Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Breno Gabriel da Silva

Modelos Mistos Aditivos Generalizados Utilizando P-*splines*: Aplicações em Dados Longitudinais

Maringá – Paraná
2020

Breno Gabriel da Silva

Modelos Mistos Aditivos Generalizados Utilizando P-*splines*: Aplicações em Dados Longitudinais

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para a obtenção de título de Mestre em Bioestatística.

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Orientadora: Prof^{a}. Dr^{a}. Terezinha Aparecida Guedes

Coorientador: Prof. Dr. Vanderly Janeiro

Maringá – Paraná

2020

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Silva, Breno Gabriel da

S586m Modelos mistos aditivos generalizados utilizando P-splines: aplicações em dados longitudinais / Breno Gabriel da Silva. -- Maringá, PR, 2020.
76 f.: il. color., figs., tabs.

Orientadora: Profa. Dra. Terezinha Aparecida Guedes.
Coorientador: Prof. Dr. Vanderly Janeiro.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2020.

1. GAMM. 2. Regressão não paramétrica. 3. Funções segmentadas. 4. Efeitos aleatórios. I. Guedes, Terezinha Aparecida , orient. II. Janeiro, Vanderly, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 23.ed. 519.5

Jane Lessa Monção - CRB 9/1173

Breno Gabriel da Silva

Modelos Mistos Aditivos Generalizados Utilizando P-*splines*: Aplicações em Dados Longitudinais

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para a obtenção de título de Mestre em Bioestatística.

Trabalho aprovado. Maringá – Paraná, 31 de janeiro de 2020:

Profª. Dra. Terezinha Aparecida Guedes
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Willian Luís de Oliveira
Universidade Estadual de Maringá – PBE/UEM

Profª. Dra. Renata Alcarde Sermarini
Escola Superior de Agricultura "Luiz de Queiroz" - ESALQ/USP

Maringá – Paraná
2020

"Este trabalho é dedicado em primeiro lugar a Deus, por ter um papel importante em minha vida, condutor do meu destino, que têm me dado saúde e forças desde sempre, a minha família, especialmente a minha mãe Mari Gerônimo da Silva, meu pai José Alves da Silva e aos meus irmãos Michel Luiz da Silva "In Memorian", Frankson Leonardo da Silva, Michelle Francielle da Silva e Bruno Gabriel da Silva, em que mesmo não entendendo este cenário não mediram esforços para que eu chegasse até esta etapa de minha vida. Ao meu companheiro Guilherme Lunardon de Oliveira por sempre me incentivar e me ouvir, aos meus avós paternos e maternos "In Memorian", responsáveis pela existência de meus pais, em que sem eles este momento único não estaria acontecendo."

Agradecimentos

- Agradeço inicialmente a Deus, por ter me guiado e me dado forças para chegar até aqui.
- Aos meus pais (Mari Gerônimo da Silva & José Alves da Silva), Avô (Antônio Gerônimo da Silva "In Memorian"), irmãos (Michel Luiz da Silva "In Memorian", Frankson Leonardo da Silva, Michelle Francielle da Silva & Bruno Gabriel da Silva) e demais familiares, que sempre me apoiaram e incentivaram todas as minhas escolhas.
- Ao meu companheiro (Guilherme Lunardon de Oliveira) que esteve ao meu lado desde o inicio desta jornada, estendendo sua mão, me apoiando e incentivando todos os dias visando a realização deste sonho.
- A minha orientadora, (Terezinha Aparecida Guedes), que sempre se manteve a disposição quanto seu papel de orientadora e pelas contribuições mais que valiosas durante a elaboração deste trabalho. Professora, você é um exemplo a ser seguido, obrigado pelas palavras e conselhos dados durante esta jornada. Agradeço ainda ao meu coorientador (Vanderly Janeiro) por se disponibilizar a me ouvir e sugerir melhorias neste trabalho, expresso gratidão a vocês por terem depositado toda a confiança em mim.
- Aos meus amigos de longa data (Aléxia Fernanda Mançan, Eloisa da Silva Brito, Géssica Elias, Géssica Cardoso, Sabrina Luiza, José Sanches) que partilharam comigo muitos momentos alegres e estiveram ao meu lado dando forças para a elaboração deste trabalho.
- Agradeço as amizades que foram construídas durante a Graduação e Pós-Graduação, especialmente (Rafaela Galo & Sthefany Volpato) que me acolheram por muitos anos durante as longas tardes de estudos, agradecer a (Thayná Guastalli, Gisele Nascimento, Letícia Souza, Ana Zampirolli, Lorena Rosa, Júlia Paiva, Heloisa Nascimento, Rebeca Seiscentos, Carol Bianchi, Raissa Oliveira e Ariane Seródio) pelos momentos juntos durante a Graduação e incentivo para a Pós-Graduação durante o período mencionado, agradecer também a (Naiara Caroline & Talita Evelin) que sempre se disponibilizaram a me auxiliar e me ouvir quando precisei, agradecer também a (Yana Borges) que foi

um presente que a Pós-Graduação me deu no qual quero partilhar contigo muitas outras viagens, trabalhos e assuntos pertinentes quanto ao passado, presente e futuro.

- Agradeço ainda as amizades que foram construídas por interseção das listadas anteriormente, especialmente a (Andressa Cazetta, Mirielly Kerolay, Michelle Ferreira, Nathan Faleiro, Theylor Faleiro) por estarem na maioria das vezes presentes em momentos de distração, trocando ideias e compartilhando sorrisos.

- Agradeço as minhas amigas que tive o prazer de participar como co-autor em trabalhos acadêmicos (Fernanda Lobo, Ana Carolina Costa, Renata Pereira e Tamiris Diniz). Vocês são responsáveis do meu crescimento quanto ao conhecimento estatístico e acima de tudo foram essenciais para que eu atingisse todos os meus objetivos durante a Pós-Graduação.

- Não poderia deixar de agradecer os Professores que tive durante a Graduação e Pós-Graduação, em especial (Luciene Parron, Alexandra Cousin, Josiane Faria, Sandra Regina, Lucieli Trivizoli, William Ferreira e Brian de Melo) por terem transmitido todo o conhecimento de vocês, vindo a auxiliar na compreensão de conceitos acadêmicos, Parabéns Professores!!.

- Agradeço à CAPES pelo suporte financeiro, em que pude conhecer vários estados brasileiros durante as participações em congressos levando trabalhos acadêmicos pertinentes ao cenário e pelo Programa de Pós-Graduação em Bioestatística pelo incentivo a pesquisa.

  Obrigado!

Tocando em frente

Ando devagar
Porque já tive pressa
E levo esse sorriso
Porque já chorei demais

Hoje me sinto mais forte
Mais feliz, quem sabe
Só levo a certeza
De que muito pouco sei
Ou nada sei

Conhecer as manhas
E as manhãs
O sabor das massas
E das maçãs

É preciso amor
Pra poder pulsar
É preciso paz pra poder sorrir
É preciso a chuva para florir

Penso que cumprir a vida
Seja simplesmente
Compreender a marcha
E ir tocando em frente

Como um velho boiadeiro
Levando a boiada
Eu vou tocando os dias
Pela longa estrada, eu vou
Estrada eu sou

Conhecer as manhas
E as manhãs
O sabor das massas
E das maçãs

É preciso amor
Pra poder pulsar
É preciso paz pra poder sorrir
É preciso a chuva para florir

Todo mundo ama um dia
Todo mundo chora
Um dia a gente chega
E no outro vai embora

Cada um de nós compõe a sua história
Cada ser em si
Carrega o dom de ser capaz
E ser feliz.

Composição: Almir Sater / Renato Teixeira

Resumo

No que tange as especificidades de um estudo longitudinal, muitas vezes as trajetórias das respostas médias de um indivíduo não admitem um comportamento linear, tornando-se assim necessária a busca por modelos que levem em consideração a não linearidade das trajetórias individuais e que descreva estas de tal maneira que associem os possíveis efeitos aleatórios a cada indivíduo. Os modelos mistos aditivos generalizados (GAMM) vem ao encontro dessa problemática, uma vez que nesta classe de modelos é possível incluir efeitos aleatórios específicos aos indivíduos, além do mais, nesta classe de modelos é possível reescrever o termo linear por soma de funções suaves não conhecidas, não especificadas de uma forma paramétrica, por meio de uma P-*spline*. Assim, o presente trabalho terá como propósito apresentar esta metodologia e aplicar a dois problemas reais: um experimento envolvendo 57 camundongos suíços infectados por *Trypanosoma Cruzi*, sendo estes observado o seu peso no período de 12 semanas e um experimento da engenharia civil que tem como objetivo verificar a evolução do peso de dois tipos de argamassas, após a absorção de água por capilaridade, considerando 5 níveis de concentração do teor aditivo, em 60 corpos-de-prova, sendo estes observados a argamassa seca no período de 1350 minutos, divididos em até os 90 minutos a cada 10 minutos, de 90 a 150 minutos a cada 15 minutos, de 150 a 360 minutos a cada 20 minutos, de 360 a 450 minutos a cada 45 minutos e de 450 a 1350 minutos a cada 60 minutos.

Palavras-chave: Efeitos aleatórios; Funções segmentadas; GAMM; Regressão não paramétrica.

Abstract

Regarding the specificities of a longitudinal study, the trajectories of an individual's average responses often do not admit a linear behavior, thus making it necessary to search for models that take into account the non-linearity of individual trajectories and that describe these from such that they associate the possible random effects to each individual. Generalized mixed additive models (GAMM) meet this problem, since in this class of models it is possible to include random effects specific to individuals, moreover, in this class of models it is possible to rewrite the linear term by adding non-smooth functions known, not specified in a parametric way, by means of a P-*spline*. Thus, the present work aims to present this methodology and apply it to two real problems: an experiment involving 57 Swiss mice infected with *Trypanosoma Cruzi*, these being observed their weight in the period of 12 weeks and an experiment of civil engineering which aims to verify the evolution of the weight of two types of mortar, after water absorption by capillarity, considering 5 levels of concentration of the additive content, in 60 specimens, these being observed the dry mortar in the period of 1350 minutes, divided into up to 90 minutes every 10 minutes, 90 to 150 minutes every 15 minutes, 150 to 360 minutes every 20 minutes, 360 to 450 minutes every 45 minutes and 450 to 1350 minutes every 60 minutes.

Keywords: Random effects; Segmented functions; GAMM; Nonparametric regression.

Lista de ilustrações

Figura 1 – Representação do modelo *spline* linear 25
Figura 2 – Representação do modelo *spline* cúbico 27
Figura 3 – Diagrama do cálculo das B-*splines* 28
Figura 4 – Bases B-*spline* com diferentes graus polinomiais 29
Figura 5 – Curvas ajustadas por meio de P-*splines* com diversos valores de λ 38
Figura 6 – Gráfico de autocorrelação 45
Figura 7 – *Box-plot* do peso dos corpos-de-prova para cada nível de concentração do pó de borracha dos pneus inservíveis e box-plot para cada tipo de argamassa. 51
Figura 8 – Perfil médio global para cada corpo-de-prova e perfil médio para cada tipo de argamassa observado no tempo 52
Figura 9 – Gráfico de dispersão do peso dos corpos-de-prova para cada tipo de argamassa e níveis de concentração do pó de borracha dos pneus inservíveis. . . 52
Figura 10 – Curva suave com 6,338 graus de liberdade efetivos (g.l.e) 55
Figura 11 – Correlograma sem considerar uma estrutura de correlação 55
Figura 12 – Curva suave com 8,97 graus de liberdade efetivos (g.l.e) considerando a estrutura de correlação AR(1) 57
Figura 13 – Correlograma considerando a estrutura de correlação AR(1) 57
Figura 14 – Curvas individuais dos sujeitos ajustadas por teor de concentração 58
Figura 15 – Box-plot da variável resposta por tratamento 60
Figura 16 – *Box-plot* da variável resposta ao longo do tempo 61
Figura 17 – Perfil médio para cada sujeito observado no tempo 61
Figura 18 – Gráfico de dispersão por tratamentos 62
Figura 19 – Curva suave com 5,89 graus de liberdade efetivos (g.l.e) considerando a estrutura de correlação AR(1) 64
Figura 20 – Correlograma considerando a estrutura de correlação AR(1) 64
Figura 21 – Suavizador para o tempo semanal utilizando GAMM. 65
Figura 22 – Suavizador para o tempo semanal em cada grupo de tratamento utilizando GAMM. 66

Figura 23 – Curvas individuais dos sujeitos ajustadas por tratamento 67
Figura 24 – Curvas estimadas por tratamentos 68

Lista de tabelas

Tabela 1 – Estrutura dos dados longitudinais 32
Tabela 2 – Medidas resumo do peso dos corpos-de-prova para cada nível de concentração. 50
Tabela 3 – Medidas resumo do peso dos corpos-de-prova para cada nível de concentração, considerando os dois tipos de argamassas. 51
Tabela 4 – Comparação de modelos utilizando o critério AIC 53
Tabela 5 – Comparação de modelos com estrutura de correlação 56
Tabela 6 – Estimativas do modelo misto aditivo generalizado utilizando P-*splines*. . . . 56
Tabela 7 – Estimativas do parâmetro λ para cada concentração. 58
Tabela 8 – Medidas resumo para cada um dos tratamentos 60
Tabela 9 – Comparação de modelos com diferentes estruturas de correlação 63
Tabela 10 – Estimativas do modelo misto aditivo generalizado utilizando termos suaves não-paramétricos. 63
Tabela 11 – Estimativas do parâmetro λ para cada tratamento. 65

Lista de abreviaturas e siglas

AR	Auto-regressiva
ARMA	Auto-regressiva com Média Móvel
ARH	Auto-regressiva Heterogênea de Primeira Ordem
cH	Escala em unidade centesimal que indica a diluição de medicamentos
CS	Simetria Composta
CSH	Simetria Composta Heterogênea
GAMM	Modelo Misto Aditivo Generalizado
GAM	Modelo Aditivo Generalizado
GLM	Modelo Linear Generalizado
HF	Huynh-Feldt
SPE	Sociedade Portuguesa de Estatística
TOEP	Toeplitz
UN	Não-Estruturada
VC	Variância Composta

Sumário

1 INTRODUÇÃO . . . 17

1.1 Panorama Estatístico . . . 17

1.2 Objetivos . . . 18

1.2.1 Geral . . . 18

1.2.2 Específicos . . . 18

1.3 Revisão Bibliográfica . . . 19

1.4 Justificativas . . . 23

2 METODOLOGIA . . . 24

2.1 *Splines* . . . 24

2.1.1 *Spline* linear . . . 25

2.1.2 *Spline* cúbica . . . 25

2.1.3 B-*spline* . . . 27

2.1.4 P-*spline* . . . 30

2.2 Dados Longitudinais . . . 31

2.3 Modelo Linear Misto . . . 32

2.4 Estimação dos Efeitos Fixos . . . 34

2.5 Predição dos Efeitos Aleatórios . . . 35

2.6 Estimação de Máxima Verossimilhança Restrita (REML) . . . 36

2.7 Graus de Liberdade Efetivos . . . 37

2.8 Critério de Seleção do Parâmetro de Suavização . . . 37

2.9 Estruturas das Matrizes de Covariância . . . 39

2.10 Autocorrelação nos Resíduos . . . 44

2.11 Modelos Mistos Aditivos Generalizados (GAMM) . . . 45

2.12 Seleção de Modelos . . . 47

2.13 Aspectos computacionais . . . 47

3 1º ENSAIO: REPRESENTAÇÃO DE P-*SPLINES* COMO MODELOS MISTOS: APLICAÇÃO EM UM ESTUDO LONGITUDINAL COM DADOS DE ARGAMASSA . . . 49

3.1 Descrição do Experimento . . . 49

3.2 Resultados . . . 50

3.3 Abordagem via Modelo Linear Misto . . . 53

3.4 Abordagem via Modelo Misto Aditivo Generalizado com P-*splines* . . . 54

4 2º ENSAIO: ANÁLISE DA EVOLUÇÃO DO PESO DE CAMUNDONGOS INFECTADOS PELO *Trypanosoma Cruzi* 59
4.1 Descrição do Experimento 59
4.2 Resultados 60
4.3 Passos Futuros 68

Referências 71

Capítulo 1

INTRODUÇÃO

1.1 Panorama Estatístico

A análise de regressão é uma técnica estatística utilizada para explicar a relação entre uma variável dependente e uma ou mais variáveis independentes. A regressão é dita linear se a curva ajustada é uma reta, caso contrário é dita não linear. Na literatura, encontra-se técnicas de linearização dos dados através de transformações matemáticas, as quais evitam a utilização da regressão não linear. Ao utilizar essas técnicas, tem-se que a regressão de dados linearizados é menos precisa do que a regressão não linear, logo é necessário que técnicas apropriadas sejam utilizadas (AZEVEDO; JÚNIOR; FERNANDES, 2016).

Destaca Henderson e Parmeter (2015) que os modelos de regressão podem ser classificados em paramétricos e não-paramétricos, onde, modelos não paramétricos diferem dos modelos paramétricos no que se refere ao preditor, uma vez que este não assume uma forma predeterminada, mas é construído de acordo com as informações derivadas dos dados, e mais, o termo "não paramétrico"não significa que falta parâmetros, mas que o número e a natureza destes são flexíveis e não fixadas antecipadamente. De acordo com Fox e Monette (2002), abordagens tradicionais incluem estimadores de núcleo, análise usando *splines*, análise de Fourier entre outras.

Menciona Keele (2008), que *splines* são funções polinomiais segmentadas, unidas por pontos de corte denominados “nós”, utilizados para ajustar uma curva a um conjunto de dados. As funções *splines*, têm sido amplamente utilizadas na análise de dados longitudinais, como por exemplo, no ajuste de curvas de lactação e crescimento. Na literatura, encontram-se várias técnicas para representar este ajuste por partes, tais como: *spline* linear, *spline* cúbico, B-*spline* entre outros.

Os autores Boor et al. (1978); Eilers e Marx (1996); Keele (2008) definem que uma B-*spline* pode ser expressa como um conjunto de funções polinomiais de ordem m para cada

intervalo limitado pelos nós, sendo as conexões nos nós suavizadas, o que não ocorre na *spline* linear, tornando-se assim uma única curva contínua. Com o intuito de fornecer maior flexibilidade ao modelo e evitar um possível *"overfitting"* Eilers e Marx (1996) propuseram uma nova técnica na qual consiste na união de B-*splines* com penalização discreta (para controlar a suavidade do ajuste) inserida na função log-verossimilhança, assim o chamam de *P-spline*.

Devido ao fato de não ser possível muitas vezes descrever um fenômeno através de um modelo de regressão linear, novas classes de modelos de regressão foram surgindo, como por exemplo, modelos lineares generalizados, modelos aditivos generalizados, etc... No que se refere as técnicas de modelagem não-linear, destaca-se os Modelos Mistos Aditivos Generalizados, em que o termo $X_j\gamma = \sum x_{ij}\gamma$ é comutado por uma função não paramétrica de suavização $\sum f_j(x_{ij})$. Currie e Durban (2002) destacam que esta classe de modelos permite estimar as curvas de diferenças individuais utilizando funções de suavização não paramétrica considerando a presença de efeitos aleatórios associados aos sujeitos classificados entre grupos de tratamentos.

Assim, neste trabalho, objetiva-se utilizar a suavização via P-*spline* associada a um modelo misto aditivo generalizado em duas aplicações: Na 1ª objetiva-se utilizar os modelos mistos aditivos generalizados a um conjunto de dados cuja variável resposta é o peso de 57 camundongos suíços (sujeitos) observados no período de 12 semanas, infectados por *Trypanosoma cruzi*, machos com 56 dias de idade, provenientes do biotério central da Universidade Estadual de Maringá com o intuito de descrever a evolução do peso ao longo do tempo. Na 2ª aplicação tem-se como objetivo verificarmos a evolução do peso de dois tipos de argamassa após a absorção de água por capilaridade, sendo esta metodologia aplicada a um conjunto de dados cuja variável resposta é o peso dos corpos-de-prova, associados aos níveis de concentração de cada teor de pó de borracha adicionado, provenientes do Laboratório de Materiais de Construção Civil da Universidade Estadual de Maringá. As análises estatísticas foram realizadas no *software* estatístico R versão 3.6.0 (R Core Team, 2019).

1.2 Objetivos

1.2.1 Geral

Utilizar os Modelos Mistos Aditivos Generalizados para descrever as trajetórias individuais ao longo do tempo.

1.2.2 Específicos

- Observar o comportamento médio da variável resposta ao longo do tempo sob a intervenção;
- Descrever, por meio dos modelos mistos aditivos generalizados, a avaliação da variável

resposta;

- Realizar a análise diagnóstica.

1.3 Revisão Bibliográfica

Nessa seção, será exposta uma síntese com as principais informações referente aos resultados de um levantamento bibliográfico de artigos, dissertações e teses sobre a presença da metodologia de modelos mistos aditivos generalizados e sobre a presença de funções não-paramétricas de suavização a nível nacional e internacional.

Currie e Durban (2002) apresentam em seu artigo a aplicação das P-*splines* a três classes de modelos com componentes suaves definidas por: modelos semiparamétricos, modelos com erros serialmente correlacionados e modelos com erros heterocedásticos. Os autores mencionam que as P-*splines* são bastante flexíveis e fornecem uma abordagem comum para problemas envolvendo erros serialmente correlacionados e heterocedásticos. Estabelecem ainda abordagens não paramétricas para a escolha do parâmetro de controle da suavização das P-*splines*, assim como, o número de nós, o grau das P-*splines* e a ordem da penalização. O método de estimação adotado foi o da máxima verossimilhança restrita (REML).

Durbán et al. (2005) modelaram o efeito a longo do tempo da terapia de radiação na altura de crianças que sofrem de leucemia linfoblástica aguda usando P-*splines* na estrutura de modelos de efeitos mistos semiparamétricos. Tem-se que curvas individuais são modeladas com P-*splines*, em que estas admitem coeficientes aleatórios. Os autores mencionam que este modelo pode ser reescrito como um modelo misto e concluem que houveram diferenças significativas entre as terapias, e ainda, a taxa de crescimento de meninas no estudo não pode ser totalmente explicada pela curva média de grupo e que curvas individuais são necessárias para refletir a resposta individual ao tratamento.

Wand e Ormerod (2008) apresentam o uso das *splines* penalizadas propostas por O'Sullivan na regressão semiparamétrica contemporânea, incluindo os modelos mistos e formulações bayesianas. Os autores definem que as *splines* penalizadas de O'Sullivan são semelhantes às P-*splines*, mas possuem a vantagem de ser uma generalização direta das *splines* de suavização. Apresentam expressões exatas para a matriz de penalidade de O'Sullivan e as comparações entre os dois tipos de *splines* revelam que as *splines* penalizadas proposta por O'Sullivan admitem o comportamento natural dos limites das *splines* de suavização.

Pulgar (2009) propôs uma extensão dos modelos mistos semiparamétricos para uma classe mais geral de modelos, o qual denomina "modelos mistos aditivos semiparamétricos com erros de contornos elípticos". O objetivo do seu trabalho foi dar flexibilidade à curtose da distribuição dos erros permitindo então escolher distribuições que admitem caudas mais leves ou pesadas do que as caudas da distribuição normal padrão. O método de estimação utilizado foi o de

máxima verossimilhança penalizada com os respectivos erros padrão aproximados. No contexto semiparamétrico, para fazer inferência para as funções não paramétricas, a função de suavização escolhida foi uma *spline* cúbica. Tem-se ainda que o parâmetro de suavização foi estimado utilizando o método da validação cruzada.

