UNIVERSIDADE ESTADUAL DE LONDRINA

FELINTO JUNIOR DA COSTA

GAMLSS ESPAÇOTEMPORAL PARA ENGENHARIA DE AVALIAÇÕES

Londrina
2022

FELINTO JUNIOR DA COSTA

GAMLSS ESPAÇOTEMPORAL PARA ENGENHARIA DE AVALIAÇÕES

Dissertação apresentada ao Departamento de Matemática da Universidade Estadual de Londrina, como requisito parcial para a obtenção do Título de MESTRE em Matemática Aplicada e Computacional.

Orientador: Prof. Dr. Rodrigo Rossetto Pescim

Londrina
2022

Ficha de identificação da obra elaborada pelo autor, por meio do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Dados Internacionais de Catalogação na Publicação (CIP)

C837 Costa, Felinto Junior da.
GAMLSS Espaçotemporal para Engenharia de Avaliações / Felinto Junior da Costa. – Londrina, 2022.
146 f. : il.

Orientador: Rodrigo Rossetto Pescim.
Dissertação (Mestrado em Matemática Aplicada e Computacional) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Matemática Aplicada e Computacional, 2022.

Inclui Bibliografia.

1. GAMLSS - Tese. 2. Estatística espaçotemporal - Tese. 3. Engenharia de avaliações - Tese. I. Pescim, Rodrigo Rossetto. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Matemática Aplicada e Computacional. III. Título.

CDU 31

FELINTO JUNIOR DA COSTA

GAMLSS ESPAÇOTEMPORAL PARA ENGENHARIA DE AVALIAÇÕES

Dissertação apresentada ao Departamento de Matemática da Universidade Estadual de Londrina, como requisito parcial para a obtenção do Título de MESTRE em Matemática Aplicada e Computacional.

BANCA EXAMINADORA

Prof. Dr. Rodrigo Rossetto Pescim -
Universidade Estadual de Londrina

Prof. Dr. Guilherme Biz
Universidade Estadual de Londrina

Prof. Dr. Luiz Ricardo Nakamura
Universidade Federal de Santa Catarina

Londrina, 7 de abril de 2022.

AGRADECIMENTOS

Pitágoras usou a palavra cosmos para expressar a harmonia e a ordem do Universo. A esse Grande Geômetra, responsável por tais ordem e harmonia, agradecemos a vida.

Ao Prof. Dr. Rodrigo Rosetto Pescim pela orientação, amizade, paciência e o crédito em meu trabalho.

Aos Profs. Drs. Guilherme Biz e Luiz Ricardo Nakamura pela coorientação e valiosas sugestões.

Aos professores do Departamento de Estatística da Universidade Estadual de Londrina, em especial ao Prof. Dr. Tiago Viana Flor de Santana às Prof.as Dr.as Mariana Ragassi Urbano pelo estímulo e amizade.

Às Prof.as Dr.as Ana Vergínia Libos Massetti, Elizabeth Strapasson e Vanderli Marino Melem pelo incentivo quando ainda no curso de Especialização em Estatística em 2014.

Àqueles que dedicaram incontáveis horas no desenvolvimento de bibliotecas suplementares escritas na linguagem de programação aqui utilizada, disponibilizando-as de modo altruísta.

A todos aqueles que de forma direta ou indireta contribuíram para a realização deste desafio.

Aos meus pais e, em especial, ao meu filho Victor.

COSTA, Felinto Junior da. **GAMLSS Espaçotemporal para Engenharia de Avaliações:**. 2022. 146 f. Dissertação (Mestrado em Matemática Aplicada e Computacional) – Universidade Estadual de Londrina, Londrina, 2022.

RESUMO

A engenharia de avaliações fornece subsídios de suma importância não apenas para órgãos públicos e o sistema legal, mas também para diversas atividades econômicas privadas, lastreando a fundamentação de decisões judiciais, garantindo um procedimento equânime nas decisões administrativas e assegurando operações financeiras baseadas em garantias reais. Todavia, a natureza heterogênea dos bens imobiliários impõe grande complexidade quando da formulação de modelos estatísticos que procurem estimar seu valor, uma consequência de três aspectos indissociáveis que introduzem grande variabilidade: endógenos (características da propriedade relacionadas a tamanhos e padrões), exógenos (a vizinha da propriedade, a presença de amenidades e serviços públicos) e temporal (o tempo de referência). Os Modelos Hedônicos de Regressão tradicionalmente adotados pela engenharia de avaliações contemplam esses três aspectos pela inclusão de um conjunto de variáveis explicativas associadas às características intrínsecas e extrínsecas mais significativas e o tempo de referência como mais uma variável, por vezes um fator assumindo tantos níveis quanto períodos temporais existirem na amostra. Entretanto, esses modelos consideram os aspectos espacial e temporal dissociadamente, contrariando a situação real observada, em que diferentes regiões de uma cidade se valorizam (ou desvalorizam) de modos distintos ao longo do tempo, não sendo assim possível admitir-se como válida, uma variabilidade temporal espacialmente homogênea. Este trabalho propõe um Modelo Hedônico de Regressão Espaçotemporal baseado na classe dos Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS: Generalized Additive Models for Location, Scale and Shape). Ele considera a variabilidade espacial e temporal de modo conjunto nos preditores dos parâmetros da distribuição teórica adotada para a resposta usando splines de produto tensor. Eles são estimados a partir das bases de duas funções suavizadoras distintas. A primeira delas modela a variabilidade puramente espacial usando thin plate splines sobre as coordenadas métricas da localização de cada elemento da amostra. A segunda modela apenas a variabilidade temporal usando splines cúbicos sobre as datas dos elementos. O modelo foi ajustado a um conjunto de dados reais composto por informações imobiliárias sobre terrenos sem benfeitorias localizados no perímetro urbano da cidade de Londrina (norte do Estado do Paraná) coletada no período de maio de 1995 a março de 2021. O modelo é capaz de estimar os valores unitários medianos em distintas localizações espaciais e referências temporais. Isso permite a geração de superfícies de valor que ilustram a variabilidade do valor na área do estudo em qualquer data. Por fim, possibilita também a utilização dessa informação como variável adicional em modelos relativos a outras tipologias de imóveis para, nesses modelos, considerar-se também o aspecto espaçotemporal de forma conjunta.

Palavras-chave: Engenharia de avaliações. Modelos *GAMLSS*. Modelo espaçotemporal. Produto tensor de bases. *Thin plate splines* bivariado. *Splines* cúbicos.

COSTA, Felinto Junior da. **Spatiotemporal GAMLSS for Real Estate Appraisal Engineering:**. 2022. 146 f. Dissertation (Master's in Applied and Computational Mathematics) – State University of Londrina, Londrina, 2022.

ABSTRACT

Appraisal engineering provides extremely important subsidies not only for public administration and the legal system, but also for several private economic activities, supporting court decisions, guaranteeing an equitable procedure in administrative decisions and ensuring mortgage-backed financial operations. However, the heterogeneous nature of real estate properties imposes great complexity when formulating statistical models that aim at estimating the properties' values as a consequence of three inseparable aspects that introduce great variability: endogenous (characteristics of the property related to sizes and building standards), exogenous (the vicinity of the property, the presence of amenities and public services) and temporal (their reference time). The Hedonic Regression Models traditionally adopted by real estate appraisal contemplate these three aspects by including a set of explanatory variables associated with the most significant intrinsec and extrinsec characteristics of the property and the reference time as another variable, sometimes a factor assuming as many levels as time periods exist in the sample (temporal). Nevertheless, these models consider spatial and temporal aspects in a dissociated way, contrary to the real situation observed, where different regions of a city value (or devalue) in different scales over time. Thus, it is not plausible to admit as valid a spatially homogeneous temporal variability. This work proposes a Hedonic Spatio-Temporal Regression Model based on the Generalized Additive Models for Location, Scale and Shape class (GAMLSS). It considers the spatial and temporal variability jointly into the predictors of the parameters of the theoretical probability distribution adopted for the response variable using tensor product splines. They are estimated combining the basis of two distinct smoothing functions. The first models only the spatial variability using thin plate splines basis at the metric coordinates of the location of each element in the sample. While the second models only the temporal variability using a cubic spline basis at the dates of the elements. The model was trained using a real dataset consisting of real estate information on land without improvements located in the urban perimeter of the city of Londrina (North of Paraná, Brazil) collected during the period of May 1995 to March 2021. The model is capable of predicting the median values in different spatial locations and temporal references. This allows the creation of value surfaces that illustrate the variability of the property's value in the study area at any date. Lastly, it also allows the use of this information as an additional variable in other models related to different real estate typologies as a way of considering the space-time aspect jointly.

Keywords: Real Estate Appraisal Engineering. GAMLSS models. Spatiotemporal model. Tensor product of basis. Bivariate thin-plate splines. Cubic splines.

LISTA DE FIGURAS

1.1 Localização, variável determinante e comum a todas as tipologias de bens imobiliários 24

2.1 Ilustração histórica do modelo de círculos concêntricos definindo áreas com diferentes valores do solo em relação ao núcleo urbano 28

3.1 Modelo econométrico da função consumo de Keynes 35
3.2 Um conjunto de *spline* e *ducks* 41
3.3 Ajuste de *splines* cúbicos (nós arbitrados em: 5, 10, 20, 30 e 40) 42
3.4 Base *B-spline* de ordem $m = 4$ 46
3.5 Efeito do parâmetro de suavização λ: equilíbrio entre *wiggliness* e *smoothness* da função de suavização estimada 48
3.6 Variações na curva da densidade da distribuição Box&Cox "t" (BCT) 67
3.7 Funções disponíveis na implementação computacional para a incorporação dos termos nos preditores dos parâmetros 70
3.8 Esquema geral dos ciclos dos algoritmos de ajuste de um *GAMLSS* 72
3.9 Esquema geral de um processo de *bootstrap* 75
3.10 Algoritmo de um processo de *bootstrap* para estimação do desvio padrão $\hat{\theta} = s(\boldsymbol{x})$ 76
3.11 Relação entre a estrutura real dos dados e a estrutura do *bootstrap* 77
3.12 Padrões sistemáticos de afastamento dos resíduos quantílicos da linha horizontal de referência em um *worm plot* 82

4.1 Localização geográfica de Londrina 87
4.2 Mapa municipal de Londrina com a localização de cada elemento amostral em Londrina (coordenadas 23,3197S;51,1662W) 89
4.3 Fluxograma da análise 93
4.4 Distribuição marginal da resposta sob a distribuição teórica Box&Cox "t" (BCTo) 96
4.5 Distribuição marginal da resposta sob a distribuição Beta generalizada tipo 2 (GB2) 96

5.1 Distribuição marginal da resposta sob a distribuição Normal (NO) 102
5.2 Dispersão da variável resposta 103
5.3 Dispersão da variável resposta 103

5.4 Frequências observadas na amostra dos níveis dos fatores “NATUREZA” e “IMPLANTACAO” 104
5.5 Frequências observadas na amostra dos níveis dos faores “PAVIMENTACAO” e “RELEVO” 104
5.6 *Box-plot* da variável resposta em relação aos níveis dos fatores “NATUREZA” e “IMPLANTACAO” 105
5.7 *Box-plot* da variável resposta em relação aos níveis dos fatores “PAVIMENTACAO” e “RELEVO” 105
5.8 Efeitos dos fatores “IMPLANTACAO” e “NATUREZA” sobre a mediana distribucional (incorporados parametricamente ao seu preditor linear) 107
5.9 Efeitos dos fatores “PAVIMENTACAO” e “RELEVO” sobre a mediana distribucional (incorporados parametricamente ao seu preditor linear) 108
5.10 Efeitos das variáveis “AT” e “DATA” sobre a mediana distribucional (incorporadas ao seu preditor linear sob a forma de funções suavizadoras unidimensionais) 109
5.11 Efeito conjunto das variáveis “UTM_X” e “UTM_Y” na mediana distribucional (incorporadas ao seu preditor linear sob a forma de uma função suavizadora bidimensional): efeito puramente espacial da localização 110
5.12 Efeito conjunto das variáveis “UTM_X”, “UTM_Y” e “DATA” na mediana distribucional (incorporadas ao seu preditor linear sob a forma de uma função suavizadora multidimensional): efeito espaçotemporal das coordenadas mostradas (UTM) e a data da informação, edf=50,95 111
5.13 Efeitos das variáveis “AT” e “DATA” sobre o coeficiente de variação da distribuição (incorporadas ao seu preditor linear sob a forma de funções suavizadoras unidimensionais) 112
5.14 Efeito conjunto das variáveis “UTM_X” e “UTM_Y” no coeficiente de variação da distribuição (incorporadas ao seu preditor linear sob a forma de uma função suavizadora bidimensional): efeito puramente espacial da localização 112
5.15 Efeitos das variáveis “AT” e “DATA” sobre o coeficiente *proxy* da assimetria da distribuição (incorporadas ao seu preditor linear sob a forma de funções suavizadoras unidimensionais) 113
5.16 Efeito conjunto das variáveis “UTM_X” e “UTM_Y” no coeficiente *proxy* da assimetria da distribuição (incorporadas ao seu preditor linear sob a forma de uma função suavizadora bidimensional): efeito puramente espacial da localização 114
5.17 Efeito da variável “AT” sobre o coeficiente *proxy* da curtose da distribuição (incorporada ao seu preditor linear sob a forma de uma função suavizadora unidimensional) edf=3,92 114
5.18 *Worm plot* do modelo proposto 115
5.19 Gráficos de probabilidade Normal dos resíduos quantílicos 118

5.20 Representação gráfica dos valores da estatística Z em cada faixa definida para as variáveis “AT” e “DATA” 119
5.21 Representação gráfica dos valores da estatística Z em cada faixa definida para as variáveis “UTM_X” e “UTM_Y” 120
5.22 Histogramas das estimativas *bootstrap* do intercepto do modelo proposto 122
5.23 Histogramas das estimativas *bootstrap* do coeficiente do fator: “NATUREZA” (se transação) 123
5.24 Histogramas das estimativas *bootstrap* do coeficiente do fator: “IMPLANTACAO” (se condomínio) 124
5.25 Histogramas das estimativas *bootstrap* do coeficiente do fator: “RELEVO” (se aclive) 125
5.26 Histogramas das estimativas *bootstrap* do coeficiente do fator: “RELEVO” (se declive) 126
5.27 Histogramas das estimativas *bootstrap* do coeficiente do fator: “PAVIMENTACAO” (se asfáltica) 127
5.28 Variograma espaçotemporal dos resíduos simples do modelo proposto 129
5.29 Variograma espaçotemporal dos resíduos simples do modelo proposto 129
5.30 Valores estimados *versus* observados 130
5.31 Alterações do valor unitário mediano na área delimitada pelo perímetro urbano de Londrina com recortes temporais definidos para jun. 2000 e jun. 2021 (sob a situação definida como paradigma) 131
5.32 Superfície de valores unitários medianos estimados (sob a situação definida como paradigma) 132
5.33 Alterações do valor unitário mediano na área delimitada pelo perímetro urbano de Londrina com recortes temporais definidos semestralmente ao longo do período de jan. 2000 a jul. 2021 (sob a situação definida como paradigma) . . . 133
5.34 Variação proporcional relativa entre o valor unitário mediano na área delimitada pelo perímetro urbano de Londrina entre as datas de jul. 2000 e jul. 2021 (sob a situação definida como paradigma) 134
5.35 *Worm plots* de modelos com estrutura idêntica à do modelo proposto, mas sob variadas transformações da variável “AT” 135
5.36 Efeitos da variável “AT” sob variadas transformações, sobre a mediana distribucional (incorporadas ao seu preditor linear sob a forma de função suavizadora unidimensional) 135

LISTA DE TABELAS

3.1 Distribuição BCT 68

4.1 Estatísticas das distribuições teóricas melhor ajustadas à variável resposta . . . 95

5.1 Medidas resumo das variáveis quantitativas incorporadas como termos nos preditores do modelo 102
5.2 Informações sobre algumas métricas do modelo proposto 106
5.3 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* 116
5.4 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada intervalo de valores da variável "AT" 116
5.5 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada intervalo de valores da variável "DATA" 116
5.6 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "RELEVO" 116
5.7 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "PAVIMENTACAO" 116
5.8 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "NATUREZA" 117
5.9 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "IMPLANTACAO" 117
5.10 Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* conjuntos de cada intervalo de valores da variável "AT" e nível do fator "NATUREZA" . . 117
5.11 Medidas descritivas dos resíduos quantílicos do modelo proposto 118
5.12 Limites do intervalo de confiança para as estimativas do intercepto do modelo proposto 122
5.13 Limites do intervalo de confiança para as estimativas do coeficiente do fator: "NATUREZA" (se transação) 123
5.14 Limites do intervalo de confiança para as estimativas do coeficiente do fator: "IMPLANTACAO" (se condomínio) 124
5.15 Limites do intervalo de confiança para as estimativas do coeficiente do fator: "RELEVO" (se aclive) 125
5.16 Limites do intervalo de confiança para as estimativas do coeficiente do fator: "RELEVO" (se declive) 126

5.17 Limites do intervalo de confiança para as estimativas do coeficiente do fator: "PAVIMENTACAO" (se asfáltica) 127
5.18 Informações sobre algumas métricas dos modelos sob variadas transformações da variável "AT" 136

LISTA DE QUADROS

3.1 Interpretação dos vários padrões sistemáticos de afastamento da linha horizontal da curva ajustada de um *worm plot* mostrados na Figura 3.12 83

4.1 Informações cadastrais dos elementos pesquisados 88
4.2 Resumo das variáveis utilizadas como termos do modelo 94

5.1 Parametrização das funções computacionais 106
5.2 Diferenças estatisticamente significativas pelo Teste de Voung ($\alpha = 0,05$) . . . 136
5.3 Diferenças estatisticamente significativas pelo Teste de Clarke ($\alpha = 0,05$) . . . 136

LISTA DE SIGLAS, ACRÔNIMOS E EXPRESSÕES ESTRANGEIRAS

ABNT	Associação Brasileira de Normas Técnicas
AIC	*Akaike Information Criteria*: Critério de Informação de Akaike (do inglês)
Box-plot	Gráfico de caixa (do inglês)
BCT	Distribuição Box&Cox "t" de probabilidade (a Distribuição Box&Cox "t" original difere apenas pela função de ligação no preditor linear do primeiro parâmetro distribucional de um *GAMLSS* na implementação computacional)
Contour map	Mapa de Contornos ou Isolinhas (do inglês)
BIC	*Bayesian Information Criteria*: Critério de Informação Bayesiano (do inglês)
CBD	*Central Business District*: centro comercial e financeiro de uma cidade (do inglês)
CONFEA	Conselho Federal de Engenharia e Agronomia
CREA	Conselho Regional de Engenharia e Agronomia
CV	*Cross validation*: Validação cruzada (do inglês)
Deviance	Afastamento de um padrão referencial de comportamento ou estado (do inglês)
Distributional regression	Regressão distribucional (do inglês)
EDF	*Effective degrees of freedom*: Graus de liberdade efetivos (do inglês)
EPSG	*European Petroleum Survey Group* foi uma organização científica ligada à indústria petrolífera europeia
EMV	Estimativa de máxima verossimilhança
EQM	Erro quadrático médio
FDP	Função densidade de probabilidade

FCP	Função de distribuição acumulada de probabilidade
GAM	*Generalized Additive Model*: Modelo Aditivo Generalizado (do inglês)
GAMLSS	*Generalized Additive Model for Location, Scale and Shape*: Modelo Aditivo Generalizado para Locação, Escala e Forma (do inglês)
GB2	Distribuição Beta Generalizada 2 (do inglês: *generalized beta 2*)
GDEV	*Global deviance*: Desvio global (do inglês)
GACI	*Generalized Akaike information criteria*: critério de Akaike generalizado (do inglês)
GLM	*Generalized Linear Model*: Modelo Linear Generalizado (do inglês)
Gray scale map	Mapa em escala de tons de cinza (do inglês)
IBGE	Instituto Brasileiro de Geografia e Estatística
LM	*Linear Model*: Modelo Linear (do inglês)
Moving window	Janela móvel (do inglês)
NBR	Norma Brasileira
Overfitting	Ajustamento excessivo aos dados (posto a palavra "sobreajuste" não constar no Vocabulário Ortográfico da Língua Portuguesa: VOLP) (do inglês)
PNAD	Pesquisa Nacional por Amostragem de Domicílios
Rought	Áspero, irregular; no contexto: função muito oscilante (do inglês)
Scatterplot	Gráfico de dispersão (do inglês)
SI	Sistema Internacional de Unidades: INMETRO
Smooth	Suave; no contexto: função com pouca oscilação (do inglês)
Smoother	Mais suave; no contexto: função que suaviza a variabilidade observada em um gráfico de dispersão (do inglês)
SQR	Soma dos quadrados dos resíduos
TRVG	Teste da razão de verossimilhanças generalizado

UTM	(Projeção) Universal Transversa de Mercator (do inglês: *Universal Transverse Mercator*)
Wiggly	Aspecto que assume de uma linha com muitas curvas; no contexto: função muito oscilante (do inglês)

LISTA DE SÍMBOLOS

Y;X;Z;V	variáveis aleatórias
$\boldsymbol{y}; \boldsymbol{x}; \boldsymbol{z}; \boldsymbol{v}$	vetores de valores observados de variáveis aleatórias
$\boldsymbol{x}_{kj}$	*j-ésimo* vetor de valores observados para o *k-ésimo* parâmetro de um *GAMLSS*
ε	erro de um modelo
Θ	espaço paramétrico de uma distribuição
θ	parâmetro de uma distribuição (θ_k representa o *k-ésimo* parâmetro de um *GAMLSS*)
$\boldsymbol{\theta}$	vetor paramétrico de uma distribuição ($\boldsymbol{\theta_k}$ representa o *k-ésimo* vetor paramétrico de um *GAMLSS*)
μ	primeiro parâmetro de uma distribuição (geralmente sua locação)
σ	segundo parâmetro de uma distribuição (geralmente associado à sua variabilidade)
ν	terceiro parâmetro de uma distribuição (geralmente associado à sua forma quanto à simetria)
τ	quarto parâmetro de uma distribuição (geralmente associado à sua forma quanto à curtose)
ϕ	parâmetro de dispersão de um *GLM* ou *GAM*
$\boldsymbol{X}_{(n,p)}$	matriz de delineamento dos efeitos fixos de p variáveis
$\boldsymbol{X}_k$	matriz de delineamento dos efeitos fixos para o k-ésimo parâmetro de um *GAMLSS*
$S_j(x)$	polinômio cúbico para o subintervalo $[x_j, x_{j+1}]$
t_i	*i-ésimo* nó de um *spline*
$a(.); b(.); f(.); g(.); h(.)$	funções (suavizadoras) de uma variável
$s_{jk}(.)$	*j-ésima* função suavizadora para o *k-ésimo* parâmetro de um *GAMLSS*

$b_j(.)$	*j-ésima* função de uma base para uma função suavizadora
$\boldsymbol{B}_{(n,j)}$	matriz de delineamento dos efeitos aleatórios de uma função suavizadora unidimensional estimada por *j-funções* de base
$\boldsymbol{Z}_{jk}$	matriz de delineamento para o *j-ésimo* suavizador do *k-ésimo* parâmetro de um *GAMLSS*
$B_{i,j}(.)$	*i-ésima* função (de ordem j) *B-spline* de uma base para uma função suavizadora
$\alpha; \beta; \delta; \epsilon; \gamma$	parâmetros de modelos (coeficientes de funções de suavização)
$\boldsymbol{\beta}_k$	vetor dos parâmetros dos efeitos fixos para o *k-ésimo* parâmetro de um *GAMLSS*
$\boldsymbol{\gamma}_{kj}$	*j-ésimo* vetor de efeito aleatório para o *k-ésimo* parâmetro de um *GAMLSS*
$\hat{\alpha}; \hat{\beta}; \hat{\delta}; \hat{\epsilon}; \hat{\gamma}$	estimativas dos parâmetros de modelos (coeficientes de funções de suavização)
$\boldsymbol{\alpha; \beta; \delta; \epsilon; \gamma}$	vetores paramétricos de modelos (coeficientes de funções de suavização)
$\hat{\boldsymbol{\alpha}}; \hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\delta}}; \hat{\boldsymbol{\epsilon}}; \hat{\boldsymbol{\gamma}}$	estimativas de vetores paramétricos de modelos (coeficientes de funções de suavização)
λ	parâmetro que regula a quantidade de suavização no ajuste de uma função de suavização (hiperparâmetro em um *GAMLSS*)
$\boldsymbol{\lambda}$	vetor de todos os hiperparâmetros de um *GAMLSS*
$\boldsymbol{D}_{n,j}$	matriz de diferenças entre dois *B-splines* adjacentes ($\boldsymbol{D}_{[d]}$ indica uma matriz de diferenças de ordem d)
$\boldsymbol{G} = \boldsymbol{D}^T\boldsymbol{D}; (\boldsymbol{S} = \boldsymbol{D}^T\boldsymbol{D})$	matrizes de penalização
$t_i(x_i, y_i)$	*i-ésima* observação tomada de um ponto (x_i, y_i) no plano
$r(\|\|(x_i, y_i) - (x_c, y_c)\|\|)$	norma entre dois pontos no plano
$\zeta(r)$	função radial aplicada à norma "r"
$\boldsymbol{T}_{(3,n)}$	matriz dos valores avaliados das funções lineares de base ϕ de um *thin-plate spline* aplicadas sobre os pontos t_i

$\boldsymbol{E}_{(n,n)}$	matriz dos valores avaliados da função radial η de um *thin-plate spline* aplicada sobre as normas r
$J_{[d]}$	funcional que quantifica a oscilabilidade de uma função suavizadora *d-dimensional*
$g(\boldsymbol{t})$	*thin plate spline* de t
$\boldsymbol{o}$	vetor de zeros
$\otimes$	operador do produto Kronecker
$\odot$	operador do produto Kronecker por linha
D	desvio de um modelo linear generalizado
$g_k(.)$	função de ligação aplicada ao *k-ésimo* parâmetro de uma distribuição em um GAMLSS
$\boldsymbol{\eta}_k$	*k-ésimo* preditor linear de um GAMLSS tal que $\boldsymbol{\eta}_k = g_k(\boldsymbol{\theta}_k)$
L	função de verossimilhança
$\hat{L}_c$	função de verossimilhança para o modelo em estudo
l	logaritmo da função de verossimilhança
l_p	logaritmo da função de verossimilhança penalizada
κ	penalidade para cada grau de liberdade no modelo aplicada no cálculo do GAIC
df	graus de liberdade (efetivos) usados em um modelo
Q	Estatística Q calculado sobre os resíduos de um modelo
Z	Estatística Z calculado sobre os resíduos de um modelo
$\mathcal{M}$	um GAMLSS estruturado como $\{\mathcal{D}; \mathcal{G}; \mathcal{T}; \mathcal{L}\}$
$\mathcal{D}$	a distribuição especificada para de um *GAMLSS*
$\mathcal{G}$	as funções de ligação para os preditores dos parâmetros distribucionais de um *GAMLSS*
$\mathcal{T}$	os termos explicativos de um *GAMLSS*
$\mathcal{L}$	a especificação dos parâmetros de suavização de um *GAMLSS*

$\boldsymbol{r} = \Phi^{-1}(.)$	vetor dos resíduos quantílicos Normalizados
$\Phi^{-1}(.)$	função quantílica da distribuição Normal padronizada
$F(y)$	função de distribuição acumulada de probabilidade de uma variável aleatória y
$f_{Y\|X}(y)$	função densidade de probabilidade condicional de uma variável aleatória Y dado X
$\mathcal{E}$	família exponencial das distribuições
$\mathcal{N}(\mu, \sigma^2)$	distribuição Normal com média μ e variância σ^2
$\mathcal{N}(0, 1)$	distribuição Normal padronizada
$\phi(.)$	função densidade de probabilidade de uma distribuição Normal padronizada
$\Phi(.)$	função de distribuição acumulada de probabilidade de uma distribuição Normal padronizada
$\bar{\beta}^*$	valor médio das replicações *bootstrap* de uma estatística qualquer
δ^*	afastamentos medidos entre uma estatística qualquer e suas replicações *bootstrap*

SUMÁRIO

Lista de Figuras viii

Lista de Tabelas xi

Lista de Quadros xiii

Lista de Siglas, Acrônimos e Expressões Estrangeiras xiv

Lista de Símbolos xvii

1 Problemática 22

2 Conceitos iniciais 27

2.1 Variabilidade do valor do espaço urbano . . . 27
2.2 Engenharia de avaliações e normatização associada . . . 29
2.3 Valor . . . 31

3 Revisão da literatura 34

3.1 Funções matemáticas e modelos estatísticos . . . 34
3.2 Modelo Clássico de Regressão Linear (*LM*) . . . 36
3.2.1 Modelos Hedônicos de Regressão . . . 37
3.3 Modelos Lineares Generalizados (*GLM*) . . . 38
3.4 Funções de suavização para métodos não paramétricos . . . 40
3.4.1 Splines . . . 41
3.4.2 *B-splines* . . . 44
3.4.3 *Splines* penalizados (*P-splines*) . . . 46
3.4.4 *Thin plate splines* . . . 49
3.4.5 *Thin plate regression splines* . . . 54
3.4.6 *Splines* de produto tensor . . . 54
3.5 Modelos aditivos . . . 57
3.6 Modelos Aditivos Generalizados (*GAM*) . . . 60
3.7 Modelos Aditivos Generalizados para Locação, Escala e Forma . . . 62
3.7.1 Introdução . . . 62
3.7.2 Definição . . . 62

3.7.3 Família de distribuições *GAMLSS* e funções de ligação 65
3.7.3.1 Distribuição Box&Cox "t" (BCT) 65
3.7.4 Preditor 69
3.7.5 Ajuste 70
3.7.5.1 Comparação de modelos 72
3.7.6 Inferência 74
3.7.6.1 Intervalos de confiança baseados em erros padrão robustos (Huber *sandwich*) 74
3.7.6.2 Inferências baseadas em *bootstrapping* 75
3.7.7 Diagnóstico 79
3.7.7.1 Resíduos quantílicos (aleatorizados) Normalizados 79
3.7.7.2 *Worm plot* dos resíduos quantílicos Normalizados 81
3.7.7.3 Estatísticas Q e Z dos resíduos quantílicos (aleatorizados) Normalizados 84

4 Materiais e métodos **86**
4.1 Materiais 86
4.1.1 Natureza da informação obtida 88
4.1.2 Descrição 88
4.1.3 Georreferenciamento 90
4.1.4 Arquivos digitais de informações espaciais 91
4.1.5 Setores censitários 91
4.2 Recursos computacionais utilizados 91
4.3 Roteiro geral da análise 92
4.4 Descrição das variáveis incorporadas como termos nos preditores do modelo . . 93
4.5 Métodos 95
4.5.1 Escolha da distribuição e funções de ligação 95
4.5.2 Incorporação de variáveis nos preditores dos parâmetros distribucionais 97
4.5.3 Geração de imagens 100

5 Resultados obtidos **102**
5.1 Análise descritiva da amostra 102
5.2 Modelo proposto 105
5.2.1 Efeitos dos termos incorporados ao preditor do primeiro parâmetro da distribuição ($\boldsymbol{\theta_1}$: mediana) 106
5.2.2 Efeitos dos termos incorporados ao preditor do segundo parâmetro da distribuição ($\boldsymbol{\theta_2}$: coeficiente de variação) 109
5.2.3 Efeitos dos termos incorporados ao preditor do terceiro parâmetro da distribuição ($\boldsymbol{\theta_3}$: coeficiente *proxy* da assimetria) 113

5.2.4 Efeitos dos termos incorporados ao preditor do quarto parâmetro da distribuição ($\boldsymbol{\theta_4}$: coeficiente *proxy* da curtose) 113
5.2.5 Análise diagnóstica do modelo proposto 115
5.2.5.1 *Worm plot* dos resíduos quantílicos 115
5.2.5.2 Gráficos de probabilidade dos resíduos quantílicos 117
5.2.5.3 Estatísticas Z 118
5.2.6 Inferências sobre coeficientes estimados 121
5.2.7 Autocorrelação espaçotemporal nos resíduos simples 128
5.2.8 Coeficiente de determinação generalizado 128
5.2.9 Superfícies de valor 130
5.2.10 Variabilidade espaçotemporal 134
5.2.11 Transformação da variável "AT" 134

6 Conclusão **137**

Referências **138**

1 PROBLEMÁTICA

Os serviços técnicos prestados pela engenharia de avaliações são de grande relevância para o desenvolvimento sócio-econômico nacional e, portanto, encontram-se submetidos a uma rígida normatização de seus procedimentos. Fornecem subsídios de suma importância não apenas para órgãos da administração pública e o sistema legal, mas também para diversas atividades econômicas privadas (CHICA-OLMO, 2007). Suas informações lastreiam a fundamentação de decisões judiciais como em processos indenizatórios de desapropriações ou na partilha de bens. Garantem um procedimento tributário equânime nas decisões administrativas de natureza fiscal. Dão segurança às operações financeiras envolvendo garantias reais, quer sejam para as aquisições individuais com o propósito de uso (residencial ou comercial), como também para operações de fundos de investimento imobiliário (CHICA-OLMO; CANO-GUERVOS, 2020) [1].

