintas (50, 100, 150 e 200 amostras).

Ao final dessas repetições, o número de grupos com melhor resultado foi 10 com 100 amostras para este banco de dados. Na Figura 9 é possível ver a variação a silhueta média em função do número de grupos para o número otimizado de amostras.

Figura 9. Silhueta média em função da variação do número de grupos.



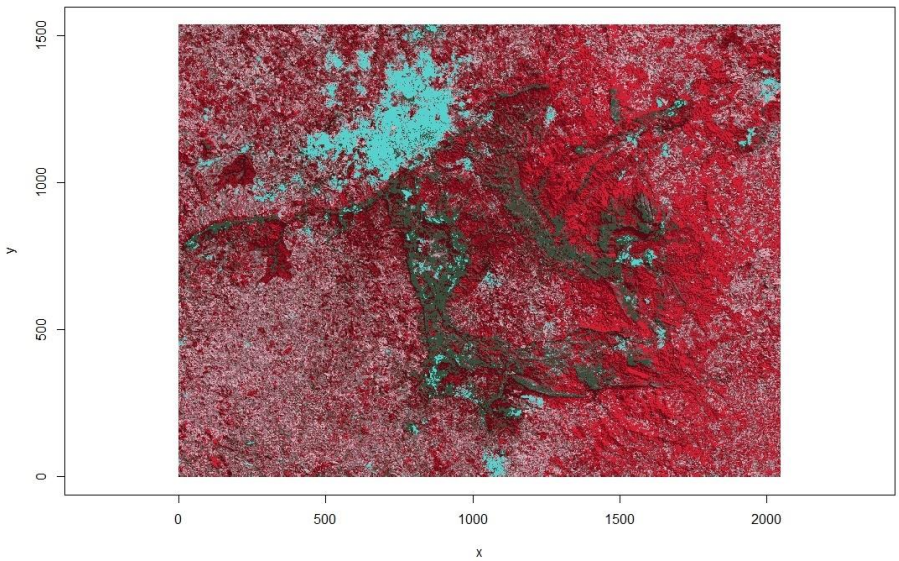
Fonte: Autores (2022).

Os demais parâmetros utilizados foram os padrões utilizados no algoritmo, essa escolha foi baseada também nas mudanças da silhueta média resultante. Como a troca desses outros parâmetros não resultou em variações significantes na otimização, então o padrão foi usado para a criação do modelo final. Com os parâmetros otimizados escolhidos, o modelo final foi criado. Esses resultados podem ser vistos no tópico a seguir.

3.3 I Modelo CLARA

O script foi executado e o primeiro resultado obtido pode ser visto na Figura 10, nela é possível ver a mesma imagem de satélite do QF. Porém com as cores baseadas nos 10 grupos selecionados na otimização descrita anteriormente. Todos os pontos classificados em um mesmo grupo estão representados com uma mesma cor escolhida pelo algoritmo.

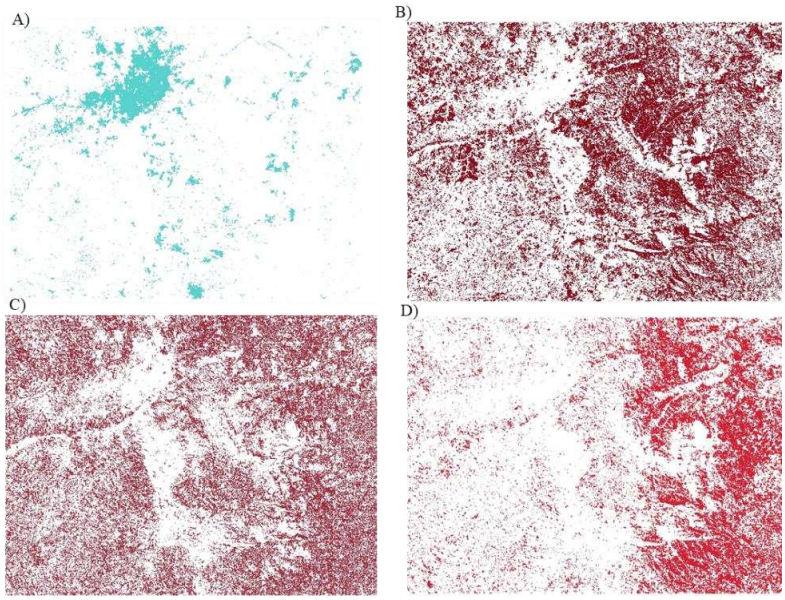
Figura 10. QF após o agrupamento da CLARA.



