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

MÉTODOS DE AGRUPAMENTO LONGITUDINAL: UMA APLICAÇÃO EM DADOS DE ONDAS ULTRASSÔNICAS

Giovanna Passos Nesterick

Trabalho de Conclusão de Curso

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

MÉTODOS DE AGRUPAMENTO LONGITUDINAL: UMA APLICAÇÃO EM DADOS DE ONDAS ULTRASSÔNICAS

Giovanna Passos Nesterick

Orientadora: Profª. Drª. Daiane Aparecida Zuanetti

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos

Novembro de 2021

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

LONGITUDINAL CLUSTERING METHODS: AN APPLICATION TO ULTRASONIC WAVE DATA

Giovanna Passos Nesterick
Advisor: Prof. Dr. 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
November 2021

Giovanna Passos Nesterick

MÉTODOS DE AGRUPAMENTO LONGITUDINAL: UMA APLICAÇÃO EM DADOS DE ONDAS ULTRASSÔNICAS

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

Aprovado em 12 de novembro de 2021

Banca Examinadora:

- Profª.Drª. Daiane Aparecida Zuanetti
- Prof. Dr. Rafael Bassi Stern
- Profª.Drª. Rosineide Fernando da Paz

Aos meus pais, Cristiana e Leandro; ao meu irmão Rennan e ao meu companheiro Matheus. Amo vocês, obrigada por tudo.

Agradecimentos

Agradeço aos meu pais, Leandro *(in memorian)* e Cristiana, por toda educação, carinho e apoio, por terem me incentivado nos estudos desde pequena, assim como me incentivaram em tudo na vida. Obrigada mãe, por me ajudar de diversas formas durante esses anos de graduação e pela confiança, mesmo em cidades distantes. Pai, espero que esteja orgulhoso!

Ao meu irmão Rennan, que sempre esteve comigo, mesmo a distância, dando conselhos, momentos de descontração e, principalmente, muito carinho. Você é um exemplo para mim.

Ao meu companheiro Matheus, pela paciência e compreensão neste último ano, por estar comigo nos momentos bons e, principalmente, nos ruins, dando todo o apoio emocional necessário. Sem seu suporte, esta fase seria muito mais difícil, muito obrigada por todo amor!

Aos meu familiares que me ajudaram dando apoio, carinho e forças.

À minha orientadora, professora Daiane Zuanetti, pela paciência, incentivo, sugestões e toda dedicação durante cada etapa deste estudo. Foi muito especial ter uma pessoa tão gentil, educada e inteligente como você trabalhando comigo, você é um exemplo!

Agradeço também aos meus colegas de curso, que se tornaram amigos para vida, por todos os momentos de descontração durante a graduação, por todas as ajudas e suportes, tornando esta fase inesquecível. Ao meu amigo inseparável de matérias e da vida, Matheus Felix, que esteve comigo em, literalmente, todos os momentos. Às minhas amigas, Luri e Alícia, com quem dividi a casa e muitos momentos memoráveis. Às minhas amigas, Júlia Beltramini, Graziela Valero, Natalia Tomazella e Pamela Peruchi e ao meu amigo Renan Rodrigues, pela companhia, risadas e bons momentos. E por fim, mas não menos importante, aos colegas do PET Estatística, por todo convívio, troca de experiências e por tornarem o projeto de extensão mais leve e tão importante.

Resumo

A preservação de construções enfrenta diversos desafios, como os desgastes causados pelo tempo, variáveis atmosféricas, assentamentos de solo ou até falta de manutenção adequada. Em alguns casos, como em patrimônios históricos, não é possível realizar testes destrutivos tradicionais para verificar a qualidade física e mecânica dos materiais e da estrutura. Assim, com o intuito de preservar a estrutura de construções, testes não destrutivos são um grande avanço para as áreas de engenharia civil e mecânica e estes testes através de ondas ultrassônicas, em especial, têm se mostrado eficientes em fornecer informações necessárias para detectar danos no interior de alvenarias e, assim, prever situações.

O objetivo deste estudo é analisar e descrever o comportamento de propagação de ondas ultrassônicas em edifícios de alvenaria. O estudo consiste em identificar por métodos não supervisionados de agrupamento, diferentes perfis no tempo de propagação de ondas em paredes de alvenaria. Os métodos utilizados são *k-means* longitudinal e GCKM (do inglês *growth curve model* combinado com *k-means*), sendo um mais conhecido e amplamente utilizado e o outro um pouco mais complexo e com bons resultados na literatura, respectivamente.

Palavras-chave: *Estabilidade de construções, GCKM, k-means longitudinal, métodos não destrutivos.*

Abstract

The preservation of buildings faces several challenges, such as the wear caused by weather, atmospheric conditions, soil settlement or even lack of proper maintenance. In some cases, such as historical heritage, it is not possible to carry out traditional destructive tests to verify the physical and mechanical quality of materials and structure. Then, in order to preserve the structure of buildings, non-destructive tests are a great advance in civil and mechanical engineering and these tests using ultrasonic waves, in particular, have been shown to be efficient in providing necessary information to detect damage in the interior of masonry and, so, predict situations.

The goal of this study is to analyze and describe the propagation behavior of ultrasonic waves in masonry buildings. The study consists of identifying, by unsupervised clustering methods, different wave propagation time profiles in masonry walls. The methods used are longitudinal *k-means* and GCKM (*growth curve model* combined with *k-means*), one being better known and widely used and the other a little more complex and with good results in the literature, respectively.

Keywords: *Building stability, GCKM, longitudinal k-means, non-destructive methods.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O patrimônio cultural é composto por um extenso e variado conjunto de bens culturais, que um país recebe (Bonilla, 2003). Contudo, sua preservação enfrenta diversos desafios. Desgastes causados pelas variáveis atmosféricas, assentamentos do solo, incêndios e inclusive a falta de manutenção contínua, fazem com que grande parte desses patrimônios sejam colocados em risco devido a problemas estruturais que afetam sua segurança e a dos indivíduos que os utilizam.

Os testes destrutivos tradicionais não são possíveis de se realizar na maioria dos edifícios históricos, por isso é necessário selecionar testes não destrutivos (NDT) que possibilitem a caracterização física e mecânica dos materiais e do comportamento da estrutura (Rubens, 2019). Devido seu diferencial, este tipo de ensaio tem sido amplamente utilizado na indústria moderna em todo mundo para a avaliação da qualidade e detecção de variações nas estruturas, pequenas falhas superficiais, presença de trincas e outras interrupções físicas, medida de espessura de materiais e determinação de algumas das propriedades de materiais industriais (Ramirez, 2015).

Uma forma de realizar o NDT é através de ondas ultrassônicas que podem fornecer informações necessárias para detectar danos no interior de paredes e, assim, prever sua situação. Um método cada vez mais utilizado é a medição da velocidade de propagação dessas ondas no interior de paredes, como podemos observar em Ramirez (2015).

Diante do exposto, a fim de contribuir para os avanços no campo da avaliação do NDT, o objetivo principal deste estudo é analisar e descrever o comportamento de propagação de ondas ultrassônicas em edifícios de alvenaria. A ideia básica é identificar, por métodos não supervisionados de agrupamento, diferentes perfis no tempo de propagação de ondas em paredes de alvenaria, melhor descrito no Capítulo 3. Caso observe-se que o compor-

tamento das ondas possa ser diferenciado através de grupos, elas podem ser utilizadas para avaliar a qualidade da alvenaria e serem aplicadas como base para engenheiros e pesquisadores da área.

Entre os métodos disponíveis para análise de agrupamento de dados longitudinais ou medidas repetidas, que é o caso dos dados analisados nesse trabalho, destacam-se o método *k-means* longitudinal (Genolini e Falissard, 2011) e o método GCKM, do inglês *growth curve model* combinado com *k-means*, (Anderson e Gerbing, 1988).

O desempenho destes dois métodos é comparado com o de outros três métodos por Den Teuling *et al.* (2020) em situações nas quais a mudança de tendências nos diferentes perfis acontece lentamente ao longo do tempo. Neste estudo, o método GCKM obteve um ótimo desempenho, sendo preferido para grandes conjuntos de dados por sua eficiência computacional e pela performance observada.

Twisk e Hoekstra (2012) mostra um estudo de caso no qual a metodologia GCKM é empregada para detectar trajetórias de desenvolvimento em um conjunto de dados epidemiológicos longitudinais. Na análise, também, é realizada a comparação com outros quatro métodos de agrupamento longitudinal.

Em outro estudo recente, Garcia (2020) utilizou de métodos de agrupamento especificamente para dados longitudinais, a fim de descobrir tendências ou padrões importantes que possam auxiliar a comunidade científica na batalha do COVID-19, sendo o *k-means* longitudinal o método com os melhores resultados. Neste estudo, três segmentos foram identificados e validados, e todos se diferem significativamente com base em variáveis psicográficas, comportamentais, geográficas e demográficas.

O trabalho está organizado da seguinte forma. O Capítulo 2 apresenta os métodos de agrupamento longitudinal estudados, separados pela Seção 2.1, referente ao *k-means* longitudinal, e a Seção 2.2, referente ao GCKM. O Capítulo 3 apresenta como o banco de dados utilizado está estruturado, assim como suas análises descritivas. Em seguida, o Capítulo 4 contém os resultados e análises sobre os agrupamentos feitos a partir dos métodos utilizados, além de suas comparações. Por fim, o Capítulo 5 apresenta as conclusões tomadas neste estudo.

Capítulo 2

Métodos de agrupamento longitudinal

Nesse capítulo, são apresentadas as duas metodologias de agrupamento de dados longitudinais ou medidas repetidas que serão utilizadas nesse trabalho. São elas: o *k-means* longitudinal e o GCKM.

2.1 *k-means* longitudinal

O *k-means* longitudinal (KML) é uma abordagem simples comumente usada (Genolini e Falissard, 2010) por assumir que as observações são condicionalmente independentes. Neste método, é pré-definido um número fixo de grupos, sendo cada grupo representado pelo seu centróide. O ponto inicial do centróide de cada grupo pode variar, o que permite estabelecer condições iniciais de dependência. O agrupamento das unidades amostrais é determinado pelo centróide mais próximo, a cada iteração os centróides são recalculados considerando as unidades amostrais de cada grupo e as observações são realocadas para o centróide mais próximo. Deste modo, o algoritmo usa uma abordagem iterativa para chegar a uma solução.

Segundo Den Teuling *et al.* (2020), o algoritmo *k-means* longitudinal visa encontrar o particionamento $I_1, I_2, ..., I_G$ com $\cup_{g=1}^{G} I_g = I$, em que I é o conjunto de todas as unidades amostrais, e $I_g \cap I_h = \emptyset$ quando $g \neq h$, minimizando a variação dentro dos grupos e

maximizando a variância entre os grupos. Os G grupos formados são obtidos por

$$\underset{I_1,I_2,...,I_G}{\operatorname{argmin}} \sum_{g=1}^{G} \sum_{i \in I_g} \|\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_g\|^2, \tag{2.1}$$

com $\hat{\boldsymbol{\mu}}_g$ sendo o vetor de médias do grupo g, ou seja, o centróide, definido por $\hat{\boldsymbol{\mu}}_g = |I_g|^{-1} \sum_{i \in I_g} \boldsymbol{y}_i$ em que a soma é executada elemento a elemento de $\boldsymbol{y}_i$. Este método assume que a variância dentro de cada grupo é igual entre os grupos, portanto, quando subgrupos nos dados têm variações diferentes, os limites estimados do agrupamento provavelmente estarão errados (Den Teuling *et al.*, 2020).

Tradicionalmente, o *k-means* pode ser executado usando várias medidas de distância. Por Genolini e Falissard (2010), têm-se a Distância Euclidiana, $Dist(\boldsymbol{y}_i, \boldsymbol{y}_l) = \sqrt{\sum_{j=1}^{t} (y_{ij} - y_{lj})^2}$, sendo a métrica mais popular para dados contínuos e que será implementada neste estudo. Existem, também, outras métricas usuais, como a Distância Manhattan, dada por $Dist_M(\boldsymbol{y}_i, \boldsymbol{y}_l) = \sum_{j=1}^{t} |y_{ij} - y_{lj}|$ e a Distância de Camberra, dada por $Dist_C(\boldsymbol{y}_i, \boldsymbol{y}_l) = \sum_{j=1}^{t} \frac{|y_{ij} - y_{lj}|}{|y_{ij} + y_{lj}|}$.

2.1.1 Diferenças entre *k-means* e *k-means* longitudinal

É possível notar que as métricas do método *k-means* tradicional e do *k-means* longitudinal são matematicamente idênticas. O que distingue estes dois métodos é que as p variáveis diferentes no *k-means* são substituídas por uma variável observada em t momentos no *k-means* longitudinal.

Portanto, tem-se o *k-means*, em que $i = 1, ..., n$ (número de unidades amostrais) e $j = 1, ..., p$ (número das diferentes variáveis observadas para cada unidade) e o *k-means* longitudinal, em que $i = 1, ..., n$ (número de unidades amostrais) e $j = 1, ..., t$ (uma mesma variável observada ao longo de t momentos). Desta forma, pode-se dizer que os métodos são similares e reforçando que o *k-means* longitudinal não trata a dependência entre diferentes observações da mesma unidade amostral de forma especial.