Florencio (2010) modelou a variabilidade nos preços de um bem imóvel no mercado levando em consideração a não-normalidade, heteroscedasticidade e heterogeneidade espacial e estrutural utilizando os modelos aditivos generalizados para posição, escala e forma (GAMLSS) propostos por Stasinopoulos, Rigby et al. (2007). O autor, considerou como variável resposta o preço unitário do terreno e como variáveis explicativas as características estruturais, locacionais e econômicas associadas ao imóvel. Modelou algumas de suas covariáveis (por exemplo, as coodenadas geográficas referentes à localização do terreno) de forma não-paramétrica utilizando suavizadores *splines* cúbicos, onde o parâmetro de suavização foi estimado utilizando o método da validação cruzada. O autor ainda realizou uma análise comparativa entre os modelos de regressão clássicos e modelos lineares generalizados em termos do ajuste dos modelos estimados, evidenciando que os modelos GAMLSS apresentam um ajuste superior aos demais.

Souza et al. (2012) modelou a intenção de compra do consumidor em diversos setores econômicos, restringindo-se ao setor de compra coletivas, levando em consideração a influência das características de segmentos de mercado. O modelo utilizado para tal abordagem foi o modelo aditivo generalizado com nove variáveis e com um termo não paramétrico, obtido a partir do método de suavização B-*splines* [1] de ordem 4 com nós nos pontos de observação para minimizar a soma de quadrados dos resíduos penalizados. O parâmetro de suavização foi estimado utilizando o método da validação cruzada. O autor concluiu que os modelos aditivos generalizados apresentam-se de maneira satisfatória, uma vez que foram capazes de explicar a propensão de indivíduos que utilizam esse tipo de serviço levando em consideração a presença de relações não lineares.

Polansky e Robbins (2013) tiveram como objetivo destacar e avaliar a metodologia dos GAMM e quantificar as mudanças intra e interanuais na produção de frutas. Assim, os autores aplicaram os modelos mistos aditivos generalizados (GAMMs) para estudar um conjunto de dados referente a presença de frutos ao longo do tempo, coletados de árvores individuais de um local de floresta tropical de alta altitude em Afromontane, dentro do Parque Nacional Impenetrável de Bwindi (BINP), Uganda. Quanto as análises, os autores realizaram estudos de simulação para verificar a utilidade prática da seleção de modelos e estimação dos termos de suavização relevantes para desvincular a variabilidade intra e interanual. Essas simulações indicaram que a estimativa de não linearidade e sazonalidade é geralmente identificada com precisão usando a teoria assintótica e concluem que a variabilidade de frutificação no nível da floresta decorre tanto da sazonalidade regular quanto da variabilidade interanual significativa, com os anos 2009-2010, em particular, mostrando um aumento significativo na presença de

[1] Método de suavização *splines* ponderados localmente.

frutos impulsionados pelo aumento da produtividade da maioria espécies e um pico anual regular associado ocorrendo ao final de uma das duas estações secas.

Relvas (2013) apresenta os modelos simétricos parcialmente lineares com uma estrutura de autocorrelação AR(1) e erros seguindo uma distribuição simétrica, generalização dos modelos parcialmente lineares, no que diz respeito, comumente, os erros seguem uma distribuição normal. O objetivo do seu trabalho foi estudar os modelos simétricos parcialmente lineares e a estrutura de autocorrelação autoregressiva de primeira ordem, no que se refere a componente não paramétrica, esta foi modelada utilizando *splines*. O método de estimação adotado foi o de máxima verossimilhança penalizada e em relação ao diagnóstico o autor apresenta o número efetivo de graus de liberdade e também resultados assintóticos, destacando-se a obtenção da curvatura normal de influência local sob diversos esquemas de pertubação e análise de resíduos.

Machado et al. (2014) propôs uma extensão dos modelos mistos não lineares, aos quais denomina "modelos mistos semiparamétricos parcialmente não lineares, com uma função não paramétrica para modelar a média da variável resposta, utilizando a técnica de suavização com P-*splines*. Em relação às estimativas dos parâmetros foi utilizado o método da máxima verossimilhança penalizada e tem-se ainda que a análise diagnóstica foi realizada por meio do método de influência local proposto por Cook (1986). Referindo-se à aplicação, o autor utilizou a metodologia para identificar os ganhos na modelagem estatística do modelo misto semiparamétrico parcialmente não linear, ao invés de se utilizar o modelo paramétrico, tem-se ainda que as covariáveis utilizadas foram os instantes de tempo $t_{ij}(h)$ e as doses $d_i(mg/kg)$ administradas ao $i-$ *ésimo* indivíduo. A variável resposta foi a concentração de $theophylline\ conc_{ij}(mg/L)$.

Shadish, Zuur e Sullivan (2014) apresentam em seu artigo a modelagem envolvendo dados de design de caso único via modelos aditivos generalizados e os modelos mistos aditivos generalizados. Os autores mencionam que esses modelos são excelentes em detectar a forma funcional entre duas variáveis, isto é, se a tendência existe, qual é a sua forma (por exemplo, linear e não linear), e ainda, descrevem que esses modelos também podem ser aplicados em outras situações, tais como: testar se o tratamento é eficaz, se os casos diferem entre si, se os efeitos do tratamento variam entre os casos e se a tendência varia entre os casos. Neste artigo, os autores ainda discutem em detalhes a sobredispersão de dados e como os modelos mistos aditivos generalizados podem ser usados para estimar modelos auto-regressivos e por fim expõem as limitações dos modelos mistos em comparação aos modelos aditivos generalizados.

Mazzini (2015) modelou a variabilidade no ganho de peso gestacional utilizando os modelos aditivos generalizados para posição, escala e forma (GAMLSS) propostos por Stasinopoulos, Rigby et al. (2007) sob a influência de algumas covariáveis, tais como: distúrbios hipertensivos, massocromia entre outras. Em seu trabalho, a autora apresenta o ajuste do modelo GAMLSS por meio da família de distribuição BCPE (Box Cox de Potência Exponencial) com função de suavização B-*splines* e ainda, conclui que o modelo obtido para a construção da curva de peso gestacional indicou que a relação entre peso gestacional e idade gestacional não é

linear. Por fim, enfatiza que a metodologia apresenta-se de maneira satisfatória em relação à modelagem, uma vez que as curvas de percentis ajustadas foram eficientes para predizer desfechos gestacionais adversos.

Mendoza (2017) modelou a variabilidade nos níveis médios da produção de cana-de-açucar utilizando os modelos mistos aditivos generalizados, sob a influência de variáveis climáticas, tais como: temperatura e precipitação, no período de 10 anos para cada mesoregião do Estado de São Paulo. A autora ainda apresenta a abordagem, metodologia e descrição dos P-*splines* como modelos mistos e como componentes das estruturas suavizadoras de variáveis unidimensionais e bidimensionais dos modelos aditivos generalizados. Propõe seu uso em uma aplicação, no que concerne sua motivação em utilizar as P-*splines*, esta se deu pelo fato de que muitas vezes não é conhecida a tendência das covariáveis climáticas mas é de conhecimento que estas influenciam a variável resposta.

Holanda (2018) apresenta os modelos lineares parciais generalizados com uma variável explicativa contínua tratada de forma não paramétrica e os modelos lineares parciais aditivos generalizados com no mínimo duas variáveis explicativas contínuas. A autora em uma de suas aplicações modelou a variabilidade no comportamento da proporção de garotas que já menstruaram, sob a influência da idade, utilizando as P-*splines* para descrever a relação da variável resposta com as variáveis explicativas contínuas. O método de estimação adotado foi o de máxima verossimilhança penalizada, em que as funções escore penalizadas e as matrizes de informação de Fisher penalizadas são desenvolvidas para a obtenção das estimativas, por meio da combinação do algoritmo *backfitting* (Gauss-Seidel) e do processo iterativo escore de Fisher para ambos os modelos.

Soares (2019) apresenta os modelos de regressão *splines* de nós-livres como uma alternativa aos modelos não lineares utilizados em curvas de crescimento multifásico, propondo uma forma empírica de determinar o número de nós da regressão *spline*, deixando que o algoritmo de busca escolha a posição dos nós. O autor em sua aplicação modela a variabilidade no comportamento médio do peso de vacas da raça Hereford, do nascimento até o período de 675 dias. O autor realizou estudos de simulação para entender melhor a posição dos nós, localidade e grau do polinômio relacionado aos modelos de regressão *spline* de nós livres e como isto afeta a qualidade do ajuste do modelo utilizado.

Assim, este levantamento bibliográfico evidencia o cenário de pesquisa no qual se insere a proposta de pesquisa aqui descrita e o que se tem produzido nos últimos anos, no sentido de auxiliar no direcionamento da pesquisa, a fim de produzir um trabalho que possa trazer novas contribuições para esse campo de pesquisa.

1.4 Justificativas

Durante a última década, observa-se um crescimento na utilização de P-*splines* no contexto de modelos mistos, uma vez que nos modelos mistos aditivos generalizados a parte fixa do modelo é reescrita por uma função de suavização, e mais, nesta classe de modelos leva-se em consideração os efeitos aleatórios associados a cada uma das unidades amostrais observadas, sendo estas classificadas em grupos de tratamento. Uma das dificuldades inerentes a metodologia de suavização via *splines* é a decisão de quantos nós serão utilizados no ajuste e a posição destes, o qual é de suma importância o que foi dito anteriormente uma vez que essa escolha influenciará a qualidade do ajuste do modelo.

Capítulo 2

METODOLOGIA

2.1 *Splines*

É comum em um estudo longitudinal que as trajetórias das respostas médias possuam um comportamento não linear, assim torna-se inviável proceder uma análise estatística por meio do ajuste de um polinômio de m ordens, uma vez que estes não são capazes de descrevê-las. Esta situação pode ser exemplificada quando a variável resposta oscila (aumenta e diminui) entre intervalos de tempo distintos. Assim, na modelagem pode-se considerar que uma ou mais variáveis "contribuem"para explicar uma variável resposta de interesse por meio de uma função não conhecida dessa variável explicativa, ou seja, a variável explicativa "entra"no modelo no contexto não paramétrico. Para estimar essa função não conhecida, utiliza-se técnicas de suavização, como por exemplo, as *splines*.

O termo *spline* é originado da Inglaterra em meados do final do século XIX onde inicialmente sua aplicabilidade era voltada ao desenvolvimento de cascos de navios, peças de aviões e à construção civil (WEGMAN; WRIGHT, 1983). Uma *spline* é definido pelo glossário Inglês-Português de Estatística e pela Sociedade Portuguesa de Estatística (SPE) como "função definida segmentadamente por polinômios (no sentido nato)". Assim, pode-se definir uma *spline* como um conjunto de funções polinomiais de ordem m, sendo estas conectadas por pontos de corte denominados de "nós", o qual são utilizadas para ajustar uma curva a um conjunto de dados. Uma das propriedades interessantes das *splines* é que tanto as funções como suas derivadas de ordem até $m-1$ são contínuas nos pontos de corte, garantindo assim a suavidade da curva.

De acordo com Ma e Racine (2013) existem diferenças entre *splines* de regressão e *splines* de suavização, em que o primeiro é posicionado os nós em pontos equidistantes e nenhuma penalidade é imposta em relação à ausência de suavidade, já as *splines* de suavização a falta de suavidade é penalizada e os próprios dados são utilizados como nós. Na literatura encontra-se diferentes técnicas de ajuste envolvendo *splines*, tais como: *spline* linear, *spline* cúbica, b-*spline*,

p-*spline* entre outras (DURBÁN, 2009).

2.1.1 *Spline* linear

Segundo Harrell e Frank (2015), essa classe de *splines* é definida como a mais simples e pode ser expressa como um conjunto de funções lineares com diferentes inclinações para cada intervalo limitado pelos nós, e mais, as conexões nos nós não são suavizadas e apresentam mudanças na direção dos segmentos de reta[2]. Assim, considere o eixo x dividido em n intervalos, definidos pelos nós $\xi_1, \xi_2, \cdots, \xi_{n-1}$, tem-se que o modelo *spline* linear é escrito por:

$$y_i = \beta_0 + \beta_1(x_i) + \beta_2(x_i - \xi_1)_+ + \cdots + \beta_{n+1}(x_i - \xi_{n-1})_+ + \epsilon_i, \tag{2.1}$$

em que, $(x_i - \xi_{n-1})_+ = \begin{cases} (x_i - \xi_{n-1}), \text{se} & x_i > \xi_{n-1} \\ 0, \text{se} & x_i \leq \xi_{n-1} \end{cases}$, os parâmetros a serem estimados são $\beta_0, \cdots \beta_{n+1}$ e x_i o tempo decorrido até o i-*ésimo* instante com $\epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Observa-se em 2.1 que com a variação crescente no eixo x, termos serão adicionados ao modelo.Tem-se ainda que quando $x \leq \xi_{n-1}$ a função que irá modelar este intervalo será $y = \beta_0 + \beta_1(x)$ e quando $x > \xi_{n-1}$, ou seja, quando x estiver entre os nós a função que irá modelar o intervalo em estudo será 2.1, supondo ξ_{n-1} nós. A Figura 1 de Soares (2019), representa uma situação envolvendo a *spline* linear, onde observa-se as retas lineares com diferentes inclinações não suavizadas.

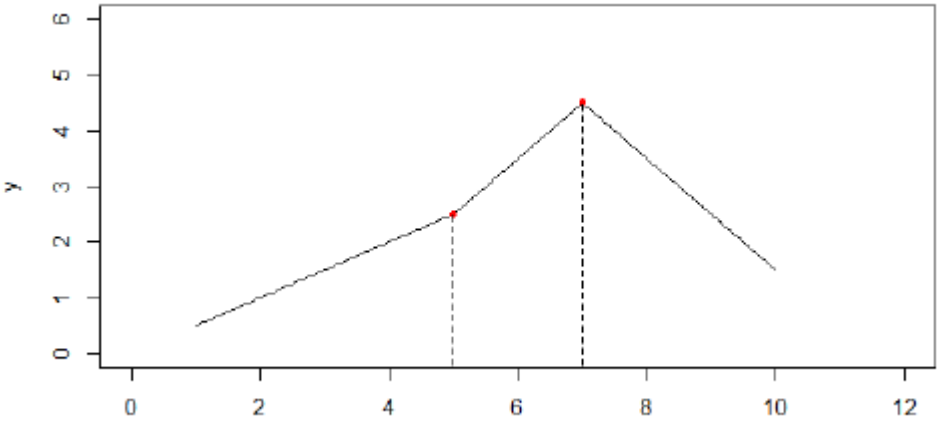

Figura 1 – Representação do modelo *spline* linear

2.1.2 *Spline* cúbica

Mencionam Stone et al. (1985) que a construção dessa classe de *splines* se difere da *spline* linear no quesito que pode ser expressa como um conjunto de funções polinomiais de ordem 3

[2] Para obter polinômios que assumam um comportamento mais suave recomenda-se olhar para polinômios de graus maiores que um.

para cada intervalo limitado pelos nós, e mais, as conexões nos nós são suavizadas o que não ocorre na *spline* linear, tornando-se assim uma única curva contínua.

A vantagem das *splines* cúbicas é a possibilidade de captar pequenas oscilações no comportamento dos dados. De acordo com Wegman e Wright (1983) as *splines* cúbicas são particionadas em duas classes: *Splines* Cúbicas Restritas e *Splines* Cúbicas não-Restritas, no qual estas se diferem no intervalo a ser modelado, em que a *spline* cúbica restrita leva em consideração as caudas (partes do polinômio antes do primeiro nó e após o último nó), sendo estas modeladas por funções lineares e a não restrita leva em consideração apenas o intervalo após o primeiro nó até o último nó. Assim, considere o eixo x dividido em n intervalos, definidos pelos nós $\xi_1, \xi_2, \cdots, \xi_{n-1}$, tem-se que o modelo *spline* cúbico restrito e não-restrito é escrito respectivamente por:

$$y_i = \beta_0 + \beta_1(x_i) + \beta_2(x_i - \xi_1)_+^3 + \cdots + \beta_{n+1}(x_i - \xi_{n-1})_+^3 + \epsilon_i, \tag{2.2}$$

e

$$y_i = \beta_0 + \beta_1(x_i) + \beta_2(x_i^2) + \beta_3(x_i^3) + \beta_4(x_i - \xi_1)_+^3 + \cdots + \beta_{n+1}(x_i - \xi_{n-1})_+^3 + \epsilon_i,$$

Logo, sua forma geral pode ser representada por:

$$y_i = \sum_{l=0}^{3} \beta_l x_i^l + \sum_{k=1}^{k} \beta_{k+3}(x_i - \xi_k)_+^3 + \epsilon_i, \tag{2.3}$$

em que $k = n - 2$. A equação 2.3, pode ser expressa por:

$$p(x_i) = \sum_{l=0}^{3} \beta_l x_i^l + \sum_{k=1}^{k} \beta_{k+3}(x_i - \xi_k)_+^3, \tag{2.4}$$

onde, $(x_i - \xi_{n-1})_+^3 = \begin{cases} (x_i - \xi_{n-1})_+^3, & x_i \geq \xi_{n-1} \\ 0, & x_i < \xi_{n-1} \end{cases}$, os parâmetros a serem estimados são $\beta_0, \cdots \beta_{n+1}$ e x_i o tempo decorrido até o i-*ésimo* instante com $\epsilon \sim^{iid} N(0, \sigma^2)$. Observa-se nas equações (2.2) e (2.3) que com a variação crescente no eixo x, termos serão adicionados ao modelo.Tem-se ainda no caso restrito que quando $x < \xi_{n-1}$ a função que irá modelar este intervalo será $y = \beta_0 + \beta_1(x)$ e no caso não-restrito a função que irá modelar este intervalo será $y = \beta_0 + \beta_1(x) + \beta_2(x_i^2) + \beta_3(x_i^3)$ e quando $x > \xi_{n-1}$, ou seja, quando x estiver entre os nós a função que irá modelar o intervalo em estudo será a função (2.2) e a função (2.3) para os casos restrito e não-restrito respectivamente, supondo ξ_{n-1} nós.

Em relação as derivadas da equação (2.4), tem-se que a primeira derivada é expressa por:

$$p(x_i)' = \beta_{01} + 2\beta_{02}(x_i) + 3\beta_{03}(x_i^2) + \sum_{j=1}^{J} 3\beta_j(x_i - \xi_{n-1})_+^2,$$

e a segunda derivada da equação *spline* (2.4) é expressa por:

$$p(x_i)'' = 2\beta_{02} + 6\beta_{03}(x_i) + \sum_{j=1}^{J} 6\beta_j(x_i - \xi_{n-1})_+,$$

Como 2.4 admite a segunda derivada contínua, esta é suave. Tem-se ainda que por ser uma função de menor ordem, sua aplicabilidade é satisfatória, uma vez que possibilita o ajuste aos dados sem ter que elevar o grau do polinômio.

A Figura 2, de Soares (2016), representa uma situação envolvendo a *spline* cúbica, cujo objetivo geral deste estudo foi de suavizar a trajetória de um míssil ao longo do tempo (em segundos),[3]. Observa-se A oscilação no comportamento de queda da tensão no período de 6 à 10 segundos e 10 à 15 segundos, sendo assim a proposta foi de ajustar um *spline* cúbico sendo o intervalo particionado em até 6,5 segundos, entre 6,5 e 13 segundos e de 13 à 20 segundos.

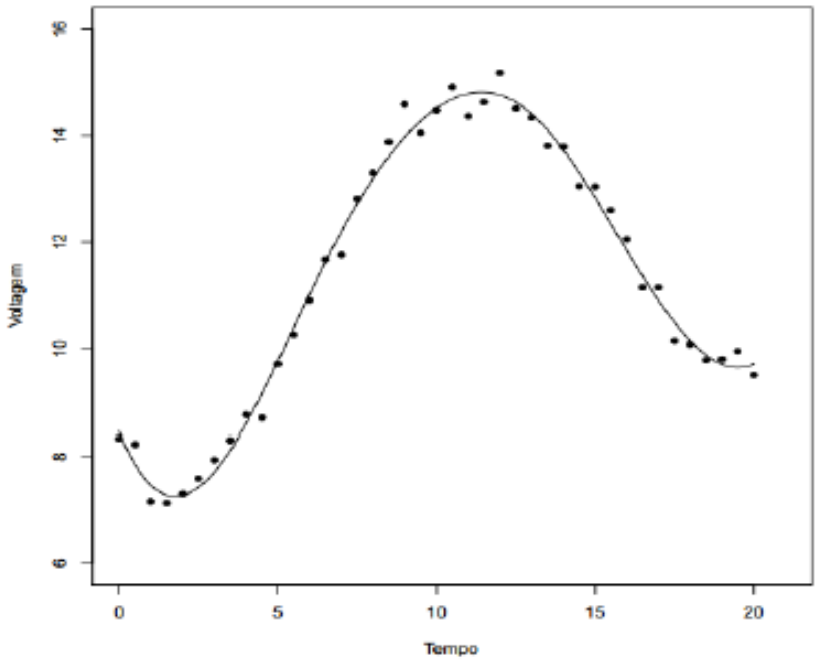

Figura 2 – Representação do modelo *spline* cúbico

2.1.3 B-*spline*

Segundo Boor et al. (1978), a construção dessa classe de *splines* se difere da *spline* linear e da *spline* cúbica no quesito que é realizada uma transformação na variável x antes do ajuste do modelo. Uma B-*spline*[4] pode ser expressa como um conjunto de funções polinomiais de ordem m para cada intervalo limitado pelos nós, e as conexões nos nós são suavizadas o que não ocorre na *spline* linear, tornando-se assim uma única curva contínua.

