Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Comparação de penetrômetros e identificação de fatores na resistência mecânica do solo à penetração via modelos mistos

Leonardo Ribeiro

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Comparação de penetrômetros e identificação de fatores na resistência mecânica do solo à penetração via modelos mistos

Leonardo Ribeiro

Orientadora: Profª. Drª. Daiane Aparecida Zuanetti

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos

Janeiro de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Comparison of penetrometers and identification of factors in soil mechanical resistance to penetration through mixed models

Leonardo Ribeiro
Advisor: Profª. Drª. Daiane Aparecida Zuanetti

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
January 2024

Leonardo Ribeiro

Comparação de penetrômetros e identificação de fatores na resistência mecânica do solo à penetração via modelos mistos

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Leonardo Ribeiro e aprovado pela banca examinadora.

Aprovado em 31 de janeiro de 2024

Banca Examinadora:

- Profª. Drª. Daiane Aparecida Zuanetti
- Prof. Dr. Afrânio Márcio Corrêa Vieira
- Prof. Dr. Ricardo Felipe Ferreira

Dedico esse Trabalho para meus pais, Marlene e Paulo Cesar, e meus irmãos, Valentina e Vinicius, que sempre estiveram ao meu lado para me apoiar.

Resumo

É de conhecimento geral que uma das maiores fontes da economia brasileira é atribuída à agricultura. No Brasil, a agricultura trata-se de uma área competitiva, sendo um meio econômico muito rico, diversificado, além de ser uma fonte de alimentação e uma fonte geradora de empregos entre os brasileiros. Para a agricultura, a principal matéria-prima é o solo que é essencial para o desenvolvimento das plantações e, devido a isso, é vital que haja um cuidado com esse recurso. Uma das técnicas mais utilizadas para a avaliação do solo é verificar a sua resistência mecânica à penetração. Essa técnica é a preferida para o processo de análise pela rapidez na verificação e pelo fácil manuseamento da ferramenta utilizada: o penetrômetro.

Neste trabalho, por meio do modelo de regressão misto, comparamos a eficiência de diferentes tipos de penetrômetros e também analisamos características e fatores que podem influenciar a resistência mecânica do solo à penetração a partir de um conjunto de dados reais do LAMAP, USP.

Palavras-chave: *Análise do solo, fatores de risco e proteção, efeitos aleatórios, resistência mecânica.*

Abstract

It is common knowledge that one of the main sources of the Brazilian economy is attributed to agriculture. In Brazil, agriculture is a competitive field, being a very rich and diversified economic sector, as well as a source of food and job creation among Brazilians. For agriculture, the main raw material is the soil, which is essential for the development of crops, and therefore, it is vital to take care of this resource. One of the most commonly used techniques for soil assessment is to verify its mechanical resistance to penetration. This technique is preferred for the analysis process due to its quick verification and easy handling of the tool used: the penetrometer.

In this work, through the mixed regression model, we compared the efficiency of different types of penetrometers and also analyzed characteristics and factors that may influence the mechanical resistance of soil to penetration using a set of real data from LAMAP, USP.

Keywords: *Mechanical resistance, random effects, risk and protective factors, soil analysis.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

É de conhecimento geral que uma das maiores fontes da economia brasileira é a agricultura. No Brasil, a agricultura trata-se de uma área competitiva, sendo um meio econômico muito rico, diversificado, além de ser uma fonte de alimentação e uma fonte geradora de empregos entre os brasileiros. Em termos econômicos, a agricultura é um dos setores mais responsáveis pelo crescimento do PIB no país, sendo que ela corresponde a 21% de todas as riquezas nacionais produzidas, um quinto de todos os empregos e 43.2% das exportações brasileiras, chegando a US$ 96.7 bilhões em 2019 (Embrapa, 2020). Com o passar dos anos, a produção alimentícia também apresentou um amplo crescimento, sendo que ao longo das últimas quatro décadas a produção de grãos exibiu um aumento de 510% (232.6 milhões de toneladas) e a produção de carnes obteve um salto de 858% (27.9 milhões de toneladas). Além dessas, outras produções também alcançaram grandes saltos de produtividade (Embrapa, 2020).

A principal matéria-prima para a agricultura é o solo que é essencial para o desenvolvimento das plantações devido à sua riqueza em nutrientes e suas funções (filtragem d'água, decomposição de resíduos, armazenamento de calor e troca de gases). Em razão dessa importância, é vital que haja um grande cuidado na conservação desse recurso para que seja possível aplicar uma agricultura produtiva e sustentável, tornando-se assim saudável para o ecossistema (Brasil, 2020).

O Brasil detém uma vasta diversidade em relação aos tipos de solo e cada um necessita de uma atenção especial para que a boa qualidade dessa matéria-prima seja mantida. Entretanto, o que vem acontecendo atualmente é a degradação desse recurso. De acordo com um relatório recente das Nações Unidas, quase um terço das terras cultiváveis do mundo desapareceu nas últimas quatro décadas. Também foi identificado que todo o solo

superficial do mundo poderá se tornar improdutivo dentro de 60 anos se as taxas atuais de perda continuarem (Brasil, 2020).

Com o desenvolvimento mecânico da agricultura, o número de máquinas de grande porte intensificou-se sobre o solo das lavouras. Por conta disso, acentuou-se uma das principais causas da degradação desse recurso, a compactação. O excesso de manipulação do solo por uso constante de máquinas agrícolas e pisoteio de animais pesados faz com que, devido à pressão causada, ocorra uma diminuição do seu volume não saturado, causando assim, a expulsão do ar do solo e, consequentemente, um aumento de densidade (Machado, 2003).

A compactação leva ao aumento da resistência mecânica do solo no crescimento das raízes das plantas. Devido ao aumento de densidade causada pela pressão no solo, seu rompimento pelas raízes das plantas na profundidade fica prejudicado e, em muitos casos, a planta não consegue crescer sua raiz totalmente, além da diminuição do ar no solo estar presente. Isso pode causar uma morte nas raízes por asfixia e também ter um baixo acesso à água e nutrientes, ocasionando no não crescimento da planta (Machado, 2003).

Quando o fenômeno da compactação acontece, muitas vezes chega a ser inviável a tentativa de reversão desse processo, pois é uma atividade extremamente custosa. Por conta disso, é mais interessante economicamente verificar com frequência a qualidade do solo e a não presença da compactação nas lavouras (Menezes, 2018).

Uma das técnicas mais utilizadas para a avaliação do solo é a análise da sua resistência mecânica do solo à penetração. Essa técnica é a preferida para esse processo pela rapidez na verificação e pelo fácil manuseamento da ferramenta utilizada: o penetrômetro (Menezes, 2018).

Vaz *et al.* (2002) diz que penetrômetros são instrumentos que medem a resistência mecânica à penetração em unidades de pressão (força/área) de um cone padrão posicionado na extremidade de uma haste de metal, quando inseridos no interior do solo. A análise feita na mesma área pode ser enviesada pela variabilidade espacial e temporal, erro do operador e tipo de equipamento, podendo causar assim, um erro de avaliação na descompactação ou não do solo. Existem diferentes tipos de penetrômetros como, por exemplo, o manual e o automático. O primeiro é mais fácil de ser transportado, porém o automático mantém constante a velocidade de penetração da sonda (Menezes, 2018).

Menezes (2018) relata um projeto e resultados conduzidos com o objetivo de comparar os equipamentos, além de investigar a influência de características texturais, estrutura e

umidade nas leituras da resistência mecânica do solo à penetração. O pesquisador realizou uma análise de variância (ANOVA) para a obtenção de seus resultados considerando independência entre os dados. Porém, os dados são dependentes, por se tratar de medições em diferentes profundidades no mesmo solo e local, e a metodologia adotada pode não ser adequada. Um dos modelos mais tradicionais e flexíveis para analisar dados dependentes de medidas repetidas ou longitudinais é o modelo misto (Singer *et al.*, 2018; Zhang, 2015; Zuanetti, 2022).

O modelo misto é um modelo de regressão que considera seus coeficientes de maneira fixa e aleatória. Pressupor esses tipos de coeficientes permite uma correlação entre as variáveis respostas de uma mesma unidade experimental, ou seja, uma dependência nos dados. Também permite variâncias heterogêneas para diferentes observações e, consequentemente, uma melhor precisão nos resultados e fácil interpretabilidade (Zuanetti, 2022; Singer *et al.*, 2018).

Dessa maneira, este trabalho tem como objetivo comparar a performance dos penetrômetros e também verificar a influência de características como umidade, tipo de solo e profundidade nas leituras da resistência mecânica do solo à penetração por meio da utilização do modelo misto. O conjunto de dados a ser analisado é o citado acima e disponível em Menezes (2018).

Este relatório está organizado como a seguir. O Capítulo 2 apresenta e descreve as principais características dos modelos mistos, metodologia estatística a ser utilizada nesse estudo. O Capítulo 3 mostra como foram coletados os dados, as variáveis a serem trabalhadas no decorrer do estudo e, por último, exibe uma análise inicial para os efeitos principais, algumas interações duplas e triplas. O Capítulo 4 apresenta os principais resultados e, finalmente, no Capítulo 5 apresentamos as conclusões e estudos futuros.

Capítulo 2

Modelos mistos

Alguns modelos estatísticos assumem que os dados não possuem correlação entre si, ou seja, há independência entre eles, e também, que os erros são provenientes de uma distribuição $Normal(0, \sigma^2)$, em que σ^2 é um número real positivo e constante. Porém, o trabalho tem como intuito analisar medições em 8 camadas de profundidade diferentes em determinados solos e locais. Conduzir esse experimento, nessas condições, não nos garante mais que os dados são independentes pois são longitudinais em relação às profundidades consideradas nas medições e esses apresentam uma certa correlação entre as camadas. Além disso, podemos ter também heterocedasticidade nas medições por causa dos níveis de profundidade e compactações do solo observadas em cada nível. Nesse cenário, a princípio o modelo misto apresenta as propriedades adequadas para descrever o conjunto de dados e também possui parâmetros de fácil interpretação.

2.1 Modelo para uma unidade amostral

O modelo misto é um modelo de regressão que apresenta efeitos fixos e aleatórios. Os efeitos fixos são vistos como parâmetros desconhecidos, entretanto, comuns a todas as observações e que impactam na média da variável resposta. Já os aleatórios, são vistos como uma variável aleatória podendo resultar em qualquer valor para cada unidade amostral, além de impactarem diretamente na variância da variável de interesse (Zuanetti, 2022). O modelo pode ser representado pela Equação (2.1) como:

$$\underset{m_i \times 1}{\mathbf{Y}_i} = \underset{m_i \times p}{\mathbf{X}_i} \times \underset{p \times 1}{\boldsymbol{\beta}} + \underset{m_i \times q}{\mathbf{Z}_i} \times \underset{q \times 1}{\mathbf{b}_i} + \underset{m_i \times 1}{\boldsymbol{\epsilon}_i}, \text{para } i = 1, ..., n, \tag{2.1}$$

em que

$\mathbf{Y}_i$: é o vetor aleatório com as variáveis respostas da i-ésima unidade amostral com dimensão $(m_i \times 1)$, em que m_i é o número de observações nessa unidade amostral;

$\boldsymbol{\beta}$: é o vetor de efeitos fixos do modelo com dimensão $(p \times 1)$;

$\mathbf{X}_i$: é a matriz de planejamento associada a $\boldsymbol{\beta}$ com dimensão $(m_i \times p)$, e contém as observações das covariáveis associadas a efeitos fixos da i-ésima unidade amostral;

$\mathbf{b}_i$: é o vetor dos efeitos aleatórios associados à i-ésima unidade amostral com dimensão $(q \times 1)$, $\mathbf{b}_i \sim Normal_q(\mathbf{0}, \mathbf{G}_q)$ e $\mathbf{G}_q$ é a matriz de covariâncias para os efeitos aleatórios;

$\mathbf{Z}_i$: é a matriz de planejamento associada a $\mathbf{b}_i$ com dimensão $(m_i \times q)$, e contém as covariáveis associadas a efeitos aleatórios da i-ésima unidade amostral; e

$\boldsymbol{\epsilon}_i$: é o vetor dos erros aleatórios da i-ésima unidade amostral com dimensão $(m_i \times 1)$, $\boldsymbol{\epsilon}_i \sim Normal_{m_i}(\mathbf{0}, \mathbf{R}_i)$ e, geralmente, $\mathbf{R}_i = \sigma^2 \mathbf{I}_{m_i}$ e $\mathbf{I}$ é a matriz identidade.

Também podemos escrever o modelo (2.1) na forma matricial da seguinte maneira,

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ Y_{im_i} \end{bmatrix} = \begin{bmatrix} X_{i11} & X_{i12} & . & . & . & X_{i1p} \\ X_{i21} & X_{i22} & . & . & . & X_{i2p} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ X_{im_i1} & X_{im_i2} & . & . & . & X_{im_ip} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} Z_{i11} & Z_{i12} & . & . & . & Z_{i1q} \\ Z_{i21} & Z_{i22} & . & . & . & Z_{i2q} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ Z_{im_i1} & Z_{im_i2} & . & . & . & Z_{im_iq} \end{bmatrix} \begin{bmatrix} b_{i1} \\ b_{i2} \\ . \\ . \\ . \\ b_{iq} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ . \\ . \\ . \\ \epsilon_{im_i} \end{bmatrix}. \tag{2.2}$$

Geralmente, adotamos a independência entre $\mathbf{b}_i$ e $\boldsymbol{\epsilon}_i$ para obtermos a média e variância do vetor de variáveis respostas facilmente. A normalidade nos erros e nos efeitos aleatórios é tradicional de se assumir por conta de ser uma distribuição muito conhecida e de possuir propriedades desejáveis. Além disso, temos um domínio algébrico sobre essa distribuição.

Aplicando a esperança em ambos os lados de (2.1) temos

$$E(\mathbf{Y}_i) = E(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i) \Rightarrow E(\mathbf{Y}_i) = E(\mathbf{X}_i\boldsymbol{\beta}) + E(\mathbf{Z}_i\mathbf{b}_i) + E(\boldsymbol{\epsilon}_i).$$

As matrizes são conhecidas por serem matrizes de planejamento e $\boldsymbol{\beta}$ são efeitos fixos, logo esses termos são constantes e o valor esperado do vetor de variáveis respostas é dado

por

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i E(\mathbf{b}_i) + E(\boldsymbol{\epsilon}_i).$$

Assumimos em (2.1) que $\mathbf{b}_i \sim Normal_q(\mathbf{0}, \mathbf{G}_q)$ e $\boldsymbol{\epsilon}_i \sim Normal_{m_i}(\mathbf{0}, \mathbf{R}_i)$, logo

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}. \tag{2.3}$$

Agora, aplicando também a variância em ambos os lados de (2.1) temos

$$V(\mathbf{Y}_i) = V(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i).$$

Como foi assumido em (2.1) que $\mathbf{b}_i$ e $\boldsymbol{\epsilon}_i$ são independentes, logo é possível aplicar propriedades da variância como

$$V(\mathbf{Y}_i) = V(\mathbf{X}_i\boldsymbol{\beta}) + V(\mathbf{Z}_i\mathbf{b}_i) + V(\boldsymbol{\epsilon}_i).$$

Assim como na esperança, $\mathbf{X}_i$ e $\boldsymbol{\beta}$ são fixos, portanto não variam e

$$V(\mathbf{Y}_i) = V(\mathbf{Z}_i\mathbf{b}_i) + V(\boldsymbol{\epsilon}_i).$$

Como $\mathbf{Z}_i$ também não se trata de um elemento aleatório,

$$V(\mathbf{Y}_i) = \mathbf{Z}_i V(\mathbf{b}_i)\mathbf{Z}_i^\top + V(\boldsymbol{\epsilon}_i).$$

Assumindo em (2.1) que $\mathbf{b}_i \sim Normal_q(\mathbf{0}, \mathbf{G}_q)$ e $\boldsymbol{\epsilon}_i \sim Normal_{m_i}(\mathbf{0}, \mathbf{R}_i)$ e se $\mathbf{R}_i = \sigma^2\mathbf{I}_{m_i}$, logo

$$V(\mathbf{Y}_i) = \boldsymbol{\Omega}_i = \mathbf{Z}_i\mathbf{G}_q\mathbf{Z}_i^\top + \sigma^2\mathbf{I}_{m_i}. \tag{2.4}$$

Através de (2.3) e (2.4) pode-se notar o que foi discutido anteriormente. O impacto na média (ou esperança) da variável resposta causado pelos efeitos fixos, já os efeitos aleatórios influenciam diretamente na variância das observações realizadas dentro da mesma unidade amostral (no nosso caso, cada perfuração onde a resistência foi medida).

Em termos matriciais, temos a variância marginal do vetor de variáveis respostas da seguinte forma:

$$V\left(\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ Y_{im_i} \end{bmatrix}\right) = \mathbf{Z}_i \begin{bmatrix} V(b_{i1}) & Cov(b_{i1}, b_{i2}) & . & . & . & Cov(b_{i1}, b_{iq}) \\ Cov(b_{i2}, b_{i1}) & V(b_{i2}) & . & . & . & Cov(b_{i2}, b_{iq}) \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ Cov(b_{iq}, b_{i1}) & Cov(b_{iq}, b_{i2}) & . & . & . & V(b_{iq}) \end{bmatrix} \mathbf{Z}_i^\top + \begin{bmatrix} \sigma^2 & 0 & . & . & . & 0 \\ 0 & \sigma^2 & . & . & . & 0 \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 0 & 0 & . & . & . & \sigma^2 \end{bmatrix}. \tag{2.5}$$

Como foram coletadas medições em 8 camadas de profundidade diferentes de solo, espera-se ter uma covariância entre as resistências observadas nessas camadas. Em termos do problema em questão $m_i = 8$ e o q, número de efeitos aleatórios, será definido mediante análise descritiva dos dados.

Neste trabalho, será necessário estimar a matriz $\mathbf{G}_q$, ou seja, estimar as variâncias e covariâncias entre os efeitos aleatórios. Existem várias estruturas possíveis para $\mathbf{G}_q$ visando estimar um modelo mais adequado mesmo com uma grande quantidade de parâmetros. Ao longo do trabalho será apresentado melhor as estruturas de $\mathbf{G}_q$, suas vantagens e desvantagens. Para simplificar a notação, chamaremos $\mathbf{G}_q$ de $\mathbf{G}$. Estruturas de covariância diferentes da homocedástica e independente também podem ser assumidas para os erros aleatórios.

Como dito anteriormente, (2.3) e (2.4) se referem às esperanças e variâncias marginais de $\mathbf{Y}_i$, respectivamente. Em (2.4), podemos notar que a variância é uma soma de dois termos. O primeiro modela a dispersão dos perfis individuais de resposta em torno de um perfil médio definido pela parte fixa do modelo (Singer *et al.*, 2018). Singer *et al.* (2018) também diz que o segundo termo da soma é relacionado com a dispersão dos valores observados em torno dos perfis individuais, isto é, a variabilidade das observações dentro de cada perfil.

