INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Marcelo de Deus Pompeu Magalhães

MAPEAMENTO DOS SOLOS BRASILEIROS EM FUNÇÃO DE PARÂMETROS GEOTÉCNICOS PARA USO EM PAVIMENTOS.

Trabalho de Graduação
2024

Curso de Engenharia Civil-Aeronáutica

CDU 625.8

Marcelo de Deus Pompeu Magalhães

MAPEAMENTO DOS SOLOS BRASILEIROS EM FUNÇÃO DE PARÂMETROS GEOTÉCNICOS PARA USO EM PAVIMENTOS.

Orientadora

Prof[a]. Dr[a]. Cláudia Azevedo Pereira (ITA)

Coorientador

Prof. Dr. José Antonio Schiavon (ITA)

ENGENHARIA CIVIL-AERONÁUTICA

São José dos Campos
Instituto Tecnológico de Aeronáutica

2024

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Magalhães, Marcelo de Deus Pompeu
Mapeamento dos Solos Brasileiros em função de parâmetros Geotécnicos para uso em pavimentos. / Marcelo de Deus Pompeu Magalhães.
São José dos Campos, 2024.
53f.

Trabalho de Graduação – Curso de Engenharia Civil-Aeronáutica– Instituto Tecnológico de Aeronáutica, 2024. Orientadora: Profª. Drª. Cláudia Azevedo Pereira. Coorientador: Prof. Dr. José Antonio Schiavon.

1. Pavimentos. 2. Mapeamento de solos. 3. Aprendizagem (inteligência artificial). 4. Modelo de mistura de gaussianas. 5. Geotecnia. 6. Engenharia civil. 7. Engenharia estrutural. I. Instituto Tecnológico de Aeronáutica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

MAGALHÃES, Marcelo de Deus Pompeu. **Mapeamento dos Solos Brasileiros em função de parâmetros Geotécnicos para uso em pavimentos.**. 2024. 53f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Marcelo de Deus Pompeu Magalhães
TITULO DO TRABALHO: Mapeamento dos Solos Brasileiros em função de parâmetros Geotécnicos para uso em pavimentos..
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2024

Marcelo de Deus Pompeu Magalhães
Rua do H8A, Ap. 113
12.228-460 – São José dos Campos–SP

MAPEAMENTO DOS SOLOS BRASILEIROS EM FUNÇÃO DE PARÂMETROS GEOTÉCNICOS PARA USO EM PAVIMENTOS.

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

Marcelo de Deus Pompeu Magalhães

Autor

Profª. Drª. Cláudia Azevedo Pereira (ITA)

Orientadora

Prof. Dr. José Antonio Schiavon (ITA)

Coorientador

Prof. Dr. Evandro José da Silva
Coordenador do Curso de Engenharia Civil-Aeronáutica

São José dos Campos, 13 de Novembro de 2024.

Dedico este trabalho à minha família, que sempre me apoiou e acreditou neste grande projeto pessoal. Aos amigos que fiz durante essa trajetória, que tornaram este caminho mais leve e tranquilo de atravessar. E aos meus professores, pela paciência e pela sabedoria transmitida que despertaram em mim a paixão pelo conhecimento. A todos vocês, o meu mais profundo agradecimento.

Agradecimentos

Primeiramente, gostaria de agradecer à minha família, principalmente aos meus pais, Thomaz e Heloísa, por terem me ensinado o que é ser uma pessoa íntegra e por me proporcionarem as oportunidades que me trouxeram até aqui. Agradeço também ao meu irmão, Felipe, que foi um grande apoio emocional durante todos esses anos de graduação. Sem eles, com certeza o ITA, até hoje, teria sido apenas um sonho muito distante.

Aos "Tiltados", Flecha, Baby Shark, Fernandão, 14 e Príncipe, e à Vica, minha canga, por terem me ajudado a chegar até aqui e por tornarem essa trajetória o mais agradável possível. Levo essas pessoas como família, e elas foram especiais em diversos momentos nesses longos 5 anos de graduação. Provavelmente, depois de formados, nossos caminhos serão distintos e poderemos até estar distantes, mas os nossos momentos juntos nunca serão esquecidos.

À família T24, uma turma da qual tive o prazer de fazer parte e com a qual criei histórias que ficarão marcadas pelo resto da minha vida. Agradeço em especial à CIVIL24, que percorreu o árduo, cansativo e desgastante percurso e, apesar de todas as dificuldades e brigas ao longo do caminho, não desistiu de chegar até o fim. Isso me deu forças para continuar.

Ao apartamento 113 e a todos os seus moradores (e agregados) pelos bons momentos de descontração e amizade que vivi no ITA. Apesar das desavenças ao longo do caminho, levo com muito carinho cada uma dessas pessoas e os momentos felizes que compartilhamos.

Por fim, gostaria de agradecer a todo o corpo docente que me ensinou e guiou durante a minha trajetória acadêmica no ITA, contribuindo para que eu me tornasse o profissional que sou hoje. Em especial, agradeço à Professora Cláudia e aos Professores Schiavon, Pi e Alex, que, além de ensinarem com maestria suas respectivas matérias, trouxeram muitas lições de vida para os seus alunos. Agradeço também pela amizade e companheirismo desenvolvidos durante a graduação, que tornaram a carga intensa de aulas mais leve e descontraída.

"Artificial intelligence is the new electricity. Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform."

— Andrew Ng

Resumo

Este trabalho de conclusão de curso aborda o ”Mapeamento dos Solos Brasileiros em Função de Parâmetros Geotécnicos para Uso em Pavimentos”. O estudo propõe uma metodologia que combina técnicas de clusterização, como o Modelo de Mistura Gaussiana (GMM) e o K-means, para classificar solos brasileiros com base em parâmetros geotécnicos. Utilizando dados de sondagens, como CPT, limites de Atterberg e SPT, o trabalho visa identificar padrões nos solos, facilitando a análise e o planejamento de pavimentações e construções. A metodologia inclui a normalização dos dados, segmentação tridimensional, e análise de clusters, proporcionando uma visão detalhada das propriedades geotécnicas. Os resultados destacam a eficiência da metodologia proposta e sugerem sua aplicação prática em diferentes cenários da engenharia civil.

Abstract

This undergraduate thesis focuses on the "Mapping of Brazilian Soils Based on Geotechnical Parameters for Pavement Use". The study proposes a methodology that combines clustering techniques, such as Gaussian Mixture Model (GMM) and K-means, to classify Brazilian soils based on geotechnical parameters. Using data from tests such as CPT, Atterberg limits, and SPT, the work aims to identify soil patterns, facilitating analysis and planning for pavements and constructions. The methodology includes data normalization, three-dimensional segmentation, and cluster analysis, providing a detailed view of geotechnical properties. The results highlight the efficiency of the proposed methodology and suggest its practical application in various civil engineering scenarios.

Lista de Figuras

Lista de Abreviaturas e Siglas

AIC	Critério de Informação de Akaike
BIC	Critério de Informação Bayesiano
CPT	Ensaio de Penetração de Cone
EM	Expectation Maximization
GMM	Modelo de Mistura Gaussiana
SIG	Sistema de Informação Geográfica
SPT	Ensaio de Sondagem à Percussão

Lista de Símbolos

μ	Média de uma distribuição
σ	Desvio Padrão de uma distribuição

Sumário

1 Introdução

Neste capítulo traz-se ao leitor uma breve contextualização sobre o assunto, bem como as motivações que levaram este Trabalho de Graduação a ser realizado. Além disso, levanta-se os objetivos deste trabalho, bem como sua organização com o intuito de dar um panorama geral do documento.

1.1 Contextualização

Os mapas têm desempenhado um papel fundamental na evolução da civilização humana, estando presentes em nosso cotidiano desde antes mesmo da invenção da escrita, como destaca Paulo Miceli (MICELI, 2013). Ao longo dos séculos, eles se consolidaram como ferramentas indispensáveis para o desenvolvimento e compreensão do mundo ao nosso redor, permitindo que sociedades interpretassem e navegassem pelos espaços geográficos de maneira eficaz.

Com uma vasta gama de possibilidades de utilização, os mapas estão integrados nos mais diversos contextos e setores. Eles são essenciais na demarcação de terras, auxiliando na gestão territorial e na resolução de disputas fundiárias. No campo da logística e do transporte, os mapas são fundamentais para o cálculo de rotas entre dois locais, otimizando trajetos e economizando recursos. Além disso, no mundo dos negócios, eles permitem a análise de potenciais clientes para uma empresa, possibilitando estratégias de mercado mais assertivas.

No setor ambiental, os mapas desempenham um papel crucial na conservação e gestão de recursos naturais. Eles facilitam o monitoramento de ecossistemas, a identificação de áreas de risco e o planejamento de ações para a preservação do meio ambiente. Por meio de mapas temáticos e sistemas de informação geográfica (SIG), é possível analisar dados climáticos, hídricos e geológicos, contribuindo para práticas sustentáveis e para a mitigação dos impactos das mudanças climáticas.

Com o advento das tecnologias modernas, os mapas evoluíram de simples representações estáticas para ferramentas dinâmicas e interativas, integradas a sistemas de in-

formação geográfica e outras plataformas digitais. Essa evolução tecnológica ampliou significativamente as aplicações dos mapas, especialmente na engenharia civil. A incorporação de métodos avançados, como a clusterização de múltiplas variáveis, permite uma análise mais profunda e detalhada de dados geoespaciais e físicos.

Na engenharia civil aeronáutica, por exemplo, a metodologia de multiclusterização possibilita o desenvolvimento de projetos mais eficientes e seguros. A utilização de técnicas de clusterização facilita a identificação de padrões e tendências que não seriam facilmente perceptíveis de outra forma. Isso contribui para um planejamento mais assertivo, otimização de recursos e mitigação de riscos em projetos de grande escala.

Portanto, a combinação de mapas com métodos avançados de análise de dados se torna uma ferramenta poderosa para enfrentar os desafios contemporâneos na engenharia civil e em outras áreas correlatas.

1.2 Motivação

A principal motivação para este estudo está enraizada na necessidade de tornar o processo de análise de solos mais eficiente e acessível. Atualmente, essa tarefa exige uma abordagem manual que consome tempo e recursos significativos, tornando-se desgastante para os profissionais e suscetível a falhas e inconsistências humanas. Em um cenário onde a demanda por análises geotécnicas precisas cresce, especialmente em áreas remotas e projetos de infraestrutura, o desenvolvimento de uma metodologia automatizada para o mapeamento de solos surge como uma solução promissora. Essa automação não apenas reduziria o tempo de análise, mas também ampliaria a precisão e consistência dos resultados, beneficiando diretamente os projetos de engenharia que dependem dessas informações.

Além disso, a incorporação de técnicas de machine learning no campo da engenharia civil alinha-se com as tendências contemporâneas de inovação tecnológica e otimização de processos. O uso de algoritmos de clusterização para identificar padrões complexos em dados geotécnicos permite a identificação de características do solo que, de outra forma, poderiam passar despercebidas em análises tradicionais. Assim, este trabalho não só visa automatizar um processo considerado monótono e repetitivo, mas também elevar a qualidade das análises, proporcionando aos engenheiros ferramentas robustas para a tomada de decisões informadas e assertivas em projetos de pavimentação e construção em grande escala.

1.3 Objetivo

O objetivo deste Trabalho de Graduação é propor uma metodologia e desenvolver uma ferramenta capaz de utilizar métodos de clusterização existentes para identificar padrões de solos em um determinado terreno, facilitando, assim, análises de viabilidade de solo de forma mais ágil e precisa.

Além disso, pretende-se incentivar futuros estudos que integrem estas áreas, que possuem um enorme potencial de crescimento e inovação, oferecendo soluções mais eficientes para a engenharia civil e outras disciplinas correlatas.

Para o desenvolvimento desta metodologia, será necessário cumprir os seguintes objetivos específicos:

- Analisar dados de solos utilizando métodos de clusterização existentes, com o intuito de identificar padrões geotécnicos que possam auxiliar na tomada de decisão para projetos de pavimentação e construção.
- Validar os padrões identificados por meio de comparações com dados reais, garantindo a precisão e confiabilidade dos modelos de clusterização aplicados.
- Propor uma ferramenta integrada, que permita aos engenheiros realizar análises de viabilidade de solo de maneira prática e rápida, utilizando os dados de clusterização.
- Facilitar o uso da ferramenta para outras áreas da engenharia e geociências, possibilitando a adaptação do método para diferentes cenários e necessidades de estudos de solo.
- Comprovar a economia de tempo e recursos ao utilizar a ferramenta proposta, demonstrando como a automação de análises de solo pode beneficiar a viabilidade econômica de projetos de infraestrutura.