Destaca Holanda (2018) que uma das vantagens de se utilizar as B-*splines* é que ajusta-se polinômios de baixo grau para cada intervalo delimitado nos nós $[\xi_1, \cdots, \xi_{n-1}]$, diferente do

[3] Para mais detalhes, veja (SOARES, 2016)
[4] Também chamada de base de De Boor.

que é feito nos modelos polinomiais, sendo assim, não exige tanto esforço computacional. Uma das principais propriedades dos B-*splines* é que além da continuidade nos nós, qualquer função de grau m pode ser expressa como uma combinação de B-*splines*, em que o caso mais comum é o B-*spline* cúbico (de grau três). De acordo com Eilers e Marx (1996), seja uma sequência $\Omega_\infty = \{\xi_j\}_{\{j\in\mathbb{Z}\}}$ não decrescente de nós, tem-se que o j-*ésimo* B-*spline* de grau m é definido por:

$$B_j^m(x) = \frac{x-\xi_j}{\xi_{j+m}-\xi_j}\, B_j^{m-1}(x) + \frac{\xi_{j+m+1}-x}{\xi_{j+m+1}-\xi_{j+1}}\, B_{j+1}^{m-1}(x),$$

sendo definido que, $x \in [a,b]$, j é a quantidade de nós definida entre $a < x_1, \cdots, x_j < b$, m é o grau das B-*splines* e $B_j^m(x)$ é o j-*ésimo* componente da base que é um B-*spline* de ordem $m+1$ no ponto x. Advoga Holanda (2018) que utiliza-se comumente na literatura $ordem = 2$, ajustando-se assim um B-*spline* cúbico para cada intervalo, em uma malha de ξ_j nós equidistantes. Tem-se ainda que:

- $B_j^m(x) = 0, \forall x \notin [x_j, x_{j+m}]$
- $B_j^m(x) > 0, \forall x \in [x_j, x_{j+m}]$

Descreve Holanda (2018) em sua dissertação um diagrama explicativo do cálculo de $B_j^m(x)$, observando-se que uma $B_j^m(x)$ depende de $B_j^{m-1}(x)$ e $B_{j+1}^{m-1}(x)$.

$$\begin{array}{ccccccccccccc}
B_1^{(0)} & & B_2^{(0)} & & B_3^{(0)} & & B_4^{(0)} & & B_5^{(0)} & & B_6^{(0)} & \dots \\
\downarrow & \swarrow & \downarrow & \swarrow & \downarrow & \swarrow & \downarrow & \swarrow & \downarrow & \swarrow & & \\
B_1^{(1)} & & B_2^{(1)} & & B_3^{(1)} & & B_4^{(1)} & & B_5^{(1)} & \dots & & \\
\downarrow & \swarrow & \downarrow & \swarrow & \downarrow & \swarrow & \downarrow & \swarrow & & & & \\
B_1^{(2)} & & B_2^{(2)} & & B_3^{(2)} & & B_4^{(2)} & \dots & & & & \\
\downarrow & \swarrow & \downarrow & \swarrow & \downarrow & \swarrow & & & & & & \\
B_1^{(3)} & & B_2^{(3)} & & B_3^{(3)} & \dots & & & & & & \\
\downarrow & \swarrow & \downarrow & \swarrow & & & & & & & & \\
B_1^{(4)} & & B_2^{(4)} & \dots & & & & & & & & \\
\downarrow & \swarrow & & & & & & & & & & \\
B_1^{(5)} & \dots & & & & & & & & & &
\end{array}$$

Figura 3 – Diagrama do cálculo das B-*splines*

Por exemplo, o cálculo para $B_1^2(x)$ é dado por:

$$B_1^2(x) = \frac{x-\xi_1}{\xi_3-\xi_1}\, B_1^1(x) + \frac{\xi_4-x}{\xi_4-\xi_2}\, B_2^1(x),$$

no qual, percebe-se que o grau do polinômio influência na definição da base. A Figura 4 apresenta B-*splines* com diferentes graus polinomiais:

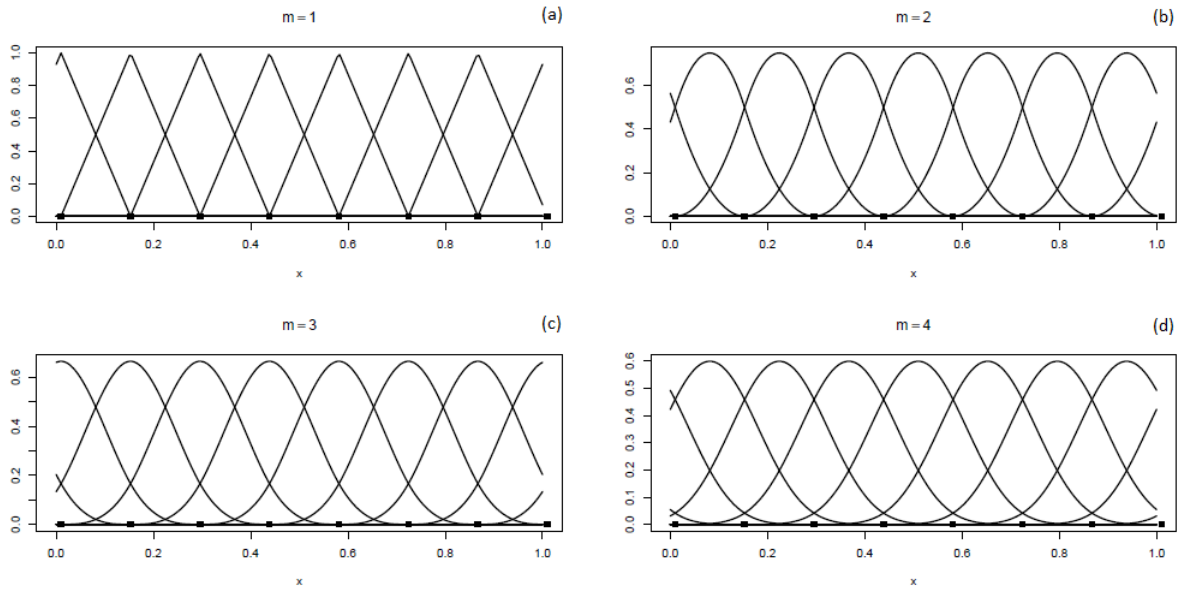

Figura 4 – Bases B-*spline* com diferentes graus polinomiais

Em relação a Figura 4 (a), de grau $m = 1$, restringindo-se no intervalo em x de $[0,01; 0,3]$, considerou-se nós em 0,01; 0,15 e 0,29. Verifica-se que uma B-*spline* de grau 1 é construída por dois segmentos lineares e três nós, sendo estes se interceptando em um nó interior, onde o primeiro segmento é formado pelos nós 0,01 e 0,15 e o segundo segmento pelos nós 0,15 e 0,29, assim, pode-se observar vários B-*splines* formados por meio das mesmas configurações. Corroborando com a propriedade de 2.5, tem-se que à esquerda do nó posicionado em 0,01 e à direita do último nó, o B-*spline* admite valor nulo.

Em relação a Figura 4 (b), de grau $m = 2$, restringindo-se no intervalo em x de $[0,01; 0,45]$, considerou-se nós em 0,01; 0,15; 0,29 e 0,43. Verifica-se que uma B-*spline* de grau 2 é construída por três componentes polinomiais de graus 2 e quatro nós, sendo estes se interceptando em dois nós internos, em que o primeiro segmento é formado pelos nós 0,01 e 0,15, o segundo segmento pelos nós 0,15 e 0,29 e o terceiro segmento pelos nós 0,29 e 0,43, assim, pode-se observar vários B-*splines* formados por meio das mesmas configurações. Corroborando com a propriedade de 2.5, tem-se que à esquerda do nó posicionado em 0 e à direita do último nó, o B-*spline* admite valor nulo.

Em relação a Figura 4 (c), de grau $m = 3$, restringindo-se no intervalo em x de $[0,01; 0,59]$, considerou-se nós em 0,01; 0,15; 0,29; 0,43 e 0,58. Verifica-se que uma B-*spline* de grau 3 é construída por quatro componentes polinomiais de graus 3 e cinco nós, sendo estes se interceptando em três nós internos, o qual o primeiro segmento é formado pelos nós 0,01 e 0,15, o segundo segmento pelos nós 0,15 e 0,29, o terceiro segmento pelos nós 0,29 e 0,43 e o quarto segmento pelos nós 0,43 e 0,58, assim, pode-se observar vários B-*splines* formados por meio das mesmas configurações. Corroborando com a propriedade de 2.5, tem-se que à esquerda do nó posicionado em 0 e à direita do último nó, o B-*spline* admite valor nulo. A

Figura 4 (d) têm sua interpretação análoga as anteriores.

Harrell e Frank (2015) enfatizam que a escolha dos nós e a quantidade destes é associada ao problema de pesquisa, uma vez que é levado em consideração a porção de dados disponíveis, porém, recomenda-se a utilização de 4 a 7 nós, fator este justificado pelo fato de que se for utilizado um número grande de nós a curva resultante fica superajustada aos dados. Portanto, com o objetivo de contornar tal situação, será utilizada uma função de penalização sobre as B-*splines*.

2.1.4 P-*spline*

Segundo Currie e Durban (2002) uma P-*spline* é definida como sendo uma combinação de B-*splines* e uma função de penalização de diferenças de ordem Δ^d associada aos coeficientes estimados das bases B-*splines* definido por γ_j, onde o uso dessa ferramenta auxilia na redução da flexibilidade das B-*splines* evitando então um superajuste da curva. Assim, a utilização das P-*splines* permite que o pesquisador fique livre na escolha do número de nós.

Os autores Eilers e Marx (1996) apresentam em seu artigo um resultado interessante a respeito da modelagem envolvendo a combinação de B-*splines*, que diz respeito a integral da segunda derivada ao quadrado pode ser expressa como uma função quadrática nos coeficientes associados ao somatório. Assim, a rugosidade de uma curva é expressa tomando uma base com vários B-*splines* associada a uma penalização de tal maneira que:

$$R = \int_a^b [f''(x)]^2 dx, \tag{2.5}$$

no qual $[a, b]$ é o intervalo de estudo, $f(x)$ é uma função que compõe somas de B-*splines* $f(x) = \sum_{j=1}^{n} a_j B_j(x)$. Porém, como ja foi mencionado, a integral definida em 2.5 pode ser expressa como uma função quadrática nos coeficientes associados ao somatório, ideia geral da P-*spline* [5], ou seja, substitui-se a 2º derivada, associada à penalização, por uma matriz de penalização discreta definida na equação (2.6):

$$S(f) = \sum_{i=1}^{m} \{f(x_i) - \sum_{j=1}^{n} \gamma_j B_j(x_i)\}^2 + \lambda \sum_{j=i+1}^{n} (\Delta^d \gamma_j)^2,$$

$$S(f) = (y - B\gamma)^T (y - B\gamma) + \lambda \gamma^T P_d \gamma, \tag{2.6}$$

onde, d é definido como a ordem do operador de diferenças Δ^d para γ_j, e ainda, este pode ser reescrito como uma matriz de dimensão $[(q + 1 + p - d) \times (q + 1 + p)]$, onde $q \in \mathbb{Z}$ e $d = p - 1$, B é definida como a função de base de regressão, λ o parâmetro de controle da suavização e tem-se ainda que Δ^d é calculado recursivamente por

$$\Delta^d = \Delta \left(\Delta^{d-1} \gamma_j \right)$$

[5] A penalização no método P-*splines* é baseada na diferença finita dos coeficientes de regressão adjacentes à base B-spline.

Para exemplificar tal situação, a função de penalização é expressa por:

$$
\begin{aligned}
& \gamma^T \Delta^T \Delta \gamma = \\
= \; & [\gamma_1, \cdots, \gamma_k, \cdots]_{1\times n} \begin{bmatrix} -1 & 0 & 0 & 0 & \cdots \\ 1 & -1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{n\times n} \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & 0 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{n\times n} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \end{bmatrix}_{n\times 1} \\
= \; & [\gamma_1, \cdots, \gamma_k, \cdots]_{1\times n} \begin{bmatrix} -1 & -1 & 0 & 0 & \cdots \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{n\times n} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \end{bmatrix}_{n\times 1} \\
= \; & [\gamma_1, \cdots, \gamma_k, \cdots]_{1\times n} \begin{bmatrix} -\gamma_1 & -\gamma_2 & 0 & 0 & \cdots \\ -\gamma_1 & 2\gamma_2 & -\gamma_3 & 0 & \cdots \\ 0 & -\gamma_2 & 2\gamma_3 & -\gamma_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{n\times n} \\
& \gamma^T \Delta^T \Delta \gamma = (\gamma_2 - \gamma_1)^2 + (\gamma_3 - \gamma_2)^2 + \cdots + (\gamma_q - \gamma_{q-1})^2,
\end{aligned}
$$

onde q é definido como a dimensão da base dos B-*splines*. Uma das vantagens de aplicar uma penalização nos coeficientes B-*splines* é a redução da dimensionalidade do problema do superajuste, pois independe do grau adotado para os B-*splines*, de tal maneira que é possível combinar qualquer ordem da penalidade com qualquer ordem das bases B-*splines*. Uma interessante propriedade inerente ao uso de P-*splines* é que podem conservar os momentos dos dados, ou seja, a média e a variância da curva estimada serão semelhante às dos dados.

2.2 Dados Longitudinais

Com o surgimento de inúmeras técnicas para resolver problemas ligados à estatística, a modelagem vem sendo utilizada para responder perguntas associadas a variáveis respostas, caracterizadas por serem mensuradas através de determinadas grandezas, tais como: tempo, distância, volume, dentre outros. De acordo com Vieira (2009), quando nos deparamos com dados dessa natureza, o planejamento apropriado é um estudo longitudinal, onde, por sua vez, observamos a variável resposta em diferentes momentos segundo a unidade amostral.

Menciona Garcia e Marder (2017) que ao nos depararmos com problemas cujos dados são de natureza longitudinal, onde a variável resposta é analisada no decorrer do tempo repetidamente em cada objeto de estudo, é possível que os pesquisadores avaliem múltiplos aspectos de um fenômeno, como: mudanças no(s) resultado(s) ao longo do tempo de acordo com os fatores de risco associados, tempo de início da doença e padrões individuais ou de grupo ao longo do tempo. Em algumas situações, descreve Almeida (2015) que as unidades observadas podem ser

dependentes, ocorrendo assim uma correlação entre as observações, na qual é preciso utilizar técnicas estatísticas apropriadas para tal situação.

Para estudos longitudinais é necessário que as observações devam estar dispostas de tal maneira que apresentem uma ligação entre as variáveis resposta e explicativas. Na Tabela 1, encontra-se um esquema para a disposição dos dados longitudinais em que diferentes observações são observadas em instantes de tempos distintos e, para cada observação, a variável resposta e as covariáveis serão observadas em cada instante de tempo.

Tabela 1 – Estrutura dos dados longitudinais

Unidade amostral	Variável Resposta	Tempo	Covariáveis X
1	y_{11}	t_{11}	x_1
1	y_{12}	t_{12}	x_1
...	...	...	...
1	y_{1p1}	t_{1p1}	x_1
2	y_{21}	t_{21}	x_2
2	y_{22}	t_{22}	x_2
...	...	...	...
2	y_{2p2}	t_{2p2}	x_2
...	...	...	...
n	y_{n1}	t_{n1}	x_n
n	y_{n2}	t_{n2}	x_n
...	...	...	...
n	y_{npn}	t_{npn}	x_n

Fonte – O autor.

Para realizar uma análise de dados longitudinais, devem ser considerados modelos estatísticos que consigam descrever a relação entre as repetidas observações na mesma unidade amostral, ou seja, modelos com efeitos aleatórios em dois estágios, como proposto por Laird, Ware et al. (1982). Modelos capazes de descrever o que foi dito anteriormente são característicos por suas distribuições de probabilidade pertencerem a uma única família, em relação às respostas dos diferentes indivíduos.

2.3 Modelo Linear Misto

De acordo com Laird, Ware et al. (1982) os modelos lineares misto são úteis quando a natureza dos dados é longitudinal, pois captam a correlação entre indivíduos relacionadas com as medidas repetidas, permitindo então ao pesquisador colher informações mais precisas a respeito do comportamento médio dos indivíduos em estudo.

Segundo Monette et al. (2010), um modelo linear misto pode ser definido na forma matricial para *N* unidades experimentais e n_i observações feitas na i-ésima unidade experimental como:

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{u_i} + \epsilon \qquad i = 1, ..., N, \tag{2.7}$$

em que $\mathbf{y_i}$ é uma matriz $(n_i \times 1)$ da variável resposta da i-ésima unidade amostral ou individuo em estudo, X_i é uma matriz $(n_i \times m)$ de posto completo[6] das variáveis regressoras dos efeitos fixos, β é uma matriz $(m \times 1)$ dos parâmetros fixos, $\mathbf{Z}_i$ é uma matriz $(n_i \times p)$, do *i*-ésimo sujeito das variáveis regressoras dos efeitos aleatórios que contém *p* preditores, u_i é um vetor$(p \times 1)$ de parâmetros de efeitos aleatórios e ϵ_i é uma matriz $(n_i \times 1)$ dos erros aleatórios intra-indivíduos, na qual:

$$\begin{aligned} u_i &\sim N(0, \mathbf{G}) \\ \epsilon_i &\sim N(0, \mathbf{R_i}), \qquad i = 1, ..., N, \end{aligned}$$

no que se refere a u_i e ϵ_i, estes são independentes devido a condição de posto completo das matrizes X e Z, G é uma matriz de variâncias e covariâncias $(p \times p)$, $\mathbf{R_i}$ é também uma matriz de variâncias e covariâncias $(n_i \times n_i)$ dependente de i.

Para qualquer conjunto de variáveis $Z_1, Z_2, ..., Z_n$ a matriz $\mathbf{G}$ de variâncias e covariâncias dos efeitos aleatórios assume uma forma não estruturada:

$$G = Var \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} \sigma^2(u_1) & \sigma(u_1, u_2) & \cdots & \sigma(u_1, u_n) \\ \sigma(u_2, u_1) & \sigma^2(u_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma(u_{n-1}, u_n) \\ \sigma(u_n, u_1) & \cdots & \sigma(u_n, u_{n-1}) & \sigma^2(u_n) \end{bmatrix};$$

onde $\sigma^2(u_i)$ denota a variância de u_i e $\sigma(u_i, u_k)$ denota a covariância de u_i e u_k. De acordo com Jones (1993) quando os $u_{i's}$ tiverem as variâncias iguais e forem independentes a matriz será diagonal. O autor citado anteriormente define a matriz de variâncias e covariâncias dos erros aleatórios $\mathbf{R_i}$ como:

$$\mathbf{R_i} = Var \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix};$$

onde σ^2 denota a variância dos erros. Ressalta Liang e Zeger (1986) que tanto a estrutura da matriz $\mathbf{G}$ como da $\mathbf{R_i}$ podem ser descritas por diversas estruturas, sendo estas empregadas individualmente ou em conjunto que é a forma mais utilizada. Assim a matriz de variância e covariância de $\mathbf{y_i}$ é definida por:

$$cov(\mathbf{y_i}) = \mathbf{V_i} = \mathbf{Z_i G Z_i^T} + \mathbf{R_i}, \tag{2.8}$$

em que pode-se concluir que a correlação entre as observações tem influência da variabilidade envolta de cada indivíduo $\mathbf{Z_i G Z_i^T}$ ou pela matriz de covariâncias intra-indivíduos $\mathbf{R_i}$. Em relação

[6] O número de linhas (ou colunas) são linearmente independentes.

a interpretação do Modelo Misto Linear definido em 2.7, sob a ótica de dados longitudinais com medidas repetidas, entende-se que $\mathbf{Y_i}$ é uma matriz que contém os valores de $\mathbf{y_{ik}}$ para o *i*-ésimo indivíduo no *k*-ésimo instante, que de acordo com Pinheiro e Bates (2006) pode ser expressa por:

$$y_{Ik}^T = [y_{11}, y_{12}, \cdots, y_{1n_1}, y_{21}, y_{22}, \cdots, y_{2n_2}, \cdots, y_{n_k}],$$

onde o *i*-ésimo sujeito é observado n_i vezes, assim entende-se que o total de observações em análise pode ser expresso por $N = \sum_{i=1}^{k} n_i$.

Se tratando das matrizes $\mathbf{Z_i}$ e $\mathbf{X_i}$, descrevem Laird, Ware et al. (1982) que $\mathbf{Z}$ é uma matriz denominada de incidência com dimensão $(n_i \times p)$ como ja foi descrito. Assim, $\mathbf{Z_i}$ é definida como uma matriz bloco diagonal da parte aleatória, onde sua diagonal contém as matrizes $\mathbf{Z_1}, \mathbf{Z_2}, \cdots, \mathbf{Z_N}$. Já a matriz $\mathbf{X_i}$, define Monette et al. (2010) que é uma matriz $(n_i \times m)$ das variáveis regressoras dos efeitos fixos que é composta por *n* valores para m preditores sendo estes assumidos sem erros.

Ainda se referindo a interpretação das componentes do modelo, como ja foi descrito, tem-se que u_i é um vetor $(p \times 1)$ de parâmetros de efeitos aleatórios. Jones (1993) descreve que $u_i = [u_{0i}, u_{1i}, \cdots, u_{(p-1)i}]$ é desconhecido e está associado a cada $\mathbf{Z_i}$ e cada $\mathbf{u_i}$ é um parâmetro de efeito aleatório específico que pode ser defindo como a diferença dos coeficientes dos *p* preditores para o *i*-ésimo indivíduo. Já a matriz $\epsilon_\mathbf{i}$ é uma matriz $(n_i \times 1)$ dos erros aleatórios $\epsilon_\mathbf{ik}$ intra-indivíduos que segundo Jones (1993) pode ser expressa por:

$$\epsilon^T = [\epsilon_{11}, \epsilon_{12}, \cdots, \epsilon_{1n_1}, \epsilon_{21}, \epsilon_{22}, \cdots, \epsilon_{2n_2}, \cdots, \epsilon_{n_k}],$$

Menciona Verbeke e Molenberghs (2009), que os parâmetros que precisam ser estimados são as variâncias e covariâncias dos efeitos aleatórios $\mathbf{G}$, as variâncias e covariâncias dos erros $\mathbf{R_i}$ e os de efeitos fixos β.

Afirmam Laird, Ware et al. (1982); Pinheiro e Bates (2006) que por meio dos erros das estimativas da média intra-indivíduos, cuja dependência é observada, que a covariância pode ser expressa, no qual nos modelos usuais os erros aleatórios são independentes e utilizados para determinar a variância. Como a variabilidade dos dados é ocasionada por fatores que são mensurados nas co-variáveis entre e intra-indivíduos, incluem-se no modelo a parte aleatória. Essa variabilidade ocorre devido a efeitos de confundimentos que são omitidos ou não observados.

2.4 Estimação dos Efeitos Fixos

Considerando o modelo 2.7, tem-se que a distribuição marginal de $\mathbf{y}$ é definida por :

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} V^{\frac{1}{2}}} exp\{-\frac{1}{2}[(\mathbf{y} - \mathbf{X}\beta)^T V^{-1} (\mathbf{y} - \mathbf{X}\beta)]\},$$

com média $E[y] = \mathbf{X}\beta$ e variância dada pela equação matricial 2.8. Assim,

$$\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{Z_i G Z_i^T} + \mathbf{R_i}),$$

A função log-verossimilhança de $(\beta, u, \sigma_u^2, \sigma_\epsilon^2)$ é expressa por:

$$L(y; \beta, u, \sigma_u^2, \sigma_\epsilon^2) = -\frac{n}{2}log(2\pi) - \frac{1}{2}log(V) - \frac{1}{2}[(\mathbf{y} - \beta)^T V^{-1}(\mathbf{y} - \beta)] \tag{2.9}$$

Derivando a equação 2.9 em relação a β e igualando a zero, tem-se que as estimativas dos efeitos fixos é

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{2.10}$$

2.5 Predição dos Efeitos Aleatórios

De acordo com Jones (1993), a predição dos efeitos aleatórios é dada pelo melhor preditor linear não viesado denominado (*Best Linear Unbiased Prediction - BLUP*), onde é imposta a suposição de que a distribuição condicional de y dado u é expressa por:

$$y|u \sim N(\mathbf{X}\beta + \mathbf{Zu}, \mathbf{R}),$$

com sua função densidade definida por:

$$f(y|u) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{R}|^{\frac{1}{2}}} exp\{-\frac{1}{2}[(y - \mathbf{X}\beta - \mathbf{Zu})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu})]\}, \tag{2.11}$$

e como foi definido na expressão 2.8, tem-se que a função densidade de u é expressa por:

$$f(u) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{G}|^{\frac{1}{2}}} exp\{-\frac{1}{2}[u^T \mathbf{G}^{-1} u]\}, \tag{2.12}$$

Robinson et al. (1991) mostram que a estimação BLUP de β e u envolve mínimos quadrados generalizados envoltos de um termo de penalização, o qual é expresso por $\beta^T \mathbf{G}^{-1} \beta$. Além disso, apresentam como é realizada esta estimação, em que, por meio da função de densidade conjunta definida por $f(y, u) = f(y|u).f(u)$ e as equações 2.11 e 2.12 têm-se:

$$f(y,u) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{R}|^{\frac{1}{2}}|\mathbf{G}|^{\frac{1}{2}}} exp\{-\frac{1}{2}[(\mathbf{y}-\mathbf{X}\beta - Zu)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zu}) + \beta^T\mathbf{G}^{-1}\beta]\},$$

e sua função de verossimilhança definida por:

$$l(y,u) = -\frac{1}{2}\{nlog(2\pi) + log(\mathbf{R}) + log(\mathbf{G}) + (\mathbf{y}-\mathbf{X}\beta-\mathbf{Zu})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zu}) + \beta^T\mathbf{G}^{-1}\beta\}, \tag{2.13}$$

e assim as soluções para β e u são dadas por:

$$\begin{aligned} BLUE(\beta) &= \tilde{\beta} = (\mathbf{X}^TV^{-1}\mathbf{X})^{-1}\mathbf{X}^TV^{-1}\mathbf{y} \\ \text{e } BLUP(u) &= \tilde{u} = \mathbf{G}\mathbf{Z}^TV^{-1}(\mathbf{y}-\mathbf{X}\tilde{\beta}) \end{aligned}$$

Advoga Schaeffer (2004) que $\tilde{\beta}$ contido no BLUP de u é substituído por $(\mathbf{X}^TV^{-1}\mathbf{X})^{-1}\mathbf{X}^TV^{-1}\mathbf{y}$ e mais, o BLUE de β é semelhante à solução de mínimos quadrados generalizados 2.10. Assim, o BLUP(β, u) é definido por:

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = [\mathbf{K^T}\mathbf{R}^{-1}\mathbf{K} + \mathbf{T}]^{-1}\mathbf{K}^T\mathbf{R}^{-1}\mathbf{y},$$

definindo-se, $\mathbf{K} = [\mathbf{X} \quad \mathbf{Z}]$ e $\mathbf{T} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^-1 \end{bmatrix}$

2.6 Estimação de Máxima Verossimilhança Restrita (REML)

Corbeil e Searle (1976) descrevem que as estimativas obtidas pelo método da máxima verossimilhança possuem viés, uma vez que não são considerados os graus de liberdade que são utilizados para a estimação dos efeitos fixos. Assim, ao invés de utilizar o método da máxima verossimilhança, utiliza-se o método da máxima verossimilhança restrita (REML) Schall (1991), no que diz respeito a este método, leva-se em consideração os graus de liberdade da estimação dos efeitos fixos. Assim, a função log-verossimilhança restrita é definida por:

$$l_{REML}(V) = l_p(V) - \frac{1}{2}log(\mathbf{X}^TV^{-1}\mathbf{X}) \tag{2.14}$$

Derivando a equação 2.14 e igualando a zero obtém-se:

$$V^{-1} = \frac{1}{\sigma_\epsilon^2}[\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + \sigma_\epsilon^2\mathbf{G}^{-1})^{-1}\mathbf{Z}^T],$$

obtendo as estimativas dos componentes de variância.