2.1.2 Escolha do número ótimo de agrupamentos

Existem diversos índices para a escolha do melhor número de agrupamentos, contudo a maioria possui princípios semelhantes em que busca-se alta variabilidade entre os

agrupamentos e baixa variabilidade dentro de cada grupo.

O critério de Caliński e Harabasz (1974) representa estes índices da seguinte forma,

$$\mathbf{B} = \sum_{g=1}^{G} n_g(\hat{\boldsymbol{\mu}}_g - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_g - \hat{\boldsymbol{\mu}})^\top$$
$$\mathbf{W} = \sum_{g=1}^{G}\sum_{m=1}^{n_g}(\boldsymbol{y}_{gm} - \hat{\boldsymbol{\mu}}_g)(\boldsymbol{y}_{gm} - \hat{\boldsymbol{\mu}}_g)^\top$$

em que, o número ótimo de agrupamentos é o valor de G que maximize

$$C(G) = \frac{\text{Traço}(\mathbf{B})}{\text{Traço}(\mathbf{W})} \cdot \frac{n-G}{G-1}. \tag{2.2}$$

É possível notar a semelhança deste índice com a estatística F de um teste ANOVA, em que é comparada a distância das curvas médias de cada grupo e a curva média geral com a distância entre as curvas individuais e a curva média do seu grupo, representadas nesta notação por $\mathbf{B}$ e $\mathbf{W}$, respectivamente.

Este critério, também, possui algumas variantes, como:

- Variante de Kryszczuk (Kryszczuk e Hurley, 2010)

$$C_K(G) = \frac{\text{Traço}(\mathbf{B})}{\text{Traço}(\mathbf{W})} \cdot \frac{n-1}{n-G}; \tag{2.3}$$

- Variante de Genolini (Genolini *et al.*, 2015) :

$$C_G(G) = \frac{\text{Traço}(\mathbf{B})}{\text{Traço}(\mathbf{W})} \cdot \frac{n-G}{\sqrt{G-1}}. \tag{2.4}$$

Neste trabalho, será utilizado o critério de Caliński e Harabasz (1974) para definir o número ótimo de agrupamentos. Além destes, existem outros critérios na literatura, tais como os propostos por Ray e Turi (1999) e Davies e Bouldin (1979), por exemplo.

2.1.3 Métodos computacionais

Para a implementação do *k-means* será utilizado o pacote *kml* (Genolini *et al.*, 2016), a partir do *software* R, o qual é gratuito e foi uma linguagem amplamente utilizada durante toda a graduação.

Todas as funções citadas na Seção 2.1 estão implementadas neste pacote. Portanto a partir dele poderá ser escolhido o número de agrupamentos, pelo critério desejado,

considerando um ponto inicial que pode ser especificado. Também pode ser definido o número de vezes que a função deve ser executada, com diferentes condições iniciais, para cada número de agrupamento. Detalhes do pacote *kml* podem ser vistos em Genolini *et al.* (2016).

2.2 GCKM

O GCKM, do inglês *growth curve model* combinado com *k-means* (Anderson e Gerbing, 1988), é um método de agrupamento em duas etapas, sendo que a primeira etapa é um pré-agrupamento, no qual as unidades amostrais são segmentadas em grupos muito pequenos e, na segunda etapa, são reagrupados formando os sub-perfis finais segundo um número ideal de grupos. Este método apresenta vantagens como robustez contra valores ausentes, capacidade de lidar com trajetórias longas e flexibilidade de lidar com trajetórias de comprimentos variados entre indivíduos ou medidas em intervalos diferentes (Wang *et al.*, 2006).

A primeira etapa se dá pela estimação dos parâmetros de um modelo misto, também conhecido por modelo de curva de crescimento ou modelo de curva latente. Segundo Den Teuling *et al.* (2020), este modelo representa o conjunto de dados longitudinais em termos de uma trajetória média, descrita por meio dos efeitos fixos ao longo do tempo, e o desvio de cada curva um relação a essa trajetória, por meio de efeitos aleatórios.

Uma trajetória descrita em termos de um polinômio de ordem K e efeitos aleatórios em todos os termos é dada por

$$\begin{aligned} \boldsymbol{Y}_i &= \sum_{k=0}^{K} \gamma_{k,i} \boldsymbol{d}_i^k + \boldsymbol{\varepsilon}_i, \\ \gamma_{k,i} &= \beta_k + b_{k,i}, \end{aligned} \tag{2.5}$$

em que $\boldsymbol{Y}_i$ representa os t valores dos tempos de propagação associados à i-ésima unidade amostral, $\boldsymbol{d}_i^k$ são as distâncias em que as medidas da i-ésima unidade amostral são observadas e $\boldsymbol{\varepsilon}_i$ denota o vetor de erros aleatórios associado às medidas da i-ésima unidade amostral, descrevendo a variabilidade dentro das unidades amostrais. Além disso, β_ks representam os coeficientes fixos de regressão que definem a trajetória média única e $b_{k,i}$s são os coeficientes aleatórios de regressão que descrevem o desvio de cada curva individual em relação à trajetória única. Os efeitos aleatórios descrevem a associação entre as

observações da mesma unidade amostral.

Portanto, β_0 é o intercepto fixo, β_1 é o coeficiente fixo que acompanha o valor da distância em cada medida realizada, β_2 é o coeficiente fixo que acompanha a distância ao quadrado e, assim, sucessivamente. Já o $b_{0,i}$ é o desvio aleatório da i-ésima unidade amostral em relação ao intercepto fixo, $b_{1,i}$ é o desvio aleatório da i-ésima unidade amostral em relação ao efeito fixo da covariável distância, $b_{2,i}$ é o desvio aleatório da i-ésima unidade amostral em relação ao efeito fixo da covariável distância ao quadrado e, assim, sucessivamente.

Assume-se, que

$$\boldsymbol{b}_i \sim N_{K+1}(\mathbf{0}, \sigma_b^2); \qquad \boldsymbol{\varepsilon}_i \sim N_t(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{I}_t), \text{ independentes entre si e de } \boldsymbol{b}_i\text{s}, \tag{2.6}$$

em que σ_b^2 é uma matriz não estruturada, com $\sigma_{ii}^2 > 0$ e $\sigma_{ij} \in \mathbb{R}$, dada por:

$$\sigma_b^2 = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1(K+1)} \\ \sigma_{12} & \sigma_{22}^2 & \sigma_{23} & \cdots & \sigma_{2(K+1)} \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_{(K+1)1} & \sigma_{(K+1)2} & \sigma_{(K+1)3} & \cdots & \sigma_{(K+1)(K+1)}^2 \end{bmatrix}.$$

O grau do polinômio define a flexibilidade do polinômio em descrever as curvas observadas. Polinômios de menor grau são menos flexíveis, mas polinômios de maior grau são mais complexos e aumentam o número de parâmetros a ser estimado. Com base em Zuanetti *et al.* (2021), em que é estudado o mesmo banco de dados deste trabalho, utilizou-se um polinômio de grau 3, por ser suficiente para descrever o comportamento das curvas dos dados analisados e não ser muito complexo. Contudo, caso o pesquisador tenha interesse em verificar o grau mais adequado para o conjunto de dados em análise, podem ser utilizados critérios de seleção de modelo, como o Critério de Informação de Akaike (AIC) ou Critério de Informação Bayesiano (BIC).

O modelo misto pode ser estimado pele método da máxima verossimilhança e os valores de $b_{k,i}$ podem ser preditos juntamente com os estimadores. Este processo é melhor detalhado na Sub-Seção 2.2.1.

Após a estimação dos parâmetros por meio de um modelo misto, a segunda etapa do método GCKM é agrupar os valores preditos para os efeitos aleatórios por meio do *k-means* longitudinal. Portanto, na Equação (2.1), $\boldsymbol{y}_i$ será dado pelo vetor $(\hat{b_{0,i}}, \hat{b_{1,i}}, \hat{b_{2,i}}, \hat{b_{3,i}})$, assim tendo 4 medidas repetidas neste estudo. Para esta etapa, optou-se por utilizar a

Distância Euclidiana e para a escolha do número ótimo de agrupamentos, o critério de Caliński e Harabasz (1974). Detalhes podem ser obtidos na Seção 2.1.

2.2.1 Estimação por máxima verossimilhança

Por Singer e Andrade (1986), o método de máxima verossimilhança consiste em obter os estimadores dos parâmetros por meio da maximização do logaritmo da verossimilhança marginal dos dados.

Com o intuito de facilitar o entendimento, o Modelo (2.5), também, pode ser escrito da seguinte forma

$$\boldsymbol{Y}_i = \sum_{k=0}^{K} \beta_k \boldsymbol{d}_i^k + \sum_{k=0}^{K} b_{k,i} \boldsymbol{d}_i^k + \boldsymbol{\varepsilon}_i. \tag{2.7}$$

Desta forma, (2.7) pode ser comparado com a notação mais conhecida de modelos mistos, dada por

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{2.8}$$

assim tendo, com as mesmas suposições de (2.6),

$$\begin{aligned} E(\boldsymbol{Y}_i) &= \boldsymbol{X}_i\boldsymbol{\beta} \\ Var(\boldsymbol{Y}_i) &= \boldsymbol{Z}_i\sigma_b^2\boldsymbol{Z}_i^\top + \sigma_\varepsilon^2\boldsymbol{I}_t, = \boldsymbol{\Omega}_i(\boldsymbol{\theta}). \end{aligned}$$

Alinhado estes pontos e seguindo Singer e Andrade (1986), tem-se que o logaritmo da verossimilhança marginal dos dados é dada por

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n} t\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\log|\boldsymbol{\Omega}_i(\boldsymbol{\theta})| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{2.9}$$

O primeiro passo para a maximização de (2.9) é igualar $\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\beta}}$ a 0, obtendo

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\sum_{i=1}^{n}\boldsymbol{X}_i^\top[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}\boldsymbol{X}_i]^{-1}[\sum_{i=1}^{n}\boldsymbol{X}_i^\top[\boldsymbol{\Omega}_i(\boldsymbol{\theta})]^{-1}\boldsymbol{y}_i]. \tag{2.10}$$

Como a derivada segunda é uma matriz definida negativa, $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ corresponde ao ponto em que $l(\boldsymbol{\beta},\boldsymbol{\theta})$ atinge o máximo. O segundo passo é substituir $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ na Equação (2.9), obtendo a função log-verossimilhança perfilada $l[\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta}]$.

O terceiro passo é igualar $\frac{\partial l[\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta}]}{\partial\boldsymbol{\theta}}$ a 0 e calcular sua derivada segunda, a fim de verificar se é definida negativa, ou seja, verificar se $\hat{\boldsymbol{\theta}}$ corresponde ao ponto máximo de

$l[\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}]$. Desta forma, sendo $\hat{\boldsymbol{\theta}}$ estimador de máxima verossimilhança de $\boldsymbol{\theta}$, ao substituir $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$ em (2.10), obtém-se o estimador de máxima verossimilhança de $\boldsymbol{\beta}$.

O valor predito do vetor $(\hat{b_{0,i}}, \hat{b_{1,i}}, ..., \hat{b_{K,i}})$ para cada unidade amostral é obtido através da distribuição condicional

$$f(\boldsymbol{b}_i|\boldsymbol{y}_i) \propto f(\boldsymbol{y}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i). \tag{2.11}$$

Assim, sabendo que $\boldsymbol{y}_i|\boldsymbol{b}_i \sim N_t(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \sigma^2_\varepsilon \boldsymbol{I}_t)$ e $\boldsymbol{b}_i \sim N_{K+1}(\boldsymbol{0}, \sigma^2_b))$, a distribuição (2.11) pode ser obtida da seguinte forma,

$$f(\boldsymbol{b}_i|\boldsymbol{y}_i) \propto exp\left\{-\frac{1}{2\sigma^2_\varepsilon}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\right)^\top \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\right)\right\} exp\left\{-\frac{1}{2}\boldsymbol{b}_i^\top (\sigma^2_b)^{-1}\boldsymbol{b}_i\right\}. \tag{2.12}$$

Trabalhando algebricamente na Expressão (2.12), tem-se que o valor predito é dado pela média da distribuição

$$\boldsymbol{b}_i|\boldsymbol{y}_i \sim N_{K+1}\left((\boldsymbol{Z}_i^\top\boldsymbol{Z}_i + \sigma^2_\varepsilon(\sigma^2_b)^{-1})^{-1}\boldsymbol{Z}_i^\top(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), (\frac{1}{\sigma^2_\varepsilon}\boldsymbol{Z}_i^\top\boldsymbol{Z}_i + (\sigma^2_b)^{-1})^{-1}\right).$$

Detalhes podem ser vistos em McCulloch e Searle (2001) e Rodrigues (2020).

2.2.2 Métodos computacionais

Para a implementação do GCKM serão utilizados dois pacotes, a partir do *software* R, o qual é gratuito e foi uma linguagem amplamente utilizada durante toda graduação.