Além da distribuição marginal, podemos escrever também a distribuição condicional de $\mathbf{Y}_i$. Se $\mathbf{b}_i$, o vetor dos efeitos aleatórios, for conhecido, teremos uma distribuição condicional de $\mathbf{Y}_i$ por $\mathbf{b}_i$. Fixar este vetor nos implica em esperanças e variâncias diferentes das vistas na distribuição marginal.

Aplicando a esperança condicional em ambos os lados de (2.1) temos:

$$E(\mathbf{Y}_i|\mathbf{b}_i) = E(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i|\mathbf{b}_i)$$

e aplicando propriedades de esperança temos:

$$E(\mathbf{Y}_i|\mathbf{b}_i) = E(\mathbf{X}_i\boldsymbol{\beta}|\mathbf{b}_i) + E(\mathbf{Z}_i\mathbf{b}_i|\mathbf{b}_i) + E(\boldsymbol{\epsilon}_i|\mathbf{b}_i).$$

As matrizes são conhecidas por serem matrizes de planejamento, $\boldsymbol{\beta}$ é o vetor de efeitos fixos e como $\mathbf{b}_i$ é conhecido, logo estes termos são constantes e o valor esperado

$$E(\mathbf{Y}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + E(\boldsymbol{\epsilon}_i|\mathbf{b}_i).$$

Assumimos em (2.1) que $\boldsymbol{\epsilon}_i \sim Normal_{m_i}(\mathbf{0}, \mathbf{R}_i)$, logo

$$E(\mathbf{Y}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i. \tag{2.6}$$

Agora, aplicando também a variância em ambos os lados de (2.1) temos:

$$V(\mathbf{Y}_i|\mathbf{b}_i) = V(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i|\mathbf{b}_i).$$

Como foi assumido em (2.1) que $\mathbf{b}_i$ e $\boldsymbol{\epsilon}_i$ são independentes, temos

$$V(\mathbf{Y}_i|\mathbf{b}_i) = V(\mathbf{X}_i\boldsymbol{\beta}|\mathbf{b}_i) + V(\mathbf{Z}_i\mathbf{b}_i|\mathbf{b}_i) + V(\boldsymbol{\epsilon}_i).$$

Como $\mathbf{X}_i$, $\boldsymbol{\beta}$, $\mathbf{Z}_i$ e $\mathbf{b}_i$ são fixos, portanto não variam,

$$V(\mathbf{Y}_i|\mathbf{b}_i) = V(\boldsymbol{\epsilon}_i).$$

Assumindo em (2.1) que $\boldsymbol{\epsilon}_i \sim Normal_{m_i}(\mathbf{0}, \mathbf{R}_i)$ e se $\mathbf{R}_i = \sigma^2\mathbf{I}_{m_i}$, logo

$$V(\mathbf{Y}_i|\mathbf{b}_i) = \sigma^2\mathbf{I}_{m_i}. \tag{2.7}$$

A apresentação da distribuição de $\mathbf{Y}_i$ condicionada em $\mathbf{b}_i$ é importante, pois Singer *et al.* (2018) mostra que o modelo (2.1) tem a característica de ser também um modelo linear em dois estágios. O primeiro estágio consiste em fixar $\mathbf{b}_i$, implicando em

$$\mathbf{Y}_i|\mathbf{b}_i \sim Normal_{m_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i). \tag{2.8}$$

Já o segundo estágio é o modelo para a distribuição marginal de $\mathbf{Y}_i$ com $\mathbf{b}_i$ independentes dado por

$$\mathbf{Y}_i \sim Normal_{m_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^\top + \mathbf{R}_i). \tag{2.9}$$

No primeiro passo, fixamos o vetor de efeitos aleatórios $\mathbf{b}_i$ para estimarmos $\mathbf{R}_i$ e $\boldsymbol{\beta}$. Na sequência, vamos para a segunda etapa e estimamos a matriz $\mathbf{G}$. Após isso, voltamos para o primeiro passo, mas agora com uma nova $\mathbf{G}$ na variância do vetor $\mathbf{b}_i$ e novamente atualizamos os valores de $\mathbf{R}_i$ e $\boldsymbol{\beta}$. Em seguida, outra vez na segunda etapa, atualizamos $\mathbf{G}$ e repetimos essas atualizações até que ocorra a convergência do processo iterativo da estimação de $\mathbf{R}_i$, $\boldsymbol{\beta}$ e $\mathbf{G}$.

Outro ponto a se destacar é na escolha dos efeitos fixos e aleatórios. Zuanetti (2022) destaca que essa decisão não é simples, nem única e depende de outros fatores, tais como o tipo de estudo, objetivo do pesquisador e contexto dos dados. Alguns itens para se destacar são:

- Os efeitos que parecem ser constantes para toda a população são considerados fixos e aqueles que devem variar entre as unidades amostrais são aleatórios;
- Colocar muitos efeitos aleatórios no modelo faz com que aumente a dificuldade para estimá-los. Por conta disso, muitos estudos consideram apenas o intercepto como efeito aleatório;
- É interessante ajustar diversos modelos e, via critério de seleção, escolher aquele que parece ser mais adequado para os dados trabalhados; e
- Uma covariável pode apresentar efeitos tanto fixos quanto aleatórios simultaneamente. Assim, em geral, atribuímos efeitos fixos a todas as covariáveis, enquanto algumas delas são designadas para ter efeitos aleatórios.

2.2 Modelo completo

Vimos anteriormente no Modelo (2.1), o modelo misto para as medidas associadas a apenas uma das perfurações (unidade amostral) feitas no solo com diferentes profundidades observadas. Singer *et al.* (2018) apresenta o modelo compactado, dessa vez com todas as perfurações em que foram coletadas as amostras do estudo e também com a suposição de que $\mathbf{b}$ é independente de $\boldsymbol{\epsilon}$. O modelo completo é definido por

$$\underset{N\times 1}{\mathbf{Y}} = \underset{N\times p}{\mathbf{X}} \times \underset{p\times 1}{\boldsymbol{\beta}} + \underset{N\times nq}{\mathbf{Z}} \times \underset{nq\times 1}{\mathbf{b}} + \underset{N\times 1}{\boldsymbol{\epsilon}}, \tag{2.10}$$

em que

$\mathbf{Y} = \left[\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_n^\top\right]^\top$ com dimensão $(N \times 1)$ em que $N = \sum_{i=1}^{n} m_i$ contém as respostas das n unidades amostrais;

$\boldsymbol{\beta}$ é o vetor de efeitos fixos do modelo com dimensão $(p \times 1)$;

$\mathbf{X} = \left[\mathbf{X}_1, \ldots, \mathbf{X}_n\right]^\top$, com dimensão $(N \times p)$, é a matriz de planejamento associada a $\boldsymbol{\beta}$ e contém as covariáveis associadas a efeitos fixos;

$\mathbf{b} = \left[\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top\right]^\top$, com dimensão $(nq \times 1)$, é o vetor dos efeitos aleatórios associados as n unidades amostrais, $\mathbf{b} \sim Normal_{nq}[\mathbf{0}, \boldsymbol{\Gamma}(\boldsymbol{\theta})]$ em que $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathbf{I}_n \otimes \mathbf{G}(\boldsymbol{\theta})$, sendo $\boldsymbol{\theta}$ o vetor que contém todos os parâmetros de variância e covariância associados a esse modelo e que dependem da especificação da matriz $\mathbf{G}$ e $\mathbf{R}_i$, em que $\otimes$ representa o produto de Kronecher, melhor descrito no Apêndice A;

$\mathbf{Z} = \oplus_{i=i}^{n} \mathbf{Z}_i$, com dimensão $(N \times nq)$, é a matriz de planejamento associada a $\mathbf{b}$ e contém as covariáveis associadas a efeitos aleatórios, em que $\oplus$ representa a soma direta de matrizes, melhor descrito no Apêndice A; e

$\boldsymbol{\epsilon} = \left[\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_n^\top\right]^\top$, com dimensão $(N \times 1)$, é o vetor dos erros aleatórios das n unidades amostrais com dimensão $(m_i \times 1)$, $\boldsymbol{\epsilon} \sim Normal_N[\mathbf{0}, \mathbf{R}(\boldsymbol{\theta})]$ com $\mathbf{R}(\boldsymbol{\theta}) = \oplus_{i=1}^{n} \mathbf{R}_i(\boldsymbol{\theta})$.

Assim como fizemos para o modelo (2.1), podemos escrever também o (2.10) na forma matricial,

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ . \\ . \\ . \\ \mathbf{Y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ . \\ . \\ . \\ \mathbf{X}_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & . & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & . & . & . & \mathbf{0} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \mathbf{0} & \mathbf{0} & . & . & . & \mathbf{Z}_n \end{bmatrix} \begin{bmatrix} \begin{pmatrix} b_{11} \\ . \\ b_{1q} \end{pmatrix} \\ . \\ . \\ . \\ \begin{pmatrix} b_{n1} \\ . \\ b_{nq} \end{pmatrix} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ . \\ . \\ . \\ \boldsymbol{\epsilon}_n \end{bmatrix}. \tag{2.11}$$

Geralmente, adotamos a independência entre $\mathbf{b}$ e $\boldsymbol{\epsilon}_i$ para obtermos a média e variância do vetor de variáveis respostas, assim como no caso para uma única unidade amostral. A escolha da distribuição Normal nos erros e nos efeitos aleatórios se dá pelo mesmo

motivo visto no modelo anterior: possuir propriedades desejáveis, ser muito conhecida e por termos um entendimento algébrico sobre ela.

Em consequência de (2.10), temos $\mathbf{Y} \sim Normal_N[\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega}(\boldsymbol{\theta})]$ em que

$$\boldsymbol{\Omega}(\boldsymbol{\theta}) = \mathbf{Z}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{Z}^\top + \mathbf{R}(\boldsymbol{\theta}).$$

No final dessa seção, apresentamos em mais detalhes a estrutura dessas matrizes.

Na Tabela 2.1, apresentamos um resumo das dimensões dos componentes do modelo na sua formulação individual e completa com o intuito de ajudar nas aplicações.

Tabela 2.1: Componentes e dimensões para as formulações do modelo misto.

Formulação individual		**Formulação compactada**	
Componente	Dimensão	Componente	Dimensão
$\mathbf{Y}_i$	$(m_i \times 1)$	$\mathbf{Y}$	$(N \times 1)$
$\mathbf{X}_i$	$(m_i \times p)$	$\mathbf{X}$	$(N \times p)$
$\boldsymbol{\beta}$	$(p \times 1)$	$\boldsymbol{\beta}$	$(p \times 1)$
$\mathbf{Z}_i$	$(m_i \times q)$	$\mathbf{Z}$	$(N \times nq)$
$\mathbf{b}_i$	$(q \times 1)$	$\mathbf{b}$	$(nq \times 1)$
$\boldsymbol{\epsilon}_i$	$(m_i \times 1)$	$\boldsymbol{\epsilon}$	$(N \times 1)$
$\mathbf{G}$	$(q \times q)$	$\boldsymbol{\Gamma}$	$(nq \times nq)$
$\mathbf{R}_i$	$(m_i \times m_i)$	$\mathbf{R}$	$(N \times N)$
$\boldsymbol{\Omega}_i$	$(m_i \times m_i)$	$\boldsymbol{\Omega}$	$(N \times N)$

2.3 Estruturas de covariância

Uma das grandes dificuldades em se modelar dados com medidas repetidas é definir a estrutura mais adequada da covariância entre os efeitos e erros aleatórios. De modo geral, essa definição deve depender do conhecimento sobre o fenômeno físico em análise e da maneira pela qual as observações foram obtidas (Singer *et al.*, 2018).

No caso de modelos mistos, foi visto anteriormente que $\mathbf{R}_i$ modela a dispersão da resposta em torno dos perfis individuais, ou seja, a variância das observações em cada perfil e a matriz $\mathbf{G}$ modela a dispersão entre os perfis individuais, isto é, a variância entre essas unidades amostrais. Logo, a matriz $\mathbf{R}_i$ combinada com a matriz $\mathbf{G}$ associadas aos

efeitos aleatórios $\mathbf{b}_i$ representam a estrutura de covariância como um todo.

Singer *et al.* (2018) diz que existem diversas estruturas de covariâncias na literatura possíveis de se utilizar. Essas podem ser empregadas para as componentes $\mathbf{R}_i$ e $\mathbf{G}$ quanto diretamente para $V(\mathbf{Y}_i)$. A seguir apresentamos algumas estruturas utilizando dimensão de 4 como exemplo:

1. **Estrutura Uniforme** $[\boldsymbol{\theta} = (\sigma^2, \tau)^\top]$:

$$\mathbf{V}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^2+\tau & \tau & \tau & \tau \\ \tau & \sigma^2+\tau & \tau & \tau \\ \tau & \tau & \sigma^2+\tau & \tau \\ \tau & \tau & \tau & \sigma^2+\tau \end{bmatrix}, \text{ em que } \sigma^2 > 0 \text{ e } \tau \in \mathbb{R}\ ; \tag{2.12}$$

2. **Estrutura AR(1)** $[\boldsymbol{\theta} = (\sigma^2, \phi)^\top]$:

$$\mathbf{V}(\boldsymbol{\theta}) = \sigma^2 \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{bmatrix}, \text{ em que } \sigma^2 > 0 \text{ e } \phi \in [-1, 1]; \tag{2.13}$$

3. **Estrutura ARMA(1,1)** $[\boldsymbol{\theta} = (\sigma^2, \gamma, \phi)^\top]$:

$$\mathbf{V}(\boldsymbol{\theta}) = \sigma^2 \begin{bmatrix} 1 & \gamma & \gamma\phi & \gamma\phi^2 \\ \gamma & 1 & \gamma & \gamma\phi \\ \gamma\phi & \gamma & 1 & \gamma \\ \gamma\phi^2 & \gamma\phi & \gamma & 1 \end{bmatrix}, \text{ em que } \sigma^2 > 0,\ \gamma \in [-1, 1] \text{ e } \phi \in [-1, 1]; \tag{2.14}$$

4. **Estrutura Toeplitz** $[\boldsymbol{\theta} = (\sigma^2, \sigma_1, \sigma_2, \sigma_3)^\top]$:

$$\mathbf{V}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}, \text{ em que } \sigma^2 > 0 \text{ e } \sigma_i \in \mathbb{R}, \text{ para } i = 1, 2, 3; \tag{2.15}$$

5. **Estrutura de Markov ou espacial** $[\boldsymbol{\theta} = (\sigma^2, \rho)^\top]$:

$$\mathbf{V}(\boldsymbol{\theta}) = \sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{bmatrix}, \text{ em que } \sigma^2 > 0,\ \rho \in [-1, 1] \text{ e } d_{i,j} \in \mathbb{Z}^*, \text{ para } i = 1, ..., 4 \text{ e } j = 1, ..., 4; \tag{2.16}$$

6. **Não-Estruturada** $[\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_{12}, \sigma_{13}, \sigma_{14}, \sigma_{23}, \sigma_{24}, \sigma_{34})^\top]$:

$$\mathbf{V}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}, \text{ em que } \sigma_i^2 > 0 \text{ e } \sigma_{i,j} \in \mathbb{R}, \text{ para } i = 1, ..., 4 \text{ e } j = 1, ..., 4. \tag{2.17}$$

Observe que estruturas mais flexíveis apresentam um número maior de parâmetros a serem estimados, enquanto estruturas mais simples e, ao mesmo tempo, com comportamentos fixos apresentam um menor número de parâmetros. Ao longo do trabalho será estudado melhor à respeito de qual estrutura é mais adequada aos dados analisados.

2.4 Estimação pelo método da máxima verossimilhança

Com o modelo apresentado em (2.10), precisamos então estimar os parâmetros associados a ele. Na literatura estatística, temos diversos métodos de estimação, um desses métodos é o de máxima verossimilhança.

Para aplicarmos o método da MV (máxima verossimilhança) no modelo misto, Singer *et al.* (2018) diz que é necessário supor que $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ e $\mathbf{R}(\boldsymbol{\theta})$ tenham estruturas conhecidas. Com essas suposições e utilizando o modelo (2.10) podemos encontrar os estimadores de $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$. A função de verossimilhança é definida como

$$L(\boldsymbol{\beta};\boldsymbol{\theta}) = (2\pi)^{-N/2}|\boldsymbol{\Omega}(\boldsymbol{\theta})|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}. \tag{2.18}$$

Aplicando o logaritmo temos

$$l(\boldsymbol{\beta};\boldsymbol{\theta}) = -\frac{N}{2}log2\pi - \frac{1}{2}log|\boldsymbol{\Omega}(\boldsymbol{\theta})| - \frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\top}\left[\boldsymbol{\Omega}(\boldsymbol{\theta})\right]^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)$$

ou também

$$l(\boldsymbol{\beta};\boldsymbol{\theta}) = -\frac{1}{2}log|\boldsymbol{\Omega}(\boldsymbol{\theta})| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}\right)^{\top}\left[\boldsymbol{\Omega}_i(\boldsymbol{\theta})\right]^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}\right). \tag{2.19}$$

Agora, assumindo $\boldsymbol{\theta}$ conhecido e aplicando $\partial l(\boldsymbol{\beta};\boldsymbol{\theta})/\partial\boldsymbol{\beta} = 0$ em (2.19), obtemos o estimador de $\boldsymbol{\beta}$ dado por

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left[\sum_{i=1}^{n}\mathbf{X}_i^{\top}[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}\mathbf{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{X}_i^{\top}[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}\mathbf{y}_i\right]. \tag{2.20}$$

Para mostrar que (2.20) é ponto de máximo, aplicamos

$$\frac{\partial^2 l(\boldsymbol{\beta};\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\top}} = -\sum_{i=1}^{n}\mathbf{X}_i^{\top}[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}\mathbf{X}_i$$

e como temos uma matriz definida negativa, $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ é o ponto de máximo para $l(\boldsymbol{\beta};\boldsymbol{\theta})$. Ou seja, quando temos $\boldsymbol{\Omega}_i(\boldsymbol{\theta})$ conhecido, (2.20) corresponde ao estimador de $\boldsymbol{\beta}$.