2.7 Graus de Liberdade Efetivos

Durbán (2018) menciona que se no modelo P-spline reescrito como modelo misto o parâmetro de suavização λ for igual a zero, então os graus de liberdade de ajuste do modelo configuram à dimensão da matriz base $\boldsymbol{B}$ subtraído o número de restrições do modelo, porém, se o parâmetro de suavização λ for muito grande, o modelo é dito ser pouco flexível e então terá pouquíssimos graus de liberdade, tem-se ainda que para definir os graus de liberdade de ajuste nesse tipo de modelo utiliza-se o traço da matriz $\boldsymbol{H}_\lambda$. Assim, considerando um modelo P-*spline* com *b* nós e diferença de ordem *d*, a dimensão é:

$$gl(\lambda) = traço(\boldsymbol{H}_\lambda) = traço[\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{P}_d)^{-1}\boldsymbol{B}^T]$$

Durban (2007) apresenta um resultado a respeito dos graus de liberdade de ajuste, o qual denomina como *graus de liberdade efetivos* em que este consiste em uma boa aproximação da dimensão do vetor de parâmetros definido pela matriz de suavização, em que o traço da matriz $\boldsymbol{H}_\lambda$ dependerá do parâmetro de suavização λ e mais, os valores positivos de λ refere-se aos graus de liberdade efetivos, ou seja

$$d + 1 < gl(\lambda) < d + 1 + b$$

ou seja, se λ é muito pequeno, tem-se que o traço($\boldsymbol{H}_\lambda$) $= d + 1 + b$. Já, se λ é muito grande o traço($\boldsymbol{H}_\lambda$) $= d + 1$.

2.8 Critério de Seleção do Parâmetro de Suavização

Segundo Ruppert, Wand e Carroll (2003) o parâmetro de suavização em qualquer método de suavização tem sua utilidade voltada para controlar a suavidade da curva, porém, nos modelos P-*splines* sua funcionalidade é penalizar os coeficientes que estão muito distantes um do outro, deixando claro que se o valor do parâmetro λ for muito grande os coeficientes da base $\boldsymbol{B}$ se aproximarão de zero, logo, o ajuste das curvas aproxima-se de um ajuste polinomial. Já quando o valor de λ for muito pequeno, utiliza-se o método dos mínimos quadrados ordinários. A Figura 5 apresenta as curvas ajustadas para diversos valores de λ:

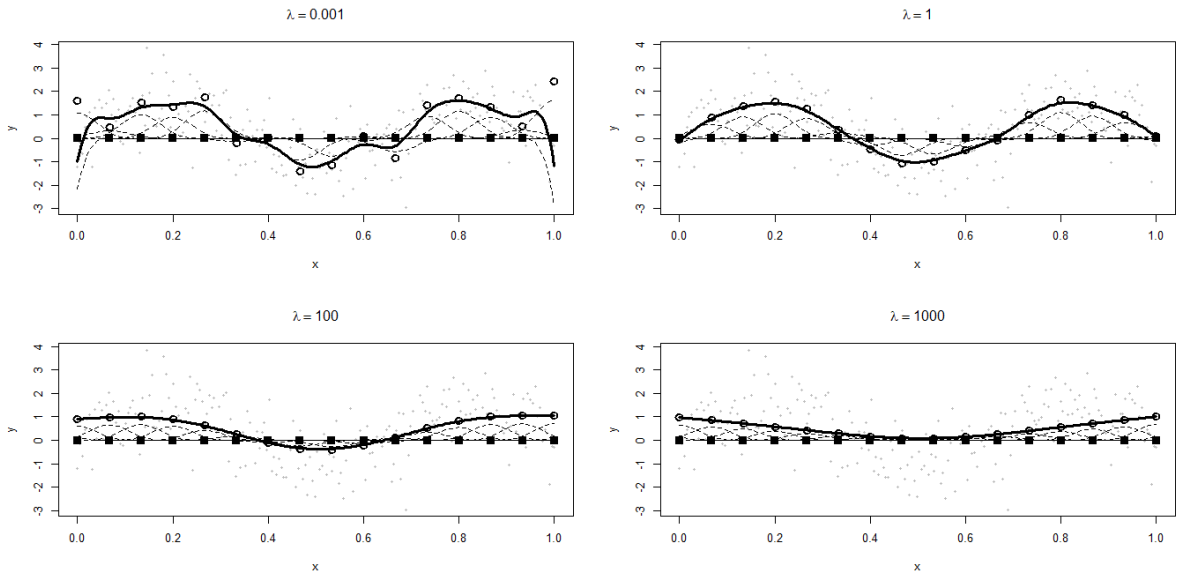

Figura 5 – Curvas ajustadas por meio de P-*splines* com diversos valores de λ

Na literatura, encontram-se diversos métodos para a seleção do parâmetro de controle da suavização, um destes é o *Método de Validação Cruzada*, definido por Friedman, Hastie e Tibshirani (2001) como:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1 - \mathbf{H}_\lambda^{ii}}\right\}^2,$$

em referência a este critério, tem-se que é uma generalização do critério de *Validação Cruzada Ord0inária*, o qual baseia-se em tirar uma observação e ajustar o modelo com as observações restantes e então calcular o quadrado da diferença entre o valor retirado e o valor predito. Assim, em oposição a realizar este processo n vezes, utiliza-se os elementos $\mathbf{H}_\lambda^{ii}$ da matriz diagonal de suavização $\boldsymbol{H}_\lambda$ (CRAVEN; WAHBA, 1978). Um resultado que generaliza o *Critério de Validação Cruzada*, denominado *Critério de Validação Cruzada Generalizada*, proposto por Friedman, Hastie e Tibshirani (2001) é definido por:

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1 - \frac{\sum_{i=1}^{n}\mathbf{H}_\lambda^{ii}}{n}}\right\}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1 - \frac{traço(\boldsymbol{H}_\lambda)}{n}}\right\}^2,$$

em que $\sum_{i=1}^{n} H_\lambda^{ii}$ é o $traço(\boldsymbol{H}_\lambda)$. Os autores Craven e Wahba (1978) ressaltam que a diferença entre as duas quantidades é ínfima e assim o melhor λ será aquele que minimiza $CV(\lambda)$ ou $GCV(\lambda)$.

2.9 Estruturas das Matrizes de Covariância

De acordo com Pinheiro e Bates (2006), as estruturas de correlação são utilizadas para modelar a matriz de covariância $\mathbf{V}_i$ referindo-se a dependência entre as observações e se tratando de modelos de efeitos mistos, elas são utilizadas para modelar a dependência entre os erros dentro do grupo. O autor citado anteriormente ainda define que as estruturas de correlação são isotrópicas[7] e a correlação geral do erro dentro do grupo é expressa para $i = 1, \cdots, k$ e k', em que $k' = 1, \cdots, n_i$, tal que:

$$cor(\epsilon_{ik}, \epsilon_{ik'}) = h[d(p_{ik}, p_{ik'}), \rho],$$

onde ρ denota o vetor de parâmetros de correlação e *h(.)* é uma função de correlação definida entre -1 e 1. Segundo Monette et al. (2010), assumindo a continuidade do vetor de parâmetros em ρ tem-se que $h(0, \rho) = 1$, ou seja, se duas observações tiverem posições idênticas, a correlação é igual a 1. Como a variância de $\mathbf{y_i}$ está relacionada com as componentes $\mathbf{Z_j G Z_i^T}$ e $\mathbf{R_i}$ é possível descrever a estrutura de $\mathbf{G}$ ou $\mathbf{R}$.

Os autores Jones (1993); Bozdogan (1987); Littell et al. (1996); Wolfinger (1993) citam que as principais estruturas de matriz de covariâncias em uma dimensão 4×4 são:

- Simetria Composta (CS);
- Não Estruturada (UN);
- Auto-regressiva de primeira ordem AR(1);
- Auto-regressiva de primeira ordem com média móvel ARMA(1,1);
- Auto-regressiva heterogênea de primeira ordem (ARH1);
- Componentes de Variância (VC);
- Toeplitz (TOEP);
- Simetria Composta Heterogênea (CSH);
- Huynh-Feldt (HF).

Simetria Composta (CS)

Esta estrutura de covariância é caracterizada pela igualdade das variâncias e covariâncias entre todos os erros dentro do grupo e é considerada a estrutura de correlação mais simples. A correlação pode ser expressa por:

$$cor(\epsilon_{ik}, \epsilon_{ik'}) = \rho, \quad \forall k \neq k', \quad h(i, \rho) = \rho, \quad i = 1, 2, \cdots, N$$

sendo o único parâmetro de correlação ρ definido como o único coeficiente de correlação intra-indivíduos. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2 + \sigma_1^2 \end{bmatrix};$$

[7] A correlação entre dois erros ϵ_{ik},$\epsilon_{ik'}$ dentro do grupo são assumidas serem dependentes .

Não Estruturada (UN)

Esta estrutura de covariância é caracterizada por impor variâncias e covariâncias distintas entre todos os erros dentro do grupo, ou seja, cada correlação é representada por um parâmetro diferente. A correlação pode ser expressa por:

$$cor(\epsilon_{ik}, \epsilon_{ik'}) = \rho_k, \quad \forall k \neq k', \quad h(i, \rho_k) = \rho_k, \quad k = 1, 2, \cdots, N$$

essa estrutura de correlação muitas vezes levará a modelos super-parametrizados. Quando há relativamente poucas observações por grupo, a estrutura de correlação geral é útil como uma ferramenta exploratória para determinar um modelo de correlação mais parcimonioso. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix};$$

Auto-regressiva de primeira ordem AR(1)

Essa família de estruturas de correlação inclui de acordo com Box, Jenkins e Reinsel (1994) diferentes classes de modelos estacionários lineares: Os modelos auto-regressivos, os modelos de média móvel e a mistura de modelos de média móvel e auto-regressivos. Estes também são chamados de modelos Box e Jenkins. A estrutura *AR (1)* é uma das mais úteis. Sua função de correlação diminui em valor absoluto exponencialmente com *lag*.[8] A correlação pode ser expressa por:

$$cor(\epsilon_{ik}, \epsilon_{ik'}) = \phi^k, \quad h(j, \phi) = \phi^k, \quad k = 0, 1, 2, \cdots, N$$

O parâmetro de correlação único, ϕ, representa a correlação de *lag*-1 e obtém valores entre -1 e 1. A estrutura *AR (1)* é caracterizada pela igualdade de variâncias e a covariância decresce à medida que as distâncias entre as medições aumentam e é uma das poucas estruturas de correlação que podem ser generalizadas para medições de tempo contínuas. Os autores Box, Jenkins e Reinsel (1994) definem que um tempo contínuo na estrutura de correlação *AR (1)* é denotado por *CAR (1)* e as componentes *k* e ϕ são maiores ou iguais a zero. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma^2 & \sigma^2\phi & \sigma^2\phi^2 & \sigma^2\phi^3 \\ \sigma^2\phi & \sigma^2 & \sigma^2\phi & \sigma^2\phi^2 \\ \sigma^2\phi^2 & \sigma^2\phi & \sigma^2 & \sigma^2\phi \\ \sigma^2\phi^3 & \sigma^2\phi^2 & \sigma^2\phi & \sigma^2 \end{bmatrix};$$

[8] A distância, ou *lag*, entre duas observações ϵ_t e ϵ_s é dada por $|t - s|$. Então *lag*-1 refere-se a observações separadas por uma unidade de tempo e assim por diante.

Auto-regressiva com média móvel ARMA(1,1)

Esta estrutura de covariância é caracterizada por assumir que a correlação dos dados é observada no tempo, cujo este é um valor inteiro e, usa-se a notação ϵ_t para se referir a uma observação feita no tempo t. Modelos auto-regressivos de média móvel, chamados modelos *ARMA*, são obtidos pela combinação de um modelo auto-regressivo e um modelo de média móvel. Antes de falarmos dos Modelos auto-regressivos de média móvel é preciso citar o modelo de média móvel cuja estrutura de correlação assume que a observação atual é uma função linear de termos de ruído [9] independentes e identicamente distribuídos.

$$\epsilon_t = \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + a_t \tag{2.15}$$

em que $[a_t]$ é o número de termos de ruído incluídos em 2.15, *q* é chamado de ordem da estrutura de covariância de média móvel, que é denotado por *MA(q)* onde existem *q* parâmetros de correlação, dado por $\theta = (\theta_1, \cdots, \theta_q)$. A correlação pode ser expressa por:

$$h(j,\theta) = \begin{cases} \dfrac{\theta_j + \theta_1\theta_{j-1} + \cdots + \theta_{j-q}\theta_q}{1 + \theta_1^2 + \cdots + \theta_q^2}, & j = 1, \cdots, q, \\ 0, & j = q+1, q+2, \cdots. \end{cases}$$

onde observações com mais de *q* unidades de tempo separadas não são correlacionadas, pois não compartilham termos comuns de ruído *[at]*. Como já foi mencionado que a estrutura de correlação *ARMA (p,q)* é obtida pela combinação de um modelo auto-regressivo e um modelo de média móvel, assim de acordo com Jones (1993), uma observação no tempo t ϵ_t é definida por:

$$\epsilon_t = \sum_{j=1}^{p} \phi_j \epsilon_{t-1} + \sum_{k=1}^{q} \theta_k a_{t-k} + a_t.$$

onde os parâmetros $p + q$ de correlação ρ correspondem a combinação dos *p* parâmetros auto-regressivos $\phi = (\phi_1, \cdots, \phi_p)$ e os *q* parâmetros de média móvel $\theta = (\theta_1, \cdots, \theta_q)$. A correlação pode ser expressa por:

$$h(i,\rho) = \begin{cases} \phi_1 h(|i-1|,\rho) + \cdots + \phi_p h(|i-p|,\rho) + \\ \theta_1 \psi(i-1,\rho) + \cdots + \theta_q \psi(i-q,\rho), & i = 1, \cdots, q, \\ \phi_1 h(|i-1|,\rho) + \cdots + \phi_p h(|i-p|,\rho), & i = q+1, q+2, \cdots. \end{cases}$$

sendo $\psi(i,\phi,\theta) = E[\epsilon_{t-i} a_t]/Var(\epsilon_t)$, onde $\psi(i,\phi,\theta) = 0, i = 1, 2, \cdots$, uma vez que neste caso ϵ_{t-i} e a_t são independentes e assim $E[a_t] = 0$. Assim, pode-se concluir que $ARMA(p,0) = AR(p)$ e $ARMA(0,q) = MA(q)$ e mais, os modelos auto-regressivo e média móvel são

[9] O ruído branco $[a_t]$ indica que as amostras são observadas como uma sequência de variáveis aleatórias não auto-correlacionadas com média zero e variância finita (ZIVOT; WANG, 2006).

exemplos particulares do modelo *ARMA (p,q)*. Tem-se ainda que a função de correlação para um modelo *ARMA (p, q)* se comporta como a função de correlação de um modelo *AR (p)* para lags maiores que *q* e como uma função de correlação *AR (p)* mais um termo relacionado à parte da média móvel do modelo entre lags 1 e *q*. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma^2 & \sigma^2\theta & \sigma^2\theta\phi & \sigma^2\theta\phi^2 \\ \sigma^2\theta & \sigma^2 & \sigma^2\theta & \sigma^2\theta\phi \\ \sigma^2\theta\phi & \sigma^2\theta & \sigma^2 & \sigma^2\theta \\ \sigma^2\theta\phi^2 & \sigma^2\theta\phi & \sigma^2\theta & \sigma^2 \end{bmatrix}; \tag{2.16}$$

Auto-regressiva heterogênea de primeira ordem ARH (1)

Destaca Davidian (2017) que esta estrutura de covariância é uma generalização da estrutura $AR(1)$ e é caracterizada pela desigualdade de variâncias e covariâncias em que ρ é o parâmetro auto-regressivo satisfazendo a seguinte condição $|\rho| < 1$. Tem-se que a estrutura *ARH(1)* é utilizada apenas se os dados são igualmente espaçados no tempo n_i e envolve $n_i + 1$ parâmetros. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix};$$

Componentes de Variância (VC)

Segundo Davidian (2017) que esta estrutura de covariância é considerada a *default* na maioria dos softwares estatísticos . Tem-se que a estrutura *VC* é definida com variâncias iguais e correlações nulas. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix};$$

Toeplitz (TOEP)

Esta estrutura de covariância de acordo com Davidian (2017) é semelhante a estrutura *AR(1)*, mas com correlações variáveis à medida que as distâncias entre os tempos crescem, ou seja, a correlação depende de um *lag*, mas não de forma exponencial. Esta estrutura pode ser definida ainda como sendo uma estrutura de médias móveis cuja ordem é igual ao tamanho da

matriz. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{13} & \sigma_{12} & \sigma^2 & \sigma_{12} \\ \sigma_{14} & \sigma_{13} & \sigma_{12} & \sigma^2 \end{bmatrix};$$

Simetria Composta Heterogênea (CSH)

Menciona Gouveia (2013) que esta estrutura de covariância é caracterizada por haver diferentes variâncias para cada elemento da diagonal principal e algumas covariâncias distintas e o coeficiente de correlação é definido de tal maneira que $|\rho| < 1$ e mais, essa estrutura contém $n_i + 1$ parâmetros. Tem-se ainda que as estruturas *CSH* e *ARH(1)* permitem variâncias distintas ao longo da diagonal principal e são extensões das estruturas básicas *CS* e *AR(1)*. O coeficiente de correlação ρ da estrutura *CSH* permanece constante, e para a estrutura *ARH(1)*, ρ é um coeficiente auto-regressivo, o que significa que as correlações vão decrescendo exponencialmente. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix};$$

Huynh-Feldt (HF)

Destaca Littell et al. (1996) que esta estrutura de covariância é caracterizada por conter $n_i + 1$ parâmetros e ser semelhante à *CSH*, uma vez que cada elemento da diagonal principal contém variâncias distintas, ou seja heterogeneidade ao longo da diagonal principal. Tem-se ainda que na estrutura *HF*, cada elemento fora da diagonal é obtido pela média aritmética de duas variâncias, subtraído do parâmetro desconhecido λ, onde λ é a diferença entre a média das variâncias e a média das covariâncias. Assim, a matriz é expressa por:

$$\mathbf{V_i} = \begin{bmatrix} \sigma_1^2 & \frac{(\sigma_1^2+\sigma_2^2)}{2} - \lambda & \frac{(\sigma_1^2+\sigma_3^2)}{2} - \lambda & \frac{(\sigma_1^2+\sigma_4^2)}{2} - \lambda \\ \frac{(\sigma_2^2+\sigma_1^2)}{2} - \lambda & \sigma_2^2 & \frac{(\sigma_2^2+\sigma_3^2)}{2} - \lambda & \frac{(\sigma_2^2+\sigma_4^2)}{2} - \lambda \\ \frac{(\sigma_3^2+\sigma_1^2)}{2} - \lambda & \frac{(\sigma_3^2+\sigma_2^2)}{2} - \lambda & \sigma_3^2 & \frac{(\sigma_3^2+\sigma_4^2)}{2} - \lambda \\ \frac{(\sigma_4^2+\sigma_1^2)}{2} - \lambda & \frac{(\sigma_4^2+\sigma_2^2)}{2} - \lambda & \frac{(\sigma_4^2+\sigma_3^2)}{2} - \lambda & \sigma_4^2 \end{bmatrix};$$

Assim, pode-se concluir que na estrutura *CS*, todas as covariâncias são iguais e, na *VC*, as (co)variâncias são nulas e que as demais estruturas consideram diferentes covariâncias para cada par de medidas. E ainda, observa-se que as estruturas que se caracterizam por haver *homogeneidade* das variâncias são: Simetria Composta (*CS*), Auto-regressiva de primeira ordem

- *AR*(1), Auto-regressiva de média móvel *ARMA* (1,1), Toeplitz (*TOEP*) e Componentes de Variância (*VC*). Já as estruturas que se caracterizam por haver *heterogeneidade* das variâncias são: Não Estruturada (*UN*), Auto-regressiva heterogênea de primeira ordem *ARH* (1), Huynh-Feldt (*HF*) e Simetria Composta Heterogênea (*CSH*).

Na literatura existem outras estruturas de variâncias e covariâncias, assim, pode-se questionar: Qual melhor estrutura de variâncias e covariâncias a ser testada? De acordo com os autores Littell et al. (1996); Vonesh e Chinchilli (1996); Davidian (2017); Lima e Demétrio (1996); Wald (2000) e Diggle (1988) é aconselhável utilizar várias estruturas de variância e covariância e que realize a escolha da estrutura mais parcimoniosa, explicativa, por meio do critério de informação de Akaike (AIC).

2.10 Autocorrelação nos Resíduos

Um importante pressuposto na análise de regressão é que os erros sejam independentes ou, pelo menos, não correlacionados. Ruppert, Wand e Carroll (2003) descrevem que se essa suposição for falsa, a estimativa dos mínimos quadrados pode ser ineficiente, e assim a principal ferramenta para detectar a dependência (correlação) dos erros é a função de autocorrelação dos resíduos, em que o k-*ésimo* lag da autocorrelação amostral é definido por $\hat{\rho}(k)$, isto é, é a correlação amostral entre e_i e e_{i-k}, em que $k+1 \leq i \leq n$.