Entretanto, a heterogeneidade dos bens imobiliários urbanos torna bastante complexa a proposição de modelos estatísticos que procurem estimar seus valores. A teoria não especifica a forma funcional exata nem tampouco as variáveis mais relevantes na formação do valor (FLORÊNCIO; CRIBARI-NETO; OSPINA, 2012).

Como exposto por vários autores, como em Florêncio (2018) e Paixão (2015), bens de natureza imobiliária diferem não somente em relação à sua destinação básica como residenciais; comerciais ou industriais, mas também quanto à sua tipologia como, por exemplo, glebas urbanizáveis, lotes urbanizados; imóveis residenciais ou comerciais. Muitas dessas tipologias ainda admitem serem implantadas tanto de modo isolado, quanto condominial, quando então áreas de uso comum são incorporadas ao empreendimento, disponibilizando aos proprietários uma gama de instalações de lazer.

Complementarmente, existem ainda diferenças estruturais dentro de uma mesma tipologia e destinação como, por exemplo, tamanhos relacionados à área edificada ou então do terreno onde está edificada; aspectos geológicos ligados à natureza do solo, seu relevo; potencial construtivo permitido pela legislação urbana; orientação solar; padrão construtivo; estado geral de conservação ou ainda quantidades associadas tais como vagas de garagem, número de quartos, salas, banheiros.

Independente da diversidade de aspectos que caracterizam um bem imobiliário específico, todos eles podem ser agrupados sob a denominação comum de aspectos intrínsecos, particularidades inerentes ao bem.

De modo análogo, outro sem número de características extrínsecas podem ser relacionadas, buscando associar as peculiaridades da região onde o bem se localiza como, por

[1] Para compreensão da relevância dos valores atribuídos a bens dados como garantia real em operações de crédito recomendamos a leitura de artigos relacionados às causas da crise mundial financeira de 2007-2008.

exemplo, a existência ou proximidade a escolas e universidades; estabelecimentos comerciais, *malls* ou *shopping centers*, unidades de saúde, hospitais e estações de transporte coletivo.

Bens imobiliários são bens econômicos e, como tais, estão sujeitos a circunstâncias temporais, um terceiro componente essencial na estimação do valor de um imóvel. Assim, qualquer modelo que busque estimar o valor de um bem imobiliário deverá conseguir considerar três aspectos simultaneamente:

- intrínsecos: características endógenas que o definem ;
- extrínsecos: características exógenas relacionadas à sua localização;
- temporal: a data a que se refere o valor.

Por estarem vinculados a parcelas do solo urbano, bens imobiliários são bens insuscetíveis de movimento. Não podem ser transportados de um lugar para o outro sem serem destruídos e, desta forma, seu valor encontra-se, em alguma proporção, relacionado ao valor daquelas e o solo não é um bem fungível nem, igualmente, transportável (CHICA-OLMO, 1994). Por essa razão, há no mercado imobiliário uma ideia tão antiga, quanto tida como verdadeira de que a localização é o fator determinante na composição do valor de um imóvel, frequentemente reconhecida no meio acadêmico: "Como qualquer avaliador lhe dirá, os três mais importantes componentes do preço de uma casa são: localização, localização e localização" (DUBIN, 1992).

Assim, a localização de um bem imobiliário acaba por ser um dos fatores mais importantes na composição de seu valor em um grau, de algum modo, relacionado à sua tipologia, como ilustrado na Figura 1.1.

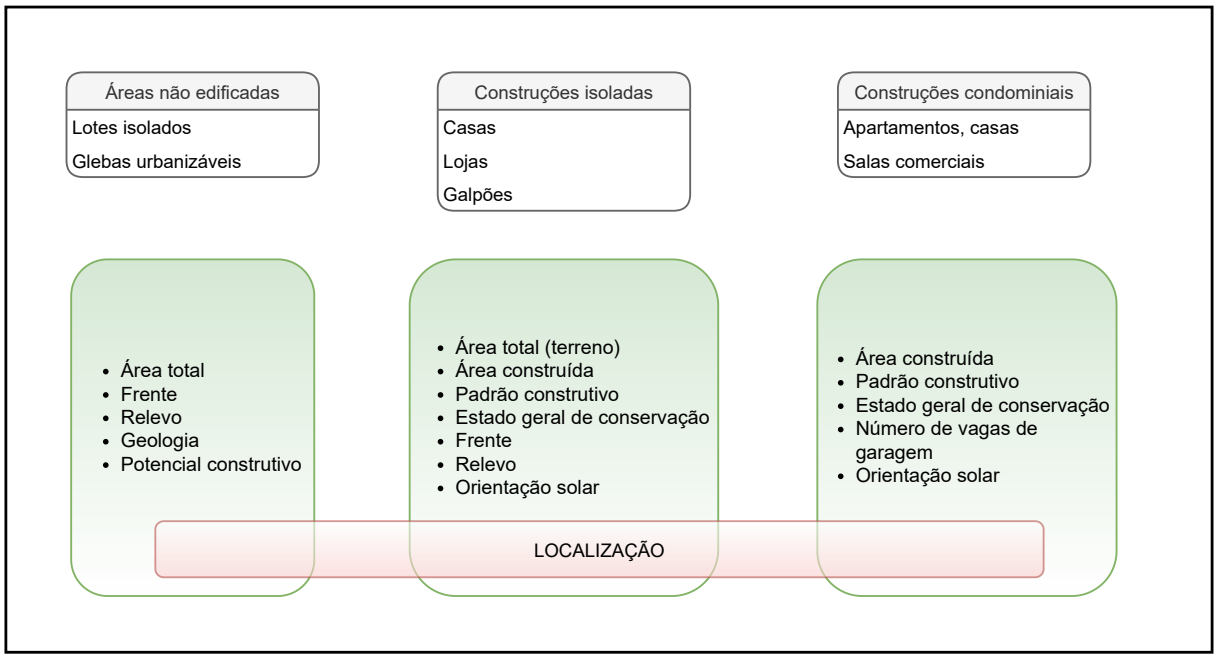
A noção de que imóveis semelhantes apresentam valores diferentes, dependendo de onde quer que eles estejam localizados numa cidade, é amplamente difundida. Equivale dizer haver consenso de que cada lugar numa cidade influi de um determinado modo, nem sempre positivo, no valor dos bens imobiliários ali existentes.

Um entendimento menos raso, mas ainda subjetivo, de que estas diferenças decorrem da influência exercida por fatores urbanos que existem na região onde o imóvel se localiza é um pouco mais restrito, limitado àqueles que operam no mercado imobiliário.

Todavia a plena compreensão, e a derivada capacidade de poder analisar e estimar objetivamente o efeito conjunto desses fatores sobre o valor de um bem imobiliário em qualquer parte da cidade, é reservada àqueles que se dedicam ao estudo científico da matéria.

A introdução da Norma Brasileira (NBR) 5676/90 abriu caminho para o abandono em definitivo dos cálculos avaliatórios baseados em simples ponderações feitas por meio dos chamados fatores empíricos de uso consagrado. Os primeiros trabalhos alicerçados na chamada metodologia científica foram publicados alguns anos antes (e republicados poucos anos depois) e propunham um Modelo Hedônico de Regressão Linear Clássico (DANTAS,

Figura 1.1: Localização, variável determinante e comum a todas as tipologias de bens imobiliários



Fonte: Próprio autor

1988) e outro (DANTAS; CORDEIRO, 1991), baseado na classe de Modelos Lineares Generalizados pouco tempo antes proposta (NELDER; WEDDERBURN, 1972).

Gomes e Monteiro (2009) mostraram as conclusões a que chegaram num estudo da contribuição marginal de atributos valorativos de propriedades residenciais. Seu modelo de formação de valor construído com dados levantados pela Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2007, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) considerava 16 variáveis como explicativas do valor do bem e formavam quatro grupos: construção, dimensões, infraestrutura e localização. A contribuição marginal obtida ao se agregar cada um desses grupos no modelo final revelou que o grupo *localização* respondia, individualmente, por 61,64 % da variabilidade do valor.

A variabilidade espacial inerente aos dados pode ser considerada nos modelos de avaliação de imóveis sob várias formas como, por exemplo:

- na forma de modelos de regressão com efeitos espaciais: *Spatial Autoregressive Model (SAM)* ou *Spatial Error Model (SER)* (DATAS et al., 2001);
- sob o prisma da geoestatística (DANTAS, 2003), (TRIVELLONI, 2005) e (CHICA-OLMO, 1994);

- pela incorporação de superfícies de tendência pela inclusão de suas coordenadas métricas da localização sob uma forma polinomial (MICHAEL, 2004) o que, frequentemente, resulta em termos altamente correlacionados (DUBIN, 1992);

- pela inclusão de variáveis que relacionem a presença de amenidades e serviços públicos nas imediações (GOMES; MONTEIRO, 2009).

A variabilidade temporal, quando considerada em um modelo, é usualmente incorporada sob a forma de mais uma variável, por vezes um fator assumindo tantos níveis quanto períodos temporais existirem na amostra, como em Wolverton (1997).

Todavia, o que se observa nos modelos assim estruturados, é que os aspectos espacial e temporal são considerados dissociadamente, contrariando o observado que diferentes regiões de uma cidade se valorizam (ou desvalorizam) de modos distintos temporalmente, não sendo plausível admitir-se como válida uma variabilidade temporal espacialmente homogênea.

Uma cidade, todavia, não se apresenta estática, imutável, por assim dizer. Está continuamente sofrendo alterações. Locais onde antes predominavam habitações térreas residenciais são ocupados por construções verticais às vezes com uso diverso daquelas. Algumas regiões, antes consideradas aprazíveis, podem se tornar degradadas e passar a serem preteridas por outras áreas que, na via reversa, antes eram recusadas e, por intervenção públicas ou privada, foram revitalizadas.

Isso corrobora a ideia de que mesmo propriedades intrinsecamente idênticas possam ser vendidas por preços muito diferentes a depender da localização e, ainda que se considere um mesmo local, possam ser negociadas por distintos valores em diferentes épocas. Assim, o valor de um bem imobiliário não pode ser estimado de forma confiável sem consideração cuidadosa de sua localização em conjunto com o tempo de referência.

Continuamente, novas e mais eficientes técnicas estatísticas de modelagem são propostas na literatura, como as que possibilitam a estimação de modelos em que as variáveis podem ser consideradas não apenas de modo paramétrico mas também não paramétrico, na forma de funções suavizadoras, assim como outros parâmetros da distribuição teórica adotada para a resposta também sejam estimados.

Os Modelos Aditivos Generalizados para Locação, Escala e Forma (*GAMLSS*: *Generalized Additive Model for Location, Scale and Shape*) são considerados como modelos de regressão da distribuição (*distributional regression*) (KLEIN et al., 2015) ao possibilitar a modelagem conjunta, de modo paramétrico, semiparamétrico ou não paramétrico, de todos os parâmetros da distribuição teórica adotada para a variável resposta

O presente trabalho propõe um Modelo Hedônico de Regressão Espaçotemporal baseado nessa classe de modelos, estimado sobre uma amostra composta de informações sobre terrenos urbanos sem benfeitorias na cidade de Londrina (norte do Estado do Paraná) coletada no período de maio de 1995 a março de 2021, que considera a variabilidade espacial e temporal de forma conjunta pela incorporação de um suavizador de produto tensor

que combina uma base e *splines* cúbicos para o tempo e uma base de *thin plate splines* para o espaço nos preditores dos parâmetros da distribuição teórica adotada para a variável resposta.

2 CONCEITOS INICIAIS

2.1 VARIABILIDADE DO VALOR DO ESPAÇO URBANO

> "Desde que o valor depende da renda econômica, a renda da localização e da conveniência e a conveniência da proximidade, nós podemos eliminar estes passos e dizer que o valor depende da proximidade. A próxima questão é: proximidade ao quê?, a qual nos remete às necessidades fundiárias dos diferentes usos, sua distribuição pela área da cidade e a consequente geração e distribuição de valores" (HURD, 1924).

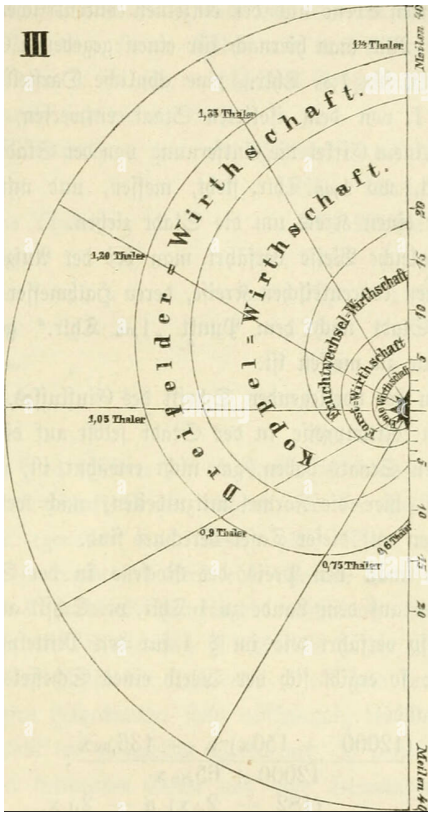
O espaço urbano apresenta diferentes aspectos em sua extensão: seja em razão de intervenções realizadas pelo homem, seja pela variação das características naturais do solo. Os fatores de maior impacto no valor de propriedades imobiliárias ao se considerar sua posição geográfica podem ser agrupados em quatro grupos distintos (CAN, 1998):

- características de acessibilidade da vizinhança;
- meio ambiente físico;
- contextos social, econômico e demográfico;
- serviços públicos essenciais.

Jean-Jacques Granelle (GRANELLE, 1970 apud CHICA-OLMO, 2007) considera que, do ponto de vista espacial, os bens urbanos sofrem influência de fatores que podem ser agrupados segundo três escalas diferentes. Sob uma escala de microlocalização encontram-se os fatores físicos do ambiente imediato como topografia, pedologia, etc., os de acessibilidade local, como a proximidade a serviços necessários como escolas, estabelecimentos comerciais, de saúde, transporte e os do ambiente socioeconômico, como nível de criminalidade, renda média, etnia. Em uma escala de macrolocalização estão aqueles relacionados à posição do bairro na cidade como acessibilidade ao centro, densidade de ocupação, potencial construtivo concedido pela legislação municipal. E, por fim, há os fatores gerais que afetam a cidade na totalidade como conjuntura econômica, questões ligadas a políticas estaduais e federais, etc.

Os primeiros modelos de estimação do valor do solo foram desenvolvidos por Johan Heinrich von Thünen (1783-1850) (CHICA-OLMO, 1994). Seu modelo de uso agrícola do solo em "O estado isolado" (THÜNEN, 1826) era genericamente composto por círculos concêntricos ao redor do centro urbano onde se localizava o polo consumidor, centro do modelo; a área imediatamente ao redor da cidade, polo de produção agrícola intensiva de frutas, hortaliças e laticínios; as florestas (fontes de combustível); na sequência as áreas destinadas à

Figura 2.1: Ilustração histórica do modelo de círculos concêntricos definindo áreas com diferentes valores do solo em relação ao núcleo urbano



Fonte: Adaptado de (THÜNEN, 1826)

colheita de grãos e, por fim, as áreas destinadas à criação de animais, como ilustrado na Figura 2.1.

Chica-Olmo afirma que Von Thünen reconhecia que o homem buscava resolver suas necessidades econômicas na região imediatamente adjacente, reduzindo seus deslocamentos ao mínimo; todavia, ele não compreendia por que dois lotes de terras com as mesmas características possuíam valores diferentes. Explicou esta diferenciação formulando um modelo em que o valor da localização era uma função do rendimento da propriedade (preços de mercado dos produtos produzidos menos os custos de produção) e das taxas de transporte (“Uma porção de cada colheita é comida pelas rodas”). A ideia central era de que os espaços urbanos eram homogêneos. Havia apenas um centro comercial e o fator diferencial era a distância a esse centro. Pela teoria das zonas concêntricas, o valor da localização era então função do custo de transporte que eram considerados iguais em todas as direções (CHICA-OLMO, 1994) e determinante para a natureza de sua ocupação.

Richard Muth mostra um modelo monocêntrico de urbanização até pouco tempo dominante, em que a densidade populacional, os valores da terra e os preços das casas caem com a distância do centro da cidade:

> "Ele acreditava que as diferenças nos gradientes de densidade entre as áreas metropolitanas deveriam ser explicadas por três conjuntos de fatores: a natureza e o custo de transporte pendular disponível para os trabalhadores do centro financeiro e comercial da cidade; a distribuição espacial do emprego e *shopping centers*; e preferências para habitação em vários pontos da cidade. Ele estimou e testou a importância de esses fatores regredindo os gradientes de densidade em várias variáveis que se acredita serem medidas desses três conjuntos de fatores" (MUTH, 1961 apud MILLS, 1970).

Em contraponto à simplicidade do modelo monocêntrico de concepção de uma cidade, teorias alternativas surgiram buscando descrever o espaço urbano como não dependente apenas de uma única localização privilegiada: "Por trás dessa simplicidade um núcleo de verdade se esconde e permite que esses modelos melhorarem nossa compreensão das cidades mais modernas" (RICHARDSON, 1975 apud CHICA-OLMO, 1994).

Uma concepção mais realista é admitir a existência de diversos polos com áreas de atratividade bem delimitadas (DUBIN, 1992), com setores destinados ao comércio, à habitação, à indústria (CHICA-OLMO, 1994). Assim, neste contexto urbano mais atual, uma cidade não possui apenas um polo de atratividade como centro comercial e financeiro.

Estes polos são áreas preferencialmente desejadas para ocupação, quer seja por serem dotadas de serviços públicos essenciais: vias com pavimentação, iluminação, drenagem superficial, redes de distribuição de energia elétrica, telefonia, água tratada e coleta de esgoto sanitário, concentrar não somente áreas de trabalho, mas também estabelecimentos comerciais e de serviços, culturais, áreas destinadas à recreação e lazer ou ainda então estarem em áreas geologicamente estáveis, preferencialmente de topografia plana e afastados de áreas degradadas. Todos esses fatores resultam em valores mais elevados no mercado imobiliário.

De modo análogo, os aspectos urbanos acima relacionados podem gerar, numa situação antagônica, um contexto desfavorável, impactando negativamente na atratividade da região. Aspectos geológicos economicamente desfavoráveis, reduzida disponibilidade de serviços públicos, grande distância de áreas demandantes de trabalhadores e proximidade de indústrias geratrizes de poluição são algumas das situações que criam um efeito contrário na atratividade de uma região urbana.

Percebe-se assim, sem dificuldade, que em cada local de uma cidade há uma variedade de fatores exercendo, de modo simultâneo, efeito no valor das propriedades naquela região urbana.

2.2 Engenharia de avaliações e normatização associada

Com a promulgação por D. Pedro II da Lei n^o 601 em 18 de setembro de 1850 (vide) ocorre a extinção do "Sistema de concessões" instituído pela Coroa portuguesa em 1375. Esta lei dispunha sobre as terras devolutas do Império e seu artigo 1^o instituía: "Ficam prohibidas as acquisições de terras devolutas por outro titulo que não seja o de compra" (sic).

A partir desse momento a incipiente sociedade capitalista teve na terra uma "forma de reserva e apropriação de capital baseados nas expectativas de valorizações e um meio de ganhos financeiros por meio de suas rendas locatícias e vendas" (ALONSO, 2007). A necessidade de se apurar esse valor determinou o surgimento da engenharia de avaliações no Brasil nas décadas seguintes (c. 1910) com a difusão dos primeiros trabalhos técnicos sobre o tema, inicialmente nos Boletins do Instituto de Engenharia ("Códigos Sanitários e Posturas Municipais sobre Habitações" na Revista Politécnica pelo Prof. Eng^o Vitor da Silva Freire, em 1918) e na revista Nacional Mackenzie.

Em 1940 surge a Associação Brasileira de Normas Técnicas (ABNT), responsável pela elaboração das Normas Brasileiras (NBR) por meio e seus Comitês Brasileiros (ABNT/CB), Organismos de Normalização Setorial (ABNT/ONS) e Comissões de Estudo Especiais (ABNT/CEE) e, nessa mesma época, o Eng^o Luis Carlos Berrini lança seu primeiro livro "Avaliação de Terrenos" (BERRINI, 1949). Logo a seguir, o Eng^o Alberto de Zagottis escreve sobre a importância do método estatístico como instrumento de avaliação pelo método científico.

Em 1954, o Eng^o Hélio de Caires e sua equipe promovem em São Paulo a 3ª Convenção Pan-americana de Avaliações, abrindo caminho para a abertura de diversos institutos estaduais, estando entre os mais antigos o Instituto de Engenharia Legal (IEL) no Rio de Janeiro, fundado em 1953 e o Instituto Brasileiro de Avaliações e Perícias de Engenharia (IBAPE) em São Paulo, em 1954 (ALONSO, 2007).

A gênese da normatização dos procedimentos avaliatórios deu-se em novembro de 1952, pelo Anteprojeto de Normas para Avaliação de Imóveis escrito pelo Eng^o Augusto Luís Duprat à ABNT (P-NB-74R/57) que, anos mais tarde, em 1977 tornou-se a 1^a Norma Brasileira sobre avaliações: NB-502/77 – "Norma para Avaliação de Imóveis Urbanos".

Em 1989 a NB-502/77 teve a sua revisão concluída, quando recebeu a nomenclatura de NB-502/89, trazendo no seu corpo enormes avanços em relação ao texto anterior, especialmente no que diz respeito aos critérios estabelecidos para o tratamento estatístico inferencial pois, até então, era permitido ao engenheiro de avaliações na homogeneização dos elementos amostrais frente ao bem avaliando a utilização de *fatores empíricos tecnicamente consagrados*, algumas vezes decorrentes de estudos específicos realizados, mas muitas vezes subjetivos e amparados apenas pela ancestralidade recorrente de sua adoção. Em 1990 é registrada no Instituto Nacional de Metrologia, Qualidade e Tecnologia (INMETRO) como NBR 5676:1990 (MAIA NETO, Francisco, 1992).

Considerando-se a natureza dos dados experimentais obtidos e a análise à qual foram submetidos, torna-se necessário verificar o atendimento às exigências das Normas Brasileiras NBR 14.653:1 (ABNT: Associação Brasileira de Normas Técnicas, 2011) e NBR 14.653:2 (ABNT: Associação Brasileira de Normas Técnicas, 2011), expostas no Apêndice A.1 do fascículo complementar disponível na seção Apêndices no repositório GitHub do autor em: (https://github.com/fjrcosta/Mestrado_PGMAC).

Atualmente os procedimentos exigidos na estimação do valor para as diferentes categorias de bens imobiliários são regulados pela NBR 14.653 em suas partes:

- NBR 14.653-1:2019 apresenta as diretrizes gerais para avaliação de bens (classificação; terminologia; definições; símbolos e abreviaturas; atividades e metodologia básica; especificação das avaliações; requisitos básicos de laudos de avaliação.
- NBR 14.653-2:2011 imóveis urbanos;
- NBR 14.653-3:2019 imóveis rurais;
- NBR 14.653-4:2002 empreendimentos;
- NBR 14.653-5:2006 máquinas, equipamentos, instalações e bens industriais;
- NBR 14.653-6:2008 recursos ambientais;
- NBR 14.653-7:2009 bens do patrimônio histórico e artístico.

As partes 1 e 2 dessa Norma estabelecem vários métodos para se estimar o valor de um bem imobiliário urbano, de seus frutos e direitos, e estão expostos no Apêndice A.3 do fascículo complementar disponível na seção Apêndices no repositório GitHub do autor em: . Algumas das imposições e enquadramentos normativos encontram-se detalhadas no Apêndice A.4, disponível no mesmo endereço.

2.3 VALOR

Segundo o dicionário Houaiss da língua portuguesa, valor é a:

> "Qualidade que confere a um objeto material a natureza de bem econômico, em decorrência de satisfazer necessidades humanas e ser trocável por outros bens, sendo uma medida variável de importância que se atribui a um objeto ou serviço necessário aos desígnios humanos e que, embora condicione o seu preço monetário, frequentemente não lhe é idêntico.´´

Tão distintas são as necessidades humanas que a percepção do valor a ser pago por um determinado serviço, bem ou ativo acaba por adquirir uma natureza pessoal e subjetiva, impondo complexidade à sua conceituação.

Historicamente, duas correntes conceituam o valor de modo diametralmente oposto (DANTAS, 1998):

- plurivalente: que correlaciona o valor de um bem com a finalidade para o qual é avaliado, podendo o mesmo atingir diversos valores (venal, potencial, comercial, mercado, contábil);

- univalente: que estabelece que o valor de um bem em determinado instante é único, independentemente do fim a que se destine.

A corrente plurivalente postula que, em tese, a cada momento e a cada transação o valor se altera, dependendo das relações entre os fatores influenciadores de valor (utilidade, escassez, desejo e acesso a meios de crédito para pagamento). Por outro lado, a escola univalente defende que o valor da propriedade é único para aquele instante, seja qual for a destinação da avaliação (FLORÊNCIO, 2018). A Norma Técnica Brasileira (ABNT: Associação Brasileira de Normas Técnicas, 2019) faz algumas distinções a respeito de valor, a depender do objetivo e finalidade da avaliação. Essas distinções estão expostas no Apêndice A.2 do fascículo complementar disponível na seção Apêndices no repositório GitHub do autor em: .

Não obstante as diferenças conceituais expostas aqui, e no Apêndice citado, ambas salientam um aspecto temporal associado ao valor: o momento ou instante fixado. Depreende-se dessa afirmação que o valor determinado, seja por uma ou outra escola, não é único no tempo e esta importante característica será convenientemente considerada nesse trabalho.

Bens de base imobiliária são assentes em parcelas do solo que, diferentemente do ambiente rural onde o solo é um meio essencialmente de produção, na zona urbana se destina mais a reserva e apropriação de capital, com diferentes mercados sendo estabelecidos em função da tipologia dos bens como, por exemplo, terrenos, casas, apartamentos ou galpões industriais.

O solo urbano é um bem bastante particular visto que, além de irreproduzível, é passível de monopolização pelos seus proprietários. "Por sua irreprodutibilidade e monopolização representa um bem escasso, e para a cidade crescer exige uma contínua expansão por meio da incorporação de terras agrícolas e a sua transformação em terras urbanas" (BARBOSA; COSTA, 2012). As diferenças de conceitos referentes a preço, valor e custo encontram-se estabelecidas no item 0.1 da norma NBR 14653-1:2019 da ABNT (ABNT: Associação Brasileira de Normas Técnicas, 2019) e, nesse contexto, Dantas coloca que "os preços praticados pelo mercado estão sempre se realizando, ora abaixo, ora acima do valor de mercado, e que, na prática, estima-se o valor de mercado como a média dos preços" (DANTAS, 1998).

Depreende-se do definido no item 0.1 da Norma que uma informação monetária vinculada a um dado imobiliário pode assumir um duplo aspecto:

- uma simples quantia monetária informada pelo vendedor ao anunciar seu bem (havendo até situações nas quais o vendedor nem está realmente interessado em levar o bem ao mercado): o "preço de oferta" é uma simples expectativa;

- a quantia monetária pela qual ambas as partes, vendedor e comprador, fecharam apropriadamente uma transação: o "preço de compra e venda" é um fato consumado.

Salienta-se, assim, essa dualidade da medida monetária associada a dados de natureza imobiliária. Podem se referir a um simples preço de oferta (uma etiqueta anexada ao bem como em qualquer outra mercadoria) ou então ao valor efetivamente estabelecido em uma transação concretizada. Enquanto o primeiro denota uma quantia monetária unilateral, por assim dizer, o segundo é o que deve ser considerado por representar o ponto de equilíbrio atingido por aqueles agentes, naquela conjuntura, e assim melhor indicar o verdadeiro valor de mercado do bem. Tal peculiaridade acaba por introduzir variabilidade adicional a dados dessa natureza quando comparados a outros procedentes de experimentos ou ensaios laboratoriais.

Na composição de uma amostra, não raras vezes o pesquisador esbarra na inexistência de um número mínimo de elementos da tipologia desejada que apresentem a informação monetária referente a uma efetiva transação e, por essa razão, a amostra final acaba por apresentar tanto elementos referentes a simples ofertas quanto efetivamente negociados e, nessa situação, a inclusão de mais uma variável torna-se necessária para realizar a distinção.

Vemos com ressalva a composição da amostra exclusivamente com elementos cuja informação monetária é extraída de bases cadastrais, como aquelas mantidas nas administrações municipais, tabelionatos e cartórios visto que não necessariamente refletem os verdadeiros valores envolvidos na transação. Não obstante, tal tipo de procedimento possa ser utilizado no desenvolvimento de novas técnicas de modelagem, seus valores estimados não devem ser considerados reais.