Primeiramente, para a estimação do modelo misto, será utilizada a função *hlme*, do pacote *lcmm* (Proust-Lima *et al.*, 2020). Nela é possível indicar, por exemplo, o grau desejado do polinômio, definir as variáveis acompanhadas de efeito aleatório e obter as estimativas de máxima verossimilhança dos parâmetros. Detalhes podem ser vistos em Proust-Lima *et al.* (2015). Para a segunda etapa, será utilizado o pacote *kml* (Genolini *et al.*, 2016), da mesma forma como explicado na Sub-Seção 2.1.3.

Capítulo 3

Banco de dados

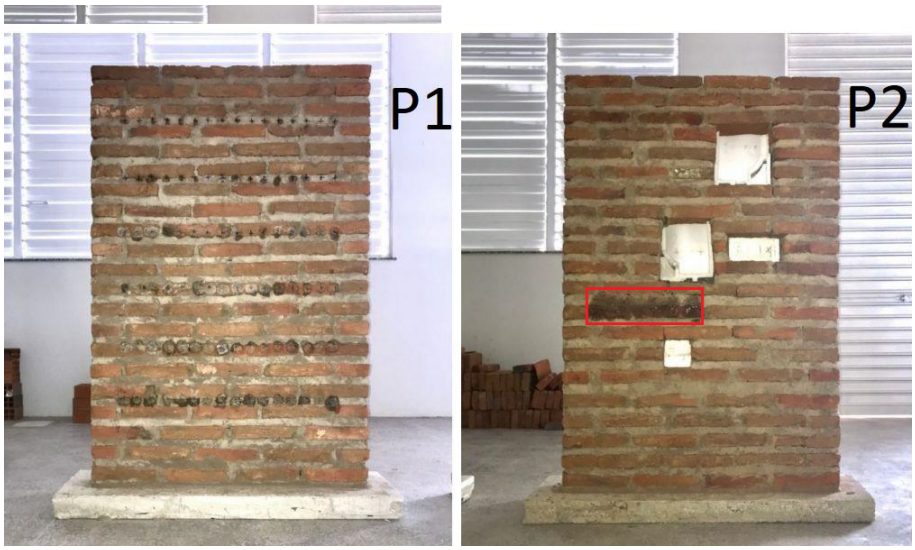
O banco de dados utilizado nesse trabalho também foi estudado por Zuanetti *et al.* (2021) e Rodrigues (2020). Sendo o primeiro uma análise do comportamento de propagação de ondas ultrassônicas através de uma abordagem Bayesiana semi-paramétrica e o segundo com foco em analisar as curvas de tempo de propagação das ondas através de modelos mistos e processos gaussianos e identificar variáveis preditoras relevantes na predição de diferentes perfis.

Os dados experimentais foram disponibilizados pelo Laboratório de Reabilitação e Durabilidade das Construções (LAREB) da Universidade Federal do Ceará (UFC) e coletados através do equipamento Pundit Lab-PROCEQ, com transdutores de 54 kHz, o qual possui um emissor de ondas ultrassônicas e calcula o tempo que elas demoram até chegar em seu receptor (Rodrigues, 2020).

As medições foram feitas em duas paredes semelhantes, que imitam construções antigas, normalmente encontradas no Estado do Ceará, ambas com a mesma dimensão: 1.50 $\times$ 1.00 $\times$ 0.135 m de altura, largura e espessura, respectivamente (Zuanetti *et al.*, 2021). Estas paredes foram elaboradas em laboratório, de forma que uma delas foi construída normalmente e a segunda com vazios internos, a fim de simular uma parede danificada com rachaduras e buracos, representadas por $P1$ e $P2$, respectivamente, conforme pode ser visto na Figura 3.1. Em seguida, ambas paredes foram revestidas com argamassa.

Seguindo as normas ABNT para construção civil, as paredes foram divididas em 12 quadrantes de 20 $\times$ 40 cm cada um. Cada parede foi demarcada com linhas horizontais em seis alturas 1.30, 1.10, 0.90, 0.70, 0.50 e 0.30 m, coincidindo com o centro de cada quadrante denotado aqui como $Q1$, $Q2$, $Q3$, $Q4$, $Q5$, $Q6$, $Q7$, $Q8$, $Q9$, $Q10$, $Q11$ e $Q12$.

Nas mesmas condições para as duas paredes, as ondas ultrassônicas foram emitidas



Fonte: Laboratório de Reabilitação e Durabilidade das Construções (LAREB).

Figura 3.1: Paredes $P1$ e $P2$ de alvenaria maciça.

da distância zero (origem da onda) e o tempo de sua propagação foi medido nas seis distâncias seguintes, na horizontal, sendo 10, 15, 20, 25, 30, 35 cm à frente do ponto de origem da onda.

Para conseguir uma maior variabilidade dos dados e controlar o erro de medição, 10 ondas (10 réplicas) foram emitidas dentro de cada quadrante de cada parede. Desta maneira, o conjunto de dados apresentaria o tempo de propagação de 240 ondas (10 réplicas x 12 quadrantes x 2 paredes) nas 6 distâncias analisadas, contudo houveram problemas nas medições da parede com vazios internos nos quadrantes $Q2$ e $Q4$, pois as ondas ultrassônicas não chegaram ao receptor, assim tendo 231 curvas observadas.

3.1 Análise descritiva

Para a primeira visão do banco de dados, foi realizada uma análise descritiva, por meio do *software* R. Mantendo o mesmo padrão, cada figura possui três gráficos, cada um para o quadrante especificado em seu título, representando as curvas das paredes com e sem vazios internos ($P1$ e $P2$) pelas cores azul e vermelha, respectivamente. No eixo das ordenadas está o tempo de propagação da onda, medido em 10^{-6} segundos e no eixo das abscissas estão as distâncias do receptor ao transmissor.

Como explicado, houveram problemas de medição da parede com vazios internos do segundo quadrante, observado na Figura 3.2, no gráfico do centro, em que apenas uma medição completa foi realizada. Mesmo assim, é possível notar que esta curva é semelhante

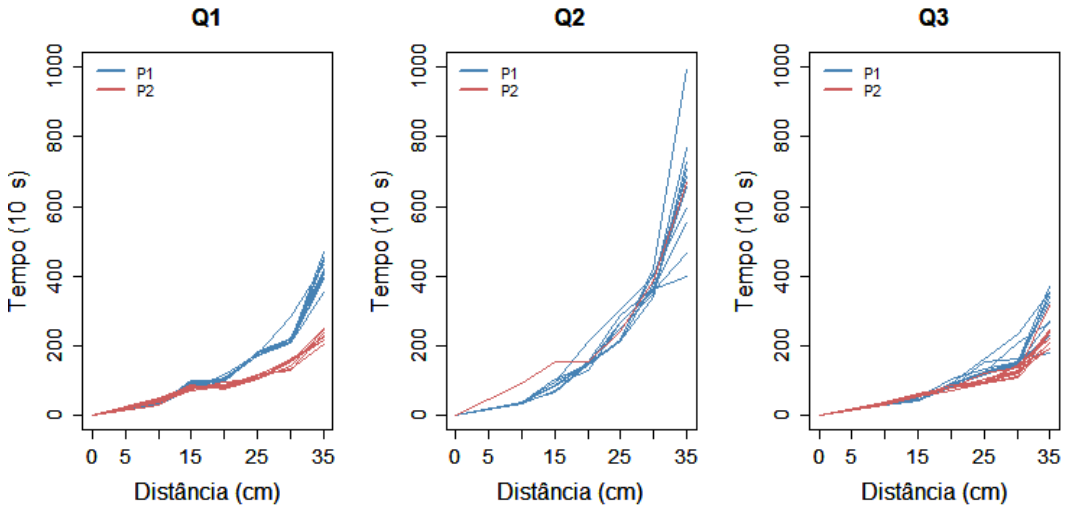


Figura 3.2: Perfis de propagação da onda para os quadrantes $Q1$, $Q2$ e $Q3$.

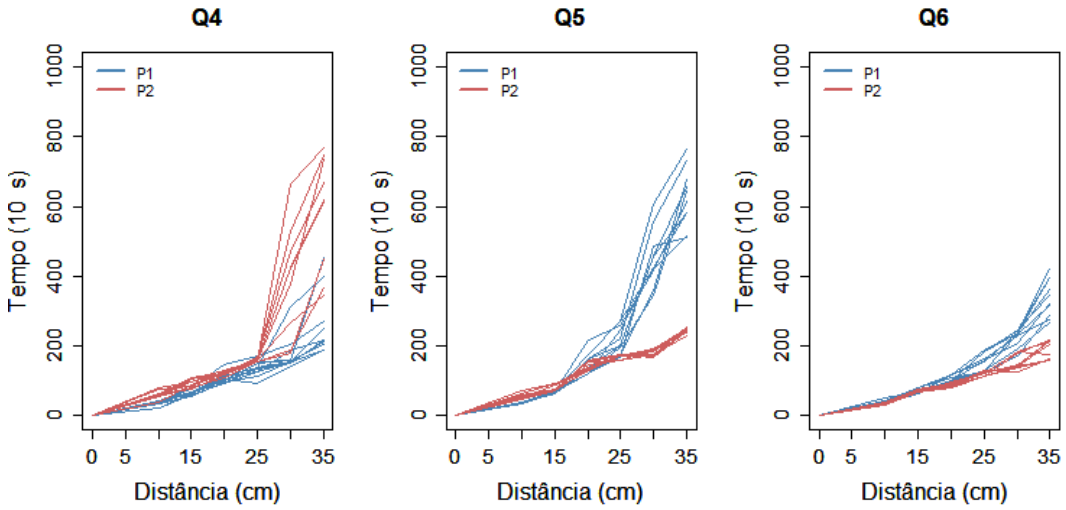


Figura 3.3: Perfis de propagação da onda para os quadrantes $Q4$, $Q5$ e $Q6$.

às curvas da parede maciça, seguindo a mesma tendência.

A onda ultrassônica apresenta velocidades distintas ao percorrer por diferentes materiais, sendo a velocidade em alvenarias maior do que no ar. Desta forma, é esperado que a onda demore menos tempo para passar na parede sem defeitos do que na parede com vazios internos. A partir da análise gráfica, é possível observar que na maioria dos quadrantes, o tempo de propagação da onda nas paredes sem vazios internos foi maior do que em paredes danificadas, não sendo o esperado. Dá-se destaque para as Figuras 3.2 e 3.3, nos quadrantes $Q1$, $Q5$ e $Q6$, as quais apresentam diferenças consideravelmente altas de tempo entre as paredes.

Alguns quadrantes, $Q3$, $Q7$ e $Q10$, nas Figuras 3.2, 3.4 e 3.5, apresentam ter tem-

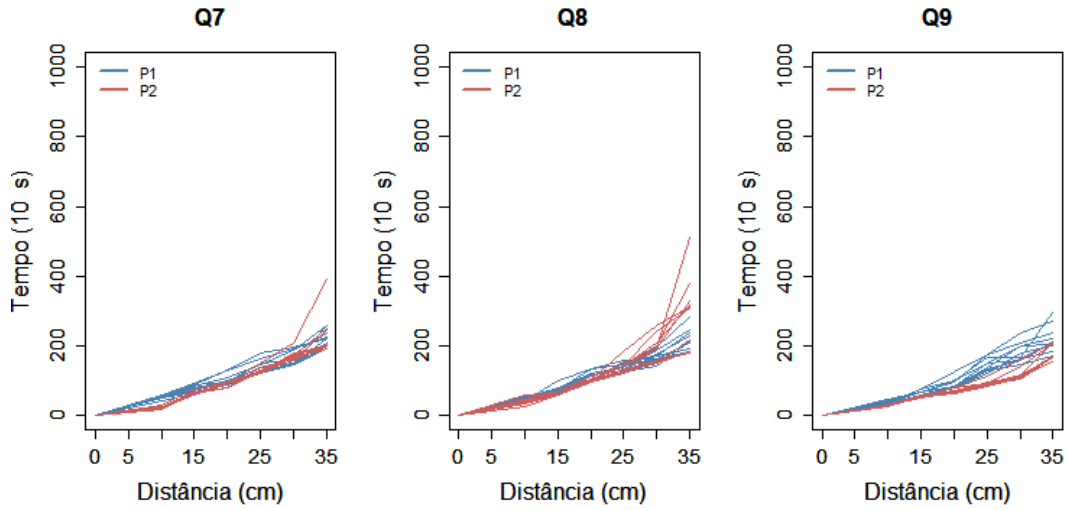


Figura 3.4: Perfis de propagação da onda para os quadrantes $Q7$, $Q8$ e $Q9$.

