JANDRESSON DIAS PIRES

MODELAGEM ESTATÍSTICA HÍBRIDA MULTIDIMENSIONAL UTILIZANDO GEOESTATÍSTICA E APRENDIZADO DE MÁQUINA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientador: Gérson Rodrigues dos Santos

VIÇOSA - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

P667m
2023

Pires, Jandresson Dias, 1988-

Modelagem estatística híbrida multidimensional utilizando geoestatística e aprendizagem de máquina / Jandresson Dias Pires. – Viçosa, MG, 2023.

1 tese eletrônica (105 f.): il. (algumas color.).

Orientador: Gerson Rodrigues dos Santos.

Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2023.

Inclui bibliografia.

DOI: https://doi.org/10.47328/ufvbbt.2023.712

Modo de acesso: World Wide Web.

1. Estatística matemática. 2. Geologia - Métodos estatísticos. 3. Krigagem. I. Santos, Gerson Rodrigues dos, 1974-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.5

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

JANDRESSON DIAS PIRES

MODELAGEM ESTATÍSTICA HÍBRIDA MULTIDIMENSIONAL UTILIZANDO GEOESTATÍSTICA E APRENDIZADO DE MÁQUINA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 01 de setembro de 2023.

Assentimento:

gov.br Documento assinado digitalmente
JANDRESSON DIAS PIRES
Data: 18/11/2023 13:23:40-0300
Verifique em https://validar.iti.gov.br

Jandresson Dias Pires
Autor

gov.br Documento assinado digitalmente
GERSON RODRIGUES DOS SANTOS
Data: 20/11/2023 07:09:36-0300
Verifique em https://validar.iti.gov.br

Gérson Rodrigues dos Santos
Orientador

À minha alma inspiradora: A Simone, que tem sido meu porto seguro ao longo desta jornada. Obrigado pelo amor incondicional, pelo apoio constante e por acreditar em mim. Sua presença é minha maior motivação.

À minha família amorosa: Aos meus avós, que me apoiaram incondicionalmente em todas as etapas da minha jornada acadêmica; aos meus pais, que sempre acreditaram em mim e me incentivaram a perseguir meus sonhos. Sem vocês, nada disso seria possível.

Aos meus amados filhos: Taylor e Aylla, vocês são a minha maior fonte de inspiração e motivação. Cada passo que dei ao longo desta jornada acadêmica foi com o objetivo de criar um futuro melhor para vocês. Agradeço pela paciência, compreensão e amor incondicional que demonstraram durante todos os momentos em que estive ocupado com minha pesquisa. Suas risadas, abraços e alegria foram a força que me impulsionaram nos dias mais difíceis. Este trabalho é dedicado a vocês, pois são a razão pela qual me esforcei tanto para alcançar este marco. Que possam sempre acreditar em si mesmos, perseguir seus sonhos e lembrar que seu pai estará aqui, orgulhoso e apoiando-os em todas as suas conquistas. Amo vocês mais do que as palavras podem expressar.

AGRADECIMENTOS

Gostaria de aproveitar esta oportunidade para expressar minha profunda gratidão a todas as pessoas que me apoiaram ao longo desta jornada e contribuíram para a conclusão bem-sucedida desta tese.

Primeiramente, gostaria de agradecer a Deus, cuja graça e orientação estiveram presentes em todos os aspectos deste trabalho. Sua infinita sabedoria e força me fortaleceram nas horas mais desafiadoras e me deram a perseverança para seguir em frente.

À minha família, quero expressar minha gratidão infinita. À minha mãe, meu pai e meu irmão, obrigado por todo o apoio emocional e encorajamento ao longo dos anos. Vocês sempre acreditaram em mim. Suas palavras de incentivo e amor foram minha força motriz.

À minha esposa e aos meus filhos, o meu profundo agradecimento por serem meu porto seguro, meu apoio incondicional e minha fonte constante de inspiração. Suas palavras de encorajamento, paciência e compreensão ao longo desta jornada foram inestimáveis. Sua presença em minha vida é um presente que valorizo além das palavras.

Agradeço ao meu orientador, professor Doutor Gérson Rodrigues dos Santos, pelo seu inestimável apoio, orientação e conhecimento compartilhado. Sua dedicação e comprometimento foram fundamentais para o desenvolvimento deste trabalho. Sou imensamente grato pela confiança que depositou em mim e por me incentivar a superar desafios e explorar novas perspectivas.

Agradeço aos membros da banca examinadora por dedicarem seu tempo e expertise na avaliação deste trabalho. Suas contribuições valiosas e comentários construtivos foram essenciais para aprimorar a qualidade desta pesquisa.

Agradeço à minha instituição de ensino, Instituto Federal do Norte de Minas Gerais - Campus Almenara, por fornecer o ambiente propício e recursos necessários para a realização deste estudo. Agradeço também ao departamento de Estatística da Universidade Federal de Viçosa por seu apoio contínuo e oportunidades de aprendizado valiosas.

Aos meus amigos, sou grato pela amizade verdadeira, pelos momentos de descontração e pelas risadas compartilhadas. Obrigado por estarem presentes em minha vida e por compreenderem as ausências e as horas dedicadas a esta pesquisa. Vocês tornaram essa jornada mais leve e memorável. Agradeço, especialmente, à minha amiga Cleonice, sou imensamente grato por sua amizade desde o período da graduação que foi a primeira pessoa que acreditou e me incentivou a ir tão longe, mesmo eu não acreditando ser possível.

Aos participantes da minha pesquisa, cuja colaboração foi fundamental em cada etapa deste trabalho, expresso minha sincera gratidão. Sem a sua generosidade e disposição em compartilhar suas experiências, este estudo não seria possível.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

A todos que contribuíram direta ou indiretamente para este trabalho, o meu sincero agradecimento.

Que minha gratidão seja sentida em cada palavra desta seção de agradecimentos.

Muito obrigado a todos.

"Suba um degrau de cada vez." (Minha Mãe, 2006)

RESUMO

PIRES, Jandresson Dias, D.Sc., Universidade Federal de Viçosa, setembro de 2023. **Modelagem Estatística Híbrida Multidimensional utilizando Geoestatística e Aprendizagem de Máquina**. Orientador: Gérson Rodrigues dos Santos.

A Modelagem Estatística Multidimensional é uma abordagem que busca representar, graficamente, dados de um determinado domínio de aplicação e fornece mecanismos interativos para a interpretação e compreensão das informações apresentadas. Nesta tese, a aplicação dessa abordagem, foi explorada em diferentes contextos, demonstrando sua eficácia na representação de informações multidimensionais. O objetivo foi a elaboração de modelos multidimensionais dos atributos físicos ou químicos do solo, bem como a predição das propriedades dos maciços rochosos, com base em técnicas de Estatística, Geoestatística e Inteligência Geográfica. Os dados utilizados foram provenientes de furos de sondagem em uma mina em Minas Gerais, Brasil, e de amostras de solo e inventário de castanhais nativos no estado do Amazonas, Brasil. Para alcançar esse objetivo, foram empregados mecanismos de aprendizado de máquina e técnicas de visualização, reconhecendo que, nenhuma técnica isolada, oferece o melhor desempenho para todas as tarefas de representação de dados multidimensionais. Portanto, uma estratégia interessante adotada foi analisar várias representações simultaneamente, mantendo uma conexão semântica entre elas, permitindo que, as ações realizadas em uma técnica, sejam refletidas, automaticamente, nas demais. Os resultados obtidos demonstraram a aplicabilidade e confiabilidade dos modelos desenvolvidos, tanto na visualização e interatividade do usuário com os resultados, quanto na qualidade das informações em si. Além disso, ressalta-se que a abordagem proposta neste trabalho pode ser aplicada em outras áreas e contextos geológico-geomecânicos, contribuindo para uma melhor compreensão e tomada de decisão, em diversos campos da engenharia e ciências ambientais. Em suma, esta tese oferece uma contribuição significativa para a Modelagem Estatística Multidimensional, mostrando sua utilidade na representação de dados complexos, como atributos do solo e propriedades geomecânicas dos maciços rochosos. Os resultados obtidos proporcionam *insights* valiosos para a comunidade científica e para os profissionais envolvidos no estudo e análise desses domínios, promovendo avanços no entendimento e gerenciamento de questões ambientais e geotécnicas.

Palavras-chave: Inteligência Geográfica. Krigagem. Ciência de Dados.

ABSTRACT

PIRES, Jandresson Dias, D.Sc., Universidade Federal de Viçosa, September, 2023. **Multidimensional Hybrid Statistical Modeling using Geostatistics and Machine Learning**. Adviser: Gérson Rodrigues dos Santos.

Multidimensional Statistical Modeling is an approach that seeks to graphically represent data from a given application domain and provide interactive mechanisms for interpreting and understanding the information presented. In this thesis, the application of this approach was explored in different contexts, demonstrating its effectiveness in representing multidimensional information. The main objective was to develop multidimensional models of the physical or chemical attributes of the soil, as well as the prediction of the properties of rock masses, based on Statistics, Geostatistics and Geographic Intelligence techniques. The data used came from drilling holes in a mine in Minas Gerais, Brazil, and from soil samples and an inventory of native chestnut trees in the state of Amazonas, Brazil. To achieve this objective, machine learning mechanisms and visualization techniques were employed, recognizing that no single technique offers the best performance for all multidimensional data representation tasks. Therefore, an interesting strategy adopted was to analyze several representations simultaneously, maintaining a semantic connection between them, allowing the actions performed in one technique to be automatically reflected in the others. The results obtained demonstrated the applicability and reliability of the models developed, both in terms of visualization and user interactivity with the results, and in the quality of the information itself. Furthermore, it is noteworthy that the approach proposed in this work can be applied in other areas and geological-geomechanical contexts, contributing to better understanding and decision-making in various fields of engineering and environmental sciences. In short, this thesis offers a significant contribution to Multidimensional Statistical Modeling, showing its usefulness in representing complex data, such as soil attributes and geomechanical properties of rock masses. The results obtained provide valuable insights for the scientific community and professionals involved in the study and analysis of these domains, promoting advances in the understanding and management of environmental and geotechnical issues.

Keywords: Geographic Intelligence. Kriging. Data Science.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AM	Aprendizado de Máquina
CART	Classification and Regression Trees
DAP	Diâmetro à Altura do Peito
ID3	Iterative Dichotomiser 3
IDW	Inverse Distance Weighted
ML	Maximum Likelihood
MSE	Erro Quadrático Médio
OLS	Ordinary Least Square e ponderados
RF	Random Forest
REML	Restricted Maximum Likelihood
RMR	Classificação de Massa Rochosa
WLS	Weight Least Squares

SUMÁRIO

1 INTRODUÇÃO

A utilização da modelagem multidimensional de um mesmo conjunto de dados permite observá-lo, a partir de diversas perspectivas, além de explorar os pontos fortes e minimizar os pontos fracos das técnicas empregadas (KEIM; KRIEGEL, 1996); (EICK; KARR, 2002). Existem várias ferramentas genéricas de visualização de informações, que oferecem diversas técnicas e que permitem o uso coordenado para a visualização de um mesmo conjunto de dados. No entanto, geralmente, essas ferramentas possuem um conjunto fixo de técnicas e formas de coordenação, que são estabelecidas durante uma pesquisa, o que limita o potencial de utilização das técnicas em diferentes contextos de aplicação e restringe a liberdade do usuário no processo de exploração.

Quando se aborda a modelagem multidimensional, o foco está na construção de um modelo voltado para a visualização exploratória, e não apenas para a análise de dados. Nesse sentido, espera-se que o modelo ofereça:

- Uma representação simplificada do objeto estudado;
- Facilidade de interpretação, permitindo que qualquer usuário, independentemente da área de estudo, possa compreendê-lo;
- Facilidade na implementação física do modelo, visando maximizar o desempenho das consultas aos dados.

A grande maioria dos modelos, que mensuram as características do solo, são construídos em duas dimensões (2D), o que pode dificultar a visualização e a compreensão da relação espacial entre as variáveis importantes para a modelagem e as intervenções de engenharia. As principais dificuldades para elaborar esses modelos, estão relacionadas com a espacialização de dados unidimensionais e a interpolação de dados pontuais para compor os dados espaciais. Nesse contexto, o presente estudo surge da necessidade de propor métodos de construção de modelos multidimensionais utilizando os atributos do solo e da rocha, incluindo suas propriedades físicas, químicas e mecânicas, com a aplicação de ferramentas geoestatísticas conhecidas e da Aprendizagem de Máquina, como suporte à elaboração desses modelos.

Assim, adota-se como hipótese principal, que motiva o desenvolvimento dessa pesquisa, a possibilidade de modelagem híbrida estatística de dados multidimensionais, utilizando as técnicas de modelagem geoestatística (modelos tradicionais para dados geoespaciais), da Inteligência Geográfica e todo o potencial das técnicas de modelagem computacional (modelos de aprendizagem de máquina).

Nos aspectos relacionados às bases de dados utilizadas neste trabalho, nos Capítulos 2 e 3, foi obtido informações de furos de sondagem, por meio de uma parceria com uma empresa de mineração. Esses dados foram utilizados, em conjunto, para a elaboração da tese de (PEREIRA, 2020), que desenvolveu um estudo para a construção de modelos geomecânicos. De acordo com PEREIRA, a mineradora em questão atua no município de Vazante, estado de Minas Gerais, e forneceu seu banco de dados geológico-geotécnico de uma das minas subterrâneas, localizadas nesse mesmo município. No Capítulo 4, em colaboração com a dissertação de (MORAES, 2021), a base de dados se refere a amostras de solo, coletadas na Propriedade do Jutica, em Tefé/AM[1]. Essas amostras foram coletadas na camada superficial de 0-20 cm de uma parcela permanente, com dimensões de 300 m x 300 m, respeitando uma distância de 50 m entre as linhas e 30 m entre os pontos, totalizando 60 amostras. Além disso, foi realizado um inventário de todas as castanheiras com DAP (Diâmetro à Altura do Peito) $\geq$ 10 cm.

Nesse contexto, levando em consideração as especificidades do banco de dados, buscou-se desenvolver um pacote metodológico que possibilitasse a visualização de informações. Ao manipular essas informações de forma dinâmica, facilita a interpretação dos dados, permitindo a percepção de padrões, tendências, relacionamentos e exceções incorporados neles.

Este trabalho está organizado da seguinte forma:

A seguir, será abordado um referencial bibliográfico das principais metodologias utilizadas neste trabalho. Destaca-se a importância da Inteligência Geográfica, do Aprendizado de Máquina, do Algoritmo *Random Forest*, da Geoestatística, da Modelagem Multidimensional e de conhecimentos inerente as especificidades dos bancos de dados utilizados na pesquisa.

No capítulo 2, consta o artigo publicado no *Journal of South American Earth Sciences - ELSEVIER* sobre a construção de modelos geomecânicos multidimensionais, com o interpolador IDW e empregando a linguagem R. Os dados utilizados são provenientes de furos de sondagem, localizados em uma mina subterrânea no estado de Minas Gerais, Brasil.

No capítulo 3, será abordada a análise da variabilidade espacial das propriedades do maciço rochoso em minas subterrâneas, por meio de um modelo simplificado de Classificação de Massa Rochosa (RMR), desenvolvido utilizando as informações obtidas a partir de furos de sondagem e o algoritmo *Random Forest*. O referido texto foi publicado como capítulo do livro *Advanced Analytics in Mining Engineering - Springer*.

No capítulo 4, que está em processo de publicação, será apresentado o artigo

[1] No trabalho desenvolvido para a dissertação, as amostras de solo foram coletadas nas Comunidades Lago das Pedras - Barcelos/AM, Propriedade do Jutica - Tefé/AM e Comunidade Jatuarana - Manicoré/AM. No entanto, neste trabalho, destacamos apenas a região que foi submetida para publicação.

intitulado "Seletividade ambiental da castanheira-da-amazônia, com a utilização da Modelagem Geoestatística e do algoritmo *Random Forest*". O objetivo deste estudo foi analisar a relação entre os atributos físicos e químicos do solo com os castanhais nativos. O banco de dados utilizado, consiste em amostras de solo coletadas em castanhais nativos do Estado do Amazonas, juntamente com o inventário de todas as castanheiras na região de estudo.

1.1 Inteligência Geográfica

A Inteligência Geográfica, também conhecida como Geointeligência, é uma disciplina que visa coletar, analisar e interpretar informações geográficas, para obter *insights* e tomar decisões informadas. Ela desempenha um papel crucial em diversas áreas, como planejamento urbano, gestão de recursos naturais, segurança e tomada de decisões estratégicas.

Segundo JOHNSON (2016), a inteligência geográfica envolve a análise de informações geoespaciais, permitindo compreender padrões espaciais e interações entre elementos geográficos. Essa compreensão aprofundada do contexto geográfico. é essencial para tomar decisões informadas e estratégicas em várias áreas.

A inclusão de informações geoespaciais na análise de dados, é uma prática que enriquece a compreensão dos fenômenos estudados. GARCIA; LEE; CHEN (2020) destacam que, a integração de dados geográficos na ciência de dados, traz desafios e oportunidades. A inclusão desses dados permite identificar relações espaciais e padrões ocultos, que podem não ser perceptíveis em análises puramente numéricas. Isso evidencia a importância de considerar os aspectos geográficos ao realizar análises de dados em diferentes áreas de estudo.

As técnicas de análise de dados geográficos, desempenham um papel fundamental na inteligência geográfica. LEE (2017) destaca que, a análise de padrões espaciais, a modelagem geográfica e a visualização de informações geográficas, são algumas das abordagens utilizadas nessa disciplina. Essas técnicas permitem compreender melhor os fenômenos geográficos, em um contexto espacial, fornecendo *insights* valiosos para a tomada de decisões.

A Inteligência Geográfica e a Ciência de Dados, embora sejam áreas de estudo distintas, podem estar relacionadas entre si. Conforme mencionado, a inteligência geográfica é uma disciplina, que se concentra na coleta, análise e interpretação de informações geográficas, para obter *insights* e tomar decisões informadas. Por outro lado, a Ciência de Dados é um campo mais amplo que se concentra na extração de conhecimento, a partir de grandes volumes de dados, independentemente de sua natureza geográfica.

Tanto a Inteligência Artificial, quanto a Inteligência Geográfica, desempenham papéis importantes na ciência de dados. Conforme mencionado por (SMITH, 2018), o avanço da inteligência artificial tem transformado várias áreas de estudo, incluindo a Ciência de Dados. Os algoritmos de aprendizado de máquina têm a capacidade de analisar grandes volumes de dados e extrair conhecimento significativo. Essa interseção, entre a Inteligência Artificial e a Ciência de Dados, demonstra como a Inteligência Artificial impulsiona o desenvolvimento de técnicas e métodos para a análise de dados em várias disciplinas.

Da mesma forma, a Inteligência Geográfica desempenha um papel crucial na ciência de dados, como mencionado por (JOHNSON, 2016). A análise de informações geoespaciais permite compreender melhor os padrões espaciais e as interações entre elementos geográficos. Ao considerar a Inteligência Geográfica no contexto da ciência de dados, é possível utilizar dados geográficos para identificar as relações espaciais e os padrões ocultos, fornecendo *insights* valiosos em análises multidimensionais. Ou seja, a Ciência de Dados tem se beneficiado, significativamente, da integração de dados geográficos em suas análises, conforme observado por (GARCIA; LEE; CHEN, 2020). A inclusão de informações geoespaciais, enriquece a compreensão dos dados, permitindo identificar relações espaciais e padrões ocultos, que podem não ser perceptíveis em análises puramente numéricas. Isso destaca a importância de considerar os aspectos geográficos, ao realizar análises de dados em diferentes áreas de estudo.

O termo *Ciência de Dados* está sendo cada vez mais utilizado para designar uma área de conhecimento voltada para o estudo e a análise de dados, onde busca-se extrair conhecimento e criar novas informações. É uma atividade interdisciplinar, que concilia, principalmente, duas grandes áreas: Ciência da Computação e Estatística. A Ciência de Dados vem sendo aplicada como apoio em outras áreas diferentes de conhecimento, tais como: Medicina, Biologia, Mineração, Meio Ambiente, Ciências Políticas, Ciências Agrárias, etc. Apesar de não ser uma área nova, o tema vem se popularizando cada vez mais, graças à explosão na produção de dados e crescente dependência dos dados para a tomada de decisão.

A produção de dados por diversos sistemas de informação e a necessidade de integração destes sistemas, tal como a inclusão de fontes de informações espaciais, demográficas e também relativas à redes sociais *on-line*, trazem à tona a necessidade de aplicação de metodologias diferenciadas, apropriadas para grandes massas de dados. Neste sentido, as áreas de ciências agrárias, mineração e meio ambiente, passam a assimilar os desafios das ciência de dados e da era *big data*, buscando novas metodologias para lidar com este tipo de dados.

Tradicionalmente, as técnicas de análise de dados, foram criadas para extrair informação a partir de poucos dados, em geral estáticos, limpos e de natureza pouco

relacional, amostrados cientificamente e seguindo suposições claras, como independência, estacionariedade e normalidade. São dados gerados e analisados para responder uma pergunta específica, formulada anteriormente à sua produção (MILLER, 2010).

O desafio de analisar *big data* é lidar com a abundância, exaustividade e variedade, sua dinâmica, desordem e incerteza, e a necessidade de lidar com dados, que não foram gerados para responder a uma questão específica (KITCHIN, 2014). O termo *big data* começou a ser utilizado no final da década de 1990 e seu significado foi se modificando, conforme as demandas e a tecnologia envolvida avançavam (WANG; HAJLI, 2017). Desta forma, a própria definição do termo *big data* ainda é vaga, existindo mais de 43 definições (HUANG et al., 2015). Na área de ciências agrárias, a quantidade de dados talvez não seja a característica mais importante para a decisão quanto ao emprego do termo *big data*, ainda mais quando comparada à quantidade de dados produzida em outras áreas, como financeira e de redes sociais (HERLAND; KHOSHGOFTAAR; WALD, 2014). Neste sentido, a definição proposta por (DEMCHENKO et al., 2013), parece melhor compreender os diversos aspectos da aplicação de *big data*, apresentando cinco conceitos chaves, que se iniciam com a letra "V": volume, velocidade, variedade, veracidade e valor. Pode-se, ainda, acrescentar a propriedade de exaustividade, proposta por (KITCHIN, 2013).

- **Volume** - Segundo a definição de (DEMCHENKO et al., 2013), volume é a característica mais importante e distinta em *big data*. Esta característica impõe uma série de desafios e requisitos específicos que, em geral, não são tratados pelas tecnologias tradicionais.

  O ganho de volume nas bases de dados se dá, em boa parte, devido aos avanços tecnológicos de instrumentação. Em diversas áreas, a tendência atual é de se coletar e armazenar dados de todos os eventos observados, de todas as atividades e sensores disponíveis, mesmo que não haja um uso previsto e direto para estes dados. Desta forma, os dados passam a ser coletados e armazenados também pelo seu uso potencial.

  As características de volume em *big data* envolvem características, como tamanho, escala, quantidade e dimensões. Naturalmente, esta característica, apresenta um certo grau de relativismo: o que é *big* em *big data*? Áreas, como física e ciências da computação, apresentam características de volume muito diferentes de áreas como demografia ou ciência do solo. Neste sentido, sugere-se que, um volume adequado para classificação como *big data*, seja a comparação dentro do mesmo campo de pesquisa.

  Por exemplo: qual é o volume de dados tipicamente coletado sobre amostras de solo e o qual seria o volume acrescentado ao se armazenar, também, dados sobre as condições climáticas, manuseio do solo, métodos de coleta, etc? Caso este acréscimo seja significante ao ponto de influenciar a escolha de novas tecnologias, que permitam

que estes dados sejam trabalhados integralmente, pode-se afirmar que trata-se de *big data*.

- **Velocidade** - Em geral, dados de *big data* são gerados em rápida velocidade. São coletados em tempo real ou em intervalos curtos de atualização. Pode-se considerar, por exemplo: dados de aerolevatamento sendo recebidos, processados e monitorados pela equipe de campo em tempo real.

- **Variedade** - Tradicionalmente, dados são coletados para uma finalidade prevista, já sendo, inclusive, armazenados, prevendo-se um método de análise e objetivos bem definidos de seu uso. Em *big data*, são coletados todos os dados, não prevendo-se, necessariamente, o seu uso antecipado: todos os dados são importantes e armazenados em seu estado mais natural.

  Desta forma, os dados podem ser do tipo "estruturados", como em planilhas e tabelas de dados, ou "não estruturados" como em textos, transcrições de falas ou sequências não lógicas de informação.

- **Valor** - Segundo DEMCHENKO et al. (2013), esta característica é dada pelo valor agregado que os dados coletados acrescentam ao processo ou atividade. Desta forma, mesmo que a coleta de certos dados aumentem os requisitos de armazenamento e processamento, estes serão coletados, caso a sua utilização seja essencial ou inovadora para o processo. Por exemplo, um sistema de informação que permita conectar dados de diversos sistemas de informações sobre as características e dados já levantados de uma determinada região, apresenta um custo considerável; contudo o valor agregado deste resultado, justifica a sua execução.

- **Veracidade** - Esta característica se divide em dois aspectos: a consistência dos dados e os métodos de processamento. Enquanto o primeiro aborda a questão da veracidade dos dados, sua correspondência com a realidade, o segundo trata de questões como confiabilidade e segurança nos processos de manipulação dos dados, que permitam a sua consistência em todo o processo, obtendo-se por fim, dados verossímeis.

- **Exaustividade** - *big data* pretende capturar por completo toda uma população ou sistema, onde n = N. Isto é, a amostra corresponde ao universo dos dados.

Em resumo, o termo *big data* pode ser aplicado quando se defronta com um grande volume de dados, produzido e atualizado em alta velocidade, com grande variedade e complexidade interna, apresentando questões específicas sobre sua veracidade, consistência e confiabilidade, tal como o valor que os dados detêm, medido pela capacidade destes gerarem novos conhecimentos e avanços (DEMCHENKO et al., 2013). Cabe aqui, ressaltar que certos conjuntos de dados gerados por sistemas de informação e outras fontes, nem sempre irão contemplar estas cinco características simultaneamente, mas isto não elimina a

necessidade de aplicação de metodologias específicas para lidar com algum tipo particular de conjunto de dados (HERLAND; KHOSHGOFTAAR; WALD, 2014).

Dessa forma, o conceito de *big data* e sua utilização, se encaixam em um campo mais amplo, conhecido como ciência de dados. Na ciência de dados, além dos dados em si, são considerados os conceitos e métodos necessários para a coleta, análise e apresentação dos resultados. Nesse sentido, uma abordagem que pode ser utilizada é o Aprendizado de Máquina (AM). O AM consiste em utilizar algoritmos e técnicas para treinar modelos computacionais, a fim de que eles possam aprender padrões nos dados e fazer previsões, ou tomar decisões com base nesses padrões. Isso permite extrair conhecimento e insights valiosos a partir dos dados, auxiliando na compreensão dos fenômenos estudados e na tomada de decisões informadas.

1.2 Aprendizagem de Máquina

Muitos dados são capturados e armazenados diariamente ao redor do mundo (MARSLAND, 2011). Estes dados podem ser fotografias, músicas, compras, etc. Com uma quantidade massiva de dados é importante, para pesquisadores e para a indústria, descobrir como extrair informações, a partir destes. Com uma quantidade pequena de dados, é possível, para os pesquisadores, a tentativa de extração de informações manualmente, mas, a medida que a quantidade de dados disponíveis para análise crescem, torna-se cada vez mais difícil para que humanos consigam fazer a análise dos dados. A partir desse ponto, utilizar uma máquina ou computador para processar e extrair informações desses dados, é interessante, uma vez que computadores podem processar dados muito mais rápido que humanos e sem se cansarem.

Aprendizado de Máquina é a capacidade de fazer o computador modificar ou adaptar suas ações (sejam para fazer predições ou controlar um robô), de maneira a torná-las mais precisas. Essa precisão é medida por quão bem as ações escolhidas refletem a resposta correta (MARSLAND, 2011). Para que seja possível o aprendizado, faz-se necessário que seja dado, como entrada, um conjunto de dados ou *dataset*, para que o computador possa aprender. Existem diversas maneiras de extrair informações de uma base de dados. Por exemplo, encontrar semelhanças para grupo de clientes de uma rede de supermercados, ou ainda, a partir de um conjunto de imagens de animais, identificar qual é o animal numa fotografia. Outra possibilidade consiste em prever o preço de uma casa com base no tamanho, localização e ano de construção. Estes problemas são tratados dentro do Aprendizado de Máquina. Existem diferentes áreas para tratar cada um destes problemas, como Aprendizagem Supervisionada e Aprendizagem Não Supervisionada.

Na Aprendizagem Supervisionada é dado um *dataset* rotulado, isto é, o valor de saída correto, já é conhecido. Neste caso, se tem a ideia que existe uma relação entre

os valores de entrada e saída. Assim, o objetivo da máquina, é encontrar uma função que represente essa relação. Os problemas da Aprendizagem Supervisionada, são ainda classificados em problemas de ”regressão”e ”classificação”. De forma resumida, em problema de regressão, espera-se prever resultados de uma saída contínua (MARSLAND, 2011). Já em problemas de classificação, deseja-se mapear as variáveis de entrada para categorias distintas.

Dois exemplos, de problemas de regressão e classificação, podem ser:

i) No caso de um conjunto de dados, contendo informações sobre o tamanho das casas e seus respectivos preços, podemos tentar prever o preço, com base nessas informações. Como o preço é uma variável contínua, esse problema se enquadra na categoria de problemas de regressão.

ii) Por outro lado, podemos ter um exemplo de classificação em um conjunto de dados contendo fotografias de homens e mulheres, no qual tentamos prever se uma determinada foto é de um homem ou de uma mulher. Nesse caso, temos duas categorias distintas, o que caracteriza um problema de classificação.

No contexto do desenvolvimento de software, o problema da estimativa de esforço pode ser definido, tanto como um problema de regressão, quanto como um problema de classificação (WEN et al., 2012); (BANIMUSTAFA, 2018). No primeiro caso, o esforço é tratado como uma variável contínua, abrangendo uma faixa de valores, que vai de 0 até infinito. Nessa abordagem, o objetivo é prever um valor numérico específico para representar o esforço necessário.

Por outro lado, ao considerar a abordagem de classificação, é necessário, primeiro, definir categorias de esforço, como ”pequeno”, ”médio”e ”grande”. A definição dessas categorias pode variar de acordo com critérios estabelecidos pelo pesquisador ou com base em algum padrão específico. Nesse contexto, a tarefa consiste em classificar o esforço em uma das categorias pré-definidas.

Existem diversas técnicas de AM para problemas de regressão e classificação, algumas podem ser utilizadas para resolver ambos os problemas. Na estimativa de esforço, diversas técnicas de AM já foram utilizadas, por exemplo: Support Vector Machine, Random Forest, XGBoost, Redes Neurais, entre outras (WEN et al., 2012); (AMARAL et al., 2019). Segundo IDRI; HOSNI; ABRAN (2016), não existe um consenso sobre qual a melhor técnica de AM para a estimativa de esforço, algumas obtém resultados satisfatórios para um *dataset*, enquanto, em outros, os resultados não são bons. Assim, começou-se a utilizar um *ensemble* de técnicas de AM, que nada mais é do que uma técnica de AM, que combina outras técnicas, ou uma única técnica, com diferentes parâmetros. A utilização de técnicas *ensemble* começou a ser empregada, de modo a conseguir estimativas satisfatórias

para o maior número de *datasets*, com a mesma técnica de AM. (IDRI; HOSNI; ABRAN, 2016).

Ao utilizar uma técnica *ensemble* os pesquisadores ainda continuam com algumas preocupações, como encontrar os melhores parâmetros para a técnica utilizada, além de qual técnica utilizar. Com objetivo de solucionar alguns destes problemas, foi criado o *Automatic Machine Learning* ou AutoML, que tem como finalidade encontrar quais os melhores parâmetros e técnicas para o dataset analisado e, posteriormente, apresentar os resultados.

A escolha, entre a abordagem de regressão e classificação para o problema da estimativa de esforço, depende do contexto específico e dos objetivos da análise. Ambas as abordagens possuem vantagens e podem fornecer informações úteis para o desenvolvimento de software.

Outra abordagem existente é a Aprendizagem Não Supervisionada. Nessa abordagem, os algoritmos buscam identificar padrões ou estruturas ocultas em conjuntos de dados não rotulados, ou seja, dados em que não há informações prévias sobre suas classes ou categorias (MARSLAND, 2011).

Um dos principais objetivos, da aprendizagem não supervisionada, é a clusterização dos dados, em que elementos semelhantes são agrupados em conjuntos ou clusters, enquanto elementos distintos são separados. Essa técnica permite descobrir grupos naturais ou segmentos latentes presentes nos dados, com base em suas características compartilhadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Existem diversos algoritmos populares para realizar a clusterização. O algoritmo K-Means é amplamente utilizado e atribui cada ponto de dados ao cluster mais próximo dos K clusters pré-definidos, minimizando a soma dos erros quadrados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Já o algoritmo DBSCAN, se baseia na densidade de pontos para agrupá-los, considerando pontos próximos uns dos outros e com densidade suficiente (ESTER et al., 1996).

Além da clusterização, a aprendizagem não supervisionada também abrange outras técnicas, como redução de dimensionalidade e descoberta de regras de associação. A redução de dimensionalidade, visa reduzir a quantidade de variáveis em um conjunto de dados, mantendo a informação relevante (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Por sua vez, a descoberta de regras de associação, busca encontrar associações frequentes entre diferentes itens em uma base de dados (AGRAWAL; SRIKANT, 1993).

A aplicação da aprendizagem não supervisionada é vasta e abrange diversas áreas, como análise de dados, segmentação de mercado, detecção de anomalias, recomendação de itens, processamento de linguagem natural, entre outras (MARSLAND, 2011). Essa abordagem é fundamental para explorar e compreender grandes volumes de dados não

rotulados e descobrir estruturas significativas, que podem fornecer *insights* valiosos para tomada de decisões e compreensão dos fenômenos estudados.

1.3 *Random Forest*

O *Random Forest* é um algoritmo de aprendizado de máquina, baseado na teoria dos conjuntos, que combina várias árvores de decisão para realizar tarefas de classificação e regressão. Foi proposto por Leo Breiman em 2001 (BREIMAN, 2001) e tem sido amplamente utilizado em diversas áreas de aplicação, como medicina, finanças, sensoriamento remoto, entre outras.

1.3.1 Árvores de Decisão

As árvores de Decisão é uma técnica de aprendizado de máquina amplamente utilizada, devido à sua simplicidade e interpretabilidade. Elas são estruturas hierárquicas, em forma de árvore, que representam um conjunto de regras de decisão sequenciais. Cada nó interno da árvore corresponde a uma decisão baseada em um atributo, e cada ramo representa o resultado dessa decisão. As folhas da árvore representam as saídas ou classes finais.

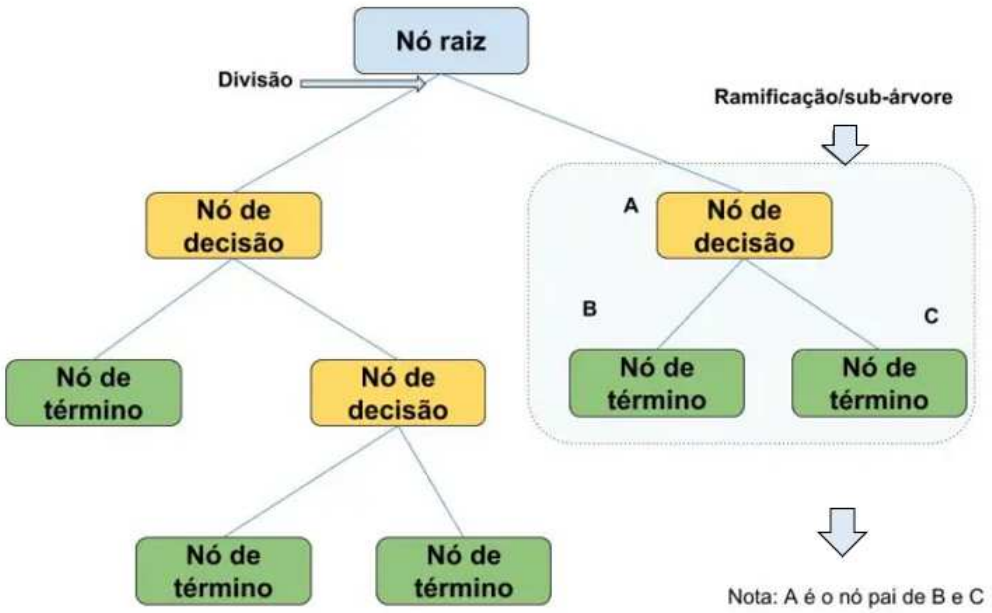


Figura 1 – Exemplo de uma árvore de decisão

Fonte: (VIDHYA, 2016)

Existem vários algoritmos para construir árvores de decisão, como o ID3 (Iterative

Dichotomiser 3), C4.5 e CART (Classification and Regression Trees). O ID3 utiliza o conceito de ganho de informação, para realizar a divisão dos dados em cada nó da árvore (QUINLAN, 1986). O ganho de informação mede a redução na incerteza (entropia), após a divisão dos dados, com base em um atributo específico. A entropia é uma medida da impureza dos dados e é definida como:

$$H(D) = -\sum_{i=1}^{c} p(i) \log_2 p(i) \tag{1.1}$$

Em que $H(D)$ é a entropia do conjunto de dados D, c é o número de classes e $p(i)$ é a proporção de instâncias da classe i em D.

O CART é um algoritmo que constrói árvores binárias de decisão, dividindo, recursivamente, o conjunto de dados, com base em diferentes atributos e critérios de divisão, como o índice Gini para problemas de classificação ou o erro quadrático médio para problemas de regressão.

1.3.1.1 Índice Gini

O índice Gini é uma medida de impureza, usada para dividir os dados em árvores de decisão. Ele mede a probabilidade de uma instância selecionada aleatoriamente ser incorretamente classificada, com base na distribuição das classes no conjunto de dados. O índice Gini é definido pela seguinte fórmula:

$$Gini(p) = 1 - \sum_{i=1}^{c} (p(i))^2 \tag{1.2}$$

Em que $Gini(p)$ é o índice Gini para um nó específico, c é o número de classes e $p(i)$ é a proporção de instâncias da classe i no nó (BREIMAN et al., 1984).