Por exemplo, as medidas de autocorrelação lag-1 e a correlação entre resíduos adjacentes, para $k = 0, 1, \cdots, n-1$ é definda por:

$$\hat{\rho}(k) = \frac{\sum_{i=1}^{n-k} e_i e_{i-k}}{\sum_{i=1}^{n} e_i^2}.$$

Logo, se os erros forem independentes, então $\hat{\rho}(k)$ é distribuído aproximadamente normal com média 0 e variância $\frac{1}{n}$. Tem-se ainda que qualquer valor de $|\hat{\rho}(k)|$ que excede $\frac{2}{\sqrt{n}}$ é significativo. Um resultado recorrente, é que mesmo havendo um grande número de autocorrelações da amostra, algumas devem ser significativas por acaso, isso acontece pelo fato de que, se existir a autocorrelação, essa deve manifestar-se em pequenos intervalos, ou seja, será significativa para $k = 1, 2$.

Por exemplo, para dados mensais, considera-se a autocorrelação em k $= 12$, onde se a autocorrelação não for significativa, é razoável assumir que os erros são independentes. Com relação a dados longitudinais, estes são frequentemente correlacionados e é necessário verificar a autocorrelação.

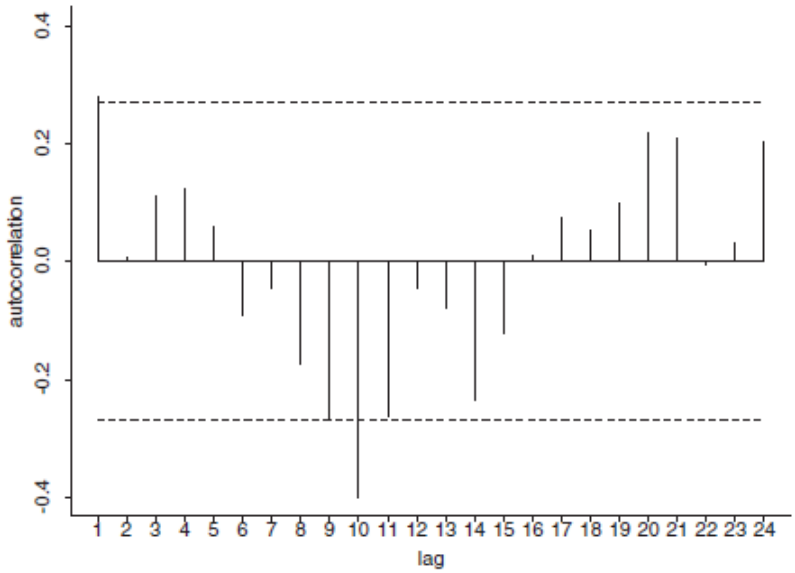

Figura 6 – Gráfico de autocorrelação

A Figura 6 apresenta os 24 primeiros valores da autocorrelação dos resíduos de um modelo ajustado em um problema de Chatterjee, Handcock e Simonoff (1995), cujo os dados são de um estudo referente ao consumo de eletricidade a partir de uma única residência ao longo do tempo. Observa-se que as linhas tracejadas horizontais estão em $\pm\frac{2}{\sqrt{n}}$, correspondente a aproximação do intervalo de confiança de 95% sob o pressuposto de independência. A autocorrelação dos valores em $k = 1$ e $k = 10$ são significativas. A autocorrelação significativa em $k = 10$ pode ser devido ao acaso como mencionado.

2.11 Modelos Mistos Aditivos Generalizados (GAMM)

Durante a última década, é visto o avanço na utilização de P-*splines* no contexto de modelos aditivos generalizados. O fato deve-se à possibilidade de reescrever o termo linear por soma de funções suaves não conhecidas, não especificadas de uma forma paramétrica, ou seja $X_j\gamma = \sum x_{ij}\gamma$ é substituído por uma função não paramétrica de suavização $\sum f_j(x_{ij})$. Neste âmbito, é possível incluir efeitos aleatórios peculiares aos sujeitos, passando a ter uma nova classe, denominada Modelos Aditivos Generalizados Mistos (GAMM).

Devido a suposição de linearidade entre a variável resposta e as variáveis regressoras em termos de uma função de ligação, os Modelos Lineares Generalizados (GLM) já não são satisfatórios referindo-se a uma possível não-linearidade. Diante disso, a utilização de modelos não paramétricos e semi-paramétricos são candidatos para tal situação, uma vez que estes são alternativas para a modelagem não-linear (MAGNANINI et al., 2010). No que se refere a essas técnicas, pode-se destacar os Modelos Aditivos Generalizados, que é uma extensão dos Modelos Lineares Generalizados, uma vez que a suposição de linearidade entre a variável

resposta e as variáveis regressoras em termos de uma função de ligação não é imposta. O Modelo Aditivo Generalizado é descrito por Durbán (2018) como:

$$\begin{aligned} g(\mu) = \eta &= \beta_0 + f(x_1) + f(x_2) + \cdots + f(x_p) = \\ &= \beta_0 + \sum f_j(x_{ij}) \end{aligned}$$

onde $f(x_j)$ são funções de suavização, e mais, esses modelos têm a característica atraente de modelar os efeitos das covariáveis sobre a resposta como uma soma de efeitos individuais.

Destaca Wood (2017) que os GAM são úteis quando:

- A relação entre a variável resposta e as variáveis regressoras é complexa, em que um modelo linear ou não-lineares canônicos não são capazes de ajustar-se;
- Não há suposições de um modelo específico;
- O objetivo é deixar os dados falarem por si só, de tal maneira que estes indiquem a forma funcional.

Durbán (2018) menciona que os Modelos semiparamétricos permitem misturas de componentes lineares (paramétricos) e não paramétricos, ou seja:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + f_1(x_{j+1}) + \cdots + f_k(x_{j+k}) + \epsilon_i$$

sendo $f_1(.); f_2(.), \cdots, f_k(.)$ funções suaves desconhecidas de $x_{j+1}, x_{j+2}, \cdots, x_{j+k}$.

Currie e Durban (2002) citam que o modelo adequado é aquele que permite estimar as curvas de diferenças específicas entre indivíduos usando termos suaves não paramétricos, ou seja, levando em consideração que há a presença do efeito aleatório a cada indivíduo, onde neste modelo é incorporado uma função $g(.)$ não paramétrica de suavização e que há a interação entre grupos de tratamentos com o preditor contínuo, com o objetivo de descrever as trajetórias individuais médias e verificar diferenças em cada um destes grupos. Assim, de acordo com Currie e Durban (2002) este modelo pode ser descrito como:

$$y_{ij} = \sum_{l=2}^{L} \gamma_l tr_{il} + f(x_{ij}) + g_i(x_{ij}) + \epsilon_{ij}, \tag{2.17}$$

no qual, γ_l corresponde a um vetor contendo os parâmetros associados aos tratamentos $(\gamma_1, ..., \gamma_l)$, $g_i(x_{ij}) = a_{i1} + a_{i2}x_{ij} + Z_i v_k$ e $f(x_{ij})$ é uma função de suavização que explica a trajetória de cada indivíduo e x_{ij} indica a i-*ésima* unidade amostral observada no j-*ésimo* instante. Tem-se ainda que a função $f(x_{ij})$ é estimada utlizando P-*splines* e mais,

- $(a_{i1}, a_{i2})' \sim N(0, \sum)$
- $\epsilon_{ij} \sim N(0, \sigma_\epsilon)$
- $tr_{il} = \begin{cases} 1, & \text{se o i-ésimo sujeito recebe o tratamento } l \\ 0, & \text{caso contrário} \end{cases}$

- $v_k \sim N(0, \sigma_v^2)$

onde as curva específicas as unidades amostrais $g_i(.)$ possuem componentes paramétricas $a_{i1} + a_{i2}x_{ij}$ e não paramétricas $\mathbf{Z_i v_k}$ sendo ambas aleatórias. Assim, pode-se escrever o modelo 2.17 na sua forma matricial como:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon$$

a matriz de efeitos aleatórios Z é definida por:

$$\mathbf{Z} = \begin{pmatrix} Z_1 & X_1 & 0 & \cdots & 0 & Z_1 & 0 & \cdots & 0 \\ Z_2 & 0 & X_2 & \cdots & 0 & 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_m & 0 & 0 & \cdots & X_m & 0 & 0 & \cdots & Z_m \end{pmatrix}$$

O vetor de efeitos aleatórios é expresso por:

$$u = (u_1, \cdots, u_K, a_{11}, a_{12}, \cdots, a_{m1}, a_{m2}, v_1, \cdots, v_K)'$$

E a matriz de covariância é escrita como:

$$\mathbf{G} = Cov(u) = \begin{pmatrix} \sigma_u^2 \mathbf{I} & 0 & 0 \\ 0 & blocodiagonal_{1 \leq i \leq m} \sum & \\ 0 & 0 & \sigma_v^2 \mathbf{I} \end{pmatrix}$$

2.12 Seleção de Modelos

A escolha ou seleção do modelo que melhor explica os dados é sem dúvida parte primordial em qualquer análise estatística inserida no contexto de modelagem. Há inúmeros testes e critérios para seleção de modelos na literatura, sendo um destes o o critério de informação de Akaike (*AIC*). A ideia por trás deste critério é comparar modelos, expresso pela seguinte equação:

$$AIC = -2l(\hat{\theta}) + 2p, \tag{2.18}$$

sendo $l(\hat{\theta})$ o logaritmo da função de verossimilhança aplicada no estimador de máxima verossimilhança de θ, $\hat{\theta}$. Assim, o modelo satisfatório será aquele que apresenta o menor valor da estatística AIC.

2.13 Aspectos computacionais

Utilizou-se o *software* R, versão 3.6.0, tanto para o tratamento dos dados como para o ajuste dos modelos utilizados. Utilizou-se a função **gamm** do pacote `mgcv` de (WOOD, 2017) para a modelagem envolvendo modelos mistos aditivos generalizados. O método de

estimação selecionado foi o `REML` por meio da função **method**, a função **lme** do pacote `nlme` de (PINHEIRO et al., 2019) foi utilizada para os ajustes dos modelos mistos. Para selecionar as bases B-splines utilizou-se a função **bs** do pacote `splines`. Para a construção dos gráficos utilizou-se os pacotes `lattice` de (SARKAR, 2008), `latticeExtra` de (SARKAR; ANDREWS, 2016) e `ggplot2` de (WICKHAM, 2016).

Capítulo 3

1º ENSAIO: REPRESENTAÇÃO DE P-*SPLINES* COMO MODELOS MISTOS: APLICAÇÃO EM UM ESTUDO LONGITUDINAL COM DADOS DE ARGAMASSA

3.1 Descrição do Experimento

O experimento que deu origem ao conjunto de dados deste trabalho foi executado por Canova (2008), em que o experimento foi desenvolvido utilizando adição de pó de borracha moída de pneus inservíveis na composição de dois tipos de argamassa de revestimento, visando a ocorrência da redução da exigência de água para a consistência padrão obtida na argamassa maturada seca.

Para avaliar o teor de pó de borracha e o fator argamassa nas propriedades dos pneus estudadas, foi utilizado um experimento fatorial completo. Isto é, foram considerados 5 níveis para a variável teor de pó de borracha (0%, 6%, 8%, 10% e 12%), 2 níveis para a variável Argamassas (argamassa maturada com pó de borracha e argamassa maturada seca com 60 dias de armazenamento com adição de pó de borracha), e 6 repetições que é igual ao número de corpos de prova ou experimentos, totalizando 60 corpos-de-prova.

A argamassa contendo como aglomerante a cal virgem e como agregado miúdo a areia fina lavada de rio recebeu a denominação de simples, como normalmente é chamada. A argamassa simples, após receber a adição de cimento, passa a ser uma argamassa composta, aqui denominada argamassa maturada, embora se saiba que a maturação é feita antes da

adição de cimento. A identificação da argamassa maturada como uma composição de referência recebeu a denominação *Am*. A argamassa simples de cal virgem e areia que passou pelo processo de maturação, foi seca em estufa e recebeu o acréscimo de adição de cimento, foi denominada argamassa seca em estufa. A identificação dessa argamassa como uma composição de referência recebeu a denominação *As* (CANOVA, 2008).

O experimento baseou-se no procedimento da NBR 9779/1995, porém observou que a ascensão capilar atingiu o topo do corpo-de-prova em menos de 5 horas. Como a Norma expõe que o ensaio não tem significado se a ascensão capilar máxima atingir a altura do corpo-de-prova, foi necessário fazer alterações nos tempos colocados pela metodologia sugerida para realização das leituras. Adotaram-se tempos para as leituras, como segue: até os 90 minutos a cada 10 minutos, de 90 a 150 minutos a cada 15 minutos, de 150 a 360 minutos a cada 20 minutos, de 360 a 450 minutos a cada 45 minutos e de 450 a 1.350 minutos a cada 60 minutos. Para acompanhar a altura da umidade superficial do corpo de prova, foram feitos marcos a cada 1 cm na altura do corpo-de-prova em três geratrizes, para facilitar as leituras. Após a umidade atingir o topo do corpo-de-prova, fez-se o rompimento diametral, para observações na parte interna do corpo-de-prova (CANOVA, 2008).

3.2 Resultados

Buscou-se relacionar a evolução do peso de dois tipos de argamassa após a absorção de água por capilaridade, considerando o nível de concentração do teor de pó de borracha adicionado.

A Tabela 2 apresenta as estatísticas resumo da variável resposta para cada um dos níveis de concentração, em que pode-se concluir que a média do peso da argamassa em relação a cada um dos teores variam de tal maneira que a média do peso dos corpos-de-prova cujo teor adicionado foi de 0% é de $358,3$g, já a média do peso dos corpos-de-prova cujo teor adicionado foi de 12% é de $316,1$g

Tabela 2 – Medidas resumo do peso dos corpos-de-prova para cada nível de concentração.

	Teores	n	Média	Desvio Padrão	Min.	Max.	Coeficiente de Variação
1	0%	12	358,3	12,40	325,6	382,1	0,034
2	6%	12	336,1	9,87	308,4	357,4	0,029
3	8%	12	327,3	8,48	304,7	345,8	0,025
4	10%	12	325,1	8,32	302,1	344,9	0,025
5	12%	12	316,1	7,76	294,8	336,6	0,024

A Tabela 3 apresenta as estatísticas resumo da variável resposta para cada um dos níveis de concentração, considerando os dos tipos de argamassas, no qual conclui-se que a média do peso da argamassa maturada com pó de borracha é superior que a média do peso da argamassa maturada seca. Quanto aos níveis dos teores adicionado, a média do peso dos corpos-de-prova

é superior em ambos os tipos de argamassa no nível de concentração de 0%, sendo $356,50$g e $360,20$g respectivamente, evidenciando que quanto maior o nível de concentração, menor é o peso, em média, dos corpos-de-prova em ambos os tipos de argamassas.

Tabela 3 – Medidas resumo do peso dos corpos-de-prova para cada nível de concentração, considerando os dois tipos de argamassas.

Teores	n		Média		D.P		Min.		Máx.		C.V	
	As	Am	As	Am	As	Am	As	Am	As	Am	As	Am
0%	6	6	356,50	360,20	11,79	12,79	325,60	327,60	375,50	382,10	0,03	0.03
6%	6	6	336,90	335,20	8,56	11,03	312,70	308,40	352,00	357,40	0,02	0.03
8%	6	6	326,30	328,60	8,27	8,60	304,70	308,40	345,00	345,80	0,02	0.02
10%	6	6	325,00	325,30	7,77	9,01	304,10	302,10	342,70	344,90	0,02	0.02
12%	6	6	314,70	318,00	6,04	9,27	299,20	294,80	327,70	336,60	0,01	0.02

D.P = Desvio Padrão; C.V = Coeficiente de Variação; As = Argamassa maturada seca; Am = Argamassa maturada com pó de borracha.

A Figura 7 apresenta o box-plot da variável resposta para cada um dos teores e o box-plot da variável resposta considerando ainda os dois tipos de Argamassas, em que pode-se concluir que nos níveis de concentração mais altos, a média do peso dos corpos-de-prova diminui.

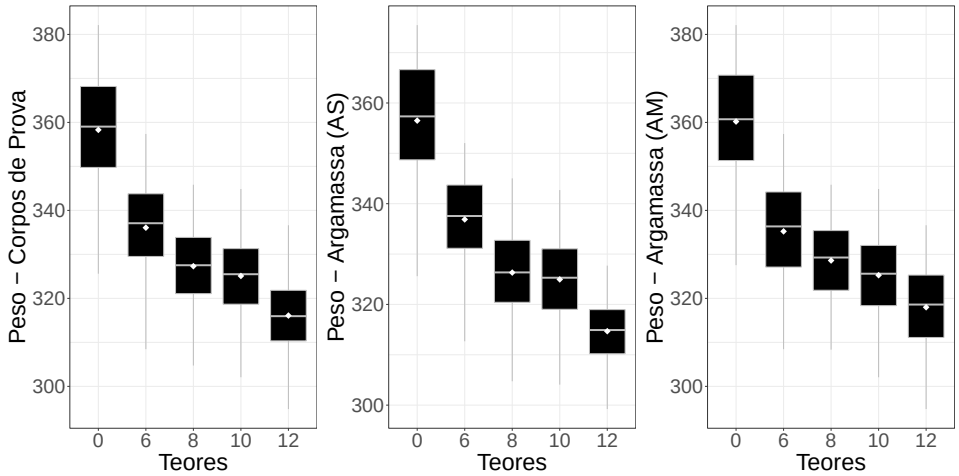

Figura 7 – *Box-plot* do peso dos corpos-de-prova para cada nível de concentração do pó de borracha dos pneus inservíveis e box-plot para cada tipo de argamassa.

A Figura 8 apresenta os gráficos de perfis médio da variável resposta para cada um dos corpos-de-prova e os gráficos de perfis médio da variável resposta para cada tipo de Argamassa, verificando-se um comportamento crescente conforme varia o tempo, e mais, vê-se ainda um comportamento oscilatório já no instante inicial.

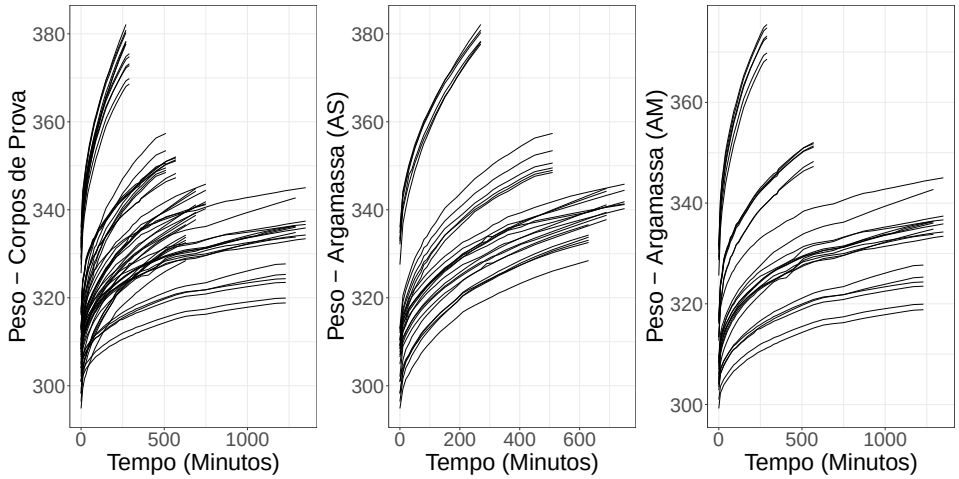

Figura 8 – Perfil médio global para cada corpo-de-prova e perfil médio para cada tipo de argamassa observado no tempo

A Figura 9, apresenta o gráfico de dispersão por teores, no qual observa-se uma curva de ajuste para cada teor levando em consideração cada tipo de argamassa e mais, verifica-se uma quantidade de observações que estão fora dessa curva de ajuste, ou seja, existe a possibilidade de existir uma variabilidade entre essas observações, justificando-se o fato de poder utilizar algum tipo de suavização que leve em consideração a variabilidade constatada.

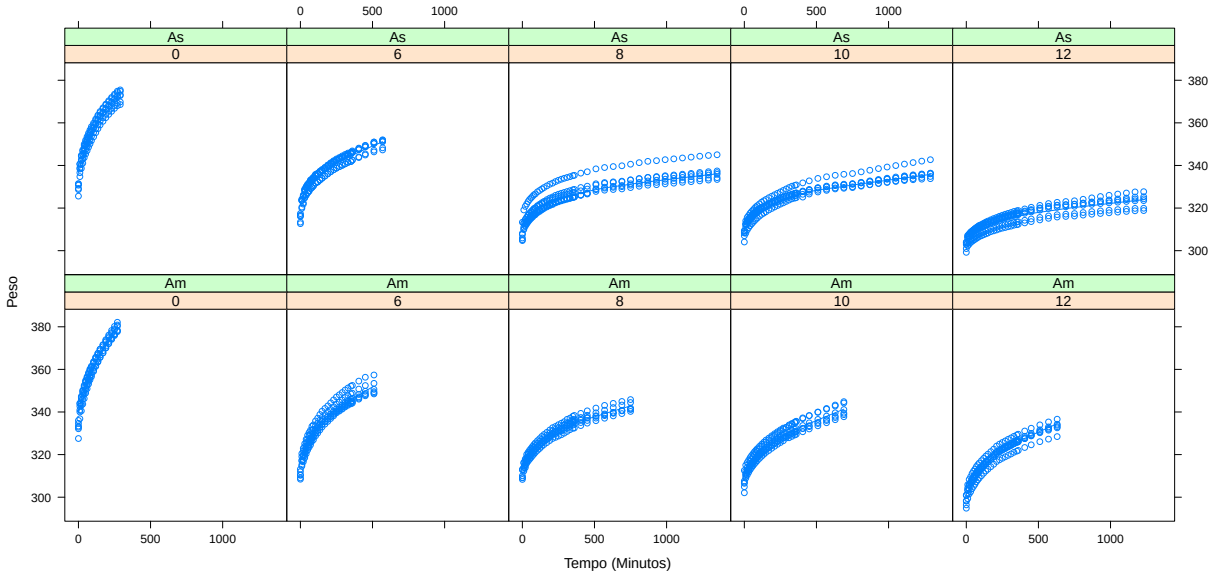

Figura 9 – Gráfico de dispersão do peso dos corpos-de-prova para cada tipo de argamassa e níveis de concentração do pó de borracha dos pneus inservíveis.

Verifica-se ainda que o peso observado após a absorção de água por capilaridade em cada tipo de argamassa, apresenta muita variação entre o intercepto e inclinação das retas que foram ajustadas, isso leva a considerar modelos mistos com efeitos aleatórios sejam eles independentes ou correlacionados entre intercepto e inclinação em relação ao tempo, neste caso para a covariável Tempo.

3.3 Abordagem via Modelo Linear Misto

Quanto a modelagem, a abordagem via modelos mistos de acordo com Laird, Ware et al. (1982) é utilizada quando a natureza dos dados é longitudinal, no que refere-se essa metodologia, é possível descrever a correlação intra indivíduos relacionadas com as medidas repetidas. A Tabela 4 contém os modelos ajustados, onde considerou-se o efeito aleatório na variável Tempo e o intercepto a_{i1} em que estes modelos são definidos com intercepto e inclinações aleatórias independentes:

Tabela 4 – Comparação de modelos utilizando o critério AIC

Modelos	AIC
$\text{Peso}_{ijk} = \beta_0 + a_{i1} + (\beta_1 + a_{i2})\text{Tempo}_{ijk} + \beta_2 \times \text{Argamassas}_{ijk} + \beta_3 \times \text{Teor}_{ijk} + \epsilon_{ijk}$	10643,7
$\text{Peso}_{ijk} = \beta_0 + a_{i1} + (\beta_1 + a_{i2})\text{Tempo}_{ijk} + \beta_2 \times \text{Argamassas}_{ijk} + \epsilon_{ijk}$	10964,1
$\text{Peso}_{ijk} = \beta_0 + a_{i1} + (\beta_1 + a_{i2})\text{Tempo}_{ijk} + \beta_3 \times \text{Teor}_{ijk} + \epsilon_{ijk}$	15350,9
$\text{Peso}_{ijk} = \beta_0 + a_{i1} + (\beta_1 + a_{i2})\text{Tempo}_{ijk} + \epsilon_{ijk}$	15512,2

Assim, o modelo linear misto considerado de acordo com a estatística AIC é expresso por:

$$\text{Peso}_{ijk} = \beta_0 + a_{i1} + (\beta_1 + a_{i2})\text{Tempo}_{ijk} + \beta_2 \times \text{Argamassas}_{ijk} + \beta_3 \times \text{Teor}_{ijk} + \epsilon_{ijk}, \quad (3.1)$$

os quais são tratados como modelos com intercepto e inclinação aleatórios independentes, daí que apresenta-se a possibilidade de modelar heterogeneidade individual tanto para os interceptos como também ao longo do tempo. Tem-se que o Peso do i-*ésimo* corpo-de-prova é explicado pelos efeitos fixos $\beta_1 \times \text{Tempo}_{ijk}$, $\beta_2 \times \text{Argamassas}_{ijk}$, $\beta_3 \times \text{Teor}_{ijk}$ e pelo efeito aleatório a_{i2} que se soma ao valor ao intercepto fixo do modelo β_0, para formar o intercepto do i-*ésimo* corpo-de-prova, tem-se ainda que o índice i está associado a a_{i1} e, assim o modelo permite que j-*ésimo* tempo apresente uma relação diferente entre o i-*ésimo* sujeito.