3 REVISÃO DA LITERATURA

Karl Pearson relembra em 1920 ter usado a expressão “curva normal” como uma estratégia de natureza diplomática para evitar uma questão internacional sobre precedência que poderia surgir no uso comum à época da denominação “Curva de Laplace-Gauss”, dois grandes matemáticos e astrônomos. Todavia, reconheceu também que a nova denominação poderia levar pessoas a incorrer no erro de entender que as demais distribuições seriam anormais (PEARSON, 1920):

> “Muitos anos atrás chamei a curva Laplace-Gauss de curva normal, uma denominação que, se por um lado evitou suscitar a questão internacional de prioridade, por outro trouxe a desvantagem de induzir as pessoas a crerem que todas as outras distribuições de frequência são, em um sentido ou outro, anormais”

Pelo extensivo uso que doravante será feito, bem como para se diferenciar do adjetivo, a “nova denominação” de Pearson será grafada como “Normal” considerando-a como um substantivo próprio.

3.1 FUNÇÕES MATEMÁTICAS E MODELOS ESTATÍSTICOS

Em 1936, John Maynard Keynes propôs em sua Teoria Geral do Emprego, do Juro e da Moeda a seguinte função para o consumo Y e renda X

$$Y = \beta_0 + \beta_1 X. \tag{3.1}$$

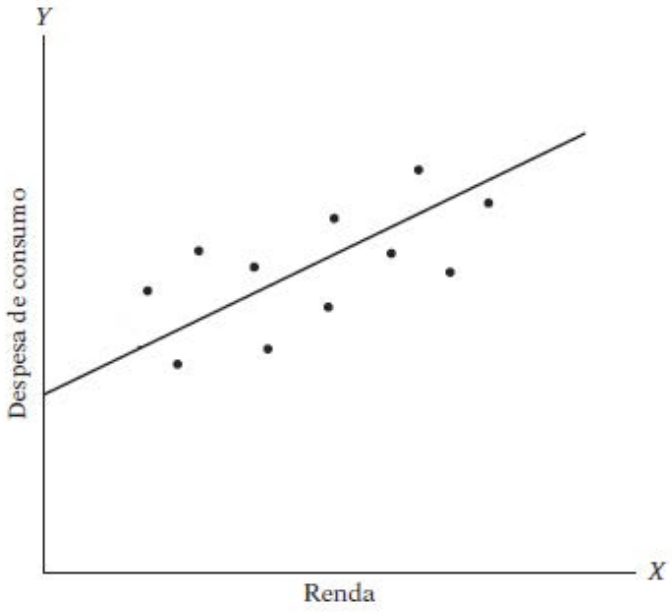
Para uma determinada renda X_i o consumo esperado, representado por Y_i, será $\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i,\ i = 1, \ldots, n$.

Nessa função matemática β_0 e β_1 são parâmetros desconhecidos em que

- β_0 é o intercepto (um consumo mínimo ou autônomo é observado mesmo nas situações em que a renda é nula, quase sempre decorrente de programas de assistência governamental);
- β_1 é a inclinação (reflete a a propensão marginal a consumir com o incremento da renda e a ela limitado): $0 < \beta_1 < 1$.

Todavia, verifica-se que em boa parte dos fenômenos estudados as relações entre as variáveis não são exatas. Observa-se sempre a presença de uma flutuação aleatória nos valores experimentais para um mesmo conjunto de valores assumido pelas variáveis explicativas como ilustrado na Figura 3.1. Assim, modelos matemáticos como o expresso em (3.1) são de aplicabilidade a um reduzido número de fenômenos cuja teoria já se encontra bem sedimentada.

Figura 3.1: Modelo econométrico da função consumo de Keynes



Fonte: Adaptado de Gujarati e Porter (2011, p. 29)

Diferentemente de uma função matemática, um modelo estatístico exprime uma relação de modo não puramente determinístico ao considerar a existência de uma aleatoriedade no processo de geração dos dados.

Em geral, modelos estatísticos podem ser classificados segundo seu propósito (STASINOPOULOS et al., 2017):

- exploratórios: propor uma teoria;
- explicativos: confirmar uma teoria;
- preditivos: estimar valores não observados.

Assim, a desempenho de um modelo está diretamente relacionada à sua capacidade explicativa ou preditiva quando aplicado em um diferente conjunto de dados. A construção de um modelo segue algumas etapas básicas:

- formulação:
  - escolha de uma distribuição de probabilidade compatível com a natureza da resposta;
  - escolha das variáveis a serem incluídas no preditor para explicar a variável estudada;
  - a função de ligação a conectar a resposta ao preditor.

- ajuste: estimação dos parâmetros de suas covariáveis;
- validação: a verificação quanto ao atendimento de pressupostos estabelecidos pela teoria da formulação.

Na escolha de um modelo, em detrimento de outros candidatos plausíveis, busca-se o melhor equilíbrio possível entre a qualidade do ajustamento aos dados, medida por alguma métrica, a parcimônia (a "navalha" de William de Ockham [1]) e a facilidade de sua interpretação (TURKMAN; SILVA, 2000).

Modelos excessivamente (ou deficitariamente) ajustados aos dados amostrais (*overfitted* ou *underfitted*, respectivamente) devem ser evitados por não terem boa capacidade preditiva (ou explicativa). A variabilidade das estimativas dos parâmetros e a presença de viés afetam os erros associados às estimativas produzidas com esses modelos. Assim modelos "sobreajustados" embora apresentem um reduzido viés, possuem uma grande variabilidade ao passo que modelos subajustados possuem reduzida variabilidade, mas um grande viés (STASINOPOULOS et al., 2017).

Modelos com propósitos preditivos requerem ainda uma criteriosa análise de seus resíduos para que os resultados obtidos com sua aplicação possam ser considerados estatisticamente válidos (TURKMAN; SILVA, 2000).

3.2 Modelo Clássico de Regressão Linear (*LM*)

Denomina-se por regressão um amplo e crescente leque de técnicas estatísticas pelas quais se estabelece a relação, e é importante salientar a distinção existente entre relação funcional e causal, entre uma variável resposta e uma ou mais variáveis explicativas: $Y = f(X_1, X_2, \ldots, X_k)$.

Houve um momento em que Johann Carl Friedrich Gauss considerou sua descoberta da regressão estatística como "trivial". O método era até certo ponto tão óbvio que não lhe parecia poder ter sido o primeiro a usá-lo e, por essa razão, não declarou publicamente sua descoberta, até muitos anos depois (GAUSS, 1857) quando seu contemporâneo Adrien-Marie Legendre publicou sobre o mesmo método. Neste momento, quando Gauss sugeriu que ele havia usado antes de Legendre deu-se início a uma das mais famosas disputas na história da ciência sobre antecedência em uma descoberta. Gauss acabou por receber a maior parte do crédito como fundador da regressão.

Em 1886, *Sir* Francis Galton propôs no artigo *Family Likeness in Stature* expressar por uma função, a relação entre estaturas de pais, seus filhos e descendentes.

[1] William de Ockham foi um frade franciscano inglês, filósofo escolástico e teólogo, que se acredita ter nascido em 1285 na aldeia de Ockham na Inglaterra (HAMILTON, 1853 apud PEARSON, 1892). Embora a expressão "navalha" seja associada ao seu nome, o próprio Ockhan afirma ser aristotélico o princípio filosófico de que a pluralidade nunca deve ser postulada sem necessidade.

Nesse artigo, Galton verificou que embora houvesse uma tendência de que pais mais altos tivessem filhos altos e, pais mais baixos, filhos baixos, a estatura média de crianças nascidas de pais com dada altura tendiam a regredir à altura média da população na totalidade. Dessa maneira surge, pela primeira vez, a expressão regressão no contexto estatístico: "É uma regra universal que um descendente desconhecido em qualquer grau de um determinado homem é mais medíocre que ele" (GALTON; DICKSON, 1886).

A *Lei da Regressão* de Galton foi referendada por Karl Pearson (*On the Laws of Inheritance*, 1903) poucos anos depois quando ao analisar os dados de milhares de registros de estatura, tamanho do antebraço e do palmo.

Embora os pais altos tendam a ter filhos altos, a altura média dos filhos de um grupo de pais altos é menor do que a altura do pai. Há uma regressão da estatura do filho em direção à média altura de todos os homens. Nas palavras de Galton isso seria uma *regressão à mediocridade*.

Em *Correlations and their Measurement, chiefly from Anthropometric Data*, apresentado à *Royal Society of London* em dezembro de 1888, ele observou aquilo que conceituaria como *co-relação* ou *correlação de estrutura*:

> "Assim, é dito que o comprimento do braço está correlacionado com o da perna, porque uma pessoa com um braço longo geralmente tem uma perna longa, e vice-versa. Se a co-relação for próxima, então uma pessoa com um braço muito longo normalmente tem uma perna muito longa; se fosse uma correlação moderada, então o comprimento de sua perna normalmente seria apenas longo, não muito longo; e se não houvesse nenhuma correlação, o comprimento de sua perna seria, em média, medíocre" (GALTON; DICKSON, 1889).

Em latim o prefixo "*co*" equivale a "*com, cum*" e esse último tem o significado de colaboração com, união com, em simultâneo. Correlação significa, portanto, uma relação mútua entre dois termos, uma correspondência.

3.2.1 Modelos Hedônicos de Regressão

Em razão de sua elevada heterogeneidade, um bem imobiliário apresenta-se ao consumidor como um pacote de incontáveis atributos qualitativos e quantitativos que, reunidos, respondem por seu valor.

Em um processo de compra o interessado pondera de modo intuitivo, segundo suas convicções e necessidades, e promove uma hierarquização desses atributos para facilitar a comparação entre dois bens assemelhados e a subsequente tomada de decisão. Quanto maior seu entendimento sobre o mercado e o bem, mais detalhada será esta decomposição. O que o interessado está a fazer, nesse momento, nada mais é do que ponderar, atribuir pesos relativos a cada um dos atributos que escolheu como mais relevantes à formação do valor, conforme seu interesse ou conveniência.

A teoria das funções hedônicas de valor fornece uma estrutura para a análise de produtos diferenciados como bens imobiliários, cujas características individuais não são observáveis nos dados de mercado. Um bem diferenciado pode ser representado como um vetor de características tendo seu valor como variável dependente. De modo implícito, o valor é expresso por uma função hedônica que relaciona às características (PAGOURTZI et al., 2003). Os modelos de regressão linear construídos nessa base teórica são denominados Modelos Hedônicos de Regressão e essa abordagem tem seu uso pioneiro atribuído a Andrew T. Court (COURT, 1939).

No campo da engenharia de avaliações os Modelos Hedônicos de Regressão ganharam destaque na avaliação de preços de residências com diversos estudos dentre os quais os de Ridker e Henning (1967) e King e Mieszkowski (1973) pela análise da relação dos valores com o grau de poluição e composição social da região, respectivamente.

Um Modelo Hedônico de Regressão para a avaliação imobiliária é uma estratégia de investigação baseada na premissa de que um bem heterogêneo pode ser compreendido em termos de um menor número de atributos ou componentes intrínsecos e extrínsecos (GOODMAN, 1978), relacionados à satisfação que auferem ao interessado, utilizados como variáveis para explicar seu valor (DUBIN, 1988).

Alguns atributos são de natureza quantitativa e diretamente associados a características mensuráveis do bem como áreas, distâncias ou quantidades; outros podem possuir uma natureza dicotômica ao se resumirem à presença ou ausência de uma determinada característica no bem (DUBIN, 1992); todavia, há também atributos qualitativos que podem ser indiretamente relacionados a outras características quantitativas do bem que sejam passíveis de medição direta como, por exemplo, o padrão construtivo que pode ser associado ao custo de construção, o estado geral de conservação ao grau de depreciação ou o relevo ao grau de inclinação do terreno e, por isso, denominadas como variáveis *proxy*. O uso de um grande número de variáveis demanda o estudo de suas inter-relações e, não raramente, resultam em algum grau de multicolinearidade.

A construção de um modelo hedônico de formação de valor requer uma estrutura de variáveis coerentes com o que se deseja explicar e demanda uma análise estatística tão criteriosa quanto à realizada em um Modelo Clássico de Regressão Linear para a confirmação do atendimento às hipóteses teóricas subjacentes para só então se decidir por sua aceitação e poder admitir como estatisticamente significantes e não viesadas as conclusões dele derivadas.

3.3 MODELOS LINEARES GENERALIZADOS (*GLM*)

Modelos clássicos de regressão linear com distribuição Normal para o erro são de aplicabilidade limitada, pois se restringem àqueles fenômenos nos quais a variável resposta segue, consequentemente, uma distribuição Normal e, não raras vezes, na etapa diagnóstica

são detectados afastamentos das premissas admitidas como válidas na formulação teórica do modelo.

Em 1972 John Nelder e Robert Wedderburn (NELDER; WEDDERBURN, 1972) apresentam uma nova classe de modelos estatísticos que generaliza o modelo linear clássico e inclui muitos outros modelos úteis em análise estatística como os modelos log-lineares para a análise de dados na forma de contagens; "probit" (*probabilit unit*) e "logit" (*logistic unit*) para dados na forma de proporções; modelos para dados contínuos com erro padrão proporcional constante e os modelos para dados de sobrevivência.

A classe proposta decorre de desenvolvimentos que perpassam dois séculos (NELDER; MCCULLAGH, 1989 apud LINDSEY, 1997) dos modelos clássicos lineares como a variável Normal com ligação identidade (Legendre e Gauss no começo do séc. XIX); a análise de variância no delineamento experimental (Ronald A. Fisher nos anos de 1920 a 1935); a função de verossimilhança (Ronald A. Fisher em 1922); os ensaios de diluição: variável binomial com ligação complemento "log-log" (Ronald A. Fisher, 1922); a família exponencial de distribuições (Ronald A. Fisher, 1934); a análise de variáveis binomiais sob ligação "probit" (Bliss, 1935); a análise de proporções sob ligação "logit" (Berkson, 1944; Dyke and Patterson, 1952); a análise de itens: distribuição de Bernoulli sob ligação "logit" (Rasch,1960); análise de dados de contagem (Birch, 1963); as superfícies de resposta com polinômios (Nelder, 1966) e a análise de sobrevivência (Feigl and Zelen, 1965; Zippin e Armitage, 1966; Glasser, 1967).

Um aspecto importante na generalização proposta é a presença, em todos os modelos, de um *preditor linear*: uma combinação linear das covariáveis explicativas da resposta. A existência de um preditor linear possibilita estender, a esta nova classe de modelos, conceitos de regressão linear clássica e modelos de análise de variância, na medida em que se referem às estimativas de parâmetros em um preditor linear (NELDER; MCCULLAGH, 1989). De modo adicional, esta nova classe também propõe uma maior flexibilidade na relação entre a resposta e o preditor pela adoção de uma função de ligação.

Os Modelos Lineares Generalizados tratam, pois, de duas novas situações de natureza menos restritiva quando comparadas ao Modelo Clássico de Regressão Linear: a resposta pode proceder de outras distribuições que não a Normal e a relação entre a resposta e as variáveis independentes não necessitam ser identicamente lineares.

Um importante conceito unificador, introduzido nesta nova classe de modelos, é o da família exponencial de distribuições, uma classe de distribuições na qual, dentre outras, incluem-se as distribuições Normal, Binomial, Binomial Negativa, Gama, Poisson, Normal inversa, Multinomial, Beta, Logarítmica, família estudada de modo pioneiro e independente por Bernard Koopman (1936), Jim Pitman (1936) e Georges Darmois (1935) (BROWN, 1986).

3.4 FUNÇÕES DE SUAVIZAÇÃO PARA MÉTODOS NÃO PARAMÉTRICOS

A análise de regressão é uma técnica estatística amplamente utilizada para predizer uma resposta a partir de um conjunto de variáveis explicativas que pode ser categorizada em três grandes grupos: paramétricas; não paramétricas e semiparamétricas.

Modelos Lineares, Modelos Lineares Generalizados e Modelos não Lineares são exemplos de regressões paramétricas, pois a função que descreve o relacionamento entre a resposta e as covariáveis é conhecida. Todavia, uma característica bastante comum em dados reais é a complexidade nas relações entre a resposta e as covariáveis e, em muitas dessas situações, o comportamento dos preditores não segue um modelo probabilístico suficientemente flexível para descrever a relação (PRICE, 2018).

Uma variável aleatória X segue um determinado modelo paramétrico F_θ se sua função de distribuição de probabilidade pertence a alguma família de distribuições com vetor de parâmetros $\boldsymbol{\theta}$ de dimensão finita k tal que

$$X \sim F \in \mathcal{F}_\Theta = \{F_\theta : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathcal{R}^k\},$$

em que $\boldsymbol{\Theta}$ é o espaço paramétrico *k-dimensional*.

Métodos não paramétricos formam um amplo conjunto de técnicas distintas orientadas aos dados (*data-driven*) e diferem dos paramétricos (*model-driven*) no aspecto de que a forma da relação funcional entre a resposta e as covariáveis não é predeterminada e pode ser ajustada de modo a capturar flutuações inesperadas dos dados. A média da resposta é modelada como uma função suavizadora não específica de covariáveis.

Os métodos semiparamétricos combinam modelos paramétricos e não paramétricos e são utilizados nas situações que um modelo totalmente não paramétrico pode não funcionar bem devido a sua sensibilidade a *outliers* ou quando um modelo paramétrico é conhecido, mas a distribuição dos erros não. Esta é uma configuração particularmente útil ao trabalhar com problemas científicos complexos.

Não obstante a importância dos métodos paramétricos, observou-se nos anos de 1980 um ressurgimento do interesse pelas chamadas técnicas de suavização da dispersão (GREEN; SILVERMAN, 1994).

Suavizadores de dispersão (*scatterplot smoothers*) desempenham um importante papel na análise estatística ao possibilitar expressar as tendências da relação entre uma variável resposta e uma ou mais variáveis explicativas sem, para isso, assumir alguma forma funcional paramétrica nessa dependência (STASINOPOULOS et al., 2017).

Considere n observações de uma variável resposta $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ e uma única variável explanatória $\boldsymbol{X} = (x_1, \ldots, x_n)^T$. Um suavizador de dispersão univariado pode ser expresso formalmente como um modelo estatístico (WOOD, 2017)

$$f(x) = \mathbb{E}(Y|x)$$

ou, de modo equivalente, como

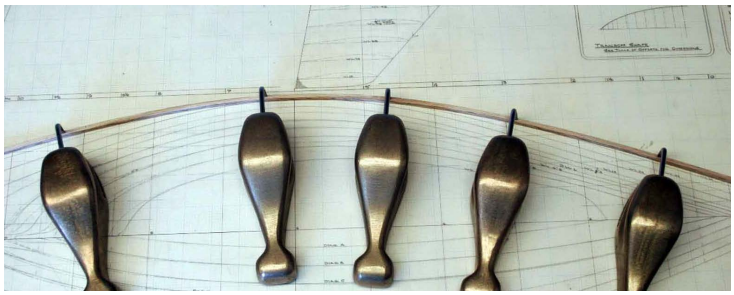
$$y_i = f(x_i) + \varepsilon_i \quad i = 1, \ldots, n.$$

A maioria dos métodos não paramétricos assume $f(\cdot)$ como o componente sistemático, uma função arbitrária que se assume existir e que possua, em algumas situações, propriedades conhecidas como continuidade e suavidade e ε_i como o *i-ésimo* componente aleatório tal que $\varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$ (WOOD, 2017), podendo ser essa última premissa relaxada (REGUERA, 2021).

3.4.1 Splines

O termo *spline* remete ao nome dado a dispositivos semirrígidos utilizados para auxiliar o desenho de curvas, uma barra flexível delgada feita de madeira ou algum outro material elástico, colocado na folha de papel milimetrado e mantido no lugar em vários pontos por meio de certos objetos pesados chamados “cães” ou “ratos”, como para assumir a forma da curva que desejamos desenhar (SCHOENBERG, 1945), como ilustrado na Figura 3.2.

Figura 3.2: Um conjunto de *spline* e *ducks*



Fonte: Imagem pública disponível na *WWW*

Um *spline* é, essencialmente, definido como sendo uma função construída por funções polinomiais em partes às quais impõe certas restrições nos pontos de união dessas (nós) com o propósito de lhe assegurar um padrão suave. Assim, uma função real $f(x)$ definida para todo $x \in \mathcal{R}$ é um *spline* de grau *k* e denotada por $\Pi_k(x)$ se ela apresenta as seguintes propriedades (SCHOENBERG, 1945)

- é composta por “arcos polinomiais” na maior parte de grau $k - 1$;
- $f(x)$ é uma função da classe $C^{(k-2)}$; *i.e.*, possui $k - 2$ derivadas contínuas;
- os únicos possíveis pontos de junção desses arcos são os pontos inteiros $x = n$ (se k for par) ou $x = {}^{(n+1)}/_2$ (se k for ímpar).

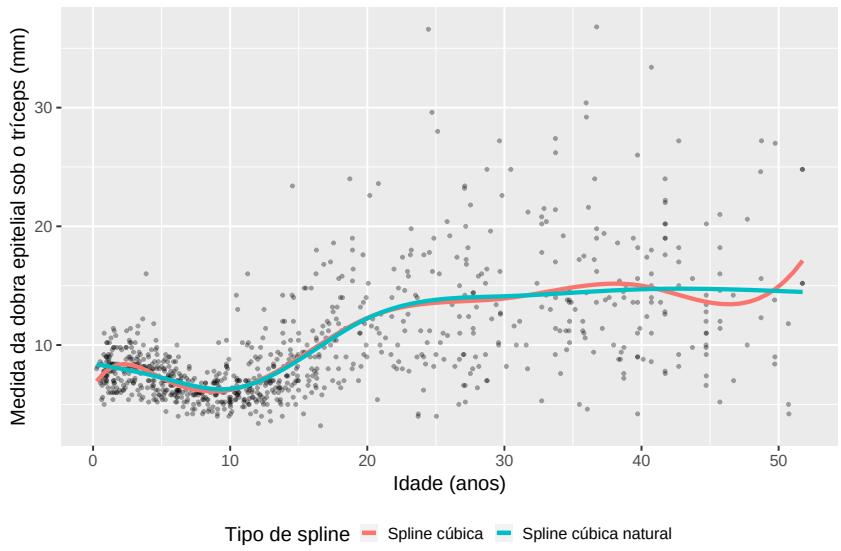
A mais comum das aproximações polinomiais por partes é a que se vale de polinômios cúbicos estimados entre cada par de nós sucessivos: o *spline* cúbico. Dada a função f definida em $[a, b]$ e em um conjunto de nós $a < t_1 < t_2 < \cdots < t_n < b$, um *spline* S para f é uma função que satisfaz às seguintes condições (BURDEN; FAIRES; BURDEN, 2017):

- $S(x)$ é um polinômio cúbico, denotado $S_j(x)$, no subintervalo $[x_j, x_{j+1}]$ para cada $j = 0, 1, \cdots, (n-1)$;
- $S_j(x_j) = f(x_j)$ e $S_j(x_{j+1}) = f(x_{j+1})$ para cada $j = 0, 1, \cdots, (n-1)$ (spline interpolador);
- $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ para cada $j = 0, 1, \cdots, (n-2)$ (continuidade nos nós);
- $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ para cada $j = 0, 1, \cdots, (n-2)$ (continuidade da inclinação);
- $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ para cada $j = 0, 1, \cdots, (n-2)$ (continuidade da curvatura).

Impondo-se condições de contorno naturais: $S''(x_0) = S''(x_n) = 0$, o *spline* cúbico será denominado natural e a curva final tem a forma de uma reta depois que passa pelos pontos de interpolação mais próximos de suas extremidades. Tal imposição resguarda o *spline* de oscilações bruscas nos limites do intervalo.

Na Figura 3.3 expomos o ajuste de dados de um estudo antropométrico realizado com 892 mulheres com menos de 50 anos em 3 aldeias do Gabão, na África (disponíveis na biblioteca *MultiKink* (WAN; ZHONG, 2020)), na forma desses dois tipos de *splines*.

Figura 3.3: Ajuste de *splines* cúbicos (nós arbitrados em: 5, 10, 20, 30 e 40)



Fonte: Próprio autor

Observa-se que os *splines* cúbicos dependem dos nós e assim, algum critério para a escolha da quantidade e sua localização é necessário.

Reguera (2021) afirma ser mais importante a escolha da quantidade que a posição dos nós e sugere, como escolha inicial, que a quantidade de nós t seja tal que $3 < t < 7$. Pode-se também determinar a quantidade de nós pela comparação dos ajustes efetuados sob diferentes quantidades de k usando-se algum critério, como o AIC. Quanto à posição, de modo usual, adota-se os postos de medidas separatrizes.

Seja $S[a, b]$ o espaço de todas as funções f definidas em $[a, b]$ que possuem duas derivadas contínuas. Reinsch (1967) demonstra que um *spline* cúbico natural é a mais suave das funções $f \in S[a, b]$ que interpola os pontos tendo nós definidos em t_i.

Depreende-se por tal propriedade que *splines* cúbicos naturais possam constituir uma boa base para a geração de funções suavizadoras para modelos estatísticos (WOOD, 2017) e uma das bases que podem ser usadas para se gerar um *spline* cúbico pode ser expressa em termos de polinômios truncados

$$f(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \sum_{t=1}^{T} \beta_k (x - b_t)^3 H(x - b_t),$$

em que $H(x - b_t)$ é a função degrau (função de Heaviside) dada por

$$H(x) = \begin{cases} 1 & \text{se } x \geq 0, \\ 0 & \text{caso contrário.} \end{cases}$$

Expressar uma função suavizadora desconhecida na forma de uma combinação linear de funções conhecidas que formem uma base $f(x) = \beta_1 b_1(x) + \cdots + \beta_p b_p(x)$ possibilita a estimação dos coeficientes por meio da resolução de um sistema linear

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n,1)} = \begin{bmatrix} b_1(x_1) & b_2(x_1) & \dots & b_p(x_1) \\ b_1(x_2) & b_2(x_2) & \dots & b_p(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_1(x_n) & b_2(x_n) & \dots & b_p(x_n) \end{bmatrix}_{(n,p)} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p,1)} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(p,1)}$$

A estimação do vetor $\boldsymbol{\beta}$ passa pela minimização da soma dos quadrados dos resíduos (SQR):

$$\text{SQR} = \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{p} \beta_p b_p(x_i) \right]^2 \text{ ou}$$

$$\hat{\boldsymbol{\beta}} = arg\, min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\beta}||^2,$$

em que $\boldsymbol{B} = [b_{ij}]$ é denominada como matriz das bases e a estimativa do vetor $\hat{\boldsymbol{\beta}}$ reside, essencialmente, na solução de um problema de regressão linear dada por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{y}.$$

3.4.2 *B-splines*

O termo *B-spline*, uma abreviação para *splines* de base (*Basis splines*), foi proposto por Schoenberg (1945) e refere-se a uma classe de funções estritamente locais definidas por sua ordem m e número de nós internos N (mais dois nos extremos: nós finais).

Bases formadas por *B-splines* fornecem uma alternativa numericamente superior às bases de potências truncadas, visto que suas funções assumem valores diferentes de zero apenas em um reduzido número de nós. De modo usual, a referência que mais se verifica é a de *B-splines* de ordem *m* como aqueles compostos por polinômios de grau $(m-1)$ e assim, uma base de *B-splines* composta por *splines* cúbicos é de ordem $m = 4$ (RACINE, 2019).

No desenvolvimento subsequente do tema, vários autores propuseram *B-splines* como base conveniente para problemas de interpolação e suavização por *splines* com nós fixos (COX, 1972) por estarem definidas apenas em intervalos específicos: cada função da base é não nula apenas sobre um intervalo compreendido entre $(m+3)$ nós adjacentes.

O perfil de suavização resultante está diretamente relacionado com o arbitramento do grau, do número de pontos de controle (parâmetros) e seu posicionamento. Não é imprescindível que os pontos de controle sejam equidistantes como em *B-splines* uniformes (STASINOPOULOS et al., 2017); não obstante Yu e Ruppert (2002) recomendem que assim sejam e fiquem posicionados nos quantis da variável preditora. Vanegas e Paula (2015) mencionam um valor padrão (*default*) de uma rotina computacional associada aos *splines* penalizados (*P-splines*): $n^{1/3} + 3$.

Seja $\boldsymbol{t} = \{t_i | i \in \mathbb{Z}\}$ uma sequência não decrescente de números ($t_i \leq t_{i+1}$) tais que

$$t_0 \leq t_1 \leq \cdots \leq t_N \leq t_{N+1}.$$

Pela natureza recursiva de sua estimação definida por De Boor (2001), estabeleça um conjunto ampliado de nós na forma

$$t_{-(m-1)} = \cdots = t_0 \leq t_1 \leq \cdots \leq t_N \leq t_{N+1} = \cdots = t_{N+m}$$

pela anexação de $n = m - 1$ vezes os nós limites inferior e superior.

Reindexando-se para o primeiro elemento do conjunto ampliado de nós de modo que os $N + 2m$ nós t_i agora passem a ser indexados por $i = 0, \ldots, N + 2m - 1$, para cada um dos nós ampliados t_i pode ser definido um conjunto de funções reais $B_{i,j}$ com $j = 0, 1, \ldots, n$ (sendo n o grau das bases *B-spline*), da forma (RACINE, 2019)

$$B_{i,j+1}(x) = \alpha_{i,j+1}(x)B_{i,j}(x) + [1 - \alpha_{i+1,j+1}(x)]\, B_{i+1,j}(x),$$

com

$$B_{i,0}(x) = \begin{cases} 1 & \text{se } t_i \leq x \leq t_{i+1}, \\ 0 & \text{caso contrário.} \end{cases}$$

e onde

$$\alpha_{i,j}(x) = \begin{cases} \frac{x-ti}{t_{i+j}-t_i} & \text{se } t_{i+j} \neq t_i, \\ 0 & \text{caso contrário.} \end{cases}$$

Assim:

- especificamente para a forma recursiva acima define-se $^0/_0 = 0$;
- o vetor $\boldsymbol{t}$ é denominado vetor de nós e cada elemento seu é um nó;
- a função $B_{i,j}$ é a *i-ésima* função *B-splines* da base, de ordem j;
- para qualquer valor não negativo de j, o espaço vetorial $V_j(\boldsymbol{t}$ sobre $\mathbb{R}$, gerado pelo conjunto de todas as funções *B-splines* da base de ordem j é *B-spline* de ordem j; ou seja, *B-spline* $V_j(\boldsymbol{t}) = span\{B_{i,j}(x)|i = 0, 1, \dots\}$ sobre $\mathbb{R}$;
- qualquer elemento de $V_j(\boldsymbol{t})$ é um *B-spline* de ordem j;
- K é a dimensão da base do *B-spline* ($K = N + m$).