Durante a construção de uma árvore de decisão no *Random Forest*, o atributo que resulta na maior redução do índice Gini, é selecionado para fazer a divisão dos dados em cada nó. Isso ajuda a criar árvores, que melhor separam as classes ou os valores de saída.

1.3.1.2 Erro Quadrático Médio (MSE)

O Erro Quadrático Médio (MSE) é uma métrica, amplamente utilizada para avaliar o desempenho de modelos de regressão. Ele mede a média dos erros ao quadrado, entre as previsões do modelo e os valores reais. O MSE é calculado pela seguinte fórmula:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1.3}$$

Em que y_i é o valor real do i-ésimo exemplo, $\hat{y}_i$ é a previsão do modelo para o i-ésimo exemplo e n é o número total de exemplos.

O MSE é uma métrica comum, porque penaliza erros maiores de forma quadrática. Isso significa que, erros grandes, terão um impacto maior no valor final do MSE. Portanto, minimizar o MSE, é uma forma de buscar previsões mais precisas e reduzir a diferença entre as previsões e os valores reais.

Segundo HASTIE; TIBSHIRANI; FRIEDMAN (2009), o MSE é uma medida de ajuste, amplamente utilizada para problemas de regressão e é amplamente interpretável, pois tem as mesmas unidades do alvo (variável dependente). No entanto, o MSE pode ser sensível a *outliers*, pois, o quadrado dos erros, amplifica sua influência no cálculo do MSE. Por isso, é importante considerar outras métricas e técnicas quando se lida com *outliers* em problemas de regressão.

1.3.2 O Algoritmo *Random Forest*

O algoritmo *Random Forest*, introduzido por Breiman (2001), é uma técnica de aprendizado de máquina, que combina várias árvores de decisão independentes, para melhorar o desempenho e a generalização do modelo. Cada árvore de decisão no *Random Forest*, é construída usando uma versão modificada do algoritmo CART (Classification and Regression Trees) (BREIMAN, 2001). O *Random Forest* ganhou popularidade devido à sua capacidade de lidar com problemas de classificação e regressão, bem como pela sua robustez e eficácia em relação ao *overfitting* (LIAW; WIENER, 2002). O processo de construção do *Random Forest*, envolve duas fontes de aleatoriedade: amostragem por *bootstrap* e seleção aleatória de atributos.

No *Random Forest*, cada árvore é construída a partir de uma amostra de treinamento, criada por meio da amostragem por *bootstrap*. O método de amostragem por *bootstrap*, é uma técnica essencial no processo de construção do *Random Forest*. Ele envolve a criação de várias amostras de treinamento, a partir do conjunto de dados original, permitindo que, cada árvore de decisão, seja treinada em uma amostra diferente. O método de amostragem por *bootstrap*, é baseado em uma abordagem de amostragem com reposição, onde cada amostra é criada, selecionando, aleatoriamente, instâncias do conjunto de dados original com substituição.

A amostragem por *bootstrap,* fornece duas vantagens principais para o *Random Forest*. Em primeiro lugar, ajuda a criar diversidade entre as árvores, já que cada amostra de treinamento é ligeiramente diferente. Em segundo lugar, permite que o *Random Forest* avalie a importância dos atributos, medindo o impacto de cada atributo nas previsões em diferentes amostras (EFRON; TIBSHIRANI, 1994).

1.3.3 Vantagens e Desvantagens

O *Random Forest* apresenta várias vantagens em relação a outros algoritmos de aprendizado de máquina:

- **Desempenho robusto**: O *Random Forest* tem um bom desempenho em uma variedade de conjuntos de dados, incluindo aqueles com ruído, dados faltantes ou *outliers*.
- **Capacidade de lidar com grandes conjuntos de dados**: O algoritmo é eficiente em termos de tempo e memória, permitindo o processamento de grandes conjuntos de dados.
- **Avaliação da importância dos atributos**: O *Random Forest* pode fornecer uma estimativa da importância de cada atributo na tarefa de previsão, o que auxilia na seleção de recursos relevantes.
- **Baixa tendência ao *overfitting***: A combinação das previsões de várias árvores, reduz a tendência ao *overfitting*, tornando o modelo mais geral e menos suscetível a ruídos nos dados de treinamento.

No entanto, também existem algumas desvantagens associadas ao uso do *Random Forest*:

- **Complexidade computacional**: O treinamento de um *Random Forest*, pode ser computacionalmente intensivo, especialmente em conjuntos de dados muito grandes ou com muitos atributos. No entanto, técnicas como a paralelização, podem ser empregadas para acelerar o processo de treinamento (LIAW; WIENER, 2002).
- **Interpretabilidade**: A combinação de várias árvores no *Random Forest*, pode dificultar a interpretação dos resultados, tornando mais desafiador entender o processo de tomada de decisão do modelo.

1.4 Geoestatística

O conhecimento espacial de características de dados georreferenciados, é. hoje, uma das grandes questões que envolvem diversas áreas do conhecimento, desde a agrícola até a área de humanas (MONTEIRO et al., 2004) . A principal ferramenta, utilizada para obter esse conhecimento, é a Geoestatística, que é o ramo da estatística, que associa o conceito de variáveis aleatórias, com o conceito de variáveis regionalizadas, gerando o conceito de funções aleatórias. Diferente da estatística clássica, a Geoestatística tem como pressuposto a dependência entre vizinhos, ou seja, observações próximas exercem influências entre si.

A Geoestatística pode ser entendida como a seção da estatística destinada a relatar o comportamento espacial dos dados, refletindo sobre as variáveis em uma perspectiva local. Essa metodologia busca discriminar a relação espacial de uma variável, provendo ferramentas estatísticas, que possibilitem uma estruturação da aleatoriedade da variável, por intermédio de uma função espacialmente correlacionada (YAMAMOTO; LANDIM, 2015). Idealizada por Daniel Gerhardus Krige (KRIGE, 1951), ao estudar dados de mineração de ouro e observar uma estrutura de dependência dos dados com a localização dos pontos amostrais, essa metodologia foi, posteriormente, formalizada matematicamente pelo francês Matheron (1963), concedendo as noções de Krige uma generalização e extensão a teoria de variáveis regionalizadas.

No contexto dos atributos físicos do solo, a Geoestatística tem sido amplamente empregada para a análise da densidade aparente, textura, porosidade e resistência à penetração (TRANGMAR; YOST; UEHARA, 1986). Estudos como os de CARVALHO; TAKEDA; FREDDI (2003), SOUZA; JÚNIOR; PEREIRA (2004), SIQUEIRA; VIEIRA; CEDDIA (2008), MONTANARI et al. (2010), SALVADOR et al. (2012), RODRIGUES et al. (2017), DALCHIAVON et al. (2017) e NOETZOLD et al. (2019) têm explorado a variabilidade espacial desses atributos em diferentes culturas agrícolas.

No que diz respeito aos atributos químicos do solo, a Geoestatística também desempenha um papel fundamental na análise da distribuição espacial de teores de nutrientes. Estudos como os de CARVALHO; TAKEDA; FREDDI (2003), SOUZA; JÚNIOR; PEREIRA (2004), SIQUEIRA; VIEIRA; CEDDIA (2008), MONTANARI et al. (2010), SALVADOR et al. (2012), RODRIGUES et al. (2017), DALCHIAVON et al. (2017) e NOETZOLD et al. (2019), têm aplicado técnicas geoestatísticas, para investigar a variabilidade espacial de nutrientes como fósforo, potássio, cálcio, magnésio, entre outros.

Uma das principais aplicações da geoestatística em maciços rochosos, é a modelagem e análise da distribuição espacial de teores minerais, como o zinco. Através da construção de semivariogramas, é possível caracterizar a dependência espacial dos teores minerais, identificando estruturas geológicas, como falhas, fraturas, veios, entre outros, que influenciam na sua distribuição (EMERY, 2013). Além disso, a Geoestatística permite a estimativa de recursos minerais por meio de técnicas, como a Krigagem, que considera a correlação espacial dos dados amostrais (CHILES; DELFINER, 2012).

A Geoestatística também desempenha um papel importante na otimização do planejamento de amostragem em maciços rochosos. Por meio da análise de variabilidade espacial, é possível identificar as áreas de maior interesse para a coleta de amostras, levando em consideração a heterogeneidade geológica do maciço. Essa abordagem, permite maximizar a eficiência da amostragem, reduzindo custos e tempo necessários para obter informações representativas (JOURNEL; HUIJBREGTS, 2012).

Além disso, a Geoestatística pode ser aplicada no estudo de atributos geotécnicos,

como resistência à compressão, densidade e permeabilidade do maciço rochoso. A análise espacial desses atributos, é fundamental para o projeto de estruturas de engenharia, como túneis, fundações, barragens, entre outros. Através da geoestatística, é possível mapear a variabilidade desses atributos e tomar decisões informadas sobre o dimensionamento e a estabilidade dessas estruturas (DEUTSCH; JOURNEL, 2011).

Dessa forma, a Geoestatística, aplicada à agricultura de precisão, meio ambiente e à mineração, oferece uma abordagem sólida para analisar a variabilidade espacial dos atributos físicos e químicos do solo, bem como dos recursos minerais em maciços rochosos. Através da modelagem da dependência espacial, por meio dos semivariogramas, é possível obter informações valiosas para a tomada de decisões e o planejamento adequado dessas atividades.

1.4.1 Semivariograma

O Semivariograma fornece um significado preciso do conceito de zona de influência de uma amostra. É uma função dada por uma curva (Figura 2). E é crescente com o aumento da distância h, que separa pares de amostras de tal forma, que, quanto mais distantes as amostras entre si, maior a diferença entre seus teores, e, portanto, menor a continuidade, ou dependência espacial, entre as mesmas (MATHERON, 1963). O Semivariograma ou variograma é uma ferramenta geoestatística, que permite a visualização do comportamento da dependência espacial, em uma determinada área (SANTOS et al., 2011) . O Semivariograma empírico ou experimental, é um gráfico bidimensional, onde o eixo x expressa a distância de separação entre pontos amostrais e o eixo y a estimativa para semivariância. A partir desses pontos, é, então, ajustada uma curva de modelo teórico (semivariograma teórico) e definidos os valores dos parâmetros do semivariograma (efeito pepita, alcance, patamar e contribuição), conforme Figura 2.

Segundo FERREIRA; SANTOS; RODRIGUES (2013), o Efeito Pepita (C_0) representa a descontinuidade na modelagem para pequenas distâncias, que, teoricamente, não deveria existir e é proveniente de fatores não controláveis no processo de coletas de dados e/ou análises. O Alcance (a) é o limiar entre dependência e aleatoriedade no contexto espacial, isto é, o valor limite para o qual a dependência de atributos, em função das distâncias de separação das amostras, pode ser modelada utilizando uma função matemática. O Patamar (C) é o valor correspondente da semivariância para uma distância de separação igual ao alcance. Já a Contribuição (C_1) é definida pela diferença entre Patamar e o Efeito Pepita ($C_1 = C - C_0$).

As estimativas de semivariância para o semivariograma empírico, são geralmente computadas pela fórmula proposta por MATHERON (1963):

Em que:

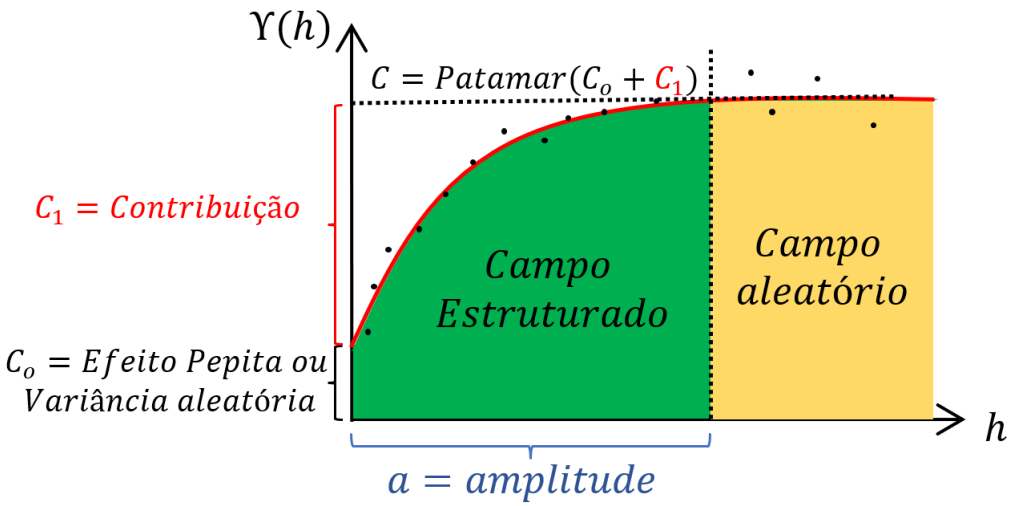


Figura 2 – Exemplo de semivariogramas teórico e empírico

Fonte: Elaborada pelo autor.

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z_i - Z_{i+h}]^2 \tag{1.4}$$

$\hat{\gamma}(h)$ é o estimador de semivariância para uma distância h;

$N(h)$ é o número de pares de pontos, que estão separados por uma distância h;

Z_i o valor da variável no ponto amostrado i;

Z_{i+h} o valor da variável no ponto amostrado $i + h$.

A escolha do melhor modelo teórico, está intimamente ligada as menores distâncias de separação entre amostras. A literatura apresenta diversos modelos teóricos, sendo os modelos esférico, exponencial e gaussiano os mais comumente utilizados. Caso os fenômenos, em estudo, apresentem maior continuidade para pequenas distâncias de separação, o modelo gaussiano é mais indicado, caso contrário, deve-se utilizar o modelo exponencial. Situações intermediárias são satisfatoriamente modeladas pelo modelo esférico.

As equações abaixo, apresentam os modelos exponencial, esféricos e gaussiano, respectivamente:

$$\gamma(h) = \begin{cases} 0, & se \quad h = 0 \\ C_0 + C_1[1 - exp(-\frac{3h}{a})], & se \quad h \neq 0 \end{cases} \tag{1.5}$$

$$\gamma(h) = \begin{cases} 0, & se \quad h = 0 \\ C_0 + C_1[1,5\frac{h}{a} - 0,5(-\frac{h}{a}^3], & se \quad 0 < h < a \\ C_0 + C_1, & se \quad h \geq a \end{cases} \tag{1.6}$$

$$\gamma(h) = \begin{cases} 0, & se \quad h = 0 \\ C_0 + C_1[1 - exp(-\frac{3h^2}{a^2}], & se \quad h \neq 0 \end{cases} \tag{1.7}$$

Os métodos utilizados para o ajuste de semivariogramas, podem ser de dois tipos: feito via semivariograma experimental: métodos de quadrados mínimos ordinários – (OLS – Ordinary Least Square) e ponderados – (WLS – Weight Least Squares); realizado diretamente a partir dos dados: métodos da máxima verossimilhança – (ML – Maximum Likelihood) – e máxima verossimilhança restrita – (REML – Restricted Maximum Likelihood). Nos métodos OLS e WLS, o ajuste das curvas de semivariograma, é realizado ao estimar os parâmetros de covariância e minimizando a expressão de soma de quadrados da diferença entre os valores observados e os estimados. As metodologias ML e REML são técnicas de estimação, onde, a partir de uma amostra, é obtido o estimador mais verossímil dos parâmetros de um certo modelo probabilístico. Após a determinação de todos parâmetros do modelo, juntamente com os pontos amostrados, pode-se, então, obter estimativas para qualquer ponto na área de estudo, de acordo com sua localização.

1.4.2 Krigagem

Para obtenção de mapas, que representem o comportamento de uma determinada variável, após a modelagem estrutural de dependência espacial, deve-se realizar um processo de interpolação a partir dos valores amostrados.

A Krigagem é um método de interpolação, que leva este nome em homenagem a Daniel G. Krige, cientista sul-africano, que foi pioneiro em estudos georreferenciados (FERREIRA; SANTOS; RODRIGUES, 2013). Segundo ISAAKS; SRIVASTAVA (1989), a krigagem é o melhor interpolador linear geoestatístico, pois produz estimativas com a menor variância do erro de estimação. Uma grande vantagem da krigagem, frente aos outros interpoladores, é a possibilidade de obtenção da variância de krigagem ou interpolação, uma medida de erro que é calculada para cada valor estimado, consequentemente, há uma medida de confiabilidade associada ao método (SOARES, 2000).

Dentre as variantes dos estimadores de krigagem, podemos citar: simples (pontual ou em bloco), ordinária, universal, cokrigagem (WEBSTER; OLIVER et al., 1990). A principal diferença entre a krigagem e outros métodos de interpolação, está na maneira como os pesos são atribuídos a diferentes amostras, sendo eles determinados por uma análise espacial, baseada no semivariograma experimental. As krigagens simples e ordinária, são processos estacionários, sendo que, na primeira, considera a média local como um constante igual a média populacional. Já as krigagens universal e cokrigagem são processos não-estacionários, sendo a krigagem simples, mais precisa que a universal (SANTOS et al., 2011).

A krigagem, de uma forma geral, pode ser entendida como parte da família de algoritmos de regressões de mínimos quadrados generalizados (GOOVAERTS et al., 1997) e a estimação é baseada no estimador definido por:

$$Z^*(u) - m(u) = \sum_{\alpha=1}^{n(u)} \lambda_\alpha(u)[Z(u_\alpha) - m(u_\alpha)] \tag{1.8}$$

Em que $n(u)$ é o número de pares de pontos amostrais, $\lambda_\alpha(u)$ são os pesos definidos para os dados $Z(u_\alpha)$, quando considerados como realizações da variável aleatória $Z(u_\alpha)$. Os valores esperados para as variáveis aleatórias $Z(u_\alpha)$ e $Z(u)$ são iguais a $m(u_\alpha)$ e $m(u)$, respectivamente. Com essa perspectiva é possível, então, obter erros de estimação para a diferença das variáveis $Z^*(u) - Z(u)$.

O parâmetro μ é assumido como constante em todo o espaço amostrado, sendo calculado como uma média dos dados. A krigagem simples é empregada para estimar os resíduos, a partir de um dado valor de referência da média, fornecido a priori, sendo, por isso, conhecida como krigagem da média. O número de pontos amostrados, utilizados para estimativa da Equação 1.8, dependerá da distância de influência encontrada no semivariograma (LI; HEAP, 2008).

Na prática, a krigagem ordinária é o tipo mais comum de krigagem. Este tipo de krigagem é semelhante à krigagem simples, todavia substitui-se o parâmetro μ por uma média local, ou seja, o valor do parâmetro será uma média obtida para uma janela de pesquisa, dentro do espaço amostrado a ser considerado. Para tal, força-se $1 - \sum_{i=1}^{N} \lambda_i$ ser igual a zero, o que resultará em $\sum_{i=1}^{N} \lambda_i$ igual a um. A krigagem ordinária calcula a média da constante local e, em seguida, executa a krigagem simples nos resíduos correspondentes e faz uso, para estimar, apenas da média estacionária, obtida para a janela de pesquisa considerada (LI; HEAP, 2008); (WESTERN; GRAYSON; BLÖSCHL, 2002).

No caso mais geral, pode-se utilizar a krigagem ordinária em uma, duas ou três dimensões, pois, a estimativa da krigagem, ainda será uma média ponderada dos dados. Tem-se que, para cada estimativa de krigagem, existe uma variância de krigagem associada, que deverá ser determinada. Assim, precisa-se determinar os pesos que minimizem essas variâncias, cuja soma deverá ser igual a um. Para tal, define-se uma função auxiliar, que contenha as variâncias a serem minimizadas, mais um termo com um multiplicador de Lagrange. A partir de então, define-se as derivadas parciais desta função, com relação aos pesos para zero. Como resultado, obtém-se um valor estimado para o ponto não amostrado, com a menor variância possível (WESTERN; GRAYSON; BLÖSCHL, 2002).

O processo de estimação da krigagem, parte da diferença $Z^*(\mathbf{u}) - Z(\mathbf{u})$, no qual minimiza-se a estimativa da variância do erro de $Z(\mathbf{u})$, que é dado por:

$$\sigma_E^2(u) = Var\{Z^*(u) - Z(u)\} \tag{1.9}$$

Note que, cada estimador de krigagem diferente (simples, ordinária, universal, cokrigagem), implica em um modelo diferente. Uma solução comum, então. é acrescentar um componente ao componente de tendência para representar $Z(\mathbf{u})$.

$$\mathbf{Z}(u) = \mathbf{R}(u) + \mu \tag{1.10}$$

Esse componente $R(\mathbf{u})$ é modelado por meio de uma função, que apresenta estacionaridade, média nula e de covariância (ou semivariância) $C(\mathbf{h})$. Dessa maneira, temos:

$$E[R(u)] = 0 \tag{1.11}$$

$$COV(R(u), R(u+h)) = E(R(u).R(u+h) \tag{1.12}$$

Note que, dessa forma, a esperança de Z no local u, corresponde ao valor médio de u analisado de forma local.

Conforme YAMAMOTO; LANDIM (2015) , para uma variável e um ponto de interesse, é possível obter uma estimava dessa variável na sua respectiva localização:

$$Z^*_{x_0} = \sum_{i=1}^{n} \mu + \lambda_i Z_{x_i} \tag{1.13}$$

Em que μ é a média do atributo, $Z_(x_0)^*$ é o valor estimado do atributo na localização, x_0, λ_i é o peso da amostra na localização, x_i, i é o indexador da amostra, x_i de localização conhecida.

O estimador linear na krigagem simples, denotado por $Z^*_{SK}(u)$, é dado por:

$$Z^*_{SK}(u) = \sum_{(\alpha=1)}^{n(u)} \lambda^{SK}_{\alpha}(u) Z(u_\alpha) + [1 - \sum_{\lambda=1}^{n(u)} \lambda^{SK}_{\alpha}(u)]m \tag{1.14}$$

Em que $\lambda^{SK}_{\alpha}(u)$ são determinados, de tal maneira, que a variância do erro seja mínima $\sigma^2_E(u) = VarZ^*_{SK}(u) - Z(u)$ mediante a condição de não tendenciosidade.

A variância de krigagem é dada pela equação:

$$\sigma^2_{KS} = C(0) - \sum_{\alpha=1}^{n(u)} \tag{1.15}$$

Maiores informações sobre os processos de krigagem, podem ser encontradas em Santos (2010).

Após a modelagem da dependência espacial e interpolação, é necessário validar o modelo geoestatístico (VIEIRA et al., 2000). A validação cruzada, é uma ferramenta que possibilita tal análise. Essa técnica, permite mensurar a incerteza da medição prévia dos dados, que tem por objetivo prever a acurácia do modelo para amostras futuras. Para isso, os dados amostrados são retirados individualmente e estimados pelo modelo utilizando os demais pontos amostrais, consequentemente é possível comparar os valores amostrados com os estimados pelo modelo e obter uma estatística de erro de estimação (ISAAKS; SRIVASTAVA, 1989).

Dentre as estatísticas de erro obtidas, podemos citar: média do erro de estimação, média do erro de estimação padronizada, raiz quadrada do erro quadrático médio, média do desvio padrão e raiz quadrada da média do erro de estimação padronizado ao quadrado (Ferreira, 2015). De acordo com VIEIRA et al. (2000), na autovalidação, devemos ter: média dos erros e dos erros padronizados nula; variância dos erros finita; variância dos erros padronizados unitária.

1.5 *Inverse Distance Weighting* (IDW)

O *Inverse Distance Weighting* (IDW), é uma técnica amplamente utilizada para interpolação, que se baseia na combinação ponderada dos valores conhecidos em pontos de amostra. Essa ponderação é inversamente proporcional à distância entre o ponto a ser interpolado e os pontos amostrados, pressupondo que, pontos próximos, têm valores similares e exercem maior influência sobre o valor interpolado, enquanto, pontos distantes, são menos influentes (SETIANTO; TRIANDINI, 2013); (SEYEDMOHAMMADI; ESMAEELNEJAD; SHABANPOUR, 2016); (MIRANDA, 2017); (GIACOMIN et al., 2014). O cálculo da média ponderada é realizado levando em consideração a distância euclidiana entre o ponto a ser interpolado e seus vizinhos.

A fórmula matemática do IDW, é dada por:

$$Z(\mathbf{u}) = \frac{\sum_{i=1}^{n} w_i(\mathbf{u}) \cdot Z_i}{\sum_{i=1}^{n} w_i(\mathbf{u})} \tag{1.16}$$

Em que $Z(\mathbf{u})$ é o valor estimado no local $\mathbf{u}$, Z_i é o valor amostrado no ponto i, $w_i(\mathbf{u})$ é o peso atribuído ao ponto i em relação ao local $\mathbf{u}$ e n é o número total de pontos amostrados.

Os pesos no IDW, são calculados utilizando a fórmula:

$$w_i(\mathbf{u}) = \frac{1}{d_i^p} \tag{1.17}$$

Em que d_i é a distância entre o ponto amostrado i e o local $\mathbf{u}$, e p é um parâmetro que controla o decaimento do peso com a distância. Valores maiores de p, atribuem mais importância aos pontos próximos, enquanto valores menores de p consideram pontos mais distantes.

O desempenho do método IDW, depende de fatores, como a função de potência, os vizinhos mais próximos e o raio de pesquisa. A função de potência, representada por p, controla a importância dos pontos conhecidos na interpolação, com base em sua distância, até o ponto de saída. A escolha do valor de p é arbitrária e pode ser influenciada pelas características dos conjuntos de dados, como variação, assimetria e curtose. Geralmente, o valor mais comumente utilizado para p é 2, resultando no método conhecido como "distância inversa ao quadrado" ou "distância quadrática inversa" (LI; HEAP, 2008); (EWIS, 2012); (GARDIMAN et al., 2012); (IKECHUKWU et al., 2017). No entanto, é possível escolher o valor de p com base na medição dos erros, visando obter o melhor resultado do IDW. À medida que o valor de p aumenta, a superfície estimada torna-se mais suave.

Essa abordagem determinística (não geoestatística), é adequada para variáveis que apresentam transições graduais, em vez de transições abruptas e limites bem definidos. No entanto, não é recomendada para situações em que as variáveis possuem limites bem definidos e transições bruscas. O IDW é especialmente útil para uma visualização ou interpretação preliminar da interpolação de uma superfície, pois permite uma avaliação geral da continuidade espacial dos dados. No entanto, é importante ressaltar que o IDW não avalia localmente os erros de previsão.

O IDW tem sido amplamente aplicado em diversas áreas, devido à sua simplicidade de implementação e interpretação. Na geologia, por exemplo, o IDW tem sido utilizado para mapear a distribuição de propriedades físicas do solo, como a concentração de minerais e a permeabilidade (DIXON; WEBSTER; LEAKE, 2010). Em estudos de qualidade da água, o IDW tem sido empregado para estimar concentrações de poluentes em locais onde não foram coletadas amostras (LARK, 2001). Além disso, o IDW tem sido utilizado em análises de dados de sensoriamento remoto para estimar características de interesse, como a temperatura da superfície terrestre e a cobertura vegetal (LI et al., 2013).

Embora o IDW seja um método rápido e amplamente utilizado na construção de modelos digitais de elevação do terreno, ele apresenta algumas limitações. Um fenômeno comum, é o chamado *"bull's eyes"*, que ocorre com frequência e resulta em valores elevados na área interpolada. Além disso, quando o tamanho da amostra não é suficientemente grande, o IDW tende a suavizar as curvas das linhas de contorno e pode criar pequenas ilhas isoladas nessas linhas. É importante observar, que o IDW fornece resultados mais satisfatórios, quando há um grande número de pontos de amostra, distribuídos uniformemente na área de interesse (JAKOB; YOUNG, 2016); (ACHILLEOS, 2008); (ACHILLEOS, 2011).

1.6 Modelagem Multidimensional

O termo "dimensionalidade" é usado para referir tanto o número de dimensões do espaço-domínio, onde determinado objeto está definido, podendo ser um espaço unidimensional (1D), bidimensional (2D) ou tridimensional (3D), quanto o número de atributos de um registro de dado, geralmente, referenciado como n-dimensional (nD), multidimensional ou multivariado. Assim, dados multidimensionais são aqueles que podem ser representados como uma tabela de dados multivariados, ou seja, que possui muitas variáveis (ou atributos), que devem ser codificadas em uma única estrutura visual 1D, 2D ou 3D (CARD, 1999). A representação de uma tabela de dados multivariados, é mostrada na Tabela 1, na qual as colunas representam variáveis e as linhas representam um vetor de valores para cada variável.

Tabela 1 – Representação de uma tabela de dados

	Variável 1	Variável 2	...	Variável i
x_1	v_{11}	v_{12}	...	v_{1i}
x_2	v_{21}	v_{22}	...	v_{2i}
...	...	...	...	...
x_j	v_{j1}	v_{j1}	...	v_{ji}

Existem várias técnicas para a visualização de dados multidimensionais, sendo descritas seguindo o critério de classificação proposto por KEIM; KRIEGEL (1996), que difere técnicas de acordo com a forma de mapeamento adotada para transformar dados em formas visuais. Este critério divide as técnicas de visualização multidimensionais, em técnicas de projeção geométrica, iconográficas e orientadas a pixel.

- **Projeção Geométrica** - São técnicas de visualização, que empregam algum tipo de projeção geométrica para mapear dados, para formas visuais. Entre as técnicas desta categoria, estão: Coordenadas Paralelas, Matriz de Scatter Plots, gráficos de linhas, gráfico de barras e histogramas, Survey Plots, Curvas de Andrews, Radviz e Coordenadas Paralelas Circulares.

- **Iconográficas** - A literatura aponta dois tipos de visualizações do tipo iconográficas: glifos e ícones. Em cada um destes, as dimensões de um conjunto de dados, são mapeadas para certas características dos glifos ou ícones. Assim, cada glifo ou ícone, representa um item de dado com suas n-dimensões.

- **Orientadas a Pixel** - Caracterizam-se por mapear o conjunto de valores de cada atributo (ou dimensão) dos dados em pixels na tela. O conjunto de valores de cada atributo, é exibido em janelas individuais, ou seja, para um conjunto de dados que possui n atributos, a tela é dividida em n janelas. Em cada uma das janelas, cada

valor do atributo é representado por um pixel, colorido conforme o valor sendo representado. A distribuição espacial dos pixels na janela, pode ser determinada de diferentes maneiras, de modo que, relações ou significados nos dados, possam ser percebidos pela análise das regiões correspondentes nas janelas (KEIM; KRIEGEL, 1996).

Além das técnicas de visualização, para uma exploração de dados efetiva, é necessário, também, o uso de algumas técnicas de interação. Estas últimas permitem que um usuário possa interagir diretamente com visualizações e alterá-las dinamicamente de acordo com seus objetivos de exploração. A possibilidade de interagir com representações visuais, pode reduzir, consideravelmente, as desvantagens e pontos fracos de algumas técnicas de visualização (principalmente, daquelas que apresentam desordem visual e sobreposição de objetos), fornecendo, ao usuário, mecanismos para manipular a complexidade de conjuntos de dados (OLIVEIRA; LEVKOWITZ, 2003).

KEIM (1997) identificou e categorizou as técnicas de interação, que são utilizadas para a exploração de conjuntos de dados multidimensionais e organizou-as em seis classes:

- **Mapeamento de dados para propriedades visuais** - Esta categoria inclui as técnicas de interação, que realizam mapeamentos de atributos (dimensões) dos dados para parâmetros de uma visualização. Estes parâmetros são atributos visuais de objetos, representando itens de dados (como cor, tamanho, transparência, orientação, etc.).

- **Projeções** - A ideia das técnicas interativas de projeção é mudar, dinamicamente, os atributos de dados (dimensões), que são projetados em eixos de uma visualização. O objetivo é explorar um conjunto de dados multidimensional, permitindo correlações entre diferentes atributos de dados.

- **Filtragem (seleção e consulta)** - Através da seleção direta do subconjunto desejado ou pela especificação de propriedades do subconjunto alvo. A seleção é realizada diretamente sobre uma representação visual, com o auxílio de um dispositivo de apontamento. Sua ênfase é na filtragem rápida, destacando um subconjunto dos dados.

- ***Brushing and linking*** - Referem-se à conexão de duas ou mais visualizações dos mesmos dados, de modo que, uma alteração na representação em uma visualização, afeta a representação na outra.

- **Zoom** - Ao representar, graficamente, conjuntos de dados com um grande número de itens de dados, é importante que estes últimos sejam exibidos de forma comprimida, com o objetivo de fornecer uma visão de *overview* do conjunto inteiro. Mas, ao mesmo tempo, é interessante fornecer também uma visualização diferenciada dos

dados, em diferentes resoluções. Uma visão *overview*, permite ao usuário detectar modelos, correlações e *outliers* no conjunto de dados. Já uma visão com nível de zoom mais alto, permite melhor a exploração de uma área de interesse, pois os itens de dados são exibidos com maior detalhe.

- **Detalhes por demanda** - Representa a possibilidade de obter, interativamente, mais detalhes sobre os dados visualizados. Estes detalhes, podem ser, por exemplo, valores de atributo de um item de dado ou ícone (ou glifo) ou informações adicionais destes.

As técnicas de interação serão as mais aplicadas, no decorrer deste trabalho, para a visualização de dados multidimensionais.

1.7 Classificação de Maciços Rochosos

Tendo em vista a complexidade e o entendimento de conceitos, a princípio é relevante destacar que, conforme apontado por (JAQUES, 2014), a Rocha é o material componente do maciço rochoso, constituído por minerais, e pode se apresentar em grande massa ou em fragmentos. O Maciço rochoso, por sua vez, é um meio descontínuo, formado pelo material rocha e pelas descontinuidades que o atravessam, incorporando a presença de água e o estado de tensões. Apresenta descontinuidades nas escalas megascópica (afloramento) e regional. BELL (2013) afirma que, os maciços rochosos, representam uma parte significativa da superfície terrestre e desempenham um papel crucial em muitos aspectos da geologia, geotecnia e engenharia civil. A compreensão das características geológicas e geotécnicas de um maciço rochoso, é essencial para avaliar sua adequação para uma variedade de aplicações, incluindo construção civil, mineração, túneis e muito mais.

A classificação do maciço rochoso, é uma etapa fundamental na engenharia geotécnica, que visa caracterizar e categorizar as propriedades geológicas e geotécnicas de uma determinada formação rochosa. Essa classificação é essencial para avaliar a adequação do maciço rochoso em projetos de construção civil, mineração, túneis, barragens e outras estruturas geotécnicas. Através de critérios geológicos e geotécnicos, como o tipo de rocha, a estrutura geológica, o grau de fraturamento e as propriedades mecânicas, a classificação do maciço rochoso fornece informações cruciais para o planejamento, projeto e implementação de projetos geotécnicos, bem como para a segurança e estabilidade dessas estruturas em ambientes rochosos. Destacam-se duas das classificações mais amplamente utilizadas na engenharia geotécnica. O RMR (Rock Mass Rating), é um sistema que avalia a qualidade do maciço rochoso, com base em vários parâmetros, enquanto a classificação de Q categoriza o maciço rochoso em classes de qualidade, levando em consideração sua estabilidade e outros fatores relevantes.

A classificação Rock Mass Rating (RMR) foi originalmente desenvolvida visando estabelecer o tempo máximo de auto sustentação de vãos livres de maciços rochosos. Ao longo dos anos, a classificação sofreu modificações, que permitiram estimar valores de resistência, como coesão e ângulo de atrito, aumentando assim a sua gama de aplicações na área mecânica das rochas. Portanto, a compreensão e a aplicação adequada de sistemas de classificação, são elementos-chave na prática da engenharia geotécnica e geologia. Dentro deste estudo, além do RMR, as variáveis Rock Quality Designation (RQD), Litologia (LITO), Fraturamento (FRAT) e Classificação Simplificada (CLASS), foram utilizadas para a construção dos modelos geomecânico.

- **RQD – Rock Quality Designation:** É um índice amplamente utilizado na engenharia geotécnica, para avaliar a qualidade de um maciço rochoso. Ele fornece informações sobre a continuidade e a qualidade das fraturas no maciço rochoso e é, especialmente, relevante em projetos de escavação de túneis, construção de barragens, mineração e outras atividades, que envolvem a interação com maciços rochosos. O RQD é expresso como uma porcentagem e é calculado dividindo o comprimento total das fraturas recuperadas, em uma amostra de testemunho de sondagem, pelo comprimento total da amostra. Quanto maior o valor do RQD, melhor a qualidade do maciço rochoso, indicando que ele é mais contínuo e menos fraturado. RQD é um indicador crítico para determinar a estabilidade do maciço rochoso durante a escavação, a capacidade de suporte de estruturas geotécnicas e a seleção de métodos de reforço apropriados.

- **LITO – Litologia:** É a descrição das características geológicas das rochas em um maciço rochoso, incluindo sua composição mineral, textura, estrutura e origem. É fundamental para entender as propriedades das rochas, como sua resistência, permeabilidade e comportamento mecânico. Diferentes tipos de litologias têm características distintas. Por exemplo, granito é uma rocha ígnea com uma textura geralmente granular e é conhecida por sua alta resistência, enquanto o calcário é uma rocha sedimentar composta, principalmente, de carbonato de cálcio e é relativamente menos resistente e suscetível à dissolução em água. A litologia desempenha um papel crucial na determinação da adequação de um maciço rochoso para determinadas aplicações geotécnicas e influencia as estratégias de engenharia, usadas em projetos, que envolvem interação com as rochas.

- **FRAT – Fraturamento:** Refere à presença e à extensão de fraturas ou descontinuidades em um maciço rochoso. Fraturas são quebras nas rochas, que podem variar em tamanho, orientação e densidade. Elas desempenham um papel fundamental na estabilidade do maciço rochoso e na maneira como ele se comporta sob carga. O fraturamento é avaliado em termos de parâmetros, como espaçamento, orientação, rugosidade, abertura e densidade das fraturas. Maciços rochosos, altamente fraturados,

podem ser menos estáveis e requerer reforços geotécnicos mais significativos. O fraturamento também afeta a permeabilidade do maciço rochoso e a infiltração de água, o que é importante em projetos que envolvem túneis, fundações e barragens.