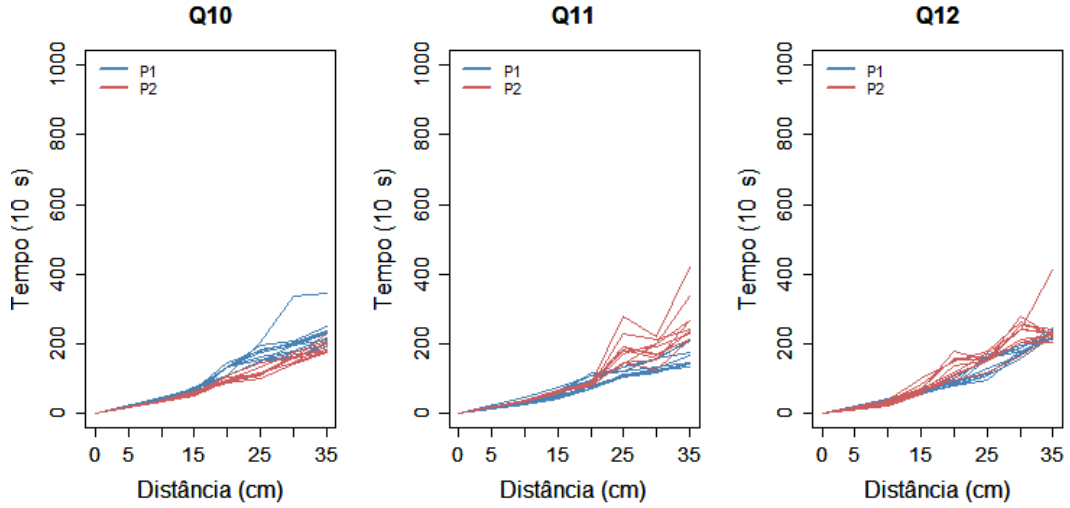


Figura 3.5: Perfis de propagação da onda para os quadrantes $Q10$, $Q11$ e $Q12$.

pos de propagação da onda muito similares entre as duas paredes, portanto a tendência de propagação parece não ser influenciada pelos vazios nestes casos. Isto pode ocorrer, também, quando ambas as paredes, naquele quadrante, não estão danificadas, pois assim não existem falhas que impactem na propagação da onda ultrassônica.

Existem casos em que o tempo de propagação das ondas nas paredes com vazios internos é maior do que nas paredes construídas normalmente, com destaque para o quadrante $Q4$, Figura 3.3, o qual apresenta uma grande variabilidade nos últimos pontos analisados.

No geral, a parede com vazios internos aparenta ter a variância entre as réplicas de um mesmo quadrante baixa, o que implica em pouco erro de medição. Alguns casos, como

nos quadrantes $Q4$, $Q8$, $Q11$ e $Q12$, Figuras 3.3, 3.4 e 3.5, merecem uma atenção maior, pois houve grande dispersão das curvas a partir de específicas distâncias (25, 25, 20, 15 cm, respectivamente), e com isso quebras na tendência.

Já na parede maciça, nos quadrantes $Q2$, $Q4$, $Q5$ e $Q6$, Figuras 3.2, 3.3, as curvas apresentam uma considerável variabilidade entre as réplicas, principalmente, a partir das maiores distâncias. Situações assim podem ocorrer, também, quando há fatores externos que estejam impactando na propagação da onda.

Portanto, alguns pontos da análise descritiva podem ser destacados. As ondas na parede sem vazios internos, geralmente, apresentam tempo de propagação maior do que na parede com imperfeições. Além disso, normalmente, quando há dispersão entre as ondas, ela acontece nas maiores distâncias.

Capítulo 4

Resultados

Neste capítulo, serão apresentados os resultados obtidos do *k-means* longitudinal e do GCKM na análise realizada com os tempos de propagação das ondas ultrassônicas através das paredes de alvenaria.

4.1 *k-means* longitudinal

A análise de dados por meio do *k-means* longitudinal envolve a definição de diversos parâmetros para o processamento do algoritmo, os quais já foram definidos na Subseção 2.1, e que são considerados pelo pacote *kml* (Genolini *et al.*, 2016) do *software* R. Neste estudo, para cada número de grupos, variando de 2 a 10, o processo de agrupamento foi executado 25 vezes com diferentes valores iniciais, a fim de garantir bons resultados. Vale ressaltar que, também, foi testado um maior número de grupos, assim como um maior número de repetições, contudo os resultados mantiveram os mesmos.

Entre as 231 curvas de medições observadas, para uma curva do quadrante $Q4$ da parede com vazios internos foram observadas apenas 4 medições. Para lidar com estes valores ausentes foi utilizado o método sugerido por Genolini e Falissard (2010), a Distância Euclidiana com ajuste *Gower*, calculada como $Dist_{Gower}(\boldsymbol{y}_i, \boldsymbol{y}_l) = \sqrt{\frac{1}{\sum w_{ilj}} \sum_{j=1}^{t} (y_{ij} - y_{lj})^2 w_{ilj}}$ para $\boldsymbol{y}_i$ e $\boldsymbol{y}_l$, em que $w_{ilj} = 0$ se y_{ij} ou y_{lj} ou ambos foram ausentes, e $w_{ilj} = 1$ caso contrário. Ou seja, essa distância descarta do seu cálculo as medições faltantes em uma ou ambas unidades amostrais. Além deste método, existem outras formas para lidar com valores ausentes no *k-means* longitudinal, que podem ser vistas em Genolini *et al.* (2015).

Para a definição do número ótimo de agrupamentos, foi utilizado o critério de Caliński e Harabasz (1974), em que o melhor número de grupos é aquele que maximiza a Expressão

(2.2). Pela Figura 4.1, observa-se um pico no valor da estatística em 3 grupos e depois um constante decaimento, o que mostra que dividir os dados em 3 grupos é a melhor escolha, segundo o critério. Além disto, pela Figura 4.2, gráfico da esquerda, nota-se que a divisão em 3 grupos se mostrou com valores maiores do que as demais quantidades de grupos na maioria das 25 repetições, o que confirma que esta é a divisão mais adequada, segundo o critério.

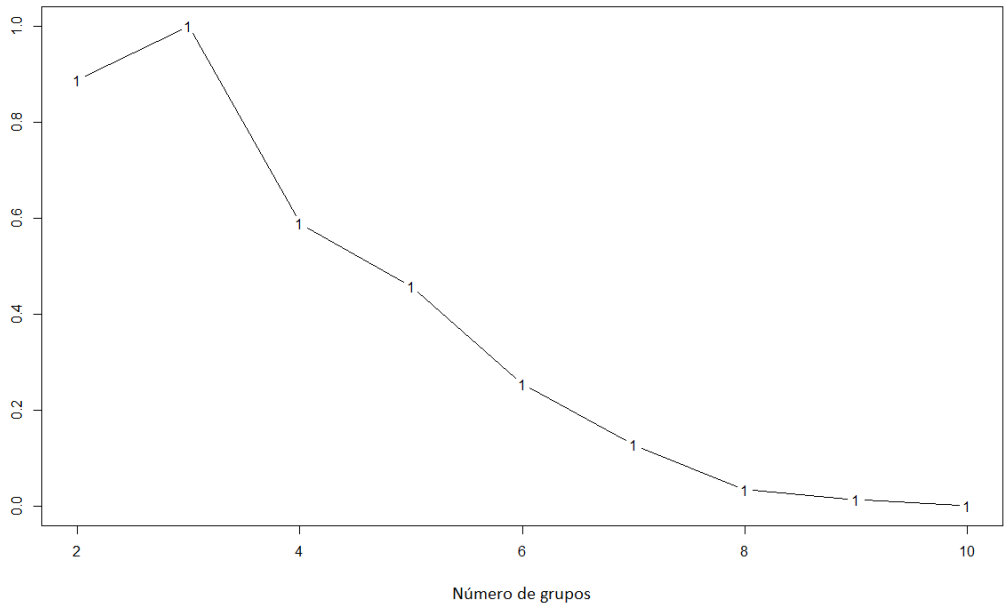


Figura 4.1: Gráfico para escolha do número ótimo de agrupamentos pelo critério de Caliński e Harabasz para o *k-means* longitudinal.

Desta forma, pela Figura 4.2, quadro da direita, tem-se as trajetórias médias de cada um dos grupos formados. Observa-se que o grupo A é composto por trajetórias lineares, com menores tempos de propagação da onda ultrassônica e o grupo B é composto por trajetórias lineares, com uma leve perda de tendência nas maiores distâncias, quando apresentam tempos mais elevados de propagação que as curvas do grupo A. Já o grupo C é composto pelas trajetórias mais distintas, com grande variabilidade, principalmente, nas maiores distâncias medidas. O grupo A é o maior, contendo 71.4% das curvas, enquanto o grupo B e C possuem tamanhos similares, com 17.7% e 10.8% das curvas, respectivamente.

Considerando as análises feitas na Subseção 3.1, é possível observar como as curvas foram alocadas nos 3 grupos, a partir das Tabelas 4.1 e 4.2. Na parede sem vazios internos, analisou-se que estas as curvas dos quadrantes $Q2$ e $Q5$ apresentam uma considerável variabilidade entre as réplicas, e pela Tabela 4.1 vê-se que a maioria delas foi alocada

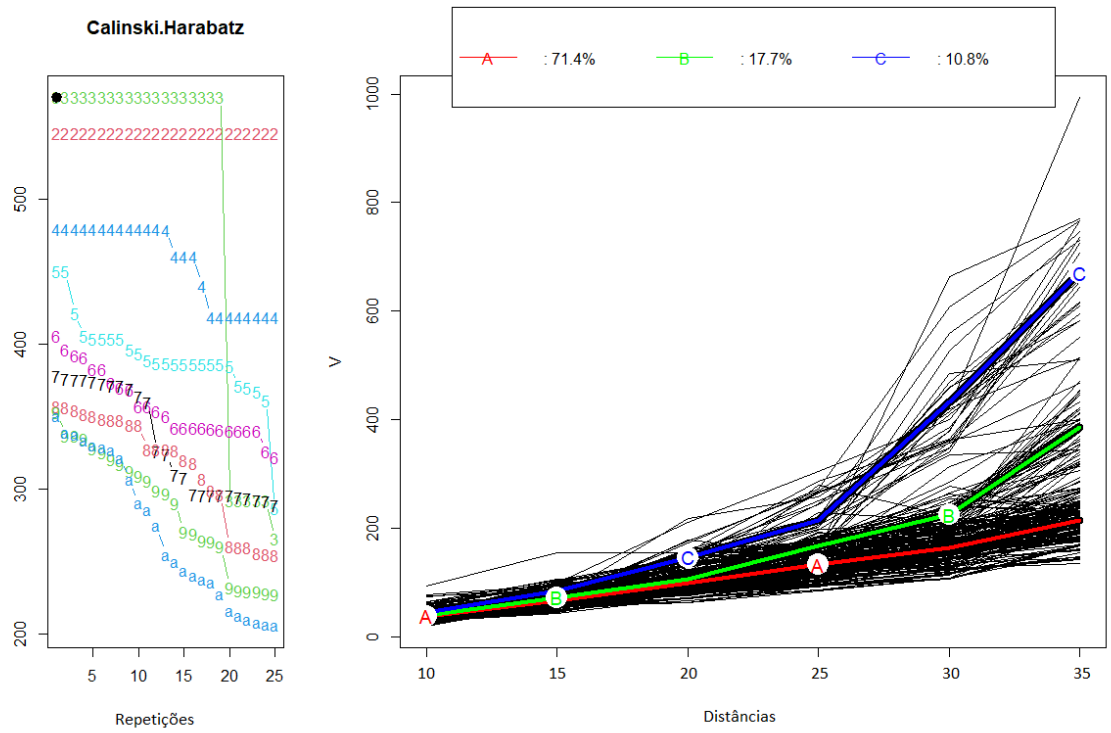


Figura 4.2: Grupos resultantes pelo critério de Caliński e Harabasz para o *k-means* longitudinal.

para o grupo C, o qual é mais atípico. Em ambas paredes, as curvas dos quadrantes $Q7$ e $Q10$, apresentam curvas bem similares, com poucas mudanças de comportamento e foram alocados, em sua maioria, no grupo A.

Tabela 4.1: Distribuição das curvas entre os grupos *versus* quadrantes na parede sem vazios internos para o *k-means* longitudinal.

Grupo	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
A	0	0	5	8	0	3	10	10	10	9	10	10
B	10	2	5	2	0	7	0	0	0	1	0	0
C	0	8	0	0	10	0	0	0	0	0	0	0

Tabela 4.2: Distribuição das curvas entre os grupos *versus* quadrantes na parede com vazios internos para o *k-means* longitudinal.

Grupo	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
A	10	0	10	0	10	10	9	4	10	10	8	9
B	0	0	0	4	0	0	1	6	0	0	2	1
C	0	1	0	6	0	0	0	0	0	0	0	0

A distribuição das curvas entre os grupos na parede com vazios internos é vista na Figura 4.3, na qual observa-se que o grupo A é o com maior número de curvas, sendo o único a conter as curvas com buracos em sua trajetória, além de ter a maioria das curvas

medidas sobre a argamassa. Portanto, as ondas que atravessam buracos ou argamassa são muito parecidas, com uma propagação mais homogênea e linear. Já as ondas propagadas em tijolos sem buracos ou com buracos perto de suas trajetórias, apresentam um comportamento heterogêneo, sendo alocadas em um número maior de grupos, como grupo C, o qual possui trajetórias mais atípicas e sendo composto apenas pelas curvas sem buracos em seu caminho.