Fonte: Autores (2022).

Por conseguir manter as características visuais da imagem original, evidenciando bem as diferentes estruturas que compõem a região do QF, é possível verificar a eficácia satisfatória da técnica utilizada. Além disso, por causa desse agrupamento, é possível filtrar esses grupos e analisar individualmente as feições da imagem. Isso pode ser visto na Figura 11, onde os grupos 3, 6, 7 e 9 estão apresentados individualmente. Na Figura 11a, é possível ver as regiões urbanas destacadas em azul onde é possível ver facilmente a cidade de Belo Horizonte (Minas Gerais) a noroeste da imagem. Além disso, nas Figuras 11b, 11c e 11d é possível ver os diferentes tipos de vegetação na região que podem ser facilmente filtrados dos diferentes grupos formados.

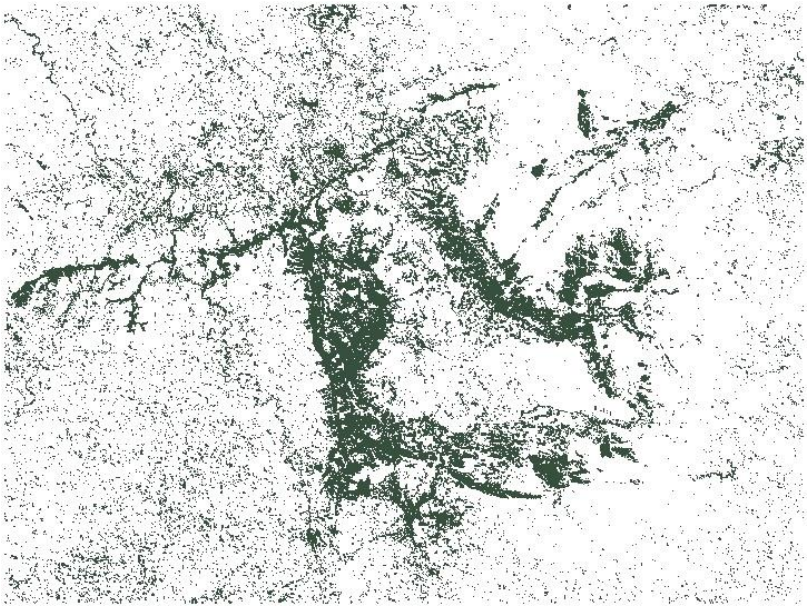
Figura 11. a) Grupo 3; **b)** Grupo 6; **c)** Grupo 7 e **d)** Grupo 9 filtrados.



Fonte: Autores (2022).

Porém, como o objetivo principal dessa pesquisa era identificar um grupo que consiga representar a geomorfologia do Quadrilátero Ferrífero, foi necessário estudar todos os 10 grupos formados. Esta validação concluiu que o Grupo 1 obteve um resultado promissor nesse objetivo. Estes dados agrupados podem ser vistos na Figura 12.