- **ALT – Alteração:** Refere-se às mudanças que ocorrem nas propriedades de um maciço rochoso, ao longo do tempo, devido a processos físicos e químicos, como intemperismo, decomposição de minerais e interações com a água. É essencial considerar a alteração, ao avaliar a qualidade e a estabilidade de um maciço rochoso, uma vez que, ela pode afetar significativamente as características geotécnicas e a capacidade de suporte do terreno. Portanto, a classificação do maciço rochoso, não apenas leva em consideração a condição inicial, mas também considera como as alterações afetam o comportamento do maciço ao longo do tempo.

- **CLASS – Classificação Simplificada:** Tendo em vista que o banco de dados não havia todas as informações necessárias para a utilização da classificação RMR, pois faltavam informações sobre as descontinuidades e a resistência à compressão uniaxial entre outras, optou-se por utilizar uma classificação RMR simplificada. Na classificação RMR simplificada, empregada neste estudo, o peso, referente à resistência da rocha intacta, foi desconsiderado e no peso associado às condições das descontinuidades, somente o valor relacionado à alteração foi considerado. Além disso, foi admitido o mesmo peso da presença de água para todos os furos de sondagem (já que é sabido que a mina apresenta quantidade significativa de água). Diante disso, a classificação RMR simplificada considerou RQD, fraturamento equivalente ao espaçamento, alteração e a presença de água. (PEREIRA, 2020)

Em resumo, o RQD é um indicador da qualidade do maciço rochoso, a litologia descreve a composição e as características das rochas, o fraturamento considera a presença, orientação e características das fraturas no maciço rochoso, a alteração aponta as mudanças das propriedades do maciço rochoso e a classificação RMR simplificada, é uma adequação, proposta por PEREIRA (2020), para a Classificação RMR, tendo em vista a ausência de informações no banco de dados. Esses fatores são vitais na avaliação e na construção de modelos geomecânicos. Compreendê-los é fundamental para garantir a segurança e o sucesso de projetos de engenharia, que envolvem interações com formações rochosas.

1.8 Solos: Definições e Características Físicas e Químicas

Tendo em vista que, além das pesquisas inerentes aos maciços rochosos, neste trabalho também foi elaborado um estudo sobre a influencia dos atributos físicos e químicos do solo na produção da castanheira-da-Amazônia, sendo assim, este tópico terá a uma exploração sobre os solos, que desempenham um papel fundamental em inúmeras

disciplinas, desde a agricultura, até à engenharia civil e ambiental. Compreender a natureza e as propriedades dos solos é essencial para avaliar sua qualidade, usabilidade e impacto ambiental. Neste contexto, esta seção se propõe a fornecer definições abrangentes de solos, bem como a analisar suas características físicas e químicas, incluindo textura, estrutura, densidade, permeabilidade, composição mineral e reações químicas, que ocorrem no solo. Aprofundar esse conhecimento, é fundamental para uma gestão sustentável dos recursos naturais e o planejamento de projetos de engenharia, que envolvam interações com o solo. Portanto, este tópico será uma contribuição valiosa para a compreensão dos solos, não apenas no decorrer deste trabalho, mas em diversas aplicações científicas e práticas.

O conceito de solo é muito amplo, pois as suas definições dependem da área em que o estuda, conforme é apontado por (LEPSCH, 2016). No entanto, nesta pesquisa, entende-se como definição do solo, as proposições feitas por OLIVEIRA; MOURA (2010), em que o solo é constituído de minerais, originados a partir da desintegração das rochas, de partículas orgânicas, formadas por restos de seres vivos, de água, que dissolve e transporta os nutrientes do solo e de ar, que ocupa os espaços entre as partículas, permitindo a respiração dos microrganismos e das raízes das plantas. E o solo como o material solto e macio, que cobre a superfície da terra e quanto em relação às suas características, tais como cor, quantidade e organização das partículas de que são compostos (argila, silte e areia), fertilidade (capacidade em suprir nutrientes, água e favorecer o crescimento das plantas), porosidade (quantidade e arranjamento dos poros), entre outras características. São constituídos de água, ar, material mineral e orgânico, contendo ainda organismos vivos. Servem como um meio natural para o crescimento das plantas. (COELHO et al., 2013)

1.8.1 Características Físicas dos Solos

De acordo com SCHAFFRATH et al. (2008) e BRADY; WEIL (2009), as características físicas do solo, descrevem a estrutura e o comportamento do solo. Essas características são essenciais para entender como o solo responde a diferentes condições, como a infiltração de água, a compactação, a resistência à erosão e a capacidade de suportar cargas. Algumas das características físicas do solo mais importantes, são:

- **Textura do solo:** A textura do solo refere-se à proporção de areia, silte e argila nas partículas do solo. Ela influencia a capacidade do solo de reter água e nutrientes, bem como a facilidade com que a água pode fluir através dele.
- **Estrutura do solo:** A estrutura se relaciona com a organização das partículas do solo em agregados. Agregados bem formados, podem melhorar a aeração, a infiltração de água e a resistência à erosão do solo.
- **Densidade do solo:** A densidade do solo é a massa de solo por unidade de volume. Isso afeta a porosidade e a capacidade do solo de suportar cargas e raízes das plantas.

- **Porosidade:** A porosidade se refere à proporção de espaços vazios (poros) no solo. Ela influencia a retenção de água e a aeração do solo, essenciais para o crescimento das plantas e a infiltração de água.

- **Permeabilidade:** A permeabilidade é a capacidade do solo de transmitir água. Solos, com alta permeabilidade, permitem a drenagem eficaz, enquanto solos, com baixa permeabilidade, retêm mais água.

1.8.2 Características Químicas dos Solos

As características químicas do solo se referem às propriedades químicas, que descrevem a composição e o comportamento químico do solo. Essas características, são fundamentais para entender como os nutrientes são retidos e disponibilizados para as plantas, bem como para avaliar o potencial de contaminação do solo. Os chamados de macronutrientes são absorvidos pelas plantas em maiores quantidades, tornando deficientes no solo antes dos demais e de micronutrientes, que são geralmente menos deficientes no solo, pois são usados em quantidades menores, porém ambos de grande importância para as plantas (ALFAIA et al., 2018). Assim:

- **Os macronutrientes primários:** Nitrogênio (N), Fósforo (P) e Potássio (K) e os macronutrientes secundários: Cálcio (Ca), Magnésio (Mg) e Enxofre (S) fazem parte de moléculas essenciais, além de possuírem função estrutural nas plantas; já os micronutrientes, como: Cobre (Cu), Ferro (Fe), Manganês (Mn), Molibdênio (Mo) e o Zinco (Zn) são necessários em menor quantidade pelos vegetais e formam as enzimas (DIAS; NEVES; SILVEIRA, 2012). Porém, as plantas precisam ainda de elementos não minerais, que são fornecidos pelo ar e pela água como o Carbono (C), o Oxigênio (O) e o Hidrogênio (H) (LOPES, 1998).

- **Os micronutrientes incluem:** O Zinco (Zn), Cobre (Cu), Boro (B), Manganês (Mn), Ferro (Fe), Molibdênio (Mo) e Cloro (Cl), considerados tão importantes para as plantas quanto os macros, suas deficiências, embora seja bem mais rara, podem ter efeitos extremos sobre a produtividade (MALAVOLTA et al., 1980).

- E o **pH** é uma medida da acidez ou alcalinidade do solo. Ele influencia a disponibilidade de nutrientes para as plantas, uma vez que diferentes nutrientes são mais ou menos solúveis em diferentes faixas de pH. O pH varia de ácido (menor que 7), neutro (igual a 7) a alcalino (maior que 7). (LOPES, 1998)

Referências

ACHILLEOS, G. Errors within the inverse distance weighted (idw) interpolation procedure. *Geocarto International*, Taylor & Francis, v. 23, n. 6, p. 429–449, 2008. Citado na página 34.

ACHILLEOS, G. The inverse distance weighted interpolation method and error propagation mechanism–creating a dem from an analogue topographical map. *Journal of spatial Science*, Taylor & Francis, v. 56, n. 2, p. 283–304, 2011. Citado na página 34.

AGRAWAL, R.; SRIKANT, R. Mining association rules between sets of items in large databases. In: ACM. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. [S.l.], 1993. p. 207–216. Citado na página 22.

ALFAIA, S. S. et al. Princípios agroecológicos para o manejo ecológico do solo e a saúde das áreas produtivas: cartilha para produtores rurais. Editora do Inpa, 2018. Citado na página 41.

AMARAL, W. et al. Using machine learning technique for effort estimation in software development. In: *Proceedings of the XVIII Brazilian Symposium on Software Quality*. [S.l.: s.n.], 2019. p. 240–245. Citado na página 21.

BANIMUSTAFA, A. Predicting software effort estimation using machine learning techniques. In: IEEE. *2018 8th International Conference on Computer Science and Information Technology (CSIT)*. [S.l.], 2018. p. 249–256. Citado na página 21.

BELL, F. G. *Engineering in rock masses*. [S.l.]: Elsevier, 2013. Citado na página 37.

BRADY, N. C.; WEIL, R. R. *Elementos da natureza e propriedades dos solos*. [S.l.]: Bookman Editora, 2009. Citado na página 40.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 23 e 25.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CRC press, 1984. Citado na página 24.

CARD, M. *Readings in information visualization: using vision to think*. [S.l.]: Morgan Kaufmann, 1999. Citado na página 35.

CARVALHO, M.; TAKEDA, E.; FREDDI, O. Variabilidade espacial de atributos de um solo sob videira em vitória brasil (sp). *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 27, p. 695–703, 2003. Citado na página 27.

CHILES, J.-P.; DELFINER, P. *Geostatistics: Modeling Spatial Uncertainty*. [S.l.]: John Wiley & Sons, 2012. Citado na página 27.

COELHO, M. R. et al. Solos: tipos, suas funções no ambiente, como se formam e sua relação com o crescimento das plantas. *MOREIRA, Fatima Maria de Souza; CARES, Juvenil Enrique; ZANETTI, Ronald*, 2013. Citado na página 40.

DALCHIAVON, F. C. et al. Variabilidade espacial de atributos químicos do solo cultivado com soja sob plantio direto. *Revista de Ciências Agroveterinárias*, v. 16, n. 2, p. 144–154, 2017. Citado na página 27.

DEMCHENKO, Y. et al. Addressing big data issues in scientific data infrastructure. In: IEEE. *2013 International conference on collaboration technologies and systems (CTS)*. [S.l.], 2013. p. 48–55. Citado 2 vezes nas páginas 18 e 19.

DEUTSCH, C. V.; JOURNEL, A. G. *Geostatistical Reservoir Modeling*. [S.l.]: Oxford University Press, 2011. Citado na página 28.

DIAS, J.; NEVES, I.; SILVEIRA, V. H. d. Nutrientes do que as plantas precisam. *Periodicidade Trimestral*, 2012. Citado na página 41.

DIXON, S. J.; WEBSTER, R.; LEAKE, J. R. Spatial prediction of soil properties in the scottish highlands using compositional kriging and landform classification. *Geoderma*, Elsevier, v. 155, n. 1-2, p. 3–13, 2010. Citado na página 34.

EFRON, B.; TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. [S.l.]: Chapman & Hall, 1994. Citado na página 25.

EICK, S. G.; KARR, A. F. Visual scalability. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 11, n. 1, p. 22–43, 2002. Citado na página 14.

EMERY, X. *Geostatistics for Engineers and Earth Scientists*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 27.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, v. 96, n. 34, p. 226–231, 1996. Citado na página 22.

EWIS, O. E.-S. Improving the prediction accuracy of soil mapping through geostatistics. *International Journal of Geosciences*, Scientific Research Publishing, v. 2012, 2012. Citado na página 34.

FERREIRA, Í. O.; SANTOS, G. R. dos; RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. *Revista Brasileira de Cartografia*, v. 65, n. 5, 2013. Citado 2 vezes nas páginas 28 e 30.

GARCIA, M.; LEE, D.; CHEN, W. The integration of geospatial data in data science: Challenges and opportunities. In: *Proceedings of the International Conference on Data Science*. [S.l.: s.n.], 2020. p. 78–85. Citado 2 vezes nas páginas 16 e 17.

GARDIMAN, B. S. et al. Análise de técnicas de interpolação para espacialização da precipitação pluvial na bacia do rio itapemirim (es) analysis of interpolation techniques for spatial rainfall distribution in river basin itapemirim (es). *Ambiência*, v. 8, n. 1, p. 61–71, 2012. Citado na página 34.

GIACOMIN, G. et al. Análise comparativa entre métodos interpoladores de modelos de superfí cies. *Revista Brasileira de Cartografia*, v. 66, n. 6, 2014. Citado na página 33.

GOOVAERTS, P. et al. *Geostatistics for natural resources evaluation*. [S.l.]: Oxford University Press on Demand, 1997. Citado na página 31.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer Science Business Media, 2009. Citado 2 vezes nas páginas 22 e 25.

HERLAND, M.; KHOSHGOFTAAR, T. M.; WALD, R. A review of data mining using big data in health informatics. *Journal of Big data*, SpringerOpen, v. 1, n. 1, p. 1–35, 2014. Citado 2 vezes nas páginas 18 e 20.

HUANG, T. et al. Promises and challenges of big data computing in health sciences. *Big Data Research*, Elsevier, v. 2, n. 1, p. 2–11, 2015. Citado na página 18.

IDRI, A.; HOSNI, M.; ABRAN, A. Systematic literature review of ensemble effort estimation. *Journal of Systems and Software*, Elsevier, v. 118, p. 151–175, 2016. Citado 2 vezes nas páginas 21 e 22.

IKECHUKWU, M. N. et al. Accuracy assessment and comparative analysis of idw, spline and kriging in spatial interpolation of landform (topography): an experimental study. *Journal of Geographic Information System*, Scientific Research Publishing, v. 9, n. 03, p. 354, 2017. Citado na página 34.

ISAAKS, E. H.; SRIVASTAVA, M. R. *Applied geostatistics*. [S.l.: s.n.], 1989. Citado 2 vezes nas páginas 30 e 33.

JAKOB, A. A. E.; YOUNG, A. F. O uso de métodos de interpolação espacial de dados nas análises sociodemográficas. *Anais*, p. 1–22, 2016. Citado na página 34.

JAQUES, D. S. Caracterização e classificação de maciços rochosos da mina de volta grande, nazareno, minas gerais. Universidade Federal de Viçosa, 2014. Citado na página 37.

JOHNSON, S. Geospatial intelligence: An overview. *International Journal of Geospatial Data Science*, v. 5, n. 1, p. 45–56, 2016. Citado 2 vezes nas páginas 16 e 17.

JOURNEL, A. G.; HUIJBREGTS, C. J. *Mining Geostatistics*. [S.l.]: Springer Science & Business Media, 2012. Citado na página 27.

KEIM, D. A. Visual techniques for exploring databases. In: *Knowledge Discovery in Databases (KDD'97)*. [S.l.: s.n.], 1997. Citado na página 36.

KEIM, D. A.; KRIEGEL, H.-P. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 8, n. 6, p. 923–938, 1996. Citado 3 vezes nas páginas 14, 35 e 36.

KITCHIN, R. Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, Sage Publications Sage UK: London, England, v. 3, n. 3, p. 262–267, 2013. Citado na página 18.

KITCHIN, R. Big data, new epistemologies and paradigm shifts. *Big data & society*, SAGE Publications Sage UK: London, England, v. 1, n. 1, p. 2053951714528481, 2014. Citado na página 18.

KRIGE, D. G. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, Southern African Institute of Mining and Metallurgy, v. 52, n. 6, p. 119–139, 1951. Citado na página 27.

LARK, R. Estimating concentrations of metals in river sediments from geochemical data: The role of uncertainty. *Environmental and Ecological Statistics*, Springer, v. 8, n. 1, p. 31–50, 2001. Citado na página 34.

LEE, R. *Spatial Analysis Techniques for Geographic Intelligence*. [S.l.]: Springer, 2017. Citado na página 16.

LEPSCH, I. F. *Formação e conservação dos solos*. [S.l.]: Oficina de textos, 2016. Citado na página 40.

LI, J.; HEAP, A. D. A review of spatial interpolation methods for environmental scientists. Geoscience Australia Canberra, 2008. Citado 2 vezes nas páginas 31 e 34.

LI, X. et al. Spatial interpolation of temperature using modis land surface temperature in the tibetan plateau. *Ecological Indicators*, Elsevier, v. 34, p. 551–556, 2013. Citado na página 34.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Citado 2 vezes nas páginas 25 e 26.

LOPES, A. S. Manual internacional de fertilidade do solo. *Tradução e Adaptação*, v. 2, 1998. Citado na página 41.

MALAVOLTA, E. et al. *Elementos de nutrição mineral de plantas*. [S.l.]: Agronômica Ceres São Paulo, 1980. v. 1. Citado na página 41.

MARSLAND, S. *Machine learning: an algorithmic perspective*. [S.l.]: Chapman and Hall/CRC, 2011. Citado 3 vezes nas páginas 20, 21 e 22.

MATHERON, G. Principles of geostatistics. *Economic geology*, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963. Citado na página 28.

MILLER, H. J. The data avalanche is here. shouldn't we be digging? *Journal of Regional Science*, Wiley Online Library, v. 50, n. 1, p. 181–201, 2010. Citado na página 18.

MIRANDA, G. H. B. Análise de amostragem e interpolação na geração de mde. Universidade Federal de Viçosa, 2017. Citado na página 33.

MONTANARI, R. et al. Aspectos da produtividade do feijão correlacionados com atributos físicos do solo sob elevado nível tecnológico de manejo. *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 34, p. 1811–1822, 2010. Citado na página 27.

MONTEIRO, A. M. V. et al. Análise espacial de dados geográficos. *Brasília: Embrapa*, 2004. Citado na página 26.

MORAES, L. J. d. Seletividade ambiental de castanheiras (bertholletia excelsa bonpl.) no estado do amazonas. Universidade Federal do Amazonas, 2021. Citado na página 15.

NOETZOLD, R. et al. Variabilidade espacial da eficiência do uso de potássio e fósforo na cultura da soja 1. *Revista Engenharia na Agricultura*, Revista Engenharia na Agricultura, v. 27, n. 6, p. 529–541, 2019. Citado na página 27.

OLIVEIRA, D. d.; MOURA, L. O solo sob nossos pés. 2010. Citado na página 40.

OLIVEIRA, M. C. F. d.; LEVKOWITZ, H. Interactivity support in visualization tools. *IEEE Computer Graphics and Applications*, IEEE, v. 23, n. 4, p. 10–13, 2003. Citado na página 36.

PEREIRA, L. C. O uso da geoestatística para elaboração de modelos geomecânicos multidimensionais. Universidade Federal de Viçosa, 2020. Citado 2 vezes nas páginas 15 e 39.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986. Citado na página 24.

RODRIGUES, R. A. S. et al. Variabilidade espacial da umidade e das frações granulométricas do solo em um plantio de bananeiras irrigado no semiárido pernambucano. *Conexões-Ciência e Tecnologia*, v. 11, n. 3, p. 134–143, 2017. Citado na página 27.

SALVADOR, M. M. S. et al. Estabilidade temporal e variabilidade espacial da distribuição da armazenagem de água no solo numa sucessão feijão/aveia-preta. *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 36, p. 1434–1447, 2012. Citado na página 27.

SANTOS, G. R. dos et al. Krigagem simples versus krigagem universal: qual o preditor mais preciso? *Energia na Agricultura*, v. 26, n. 2, p. 49–55, 2011. Citado 2 vezes nas páginas 28 e 30.

SCHAFFRATH, V. R. et al. Variabilidade e correlação espacial de propriedades físicas de solo sob plantio direto e preparo convencional. *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 32, p. 1369–1377, 2008. Citado na página 40.

SETIANTO, A.; TRIANDINI, T. Comparison of kriging and inverse distance weighted (idw) interpolation methods in lineament extraction and analysis. *Journal of Applied Geology*, v. 5, n. 1, 2013. Citado na página 33.

SEYEDMOHAMMADI, J.; ESMAEELNEJAD, L.; SHABANPOUR, M. Spatial variation modelling of groundwater electrical conductivity using geostatistics and gis. *Modeling earth systems and environment*, Springer, v. 2, n. 4, p. 1–10, 2016. Citado na página 33.

SIQUEIRA, G. M.; VIEIRA, S. R.; CEDDIA, M. B. Variabilidade de atributos físicos do solo determinados por métodos diversos. *Bragantia*, SciELO Brasil, v. 67, p. 203–211, 2008. Citado na página 27.

SMITH, J. Artificial intelligence and its impact on data science. *Journal of Data Science*, v. 10, n. 2, p. 123–135, 2018. Citado na página 17.

SOARES, A. *Geoestatística para as ciências da terra e do ambiente*. [S.l.]: Instituto Superior Técnico, 2000. Citado na página 30.

SOUZA, Z. d.; JÚNIOR, J. M.; PEREIRA, G. Variabilidade espacial de atributos físicos do solo em diferentes formas do relevo sob cultivo de cana-de-açúcar. *Revista Brasileira de Ciência do Solo*, SciELO Brasil, v. 28, p. 937–944, 2004. Citado na página 27.

TRANGMAR, B. B.; YOST, R. S.; UEHARA, G. Application of geostatistics to spatial studies of soil properties. *Advances in agronomy*, Elsevier, v. 38, p. 45–94, 1986. Citado na página 27.

VIDHYA, A. C. T. *A complete tutorial on decision tree-based modeling (R and Python code)*. 2016. Urlhttps://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/. Citado na página 23.

VIEIRA, S. R. et al. Geoestatística em estudos de variabilidade espacial do solo. *Tópicos em ciência do solo. Viçosa: Sociedade Brasileira de Ciência do Solo*, v. 1, p. 1–53, 2000. Citado na página 33.

WANG, Y.; HAJLI, N. Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, Elsevier, v. 70, p. 287–299, 2017. Citado na página 18.

WEBSTER, R.; OLIVER, M. A. et al. *Statistical methods in soil and land resource survey*. [S.l.]: Oxford University Press (OUP), 1990. Citado na página 30.

WEN, J. et al. Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology*, Elsevier, v. 54, n. 1, p. 41–59, 2012. Citado na página 21.

WESTERN, A. W.; GRAYSON, R. B.; BLÖSCHL, G. Scaling of soil moisture: A hydrologic perspective. *Annual Review of Earth and Planetary Sciences*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 30, n. 1, p. 149–180, 2002. Citado na página 31.

YAMAMOTO, J. K.; LANDIM, P. M. B. *Geoestatística: conceitos e aplicações*. [S.l.]: Oficina de textos, 2015. Citado 2 vezes nas páginas 27 e 32.

2 CONSTRUCTION OF MULTIDIMENSIONAL GEOMECHANICAL MODELS WITH IDW AND USING R LANGUAGE - ELSEVIER ENHANCED READER

Journal of South American Earth Sciences 116 (2022) 103775

Contents lists available at ScienceDirect

Journal of South American Earth Sciences

journal homepage: www.elsevier.com/locate/jsames

Construction of multidimensional geomechanical models with IDW and using R language

Luana Cláudia Pereira [a,*], Gérson Rodrigues dos Santos [b], Eduardo Antonio Gomes Marques [a], Jandressom Dias Pires [c], Rodolfo Renó [d]

[a] Federal University of Vicosa, Civil Engineering Department, Minas Gerais State, Brazil
[b] Federal University of Vicosa, Statistical Department, Minas Gerais State, Brazil
[c] Federal Institute of Northern Minas Gerais, Minas Gerais State, Brazil
[d] Nexa Resources, Rock Mechanics Consultant, Minas Gerais State, Brazil

ARTICLE INFO

Keywords:
Geomechanical model
IDW
R program
Boreholes

ABSTRACT

This article aims to present a methodology for building geomechanical models using existing drillhole data and using a free programming language - the R program. The geomechanical models were developed using the Inverse Distance Weighting (IDW) interpolator and using the RQD variable and a simplified RMR classification, as a basis. Models were built for three sectors of an underground mine located in the state of Minas Gerais, Brazil. In total, data from about 370 drillholes were used, with depths of a few meters up to 650 m. The entire construction of the model was carried out in the R program and, for this, an original script was developed as a result of the research. After building the models for the three sub-sectors previously defined for the mine, the results obtained were compared with data from the technical literature of the region, and have shown to be consistent and reliable with the local reality. Thus, the methodology proposed in this study provides a useful visualization and allows the users to easily interact and evaluate the results and the quality of the information itself. Although, it must be emphasized that the proposed approach needs to be applied and tested in other areas with different rock masses, to assess its applicability in other geological-geomechanical contexts.

1. Introduction

The elaboration of a geomechanical model is not a trivial task, considering that the integration of specific technical knowledge on the subject is required. Specific skills are needed: assessment of whether the data are sufficient in quantity and quality; understanding which variables have the greatest and least influence on the behavior of the studied rock mass; having mathematical/statistical knowledge to make the model itself, and mastery of computational tools to integrate and enable the construction of the model.

The elaboration of a geomechanical model necessarily involves the use of interpolation techniques, given that there will always be places where there will be no information collected in the field (primary information), and it is necessary to know how a given variable behaves. Thus, interpolation techniques are essential to understanding the behavior of a given parameter, from information gathered about it in the area being studied.

A basic assumption is that the information gathered from locations on the surface of the phenomenon, where sampling has been performed, has spatial dependence (autocorrelation) at a satisfactory level, and that the mathematical function used is close to the continuous phenomenon that is analyzed.

Thus, each interpolation method can result in different representations of the same data set. The use of a particular interpolator depends on a priori knowledge, both of the input data set and the intrinsic characteristics of the interpolator. Each interpolator has a particularity and, therefore, its selection must be carefully evaluated before its application (Silva et al., 2007; Achilleos, 2008; Sajid et al., 2013; Setianto and Triandini, 2013).

Development in the field of computer science has given Earth science scientists the opportunity and the possibility to use the interpolation method or technique they want, and the amount of data they want, without worrying about processing time and capacity necessary for the procedure. In addition, scientific research has moved towards other

* Corresponding author.
E-mail addresses: luanac_pereira@hotmail.com (L.C. Pereira), gerson.santos@ufv.br (G.R. Santos), emarques@ufv.br (E.A.G. Marques), jandresson.pires@ifnmg.edu.br (J.D. Pires), rodolfo.reno@nexaresources.com (R. Renó).

https://doi.org/10.1016/j.jsames.2022.103775
Received 26 April 2021; Received in revised form 16 March 2022; Accepted 16 March 2022
Available online 31 March 2022
0895-9811/

factors that affect the interpolation procedure and that have not yet been sufficiently examined (Achilleos, 2008).

Therefore, this article discusses the use of the IDW interpolator for constructing geomechanical models of a rock mass located in an underground mining area, using input data from existing drillholes and using the programming language, R which can be accessed for free. The geomechanical models were built considering two main variables, RQD and a Simplified Rock Mass Classification Value, as it will be later presented. The first variable is continuous while the rock mass class is categorical.

2. Geomechanical models

The elaboration of a geomechanical model of a given rock mass, first, requires the existence of reliable local data, in-depth knowledge of its geological and geomechanical properties, as well as an understanding of how these properties are related, and control the behavior of the rock mass. Once this understanding is acquired, it becomes necessary to spatialize it, interpolate it (using deterministic interpolators, such as the IDW, or geostatistics, i.e.; the different types of Kriging) and visualize it to, subsequently, use it for different purposes. Some authors have already used different interpolators in the field of rock mechanics to develop several models, among them: lithological studies (Rosenbaum et al., 1997) to determine the Rock Quality Designation - RQD (Ozturk and Nasuf, 2002; Ellefmo and Eidsvik, 2009; Esfahani and Asghari, 2013; Ozturk and Simdi, 2014); for Rock Mass Rating (RMR) geomechanical classification (Oh et al., 2004; Stavropoulou et al., 2007; Exadaktylos and Stavropoulou, 2008; Kaewkongkaew et al., 2015; Egaña and Ortiz, 2013; Yi et al., 2014); Ferrari, 2014) and for determine the Geological Strength Index (GSI) (Ozturk and Simdi, 2014; Deisman et al., 2013).

For the determination of RQD, it is possible to highlight some works in the literature, among them: Ozturk and Nasuf (2002) which used Kriging to determine properties of a rock mass and, consequently, how a cutting machine would perform in creating a sewer tunnel in Turkey. Compressive strength, RQD, Joint Compression Strength (JCS) and cut rate were analyzed. Ellefmo and Eidsvik (2009), using drillhole data from a Norwegian iron mine, geostatistical techniques were employed to quantify the local and spatial frequency of the joints and then determine the RQD. Esfahani and Asghari (2013) used sequential Gaussian simulation to identify fractured areas and, later, to define the mining method. The fractured zones were defined by determining the RQD in an apatite deposit in Iran. Regions with RQD below 20% were considered fractured zones, and areas with high values, as being less fractured. Ozturk and Simdi (2014) used the Kriging technique to estimate geotechnical variables in a 13 km stretch of the Istanbul metro. RQD, GSI, uniaxial compression strength and elasticity modules of intact rock and rock mass were determined from data from results of laboratory tests and previous studies. The results obtained were used to define the drilling pattern to be carried out for the subsequent studies of opening the tunnel.

As for the RMR classification, some authors, who determined the classification using interpolators, stand out, among them: Oh et al. (2004) associated resistivity and borehole data to estimate the RMR classification along a tunnel using Kriging techniques. Stavropoulou et al. (2007) made use of geological-geotechnical data from the exploratory phase of a tunnel project to study the variability of the quality of the massif, in terms of RMR. To this end, they associated 3D geological data with geostatistical information and employed Kriging techniques. Exadaktylos and Stavropoulou (2008) used Kriging to assess the spatial variability of the RMR classification in an underground excavation. In addition to the RMR, these authors also investigated the deformation and strength of the rock mass. Kaewkongkaew et al. (2015), using ordinary Kriging, estimated the RMR of a rock mass located in Thailand, based on drillhole data from two previous studies, namely: a storage project located in a region of sedimentary rocks, and another at a dam located in a region of volcanic rock. The results obtained by these authors were compared with data from exploratory tunnels and, in the first case, the results were satisfactory. In the region of the dam, the condition of complex geology interfered negatively in the results. Egaña and Ortiz (2013) used one of these estimates to determine the RMR in an underground mining project located in northern Chile. The methodology consisted of using Gaussian simulation to individually estimate all the variables that make up the RMR classification, namely resistance to uniaxial compression, fracture frequency, RQD, joint condition and water. The results obtained were compared using the jack-knife technique, and compared with the average RMR values of the geotechnical units. The results of the methodology were satisfactory. Yi et al. (2014) estimated the RMR values in a rock mass in South Korea, using data from boreholes and electrical resistance. Indicative Kriging was used. Ferrari (2014) estimated the RMR classification of a rock mass in Italy, using geostatistical techniques. To this end, geomechanical field surveys were carried out to characterize the rock masses and, subsequently, to define their class. These authors concluded that geological and structural history is what defines the quality of a rock mass.

Pinheiro et al. (2016) state that for determining the dimensions of geotechnical mining works, geomechanical parameters of rock formations are defined based on drilling campaigns, in situ characterization tests and laboratory tests. According to the results of these campaigns, geotechnical zoning is established and a set of geomechanical parameters is assigned to each location. This procedure is very subjective, but it is important in the initial phases of the project. However, issues such as intrinsic spatial variability and heterogeneity of rock masses are not duly considered in the formulation of this zoning. Thus, the elaboration of geomechanical models is a tool to meet this demand. In addition, models can assist in the representation of geological information that, as mentioned by Schneeberger et al. (2017), is inherently three-dimensional, but it is usually represented in two-dimensions.

Pakyuz-Charrier et al. (2018) argue that 3D geological models aim to represent an extremely complex reality in a condensed format for the end user. Schneeberger et al. (2017) report that, with the increase in computational capacity, 3D modeling has spread, as they can be used to build/manipulate models on ordinary computers. Thus, the development and use of these geomechanical models has seen a significant increase in Geotechnics in recent years, having been greatly driven by advances in computer science, both in programs and in hardware. Despite this, according to Hack et al. (2006) the implementation of numerical models in the area of geomechanics is still time-consuming, mainly due to two factors: the first is related to the lack of input data in sufficient quantity and quality to construct the models. The second is the need to use more than one program to carry out all the necessary activities in a geotechnical project with the inherent difficulties regarding compatibility among these programs.

3. Inverse Distance Weighting (IDW)

Inverse Distance Weighting (IDW) uses the combination of all values, weighted and inversely proportional to the distance from the sampled point to the point to be interpolated. Thus, this interpolator considers that closer points have similar values and have greater influence on the value to be interpolated, while more distant points are independent and have less influence on the result. The calculation of the average is weighted by the Euclidean distance between the point to be interpolated and its neighbors. Thus, normally, the weight attributed to the distance is adjusted by an exponent; thus, the greater the exponent, the greater the influence of distance. The user can predetermine this weight, and when assigning higher values, the less the influence of the more distant points will be, and vice versa. The best interpolation results are obtained when the sampling is dense, in relation to the attempted simulation of the local variation, whereas if the sampling is insufficient or the spatial distribution is inappropriate to the characteristic of the surface to be modeled, the results may form an insufficient representation (Setianto

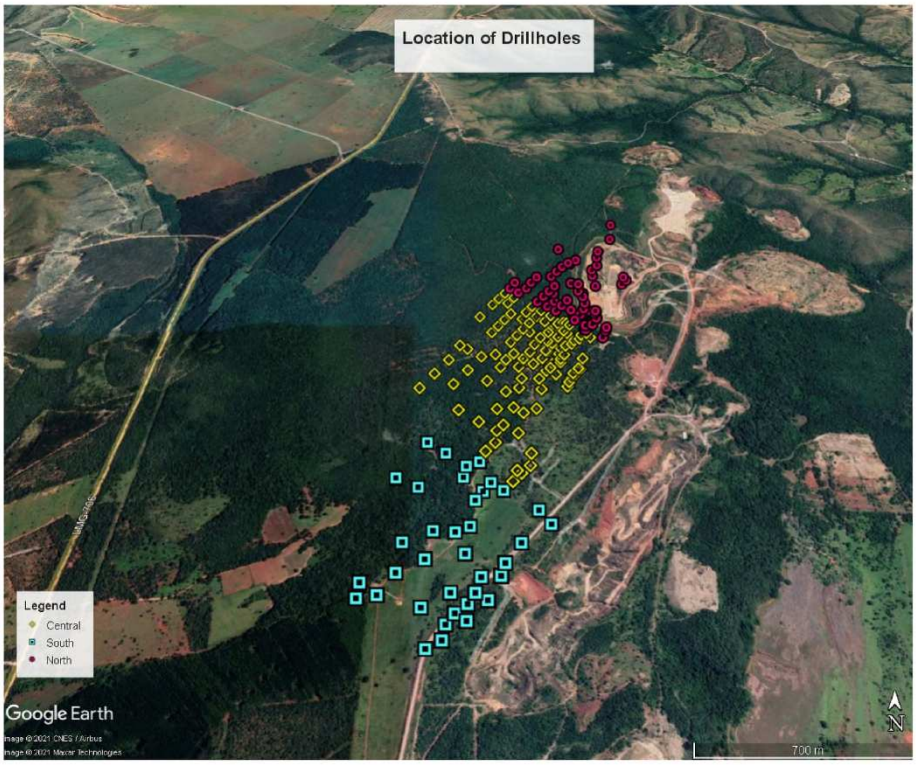


Fig. 1. Location of the drilling holes in the mine.

and Triandini, 2013; Seyedmohammadi et al., 2016; Miranda, 2017).

The performance of the IDW method depends on the exponential function, closest neighbors and search radius. The power function, p, controls the significance of the known points in the interpolated values, based on their distance from the exit point. The choice of the p-value is arbitrary and the coefficients of variation, asymmetry and kurtosis of the data sets can affect the optimal p-value. The most popular choice for p is 2 and the resulting method is often called inverse square distance or inverse squared distance. The power parameter can also be chosen based on the measurement of errors (for example, minimum absolute mean error, resulting in the ideal IDW). The smoothness of the estimated surface increases as the power parameter increases. It was found that the estimated results become less satisfactory when $p = 1$ or 2, compared to when p is equal to 4. The IDW will never predict values above the maximum measured value or below the minimum measured value (Li and Heap, 2008; El-Sayed, 2012; Gardiman Junior et al., 2012; Sajid et al., 2013; Karami and Afzal, 2015; Ikechukwu et al., 2017).

IDW is a relatively quick and widely used method for making digital terrain elevation models. However, it presents certain deficiencies, for example, what is known as the "Bull's Eyes Phenomenon", which frequently appears to generate high values in the interpolated area. In addition, it tends to smooth out the contour line curves when the sample size is not large enough, and also sometimes creates small isolated islands of contour lines. It is important to mention that the IDW provides satisfactory results when the number of sampling points in an area is large and is uniformly distributed (JAKOB; YOUNG, 2006; Achilleos, 2008; Achilleos, 2011). However, in the area of rock mechanics, specifically for the construction of geomechanical models, IDW is not widely used.

4. Study area and methodology

For the present research, data from existing drillholes in a sector of an underground mine located in the state of Minas Gerais, Brazil, called the Extreme North were used. The mine is, in turn, subdivided into three sectors, South, Central and North. 70 drillholes were used for the South sector, 255 holes for the Central and 149 holes for the North sector, Fig. 1. For all the sectors there were eight quantitative variables - geographic coordinates, depth, local geomechanical classification, simplified RMR classification, RQD, fracturing and alteration - and a qualitative variable - lithology.

Vazante mine is located in a shear zone, throughout mineralized fluids percolated, so originating the crystallization of zinc ore and associated minerals. The mine is inserted in the so-called Vazante Group, with controversial age, in which great deformations associated to variable metamorphic degrees can be observed. The host rocks of zinc ore are mainly pink dolomites interbedded with metapelitic – generally slate and marl; grey dolomites, also interbedded with metapelitics and marls; and dolomitic breccia and siderite and ankerite. Pink dolomites belong to Serra do Poço Verde Formation, Pamplona Member and are found preferably at hanging wall of Vazante fault; while the grey dolomites (Morro do Pinheiro Member) are found on the footwall of the fault. The combination of metamorphised lithologies, highly deformed, with high geotechnical complexity, associated with an expressive hydrogeological system associated to a karstic and structural region (specially the structures with NW-SE direction) provide the ideal conditions to failure of rock masses. This geological complex environment, intensified by weathering and karstification, generates block fall and sliding during mine operation and cavities formation, which can hold weathered material and water under pressure (Dardene, 2000; Bhering, 2009; BITTENCOURT; REIS NETO, 2012).