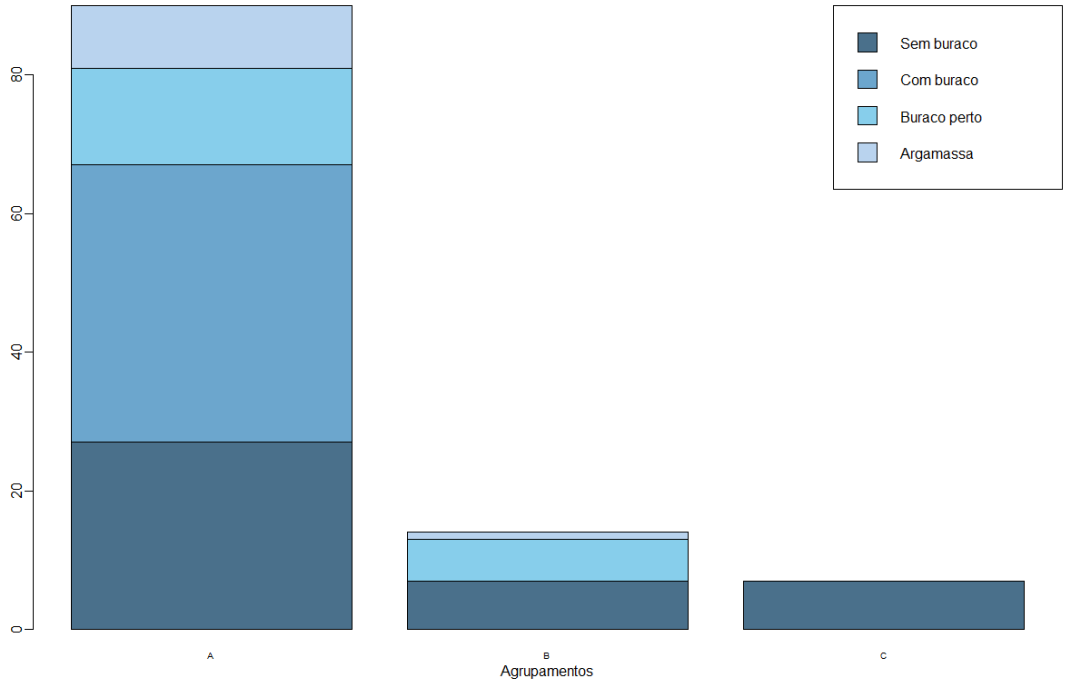


Figura 4.3: Gráfico de distribuição das curvas entre os grupos na parede com vazios para o *k-means* longitudinal.

Portanto, a partir deste agrupamento pelo *k-means* longitudinal, é possível classificar as curvas considerando o grupo que ela pertence. Por exemplo, quando uma curva é alocada para o grupo C, existe uma grande possibilidade de que a onda foi propagada em tijolos sem buracos em suas trajetórias. Já quando a onda é alocada para o grupo A, é preciso tomar maiores cuidados, pois pode haver qualquer um dos 4 casos, com maiores chances para quando a onda foi propagada em tijolos com buracos em suas trajetórias ou na argamassa.

Nota-se que os resultados são, a princípio, estranhos, pois esperava-se que as ondas que não possem buracos em suas trajetórias tivessem uma trajetória linear e, consequentemente, um comportamento mais homogêneo. Contudo isso não ocorre, provavelmente, pela heterogeneidade de material na composição dos tijolos, o que torna o meio de propagação menos homogêneo e faz com que os tempos das curvas que passam apenas pelos

tijolos também não sejam homogêneos.

Analisando as ondas que possuem buracos em suas trajetórias, acredita-se que elas possuem um comportamento mais homogêneo e que o tempo de propagação seja mais rápido (contrariando as leis da física), pois ao se deparar com as danificações na parede, a onda muda sua trajetória e se propaga pela argamassa do revestimento, a qual é mais homogênea do que os tijolos. Essa ideia também reafirma quando verificamos que o comportamento das ondas propagadas na argamassa ou nos tijolos com vazio interno na trajetória, são semelhantes. Em seu estudo, Carelli *et al.* (2014) analisa o comportamento de ondas ultrassônicas em elementos fissurados de concreto e argamassa, realizando testes para diferentes parâmetros, apresentando um maior foco para as engenharias civil e mecânica.

4.2 GCKM

O processo de agrupamento pelo método de GCKM envolve duas etapas, em que primeiro é ajustado um modelo misto, a fim de obter os valores preditos dos coeficientes aleatórios. No modelo misto definido neste estudo, tem-se o tempo de propagação da onda ultrassônica como variável resposta e a distância entre o transmissor e receptor da onda como variável preditora, pois acredita-se que a distância nas medições resulta em diferentes efeitos em cada curva observada. Além disto, como dito na Seção 2.2, será adotado um polinômio de terceiro grau para descrever a relação entre o tempo e a distância de forma adequada.

O modelo misto foi ajustado a partir da função *hlme* do pacote *lcmm*, pelo *software* R (Proust-Lima *et al.*, 2020). Desta forma, foi obtido o vetor predito $(\hat{b_{0,i}}, \hat{b_{1,i}}, \hat{b_{2,i}}, \hat{b_{3,i}})$ para cada unidade amostral, em que $\hat{b_{0,i}}$ representa os valores preditos para o intercepto do modelo, $\hat{b_{1,i}}$ representa o coeficiente angular da distância entre o transmissor e receptor da onda, $\hat{b_{2,i}}$ representa o coeficiente da distância ao quadrado e $\hat{b_{3,i}}$ representa o coeficiente da distância ao cubo.

Deste modo, com os valores preditos dos desvios aleatórios, foi feito o agrupamento pelo método de *k-means* longitudinal, implementado pelo pacote *kml* no *software* R. Para cada número de grupos, variando de 2 a 26, o processo de agrupamento foi executado 25 vezes com pontos iniciais diferentes, a fim de garantir bons resultados. Vale ressaltar que 26 é a quantidade máxima de grupos permitida nesta função, portanto foram testadas

todas as opções.

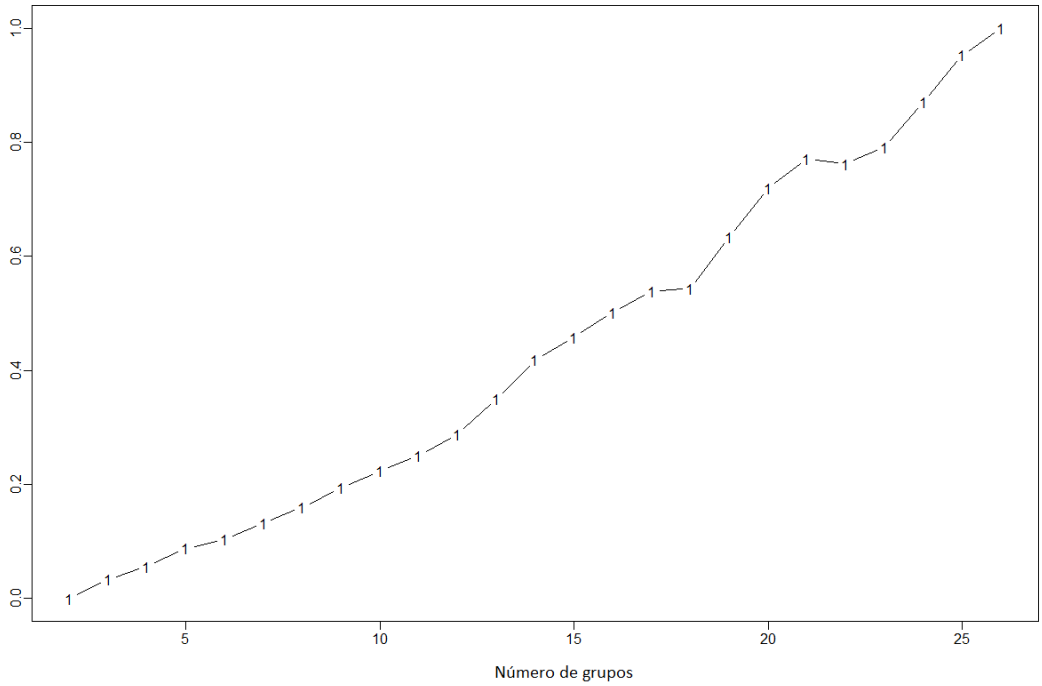


Figura 4.4: Gráfico para escolha do número ótimo de agrupamentos pelo critério de Caliński e Harabasz para o GCKM.

Para a definição do número ótimo de agrupamentos, foi utilizado do critério de Caliński e Harabasz (1974). Pela Figura 4.4, observa-se que quanto maior a quantidade de grupos, maior o valor da estatística, portanto tem-se que a divisão em 26 grupos é a mais adequada para os dados. Desta forma, nota-se que utilizar os valores preditos dos efeitos aleatórios como variáveis de agrupamento, fez com que o critério de Caliński e Harabasz apresentasse uma sensibilidade muito grande no momento de agrupar as curvas, captando diferenças mínimas entre as curvas.

Como o método de Caliński e Harabasz resultou que o melhor número de grupos é 26, optou-se por utilizar o método da soma de quadrados dentro dos grupos relativa a soma de quadrados total como outro método para estabelecer o número ótimo de agrupamentos. Este método é tradicionalmente utilizado no *k-means*, em que escolhe-se como o melhor número de grupos, o valores em que se estabiliza o decaimento da soma de quadrados relativa. Na Figura 4.5, observa-se que a partir de 5 grupos a curva torna-se quase que constante, tendendo a ser paralela ao eixo da abcissa, ou seja, os agrupamentos com mais de 5 grupos não mais reduzem a variabilidade dentro dos grupos de maneira relevante.

Pela Tabela 4.3, tem-se o cruzamento entre os grupos obtidos pelo método de Caliński e Harabasz, de A a Z na vertical, e os grupos obtidos pelo método da soma de quadrados

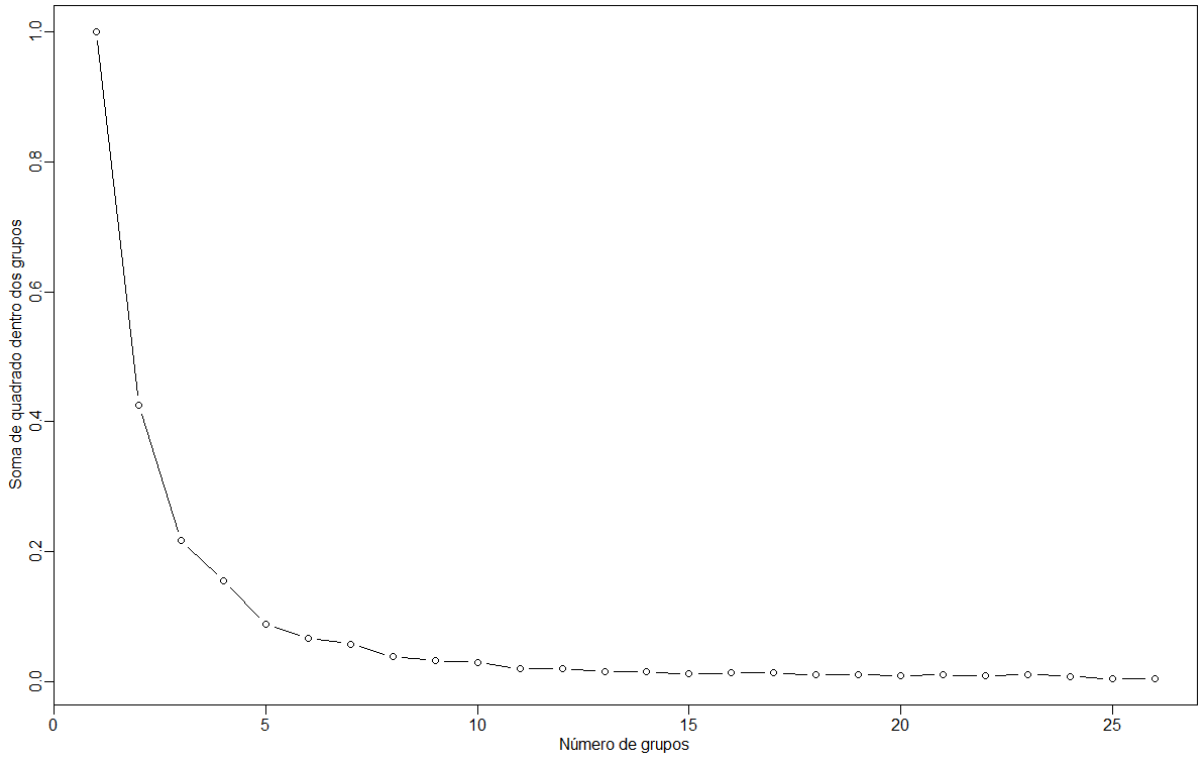


Figura 4.5: Gráfico para escolha do número ótimo de agrupamentos pelo critério da soma de quadrados relativa dentro dos grupo para o GCKM.

relativa dentro dos grupo, de A* a E* na horizontal. Observa-se que pelo primeiro método existem grupos que apresentam apenas uma ou poucas observações, o que não é interessante para o estudo, já que o objetivo é identificar grupos de curvas com características em comum e não curvas isoladas com comportamentos únicos. Desta forma, podem ser vistos que grupos pequenos, como de R a Z, foram realocados para grupos maiores, que também identificassem diferenças, mas de forma geral. Portanto, foi preferido prosseguir o estudo do GCKM, definindo a divisão em 5 grupos.