1.4 Organização do trabalho

Os capítulos deste estudo foram organizados da seguinte maneira:

- Capitulo 1 : Introdução - Contextualiza e traz as motivações que levaram à realização desta Tese de Graduação;
- Capitulo 2: Revisão Bibliográfica - Apresenta uma breve discussão sobre os conhecimentos existentes nos assuntos abordados que serão de grande relevância para o desenvolvimento deste Trabalho de Graduação;

- Capitulo 3: Metodologia - Explica como o conhecimento apresentado no Capítulo anterior foi desenvolvido para chegar no objetivo proposto no Capítulo 1;
- Capitulo 4: Resultados e Discussões - Expõe e interpreta os resultados obtidos; e
- Capitulo 5: Considerações finais - Faz uma conclusão geral do trabalho, sugere futuras implementações para futuros trabalhos que possam complementar a tese.

2 Revisão Bibliográfica

Neste capítulo, apresenta-se uma síntese dos estudos e pesquisas relevantes relacionados ao tema desta tese de graduação. Abordamos os principais conceitos, teorias e avanços recentes na área, proporcionando uma base teórica sólida para o desenvolvimento do trabalho.

2.1 A Importância dos Dados na Engenharia Civil

A coleta, análise e interpretação de dados são fundamentais em todas as fases dos projetos de engenharia civil, desde o planejamento inicial até a manutenção das estruturas concluídas. No planejamento, dados geológicos, topográficos e climáticos são essenciais para a seleção adequada do local e para o dimensionamento correto das obras (EASTMAN *et al.*, 2018). Por exemplo, a análise de dados climáticos históricos influencia o design estrutural para resistir a eventos extremos (LI *et al.*, 2019). Durante a fase de construção, o monitoramento em tempo real dos dados permite ajustes imediatos, garantindo a qualidade e a segurança do projeto (BILAL *et al.*, 2016). Na manutenção, a análise contínua de dados estruturais auxilia na identificação precoce de possíveis falhas, prolongando a vida útil das edificações (NI; WONG, 2015).

Entretanto, a gestão de grandes volumes de dados apresenta desafios significativos. A heterogeneidade dos dados, provenientes de diversas fontes e em diferentes formatos, dificulta a integração e a análise eficiente (BILAL *et al.*, 2016). Problemas relacionados à qualidade dos dados, como inconsistências, lacunas e informações desatualizadas, podem comprometer a confiabilidade das análises e das decisões tomadas (LOVE *et al.*, 2018). Esses desafios impactam diretamente a eficiência dos projetos, podendo resultar em atrasos, aumento de custos e riscos à segurança das estruturas (SUN *et al.*, 2017).

Com o avanço tecnológico, ferramentas de Big Data e técnicas avançadas de análise de dados têm contribuído para superar esses obstáculos. Algoritmos de aprendizado de máquina e inteligência artificial permitem processar grandes volumes de dados complexos, identificando padrões que auxiliam na tomada de decisão (LI *et al.*, 2019). Além disso, algoritmos de clusterização emergiram como ferramentas diferenciais na análise de grandes

conjuntos de dados. Esses algoritmos permitem agrupar dados semelhantes, facilitando a identificação de padrões ocultos e tendências (XU; TIAN, 2015). Por exemplo, o uso de sensores IoT em pontes e edifícios permite a coleta contínua de dados sobre vibração e deformação (GUO *et al.*, 2018). Esses dados, quando analisados com técnicas de Big Data, podem prever falhas estruturais e orientar ações de manutenção preventiva (NI; WONG, 2015).

2.1.1 Análise de Dados Geotécnicos no Mapeamento e Previsão de Riscos

Dentro desse contexto, a análise de dados geotécnicos destaca-se no mapeamento e previsão de riscos, sendo crucial para a segurança das construções. A compreensão das propriedades do solo e das condições geológicas permite decisões informadas durante todas as fases do projeto (FENG; ZHOU, 2016). A heterogeneidade e complexidade dos solos tornam a caracterização do terreno uma tarefa desafiadora, exigindo abordagens robustas de coleta e interpretação de dados (ZHANG *et al.*, 2015).

Técnicas como Sistemas de Informação Geográfica (SIG), aprendizado de máquina e algoritmos de clusterização têm sido aplicadas para superar esses desafios. Park e Kim (PARK; KIM, 2019) demonstram que o uso de SIG facilita o mapeamento de áreas suscetíveis a riscos geotécnicos. Além disso, Lee e Kim (LEE; KIM, 2018) aplicaram redes neurais artificiais para prever a suscetibilidade a deslizamentos de terra, obtendo maior precisão preditiva em comparação com métodos tradicionais. A clusterização também tem sido utilizada para identificar zonas com características geotécnicas semelhantes, auxiliando na criação de mapas detalhados que identificam áreas propensas a instabilidades (XU *et al.*, 2019).

2.2 Aplicações de Técnicas de Clusterização na Engenharia Civil

A clusterização, também conhecida como agrupamento, é uma técnica de aprendizado de máquina não supervisionado que tem como objetivo organizar um conjunto de dados em grupos (clusters) com características semelhantes (AGGARWAL; REDDY, 2014). Essa metodologia é essencial na análise de dados complexos, pois permite a identificação de padrões e estruturas ocultas sem a necessidade de rótulos pré-definidos (XU; TIAN, 2015). Ao segmentar grandes volumes de informações em clusters significativos, a clusterização facilita a interpretação dos dados e a tomada de decisões informadas em diversos campos da engenharia.

2.2.1 Clusterização Espacial na Engenharia Civil

A clusterização espacial é uma extensão das técnicas tradicionais que incorpora a dimensão geográfica aos processos de agrupamento (SHI *et al.*, 2015). Essa abordagem é particularmente relevante na engenharia civil, onde a localização espacial dos fenômenos influencia diretamente o planejamento e a execução de projetos. A clusterização espacial permite identificar padrões geográficos, como a concentração de patologias em edificações, áreas com alto índice de acidentes ou regiões com demanda crescente por serviços públicos (MILLER; GOODCHILD, 2015).

Apesar de seu potencial, a clusterização espacial ainda é pouco explorada na engenharia civil. Isso pode ser atribuído à complexidade dos dados espaciais, que exigem métodos computacionalmente mais robustos e a integração de sistemas de informação geográfica (SIG) (SHI *et al.*, 2015). Além disso, a necessidade de profissionais com conhecimentos multidisciplinares pode representar uma barreira para a ampla adoção dessas técnicas.

2.2.2 Relevância na Análise de Dados Complexos

Com o advento do big data, a quantidade de informações disponíveis aumentou exponencialmente, tornando indispensável o uso de técnicas avançadas para sua análise (CHEN *et al.*, 2014). A clusterização oferece ferramentas para lidar com essa complexidade, possibilitando a descoberta de insights valiosos a partir de dados não estruturados ou parcialmente estruturados (MINELLI *et al.*, 2014). Na engenharia civil, onde os projetos frequentemente envolvem múltiplas variáveis e condições ambientais, a clusterização contribui para uma compreensão mais aprofundada dos fenômenos estudados.

2.2.3 Aplicações em Sensoriamento Remoto e Classificação de Solos

No campo do sensoriamento remoto, a clusterização é amplamente utilizada para a interpretação de imagens de satélite e aéreas (ZHANG *et al.*, 2016). Por meio do agrupamento de pixels com propriedades espectrais semelhantes, é possível identificar diferentes tipos de cobertura e uso do solo. Essa abordagem auxilia no monitoramento ambiental, no planejamento urbano e na gestão de recursos naturais. Por exemplo, técnicas como o algoritmo *k-means* permitem classificar áreas urbanas, agrícolas e florestais, contribuindo para estudos de ocupação do solo e mudanças ambientais (MA *et al.*, 2019).

Na classificação de solos, a clusterização facilita o agrupamento de amostras com características físico-químicas similares (PADARIAN *et al.*, 2019). Esse processo é fundamental para a elaboração de mapas de solos, que são essenciais em projetos de engenharia geotécnica, agrícolas e ambientais, como os desenvolvidos por Robertson (ROBERTSON, 1990),

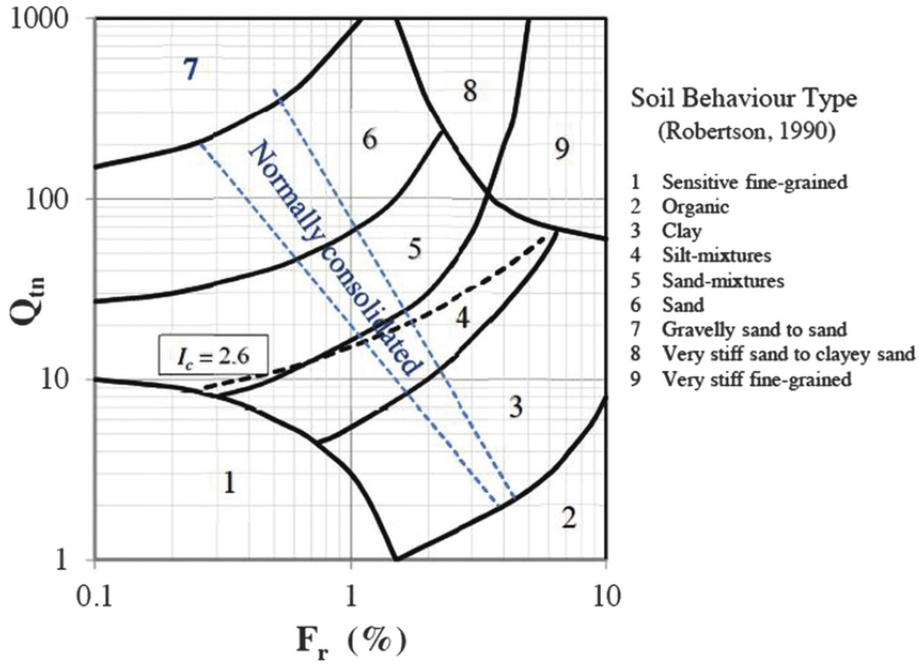


FIGURA 2.1 – Distribuição e classificação do solo por parâmetros geofísicos da matéria feito por Robertson (ROBERTSON, 1990)

ilustrados na Figura 2.1. Ao identificar clusters de solos com propriedades homogêneas, os engenheiros podem prever comportamentos mecânicos e hidráulicos, otimizando o design e a execução de obras (HENGL *et al.*, 2018; ROBERTSON, 1990), e aproveitar o solo da forma mais otimizada possível, utilizando-os como materiais de empréstimo.

2.3 Normalização de Bases de Dados na Análise de Engenharia

A normalização de bases de dados é um passo crítico no pré-processamento de informações, especialmente na engenharia civil, onde múltiplas variáveis são frequentemente coletadas em diferentes escalas ou unidades de medida. Este processo ajusta os dados para que variáveis distintas possam ser comparadas diretamente, eliminando disparidades causadas por amplitudes numéricas variadas. Sem a normalização, variáveis com magnitudes maiores podem dominar análises estatísticas ou modelos de aprendizado de máquina, levando a resultados enviesados ou interpretações errôneas (IOFFE; SZEGEDY, 2015).

A normalização evita vieses nos modelos de análise ao garantir que cada variável contribua proporcionalmente para o resultado final. Em modelos estatísticos e algoritmos de aprendizado de máquina, como regressão linear ou redes neurais, a escala das variáveis influencia diretamente os coeficientes calculados. Variáveis não normalizadas podem levar a coeficientes desbalanceados, prejudicando a capacidade do modelo de aprender

padrões significativos (SINGH; SINGH, 2020). Portanto, a normalização é essencial para obter modelos mais robustos e generalizáveis.

2.3.1 Normalização Min-Max

A técnica de normalização Min-Max transforma os dados para um intervalo pré-definido, geralmente entre 0 e 1, aplicando a fórmula:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{2.1}$$

Em que:

- X é o valor original.
- X_{min} é o valor mínimo da variável.
- X_{max} é o valor máximo da variável.

As vantagens desse método incluem a preservação das relações entre os valores originais e sua simplicidade de implementação (GARCíA *et al.*, 2015). Contudo, sua principal desvantagem surge na presença de outliers, que podem distorcer os dados normalizados. É mais indicada quando não se conhece a distribuição dos dados trabalhados.

2.3.2 Normalização Z-score

A normalização por Z-score, ou padronização, transforma os dados para que tenham média zero e desvio padrão igual a um. A fórmula utilizada é:

$$Z = \frac{X - \mu}{\sigma} \tag{2.2}$$

Em que:

- X é o valor original.
- μ é a média da variável.
- σ é o desvio padrão.

As vantagens desse método são a adequação para dados com distribuição aproximadamente normal e a redução do impacto de outliers (PATRO; SAHU, 2015). Entretanto,

sua desvantagem é a possibilidade de não preservar os limites originais dos dados, o que pode dificultar a interpretação. É mais indicada quando se assume distribuições normais dos dados.