Before using the IDW interpolator, a database treatment was conducted, in order to correct discrepancies, such as negative RQD values. In addition, a previous study was carried out, via machine learning, to verify which variable (or variables) most affected the geomechanical behavior of the rock mass of each of the sectors of the Extreme North Mine. As a result, among the variables considered, the RQD variable was identified as the one that best explained the geomechanical behavior of the rock masses under study. Thus, multidimensional geomechanical models were built for RQD, a continuous variable, and for a simplified

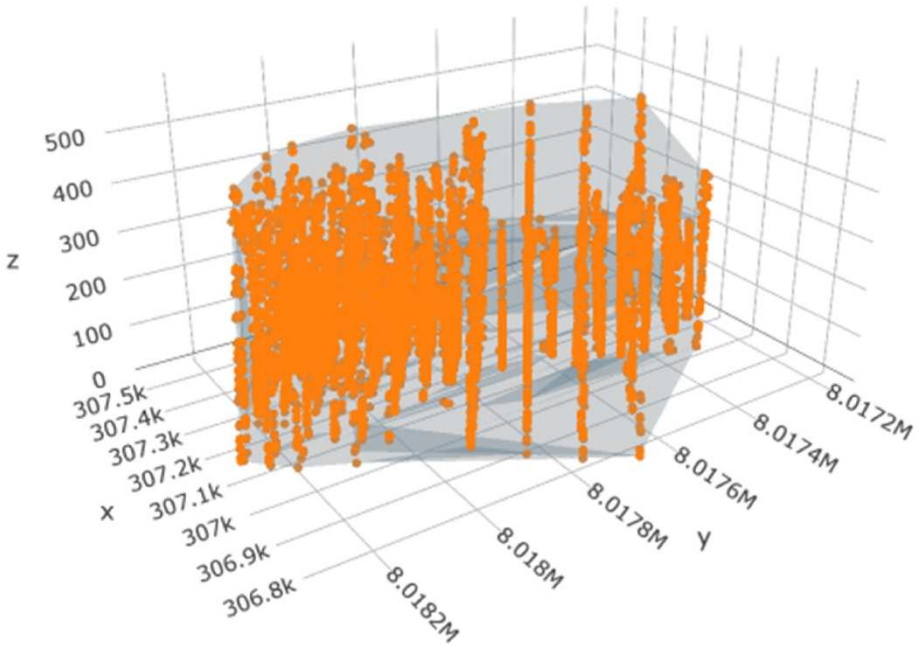


Fig. 2. Example of a view of the volume of the geomechanical model grid with the position of drillholes used on modelling.

Table 1
Information about grids sectors.

	Axes						Area (km^2)
	X		Y		Z		
	From (m)	To (m)	From (m)	To (m)	From (m)	To (m)	
South	306,570	307,315	8,016,535	8,017,415	0	640	0.7
Central	306,740	307,540	8,017,190	8,018,290	0	580	0.6
North	307,135	307,790	8,018,010	8,018,890	0	515	0.6

RMR classification using IDW, a categorical variable.

The simplified RMR classification, whose name given in the present study to the variable was CLASS, was made in order to categorize the rock mass, given that the classification provided in the database from the drillholes was developed locally. However, as the database did not include all the information necessary to use the aforementioned classification, lacking information on discontinuities and uniaxial compressive strength, among other factors, it was decided to use a simplified RMR classification.

In the simplified RMR classification used in the present study, the weight related to the resistance of the intact rock was disregarded and, in the weight associated with the conditions of the discontinuities, only the value related to alteration was considered. In addition, the same weight in regard to the presence of water was conceded for all drillholes (equal to 4, since it is known that the mine has a significant amount of water). Therefore, the simplified RMR classification considered RQD, fracturing equivalent to spacing, alteration and the presence of water.

In making the geomechanical model, a script was developed in the R programming language, in which all analyzes were made. In the script, after reading the database, multidimensional analysis began by creating a regular cubic visualization grid. In this grid, the x and y-axes are the geographic coordinates and the z-axis is the depth of the borehole in meters, with the source of this information coming from the boreholes drilled in the region. It is noteworthy that the dimensions of the grid were defined based on the geographic location of the holes, that is, each sector has its own analysis dimensions. One example of the view of the grid with the drillholes is shown in Fig. 2.

In Table 1 there is information about the grids of the mining sectors – South, Central and North –and the R packages used in the model is presented in Table 2.

5. Results and discussion

The traditional Rock Mass Rating (RMR), proposed by Bieniawski (1974) and latter modified by the author (1989) provides a general evaluation of rock masses mechanical behavior through a final score varying between 0 to 100, where the most higher values are associated with higher rock mass strengths. This range is based on 5 (five) different parameters: uniaxial compressive strength, RQD, discontinuities spacing, discontinuities condition (opening, rugosity, type of filling, weathering and strength of discontinuities walls) and water condition. For each parameter there is a specific evaluation, quantitative and qualitative, resulting in a numerical value, and the final RMR value is given by the sum of the values for each one of these five parameters. A more detailed description of RMR can be found in the work of Bieniawski (1989). On the present research, the variable CLASS, is a simplification of the general RMR, as some of the parameters previously described were not available on drillholes description.

The simplified RMR proposed for the rock mass of the Extreme North Mine can be seen in Table 3. In this, the values of the RQD, fracturing and the presence of water weights from the simplified classification, developed in the present research, were kept equal to the original RMR classification, as well as in relation to the weight of weathering. However, the other values were ignored. Thus, once the weights were added, they were totaled, and the intervals for the simplified RMR classification were obtained – for the rock masses with a sum inferior to 21, a value of RMR equal to 5 – very poor rock mass, was used; Modified RMR between 22 to 29 were classified as Class 4 – poor; values of modified RMR

Table 2
R packages used in the elaboration of geomechanical models.

Package name	Description	Reference
sp	Classes and methods for spatial data; the classes document where the spatial location information resides, for 2D or 3D data.	Pebesma and Bivand (2005)
spacetime	Classes and methods for spatio-temporal data, including space-time regular lattices, sparse lattices, irregular data, and trajectories; utility functions for plotting data as map sequences (lattice or animation) or multiple time series.	Pebesma (2012)
raster	Reading, writing, manipulating, analyzing and modeling of spatial data. The package implements basic and high-level functions for raster data and for vector data operations such as intersections.	Hijmans (2020)
rgeos	Interface to Geometry Engine - Open Source ('GEOS') for topology operations on geometries.	Bivand and Rundel (2020)
rgdal	Provides bindings to the 'Geospatial' Data Abstraction Library and access to projection/transformation operations from the 'PROJ' library.	BIVAND; KEITT; ROWLINGSON (2019)
lattice	A powerful and elegant high-level data visualization system inspired by Trellis graphics, with an emphasis on multivariate data.	Sarkar (2008)
moments	Functions to calculate: moments, Pearson's kurtosis, Geary's kurtosis and skewness	Komsta and Novomestky (2015)
plotKML	Writes sp-class, spacetime-class, raster-class and similar spatial and spatio-temporal objects to KML following some basic cartographic rules.	Hengl et al. (2015)
GSIF	Global Soil Information Facilities - tools (standards and functions) and sample datasets for global soil mapping.	Hengl (2020)
geoR	Geostatistical analysis including variogram-based, likelihood-based and Bayesian methods.	Ribeiro et al. (2020)
plotly	Create interactive web graphics from 'ggplot2' graphs and/or a custom interface to the (MIT-licensed) JavaScript library 'plotly.js' inspired by the grammar of graphics.	Sievert (2020)
DescTools	A collection of miscellaneous basic statistic functions and convenience wrappers for efficiently describing data.	Signorell et al. (2020)
readxl	Import excel files into R.	Wickham and Bryan (2019)
psych	A general purpose toolbox for personality, psychometric theory and experimental psychology.	Revelle (2019)
ggplot2	A system for 'declaratively' creating graphics, based on "The Grammar of Graphics".	Wickham (2016)
dplyr	A fast, consistent tool for working with data frame like objects, both in memory and out of memory.	Wickham et al. (2020)
caret	Misc functions for training and plotting classification and regression models.	Kuhn (2020)
corrplot	Provides a visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables.	Wei and Simko (2017)
spatstat	Comprehensive open-source toolbox for analyzing Spatial Point Patterns.	Baddeley et al. (2015)
maptools	Set of tools for manipulating geographic data.	Bivand and Lewin-Koh (2019)
scatterplot3d	Plots a three dimensional (3D) point cloud.	Ligges & Mächler (2003)
tcltk2	A series of additional Tcl commands and Tk widgets with style and various functions to supplement the tcltk package	Grosjean (2019)
doParallel	Provides a parallel backend.	MICROSOFT (2019)
GGally	Plotting system based on the grammar of graphics.	Schloerke et al. (2020)
e1071	Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, generalized k-nearest neighbor.	Meyer et al. (2019)
mlbench	A collection of artificial and real-world machine learning benchmark problems.	Leisch and Dimitriadou (2010)
MASS	Functions and datasets to support Venables and Ripley.	Venables and Ipley (2002)
Lahman	Provides the tables from the 'Sean Lahman Baseball Database' as a set of R data.frames.	Friendly (2019)
htmlwidgets	A framework for creating HTML widgets that render in various contexts including the R console	Vaidyanathan et al. (2019)

between 30 to 37 were classified as Class 3 – reasonable (regular) rock mass; values between 38 to 48 – Class 2 (good rock mass); and finally, values between 49 to 61 defines a very good rock mass, Class 1.

RQD and the CLASS variables (simplified RMR classification) were interpolated using the IDW for all three sectors studied. The p-value equal to 2 is the most usual value, as cited by Babak (2014). Alghanlandis & Afzal (2011) report that an exponent equal to 2 is considered an optimal exponent for IDW interpolation. Despite that, we have tested different p-value, between 1 to 5 (sensitivity analysis). On South Sector, for both variables, the lower errors values for RMSE were associated to p-value between 1 and 2. Maps were generated with these values, but the outcome did not show good results when compared to regional geology information, as observed for p-values equal to 2. Based on this sensitivity analysis, and as for the other two sectors the best results were obtained by using p-value equal to 2, this value was defined for all sectors to be used in IDW interpolation.

As a result of the interpolation, a 3D model was obtained with two different views - a horizontally sliced view, in which a series of maps were generated for different depths and a multidimensional view, which is the 3D model itself.

It is noteworthy that the entire explanation of the results of the models was made using the images generated for the plan model, for the sake of ease. Visualizing a multidimensional model on a plane of paper is difficult, as it cannot be rotated, for example. Thus, it was decided to make all of the explanations using the planned model. However, for the day-to-day activities of a mining company, the multidimensional graphics generated will be more practical and usual than the planned model, which is presented as complementary material. It is also notable that some directions of discontinuities (based on technical literature) were indicated by solid lines in the images of the models presented.

5.1. South sector

Fig. 3 shows the results of the interpolation of the RQD variable by the IDW interpolator for the South sector. It is noted that there are low values (<40) of RQD up to a depth of 530 m, and up to a depth of 460 m these can be observed in an expressive way. For the first 170 m of depth, it is observed that the low RQD values have a preferential spatial orientation - NE-SW - and from the surface to the depth of 50 m this region has a larger area. It is noteworthy that this direction is in line with the local and regional structural geology, coinciding with the same direction of Vazante fault, as will be demonstrated later. Between depths of 180 and 290 m it is possible to observe that the low RQD values

Table 3
Simplified RMR classification (CLASS variable) elaborated in the present research.

Parameter		Range of values				
1	Strength of intact rock material	Not considered				
2	RQD	91–100	76–90	51–75	26–50	<25
	Rating	**20**	**17**	**13**	**8**	**3**
3	Fracturing	>2 m	0,6-2 m	200–600 mm	60–200 mm	<60 mm
	Rating	**20**	**15**	**10**	**8**	**5**
4	Weathering	Unweathered	Slightly weathered	Moderately weathered	Highly weathered	Decomposed
	Rating	**6**	**5**	**3**	**1**	**0**
5	Groundwater	Completely dry	Damp	Wet	Dripping	Flowing
	Rating	**15**	**10**	**7**	**4**	**0**
Global Rating		**49–61**	**38–48**	**30–37**	**22–29**	**< 21**
Simplified RMR Classification (CLASS variable)		**1**	**2**	**3**	**4**	**5**

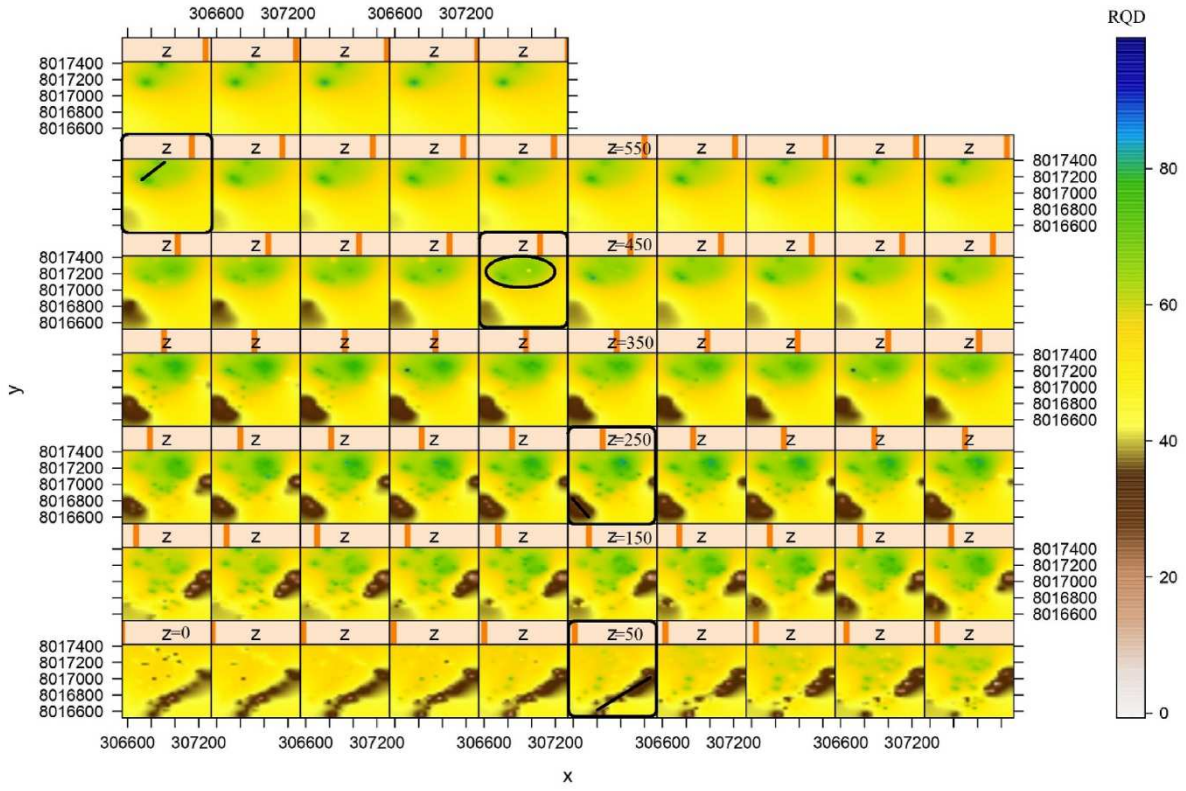


Fig. 3. Geomechanical model of the RQD variable obtained using the IDW interpolator - South sector.

remain, however in a much smaller region of the maps.

From a depth of 170–490 m, it is possible to visualize the tendency of a second direction of low RQD values – along the NW-SE axis; this orientation is also in agreement with the regional structural geology, since a family of fractures is observed in this direction in the Vazante mine massifs. Its greatest representation is between depths of 190 and 390 m. Thus, between 170 and 280 m the presence of the two directions mentioned occurs. Still, at a depth of 250 m, it is noted that for the northernmost region of the geomechanical model there are nuclei that have high RQD values (>80), and this behavior can be observed at a depth of 210 m. However, at depths around 440 m and 450 m there is a presence of low RQD values within these nuclei. Finally, from the depth of 460 m to the end of the model at a depth of 640 m, the SW preferential trend is, again, distinguishable, although only at the left end of the geomechanical model.

It has been observed that after 530 m of depth the only information available comes from the center-north end of the model, with the remainder showing only the average value (RQD between 40 and 60). This is because, as observed in the exploratory analysis, most drillholes reach, on average, up to 400 m in depth. Therefore, below this depth, there is not much information that allows the construction of the model. In view of the reduction of data from the depth cited, the interpolator presents an average value for these locations, thus indicating that the model's interpolation quality is high.

Fig. 4 shows an image of the 3D geomechanical model (x, y, z and RQD) considering the RQD variable obtained with the use of the IDW interpolator, making it possible to visualize regions with low RQD values. In this model, it is possible to view the x and y-axes, which are the geographic coordinates and depth - this information appears when the mouse hovers over the model; - the RQD value can be checked using the grid colors. It is possible to view RQD information across the grid for which the North-South sector model was made, as well as rotate the model and zoom in and out. Evidently, the results are the same as shown in Fig. 3. It is emphasized that the construction of this model, with the image quality that it presents, combined with the ease of viewing the information, the good interactivity and, mainly, the fact of that it had been generated through the use of a free programming language (the R) is something unique. In addition, the methodology for creating this model allows the user to evaluate the entire interpolation calculation process, that is, it is not a "black box" that releases a final result without

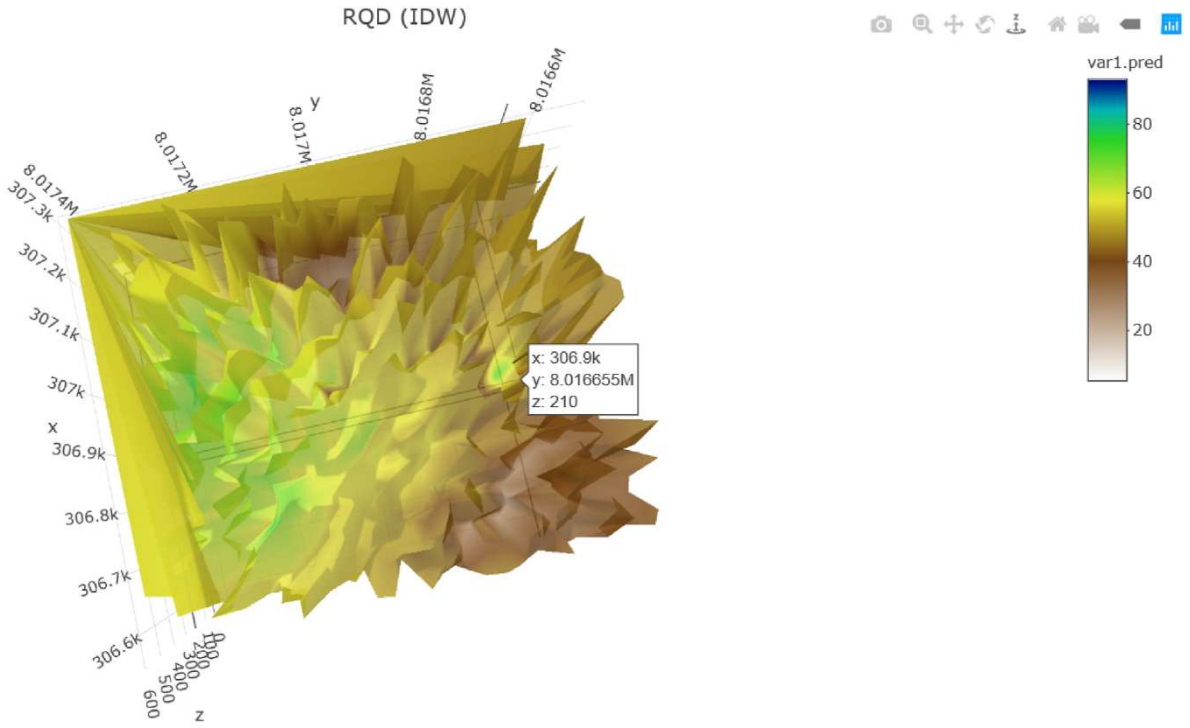


Fig. 4. Visualization of the 3D graph of the geomechanical model of the RQD variable using the IDW - South sector.

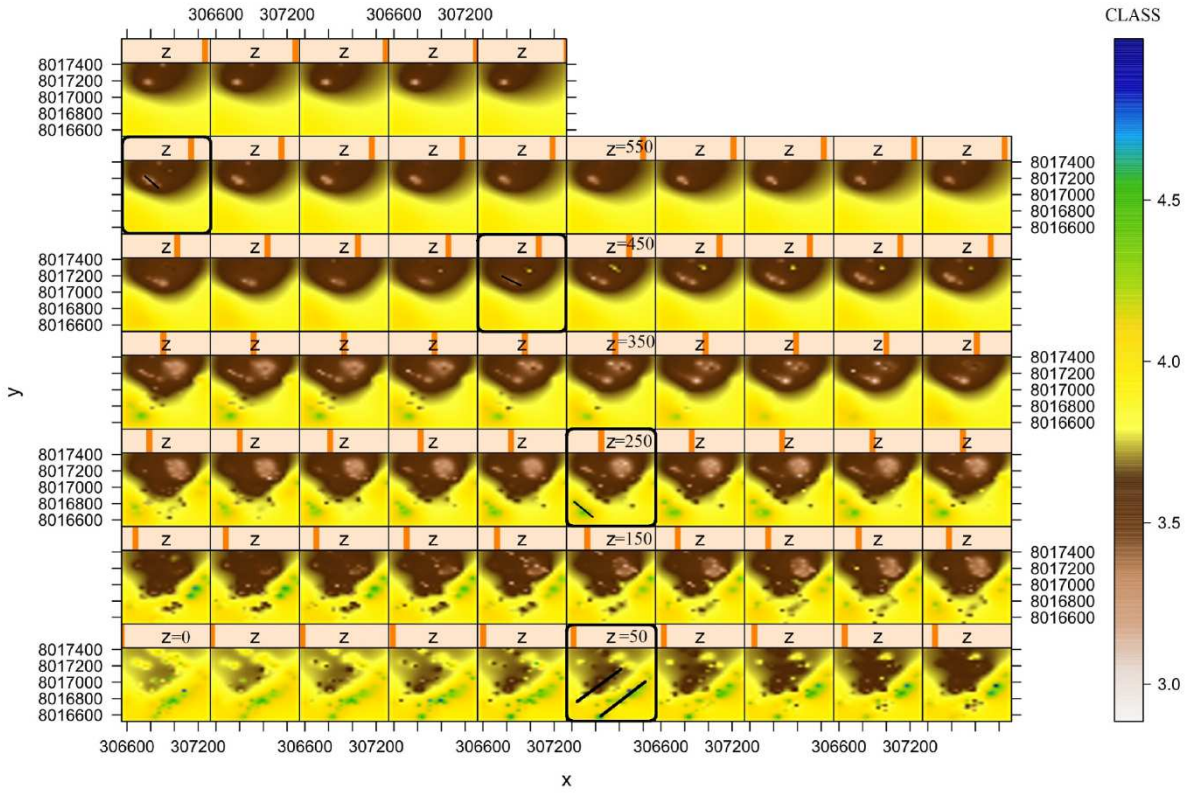


Fig. 5. Geomechanical model of the CLASS variable obtained using the IDW interpolator - South sector.

permitting analysis of how it was obtained.

For the geomechanical model of the simplified RMR classification variable (CLASS), presented on Fig. 5, it is noted, in general, a rock mass with a rating around 4.0 (CLASS equal to 4), that is, a rock mass with reasonable to low mechanical resistance. From the surface up to 140 m in depth, there is an alignment of points with NE-SW direction, with a poor geomechanical classification (CLASS equal to 4.5). After 140 m of depth, one can see points, without alignment, within the same aforementioned classification Additionally, after 210 m, there is an inversion of the alignment of the points classified as bad. The points start to show

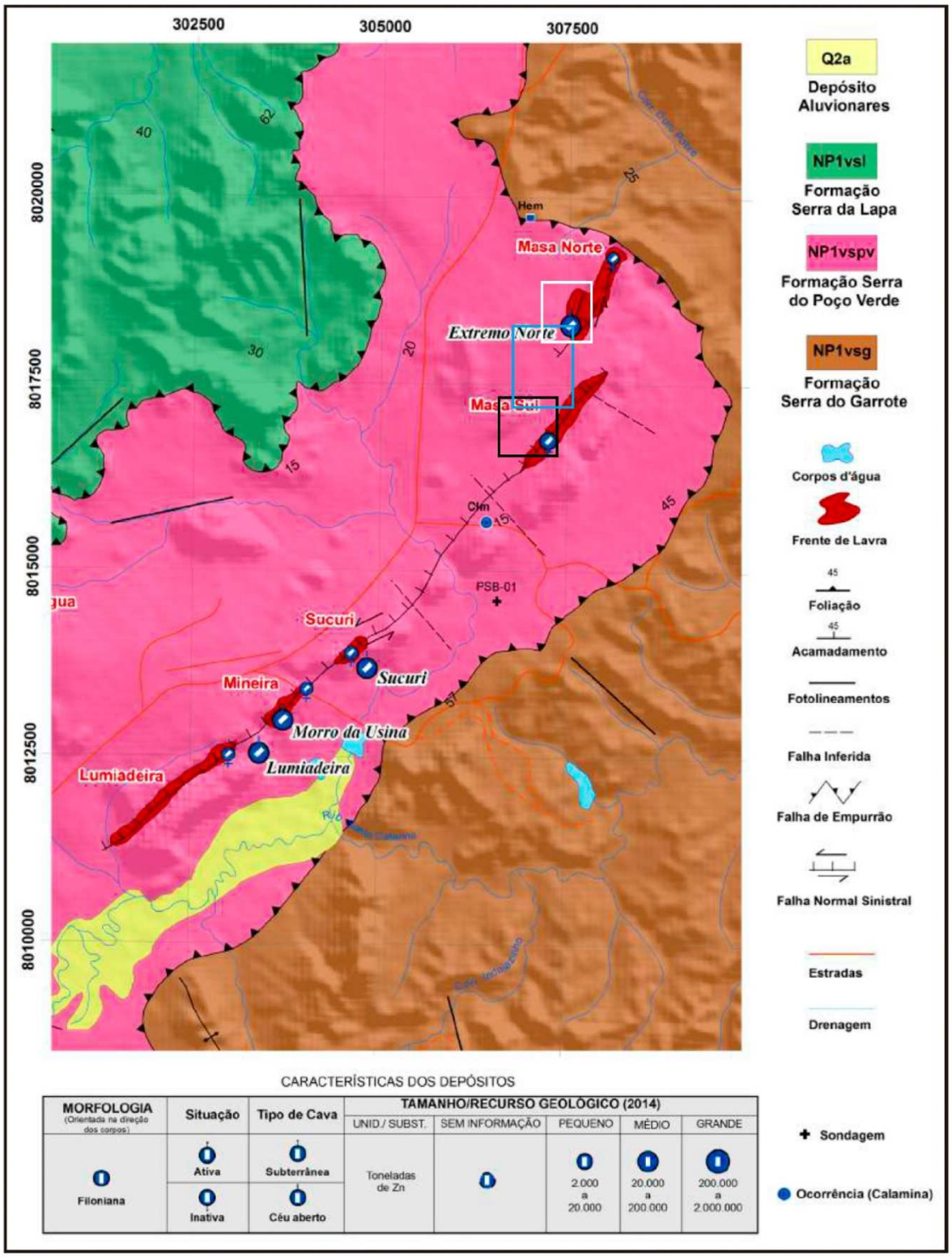


Fig. 6. Metallogenic map of Vazante Zinc District with indication of subsectors; scale 1:75.000. South sector area is indicated in black, the Central sector in blue, and the North sector in white (CPRM, 2015).

alignment with the NW-SE direction up until 290 m depth, again, showing agreement with the structural geology of the area. Between 300 and 370 m in depth, points of CLASS equal to 4.5 can be observed without orientation.

From a depth equal to 20 m there is an alignment, in a NE-SW direction, of points of better rock material, identified as CLASS between 3 and 3.5. This alignment continues in depth and gives rise to a more notable region starting at 80 m, and it is even possible to notice such a region until the final depth of the model. The depths at which the region is most expressive occur between 80 and 350 m. Within this region of better geomechanical material, it is possible to notice an alignment of points with a NW-SE direction, of material with a geomechanical class

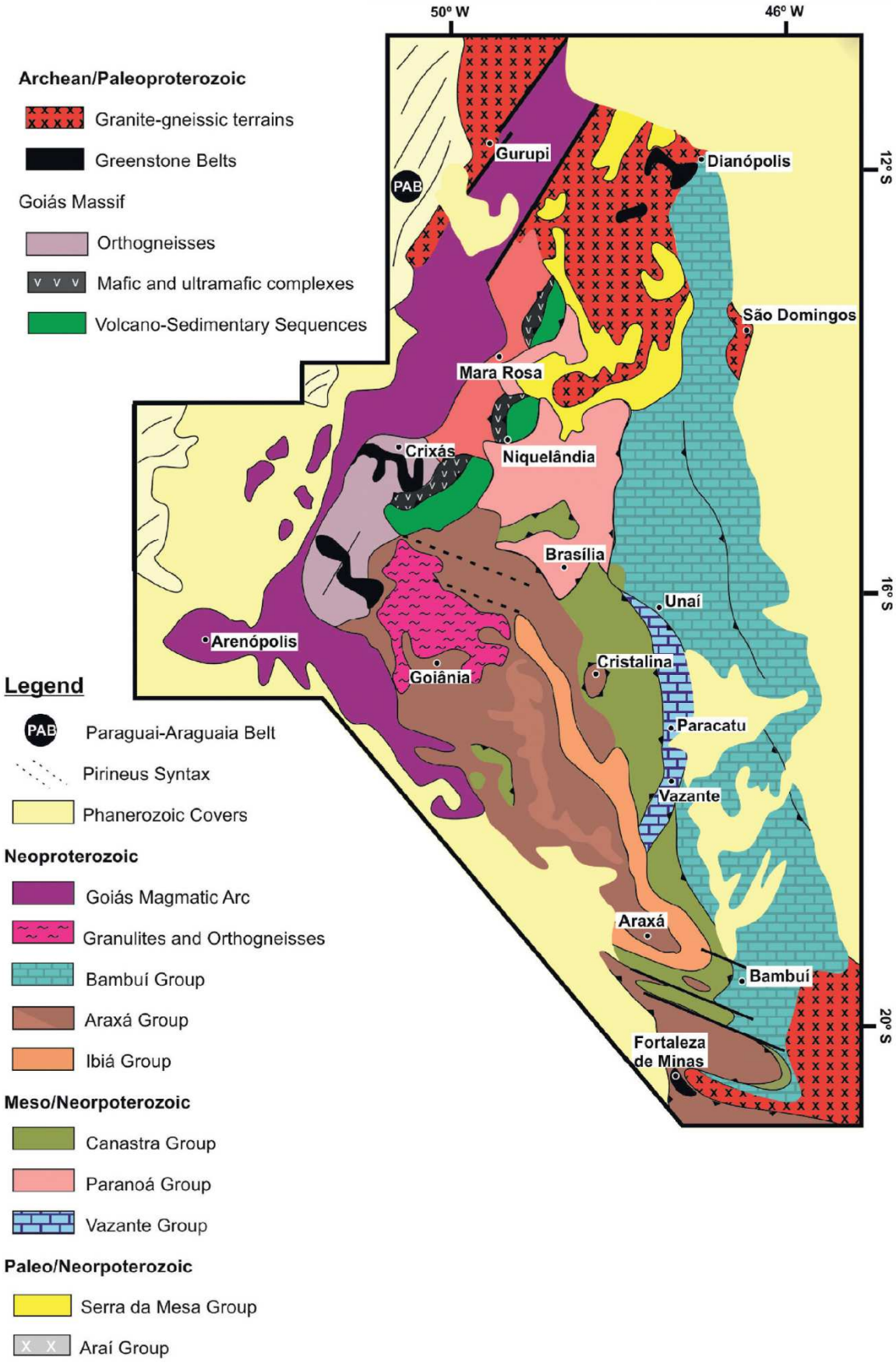


Fig. 7. Geotectonic configuration of southern portion of Brasília belt. See Vazante area (Monteiro, 2002).

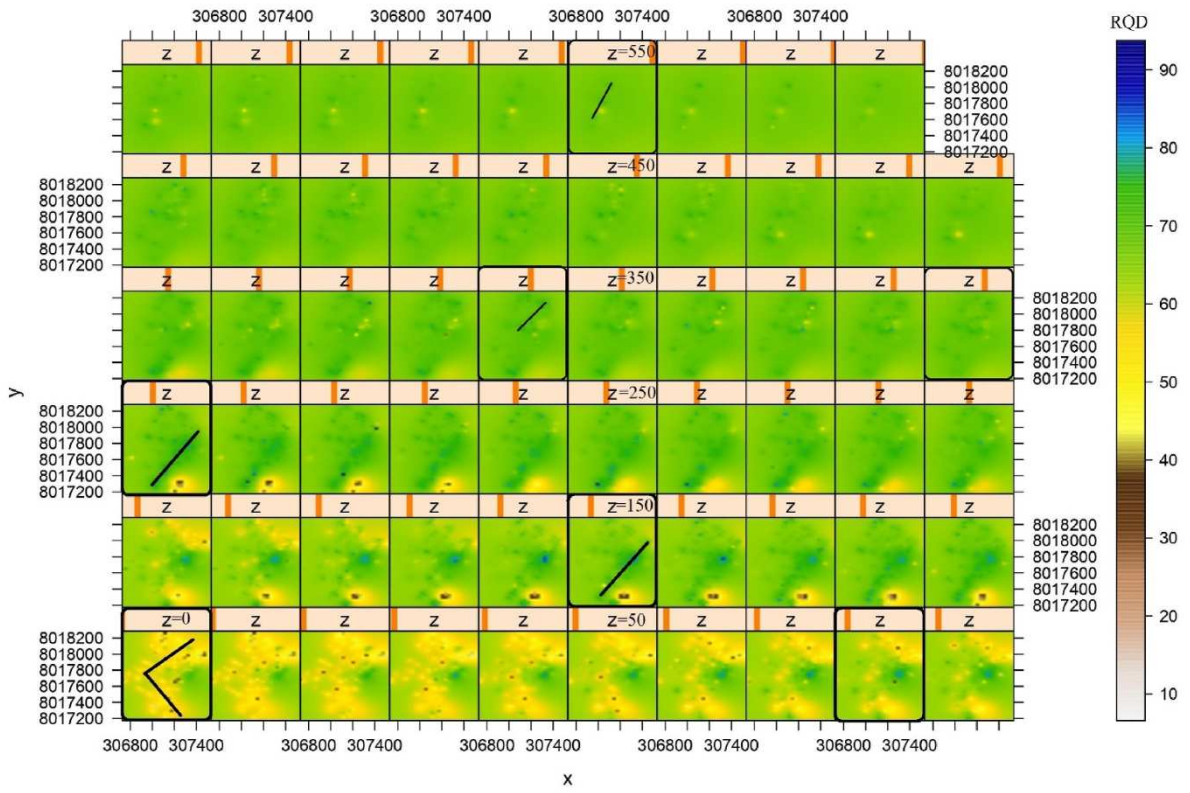


Fig. 8. Geomechanical model of the RQD variable obtained using the IDW interpolator - Central sector. Black lines are alignments of RQD values related to regional structural geology.

even better than the material that surrounds it. This alignment can be observed between depths of 130–500 m. Below this depth there are still unaligned points, also of excellent geomechanical quality. The same situation occurs both for CLASS variable and for RQD, that is, after 530 m of depth, the model presents average values of the CLASS variable due to the significant reduction of the number of drillholes from this depth. Finally, the entire model of the Norte_Sul sector has reasonable mechanical strength.

The low RQD values (RQD <40) are mostly related to the presence of void spaces (FD) from the dissolution process of carbonate rocks present within the mine, with material classified as unrecovered (S/R) during the process of drilling the boreholes, in addition to the presence of soil in the region. Analyzing the database, there are void spaces from the surface to a depth of 450 m, and they appear more expressively between 50 and 100 m. Bittencourt and Reis Neto (2012) mention that in the Vazante region, void spaces were found up to a depth of 600 m. Up to 50 m in depth, the results obtained for the RQD model can also be related to the presence of the unrecovered material (S/R or R/S) and of the soil. These authors mention that there is a layer of colluvium and alluvial soil in the study region, which varies from a few centimeters to hundreds of meters. The database states that between 300 and 350 m of depth there is the presence of SOLO (soil layers). It is believed that this fact may be related to the presence of filling material, as reported by Bittencourt and Reis Neto (2012). In relation to the higher RQD values (RQD> 80), it is found that these are related to the sound breccia and dolomites. It is noticeable that abrupt contacts between weathered and sound rock material is quite common for carbonatic weathering profiles.

From a structural point of view, there is no local information available for comparison with the results of the geomechanical models. However, comparing the results with regional structural geology studies, it is observed that the South sub-sector crosses the Vazante geological fault, which has NE-SW direction, according to CPRM (2015), Fig. 6; on Fig. 7 shows the regional trend of this regional fault. This fact explains the presence of this direction both in the RQD model, as previously described, and as reported in scientific works carried out in the Vazante region, such as Lemos (2011), Bittencourt and Reis Neto (2012) and Charbel (2015). The NW-SE direction, on the other hand, is the direction of a secondary, water-conducting fracturing system, as described by Bhering (2009). It makes sense that near this direction, there is a region of the massif with less mechanical resistance, since the water causes dissolution of the carbonate rocks present in the place. It has been made clear that the mining company that provided the drillhole data did not provide the local geological and structural data.