A Figura 3.4 apresenta um *B-spline* de ordem $m = 4$ composta por 7 *splines* cúbicos denotados como $B_{0,3}$; $B_{1,3}$; $B_{2,3}$; $B_{3,3}$; $B_{4,3}$; $B_{5,3}$; $B_{6,3}$ estimado sobre o vetor ampliado de nós $\boldsymbol{t} = (0; 0; 0; 0, 25; 0, 50; 0, 75; 1; 1; 1)$ tendo três nós internos fixados em 0,25, 0,50 e 0,75 mais dois nós extremos em 0 e 1.

Um *B-spline* de ordem $m = n + 1$ é a curva paramétrica composta pela combinação linear das bases *B-splines* $B_{i,j}(x)$ de grau n dada por

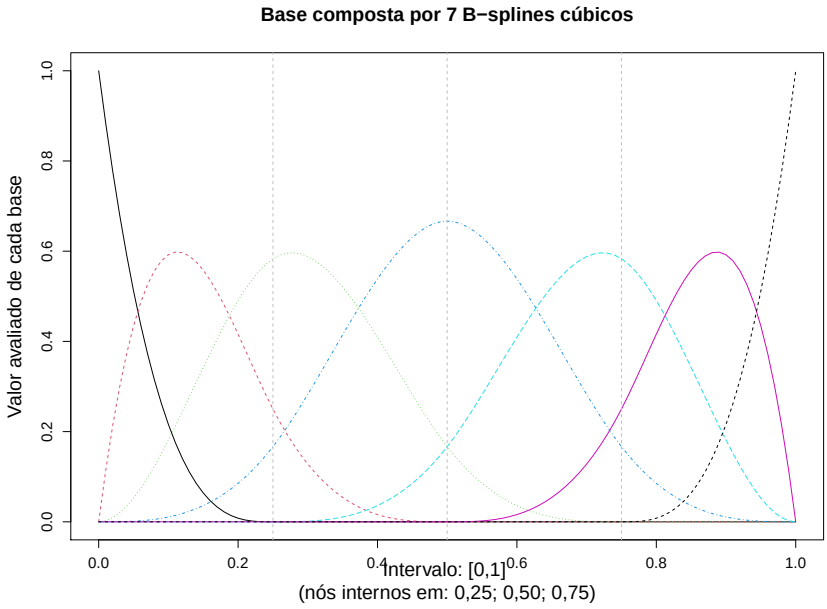
$$f(x) = \sum_{i=0}^{N+m} \beta_i B_{i,n}(x), \quad x \in [t_0, t_{N+1}], \tag{3.2}$$

A estimação do vetor $\boldsymbol{\beta} = \beta_1, \dots, \beta_{N+m}$ passa pela minimização da função:

$$\text{SQR} = \sum_{j=1}^{n} \left[y_j - \sum_{i=1}^{N+m} \beta_i B_{i,n}(x_j) \right]^2 = ||\boldsymbol{y} - \boldsymbol{\beta B}||^2, \tag{3.3}$$

em que $\boldsymbol{B}_{(N+m,j)}$ é denominada matriz das funções da base e cujos elementos são os valores avaliados dessas funções em cada observação e assim, a solução da Equação (3.3) reside,

Figura 3.4: Base *B-spline* de ordem $m = 4$



Fonte: Próprio autor

essencialmente, em um problema de regressão linear e a estimativa de mínimos quadrados para $\boldsymbol{\beta}$ é dada por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{y}.$$

Funções assim estabelecidas constituem bases para se gerar o espaço vetorial de funções não conhecidas *a priori*. O número de funções da base determina o quão exata será a aproximação realizada pelo *spline*; todavia, a adoção de uma base de ordem mais elevada do que o necessário resultará em *overfitting* dos dados e não somente a tendência geral estará sendo modelada por $f(x)$ mas também os erros.

3.4.3 *Splines* penalizados (*P-splines*)

O uso de *B-splines* como base na estimação de um *spline* suavizador requer que alguns parâmetros sejam arbitrados de modo preliminar, como o grau da função que se deseja avaliar; a quantidade de nós e seu posicionamento.

Uma maneira informal de se olhar para os *splines* penalizados é dar o papel principal aos parâmetros, com os *B-splines* meramente adicionando a "carne aos ossos" (EILERS; MARX; REGUERA, 2015). Esses parâmetros de ajuste têm um impacto determinante no perfil de suavização do *spline* resultante: um grande número de nós implica em alta flexibilidade do *spline* mas, por vezes, muito além daquela necessária para modelar a tendência da resposta e assim resultar no sobre-ajustamento (*overfitting*) aos dados. Equivale

dizer que tanto a tendência quanto os erros serão suavizados pela Equação (3.2); por outro lado, um reduzido número de nós pode resultar em uma estimativa imprecisa e enviesada.

Realizar esse controle pela imposição de uma penalidade na minimização da soma dos quadrados dos resíduos (SQR) requer que alguma medida de aspereza seja utilizada como padrão da cuva da função final estimada e o quadrado integrado da segunda derivada tem sido a mais usada (HANDSCOMB, 1966) e (REINSCH, 1967 apud EILERS; MARX, 1996).

A imposição de uma penalização evita os problemas relacionados à seleção de nós a partir de um conjunto maximal de nós com a complexidade do ajuste sendo regulada pelo parâmetro de suavização. Assim, deve-se buscar, dentre todas as possíveis funções com duas derivadas contínuas, uma que minimize a soma penalizada dos quadrados dos resíduos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

$$\text{SQR}(f, \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int \left[f''(x)\right]^2 \, dx. \tag{3.4}$$

A primeira estrutura em (3.4), quadrado médio dos resíduos, é a medida da proximidade do ajuste aos dados pela função de suavização $f(\cdot)$ enquanto a outra estrutura apresenta a segunda derivada da função f em relação a x como a medida pontual da curvatura da função, elevada ao quadrado posto não ser relevante se é côncava ou convexa, integrada por toda a faixa de valores de x para expressar quão curva é no intervalo, multiplicada pelo parâmetro de suavização $\lambda > 0$.

Essa segunda estrutura, ao ser adicionada à primeira, está aplicando uma penalização à SQR baseada na curvatura da função e assim, se $\lambda \to 0$, a função f pode ser qualquer uma que interpole os dados, um *spline* interpolador que irá passar por todos os valores x_i. Se $\lambda \to \infty$, a função será idêntica àquela de uma regressão linear estimada por mínimos quadrados, como ilustrado na Figura 3.5.

Para quaisquer valores intermediários do parâmetro de suavização λ, f será uma função que se mostra equilibrada entre aproximar todos os pontos tendo uma baixa curvatura.

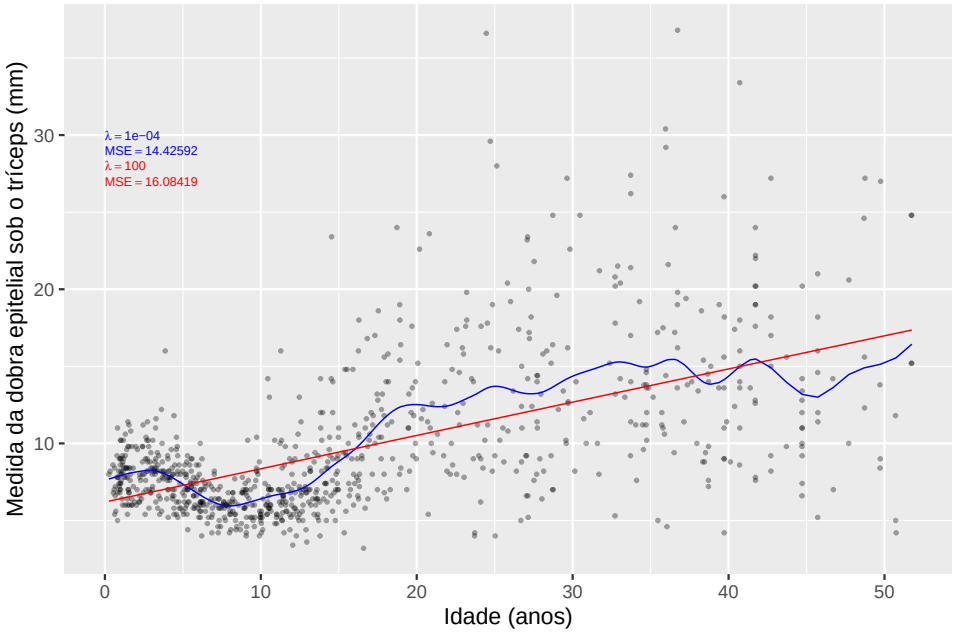
Segundo Hastie, Tibshirani e Friedman (2009), pode-se demonstrar que o critério estabelecido em (3.4) tem como único e específico minimizador de dimensão finita, um *spline* cúbico com nós em cada valor único $x_i, \quad i = 1, \ldots, n$. Para mais detalhes, ver em Green e Silverman (1994).

Assim, sendo $f(\cdot)$ em (3.4) um *spline* cúbico, poderá ser expresso na forma

$$f(x) = \sum_{j=1}^{n} B_j(x)\beta_j,$$

em que $B_j(x)$ são um conjunto *n-dimensional* de funções de base *B-splines* cúbicos da família dos *splines* naturais.

Figura 3.5: Efeito do parâmetro de suavização λ: equilíbrio entre *wiggliness* e *smoothness* da função de suavização estimada



Fonte: Próprio autor

P-splines (*penalized splines*) propostos por Eilers e Marx (1996) contornam a dificuldade no cálculo ao usar como penalização as diferenças finitas de ordem superior entre os coeficientes de *B-splines* adjacentes como, por exemplo, para ordem 2:

$$\begin{bmatrix} \beta_1 - 2\beta_2 + \beta_3 \\ \beta_2 - 2\beta_3 + \beta_4 \\ \beta_3 - 2\beta_4 + \beta_5 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & -2 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & -2 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & -2 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta \\ \beta_5 \\ \vdots \end{bmatrix}$$

Essa abordagem reduz a dimensionalidade do problema, anteriormente da ordem do número de observações, para o número de *B-splines*, apresentando mais estabilidade numérica e maior facilidade de implementação e os coeficientes da função pretendida podem ser obtidos pela resolução do sistema matricial

$$||\boldsymbol{y} - X\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{\beta},$$

em que $\boldsymbol{S}$ é uma matriz de diferenças de ordem (geralmente 2).

Escrevendo-se $\boldsymbol{S} = \boldsymbol{D}^T \boldsymbol{D}$ e denotando-a como matriz de penalização, a estimação do vetor $\boldsymbol{\beta}$ passa pela minimização da função:

$$\hat{\boldsymbol{\beta}} = arg\, min_{\boldsymbol{\beta}|\lambda} \left[||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} \right],$$

cuja solução reside em um problema de regressão linear:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{S})^{-1} \boldsymbol{X}^T \boldsymbol{y},$$

para um valor fixo de lambda.

Uma das maneiras de se determinar o melhor valor para o parâmetro de suavização λ é pela validação cruzada (*leave one out*) dos resultados das sucessivas soma de quadrados dos resíduos para diferentes valores desse parâmetro (GOLUB; HEATH; WAHBA, 1979).

Nesse procedimento deixa-se, a cada ciclo, um ponto (x_i, y_i) fora do conjunto e avalia-se seu valor pelo *spline* ajustado ao conjunto de observações restantes e calcula-se assim a soma dos quadrados médios dos resíduos sob um valor inicial arbitrado para λ.

Repetindo-se para outros valores de λ o valor ideal será o que apresentar a menor média da soma de quadrados de resíduos penalizados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

$$CV_\lambda = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \hat{f}_\lambda^{-i}(x_i) \right]^2.$$

Nos *P-splines* os nós estão posicionados de modo equidistante, não obstante alguns autores recomendem posicioná-los nos quantis da variável. Quanto à quantidade, Reguera (2021) cita como regra comum $min[40, (\frac{\text{valores únicos de x}}{4})]$.

3.4.4 *Thin plate splines*

Várias áreas do conhecimento como demandam análise e modelagem estatística da variabilidade e incerteza de dados coletados sobre um domínio espacialmente delimitado e os valores observados da variável em estudo encontram-se georreferenciados. Como exemplo, temos as citadas por De Bastiani (2016):

- ciências ambientais: geologia (teores de minerais componentes do solo); hidrologia; climatologia; ecologia; oceanografia ou agronomia;
- ciências sócio-econômicas: geografia humana; econometria ou planejamento urbano;
- ciências médicas: epidemiologia.

Thin plate splines foram propostos como solução geral para modelagem de relações complexas entre preditores contínuos e a variável resposta, situações nas quais se deseja estimar uma função suavizadora baseada em múltiplas variáveis preditoras, uma

generalização para espaços vetoriais gerados por dois ou mais *splines* (DUCHON, 1977 apud WOOD, 2017).

Sejam $z_1, \ldots, z_n$ observações da altura de uma superfície tomadas nos pontos $t_1(x_1, y_1), \ldots, t_n(x_n, y_n)$, em que $t_1, \ldots, t_n$ são pontos em $\mathcal{R}^2$. Um modelo observacional para esse fenômeno pode ser expresso na forma:

$$z_i = g(\boldsymbol{t}_i) + \varepsilon_i \quad 1 \leq i \leq n,$$

em que z_i é o valor observado na *i-ésima* localização, ε_i são componentes aleatórios usualmente associados a erros de medição admitidos não correlacionados e g é a função de interesse que retorna as estimativas z_i.

Thin plate splines são construídos tendo funções radiais como base, as quais são funções de valores reais $\zeta : [0, \infty] \to \mathcal{R}$ definidas em termos da distância a um certo ponto

$$\begin{aligned} t(x, y) &= \delta\zeta\left(\sqrt{(x - x_c)^2 + (y - y_c)^2}\right) \\ &= \delta\zeta(||\boldsymbol{t} - \boldsymbol{c}||) \\ &= \delta\zeta(r), \end{aligned}$$

em que δ é um ponderador, $c = (x_c, y_c)$ são as coordenadas do centro e $r = ||\boldsymbol{t} - \boldsymbol{c}||$ a distância euclidiana de qualquer ponto em $\mathcal{R}^2$ a c.

Assim, os valores de uma função radial dependem apenas de uma distância entre o valor de entrada e algum valor fixo de referência para operarem a transformação e, por esta razão, apresentam a propriedade de serem invariantes a todas as transformações euclidianas (translações, rotações e projeções).

De particular interesse para os *thin plate splines* são as funções radiais da família dos *splines* poliharmônicos

$$\zeta(r) = r^2 \log(r).$$

Quando aplicadas a um espaço vetorial $||.|| : V \to [0, \infty]$ a função $\zeta = \zeta(||\boldsymbol{t} - \boldsymbol{c}||)$ é denominada como um *kernel* radial centrado em c.

Uma função radial e seu *kernel* radial são denominadas como funções radiais de base se, para qualquer conjunto de nós $\{t_k\}_{k=1}^n$, os *kernels* $\zeta(t_1), \zeta(t_2), \ldots, \zeta(t_n)$ forem linearmente independentes e a matriz definida por

$$\begin{bmatrix} \zeta(||t_1 - t_1||) & \zeta(||t_2 - t_1||) & \ldots & \zeta(||t_n - t_1||) \\ \zeta(||t_1 - t_2||) & \zeta(||t_2 - t_2||) & \ldots & \zeta(||t_n - t_2||) \\ \vdots & \vdots & \ddots & \vdots \\ \zeta(||t_1 - t_n||) & \zeta(||t_2 - t_n||) & \ldots & \zeta(||t_n - t_n||) \end{bmatrix}_{(n,n)}$$

for não singular.

Para uma definição formal de *thin plate splines* admita uma função radial $\zeta(r)$ na forma (GREEN; SILVERMAN, 1994)

$$\begin{aligned} \zeta(r) &= \frac{1}{16\pi} r^2 \log r^2 \quad r > 0; \\ \zeta(r) &= 0 \qquad\qquad\qquad\quad r = 0 \end{aligned} \tag{3.5}$$

e considere ϕ_j como j funções que geram o espaço d-dimensional de todos os polinômios de grau menor que m e, dessa forma, uma base para representar qualquer ponto $t(x, y)$ em $\mathcal{R}^2$ pode ser composta por três funções $\phi_1(x, y) = 1,\ \phi_2(x, y) = x$ e $\phi_3(x, y) = y$ que permitem expressar qualquer ponto como uma combinação linear de seus valores.

Define-se a matriz $\boldsymbol{T}_{(3,n)}$ com elementos $t_{(j,k)} = \phi_j(t_k)$ em que t_k é o *k-ésimo* ponto com coordenadas (x_k, y_k) na forma

$$\boldsymbol{T} = \begin{bmatrix} \phi_1(t_1) & \phi_1(t_2) & \dots & \phi_1(t_n) \\ \phi_2(t_1) & \phi_2(t_2) & \dots & \phi_2(t_n) \\ \phi_3(t_1) & \phi_3(t_2) & \dots & \phi_3(t_n) \end{bmatrix}_{(3,n)} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_n \end{bmatrix}_{(2,n)}$$

e expressa-se a norma euclidiana do vetor **t** como $||\boldsymbol{t}||^2 = \boldsymbol{t}^T\boldsymbol{t}$.

Uma função *g(t)* é um *thin plate spline* sobre o conjunto de pontos $t_1(x_1, y_1), \dots, t_n(x_n, y_n)$ em $\mathcal{R}^2$ se, e somente se, for expressa na forma

$$g(\boldsymbol{t}) = \sum_{i=1}^{n} \delta_i \zeta\left(||\boldsymbol{t} - \boldsymbol{t}_i||\right) + \sum_{j=1}^{3} a_j \phi_j(\boldsymbol{t}), \tag{3.6}$$

em que $\phi_j(t)$ correspondem a funções que geram $\mathcal{R}^2$ e $\zeta(||\boldsymbol{t} - \boldsymbol{t}_i||)$ são as funções radiais avaliadas nas distâncias euclidianas entre um ponto qualquer t e os demais pontos observados t_i (pontos de controle), ponderadas por δ_i para modelar as não linearidades não ajustadas pelas funções ϕ.

Uma interpretação menos formal é que os coeficientes (a_1, a_2, a_3) definem o plano que melhor aproxima quaisquer pontos (x, y) dos pontos de controle (x_i, y_i) e a função radial aplicada às distâncias euclidianas $(||(x, y) - (x_i, y_i)||)$, ponderadas por δ_i, expressam o afastamento de um ponto qualquer a todos os pontos de controle.

Para a estimação dos vetores coluna $\boldsymbol{\delta} = \delta_1, \dots, \delta_n$ e $\boldsymbol{a} = a_1, a_2, a_3$, define-se a matriz $\boldsymbol{E}_{(n,n)}$ com elementos

$$E_{(i,j)} = \zeta(||\boldsymbol{t}_i - \boldsymbol{t}_j||), \tag{3.7}$$

e $E_{(i,i)} = 0$, tal como exposta em (3.5).

Impondo-se as igualdades

$$\sum_{i=1}^{n} \delta_i = 0;\ \sum_{i=1}^{n} \delta_i x_i \ = 0;\ \sum_{i=1}^{n} \delta_i y_i = 0, \tag{3.8}$$

equivalente a se escrever matricialmente $\boldsymbol{T}\,\boldsymbol{\delta} = \mathbf{0}$, segue-se (GREEN; SILVERMAN, 1994)

$$g(\boldsymbol{t}_i) = \sum_{j=1}^{n} \delta_j \zeta(||\boldsymbol{t} - \boldsymbol{t}_j||) + \sum_{k=1}^{3} a_k \phi_k(\boldsymbol{t_j}) = (\boldsymbol{E\delta} + \boldsymbol{T}^T \boldsymbol{a})_i,$$

equivalente ao sistema matricial

$$\begin{bmatrix} \zeta||t_1 - t_1|| & \dots & \zeta||t_1 - t_n|| & 1 & x_1 & y_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \zeta||t_n - t_1|| & \dots & \zeta||t_n - t_n|| & 1 & x_n & y_n \\ 1 & \ddots & 1 & 0 & 0 & 0 \\ x_1 & \dots & x_n & 0 & 0 & 0 \\ y_1 & \dots & y_n & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} z_1 \\ \vdots \\ z_n \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\left[\begin{array}{c|c} \boldsymbol{E} & \boldsymbol{T}^T \\ \hline \boldsymbol{T} & \mathbf{0} \end{array}\right] \cdot \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{o} \end{bmatrix}, \tag{3.9}$$

em que $E_{(i,j)} = \zeta(||(x_i, y_i) - (x_j, y_j)||)$, a *i-ésima* linha de $\boldsymbol{T}$ é $(1, x_i, y_i)$, $\mathbf{0}$ é uma matriz de zeros com dimensão 3×3, $\boldsymbol{o}$ é um vetor de zeros com dimensão 3×1 e $\boldsymbol{z}$ é um vetor coluna dos valores estimados $z_1, \dots, z_n$.

As condições impostas asseguram o crescimento quase linear do *thin plate spline* à medida que a norma cresça e também asseguram que o sistema linear a ser solucionado tenha número de equações igual ao número de incógnitas, independentemente do número de pontos de controle.

Para a estimação da mais suave das funções tal que $y_i = g(\boldsymbol{t}_i)$ para $i = 1, \dots, n$ torna-se necessário, antes, definir um funcional para penalização.

Para os suavizadores unidimensionais o funcional foi definido baseado na medida das oscilações da função suavizadora em termos do quadrado integrado de sua segunda derivada. Na situação bidimensional, por analogia, o conceito será o de quantificar o "empenamento" de uma superfície pela definição de um funcional $J_{[2]}(g)$.

O funcional $J_{[2]}(g)$ mede as ondulações de g (sua curvatura no espaço) e escrito em termos das coordenadas de um ponto em $\mathcal{R}^2$ pode ser expresso como (GREEN; SILVERMAN, 1994):

$$J_{[2]}(g) = \int_{\mathcal{R}} \int_{\mathcal{R}} \left[\left(\frac{\partial^2 g}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial x \partial y} \right) + \left(\frac{\partial^2 g}{\partial y^2} \right)^2 \right] dx\, dy.$$

Segundo Green e Silverman (1994), $J_{[2]}(g)$ será finito e o quadrado das segundas derivadas de g integráveis em $\mathcal{R}^2$ se, e somente se, $g(t)$ for um *thin plate spline* natural. Se $g(t)$ é um *thin plate spline* natural, então por (3.7) e (3.8) segue-se

$$J_{[2]}(g) = \boldsymbol{\delta}^T \boldsymbol{E\delta}.$$

Observa-se que o funcional $J_{[2]}(g)$ possui uma interpretação prática. Ele quantifica a energia necessária para se empenar uma chapa fina e elástica quando se tenta lhe dar conformação de uma função g. Quanto maior a curvatura necessária para acomodar grandes flutuações locais de g , maior será a energia requerida nesse processo (GREEN; SILVERMAN, 1994) e assim, dependendo da rigidez mecânica dessa chapa, o resultado poderá oscilar entre uma superfície plana (uma chapa muito rígida, mantendo a analogia explicada) ou uma superfície bastante irregular seguindo as varições de g (uma chapa mecanicamente muito flexível).

Assim, a estimação de g é realizada pela minimização da soma dos quadrados dos resíduos penalizados

$$\begin{aligned} \text{SQR}(g,\lambda) &= \sum_{i=1}^{N}\left[y_i - g(t_i)\right]^2 + \lambda + \int_{\S}\int_{\dagger}\left[\left(\frac{\partial^2 g}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 g}{\partial x \partial y}\right) + \left(\frac{\partial^2 g}{\partial y^2}\right)^2\right] dxdy, \\ &= \sum_{i=1}^{N}\left[y_i - g(t_i)\right]^2 + \lambda J_{[2]}(g). \end{aligned} \tag{3.10}$$

Observa-se em (3.10) que, se $\lambda = 0$, não será imposta nenhuma penalização e, em termos da analogia mecânica estabelecida, a chapa será muito flexível e se empenará de modo a se aproximar de uma interpolação dos pontos y_i. Para $\lambda = \omega$ $(\omega \neq 0)$ uma penalização será imposta e, ainda sob a mesma analogia, a chapa será um pouco mais rígida e exibirá uma curvatura menos oscilante, deixando de interpolar todos os pontos e exibindo uma tendência geral das observações.

O sistema matricial expresso em (3.9) apresenta solução única, considerando que a matriz composta por $\boldsymbol{E}$, $\boldsymbol{T}$ e $\mathbf{0}$, de dimensões $(n+3)\times(n+3)$, é de posto completo (não singular) (GREEN; SILVERMAN, 1994).

Assim, o *thin plate spline* expresso por (3.6) que minimiza a soma de quadrados penalizados em (3.10) pode ser expresso como

$$g(\boldsymbol{t}) = (\boldsymbol{Y} - \boldsymbol{E\delta} - \boldsymbol{T}^T\boldsymbol{a})^T(\boldsymbol{Y} - \boldsymbol{E\delta} - \boldsymbol{T}^T\boldsymbol{a}) + \lambda\boldsymbol{\delta}^T\boldsymbol{E\delta}.$$

Green e Silverman (1994) demonstraram que as estimativas $\hat{\boldsymbol{\delta}}$ e $\hat{\mathbf{a}}$ sob um parâmetro de suavização fixo λ são únicas e obtidas pela resolução do sistema matricial

$$\left[\begin{array}{c|c} \boldsymbol{E}^2 + \lambda\boldsymbol{E} & \boldsymbol{E}\boldsymbol{T}^T \\ \hline \boldsymbol{T}\boldsymbol{E} & \boldsymbol{T}\boldsymbol{T}^T \end{array}\right] \cdot \begin{bmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{E} \\ \boldsymbol{T} \end{bmatrix} \boldsymbol{Y}.$$

A dificuldade inerente à resolução desse sistema é o custo computacional envolvido, visto que funções suavizadoras estimadas com essa base têm tantos parâmetros desconhecidos, quanto observações ou, estritamente, seu número de combinações únicas.

3.4.5 *Thin plate regression splines*

Wood (2003) propôs uma elegante solução para se reduzir a complexidade da estimação imposta pelo uso de tantas funções na base ao demonstrar que, pela decomposição da base integral do *thin-plate spline* ao tomar apenas *k* autovetores com *k* maiores autovalores e com estes criar uma base truncada para o suavizador desejado (*low rank smoother*), o espaço vetorial gerado conseguia preservar muito da base original com dimensionalidade bastante menor (WOOD, 2017). A implementação computacional dos *thin plate regression splines* encontra-se disponível na biblioteca mgcv (WOOD, 2022) escrita em R (R Core Team, 2021).

3.4.6 *Splines* de produto tensor

Thin plate splines são invariantes sob rotação e translação e essa é uma propriedade desejável ao modelar uma função suavizadora que tenha coordenadas geográficas como covariáveis (especificamente neste contexto, esta propriedade é usualmente denominada isotropia).

Todavia, existem diversas outras situações nas quais as variáveis usadas como argumentos na função suavizadora são medidas em diferentes unidades como, por exemplo, nas situações com coordenadas espaciais e temporais e nessas a importância relativa implícita da suavidade no tempo (medido em horas) *versus* suavidade no espaço (medido em metros) será muito diferente de uma outra situação na qual as unidades de medida forem anos-luz e nanossegundos (WOOD, 2017).

O produto tensor de famílias unidimensionais de suavizadores é uma outra abordagem destinada a gerar funções suavizadoras em espaços de maior dimensão e as referências inciais a *splines* assim construídas remontam a De Boor (2001).

Para fins de simplificação considere uma área retangular $T \times U$ em $\mathcal{R}^2$ em que *T* e *U* são intervalos em $\mathcal{R}$.

Dadas duas funções suavizadoras unidimensionais f e g definidas em *T* e *U*: $\{f_{j_1} : j_1 = 1, 2, \ldots, q_1\}$ e $\{g_{j_2} : j_2 = 1, 2, \ldots, q_2\}$ respectivamente, tais que $f : T \to \mathcal{R}$ e $g : U \to \mathcal{R}$, o produto tensor de *splines* é o conjunto de todas as combinações lineares geradas sistematicamente a partir do produto tensor de todas as combinações lineares possíveis de funções suavizadoras unidimensionais e resulta em funções suavizadoras de espaços de maior dimensão (GREEN; SILVERMAN, 1994), representado por

$$f_{f,g}(t,u) = \left\{ \sum_r \beta_r \left[\sum_{j_1=1}^{q_1} \delta_{rj_1} f_{j_1} \right] \otimes \left[\sum_{j_2=1}^{q_2} \epsilon_{rj_2} g_{j_2} \right] \right\} = \left\{ \sum_{j_1=1}^{q_1} \sum_{j_2=1}^{q_2} \alpha_j f_{j_1} \otimes g_{j_2} \right\},$$

em que o índice j denota o par (j_1, j_2) e as $q_1 q_2$ funções $f_{j_1} \otimes g_{j_2}$ são linearmente independentes e formam uma base para $\mathcal{G}$. O produto tensor de f por g é definido como a operação $f \otimes g : T \times U \to \mathcal{R}$ definida por

$$(f \otimes g)(t,u) = f(t)g(u),$$

em que $\otimes$ é o operador do produto Kronecker.

Para o desenvolvimento de uma função suavizadora a partir de três outras funções suavizadoras de distintas variáveis, estimadas a partir de distintas bases como

$$f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x);\ f_z(z) = \sum_{l=1}^{L} \delta_l d_l(z); \text{e } f_v(v) = \sum_{k=1}^{K} \beta_k b_k(v),$$

em que $\alpha_i, \delta_l, \beta_k$ são os parâmetros e $a_i(x), d_l(z), b_k(v)$ são as funções de suas bases, considere converter $f_x(x)$ em uma função suavizadora baseada em x e z.

Para tanto, é necessário que f_x varie suavemente com z, o que é conseguido permitindo-se que seu parâmetro α_i varie suavemente com z. Expressando-o em termos das combinações lineares da base usada para gerar z tem-se

$$\alpha_i = \sum_{l=1}^{L} \delta_{il} d_l(z)$$

e a função convertida nessa etapa passa a ser expressa por

$$f_{xz}(x, z) = \sum_{i=1}^{I} \sum_{l=1}^{L} \delta_{il} d_l(z) a_i(x). \tag{3.11}$$

Na sequência, uma função suavizadora de x, z e v pode ser criada permitindo que f_{xz} varie suavemente com v e, de modo análogo, isso é feito permitindo-se que os parâmetros de $f_{x,z}$ variem suavemente com v. Expressando-a na forma semelhante a (3.11) (WOOD, 2016) tem-se

$$f_{xzv}(x, z, v) = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{k=1}^{K} \beta_{ilk} b_k(v) d_l(z) a_i(x). \tag{3.12}$$

Wood (2017) demonstra que para qualquer observação x, z e v em particular há uma única relação entre a matriz de delineamento $\boldsymbol{B}$ da função suavizadora f_{xzv} e as matrizes de delineamento das funções suavizadoras marginais B_x, B_z e B_v avaliadas na mesma observação

$$\boldsymbol{B}_i = \boldsymbol{B}_{x_i} \odot \boldsymbol{B}_{z_i} \odot \boldsymbol{B}_{v_i},$$

em que $\odot$ é o produto Kronecker das bases marginais pela *i-ésima* linha.