2.4 Técnicas de Clusterização

A clusterização é uma técnica fundamental em análise de dados que visa agrupar objetos com características semelhantes em clusters ou grupos. Esse processo é essencial em diversas áreas do conhecimento, incluindo a engenharia civil, onde a identificação de padrões em grandes conjuntos de dados auxilia no planejamento urbano, gestão de recursos e análise de materiais. Dentre os métodos de clusterização, destacam-se os Modelos de Mistura Gaussiana (GMM) e os algoritmos baseados em centróides, como o K-means. Este trabalho aborda esses dois métodos, descrevendo detalhadamente seu funcionamento matemático, principais características e aplicações práticas (HASTIE *et al.*, 2009).

2.4.1 Modelo de Mistura Gaussiana (GMM)

Os Modelos de Mistura Gaussiana (GMM) são uma abordagem probabilística para a clusterização que assume que os dados são gerados a partir de uma combinação de distribuições gaussianas (BISHOP, 2006). Matematicamente, a densidade de probabilidade de um GMM é expressa como:

$$p(x) = \sum_{k=1}^{K} \pi_k \, N(x|\mu_k, C_k) \tag{2.3}$$

Em que:

- x é um vetor de observações;
- K é o número de componentes (clusters);
- π_k é o peso ou proporção do k-ésimo componente, com $\sum_{k=1}^{K} \pi_k = 1$ e $0 \leq \pi_k \leq 1$;
- $N(x|\mu_k, C_k)$ é a função densidade de probabilidade de uma distribuição gaussiana multivariada com média μ_k e matriz de covariância C_k.

O objetivo do GMM é estimar os parâmetros $\{\pi_k, \mu_k, C_k\}$ que melhor se ajustam aos dados observados. Para isso, utiliza-se o algoritmo de *Expectation-Maximization* (EM) (BISHOP, 2006), que consiste em duas etapas iterativas:

Etapa E (Expectation): Calcula a probabilidade posterior de que cada observação x_i pertença ao componente k, dada pelos pesos de responsabilidade (BERNSTEIN, 2020):

$$\gamma_{ik} = \frac{\pi_k \, N(x_i|\mu_k, C_k)}{\sum_{j=1}^{K} \pi_j \, N(x_i|\mu_j, C_j)} \tag{2.4}$$

Etapa M (Maximization): Atualiza os parâmetros dos componentes com base nos pesos calculados:

$$\pi_k^{\text{novo}} = \frac{N_k}{N} \tag{2.5}$$

$$\mu_k^{\text{novo}} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_i \tag{2.6}$$

$$C_k^{\text{novo}} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k^{\text{novo}})(x_i - \mu_k^{\text{novo}})^\top \tag{2.7}$$

Em que:

- N é o número total de observações;
- $N_k = \sum_{i=1}^{N} \gamma_{ik}$ é o número de observações atribuídas ao componente k.

O processo é repetido até a convergência dos parâmetros.

A Figura 2.2 ilustra as iterações do GMM até a convergência dos parâmetros para um número definido de clusters. Observa-se que o GMM ajusta distribuições gaussianas aos dados, permitindo a identificação de clusters com diferentes formas e orientações.

2.4.2 Clusterização baseada em centróides

O K-means é um método de clusterização por centróides que busca particionar os dados em K clusters, minimizando a variância dentro de cada grupo (HASTIE *et al.*, 2009). O algoritmo funciona através dos seguintes passos:

- **Inicialização**: Seleciona aleatoriamente K centróides iniciais $\{\mu_1, \mu_2, ..., \mu_K\}$.
- **Atribuição de Clusters**: Para cada observação x_i, atribui ao cluster cujo centróide está mais próximo, de acordo com a distância euclidiana:

$$c_i = \arg\min_k \|x_i - \mu_k\|^2 \tag{2.8}$$

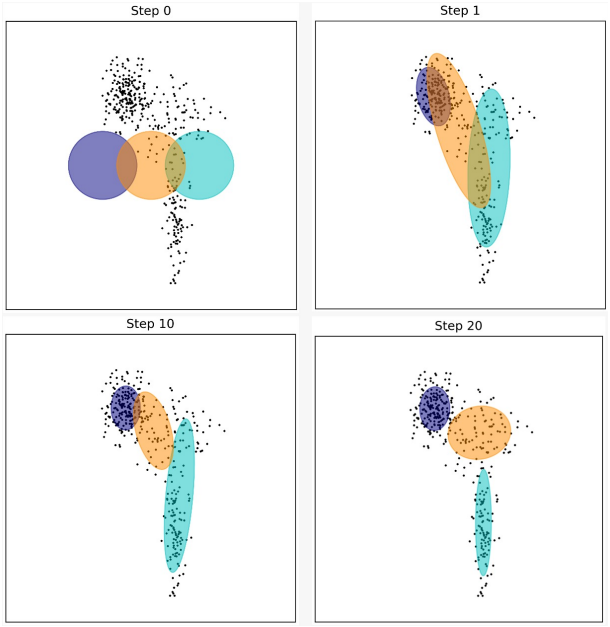


FIGURA 2.2 – Iterações do GMM até a convergência dos parâmetros para um número definido de clusters. Fonte: (BERNSTEIN, 2020)

- **Atualização dos Centróides**: Recalcula os centróides como a média das observações atribuídas a cada cluster:

$$\mu_k = \frac{1}{N_k} \sum_{i:c_i=k} x_i \tag{2.9}$$

- **Convergência**: Repete os passos 2 e 3 até que as atribuições não mudem ou até atingir um número máximo de iterações.

O objetivo é minimizar a soma total das variâncias intra-cluster, definida como:

$$J = \sum_{k=1}^{K} \sum_{i:c_i=k} \|x_i - \mu_k\|^2 \tag{2.10}$$

A Figura 2.3 demonstra as iterações do K-means até a convergência dos parâmetros para um número definido de clusters. Nota-se que o K-means agrupa os dados com base na proximidade aos centróides.

2.5 Métodos de Otimização de Clusterização

Na análise de dados em engenharia civil, a clusterização é fundamental para agrupar dados com características semelhantes, facilitando a identificação de padrões e tendências relevantes. No entanto, determinar o número ideal de clusters e avaliar a qualidade da clusterização são desafios que exigem métodos de otimização eficazes. A otimização

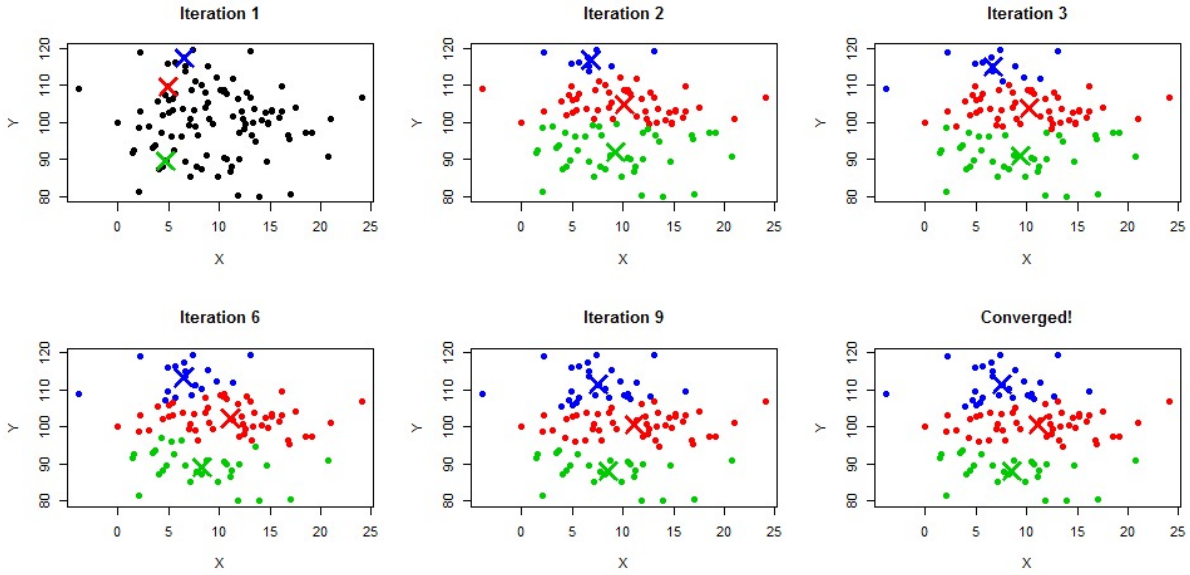


FIGURA 2.3 – Iterações do K-means até a convergência dos parâmetros para um número definido de clusters. Fonte: (JONES, 2024).

aprimora a precisão dos modelos de clusterização, evitando problemas como overfitting e garantindo que os clusters gerados sejam significativos para a interpretação dos dados (AGGARWAL; REDDY, 2014).

2.5.1 Otimização do GMM

Para a otimização da clusterização GMM, dois métodos principais são comumente usados: Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC).

2.5.1.1 AIC

O Critério de Informação de Akaike (AIC) é um método estatístico utilizado para a seleção de modelos, focado em estimar a qualidade relativa dos modelos estatísticos para um conjunto de dados. No contexto dos GMM, o AIC é empregado para determinar o número ideal de componentes (clusters) no modelo (HASTIE *et al.*, 2009).

A fórmula do AIC é:

$$\text{AIC} = 2k - 2\ln(L) \tag{2.11}$$

Onde:

- k é o número de parâmetros estimados no modelo;

- $\ln(L)$ é o logaritmo da função de verossimilhança máxima do modelo.

O modelo final escolhido é aquele com o menor AIC, correspondendo ao melhor balanceamento entre qualidade de ajuste e complexidade. Contudo, o AIC tende a preferir modelos mais complexos.

2.5.1.2 BIC

O Critério de Informação Bayesiano (BIC) é outro método estatístico utilizado para selecionar modelos que melhor se ajustam aos dados, penalizando a complexidade para evitar overfitting. No contexto dos GMM, o BIC é empregado para determinar o número ideal de componentes (clusters) no modelo (HASTIE *et al.*, 2009).

Matematicamente, o BIC é definido como:

$$\text{BIC} = k \ln(n) - 2 \ln(L) \tag{2.12}$$

Em que:

- k é o número de parâmetros estimados no modelo;
- $\ln(L)$ é o logaritmo da função de verossimilhança máxima do modelo;
- n é o tamanho da amostra (número total de observações).

O modelo final escolhido é aquele com o menor BIC, balanceando qualidade de ajuste com simplicidade. Comparado ao AIC, o BIC penaliza mais severamente modelos complexos, especialmente quando a amostra é grande, sendo útil para evitar overfitting (BURNHAM; ANDERSON, 2014).

2.5.2 Otimização para Clusterização por Centróides

Para a otimização da clusterização por centróides, dois métodos principais são utilizados: Coeficiente de Silhueta e o Método do Cotovelo.

2.5.2.1 Coeficiente de Silhueta

O coeficiente de silhueta é uma métrica que avalia a qualidade de uma clusterização, medindo quão semelhante um objeto é ao seu próprio cluster em comparação com outros clusters. Esta métrica varia de -1 a 1, onde valores próximos de 1 indicam uma boa atribuição (ROUSSEEUW, 1987).

Para cada objeto i, o coeficiente de silhueta é calculado por:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.13}$$

Em que:

- $a(i)$ é a distância média entre o objeto i e todos os outros objetos do mesmo cluster;
- $b(i)$ é a menor distância média entre o objeto i e todos os objetos de qualquer outro cluster do qual i não faz parte.

O número ótimo de clusters k é aquele que maximiza o valor médio de $s(i)$.

2.5.2.2 Método do Cotovelo

O método do cotovelo é uma técnica gráfica utilizada para determinar o número ótimo de clusters, analisando a variância interna em função do número de clusters (JR.; SHOOK, 1996).

Para implementá-lo, varia-se o número de clusters k e calcula-se a inércia do sistema, dada pela soma das distâncias quadradas dentro dos clusters:

$$\text{Inércia}(k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2.14}$$

Em que:

- C_i é o i-ésimo cluster;
- μ_i é o centróide do i-ésimo cluster;
- x são os pontos de dados no cluster C_i.

Ao plotar o gráfico de inércia versus k, identifica-se o ponto em que a redução da inércia se torna menos significativa com o aumento de clusters (o "cotovelo"), como ilustrado na Figura 2.4.