5.2. Central sector

Fig. 8 shows the results of the interpolation of the RQD variable by the IDW interpolator for the Central sector. It is noted that, from the surface, there is an extensive area with median RQD values (RQD between 50 and 60) with a few points with low RQD values inside of it. This behavior can be noticed up to a depth of 60 m. Between 70 and 130 m of depth, it appears that there is a reduction of the area with median RQD values, and there is an increase of the area with high values of RQD> 70. In addition, there is a decrease in points with low RQD values and an increase in points with very high RQD values, including points with RQD around 90. Between 140 and 230 m in depth there is a decrease of the area with average values of RQD, and in the interior of these remaining areas there is the presence of low values of RQD (RQD <40), which disappear at a depth of 240 m. In addition, between the depths of 140 and 330 m, it is possible to see points with high RQD values (RQD> 80) aligned in accordance with the NE-SW direction (black lines in Fig. 8). From a depth of 340 m, it can be seen that the alignment of the points with high RQD values tends towards the N–S direction, with a reduction in the alignment to the angle with a northern orientation. Lemos (2011) reports that the fault zone of the Extreme North region has a preferential N–S direction, so it is believed that this

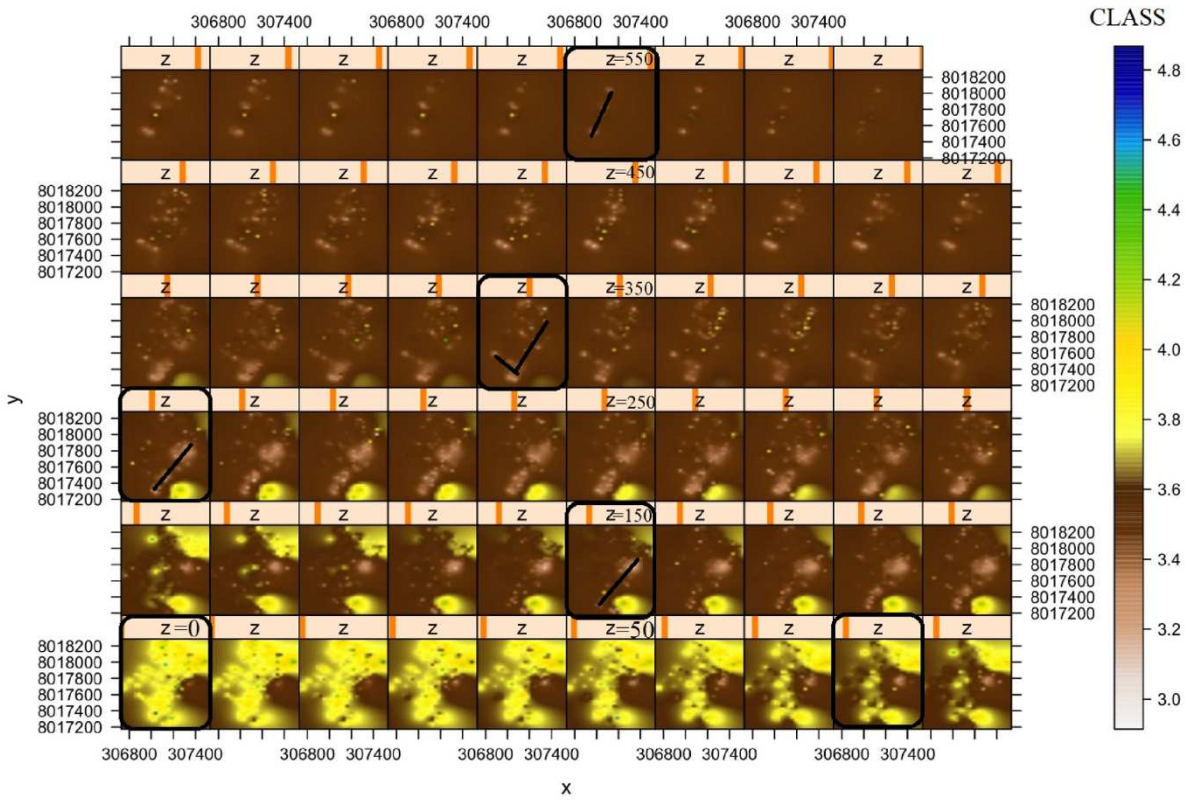


Fig. 9. Geomechanical model of the CLASS variable obtained using the IDW interpolator - Central sector.

rotation of RQD values is related to the influence of this regional structure. Finally, at a depth of 390 m, the region with median RQD values is no longer observed.

Analyzing Fig. 9, it can be noted that the CLASS variable model (simplified RMR classification), obtained through the IDW interpolator, presented a rock mass with medium geomechanical resistance, including

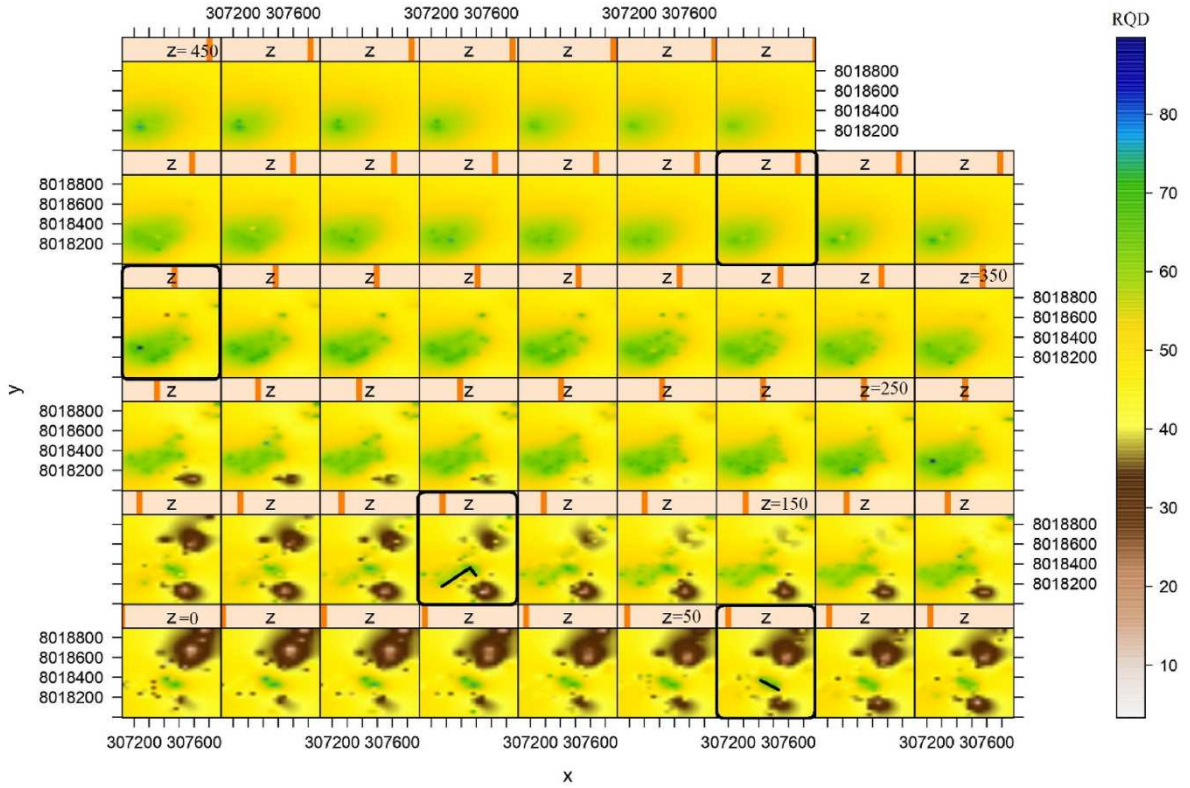


Fig. 10. Geomechanical model of the RQD variable obtained using the IDW interpolator - North sector.

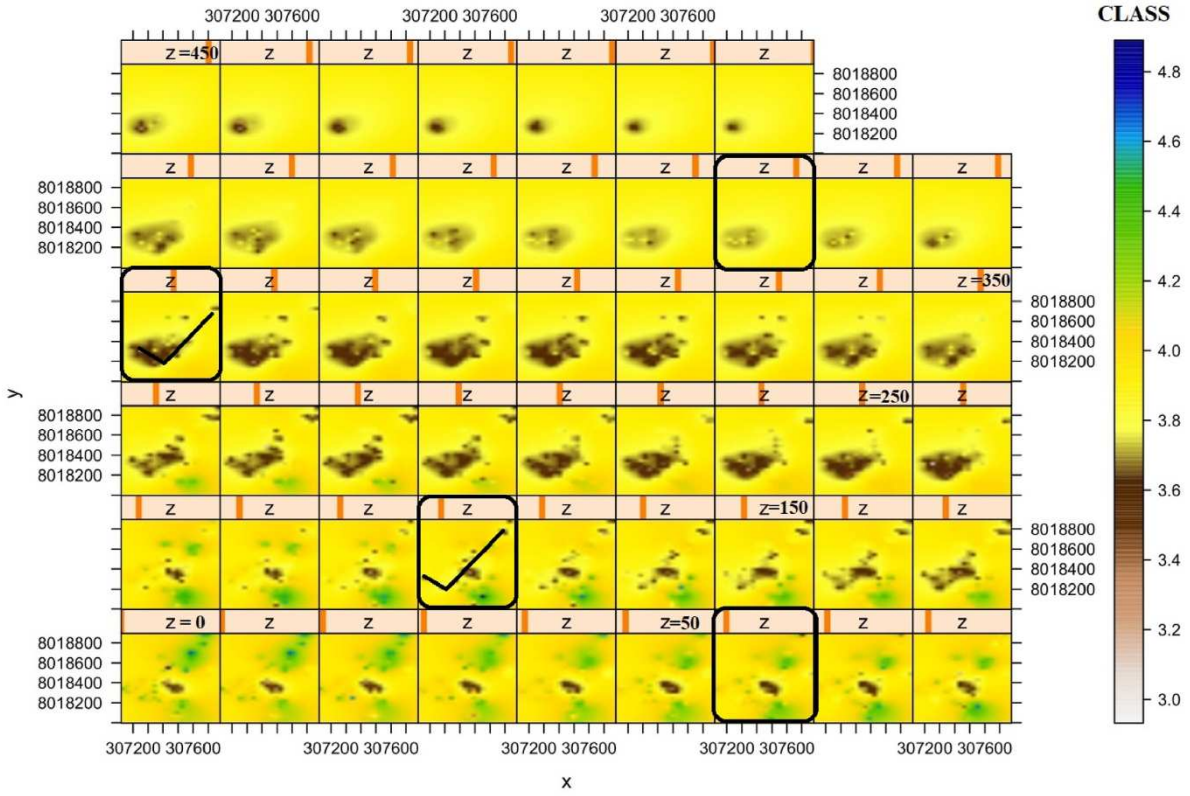


Fig. 11. Geomechanical model of the CLASS variable obtained using the IDW interpolator - North sector.

extensive areas with CLASS equal to 3.6 from the surface to the final depth analyzed. Between 80 and 190 m of depth, the extension of the aforementioned area is very expressive, with only small areas at the edges of the model with CLASS around 4. Additionally, between 200 and 290 m of depth, this area of CLASS is around 4 it is only seen at the bottom of the model. Specifically, the model's surface is composed of an area with CLASS equal to 4, with points of less resistance inserted into it. At the depth of 80 m, a core of medium resistance (CLASS = 3) is observed within the region of the worst geomechanical class. Between 120 and 390 m there is an alignment of points of a good geomechanical class with NE-SW direction. At a depth of 290 m, there is a second alignment of points with good resistance, however in the NW-SE direction. From a depth of 450 m, it can be seen that the alignment of the points of good geomechanical resistance shows a steep declination, tending to the N–S direction, for the same reasons previously explained. Finally, throughout the model, there are points of the CLASS variable around 4 within the areas with best geomechanical behaviour, and this pattern remains even as the depth advances.

In confronting the two models made, both the RQD and the classification model, it appears that, although consistent, there was no significant variation in results from 240 to 250 m in depth, showing an improvement in the rock mass from this depth and without major lateral and vertical variations. For the model of the variable RQD, a significant variation was observed in a considerable range up to a depth of 150 m, varying from very low values to an RQD approximately equal to 70. Similar behavior was noted in the classification model. Observation of Figs. 8 and 9 show strong correlation between the values of RQD and those of simplified RMR classification (CLASS), since the areas of lower RQD are in agreement with those of the worst category. The same can be stated for the regions with the highest RQD and the best geomechanical class. The same directions were observed for the alignment of points with good geomechanical resistance.

The worst RQD values may be related to the presence of void spaces. For the Central sub-sector, it was noted that the void spaces were observed from the surface up to a depth of 400 m, while between 50 and 100 m they were more expressive. However, for the first 50 m of depth, the results obtained are probably related to the presence of the unrecovered material (S/R or R/S), which, in turn, may be related to the soil cover present in the region, as reported by Bittencourt and Reis Neto (2012) and which are common in the region where carbonate rocks occur.

From the point of view of structural geology, in Fig. 6, it appears that the Central sector does not have structural control at the regional level. This fact may explain the decrease in the presence of void spaces in the site and, consequently, the increase in regions with high RQD values.

5.3. North sector

Fig. 10 shows the results of the model of the variable RQD obtained through the use of the IDW interpolator for the North sector. From the surface (z = 0) to a depth of 130 m, there is an extensive area with low values (especially in the northern portion of the area) and median RQD (RQD around 50) representing almost the entire analysis grid of this subsector. In addition, there is a small area, also with low RQD values at the bottom of the model, and in the central region, and alignments of high RQD points (RQD> 60) in the SE-NW direction, which gradually increases in area, as the depth increases. It is noted that with the advance of the depth (within this range that is from 0 to 130 m) there is a reduction in the area of low RQD values in the North region and an increase in low RQD area to the South. Beyond this, there is a consistent appearance of high RQD points aligned in a second direction, NE-SW. From a depth of 120 m, the high RQD points of the two directions cited form an area whose dimensions increase progressively with the increase of depth. In addition, this area is surrounded by an edge with median RQD values over the entire depth of the model, with an increase in the edge, as the area increases. In the range of 140–220 m in depth, there is a gradual reduction of the area with low RQD values until its disappearance. Between the depths of 230–270 m there are points with

very high RQD values (RQD> 80) inserted into this area with RQD around 70 and points with low RQD values (RQD <30) inserted into the area with RQD around 50. Finally, from a depth of 280 m, there is an inversion of this behavior, with a decrease in the area with high RQD values, in contrast to the increase in the region with median values of the analyzed variable.

The model for the simplified RMR classification of the North-North sector is shown in Fig. 11. In the first 20 m of depth, there is an alignment of points with a NE-SW direction of very low geomechanical resistance, CLASS around 4.8. This alignment persists up to 40 m in depth, however there is a slight improvement in resistance with advancing depth, reflected by the increase in the CLASS variable. From the surface (z equal to zero) up to 80 m there is a cluster of points with a good geomechanical class (CLASS <3.6) within the classification of the region around 4. Between 90 and 280 m in depth there is an NE -SW alignment of points with good geomechanical resistance, and in the range of 100 and 200 m there is also an alignment in the NW-SE direction of points of good resistance. Between 170 and 350 m, it is noted, again, that there is a nucleus of points of good resistance within a Class of around 4. This region loses expressiveness with the depth increase, however, it can be observed along the entire model.

By comparing the two models (RQD and CLASS) through use of the IDW interpolator for the North sub-sector, results present a similar pattern. Significant areas of median values are observed for both variables analyzed and regions with high RQD values and good classification values are aligned in two different directions: SW-NE and NW-SE.

The results found in both models, specifically, the low RQD values and areas with poor geomechanical classifications (CLASS above 4.2) may be associated with the presence of void spaces and with sections classified as S/R inside the massif. For the North sub-sector it was noted that the void spaces were observed from the surface up to the depth of 250 m, and between 100 and 150 m they were more significant. However, for the first 50 m of depth, the results obtained are probably related to the presence of soil cover that may have given rise to unrecovered material (S/R). This material, probably, could have another source other than the soil, since it was observed up to depths between 250 and 300 m.

The 3D models of the RQD and CLASS variables using the IDW interpolator can be viewed at https://drive.google.com/drive/folders/1hHxxYp7WQ-IjFeA6qwI8SUW_8zgko8ar?usp=sharing. These allow the user to analyze the behavior of the variable and instantly find out its value within the analysis grid.

Comparing the results of the RQD models of the three sectors - Figs. 3, 8 and 10 - it is observed that the South is the sector that presents areas with the lowest RQD values, and these low values can be observed from the surface to around 500 m of depth. In addition, in this sector there is a notable alignment of these points along the NE-SW direction. The Central sector, in turn, is the one with the highest RQD values (this result had already been identified in the exploratory analysis of the data), and it is also possible to observe the alignment of points of median values of the RQD along the NE-SW and NW-SE axes. The North sector, on the other hand, also presents extensive areas with low RQD values, but only within the first 150 m of depth.

In relation to the simplified RMR classification models, when comparing the three sectors - Figs. 5, 9 and 11 - it is found that the Central sector is composed almost entirely by rock mass with CLASS around 3, mainly at depths greater than 150 m. The North and South sectors have some points' alignment, and some cores close to CLASS 4, with those in the North presenting larger areas and those in the South continuing to greater depths.

As the mining company presented no local structural and geological data, the structural behavior observed as a result of the performed classification can only be compared to information on a regional scale. This fact made it difficult to carry out a more detailed interpretation of the results obtained, since the structural control over the mechanical behavior of the massif is notorious. In the technical literature, the only specific work found on the northernmost area of the Vazante Mine was by Lemos (2011). This author mentions that the zinc deposit area in the Extreme North is characterized by rocks belonging to the Serra de Garrote and Serra do Poço Verde formations of the Vazante Group. Still, according to Lemos (2011), the region of Extreme North presents quite fractured rocks, and the main structures responsible for this behavior are represented by beddings and bandings; by the plains related to the Vazante Fault; by the brittle structures of NW direction, as well as by low dip structures. It is known that the structures related to the Vazante Fault have a N–S direction due to the proximity of the contact between the Serra do Poço Verde and Serra do Garrote formations. This direction was observed in alignments of points of high RQD and a good geomechanical CLASS in the North-Central sub-sector. Regarding the degree of fracturing in the Vazante region, as a whole, Bhering (2009) states that the main factors responsible for the high degree of fracturing are related to the different states of tension in which the rock massif was subjected to over geological time, to brittle tectonics - as the most striking structural characteristic and to excessive hydrothermalism. Lemos (2011) also reports that the main water-conducting discontinuities, relevant to the understanding of the hydrogeological and karst issue of the entire Vazante region, belong to the NW-SE family of fractures. This direction was also observed in the results of the geomechanical models in the present study, associated with the alignment of low values of RQD and worse geomechanical classes for the South and North subsectors.

From the point of view of regional structural geology, the South sector "cuts" a normal fault, probably the Vazante Fault, Figs. 6 and 7. The North sector, in addition to the fault, also presents lineaments in its analysis grid. Therefore, the presence of these structural controls justifies the fact that these sub-sectors have presented extensive regions with low RQD values. The grid region of the Central sector, on the other hand, does not intercept the Vazante Fault zone and does not present lineaments within it. Thus, this may explain the fact that this sector had the highest RQD values, indicating a rock mass with good geomechanical resistance, that is, values of the CLASS variable below 3.5.

In relation to local geology, the lithotypes found in all three sectors, according to the drillhole data, do not vary, being basically composed of dolomites and breccia. What changes from one sector to another is the greater or lesser presence of void spaces (classified as FD), unrecovered material (S/R or R/S) and the presence of soil. Comparing the percentage of void spaces in the Central sector with the South at the same depths, there is a significant reduction in the presence of these void spaces for the Central sector at all depths. Specifically, there was a significant reduction in the presence of void spaces for depths below 200 m. For example, between 250 and 300 m in depth, in the South sector, 19.4% of the massif is made up of void spaces, while for the Central sector this percentage is only 1.8%. In relation to the first 50 m of depth, the percentage of the unrecovered material in both sectors did not vary. As for the soil, a decrease in the percentage was also observed for the Central sub-sector. Finally, comparing the percentage of void spaces in the North sector with the other two sectors, it is noted that the values observed for this sector have the same pattern for the void spaces as observed in the Central sector. Therefore, the subsector with the greatest presence of void spaces is the South. Therefore, in addition to the structural control represented by the Vazante Fault, the South sector has a higher presence of void spaces. This association of factors may help to clarify the fact that this sector had lower RQD values, mainly at greater depths, when compared to the North sector, which also presents structural control, but has a lesser presence of void spaces.

The association between geology and structural control on a regional scale allows analysis of some of the results obtained. As the Central sector is the sector that has the lower void spaces content and has no structural control, it was expected that the geomechanical class of the sector would be better than that of the other two sub-sectors, which was verified by the model.

With the aim of evaluate the accuracy of the interpolation provided by using IDW, a cross-validation technique was used, in which a graph

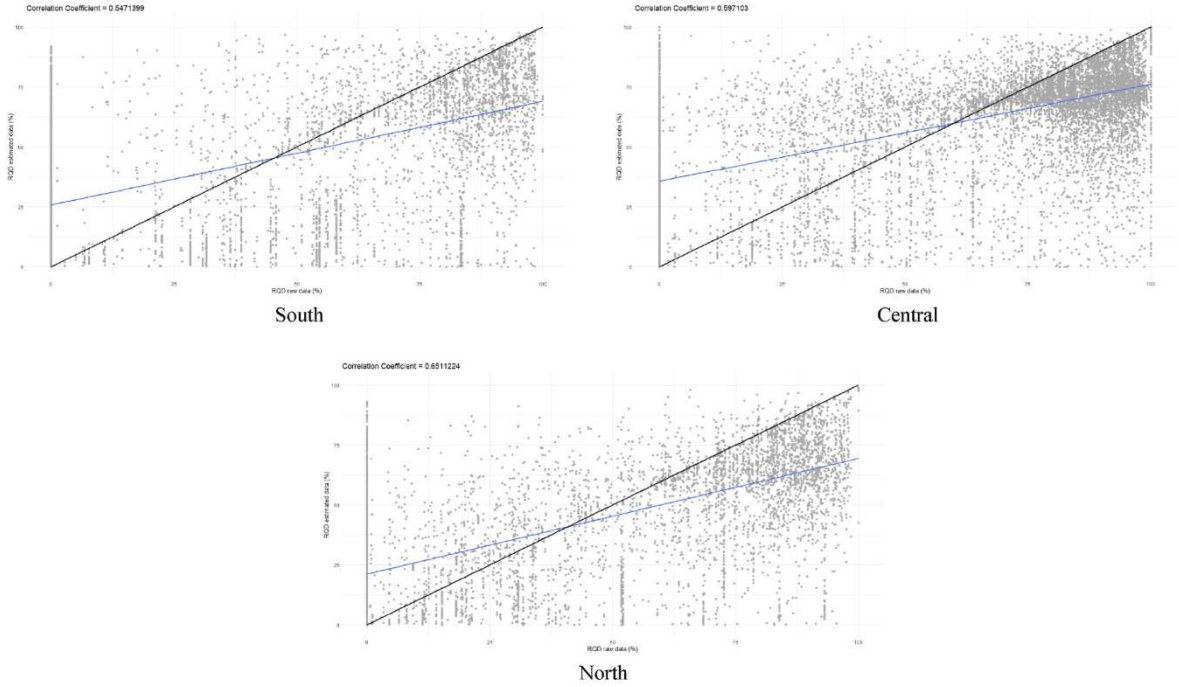


Fig. 12. Correlation coefficient for RQD on the three mine sectors.

relating observed and estimated values was produced for each sector – South, central and North. Low to regular correlation coefficients – 0.55 to 0.65, as shown in Fig. 12, were found. These values can indicate that particularities of the rock masses under study – as the presence of dissolution cracks (low RQD values in depth), difficult the interpolation and that IDW cannot the most adequate technique to be used.

As CLASS variable is categorical, it was not possible to prepare similar graphs. In this case validation was based on its comparison to regional trends, as already discussed. So, kriging was used to provide another approach to deal with the problem. The results of such evaluation will not be presented on this manuscript but can be found in Pereira (2020).

Finally, the methodology developed for the creation of 3D geomechanical models using drillholes data and using free programming language is unique, having enormous potential for use, and is relevant in the context of geomechanics. However, the results obtained still need to be improved, such as its use in other types of rock masses and the need for a careful comparison with local field data.

6. Conclusions

The methodology created for the development of 3D geomechanical models using drillholes and free programming language (the R program) proved to be satisfactory, both in terms of visualization and user interactivity. Despite the regular R^2 values of RQD graphs and the fact that CLASS variable is categorical, the geomechanical model maps have resulted in satisfactory results when compared to local and regional geology. Although, models' results should be compared to real field data, in the future. Yet, the potential of the tool created for the 3D geomechanical classification of rock masses is very interesting and should be better explored in the future.

From the information we have about the extreme North region of the mine studied, it can be noted that the geomechanical models of the RQD and the CLASS variable using the IDW interpolator have resulted in satisfactory outcomes. In general, the Norte_Central sector has the most resistant rock mass, with higher RQD values. This fact was also observed in the exploratory analysis of the data and it is in line with regional data. The models of the South and North sectors showed regions with lower RQD values, and for the South this behavior was observed up to great depths. The justification for the results obtained is certainly related to the association between lithology, weathering and structural geology. The presence of the Vazante Fault, of other structural lineaments, and of the greater or lesser presence of the void spaces arising from the dissolution of the carbonate rocks in the interior of the analysis grids, justify the greater or lesser values of the RQD observed.

CRediT authorship contribution statement

Luana Cláudia Pereira: Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation, Conceptualization. **Gérson Rodrigues dos Santos:** Software, Methodology, Conceptualization. **Eduardo Antonio Gomes Marques:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Jandressom Dias Pires:** Writing – review & editing, Validation, Software. **Rodolfo Renó:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsames.2022.103775.

References

Achilleos, G., 2008. Errors within the inverse distance weighted (IDW) interpolation procedure. Geocarto Int. 23 (6), 429–449.
Achilleos, G.A., 2011. The Inverse Distance Weighted interpolation method and error propagation mechanism - creating a DEM from an analogue topographical map. Spatial Sci. 56 (2), 283–304.
Alghalandis, Y.F., Afzal, P., 2011. Important Considerations on the Application of IDW Interpolation Method. Int. 2nd Conference of Mine and Industry, Tehran, Iran.
Babak, O., 2014. Inverse distance interpolation for facies modeling. Stoch. Environ. Res. Risk Assess. 28, 1373–1982.

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London, 2015, URL. http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/.

Bhering, A.P., 2009. Classificação do maciço rochoso e caracterização das brechas da mina subterrânea de Vazante – MG. Dissertação (Departamento Engenharia Civil) – Universidade Federal de Viçosa, Viçosa, p. 185, 2009.

Bieniawski, Z.T., 1974. Engineering classification of jointed rock masses. Trans. S. Afr. Inst. Civ. Eng. 15, 335–344.

Bieniawski, Z.T., 1989. Engineering Rock Mass Classifications. Wiley, New York.

Bittencourt, C., Reis Neto, J.M.D.O., 2012. O sistema cárstico de Vazante - carste em profundidade em metadolomitos do Grupo Vazante - MG. Rev. Bras. Geociencias 42 (1), 1–10, 2012.

Bivand, R., Lewin-Koh, N., 2019. Maptools: Tools for Handling Spatial Objects. R package version 0.9-9. https://CRAN.R-project.org/package=maptools.

Bivand, R., Rundel, C., 2020. Rgeos: interface to geometry engine - open source ('GEOS'). R package version 0.5-3. https://CRAN.R-project.org/package=rgeos.

Bivand, R., Keitt, T., Rowlingson, B., 2019. Rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.4-8. https://CRAN.R-project.org/package=rgdal.

Charbel, P.A., 2015. Gerenciamento de risco aplicado à diluição de minério. 2015. 448 f. Tese (Doutorado em Geotecnia) - Faculdade de Tecnologia, Departamento de Engenharia Civil e Ambiental. Universidade de Brasília, Brasília/DF.

Companhia de Pesquisa de Recursos MINERAIS – CPRM/Serviço Geológico do Brasil, 2015. Metalogenia das Províncias Minerais do Brasil: Distrito Zincífero de Vazante, MG – Estado de Minas Gerais: texto e mapa metalogenético, escala 1:75.000. In: Dias, Paulo H.A., Marinho, Marcelo de S., Vilela, Francisco T., Sotero, Marcus P., Matos, Caio A., Marques, Eduardo D. (Eds.). CPRM, Belo Horizonte.

Microsoft Corporation; Weston, S., 2019. doParallel: for Each Parallel Adaptor for the 'parallel' Package. R package version 1.0.15. https://CRAN.R-project.org/package=doParallel.

Dardene, M.A., 2000. The brasilia fold belt. In: Cordani, U.G., Milani, E.J., Thomaz Filho, A., Campos, D.A. (Eds.), Tectonic Evolution of South America. 31 St International Geological Congress. SBG, Rio de Janeiro, pp. 231–263.

Deisman, N., Khajeh, M., Chalaturnyk, R.J., 2013. Using geological strength index (GSI) to model uncertainty in rock mass properties of coal for CBM/ECBM reservoir geomechanics. Int. J. Coal Geol. 112, 76–86. https://doi.org/10.1016/j.coal.2012.10.015.

Egaña, M., Ortiz, J., 2013. Assessment of rmr and its uncertainty by using geostatistical simulation in a mining project. J. GeoEng. 8 (3), 83–90.

El-Sayed, E.O., 2012. Improving the prediction accuracy of soil mapping through geostatistics. Int. J. Geosci. 3, 574–590. July.

Ellefmo, S.L., Eidsvik, J.O., 2009. Local and spatial joint frequency uncertainty and its application to rock mass characterization. Rock Mech. Rock Eng. 42 (4), 667–688.

Esfahani, N.M., Asghari, O., 2013. Fault detection in 3D by sequential Gaussian simulation of Rock Quality Designation (RQD). Case study: gazestan phosphate ore deposit, Central Iran. Arabian J. Geosci. 6, 3737–3747.

Exadaktylos, G., Stavropoulou, M., 2008. A specific upscaling theory of rock mass parameters exhibiting spatial variability: analytical relations and computational scheme. Int. J. Rock Mech. Min. Sci. 45 (7), 1102–1125.

Ferrari, F., 2014. Rock Mass Characterization and Spatial Estimation of Geomechanical Properties through Geostatistical Techniques. Tese (Doutorado em Ricerca in Scienze della Terra) - Università Degli Studi di, Milano, 2014. 232 f.

Friendly, M., Dalzell, C., Monkman, M., Murphy, D., Foot, V., Zaki-Azat, J., 2019. Lahman: Sean 'Lahman' Baseball Database. R Package Version 7.0-1. https://CRAN.R-project.org/package=Lahman.

Gardiman Junior, B.S., Magalhães, I.A., De Freitas, C.A.A., Cecílio, R.A., 2012. Análise de técnicas de interpolação para espacialização da precipitação pluvial na bacia do rio Itapemirim (ES)/Analysis of interpolation techniques for spatial rainfall distribution in river basin Itapemirim (ES). Revista Ambiência 8 (1), 61–71.

Grosjean, Ph, 2019. SciViews: a gui api for R. umons, mons, Belgium. URL. http://www.sciviews.org/SciViews-R.

Hack, R., Orlic, B., Ozmutlu, S., Zhu, S., Rengers, N., 2006. Three and more dimensional modelling in geo-engineering. Bull. Eng. Geol. Environ. 65 (2), 143–153.

Hengl, T., 2020. GSIF: Global Soil Information Facilities. R package version 0.5-5.1. https://CRAN.R-project.org/package=GSIF.

Hengl, T., Roudier, P., Beaudette, D., Pebesma, E., 2015. plotKML: scientific visualization of spatio-temporal data. J. Stat. Software 63 (5), 1–25. URL. http://www.jstatsoft.org/v63/i05/.

Hijmans, R.J., 2020. Raster: geographic data analysis and modeling. R package version 3.1-5. https://CRAN.R-project.org/package=raster.

Ikechukwu, M.N., Ebinne, E., Idorenyin, U., Raphael, N.I., 2017. Accuracy assessment and comparative analysis of IDW, spline and kriging in spatial interpolation of landform (topography): an experimental study. J. Geogr. Inf. Syst. 9 (3), 354–371.

Jakob, A.A.E., Young, A.F., 2006. O uso de métodos de interpolação espacial de dados nas análises sociodemográficas. In: Encontro Nacional de Estudos Populacionais, 15. Caxambu. Anais... Caxambu, p. 22, 2006.

Kaewkongkaew, K., Phien-Wej, N., Kham-Ai, D., 2015. Prediction of rock mass along tunnels by geostatistics. KSCE J. Civ Eng 19, 81–90. https://doi.org/10.1007/s12205-014-0505-3.

Karami, R., Afzal, P., 2015. Estimation of element distributions by combining artificial neural network and inverse distance weighted (IDW) based on lithogeochemical data in Kahang Porphyry Deposit, Central Iran. Univ. J. Geosci. 3 (2), 59–65.

Komsta, L., Novomestky, F., 2015. Moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests. R package version 0.14. https://CRAN.R-project.org/package=moments.

Kuhn, M., 2020. Caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret.

Leisch, F., Dimitriadou, E., 2010. Mlbench: Machine Learning Benchmark Problems. R package version 2.1-1.

Lemos, M.G., 2011. Caracterização geológica e tecnológica do minério de zinco do Extremo Norte da Mina de Vazante, Minas Gerais. 2011. 193 f. Dissertação (Mestrado) – Instituto de Geociências, Universidade Estadual de Campinas, Campinas.

Li, J., Heap, A.D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists, vol. 137. Geoscience Australia, Canberra, p. 154, 2008/23.

Ligges, U., Mächler, M., 2003. Scatterplot3d - an R package for visualizing multivariate data. J. Stat. Software 8 (11), 1–20.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., Lin, C., 2019. e1071: Misc Functions of the Department of Statistics. In: Probability Theory Group (Formerly:E1071), TU Wien. R package version 1.7-3. https://CRAN.R-project.org/package=e1071.

Miranda, G.H.B., 2017. Análise de amostragem e interpolação na geração de MDE. 2017. 65f. Dissertação (Mestrado em Engenharia Civil). Universidade Federal de Viçosa, Viçosa.

Monteiro, L.V.S., 2002. Modelamento metalogenético dos depósitos de zinco de Vazante, Fagundes e Ambrósia, associados ao Grupo Vazante, Minas Gerais. 2002, f.362. Tese (Doutorado em Hidrogeologia e Recursos Hídricos). – Universidade de São Paulo, São Paulo.

Oh, S., Chung, H., Lee, D.K., 2004. Geostatistical integration of MT and borehole data for RMR evaluation. Environ. Geol. 46 (8), 1070–1078 spec. iss.

Öztürk, C.A., Nasuf, E., 2002. Geostatistical assessment of rock zones for tunneling. Tunn. Undergr. Space Technol. 17 (3), 275–285.

Öztürk, C.A., Simdi, E., 2014. Geostatistical investigation of geotechnical and constructional properties in Kadikoy-Kartal subway, Turkey. Tunn. Undergr. Space Technol. 41 (1), 35–45.

Pakyuz-Charrier, E., Giraud, J., Ogarko, V., Lindsay, M., Jessell, M., 2018. Drillhole uncertainty propation for three-dimensional geological modeling using Monte Carlo. Tectonophysics 747–748, 16–39. https://doi.org/10.1016/j.tecto.2018.09.005.

Pebesma, E.J., 2012. Spacetime: spatio-temporal data in R. J. Stat. Software 51 (7), 1–30. http://www.jstatsoft.org/v51/i07/.

Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R. News 5 (2). https://cran.r-project.org/doc/Rnews/.

Pereira, L.C., 2020. O uso da geoestatística para elaboração de modelos geomecânicos multidimensionais. 2020. 194 f. Tese (Doutorado em Engenharia Civil). Universidade Federal de Viçosa, Viçosa.

Pinheiro, M., Vallejos, J., Miranda, T., Emery, X., 2016. Geostatistical simulation to map the spatial heterogeneity of geomechanical parameters: a case study with rock mass rating. Eng. Geol. 205, 93–103.

Revelle, W., 2019. Psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych Version = 1.9.12.

Ribeiro Jr., P.J., Diggle, P.J., Christensen, O., Schlather, M., Bivand, R., Ripley, B., 2020. geoR: Analysis of Geostatistical Data. R package version 1.8-1. https://CRAN.R-project.org/package=geoR.

Sajid, A.H., Rudra, R.P., Parkin, G., 2013. Systematic evaluation of kriging and inverse distance weighting methods for spatial analysis of soil bulk density. Can. Biosyst. Eng./Le Genie des Biosyst. Canada 55 (1983), 1–13.

Sarkar, D., 2008. Lattice: Multivariate Data Visualization with R. Springer, New York, ISBN 978-0-387-75968-5.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elber, A., Toomet, O., Crowley, J., Hofmann, H., Vickha, H., 2020. GGally: Extension to 'ggplot2'. R Package Version 1.5.0. https://CRAN.R-project.org/package=GGally.

Schneeberger, R., La Varga, M. de, Egli, D., Berger, A., Kober, F., Wellmann, F., Herwegh, M., 2017. Methods and uncertainty-estimations of 3D structural modelling in crystalline rocks: a case study. Solid Earth Discuss 8, 987–1002. https://doi.org/10.5194/se-2017-47.

Setianto, A., Triandini, T., 2013. Comparison of Kriging and Inverse Distance Weighted (IDW) interpolation methods in lineament extraction and analysis. J. Southeast Asian Appl. Geol. 5 (1), 21–29.

Seyedmohammdi, J., Esmaeelnejad, L., Shabanpour, M., 2016. Spatial variation modelling of groundwater electrial conductivity using geostatistics and GIS. Model Earth Syst. Environ. 2, 1–10. https://doi.org/10.1007/s40808-016-0226-3.

Sievert, C., 2020. Interactive Web-Based Data Visualization with R, Plotly, and Shiny. Chapman and Hall/CRC Florida.

Signorell, A., et al., 2020. DescTools: tools for descriptive statistics. R package version 0 99, 34.

silva, C.D.R., Lopes, M.C.Q., Centeno, J.A.S., 2007. Estudo do método de interpolação do inverso da distância a uma potência. In: Simpósio Brasileiro de Geomática, 2, Presidente Prudente. Anais. Presidente Prudente, pp. 57–62, 2007.