Desta forma, pela Figura 4.6, tem-se as trajetórias médias de cada um dos grupos formados. É importante ressaltar que as trajetórias são descritas pelos valores dos efeitos aleatórios, por isso apresentam uma visão diferente da apresentada no método de *k-means* longitudinal. Observa-se que o grupo C é composto por trajetórias muito parecidas à curva média estimada pelos efeitos fixos, com efeitos aleatórios preditos próximos de zero. O grupo B é composto por trajetórias que apresentam valores positivos preditos para o intercepto, seguidos de valores negativos para a distância, a qual vai se aproximando de zero nas distâncias ao quadrado e ao cubo. O comportamento do grupo D é semelhante ao grupo B, porém nele os valores positivos para o intercepto, assim como os valores negativos da inclinação linear, são mais acentuados. Os grupos E e A apresentam comportamentos

Tabela 4.3: Comparação da distribuição das curvas entre os grupos pelo critério de Caliński e Harabasz *versus* soma de quadrados relativa dentro dos grupos para o GCKM.

	A*	B*	C*	D*	E*
A	0	0	22	0	0
B	0	0	17	0	0
C	0	0	16	0	0
D	0	0	15	0	0
E	0	0	15	0	0
F	0	0	13	0	0
G	0	13	0	0	0
H	0	12	0	0	0
I	0	11	0	0	0
J	0	0	0	0	11
K	0	0	0	0	10
L	0	0	0	0	10
M	0	0	0	10	0

	A*	B*	C*	D*	E*
N	0	8	0	0	0
O	0	0	0	0	7
P	0	7	0	0	0
Q	0	0	0	0	7
R	5	0	0	0	0
S	0	0	0	0	5
T	4	0	0	0	0
U	0	0	0	4	0
V	0	0	0	3	0
W	0	0	0	2	0
X	2	0	0	0	0
Y	0	0	0	1	0
Z	1	0	0	0	0

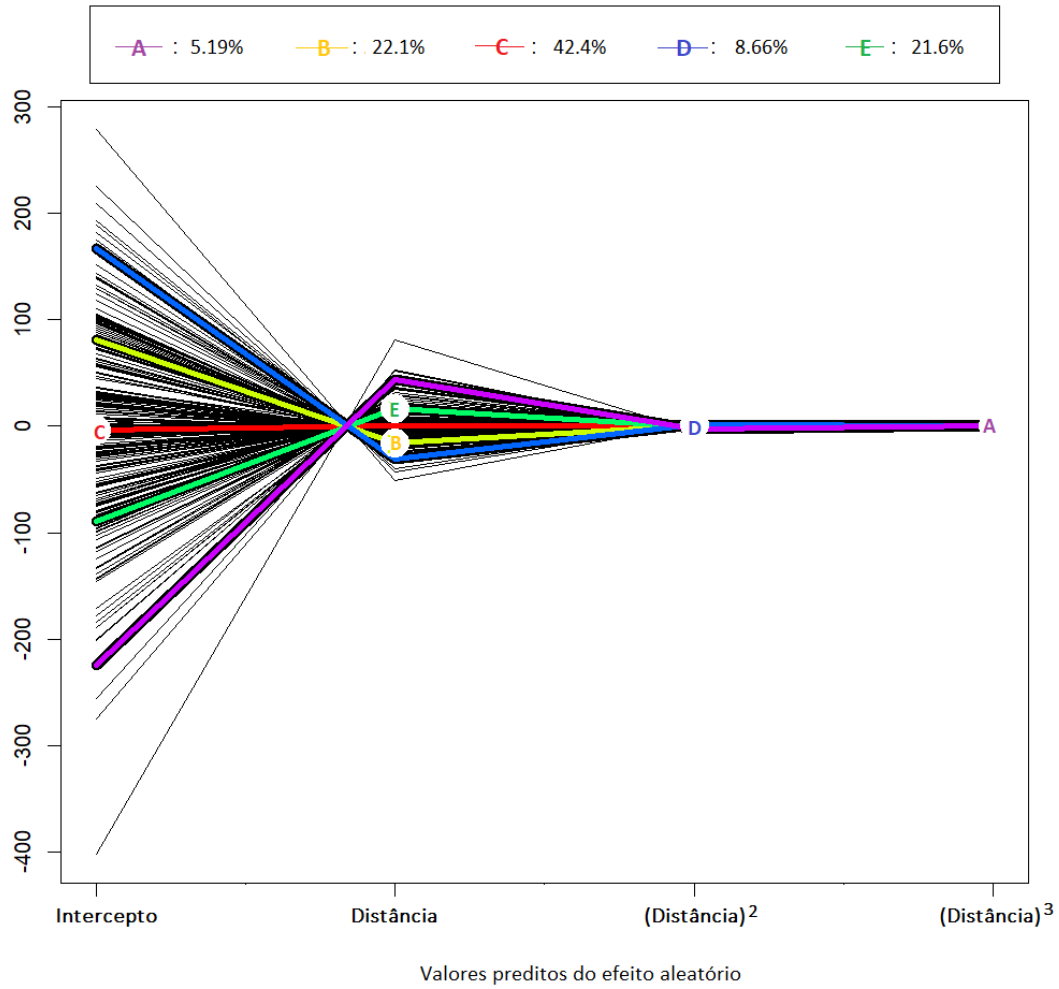


Figura 4.6: Trajetórias médias dos valores preditos por grupos resultantes pelo critério de soma de quadrados relativa dentro dos grupo para o GCKM.

parecidos com os grupos B e D, respectivamente, porém com a visão invertida. Vale ressaltar que os efeitos aleatórios que acompanham a distância ao quadrado e ao cubo são muito próximos de zero quando comparados aos efeitos aleatórios de intercepto e inclinação linear. O grupo C é o maior, contendo 42.4% das curvas, os grupos B e E possuem tamanhos similares, com 22.1% e 21.6% das curvas, respectivamente, enquanto os menores grupos são o D, com 8.66% das curvas, e o A, com 5.19% das curvas.

A Figura 4.7 mostra, por grupo identificado, as curvas observadas pelas diferentes distâncias analisadas, assim como no *k-means* longitudinal. Nela observa-se que o método de GCKM não diferenciou de forma evidente as curvas analisadas, contendo, por exemplo, tendo curvas atípicas, com altos valores de tempo, em todos os grupos criados. Contudo, pelos gráficos de cada grupo, nota-se que eles parecem separar as curvas pela distância em que as suas trajetórias mudam de tendência. No grupo A e E, em geral, é a partir da distâncias 25 ou 30 cm que o tempo de propagação muda de tendência e fica maior, especialmente no Grupo A. O grupo C, por sua vez, é o que contém, em geral, a maioria das trajetórias lineares que não mudam de tendência. Já nos grupos B e D, observamos que a tendência das trajetórias muda, geralmente, em torno da distância 15 ou 20 cm.

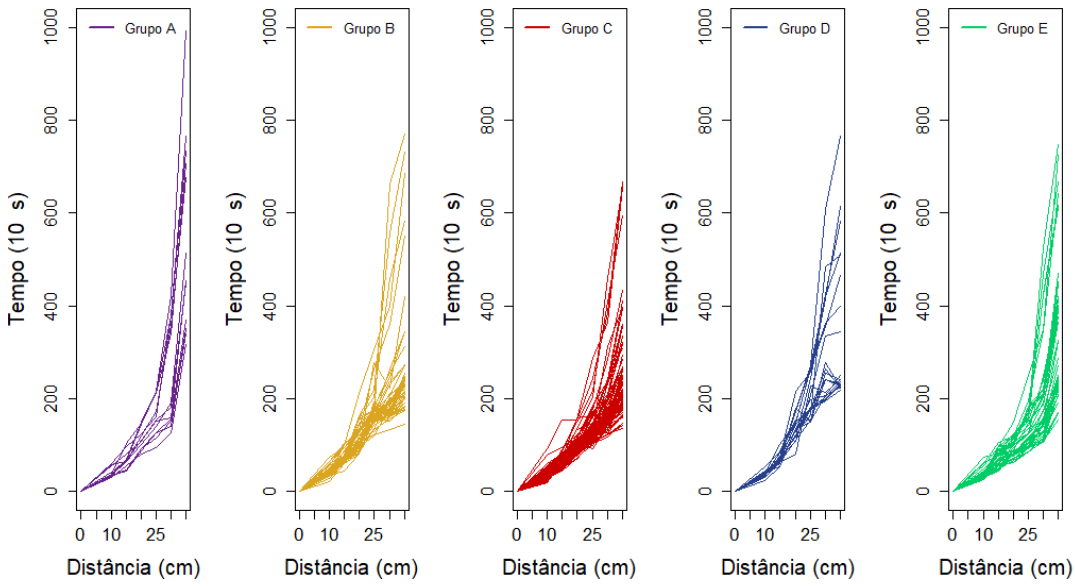


Figura 4.7: Trajetórias das ondas ultrassônicas por grupos resultantes pelo critério de soma de quadrados relativa dentro dos grupos para o GCKM.

Considerando as análises feitas na Subseção 3.1, é possível observar como as curvas foram alocadas nos 5 grupos, a partir das Tabelas 4.4 e 4.5. Na parede com vazios internos, analisou-se que as curvas dos quadrantes $Q6$, $Q7$ e $Q10$, apresentam um comportamento

linear, muito semelhante entre as curvas, e pela Tabela 4.5 nota-se que foram alocadas, em sua maioria, para o grupo C, o qual apresenta valores preditos próximos de zero. Observa-se que os grupos A e D são os que apresentam curvas mais atípicas, como exemplo na parede com vazios internos analisou-se que as curvas do quadrante $Q12$ apresentam grande dispersão para maiores distâncias e pela Tabela 4.5 vê-se que a maioria delas foi alocada para o grupo D.

Tabela 4.4: Distribuição das curvas entre os grupos *versus* quadrantes na parede sem vazios internos para o GCKM.

Grupo	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**
A	0	3	3	1	1	0	0	0	0	0	0	0
B	0	2	1	2	2	2	1	4	4	5	2	4
C	2	2	3	6	1	5	8	6	6	0	8	5
D	0	2	0	0	5	0	1	0	0	5	0	0
E	8	1	3	1	1	3	0	0	0	0	0	1

Tabela 4.5: Distribuição das curvas entre os grupos *versus* quadrantes na parede com vazios internos para o GCKM.

Grupo	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**
A	0	0	1	2	0	0	0	1	0	0	0	0
B	0	0	0	2	10	1	0	2	0	0	5	2
C	4	1	1	1	0	8	9	5	1	10	4	2
D	0	0	0	0	0	0	0	0	0	0	1	6
E	6	0	8	5	0	1	1	2	9	0	0	0

A distribuição das curvas entre os grupos na parede com vazios internos é vista na Figura 4.8, em que observa-se comportamentos semelhantes aos da Figura 4.3. O grupo C é o com maior número de curvas, contendo quase todas as curvas que passam pela argamassa e uma grande quantidade das que têm buracos em suas trajetórias. O grupo E contém curvas que apresentam os 4 tipos de trajetórias diferentes, assim como grupo C, sendo eles muito semelhantes. As curvas sem buracos em suas trajetórias, apresentam um comportamento heterogêneo, estando contidas em todos os grupos, com destaque para o grupo D, no qual é a única presente. Além disso, as curvas com buracos perto de suas trajetórias, também, apresentam um comportamento heterogêneo, estando contidas em 4 dos 5 grupos apresentados.