2.6 Utilização de Propriedades Físicas Médias do Solo como Intervalos na Engenharia Civil

A caracterização precisa das propriedades físicas do solo é fundamental para o sucesso de projetos de engenharia civil. Devido à natureza heterogênea dos solos, suas pro-

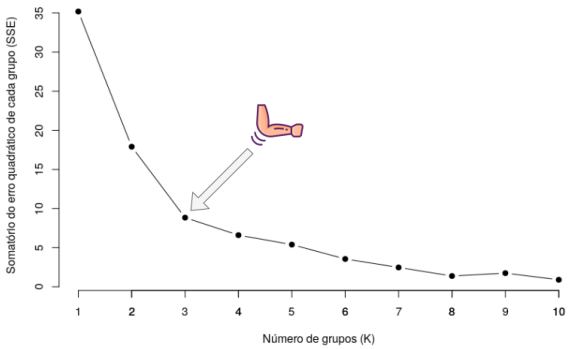


FIGURA 2.4 – Exemplo do método do cotovelo para determinação do número ideal de clusters. Fonte: (DATAAT, 2024).

priedades físicas, como resistência ao cisalhamento, permeabilidade e compressibilidade, apresentam variabilidade natural. A utilização da média de uma distribuição dessas propriedades permite representar um intervalo específico, facilitando a análise e classificação dos solos. Essa abordagem estatística simplifica a interpretação dos dados e fornece uma base sólida para o dimensionamento de estruturas geotécnicas.

Ao considerar a média como representativa de um intervalo, os engenheiros podem associar propriedades médias a intervalos de confiança, o que é essencial para o planejamento e a segurança dos projetos. Por exemplo, ao determinar a capacidade de carga de uma fundação, é crucial conhecer não apenas o valor médio da resistência do solo, mas também a variação em torno desse valor. A aplicação de intervalos de confiança permite incorporar a incerteza inerente aos parâmetros do solo, resultando em projetos mais seguros e economicamente viáveis.

Os métodos estatísticos, como a análise de variância (ANOVA) e o cálculo de intervalos de confiança, são ferramentas essenciais para calcular médias e intervalos de propriedades do solo. Esses métodos permitem quantificar a variabilidade dos dados e estimar a probabilidade de ocorrência de determinados valores. A norma NBR 6484 (ABNT, 2001) da Associação Brasileira de Normas Técnicas, por exemplo, orienta sobre a execução de sondagens de simples reconhecimento dos solos, enfatizando a importância da análise estatística dos resultados obtidos.

Aplicações práticas dessa abordagem incluem o dimensionamento de fundações superficiais e profundas, onde a variabilidade das propriedades do solo influencia diretamente a segurança e o desempenho das estruturas. Estudos acadêmicos, como o apresentado por (DUNCAN, 2000) e (AOKI; VELLOSO, 1975), demonstram que a consideração de propriedades médias com intervalos de confiança reduz significativamente as incertezas nos projetos geotécnicos. Isso se traduz em estruturas mais seguras e na otimização de recursos, beneficiando tanto os profissionais da área quanto a sociedade em geral.

3 Metodologia

Este capítulo apresenta, de maneira sequencial e estruturada, as etapas da metodologia proposta para o mapeamento de solos brasileiros, utilizando técnicas de clusterização. A Figura 3.1 exibe o fluxograma das etapas envolvidas no desenvolvimento da metodologia e do software implementado para este estudo.

3.1 Aquisição e Leitura dos Dados

Para validar o método proposto, foram utilizadas bases de dados fornecidas por um professor do instituto e ensaios específicos realizados por uma empresa que possuia dados de sondagem. Estas bases incluem diferentes tipos de ensaios geotécnicos, como:

- Ensaios de Penetração de Cone (CPTUs e CPTs), como ilustrado na Figura 3.2;
- Ensaios de limites de Atterberg, como ilustrado na Figura 3.3;
- Ensaio de Sondagem à Percussão (SPTs), como presente na Figura 3.4.

A leitura dos dados foi realizada com a biblioteca *pandas* no Python através da função `read_database`, escolhida pela sua capacidade de lidar com múltiplos formatos de arquivos, como *.xlsx* e *.csv*, principais formatos encontrados neste estudo. Esses arquivos foram importados integralmente como *dataframes*, facilitando o processamento e a manipulação dos dados. Um exemplo do funcionamento desta função está presente na Figura 3.5.

3.2 Tratamento dos Dados

Após a leitura dos dados, foi necessário realizar uma série de tratamentos. Dada a diversidade dos dados coletados, optou-se por definir as variáveis relevantes para a análise em uma lista chamada `columns_of_interest`, que identifica apenas as colunas de interesse, eliminando informações irrelevantes. Esse tratamento incluiu a padronização dos valores para evitar erros e inconsistências.

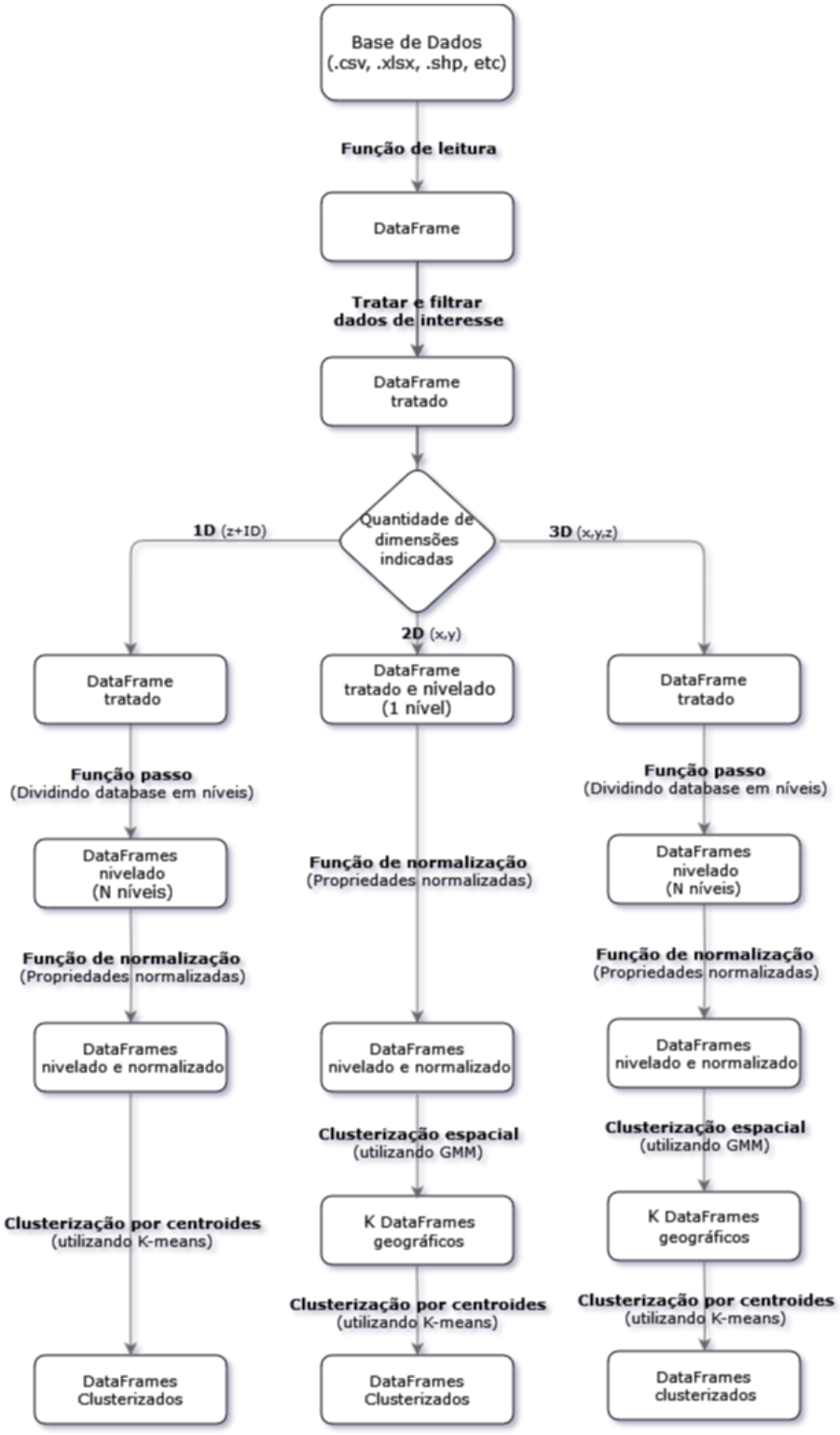


FIGURA 3.1 – Fluxograma das etapas metodológicas do trabalho.

A implementação desse tratamento foi feita usando a biblioteca *pandas* para manipulação e filtragem, com suporte do pacote *re* para lidar com valores em diferentes formatos (como porcentagens). A função `tratar_dados` aplica essas operações de limpeza e conversão, preparando o *dataframe* para as etapas de normalização e clusterização. A escolha

FIGURA 3.2 – Exemplo da execução de uma sondagem CPT. Fonte: (SOLO, 2024).

FIGURA 3.3 – Exemplo da execução de um ensaio de limites. Fonte: (ENGENHARIA, 2024a).

FIGURA 3.4 – Exemplo da execução de uma sondagem SPT. Fonte: (ENGENHARIA, 2024b)

da *pandas* se deu pela sua versatilidade na manipulação de dados e pela facilidade de trabalhar com grandes volumes de dados de forma eficiente. Um exemplo do funcionamento desta função está presente na Figura 3.6.

```
df = read_database('completa_corrigida_csv.csv')
df.head(50)
```

✓ 0.4s Python

	No	ID	Depth_m	qc_MPa	fs_kPa	u_kPa	Other	qt_MPa	Rf_p	SBT	...	SBTn	n	Cn	Icn	Qtn
0	1	1	0.05	1.13	33.21	14.45	0	1.14	2.92	3	...	8	0.63	19.18	2.03	217.66
1	2	1	0.1	1.09	32.07	14.34	0	1.09	2.94	3	...	5	0.66	14.2	2.12	154.48
2	3	1	0.15	1	30.22	14.26	0	1.01	3	3	...	5	0.69	11.88	2.2	119.42
3	4	1	0.2	0.91	27.79	14.18	0	0.91	3.04	3	...	5	0.71	10.62	2.26	96.73
4	5	1	0.25	0.77	23.58	14.08	0	0.78	3.04	3	...	5	0.74	9.88	2.33	76.13
5	6	1	0.3	0.66	16.29	14	0	0.67	2.44	3	...	5	0.74	8.79	2.35	58.17
6	7	1	0.35	0.63	12.75	13.92	0	0.63	2.02	3	...	5	0.75	7.85	2.35	49.08
7	8	1	0.4	0.63	10.99	13.82	0	0.63	1.74	3	...	5	0.74	7.07	2.34	44.24
8	9	1	0.45	0.63	10.5	13.74	0	0.63	1.66	3	...	5	0.75	6.57	2.35	41.11
9	10	1	0.5	0.63	11.09	13.66	0	0.63	1.75	3	...	5	0.76	6.3	2.38	39.45
10	11	1	0.55	0.64	11.77	13.55	0	0.64	1.83	3	...	5	0.77	5.99	2.41	38
11	12	1	0.6	0.65	12.66	13.47	0	0.66	1.93	3	...	5	0.78	5.71	2.43	36.95
12	13	1	0.65	0.67	13.36	13.39	0	0.67	1.99	3	...	5	0.79	5.45	2.45	35.94
13	14	1	0.7	0.68	13.98	13.29	0	0.68	2.04	3	...	5	0.8	5.21	2.46	35
14	15	1	0.75	0.69	14.65	13.21	0	0.69	2.12	3	...	5	0.8	5.01	2.48	33.99

FIGURA 3.5 – Exemplo de um *database* lido pela função `read_database`

```
df2 = tratar_dados(df,columns_of_interest,axis,id = id)
df2.head(10)
```

[13] ✓ 0.8s

	ID	Depth_m	qc_MPa	fs_kPa	u_kPa	qt_MPa	Rf_p	SBT
0	1	0.05	1.13	33.21	14.45	1.14	2.92	3.0
1	1	0.10	1.09	32.07	14.34	1.09	2.94	3.0
2	1	0.15	1.00	30.22	14.26	1.01	3.00	3.0
3	1	0.20	0.91	27.79	14.18	0.91	3.04	3.0
4	1	0.25	0.77	23.58	14.08	0.78	3.04	3.0
5	1	0.30	0.66	16.29	14.00	0.67	2.44	3.0
6	1	0.35	0.63	12.75	13.92	0.63	2.02	3.0
7	1	0.40	0.63	10.99	13.82	0.63	1.74	3.0
8	1	0.45	0.63	10.50	13.74	0.63	1.66	3.0
9	1	0.50	0.63	11.09	13.66	0.63	1.75	3.0

FIGURA 3.6 – Exemplo de um *database* tratada pela função `tratar_dados`

3.3 Função Passo

Para estruturar os dados de forma nivelada e permitir uma análise mais detalhada e precisa, foi utilizada uma função chamada `divide_database`. Essa função introduz o conceito de "passo"para segmentar o banco de dados em diferentes níveis de profundidade, como visto na Figura 3.7, tratando o valor de uma propriedade para um intervalo predefinido como a média dos valores distribuídos dentro desse intervalo, como descrito nas equações 3.1 e 3.2. Essa abordagem, comumente utilizada em outras áreas da engenharia civil, facilita a análise ao criar seções específicas das coordenadas espaciais, representando, por exemplo, a média das características geotécnicas de uma camada de solo. A função divide o conjunto em dois *dataframes* principais: um que armazena as coordenadas espaciais e outro dedicado às propriedades geotécnicas, ambos organizados conforme o intervalo de "passo".