Escrevendo-se o vetor dos coeficientes β_{ilk} na ordem: $\boldsymbol{\beta}^T = (\beta_{111}, \beta_{112}, \ldots, \beta_{11K}, \beta_{121}, \beta_{122}, \ldots, \beta_{1LK}, \beta_{ILK})$ (com I,L,K sendo as dimensões de cada uma das bases), para qualquer observação x, z e v, a matriz $\boldsymbol{B}$ pode ser gerada e irá mapear os parâmetros β_{ilk} aos valores avaliados da função nesses valores.

Definida a função suavizadora como em (3.12), um funcional que quantifique sua variabilidade (*wiggliness*) pode ser derivado a partir dos funcionais das funções marginais das quais foi derivada.

Considere $f_{x|zv}(x)$ e $f_{v|xv}(v)$ como funções suavizadoras apenas de x e v, quando se mantêm constantes z e v, x e v, respectivamente. Assim, o funcional desejado pode ser expresso na forma básica

$$J_{[3]}(f) = \lambda_x \int_{z,v} J_x(f_{x|zv})dzdv + \lambda_z \int_{x,v} J_z(f_{z|xv})dxdv + \lambda_v \int_{x,z} J_v(f_{v|xz})dxdz, \qquad (3.13)$$

em que $\lambda_\cdot$ são os parâmetros de suavização que mediam o equilíbrio entre interpolação e suavização nas diferentes direções, permitindo que a penalização seja invariante com relação à escala das covariáveis.

Wood (2017) estende o desenvolvimento do funcional exposto em (3.13) a partir dos funcionais das funções marginais de suavização expressos em suas formas quadráticas. Mais detalhes em (EILERS; MARX, 2003), (CURRIE; REGUERA; EILERS, 2004) e (REGUERA, 2021) .

3.5 MODELOS ADITIVOS

Embora conceitualmente simples e bastante úteis por possibilitar descrever as relações entre uma variável e um conjunto de covariáveis e também permitir realizar inferências sobre as contribuições de cada uma delas, os Modelos de Regressão Linear, muitas vezes, mostram-se inaplicáveis a diversas situações reais. Uma alternativa é a utilização de Modelos Aditivos que apresentam como característica relacionada à sua interpretação, que o efeito na resposta devido a uma das covariáveis - mantidas todas as demais fixas - não depende dessas.

Admita-se um conjunto de observações y_i com $i = 1, \ldots, n$, de uma variável resposta explicada por um conjunto de p covariáveis $\{x_{i1}, \ldots, x_{ip}\}$. Modelos Aditivos são, de certo modo, uma generalização dos Modelos Lineares e assim, um modelo não paramétrico pode ser definido por

$$Y = f(X_1, \ldots, X_p) + \varepsilon;$$

e

$$\mathbb{E}[Y|(X)] = \alpha + \sum_{j=1}^{p} f_j(x_{ji}) + \varepsilon_i, \tag{3.14}$$

em que α é o intercepto comum (em (3.14) está implícito que $\mathbb{E}(f_j(\boldsymbol{X}_j)) = 0$ de modo que exista apenas um intercepto); ε_i o erro aleatório tal que $\varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$ e $f_j(x_{ji})$ são funções suavizadoras univariadas não conhecidas para cada uma das covariáveis.

Considere a seguir a situação mais simples com duas covariáveis (*x,v*) tal como $y_i = \alpha + f_1(x_i) + f_2(v_i)$, em que cada função suavizadora pode ser expressa na forma

$$f_1 = \sum_{j=1}^{k_1} \delta_j b_j(x); \text{ e } f_2 \ = \sum_{j=1}^{k_2} \gamma_j c_j(v),$$

em que δ_j são os coeficientes desconhecidos das funções de base $b_j(x)$ definidas, usando-se uma sequência de k_1 nós x_j^* equidistantes uns dos outros sobre a faixa de valores assumidos pela variável x e γ_j são os coeficientes desconhecidos das funções de base $c_j(v)$ definidas, usando-se uma sequência de k_2 nós v_j^* equidistantes uns dos outros sobre a faixa de valores assumidos pela variável v.

Escrevendo-se vetorialmente f_1 e f_2 e definindo-se as matrizes $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ composta pelos valores avaliados de cada uma das funções de base nos valores observados, na forma

$$\boldsymbol{X_1} = \begin{bmatrix} b_1(x_1) & \ldots & b_{k_1}(x_1) \\ b_1(x_2) & \ldots & b_{k_1}(x_2) \\ \vdots & \vdots & \ddots \\ b_1(x_n) & \ldots & b_{k_1}(x_n) \end{bmatrix}_{(n,k_1)} ; \boldsymbol{X_2} = \begin{bmatrix} c_1(x_1) & \ldots & c_{k_2}(x_1) \\ c_1(x_2) & \ldots & c_{k_2}(x_2) \\ \vdots & \vdots & \ddots \\ c_1(x_n) & \ldots & c_{k_2}(x_n) \end{bmatrix}_{(n,k_2)},$$

os vetores $\boldsymbol{\delta}$ e $\boldsymbol{\gamma}$ como

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{k_1} \end{bmatrix}_{(k_1,1)} ; \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{k_2} \end{bmatrix}_{(k_2,1)},$$

segue que $\boldsymbol{f}_1 = \boldsymbol{X}_1\boldsymbol{\delta}$ e $\boldsymbol{f}_2 = \boldsymbol{X}_2\boldsymbol{\gamma}$.

Definindo-se a matriz $\boldsymbol{D}$ das diferenças de ordem $d = 2$ entre duas bases adjacentes como descrito por Reguera (2021) na forma

$$\boldsymbol{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \dots \\ 0 & 1 & -2 & 1 & 0 & \dots \\ 0 & 0 & 1 & -2 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{(j-2,j)}.$$

as penalizações para cada uma das duas funções podem ser escritas como uma forma quadrática (WOOD, 2017)

$$\boldsymbol{\delta}^T\boldsymbol{D}_1^T\boldsymbol{D}_1\boldsymbol{\delta} = \boldsymbol{\delta}^T\boldsymbol{G}_1\boldsymbol{\delta};$$

e

$$\boldsymbol{\gamma}^T\boldsymbol{D}_2^T\boldsymbol{D}_2\boldsymbol{\gamma} = \boldsymbol{\gamma}^T\boldsymbol{G}_2\boldsymbol{\gamma},$$

em que $\boldsymbol{G}_1 = \boldsymbol{D}_1^T\boldsymbol{D}_1$ e $\boldsymbol{G}_2 = \boldsymbol{D}_2^T\boldsymbol{D}_2$.

A incorporação de funções suavizadoras em modelos aditivos envolve a imposição de restrições de identificabilidade pela impossibilidade de se estimar um intercepto para cada uma delas $\boldsymbol{\delta}$ e $\boldsymbol{\gamma}$. O modo mais conveniente é reparametrizando as funções de modo que se tornem ortogonais em relação ao intercepto (WOOD, 2017):

$$\sum_{i=1}^{n} f_1(x_i) = 0; \text{ e } \sum_{i=1}^{n} f_2(v_i) \ = 0, \tag{3.15}$$

em que (3.15) são equivalentes a $\mathbf{1}^T\boldsymbol{X}_1\boldsymbol{\delta} = \mathbf{1}^T\boldsymbol{f}_1 = 0$ e $\mathbf{1}^T\boldsymbol{X}_2\boldsymbol{\gamma} = \mathbf{1}^T\boldsymbol{f}_2 = 0$. As versões reparametrizadas (restritas) de $\boldsymbol{f}$ e $\boldsymbol{X}$ e $\boldsymbol{D}$ (por conseguinte) ao final assumem a forma

$$\tilde{\boldsymbol{X}}_1 = \boldsymbol{X}_1 - \mathbf{1}\,\mathbf{1}^T\,\boldsymbol{X}_1\,(1/n);$$

e

$$\tilde{\boldsymbol{X}}_2 = \boldsymbol{X}_2 - \mathbf{1}\,\mathbf{1}^T\,\boldsymbol{X}_2\,(1/n),$$

$$\tilde{\boldsymbol{f}}_1 = \tilde{\boldsymbol{X}}_1\boldsymbol{\delta} = \boldsymbol{X}_1\boldsymbol{\delta} - \mathbf{1}\,\mathbf{1}^T\,\boldsymbol{X}_1\boldsymbol{\delta}\,(1/n) = \boldsymbol{X}_1\boldsymbol{\delta} - \mathbf{1}\,c = \boldsymbol{f}_1 - c_1;$$

e

$$\tilde{\boldsymbol{f}}_2 = \tilde{\boldsymbol{X}}_2\boldsymbol{\gamma} = \boldsymbol{X}_2\boldsymbol{\gamma} - \mathbf{1}\,\mathbf{1}^T\,\boldsymbol{X}_2\boldsymbol{\gamma}\,(1/n) = \boldsymbol{X}_2\boldsymbol{\gamma} - \mathbf{1}\,c = \boldsymbol{f}_2 - c_2,$$

com os escalares $c_1 = \mathbf{1}^T \boldsymbol{X}_1 \boldsymbol{\delta}\ (1/n)$ e $c_2 = \mathbf{1}^T \boldsymbol{X}_2 \boldsymbol{\gamma}\ (1/n)$. Com as bases restritas já definidas é possível expressar o modelo como

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \tag{3.16}$$

em que $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2)$ e $\boldsymbol{\beta} = (\alpha, \boldsymbol{\delta}^T, \boldsymbol{\gamma}^T)$.

A estimação do vetor $\hat{\boldsymbol{\beta}}$ em (3.16) é promovida pela minimização da expressão da soma de quadrados penalizada (para λ_1 e λ_2 fixos)

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_1 \boldsymbol{\beta}^T \boldsymbol{G}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \boldsymbol{G}_2 \boldsymbol{\beta},$$

e uma estimativa para β será dada por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda_1 \boldsymbol{G}_1 + \lambda_2 \boldsymbol{G}_2)^{-1} \boldsymbol{X}^T \boldsymbol{y};$$
$$\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y},$$

em que $\boldsymbol{H}$ é a matriz de projeção

$$\boldsymbol{H} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda_1 \boldsymbol{G}_1 + \lambda_2 \boldsymbol{G}_2)^{-1} \boldsymbol{X}^T.$$

3.6 MODELOS ADITIVOS GENERALIZADOS (*GAM*)

Em meados da década de 1980, Hastie e Tibshirani (1986) introduziram uma classe de modelos na qual estendem uma coleção de modelos baseados na verossimilhança a outros, mediante a substituição do preditor linear $\sum \beta_j X_j$ por um preditor aditivo $\sum s_j(X_j)$, isto é, uma soma de funções suavizadores não paramétricas que se mostrassem úteis para revelar efeitos não lineares das covariáveis e a denominaram Modelos Generalizados Aditivos (GAM: *Generalized Additive Models*).

Modelos Aditivos Generalizados incorporam à estrutura dos modelos aditivos a pressuposição de que a resposta observada segue alguma distribuição de probabilidade pertencente à família exponencial, permitindo, assim, a inclusão de um termo paramétrico na estrutura do modelo e possibilitando sua estimação conjunta pela maximização do logaritmo da função de verossimilhança (WOOD, 2017). Se a variável resposta pode ser modelada por uma das distribuições pertencente à família exponencial mas apresenta alguma relação entre sua média e variância, é possível a realização de inferências pela quase-verossimilhança (WOOD, 2008).

Não obstante essa classe de modelos permita uma especificação bem mais flexível da relação de dependência da resposta com as covariáveis, tal flexibilidade e conveniência vêm incorporar dois novos problemas teóricos: como representar e determinar quão suaves são as funções suavizadoras (STASINOPOULOS et al., 2017).

Um modelo generalizado aditivo semiparamétrico pode ser expresso na forma (WOOD, 2017):

$$\begin{aligned} g(\mu_i) &= \boldsymbol{A}_i\boldsymbol{\gamma} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4,i}) + \ldots, \\ g(\mu_i) &= \boldsymbol{A}_i\boldsymbol{\gamma} + \sum_j f_j(x_{ji}), \quad y_i \overset{ind}{\sim} \mathcal{E}\left(\mu_i, \phi\right), \end{aligned} \tag{3.17}$$

em que Y é a variável resposta; $\mathcal{E}(\boldsymbol{\mu}, \phi)$ denota uma distribuição pertencente à família exponencial com vetor de médias $\boldsymbol{\mu}$ e parâmetro de dispersão ϕ; $\boldsymbol{A}_i$ é a *i-ésima* linha da matriz de delineamento para qualquer termo estritamente paramétrico do modelo e $\boldsymbol{\gamma}$ o vetor de parâmetros correspondentes à matriz de delineamento e f_j são funções suavizadoras não paramétricas pertencentes ao espaço *infinito-dimensional* das funções duplamente contínuas e diferenciáveis e duplamente integráveis aplicadas às covariáveis x_j.

Cada uma das funções suavizadoras definidas em (3.17) é expressa como uma combinação linear de Q bases conhecidas $b_{jq}(x_{ji})$ por β_{jq} parâmetros a serem estimados

$$f_j(x_{ji}) = \sum_{q=1}^{Q_j} \beta_{jq} b_{jq}(x_{ji}).$$

Assim, cada uma das funções suavizadoras f_j terá associadas matrizes de delineamento e de penalização específicas, $\boldsymbol{X}^{[j]}$ e $\boldsymbol{S}^{[j]}$ respectivamente e, se b_{jq} é a *q-ésima* função da base de f_j, então $X_{iq}^{[j]} = b_{jq}(x_{ji})$.

De modo análogo à estimação dos parâmetros do modelo aditivo, restrições para identificabilidade da mesma forma que em (3.15) deverão ser aplicadas para que o intercepto de cada uma das funções suavizadoras não se confunda com o intercepto da parte paramétrica em **A**.

Definindo $\boldsymbol{\mathcal{X}}^{[j]}$ como a matriz de delineamento para f_j, após a reparametrização ela pode ser acrescida como colunas à matriz de delineamento da parte paramétrica **A**, criando-se uma matriz de delineamento global para o modelo na forma

$$\boldsymbol{X} = (\boldsymbol{A} : \boldsymbol{\mathcal{X}}^{[1]} : \boldsymbol{\mathcal{X}}^{[2]} : \dots).$$

O vetor $\boldsymbol{\beta}$ de coeficientes da matriz de delineamento global contém os parâmetros $\boldsymbol{\gamma}$ e, em sequência, os coeficientes de cada uma das funções suavizadoras.

O termo de penalização global é expresso na forma

$$\sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}_{[j]} \boldsymbol{\beta},$$

em que λ_j são os parâmetros de suavização arbitrados que regulam a qualidade do ajuste *versus* a suavidade final do modelo e $\boldsymbol{S}_{[j]}$ é uma matriz composta pelas matrizes de penalização reparametrizadas de cada função suavizadora f_j adicionadas diagonalmente, tendo zero nas suas outras entradas.

Com essa reparametrização o modelo (3.17) assume a estrutura de um Modelo Linear Generalizado expresso na forma:

$$g(\mu_i) = \boldsymbol{X}_i \boldsymbol{\beta}, \;\; y_i \overset{ind}{\sim} \mathcal{E}\,(\mu_i, \phi).$$

Ao se estipular *a priori* um grande número de funções de bases para cada uma das funções suavizadoras, a estimação de $\boldsymbol{\beta}$ pela simples minimização dos desvios acarretará em um "sobreajustamento" e, por esta razão, os Modelos Aditivos Generalizados são estimados definindo-se $D(\boldsymbol{\beta}) = 2\{l_{max} - l(\boldsymbol{\beta})\}$, em que l_{max} é o valor do logaritmo da função de verossimilhança do modelo saturado.

Minimizando-se os desvios penalizados (WOOD, 2008) tem-se que

$$\hat{\boldsymbol{\beta}} = arg.min_{\boldsymbol{\beta}} \left[D(\boldsymbol{\beta}) + \sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}_{[j]} \boldsymbol{\beta} \right],$$

o que equivale a maximizar o logaritmo da função de verossimilhança penalizada para um valor fixo de λ, isto é

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi} \sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}_{[j]} \boldsymbol{\beta}.$$

3.7 Modelos Aditivos Generalizados para Locação, Escala e Forma

3.7.1 Introdução

No XVI *Workshop* Internacional de Modelagem Estatística, promovido pela Sociedade de Modelagem Estatística (*SMS Statistical Modelling Society*) e realizado na Dinamarca em 2001, Robert A. Rigby e Dimitrius Mikis Stasinopoulos apresentaram como contribuição o projeto *GAMLSS*: uma abordagem flexível para modelagem estatística (*The GAMLSS project: a flexible approach to statistical modelling*). Essa nova classe de modelos, que poucos anos mais tarde (RIGBY; STASINOPOULOS, 2005) seria formalmente introduzida, viria a permitir que também a variabilidade e a forma da distribuição da variável resposta fossem modeladas de modo explícito a partir das covariáveis, superando uma peculiaridade de modelos de outras classes como Modelos Lineares Generalizados e Modelos Aditivos Generalizados, em que apenas a média μ era modelada diretamente e a variabilidade da resposta $\text{Var}(y) = \phi V(\mu)$ dependia de modo implícito de uma função de variância sobre μ e de uma constante (parâmetro de dispersão ϕ) (STASINOPOULOS et al., 2017).

Os *GAMLSS* de Rigby e Stasinoplous (e outros que vieram a compor a equipe: *GAMLSS* Team) enquadram-se na técnica que veio a ser denominada como regressão da distribuição (*distributional regression*) por possibilitar que cada parâmetro de uma distribuição potencialmente complexa assumida para a resposta possa ser estruturado na forma de um preditor aditivo composto por funções paramétricas e/ou semiparamétricas (funções suavizadoras) e/ou efeitos aleatórios estimadas a partir de subconjuntos de covariáveis particulares para cada um dos parâmetros (KLEIN et al., 2015).

A estrutura computacional para modelagem dessa classe estatística de modelos encontra-se implementada em uma série de bibliotecas escritas em linguagem *R* e disponíveis em: *The Comprehensive R Archive Network*.

3.7.2 Definição

Seja $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ o vetor das observações realizadas de uma variável resposta com função densidade de probabilidade $f(y|\boldsymbol{\theta})$ e vetor paramétrico $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, em que as observações y_i para $i = (1, \ldots, n)$ são independentes e condicionais a θ^i, com função densidade de probabilidade $f(y_i|\boldsymbol{\theta}^i)$ e vetor paramétrico $\boldsymbol{\theta}^i = (\theta_{i1}, \ldots, \theta_{ip})^T$ relacionado às variáveis explicativas e a efeitos aleatórios.

Ao se considerar uma distribuição de probabilidade com quatro parâmetros para modelar uma variável resposta *Y*, um *GAMLSS* semiparamétrico pode ser expresso como $\boldsymbol{Y} \overset{ind}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ e cada parâmetro distribucional estimado por um preditor específico

$$
\begin{aligned}
\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + s_{11}(\boldsymbol{x}_{11}) + \cdots + s_{1J_1}(\boldsymbol{x}_{1J_1});\\
\boldsymbol{\eta}_2 &= g_2(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + s_{21}(\boldsymbol{x}_{21}) + \cdots + s_{2J_2}(\boldsymbol{x}_{2J_2});\\
\boldsymbol{\eta}_3 &= g_3(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 + s_{31}(\boldsymbol{x}_{31}) + \cdots + s_{3J_3}(\boldsymbol{x}_{3J_3});\\
\boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4 + s_{41}(\boldsymbol{x}_{41}) + \cdots + s_{4J_4}(\boldsymbol{x}_{4J_4}),
\end{aligned}
\tag{3.18}
$$

em que $\mathcal{D}(\mu, \sigma, \nu, \tau)$ é a distribuição adotada para a variável resposta $\mathbf{Y}$; $g_1(.), g_2(.), g_3(.), g_4(.)$ são as funções de ligação dos parâmetros da distribuição aos preditores lineares η_1, η_2, η_3, η_4 (μ, σ, ν e τ respectivamente); $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$ são as matrizes de delineamento incorporando, de modo aditivo, termos lineares à parte paramétrica do modelo; $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4$ são os vetores dos coeficientes lineares da parte paramétrica do modelo e, para $k = 1, 2, 3, 4$ e $j = 1, 2, ..., j_k$, $s_{kj}(\boldsymbol{x}_{kj})$ são funções suavizadoras não conhecidas das variáveis explanatórias x_{kj} incorporadas aditivamente a cada preditor dos parâmetros da distribuição escolhida para o modelo.

Cada função suavizadora $s_{kj}(\boldsymbol{x}_{kj})$ pode ser escrita na forma $s(\boldsymbol{x}) = \boldsymbol{Z}\boldsymbol{\gamma}$ em que $\boldsymbol{Z}$ é a matriz formada pelas funções da base da função suavizadora e $\boldsymbol{\gamma}$ é o vetor dos coeficientes a serem estimados e sujeitos a uma penalização expressa na forma quadrática $\lambda\boldsymbol{\gamma}^T\boldsymbol{G}\boldsymbol{\gamma}$ em que λ é o parâmetro de suavização, $\boldsymbol{G}$ é a matriz de penalização conhecida que pode ser expressa na forma $\boldsymbol{G} = \boldsymbol{D}^T\boldsymbol{D}$ e $\boldsymbol{D}$ é uma matriz de diferenças, em geral, de ordem $d = 2$ (REGUERA, 2021).

Sob essa reformulação, o modelo expresso em (3.18) pode ser generalizado para

$$
\begin{aligned}
\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_{11}(\boldsymbol{\gamma}_{11}) + \cdots + \boldsymbol{Z}_{1J_1}(\boldsymbol{\gamma}_{1J_1});\\
\boldsymbol{\eta}_2 &= g_2(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_{21}(\boldsymbol{\gamma}_{21}) + \cdots + \boldsymbol{Z}_{2J_2}(\boldsymbol{\gamma}_{2J_2});\\
\boldsymbol{\eta}_3 &= g_3(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \boldsymbol{Z}_{31}(\boldsymbol{\gamma}_{31}) + \cdots + \boldsymbol{Z}_{3J_3}(\boldsymbol{\gamma}_{3J_3});\\
\boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \boldsymbol{Z}_{41}(\boldsymbol{\gamma}_{41}) + \cdots + \boldsymbol{Z}_{4J_4}(\boldsymbol{\gamma}_{4J_4}),
\end{aligned}
\tag{3.19}
$$

em que $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T, \boldsymbol{\beta}_4^T)$ são os vetores dos parâmetros dos efeitos fixos e $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}^T, \boldsymbol{\gamma}_{12}^T, \ldots, \boldsymbol{\gamma}_{1J_1}^T, \boldsymbol{\gamma}_{21}^T, \boldsymbol{\gamma}_{22}^T, \ldots, \boldsymbol{\gamma}_{2J_2}^T, \boldsymbol{\gamma}_{31}^T, \boldsymbol{\gamma}_{32}^T, \ldots, \boldsymbol{\gamma}_{3J_3}^T, \boldsymbol{\gamma}_{41}^T, \boldsymbol{\gamma}_{42}^T, \ldots, \boldsymbol{\gamma}_{4J_4}^T)$ são os coeficientes dos efeitos aleatórios.

Assim, (3.19) pode ser escrito, de modo equivalente, como

$$\begin{aligned}
\boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1};\\
\boldsymbol{\eta}_2 &= g_2(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2};\\
\boldsymbol{\eta}_3 &= g_3(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3};\\
\boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}.
\end{aligned} \tag{3.20}$$

Denotando como $g(\cdot)_k$ para $k = 1, \ldots, p$, uma função monotônica de ligação conhecida que relaciona o *k-ésimo* parâmetro $\boldsymbol{\theta}_k$ às covariáveis e efeitos aleatórios por meio de uma estrutura aditiva, o GAMLSS expresso em (3.20) pode ser expresso em sua forma mais sintética

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{3.21}$$

em que $\boldsymbol{\theta}_k = (\theta_{1k}, \ldots, \theta_{nk})^T$ é o vetor referente ao *k-ésimo* parâmetro da distribuição adotada pela variável $\mathbf{Y}$ vinculados ao vetor $\boldsymbol{\eta}_k$ (seu preditor linear aditivo) de igual dimensão sob a função de ligação $g_k(.)$.

Os termos $\boldsymbol{X}_k\boldsymbol{\beta}_k$ compõem a estrutura paramétrica do preditor aditivo em que $\boldsymbol{X}_k$ é a matriz de delineamento conhecida das covariáveis consideradas, com dimensão $n \times J_k^{'}$ e $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{J_k^{'}k})^T$ o correspondente vetor de parâmetros dos efeitos fixos a serem estimado.

Os termos $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$ compõem a estrutura de efeitos aleatórios do preditor aditivo formada por J_k termos, em que $\boldsymbol{Z}_{jk}$ é a *j-ésima* matriz de delineamento conhecida e fixa de dimensão $n \times q_{jk}$ e $\boldsymbol{\gamma}_{jk} = (\gamma_{1k}, \ldots, \gamma_{J_k k})^T$ são vetores de variáveis aleatórias que podem ser combinados (*empilhados*) em um único vetor $\boldsymbol{\gamma}_k$ associado a uma única matriz $\boldsymbol{Z}_{(k)}$ (expandida em colunas).

Se, para todos os parâmetros da distribuição adotada para a variável resposta ($k = 1, \ldots, p$), não há nenhum termo de efeito aleatório ($J_k = 0$) o modelo (3.21) fica reduzido à sua forma estritamente paramétrica dada por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}\boldsymbol{\beta}_k. \tag{3.22}$$

Caso $\boldsymbol{Z}_{jk} = \boldsymbol{I}_{(n)}$, sendo $\boldsymbol{I}_{(n)}$ é uma matriz identidade $n \times n$ e $\gamma_{jk} = \boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ para todas as combinações *j* e *k* em (3.21), então o *GAMLSS* assume sua forma semiparamétrica (efeitos fixos e aleatórios)

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\boldsymbol{x}_{jk}),$$

em que $h_{jk}(.)$ é uma função não conhecida da variável explicativa X_{jk}; $\boldsymbol{x}_{jk}$ são vetores explanatórios de comprimento *n* considerados como conhecidos e $\boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ é um vetor da função $h_{jk}(.)$ avaliada em $\boldsymbol{x}_{jk}$, para $j = 1, \ldots, J_k$ e $k = 1, \ldots, p$.

Assim, um *GAMLSS* anunciado como $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \Lambda\}$ é tal que $\mathcal{D}$ é a distribuição da variável resposta; $\mathcal{G}$ são as funções de ligação dos preditores: $g_1(.), \ldots, g_4(.)$; $\mathcal{T}$ são os termos considerados nos preditores dos parâmetros distribucionais e Λ são os hiperparâmetros das funções consideradas no modelo.

Os dois primeiros parâmetros distribucionais θ_1 e θ_2 representam medidas de locação e variabilidade da distribuição (usualmente μ e σ) e os demais parâmetros θ_3 e θ_4, caso existam, definem aspectos relacionados à sua forma (usualmente simetria e curtose).

3.7.3 Família de distribuições *GAMLSS* e funções de ligação

Na estrutura de um *GAMLSS* a função densidade (ou probabilidade) da variável resposta $\mathbf{Y}$ $f(y|\boldsymbol{\theta})$ é intencionalmente apresentada de modo genérico. A única exigência do algoritmo de estimação é que $f(y|\boldsymbol{\theta})$ e sua primeira derivada com relação a cada um dos parâmetros $\boldsymbol{\theta}$ sejam possíveis de estimação numérica (STASINOPOULOS et al., 2017).

3.7.3.1 Distribuição Box&Cox "t" (BCT)

As transformações de variáveis propostas por Box&Cox (BOX; COX, 1964) produzem uma nova família de distribuições Normais de potências truncadas à qual pertencem a própria distribuição Normal e a log-Normal.

A distribuição BCT proposta por Rigby e Stasinopoulos (2006) é uma generalização da família de distribuições Normais Box&Cox para as situações nas quais a modelagem tanto da assimetria quanto do excesso de curtose (distribuições assimétricas e leptocúrticas) torna-se necessária.

Seja *Y* uma variável aleatória contínua ($y > 0$) que segue uma distribuição BCT denotada por BCT(μ, σ, ν, τ) e definida por meio da variável aleatória transformada *Z* dada por

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right] & \text{se } \nu \neq 0 \\ \text{ou} & \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right) & \text{se } \nu = 0, \end{cases} \tag{3.23}$$

com $\mu > 0$, $\sigma > 0$ e $-\infty < \nu < \infty$.

A variável aleatória transformada *Z* segue uma distribuição truncada "t" com $\tau > 0$ graus de liberdade e a função da variável *Y* será dada por (RIGBY et al., 2020)

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{y^{\nu-1} f_T(z)}{\mu^{\nu}\sigma F_T\left(\frac{1}{\sigma|\nu|}\right)}, \tag{3.24}$$

em que $f_T(t)$ e $F_T(t)$ são, respectivamente, as funções densidade e de distribuição acumulada de uma variável aleatória $T \sim t[0,1]_{(\tau)}$ (distribuição "t" padronizada), dadas por

$$f_T(z) = \frac{\Gamma[(\tau+1)/2]}{\Gamma(1/2)\Gamma(\tau/2)\tau^2}\left[1+\frac{z^2}{\tau}\right]^{-\frac{(\tau+1)}{2}} = \frac{1}{B(\frac{1}{2},\frac{\tau}{2})\tau^{\frac{1}{2}}}\left[1+\frac{z^2}{\tau}\right]^{-\frac{(\tau+1)}{2}}$$

e

$$F_T(z) = \frac{1}{2} + z\Gamma[(\tau+1)/2] \times \frac{2^{F_1(\frac{1}{2},\frac{(\tau+1)}{2};\frac{3}{2};-\frac{z^2}{\tau})}}{\Gamma(1/2)\Gamma(\tau/2)\tau^2},$$

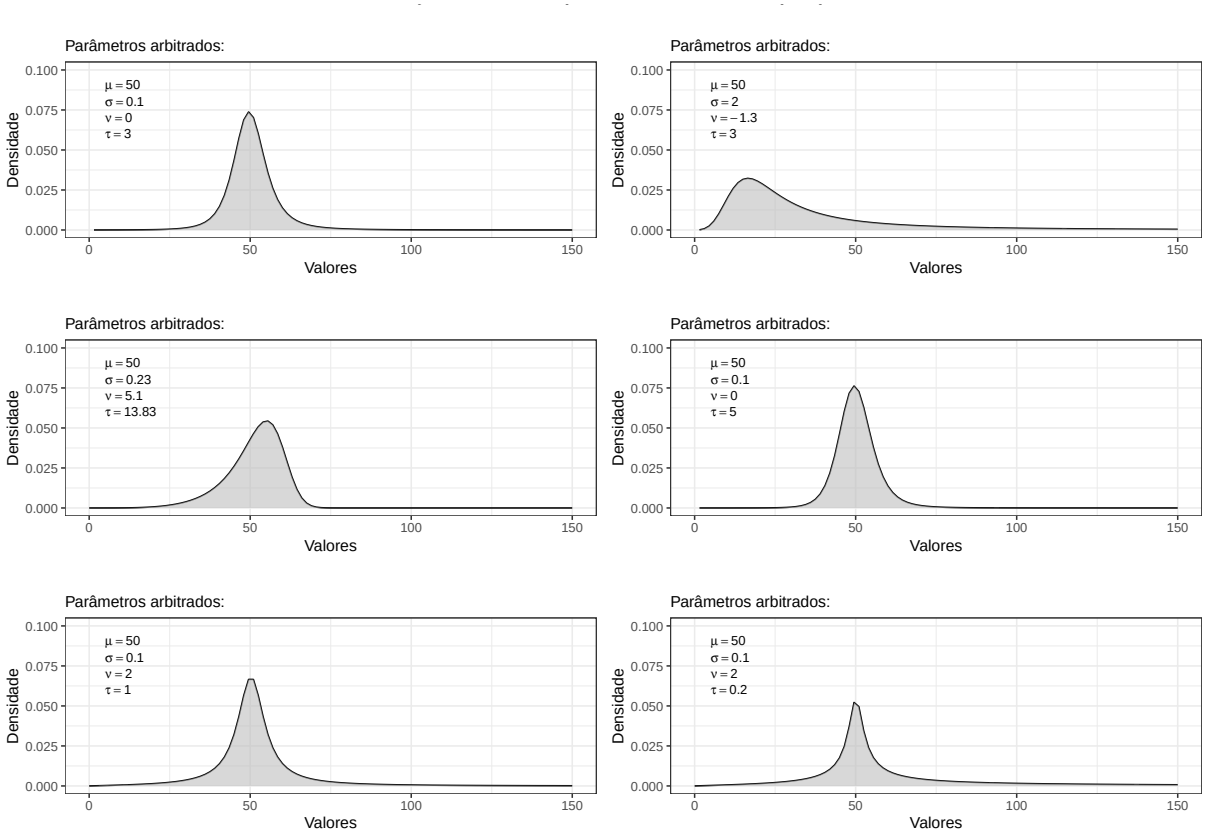
em que z é dado por (3.23), 2^{F_1} é a função hipergeométrica e $\Gamma(\cdot)$ e $B(\cdot)$ são, respectivamente, as funções gama e beta.