Stavropoulou, M., Exadaktylos, G., Saratsis, G., 2007. A combined three-dimensional geological-geostatistical-numerical model of underground excavations in rock. Rock Mech. Rock Eng. 40, 213–243.

Vaidyanathan, R., Xie, Y., Allaire, J.J., Cheng, J., Sievert, C., 2019. Htmlwidgets: HTML Widgets for R. R package version 1.5.1. https://CRAN.R-project.org/package=htmlwidgets.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York, ISBN 0-387-95457-0.

Wei, T., Simko, V., 2017. R Package "corrplot": Visualization of a Correlation Matrix. Available from:, Version 0.84. https://github.com/taiyun/corrplot.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, 2016.

Wickham, H., Bryan, J., 2019. Readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl.
Wickham, H., et al., 2020. Dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr.
Yi, H., Choi, Y., Park, H.D., 2014. Application of multiple indicator Kriging for RMR value estimation in areas of new drift excavation during mine site redevelopment. Environ. Earth Sci. 71, 4379–4386.

3 ADVANCED ANALYTICS IN MINING ENGINEERING

Chapter 24
Advanced Analytics for Spatial Variability of Rock Mass Properties in Underground Mines

Luana Cláudia Pereira, Eduardo Antonio Gomes Marques, Gérson Rodrigues dos Santos, Marcio Fernandes Leão, Lucas Bianchi, and Jandresson Dias Pires

Abstract One of the critical features required for rock engineering is achieving a reliable estimate of rock masses properties. However, representing the geomechanical properties of rock mass remains a challenge in rock mechanics, and determining these parameters directly is time-consuming, expensive, and the reliability of the results of these tests is sometimes questionable. Therefore, this chapter aims to predict the rock mass properties via Random Forests by a Case Study. Random Forest is an algorithm that can act as a classifier and regressor, using a collection of decision trees. To do this, a simplified Rock Mass Rating (RMR) model was developed using the information obtained from the drill hole. Also, a data treatment method was used to prevent information quality harm by conflicting information, and a qualitative analysis was performed. Finally, the proposed method results are satisfactory and showed that by validating and calibrating the database, using this method for geomechanical modeling can be successful.

Keywords Underground mining · Artificial intelligence · Rock mass properties · Random Forest

Introduction

The comprehension of a rock mass behavior requires the knowledge of its mechanical parameters. These data are typically obtained in situ or through laboratory tests, in samples collected on drill holes, which information is only punctual. Due to rock genesis and discontinuity distribution, rock masses can be highly complex, which representativeness can be simplified as a function of the applied geomechanical characterization method. For rocks that present dissolution susceptibility, geomechanical behavior is even more complex, as this dissolution easiness is a function of rock mass discontinuities distribution.

L. C. Pereira ( ) · E. A. G. Marques · G. R. dos Santos · M. F. Leão · L. Bianchi · J. D. Pires
Federal University of Viçosa, Viçosa, Brazil

A. Soofastaei (ed.), *Advanced Analytics in Mining Engineering*,
https://doi.org/10.1007/978-3-030-91589-6_24

On ideal elastic conditions, the acquisition of material parameters, be they soils, rocks, or intermediate materials (weathered rocks), on statistic (stress–strain relationships), or dynamic (elastic wave velocity), should arrive at the same result. Nevertheless, rock masses are not usually elastic, mainly composed of heterogeneous and anisotropic, such as phyllites. This factor can be intensified considering the mother rock genesis, resulting in differences in elasticity moduli values, estimated by the methods mentioned earlier, influenced by constitutive properties of rock materials, such as porosity.

In this context, understanding and representing the spatial variability of rock mass geomechanical parameters are still a challenge for rock mechanics, and, in this process, the construction of geomechanical models is of great interest.

A model consists of a simplified representation of a natural environment. According to Landim [1], the quantitative management of geological data requires simplification. Kaufmann and Martin [2] sustain that the first step for constructing a 3D geological model is collecting, organizing, and selecting the data to be used. According to these authors, the initial effort is considerable but essential to guarantee a reliable model. In the sequence, data need to be processed and stored in a consistent database. Still, according to these same authors, selected data need to be referred in conformity to a geographical system and need to have a working scale defined. Finally, due to natural variability and guaranteeing the quality of data, a thorough validation is necessary.

Kaufmann and Martin [2] report that, in many cases, the primary source of geological data is geological mapping and that punctual data should be obtained from drill holes. These authors also describe a method for building a geological model composed of several steps, as shown in Fig. 24.1.

Among these steps, one consists of automation so that agility of data processing is provided, allowing short time updating. Another step is the reinterpretation and validation of data.

Hadjigeorgiou [3] reports that data must be analyzed, summarized, and transformed into useful information for many geotechnical projects. Otherwise, its collection will have been the last activity itself.

The construction of geomechanical models is still a challenge, as there is a necessity to encompass several parameters to guarantee the representativeness of field reality, which is not a simple task. Furthermore, selecting one variable (or some variables) is necessary to construct geomechanical models in some cases. This choice needs to be trustworthy to field observations, and also, redundant or irrelevant variables should be left out of the model to reduce computational cost. So, to reduce subjectivity, the error on the selection process and improving the prevision precision of the built model, one can use some methods to rank the variables available in the data set, and multivariate and Bayesian statistical and regression analyses and machine learning, using Random Forest, are the most used techniques.

The regression analyses use methods to investigate the association between some observable quantities and, in the affirmative case, the nature of such associations. Multivariate statistics can be applied to analyses in which the data set encompasses simultaneous measurements of several variables, whose purpose is to measure,

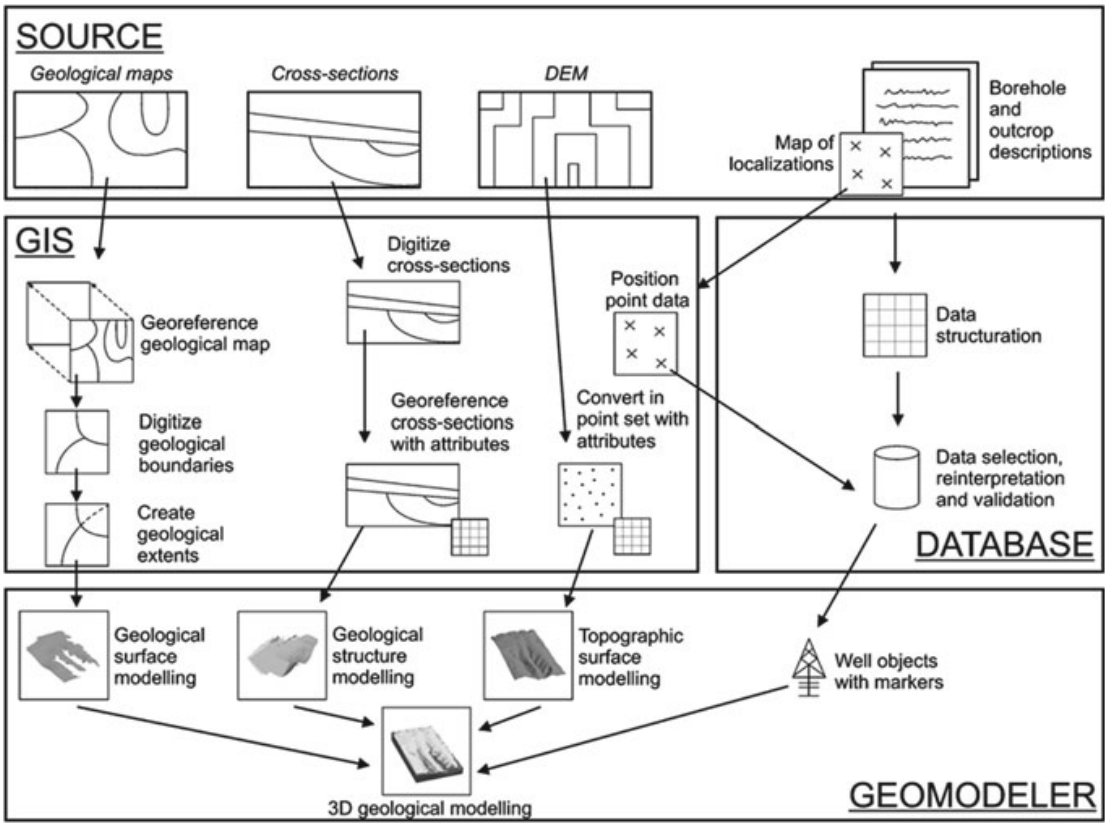


Fig. 24.1 Organogram of the method proposed by Kaufmann and Martin [2] for building a 3D geological model

explain, and predict the interrelationship degree between these variables. Bayesian statistics can also be used to quantify the uncertainties of a selected variable based on probability. There is necessary to test theoretical assumptions for each of these methods, which may not be determined in practice. So, in this context, the option for using machine learning techniques, such as Random Forest, can be used, as in the present study.

The purpose of this chapter is to present the results of a method developed in R language [4] to perform drill holes data treatment, classical exploratory data analyses, spatial exploratory data analyses, univariate interpolation analyses through kriging, multivariate kriging through machine learning (Random Forest), and kriging, and, finally, to rank variables used for simplified classification of rock masses (4D interactive maps).

Application of Multivariate Analyses of Rock Masses Classification

The construction of a geomechanical model necessarily uses interpolation techniques, as there will always be places where no field (primary) information is available, so it is necessary to know how a specific variable behaves in these areas. So, interpolation techniques are essential to understand the behavior of a specific parameter from information collected in the study area.

However, when there are many variables to construct a model, working with all these variables can be challenging. Besides, one can have variables that will not influence the results in a significant manner, and that should not be, in theory, analyzed, or considered. With this in mind, identifying the variables with a bigger (or smaller) influence on a determined area using multivariate analyses is a significant activity for building a geomechanical model. In order to define which variable has a significant influence on a rock mass behavior and, consequently, which one will be most important for the construction of the geomechanical model, it is necessary to provide joint analyses of those variables, so defining an order of importance. One way to perform this sensitivity analysis is to use a machine learning technique, such as the Random Forest algorithm. With this method, selecting the explanatory variables essential to describe a problem sets aside the less relevant ones [5].

Once known and defined as the variable (or more than one variable) that controls the rock mass behavior, mathematic, and statistic methods, such as interpolation, can be used to enable the construction of a geomechanical model. The production of secondary information is based on a determined model governed by assumptions and conditions. One basic assumption is that the information on the surface places of the phenomenon in which sampling was performed presents a satisfactory level of spatial dependency (autocorrelation). The mathematic function used is approaching the continuous phenomenon that is being analyzed. So, each interpolation method can result in different representations of the same set of data. The use of a determined interpolator depends on the input data set from the interpolator's intrinsic characteristics. Each interpolator has a particularity, so it must be carefully observed before its application [6–9].

Random Forest (RF)

Random Forest is a deterministic interpolation method that integrates what is known as machine learning, which basic principle states that systems can learn from analyzed data, identify patterns, and, finally, making decisions with or without a minimum of human intervention. Trees are classification or regression models that allow the use of continuum and/or discrete variables. These models are adjusted by dividing successively a data set into more and more homogeneous groups. The classification can be translated as a categorical or discrete value prediction and aims to construct

models and define rules, from a set of correctly pre-classified examples, for further classification of new and unknown examples. On the other hand, regression seeks to find a function that can map a data item for a variable prediction with a continuous numerical value [10, 11].

Random Forest (RF) is an algorithm that can act as a classifier and regressor, using a collection of decision trees initially developed by Breiman [12]. It is a random decision tree model for non-linear prediction between statistical variables whose main characteristics include calculating essential measures of the variable (variable ranking), automatic calculation of errors, automatic treatment of missing data, and the possibility of verification of variable dependency. They also avoid overfitting and are less sensitive to noise. It can and has been used in several fields of knowledge [13–15].

As they are also known, binary decision trees use binary recursive partitioning, in which data from a primary variable are successfully divided along the gradient of the explanatory variables into two subsets, or nodes, descending. These divisions occur in a way that, for any node, the division is selected to maximize the difference between two sets or ramifications. The mean value of the primary variable in each node can then be used to map the variable through the region of interest [10].

The RF aims to perform the conception of several decision trees using a subset of randomly selected attributes with reposition around 2/3 of the original set and reserving the other 1/3, named out-of-bag, for validation. So, there is the use of two main approaches.

The first approach uses initial aggregation known as bagging (bootstrap sample), ensuring that each new selected set could have some registers included more than once and others not included. A sample $L(\theta)$ of "n" size is selected from the test set (L) of N size as a training set modified for each new tree. Each predictor $T_{L(\theta)}$ is dependent on a random vector θ that indicates that the samples selected belong to the entire set (L). In the end, the predictor (f) is the majority vote—in other words, the one with the highest number of votes on the tree—or the mean of the trees (with $y_{\eta}^{'}$ as a predicted answer for samples x_{η} and with K equal to the size of the sample).

The second approach is related to the restriction of variables, randomly selected, in each node. Thus, the vector θ indicates both the primary selection and the randomization [11, 13, 15–18].

The CART algorithm was proposed for the learning of classification and regression trees. The algorithm stated that given a training sample L with N samples, formed by M predictor variables x_i ($i = 1, 2, \ldots, M$) as the entrance space X and one response variable y, the CART recursively divide the entrance space to obtain a predictor for the T_L tree (with y' as the response, variable), according to Eq. (24.1). Having all the entrance space, the algorithm tries to find a binary partition to maximize the purity of the response on the subspace formed by the suggested partition. The algorithm depends on the homogeneity of the response classes; a commonly used measure to classify it is the Gini's impurity measure, while the mean error is used for regression. The binary partitioning is repeated on each new subspace up to homogeneity of sub-spatial response is achieved. The subspace estimate for a particular point is the majoritarian value, classification, or the mean, for regression, of responses of training on the subspace [13, 19].

$$Y' = T_L(X) \tag{24.1}$$

The decision trees are algorithms wherein the main idea is founded in principle "*divide – and – conquer*," in which data are divided in a training dataset in a recursively way. They present characteristics such as: are non-parametric, work both with homogeneous or heterogeneous data (or both), are easy to interpret, and are relatively resistant to the presence of outliers [20, 21].

As already mentioned, Random Forest is an algorithm that uses the majority voting (for classification) or the mean (for regression) to make predictions. For a set to be more precise than its members, two conditions must be met: the set member must have better individual predictions than the random ones and the members of the set must be diverse, meaning that the errors in prevision are not correlated [13].

Equations (24.2) and (24.3) present the calculations for the use of RF for regression and classification, respectively.

$$y' = f\left(x_\eta\right) = \frac{1}{K}\sum_{K=1}^{K} T_{L(\theta_K)}(x_\eta)_1^K \tag{24.2}$$

$$y' = f\left(x_\eta\right) = \text{voto majoritario}\left\{T_{L(\theta_K)}(x_\eta)\right\}_1^K \tag{24.3}$$

In Fig. 24.2, the structures of decision trees, showing the "*if–then*" for each branch, with the partitioning of the subspace associated with a hypothetic two-dimensional space. The individual predictions of all trees are collected and combined as a unique prevision by voting or by mean [13]. The key parameters of the RF model are *mtry*,

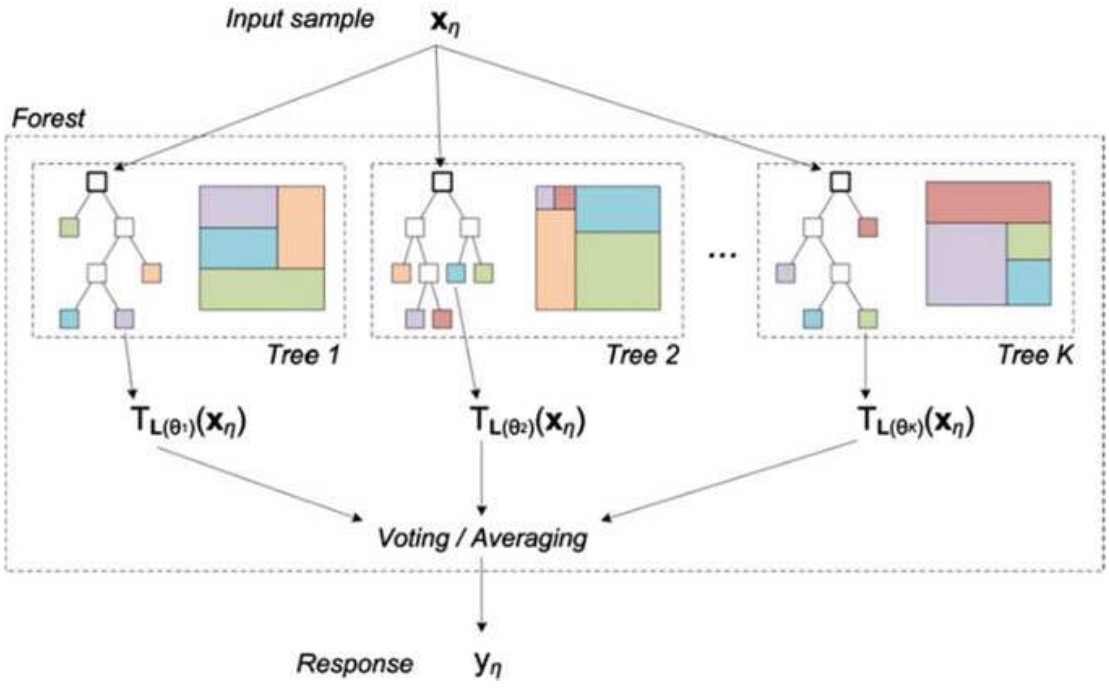


Fig. 24.2 Scheme of the Random Forest proposed by Auret and Aldrich [7]

and *ntree*, in which *mtry* represents the square root of the number of factors and *ntree* represents the number of trees in the forest.

For the growing of trees, one has a tree that can present high growth, causing *overfitting*, which can be explained as the decomposition of the generalization error in terms of bias and variance [21]. Oshiro [22] argue that to avoid this problem, it is necessary to use a technique known as *pruning*, in which there is the generation of a more generic hypothesis from the training dataset. There are two types of *pruning*:

- a pre-pruning, in which the tree construction is interrupted during its growth due to a pre-established criterion; and
- the post-pruning, in which a complete tree is built and, later, sub-trees are removed, based on the estimate error calculation of a determined node. The first type can induce more errors.

Other relevant characteristics in using RF are the no need to perform cross-validation or a separated test to obtain an impartial test error estimate. This fact is caused by the fact that each tree is built using one bootstrap sample, different from the original data, and the third part of the omitted cases is not used in tree construction, is set under the tree. Instead, a classification test is performed and, later, the out-of-bag error is estimated. In this way, it is possible to note that the error rate is related to the correlation between any two trees of the forest and each tree's strength on it [16–18, 20–23].

Girolamo Neto [11] endorses that several evaluation and performance measures can be applied to analyze the behavior of an RF, such as performance measures derived from a matrix that represents matchings and errors of the model (named confusion matrix), followed by a graphical analysis of ROC (Receiver Operating Characteristics) type and the agreement index, Kappa. However, it is not advisable to use the index.

Finally, Matin et al. [24] quote that there was an increasing interest in RF models in the last decade, but they have not yet appeared expressively in rock mechanics literature, for example.

Use of R Language to Rock Mass Geomechanical Classification—Case Study

The origin of R statistical, computational environment, and language [4] is linked to the implementation of S computational language, developed by ATT&T Bell Laboratories. It integrates the GNU Project (GNU's Not Unix), a world project that has developed the GNU computational language, aiming to create and maintain a free complete software system (available as a Free Software as a font code). It is compatible with several platforms, such as Linux, Windows, MacOs. R is a statistical calculation system and graphic construction that uses its programing language. It calculates mathematic, statistic, and data management functions, and its functions

operate on data structure, including lists, vectors, lettering, factors, and matrices. R language is a functional programming environment in which the users can use internal functions or write their functions [25] in a free and applied way in several fields of knowledge.

From the following R packages: gstat [26], sp [27], spacetime [28], raster [29], rgeos [30] (2020), rgdal [30], lattice [31], moments [32], plotKML [33], GSIF [34], ranger [35], geoR [36], plotly [37], DescTools [38], readxl [39], psych [40], ggplot2 [41], dplyr [42] caret [43], corrplot [44], spatstat [45], maptools [30], scatterplot3d [46], tcltk2 [47], doParallel [48], GGally [49], e1071 [50], rpart [51], mlbench [52], randomForest [53], party [54], MASS [55], nycflights13 [56], gapminder [56], Lahman [57], and htmlwidgets [58] it was possible to, through R language, to develop the geomechanical model based on drillhole data, defining a script that better represented the geomechanical characteristics of rock masses in a underground mine. On Fig. 24.3, it is presented the area in which the proposed methodology was applied.

The study area is located in the middle of a shear zone where mineralized hydrothermal fluids have percolated, originating zinc, and other associated minerals. The host rocks of the zinc deposit are composed of political carbonates, very susceptible to dissolution and weathering, generating cavities into the rock mass, which can be filled with many different weathering and soil-like materials. According to the previous studies and in situ information, the rock mass is fractured, with intense water percolation, that originates thick dissolution and weathered zones. This process is responsible for cavity formations, which, usually, are filled with clay plastic material with rock blocks. A thick shear zone where several anastomosed surfaces cross one to another composes the mineralized body, isolating lenticular bodies with metric decametric dimensions. In addition, to mass failures, it can be

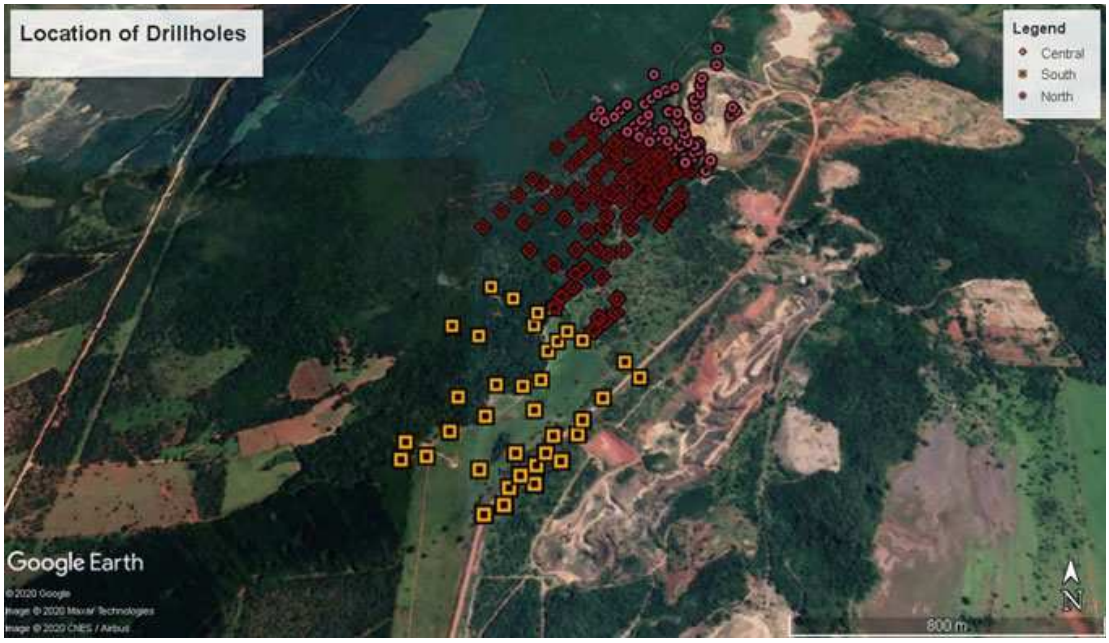


Fig. 24.3 Location of drill holes on north sector of the underground mine from which the database was collect for the study

noted that those are commonly controlled by shear zones, originating discontinuities into the rock masses. These discontinuities have an irregular spacing, form metric to a decametric, low strength (plan to low undulated), and are generally filled by clay material with millimetric thickness.

RMR (Rock Mass Rating) and Q (Tunneling Quality Index) were insufficient to identify those subtle differences in rock mass quality, and a specific and local geomechanical classification was developed for the mine and is presented in Table 24.1.

For the design of this local classification, the empirical knowledge of the mine rock mass and drill hole and laboratory tests data were used. Based on this information, it was concluded that the most relevant parameters for local rock mass classification should be: fracturing and weathering degree, structural pattern, crack presence, drill hole recovery, and RQD. A rock bolting design was defined for each geotechnical rock mass class possessing the structural pattern and back analyses of observed failures. The support type indicated for each class was defined by a trial and error procedure based on a theoretical and practical foundation.

Data Bank: Treatment and Analyses

In practical mining activities, the estimate of mineral resources uses drill hole data, in which information such as extension and geometry of mineralized ore bodies and physical and chemical data [59]. The drilling of drill holes is a routine activity in which, once drilled, the drill holes are made a lithological and geomechanical description of collected samples. This description is made for segments of similar behavior, from the borehole mouth to its end. Usually, a description of different lithotypes, geological structures, and geomechanical characteristics is performed. All this information is stored in a databank and later used for different purposes and designs, such as geological modeling, block modeling, or to construct geological-geotechnical cross-sections. Jakubec and Esterhuizen [60] alert that misconceptions can occur during description, such as difficulties in differentiating natural and mechanical fractures, recognizing filling materials, and others. So, this activity must be performed in a criterion manner by specialized technicians.

Schepers et al. [61] state that drill holes supply the need for precise information on physical parameters of the rock formations, but this information is limited to a limited distance from the drill hole. Data obtained from drill holes are fundamental for decision-making in mining. Definitions on which direction the ore body extends; the ore content; the best technique to extract this ore; and which are the geomechanical characteristics of the sampled rock masses are possible due to drill hole data. So, the most reliable is drill hole data. The best will be the quality of geological and geotechnical models based on such information. This quality is related to drilling the drill hole itself and mainly to the quality of the description of the information obtained from samples collected in the drill hole.

Table 24.1 Adapted geomechanical classification for the rock mass under study

Geomechanical class									
Rock mass class	Weathering degree	Fracturing degree	Rock quality	Structural pattern	Lithology	Crack presence	Block fragmentation	Recovery	RQD (%)
II-A	A2	F2-F3	Very good	Floor slab	Dolomites	Do not occur	m^3	>95	>60
II-B	A2	F2-F3	Very good	Wedge	Breccia	Do not occur	m^3	>95	>60
III-A	A2-A3	F3	Very good	Floor slab	Dolomite	Do not occur	dm^3 a m^3	90 a 95	50 a 75
III-B	A2-A3	F3	Very good	Wedge	Breccia	Do not occur	dm^3 a m^3	90 a 95	50 a 75
IV-A	A3	F3-F4	Good/average			Thick cm to dm	cm^3 a dm^3	>90	25 a 50
IV-B	A3-A4	F3-F4	Average/poor			Thick cm to dm	cm^3 a dm^3	75 a 95	25 a 50
V	A3-A4	F4	Very poor			Thick cm to dm	cm^3 a dm^3	50 a 75	<25
VI	A2	F4-F5	Very poor	Floor slab	Phyllite	Do not occur	cm^3 a dm^3	>95	<25
VII	A4	F4-F5	Very poor			Metric thickness		<50	<25
Pillar	Regardless of the above parameters, the effect of induced stresses must be considered								

Weathering degree

A2—Little weathering, some minerals are discolored. Rock matrix loses very little strength to hammer blow, and it is not friable

A3—Very discolored rock; mineral alteration is very penetrative, assuming colors predominately orange and reddish. Material is not friable—moderate strength to hammer blow

A4—Completely weathered rock with well-developed clay portions. Friable fragments are present and show the original rock structure

A5—Soil. The original rock texture is wholly destroyed, and the soil generally presents reddish color

Fracturing degree

F1—Low fractured: average spacing of fractures is bigger than 2 m, forming rock blocks with a volume equal to cubic meter

F2—Low fractured: average spacing of fractures from 0.6 to 2 m, forming rock blocks with a volume equal to cubic meter

F3—Moderately fractured: average spacing of fractures from 0.06 to 0.2 m, forming rock blocks with volume varying from cubic centimeters to cubic decimeters

F4—Very fractured: average spacing of fractures from 0.06 to 0.2 m, forming blocks with volume from cubic centimeters to cubic decimeters

F5—Very fractured: average spacing of fractures lower than 0.06 m, forming rock blocks with volumes lower than or up to some cubic centimeters

Bieniawski [62] advocates that geomechanical need to be collected in sufficient quantity and quality and that the interpretation of this data should be specific for each project.

In this context, to support the geomechanical model of the study area, a total of 513 drill holes, several laboratory tests, and local (adapted) geomechanical classification data were used. Nevertheless, no lithological, geomechanical, and structural maps/models were provided. The mining sector used for this study is subdivided into three subsectors: South, Central, and North. In Fig. 24.3, it is presented the drill holes available for each of these subsectors—73 for South, 170 for Central, and 270 for North.

For each one of the subsectors, the following data was provided: (i) drill hole coordinates, both for drill hole "mouth" and in-depth, allowing the definition of the actual position of each drill hole into space; (ii) geomechanical parameters, for each segment of the drill hole, such as rock mass local classification, weathering, fracturing, recovery, and RQD; (iii) lithotypes for each segment of the drill hole; and (iv) the inclination of each drill hole and the geological-geomechanical cross-section that each one intercept.

The method used for the databank construction was partially based on the proposal by Kaufmann and Martin [2] and has been considered input data for the drill holes. All data were used from the information collected from drill holes, except the variable recovery, because inconsistencies were observed, such as high values of recovery associated with high weathering and fracturing levels and high RQD values.

Data treatment has begun with the adequacy of the coordinate system that was all modified from the local coordinate system to UTM, Datum Córrego Alegre 23 S.

A script was developed in R [4] to compile drill hole position information and all its geomechanical classification parameters—weathering, fracturing, recovery and RQD, rock mass classification, and lithology. The script has furnished, as an answer, a final datasheet with the following columns: one column with drill hole id and its name; three columns with x, y, and z coordinates, respectively; two columns with the geomechanical classification of the rock mass, one with the local classification and the second with a simplified RMR classification; one column with the lithotypes; and, finally, three columns with weathering, fracturing, and RQD. In Table 24.2, it is presented an example of a final spreadsheet, without the id and coordinates data, for confidentiality reasons.

The joint evaluation of variables recovery and RQD would aim to identify the occurrence of gaps (cracks—FD) associated with the regions of the rock mass with high RQD values. However, on the provided databank, it was possible to observe that for some depth segments, there was a low value of recovery associated with high values of RQD and low fracturing and weathering values, which does not make sense. So, because of such inconsistencies, the variable recovery was excluded from the analyses.

The quality and quantity of collected data are defined by the geotechnical project's purposes to be developed. However, on the case study presented here, the type and the quantity of data needed a thorough analysis to guarantee that they would fit the necessary quality for the proposed use.

Table 24.2 Example of a final spreadsheet with final databank

ID	X	Y	Z	Class	Class local	Lithology	Weathering	Fracturing	RQD
Drillhole 1	X_0	Y_0	0	5	0	S/R	0	0	0
Drillhole 1	$X_0 + 0.01$	$Y_0 + 0.01$	44.95	5	0	S/R	0	0	0
Drillhole 1	X_1	Y_1	44.95	3	2.5	DORO	2.5	2	84.72
Drillhole 1	$X_1 + 0.01$	$Y_1 + 0.01$	51.1	3	2.5	DORO	2.5	2	84.72
Drillhole 1	X_2	Y_2	51.1	4	6	DORO	3.5	2	39.02
Drillhole 1	$X_2 + 0.01$	$Y_2 + 0.01$	55.2	4	6	DORO	3.5	2	39.02
Drillhole 1	X_3	Y_3	55.2	5	9	DORO	3.5	1	13.6
Drillhole 1	$X_3 + 0.01$	$Y_3 + 0.01$	87.7	5	9	DORO	3.5	1	13.6
Drillhole 1	X_4	Y_4	87.7	4	2.5	DORO	3	2	69.23

So, in possession of the compiled spreadsheet, data were analyzed and treated. Initially, the quality of data was sought to allow its use for multidimensional geomechanical modeling. So, some criterion was used to grant its quality and also to allow its reading and analyses.

The first criterion was compatible with the lithological and geomechanical information segments to obtain equal intervals that contained both information. Then, each drill hole segment's geographical coordinates (x and y) were obtained based on the drill hole plunge azimuth and orientation; i.e., a correction of drill hole inclination was performed. Six drill holes were excluded from the databank because they do not have any information on its geomechanical parameters. Some others were excluded, as they do not have any information of its azimuth and plunge.

The initial and final value of each segment was considered. The final value of each segment was coincident with the initial one from the following segment. So, to avoid duplicate information at one only depth, a value equal to 0.01 was added for x and y variables to the first point of each segment, starting with the second segment up to the end of the drill hole, as shown in Table 24.3.

The last criterion is RQD, as the value provided on the drill holes description was used on the databank. All other geomechanical parameters—weathering, fracturing, and local geomechanical classification were reclassified according to a numerical scale, as presented in Tables 24.4, 24.5 and 24.6. For the cases in which, for example, the weathering grade was classified as A2/A3, a value equal to 2.5 was considered on

Table 24.3 Example of the artifice used for definition of drillhole depth

ID	X	Y
Drillhole 1	X_0	Y_0
Drillhole 1	$X_0 + 0.01$	$Y_0 + 0.01$
Drillhole 1	X_1	Y_1
Drillhole 1	$X_1 + 0.01$	$Y_1 + 0.01$
Drillhole 1	X_2	Y_2
Drillhole 1	$X_2 + 0.01$	$Y_2 + 0.01$
Drillhole 1	X_3	Y_3
Drillhole 1	$X_3 + 0.01$	$Y_3 + 0.01$
Drillhole 1	X_4	Y_4

Table 24.4 Numerical scale adopted for weathering

Weathering	
Drillhole description	Numerical scale
A1	1
A2	2
A3	3
A4	4
A5	5

Table 24.5 Numerical scale adopted for fracturing

Fracturing	
Drillhole description	Numerical scale
F1	1
F2	2
F3	3
F4	4
F5	5

Table 24.6 Numerical scale adopted for local geomechanical classification

Local geomechanical classification	
Drillhole description	Numerical scale
II-A	1
II-B	2
III-A	3
III-B	4
IV-A	5
IV-B	6
V	7
VI	8
VII	9

Table 24.7 Legend of lithotypes present on the databank

Legend	Lithotypes
ARD	Slate
BXD	Hydrothermal breccia
BXD/Dcz	Dolomitic breccia/gray dolomite
BXD/Doro	Dolomitic breccia/pink dolomite
BXH	Hematitic breccia
BXW	Willemitic breccia
DCZ	Gray dolomite
DCZ/bx	Gray dolomite
DO	Dolomite
DORO	Pink dolomite
DORObd	Pink dolomite
DORObx	Pink dolomite
DOROfr	Pink dolomite
DRA	Dolarenite
DRU	Dolorudite
FD	Crack
FIL	Phyllite
FO	Shale
HC	Compact hematite
MB	Metabasic rock
MGA	Marl
MT	Terrain material
MT/CAL	Terrain material/calamine/calamine
MTARG	Clay material
R/S	Destroyed
S/R	Not recovered
SOLO	Soil

the numerical scale. The same reasoning was used for fracturing and local geomechanical classes. Due to the absence of drill hole bulletins, RQD was determined for equal fracturing segments, not maneuver segments. Lithotypes present on the data bank are listed in Table 24.7, and no modification of such data was done.

After finishing data treatment, a quality analysis was performed, as for some drill hole segments, no geomechanical data (weathering, fracturing, and/or RQD) was available. The lack of information for some segments can be related to not recovering rock material during drilling. In this case, an S/R or R/S symbol was used. In addition, geological factors such as unconsolidated material, discontinuities intersection, secondary porosity, fault zones, and others can affect the material recovering

on drill holes. Another possible cause can be inability during drill hole description, as trained technical staff must be used for this procedure.

Fe and FD symbols on some drill hole segments indicate voids/cracks/gaps into rock masses. Therefore, for these segments, geomechanical parameters were also considered to be equal to zero.

For all segments in which the lithotypes were classified as Soil, R/S (destroyed), and S/R (Without Recovery), geomechanical values equal to zero were given, as there is no sense in giving rock mechanical parameters for soil-like materials.

Pells et al. [63] argue that there are problems in the use and determination of RQD around the globe and that the use of such an index must be carefully performed. The problems mentioned above go from errors during samples drill hole samples description to misunderstanding the index itself and comprise the lack of ability of the professional responsible for its quantification. However, despite all these problems, RQD is still a used parameter used in rock mechanics worldwide, and its use was kept for the present study.

In order to classify the rock mass understudy, a modified RMR was used, as there was not all the information needed to the traditional RMR—no information on discontinuities characteristics and uniaxial compressive strength. Therefore, on this modified RMR classification, the weight of rock matrix strength was not considered, and the weight related to discontinuities, only the weight for weathering was considered. Also, the same weight for water presence was used for all drill holes (equal to 4, based on field information). So, the modified RMR classification was based on the following parameters: RQD, fracturing equivalent to spacing, weathering, and water presence.

As presented in Table 24.8, the weight for RQD, weathering, fracturing, and water presence was kept equal to the original RMR classification, but the final sum of weights was reorganized for each RMR class.

By comparing the simplified RMR (Table 24.7) to the local classification (Table 24.1), one can note that the differences between the classifications are mainly because the local classification considers, besides the RMR criterion, the rock quality, the structural pattern, and the structural pattern the block fragmentation. These data are not available on the available data bank and had to be discarded.

The modifications made to the data bank, previously described, allowed an improvement in data quality and reliability. At the end of the analysis, a CLASS variable corresponding to the simplified RMR value was included in the data bank. All this data treatment process was automatically performed by using R language and manually revised lately. So, after this process, the data bank for the South subsector comprises a total of 70 drill holes; the Central subsector a total of 255 drill holes; and the North subsector a total of 149 drill holes, with a drill hole reduction of 4.1%, 5.6%, and 12.4%, respectively.