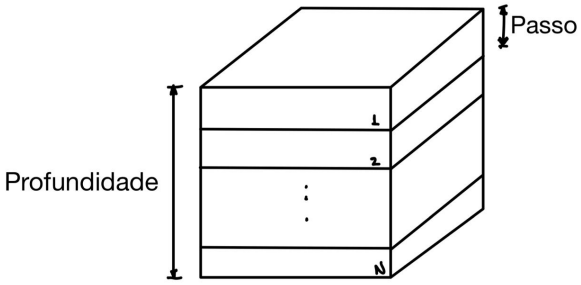
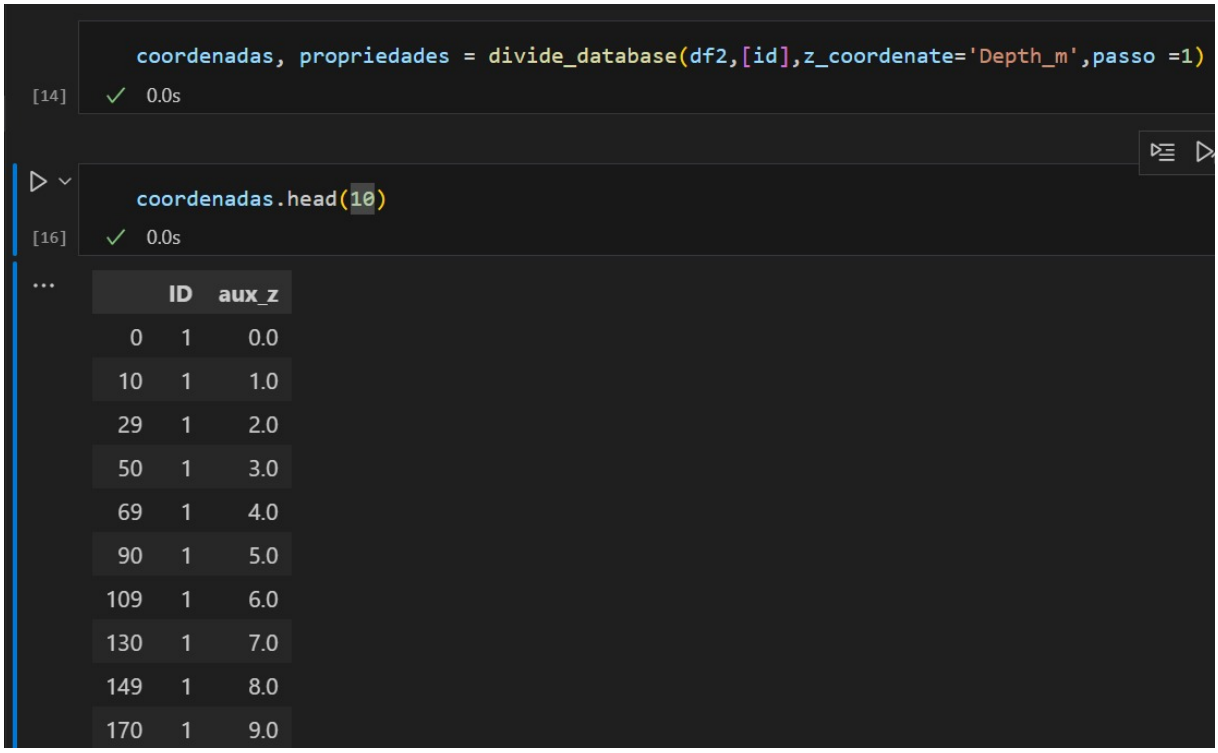


FIGURA 3.7 – Divisão do sistema tridimensional em níveis pela função "passo"

$$N_{camadas} = \left\lfloor \frac{Profundidade}{Passo} \right\rfloor \tag{3.1}$$

$$\left(\overline{Prop}\right)_i = \frac{\sum_{i=1}^{n} Prop_i}{n} \tag{3.2}$$

A implementação utiliza a biblioteca *pandas*, em conjunto com parâmetros como `passo` e `z_coordenate`, para estruturar o banco de dados conforme a segmentação espacial e a profundidade necessária. Essa segmentação permite isolar as coordenadas, exemplificada na Figura 3.8, para uma análise espacial detalhada, enquanto as propriedades geotécnicas, exemplificada na Figura 3.9, são organizadas em um segundo conjunto de dados para facilitar o processamento posterior. A função `divide_database` é, portanto, uma etapa crucial para preparar os dados para a análise de clusterização, garantindo uma integração eficiente com as próximas etapas e permitindo uma análise robusta e alinhada aos objetivos geotécnicos do estudo.

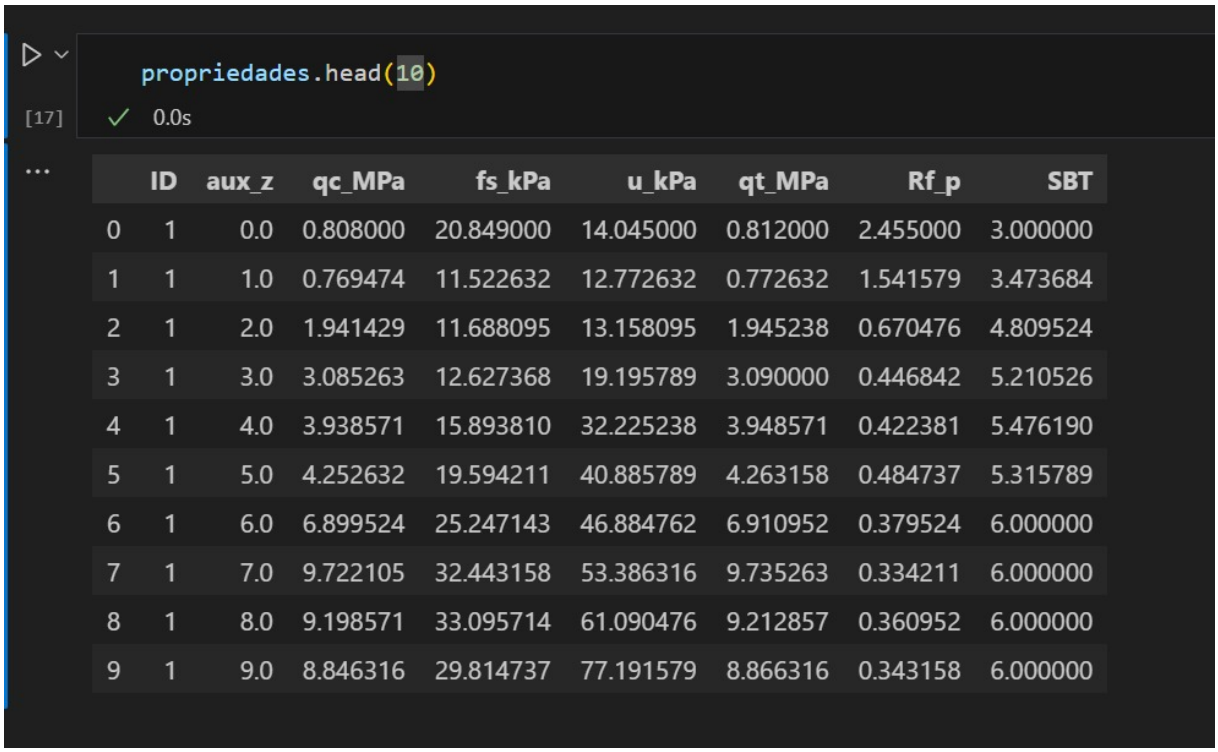


```
[14]  coordenadas, propriedades = divide_database(df2,[id],z_coordenate='Depth_m',passo =1)
      ✓ 0.0s

[16]  coordenadas.head(10)
      ✓ 0.0s
```

	ID	aux_z
0	1	0.0
10	1	1.0
29	1	2.0
50	1	3.0
69	1	4.0
90	1	5.0
109	1	6.0
130	1	7.0
149	1	8.0
170	1	9.0

FIGURA 3.8 – Exemplo de um *database* de coordenadas gerado pela função `divide_database`



```
[17]  propriedades.head(10)
      ✓ 0.0s
```

	ID	aux_z	qc_MPa	fs_kPa	u_kPa	qt_MPa	Rf_p	SBT
0	1	0.0	0.808000	20.849000	14.045000	0.812000	2.455000	3.000000
1	1	1.0	0.769474	11.522632	12.772632	0.772632	1.541579	3.473684
2	1	2.0	1.941429	11.688095	13.158095	1.945238	0.670476	4.809524
3	1	3.0	3.085263	12.627368	19.195789	3.090000	0.446842	5.210526
4	1	4.0	3.938571	15.893810	32.225238	3.948571	0.422381	5.476190
5	1	5.0	4.252632	19.594211	40.885789	4.263158	0.484737	5.315789
6	1	6.0	6.899524	25.247143	46.884762	6.910952	0.379524	6.000000
7	1	7.0	9.722105	32.443158	53.386316	9.735263	0.334211	6.000000
8	1	8.0	9.198571	33.095714	61.090476	9.212857	0.360952	6.000000
9	1	9.0	8.846316	29.814737	77.191579	8.866316	0.343158	6.000000

FIGURA 3.9 – Exemplo de um *database* de propriedades gerado pela função `divide_database`

3.4 Normalização dos Dados

Para garantir a comparabilidade entre variáveis com diferentes escalas e unidades, foi adotada a normalização dos dados. A escolha do método de normalização *Min-Max* visa ajustar as variáveis para um mesmo intervalo, eliminando a influência de disparidades numéricas que possam comprometer a análise. Colunas já normalizadas foram identificadas em uma lista (`percentage_columns`), prevenindo a duplicação do processo.

A normalização foi realizada com a biblioteca *sklearn.preprocessing* por meio da função `normalization_function`, mostrada na Figura 3.10. Essa função verifica a presença de colunas já normalizadas e aplica os métodos de normalização selecionados às colunas restantes. Esse processo é essencial para uniformizar as escalas dos dados e garantir que todas as variáveis tenham um impacto proporcional na etapa de clusterização.

```
normalizado = normalization_function(propriedades, to_not_normalize_col=percentage_columns+[id]+['aux_z'
normalizado.head(10)
```

[18] ✓ 0.0s Python

	ID	aux_z	qc_MPa	fs_kPa	u_kPa	qt_MPa	Rf_p	SBT
0	1	0.0	0.013524	0.019670	0.014165	0.013567	0.133558	0.333333
1	1	1.0	0.012871	0.010845	0.014044	0.012900	0.083662	0.385965
2	1	2.0	0.032746	0.011002	0.014081	0.032773	0.036078	0.534392
3	1	3.0	0.052143	0.011891	0.014652	0.052175	0.023862	0.578947
4	1	4.0	0.066614	0.014981	0.015886	0.066726	0.022526	0.608466
5	1	5.0	0.071940	0.018482	0.016706	0.072058	0.025932	0.590643
6	1	6.0	0.116827	0.023831	0.017274	0.116933	0.020185	0.666667
7	1	7.0	0.164694	0.030639	0.017890	0.164800	0.017710	0.666667
8	1	8.0	0.155816	0.031257	0.018619	0.155946	0.019171	0.666667
9	1	9.0	0.149842	0.028152	0.020143	0.150073	0.018199	0.666667

FIGURA 3.10 – Exemplo do *database* de propriedades normalizado após aplicarmos a função `normalization_function`

3.5 Clusterização Espacial

A clusterização espacial foi realizada com o uso do Modelo de Mistura Gaussiana (GMM), que calcula a probabilidade de cada ponto pertencer a um cluster específico. O Critério de Informação Bayesiano (BIC) foi utilizado para determinar o número ideal de clusters, minimizando a complexidade do modelo sem comprometer a precisão da análise. Esta abordagem é particularmente útil para capturar padrões espaciais e facilitar a visualização dos grupos geográficos.