A Tabela 3.1 resume as características da distribuição Box&Cox "t" (BCT) e a Figura 3.6 ilustra os diversos perfis assumidos pela curva de densidade da distribuição Box&Cox "t" (BCT), sob diferentes valores paramétricos arbitrados [2]

A distribuição de probabilidade Box&Cox "t" original (BCTo) também tem sua função densidade dada por (3.24) e difere da distribuição BCT apenas pela função de ligação *default* adotada para seu parâmetro de locação μ (mediana) na implementação computacional: função logaritmo na primeira e identidade nessa última.

[2] À guisa de complementação, está disponibilizada na seção BCTo no repositório GitHub do autor em: uma função que permite a visualização do perfil assumido pela curva de densidade da distribuição Box&Cox "t" (BCT) para valores dos parâmetros livremente arbitrados pelo usuário.

Figura 3.6: Variações na curva da densidade da distribuição Box&Cox “t” (BCT)



Fonte: Próprio autor

Tabela 3.1: Distribuição BCT

Faixa de valores	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$
σ	$0 < \sigma < \infty$
ν	$-\infty < \nu < \infty$
τ	$0 < \tau < \infty$
Medidas da distribuição	
-	média
μ	mediana
-	moda
-	variância
τ	coef. associado à curtose (graus de liberdade)
ν	coef. associado à assimetria (transf. pot. pelo parâmetro λ)
$\sigma(\frac{\tau}{\tau-2})^{0,50}$	coef. de variação
Funções	
função geratriz de momentos	-
FDP	$\frac{y^{\nu-1} f_T(z)}{\mu^{\nu}\sigma F_T(\frac{1}{\sigma\lvert\nu\rvert})}$ sendo $T \sim t_\tau = TF[0,1]_{(\tau)}$
FCP	$F_T(z)$ sendo $T \sim t_\tau$ e z é dado por (3.23)
inversa da FCP (y_p)	$\begin{cases} \mu(1+\sigma t_{(p,\tau)})^{\frac{1}{\nu}} & \text{se } \nu \neq 0 \\ \mu\, e^{[\sigma t_{(p,\tau)}]} & \text{se } \nu = 0 \\ \text{com } t_{(p,\tau)} = F_T^{-1}(p) \text{ e } T \sim t_r \end{cases}$

Fonte: Adaptado de Rigby et al. (2020)

3.7.4 Preditor

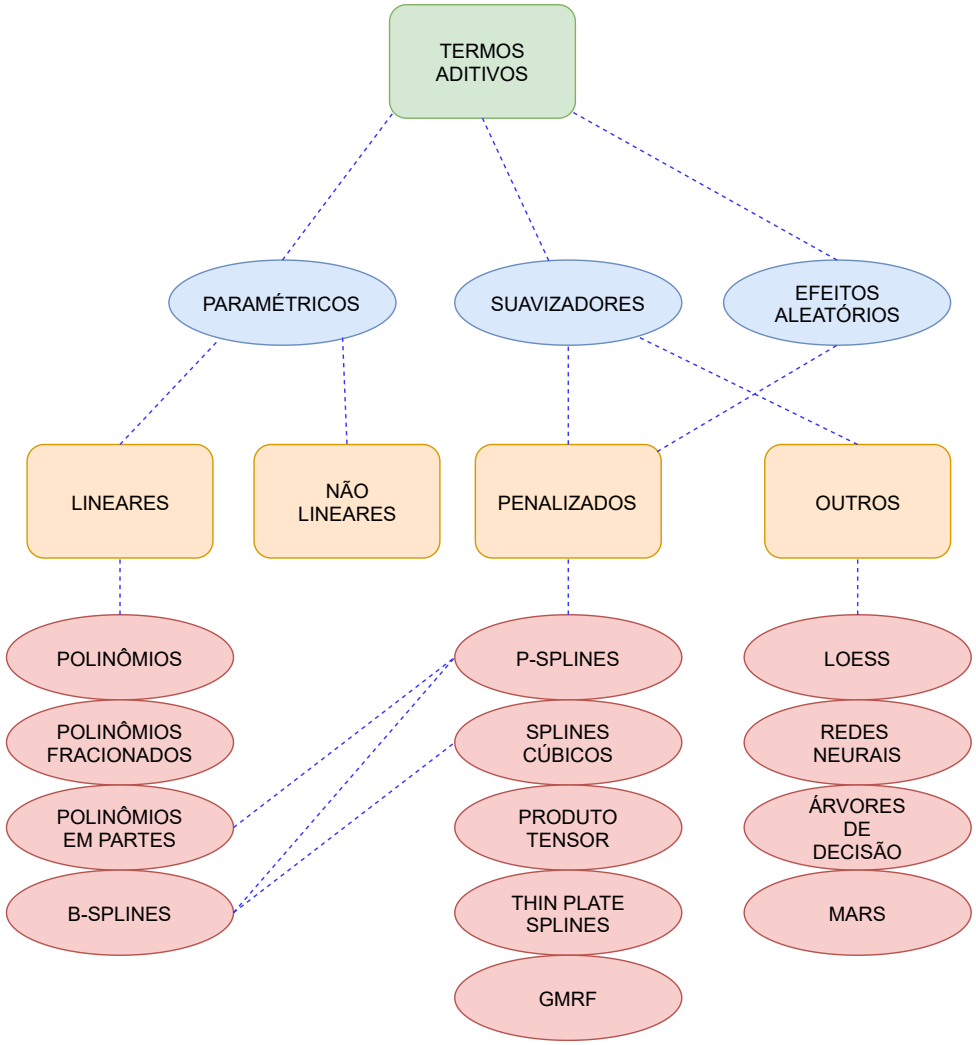
Diferentemente dos modelos clássicos de regressão linear, *GAMLSS* permitem modelar os vários parâmetros da distribuição teórica adotada por meio de variáveis explanatórias, usualmente denominadas como termos em um *GAMLSS*.

Os termos são incorporados aditivamente a cada uma das duas estruturas (3.21) que compõem os preditores lineares η_k dos parâmetros distribucionais de um *GAMLSS* sob variadas formas funcionais, como mostrado na Figura 3.7 e, portanto, o efeito conjunto dos termos em um preditor linear relativo a um parâmetro específico da distribuição da variável resposta será a soma dos efeitos individuais de cada um desses termos:

- paramétrica;
  - lineares: polinômios, polinômios fracionários, polinômios por partes, *B-splines*; ou,
  - não lineares.
- não paramétrica (funções suavizadoras):
  - penalizadas: *P-splines*, *splines* cúbicos, produto tensor, *thin plate splines*, campos aleatórios gaussianos Markovianos;
  - outras: regressão local (LOESS/LOWESS), redes neurais, árvores de decisão e *splines* de regressão adaptativa multivariada.
- efeitos aleatórios.

Torna-se importante destacar haver na biblioteca gamlss.add (STASINOPOULOS et al., 2020) funções que permitem a inclusão dos termos sob variadas formas não paramétricas, como a função de interface ga() a qual permite a utilização de todas as funções suavizadoras existentes na biblioteca mgcv (WOOD, 2022), assim como as funções nn() para a biblioteca nnet (RIPLEY; VENABLES, 2022) e tr() para a biblioteca rpart (THERNEAU; ATKINSON; RIPLEY, 2022), todas escritas em (R Core Team, 2021).

Figura 3.7: Funções disponíveis na implementação computacional para a incorporação dos termos nos preditores dos parâmetros



Fonte: Adaptado de Stasinopoulos et al. (2017)

3.7.5 Ajuste

O ajuste de um *GAMLSS*, tal como apresentado em (3.20), requer que sejam estimados:

- os parâmetros dos efeitos fixos: $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t, \boldsymbol{\beta}_3^t, \boldsymbol{\beta}_4^t)$;
- coeficientes dos efeitos aleatórios:
  $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}^T, \ldots, \boldsymbol{\gamma}_{1J_1}^T, \boldsymbol{\gamma}_{21}^T, \ldots, \boldsymbol{\gamma}_{2J_2}^T, \boldsymbol{\gamma}_{31}^T, \ldots, \boldsymbol{\gamma}_{3J_3}^T, \boldsymbol{\gamma}_{41}^T, \ldots, \boldsymbol{\gamma}_{4J_4}^T)$;
- os parâmetros de suavização:
  $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{11}^T, \ldots, \boldsymbol{\lambda}_{1J_1}^T, \boldsymbol{\lambda}_{21}^T, \ldots, \boldsymbol{\lambda}_{2J_2}^T, \boldsymbol{\lambda}_{31}^T, \ldots, \boldsymbol{\lambda}_{3J_3}^T, \boldsymbol{\lambda}_{41}^T, \ldots, \boldsymbol{\lambda}_{4J_4}^T)$.

Admita que os $\gamma_{kj}^{'}$ em (3.19) sejam independentes e tenham distribuição

$$\gamma_{kj} \overset{ind}{\sim} \mathcal{N}(0, [\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})]^{-1}), \tag{3.25}$$

em que $[\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})]^{-1}$ é a inversa generalizada da matriz simétrica $\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})$ de dimensões $q_{kj} \times q_{kj}$ a qual, talvez, dependa do vetor de parâmetros $\boldsymbol{\lambda}_{kj}$. Se $\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})$ é uma matriz singular, então γ_{kj} é entendido como tendo uma função densidade imprópria proporcional a $\exp(\frac{1}{2}\gamma^T \boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj})\gamma)$. Uma importante situação relativa a (3.25) é quando $\boldsymbol{G}_{kj}(\boldsymbol{\lambda}_{kj}) = \lambda_{kj}\boldsymbol{G}_{kj}$, e $\boldsymbol{G}_{kj}$ é uma matriz conhecida para todo *k*, *j* (STASINOPOULOS et al., 2017).

Sob a premissa de que as observações sejam independentes, um *GAMLSS* estritamente paramétrico como (3.22) é estimado pela maximização do logaritmo de sua função de verossimilhança (STASINOPOULOS et al., 2017)

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i), \tag{3.26}$$

ao passo que um modelo não paramétrico como (3.20), pela maximização do logaritmo de sua função de verossimilhança penalizada

$$l_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = l(\boldsymbol{\theta}) - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \lambda_{kj}\boldsymbol{\gamma}_{kj}^T \boldsymbol{G}_{kj}\boldsymbol{\gamma}_{kj}.$$

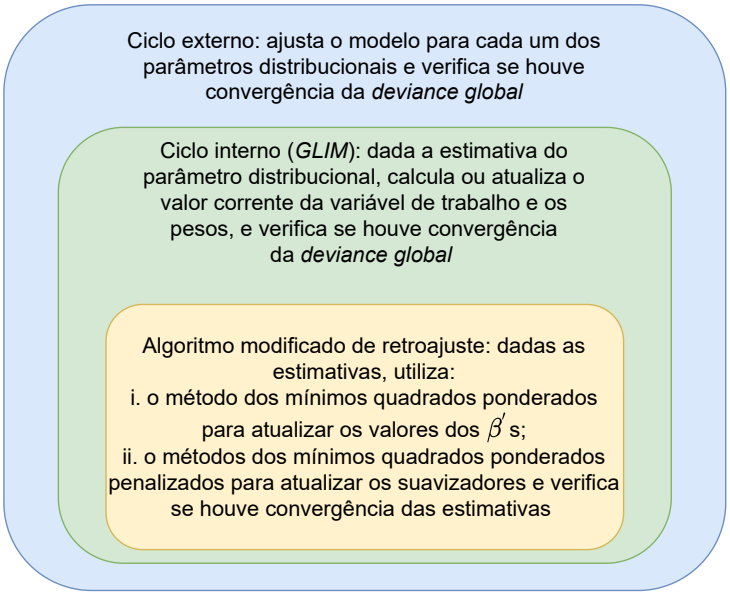
em que $l(\boldsymbol{\theta})$ é o logaritmo da função de verossimilhança (3.26); $\boldsymbol{G}_{kj}$ é uma matriz de penalização simétrica de dimensões $q_{kj} \times q_{kj}$ definida para o modelo (3.20) cuja inversa generalizada é a matriz de variâncias e covariâncias dos efeitos aleatórios γ_{kj}

Stasinopoulos et al. (2017) apresentam dois algoritmos para o ajuste de um *GAMLSS*:

- RS: uma versão generalizada da originalmente proposta por Rigby e Stasinopoulos (1996a) e Rigby e Stasinopoulos (1996b) para o ajuste de Modelos Aditivos Generalizados para Média e Dispersão, geralmente mais estável e, na maior parte das situações, mais rápida;
- CG: uma generalização do algoritmo proposto por Cole e Green (1992).

Rigby e Stasinopoulos (2005) demonstram que, nos modelos estritamente paramétricos, os algoritmos retornam estimativas de máxima verossimilhança (EMV) para $\boldsymbol{\beta}$ ao passo que, nos modelos não paramétricos, são obtidas as estimativas máximas *a posteriori* de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ para valores fixos de $\boldsymbol{\lambda}$. Os ciclos externo, interno e de retroajuste dos algoritmos citados encontram-se ilustrados na Figura 3.8).

Figura 3.8: Esquema geral dos ciclos dos algoritmos de ajuste de um *GAMLSS*



Fonte: Adaptado de Thomas (2017)

3.7.5.1 Comparação de modelos

O critério de convergência adota como métrica o desvio global que, nos *GAMLSS*, é definido de um modo ligeiramente diferente do que nos Modelos Lineares Generalizados (STASINOPOULOS et al., 2017)

$$D_{GAMLSS} = -2\log(\hat{L}_c),$$

em que $\hat{L}_c$ é a função de verossimilhança estimada do modelo em estudo e D_{GAMLSS} doravante será denominada apenas por desvio global (*GDEV*).

Dois *GAMLSS* paramétricos aninhados $\mathcal{M}_0$ e $\mathcal{M}_1$, com desvios globais $GDEV_0$, $GDEV_1$ e graus de liberdade dos resíduos df_{e0} e df_{e1} respectivamente, podem ser comparados pelo teste da razão de verossimilhanças generalizado (TRVG), dado por $\Lambda = GDEV_0 - GDEV_1$ o qual, se certas condições de regularidade são atendidas, segue uma distribuição $\chi^2_{(df_{e0}-df_{e1})}$.

A comparação entre dois *GAMLSS* quaisquer pelo Critério de Informação de Akaike (AIC) assume a expressão

$$\text{GAIC}_\kappa = -2\log(\hat{L}_c) + (\kappa \times df),$$

em que *df* são os graus de liberdade efetivos do modelo proposto (*i.e.* o número efetivo de parâmetros) e κ é uma penalização aplicada para cada grau de liberdade utilizado tal que, com $\kappa = 2$, a medida resultante será a do Critério de Informação de Akaike (AIC) e tomando-se

$\kappa = \log(n)$ a medida será a do Critério de Informação Bayesiano (BIC). A escolha entre dois modelos ajustados que possuam *GAIC* ligeiramente diferentes, suscita dúvida sobre a significância associada à diferença existente entre os valores daquela métrica.

Os testes propostos por Vuong (1989) e Clarke (2007) são baseados na razão de verossimilhança para uma seleção de modelos baseados no critério de informação de Kullback-Leibler e podem ser utilizados para a comparação de dois modelos não necessariamente aninhados. Esses testes encontram-se implementados computacionalmente na função VC.test() da biblioteca *gamlss* (STASINOPOULOS et al., 2022).

A estatística do teste de Vuong segue assintoticamente uma distribuição Normal padrão e a hipótese nula é de que não há diferença estatisticamente significante entre os dois modelos.

A estatística do teste de Clarke segue assintoticamente uma distribuição binomial com parâmetro 0,5 e, se não há diferença estatisticamente significativa entre os dois modelos, as razões de log-verossimilhança das observações devem ser distribuídas uniformemente em torno de zero e cerca de metade das razões deve ser maior que zero.

Stasinopoulos et al. (2017) mencionam outros modos de se promover a escolha de modelos, como pela divisão do conjunto de dados em dois subconjuntos: treinamento e teste. Os modelos são então novamente estimados usando o conjunto de dados destinado ao treinamento e, em seguida, têm sua capacidade preditiva aferida por alguma métrica quando aplicados na estimação dos dados do conjunto destinado a teste. A subdivisão pode se dar uma única vez, no início do procedimento, quando então se arbitram as proporções dos dados que irão compor cada um dos dois subconjuntos (frequentemente escolhem-se proporções como 80 % para treinamento e 20 % para teste).

Já no procedimento conhecido como validação cruzada (*k-fold cross validation*), o parâmetro k, que se refere ao número de subconjuntos que serão gerados aleatoriamente a partir do conjunto original dos dados, é arbitrado (frequentemente adota-se $k = 10$). Para cada um dos k-grupos, o modelo é estimado sobre os dados formados pelos grupos restantes e avaliado sobre o grupo destacado para teste, assinalando-se alguma métrica ao resultado dessa etapa. Tal processo é repetido para todos os k-grupos e, ao final, um resumo dos valores da métrica adotada em cada uma das avaliações pode ser apresentado para análise da *performance* geral do modelo. É um método bastante popular porque além de ser de fácil compreensão, geralmente resulta em uma estimativa menos tendenciosa ou menos otimista da capacidade do modelo do que outros métodos como o mencionado no parágrafo anterior.

Há também na implementação computacional dos *GAMLSS* a função Rsq() que calcula um pseudo coeficiente de determinação (R^2). O coeficiente de determinação está bem estabelecido na análise de modelos clássicos de regressão e sua definição como a proporção da variabilidade explicada pelo modelo de regressão a torna útil como uma medida de sucesso em se predizer a variável resposta a partir de um conjunto de covariáveis explicativas.

Nagelkerke (1991) apresenta uma definição do R^2 para modelos mais gerais ligeiramente diferente daquela proposta por (COX; SNELL, 1989).

$$R^2 = 1 - \exp\left[-\frac{2}{n}\left\{l(\hat{\beta}) - l(0)\right\}\right] = 1 - \left\{\frac{L(0)}{L(\hat{\beta})}\right\}^{\frac{2}{n}},$$

em que $l(\hat{\beta}) = \log\ L(\hat{\beta})$ e $l(0) = \log\ L(0)$ são, respectivamente, os logaritmos das funções de verossimilhança para o modelo ajustado e nulo.

3.7.6 Inferência

Stasinopoulos et al. (2017) apresentam uma lista de questões inferenciais que podem ser suscitadas pelo pesquisador:

- sobre valores de parâmetros da distribuição da resposta estimados como constantes;
- coeficientes dos termos paramétricos presentes nos preditores do modelo;
- funções suavizadoras: seu perfil, erros padrão e parâmetros de suavização associados;
- predição de valores futuros dos parâmetros μ, σ, ν e τ da distribuição da resposta para novas observações;
- predição da distribuição assumida para uma resposta futura;
- seleção de modelos.

As estimativas $\hat{\boldsymbol{\theta}}$ dos parâmetros de um *GAMLSS* são obtidas pela maximização do logaritmo da função de verossimilhança (penalizada, quando na presença de funções suavizadoras). Assim, espera-se que para um modelo validado por uma criteriosa análise diagnóstica dos resíduos, elas apresentem propriedades (COX; HINKLEY, 1974 apud STASINOPOULOS et al., 2017) que assegurem que inferências realizadas com estes resultados possam ter um nível de significância associado corretamente dimensionado.

3.7.6.1 Intervalos de confiança baseados em erros padrão robustos (Huber *sandwich*)

O método usual de estimação dos erros padrão de um coeficiente é expresso por

$$\text{Var}(\hat{\beta}) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}, \tag{3.27}$$

em que $\boldsymbol{X}$ é a matriz de delineamento e $\boldsymbol{\Omega} = \sigma^2\boldsymbol{I}_n$.

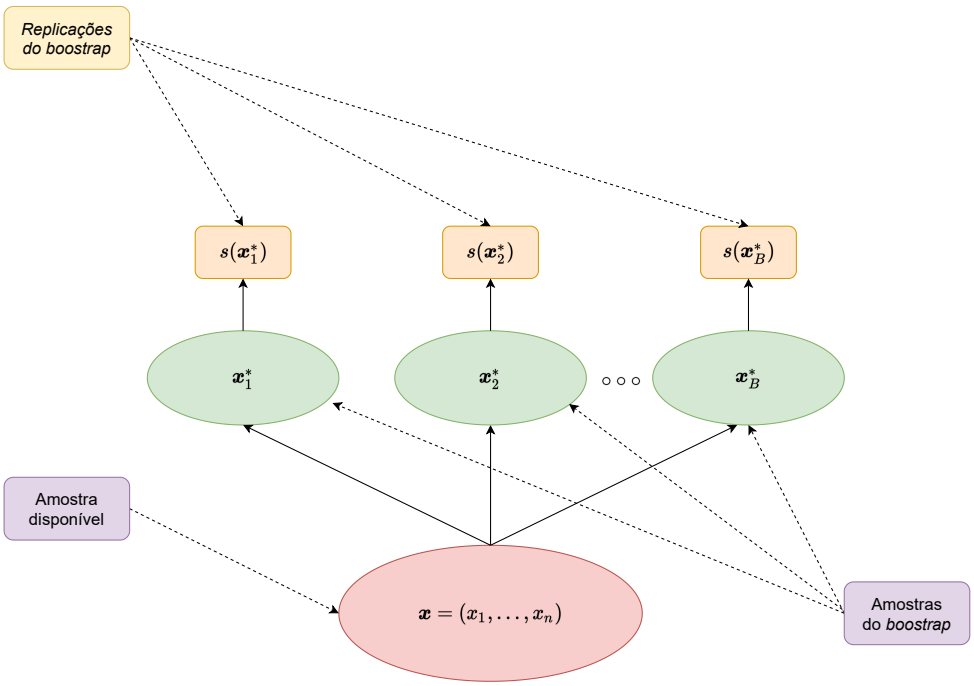
Sob a suspeita de que a relação funcional adotada para a variância não é a correta (como a presença de heterocedasticidade ou autocorrelação serial nos resíduos em

um Modelo Linear), os erros padrão robustos propostos por Huber (1967) e White (1980) substituem o núcleo da expressão[3] (3.27) e, em geral, resultam em estimativas mais confiáveis e, geralmente, mostram-se superiores aos estimados convencionalmente .

3.7.6.2 Inferências baseadas em *bootstrapping*

O *bootstrap* é uma técnica estatística que se vale da capacidade computacional de conduzir um processo intensivo de reamostragem e recálculo, com o propósito de extrair informações sobre alguma característica da distribuição de uma amostra original finita, ou dos parâmetros estimados por um modelo ajustado sobre essa. Um dos propósitos da teoria do *bootstrap* é produzir bons intervalos de confiança automaticamente (EFRON; TIBSHIRANI, 1993). Admita para tanto uma estatística de interesse $s(\boldsymbol{x})$ como ilustrado na Figura 3.9:

Figura 3.9: Esquema geral de um processo de *bootstrap*



Fonte: Adaptado de Efron e Tibshirani (1993)

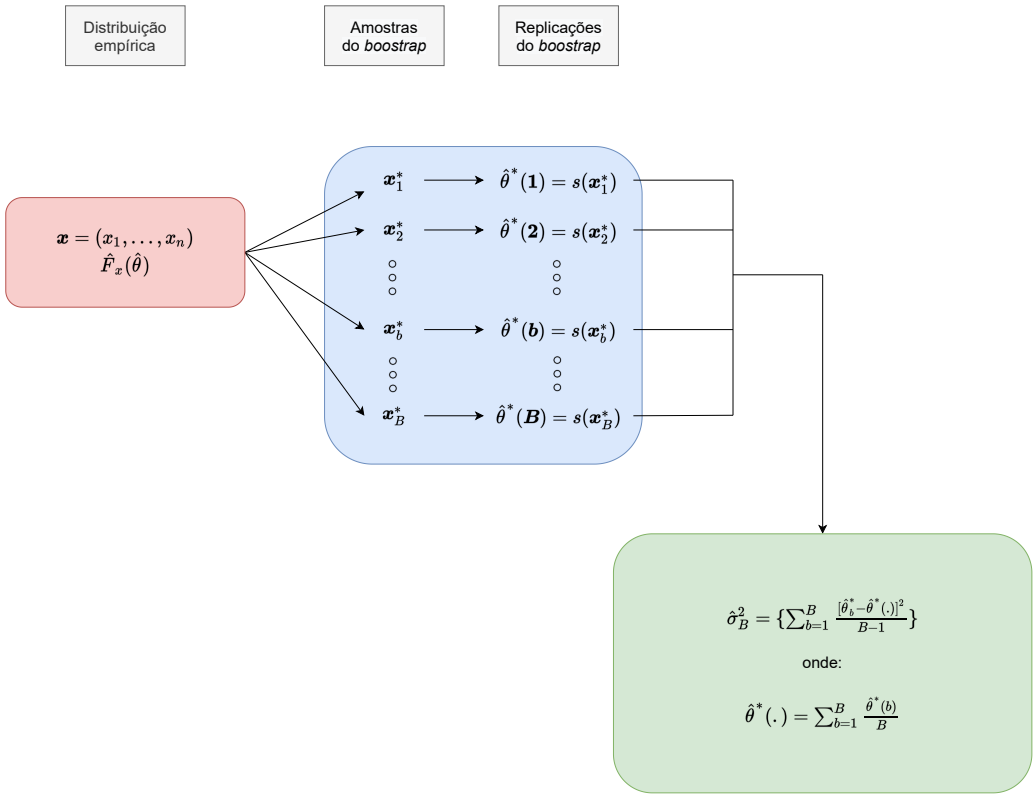
A construção de um intervalo de confiança exposto de modo esquemático na Figura 3.10 para $s(\boldsymbol{x})$ pode ser feito a partir do desvio padrão do processo de *bootstrap* realizado sobre a amostra finita $\boldsymbol{x} = (x_1, \ldots, x_n)$ pela geração de B amostras de tamanho n (com reposição) da amostra original $\boldsymbol{x}$ (as amostras do *bootstrap*) e o cálculo da estatística $s(\boldsymbol{x})$ de interesse em cada uma delas (as replicações do *bootstrap*)

[3] A expressão (3.27) é frequentemente denotada como um *sandwich* pois apresenta estruturas idênticas lado-a-lado de seu interior (*meat*)

$$\hat{\sigma}_B^2 = \sum_{b=1}^{B} \frac{[s(\boldsymbol{x}_b^*) - s(.)]^2}{(B-1)},$$

em que $s(.) = \sum_{b=1}^{B} \frac{s(\boldsymbol{x}_b^*)}{B}$. Valores típicos sugeridos para B situam-se na faixa de 50 a 200 (EFRON; TIBSHIRANI, 1993).

Figura 3.10: Algoritmo de um processo de *bootstrap* para estimação do desvio padrão $\hat{\theta} = s(\boldsymbol{x})$



Fonte: Adaptado de Efron e Tibshirani (1993)

Um intervalo de confiança de uma estatística de interesse $\hat{\theta}$ estimada pode ser aproximado a partir de seu desvio padrão $\hat{\sigma}$ e do valor da estatística "t" de *Student* para $(n-1)$ graus de liberdade, em que n é o tamanho da amostra processada. Esse intervalo é derivado da pressuposição de que a distribuição de uma estatística T é tal que

$$T = \frac{(\hat{\theta} - \theta)}{\hat{\sigma}} \sim \mathcal{N}(0,1),\ n \to \infty, \tag{3.28}$$

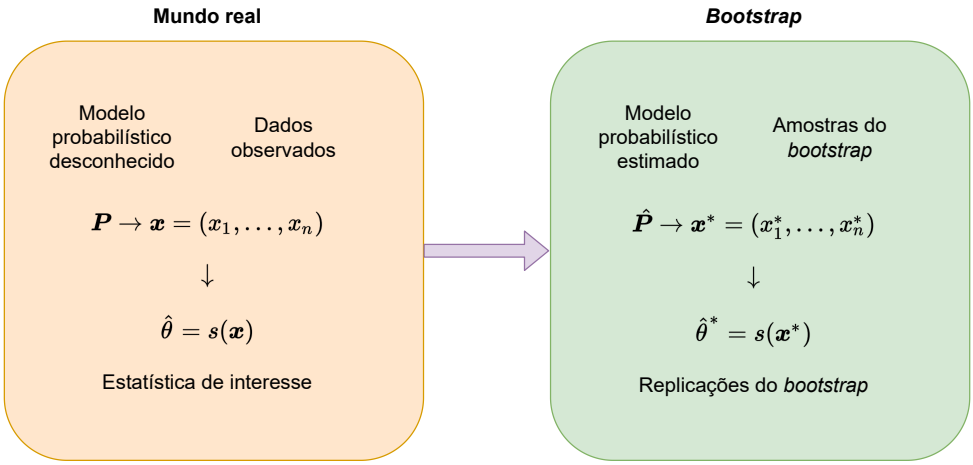
que pode ser aproximada para pequenas amostras por

$$T = \frac{(\hat{\theta} - \theta)}{\hat{\sigma}} \sim t_{n-1},$$

que resulta no intervalo de confiança dado por

$$IC_t = [\hat{\theta} - t_{(n-1),(1-\alpha)}\hat{\sigma}, \hat{\theta} - t_{(n-1);(\alpha)}\hat{\sigma}].$$

Figura 3.11: Relação entre a estrutura real dos dados e a estrutura do *bootstrap*



Fonte: Adaptado de Efron e Tibshirani (1993)

Um intervalo de confiança "t" de *bootstrap* (EFRON; TIBSHIRANI, 1993) pode ser construído sem a necessidade de se verificar a pressuposição assumida em (3.28) uma vez que a distribuição de T passa a ser estimada diretamente dos dados como mostra a Figura 3.11.