Tem-se ainda que a_{i1} representa o intercepto aleatório individual para cada curva, assumido com distribuição normal com média 0 e variância $\sigma^2_{corpo-de-prova}$, a_{i2} representa as inclinações aleatórias individuais para cada curva, com distribuição normal de média 0 e variância σ^2_{Tempo}, assim observa-se que $(a_{i1}, a_{i2})'$ segue distribuição normal com vetor de médias 0 e matriz de variância-covariância para os interceptos e inclinações aleatórias denotada por G, tem-se ainda que os resíduos ϵ_{ijk} tem distribuição normal com média 0 e variância σ^2_ϵ e que os termos aleatórios são assumidos independentes.

Ao utilizar modelos somente com intercepto aleatório, têm-se que todas as curvas individuais são paralelas, ou seja, os modelos utilizados não seriam capazes de modelar cada curva com a sua própria característica e de maneira individual (DURBÁN, 2009). Logo, é por essa razão de propor e utilizar o modelo proposto por Durbán et al. (2005) com diferentes curvas individuais, isto é, que cada curva apresente uma determinada inclinação o que permite que cada curva individual seja diferente. Têm-se que as estimativas dos parâmetros dos efeitos fixos e das componentes de variância foram obtidas utilizando o *método REML*.

3.4 Abordagem via Modelo Misto Aditivo Generalizado com P-*splines*

Nessa seção, encontram-se os resultados considerando a função de suavização no tempo para cada teor e tipo de argamassa dentro do modelo aditivo, em que a vantagem de utilizar as funções de suavização, é que é possível descrever as trajetórias da evolução do peso dos tipos de argamassa após a absorção de água por capilaridade ao longo do tempo e o comportamento não linear dos efeitos e mais, considerando a possível existência de heterocedasticidade. Será definido uma estrutura de correlação, fornecendo então as estimativas das trajetórias em função do tempo e reduzindo a estrutura dos erros.

O modelo considerado é descrito por:

$$\textit{Peso}_{ij} = \sum_{l=1}^{L} \gamma_l Teor_{il} + f_l(Tempo_{ij}) + \beta_1 \times Argamassas_{ij} + a_{i1} + a_{i2} Tempo_{ij} + \epsilon_{ij},$$

$$Teor_{il} = \begin{cases} 1, & \text{se o i-ésimo corpo-de-prova recebe o } l\text{-ésimo teor de pó de borracha} \\ 0, & \text{caso contrário} \end{cases} \tag{3.2}$$

em que, γ_l corresponde a um vetor contendo os parâmetros associados aos tratamentos $(\gamma_1, ..., \gamma_l)$, $Peso_{ij}$ o peso do i-*ésimo* corpo-de-prova no j-*ésimo* instante de tempo, onde $i = 1, \cdots, 60$ indivíduos, j = 1, $\cdots$, 1350 minutos, divididos em até os 90 minutos a cada 10 minutos, de 90 a 150 minutos a cada 15 minutos, de 150 a 360 minutos a cada 20 minutos, de 360 a 450 minutos a cada 45 minutos e de 450 a 1350 minutos a cada 60 minutos e l=1,$\cdots$,5 teores de pó de borracha. Tem-se ainda que f_1 denota a curva média para os corpos-de-prova que contém 0% do teor de pó, f_2 denota a curva média para os corpos-de-prova que contém 6% do teor de pó, f_3 denota a curva média para os corpos-de-prova que contém 8% do teor de pó, f_4 denota a curva média para os corpos-de-prova que contém 10% do teor de pó e f_5 denota a curva média para os corpos-de-prova que contém 12% do teor de pó, a_{i1} e a_{i2} denotam os efeitos aleatórios correspondentes aos interceptos e inclinações das curvas.

A $f(Tempo_{ij})$ é a função de suavização desconhecida da variável *Tempo*, que descreve a trajetória dessa variável em relação a evolução do peso dos corpos-de-prova, o qual estima-se $f(.)$ por meio de P-*splines*, com função de base B-*splines* de terceiro grau polinomial, e uma penalização discreta de segunda ordem, a_i indica o intercepto aleatório que segue distribuição Normal com média 0 e variância $\sigma^2_{sujeito}$ e os resíduos ϵ_{ij} tem distribuição normal com média 0 e variância σ^2_{ϵ} e que os termos aleatórios são assumidos independentes.

A Figura 10 apresenta a tendência da covariável *Tempo* sem considerar uma estrutura de correlação nos resíduos, em que pode-se verificar uma curva suave com quantidade de suavização de 6,338 graus de liberdade efetivos (g.l.e).

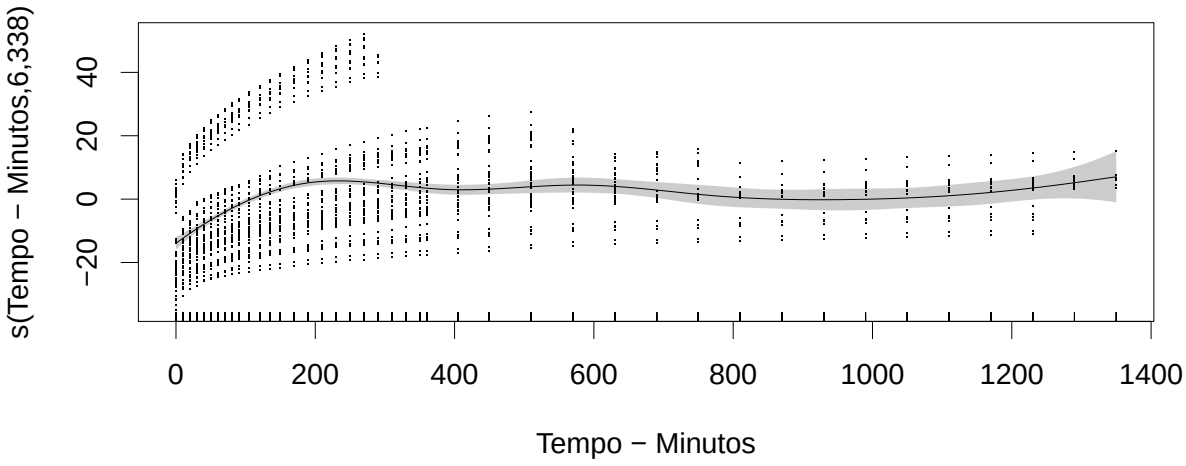

Figura 10 – Curva suave com 6,338 graus de liberdade efetivos (g.l.e)

A Figura 11 apresenta o correlograma, utilizado para verificar se os resíduos estão ou não correlacionados. Observando a figura conclui-se que os resíduos estão correlacionados, uma vez que a uma série de picos fora do limite de aceitação, enquanto que o comportamento decrescente dos valores de autocorrelação ocorrem por conta de uma tendência na evolução da absorção de água por capilaridade nos corpos-de-prova, evidenciando assim a violação da pressuposição de independência. Para contornar tal situação, pode-se definir uma estrutura de correlação nos resíduos.

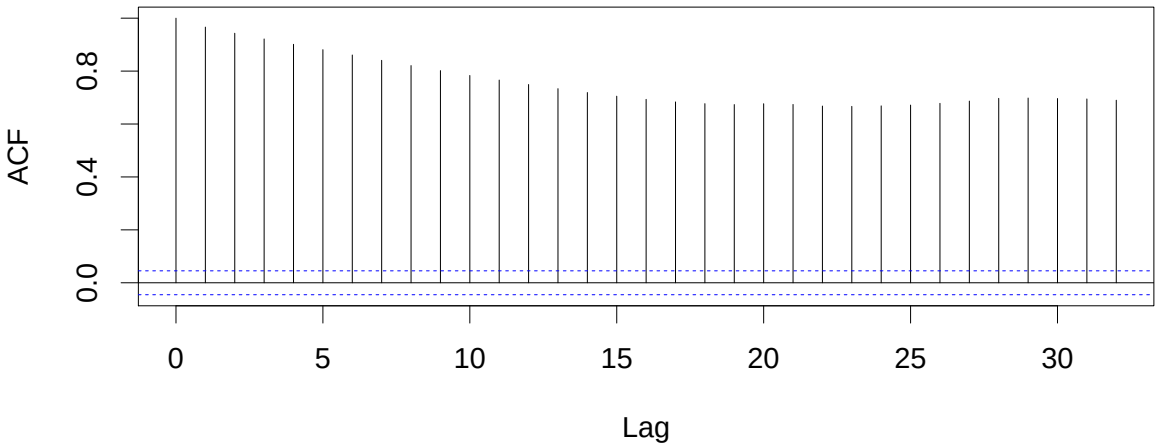

Figura 11 – Correlograma sem considerar uma estrutura de correlação

A Tabela 5, apresenta os modelos ajustados, definidos por: a estrutura de correlação AR(1),

CS e ARMA(1,1), considerando a função não-paramétrica de suavização na covariável *Tempo*. Para selecionar o modelo utilizou-se o critério de Akaike.

Tabela 5 – Comparação de modelos com estrutura de correlação

Modelo	AIC
Estrutura AR(1)	4314,40
Estrutura CS	5443,40
Estrutura ARMA(1,1)	10527,56

Assim, ao observar a Tabela 5 conclui-se que o modelo selecionado é aquele com a estrutura de correlação AR(1), uma vez que este admite o menor valor da estatística AIC.

A Tabela 6 contém as estimativas do modelo referente aos coeficientes paramétricos e as estimativas dos termos suaves, em que verifica-se que os Teores não foram significativos (valor-p $>$ 0,05), porém a Argamassa (As) foi significativa (valor-p $=$ 0,006) e ainda, que os termos suaves foram significativos (valor-p $<$ 0,001).

Tabela 6 – Estimativas do modelo misto aditivo generalizado utilizando P-*splines*.

Coeficientes Paramétricos:				
Efeito	Estimativa	Erro Padrão	Estatística de Teste	Valor-p
Intercepto	255,88	47,02	5,44	<0,001
Teor - 6%	88,89	47,42	1,87	0,06
Teor - 8%	71,50	47,03	1,52	0,12
Teor - 10%	69,61	47,03	1,48	0,13
Teor - 12%	61,77	47,03	1,31	0,18
Argamassa (As)	-1,77	0,65	-2,72	0,00
Significância dos termos suaves				
Efeito	gle	Ref.gl	Estatística de Teste	Valor-p
s(Tempo): Teor 0%	3,95	3,95	1589,60	<0,001
s(Tempo): Teor 6%	4,97	4,97	1079,70	<0,001
s(Tempo): Teor 8%	8,85	8,85	709,40	<0,001
s(Tempo): Teor 10%	8,38	8,38	605,60	<0,001
s(Tempo): Teor 12%	7,93	7,93	437,20	<0,001
R^2.(adj) $=$ 0,93		Deviance explicada $=$ 94,2%		
UBRE score $=$ 1,52		Parâmetro de Escala est. $=$ 0,64		n $=$ 1884

Em relação aos graus de liberdade efetivos (gle), estes indicam uma tendência não-linear, onde estes variam de (gle $=$ 3,95) para o teor de 0% à (gle $=$ 8,85) para o teor de 8%.

A Figura 12 apresenta a tendência da covariável *Tempo* considerando a estrutura de correlação AR(1) nos resíduos. Observa-se uma curva suave com quantidade de suavização de 8,97 graus de liberdade efetivos (g.l.e). Verifica-se ainda que nos instantes iniciais os corpos-de-prova atingem o maior peso em média aos 250 minutos $s(Tempo) = 241,13$ e a partir desse período verifica-se uma oscilação referente ao pesos destes, ou seja, no início do experimento observou-se que os corpos-de-prova demoram um tempo maior quanto a absorção de água por capilaridade.

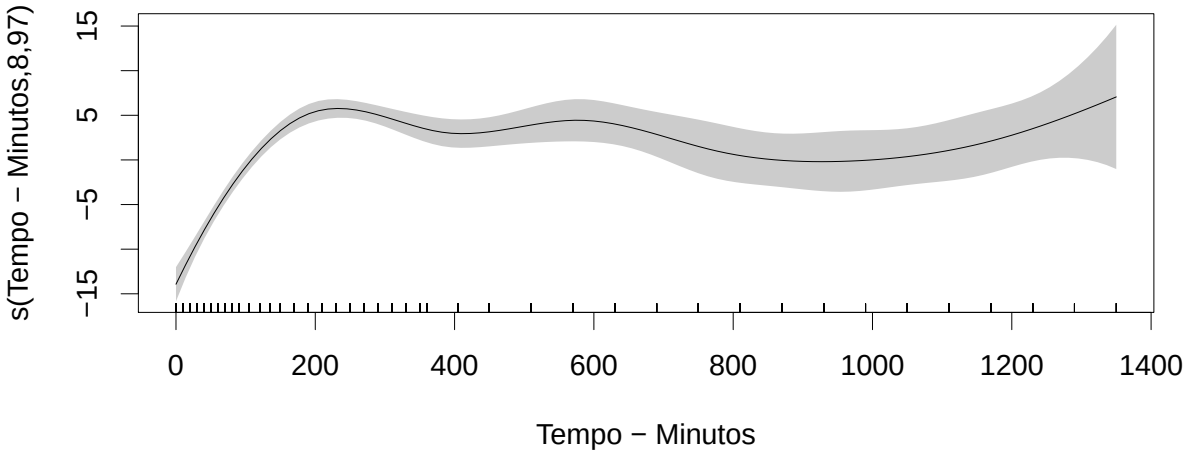

Figura 12 – Curva suave com 8,97 graus de liberdade efetivos (g.l.e) considerando a estrutura de correlação AR(1)

Observa-se o intervalo de confiança de 95% da evolução do peso dos corpos-de-prova em função do tempo, concluindo que o efeito dessa covariável segue um padrão não linear.

A Figura 13 apresenta o correlograma com a estrutura AR(1) nos resíduos, em que observa-se a redução da problemática a respeito da violação de uma das pressuposições dos erros.

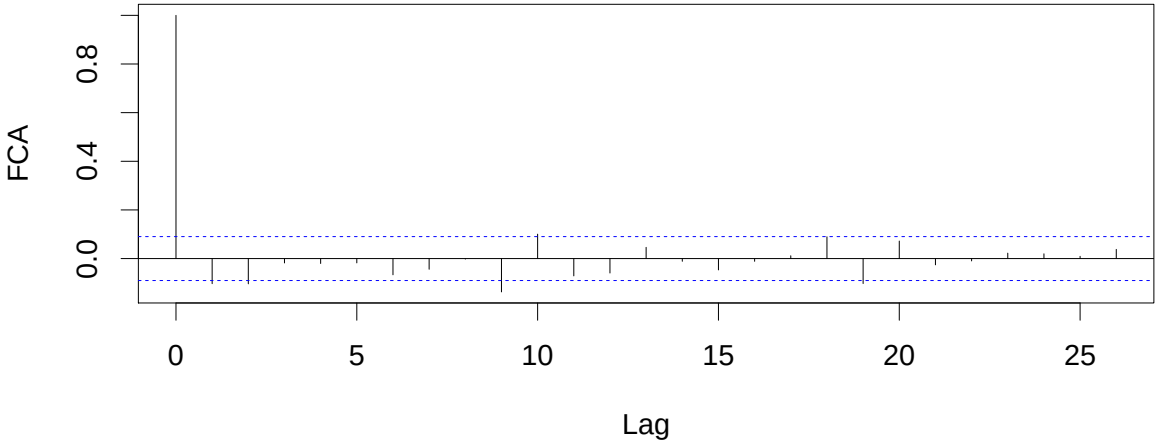

Figura 13 – Correlograma considerando a estrutura de correlação AR(1)

Com o intuito de verificar a hipótese de independência, realizamos o teste de Durbin-Watson, em que a estatística de teste resultante foi de 1,92 e o (valor-p = 0,22), isto é, não há a presença de autocorrelação nos resíduos.

A Tabela 7 contém as estimativas do parâmetro λ, com seus respectivos graus de liberdade efetivos (g.l.e), teste F e o p-valor associado. Conclui-se que a suavização da a covariável *Tempo* para cada *Teor* foi altamente significativa em relação a evolução do Peso.

Tabela 7 – Estimativas do parâmetro λ para cada concentração.

Suavizador	g.l.e	Parâmetro (λ)	F	p-valor
s(tempo):Teor 0%	3,95	$4{,}0\times10^{-4}$	865,17	<0,001
s(tempo):Teor 6%	4,97	$5{,}0\times10^{-3}$	271,60	<0,001
s(tempo):Teor 8%	8,91	$2{,}3\times10^{-2}$	86,99	<0,001
s(tempo):Teor 10%	8,75	$2{,}9\times10^{-2}$	81,25	<0,001
s(tempo):Teor 12%	8,00	$3{,}5\times10^{-2}$	75,96	<0,001

A Figura 14 contém as curvas estimadas específicas aos corpos-de-prova alocados em suas respectivas concentrações de teores separadas pelos tipos de argamassa testados, em que o modelo 2.17 descreve estas levando em consideração a não-linearidade e os efeitos aleatórios associados. Observa-se ainda a capacidade de ajuste do modelo utilizado com 9 nós, apresentando-se de maneira satisfatória quanto a interpolação.

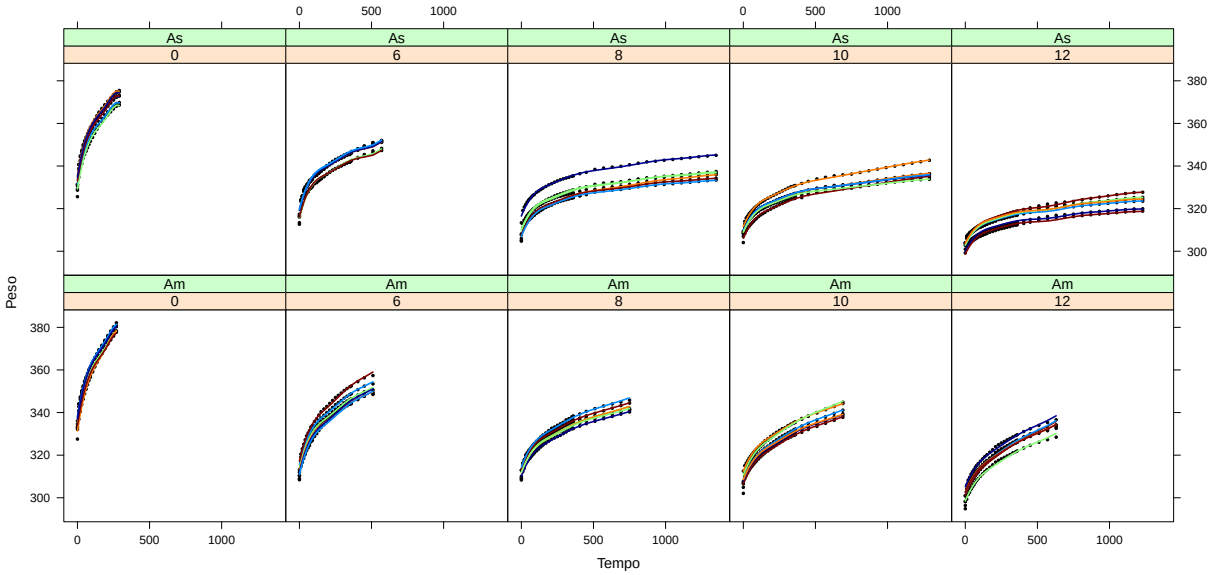

Figura 14 – Curvas individuais dos sujeitos ajustadas por teor de concentração

Ao observar a Figura 14 verifica-se que os corpos-de-prova com argamassa maturada adicionados os teores em boa parte do estudo admitem uma evolução superior quanto a velocidade de absorção de água por capilaridade, sendo observado que em todas as concentrações de teores, em média, os corpos-de-prova não chegaram até o tempo final do estudo. Em relação aos corpos-de-prova do tipo de argamassa seca em boa parte do estudo admitem uma evolução inferior quanto a velocidade de absorção de água por capilaridade, sendo observado que nas concentrações de teores mais altas, em média, os corpos-de-prova chegaram até o tempo final do estudo.

Capítulo 4

2º ENSAIO: ANÁLISE DA EVOLUÇÃO DO PESO DE CAMUNDONGOS INFECTADOS PELO *Trypanosoma Cruzi*

4.1 Descrição do Experimento

O conjunto de dados deste trabalho foi coletado por Ferreira et al. (2018), com o objetivo de avaliar o efeito do bioterápico de soro de galinha em roedores experimentalmente infectados pelo *Trypanosoma cruzi*, considerando parâmetros clínicos, imunológicos e parasitológicos. O projeto para a execução dos experimentos foi aprovado pelo Comitê de Ética em Pesquisa com Animais da Universidade Estadual de Maringá, Paraná, parecer CEUA 2401220716/2016.

Segundo Ferreira et al. (2018), os experimentos foram realizados por ensaios cegos, controlados e aleatorizados. Foram utilizados 57 camundongos suíços machos, com 56 dias de idade, provenientes do Biotério central da Universidade Estadual de Maringá. Os camundongos foram distribuídos em grupos de tratamentos de modo que a média inicial dos pesos dos animais em cada grupo fossem aproximadamente iguais. Os grupos experimentais (tratamentos) foram os seguintes:

- G1: (Controle não infectado). Animais não infectados e não tratados (n=5).
- G2: Animais tratados com bioterápico de soro de galinha na diluição 13cH[10] (n=13).

[10] Nomenclatura utilizada na homeopatia para denominar a escala de diluição dos medicamentos, em que são utilizadas as seguintes escalas: decimal (DH), centesimal (CH) e cinquenta milésimas (LM). Na escala centesimal, a proporção da diluição é de 1:100 (uma parte de insumo ativo para 100 partes de insumo inerte) , ou seja, uma parte de soluto para 99 partes de solvente.

- G3: Animais tratados com bioterápico de soro de galinha na diluição 6cH (n=13).
- G4: (Controle infecção). Os animais foram infectados e não receberam tratamento (n=13).
- G5: Animais tratados com bioterápico de soro de galinha na diluição 3cH (n=13).

A observação do peso corporal foi realizada durante 12 semanas, mensurado em balança semi analítica (Balance BEL®).

4.2 Resultados

A Tabela 8 apresenta as estatísticas resumo da variável resposta para cada um dos tipos de tratamentos, em que pode-se concluir que a média dos pesos em relação a cada um dos tratamentos variam de tal maneira que a média dos pesos dos indivíduos do grupo de tratamento (Controle não infectado) é de 45,49g, já a dos indivíduos que foram alocados no grupo de tratamento homeopático na diluição 3cH é de 41,44g.