Para obtermos o estimador de $\boldsymbol{\theta}$, substituímos $\hat{\boldsymbol{\beta}}$ em (2.19) e assim temos a função log-verossimilhança perfilada $l[\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta}]$. Aplicando $\partial l(\hat{\boldsymbol{\beta}};\boldsymbol{\theta})/\partial\boldsymbol{\theta} = 0$, obtemos o estimador de $\boldsymbol{\theta}$ dado por

$$\hat{\boldsymbol{\theta}} = -\frac{1}{2}\sum_{i=1}^{n}tr\left[[\boldsymbol{\Omega}_i(\hat{\boldsymbol{\theta}})]^{-1}\dot{\boldsymbol{\Omega}}_i(\hat{\boldsymbol{\theta}})\right] - \frac{1}{2}\sum_{i=1}^{n}tr\left[\partial Q_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_j|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right], \tag{2.21}$$

$j = 1,\ldots,t$, em que t é o número de parâmetros em $\boldsymbol{\theta}$,

$$\dot{\boldsymbol{\Omega}}_i(\hat{\boldsymbol{\theta}}) = [\partial \boldsymbol{\Omega}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}_j]^{\top}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

e

$$Q_i(\boldsymbol{\theta}) = [\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]^{\top}[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}[\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})].$$

Singer *et al.* (2018) mostra com mais detalhes que (2.21) é um ponto de máximo através da aplicação de $\partial^2 l(\boldsymbol{\beta};\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}$. Logo em (2.21), temos o estimador que maximiza a função $l(\boldsymbol{\beta};\boldsymbol{\theta})$ para $\boldsymbol{\theta}$.

Foi dito anteriormente que substituímos $\hat{\boldsymbol{\beta}}$ em $\boldsymbol{\beta}$ na Equação (2.19) para encontrarmos o estimador de MV para $\boldsymbol{\theta}$. O mesmo ocorre para o caso contrário, ou seja, se substituirmos $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$ em (2.19), obtemos o estimador MV de $\boldsymbol{\beta}$. Por conta disso, podemos dizer que o modelo misto possui um processo iterativo na estimação de seus parâmetros, em que primeiro estimamos $\boldsymbol{\beta}$, na sequência $\boldsymbol{\theta}$ e repetimos esse processo até a convergência de seus valores. Logo, para estimarmos um modelo misto pelo método de máxima verossimilhança, precisamos:

1. Escolher a estrutura das duas matrizes de variâncias e covariâncias específicas ou da combinada;
2. Atribuir valores iniciais para os parâmetros da estrutura;
3. Estimar $\boldsymbol{\beta}$;
4. Estimar $\boldsymbol{\theta}$; e
5. Repetir os passos 3 e 4 até que os parâmetros convirjam para determinados valores estimados.

De modo geral, na estatística é mais difícil de conseguirmos a convergência de parâmetros de variância e covariância. Por conta disso, pode ser custoso obtermos a convergência de $\boldsymbol{\theta}$ ou, em alguns casos, não conseguimos sua convergência devido a complexidade da estrutura de covariância. Afim de obter um processo de estimação mais eficaz para $\boldsymbol{\theta}$, muitos autores recomendam a utilização do método de máxima verossimilhança restrita (MVR), proposto por Patterson e Thompson (1971). O método consiste em maximizar a verossimilhança de uma transformação linear ortogonal dos dados com o objetivo de estimar termos da variância.

A transformação utilizada é $\mathbf{Y}^{\dagger} = \mathbf{U}^{\top}\mathbf{Y}$ em que, no geral, usamos $\mathbf{U} = \mathbf{I} - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}$ e com dimensão $(N \times N)$ para se obter a verossimilhança restrita. Além disso, temos $E(\mathbf{Y}^{\dagger}) = \mathbf{0}$ e $\mathbf{U}^{\top}\mathbf{X} = \mathbf{0}$. De fato,

$$E(\mathbf{Y}^{\dagger}) = E(\mathbf{U}^{\top}\mathbf{Y}) = E(\mathbf{U}^{\top}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} + \boldsymbol{\epsilon}]) = E(\mathbf{U}^{\top}\mathbf{X}\boldsymbol{\beta} + \mathbf{U}^{\top}\mathbf{Z}\boldsymbol{b} + \mathbf{U}^{\top}\boldsymbol{\epsilon})$$

$$= E(\mathbf{U}^{\top}\mathbf{X}\boldsymbol{\beta}) + E(\mathbf{U}^{\top}\mathbf{Z}\boldsymbol{b}) + E(\mathbf{U}^{\top}\boldsymbol{\epsilon}) = \mathbf{U}^{\top}\mathbf{X}\boldsymbol{\beta} + \mathbf{U}^{\top}\mathbf{Z}E(\boldsymbol{b}) + \mathbf{U}^{\top}E(\boldsymbol{\epsilon}).$$

Assumimos em (2.10) que $\mathbf{b} \sim Normal_{nq}(\mathbf{0}, \boldsymbol{\Gamma}(\boldsymbol{\theta}))$ e $\boldsymbol{\epsilon} \sim Normal_{N}(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}))$, logo

$$E(\mathbf{Y}^{\dagger}) = \mathbf{U}^{\top}\mathbf{X}\boldsymbol{\beta}.$$

Substituindo $\mathbf{U}$, temos

$$E(\mathbf{Y}^{\dagger}) = (\mathbf{I} - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top})^{\top}\mathbf{X}\boldsymbol{\beta}.$$

Transpondo $\mathbf{U}$, temos $\mathbf{U} = \mathbf{U}^{\top}$, logo

$$E(\mathbf{Y}^{\dagger}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}. \tag{2.22}$$

Já a segunda expressão pode ser obtida por

$$\mathbf{U}^{\top}\mathbf{X} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top})\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{I} = \mathbf{0}. \tag{2.23}$$

Também temos $V(\mathbf{Y}^{\dagger}) = \mathbf{U}\boldsymbol{\Omega}(\boldsymbol{\theta})\mathbf{U}^{\top}$, pois

$$V(\mathbf{Y}^{\dagger}) = V(\mathbf{U}^{\top}\mathbf{Y}) = V(\mathbf{U}^{\top}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} + \boldsymbol{\epsilon}]).$$

Assumindo em (2.10) que $\mathbf{b}$ e $\boldsymbol{\epsilon}$ são independentes,

$$V(\mathbf{Y}^{\dagger}) = V(\mathbf{U}^{\top}\mathbf{X}\boldsymbol{\beta}) + V(\mathbf{U}^{\top}\mathbf{Z}\boldsymbol{b}) + V(\mathbf{U}^{\top}\boldsymbol{\epsilon}) = (\mathbf{U}^{\top}\mathbf{Z})V(\boldsymbol{b})(\mathbf{Z}^{\top}\mathbf{U}) + \mathbf{U}^{\top}V(\boldsymbol{\epsilon})\mathbf{U}.$$

Assumindo em (2.10) que $\mathbf{b} \sim Normal_{nq}(\mathbf{0}, \boldsymbol{\Gamma}(\boldsymbol{\theta}))$ e $\boldsymbol{\epsilon} \sim Normal_{N}(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}))$,

$$V(\mathbf{Y}^{\dagger}) = (\mathbf{U}^{\top}\mathbf{Z})\boldsymbol{\Gamma}(\boldsymbol{\theta})(\mathbf{Z}^{\top}\mathbf{U}) + \mathbf{U}^{\top}\mathbf{R}(\boldsymbol{\theta})\mathbf{U}. \tag{2.24}$$

Podemos simplificar (2.24) deixando $\mathbf{U}^{\top}$ e $\mathbf{U}$ em evidência. Logo,

$$V(\mathbf{Y}^{\dagger}) = \mathbf{U}^{\top}(\mathbf{Z}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{Z}^{\top} + \mathbf{R}(\boldsymbol{\theta}))\mathbf{U} = \mathbf{U}^{\top}\boldsymbol{\Omega}(\boldsymbol{\theta})\mathbf{U}. \tag{2.25}$$

Por consequência, temos $\mathbf{Y}^{\dagger} \sim Normal_N[\mathbf{0}, \mathbf{U}^{\top}\boldsymbol{\Omega}(\boldsymbol{\theta})\mathbf{U}]$, ou seja, $\mathbf{Y}^{\dagger}$ não depende de $\boldsymbol{\beta}$.

Patterson e Thompson (1971) provaram que quando utilizamos a transformação $\mathbf{Y}^{\dagger}$ ao invés de $\mathbf{Y}$, nenhuma informação de $\boldsymbol{\theta}$ é perdida quando há ausência de $\boldsymbol{\beta}$. Isso quer dizer que utilizando uma verossimilhança restrita, nós podemos estimar diretamente $\boldsymbol{\theta}$. Com isso, como iremos estimar apenas $\boldsymbol{\theta}$, a convergência desse estimador fica mais facilitada. O processo de estimação por máxima verossimilhança restrita pode ser definido como:

1. Escolher a estrutura da(s) matriz(es) de variância e covariância;
2. Estimar $\boldsymbol{\theta}$ utilizando MVR (o texto de Zhang (2015) traz mais detalhes de como esse processo é realizado);
3. Estimar $\boldsymbol{\beta}$ utilizando MV com base no $\boldsymbol{\theta}$ estimado no passo 2.

Singer *et al.* (2018) diz que é possível também obter os preditores dos efeitos aleatórios por meio da distribuição conjunta de $\mathbf{Y}$ e $\mathbf{b}$ dada por

$$f(\mathbf{y}, \mathbf{b}) = f(\mathbf{y}|\mathbf{b})f(\mathbf{b}), \tag{2.26}$$

em que $\mathbf{y}|\mathbf{b} \sim Normal_N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R}(\boldsymbol{\theta}))$ e $\mathbf{b} \sim Normal_{nq}(\mathbf{0}, \boldsymbol{\Gamma}(\boldsymbol{\theta}))$.

O estimador de $\boldsymbol{\beta}$ (como já vimos antes) é dado por

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\mathbf{X}^{\top}[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}\mathbf{X}]^{-1}\mathbf{X}^{\top}[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}\mathbf{y} \tag{2.27}$$

e

$$\hat{\mathbf{b}}(\boldsymbol{\theta}) = \boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{Z}^{\top}[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})] = \boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{Z}^{\top}\mathbf{Q}(\boldsymbol{\theta})\mathbf{y} \tag{2.28}$$

em que

$$\mathbf{Q}(\boldsymbol{\theta}) = [\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1} - [\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}\mathbf{X}[\mathbf{X}^{\top}[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}\mathbf{X}]^{-1}\mathbf{X}^{\top}[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}. \tag{2.29}$$

No software R, um modelo misto pode ser estimado por MV ou MVR usando o pacote *nlme* de Pinheiro *et al.* (2023). Geralmente, se a convergência for alcançada, os métodos apresentam soluções muito semelhantes.

2.5 Teste de significância dos efeitos fixos

Para análise de modelos lineares e modelos lineares generalizados, no geral, um dos testes mais utilizados para verificar a significância dos coeficientes de regressão (aqui dos efeitos fixos) é o teste de Wald.

O teste de Wald avalia a distância ponderada entre a estimativa do parâmetro e o valor postulado sob a hipótese nula. Quanto mais distante de 0 for o valor da distância ponderada, menor é a chance da hipótese de igualdade ser verdadeira, ou seja, do valor verdadeiro do coeficiente ser igual ao valor postulado (de Freitas, 2018).

O teste individual para os coeficientes do modelo tem como hipóteses

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \quad \text{com } j = 1, ..., p, \end{cases}$$

em que falhar na rejeição da hipótese nula é um indicativo da falta de significância do coeficiente. Já se a hipótese nula for rejeitada, o coeficiente aparenta ser significativo para o modelo em relação à amostra observada.

Montgomery *et al.* (2012) representa a estatística de Wald por

$$Z_0 = \frac{\hat{\beta}_j}{ep(\hat{\beta}_j)}, \tag{2.30}$$

em que ep representa o erro padrão associado ao estimador e $Z_0 \sim Normal(0, 1)$ de maneira exata ou assintoticamente aproximada. O valor-p associado a esse teste é utilizado para rejeitar ou não H_0.

O teste de Wald será de suma importância para as análises que serão apresentados posteriormente para definirmos quais fatores impactam na resistência mecânica média do solo. Para verificarmos a significância dos efeitos fixos, neste trabalho, utilizamos um nível de 5% de significância.

2.6 Análise de diagnóstico

Anteriormente, vimos que nos modelos (2.1) e (2.10) temos algumas suposições em relação a distribuição dos erros e dos efeitos aleatórios. Isso acontece para que certas propriedades sejam atendidas, além de termos algumas outras vantagens matemáticas.

Para verificar a adequabilidade do modelo em descrever o comportamento dos dados analisados e confiarmos nos resultados obtidos, será necessário validar essas suposições na modelagem. Nobre (2004) sugere, para analisar as suposições dos erros, padronizar os resíduos condicionais porque os resíduos ordinais $\hat{\boldsymbol{\epsilon}}$ podem ter variâncias distintas. Com essa padronização é avaliada a homocedasticidade (através do gráfico dos resíduos condicionais padronizados vs preditos) e também a presença ou não de *outliers* nas observações (através do gráfico dos resíduos condicionais padronizados vs índices das observações). Utilizando (2.10) e com base em (2.6), os resíduos condicionais podem ser definidos como

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{E}[\mathbf{y}|\mathbf{b}] = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}, \tag{2.31}$$

e com isso, Nobre e Singer (2007) definem os resíduos condicionais padronizados como

$$\hat{\epsilon}^*_{ij} = \frac{\hat{\epsilon}_{ij}}{diag_{ij}(\mathbf{RQR})^{1/2}},$$

em que $diag_{ij}(\mathbf{RQR})$ se refere à j-ésima observação da i-ésima unidade amostral e $\mathbf{Q}$ é definida na Equação (2.29).

Os preditos para a variável resposta também podem ser definidos através de (2.6), como

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}.$$

Através de (2.31) temos que os resíduos condicionais e os efeitos aleatórios podem estar confundidos. Singer *et al.* (2018) diz que isto implica que $\hat{\boldsymbol{\epsilon}}$ pode não ser adequado para avaliar a suposição de normalidade de $\boldsymbol{\epsilon}$ porque quando $\mathbf{b}$ não atende a suposição de normalidade, $\hat{\boldsymbol{\epsilon}}$ pode não apresentar comportamente Gaussiano mesmo quando $\boldsymbol{\epsilon}$ tenha distribuição gaussiana. Por conta disso, a suposição de normalidade dos erros (através do gráfico QQ-plot) pode ser verificada através dos resíduos com confundimento mínimo padronizados.

Hilden-Minton (1995) sugere utilizar uma transformação linear do $\hat{\boldsymbol{\epsilon}}$, que minimize o

confundimento. Para maiores detalhes ver Hilden-Minton (1995).

Já em relação às suposições dos efeitos aleatórios, analisaremos a normalidade desses efeitos, além de verificar a presença ou não de *outliers* em relação às unidades amostrais. Para essas análises utilizaremos o EBLUP (BLUP empírico, ou *Empirical Best Linear Unbiased Predictor*). Singer *et al.* (2018) mostra que o BLUP para o modelo misto é definido por (2.28) e seu EBLUP ocorre quando substituímos $\boldsymbol{\theta}$ por seus estimadores. Então para essas investigações, predizemos os efeitos aleatórios para cada unidade amostral e por meio de gráficos de boxplots e histogramas são feitas as análises.

Nobre (2004) mostra outros tipos de diagnósticos mais sofisticados, porém não iremos abrangê-los ao longo do trabalho. Além disso, foi utilizada uma função desenvolvida pelos autores Francisco Marcelo M. Rocha, Juvencio S. Nobre e Julio M. Singer chamada de *residdiag_nlme* com o intuito de obter os resíduos condicionais padronizados, os resíduos com confundimento mínimo padronizados e alguns gráficos.

Capítulo 3

Dados de solo e penetrômetros

O experimento através do qual os dados analisados foram extraídos foi realizado na Faculdade de Zootecnia e Engenharia de Alimentos (FZEA) da Universidade de São Paulo (USP), localizada no município paulista de Pirassununga. Cinco áreas quadradas distribuídas na PUSP-FC, com lados de 5 metros, diferentes (1 a 5) e classificadas como latossolo vermelho eutroférrico ricas em óxido férrico, altamente férteis e com histórico de cultivo foram utilizadas com o intuito de avaliar a resistência mecânica do solo à penetração em MPa (variável resposta de interesse). A Figura 3.1 ilustra a localização das 5 áreas analisadas.

Figura 3.1: Localização das áreas do experimento na PUSP-FC. (Imagem: Google Earth, 13/05/2023).

Os dados foram coletados entre outubro de 2016 e março de 2017. Nessa coleta, em cada área, foram utilizados três níveis de umidade (S - seco, U - úmido e E - encharcado) para verificar a resistência mecânica do solo em situações diferentes do dia a dia. Além disso, três tipos diferentes de penetrômetos (I - impacto, M - manual e A - automático) foram usados nas medições em 8 camadas de profundidade distantes em $5cm$ entre 0 e $40cm$ (P_a a P_h). O experimento apresenta 30 repetições por combinação de área, penetrômetro, umidade do solo e camada de profundidade, totalizando 10800 amostras (Menezes, 2018).

Primeiramente, em cada área foi realizada uma análise detalhada para diferenciar as características entre elas. As áreas foram classificadas texturalmente da seguinte maneira: 1 - Franco-argiloarenosa, 2 - Argiloarenosa, 3, 4 e 5 - Argilosa. Apesar das três últimas áreas apresentarem classificações iguais, elas possuem outras diferenças entre si. Menezes (2018) apresenta com mais detalhes em seu trabalho essas diferenças.

Em relação à umidade, foi necessária uma classificação de seus níveis para a realização do experimento, na qual as mudanças nos níveis desse fator ocorreram via precipitação natural. Essa classificação foi feita da seguinte maneira: abaixo de $0.20cm^3cm^{-3}$ o solo é considerado seco, entre $0.20cm^3cm^{-3}$ (inclusive) e $0.30cm^3cm^{-3}$ o solo possui nível úmido e acima de $0.30cm^3cm^{-3}$ (inclusive) o solo é classificado como encharcado.

Para as medições, foram utilizados três tipos diferentes de penetrômetros. O de impacto (I) é um penetrômetro dinâmico, acionado manualmente, com taxa de penetração constante, registro de dados manual e dados analógicos; o manual (M) é um penetrômetro estático, acionado manualmente, com taxa de penetração variável, registro automático dos dados e dados digitais; o automático (A) é um penetrômetro estático, acionado hidraulicamente, com taxa de penetração constante, registro automático dos dados e dados digitais.

Como dito anteriormente, o experimento totaliza 10800 observações por apresentar 30 repetições por combinação, sendo assim, um experimento balanceado. Porém, o experimento não foi balanceado, pois foram descartadas todas as réplicas com pelo menos uma camada com resistência do solo à penetração igual ou superior a 6.5 MPa e também as medições do penetrômetro de impacto com resultado igual a 0.