Table 24.8 Simplified RMR classification

Parameters	Coefficients					
1	Rock matrix strength	Not considered				
2	RQD	91–100	76–90	51–75	26–50	<25
	Weight	20	17	13	8	3
3	Fracturing (interval—from drillholes, Table 24.3)	>2 m (0–1.5)	0.6–2 m (1.6–2.5)	200–600 mm (2.6–3.5)	60–200 mm (3.6–4.5)	<60 mm (4.6–5)
	Weight	20	15	10	8	5
4	Weathering (interval from drillholes, Table 24.3)	Not weathered (0–1.5)	Slightly weathered (1.6–2.5)	Moderately weathered (2.6–3.5)	Very weathered (3.6–4.5)	Decomposed (4.6–5)
	Weight	6	5	3	1	0
5	Presence of water	Completely dry	Interstitial water	Humid	Low flow	Water inflow
	Weight	15	10	7	4	0
Global weight		49–61	38–48	30–37	22–29	<21
Simplified RMR classification		I	II	III	IV	V

Application of Random Forest to the Data Bank

Random Forest was used to listing which variable (or variables) would be the most important to explain the mechanical behavior of the rock mass understudy on the underground mining in the North sector. Kaufmann and Martin [2] argue that the primary goal of selecting variables is: to improve interpretation capacity, eliminating irrelevant input to create a model with a small number of variables; to reduce noise, avoiding overfitting risk; to accelerate modeling time, improving model precision; and, finally, to furnish a better definition of the process of data generation. Still, according to those authors, the use of RF to measure the importance of variables has several advantages in comparison to other similar methods, such as it can determine the impact of each predictor variable individually, as well as on multivariate interactions with other predictors; it grants a measure of impartial importance and can be applied to several fields of knowledge.

RF provides two measures of importance—the Mean Decrease Accuracy (*IncMSE*) and the Mean Decrease Gini (*IncNodePurity*). These two parameters can be used to classify and select variables. The IncMSE quantifies the importance of a variable by measuring the change into prediction precision when this variable's values are randomly commuted compared to the original observations. It is a measure of how much the precision reduces if a variable is excluded from the model. The IncNodePurity is the sum of all the reductions on Gini impurity due to a determined

variable (whenever this variable is used to form a division into RF), normalized by the number of trees. The IncNodePurity measures the contribution of each predictor variable to the impurity of the trees resulting from the RF model and reflects how much each variable contributes to the homogeneity of the nodes and leaves [64–67].

The results obtained for all subsectors are presented in Fig. 24.4 and Table 24.9. For all three subsectors South, Central, and North, a regression of variables was performed throughout RF. As a result, the number of used trees in all subsectors was equal to 1000 ($ntree = 1000$), and the number of nodes of the analyzed trees was equal to 6 ($mtry = 6$).

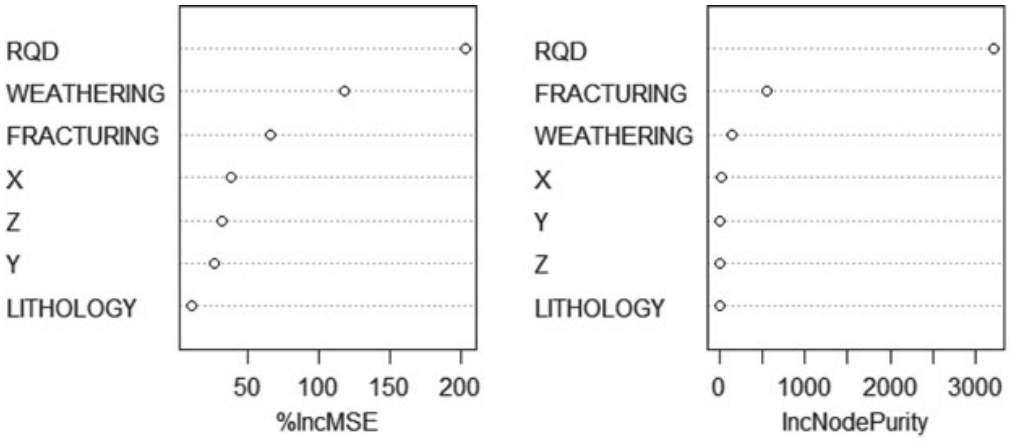
Table 24.9 presents a resume of the determination coefficients, obtained errors for predicted and actual values, as well as the percentage of each variable explained by RF. As observed in Table 24.9, it is found that more than 99% (*% var explained*) of the rock mass behavior of the North Edge of the underground mine can be explained by the results obtained throughout RF. Concomitantly, the errors (RMSE and MAE) were very low, and the determination coefficient (*R-squared*) was very close to 1. So, it can be concluded that the ranking of the importance of the variables by using RF is reliable to explain the behavior of the studied rock mass.

The IncMSE of the South sector presented the variables RQD, weathering, and fracturing as the ones that most influence rock mass strength and lithology is the less important. For the Central subsector, the variables that better explain the strength behavior of the rock mass are RQD, fracturing, and weathering, and again, the less critical variable was lithology. Finally, the most important variables were observed in the South subsector—RQD, weathering and fracturing, and lithology was the less important one for the North subsector. Based on these results, it became clear that RQD is the variable that can better explain the behavior of the rock masses on all analyzed subsectors, followed by weathering and fracturing. Based on this result, RQD should be considered the independent variable to construct geomechanical data of the underground mine. So, RQD and RMR simplified maps were generated for each subsector.

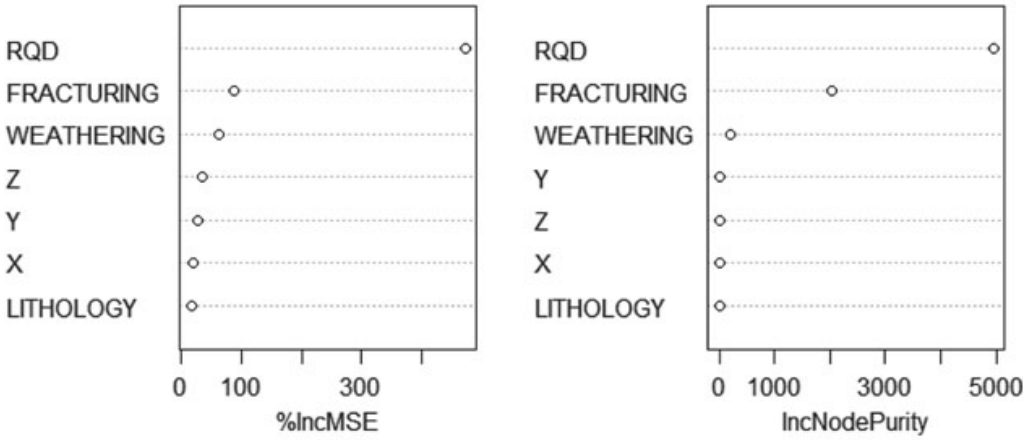
The relevance of RQD was already expected and has proved to be coherent, as this parameter compiles weathering and fracturing characteristics in only one index and is a measure of drill hole recovery.

By comparing the results of RF with data from exploratory analyses, it can be noted that the mean RQD value is higher for the Central subsector. This is because this region has a minor structural control compared to the other two subsectors, as this subsector is out of the fault zone. On the other hand, the North subsector presents the lower mean value for RQD, and this result can be linked to several faults and structural lineaments in this area.

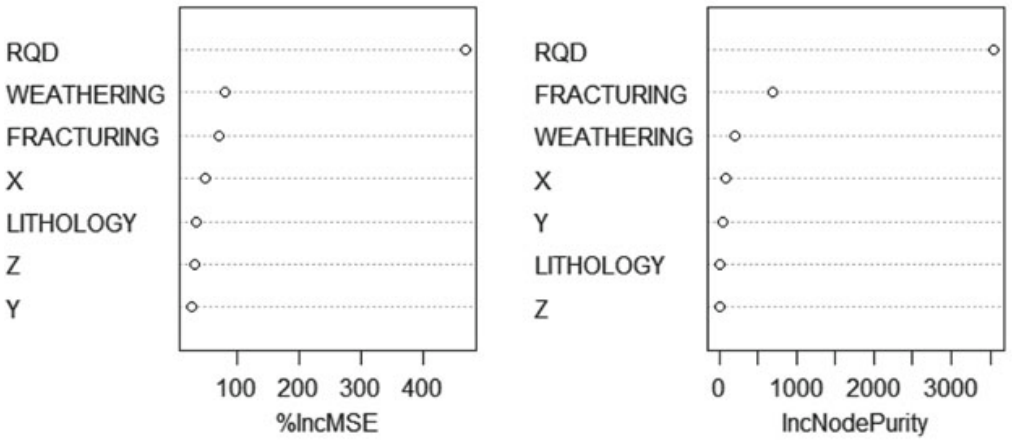
Among the results of RF, it was expected that, with the increase of depth, there would be an increase in strength and, consequently, a better rock mass behavior. However, as the region has an expressive presence of carbonatic rocks and underground water, the association of these two geological characteristics explains the fact that the increase of depth does not present a direct and transparent relationship with the strength increase, as even for very high depths (up to 600 m) there is the presence of dissolution voids/gaps/cracks, filed or not with soil-like materials.



A - South Subsector



B - Central Subsector



C - North Subsector

Fig. 24.4 Ranking of the importance of each variable of the rock mass. On the abscissae axis, the importance of each variable; and on the ordinate axis, the variables, with information about the mean decrease precision (InsMSE) and of Gini mean decrease (IncNodePurity)

Table 24.9 Resume of determination coefficients and errors obtained through RF

Subsector	RMSE (average square error)	R-squared (determination coefficient)	MAE (absolute square error)	% of explained variable
South	0.0157	0.9999	0.0018	99.96
Central	0.0454	0.9987	0.0041	98.86
North	0.0286	0.9997	0.0021	99.81

The technical literature related to geomechanical behavior consulted by the authors do not present works in which the RF was used to predict the importance of variables and its later use for geomechanical modeling. However, some studies use RF to predict one variable's behavior based on a set of variables for both rock matrix and rock masses. Among the relevant contributions consulted, Tiryaki [68] has used neural networks, regression trees techniques, principal components, and factor analyses to predict the uniaxial rock strength and the elasticity module of three different rock types. Dong et al. [69] have used RF to classify if the rock-burst phenomenon would occur and its intensity in underground mining. The authors have analyzed factors that can start the phenomena, such as: in situ stresses, uniaxial compressive strength, tension strength, and others. Kaufmann and Martin [2] has used RF to predict the elasticity module and uniaxial compressive strength based on some selected variables: porosity, Point Load Index (Is_{50}), wave propagation velocity (V_p), and Schmidt accelerometry. Finally, Xie and Peng [70] have used regression RF to predict damage zones on excavations in an underground coal mine, considering that these zones were affected by disturbs caused by explosives, stress relief due to excavation process, and by adjustments of in situ stresses.

Joint Analyses

5D iterative graphics and animations were built and allowed joint analyses of different variables of the databank. Among the interactive graphics, a graphic is presented in Fig. 24.5, in which it is possible to visualize the simplified RMR classification for each lithotype. On this graphic, the variable CLASS was interpolated using Ordinary Kriging, and each colored dot indicates which lithotype was observed in the field. So, there is an association of predicted values for the CLASS variable with the sampled lithology points obtained from drill holes. So, in Fig. 24.5, only the visualization of voids/gaps/cracks can be seen, and it allows observing that some low RMR regions (Class 4) can be associated with these structures' presence.

For the CLASS × RQD graph (Fig. 24.6), the same behavior of Fig. 24.5 can be observed, as the CLASS variable is interpolated with RQD, and it can be seen that high RQD regions are related to good RMR (Class 3) materials for South subsector.

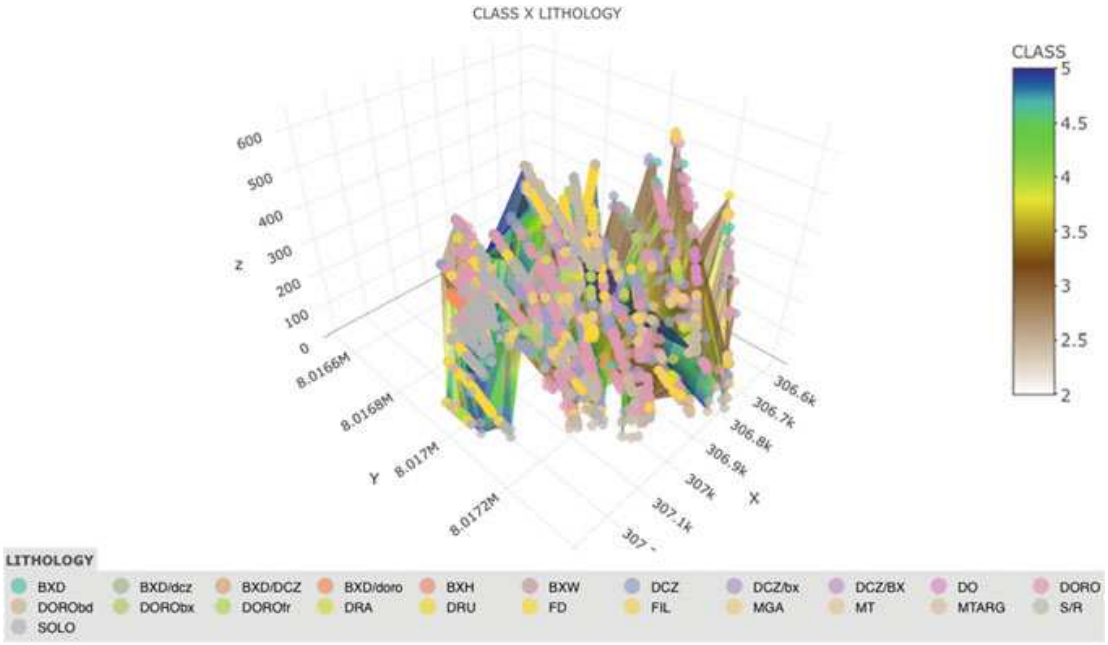


Fig. 24.5 Simplified RMR classification and its interaction with lithotypes

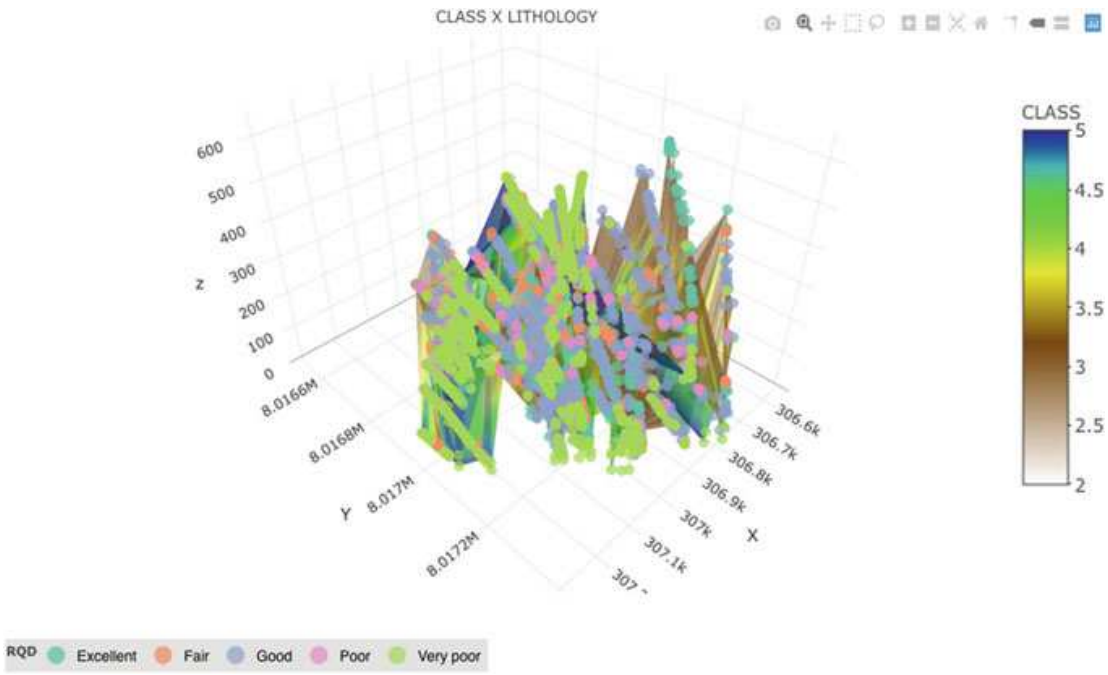


Fig. 24.6 Simplified RMR classification and its interaction with RQD values

In Fig. 24.7, the relationship between CLASS and weathering can be seen, and one can note that weathering materials classified as equal to 2 are related to several RMR classes for the North subsector.

Iterative graphics were also generated for joint analyses of the CLASS variable with depth. The same iterative graphs mentioned for the rock mass classification were also generated for RQD, as, for example, RQD × Weathering.

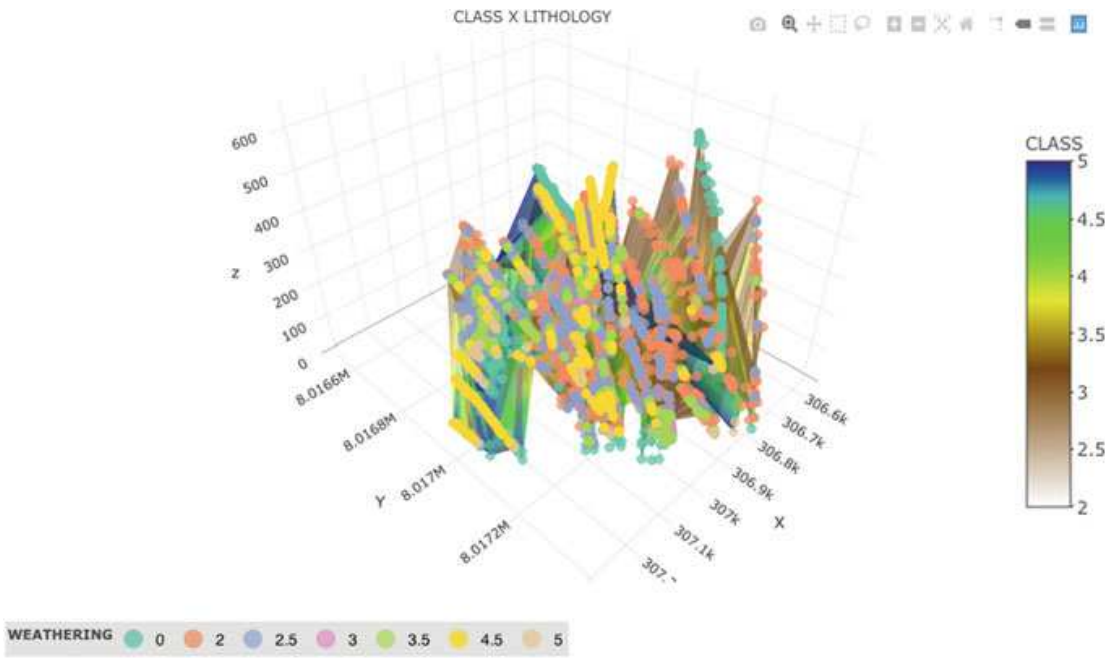


Fig. 24.7 Simplified RMR classification and its interaction with weathering

Still, as an analyses tool, animations in which three or more variables were jointly analyzed. For example, lithology × fracturing × CLASS (Fig. 24.8) or lithology × fracturing × RQD (Fig. 24.9). It must be highlighted that to present/explain this kind of toll only in 2D harms its understanding, as only one frozen image can be shown and not the role animation itself.

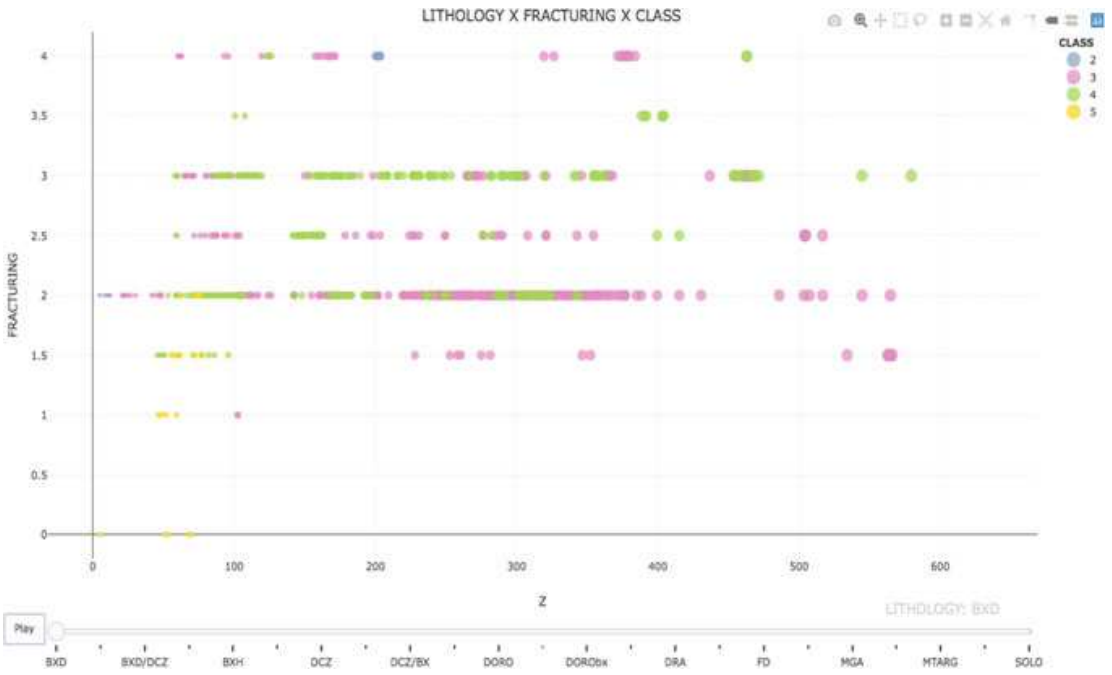


Fig. 24.8 Image of an animation correlating three variables: lithology × fracturing × CLASS

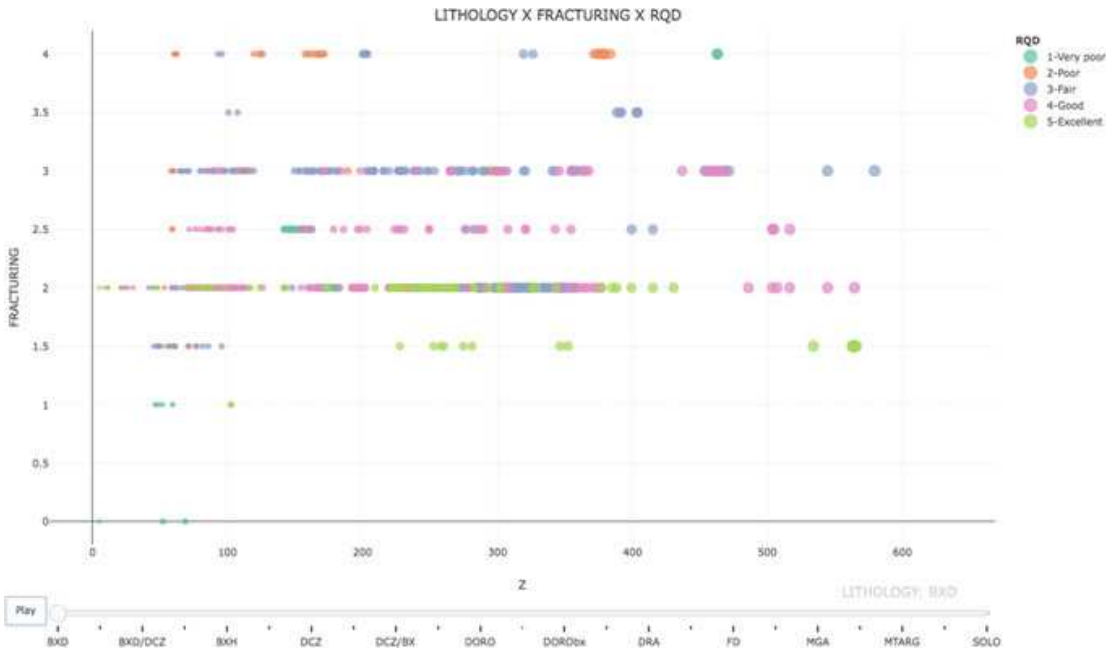


Fig. 24.9 Image of an animation correlating three variables: lithology × fracturing × RQD

The iterative graphs have two significant advantages on geological-geotechnical modeling. The first is to analyze the essential variables that govern the geomechanical behavior in an easy, rapid, and iterative way. On the graphs that correlate RQD × Weathering, as an example, it is possible to visualize points/regions in the interior of rock masses that have low RQD values. So, it is possible to verify how Weathering behavior in these same places. The second advantage is to allow a better comprehension by professionals of correlate areas. Herwanger [71] advocate that 3D and 4D geomechanical models integrate data from several disciplines and allow that scientist and engineers collaborate and communicate with each other on the way for more successful decision-making.

Conclusions

The information collected on drill holes was used for developing a simplified RMR classification named CLASS. On this classification, those segments described as voids/cracks/gaps (named FD) without recovering (S/R or R/S) or soil were classified as having a value equal to zero. In addition, previous treatment of the used databank was performed to withdraw any conflicting information that could harm the quality of the geomechanical modeling.

Using a machine learning technique (Random Forest), it was possible to identify, among all considered variables for rock mass classification, the one that better explains the geomechanical behavior of the rock mass of all subsectors of the North of an underground mine is the RQD. Despite being already expected, this result did not

result in disregarding the other variables for the analysis of the observed mechanical behavior.

The proposed method for developing geomechanical modeling using drill hole data and free programming software (R software) was satisfactory, both for the visualization of the results and the quality of the information itself. However, it is still necessary for the specific study area to calibrate and validate the results with field data or in a new drill hole database. Meanwhile, it became clear that the potential of the developed tool for geomechanical modeling is encouraging and should be better explored in future.

References

1. Landim, P. 2003. *Statistical analysis of geological data.* Sao Paulo: UNESP.
2. Kaufmann, O., and T. Martin. 2009. Reprint of "3D geological modelling from boreholes, cross-sections and geological maps, application over former natural gas storages in coal mines" [Comput. Geosci. 34 (2008) 278–290]. *Computers & Geosciences* 35 (1): 70–82.
3. Hadjigeorgiou, J. 2012. Where do the data come from? *Mining Technology* 121 (4): 236–247.
4. Team, R.C. 2013. *R: A language and environment for statistical computing.*
5. Bhattacharya, S., et al. 2019. Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: A case study from the Marcellus Shale, North America. *Journal of Petroleum Science and Engineering* 176: 702–715.
6. Silva, C.D.R.L., and J.A.S. Centeno. 2007. Study of the interpolation of the inverse distance to a power method. In *Simpósio Brasileiro de Geomática*, 57–62. São Paulo: Presidente Prudente.
7. Auret, L., and C. Aldrich. 2012. Interpretation of non-linear relationships between process variables by use of random forests. *Minerals Engineering* 35: 27–42.
8. Sajid, A., R. Rudra, and G. Parkin. 2013. Systematic evaluation of kriging and inverse distance weighting methods for spatial analysis of soil bulk density. *Canadian Biosystems Engineering* 55.
9. Setianto, A., and T. Triandini. 2013. Comparison of kriging and inverse distance weighted (IDW) interpolation methods in lineament extraction and analysis. *Journal of Applied Geology* 5 (1).
10. Li, J., and A.D. Heap. 2008. A review of spatial interpolation methods for environmental scientists.
11. Girolamo Neto, C.D. 2014. Potential of data mining for mapping of coffee plantation areas. Monography for SER-300 course. Cited 2021 June 4. Available from: http://wiki.dpi.inpe.br/lib/exe/fetch.php?media=ser300:alunos2014:cesare_monografia.pdf.
12. Breiman, L. 2001. Random forests. *Machine Learning* 45 (1): 5–32.
13. Strobl, C., et al. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8 (1): 1–21.
14. Harris, J.R., et al. 2014. A comparison of different remotely sensed data for classifying bedrock types in Canada's Arctic: Application of the robust classification method and random forests. *Geoscience Canada: Journal of the Geological Association of Canada/Geoscience Canada: Journal de l'Association Géologique du Canada* 41 (4): 557–584.
15. Lorenzett, C.D.C.T.A. 2016. Comparative study between algorithms of data mining, random forest and J48 for decision making. *Simpósio de Pesquisa e Desenvolvimento em Computação* 2 (1).
16. Couteiro, T.A.F. 2010. Relational statistics methods for prevision of medical results. In *Integrated master in engineering, IT and computing*, 67. Porto: Faculdade de Engenharia, Universidade do Porto.

17. Ou, C.-M., W.-J. Hwang, and S.-M. Yang. 2013. FPGA-based online learning hardware architecture for kernel fuzzy c-means algorithm.
18. Harris, J., and E.C. Grunsky. 2015. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Computers & Geosciences* 80: 9–25.
19. Breiman, L. 1993. *Classification and regression trees*. New York: Chapman & Hall. ISBN 0412048418.
20. Carvalho, H.M. 2014. Machine learning applied to data mining: Decision trees. Undergraduation conclusion thesis, in software engineering, Universidade de Brasília, Brasília, 90.
21. Louppe, G. 2014. Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502.
22. Oshiro, T.M. 2013. An approach for building a single tree from a random forest for classifying the basis of genetic expression. In *Biocomputing*. São Paulo: Universidade de São Paulo.
23. Soofastaei, A. 2020. *Data analytics applied to the mining industry*. CRC Press.
24. Matin, S., et al. 2018. Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest. *Applied Soft Computing* 70: 980–987.
25. Grunsky, E. 2002. R: A data analysis and statistical programming environment—An emerging tool for the geosciences. *Computers & Geosciences* 28 (10): 1219–1222.
26. Pebesma, E.J. 2004. Multivariable geostatistics in S: The gstat package. *Computers & Geosciences* 30 (7): 683–691.
27. Pebesma, E. 2012. spacetime: Spatio-temporal data in R. *Journal of Statistical Software* 51 (7): 1–30.
28. Pebesma, E., and R.S. Bivand. 2005. S classes and methods for spatial data: The sp package. *R News* 5 (2): 9–13.
29. Hijmans Robert, J. 2020. *raster: Geographic data analysis and modeling*. R package version 3.1-5.
30. Bivand, R., and C. Rundel. 2017. *RGeos: Interface to geometry engine-open source ('GEOS')*. R package version 0.3–26.
31. Sarkar, D. 2008. *Lattice: Multivariate data visualization with R*. Springer Science & Business Media.
32. Komsta, L., and F. Novomestky. 2015. *Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.14.
33. Hengl, T., et al. 2015. plotKML: Scientific visualization of spatio-temporal data. *Journal of Statistical Software* 63 (5): 1–25.
34. Hengl, T., et al. 2017. *GSIF: Global soil information facilities*. R package version 0.5–4 [Software].
35. Wright, M.N., and A. Ziegler. 2015. ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.
36. Ribeiro Jr., P.J., et al. 2020. *Package 'geoR'*.
37. Sievert, C. 2020. *Interactive web-based data visualization with R, plotly, and shiny*. CRC Press.
38. Signorell, A., et al. 2020. *DescTools: Tools for descriptive statistics*. R package version 0.99.36.
39. Wickham, H. 2018. *nycflights13: Flights that departed NYC in 2013*. R package version 1(0).
40. Revelle, W. 2019. *psych: Procedures for personality and psychological research*. Northwestern University. Version 1: 12.
41. Hadley, W. 2016. *Ggplot2: Elegant graphics for data analysis*. Springer.
42. Wickham, H., R. François, L. Henry, and K. Müller. 2020. *dplyr: A grammar of data manipulation*. R package version 0.8.4.
43. Kuhn, M., et al. 2020. *caret: Classification and regression training*. R package version 6.0-86. Available at: https://cran.r-project.org/web/packages/caret/caret.pdf. Accessed 20 Mar 2020.
44. Wei, T., and V. Simko. 2017. *R package "corrplot": Visualization of a correlation matrix* (version 0.84).
45. Baddeley, A., E. Rubak, and R. Turner. 2015. *Spatial point patterns: Methodology and applications with R*. CRC Press.

46. Ligges, U., and M. Mächler. 2002. Scatterplot3d—An R package for visualizing multivariate data. Technical report.
47. Grosjean, P., E. Lecoutre, and J.C. Faria. 2019. *SciViews-R*. Citeseer.
48. Calaway, R., et al. 2019. *doParallel: Foreach parallel adaptor for the 'parallel' package*. R package version 1.0. 14. Vienna: Comprehensive R Archive Network.
49. Schloerke, B., et al. 2020. *GGally: extension to 'ggplot2'*. R package version 1.5.0. R Foundation for Statistical Computing.
50. Meyer, D., et al. 2019. *Misc functions of the statistics department, probability theory group*. R package version 1.
51. Therneau, T., et al. 2015. Package 'rpart'. Available online: http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf. Accessed 20 Apr 2016.
52. Leisch, F., and E. Dimitriadou. 2010. *mlbench: Machine learning benchmark problems*. R package version 2.0-2010. URL http://CRAN.R-project.org/package=mlbench.
53. Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. *R News* 2 (3): 18–22.
54. Hothorn, T., K. Hornik, and A. Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15 (3): 651–674.
55. Venables, W., and B. Ripley. 2002. *Modern applied statistics with S*, 4th ed. World.
56. Wickham, H., et al. 2019. *Package 'readxl'*.
57. Friendly, M., et al. 2014. *Lahman: Sean Lahman's baseball database*. R package version 3.0-1.
58. Vaidyanathan, R., et al. 2019. *htmlwidgets: HTML widgets for R*. R package version 1.5.1.
59. Webber, T., J.F.C.L. Costa, and P. Salvadoretti. 2013. Using borehole geophysical data as soft information in indicator kriging for coal quality estimation. *International Journal of Coal Geology* 112: 67–75.
60. Jakubec, J., and G.S. Esterhuizen. 2007. Use of the mining rock mass rating (MRMR) classification: Industry experience. In *Proceedings of the International Workshop on Rock Mass Classification in Underground Mining*.
61. Schepers, R., et al. 2001. Application of borehole logging, core imaging and tomography to geotechnical exploration. *International Journal of Rock Mechanics and Mining Sciences* 38 (6): 867–876.
62. Bieniawski, Z. 1988. Towards a creative design process in mining. *Mining Engineering* 40 (11).
63. Pells, P., et al. 2017. Rock quality designation (RQD): Time to rest in peace. *Canadian Geotechnical Journal* 54 (6): 825–834.
64. Calle, M.L., and V. Urrea. 2011. Letter to the editor: Stability of random forest importance measures. *Briefings in Bioinformatics* 12 (1): 86–89.
65. Verikas, A., A. Gelzinis, and M. Bacauskiene. 2011. Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 44 (2): 330–349.
66. Mohammady, M., H.R. Pourghasemi, and M. Amiri. 2019. Land subsidence susceptibility assessment using the random forest machine-learning algorithm. *Environmental Earth Sciences* 78 (16): 1–12.
67. Zaimes, G.N., D. Gounaridis, and E. Symenonakis. 2019. Assessing the impact of dams on riparian and deltaic vegetation using remotely-sensed vegetation indices and Random Forests model. *Ecological Indicators* 103: 630–641.
68. Tiryaki, B. 2008. Predicting intact rock strength for mechanical excavation using multivariate statistics, artificial neural networks, and regression trees. *Engineering Geology* 99 (1–2): 51–60.
69. Dong, L.-J., X.-B. Li, and K. Peng. 2013. Prediction of rockburst classification using Random Forest. *Transactions of Nonferrous Metals Society of China* 23 (2): 472–477.
70. Xie, Q., and K. Peng. 2019. Space-time distribution laws of tunnel excavation damaged zones (EDZs) in deep mines and EDZ prediction modeling by random forest regression. *Advances in Civil Engineering* 2019.
71. Herwanger, J. 2019. 4D geomechanical simulations for field development planning. *Oil Geomechanic* 3 (1): 34–44.

4 A SELETIVIDADE AMBIENTAL DA CASTANHEIRA-DA-AMAZÔNIA COM A UTILIZAÇÃO DA MODELAGEM GEOESTATÍSTICA E DO ALGORITMO RANDOM FOREST.

Resumo

A Castanheira-da-Amazônia representa uma das espécies florestais da Amazônia de grande relevância para a região, contribuindo na conservação das florestas tropicais e servindo de importante fonte de renda e alimentação para as famílias extrativistas através de suas amêndoas e derivados. O estudo teve o objetivo de analisar a relação entre os atributos físicos e químicos do solo com castanhais nativos. As amostras de solo foram coletadas na Propriedade do Jutica – Tefé/AM. Foram coletadas amostras de solo na camada superficial de 0-20cm de uma parcela permanente (300m x 300m), obedecendo as distâncias de 50m entre linhas e 30m entre pontos, totalizando 60 amostras. Foi realizado o inventário de todas as castanheiras com DAP $\geq$ 10cm. Os dados de solo foram avaliados através da estatística descritiva e Geoestatística. O DAP foi interpolado por meio da Krigagem Indicativa com um fatiamento do DAP $\geq$ 50cm e DAP $<$ 50cm. Posteriormente, realizou-se o processo multivariado de regressão espacial com o Algoritmo Random Forest para as predições espaciais. Todas as análises dos dados foram realizadas pelo software R. Os resultados apontaram que a maioria dos atributos estudados apresentam moderado grau de heterogeneidade. O modelo gerou mapas de probabilidade para caracterizar e compreender a distribuição espacial das castanheiras com DAP $\geq$ 50cm e DAP $<$ 50 cm na área e, por fim, indicou os principais atributos do solo na região em estudo.

Palavras-chave: Castanheira-da-amazônia, solo, geoestatística, aprendizagem de máquina.

Abstract

Amazonia nut is one of the most important forest species in the Amazon for the region, contributing to the conservation of tropical forests and serving as an important source of income and food for extractive families through its almonds and derivatives. The study aimed to analyze the relationship between the physical and chemical attributes of the

soil with native chestnut groves. Soil samples were collected at the Jutica property – Tefé/AM. Soil samples were collected in the 0-20cm surface layer of a permanent plot (300m x 300m), obeying distances of 50m between rows and 30m between points, totaling 60 samples. An inventory of all Brazil nut trees with DBH $\geq$ 10cm was carried out. Soil data were evaluated through descriptive statistics and Geostatistics. DBH was interpolated using Indicative Kriging with DBH slicing $\geq$ 50cm and DBH $<$ 50cm. Subsequently, the multivariate spatial regression process was carried out with the Random Forest Algorithm for spatial predictions. All data analyzes were performed using the R software. The results showed that most of the attributes studied present a moderate degree of heterogeneity. The model generated probability maps to characterize and understand the spatial distribution of Brazil nut trees with DBH $\geq$ 50 cm and DBH $<$ 50 cm in the area and, finally, indicated the main soil attributes in the region under study.