Tabela 8 – Medidas resumo para cada um dos tratamentos

	Grupos de Tratamento	n	Média	Desvio Padrão	Min.	Max.	Coeficiente de Variação
1	Controle não infectado	5	45,49	3,48	39,20	53,40	0,076
2	Tratados (13cH)	13	44,34	4,71	31,60	54,40	0,106
3	Tratados (6cH)	13	42,09	2,83	35,40	49,20	0,067
4	Controle infecção	13	42,85	4,52	33,80	55,00	0,105
5	Tratados (3cH)	13	41,44	3,84	31,00	49,90	0,092

A Figura 15 apresenta o comportamento da variável resposta para cada um dos tratamentos, em que pode-se concluir que o grupo de tratamento homeopático na diluição de 3cH admitiu um nível médio dos pesos inferior aos demais grupos de tratamento, enquanto o grupo controle não infectado admite o maior nível médio dos pesos em relação aos demais grupos de tratamento.

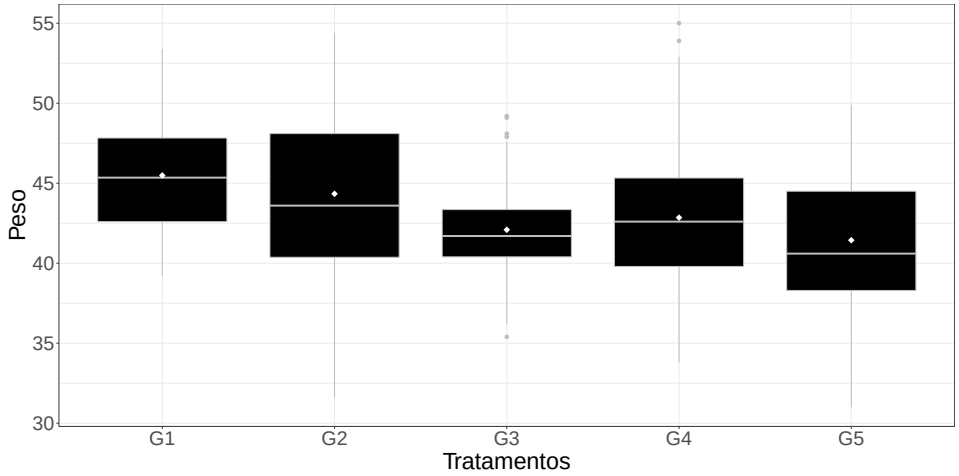

Figura 15 – Box-plot da variável resposta por tratamento

A Figura 16 apresenta o comportamento da variável resposta no tempo, considerando as 12 semanas do estudo, verificando-se assim um comportamento crescente do peso conforme varia o tempo, e mais, vê se uma oscilação deste a partir da sexta semana. Outro fato notório é que observa-se medidas candidatas a *outliers* na primeira semana e na quarta semana de acompanhamento.

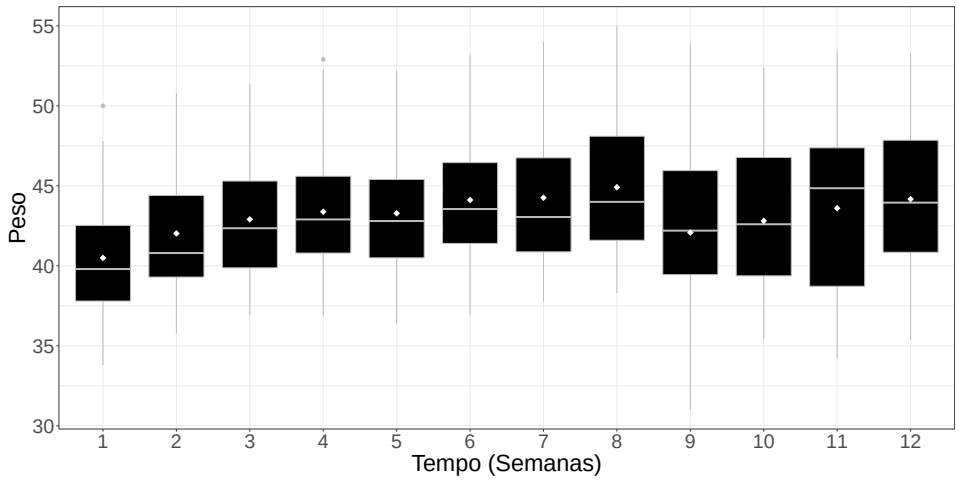

Figura 16 – *Box-plot* da variável resposta ao longo do tempo

Em relação ao gráfico de perfis, observa-se na Figura 17 um comportamento crescente a partir da segunda semana em relação ao peso conforme varia o tempo. Nota-se ainda um comportamento oscilatório no peso já na primeira semana, fator este justificado por ser o inicio do tratamento, e mais, observa-se um comportamento aparentemente linear da variável resposta da segunda semana à sexta semana para cada sujeito em relação aos grupos de tratamento conforme varia o tempo.

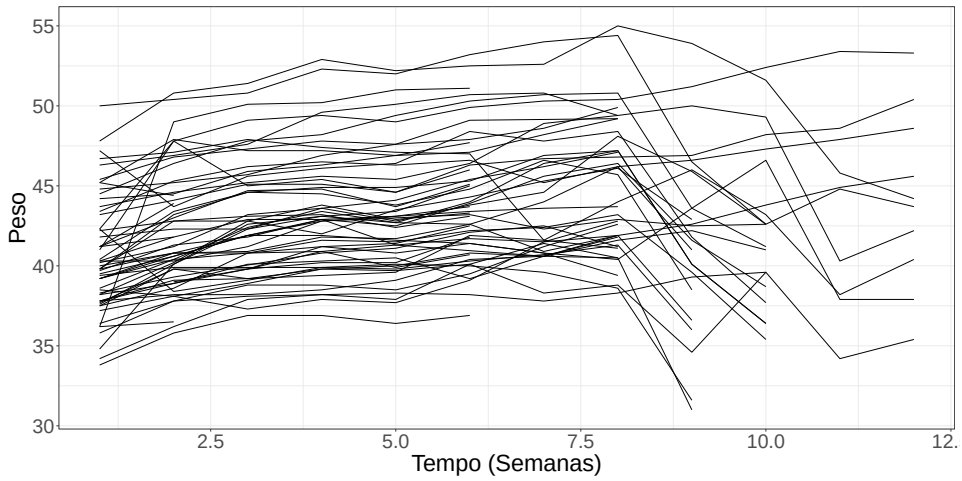

Figura 17 – Perfil médio para cada sujeito observado no tempo

A Figura 18, apresenta o gráfico de dispersão por tratamento, no qual existe a possibilidade de uma variabilidade entre essas observações, justificando-se o fato de poder utilizar algum tipo de suavização que leve em consideração a variabilidade constatada. Vê-se ainda que o

ganho de peso dos camundongos em cada grupo de tratamento apresenta muita variação entre o intercepto e inclinação das curvas que foram ajustadas, isso leva a considerar modelos mistos com efeitos aleatórios sejam eles independentes ou correlacionados entre intercepto e inclinação em relação ao tempo, neste caso para a covariável Tempo.

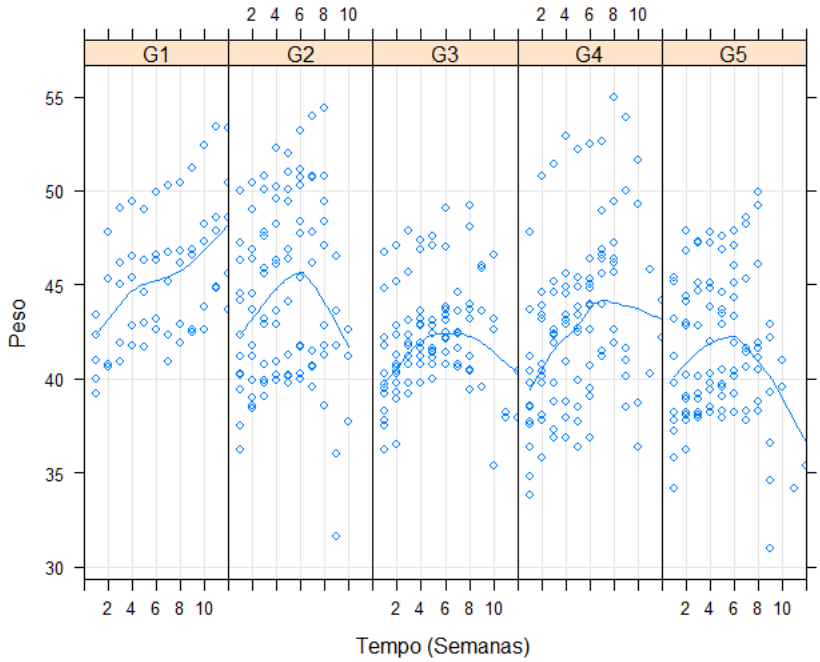

Figura 18 – Gráfico de dispersão por tratamentos

Considerando a função de suavização no tempo, o modelo considerado é descrito por:

$$\textit{Peso}_{ij} = \sum_{l=2}^{L} \gamma_l G_{il} + f_l(Tempo_{ij}) + a_{i1} + a_{i2} Tempo_{ij} + \epsilon_{ij}, \tag{4.1}$$

$$G_{il} = \begin{cases} 1, & \text{se o i-ésimo sujeito recebe o tratamento } l \\ 0, & \text{caso contrário} \end{cases}$$

em que, $Peso_{ij}$ denota o peso do i-*ésimo* sujeito com a j-*ésima* semana, para $i = 1, \cdots, 57$, $j = 1, \cdots, 12$. Os grupos homeopáticos de tratamento variam de $l = 1, \cdots, 5$. Tem-se ainda que f_1 denota a curva média para os indivíduos alocados no grupo (Controle não infectado), f_2 denota a curva média para os sujeitos que receberam o tratamento homeopático na diluição de 13cH, f_3 denota a curva média para os indivíduos que receberam o tratamento homeopático na diluição de 6cH, f_4 denota a curva média para os indivíduos alocados no grupo (Controle infecção) e f_5 denota a curva média para os indivíduos que receberam o tratamento homeopático

na diluição de 3cH, a_{i1} e a_{i2} denotam os efeitos aleatórios correspondentes aos interceptos e inclinações das curvas.

Tem-se ainda que $f_l(Tempo)$ é a função de suavização desconhecida, que descreve a trajetória dessa variável em relação a evolução do Peso, em que estima-se $f(.)$ por meio de P-*splines*, com função de base B-*splines* de terceiro grau polinomial, e uma penalização discreta de segunda ordem, considerando 10 nós. A variabilidade entre os sujeitos é determinada através do efeito aleatório a_i e um efeito aleatório na inclinação para a variável Tempo em que $(a_{i1}, a_{i2})' \sim N(0, \sum)$ e $\epsilon_{ij} \sim N(0, \sigma_\epsilon)$, assumindo a independência dos termos aleatórios.

Na Tabela 9, são apresentados os valores do AIC para 4 modelos ajustados, definidos por: um modelo sem considerar uma estrutura de correlação nos resíduos e outros três modelos utilizando a estrutura de correlação AR(1), CS e ARMA(1,1). Para selecionar o modelo utilizou-se o critério de Akaike, em que verificou-se que o modelo selecionado é aquele com a estrutura de correlação AR(1), uma vez que este proporciona o menor valor da estatística AIC.

Tabela 9 – Comparação de modelos com diferentes estruturas de correlação

Modelo	AIC
Não Estruturada	2666,09
Estrutura AR(1)	1999,07
Estrutura CS	2045,82
Estrutura ARMA(1,1)	2002,60

A Tabela 10 contém as estimativas do modelo referente aos coeficientes paramétricos e as estimativas dos termos suaves, em que verifica-se que os grupos de tratamento não foram significativos (valor-p $>$ 0,05) e que os termos suaves foram significativos (valor-p $<$ 0,001).

Tabela 10 – Estimativas do modelo misto aditivo generalizado utilizando termos suaves não-paramétricos.

Coeficientes Paramétricos:				
Efeito	Estimativa	Erro Padrão	Estatística de Teste	Valor-p
Intercepto	44,85	1,55	28,77	<0,001
Grupo 2 - 13cH	-1,83	1,84	-0,99	0,32
Grupo 3 - 6cH	-3,03	1,84	-1,66	0,09
Grupo 4 - Controle Infecção	-2,68	1,83	-1,45	0,14
Grupo 5 - 3cH	-2,74	1,84	-1,49	0,13
Significância dos termos suaves				
Efeito	gle	Ref.gl	Estatística de Teste	Valor-p
s(Tempo): Grupo 1 - Controle	3,75	3,75	7,76	<0,001
s(Tempo): Grupo 2 - 13cH	4,76	4,76	43,24	<0,001
s(Tempo): Grupo 3 - 6cH	2,72	2,72	15,68	<0,001
s(Tempo): Grupo 4 - Controle Infecção	6,86	6,86	41,41	<0,001
s(Tempo): Grupo 5 - 3cH	2,00	2,00	19,89	<0,001
R^2.(adj) = 0,72		Deviance explicada = 68,7%		
UBRE score = 1,70		Parâmetro de Escala est. = 1		n = 471

Em relação aos graus de liberdade efetivos (gle), estes indicam uma tendência não-linear, destacando-se o Grupo 4 - Controle Infecção (gle = 6,86).

A Figura 19 contém o gráfico da curva suavizada da covariável Tempo após a determinação da estrutura de correlação, verificando-se que a evolução do Peso se comportou de forma crescente até a 7º semana aproximadamente e a partir desse período apresentou um decrescimento até a 12º, concluindo que o efeito dessa covariável segue um padrão não linear.

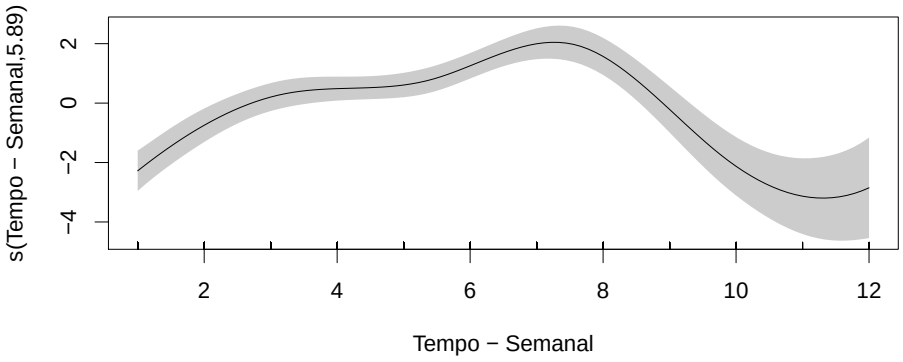

Figura 19 – Curva suave com 5,89 graus de liberdade efetivos (g.l.e) considerando a estrutura de correlação AR(1)

A Figura 20 apresenta o correlograma com a estrutura AR(1) nos resíduos, em que observa-se a redução da problemática encontrada, quanto a pressuposição de independência, nos erros do modelo sem considerar uma estrutura de correlação, visto que a maioria dos lags analisados, as funções da autocorrelação são menores em valores absoluto do que a linha tracejada, que corresponde ao limite de não rejeição de autocorrelação igual a 0 (RELVAS, 2013).

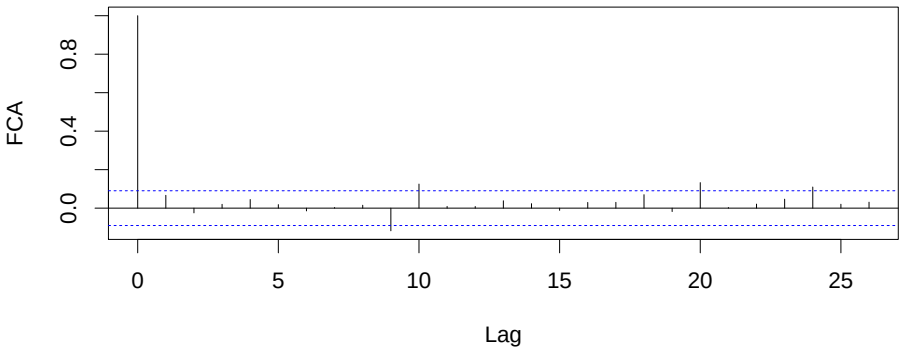

Figura 20 – Correlograma considerando a estrutura de correlação AR(1)

Com o intuito de verificar a hipótese de independência, realizamos o teste de Durbin-Watson,

em que a estatística de teste resultante foi de 1,69 e o (valor-p = 0,35), isto é, não há a presença de autocorrelação nos resíduos.

A Tabela 11 contém as estimativas do parâmetro λ, com seus respectivos graus de liberdade efetivos (g.l.e), teste F e o p-valor associado. Conclui-se ao observar a Tabela 11 que a suavização da a covariável Tempo para cada grupo homeopático de tratamento foi altamente significativa em relação a evolução do Peso.

Tabela 11 – Estimativas do parâmetro λ para cada tratamento.

Suavizador	g.l.e	Parâmetro (λ)	F	p-valor
s(tempo):Grupo 1 - Controle	3,751	$2{,}334\times10^{-}4$	7,764	<0,001
s(tempo):Grupo 2 - 13cH	4,763	$2{,}214\times10^{-}3$	43,248	<0,001
s(tempo):Grupo 3 - 6cH	2,726	$3{,}011\times10^{-}5$	15,681	<0,001
s(tempo):Grupo 4 - Controle Infecção	6,864	$4{,}662\times10^{-}2$	41,414	<0,001
s(tempo):Grupo 5 - 3cH	2,000	$7{,}176\times10^{-10}$	19,867	<0,001

A Figura 21 apresenta o comportamento da função não paramétrica de suavização estimada do modelo misto aditivo generalizado, no qual a linha contínua é a função de suavização estimada e a faixa cinza são as bandas com 95% de confiança, em que observa-se o efeito da covariável com comportamento não linear. Tem-se ainda que o eixo horizontal contém a variável Tempo e no eixo vertical a contribuição da função de suavização para os valores da variável resposta Peso. Assim, como o efeito da variável Tempo não é linear, a função de suavização estimada é significativa com 2,79 graus de liberdade efetivos.

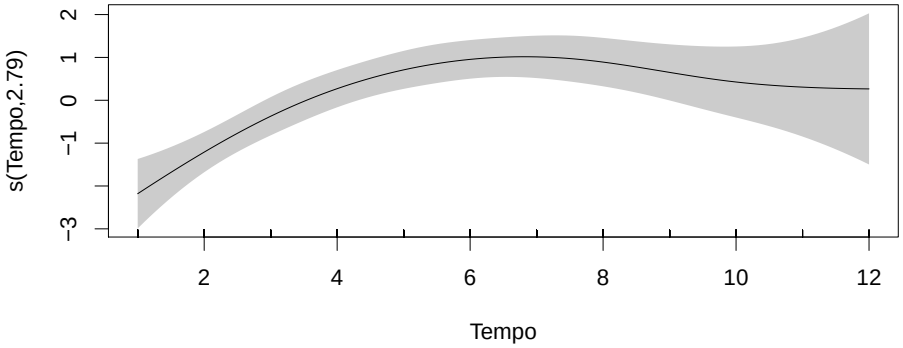

Figura 21 – Suavizador para o tempo semanal utilizando GAMM.

A Figura 22 apresenta o comportamento das funções não paramétricas de suavização estimadas do modelo misto aditivo generalizado para cada grupo de tratamento, no qual a linha contínua é a função de suavização estimada e a faixa cinza são as bandas com 95% de confiança.

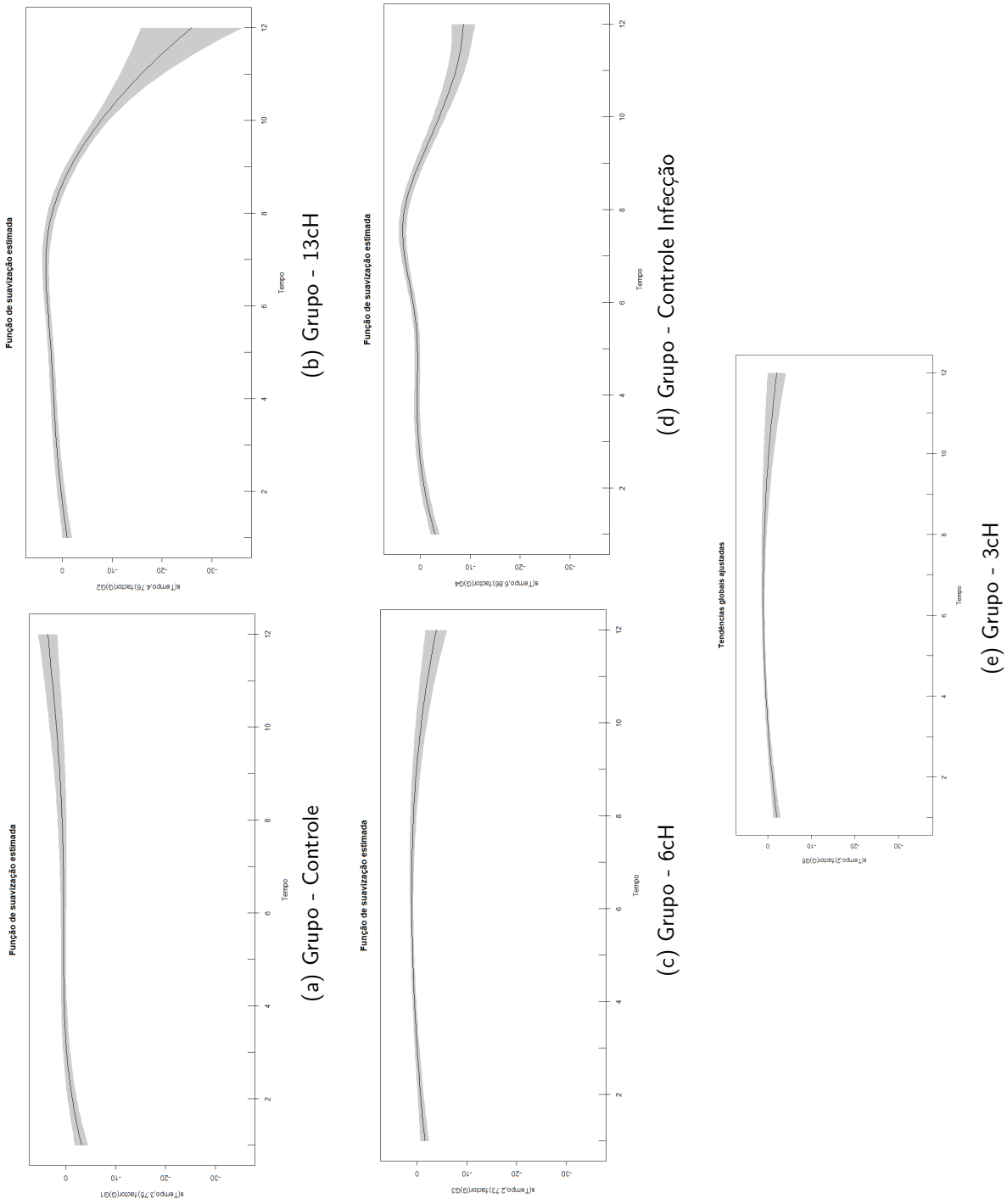

Figura 22 – Suavizador para o tempo semanal em cada grupo de tratamento utilizando GAMM.