Com isso, no experimento temos 8824 observações ($N = 8824$). Podemos verificar, através da Tabela 3.1, que ocorre um balanceamento no número de unidades amostrais ou experimentais (n) em relação as profundidades pois todas possuem 1103 unidades

amostrais.

Tabela 3.1: Frequência das unidades amostrais por profundidade.

	Profundidades (em centímetros)							
	5	10	15	20	25	30	35	40
Unidades experimentais	1103	1103	1103	1103	1103	1103	1103	1103

Tabela 3.2: Número de observações para diferentes combinações de umidade (E, S e U), área (1 a 5) e penetrômetro (A, I e M).

(a) Umidade = E

	Penetrômetros		
Área	A	I	M
1	240	240	240
2	240	232	240
3	240	240	240
4	0	240	240
5	240	208	240

(b) Umidade = S

	Penetrômetros		
Área	A	I	M
1	240	240	240
2	176	16	224
3	184	40	240
4	144	80	136
5	216	40	240

(c) Umidade = U

	Penetrômetros		
Área	A	I	M
1	240	232	240
2	240	216	240
3	240	160	240
4	192	152	216
5	152	72	216

A Tabela 3.2 mostra a distribuição das 8824 observações por cada combinação de área, umidade e penetrômetro. Nela observamos que, com exceção da combinação área 4 e penetrômetro automático (A) na qual não tivemos observações, a umidade E (encharcado) apresenta a maior quantidade de observações. O inverso acontece para seco (S). Logo, temos um indício de que quanto maior a umidade, menos observações foram perdidas devido a pelo menos uma camada com resistência mecânica do solo à penetração igual ou superior a 6.5 MPa. Nota-se também que grande parte das amostras perdidas envolvem o penetrômetro de impacto (I) e, principalmente, em solos com baixa umidade (seco), dando indícios de que na coleta das observações por esse penetrômetro somado com o tipo de umidade gerou uma perda de informações em todas as áreas com exceção da 1 que não foi afetada.

Os fatores e covariáveis disponíveis para a análise são, então:

- Área: local onde o experimento é realizado (1 a 5);
- Umidade (em cm^3cm^{-3}): nível de quantidade de água no solo (seco, úmido ou encharcado);

- Penetrômetro: ferramenta utilizada nas medições da resistência mecânica do solo à penetração (impacto, manual ou automático); e

- Profundidade (em centímetros): nível de profundidade na qual ocorreu a medição em relação ao solo (5, 10, 15, 20, 25, 30, 35 e 40).

3.1 Análise descritiva dos dados

Inicialmente, aplicamos uma análise descritiva através do *software* R com o intuito de entender os dados experimentais sobre as variáveis do estudo. Isso foi feito por meio de boxplots para as covariáveis individualmente, e também utilizamos gráficos de linhas para algumas combinações (ou interações) entre as covariáveis. Essas combinações foram feitas duas a duas e três a três. No primeiro caso, as interações realizadas foram: área-profundidade, umidade-profundidade e penetrômetro-profundidade. A ideia é obter indicativos de como as variáveis área, umidade e penetrômetro impactam a resistência mecânica à penetração à medida que a profundidade do solo aumenta. Já no segundo caso, as interações triplas realizadas foram: área-umidade-profundidade, área-penetrômetro-profundidade e umidade-penetrômetro-profundidade. O intuito também foi de obter indicativos de como as variáveis área, umidade e penetrômetro impactam a resistência mecânica à penetração à medida que a profundidade do solo aumenta.

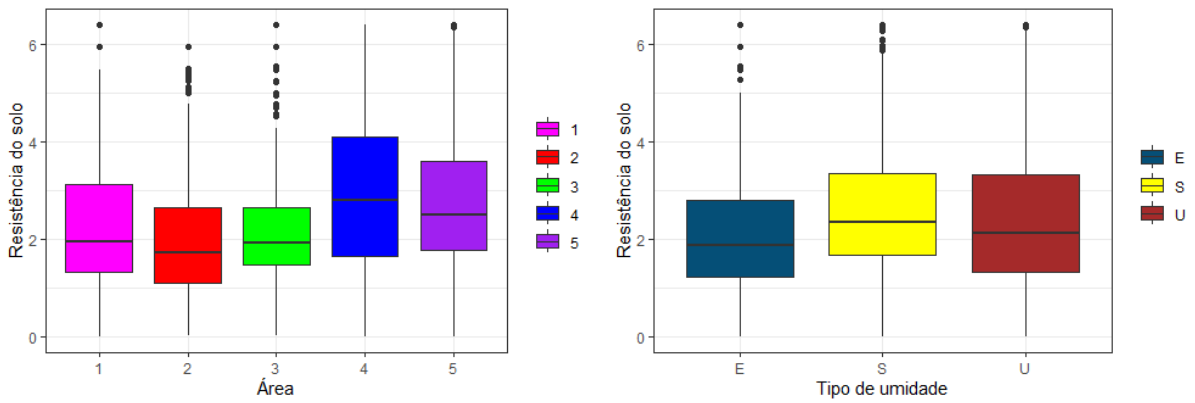

Figura 3.2: Boxplots da resistência mecânica do solo à penetração por áreas (1 a 5) e por umidades (E, S e U), respectivamente.

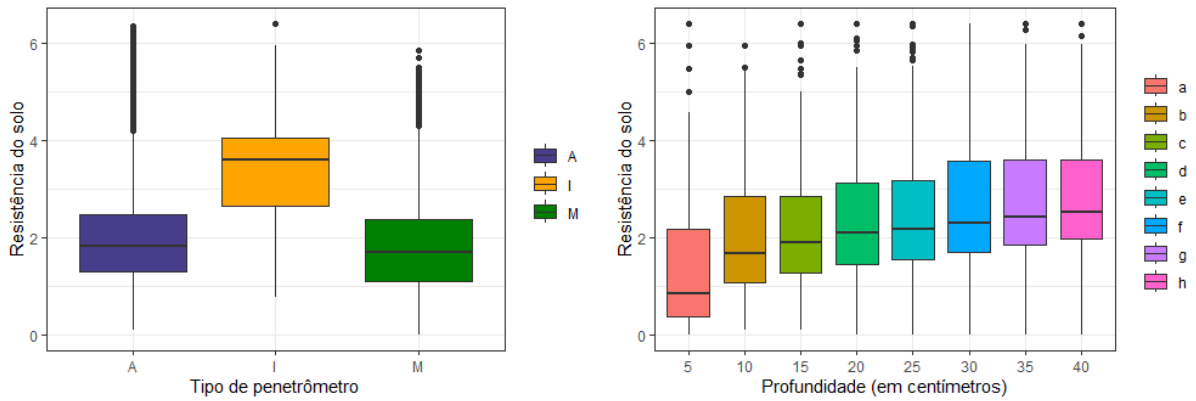

Figura 3.3: Boxplots da resistência mecânica do solo à penetração por diferentes penetrômetros (A, I e M) e por profundidades (a a h), respectivamente.

Pela Figura 3.2, temos um indicativo de que, em média, as áreas apresentam resistências mecânicas do solo à penetração iguais. É de se observar também que as áreas argilosas (3 a 5) possuem variações diferentes, sendo a primeira a com menor variação e com mais pontos discrepantes dentre todas. Em relação ao tipo de umidade, temos um indicativo de que essas categorias (E, S e U) são muito semelhantes em média, variância e até mesmo por suas medianas.

Quanto a Figura 3.3, temos um indicativo de que, em média, o penetrômetro de impacto (I) apresenta uma resistência mecânica do solo à penetração maior do que os demais. Esse comportamento pode ser fruto de uma possível técnica de manuseio ou outro comportamento diferente para esse penetrômetro. Em relação às profundidades, temos um indicativo de que essas categorias (a a h) são semelhantes, porém podemos notar que conforme a profundidade aumenta (em centímetros), a resistência mecânica do solo à penetração também aumenta em um comportamento quase logarítmico.

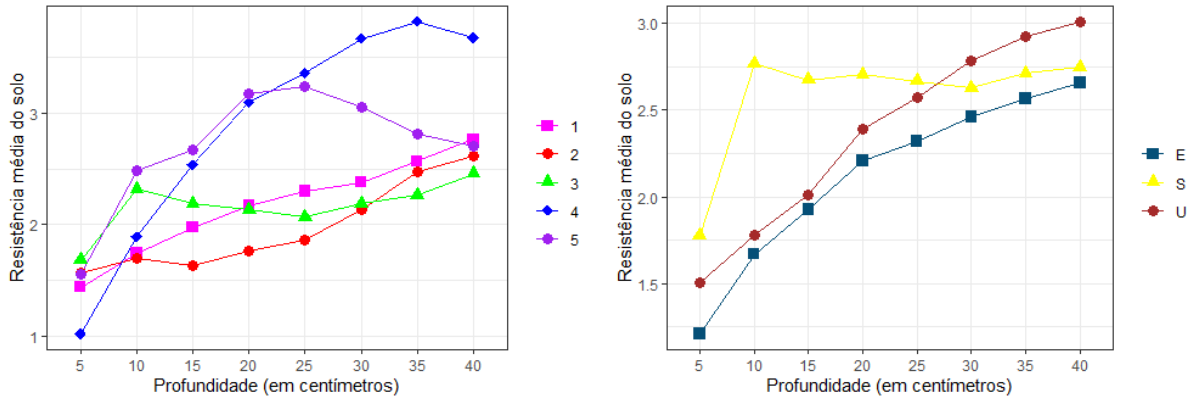

Figura 3.4: Gráfico de linhas para a resistência mecânica média do solo à penetração entre as interações duplas envolvendo as profundidades (5 a 40) com as áreas (1 a 5) e umidades (E, S e U), respectivamente.

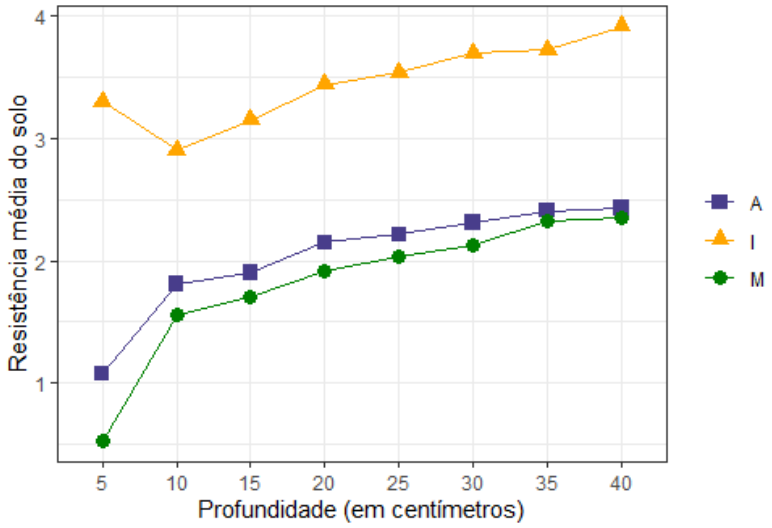

Figura 3.5: Gráfico de linhas para a resistência mecânica média do solo à penetração entre as interações duplas envolvendo as profundidades (5 a 40) com os penetrômetros (A, I e M).

Pela Figura 3.4, temos um indicativo de que, em média, as áreas 1,2 e 3 apresentam um comportamento parecido conforme a profundidade (em centímetros) aumenta e aparentemente linear ou levemente de forma logarítmica. A área 4 possui uma resistência mecânica média do solo à penetração maior em maiores profundidades e a área 5 possui resistência mecânica média maior do que 1, 2 e 3 entre 15 e 30 centímetros, a partir disso as medidas são praticamente iguais.

Ainda na Figura 3.4, temos um indicativo de que o solo seco (S) inicialmente apresenta uma resistência mecânica média bem maior que os outros tipos de umidade, porém, essa resistência mecânica média fica constante a partir de 10 centímetros de profundidade e, por conta disso, a partir dos 25 centímetros o solo seco apresenta resistência mecânica média parecida com os demais e, também, nos pontos de maior profundidade possui valores menores que o solo umido (U). Já os solos úmido (U) e encharcado (E) possuem o mesmo comportamento crescente e linear na resistência mecânica média do solo à penetração com o aumento da profundidade, sendo o solo úmido o que sempre apresenta resistência mecânica média maior.

E por último, na Figura 3.5, temos um indicativo que, em média, o penetrômetro de impacto (I) possui resistência mecânica do solo à penetração muito maior do que os outros dois em qualquer camada de profundidade. Esse comportamento já era notório pelo boxplot da esquerda na Figura 3.3. Já os penetrômetros automático e manual (A e M, respectivamente) possuem comportamentos médios semelhantes conforme aumento

de profundidade.

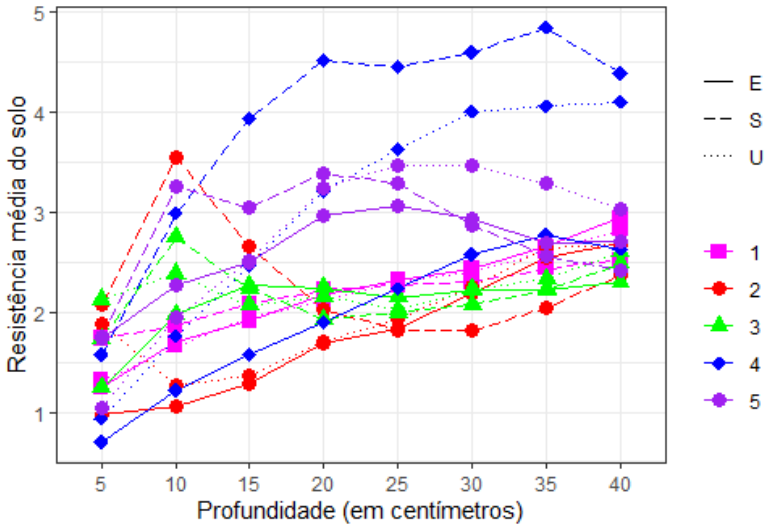

Figura 3.6: Gráfico de linhas para a resistência mecânica média do solo à penetração entre as interações triplas envolvendo as profundidades (5 a 40), áreas (1 a 5) e umidades (E, S e U).

Com a Figura 3.6, não conseguimos ter um indicativo claro quanto às interações triplas entre as variáveis. Porém pode-se destacar que, com uma profundidade de 10 centímetros, a área 2 com o solo seco (S) possui a maior resitência mecânica média do solo à penetração, mas conforme a profundidade aumenta, essa resistência mecânica média cai bastante.

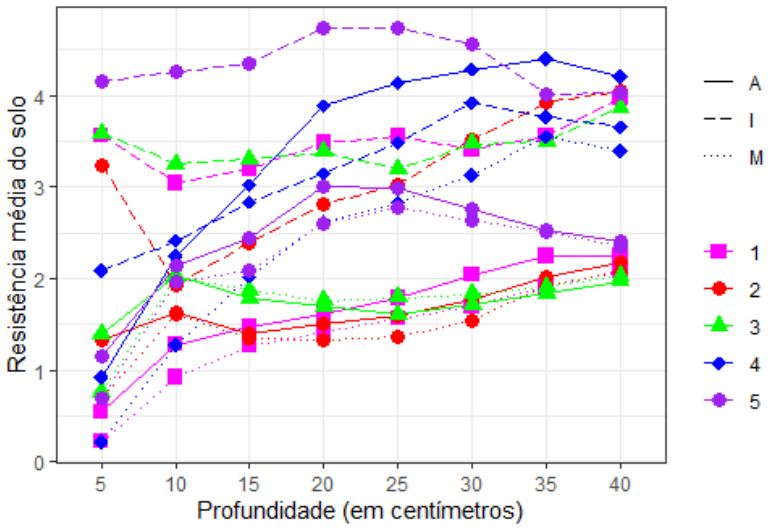

Figura 3.7: Gráfico de linhas para a resistência mecânica média do solo à penetração entre as interações triplas envolvendo as profundidades (5 a 40), áreas (1 a 5) e penetrômetros (A, I e M).

Pela Figura 3.7, não conseguimos também ter muitos indicativos no que se refere às interações triplas entre as variáveis. Porém nota-se que, o penetrômetro de impacto (I)

demonstra ser bastante influente, pois todas as áreas, com exceção da 5, apresentam um grande desequilíbrio na resistência mecânica média do solo à penetração em relação aos outros tipos de penetrômetro e suas profundidades.

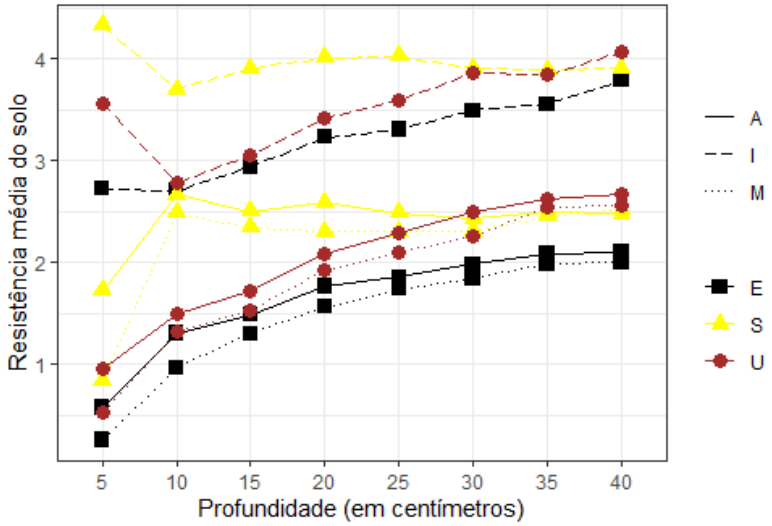

Figura 3.8: Gráfico de linhas para a resistência mecânica média do solo à penetração entre as interações triplas envolvendo as profundidades (5 a 40), umidades (E, S e U) e penetrômetros (A, I e M).

Através da Figura 3.8, temos um indicativo de que as umidades E e U (encharcado e úmido, respectivamente) quando combinadas com os tipos de penetrômetros aparentam não apresentar efeito de interação na medida em que a profundidade aumenta, pois apresentam comportamentos paralelos. Temos também que as interações envolvendo o solo seco (S) possuem uma resistência mecânica média do solo à penetração maior do que as interações com os outros tipos de umidade, porém a partir de 10 centímetros, a resistência mecânica média dessas interações com S ficam praticamente constantes.

Vale ressaltar que, exceto em alguns casos particulares tais como: área 4 e 5, solo seco e algumas interações, a resistência mecânica média parece aumentar linearmente ou de forma ligeiramente logarítmica em relação ao aumento na profundidade e, apesar de alguns cruzamentos entre curvas médias, geralmente as curvas médias são paralelas e apresentam comportamento parecido, embora tenham locação diferente. Mais a frente, vamos discutir mais sobre esses aspectos.