A clusterização espacial foi realizada utilizando a função `gmm_clustering`, como vista na Figura 3.11 desenvolvida para identificar agrupamentos geoespaciais nos dados de coordenadas e propriedades. Essa função utiliza o modelo `GaussianMixture` da biblioteca *sklearn.mixture*, com o objetivo de otimizar o número de clusters por meio do critério BIC (Bayesian Information Criterion). Inicialmente, as coordenadas são normalizadas com `StandardScaler`, e o modelo seleciona automaticamente o número ideal de clusters baseado na menor pontuação BIC. Em seguida, os clusters obtidos são mapeados de volta para os dados originais, permitindo uma análise de agrupamento eficiente para variações geoespaciais.

```
coordenadas_clusterizado,propriedades_clusterizado =gmm_clustering(coordenadas,normalizado,axis,z_coordinate='Depth_m')
coordenadas_clusterizado
```

✓ 0.0s

	ID	aux_z	cluster_espacial
0	1	0.0	1
10	1	1.0	1
29	1	2.0	1
50	1	3.0	1
69	1	4.0	1
...	...	...	...
85377	111	16.0	1
85418	111	17.0	1
85457	111	18.0	1
85498	111	19.0	1
85537	111	20.0	1

2781 rows × 3 columns

FIGURA 3.11 – Exemplo do *database* de clusterizado espacialmente com o uso do GMM após aplicarmos a função `gmm_clustering`

3.6 Clusterização por Centroides

Além da clusterização espacial, aplicou-se o algoritmo K-means para agrupar os dados em clusters de centróides, uma abordagem clássica e eficaz para identificar padrões gerais. O número ideal de clusters foi determinado com base no Coeficiente de Silhueta, proporcionando uma visão detalhada das propriedades geotécnicas analisadas.

A função `kmeans_optimal_clusters`, como vista na Figura 3.12, utiliza a biblioteca *sklearn.cluster* para implementar o algoritmo K-means. O Coeficiente de Silhueta foi aplicado para escolher o número ideal de clusters, permitindo a criação de agrupamentos representativos, sendo necessário de pelo menos 3 pontos na amostra para criação dos clusters, o que pode ser visto na Figura 3.13. Este método facilita a identificação de padrões entre as propriedades do solo e auxilia na análise das características geotécnicas dos dados. Um exemplo de output esperado para a função é a apresentada na Figura 3.14

```
k=0
j = 0
d={}
optimal_clusters = list()
labels = list()
for k in range(len(it1)):
    if len(it1) == 1:
        aux = propriedades_clusterizado.copy()
    else:
        aux = propriedades_clusterizado[propriedades_clusterizado['cluster_espacial']==it1[k]].copy()

    for j in range(len(list(it2))):
        padrao=aux[aux['aux_z']==it2[j]].copy()
        print(it1[k],it2[j])
        print(padrao)
        a,b,c,d = kmeans_optimal_clusters(padrao,axis+['aux_z'],n_max=10,n_min = 2)
```

55.0s Python

```
1 0.0
      ID  aux_z    qc_MPa    fs_kPa     u_kPa    qt_MPa      Rf_p       SBT  \
0      1    0.0  0.013524  0.019670  0.014165  0.013567  0.133558  0.333333
30     2    0.0  0.057949  0.020304  0.012165  0.054283  0.037014  0.542222
77     3    0.0  0.098910  0.073639  0.012799  0.098723  0.074290  0.577778
156    5    0.0  0.061619  0.006714  0.013440  0.069123  0.010980  0.633333
314    8    0.0  0.002179  0.009407  0.013002  0.002178  0.390513  0.222222
...  ...    ...       ...       ...       ...       ...       ...       ...
2676 107    0.0  0.114342  0.194595  0.011571  0.113680  0.169610  0.511111
2697 108    0.0  0.015322  0.085016  0.013710  0.015338  0.534314  0.272222
2723 109    0.0  0.018324  0.087566  0.012875  0.018304  0.439020  0.333333
2736 110    0.0  0.018765  0.091185  0.015338  0.018829  0.475264  0.316667
2760 111    0.0  0.017512  0.055945  0.014260  0.017137  0.312512  0.327485
```

FIGURA 3.12 – Exemplo do *database* de clusterizado espacialmente com o uso do Kmeans após aplicarmos a função `kmeans_optimal_clusters`

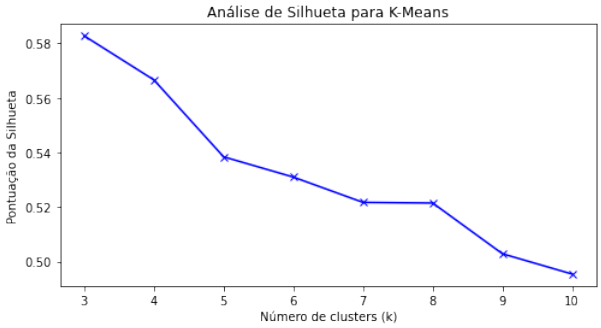


FIGURA 3.13 – Exemplo de gráfico de coeficiente de Silhueta após aplicarmos a função `kmeans_optimal_clusters`

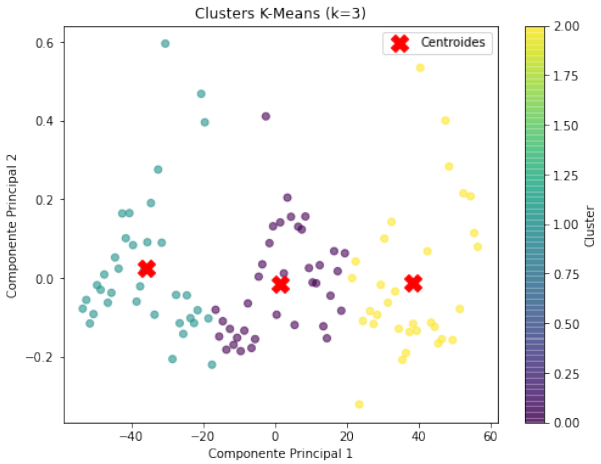


FIGURA 3.14 – Exemplo de output gráfico esperado após aplicarmos a função `kmeans_optimal_clusters`

4 Resultado e Discussões

Neste capítulo, será apresentado uma síntese dos resultados obtidos após a aplicação da metodologia instituída no Capítulo 3. Serão destacados e discutidos, individualmente, os pontos relevantes que puderam ser concluídos na aplicação da metodologia.

4.1 Base de Dados Unidimensional

Utilizou-se a base disponibilizada pelo professor para avaliar a eficácia da metodologia proposta em bases unidimensionais. Os resultados foram comparados ao estudo desenvolvido por Robertson (ROBERTSON, 1990), como visto na Figura 2.1, no qual utiliza as variáveis *cone resistance* (Qt) e *friction ratio* (Fr), obtidas do ensaio CPT, definindo nove possíveis classificações de solo. O uso das funções `read_database` e `tratar_dados` permitiu a preparação adequada dos dados para análise.

Em seguida, dividiu-se a base de dados utilizando a função `divide_database` de acordo com a sua profundidade, gerando dois *dataframes*: o de *coordenadas* e o de *propriedades*. Para o último aplicamos a função passo nas propriedades que se esta estudando, considerando, assim, a média da propriedade como a propriedade da camada estudada.

Com os *dataframes* divididos, pode-se normalizar as propriedades utilizando a função `normalization_function`. Esse passo é crucial para a análise uma vez que, sem a normalização, as propriedades podem enviesar o estudo. Uma parte do dataframe e as etapas seguidas até então se encontram na Figura 4.1

Por se tratar de uma base unidimensional, ao aplicar-se a função `gmm_clustering`, a única mudança realizada aos *dataframes* foi a inserção da coluna `cluster_espacial` que é constante e igual a 1, o que pode ser visto na Figura 4.2. Isso se deve ao fato de não ser necessário a clusterização espacial quando tratamos de dados unidimensionais.

Por fim, aplica-se, iterativamente, a função `kmeans_optimal_clusters` para todo cluster espacial, que para esse caso é somente um, e para todo nível que o database foi dividido, presente na Figura 4.3.

Tendo como exemplo os *outputs* da função para o nível 0 (entre 0 e 1 metros), percebe-

```
# columns_of_interest = ['qc_MPa','fs_kPa','u_kPa','qt_MPa','Rf_p', 'SBT']

columns_of_interest= ['qt_MPa', 'Fr_p']
percentage_columns = ['Fr_p']# ['LL (%)','LP (%)','IP (%)','C.B.R.','ISC']
axis = ['Depth_m'] #'Lat','Long',
id ='ID'
✓ 0.0s
```

```
df = read_database('completa_corrigida_csv3d.csv')
df2 = tratar_dados(df,columns_of_interest,axis,id = id) # id = 'ID'
✓ 0.7s
```

```
coordenadas, propriedades = divide_database(df2,[id],z_coordenate='Depth_m',passo =1) #['ID']
# coordenadas.head(5)
✓ 0.0s
```

```
propriedades['Fr_p']= propriedades['Fr_p']/100
✓ 0.0s
```

```
normalizado = normalization_function(propriedades, to_not_normalize_col=percentage_columns+[id]+['aux_z']) #axis+
normalizado.head(5)
✓ 0.0s
```

	ID	aux_z	qt_MPa	Fr_p
0	1	0.0	0.013567	0.024710
1	1	1.0	0.012900	0.015779
2	1	2.0	0.032773	0.006857
3	1	3.0	0.052175	0.004563
4	1	4.0	0.066726	0.004319

FIGURA 4.1 – Aplicação das funções `read_database`, `tratar_dados`,`divide_database` e `normalization_function` na base unidimensional.

se que a função encontrou um máximo coeficiente de silhueta para 3 *clusters*, como visto na Figura 4.4. Sua distribuição não espacial se encontra na Figura 4.5, em que cada ponto presente na Figura representa uma amostra estudada e cada “x”representa um centroíde de cluster, sendo este seu ponto representativo.

Analisando outros *outputs* da função para o *database* unidimensional, precebe-se que o número máximo de *clusters* encontrados é 3, diferente do esperado pelo estudo empírico de Robertson. Mesmo com tal divergência, o resultado não se mostra absurdo uma vez que Robertson divide as principais classificações de solo (arenoso, siltoso e argiloso) em subgrupos, mostrando que o resultado encontrado é valido. Além disso, outro ponto que pode estar levando a essa divergência é o nível de precisão do algoritmo utilizado. Dessa forma, pode-se afirmar que a metodologia está trazendo resultado de acordo com o esperado.

```
coordenadas_clusterizado,propriedades_clusterizado =gmm_clustering(coordenadas,normalizado,['ID'],z_coordinate='Depth_m')
coordenadas_clusterizado.head(5)
```

✓ 0.0s

	ID	aux_z	cluster_espacial
0	1	0.0	1
10	1	1.0	1
29	1	2.0	1
50	1	3.0	1
69	1	4.0	1

```
propriedades_clusterizado.head(5)
```

✓ 0.0s

	ID	aux_z	qt_MPa	Fr_p	cluster_espacial
0	1	0.0	0.013567	0.024710	1
1	1	1.0	0.012900	0.015779	1
2	1	2.0	0.032773	0.006857	1
3	1	3.0	0.052175	0.004563	1
4	1	4.0	0.066726	0.004319	1

FIGURA 4.2 – *Dataframes* resultantes da utilização da função `gmm_clustering`, clusterizando espacialmente o *database* unidimensional.

```
it1 = list(propriedades_clusterizado['cluster_espacial'].drop_duplicates())
it1 = sorted(it1)

it2 = list(propriedades_clusterizado['aux_z'].drop_duplicates())
it2 = sorted(it2)
print(it1,it2)
```

✓ 0.0s

```
[1] [0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0, 16.0, 17.0, 18.0, 19.0, 20.0, 21.0, 22.0, 23.0, 24.0, 25.0, 26.0, 27.0, 28.0, 29.0, 30.0, 31.0, 32.0, 33.0, 34.0, 35.0, 36
```

```
k=0
j = 0
d={}
optimal_clusters = list()
labels = list()
for k in range(len(it1)):
    if len(it1) == 1:
        aux = propriedades_clusterizado.copy()
    else:
        aux = propriedades_clusterizado[propriedades_clusterizado['cluster_espacial']==it1[k]].copy()

    for j in range(len(list(it2))):
        padrao=aux[aux['aux_z']==it2[j]].copy()
        print(it1[k],it2[j])
        print(padrao)
        d[f'model_{it1[k]}_{it2[j]}'],d[f'cluster_{it1[k]}_{it2[j]}'], d[f'centroid_{it1[k]}_{it2[j]}'],d[f'optimal_n_{it1[k]}_{it2[j]}'] = kmeans_optimal_clusters(padrao,axis+['aux_z'],n_max=10,n_min = 2)
```

✓ 55.0s

FIGURA 4.3 – Implementação do código iterativo do `kmeans_optimal_clusters` para cada nível e cada cluster espacial

4.2 Base de Dados Bidimenionais e Tridimensional

Por falta de bases de dados tridimensionais, adaptou-se a base disponibilizada pelo professor a fim de comprovar a efetividade da metodologia proposta para bases de dados bidimensionais e tridimensionais, pois pode-se considerar uma base bidimensional como uma base tridimensional com 1 nível de profundidade. Similar ao realizado para a base unidimensional, compararam-se os resultados encontrados ao estudo de Robertson (ROBERTSON, 1990). O uso das funções `read_database` e `tratar_dados` permitiu a preparação adequada dos dados para análise.