Portanto, a partir deste agrupamento pelo GCKM, é possível classificar as curvas considerando o grupo que ela pertence. Por exemplo, quando uma curva é alocada para o grupo D, existe uma grande possibilidade de que a onda foi propagada em tijolos sem

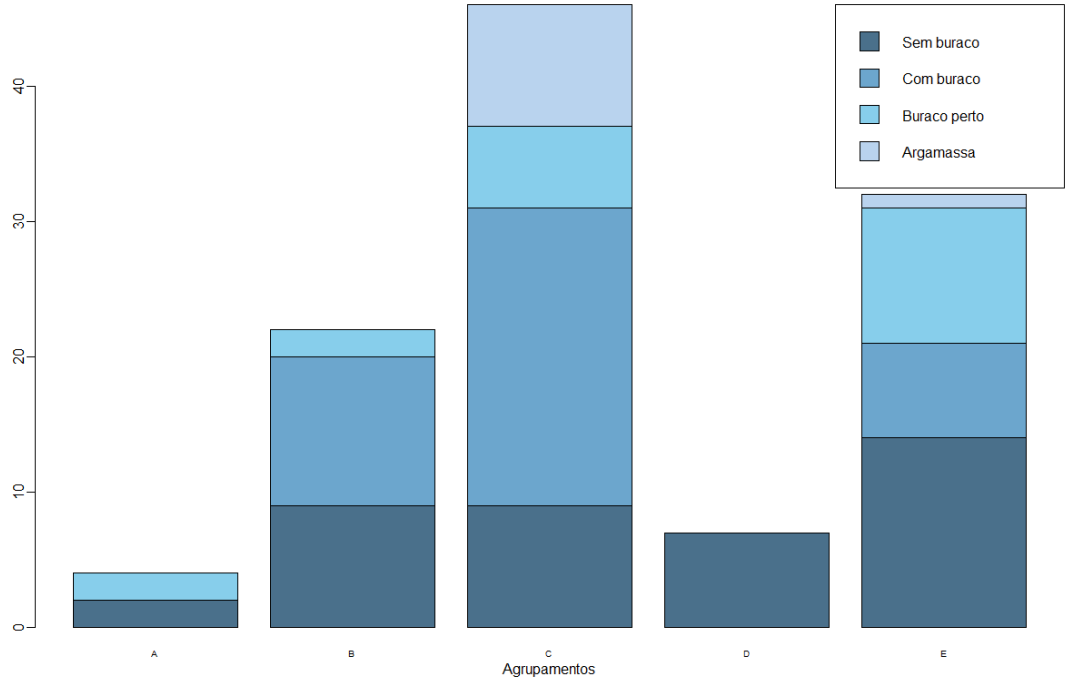


Figura 4.8: Gráfico de distribuição das curvas entre os grupos na parede com vazios internos para o GCKM.

buracos em suas trajetórias. Já quando a onda é alocada para o grupo C ou E, é preciso tomar maiores cuidados, pois pode haver qualquer um dos 4 casos, com maiores chances para quando a onda foi propagada em tijolos com buracos em suas trajetórias ou na argamassa no grupo C.

Assim como no *k-means* longitudinal, os resultados obtidos pelo GCKM também são, a princípio, estranhos devido ao fato das ondas que não possem buracos em suas trajetórias terem um comportamento heterogêneo e o tempo de propagação mais lento, enquanto as ondas que possuem buracos em suas trajetórias apresentam um comportamento e tempo de propagação mais rápido, em geral. As justificativas são as mesmas apresentadas na Seção 4.1.

4.3 Comparação entre os métodos

Ao longo deste estudo, foram analisadas duas metodologias para agrupamento de dados longitudinais ou medidas repetidas, o *k-means* longitudinal e o GCKM. Nesta seção será feita a comparação entre as metodologias em termos de facilidade de entendimento, esforço computacional e distribuição das curvas entre os grupos formados.

O *k-means* longitudinal é um método simples, de fácil entendimento e aplicação. O

algoritmo atribui cada unidade amostral a um dos G grupos e converge para uma solução relativamente rápido. Já o GCKM é um pouco mais complexo, pois exige uma primeira etapa antes da aplicação do *k-means*, que é ajustar um modelo misto, a fim de obter os valores preditos dos coeficientes aleatórios. Isto faz com que esta metodologia necessite maior conhecimento estatístico e, consequentemente, maior tempo para sua aplicação. Contudo, apesar de exigir maior esforço, o GCKM possui uma grande vantagem em relação ao *k-means* longitudinal, pois considera a dependência entre as observações da mesma unidade amostral, por meio dos efeitos aleatórios, o que já o torna mais adequado para descrever dados longitudinais ou medidas repetidas.

Tanto para o *k-means* longitudinal, quanto para o GCKM, existem pacotes e funções prontas e publicamente disponíveis para uso que facilitam as aplicações destes métodos. Comparando-os computacionalmente, em ambas metodologias não houve muito esforço computacional, principalmente devido ao pequeno tamanho da base de dados, com apenas 1615 observações. Isto nos permitiu testar diferentes parâmetros para cada método, sem demandar muito tempo. Contudo, o GCKM exige um esforço maior por ter duas etapas, necessitando de um tempo para estimar o modelo misto, antes da segunda etapa do agrupamento pelo *k-means*.

Ambas metodologias obtiveram resultados satisfatórios para o conjunto de dados estudado, sendo possível notar com clareza, em alguns grupos, os fatores que influenciaram nas divisões estabelecidas. Em cada um dos métodos foi especificado como seria feita a divisão entre os grupos, sendo que no *k-means* longitudinal é o comportamento das curvas nas diferentes distâncias analisadas e no GCKM é o comportamento dos valores preditos dos efeitos aleatórios do modelo misto estabelecido.

Durante os estudos, notou-se que o agrupamento a partir dos valores preditos se mostrou muito mais sensível às variações de comportamento do que o agrupamento pelas trajetórias das curvas. Pelo GCKM, primeiramente, foram propostos 26 grupos, sendo que a maioria possuía poucas curvas ou apenas uma muito atípica. Deste modo, existem indícios de que ao trabalhar com os valores preditos dos efeitos aleatórios, é possível identificar mais detalhadamente as diferenças entre as curvas..

A comparação da distribuição das curvas entre os grupos é vista na Tabela 4.6, na qual observa-se que as curvas alocadas no grupo A pelo *k-means* longitudinal se distribuem pelos 5 grupos do GCKM, com a maioria sendo alocada no grupo C*. Já as curvas alocadas no grupo B pelo *k-means* se distribuem majoritariamente nos grupos C* e E* do

Tabela 4.6: Comparação da distribuição das curvas entre os grupos pelo *k-means* longitudinal *versus* GCKM.

	A*	**B***	**C***	**D***	**E***
A	10	38	69	17	31
B	2	8	15	1	15
C	0	5	14	2	5

GCKM, indicando uma concordância entre os métodos. O mesmo acontece com as curvas alocadas no grupo C, as quais são destinadas, em sua maioria, no grupo C* pelo GCKM.

Capítulo 5

Conclusão

Neste estudo foram introduzidos dois métodos para o agrupamento de dados longitudinais ou medidas repetidas, com o objetivo de analisar e descrever o comportamento de propagação de ondas ultrassônicas em edifícios de alvenaria. O *k-means* longitudinal é uma abordagem simples, que assume independência entre as unidades amostrais, sendo muito similar ao *k-means* tradicional, o qual é mais conhecido na literatura. Já o GCKM é um método de agrupamento em duas etapas, em que primeiro é ajustado um modelo misto e em seguida os valores preditos dos coeficientes aleatórios são aplicados no *k-means*, a fim de realizar os agrupamentos.

O método de *k-means* longitudinal resultou em uma divisão em 3 grupos, enquanto o GCKM apresentou 5 grupos. Em ambos os casos foi observado que o maior grupo continha, principalmente, curvas com vazio em sua trajetória ou próximo ou que são medidas na argamassa. Essas curvas, geralmente, apresentam um comportamento mais homogêneo e o tempo de propagação praticamente é linear à distância percorrida. Curvas sem vazios em sua trajetória apresentam um comportamento mais heterogêneo, sendo alocadas em um número maior de grupos, geralmente com perfis atípicos.

Provavelmente, este comportamento é devido à heterogeneidade de material na composição dos tijolos, que torna o meio de propagação menos homogêneo e faz com que os tempos das curvas que passam apenas pelos tijolos também não sejam homogêneos. Analisando as ondas que possuem buracos em suas trajetórias, acredita-se que elas possuem um comportamento mais homogêneo e um tempo de propagação mais rápido pois, ao se deparar com as danificações na parede, a onda muda sua trajetória e se propaga pela argamassa do revestimento, que é mais homogênea do que os tijolos, resultando em ondas com tempos mais homogêneos de propagação.

Em seu estudo, Zuanetti *et al.* (2021) propôs um reconhecimento de padrão não supervisionado para o tempo de propagação de ondas em paredes de alvenaria por análise de agrupamento, aplicando sua metodologia no mesmo conjunto de dados utilizado neste trabalho. Nesta nova abordagem, as curvas foram alocadas em 19 grupos, em que 61,8% pertenciam a 2 grupos maiores, enquanto as demais pertenciam a grupos com poucas curvas ou apenas uma muito atípica. Foi observado que o maior grupo continha, principalmente, curvas com vazio em sua trajetória ou próximo ou que foram medidas na argamassa. Além disso, foi observado que as curvas sem vazios em seu caminho tinham um comportamento mais heterogêneo, sendo alocadas em um número maior de grupos, geralmente com perfis atípicos.

Desta forma, nota-se que os agrupamentos e análises realizados neste trabalho condizem com os do artigo de Zuanetti *et al.* (2021), já que as mesmas observações citadas foram feitas neste estudo, a partir dos grupos formados por ambos os métodos. Vale ressaltar que, se pelo método do GCKM fosse considerada a divisão em 26 grupos, como indicado pelo critério de Caliński e Harabasz, Tabela 4.3, haveriam muitos grupos com poucas ou apenas uma curva muita atípica, assim como em Zuanetti *et al.* (2021).

Portanto, foi possível analisar e descrever o comportamento de propagação de ondas ultrassônicas em edifícios de alvenaria, trazendo diferenciações através dos grupos formados em ambos os métodos. A partir das análises feitas neste estudo, foram registradas particularidades nos resultados de cada método, em que o *k-means* longitudinal apresentou diferenciações evidentes entre as trajetórias médias das curvas ultrassônicas, enquanto o GCKM captou as particularidades das curvas, principalmente, nas maiores distâncias. Deste modo, os agrupamentos feitos podem ser utilizados para avaliar a qualidade da alvenaria e serem aplicadas como base para engenheiros e pesquisadores da área. Os resultados não esperados em algumas situações também servem de alerta e insumo para novas pesquisas sobre como a onda ultrassônica se propaga em diferentes materiais e condições.

Referências Bibliográficas

Anderson, J. C. e Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, **103**(3), 411.

Bonilla, J. A. T. (2003). La importancia del patrimonio arquitectónico como documento histórico. *Cuadernos de Arte de la Universidad de Granada*, **34**, 195–206.

Caliński, T. e Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, **3**(1), 1–27.

Carelli, J. M. *et al.* (2014). Análise do comportamento de ondas ultrassônicas em elementos fissurados de concreto e argamassa.

Davies, D. L. e Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-1**(2), 224–227.

Den Teuling, N., Pauws, S. e van den Heuvel, E. (2020). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics-Simulation and Computation*.

Garcia, A. A. d. S. G. (2020). *Clustering of longitudinal data: Application to COVID-19 data*. Trabalho de Conclusão de Curso, Faculdade de Ciências da Universidade de Porto.

Genolini, C. e Falissard, B. (2010). Kml: k-means for longitudinal data. *Computational Statistics*, **25**(2), 317–328.

Genolini, C. e Falissard, B. (2011). Kml: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, **104**(3), 112–121.

Genolini, C., Alacoque, X., Sentenac, M., Arnaud, C. *et al.* (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, **65**(4), 1–34.

Genolini, C., Falissard, B. e Genolini, M. C. (2016). Instructions manual of package 'kml'.

Kryszczuk, K. e Hurley, P. (2010). Estimation of the number of clusters using multiple clustering validity indices. Em *International Workshop on Multiple Classifier Systems*, páginas 114–123. Springer.

McCulloch, C. E. e Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Proust-Lima, C., Philipps, V. e Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes: the r package lcmm. *Journal of Statistical Software*.

Proust-Lima, C., Philipps, V., Diakite, A., Liquet, B., Proust, M. C. e Lima, P. (2020). Instructions manual of package 'lcmm'.

Ramirez, F. C. (2015). *Detecção de danos em estruturas de concreto por meio de tomografia ultrassônica.*. Tese de doutorado, Universidade de São Paulo.

Ray, S. e Turi, R. H. (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. Em *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, páginas 137–143. Citeseer.

Rodrigues, R. V. (2020). *Análise estatística de dados ultrassônicos para a seleção de características da alvenaria que comprometem sua estabilidade*. Trabalho de Conclusão de Curso, Universidade Federal de São Carlos.

Rubens, T. D. M. (2019). *Caracterização ultrassônica de blocos históricos cearenses*. Trabalho de Conclusão de Curso, Universidade Federal do Ceará.

Singer, J. M. e Andrade, D. F. (1986). Análise de dados longitudinais. *Simpósio Nacional de Probabilidade e Estatística*, **7**.

Twisk, J. e Hoekstra, T. (2012). Classifying developmental trajectories over time should be done with great caution: a comparison between methods. *Journal of Clinical Epidemiology*, **65**(10), 1078–1087.

Wang, X., Smith, K. e Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, **13**(3), 335–364.

Zuanetti, D. A., da Paz, R. F., Rodrigues, T. e Mesquita, E. (2021). Clustering ultrasonic waves propagation time: A hierarchical polynomial semiparametric approach. *Applied Stochastic Models in Business and Industry*, **37**(5), 894–907.