Portanto, concluímos que, como principais pontos, o penetrômetro de impacto (I) demonstra uma resistência maior em comparação com os demais penetrômetros. Além disso, observamos que, à medida que a profundidade aumenta em centímetros, a resistência apresenta um crescimento que segue um padrão quase logarítmico. Notamos também que

as áreas 1,2 e 3, em média, possuem um comportamento próximo e linear ou levemente logarítmico em relação à resistência conforme a profundidade aumenta. Por outro lado, as áreas 4 e 5 mostram uma resistência média maior em maiores profundidades e um comportamento não linear.

Capítulo 4

Resultados

4.1 Definição do modelo principal

Inicialmente, optamos por trabalhar com $\log(\mathbf{Y}_i)$ por conta do comportamento médio não linear das áreas 4 e 5 em relação às profundidades visto no Gráfico 1 da Figura 3.4. Porém o banco de dados conta com observações tendo resistências iguais a zero, logo aplicamos o logaritmo em $\mathbf{Y}_i + 1$ como transformação para a variável resposta. Com isso, o modelo para o logaritmo de cada perfuração do solo pode ser representado através de (2.1) como

$$\underset{m_i \times 1}{\log(\mathbf{Y}_i + 1)} = \underset{m_i \times p}{\mathbf{X}_i} \times \underset{p \times 1}{\boldsymbol{\beta}} + \underset{m_i \times q}{\mathbf{Z}_i} \times \underset{q \times 1}{\mathbf{b}_i} + \underset{m_i \times 1}{\boldsymbol{\epsilon}_i}, \text{para } i = 1, ..., n. \tag{4.1}$$

Em relação à escolha das covariáveis para os efeitos fixos, primeiramente, devido a problemas de singularidade e com base no Gráfico 1 da Figura 3.4, juntamos as áreas 1,2 e 3 em apenas uma categoria por apresentarem comportamento médio muito parecido. Ainda no Gráfico 1 da Figura 3.4, podemos observar que as áreas 4 e 5 apresentam comportamentos distintos das demais a partir de 20 centímetros de profundidade. A área 4 possui um crescimento médio em grande escala até 35 centímetros e depois com um leve decréscimo em 40 centímetros. Já a área 5 apresenta um leve salto de 15 para 20 centímetros, permanece constante até 30 centímetros e depois um leve decaimento até possuir uma resistência média próxima das áreas 1,2 e 3 com 40 centímetros de profundidade.

Por conta desse comportamento não linear entre profundidade e resistência que é assumido no modelo, criamos varíaveis *dummies* para as interações entre as áreas 4 e 5 e as

profundidades a partir de 20 centímetros, quando ambas as curvas começam a apresentar um comportamento diferente. Desse modo, temos um total 10 interações duplas. Um exemplo dessas variáveis de interação é definido como

$$X_4X_{25} = \begin{cases} 1, & \text{se a observação pertence à área 4 e profundidade 25 cm} \\ 0, & \text{caso contrário.} \end{cases}$$

Além dessas interações duplas temos também os efeitos principais das variáveis vistas no capítulo anterior. Portanto, as covariáveis consideradas na matriz de planejamento dos efeitos fixos são:

- Áreas em que foram coletadas as observações (varíavel com 3 categorias: 1 - áreas (1,2 e 3), 4 - área 4 e 5 - área 5);
- Níveis de umidades no solo (variável com 3 categorias: seco, úmido e encharcado);
- Tipos de penetrômetros utilizados na coleta das observações (variável com 3 categorias: impacto, manual ou automático);
- Profundidade em centímetros;
- As variáveis de interação comentadas anteriormente; e
- A coluna de valores iguais a 1 para considerarmos o intercepto.

Exceto para a variável profundidade que é numérica, representamos todas as variáveis categóricas através de variáveis *dummies*. Logo, a matriz $\mathbf{X}_i$ será associada ao vetor dos efeitos fixos $\boldsymbol{\beta}$ com $p = 18$, ou seja, $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_{17})^\top$, em que

- $\boldsymbol{\beta}_0$: intercepto;
- $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$: coeficientes associados às categorias das áreas 4 e 5, respectivamante (como estamos trabalhando com *dummies*, a categoria 1, com as áreas 1, 2 e 3 agrupadas, é a categoria de referência);
- $\boldsymbol{\beta}_3$ e $\boldsymbol{\beta}_4$: coeficientes ligados aos penetrômetros de impacto e manual, respectivamente (automático como categoria de referência);
- $\boldsymbol{\beta}_5$ e $\boldsymbol{\beta}_6$: coeficientes relacionados às categorias de umidades seco e umido, respectivamente (encharcado como categoria de referência);

- $\boldsymbol{\beta}_7$: coeficiente ligado à profundidade;

- $\boldsymbol{\beta}_8$ a $\boldsymbol{\beta}_{12}$: coeficientes das interações duplas entre a área 4 e as profundidades de 20 a 40, respectivamente; e

- $\boldsymbol{\beta}_{13}$ a $\boldsymbol{\beta}_{17}$: coeficientes das interações duplas entre a área 5 e as profundidades de 20 a 40, respectivamente.

Com a matriz $\mathbf{X}_i$ obtida, temos também a matriz de planejamento que, em suas colunas, contém as observações das covariáveis associadas aos efeitos aleatórios, a matriz $\mathbf{Z}_i$. Os efeitos aleatórios são aqueles que impactam na variabilidade da resistência mecânica do solo à penetração e as perfurações (unidade amostral) possuem estimativas individualmente diferentes para cada efeito aleatório, ou seja, os efeitos aleatórios estão relacionados à diferença entre as perfurações.

Foram feitos vários ajustes de modelos em busca da melhor escolha para os efeitos aleatórios. Dois ajustes obtiveram os melhores resultados com relação ao diagnóstico do modelo e com valores condizentes para as estimativas dos parâmetros fixos com base na análise descritiva.

O primeiro modelo contém apenas o intercepto aleatório e o segundo com o intercepto e a profundidade aleatórios, porém notamos que as predições dos efeitos aleatórios associados à profundidade eram praticamente zero e a exclusão desse efeito não impactava de maneira relevante os resultados do modelo. Logo, em busca de um modelo mais parcimonioso, o primeiro foi o selecionado, ou seja, temos apenas um intercepto aleatório para cada unidade amostral (perfuração) no vetor $\mathbf{b}_i$ com $q = 1$ e Z_i se resume a um vetor coluna do valor 1. Com isso, teremos uma estimativa de intercepto para cada uma das unidades experimentais, ou seja, totalizando 1103 efeitos aleatórios estimados no modelo.

Por último, temos que $\boldsymbol{\epsilon}_i$ é o vetor dos erros aleatórios da i-ésima perfuração. Como visto em (2.1), geralmente $\mathbf{R}_i = \sigma^2\mathbf{I}_{m_i}$, porém, para a modelagem do estudo implementamos uma estrutura de covariância autoregressiva, definida na Eq. (2.13), que se mostrou mais adequada aos dados do que a matriz que assume homocedasticidade e independência entre os erros aleatórios. A escolha dessa estrutura foi por conta da ideia de que as medições em profundidades mais próximas, em tese, são mais correlacionadas do que as medições em profundidades mais distantes. Considerando os efeitos aleatórios, como temos apenas o intercepto, precisamos estimar apenas uma variância associada a ele. Com

isso, temos que $\boldsymbol{\theta} = (\sigma^2, \phi, \sigma_1^2)^\top$, ou seja, será necessário estimar dois parâmetros para a matriz de covariâncias dos erros aleatórios e uma variância para os interceptos aleatórios.

4.2 Ajuste do modelo

Através do *software* R e do pacote *nlme* de Pinheiro *et al.* (2023), aplicamos o modelo (4.1) com o intuito de modelar o logaritmo da resistência mecânica do solo à penetração adicionando 1 unidade de MPa para as 1103 unidades experimentais do banco de dados. Utilizando o método de estimação de máxima verossimilhança restrita (MVR), os resultados em relação aos efeitos fixos podem ser visto abaixo.

Tabela 4.1: A tabela contempla estimativas pontuais e o valor da estatística e do valor-p para o teste Wald de significância de cada efeito.

	Estimativa	Erro padrão	Graus de liberdade	Estatística teste	Valor-p
Intercepto	0.4872	0.0162	7710	30.1211	$< 2 \times 10^{-16}$
Área 4	0.0251	0.0206	1096	1.2164	0.2241
Área 5	0.2413	0.0194	1096	12.4221	$< 2 \times 10^{-16}$
Penet. Impacto	0.5017	0.0151	1096	33.2354	$< 2 \times 10^{-16}$
Penet. Manual	-0.1139	0.0136	1096	-8.3736	$< 2 \times 10^{-16}$
Umid. Seco	0.2438	0.0146	1096	16.6773	$< 2 \times 10^{-16}$
Umid. Úmido	0.1300	0.0137	1096	9.5013	$< 2 \times 10^{-16}$
Profundidade	0.0155	0.0004	7710	39.0851	$< 2 \times 10^{-16}$
Área 4 e Prof. 20	0.1361	0.0168	7710	8.0997	6.35×10^{-16}
Área 4 e Prof. 25	0.1655	0.0219	7710	7.5590	4.53×10^{-14}
Área 4 e Prof. 30	0.2027	0.0250	7710	8.1115	5.77×10^{-16}
Área 4 e Prof. 35	0.1982	0.0272	7710	7.2877	3.47×10^{-13}
Área 4 e Prof. 40	0.1068	0.0289	7710	3.6919	0.0002
Área 5 e Prof. 20	0.0937	0.0156	7710	5.9949	2.13×10^{-9}
Área 5 e Prof. 25	0.0576	0.0204	7710	2.8208	0.0048
Área 5 e Prof. 30	-0.0410	0.0234	7710	-1.7534	0.0796
Área 5 e Prof. 35	-0.1640	0.0255	7710	-6.4345	1.31×10^{-10}
Área 5 e Prof. 40	-0.2614	0.0272	7710	-9.6063	$< 2 \times 10^{-16}$

Pela Tabela 4.1, temos evidências de que, com um nível de significância de 5% e utilizando o teste de Wald, apenas o efeito principal para a categoria Área 4 e a interação dupla Área 5 e Prof. 30 não foram significativos para explicar o logaritmo da variável

resposta. Em contrapartida, os demais efeitos foram fortemente significativos, visto que cada efeito possui um valor-p praticamente igual a 0. Além disso, podemos ter as seguintes interpretações para as estimativas dos efeitos fixos:

- A resistência mecânica à penetração média é menor nas áreas 1, 2 e 3, enquanto que a área 5 é a que apresenta maior resistência média;
- O penetrômetro manual é o que apresenta, em média, os menores valores de resistência mecânica, ao passo que o de impacto apresenta, em médias, os maiores valores;
- O solo encharcado apresenta as menores resistências médias, acompanhado pelos solos úmidos e seco;
- A resistência mecânica média aumenta com a profundidade; e
- Os efeitos de interação entre área 4 e 5 e profundidade acompanham os comportamentos de aumento e redução observados na análise descritiva. Provavelmente, é devido à presença desses efeitos de interação que o efeito principal da área 4 não foi significativo, considerando que nas primeiras profundidades, seu comportamento médio se assemelha ao das áreas 1, 2 e 3.

Os resultados das estimativas para os efeitos fixos podem ser comparados com a análise descritiva. Na Figura 3.2, tivemos um indicativo de que, em média, as áreas possuíam resistências iguais; no entanto, observa-se que as caixas das áreas 4 e 5 estão ligeiramente acima das caixas 1, 2 e 3, assim como as curvas médias na Figura 3.4. Portanto, isso está de acordo com o fato de que as estimativas apresentadas pelo modelo para as áreas 4 e 5 são maiores que zero em relação à categoria criada para as áreas 1, 2 e 3.

O mesmo acontece em relação às umidades. Ainda na Figura 3.2, tivemos indicativos de igualdade para as resistências em média, porém as caixas das umidades seca e úmida estão levemente acima da encharcada. As curvas médias na Figura 3.4 evidenciam o mesmo comportamento. Logo, condiz com as estimativas vistas na Tabela 4.1, em que as duas categorias possuem estimativas positivas e, portanto, em comparação a encharcado impactam mais na resistência do solo.

Através da Figura 3.3, tivemos um indicativo de que, em média, o penetrômetro de impacto apresentava resistência do solo maior do que os demais e que os penetrômetros

dos tipos automático e manual possuíam a mesma resistência média. No entanto, nota-se que a caixa do tipo automático estava ligeiramente acima da caixa do penetrômetro manual. As curvas médias na Figura 3.4 evidenciam o mesmo comportamento. Logo, tem sentido com as estimativas do modelo, visto que na Tabela 4.1 o penetrômetro de impacto tem a maior estimativa positiva e o penetrômetro manual possui estimativa negativa, em relação ao penetrômetro automático.

E por último, temos também que o resultado é condizente com a análise descritiva para a variável profundidade. Na Figura 3.3, tivemos um indicativo de que a resistência, em média, ficava maior a medida em que a profundidade aumentava. Como a estimativa obtida para a profundidade é maior que zero, ela impacta positivamente na resistência mecânica em média, sendo assim, coerente com a análise vista anteriormente.

Com os resultados para os efeitos fixos apresentados e analisados, agora vamos exibí-los para os efeitos aleatórios.

Tabela 4.2: Resultados para os efeitos aleatórios.

	Intercepto	Resíduo
Desvio Padrão	2.6832×10^{-5}	0.2876

Vimos anteriormente que os efeitos aleatórios impactam diretamente na variância da variável resposta. O intuito é o efeito aleatório não ter um desvio padrão muito próximo de zero (porque nesse caso os valores preditos são todos praticamente nulos) e não ser muito inferior ao desvio dos resíduos, pois assim ele pode não ser relevante para explicar a associação entre as observações. Com isso, verificamos pela Tabela 4.2 que o desvio padrão do intercepto é bem baixo em relação ao desvio dos resíduos, e isso talvez aconteceu por conta da estrutura de covariância (2.13) escolhida para os erros.

Durante o ajuste do modelo, foram realizados vários testes na busca de modelos diferentes e, talvez, mais adequados. Apesar de não serem mostrados aqui, um modelo sem a estrutura AR na covariância dos erros aleatórios foi estimado, o desvio padrão do intercepto aumentou consideravelmente, mas outras métricas de qualidade ficaram piores. De qualquer maneira, o intercepto aleatório (apesar de apresentar baixo valor de desvio padrão estimado) ainda é relevante para esse modelo.

Na seção anterior, vimos que cada uma das perfurações possui uma estimativa de intercepto única, representando, portanto, a aleatoriedade dessas perfurações no conjunto de dados. Abaixo temos as estimativas dos interceptos aleatórios nas perfurações:

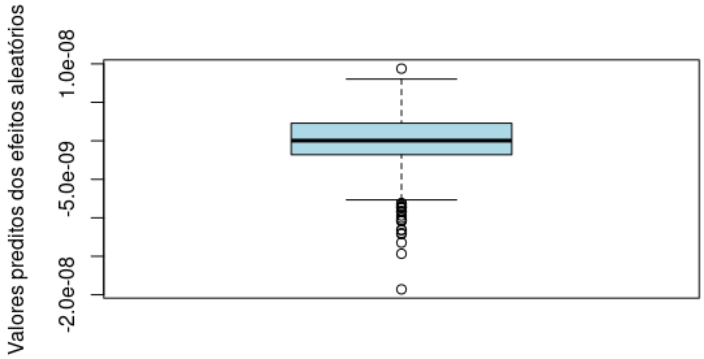

Figura 4.1: Valores preditos dos interceptos aleatórios nas perfurações.

Pela Figura 4.1, nota-se que as estimativas dos efeitos aleatórios são bem próximas de zero. Além disso, temos estimativas tanto positivas quanto negativas.

No Capítulo 2, vimos que seria necessário estimar a matriz de variâncias e covariâncias entre os efeitos aleatórios, $\mathbf{G}$. Como o modelo possui apenas o intercepto como efeito aleatório, $\mathbf{G}$ terá dimensão (1×1), ou seja, será composta apenas da variância estimada do intercepto aleatório. Temos na Tabela 4.2 o desvio padrão, $\hat{\sigma}(b_{i1})$, estimado do intercepto aleatório, logo a variância estimada será $\hat{\sigma}^2(b_{i1}) = 7.1996 \times 10^{-10}$.

Por último, estimamos a matriz de variâncias e covariâncias para o vetor dos erros aleatórios, $\mathbf{R}_i(\boldsymbol{\theta})$. Foi dito previamente que implementamos a estrutura de covariância autoregressiva, com isso, para estimarmos $\mathbf{R}_i(\boldsymbol{\theta})$ é necessário estimar os parâmetros de variância e covariância contidos no vetor $\boldsymbol{\theta}$, ou seja, o parâmetro referente a variância e o coeficiente de correlação, σ^2 e ϕ respectivamente.

Tabela 4.3: Estimativas dos parâmetros para os erros aleatórios.

	σ^2	ϕ
Estimativas	0.0827	0.6960

A Tabela 4.3 mostra as estimativas desses parâmetros. A estimativa da correlação evidencia uma correlação positiva moderada entre os erros de medições adjacentes.

Tabela 4.4: Estimativa da matriz $\mathbf{R}_i$.

	1	2	3	4	5	6	7	8
1	0.0827	0.0576	0.0401	0.0279	0.0194	0.0135	0.0094	0.0065
2	0.0576	0.0827	0.0576	0.0401	0.0279	0.0194	0.0135	0.0094
3	0.0401	0.0576	0.0827	0.0576	0.0401	0.0279	0.0194	0.0135
4	0.0279	0.0401	0.0576	0.0827	0.0576	0.0401	0.0279	0.0194
5	0.0194	0.0279	0.0401	0.0576	0.0827	0.0576	0.0401	0.0279
6	0.0135	0.0194	0.0279	0.0401	0.0576	0.0827	0.0576	0.0401
7	0.0094	0.0135	0.0194	0.0279	0.0401	0.0576	0.0827	0.0576
8	0.0065	0.0094	0.0135	0.0194	0.0279	0.0401	0.0576	0.0827

Na Tabela 4.4 temos nas linhas e nas colunas as profundidades em que foram feitas as medições das amostras sendo 1 a profundidade de 5 centímetros e 8 a de 40 centímetros. Observa-se que $\mathbf{R}_i$ apresenta maiores covariâncias entre os erros de profundidades mais próximas quando observado linha a linha, e menores covariâncias entre erros de medições mais distantes. Esse comportamento era o desejado na escolha da estrutura e pela análise de diagnóstico, a ser mostrado em seguida, parece ser adequada.

Basta agora verificarmos se as suposições do modelo são satisfatoriamente atendidas por meio de uma análise de diagnóstico para confiarmos nos resultados obtidos.