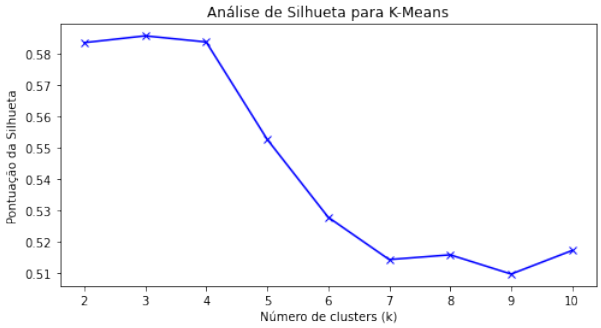


FIGURA 4.4 – Resultado gráfico do método de Shilhueta para otimização do número de clusters, indicando um número ótimo de 3 clusters para a nuvem de pontos específica

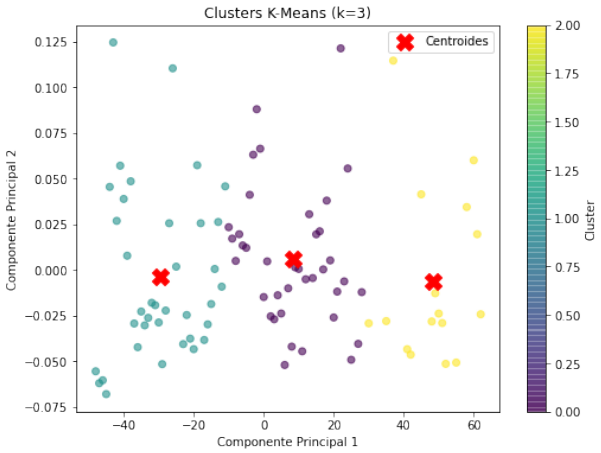


FIGURA 4.5 – Visualização gráfica dos pontos clusterizados

Em seguida, dividiu-se a base de dados utilizando a função `divide_database` de acordo com a sua profundidade, gerando dois *dataframes*: o de *coordenadas* e o de *propriedades*. Para o último, aplicou-se a função passo nas propriedades em estudo, considerando, assim, a média da propriedade como a propriedade da camada analisada.

Com os *dataframes* divididos, é possível normalizar as propriedades utilizando a função `normalization_function`. Esse passo é crucial para a análise, uma vez que, sem a normalização, as propriedades podem enviesar o estudo. Uma parte do *dataframe* e as etapas seguidas até então encontram-se na Figura 4.6.

Aplica-se, depois, a função `gmm_clustering` com a finalidade de clusterizar espacialmente as amostras estudadas, gerando os dataframes clusterizados, presentes na Figura 4.7. Perceba que, pela Figura 4.8, é possível ver com clareza os grupos aos quais os dados pertencem.

Por fim, aplica-se, iterativamente, a função `kmeans_optimal_clusters` para todo cluster espacial e para todo nível que o database foi dividido, presente na Figura 4.9.

Tendo como exemplo os *outputs* da função para o nível 21 (entre 21 e 22 metros) e para o cluster espacial 0, percebe-se que a função encontrou um coeficiente de silhueta máximo

```
# columns_of_interest = ['qc_MPa','fs_kPa','u_kPa','qt_MPa','Rf_p', 'SBT']

columns_of_interest= ['qt_MPa', 'Fr_p']
percentage_columns = ['Fr_p']# ['LL (%)','LP (%)','IP (%)','C.B.R.','ISC']
axis = ['Lat','Long','Depth_m'] #
# id ='ID'
```
✓ 0.0s

```
df = read_database('completa_corrigida_csv3d.csv')
df2 = tratar_dados(df,columns_of_interest,axis) # id = 'ID'
```
✓ 1.0s

+ Code + Markdown

```
coordenadas, propriedades = divide_database(df2,axis,z_coordenate='Depth_m',passo =1) #['ID']
# coordenadas.head(5)
```
✓ 0.1s

```
propriedades['Fr_p']= propriedades['Fr_p']/100
```
✓ 0.0s

```
normalizado = normalization_function(propriedades, to_not_normalize_col=percentage_columns+[id]+['aux_z']) #axis+
normalizado.head(5)
```
✓ 0.0s

	ID	qt_MPa	Fr_p	aux_z
0	1	0.013567	0.024710	0.0
1	2	0.012900	0.015779	1.0
2	3	0.032773	0.006857	2.0
3	4	0.052175	0.004563	3.0
4	5	0.066726	0.004319	4.0

FIGURA 4.6 – Aplicação das funções `read_database`, `tratar_dados`,`divide_database` e `normalization_function` na base tridimensional.

para 2 *clusters*, como visto na Figura 4.10. Sua distribuição não espacial está representada na Figura 4.11, onde cada ponto na Figura corresponde a uma amostra estudada e cada "x" representa o centróide de um cluster, sendo este o seu ponto representativo.

Ao analisar outros *outputs* da função para o *database* tridimensional, observa-se que o número máximo de *clusters* identificados é três, diferindo do esperado com base no estudo empírico de Robertson. No entanto, assim como os resultados obtidos para o *dataframe* unidimensional, essa diferença não é necessariamente surpreendente, pois Robertson subdivide as classificações principais de solo (arenoso, siltoso e argiloso) em subgrupos, o que valida a consistência dos resultados encontrados. Ademais, uma possível causa dessa variação pode ser o nível de precisão do algoritmo empregado. Com isso, pode-se concluir que a metodologia está fornecendo resultados alinhados com as expectativas.

```
coordenadas_clusterizado,propriedades_clusterizado =gmm_clustering(coordenadas,normalizado,axis,z_coordinate='Depth_m')
coordenadas_clusterizado.head(5)
```

✓ 0.1s

	Lat	Long	aux_z	ID	cluster_espacial
0	81.363635	141.047191	0.0	1	1
1	81.363635	141.047191	1.0	2	1
2	81.363635	141.047191	2.0	3	1
3	81.363635	141.047191	3.0	4	1
4	81.363635	141.047191	4.0	5	1

+ Code + Markdown

```
propriedades_clusterizado.head(5)
```

✓ 0.0s

	ID	qt_MPa	Fr_p	aux_z	cluster_espacial
0	1	0.013567	0.024710	0.0	1
1	2	0.012900	0.015779	1.0	1
2	3	0.032773	0.006857	2.0	1
3	4	0.052175	0.004563	3.0	1
4	5	0.066726	0.004319	4.0	1

FIGURA 4.7 – *Dataframes* resultantes da utilização da função `gmm_clustering`, clusterizando espacialmente o *database* tridimensional.

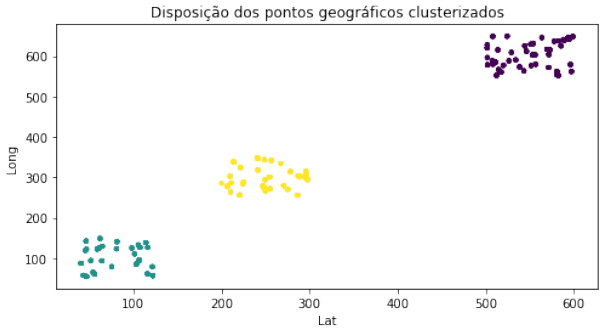


FIGURA 4.8 – Representação gráfica geográficas do *dataframe* de coordenadas clusterizado indicando a presença de três nuvens de pontos

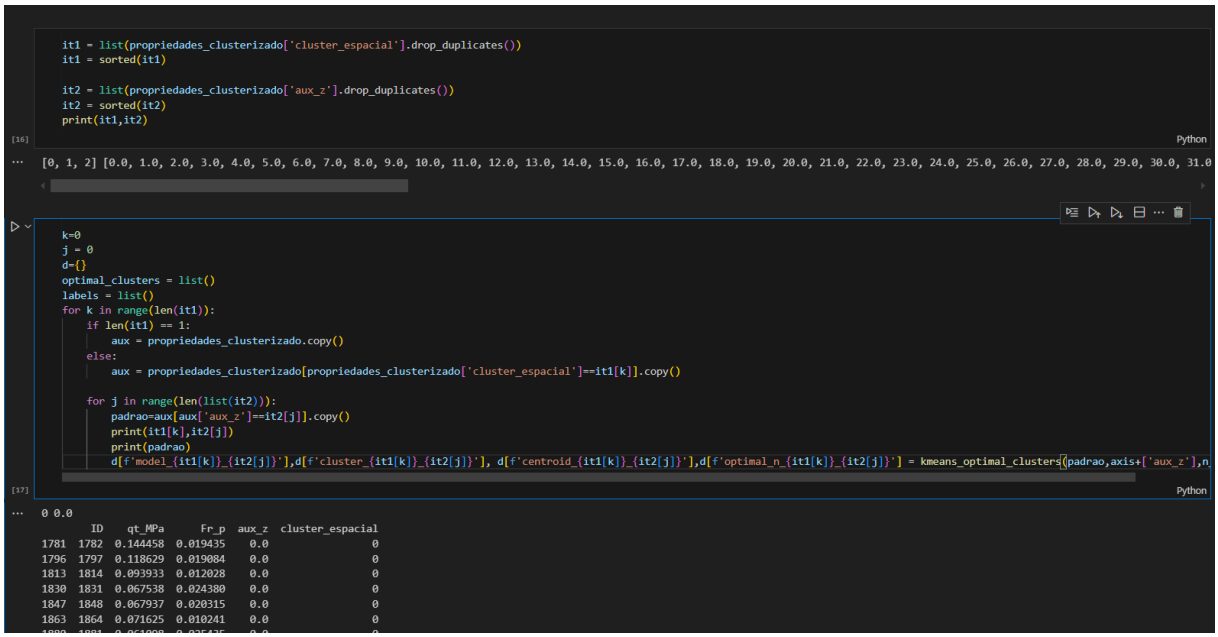


FIGURA 4.9 – Implementação do código iterativo do `kmeans_optimal_clusters` para cada nível e cada cluster espacial

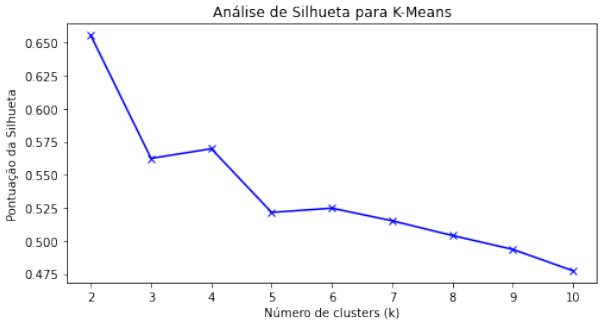


FIGURA 4.10 – Resultado gráfico do método de Shilhueta para otimização do número de clusters, indicando um número ótimo de 3 clusters para a nuvem de pontos específica

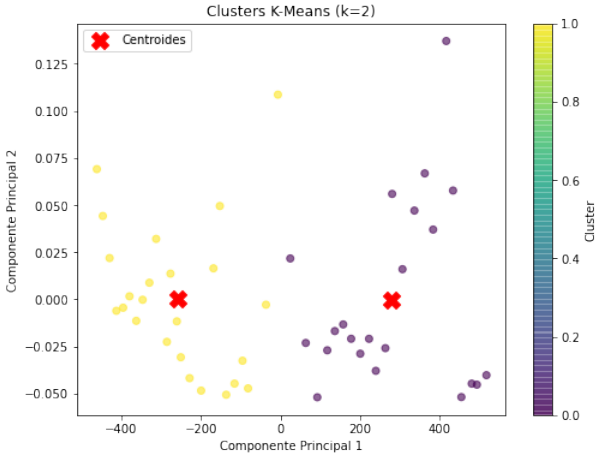


FIGURA 4.11 – Visualização gráfica dos pontos clusterizados

5 Considerações Finais

Neste capítulo, será apresentado uma visão geral do trabalho e sugestões de temas para estudos futuros.

5.1 Considerações Finais

Neste trabalho, foi desenvolvida uma metodologia inicial com o objetivo de aplicar técnicas de *machine learning*, especificamente métodos de aprendizado não supervisionado, no mapeamento de solos. A utilização de algoritmos de clusterização, aliada ao uso de ferramentas computacionais modernas, buscou facilitar o processo de análise de solos, que atualmente é conduzido de forma majoritariamente manual, além de permitir a identificação de padrões mais complexos e robustos.

Para alcançar esse objetivo, foi necessário desenvolver uma metodologia concisa, robusta e replicável, além de entender como se deve tratar as propriedades de camadas do solo. Também foi fundamental comprovar o seus resultados a partir de estudo empíricos e com grande respaldo literário, como o de Robertson (ROBERTSON, 1990).