Para cada uma das B amostras *bootstrap* $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_B^*$ calcula-se $\hat{\theta}^*(b) = s(\boldsymbol{x}_b^*)$ (o valor $\hat{\theta}$ para a *b-ésima* amostra *bootstrap*) e seu desvio padrão $\hat{\sigma}^*(b)$. Assim a estatística $T^*(b)$ é tal que

$$T^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\hat{\sigma}^*(b)}.$$

O valor de $T^*(b)$ correspondente ao α-ésimo percentil determina o valor $\hat{t}_\alpha^*$ tal que

$$\frac{\{T^*(b) \leq \hat{t}_\alpha\}}{B} = \alpha,$$

e o intervalo de confiança será dado por

$$IC_{\hat{t}} = [\hat{\theta} - \hat{t}_{(1-\alpha)}\hat{\sigma}; \hat{\theta} - \hat{t}_{(\alpha)}\hat{\sigma}].$$

Intervalos de confiança podem ser estabelecidos diretamente com base nos percentis da distribuição das replicações do *bootstrap*.

Em uma situação ideal de *bootstrap* na qual o número de amostras do *bootstrap* $B \rightarrow \infty$ fica assegurado que o uso de percentis da distribuição de $\hat{\theta}^*$ para determinação de um intervalo de confiança para $\hat{\theta}$ se verifica e assim, definindo-se $\hat{G}$ como

a distribuição acumulada das replicações do *bootstrap* $\hat{\theta}^*$, o intervalo percentil estabelecido por α e $(1-\alpha)$ de $\hat{G}$ será (EFRON; TIBSHIRANI, 1993)

$$[\hat{\theta}_{inf}, \hat{\theta}_{sup}] = [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)].$$

Como, por definição, $\hat{G}^{-1}(\alpha) = \hat{\theta}^*$ (o α-percentil da distribuição *bootstrap*), o intervalo anterior pode ser expresso como

$$[\hat{\theta}_{inf}, \hat{\theta}_{sup}] = [\hat{\theta}^*_{\alpha}, \hat{\theta}^*_{(1-\alpha)}].$$

Para análises que envolvam modelos de regressão (sejam lineares ou não) há dois modos de se aplicar a técnica de *bootstrap* e a escolha por um ou outro procedimento irá depender dos pressupostos estabelecidos pela natureza do modelo envolvido:

- reamostrando os resíduos gerados pelo modelo; ou,
- reamostrando os *pares*: resposta e o vetor de covariáveis associadas.

Efron e Tibshirani (1993) afirmam que, em geral, o *bootstrap* sobre os pares é menos sensível às hipóteses assumidas pelo modelo de regressão do que o *bootstrap* sobre os resíduos.

Especificamente para os *GAMLSS*, Stasinopoulos et al. (2017) propõem três modos de se aplicar a técnica de *bootstrap* assim nominados: paramétrico; semiparamétrico e não paramétrico.

Considere $\mathcal{M}_0$ como um *GAMLSS* ajustado, no qual a distribuição estimada é dada por $\mathcal{D}(\boldsymbol{y}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\tau}})$, com $\mathbf{y}$ sendo o vetor da variável resposta observada e $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}$ os vetores das estimativas dos parâmetros dessa distribuição. Se $\hat{\boldsymbol{y}}$ for o vetor das estimativas realizadas com o modelo $\mathcal{M}_0$, pode-se definir $\mathbf{r}$ como o vetor dos resíduos quantílicos Normalizados ($\boldsymbol{r} = \Phi^{-1}(u)$, com $\boldsymbol{u} = \mathcal{F}(\boldsymbol{y}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\tau}})$). Denotando a amostra original como D_0 e por $D_{[b]}$ a *b-ésima* amostra *bootstrap* de D_0, teremos associado a esta última um vetor das repostas observadas $\boldsymbol{y}_{[b]}$ e das estimativas $\hat{\boldsymbol{y}}_{[b]}$.

O processo paramétrico de *bootstrap* fica definido como sendo a reestimação do modelo $\mathcal{M}_0$ com base na *b-ésima* amostra $D_{[b]}$ *bootstrap* com as respostas $y_{[b]i}$ dadas pela operação $s(.)$, definida pela função geratriz aleatória $\mathcal{D}_r$ de números da distribuição $\mathcal{D}$ adotada para a resposta $\mathbf{y}$ sob os parâmetros $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}$ estimados por $\mathcal{M}_0$ tal que

$$y_{[b]i} = \mathcal{D}_r(\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)).$$

O modelo $\mathcal{M}_{[b]}$, referente à *b-ésima* amostra *bootstrap*, pode então ser estimado, e qualquer métrica de interesse ser levantada.

O processo não paramétrico de *bootstrap* é definido como sendo a reestimação do modelo $\mathcal{M}_0$ com base na *b-ésima* amostra $D_{[b]}$ *bootstrap*, obtida de modo aleatório e com reposição sobre a amostra original, D_0 possibilitando a geração do modelo $\mathcal{M}_{[b]}$.

Já no processo semiparamétrico de *bootstrap*, tal como definido por Stasinopoulos et al. (2017), calcula as probabilidades associadas aos resíduos quantílicos Normalizados $r_{[b]i}$ da *b-ésima* amostra *bootstrap* via função de distribuição acumulada da distribuição Normal ($\hat{u}_{[b]i} = \Phi(r_{[b]i})$). Os quantis associados da distribuição adotada pela resposta tornam-se as respostas $y_{[b]i}$ simuladas da *b-ésima* amostra *bootstrap*

$$y_{[b]i} = \mathcal{F}^{-1}(\hat{u}_{[b]i}|\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i).$$

O modelo $\mathcal{M}_{[b]}$, referente à *b-ésima* amostra *bootstrap*, pode então ser estimado, e qualquer métrica de interesse ser levantada.

3.7.7 Diagnóstico

É certo que todo o processo de investigação de prováveis modelos pode ser facilitado pelo uso de funções embutidas nas mais variadas implementações computacionais disponíveis que automatizam, de algum modo, a busca do melhor subconjunto de covariáveis adotando alguma métrica para comparação, coma a função stepGAIC() da biblioteca *gamlss*, por exemplo. Todavia, a ação discricionária do pesquisador na análise individual dos efeitos de cada termo e na forma como são incorporados aos preditores do modelo ainda está um tanto longe de ser, de todo, suplantada.

3.7.7.1 Resíduos quantílicos (aleatorizados) Normalizados

Resíduos nada mais são do que as diferenças entre os valores estimados por um modelo e os valores observados. Diferentes construções buscam transformar essa diferença tentando incorporar boas propriedades para tornar mais robusta sua utilização no diagnóstico do modelo (frequentemente para estabilizar a variância ou induzir simetria na sua distribuição, a fim de garantir sua comparabilidade e possibilitar a detecção de valores discrepantes) resultando em vários tipos de resíduos, sendo os mais comuns:

- simples (ordinários);
- de Pearson;
- de Pearson “estudentizados” internamente;
- componentes do desvio;
- componentes do desvio “estudentizados” internamente.

O diagnóstico de um *GAMLSS* é feito, fundamentalmente, pela análise dos resíduos quantílicos (aleatorizados) Normalizados (STASINOPOULOS et al., 2017).

Resíduos quantílicos (aleatorizados) possuem a característica muito atraente de que, independentemente da distribuição adotada para a variável resposta, os verdadeiros

resíduos sempre terão uma distribuição Normal padronizada quando o modelo proposto estiver bem ajustado. São obtidos invertendo-se a distribuição ajustada para cada valor de resposta e determinando-se o quantil Normal padrão equivalente (DUNN; SMITH, 1996).

Admita que a $F(y|\boldsymbol{\theta})$ seja a função de distribuição acumulada de uma variável aleatória *Y*. Se *Y* é contínua, o Teorema da Inversa da Função de Distribuição assegura que $U_i = F(y_i|\theta_i)$ é uniformemente distribuída no intervalo $[0, 1]$. Assim, o resíduo quantílico Normalizado de um *GAMLSS* estimado é dado por

$$r_i^q = \Phi^{-1}\left(F\left(y_i|\hat{\boldsymbol{\theta}}_i\right)\right) \quad i = 1, \ldots, n,$$

em que $\Phi^{-1}(.)$ é a função quantílica da distribuição Normal padronizada; $F(y|\hat{\boldsymbol{\theta}})$ é a função de distribuição estimada e $\hat{\boldsymbol{\theta}}_i = (\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)^T$ são as estimativas dos parâmetros do modelo.

Os resíduos quantílicos Normalizados são calculados de maneira diferente se *Y* é uma variável aleatória discreta. Neste caso a variável *Y* é transformada em uma nova variável aleatória *U*, com distribuição uniforme no intervalo (u_1, u_2), e um valor observado y_i será mapeado para um valor u_i compreendido entre

$$u_{1i} = F(y_i - 1|\hat{\boldsymbol{\theta}}_i);$$
$$u_{2i} = F(y_i|\hat{\boldsymbol{\theta}}_i).$$

Nesse intervalo, u_i é aleatoriamente escolhido e o resíduo quantílico Normalizado dado por

$$r_i^q = \Phi^{-1}(u_i).$$

Considerando que a verificação dos pressupostos de um modelo pela Normalidade de seus resíduos está bem estabelecida na literatura, os resíduos quantílicos Normalizados permitem uma maneira de verificar a adequação de um *GAMLSS* ajustado por meio de técnicas gráficas bem familiares:

- resíduos *versus* valores ajustados do parâmetro μ;
- resíduos *versus* alguma ordem de coleta, variável explicativa específica considerada no modelo ou ainda alguma variável não incluída no modelo;
- a densidade dos resíduos;
- gráfico de probabilidade dos quantis da distribuição Normal padronizada;

Se os parâmetros do modelo proposto foram consistentemente estimados, então $r^q \sim \mathcal{N}(0, 1)$, aproximadamente simétrica e mesocúrtica.

3.7.7.2 *Worm plot* dos resíduos quantílicos Normalizados

Worm plots (literalmente gráficos de "minhoca") foram introduzidos por Van Buuren e Fredriks (2001) como uma ferramenta gráfica auxiliar no diagnóstico de resíduos para análise de um modelo estatístico. Consistem em um gráfico de probabilidade quantil-quantil do qual foi removida a tendência.

Um gráfico dos quantis teóricos da distribuição Normal padronizada *versus* os quantis amostrais dos resíduos sem tendência difere do gráfico usual por apresentar uma linha horizontal que passa pela origem, representando os quantis unitários teóricos que se esperaria observar, se os dados fossem exatamente Normais.

A dispersão de pontos representa os resíduos ordenados. A magnitude e a direção do desvio dos quantis observados e teóricos é calculada subtraindo-se o quantil esperado do quantil observado e encontra-se assinalada no eixo vertical. Desse modo, se um ponto estiver abaixo da linha de tendência no gráfico quantil-quantil tradicional, ele aparecerá acima da linha horizontal no gráfico sem tendência, posto o resíduo observado ser superior ao esperado. De modo acessório, podem ser acrescentadas faixas que delimitem um intervalo de confiança, construídas sob determinado nível de significância.

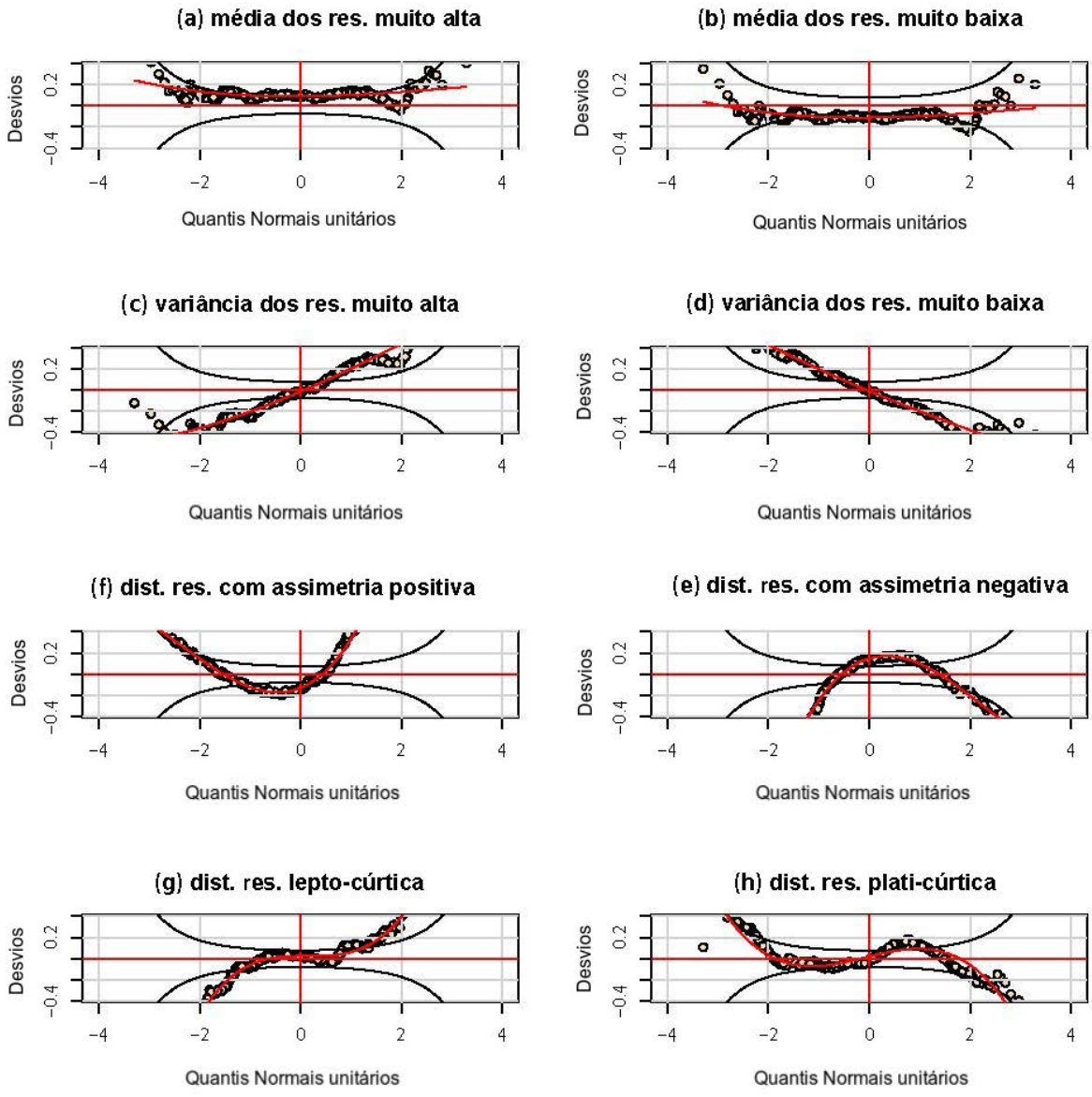
Situações em que se observa uma grande quantidade de pontos situados fora das bandas de confiança ou que o perfil da dispersão mostre algum padrão sistemático de afastamento da linha horizontal, indicam que a distribuição ajustada ou os termos considerados no ajuste do modelo são inadequados para explicar a variável resposta sob análise. Pontos isolados que se situem muito afastados das faixas de confiança devem ser atenciosamente analisados sob a suspeita de serem *outliers*.

Depreende-se assim que, se a média dos resíduos quantílicos está muito alta, o preditor do parâmetro da locação da distribuição adotada deve ser mais flexível, o que pode ser obtido pela inclusão de outros termos ou pela reconsideração sob outra forma dos termos presentes.

Van Buuren e Fredriks (2001) propuseram quantificar alguns aspectos básicos do *worm plot* ajustando um polinômio cúbico $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$ (em que $\boldsymbol{Y}$ e $\boldsymbol{X}^T$ representam os vetores das observações e dos quantis observados) aos dos pontos do gráfico. O perfil assumido por essa curva ajustada sugere variadas inadequações no modelo e associaram os valores absolutos dos coeficientes estimados $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\beta}_3$, os termos constante, linear, quadrático e cúbico do ajuste polinomial promovido aos pontos do *worm plot*, respectivamente como desajustamentos do modelo na estimação da locação, variabilidade, simetria e curtose em relação ao modelo teórico adotado. Como valores limite, em módulo, sugeriram 0,10; 0,10; 0,05 e 0,03.

A Figura 3.12 ilustra, com uma linha vermelha, o polinômio cúbico em cada uma das situações nas quais o ajuste do modelo não se mostra adequado e o Quadro 3.1 apresenta um diagnóstico para cada uma dessas possíveis situações.

Figura 3.12: Padrões sistemáticos de afastamento dos resíduos quantílicos da linha horizontal de referência em um *worm plot*



Fonte: Adaptado de Stasinopoulos et al. (2017)

Quadro 3.1: Interpretação dos vários padrões sistemáticos de afastamento da linha horizontal da curva ajustada de um *worm plot* mostrados na Figura 3.12

Perfil	Situação	Resíduos quantílicos	Diagnóstico do ajuste do modelo
intercepto	acima da origem	média muito alta	a locação ajustada está baixa
	abaixo da origem	média muito baixa	a locação ajustada está alta
inclinação	positiva	variância muito alta	a escala ajustada é muita baixa
	negativa	variância muito baixa	a escala ajustada é muito alta
parábola	concavidade p/cima	assimetria positiva	o coef. de assimetria ajustado é muito baixo
	concavidade p/baixo	assimetria negativa	o coef. de assimetria ajustado é muito alto
curva em "S"	extrem. esq. para baixo	leptocurtose	as caudas da distribuição ajustada são muito leves
	extrem. esq. para cima	platicurtose	as caudas da distribuição ajustada são muito pesadas

Fonte: Adaptado de Stasinopoulos et al. (2017)

3.7.7.3 Estatísticas Q e Z dos resíduos quantílicos (aleatorizados) Normalizados

A Normalidade dos resíduos pode também ser analisada por meio das estatísticas Q e Z (STASINOPOULOS et al., 2017). Para tanto considere que uma determinada variável explicativa contínua, presente no modelo, seja subdividida em G grupos e que $\{r_{gi}, i = 1, \ldots, n_i\}$ sejam os resíduos do *g-ésimo* grupo formado, tendo média $\bar{r}_g$ e desvio padrão s_g para $g = 1, \ldots, G$.

As estatísticas Z_{g1}, Z_{g2}, Z_{g3} e Z_{g4} podem ser calculadas e utilizadas para se verificar se os resíduos do *g-ésimo* grupo possuem média populacional 0, variância 1, distribuição simétrica (coeficiente de assimetria igual a 0) e mesocúrtica (coeficiente de curtose igual a 3); ou seja, se possuem uma distribuição Normal padronizada.

As duas primeiras são dadas por (ROYSTON; WRIGHT, 2000)

$$Z_{g1} = n_g^{\frac{1}{2}} \bar{r}_g$$

e

$$Z_{g2} = \frac{s_g^{\frac{2}{3}} - [1 - \frac{2}{(9n_g - 9)}]}{\{\frac{2}{(9n_g - 9)}\}^{\frac{1}{2}}}.$$

D'Agostino, Belanger e D'Agostino Jr. (1990) derivaram as duas últimas $Z(\sqrt{b_1})$ e $Z(b_2)$ (respectivamente pelas expressões [13] e [19] nas páginas do citado trabalho) que juntas formam a estatística K^2 de D'Agostino para um teste de hipóteses unilateral conjunto quanto à assimetria e ao excesso de curtose pela medida

$$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2),$$

em que $\sqrt{b_1}$ e b_2 são as estimativas amostrais de $\sqrt{\beta_1}$ e β_2 propostas por (PEARSON, 1895) como

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}}$$

e

$$b_2 = \frac{m_4}{m_2^2},$$

sendo $m_k = \sum (X_i - \bar{X})^k / n$.

A estatística Q é assim expressa como

$$Q_j = \sum_{g=1}^{G} Z_{gj}^2 \; j = 1, 2, 3, 4.$$

Royston e Wright (2000) discorrem que a distribuição da estatística Q para um teste de hipóteses sob a premissa de que os verdadeiros resíduos provêm de uma distribuição Normal é

$$Q_1 \sim \chi^2_{(G-df_\mu)};$$
$$Q_2 \sim \chi^2_{(G-\frac{(df_\sigma+1)}{2})};$$
$$Q_3 \sim \chi^2_{(G-df_\nu)}$$
e
$$Q_4 \sim \chi^2_{(G-df_\tau)},$$

em que df_μ, df_σ, df_ν e df_τ são os graus de liberdade dos preditores dos parâmetros da distribuição adotada para a resposta. Valores significantes das estatísticas Q_1, Q_2, Q_3 e Q_4 seriam indicativos de inadequações nos preditores de μ, σ, ν e τ do modelo ajustado, contornáveis possivelmente pelo aumento dos graus de liberdade das funções suavizadoras consideradas em cada um desses preditores.

Uma representação gráfica que mostre as estatísticas Z^2_{gj}, em linhas para cada um dos g grupos e colunas (estatística Z_j associada, facilita a visualização da contribuição dada por cada grupo à estatística Q_j, sugerindo quais grupos da variável não estão bem ajustados) (STASINOPOULOS et al., 2017).

4 MATERIAIS E MÉTODOS

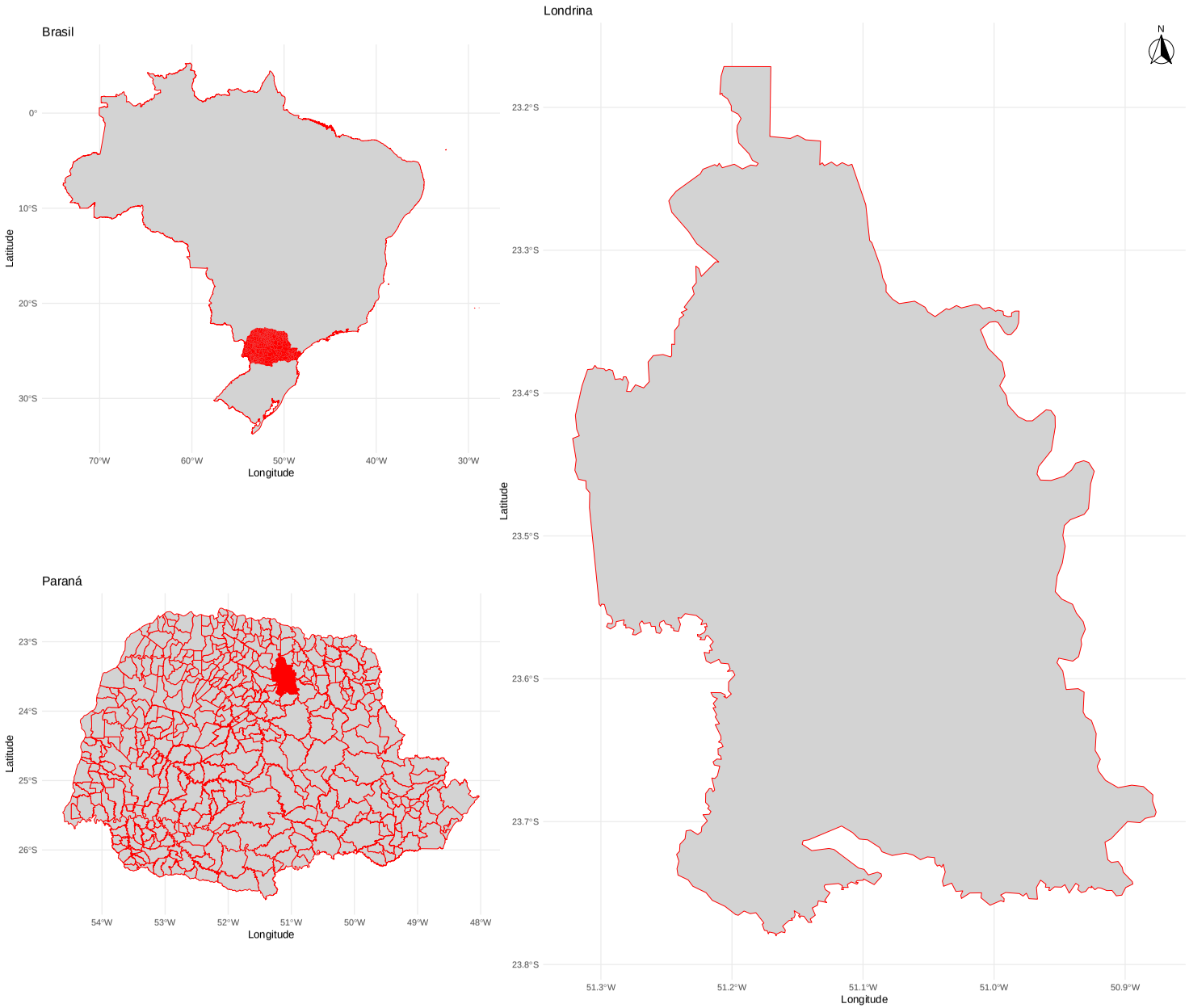
4.1 MATERIAIS

O conjunto de dados utilizado nesse trabalho pertence à base de informações imobiliárias de uso profissional do autor e é composto por 746 informações imobiliárias referentes a terrenos localizados no perímetro urbano - efetivamente transacionados ou simplesmente ofertados a mercado - coletadas no período compreendido entre maio de 1995 e março de 2021 no município de Londrina, cuja localização geográfica está ilustrada na Figura 4.1.

Destacamos, preliminarmente, que alguns elementos pertencentes ao conjunto de dados foram suprimidos em razão de se referirem a terrenos com uma área total muito grande, superior a 50.000 m^2 e, por essa razão, dificilmente poderiam ser caracterizados como lote urbano. Assim, a amostra efetivamente utilizada é composta por 730 elementos.

Os dados analisados podem ser obtidos na seção Dados no repositório GitHub do autor em: e, no Apêndice C.1 do fascículo complementar disponível na seção Apêndices no mesmo repositório, estão disponibilizadas informações complementares acerca da colonização, história e desenvolvimento de Londrina.

Figura 4.1: Localização geográfica de Londrina



Fonte: Próprio autor

4.1.1 Natureza da informação obtida

Informações sobre dados imobiliários meramente ofertados ao mercado são abundantes e podem ser facilmente reunidas por diversos meios de pesquisa. Consulta direta ao anunciante, proprietário ou empresa imobiliária cujo nome é ostentado nas placas tradicionalmente afixadas nos terrenos. Busca em anúncios publicados na mídia impressa ou em domínios virtuais mantidos por empresas imobiliárias junto à rede mundial de computadores, nos quais pode-se realizar a pesquisa sob vários critérios como, por exemplo, estado, cidade, bairro de localização, dimensão.

Por outro lado e, de modo totalmente diverso da anterior, informações sobre transações efetivamente realizadas são obtidas apenas por meio de diligente trabalho investigativo junto às partes diretamente envolvidas em uma negociação realizada: comprador e vendedor (ou intermediador). O simples levantamento de informações arquivadas em Cartório de Registros Imobiliários, por serem meras transcrições de escrituras públicas lavradas sob a vontade das partes envolvidas, nem sempre refletem (por motivos que não pertencem ao escopo desse trabalho) os reais termos acordados na negociação e modelos estimados com esses dados nem sempre são úteis para o propósito de se avaliar o valor de mercado (ZYGA, 2017).

4.1.2 Descrição

Cada um dos elementos da amostra se encontra identificado tanto em termos cadastrais, quanto por suas caraterísticas estruturais físicas, temporais, espaciais e econômicas. O Quadro 4.1 apresenta a identificação cadastral associada a cada informação coletada. Estando os dados disponíveis para acesso público no repositório já informado, por razões de sigilo, as colunas que apresentam informações sobre o nome e telefone dos informantes foram suprimidas. A Figura 4.2 ilustra a disposição espacial de cada elemento da amostra.