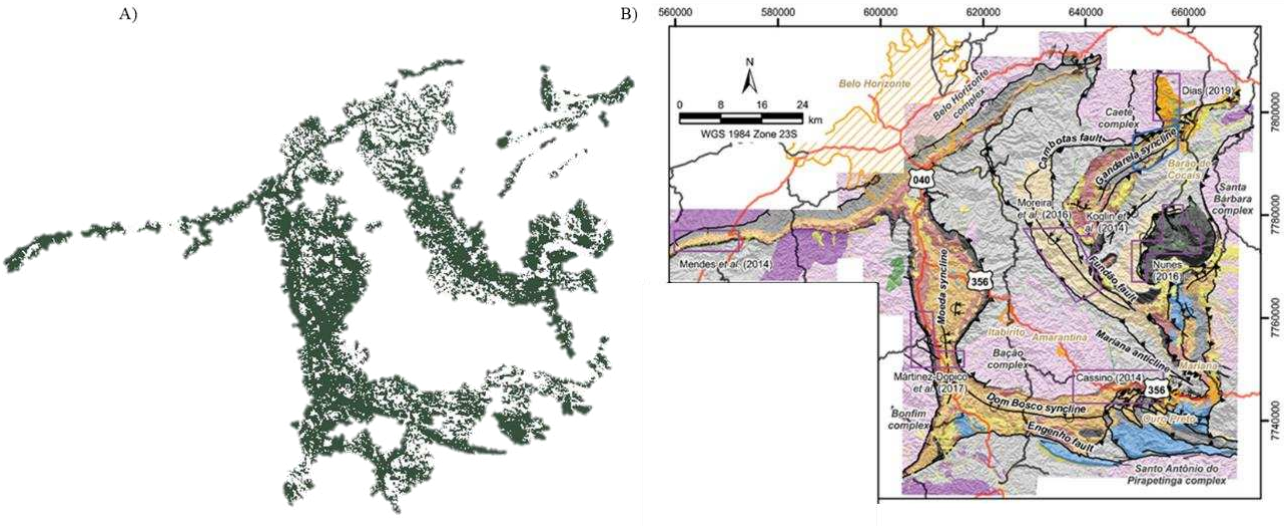
Figura 12. Representação do Grupo 1 de dados agrupados.



Fonte: Autores (2022).

Para uma melhor comparação, o resultado obtido teve os ruídos removidos e os demais pontos dispostos ao lado do mapa do Quadrilátero Ferrífero como pode ser visto na Figura 13. Assim, o resultado obtido pela CLARA foi satisfatório conseguindo representar bem as estruturas geológicas da região.

Figura 13. Comparação entre **a)** Resultado do agrupamento de dados filtrado do CLARA e o **b)** Mapa da Geologia do Quadrilátero Ferrífero.



Fonte: Autores (2022) e Dutra, Martins & Lana, (2019) baseado em Lobato et al. (2005).

4. Conclusão

A partir do estudo foi possível apresentar a utilização de uma técnica de machine learning para identificação de estruturas de particular interesse geológico, a partir de imagens de satélite. A técnica aplicada foi o Clustering Large

Applications (CLARA) que é um algoritmo não-supervisionado para agrupamento de dados, com alta performance em banco de dados massivos.

A área utilizada como estudo de caso foi o Quadrilátero Ferrífero, uma das maiores províncias minerais do planeta, localizada no estado de Minas Gerais, Brasil. Os resultados do modelo CLARA permitiram delinear todas as feições que formam o Quadrilátero Ferrífero.

Neste contexto acredita-se que esta possa ser uma boa ferramenta para seleção de alvos exploratórios reduzindo incerteza e risco aos investidores. O que propicia não somente a atração de novas empresas para pesquisa mineral, além da ampliação das reservas dos recursos minerais brasileiros. Para futuros trabalhos sugere-se a utilização de outras técnicas de machine learning, tais como as técnicas PAM (Partitioning Around Medoids) e K Means na região do QF e também para identificação de outras estruturas a partir de imagens de satélite.

Agradecimentos

Os autores agradecem ao CEFET-MG e ao CIDENG-CNPq pelo apoio durante a pesquisa.

Referências

Alves, H. J. De P.; Fernandes, F. A..; Lima, K. P. De; Batista, B. D. De O.; & Fernandes, T. J. (2020). The COVID-19 pandemic in Brazil: an application of the k-means clustering method. *Research, Society and Development*, 9(10), e5829109059, 2020. 10.33448/rsd-v9i10.9059.

Cabral, A. R., Zeh, A., Koglin, N., Seabra Gomes, A. A., Viana, D. J., & Lehmann, B. (2012). Dating the Itabira iron formation, Quadrilátero Ferrífero of Minas Gerais, Brazil, at 2.65Ga: Depositional U–Pb age of zircon from a metavolcanic layer. *Precambrian Research*, 204-205, 40–45. https://doi.org/10.1016/j.precamres.2012.02.006

Cloutis, E. A. (1996). Review Article Hyperspectral geological remote sensing: evaluation of analytical techniques. *International Journal of Remote Sensing*, 17(12), 2215–2242. doi:10.1080/01431169608948770.