4.3 Diagnóstico

Vimos anteriormente no Capítulo 2 métodos para se verificar a análise das suposições em relação aos erros e aos efeitos aleatórios. Primeiramente, vamos analisar as suposições de homocedasticidade, normalidade e observar também a presença ou não de *outliers* para os erros através de resíduos. Isto é, se os resíduos atenderam essas suposições, elas serão validadas para os erros aleatórios do modelo. Para isso então, usamos os resíduos condicionais padronizados e os resíduos com confundimento mínimo padronizados.

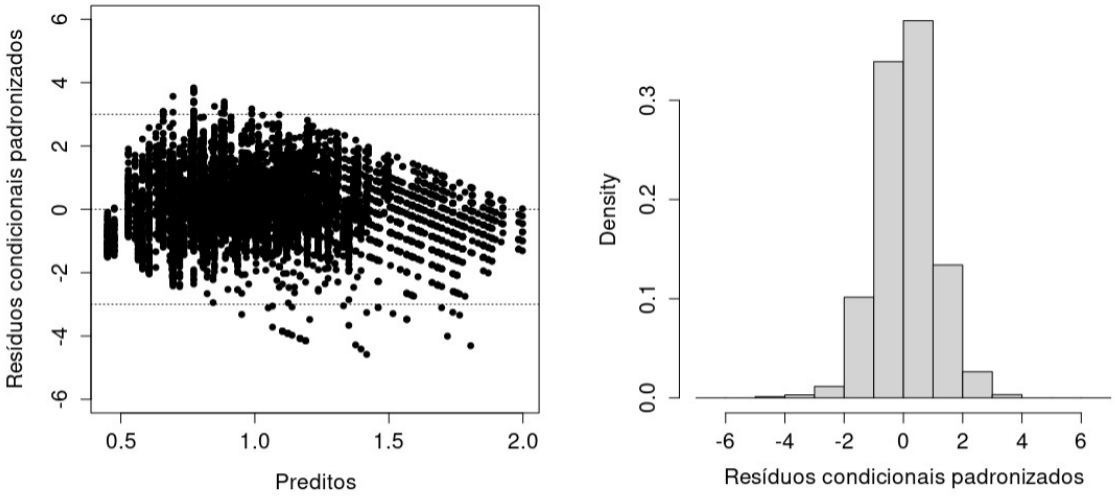

Figura 4.2: Gráfico dos resíduos condicionais padronizados *versus* preditos e histograma dos resíduos condicionais padronizados, respectivamente.

Com base na Figura 4.2, observamos que os resíduos condicionais padronizados no geral estão bem distribuídos em torno do zero. Exceto nas extremidades dos valores preditos, que possuem poucas observações, os pontos estão alocados dentro de uma mesma faixa de dispersão. Pelo histograma, na direita, nota-se que a distribuição dos resíduos condicionais padronizados seguem uma simetria. Portanto, podemos dizer que a suposição de homocedasticidade para os erros está satisfeita.

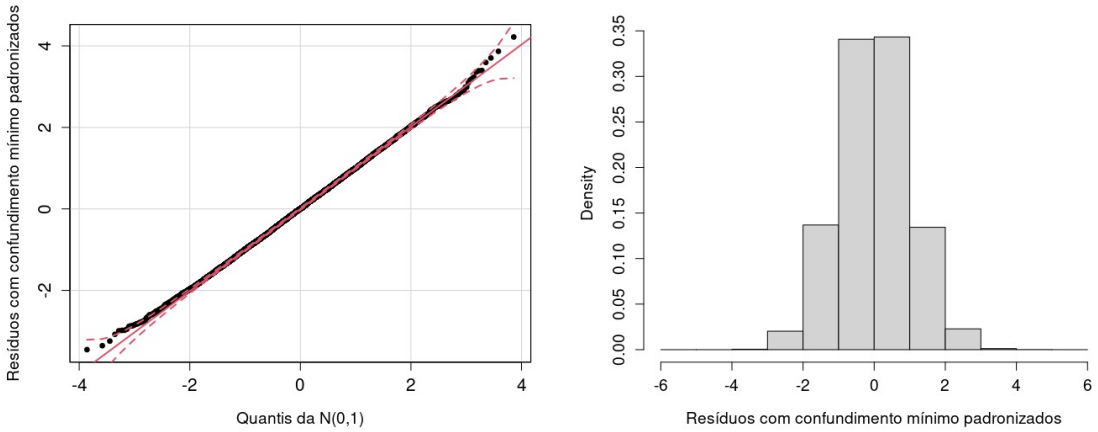

Figura 4.3: Gráfico Quantil-Quantil para a distribuição Normal com um envelope de 95% de confiança e histograma dos resíduos com confundimento mínimo padronizados, respectivamente.

Pela Figura 4.3, no gráfico da esquerda, observa-se que os resíduos com confundimento mínimo padronizados estão todos dentro do envelope com 95% de confiança e no histograma, pelo gráfico da direita, podemos ver que os resíduos são bastante simétricos para a densidade Normal. Logo, concluímos que os resíduos com confundimento mínimo

padronizados possuem distribuição Normal e, com isso, a suposição de normalidade para os erros está atendida.

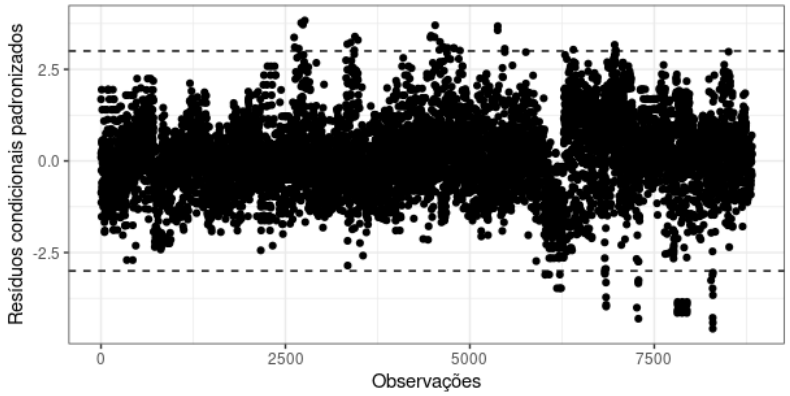

Figura 4.4: Gráfico dos resíduos condicionais padronizados *versus* indice das observações.

Com base na Figura 4.4, nota-se que alguns pontos entre as observações 2500 e 5000 são pontos discrepantes, pois seus resíduos condicionais padronizados são superiores a 3. Outros pontos discrepantes aparecem a partir da observação 6500, aproximadamente, e os mais atípicos estão por volta da observação 7500.

Como temos alguns pontos discrepantes, ou seja, *outliers*, na próxima seção, iremos analisá-los para verificar se há algum padrão ou problema em relação às observações da amostra. Apesar de termos alguns *outliers*, isso não é tão problemático, uma vez que possuímos um grande número de observações.

Com as suposições dos erros atendidas e com a presença de alguns *outliers*, vamos agora verificar se as suposições para os efeitos aleatórios estão atendidas. Aqui vamos nos ater a verificar a normalidade dos efeitos preditos. Também será verificado a presença ou não de *outliers* nas perfurações (unidades amostrais). Para a investigação vamos utilizar as predições dos efeitos aleatórios.

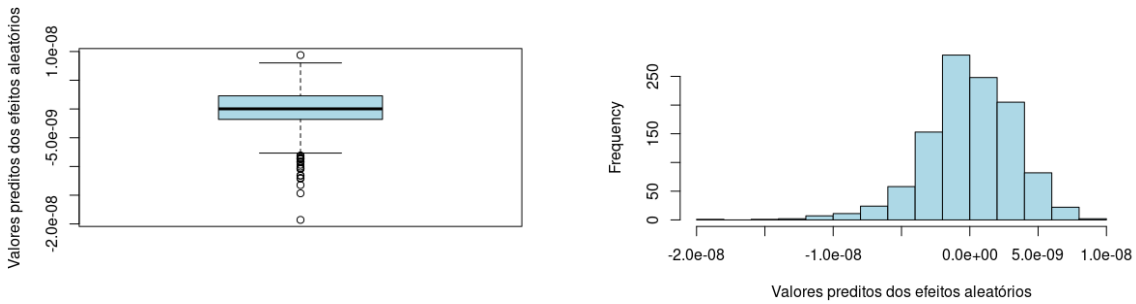

Figura 4.5: Boxplot (na esquerda) e histograma (na direita) das predições para os efeitos aleatórios.

Na Figura 4.5, observa-se que pelo boxplot e histograma, temos uma leve assimetria por conta da cauda inferior ser um pouco mais pesada, porém podemos assumir que os efeitos aleatórios possuem uma distribuição simétrica satisfatória, aqui a Normal.

Ainda na Figura 4.5, encontra-se alguns *outliers* no boxplot. A perfuração (unidade amostral) com a estimativa de valor máximo é um ponto discrepante e, além dela, temos alguns pontos negativos também atípicos na cauda inferior. Como o conjunto de dados é grande, com 1103 unidades amostrais, é natural termos alguns (mas poucos) pontos atípicos, como é o caso aqui observado.

4.4 Análise de *outliers*

Outliers, ou valores atípicos, são pontos que apresentam grandes diferenças entre as demais observações, ou que apresentam inconsistências. Neste trabalho, com os *outliers* já observados na seção anterior, apresentamos uma análise mais detalhada sobre eles via tabelas de frequência que apresentam esses valores discrepantes.

Das 8824 observações do banco de dados, a análise de diagnóstico realizada apresentou 66 resíduos discrepantes, sendo as observações: 2618, 2626, 2682, 2714, 2738, 2762, 3338, 3386, 3434, 3442, 3482, 4466, 4474, 4522, 4530, 4586, 4595, 4642, 4698, 4778, 4786, 4866, 5377, 5378, 5473, 6021, 6076, 6181, 6220, 6221, 6237, 6403, 6830, 6831, 6845, 6846, 6847, 6848, 6963, 6971, 6979, 7264, 7282, 7284, 7285, 7287, 7813, 7814, 7815, 7816, 7877, 7878, 7879, 7880, 7941, 7942, 7943, 7944, 8269, 8289, 8291, 8292, 8293, 8294, 8295 e 8296.

Tabela 4.5: Frequência das observações *outliers* em relação às áreas.

	Áreas				
	1	2	3	4	5
Nº de *outliers*	0	11	14	16	25

Pela Tabela 4.5, nota-se que a área 1 não apresenta valor atípico. Além disso, temos que o número de observações discrepantes para as de áreas 2 a 4, no geral, está razoavelmente uniforme, enquanto a área 5 apresenta um leve aumento no número de *outliers*.

Tabela 4.6: Frequência das observações *outliers* em relação às umidades.

	Umidades		
	E	*S*	*U*
Nº de *outliers*	11	30	25

Na Tabela 4.6, observa-se que as umidades seca e úmida apresentam mais valores aberrantes do que o nível encharcado. Além disso, percebe-se que todas os níveis apresentam *outliers*.

Tabela 4.7: Frequência das observações *outliers* em relação aos penetrômetros.

	Penetrômetros		
	A	*I*	*M*
Nº de *outliers*	19	11	36

A Tabela 4.7 mostra que o penetrômetro do tipo manual se destoa levemente dos demais na frequência de valores discrepantes e isso, talvez, pelo fato do desempenho desse penetrômetro depender muito do funcionário que o utiliza.

Tabela 4.8: Frequência das observações *outliers* em relação às profundidades.

	Profundidades (em centímetros)							
	5	10	15	20	25	30	35	40
Nº de *outliers*	3	23	6	4	11	6	7	6

Pela Tabela 4.8, nota-se que a profundidade de 10 centímetros se destaca dentre as outras com uma quantidade maior de valores atípicos. Apesar disso, no geral, as profundidades apresentam uma certa constância no número de *outliers* em seus níveis de coleta.

Tabela 4.9: Frequência das observações *outliers* em relação às áreas e profundidades.

	Profundidades (em centímetros)							
Área	5	10	15	20	25	30	35	40
1	0	0	0	0	0	0	0	0
2	0	11	0	0	0	0	0	0
3	2	11	1	0	0	0	0	0
4	0	0	4	2	5	2	2	1
5	1	1	1	2	6	4	5	5

Na Tabela 4.9, podemos observar que todos os valores aberrantes da área 2 foram coletados em 10 centímetros de profundidade e, além disso, praticamente todos os *outliers* da área 3 também estão presentes na profundidade de 10 centímetros. Temos também que a área 5 apresenta valores atípicos em todas as profundidades, sendo que, a partir de 25 centímetros, a frequência desses valores aumenta levemente. E por último, podemos ver que de 15 centímetros em diante a área 4 apresenta números discrepantes em todos os níveis de coleta.

Em resumo, apesar de termos observações atípicas, elas representam 0.75% da quantidade de observações e geralmente acontecem em combinação de áreas e profundidades que, pela análise descritiva, já mostravam um comportamento diferente do comportamento médio das demais observações.

Considerando agora os *outliers* nos efeitos aleatórios, observamos que 23 dos 1103 valores preditos são discrepantes, sendo eles as predições das perfurações: 589, 760, 764, 766, 768, 773, 775, 776, 777, 778, 780, 782, 784, 854, 856, 908, 911, 970, 971, 977, 985, 993 e 1037.

Pela Tabela 4.10, nota-se que a maior parte das unidades amostrais (perfurações) discrepantes são feitas na área 4 com umidade encharcada e penetrômetro do tipo impacto. Temos também que a presença de 6 valores atípicos na área 5 com umidade do tipo úmida.

Novamente, percebemos que as unidades amostrais mais diferentes estão presentes em áreas que pela análise descritiva já apresentavam um comportamento especial. Seria interessante avaliar se elas possuem alguma característica diferente ou se tivemos algum problema de medição.

Tabela 4.10: Frequência dos efeitos aleatórios *outliers* com diferentes combinações de umidade (E, S e U), área (1 a 5) e penetrômetro (A, I e M).

(a) Umidade = E

	Penetrômetros		
Área	A	I	M
1	0	0	0
2	0	0	0
3	0	0	0
4	0	12	2
5	0	0	0

(b) Umidade = S

	Penetrômetros		
Área	A	I	M
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	1	0	1

(c) Umidade = U

	Penetrômetros		
Área	A	I	M
1	0	0	0
2	0	0	0
3	0	0	1
4	0	0	0
5	0	2	4

4.5 Modelo misto com distribuição Gama

Neste trabalho, além de apresentarmos o modelo linear misto ajustado e analisado assumindo distribuição Normal para as variáveis respostas, exploramos brevemente também um modelo linear generalizado misto que adota uma distribuição gama para elas.

A opção por este modelo foi motivada pelo fato de a variável resposta (resistência mecânica do solo) apresentar exclusivamente valores positivos. Além disso, as Figuras 3.2 e 3.3 da análise descritiva indicam uma assimetria nas caudas de forma geral.

O intuito desse breve ajuste é apenas em comparar as suas estimativas e valores preditos para os efeitos aleatórios com as estimativas e valores preditos para os efeitos aleatórios do modelo anterior. Então modelamos as observações dos vetores aleatórios $(\mathbf{Y}_i + 1)$ usando a função de ligação logaritma. Foi necessário somar uma unidade em todas as observações pois tínhamos valores zero que não são modelados pela distribuição Gama. Além disso, aplicar uma transformação na variável profundidade para que o modelo congervisse. Passamos as profundidades de centímetros para metros, ou seja, ao invés de termos as profundidade de 5 a 40 centímetros, elas estão representadas de 0.05 a 0.40 metros. Esse modelo também não terá a estrutura de covariância (2.13) para os erros, pois no pacote utilizado não foi possível adicionar essa estrutura para os erros. Para detalhes técnicos sobre o modelo Gama ver Singer *et al.* (2018) e Ribeiro *et al.* (2019).

Utilizando o pacote *lme4* de Bates *et al.* (2023), aplicamos o modelo *log*-Gama para modelar a resistência mecânica do solo adicionando 1 unidade de MPa. Os resultados para os efeitos fixos seguem abaixo.

Tabela 4.11: Resultados obtidos dos efeitos fixos para o modelo log-Gama.

	Estimativa	Erro padrão	Estatística teste	Valor-p
Intercepto	0.6284	0.0182	34.592	$< 2 \times 10^{-16}$
Área 4	-0.0011	0.0222	-0.052	0.9586
Área 5	0.2083	0.0209	9.945	$< 2 \times 10^{-16}$
Penet. Impacto	0.4420	0.0193	22.909	$< 2 \times 10^{-16}$
Penet. Manual	-0.0870	0.0173	-5.013	5.35×10^{-7}
Umid. Seco	0.2454	0.0186	13.159	$< 2 \times 10^{-16}$
Umid. Úmido	0.1194	0.0174	6.843	7.78×10^{-12}
Prof (em metros)	1.1236	0.0250	44.888	$< 2 \times 10^{-16}$
Área 4 e Prof. 20	0.2778	0.0192	14.441	$< 2 \times 10^{-16}$
Área 4 e Prof. 25	0.2872	0.0194	14.763	$< 2 \times 10^{-16}$
Área 4 e Prof. 30	0.3098	0.0197	15.717	$< 2 \times 10^{-16}$
Área 4 e Prof. 35	0.3015	0.0200	15.039	$< 2 \times 10^{-16}$
Área 4 e Prof. 40	0.2196	0.0205	10.728	$< 2 \times 10^{-16}$
Área 5 e Prof. 20	0.1662	0.0179	9.296	$< 2 \times 10^{-16}$
Área 5 e Prof. 25	0.1277	0.0181	7.054	1.74×10^{-12}
Área 5 e Prof. 30	0.0308	0.0184	1.676	0.0937
Área 5 e Prof. 35	-0.0782	0.0188	-4.162	3.15×10^{-5}
Área 5 e Prof. 40	-0.1610	0.0192	-8.371	$< 2 \times 10^{-16}$

A Tabela 4.11 apresenta resultados bem semelhantes comparada com a Tabela 4.1. Temos indícios de que, com um nível de significância de 5% e com o uso do teste de Wald, a categoria Área 4 e a interação dupla Área 5 e Prof. 30 foram os únicos efeitos fixos não significativos, ou seja, um resultado coincidente ao do modelo misto com distribuição Normal.

Assim como no primeiro modelo, as estimativas dos efeitos possuem valores baixos, porém os desvios padrões associados são extremamente pequenos. No geral, as estimativas apresentadas na Tabela 4.11 são bem próximas com as do modelo anterior. As únicas diferenças são em relação a categoria Área 4 que passou a ter estimativa negativa, mas praticamente zero, e a interação dupla Área 5 e Prof. 30 que antes tinha estimativa

negativa e agora apresentou um valor positivo. Porém, em ambos os modelos, as categorias são não significativas.