Assim, verifica-se na Figura 22 que como o efeito da variável Tempo para cada grupo de tratamento não é linear, as funções de suavização estimadas são significativas com 3,751; 4,763; 2,726; 6,864 e 2,000 graus de liberdade efetivos. Vê-se ainda que a evolução do peso ao longo do tempo admite um período crescente até a 7º semana aproximadamente nos grupos de tratamento homeopático na diluição 13cH: Figura 22b; na diluição 6ch: Figura 22c e na diluição 3cH: Figura 22e) conforme o tempo aumenta, já para o grupo (Controle não infectado): Figura 22a e (Controle Infecção): Figura 22d) observa-se um leve crescimento no inicio do tratamento, passando por uma oscilação a partir da 2º semana. Vê-se ainda nas Figuras 22b e 22d um decrescimento notório, evidenciando assim o emagrecimento dos animais após o período citado.

A Figura 23 contém as curvas estimadas específicas aos sujeitos alocados em seus grupos de tratamento, em que o modelo 2.17 descreve estas levando em consideração a não-linearidade e os efeitos aleatórios associados. Observa-se ainda a capacidade de ajuste do modelo utilizado com 10 nós, em que este se apresenta de maneira satisfatória quanto a interpolação.

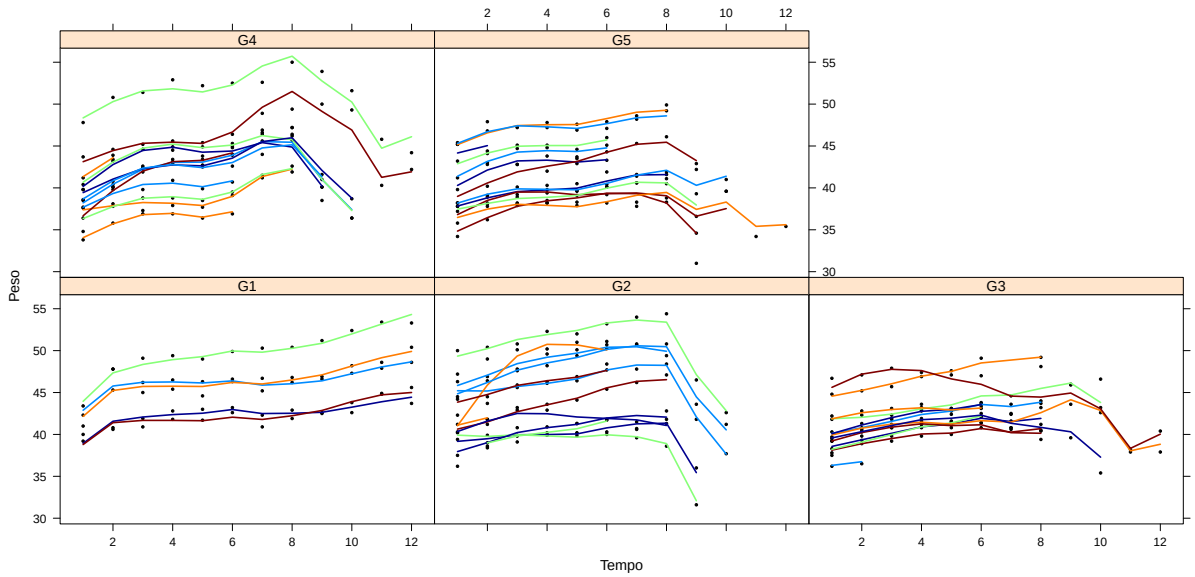

Figura 23 – Curvas individuais dos sujeitos ajustadas por tratamento

Assim, com o modelo definido em 2.17 pode-se descrever as trajetórias médias por grupos de tratamento em que a Figura 24 apresenta as curvas estimadas por grupos de tratamento, concluindo-se que os animais alocados no grupo controle em boa parte do estudo admitem um peso médio superior aos demais grupos de tratamento. Em relação ao grupo de tratamento homeopático na diluição de 13cH nota-se que nenhum dos animais chegaram até o final do estudo, em que neste grupo foi utilizado a maior diluição do soro bioterápico para o tratamento da doença de chagas. Por fim, observa-se ainda que em média os sujeitos alocados nos grupos de tratamento homeopático considerando as diluições de 13cH, 6cH, 3cH e Controle infecção admitem um decrescimento no peso a partir da sétima semana.

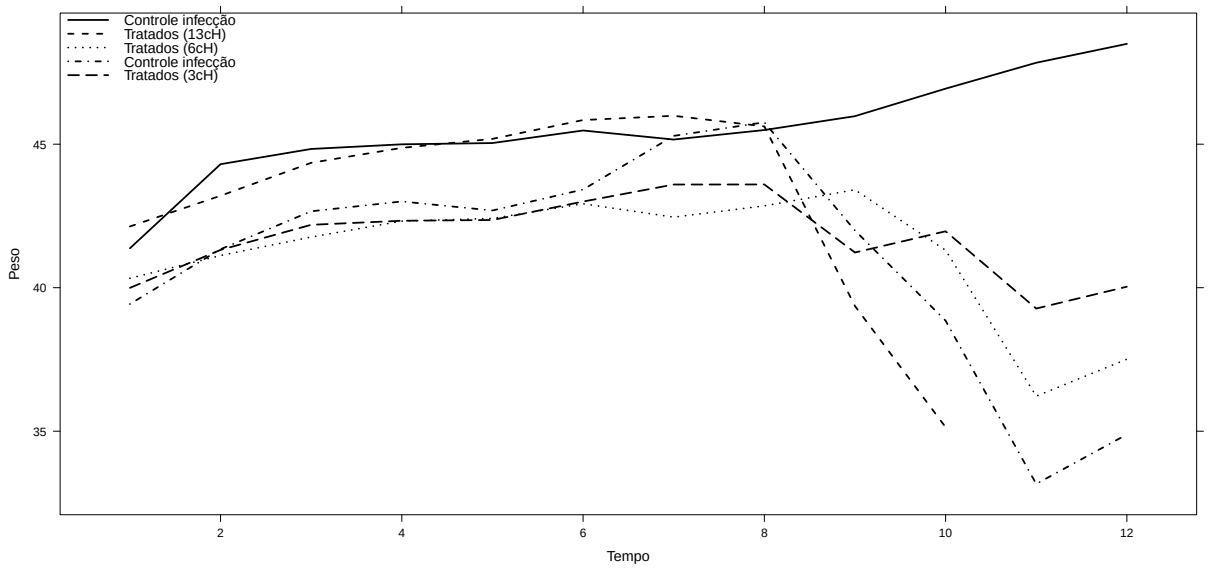

Figura 24 – Curvas estimadas por tratamentos

4.3 Passos Futuros

Este trabalho, em sua natureza, buscou não só expor a compreensão inicial a respeito do tema de um modelo P-*spline* reescrito como modelo misto, mas também levantar questões adicionais a serem estudadas posteriormente. Dado este caráter e a falta de estudos sobre o tema a nível nacional, acreditou-se que este trabalho é de suma importância. Sendo assim, em seguida são apresentadas considerações para elaboração final deste trabalho. Acredita-se que os seguintes pontos são interessantes de serem explorados:

- Realizar a modelagem conjunta de média e variância via pacote *gamlss*, estimando a função não-paramétrica por meio de uma P-*spline*;
- Aplicar os modelos do tipo P-*splines* como modelos mistos a um conjunto de dados reais e realizar o estudo dos resíduos do modelo;
- Verificar a capacidade preditiva das curvas estimadas.

Conclusão

Durante a construção deste trabalho, conclui-se que ao estudar as complexidades existentes ao trabalhar com modelos lineares mistos, estes podem não ser apropriados em algumas situações, por exemplo, quando as trajetórias individuais são uma função não linear do tempo, porém, os modelos P-*splines* reescritos como modelos mistos trazem maior flexibilidade em termos de modelagem, devido ao fato de que a introdução de efeitos aleatórios é levada em consideração para cada uma das curvas estimadas por tratamento por meio da função não paramétrica de suavização, concedendo então informações a respeito do comportamento médio de um determinado objeto classificado em grupos de observação.

Quanto as potencialidades observadas na utilização desta classe de modelos pode-se mencionar ainda a possibilidade de incluir estruturas complexas nos modelos usuais de suavização. No que se refere a dados correlacionados, se for realizado o ajuste de uma curva independentemente da correlação nos dados, vê-se que os métodos de seleção de parâmetros de suavização escolherão um valor mais baixo e, portanto, a curva não será suave.

Outro fato notório é que se o parâmetro de suavização for estimado, ou seja, os parâmetros que determinam a estrutura de correlação, a estimação seria de forma iterativa, o que tornaria o resultado final muito sensível à escolha dos parâmetros iniciais, porém, com os Modelos Mistos Aditivos Generalizados, ao introduzir uma estrutura de correlação e estimar simultaneamente a curva suave tem-se que estes se apresentam de maneira satisfatória.

Quanto aos resultados da análise estatística, observou-se na primeira aplicação que houveram diferenças significativas na trajetória da evolução do peso em ambos tipos de argamassas, concluindo que a argamassa do tipo maturada possui uma velocidade de absorção de água por capilaridade superior a argamassa do tipo seca, tem-se ainda que as suposições necessárias para a validade do modelo foram satisfeitas. Em relação a segunda aplicação, conclui-se que o peso, em média, até o período da oitava semana dos animais tratados com nível de diluição de (13cH) do soro bioterápico é superior aos demais grupos de tratamento. Já em relação aos animais tratados com nível de diluição de (3cH) do soro bioterápico, em média, estes chegaram até o final do estudo, tendo em média os pesos superiores aos animais tratados com nível de

diluição de (6cH) do soro bioterápico.

Em geral, a natureza destes dados correspondem a medidas mensuradas para vários indivíduos em momentos diferentes, e são caracterizadas pela dependência entre medidas repetidas feitas para o mesmo indivíduo. No que se refere as P-*splines*, quanto ao uso de bases B-*splines*, essa trás potencialidades no que se refere ao ganho computacional. No entanto, Durbán (2018), Eilers e Marx (1996) & Wood (2017) apresentam outras bases que podem ser utilizadas, como por exemplo, *thin plate splines*, porém exigem um alto custo computacional.

Referências

ALMEIDA, E. P. d. Um estudo longitudinal de dados de huanglongbing em pomares. 2015. Citado na página 31.

AZEVEDO, A. M.; JÚNIOR, V. C. A.; FERNANDES, J. S. Transformação box-cox na homocedasticidade e normalidade uni e multivariada em experimentos de batata-doce. *Horticultura Brasileira*, SciELO Brasil, v. 34, n. 1, p. 93–101, 2016. Citado na página 17.

BOOR, C. D.; BOOR, C. D.; MATHÉMATICIEN, E.-U.; BOOR, C. D.; BOOR, C. D. *A practical guide to splines*. [S.l.: s.n.], 1978. v. 27. ISBN 978-0-387-95366-3. Citado 2 vezes nas páginas 17 e 27.

BOX, G.; JENKINS, G.; REINSEL, G. *Time series analysis, forecasting and control. Englewood Clifs*. [S.l.]: NJ: Prentice-Hall, 1994. ISBN 978-1-118-67502-1. Citado na página 40.

BOZDOGAN, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, Springer, v. 52, n. 3, p. 345–370, 1987. doi: <10.1007/BF02294361>. Citado na página 39.

CANOVA, J. Influência da adição de pó de borracha em argamassa de revestimento. *PhD Tesis, State University of Maringá, Maringá, Brazil*, 2008. Citado 2 vezes nas páginas 49 e 50.

CHATTERJEE, S.; HANDCOCK, M. S.; SIMONOFF, J. S. *A casebook for a first course in statistics and data analysis*. [S.l.: s.n.], 1995. ISBN 978-0471110309. Citado na página 45.

COOK, R. D. Assessment of local influence. 1986. Citado na página 21.

CORBEIL, R. R.; SEARLE, S. R. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, Taylor & Francis, v. 18, n. 1, p. 31–38, 1976. doi: <10.1080/00401706.1976.10489397>. Citado na página 36.

CRAVEN, P.; WAHBA, G. Smoothing noisy data with spline functions. *Numerische mathematik*, Springer, v. 31, n. 4, p. 377–403, 1978. doi: <10.1007/bf01404567>. Citado na página 38.

CURRIE, I. D.; DURBAN, M. Flexible smoothing with p-splines: a unified approach. *Statistical Modelling*, Sage Publications Sage CA: Thousand Oaks, CA, v. 2, n. 4, p. 333–349, 2002. doi: <10.1191/1471082x02st039ob>. Citado 4 vezes nas páginas 18, 19, 30 e 46.

DAVIDIAN, M. *Nonlinear models for repeated measurement data*. [S.l.]: Routledge, 2017. ISBN 9780412983412. Citado 2 vezes nas páginas 42 e 44.

DIGGLE, P. J. An approach to the analysis of repeated measurements. *Biometrics*, JSTOR, p. 959–971, 1988. doi: <10.2307/2531727>. Citado na página 44.

DURBAN, M. *Splines con penalizaciones: Teoría y aplicaciones*. [S.l.]: Universidad Pública de Navarra, 2007. Citado na página 37.

DURBÁN, M. An introduction to smoothing with penalties: P-splines. *Boletín de Estadística e Investigación Operativa*, v. 25, n. 3, p. 195–205, 2009. Citado 2 vezes nas páginas 25 e 53.

DURBÁN, M. Generalized additive models with p-splines. *II Workshop em Bioestatística*, p. 1–90, 2018. Citado 3 vezes nas páginas 37, 46 e 70.

DURBÁN, M.; HAREZLAK, J.; WAND, M.; CARROLL, R. Simple fitting of subject-specific curves for longitudinal data. *Statistics in medicine*, Wiley Online Library, v. 24, n. 8, p. 1153–1167, 2005. doi: <10.1002/sim.1991>. Citado 2 vezes nas páginas 19 e 53.

EILERS, P. H.; MARX, B. D. Flexible smoothing with b-splines and penalties. *Statistical science*, JSTOR, p. 89–102, 1996. doi: <10.1214/ss/1038425655>. Citado 5 vezes nas páginas 17, 18, 28, 30 e 70.

FERREIRA, É. C.; CIUPA, L.; PORTOCARRERO, A. R.; BRUSTOLIN, C. F.; MASSINI, P. F.; ALEIXO, D. L.; ARAÚJO, S. M. de. Phosphorus protects cardiac tissue by modifying the immune response in rats infected by trypanosoma cruzi. *Cytokine*, Elsevier, v. 102, p. 102–106, 2018. doi: <10.1016/j.cyto.2017.07.017>. Citado na página 59.

FLORENCIO, L. de A. *Engenharia de avaliações com base em modelos GAMLSS*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2010. Citado na página 20.

FOX, J.; MONETTE, G. *An R and S-Plus companion to applied regression*. [S.l.]: Sage, 2002. Citado na página 17.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001. v. 1. ISBN 978-0387848570. Citado na página 38.

GARCIA, T. P.; MARDER, K. Statistical approaches to longitudinal data analysis in neurodegenerative diseases: huntington's disease as a model. *Current neurology and neuroscience reports*, Springer, v. 17, n. 2, p. 14, 2017. doi: <10.1007/s11910-017-0723-4>. Citado na página 31.

GOUVEIA, J. F. Modelos volumétricos mistos aplicados em clones de eucalyptus no polo gesseiro do araripe-pe. 2013. doi: <10.5380/rf.v45i3.36844>. Citado na página 43.

HARRELL, J.; FRANK, E. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. [S.l.]: Springer, 2015. ISBN 978-1-4757-3462-1. Citado 2 vezes nas páginas 25 e 30.

HENDERSON, D. J.; PARMETER, C. F. *Applied nonparametric econometrics*. [S.l.: s.n.], 2015. ISBN 978-1-107-01025-3. Citado na página 17.

HOLANDA, A. A. *Modelos lineares parciais aditivos generalizados com suavização por meio de P-splines*. Tese (Doutorado) — Universidade de São Paulo, 2018. Citado 3 vezes nas páginas 22, 27 e 28.

JONES, R. H. *Longitudinal data with serial correlation: a state-space approach*. [S.l.]: Chapman and Hall/CRC, 1993. ISBN 9780367450083. Citado 5 vezes nas páginas 33, 34, 35, 39 e 41.

KEELE, L. J. *Semiparametric regression statistic for the science*. [S.l.]: John Wiley and Sons, 2008. ISBN 978-0-470-99812-0. Citado na página 17.

LAIRD, N. M.; WARE, J. H. et al. Random-effects models for longitudinal data. *Biometrics*, v. 38, n. 4, p. 963–974, 1982. doi: <10.1007/978-3-642-11760-2_2>. Citado 3 vezes nas páginas 32, 34 e 53.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, Oxford University Press, v. 73, n. 1, p. 13–22, 1986. doi: <10.1177/1094428104263672>. Citado na página 33.

LIMA, C. G. d.; DEMÉTRIO, C. G. B. Análise de dados longitudinais provenientes de experimentos em blocos casualizados. 1996. Citado na página 44.

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D. *SAS system for mixed models*. [S.l.]: SAS institute Cary, NC, 1996. ISBN 1-55544-779-1. Citado 3 vezes nas páginas 39, 43 e 44.

MA, S.; RACINE, J. S. Additive regression splines with irrelevant categorical and continuous regressors. *Statistica Sinica*, JSTOR, p. 515–541, 2013. doi: <10.5705/ss.2011.096>. Citado na página 24.

MACHADO, R. J. M. et al. Modelos mistos semiparamétricos parcialmente não lineares. Universidade Federal de São Carlos, 2014. Citado na página 21.

MAGNANINI, M. M. F. et al. *Modelagem de dados longitudinais aplicada a uma coorte de pacientes hipertensos resistentes*. Tese (Doutorado), 2010. Citado na página 45.

MAZZINI, A. R. d. A. Construção de curva de peso gestacional em uma coorte de gestantes brasileiras eutróficas usando modelos aditivos generalizados de localização, escala e forma. 2015. Citado na página 21.

MENDOZA, N. V. R. *Estruturas unidimensionais e bidimensionais utilizando P-splines nos modelos mistos aditivos generalizados com aplicação na produção de cana-de-açúcar*. Tese (Doutorado) — Universidade de São Paulo, 2017. Citado na página 22.

MONETTE, G. et al. *Mixed Models with R: Longitudinal Data Analysis with Mixed Models*. [S.l.]: USP/ESALQ, 2010. Citado 3 vezes nas páginas 32, 34 e 39.

PINHEIRO, J.; BATES, D. *Mixed-effects models in S and S-PLUS*. [S.l.]: Springer Science & Business Media, 2006. ISBN 978-0-387-22747-4. Citado 2 vezes nas páginas 34 e 39.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D.; R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. [S.l.], 2019. R package version 3.1-139. Disponível em: <https://CRAN.R-project.org/package=nlme>. Citado na página 48.

POLANSKY, L.; ROBBINS, M. M. Generalized additive mixed models for disentangling long-term trends, local anomalies, and seasonality in fruit tree phenology. *Ecology and Evolution*, Wiley Online Library, v. 3, n. 9, p. 3141–3151, 2013. Citado na página 20.

PULGAR, G. M. I. *Modelos mistos aditivos semiparamétricos de contornos elípticos*. Tese (Doutorado) — Universidade de São Paulo, 2009. Citado na página 19.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>. Citado na página 18.

RELVAS, C. E. M. *Modelos parcialmente lineares com erros simétricos autoregressivos de primeira ordem*. Tese (Doutorado) — Universidade de São Paulo, 2013. Citado 2 vezes nas páginas 21 e 64.

ROBINSON, G. K. et al. That blup is a good thing: the estimation of random effects. *Statistical science*, Institute of Mathematical Statistics, v. 6, n. 1, p. 15–32, 1991. doi: <10.1214/ss/1177011930>. Citado na página 35.

RUPPERT, D.; WAND, M. P.; CARROLL, R. J. *Semiparametric regression*. [S.l.]: Cambridge university press, 2003. doi: <10.1214/09-EJS525>. Citado 2 vezes nas páginas 37 e 44.

SARKAR, D. *Lattice: Multivariate Data Visualization with R*. New York: Springer, 2008. ISBN 978-0-387-75968-5. Disponível em: <http://lmdvr.r-forge.r-project.org>. Citado na página 48.

SARKAR, D.; ANDREWS, F. *latticeExtra: Extra Graphical Utilities Based on Lattice*. [S.l.], 2016. R package version 0.6-28. Disponível em: <https://CRAN.R-project.org/package=latticeExtra>. Citado na página 48.

SCHAEFFER, L. Application of random regression models in animal breeding. *Livestock Production Science*, Elsevier, v. 86, n. 1-3, p. 35–45, 2004. doi: <10.1016/S0301-6226(03)00151-9>. Citado na página 36.

SCHALL, R. Estimation in generalized linear models with random effects. *Biometrika*, Oxford University Press, v. 78, n. 4, p. 719–727, 1991. doi: <10.1093/biomet/78.4.719>. Citado na página 36.

SHADISH, W. R.; ZUUR, A. F.; SULLIVAN, K. J. Using generalized additive (mixed) models to analyze single case designs. *Journal of school psychology*, Elsevier, v. 52, n. 2, p. 149–178, 2014. Citado na página 21.

SOARES, A. R. *Regressão Spline*. [S.l.]: LCE - Laboratório de Ciências Exatas, ESALQ - USP, 2016. Citado na página 27.

SOARES, A. R. *Regressão spline de nós livres para modelagem de curvas de crescimento multifásica*. Tese (Doutorado) — Universidade de São Paulo, 2019. Citado 2 vezes nas páginas 22 e 25.

SOUZA, E. L. d. et al. Modelos aditivos generalizados para a avaliação da intenção de compra de consumidores. Universidade Federal da Paraíba, 2012. Citado na página 20.

STASINOPOULOS, D. M.; RIGBY, R. A. et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007. doi: <10.18637/jss.v023.i07>. Citado 2 vezes nas páginas 20 e 21.

STONE, C. J. et al. Additive regression and other nonparametric models. *The annals of Statistics*, Institute of Mathematical Statistics, v. 13, n. 2, p. 689–705, 1985. Citado na página 25.

VERBEKE, G.; MOLENBERGHS, G. *Linear mixed models for longitudinal data*. [S.l.]: Springer Science & Business Media, 2009. ISBN 978-0-387-22775-7. Citado na página 34.

VIEIRA, M. de T. Desafios na análise de dados longitudinais amostrais complexos. 2009. Citado na página 31.

VONESH, E.; CHINCHILLI, V. M. *Linear and nonlinear models for the analysis of repeated measurements*. [S.l.]: CRC press, 1996. ISBN 9780824782481. Citado na página 44.

WALD, V. *A metodologia de modelos mistos não lineares aplicados à análise de dados longitudinais em plantas forrageiras*. Tese (Doutorado) — Dissertation (MS in Zootechnique), Federal University of Rio Grande do Sul , 2000. Citado na página 44.

WAND, M.; ORMEROD, J. On semiparametric regression with o'sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, Wiley Online Library, v. 50, n. 2, p. 179–198, 2008. Citado na página 19.

WEGMAN, E. J.; WRIGHT, I. W. Splines in statistics. *Journal of the American Statistical Association*, Taylor & Francis, v. 78, n. 382, p. 351–365, 1983. doi: <10.1080/01621459.1983.10477977>. Citado 2 vezes nas páginas 24 e 26.

WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>. Citado na página 48.

WOLFINGER, R. Covariance structure selection in general mixed models. *Communications in statistics-Simulation and computation*, Taylor & Francis, v. 22, n. 4, p. 1079–1106, 1993. doi: <10.1080/03610919308813143>. Citado na página 39.

WOOD, S. N. *Generalized additive models: an introduction with R*. [S.l.]: Chapman and Hall/CRC, 2017. ISBN 9781498728331. Citado 3 vezes nas páginas 46, 47 e 70.

ZIVOT, E.; WANG, J. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-Plus®*, Springer, p. 385–429, 2006. doi: <10.1007/978-0-387-21763-5_11>. Citado na página 41.