Dorr II, J. V. N. (1969). Physiographic, stratigraphic and structural development of the Quadrilatero Ferrífero, Minas Gerais, Brazil, U.S. *Geological Survey Professional Paper*. 641-A, 110 pp.

Dutra, L. F., Martins, M., & Lana, C. (2019). Sedimentary and U-Pb detrital zircons provenance of the Paleoproterozoic Piracicaba and Sabará groups, Quadrilátero Ferrífero, Southern São Francisco craton, Brazil. *Brazilian Journal of Geology*, 49(2). doi:10.1590/2317-4889201920180095

ENGESAT (2015). Sentinel-2. Curitiba-PR. http://www.engesat.com.br/sentinel-2/.

ESA (2021). Sentinel Online. European Space Agency (ESA). https://sentinel.esa.int/web/sentinel/home.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27. 10.1016/j.rse.2017.06.031

Kaufman, L., & Rousseeuw, P. J. (1987), Clustering by means of medoids. *Statistical Data Analysis based on the L, Norm*, edited by Y. Dodge, Elsevier/North-Holland, Amsterdam, pp. 405-416.

Kaufman, L., & Peter Rousseeuw. (1990a). Finding Groups in Data: An Introduction to Cluster Analysis. Introduction. (n.d.). *Wiley Series in Probability and Statistics,* 1–67. 10.1002/9780470316801.ch1

Kaufman, Leonard, & Peter Rousseeuw. (1990b). Finding Groups in Data: An Introduction to Cluster Analysis. Partitioning Around Medoids (Program PAM). (n.d.). *Wiley Series in Probability and Statistics*, 68–125. 10.1002/9780470316801.ch2

Kaufman, L., & Peter Rousseeuw. (1990c). Finding Groups in Data: An Introduction to Cluster Analysis. Clustering Large Applications (Program CLARA). (n.d.). *Wiley Series in Probability and Statistics*, 126–163. 10.1002/9780470316801.ch3

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1), 3–10. 10.1016/j.gsf.2015.07.003

Lima, N. P., Ferreira, M. T. S., Ruffeil, M., Ferreira, R. F., Pirette, W. & Galbiatti, H. F. (2020). *Quadrilátero Ferrífero: Cinco Décadas de Histórias, Descobertas, Importância Econômica e Tecnológica e Novas Fronteiras para a Mineração de Ferro*. In: Paulo de Tarso Amorim Castro; Issamu Endo; Antonio Luciano Gandini. (2020). Quadrilátero Ferrífero: Avanços do Conhecimento nos Últimos 50 Anos. Belo Horizonte: 3i. 1, 318-41

Lobato L.M., Baltazar O.F., Reis L.B., Achtschin A. B., Baars F.J., Timbó M.A., Berni G.V., Mendonça B.R.V., & Ferreira D. (2005). *Projeto Geologia do Quadrilátero Ferrífero - Integração e Correção Cartográfica em SIG com Nota Explicativa*. Belo Horizonte, 68 p.

Nascimento, E. R. Do.; Albuquerque, M. A. De.; Barros, K. N. N. De O..; & Barros, P. S. N. Cluster analysis applied to the Human Development Index (HDI) of Brazilian States. Research, Society and Development, [S. l.], 11(2), e18011225747, 2022. 10.33448/rsd-v11i2.25747.

Sabins, F. F. (1997). *Remote Sensing — Principles and Interpretation*, 3rd edn., W.H. Freeman, New York, NY., 494 pp

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/.

Reis, A. R. S.; Silva, K. P. da; Chagas, D. R. das. Analysis of leaf surface and clustering of 10 tree species: a tool in the identification of Amazonian species. *Research, Society and Development*, [S. l.], v. 10, n. 2, p. e58810212961, 2021. 10.33448/rsd-v10i2.12961.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

USGS. (2017). *Mineral commodity summaries 2016*. United States Geological Survey (USGS). U.S. Geological Survey, 202 p.90-1