Ao desenvolver essa metodologia, observou-se que a aplicação de algoritmos de clusterização como o Modelo de Mistura Gaussiana (GMM) e o K-means permitiu uma análise detalhada das propriedades geotécnicas, facilitando a divisão de padrões em diferentes tipos de solo. Essa abordagem, além de fornecer uma ferramenta útil para o mapeamento de solos, mostrou-se capaz de adaptar-se a diferentes bases de dados e propriedades geológicas.

Para assegurar a correspondência entre os dados e os clusters identificados com o gráfico empírico de Robertson, seria ideal realizar a associação dos grupos obtidos com as categorias características descritas no gráfico, além de compará-los com a classificação real dos solos presentes na base de dados. Contudo, devido à ausência dessa classificação real na base utilizada, não foi possível efetuar essa validação direta, o que limita a análise da verossimilhança com o comportamento real esperado.

Por fim, a metodologia proposta demonstrou potencial para automação e padronização

dos processos de análise de solos. No entanto, devido à natureza exploratória do trabalho, espera-se que os métodos aqui introduzidos possam ser aprimorados em estudos futuros, de forma a elevar a precisão e a aplicabilidade dos resultados.

5.2 Sugestões para Trabalhos Futuros

Este trabalho representa um passo inicial na aplicação de *machine learning* para o mapeamento geotécnico. A seguir, são apresentadas algumas sugestões para pesquisas futuras que podem contribuir para a evolução dessa área:

- **Exploração de Outros Métodos de Clusterização e Normalização**: A implementação de diferentes algoritmos de clusterização e estratégias de normalização pode aprimorar a segmentação e interpretação dos dados, permitindo a comparação dos resultados com aqueles obtidos no presente estudo. Testes com modelos alternativos, como *DBSCAN* ou *Agglomerative Clustering*, podem fornecer insights adicionais e adequar-se melhor a distribuições específicas dos dados.

- **Consideração de Dados Topográficos**: Incorporar dados topográficos nas análises permitiria levar em conta não apenas a profundidade dos dados, mas também a sua localização geográfica. Esse aspecto é fundamental, pois a variação topográfica pode impactar significativamente a interpolação e a distribuição das camadas de solo, especialmente em terrenos irregulares. Estudos futuros poderiam incorporar a topografia como uma dimensão adicional de análise.

- **Uso da Biblioteca *GeoPandas* para Dados GIS**: A inclusão da biblioteca *GeoPandas* permitiria a integração de arquivos GIS no processo de análise, facilitando o tratamento de dados espaciais e georreferenciados. Esse recurso pode ser particularmente útil na visualização de clusters em mapas e na análise espacial detalhada, proporcionando uma abordagem mais rica e precisa para o mapeamento geotécnico.

- **Análise da Eficiência do Método**: Realizar uma análise comparativa dos resultados obtidos com outros métodos de análise geotécnica para validar a eficiência e robustez do modelo proposto. Essa comparação pode ser feita em relação a métodos tradicionais e modernos, com o objetivo de verificar a acurácia, aplicabilidade e ganho de eficiência do modelo baseado em *machine learning*.

- **Implementações Matemáticas ou Computacionais Adicionais**: A incorporação de modelos matemáticos e computacionais mais avançados, como técnicas de interpolação sofisticadas e redes neurais, pode ampliar significativamente a capacidade do sistema de identificar e interpretar padrões complexos nos dados geotécnicos.

Além disso, abordar a questão da dependência linear entre as propriedades, por meio da identificação de componentes principais, pode reduzir o custo computacional de análises mais complexas e melhorar a eficiência e a precisão dos processos de clusterização. Essas melhorias não apenas aprimoram a qualidade dos resultados obtidos, mas também podem fornecer novos insights sobre as propriedades e o comportamento dos solos analisados.

Essas sugestões visam promover o desenvolvimento de uma metodologia mais robusta e abrangente para o mapeamento de solos, explorando o potencial do *machine learning* e ferramentas computacionais avançadas na área de geotécnica.

Referências

ABNT. **Solo - Sondagens de Simples Reconhecimento com SPT - Método de Ensaio**. Rio de Janeiro, 2001.

AGGARWAL, C. C.; REDDY, C. K. **Data Clustering: Algorithms and Applications**. Boca Raton: CRC Press, 2014.

AOKI, N.; VELLOSO, D. A. An approximate method to estimate the bearing capacity of piles. In: **Proceedings of the 5th Pan-American Conference on Soil Mechanics and Foundation Engineering**. Buenos Aires, Argentina: [s.n.], 1975. p. 367–376.

BERNSTEIN, M. N. **Gaussian mixture models**. Boston, MA, 2020. Disponível em: <https://mbernste.github.io/posts/gmm_em/>. Acesso em: 29 oct. 2024.

BILAL, M.; OYEDELE, L. O.; QADIR, J.; MUNIR, K.; AJAYI, S. O.; AKINADE, O. O.; PASHA, M. Big data in the construction industry: A review of present status, opportunities, and future trends. **Advanced Engineering Informatics**, v. 30, n. 3, p. 500–521, 2016.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006.

BURNHAM, K. P.; ANDERSON, D. R. **Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach**. 2. ed. New York: Springer, 2014.

CHEN, M.; MAO, S.; LIU, Y. Big data: A survey. **Mobile Networks and Applications**, v. 19, n. 2, p. 171–209, 2014.

DATAAT. **Introdução ao Machine Learning: Agrupamento**. 2024. `https://dataat.github.io/introducao-ao-machine-learning/agrupamento.html`. Acessado em: 29 de outubro de 2024.

DUNCAN, J. M. Factors of safety and reliability in geotechnical engineering. **Journal of Geotechnical and Geoenvironmental Engineering**, v. 126, n. 4, p. 307–316, 2000.

EASTMAN, C. M.; TEICHOLZ, P.; SACKS, R.; LISTON, K. **BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors**. 3. ed. Hoboken, NJ: John Wiley & Sons, 2018.

ENGENHARIA, H. **Áreas de Atuação - Habisolute Engenharia**. 2024. `https://habisoluteengenharia.com.br/area-de-atuacao/12`. Acessado em: 29 de outubro de 2024.

ENGENHARIA, R. S. **Sondagem SPT**. 2024. https://www.renansousaeng.com.br/post/57391-sondagem-SPT. Acessado em: 29 de outubro de 2024.

FENG, X.; ZHOU, H. Dynamics and challenges in geotechnical engineering: state of the art and practice. **Journal of Rock Mechanics and Geotechnical Engineering**, v. 8, n. 3, p. 241–248, 2016.

GARCíA, S.; LUENGO, J.; HERRERA, F. **Data Preprocessing in Data Mining**. [S.l.]: Springer, 2015.

GUO, H.; LI, H.; SKITMORE, M.; HUANG, T.; LUO, X. Using big data and machine learning to predict safety hazard levels for construction projects. **Automation in Construction**, v. 93, p. 203–213, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. [S.l.]: Springer, 2009.

HENGL, T.; NUSSBAUM, M.; WRIGHT, M. N.; HEUVELINK, G. B. M.; GRäLER, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2015. p. 448–456.

JONES, M. J. T. **K-Means Clustering**. 2024. https://www.learnbymarketing.com/methods/k-means-clustering/. Acessado em: 29 de outubro de 2024.

JR., D. J. K.; SHOOK, C. L. The application of cluster analysis in strategic management research: an analysis and critique. **Strategic Management Journal**, v. 17, n. 6, p. 441–458, 1996.

LEE, S.; KIM, Y.-J. Landslide susceptibility mapping using convolutional neural network and gis. **Landslides**, v. 15, n. 2, p. 397–408, 2018.

LI, J.; SHAN, M.; HWANG, B.-G. Big data analytics in the construction industry: A review. **Engineering, Construction and Architectural Management**, v. 26, n. 10, p. 2346–2367, 2019.

LOVE, P. E. D.; MATTHEWS, J.; SING, M. C. P.; ZHOU, J. Z. A framework for reconciliation of clashes in building information models. **Automation in Construction**, v. 85, p. 224–234, 2018.

MA, L.; LIU, Y.; ZHANG, X.; YE, Y.; YIN, G.; JOHNSON, B. A. Deep learning in remote sensing applications: A meta-analysis and review. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 152, p. 166–177, 2019.

MICELI, P. **História: A história da cartografia e a importância dos mapas**. outubro 2013. YouTube. YouTube, acesso em: 3 de novembro de 2024. Disponível em: <https://www.youtube.com/watch?v=Ls-DTif6QKg>.

MILLER, H. J.; GOODCHILD, M. F. Data-driven geography. **GeoJournal**, v. 80, n. 4, p. 449–461, 2015.

MINELLI, M.; CHAMBERS, M.; DHIRAJ, A. **Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses**. Hoboken: John Wiley & Sons, 2014.

NI, Y. Q.; WONG, K. Y. The state-of-the-art of structural health monitoring in hong kong infrastructures. **Structural Monitoring and Maintenance**, v. 2, n. 1, p. 1–32, 2015.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Machine learning and soil sciences: A review aided by machine learning tools. **Soil**, v. 5, n. 1, p. 35–52, 2019.

PARK, N.-W.; KIM, J. Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. **Applied Sciences**, v. 9, n. 5, p. 942, 2019.

PATRO, S. G. K.; SAHU, K. K. Normalization: A preprocessing stage. **arXiv preprint arXiv:1503.06462**, 2015.

ROBERTSON, P. K. Soil classification using the cone penetration test. **Canadian Geotechnical Journal**, v. 27, n. 1, p. 151–158, 1990.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987.

SHI, W.; YUAN, H.; YE, Q. Spatial data clustering: A survey of methods in spatial data mining. **Geo-spatial Information Science**, v. 18, n. 2, p. 83–91, 2015.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. **Applied Soft Computing**, Elsevier, v. 97, p. 105524, 2020.

SOLO, C. **Ensaios de CPT e CPTU**. 2024. `https://www.civilsolo.com.br/ensaios-de-cpt-e-cptu/`. Acessado em: 29 de outubro de 2024.

SUN, C.; JIANG, S.; SKIBNIEWSKI, M. J.; MAN, Q.; SHEN, L.-Y. A literature review of the factors limiting the application of bim in the construction industry. **Technological and Economic Development of Economy**, v. 23, n. 5, p. 764–779, 2017.

XU, C.; DAI, F.-d.; TU, X.-b.; THAM, L. G. Application of cluster analysis and discriminant analysis to spatial prediction of landslides at regional scale: a case study in the zigui segment of the yangtze three gorges, china. **Bulletin of Engineering Geology and the Environment**, v. 78, n. 6, p. 4035–4052, 2019.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. **Annals of Data Science**, v. 2, n. 2, p. 165–193, 2015.

ZHANG, J.; ZHAO, Z.; LI, M.; YU, J. Big data mining in geotechnical engineering: Recent advances and future trends. **KSCE Journal of Civil Engineering**, v. 19, n. 5, p. 1169–1176, 2015.

ZHANG, L.; ZHANG, L.; DU, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. **IEEE Geoscience and Remote Sensing Magazine**, v. 4, n. 2, p. 22–40, 2016.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
TC	14 de novembro de 2024	DCTA/ITA/TC-084/2024	53

5. TÍTULO E SUBTÍTULO:

Mapeamento dos Solos Brasileiros em função de parâmetros Geotécnicos para uso em pavimentos.

6. AUTOR(ES):

Marcelo de Deus Pompeu Magalhães

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Mapeamento; Geotecnia; Machine Learning

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Pavimentos; Mapeamento de solos; Aprendizagem (inteligência artificial); Modelo de mistura de gaussianas; Geotecnia; Engenharia civil; Engenharia estrutural.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA, São José dos Campos, Curso de Engenharia Civil-Aeronáutica. Orientadora: Cláudia Azevedo Pereira, Coorientador: José Antonio Schiavon; Apresentação em 13/11/2024. Publicada em 2024.

11. RESUMO:

Este trabalho de conclusão de curso aborda o "Mapeamento dos Solos Brasileiros em Função de Parâmetros Geotécnicos para Uso em Pavimentos". O estudo propõe uma metodologia que combina técnicas de clusterização, como o Modelo de Mistura Gaussiana (GMM) e o K-means, para classificar solos brasileiros com base em parâmetros geotécnicos. Utilizando dados de sondagens, como CPT, limites de Atterberg e SPT, o trabalho visa identificar padrões nos solos, facilitando a análise e o planejamento de pavimentações e construções. A metodologia inclui a normalização dos dados, segmentação tridimensional, e análise de clusters, proporcionando uma visão detalhada das propriedades geotécnicas. Os resultados destacam a eficiência da metodologia proposta e sugerem sua aplicação prática em diferentes cenários da engenharia civil.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**