Apêndice A

Códigos do *k-means* longitudinal

```
#########
## KML ##
#########

## Pacote
library(kml)

## Chamar os dados
dados_kml <- read.csv2("C:/Users/gnesterick/Desktop/TCC/Dados/Dados_kml.csv")
colnames(dados_kml) <- c("ID", "t10", "t15", "t20", "t25", "t30", "t35")

## Ajustando os dados para o KML
dados <- cld(dados_kml, timeInData = 2:7)

## Ver os dados
dados["traj"]
dados["idAll"]
dados["time"]

## Colocar semente (para obter sempre os mesmo resultados)
set.seed(57)

## Roda o KML, testando de 1 a 10 grupos, em que o k-means deve ser
## executado 25 vezes (com diferentes condições iniciais) para cada
```

```
## número de grupos
kml(dados, 1:10, nbRedraw = 25)

## Ver todos os critérios
plotAllCriterion(dados)

## Ver apenas critério Calinski Harabatz
plotAllCriterion(dados,criterion=CRITERION_NAMES[1])

## Graficos: valores para o critério Calinski Harabatz e grupos formados
dados@criterionActif <- "Calinski.Harabatz"
choice(dados)

## Colocar o nome do grupo em cada observacao
dados_kml$clusters <- getClusters(dados, 3)

## Verificar agrupamnetos
table(dados_kml$clusters)



######################
## KML - histograma ##
######################

## Chamar os dados
dados_completo <- read.csv2("C:/Users/gnesterick/Desktop/TCC/Dados/
dados_total_agrupados.csv")
table(dados_completo$Grupo)

## Deixar o quadrante como numerico
dados_completo$quadrante_n <-
ifelse(dados_completo$Quadrante == "Q1", 1,
ifelse(dados_completo$Quadrante == "Q2", 2,
ifelse(dados_completo$Quadrante == "Q3", 3,
ifelse(dados_completo$Quadrante == "Q4", 4,
```

```
ifelse(dados_completo$Quadrante == "Q5", 5,
ifelse(dados_completo$Quadrante == "Q6", 6,
ifelse(dados_completo$Quadrante == "Q7", 7,
ifelse(dados_completo$Quadrante == "Q8", 8,
ifelse(dados_completo$Quadrante == "Q9", 9,
ifelse(dados_completo$Quadrante == "Q10", 10,
ifelse(dados_completo$Quadrante == "Q11", 11,
ifelse(dados_completo$Quadrante == "Q12", 12, "erro")))))))))))

dados_completo$quadrante_n <- as.numeric(dados_completo$quadrante_n)

## Criar tabela de dados para cada curvinha (231 curvas)
library(tidyr)
library(dplyr)
dados <- dados_completo %>% group_by(ID_Amostra, Grupo, parede_n, quadrante_n,
replicas) %>% summarise()

## Estabelecer onde estão os buracos
nao_buraco<-which((dados$quadrante_n %in% c(2,4,9,12,11) & dados$parede_n==2) |
(dados$quadrante_n %in% c(4,9,12,5,6,7,8,10) & dados$parede_n==1))
argamassa<-which((dados$quadrante_n %in% c(1,2,3,11) & dados$parede_n==1) |
(dados$quadrante_n %in% c(7) & dados$parede_n==2))
buraco<-which(dados$quadrante_n %in% c(1,5,6,10) & dados$parede_n==2)
buraco_p<-which(dados$quadrante_n %in% c(3,8) & dados$parede_n==2)

## Criar as categorias
categoria<-NULL
categoria[nao_buraco]<-"Sem_buraco"
categoria[buraco]<-"Buraco"
categoria[argamassa]<-"Argamassa"
categoria[buraco_p]<-"Buraco_perto"
#
categoria<-NULL
categoria[nao_buraco]<-0
categoria[buraco]<-1
```

```
categoria[argamassa]<-3
categoria[buraco_p]<-2

## Criar os grupos
Grupo_A <- which(dados$Grupo == "A")
Grupo_B <- which(dados$Grupo == "B")
Grupo_C <- which(dados$Grupo == "C")
#
grupo<-NULL
grupo[Grupo_A]<-"A"
grupo[Grupo_B]<-"B"
grupo[Grupo_C]<-"C"

#####################################
## PARA A PAREDES 2 - COM VAZIOS -----
#
## Criar as frequencias
#
freq_absol<-table(categoria[dados$parede_n==2],grupo[dados$parede_n==2])
# calculo das freq relativas por linha
freq_linha<-round(prop.table(freq_absol,1),3)*100
# calculo das freq relativas por coluna
freq_coluna<-round(prop.table(freq_absol,2),3)*100

## Criar o gráfico
par(mar=c(3,3,0.5,0.5),mgp=c(1.5,0.5,0))
barplot(freq_absol,
        main = " ",
        xlab = "Agrupamentos",
        col = c("skyblue4","skyblue3","skyblue","slategray2"),cex.names=0.60
)
legend("topright",
       c("Sem buraco","Com buraco","Buraco perto","Argamassa"),
       fill = c("skyblue4","skyblue3","skyblue","slategray2")
)
```

```
## VER OS AGRUPAMENTOS POR TABELAS
table(dados_kml_agrupados_completo$Grupo, dados_kml_agrupados_completo$Quadrante)/7
table(dados_kml_agrupados_completo$Grupo, dados_kml_agrupados_completo$Quadrante,
dados_kml_agrupados_completo$parede_p)/7
```

Apêndice B

Códigos do GCKM

```
##################
## Modelo Misto ##
##################

library(readxl)
library(dplyr)
library(lcmm)

Dados <- read_excel("C:/Users/gnesterick/Desktop/TCC/Dados/Dados.xlsx")

## ARRUMAR O ID (tem que ser numérico)
dados_aux <- Dados %>% group_by(ID_Amostra) %>% summarise()
dados_aux$ID_aux <- as.numeric(1:nrow(dados_aux))

Dados <- Dados %>% left_join(dados_aux, by=c("ID_Amostra"="ID_Amostra"))

## Arrumar os dados para o hlme
dados_hlme <- Dados %>% select(ID_aux ,Valor, D_m)
dados_hlme <- dados_hlme %>% filter(dados_hlme$D_m != 0)
table(dados_hlme$ID_aux)

## Rodar o modelo
mod_hlme <- hlme(Valor ~ 1 + D_m +  I(D_m^2) +  I(D_m^3) ,
random = ~ 1 + D_m +  I(D_m^2) +  I(D_m^3),
```

```
subject = "ID_aux", ng = 1, data = dados_hlme)

## Verificar resultados
mod_hlme$best
summary(mod_hlme)

## Colacar resultados na tabela
hlme_pred <- data.frame(dados_aux$ID_Amostra, mod_hlme$predRE)
hlme_pred$ID_aux <- NULL
colnames(hlme_pred) <- c("ID_Amostra","Intecepto", "D_m", "D_m_2", "D_m_3")

#########
## KML ##
#########

## Pacote
library(kml)

## Chamar os dados
load("C:/Users/mbrod/OneDrive/Área de Trabalho/TCC_Gi/Scripts_finais/
GCKM/hlme_pred.RData")

## Ajustando os dados para o KML
dados <- cld(hlme_pred, timeInData = 2:5)

## Colocar semente (para obter sempre os mesmo resultados)
set.seed(57)

## Roda o KML, testando de 1 a 26 grupos, em que o k-means deve ser
## executado 25 vezes (com diferentes condições iniciais) para cada número de grupos
kml(dados, 1:26, nbRedraw = 25)

## Ver apenas critério Calinski Harabatz
plotAllCriterion(dados,criterion=CRITERION_NAMES[1])
```

```
## Graficos: valores para o critério Calinski Harabatz e grupos formados
choice(dados)

## Colocar o nome do grupo em cada observacao
hlme_pred_kml <- hlme_pred
hlme_pred_kml$clusters <- getClusters(dados, 26)

## Verificar agrupamnetos
table(hlme_pred_kml$clusters)

####### Agrupamento via K-médias #########

## Chamar os dados
load("C:/Users/mbrod/OneDrive/Área de Trabalho/TCC_Gi/Scripts_finais/
GCKM/hlme_pred.RData")

## selecionando o melhor número de grupos ##

set.seed(100)
mydata <- hlme_pred[,-1]
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:26) wss[i] <- sum(kmeans(mydata,centers=i)$withinss)
plot(1:26, wss/wss[1], type="b", xlab="Número de grupos",
ylab="Soma de quadrado dentro dos grupos")

## assumindo apenas 5 grupos
set.seed(100)
clus_med_1<-kmeans(mydata,centers=5)

length(clus_med_1[[1]])
length(hlme_pred_kml$clusters)
table(clus_med_1[[1]])
table(hlme_pred_kml$clusters, clus_med_1[[1]])

hlme_pred$grupo <- clus_med_1[[1]]
```

```
## número de grupos pelo kml para ter o grafico
kml(dados, 2:5, nbRedraw = 25)

hlme_pred_kml_5 <- hlme_pred
hlme_pred_kml_5$clusters <- getClusters(dados, 5)

## Verificar agrupamnetos
table(hlme_pred_kml_5$clusters)
table(hlme_pred_kml_5$clusters, clus_med_1[[1]])

## Graficos: valores para o critério Calinski Harabatz e grupos formados
choice(dados)

######################
## KML - histograma ##
######################

library(readxl)
load("C:/Users/mbrod/OneDrive/Área de Trabalho/TCC_Gi/Scripts_finais/
GCKM/hlme_pred_grupos.RData")
dados_com_e_sem_vazios <- read_excel("C:/Users/mbrod/OneDrive/
Área de Trabalho/TCC_Gi/Dados/Dados.xlsx")

library(dplyr)
hlme_pred <- hlme_pred %>% select(ID_Amostra, grupo)
dados_completo <- dados_com_e_sem_vazios %>%
left_join(hlme_pred, by = c("ID_Amostra"="ID_Amostra"))

table(dados_completo$grupo)/7

## Deixar o quadrante como numerico
dados_completo$quadrante_n <-
ifelse(dados_completo$Quadrante == "Q1", 1,
ifelse(dados_completo$Quadrante == "Q2", 2,
```

```
ifelse(dados_completo$Quadrante == "Q3", 3,
ifelse(dados_completo$Quadrante == "Q4", 4,
ifelse(dados_completo$Quadrante == "Q5", 5,
ifelse(dados_completo$Quadrante == "Q6", 6,
ifelse(dados_completo$Quadrante == "Q7", 7,
ifelse(dados_completo$Quadrante == "Q8", 8,
ifelse(dados_completo$Quadrante == "Q9", 9,
ifelse(dados_completo$Quadrante == "Q10", 10,
ifelse(dados_completo$Quadrante == "Q11", 11,
ifelse(dados_completo$Quadrante == "Q12", 12, "erro")))))))))))

dados_completo$quadrante_n <- as.numeric(dados_completo$quadrante_n)

## Criar tabela de dados para cada curvinha (231 curvas)
library(tidyr)
dados <- dados_completo %>%
group_by(ID_Amostra, grupo, parede_n, quadrante_n, replicas) %>%
  summarise()

## Estabelecer onde estão os buracos
nao_buraco<-which((dados$quadrante_n %in% c(2,4,9,12,11) & dados$parede_n==2) |
(dados$quadrante_n %in% c(4,9,12,5,6,7,8,10) & dados$parede_n==1))
argamassa<-which((dados$quadrante_n %in% c(1,2,3,11) & dados$parede_n==1) |
(dados$quadrante_n %in% c(7) & dados$parede_n==2))
buraco<-which(dados$quadrante_n %in% c(1,5,6,10) & dados$parede_n==2)
buraco_p<-which(dados$quadrante_n %in% c(3,8) & dados$parede_n==2)

## Criar as categorias
categoria<-NULL
categoria[nao_buraco]<-"Sem_buraco"
categoria[buraco]<-"Buraco"
categoria[argamassa]<-"Argamassa"
categoria[buraco_p]<-"Buraco_perto"
#
categoria<-NULL
```

```
categoria[nao_buraco]<-0
categoria[buraco]<-1
categoria[argamassa]<-3
categoria[buraco_p]<-2

## Criar os grupos
Grupo_A <- which(dados$grupo == 1)
Grupo_B <- which(dados$grupo == 2)
Grupo_C <- which(dados$grupo == 3)
Grupo_D <- which(dados$grupo == 4)
Grupo_E <- which(dados$grupo == 5)
#
grupo<-NULL
grupo[Grupo_A]<-"A"
grupo[Grupo_B]<-"B"
grupo[Grupo_C]<-"C"
grupo[Grupo_D]<-"D"
grupo[Grupo_E]<-"E"

#####################################
## PARA A PAREDES 2 - COM VAZIOS -----

## Criar as frequencias

freq_absol<-table(categoria[dados$parede_n==2],grupo[dados$parede_n==2])
# calculo das freq relativas por linha
freq_linha<-round(prop.table(freq_absol,1),3)*100
# calculo das freq relativas por coluna
freq_coluna<-round(prop.table(freq_absol,2),3)*100

## Criar o gráfico
par(mar=c(3,3,0.5,0.5),mgp=c(1.5,0.5,0))
barplot(freq_absol,
        main = " ",
        xlab = "Agrupamentos",
```

```
        col = c("skyblue4","skyblue3","skyblue","slategray2"),cex.names=0.60
)
legend("topright",
       c("Sem buraco","Com buraco","Buraco perto","Argamassa"),
       fill = c("skyblue4","skyblue3","skyblue","slategray2")
)

## VER OS AGRUPAMENTOS POR TABELAS
table(dados_completo$grupo, dados_completo$Quadrante)/7
table(dados_completo$grupo, dados_completo$Quadrante, dados_completo$parede_p)/7
table(dados_completo$grupo, dados_completo$parede_p)/7
```