Keywords: Brazil nut, soil, geostatistics, machine learning.

4.1 INTRODUÇÃO

A utilização da floresta tropical de forma sustentável tem sido abordada como alternativas de suma importância para a conservação dos ambientes naturais na Amazônia (WADT et al., 2017). O uso sustentável dos PFNMsProdutos Florestais Não Madeireiros tem sido essenciais para a manutenção dessas florestas, constituindo fonte de renda e alimentação para sobrevivência de inúmeras famílias (SFB, 2014). Entre os produtos florestais não madeireiros, temos as amêndoas da Castanheira-da-Amazônia (*Bertholletia excelsa, Bonpl.*) também conhecida como castanha-do-brasil, castanha-da-Amazônia e castanha-do-pará, representando um dos principais produtos florestais extrativistas da região amazônica de relevante interesse, sendo um dos produtos mais apreciados e com alto teor nutricional (TONINI e BALDONI, 2019). A Castanheira-da-Amazônia é definida como a única espécie do gênero *Bertholletia*, pertencente à família *Lecythidaceae*. Denominada como uma árvore símbolo da Amazônia, representante das florestas tropicais não somente pelo seu grande porte e exuberância, mas devido a sua importância social, econômica e ambiental (WADT et al., 2019). Encontrada em florestas de terra firme e em determinados ambientes constituem os castanhais, associada a outras espécies florestais (BRASIL, 2017). As populações de castanhais nativos encontram-se em terras firmes principalmente em solos pobres, bem estruturados e drenados, argilosos ou argilo-arenosos com mais ocorrência em solos de textura média a pesada. (SPERA et al., 2019). Uma floresta exuberante nem sempre representa um solo de alta fertilidade, válido para a floresta amazônica que em sua maioria, apresenta solos naturalmente ácidos, de baixos níveis de fertilidade (WADT et al., 2017). Existe uma carência de literatura a respeito dos ambientes naturais de ocorrência de castanheiras, apesar da castanha-da-amazônia se apresentar como objeto de muitos estudos; no entanto, na maioria destes, abordam aspectos socioeconômicas do

extrativismo e poucos entram efetivamente nos ambientes de castanhais. Especialmente em solos que abrangem as regiões mais distantes do Estado do Amazonas, pouco se sabe sobre a distribuição da espécie, e os fatores que determinam ou contribuem para a sua ocorrência. Através das características físicas e nutricionais do solo é possível conhecer a estrutura e a disponibilidade de nutrientes nessas áreas de castanhais. Diante desse contexto, é de suma importância conhecer a contribuição das variáveis ambientais, em especial atributos do solo dos castanhais em áreas de ocorrência natural, as quais podem auxiliar nas práticas de manejo, dando apoio à conservação desses ambientes e/ou expansão da produtividade.

4.2 MATERIAL E MÉTODOS

4.2.1 Caracterização da área de estudo

Para a realização deste estudo foram obtidas amostras de solo de castanhal nativo do Estado do Amazonas na Propriedade do Jutica (município de Tefé), conforme Figura 3. Nesta localidade, as coletas das amostras de solo e o inventário florestal foram realizadas em uma parcela permanente de 300 m x 300 m, segundo metodologia proposta por WADT et al. (2017).

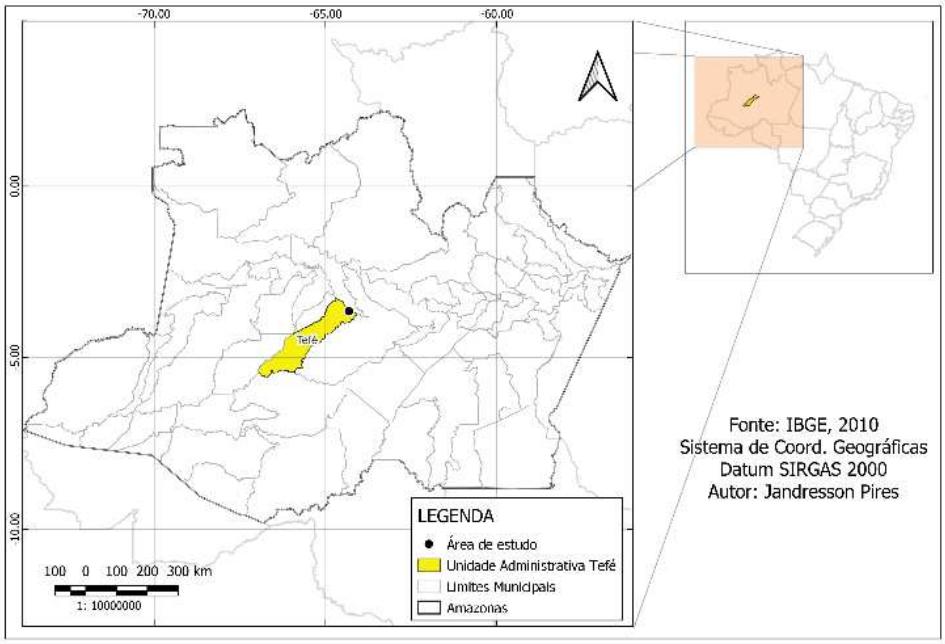


Figura 3 – Mapa de localização da área de estudo: Castanhal – Propriedade do Jutica em Tefé/AM.

O castanhal pertence à Propriedade do Jutica – Tefé/AM, compreendendo as coordenadas 3° 38’ 8,323”S / 64° 18’ 42,648”W. Região de clima que naturalmente apresenta condições de alta temperatura e umidade. As temperaturas médias são altas anualmente e

com umidade relativa do ar sempre elevada, correspondendo os meses de maior umidade (84% a 90%) aos de maior incidência de chuvas (RADAMBRASIL, 1978).

4.2.2 Coleta e análise dos dados

As amostras de solo foram coletadas em 2017, no âmbito do projeto MapCast (Mapeamento de castanhais nativos e caracterização socioambiental e econômica de sistemas de produção da Castanha-da-amazônia na Amazônia) financiados pela Empresa Brasileira de Pesquisa Agropecuária - Embrapa. Em cada área do experimento foram coletadas amostras em uma área de 9 ha (300 x 300 m), utilizando uma amostragem sistemática visando uma maior representação, onde foram adotados um grid de espaçamento regular, totalizando 60 pontos, obedecendo as distâncias de 50 m entre linhas (totalizando seis linhas) e 30 m entre pontos (Figura 4).

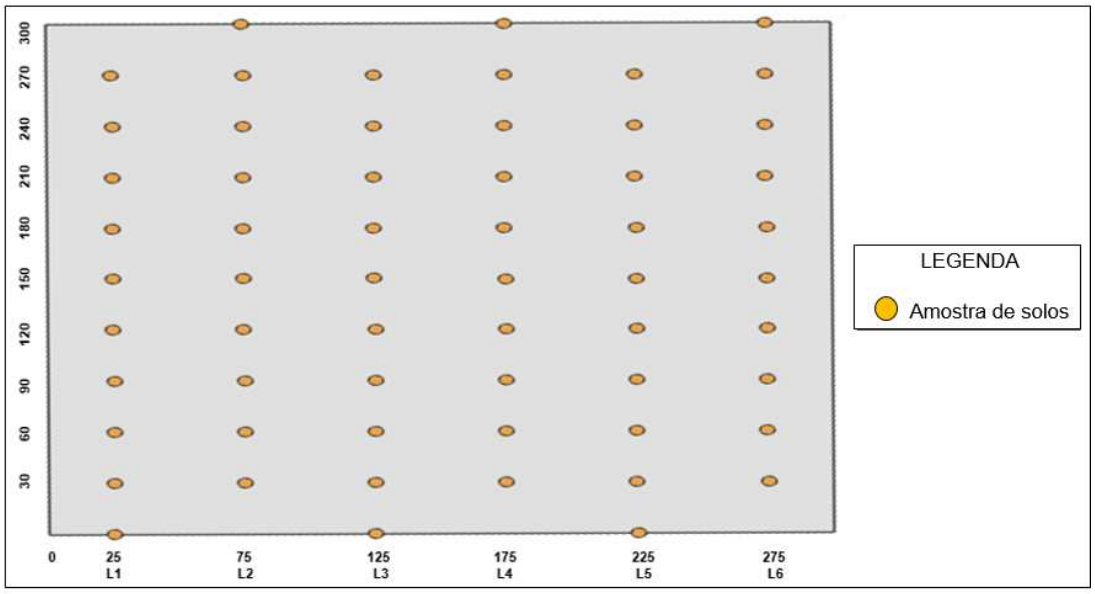


Figura 4 – Amostragem Sistêmica das coletas de solo.

As amostras de solo foram coletadas com trado holandês na profundidade de 0-20 cm. A localização das amostras foram georreferenciadas em coordenadas UTM, Datum WGS84. Após as coletas, as amostras de solos foram armazenadas em sacos plásticos identificados, e posteriormente encaminhadas para o laboratório da Embrapa Amazônia Ocidental-Manaus/AM, onde foram analisadas e identificadas as características físicas e químicas do solo conforme a EMBRAPA (2017). As variáveis de solo utilizadas no trabalho foram: Análises físicas: Areia Total, Silte, Argila, Classificação textural, densidade do solo, VTP%, microporosidade (%), macroporosidade (%). Análises químicas: pH (potencial Hidrogeniônico), MO (matéria orgânica), N (nitrogênio), P (fósforo), K (Potássio), Na (sódio), Ca (cálcio), Mg (magnésio), Al (alumínio), H+Al (Hidrogênio + alumínio = acidez

potencial), Fe (ferro), Zn (zinco), Mn (manganês) e Cu (cobre). Na área em estudo foi realizado o inventário de todas as castanheiras com diâmetro a altura do peito DAP $\geq$ 10 cm, possuindo as dimensões de 9 ha (300 x 300 m).

4.2.3 Geoestatística

A Geoestatística pode ser definido como parte da Estatística que estuda a Teoria das variáveis regionalizadas, ou seja, há a pressuposição de que existe correlação entre os valores observados e a distância de onde os mesmos foram obtidos. Esse conceito é utilizado na Geoestatística para predizer valores não amostrados utilizando principalmente as ferramentas essenciais, como o semivariograma e a interpolação por krigagem (YAMAMOTO e LANDIM, 2013). O semivariograma é utilizado para estudar a similaridade entre amostras vizinhas, de forma que, observações mais próximas, são mais semelhantes entre si do que aquelas que estão separadas por distâncias maiores. Estimando as semivariâncias para cada distância h, é possível descobrir o comportamento espacial da variável em estudo, o que pode ser visualizado em forma de gráfico, denominado por variograma ou semivariograma. O semivariograma pode ser construído pelo semivariograma experimental que é obtido através da amostragem; e pelo semivariograma teórico, que é obtido através do ajuste de modelos teóricos principais, tais como exponencial, esférico, gaussiano. A partir do ajuste do modelo teórico ao semivariograma, pode-se proceder à interpolação geoestatística conhecida como krigagem. Que consiste em ponderar os vizinhos mais próximos do ponto a ser estimado, obedecendo os critérios de não tendenciosidade, que significa que em média a diferença entre valores estimados e observados para o mesmo ponto dever ser nula e ter mínima variância, ou seja, que os estimadores possuam a menor variância dentre todos os estimadores não tendenciosos (OLIVEIRA et al., 2015).

4.2.4 Random Forest

Random Forest (RF) é um algoritmo de aprendizado de máquina que foi desenvolvido por Breiman et al. (1984). O destaque desse algoritmo está na utilização para predição de variáveis em diversos estudos, devido sua capacidade de seleção e classificação de preditores por importância (Suchetana et al., 2017). Baker et al. (2014) diz que o RF consiste em agrupar várias árvores de decisão obtidas a partir de um conjunto de dados de treinamento selecionados aleatoriamente utilizando a metodologia de divisão binária nas variáveis preditoras para classificar ou predizer valores de uma variável resposta. O Random Forest para regressão representa uma combinação de várias árvores aleatórias, independentes e com mesma distribuição, em que o resultado final é a média geral de todos os resultados gerados. Essas árvores são construídas inicialmente por um único nó que se divide em prováveis resultados e, cada resultado se ramifica em outros nós, gerando outras possibilidades. Viana (2019), afirma que o RF é uma técnica de aprendizagem de

máquina forte, que apresenta bom desempenho preditivo quando aplicado às observações não pertencentes aos conjuntos de treinamento, o que reduz o erro de generalização. Ela estima k árvores de regressão, sem poda, por meio de k novos conjuntos de treinamentos distintos, com o mesmo tamanho da amostra original, através de um sorteio aleatório com reposição (reamostragem bootstrap).

4.3 RESULTADOS

Os valores obtidos com os resultados das análises descritivas dos atributos físicos e químicos do solo podem ser encontrados nas tabelas 2 e 3. Os coeficientes de variação (CV) obtidos para a maioria das variáveis físicas das amostras de solo na região de estudo foram considerados de média variabilidade (12%< CV <60%) com exceção da microporosidade (7,21%) e o pH (3,41%) que demonstraram baixa variabilidade. Os parâmetros analisados foram de acordo com a classificação da variabilidade dos atributos definidos por Warrick e Nielson (1980), onde: (CV $<$ 12 %) considerados baixos, (12 % $<$ CV $<$ 60 %) média e (CV $>$ 60 %) alta variabilidade. Dentre as maiores variabilidades encontradas nas variáveis de Jutica - Tefé, foram Cu (45.43%), Ca (42,51%) e Mn (40,25%).

Os valores obtidos para as variáveis físico-químicas foram submetidos à análise geoestatística a fim de verificar a dependência espacial das mesmas. Os resultados das interpolações da análise geoestatística mostraram que os atributos físico-químicos do solo apresentaram dependência espacial, sendo que o mesmo foi observado para os dados do DAP (Tabela 4). Assim foram ajustados e definidos os melhores modelos dos semivariogramas para as variáveis. Nesta análise, para as principais variáveis físicas e químicas, houve predominância do ajuste ao semivariograma do modelo esférico.

Após as análises univariadas dos atributos do solo e do DAP, foi utilizado o processo multivariado de regressão espacial, empregando a Aprendizagem de Máquina por meio do algoritmo Random Forest, para realização das predições espaciais, cujo resultado final consiste em gerar mapas de probabilidade de ocorrência da castanheira e mapas para as variáveis de solo mais importantes. A predição do algoritmo Random Forest demonstrou a influência em determinadas áreas com cores verdes, que representam zonas com maiores probabilidades de ocorrer indivíduos maior ou igual a 50 cm ($\geq$ 50 cm) de DAP, classificados como indivíduos produtivos. Zonas amarelas, medianas probabilidades de ocorrer indivíduos $\geq$ 50 cm; áreas em salmão são zonas com relativa probabilidade e zonas brancas baixa probabilidade (Figura 5). Os Pontos na coloração preta são definidos como DAP maior ou igual a 50 cm (DAP $\geq$ 50cm) e pontos na coloração branca são definidos como DAP menor que 50 cm (DAP $<$ 50cm). Observa-se ainda que não há pontos pretos em zonas brancas, pois em zonas brancas são considerados de baixa probabilidade de DAP $\geq$ 50 cm. Assim como em zonas verdes (alta probabilidade de ocorrer DAP $\geq$

Tabela 2 – Estatística descritiva para as variáveis físicas: Argila, silte, areia total, densidade do solo, volume total de poros (VTP), microporosidade, macroporosidade, na profundidade de 0-20 cm do solo da Propriedade do Jutica – Tefé/AM.

Variável	Média	Mediana	Valor Mín.	Valor Max.
Areia Total (g kg-1)	417,74	424,42	274,92	594,41
Silte (g kg-1)	299,86	289,41	197,85	482,48
Argila (g kg-1)	282,40	284,75	195,00	383,50
Dens. do solo (g cm-3)	0,87	0,87	0,60	1,11
VTP (%)	43,76	42,38	35,20	62,89
Microporosidade (%)	66,57	66,44	57,45	76,85
Macroporosidade (%)	22,81	23,29	6,79	32,53

Coeficientes			Desvio Padrão
Variação(%)	Assimetria	Curtose	
16,77	0,14	-0,58	70,03
21,40	0,96	0,58	64,17
14,48	0,07	-0,46	40,89
14,37	-0,18	-0,59	0,12
13,31	1,12	0,86	5,82
7,21	0,18	-0,59	4,80
22,30	-0,61	0,57	5,09

Tabela 3 – Estatística descritiva para as variáveis químicas: pH, Matéria Orgânica (MO), Nitrogênio (N), Fósforo (P), Potássio (K), Sódio (Na), Magnésio (Mg), Hidrogênio + Alumínio (H+Al), Ferro (Fe), Zinco (Zn), Manganês (Mn), Cobre (Cu), na profundidade de 0-20 cm do solo da Propriedade do Jutica – Tefé/AM.

Variável	Média	Mediana	Valor Mín.	Valor Max.
pH (H2O)	4,07	4,07	3,75	4,45
MO (g kg-1)	35,02	33,63	15,75	66,64
N (g kg-1)	1,51	1,46	0,80	2,48
P (mg dm-3)	3,00	2,87	1,72	4,59
K (mg dm-3)	28,97	27,00	15,00	68,00
Na (mg dm-3)	2,08	2,00	1,00	9,00
Ca (cmolc dm-3)	0,03	0,02	0,01	0,25
Mg (cmolc dm-3)	0,11	0,10	0,07	0,20
H+Al (cmolc dm-3)	10,55	10,65	6,72	14,29
Fe (mg dm-3)	313,92	299,00	180,00	604,00
Zn (mg dm-3)	0,45	0,41	0,24	1,19
Mn (mg dm-3)	0,68	0,63	0,31	1,87
Cu (mg dm-3)	0,56	0,49	0,23	1,28

Coeficientes			Desvio Padrão
variação (%)	Assimetria	Curtose	
3,41	0,40	0,77	0,14
26,38	0,76	1,25	9,24
21,14	0,44	0,63	0,32
19,30	0,36	-0,43	0,58
33,72	1,60	3,35	9,77
32,22	0,90	1,63	0,67
42,51	0,74	0,03	0,01
27,97	0,97	-0,14	0,03
13,95	-0,03	-0,15	1,47
28,26	0,86	0,57	88,72
39,43	1,42	2,94	0,18
40,25	2,49	8,09	0,27
45,42	0,86	0,24	0,25

Tabela 4 – Resultados da Análise Geoestatística para os principais atributos físicos e químicos de solo e do Diâmetro à Altura do Peito (DAP) da Propriedade do Jutica – Tefé/AM.

variável			Patamar	Alcance
Ca (cmolc dm-3)	Exponencial	0	0,00015	23,3784
Na (mg dm-3)	Exponencial	0	0,3841	29,9732
MO (g kg-1)	Esférico	72,7534	83,7375	39,0512
Cu (mg dm-3)	Exponencial	0	0,054	29,7518
Areia Total (g kg-1)	Esférico	1605,27	4704,13	39,0512
N (g kg-1)	Esférico	0,0824	0,1001	39,0512
DAP (cm)	Esférico	0	0,2115	55,027

50 cm) não encontramos pontos brancos. Observou-se que existem alguns pontos pretos (DAP $\geq$ 50cm) em zonas amarelas, pois talvez a castanheira ainda não tenha DAP de 50 cm, porém, esses indivíduos situados nesse ambiente são possíveis chegar a essas medidas (Figura 5). Pode-se considerar que o Random Forest se aplica bem para estimativa da probabilidade de ocorrência de castanheiras, pois a maioria dos indivíduos produtivos (pontos pretos), encontram-se em zonas de média a alta probabilidade de ocorrência, estimados pelo método, o que foi efetivo para a localidade de Jutica - Tefé, que apresentou MSE de 0,0049, na qual (R^2) 77.16% da variável foi explicada (Figura 5).

Em nosso estudo, os mapas das variáveis de solos que mais contribuíram para o modelo foram gerados pela aprendizagem de máquina - RF, onde elas foram analisadas de forma individual, e definidas através de uma medida de importância baseada no índice de impureza de Gini (Figuras 6). Nos mapas das variáveis mais importantes classificada pelo modelo criado pelo Random Forest na região do Jutica - Tefé, apresentam que em áreas verdes são zonas com maiores teores dos atributos no solo e zonas brancas foram definidas

como áreas com pouca influência dos atributos. As variáveis mais importantes baseadas no índice de Gini para essa área, foram: Cálcio, Sódio, Matéria Orgânica, Cobre, Areia Total e Nitrogênio (Figuras 6).

4.4 DISCUSSÃO

Entre as variáveis físicas analisadas a macroporosidade foi o maior CV (22,30%) (Tabela 2), porém, sendo ainda classificada como média. Aquino et al. (2014) em suas análises observaram que, apesar de o CV ser considerado baixo ou alto, ocorre maior predominância de variabilidade moderada dos atributos físicos tanto no ambiente de floresta nativa como no de pastagem na região de Manicoré – AM. Guerreiro et al. (2017) analisando o solo de castanhal nativo na Flona do Tapajós observou que as variáveis físicas foram consideradas de maior variabilidade para microporosidade e as variáveis químicas: potássio e Iodo. E com menor variabilidade para as variáveis químicas, como zinco, cobre e pH. No estudo houve predomínio da macroporosidade, dentre as variáveis com maior variabilidade na área (Tabela 2) e valores mínimos e máximos de pH variaram entre 3,75 a 4,45 (Tabela 3), considerados solos ácidos. Estudando solos de castanhais nativos da Flora do Tapajós – Pará, observou que as variáveis químicas obtiveram um comportamento mais heterogêneo e com o pH apresentando baixa variação (GUERREIRO et al., 2017). Dentre as principais variáveis que apresentou maiores variabilidades foram Cálcio, Cobre e Manganês (Tabela 3). Semelhantes aos resultados encontrados por Guerreiro et al. (2017), em que as variaveis Cálcio e Manganês apresentaram alta variabilidade (CV $>$ 60%). Costa et al. (2017) estudando a relação dos atributos de solos de Roraima com a produção de castanha, observaram que a espécie da Castanheira (Bertholletia excelsa Bonpl.) é exigente em cálcio, afirmando a importância de técnicas de manejo visando a reposição do nutriente. A correlação linear entre os atributos físico-químicos do solo, mostrou que a maioria das variáveis apresentaram baixa correlação, com exceção das variáveis C e MO, C e N e, por fim, N e MO que obtiveram alta significância. Em destaque, as variáveis que apresentaram uma correlação negativa, ou seja, são inversamente proporcionais são densidade do solo com microporosidade. Já as variáveis de solo que apresentaram os maiores destaques na correlação com o DAP neste estudo foram Cu com 0,39 e o Zn com -0,047.

CONCLUSÃO

- A maioria dos atributos estudados apresentam moderado grau de heterogeneidade. Dentre os principais atributos físicos que apresentaram maiores variabilidades nos dados estão o Silte e macroporosidade. E os atributos químicos que apresentaram maior valor foram o Ca, Cu e Mn sendo estas as variáveis consideradas mais heterogêneas nas áreas.

- Nas análises Geoestatísticas, os resultados demonstraram que os semivariogramas modelados e ajustados auxiliaram para reproduzir de forma satisfatória o comportamento e a distribuição espacial dos atributos físico e químicos e a distribuição espacial das castanheiras (DAP $<$ 50 cm e de DAP $\geq$ 50 cm) gerando mapas univariados.

- A predição realizada através do modelo criado pelo Algoritmo RF de predição gerou mapas de probabilidade de ocorrências de castanheiras, contribuindo assim para caracterizar e compreender como estão distribuídas espacialmente as castanheiras nas áreas e em quais regiões ou zonas está mais propicio encontrar castanheiras com DAP $\geq$ 50cm e DAP $<$ 50 cm.

- O modelo indicou os principais atributos do solo na área em estudo, gerando mapas das variáveis mais importantes na área, onde, de acordo com o algoritmo, os atributos mais relevantes nas são a cálcio, sódio e matéria orgânica.

- Os resultados apresentados neste estudo dispõem de informações base que podem ser relacionadas com outras áreas produtivas de castanheira-da-amazônia da região amazônica e sua relação com variáveis ambientais, em especial atributos do solo.

- Sugere-se que os próximos trabalhos possam comparar os resultados obtidos com o modelo predito pelo RF em Tefé com outras regiões, uma vez que o presente trabalho é inédito e pode servir de base de comparação para futuras avaliações em outros locais da Amazônia.

REFERÊNCIAS

AQUINO, R. E.; CAMPOS, M. C. C.; MARQUES JR, J.; OLIVEIRA, I. A.; MANTOVANELI, B. C.; SOARES, M. D. R. Geoestatística na avaliação dos atributos físicos em Latossolo sob floresta nativa e pastagem na região de Manicoré, Amazonas. Revista Brasileira de Ciência do Solo, Viçosa, v. 38, p. 397-406, 2014.

BAKER, P.T., CAUDILL S., HODGE, K. A., TALUKDER, D., CAPANO, C., CORNISH, N.J., Multivariate classification with random forests for gravitational wave searches of black hole binary coalescence. Physical Review D, v.91, n. 6, 2014.

BRASIL. Ministério do Meio Ambiente – MMA. Secretaria de Extrativismo e Desenvolvimento Rural Sustentável. Departamento de Extrativismo. Castanha-do-Brasil: boas práticas para o extrativismo sustentável orgânico. Brasília, 1 ed. 2017, 55 p.

COSTA, M. G. C.; TONINI, H.; MENDES FILHO, P. M. Atributos do Solo Relacionados com a Produção da Castanheira-do-Brasil (Bertholletia excelsa). Revista Floresta e Ambiente. Rio de janeiro, p. 1-10, 2017.

EMBRAPA - EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. Manual de métodos de análise de solo. Brasília, 3 ed. 2017, 574 p.

GUERREIRO, Q. L. de M.; OLIVEIRA JUNIOR, R. C.; SANTOS, G. R.; RUIVO, M. L. P.; BELDINI, T. P.; CARVALHO, E. J. M.; SILVA, K. E.; GUEDES, M. C.; SANTOS, R. B. Spatial variability of soil physical and chemical aspects in a Brazil nut tree stand in the Brazilian Amazon. African Journal of Agricultural Research, v. 12, p. 237¬-250, 2017.

OLIVEIRA, R. P.; GREGO, C. R.; BRANDÃO, Z. N. Geoestatística aplicada na agricultura de precisão utilizando o vesper. Brasília: Embrapa, 1 ed. 2015, 159 p. RADAMBRASIL. Folha SA.20 Manaus: Geologia, Geomorfologia, Pedologia, Vegetação, uso Potencial da Terra. Rio de Janeiro: Departamento Nacional da Produção Mineral, 1978, 747 p.

SERVIÇO FLORESTAL BRASILEIRO - SFB. Manejo da Castanha-do-Brasil (Bertholletia Excelsa) Orientações para as boas práticas de manejo, coleta e pós coleta das castanha-do-brasil, Bioma Brasileiro. Brasília, 2014, 44 p.

SPERA, S. T.; MAGALHÃES, C. A. S.; BALDONI, A. B.; CALDERANO, S. B. Caracterização pedológica de locais de estudo de populações naturais de castanheira-do-brasil no estado de Mato Grosso. Pesquisas Agrárias e Ambientais. Nativa, Sinop, v. 7, n. 2, p. 145-161, 2019.

SUCHETANA, B., RAJAGOPALAN, B., SILVERSTEIN, J. Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a regression tree model. Science of The Total Environment, v. 598, p. 249-257, 2017.

TONINI, H.; BALDONI, A. B. Estrutura e regeneração de Bertholletia excelsa Bonpl. em castanhais nativos da Amazônia. Ciência Florestal de Santa Maria, v. 29, n. 2, p. 607-621, 2019.

WADT, L. H. O.; SANTOS, L.M.H.; MAROCOLO, F.J.; REGO, D. S.G; EMIDIO, K. Panorama geral da produção extrativista de castanha-da-amazônia no Estado de Rondônia. Rondônia: Embrapa. Doc.166, 2019, 39 p.

WADT, L. H. O; SANTOS, L. M. H. BENTES, M. P. M.; OLIVEIRA, V. B. V. Avaliação edáfica e nutricional em espécies arbóreas. Produtos florestais não madeireiros: guia metodológico da Rede Kamukaia. Brasília: EMBRAPA, 1 ed. 2017, 133 p.

WARRICK, A.W.; NIELSEN, D.R. Spatial variability of soil physical properties in the field. In: HILLEL, D., ed. Applications of soil physics. New York, Academic Press, 1980. 350 p.

YAMAMOTO, J. K.; LANDIM, P. M. B. Geoestatística: conceitos e aplicações. São Paulo: Oficina de texto, 1 ed. 2013, 215 p.

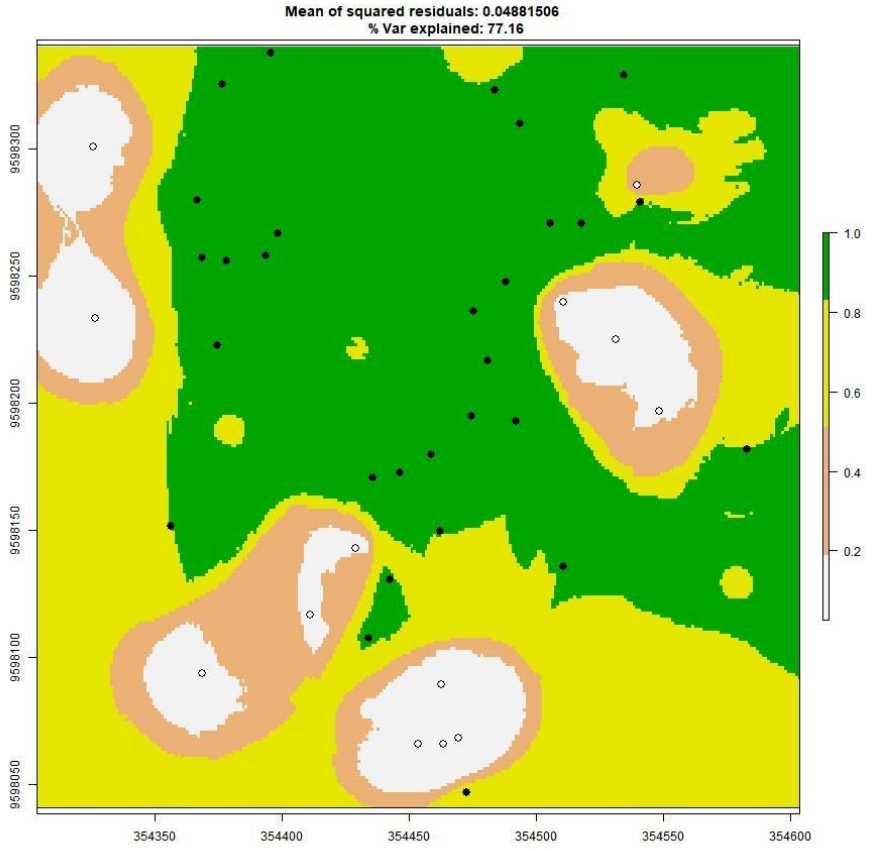


Figura 5 – Mapa de probabilidade de ocorrência de DAP < 50 cm e DAP ≥ 50 cm nos dados de Jutica - Tefé/AM, gerado pela predição realizada pela Aprendizagem de Máquina - Random Forest.

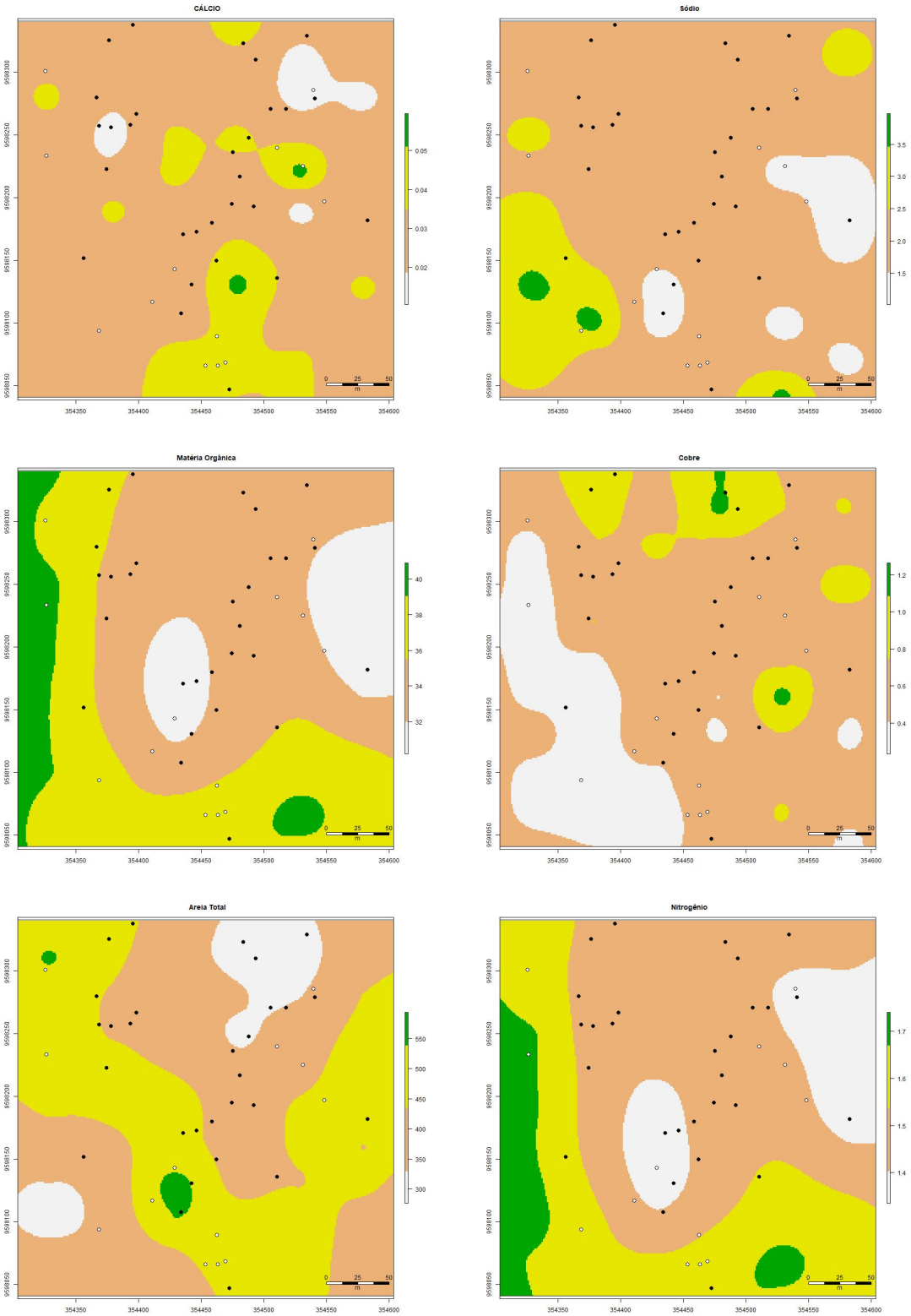


Figura 6 – Mapas das variáveis mais importantes para os dados do Jutica – Tefé/ AM, indicado pela predição realizada pela Aprendizagem de Máquina - Random Forest junto com DAP.

5 CONCLUSÕES GERAIS

Os estudos apresentados nessa tese, demonstram a aplicação da Modelagem Estatística Híbrida Multidimensional em diferentes contextos, visando a representação e compreensão de informações em diversos domínios de aplicação. A modelagem estatística aconteceu de várias maneiras, quer seja na forma tradicional, através de modelos matemáticos, quer seja na forma computacional, através dos algoritmos de aprendizagem de máquina, atingindo, assim, sua forma híbrida. A elaboração de modelos multidimensionais dos atributos físicos ou químicos do solo e a predição das propriedades dos maciços rochosos, são exemplos de como essa abordagem pode ser utilizada de forma eficaz. Vale salientar a elevada interatividade da interface dos resultados, haja vista, que é possível, inclusive, analisar duas, três ou mais variáveis do banco de dados, simultaneamente, permitindo uma análise rápida das relações/associações (ou não), das variáveis analisadas nos gráficos.

No segundo capítulo, foi proposta uma metodologia para a construção de modelos geomecânicos, utilizando dados de furos de sondagem e a linguagem de programação R. Os modelos geomecânicos elaborados, foram confrontados com dados da literatura técnica, evidenciando a sua confiabilidade e adequação para representar as características dos maciços rochosos. A metodologia proposta, demonstrou ser satisfatória em termos de modelagem, visualização, interatividade e na qualidade das informações obtidas.

No terceiro capítulo, o algoritmo *Random Forest* foi utilizado para predizer as propriedades do maciço rochoso. O uso desse algoritmo, juntamente com um modelo simplificado do Índice de Qualidade do Maciço Rochoso, permitiu estimar, de forma confiável, as propriedades geomecânicas dos maciços rochosos. Os resultados obtidos com esse método, demonstraram a viabilidade de sua aplicação na modelagem geomecânica.

O quarto capítulo analisou a relação entre os atributos físicos e químicos do solo, com os castanhais nativos da Amazônia. Por meio da coleta de amostras de solo e inventário das castanheiras, foi possível identificar os principais atributos do solo, que estão correlacionados com a distribuição espacial das plantas. O uso da modelagem híbrida proporcionou a geração de mapas de inteligência geográfica, que caracterizaram a distribuição das castanheiras na área estudada.

Em conclusão, os estudos realizados nessa tese, demonstraram a eficácia da Modelagem Estatística Híbrida Multidimensional na representação e compreensão de informações, em diferentes domínios. A utilização de técnicas estatísticas, geoestatísticas e de aprendizagem de máquina, aliadas a mecanismos de aprendizado de máquina e visualização, possibilitaram a elaboração de modelos confiáveis e a identificação de padrões e relações entre os dados. Essa abordagem mostra-se promissora e pode ser aplicada em outras

áreas e contextos geológico-geomecânicos, contribuindo para uma melhor compreensão e tomada de decisão, em diversas áreas da engenharia e ciências ambientais.