Tabela 4.12: Resultados para os efeitos aleatórios do modelo *log*-Gama.

	Intercepto	Resíduo
Desvio Padrão	0.1258	0.2325
Variância	0.0158	0.0541

Pela Tabela 4.12, temos o mesmo comportamento da Tabela 4.2, ou seja, que o desvio padrão do intercepto é menor que desvio dos resíduos, porém dessa vez o desvio do intercepto é bem maior comparado com o valor encontrado na Tabela 4.2. Esse aumento se deve, provavelmente, à falta da estrutura (2.13) no modelo *log*-Gama.

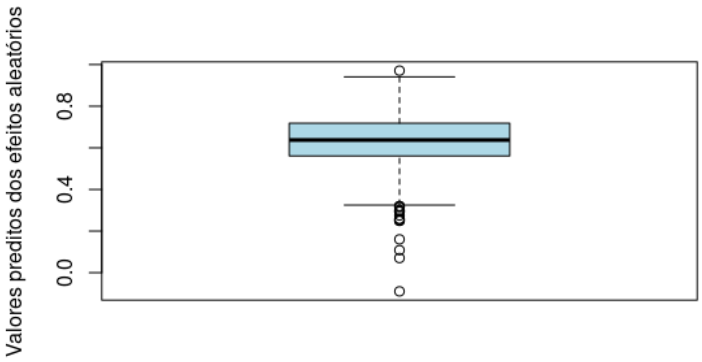

Figura 4.6: Valores preditos dos interceptos aleatórios nas perfurações através do modelo *log*-Gama.

A Figura 4.6, quando comparada com a Figura 4.1 apresenta predições muito maiores. Além disso, foi verificado que apenas a perfuração 1037 possui predição negativa, enquanto que, no modelo misto com distribuição Normal, 544 perfurações possuem valores preditos negativos. Isso provavelmente se deve ao fato da distribuição escolhida no ajuste.

Neste trabalho não iremos entrar em mais detalhes em relação aos modelos lineares generalizados mistos (GLMMs), porém foi visto que os coeficientes fixos que foram significativos do modelo *log*-Gama coincidem com o resultado do modelo linear misto e, em termos de estimativas, os valores apresentados foram próximos. Na análise de diagnóstico, não mostrada aqui, o modelo misto linear também se apresentou mais adequado do que o modelo Gama.

4.6 Comparação entre o modelo misto e o modelo linear de efeitos fixos

Na introdução desse trabalho, discutiu-se que em um estudo anterior para a aplicação desse conjunto de dados, o modelo de regressão linear apenas com efeitos fixos e assumindo independência entre as medições foi implementado (ANOVA). A princípio, essa metodologia não parece ser adequada pois as medições não são independentes. Portanto, para efeitos de comparação, ajustamos uma regressão linear com os mesmos efeitos fixos usados em (4.1) para o logaritmo da resistência do solo.

O objetivo desta seção é comparar, por meio da verificação da suposição de normalidade dos erros através dos resíduos, se um modelo mais simples é também adequado. Em outras palavras, na análise da regressão linear, vamos verificar se os resíduos seguem uma distribuição Normal. Se isso ocorrer, podemos concluir que os erros são provenientes de uma distribuição Normal e a suposição foi satisfatória e, apesar das medições não serem independentes, as suposições da regressão linear não são totalmente violadas.

Através do pacote *hnp* de Moral *et al.* (2018), temos o gráfico quantil-quantil para os resíduos da regressão linear e um envelope para a distribuição Normal com 95% de confiança que foram usados para essa verificação.

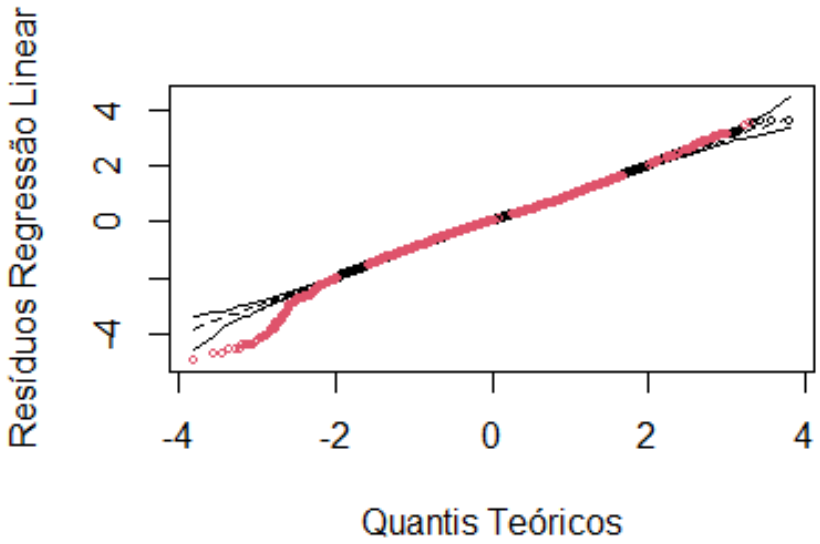

Figura 4.7: Gráfico Quantil-Quantil para a distribuição Normal com um envelope de 95% de confiança para a regressão linear (em vermelho, temos os pontos fora do envelope e, em preto, os que estão dentro).

Pela Figura 4.7, observa-se que vários resíduos estão fora do envelope com 95% de confiança. Ou seja, concluímos que, por meio de uma visualização gráfica, os resíduos não possuem distribuição Normal. Logo, a suposição de normalidade para os erros do modelo não é satisfatória.

Vimos na Figura 4.3 que, em relação ao modelo misto linear, a suposição foi satisfeita. Com isso, podemos concluir que a metodologia do modelo misto apresentado ao longo do trabalho foi mais adequada do que a regressão linear para esse conjunto de dados.

Capítulo 5

Conclusões e estudos futuros

No trabalho apresentamos a metodologia de modelos mistos, além de toda sua estrutura matemática e de estimação para o uso em uma aplicação de dados de solo. Essa técnica de modelagem foi utilizada por conta das características das observações feitas em cada perfuração, que se tratam de dados longitudinais medidos em níveis de profundidades distintos porém na mesma posição do solo.

O modelo linear misto foi ajustado e, através dele, concluímos que os fatores: área, umidade e penetrômetro e as interações duplas entre às áreas 4 e 5 com as profundidades de 20 a 40, com exceção da interação entre área 5 e profundidade 30, foram significativos para explicar o logaritmo da resistência mecânica do solo à penetração.

Pelas estimativas associadas aos penetrômetros de impacto e manual, podemos constatar que os três tipos da ferramenta são diferentes em relação à resistência mecânica média observada. O penetrômetro de impacto é o que apresenta, em média, os maiores valores de resistência mecânica e o manual os menores valores.

Além disso, o modelo apresentou um bom ajuste para o logaritmo da resistência do solo à penetração, uma vez que pela a análise de diagnóstico, observamos que suas suposições foram satisfatoriamente atendidas. Observações e perfurações atípicas e discrepantes também foram identificadas, mas elas representam uma quantidade muito pequena dentro do conjunto de dados analisado. Geralmente, elas estão presentes na área 4 e 5 e seria interessante avaliar se essas áreas possuem, de fato, alguma característica diferente ou se tivemos algum problema de medição.

Por fim, implementamos brevemente um modelo linear generalizado misto (GLMM) apenas para comparação com o modelo linear misto em relação aos efeitos fixos e aleatórios. Essa aplicação se deu por conta de comportamentos específicos na análise descritiva. O

GLMM ajustado foi o *log*-Gama e apresentou resultados muito próximos ao primeiro modelo.

Para estudos futuros, temos como objetivos desenvolver mais a parte de modelos lineares generalizados mistos e também aplicar novas técnicas de dados longitudinais para relatar possíveis melhores resultados.

Referências Bibliográficas

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N. e Tanaka, E. (2023). *Linear Mixed-Effects Models using 'Eigen' and S4*. https://CRAN.R-project.org/package=lme4. Version 1.1-35.1.

Brasil, C. (2020). O solo: patrimônio essencial da agricultura. https://croplifebrasil.org/noticias/o-solo-patrimonio-essencial-da-agricultura/. Acessado em: 13-07-2023.

de Freitas, L. A. C. (2018). Teste Wald. https://lineu96.github.io/msc_page/2_wald_test.html#introdu%C3%A7%C3%A3o. Acessado em: 27/12/2023.

Embrapa (2020). *VII Plano Diretor da Embrapa 2020-2030*. Embrapa Informação Tecnológica, Brasilia, DF, first edition.

Hilden-Minton, J. A. (1995). *Multilevel diagnostics for mixed and hierarchical linear models*. University of California, Los Angeles.

Machado, P. L. O. d. A. (2003). *Compactação do solo e crescimento de plantas: como identificar, evitar e remediar*. Embrapa Solos, Rio de Janeiro, RJ, first edition.

Menezes, T. A. V. (2018). *Comparação entre três penetrômetros na avaliação da resistência mecânica do solo à penetração de um latossolo vermelho eutroférrico*. Tese de doutorado, Universidade de São Paulo Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, SP.

Montgomery, D. C., Peck, E. A. e Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons, fifth edition.

Moral, R. d. A., Hinde, J. e Demétrio, C. G. B. (2018). *Half-Normal Plots with Simulation Envelopes*. https://CRAN.R-project.org/package=hnp. Version 1.2-6.

Nobre, J. S. (2004). *Métodos de Diagnóstico para Modelos Lineares Mistos*. Tese de mestrado, Universidade de São Paulo - Instituto de Matématica e Estatística, São Paulo.

Nobre, J. S. e Singer, J. d. M. (2007). Residual analysis for linear mixed models. *Biometrical Journal*, **49**(6), 863–875.

Patterson, H. D. e Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**(3), 545–554.

Pinheiro, J., Bates, D., DebRoy, S., Deepayan, S., authors, E., Heisterkamp, Siem Willigen, B. V., Ranke, J. e Team, R. C. (2023). *Linear and Nonlinear Mixed Effects Models*. R Core Team (R-core at R-project.org), https://svn.r-project.org/R-packages/trunk/nlme/. Version 3.1-162.

Ribeiro, M. H. D. M., Santiago, A. N., Oliveira, R. M. W. d., Milani, H. e Previdelli, I. (2019). Longitudinal modeling using log-gamma mixed model: case of memory deterioration after chronic cerebral hypoperfusion associated with diabetes in rats. *Acta Scientiarum. Technology*, **41**(1), e35789. DOI: https://doi.org/10.4025/actascitechnol.v41i1.35789.

Singer, J. M., Nobre, J. S. e Rocha, F. M. (2018). Análise de dados longitudinais. São Paulo, SP. Universidade de São Paulo - Departamento de Estatística.

Vaz, C. M. P., Primavesi, O., Patrizzi, V. C. e Iossi, M. d. F. (2002). Influência da umidade na resistência do solo medida com penetrômetro de impacto. Embrapa, São Carlos, SP.

Zhang, X. (2015). A tutorial on restricted maximum likelihood estimation in linear regression and linear mixed-effects model. *https://xiuming.info/docs/tutorials/reml.pdf*.

Zuanetti, D. A. (2022). Notas de aula da disciplina de Estatística Aplicada no Bacharelado em estatística da UFSCar - modelos mistos lineares.

Apêndice A

Soma direta de matrizes e produto de Kronecher

A.1 Soma direta

Sejam $\mathbf{A}$ uma matriz de dimensão $(m \times p)$ e $\mathbf{B}$ uma matriz de dimensão $(n \times q)$. A soma direta é uma matriz de ordem $(m \times p) \times (n \times q)$ definida como

$$\mathbf{A} \oplus \mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}.$$

Em geral, se temos matrizes $\mathbf{A}_i$ com dimensão $(n_i \times m_i)$, a soma direta delas é a matriz com dimensão $(\sum_{i=1}^{n} n_i \times \sum_{i=1}^{m} m_i)$ dada por

$$\bigoplus_{i=1}^{n} \mathbf{A}_i = \mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_n = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{A}_n \end{bmatrix}.$$

A.2 Produto de Kronecher

Se $\mathbf{A}$ é uma matriz com dimensão $(m \times n)$ e $\mathbf{B}$ uma matriz com dimensão $(p \times q)$. O produto de Kronecher das matrizes é uma matriz com dimensão $(mp \times nq)$, definida por:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}.$$

No geral, $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$.

Apêndice B

Códigos

B.1 Função residdiag_nlme

```
# plot 1: Resíduos condicionais padronizados versus preditos e histograma
# plot 2: Normal QQ plot e hist para os resíduos com confund min padronizados

####################################
residdiag.nlme = function(fit, limit,plotid=NULL) {
  require(MASS)
  require(Matrix)
  #require(car)

  ######################################
  ## This function obtains the square root of a matrix
  ######################################

  sqrt.matrix <- function(mat) {
    mat <- as.matrix(mat)
    singular_dec <- svd(mat)
    U <- singular_dec$u
    V <- singular_dec$v
    D <- diag(singular_dec$d)
    sqrtmatrix <- U %*% sqrt(D) %*% t(V)
```

```
}

#####################################
## This function extracts various objects of the function lme
#####################################

extract.lmeDesign2 <- function(m){
  start.level = 1
  data <- getData(m)
  grps <- nlme::getGroups(m)
  n <- length(grps)
  X <- list()
  grp.dims <- m$dims$ncol
  Zt <- model.matrix(m$modelStruct$reStruct, data)
  cov <- as.matrix(m$modelStruct$reStruct)
  i.col <- 1
  n.levels <- length(m$groups)
  Z <- matrix(0, n, 0)
  if (start.level <= n.levels) {
    for (i in 1:(n.levels - start.level + 1)) {
      if (length(levels(m$groups[[n.levels - i + 1]])) != 1)
      {
        X[[1]] <- model.matrix(~m$groups[[n.levels - i +
                                            1]] - 1,
                               contrasts.arg = c("contr.treatment",
                                                 "contr.treatment"))
      }
      else X[[1]] <- matrix(1, n, 1)
      X[[2]] <- as.matrix(Zt[, i.col:(i.col + grp.dims[i] -
                                        1)])
      i.col <- i.col + grp.dims[i]
      Z <- cbind(mgcv::tensor.prod.model.matrix(X),Z)
    }
```

```
    Vr <- matrix(0, ncol(Z), ncol(Z))
    start <- 1
    for (i in 1:(n.levels - start.level + 1)) {
      k <- n.levels - i + 1
      for (j in 1:m$dims$ngrps[i]) {
        stop <- start + ncol(cov[[k]]) - 1
        Vr[ncol(Z) + 1 - (stop:start),ncol(Z) + 1 - (stop:start)] <- cov[[k]]
        start <- stop + 1
      }
    }
  }
  X <- if (class(m$call$fixed) == "name" &&  !is.null(m$data$X)) {
    m$data$X
  } else   {
    model.matrix(formula(eval(m$call$fixed)),data)
  }
  y <- as.vector(matrix(m$residuals,
  ncol = NCOL(m$residuals))[,NCOL(m$residuals)] +
                  matrix(m$fitted, ncol = NCOL(m$fitted))[,NCOL(m$fitted)])
  return(list(
    Vr = Vr,
    X = X,
    Z = Z,
    sigmasq = m$sigma ^ 2,
    lambda = unique(diag(Vr)),
    y = y,
    k = n.levels
  )
  )
}

####################################
## Extracting information from lme fitted model and dataset
```

```
#####################################

data.fit <- extract.lmeDesign2(fit)
data <-    getData(fit)
y <- data.fit$y
X <- data.fit$X
N <- length(y)
# Number of observations
id <-  sort(as.numeric(getGroups(fit, level = 1)), index.return = TRUE)$x
# as.numeric(getGroups(fit, level = 1))
subject <- as.numeric(unique(id))
n <- length(as.numeric(names(table(id))))
# Number of units
vecni <- (table(id))
# Vector with number of observations per unit
p <- ncol(X)
# Number of fixed parameters
n.levels <- length(fit$groups)
# Number of levels of within subject factors
start.level <- 1
Cgrps <- nlme::getGroups(fit, level = start.level)
# Level = 1
CCind <- levels((Cgrps))
# Indices of the observations
sigma2 <- fit$sigma^2
obs <- numeric()

for (i in 1:n)
{
  obs <- append(obs,1:vecni[i])
  # Labels for observations
}
###################################### Construction of the Z and Gam matrices
```

```
if (n.levels > 1) {
  lZi <- list()
  lgi <- list()
  numrow <- numeric()

  mgroups <- fit$groups
  for (n in 1:length(CCind)) {
    dgi <- data.frame(as.matrix(mgroups[mgroups == CCind[n], ]))
    nrowzi <- dim(dgi)[1]
    ncolzi <- 0

    # Number of repetitions of the variance components
    # to construct the Gi matrix
    girep <- as.numeric(length(levels(dgi[,1])))
    for (k in 2:n.levels) {
      girep <- c(girep,as.numeric(length(levels(dgi[,k]))))
    }
    # Number of columns of the Zi matrix
    for (k in 1:n.levels) {
      ncolzi <- ncolzi + as.numeric(length(levels(dgi[,k])))
    }
    # Numbers of one's by columns of the Zi matrix
    auxi <- as.vector(table(dgi[,1]))
    for (i in 2:n.levels) {
      auxi <- c(auxi,as.vector(table(dgi[,i])))
    }
    # Matrix Zi
    l <- 1
    Zi <- matrix(0,nrowzi,ncolzi)
    # Inserting elements in Zi
    for (j in 1:ncolzi) {
      Zi[l:(l + auxi[j] - 1),j] <- rep(1,auxi[j])
```

```
    l <- l + auxi[j]
    if (l == (nrowzi + 1)) l <- 1
  }

  lZi[[n]] <- Zi

  numrow[n] <- dim(Zi)[1]

  # Matrix Gi
  comp.var <- as.matrix(fit1$modelStruct$reStruct)
  auxg <- rep(as.numeric(comp.var[1])*sigma2,girep[1])
  for (i in 2:length(girep)) {
    auxg <- c(auxg,rep(as.numeric(comp.var[i])*sigma2,girep[i]))
  }
  lgi[[n]] <- diag(auxg)
}
q <- dim(lgi[[1]])[1]                    # Dimensions of Gi matrices
for (h in 2:length(CCind)) {
  q <- c(q,dim(lgi[[h]])[1])
}
Z <- lZi[[1]]
for (k in 2:length(CCind)) {
  Z <- bdiag(Z,(lZi[[k]]))
}
Z <- as.matrix(Z)
nrowZi <- lZi[[1]]                       # Dmensions of Zi matrices
for (h in 2:length(CCind)) {
  nrowZi <- c(nrowZi,dim(lZi[[h]])[1])
}

Gam <- lgi[[1]]
for (k in 2:length(CCind)) {
  Gam <- bdiag(Gam,(lgi[[k]]))
```

```
  }
  Gam <- as.matrix(Gam)
}else{
  mataux <- model.matrix(fit$modelStruct$reStruct,data)
  mataux <- as.data.frame(cbind(mataux,id))
  lZi <- list()
  lgi <- list()

  for (i in (as.numeric(unique(id)))) {
    lZi[[i]] <- as.matrix((subset(split(mataux,id == i,
                                        drop = T)$`TRUE`,select = -id)))
    lgi[[i]] <- getVarCov(fit,type = "random.effects")
  }
  Z <- as.matrix(bdiag(lZi))
  # for (i in 2:7) {
  #   Z <- as.matrix(bdiag(Z,lZi[[i]]))
  # }
  g <- getVarCov(fit,type = "random.effects")
  q <- dim(g)[1]
  # Total number of random effects
  Gam <- as.matrix(kronecker(diag(length(as.numeric(unique(id)))),g))
}
####################################
## Estimate of the covariance matrix of conditional errors
##(homoskedastic conditional independence model)   ##
####################################

if (n.levels > 1) {
  if (!inherits(fit, "lme"))
    stop("object does not appear to be of class lme")
  grps <- nlme::getGroups(fit)
  n <- length(grps)
  n.levels <- length(fit$groups)
```

```
if (is.null(fit$modelStruct$corStruct))
  n.corlevels <- 0
else
  n.corlevels<-length(all.vars(nlme::getGroupsFormula(fit$modelStruct$corStruct)))
# Levels of the repeated measures
if (n.levels < n.corlevels) {
  getGroupsFormula(fit$modelStruct$corStruct)
  vnames <- all.vars(nlme::getGroupsFormula(fit$modelStruct$corStruct))
  lab <- paste(eval(parse(text = vnames[1]), envir = fit$data))
  if (length(vnames) > 1)
    for (i in 2:length(vnames)) {
      lab <- paste(lab, "/", eval(parse(text = vnames[i]),
                                  envir = fit$data), sep = "")
    }
  grps <- factor(lab)
}
if (n.levels >= start.level || n.corlevels >= start.level) {
  if (n.levels >= start.level)
    Cgrps <- nlme::getGroups(fit, level = start.level)
    # Level = 1
  else Cgrps <- grps
  Cind <- sort(as.numeric(Cgrps), index.return = TRUE)$ix
  # Indices of the observations
  rCind <- 1:n
  rCind[Cind] <- 1:n
  Clevel <- levels(Cgrps)
  # Levels of the first nesting level
  n.cg <- length(Clevel)
  size.cg <- array(0, n.cg)
  for (i in 1:n.cg) size.cg[i] <- sum(Cgrps == Clevel[i])
  # Number of the observations by subject
}
else {
```

```
  n.cg <- 1
  Cind <- 1:n
}
if (is.null(fit$modelStruct$varStruct))
  w <- rep(fit$sigma, n)
else {
  w <- 1/nlme::varWeights(fit$modelStruct$varStruct)
  group.name <- names(fit$groups)
  order.txt <- paste("ind<-order(data[[\"", group.name[1],
                     "\"]]", sep = "")
  if (length(fit$groups) > 1)
    for (i in 2:length(fit$groups)) order.txt <- paste(order.txt,
            ",data[[\"", group.name[i], "\"]]", sep = "")
  order.txt <- paste(order.txt, ")")
  eval(parse(text = order.txt))
  w[ind] <- w
  w <- w * fit$sigma
}
w <- w[Cind]
if (is.null(fit$modelStruct$corStruct))
  lR <- array(1, n)
else {
  c.m <- nlme::corMatrix(fit$modelStruct$corStruct)
  if (!is.list(c.m)) {
    lR <- c.m
    lR <- lR[Cind, ]
    lR <- lR[, Cind]
  }
  else {
    lR <- list()
    ind <- list()
    for (i in 1:n.cg) {
      lR[[i]] <- matrix(0, size.cg[i], size.cg[i])
```

```
      ind[[i]] <- 1:size.cg[i]
    }
    Roff <- cumsum(c(1, size.cg))
    gr.name <- names(c.m)
    n.g <- length(c.m)
    j0 <- rep(1, n.cg)
    ii <- 1:n
    for (i in 1:n.g) {
      Clev <- unique(Cgrps[grps == gr.name[i]])
      if (length(Clev) > 1)
        stop("inner groupings not nested in outer!!")
      k <- (1:n.cg)[Clevel == Clev]
      j1 <- j0[k] + nrow(c.m[[i]]) - 1
      lR[[k]][j0[k]:j1, j0[k]:j1] <- c.m[[i]]
      ind1 <- ii[grps == gr.name[i]]
      ind2 <- rCind[ind1]
      ind[[k]][j0[k]:j1] <- ind2 - Roff[k] + 1
      j0[k] <- j1 + 1
    }
    for (k in 1:n.cg) {
      lR[[k]][ind[[k]], ] <- lR[[k]]
      lR[[k]][, ind[[k]]] <- lR[[k]]
    }
  }
}
if (is.list(lR)) {
  for (i in 1:n.cg) {
    wi <- w[Roff[i]:(Roff[i] + size.cg[i] - 1)]
    lR[[i]] <- as.vector(wi) * t(as.vector(wi) * lR[[i]]) # Matrix lR
  }
}
else if (is.matrix(lR)) {
  lR <- as.vector(w) * t(as.vector(w) * lR)
```

```
  }
  else {
    lR <- w^2 * lR
  }
  if (is.list(lR)) {
    R <- lR[[1]]
    for (k in 2:n.cg) {
      R <- bdiag(R,lR[[k]])
    }
    R <- as.matrix(R)
  }
  else{
    R <- diag(lR)
  }
}else{
  R <- getVarCov(fit,type = "conditional",individual = 1)[[1]]
  for (i in 2:length(as.numeric(unique(id)))) {
    R <- as.matrix(bdiag(R,getVarCov(fit,
           type = "conditional",individual = i)[[1]] ) )
  }
}
########################################
## Construction of covariance matrix of Y (Here denoted as V;
##in the paper it is denoted \Omega)
####################################

V <- (Z %*% Gam %*% t(Z)) + R
iV <- solve(V)

####################################
## Construction of the Q matrix
####################################
```

```
varbeta <- solve((t(X) %*% iV %*% X))
Q <- (iV - iV %*% X %*% (varbeta) %*% t(X) %*% iV )
zq <- t(Z) %*% Q
norm.frob.ZtQ <- sum(diag(zq %*% t(zq)))

###################################
## EBLUE and EBLUP
###################################

eblue <- as.vector(fixef(fit))
eblup <- Gam %*% t(Z) %*% iV %*% (y - X %*% eblue)

###################################
## Residual analysis
###################################

predm <- X %*% eblue
# Predicted values for expected response
predi <- X %*% eblue + Z %*% eblup         # Predicted values for units
resc <- (y - predi)                        # Conditional residuals

###################################
## Variance of conditional residuals
###################################

var.resc <- R %*% Q %*% R

###################################
## Standardized conditional residuals
###################################

rescp <- resc/sqrt(diag(var.resc))
```

```
##################################
## Least confounded residuals
##################################

R.half <- sqrt.matrix(R)
auxqn <- eigen((R.half %*% Q %*% R.half), symmetric = T, only.values = FALSE)

lt <- sqrt(solve(diag((auxqn$values[1:(N-p)])))) %*%
t(auxqn$vectors[1:N,1:(N-p)]) %*% solve(sqrt.matrix(R[1:N,1:N]))

var.resmcp <- lt %*% var.resc[1:N,1:N] %*% t(lt)
resmcp <- (lt %*% resc[1:N] )/sqrt(diag(var.resmcp))

################################
## This function constructs QQ plots for normality of random eefects
################################

qqPlot2 <- function(x, distribution="norm", ..., ylab=deparse(substitute(x)),
                  xlab=paste(distribution, "quantiles"), main = NULL,
                  las = par("las"),
                  envelope = .95,
                  col = palette()[1],
                  col.lines = palette()[2], lwd = 2, pch = 1, cex = par("cex"),
                  cex.lab = par("cex.lab"), cex.axis = par("cex.axis"),
                  line = c("quartiles", "robust", "none"),
                  labels = if (!is.null(names(x))) names(x) else seq(along = x),
                  id.method = "y",
                  id.n = if (id.method[1] == "identify") Inf else 0,
                  id.cex = 1, id.col=palette()[1], grid = TRUE)
{
  line <- match.arg(line)
  good <- !is.na(x)
  ord <- order(x[good])
```

```
ord.x <- x[good][ord]
ord.lab <- labels[good][ord]
q.function <- eval(parse(text = paste("q", distribution, sep = "")))
d.function <- eval(parse(text = paste("d", distribution, sep = "")))
n <- length(ord.x)
P <- ppoints(n)
z <- q.function(P, ...)
plot(z, ord.x, type = "n", xlab = xlab, ylab = ylab, main = main, las = las,
cex.lab = cex.lab, cex.axis = cex.axis)
if (grid) {
  grid(lty = 1, equilogs = FALSE)
  box()}
points(z, ord.x, col = col, pch = pch, cex = cex)
if (line  == "quartiles" || line == "none") {
  Q.x <- quantile(ord.x, c(.25,.75))
  Q.z <- q.function(c(.25,.75), ...)
  b <- (Q.x[2] - Q.x[1])/(Q.z[2] - Q.z[1])
  a <- Q.x[1] - b*Q.z[1]
  abline(a, b, col = col.lines, lwd = lwd)
}
if (line == "robust") {
  coef <- coef(rlm(ord.x ~ z))
  a <- coef[1]
  b <- coef[2]
  abline(a, b)
}
conf <- if (envelope == FALSE) .95 else envelope
zz <- qnorm(1 - (1 - conf)/2)
SE <- (b/d.function(z, ...))*sqrt(P*(1 - P)/n)
fit.value <- a + b*z
upper <- fit.value + zz*SE
lower <- fit.value - zz*SE
if (envelope != FALSE) {
```

```
      lines(z, upper, lty = 2, lwd = lwd, col = col.lines)
      lines(z, lower, lty = 2, lwd = lwd, col = col.lines)
    }
  }

  ##################################
  ## This function constructs the diagnostic plots
  ##################################

  plotg = function(plotid){
    cat("\n To select the graphic use plotid \n
1- Resíduos condicionais padronizados versus preditos e histograma
2- Normal QQ plot e histograma para os resíduos com confund min padronizados
\n")
    cat("\n Graph plotting", plotid)

    if (plotid == 1)
    {
      par(mfrow = c(1,2), mar = c(11, 5, 1, 2))
      plot(predi, rescp, xlab = expression(paste("Preditos")),
           cex = 1.2, cex.lab = 1.3, cex.axis = 1.3,
           ylab = expression(paste("Resíduos condicionais padronizados")),
           pch = 20, ylim = c(-1.3*max(abs(range(rescp))),
                              1.3*max(abs(range(rescp)))))
      abline(h = 0,lty = 3)
      abline(h = limit,lty = 3)
      abline(h = -limit,lty = 3)
      hist(rescp, freq = F,breaks = c(-7:7), main = "",
           xlab = expression(paste("Resíduos condicionais padronizados")),
           cex = 1.0, cex.lab = 1.3, cex.axis = 1.3)
    }

    if (plotid == 2)
```

```
    {
      par(mfrow = c(1,2), mar = c(11, 5, 3, 2))
      qqPlot2(resmcp, ylab = "Resíduos com confundimento mínimo padronizados",
              xlab = "Quantis da N(0,1)", pch = 20, cex = 1.2, cex.lab = 1.2,
              cex.axis = 1.3)
      hist(resmcp, freq = F,breaks = c(-6:6),
           xlab = "Resíduos com confundimento mínimo padronizados",
           main = "", cex = 1.0, cex.lab = 1.2, cex.axis = 1.2, pch = 20)
    }
   }
  if (is.null(plotid)) {
    cat("\n To choose plot, select plotid \n
1- Resíduos condicionais padronizados versus preditos e histograma correspondente
2- Normal QQ plot e histograma para os resíduos com confundimento mínimo padronizados
        \n")
    return(1);
  }

  ############################
  # Generation of diagnostic plots
  ############################
  for (g in plotid) {
    plotg(g)
    cat("\n Press ENTER to continue...")
    readline()
  }

  useful.results <- list(
    std.conditional.residuals = cbind(Subject = id,Predicted = as.numeric(rescp)),
    least.confounded.residuals = cbind(l.c.r = as.numeric(resmcp)))
}
```

B.2 Organizando os dados e seus efeitos fixos

```
#Pacotes:

library(dplyr)
library(nlme)
library(ggplot2)
library(lme4)

####################DADOS:

dados <- read.table(paste("dados.csv", sep = ""), h=T, sep = ",")

dados$area <- recode(dados$area,
                     "A" = 1, "B" = 2, "C" = 3, "D" = 4, "E" = 5)

dados$prof <- recode(dados$prof,
                     "a" = 5, "b" = 10, "c" = 15, "d" = 20,
                     "e" = 25, "f" = 30, "g" = 35, "h" = 40)

#####criando replicas:

table(dados$prof) #1103 individuos com medidas completas
indiv <- rep(1:1103,each = 8)

dados <- dados %>% mutate(indiv = indiv)
dados$area <- as.factor(dados$area)
dados$penet <- as.factor(dados$penet)
dados$umid <- as.factor(dados$umid)

attach(dados)
```

```
area.new <- as.numeric(area)
area.new[area == 2] <- 1
area.new[area == 3] <- 1 #(juntando 1, 2 e 3: grafico de linhas descritiva)

area.new <- factor(area.new, levels = c(1,4,5))

############ criando as dummies das interacoes duplas:

area4_prof20 <- as.matrix(ifelse(dados$area == 4 & dados$prof == 20,1,0))
area4_prof25 <- as.matrix(ifelse(dados$area == 4 & dados$prof == 25,1,0))
area4_prof30 <- as.matrix(ifelse(dados$area == 4 & dados$prof == 30,1,0))
area4_prof35 <- as.matrix(ifelse(dados$area == 4 & dados$prof == 35,1,0))
area4_prof40 <- as.matrix(ifelse(dados$area == 4 & dados$prof == 40,1,0))

area5_prof20 <- as.matrix(ifelse(dados$area == 5 & dados$prof == 20,1,0))
area5_prof25 <- as.matrix(ifelse(dados$area == 5 & dados$prof == 25,1,0))
area5_prof30 <- as.matrix(ifelse(dados$area == 5 & dados$prof == 30,1,0))
area5_prof35 <- as.matrix(ifelse(dados$area == 5 & dados$prof == 35,1,0))
area5_prof40 <- as.matrix(ifelse(dados$area == 5 & dados$prof == 40,1,0))
```

B.3 Modelo misto Normal

```
##############Modelo Normal:

modelo<-lme(log(resist+1) ~ area.new+penet+umid+prof+area4_prof20+
              area4_prof25+area4_prof30+area4_prof35+area4_prof40+
              area5_prof20+area5_prof25+area5_prof30+area5_prof35+
              area5_prof40, random = ~ 1 | indiv,
            correlation = corAR1(form = ~ 1 | indiv),na.action=na.omit,
            data=dados)
```

```
summary(modelo)

beta<-round(modelo[[4]]$fixed,3) # estimativa dos efeitos fixos
beta
b<-modelo[[4]]$random$indiv # predicoes efeitos aleatorios para cada perfuracao
b

boxplot(b, ylab = "Valores preditos dos efeitos aleatórios", col = "lightblue",
        xaxt = "n")

R<-getVarCov(modelo, type="conditional") # matriz de var-cov dos erros
R

#Diagnostico:

modelo_diag <- residdiag.nlme(modelo,limit = 3, plotid = 1:2)
#usando a funcao residdiag
#limit sao os limites para o grafico resid vs pred

#o plot 1: resid cond padronizados vs pred
#o plot 2: normalidade usando os res com confundimento min padronizados

residuos_cond <- modelo_diag$std.conditional.residuals[,2]
observ <- 1:nrow(dados)

ggplot(dados, aes(x = observ, y = residuos_cond)) +
  geom_point() +
  labs(
    title = "",
    x = "Observações",
    y = "Resíduos condicionais padronizados"
  ) + geom_hline(yintercept = 3, linetype = "dashed", color = "black") +
  geom_hline(yintercept = -3, linetype = "dashed", color = "black") +
```

```
  theme_bw()

par(mfrow = c(1,1), mar = c(4, 4, 4, 4))
boxplot(b, ylab = "Valores preditos dos efeitos aleatórios", col = "lightblue",
        xaxt = "n")
hist(b, xlab = "Valores preditos dos efeitos aleatórios", col = "lightblue",
     main = "")

#Analisando outliers erros:

residuos_ruins <- which(residuos_cond <= -3 | residuos_cond >= 3)
residuos_ruins
dados_ruins <- dados[residuos_ruins,]

table(area[residuos_ruins])
table(dados_ruins$umid)
table(dados_ruins$penet)
table(dados_ruins$prof)

table(area[residuos_ruins], dados_ruins$prof)

#Analisando outliers efeitos:

# Limites do boxplot
LIR <- boxplot.stats(b)$stats[1]
LSR <- boxplot.stats(b)$stats[5]

efeitos_ruins <- which(b < LIR | b > LSR) #coletando os outliers
efeitos_ruins
dados_ruins_2 <- dados[match(efeitos_ruins, dados$indiv), ]
#selecionando os indiv que são outliers
dados_ruins_2 <- dados_ruins_2[,- c(4:5)]
#excluindo a coluna "resist" e "prof"
```

```
table(dados_ruins_2$area,dados_ruins_2$penet, dados_ruins_2$umid)
#verificando os outliers
```

B.4 Modelo misto Gama

```
##########Modelo Gama:

profcm <- prof/100 #colocar a prof em centimetros (para convergir)

modelo.gama <- glmer((resist+1) ~ area.new+penet+umid+profcm+area4_prof20+
                       area4_prof25+area4_prof30+area4_prof35+area4_prof40+
                       area5_prof20+area5_prof25+area5_prof30+area5_prof35+
                       area5_prof40 + (1 | indiv),
                     data = dados, family = Gamma(link = "log"),
                    glmerControl(optimizer = "bobyqa",
                                 optCtrl = list(maxfun = 100000)))

#escolhendo o otimizador "bobyqa" e aumentando a quantidade de iteracoes
#para o modelo convergir

summary(modelo.gama, correlation = TRUE)
## coeficientes da estrutura aleatoria
randcoef <- coef(modelo.gama)$indiv[1] #intercepto aleatorio
randcoef
boxplot(randcoef, ylab = "Valores preditos dos efeitos aleatórios",
col = "lightblue", xaxt = "n")

## coeficientes da estrutura fixa do modelo
fixcoef  <- fixef(modelo.gama)
fixcoef
```