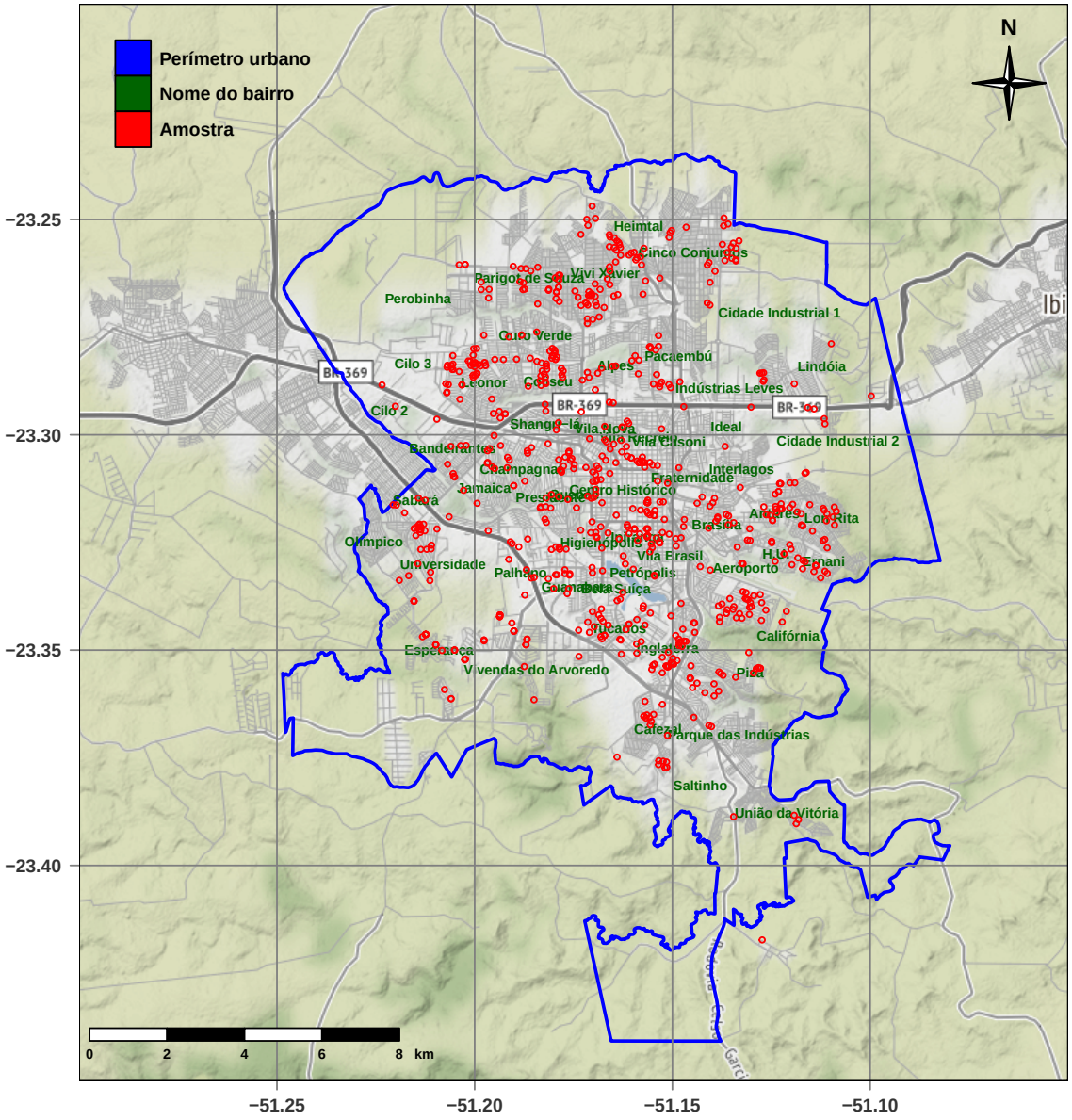
Quadro 4.1: Informações cadastrais dos elementos pesquisados

Abreviatura	**Descrição**
REF	identificação única do elemento amostral no banco de dados
FONTE	origem da informação
DIVULGAÇÃO	veículo de divulgação da informação
INF	responsável pela informação (empresa ou pessoa física)
TEL	telefone do responsável pela informação
ENDEREÇO	nome e informações complementares do logradouro
BAIRRO	nome do bairro
SETOR	identificação do setor censitário (IBGE censo 2010)

Fonte: Próprio autor

Na Engenharia de Avaliações, mais especificamente para o caso da tipologia em estudo, a modelagem geralmente é promovida usando-se o valor unitário, simples quociente

Figura 4.2: Mapa municipal de Londrina com a localização de cada elemento amostral em Londrina (coordenadas 23,3197S;51,1662W)



Fonte dos dados: Próprio autor (imagem de base: Stamen Maps 2021)

do valor total informado pela área total do terreno. Assim, nesse trabalho, tem-se como variável resposta o valor unitário e os seguintes atributos de natureza estrutural, espacial, econômica e temporal:

- Valor: variável quantitativa contínua, expressa em R$ (Reais) e que corresponde ao preço pedido (ou valor transacionado) do elemento amostral;
- Área: variável quantitativa contínua, medida em m^2 (metros quadrados) e que corresponde à área do elemento amostral;

- Valor unitário: variável quantitativa contínua, expressa em R$/m^2 e resultante do quociente do valor pela área do elemento amostral;
- Data: variável quantitativa discreta, admitida contínua e que corresponde ao número de dias transcorridos entre a data de referência inicial adotada (01/01/1970=1) e a data na qual a informação do elemento amostral (Valor) foi obtida;
- Renda: variável quantitativa contínua, medida em R$ (Reais) e que corresponde à renda familiar média do setor censitário onde o elemento amostral se situa. Para dados com data anterior a 01/01/2010 foram usadas as informações do Censo de 2000 e para aqueles em data posterior a 31/12/2009, as informações do Censo 2010;
- Relevo: variável qualitativa que corresponde ao relevo do elemento amostral pesquisado para um referencial posicionado defronte a ele, no logradouro. Esse fator assume os seguintes níveis: plano (1), aclive (2) ou declive (3);
- Natureza: variável qualitativa que corresponde à natureza econômica da variável VALOR do elemento amostral. Esse fator assume os seguintes níveis: oferta (1) ou transação (2);
- Pavimentação: variável qualitativa que corresponde ao tipo de pavimentação existente defronte ao elemento amostral. Esse fator assume os seguintes níveis: outra (1) ou pav. asfáltica (2);
- Implantação: variável qualitativa que corresponde ao tipo de implantação do elemento amostral, se lote isolado ou na forma condominial, pertencente a um empreendimento dotado de uma estrutura de amenidades para seus moradores. Esse fator assume os seguintes níveis: lote isolado (1) ou condomínio (2);
- Coordenada geográfica X de localização: variável quantitativa contínua, medida em m (metro) e que correspondente à coordenada métrica longitudinal (UTM SAD69/EPSG 29192) da localização do elemento amostral;
- Coordenada geográfica Y de localização: variável quantitativa contínua, medida em m (metro) e que correspondente à coordenada métrica latitudinal (UTM SAD69/EPSG 29192) da localização do elemento amostral.

4.1.3 Georreferenciamento

Os dados de natureza espacial podem ser geograficamente identificados em relação à posição que ocupam por meio de suas coordenadas, num processo denominado georreferenciamento. O georreferenciamento identifica qualquer ponto localizado na superfície terrestre por meio de suas coordenadas expressas sob um sistema de referência. Nesse trabalho os dados encontram-se georreferenciados na projeção Universal Transversa de Mercator

(UTM). Por utilizar o metro (m) como medida, esse sistema também é conhecido como sistema de coordenadas planas ou métricas. Detalhes complementares podem ser obtidos no Apêndice C.2 do fascículo complementar disponível na seção Apêndices no repositório GitHub do autor em: .

4.1.4 Arquivos digitais de informações espaciais

Arquivos digitais com a extensão *shp* foram desenvolvidos e padronizados pela empresa *ESRI: Environmental Systems Research Institute, Inc.* como uma especificação aberta para interoperabilidade de dados entre seu próprio *software* de mapeamento e outros. Apresentam informações geoespaciais de dados em formato vetorial e possibilitam a manipulação desses dados em sistemas de informações geográficas (*GIS: Graphical information system*).

4.1.5 Setores censitários

Como já afirmado anteriormente, o propósito dessa dissertação é o da estimação dos parâmetros de um Modelo Hedônico de Regressão Espaçotemporal baseado na classe dos Modelos Aditivos Generalizados para Locação, Escala e Forma (*GAMLSS*) que contemple, de modo conjunto, a variabilidade espacial e temporal presente na variável resposta e dentre as variáveis explicativas pesquisadas encontra-se a "RENDA". As informações relativas a essa variável, referentes aos censos de 2000 e 2010, são acessíveis a partir dos arquivos digitais dos setores censitários de Londrina e planilhas de dados agregados, ambos disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em IBGE: agregados dos setores censitários (censo 2000); IBGE: agregados dos setores censitários do Paraná (censo 2010) e IBGE: documentação dos setores censitários (censo 2010). Detalhes complementares podem ser obtidos no Apêndice C.3 do fascículo complementar disponível na seção Apêndices no repositório GitHub do autor em: .

Ao leitor mais atento, não deve escapar que essa variável encontra-se relacionada tanto à localização quanto ao tempo. Todavia, os valores dessa variável resultam da tabulação dos dados coletados nos censos realizados apenas em 2000 e 2010, posto não ter ocorrido, por razões de saúde pública, o censo programado para 2020 [1]. Assim, a incorporação dessa variável ao modelo resultaria na introdução de distorções não só de natureza temporal por estar restrita a períodos específicos, como também espacial, por admitir como homogênea a renda nos setores censitários. Por essa razão, o modelo proposto não se vale dessa variável.

4.2 Recursos computacionais utilizados

Os cálculos computacionais requeridos nos procedimentos a seguir descritos foram realizados por meio de diversas bibliotecas escritas em R. O software R foi criado por

[1] Para compreensão da pandemia mundial de 2020 acessar Organização Mundial da Saúde.

Ross Ihaka e Robert entleman na Universidade de Auckland, Nova Zelândia e atualmente é mantido e desenvolvido por R Core Team (2021).

De modo complementar à biblioteca principal gamlss, a biblioteca gamlss.add com rotinas suplementares foi necessária. Essa biblioteca fornece uma interface para a função gam() da biblioteca mgcv na qual muitos suavizadores encontram-se implementados. O código da programação realizada pode ser acessado na seção Códigos no repositório GitHub do autor em: .

4.3 Roteiro geral da análise

Como exposto nas seções anteriores, na implementação computacional dos *GAMLSS* (e bibliotecas auxiliares) várias funções não lineares podem ser empregadas na suavização dos termos incorporados ao modelo de modo não paramétrico:

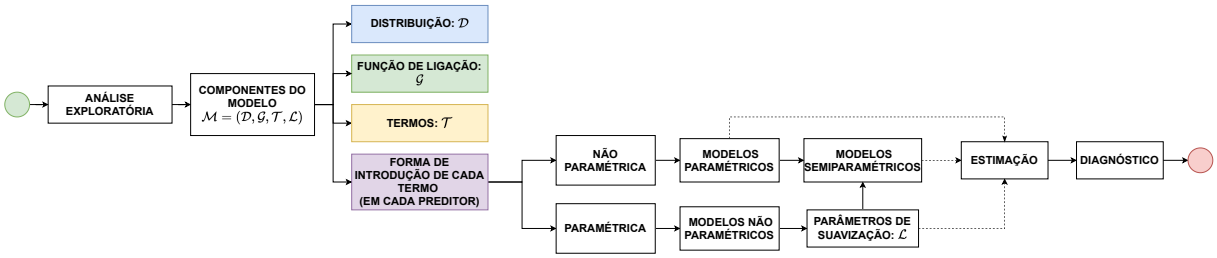
- curvas *LOESS* (CLEVELAND, 1979);
- *splines* cúbicos (GREEN; SILVERMAN, 1994);
- bases de *splines* (*basis splines*) (SCHOENBERG, 1945);
- *splines* penalizados (EILERS; MARX, 1996);
- *thin plate splines* (DUCHON, 1977);
- *tensor product splines* (DE BOOR, 2001).

Qualquer variável pode ser aditivamente incorporada sob a forma de efeitos aleatórios de modo conjunto, a outras, considerados parametricamente, em cada um dos preditores lineares dos parâmetros da distribuição adotada para o modelo (θ_k).

A distribuição para a variável resposta Y pode ser escolhida a partir de uma ampla gama de distribuições discretas e contínuas disponíveis na implementação computacional dos *GAMLSS*, incluindo distribuições altamente assimétricas, leptocúrticas ou platicúrticas (STASINOPOULOS et al., 2017).

A Figura 4.3 ilustra de modo esquemático o roteiro geral da análise realizada.

Figura 4.3: Fluxograma da análise



Fonte: Próprio autor

4.4 DESCRIÇÃO DAS VARIÁVEIS INCORPORADAS COMO TERMOS NOS PREDITORES DO MODELO

O Quadro 4.2 apresenta, resumidamente, a terminologia adotada para denominação das variáveis, bem como algumas de suas características.

Quadro 4.2: Resumo das variáveis utilizadas como termos do modelo

Variável	Código	Classificação I	Classificação II	Classificação III	Domínio	Faixa de valores (níveis)
Preço unitário	UNIT	Dependente	Quantitativa	Contínua	$\mathbb{R}_+^*$	R$ 5,00 - R$ 3.348,21
Área	AT	Independente	Quantitativa	Contínua	$\mathbb{R}_+^*$	157,00 m^2 - 43.088,00 m^2
Data	DATA	Independente	Quantitativa	(admitida) Contínua	$\mathbb{R}_+^*$	9.279-18.697 (1)
Relevo	RELEVO	Independente	Qualitativa	Nominal	-	Plano: 1 Aclive: 2 Declive: 3
Natureza	NATUREZA	Independente	Qualitativa	Nominal	-	Oferta: 1 Transação: 2
Pavimentação	PAVIMENTACAO	Independente	Qualitativa	Nominal	-	Outra: 1 Pav. asfáltica: 2
Implantação	IMPLANTACAO	Independente	Qualitativa	Nominal	-	Lote isolado: 1 Condomínio: 2
Coordenada geográfica X	UTM_X	Independente	Quantitativa	Contínua	$\mathbb{R}_+^*$	477.200- 489.853
Coordenada geográfica Y	UTM_Y	Independente	Quantitativa	Contínua	$\mathbb{R}_+^*$	7.410.324-7.429.181

(1): 29/05/1995-11/03/2021.

Fonte: Próprio autor

4.5 MÉTODOS

4.5.1 Escolha da distribuição e funções de ligação

Conforme apresentado no Quadro 4.2, a faixa de valores assumidos pela variável resposta "UNIT" ($\mathbb{R}_+^*$) restringe o leque para escolha de uma das distribuições teóricas ($\mathcal{D}$) e a implementação computacional oferece algumas funções na biblioteca *gamlss* para auxiliar essa escolha:

- fitDist(): esta função ajusta todas as distribuições paramétricas implementadas nesta classe de modelos especificadas em um de seus argumentos a um único vetor de dados (a variável resposta sem covariáveis). A distribuição marginal final pode ser selecionada pelo *GAIC* que, para $\kappa = 2$, retorna o AIC;
- chooseDist() : esta função ajusta todas as distribuições paramétricas implementadas nesta classe de modelos, especificadas em um de seus argumentos, a um determinado *GAMLSS* ajustado. A distribuição condicional pode ser selecionada pelo *GAIC* que, para $\kappa = 2$, retorna o AIC.

Enquanto a função fitDist() retorna as distribuições teóricas que melhor representam a resposta de modo não condicional às demais covariáveis, a função chooseDist() parte de um modelo ajustado sob uma distribuição teórica com a inclusão de covariáveis preliminarmente arbitradas para retornar a distribuição que melhor represente a resposta condicionada aos termos considerados, especificado *a priori* a faixa de valores possíveis para a resposta.

Para esta última simulação, o modelo protótipo considerado foi proposto sob a distribuição Normal para a resposta e os seguintes termos considerados para a estimação dos preditores dos dois parâmetros ($\theta_1 = \mu$ e $\theta_2 = \sigma$): "AT"; "NATUREZA"; "IMPLANTACAO"; "RELEVO"; "PAVIMENTACAO"; "UTM_X´, "UTM_Y" e "DATA". Os resultados obtidos para a métrica adotada na comparação estão expostos na Tabela 4.1 .

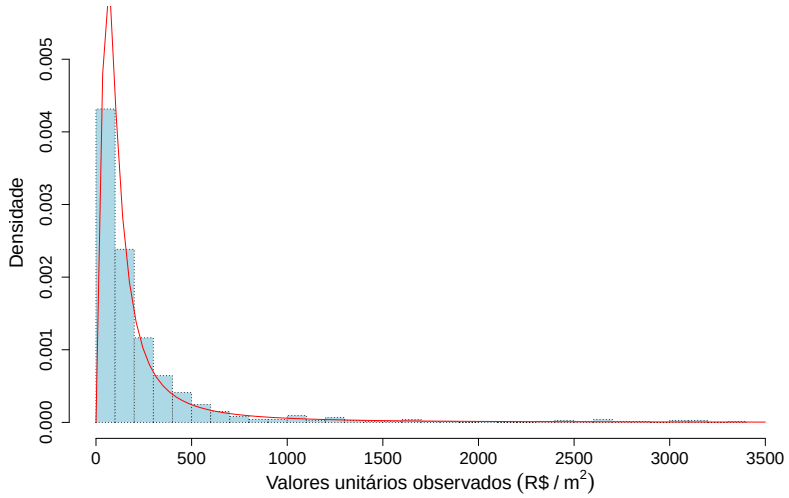
Tabela 4.1: Estatísticas das distribuições teóricas melhor ajustadas à variável resposta

Função	Distribuição teórica	*AIC*
fitDist()	Beta generalizada tipo 2 (GB2)	9.262,295
	Box&Cox "t" (BCTo)	9.264,840
chooseDist()	Box&Cox"t" (BCTo)	8.581,155
	Beta generalizada tipo 2 (GB2)	8.650,552

Fonte: Próprio autor

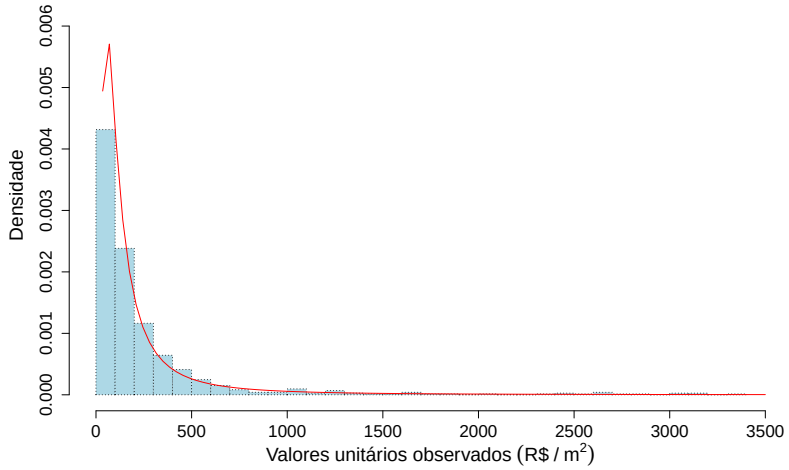
As Figuras 4.4 e 4.5 mostram as duas distribuições teóricas consideradas sobrepostas ao histograma da variável resposta (a rigor a distribuição marginal da variável resposta).

Figura 4.4: Distribuição marginal da resposta sob a distribuição teórica Box&Cox “t” (BCTo)



Fonte: Próprio autor

Figura 4.5: Distribuição marginal da resposta sob a distribuição Beta generalizada tipo 2 (GB2)



Fonte: Próprio autor

No processo de refinamento dos diversos modelos propostos, os resultados dos ajustes intermediários, medidos sob o *GAIC* e verificados liminarmente pelo *worm-plot* mostraram-se sempre favoráveis à distribuição BCTo, razão pela qual a escolha recaiu sobre essa distribuição teórica para a resposta.

A distribuição $\mathcal{D} = BCTo$, adotada para o modelo, permite que não apenas a estimação dos parâmetros da locação (mediana) e variabilidade (coef. de variação aproximado) ($\theta_1 = \mu$ e $\theta_2 = \sigma$, respectivamente) sejam modelados, como também sua forma no que tange à sua assimetria e curtose ($\theta_3 = \nu$ e $\theta_4 = \tau$), como apresentado na Tabela 3.1.

As funções de ligação adotadas para os quatro preditores dos parâmetros distribucionais do modelo tal que $\boldsymbol{Y} \overset{ind}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ estão indicadas em (4.1)

$$\begin{aligned} \boldsymbol{\eta}_1 &= g_1(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}); \\ \boldsymbol{\eta}_2 &= g_2(\boldsymbol{\sigma}) = \log(\boldsymbol{\sigma}); \\ \boldsymbol{\eta}_3 &= g_3(\boldsymbol{\nu}) = (\boldsymbol{\nu}); \\ \boldsymbol{\eta}_4 &= g_4(\boldsymbol{\tau}) = \log(\boldsymbol{\tau}), \end{aligned} \tag{4.1}$$

e, posto a função de ligação g_1 adotada no preditor linear η_1 do parâmetro μ ser a função logaritmo, na implementação computacional esta variante recebe a denominação "o" de original (BCTo).

4.5.2 Incorporação de variáveis nos preditores dos parâmetros distribucionais

Para a incorporação das variáveis explicativas como termos nos quatro preditores aditivos dos parâmetros da distribuição teórica, adotou-se o seguinte critério:

- variáveis qualitativas: incorporadas como fatores;
- variáveis quantitativas: incorporadas por meio de funções suavizadoras estimadas sobre bases apropriadas.

Desse modo, os fatores "RELEVO", "NATUREZA", "PAVIMENTAÇÃO" e "IMPLANTACAO" - *quando presentes em um preditor de algum dos parâmetros distribucionais* - indicam diretamente a alteração sofrida por esse parâmetro na presença ou ausência (ou no nível assumido pelo fator).

A variável "AREA" e o efeito principal da variável, admitida como quantitativa contínua, "DATA" - *quando presentes em um preditor de algum dos parâmetros distribucionais* - serão incorporadas ao modelo por meio das funções suavizadoras s_{k1} e s_{k2} estimadas sobre um espaço vetorial gerado por uma base de funções apropriadas e que estão sujeitas, ou não, a uma penalização nos coeficientes dessa base, a fim de controlar o grau final de suavização evitando assim o "sobreajustamento" da função à variável.

Variáveis explicativas podem ser incorporadas sob a forma de funções suavizadoras (penalizadas ou não) como termos nos preditores dos parâmetros distribucionais de um *GAMLSS* de modo direto ou indireto.

O modo *direto* se dá pela inclusão, sob a forma da função suavizadora escolhida, diretamente nas estruturas especificadas para os preditores dos parâmetros distribucionais definidas na fórmula especificada na função gamlss() da biblioteca gamlss.

Já no modo *indireto*, a variável a ser incorporada sob a forma de uma função suavizadora na estrutura dos preditores dos parâmetros distribucionais é especificada antes, tal como se faz em um modelo *GAM*, por meio das funções s() ou e te()/ti() diretamente na estrutura definida na fórmula especificada na função gam() da biblioteca *mgcv*. Só então ela é incluída como termo na estrutura das fórmulas dos preditores dos parâmetros distribucionais de um *GAMLSS* envolvida (*wrapped*) pela função ga(), a qual estende o uso das funções suavizadoras implementadas na biblioteca *mgcv* para os da classe dos *GAMLSS* (STASINOPOULOS et al., 2017).

Para a situação em estudo as funções suavizadoras unidimensionais consideradas na implementação computacional dos *GAMLSS* (biblioteca *gamlss*) foram:

- pb() : baseada em *P-splines*;
- cs() : baseada em *splines* cúbicos penalizados;
- bs() : baseada em *B-splines*.

As funções s() e te()/ti() da biblioteca *mgcv* exigem a especificação de uma base por meio do argumento *bs()* na função $gam(\sim s(...bs()...))$:

- *thin plate regression splines*: *bs="tp"* (base *default*);
- *splines* cúbicos naturais (*Natural cubic regression splines*): *bs="cr"* com os nós da função posicionados nos quantis; ou,
- *B-splines*: *bs="bs"* com os nós da função igualmente espaçados (para *splines* penalizados a posição exata não é relevante desde que seja fornecida uma base de dimensão razoavelmente grande, o argumento "k" das funções, e a distribuição dos nós propicie uma boa cobertura),

dentre diversas outras bases disponíveis para circunstâncias específicas.

Em termos gerais, os *thin plate regression splines* tendem a fornecer um melhor desempenho quanto ao erro quadrático médio (EQM) sendo seguidos dos *splines* cúbicos com penalidades baseadas em derivadas, embora computacionalmente sejam mais lentos para se configurar do que as outras bases.

A função ti() da biblioteca *mgcv* (quando utilizada para se estimar uma função suavizadora unidimensional) apresenta os mesmos resultados que a função s() se as funções de base adotadas forem *splines* cúbicos naturais (*bs="cr"*) e suas dimensões k forem as mesmas .

Por outro lado, se as funções de base escolhidas forem *thin plate regression splines*, mas ainda mantendo uma mesma dimensão k, a função ti() só apresentará os mesmos resultados que a função $s()$ se não for feita a reparametrização Normal (*Normal parametrization*) (WOOD, 2017)), o que pode ser obtido por meio do argumento interno *ti(...,np=FALSE,...)*.

A variabilidade de natureza puramente espacial - *quando presente em um preditor de algum dos parâmetros distribucionais* - será incorporada ao modelo por meio de uma função suavizadora (s_{k3}) estimada sobre o espaço vetorial gerado pelo produto tensor dos espaços gerados pelas funções que formam as bases marginais apropriadas (*thin plate regression splines*) para cada uma das variáveis "UTM_X" e "UTM_Y" e sujeitas a penalizações marginais individuais. A função s() também permite a construção de uma superfície de suavização com a diferença, em relação à função ti(), que adota uma penalização única para as bases marginais (isotropismo). A escolha da introdução da variabilidade espacial, somente sob a forma da interação entre as variáveis "UTM_X" e "UTM_Y", resulta da conformação espacial observada na cidade - que não privilegia um dos eixos cartesianos em detrimento do outro - tal que a variabilidade espacial deve considerar as coordenadas de modo conjunto.

Por fim, a variabilidade espacial e temporal conjunta - *quando presente em um preditor de algum dos parâmetros distribucionais* - será incorporada ao modelo por meio de um função suavizadora (s_{k4}) estimada sobre o espaço vetorial gerado pelo produto tensor do espaço gerado das funções de bases dos *thin plate regression splines* sobre as variáveis "UTM_X" e "UTM_X", pelo espaço gerado pelas funções de base do *cubic regression splines* sobre a variável "DATA", estando tanto o primeiro, quanto o segundo espaço vetorial citados, sujeitos a penalizações individuais na estimação dos *splines* de produto tensor.

Funções suavizadoras bidimensionais são incorporadas à estrutura de um modelo *GAM* por meio da função ti() tendo igual parametrização para as bases marginais em seu argumento *bs()*: $ga(\sim ti(..., bs = \text{“}tp\text{”}, ...), ...)$ (WOOD, 2017) as quais (indiretamente) podem, então, ser incluídas nas estruturas das fórmulas dos preditores do parâmetros distribucionais de um *GAMLSS* definidas na função *gamlss()* por meio da função ga().

Funções suavizadoras multidimensionais são incorporadas à estrutura de um *GAM* por meio da função ti() com uma especificação adicional no argumento *d()*: $ga(\sim ti(..., bs = c(\text{“}tp\text{”}, \text{“}cr\text{”}), d = (2, 1), ...))$ para indicar que o suavizador multidimensional final será estimado pelo produto tensor de uma base bidimensional, por uma base unidimensional (WOOD, 2017) as quais, indiretamente, podem então ser incluídas nas estruturas das fórmulas dos preditores dos parâmetros distribucionais de um *GAMLSS* por meio da função ga().

Assim, as estimativas dos parâmetros distribucionais serão dadas por seus preditores lineares que assumem as seguintes estruturas gerais aditivas no modelo maximal exposto em (4.2).

$$
\begin{aligned}
\log(\hat{\mu}) =& \beta_{10} + s_{11}(\text{AT}) + s_{12}(\text{DATA}) + s_{13}(\text{UTM\_X,UTM\_Y}) + s_{14}(\text{UTM\_X,UTM\_Y; DATA}) \\
& \beta_{11}(\text{se RELEVO=}\textit{“Aclive”}) + \beta_{12}(\text{se RELEVO=}\textit{“Declive”}) + \\
& \beta_{13}(\text{se NATUREZA=}\textit{“Transação”}) + \\
& \beta_{14}(\text{se PAVIMENTACAO=}\textit{“Pav. asfáltica”}) + \\
& \beta_{15}(\text{se IMPLANTACAO=}\textit{“Condomínio”}) \\
\log(\hat{\sigma}) =& \beta_{20} + s_{21}(\text{AT}) + s_{22}(\text{DATA}) + s_{23}(\text{UTM\_X,UTM\_Y}) + s_{24}(\text{UTM\_X,UTM\_Y; DATA}) \\
& \beta_{21}(\text{se RELEVO=}\textit{“Aclive”}) + \beta_{22}(\text{se RELEVO=}\textit{“Declive”}) + \\
& \beta_{23}(\text{se NATUREZA=}\textit{“Transação”}) + \\
& \beta_{24}(\text{se PAVIMENTACAO=}\textit{“Pav. asfáltica”}) + \\
& \beta_{25}(\text{se IMPLANTACAO=}\textit{“Condomínio”}) \\
\hat{\nu} =& \beta_{30} + s_{31}(\text{AT}) + s_{32}(\text{DATA}) + s_{33}(\text{UTM\_X,UTM\_Y}) + s_{34}(\text{UTM\_X,UTM\_Y; DATA}) \\
& \beta_{31}(\text{se RELEVO=}\textit{“Aclive”}) + \beta_{32}(\text{se RELEVO=}\textit{“Declive”}) + \\
& \beta_{33}(\text{se NATUREZA=}\textit{“Transação”}) + \\
& \beta_{34}(\text{se PAVIMENTACAO=}\textit{“Pav. asfáltica”}) + \\
& \beta_{35}(\text{se IMPLANTACAO=}\textit{“Condomínio”}) \\
\log(\hat{\tau}) =& \beta_{40} + s_{41}(\text{AT}) + s_{42}(\text{DATA}) + s_{43}(\text{UTM\_X,UTM\_Y}) + s_{44}(\text{UTM\_X,UTM\_Y; DATA}) \\
& \beta_{41}(\text{se RELEVO=}\textit{‘Aclive”}) + \beta_{42}(\text{se RELEVO=}\textit{“Declive”}) + \\
& \beta_{43}(\text{se NATUREZA=}\textit{“Transação”}) + \\
& \beta_{44}(\text{se PAVIMENTACAO=}\textit{“Pav. asfáltica”}) + \\
& \beta_{45}(\text{se IMPLANTACAO=}\textit{“Condomínio”})
\end{aligned}
\tag{4.2}
$$

4.5.3 Geração de imagens

Admitindo-se como fronteiras amostrais a área delimitada pelo perímetro urbano da cidade de Londrina e o intervalo temporal dos dados, o modelo proposto possibilita não apenas a estimação do valor unitário mediano em qualquer referência espacial ou temporal desejada, como também permite a geração de uma superfície de valor que represente graficamente a variação do valor unitário mediano em toda a extensão da área do estudo, para uma determinada data arbitrada, bastando que seja estabelecida uma situação paradigma das características do imóvel a ser estimado (área total, relevo, implantação, natureza da informação e data de referência) e essa estimação seja replicada em cada localização definida em um *grid* reticulado de localizações espaciais que, no estudo foi definido com um espaçamento entre pontos de 50 m em cada direção cartesiana.

A produção de imagens e animações, em duas ou três dimensões, que espelhem a variabilidade do valor mediano no perímetro urbano de Londrina referente a

um período temporal específico ou a um intervalo entre datas, pode ser realizada pelas funções disponíveis em variadas bibliotecas como *ggplot2* (ORCID et al., 2021); *plotKML* (HENGL et al., 2021); *tmap* (TENNEKES et al., 2022); *mapview* (APPELHANS et al., 2021); *mapdeck* (COOLEY, 2020); *maps* (BECKER et al., 2021); *maptools* (BIVAND et al., 2022); *inlabru* (LINDGREN et al., 2022); *spaceTime* (PEBESMA et al., 2022); *spData* (BIVAND et al., 2021); *plotGoogleMaps* (KILIBARDA; SEKULIC, 2015) e *sp* (PEBESMA et al., 2021),sob variados formatos de apresentação (jpeg, html, kml, mp4, m4v), muitas das quais demandando a transformação do objeto que contém a informação para outras classes como *SpatialPixelsDataFrame* ou um simples *dataframe* (para tanto valendo-se de bibliotecas auxiliares como ; *geobr* (PEREIRA; GONÇALVES, 2022); *sf* (PEBESMA et al., 2022); *spatialEco* (EVANS; MURPHY; RAM, 2021); *GISTools* (BRUNSDON; CHEN, 2014); *geoR* (JR et al., 2020); *raster* (HIJMANS et al., 2022)).

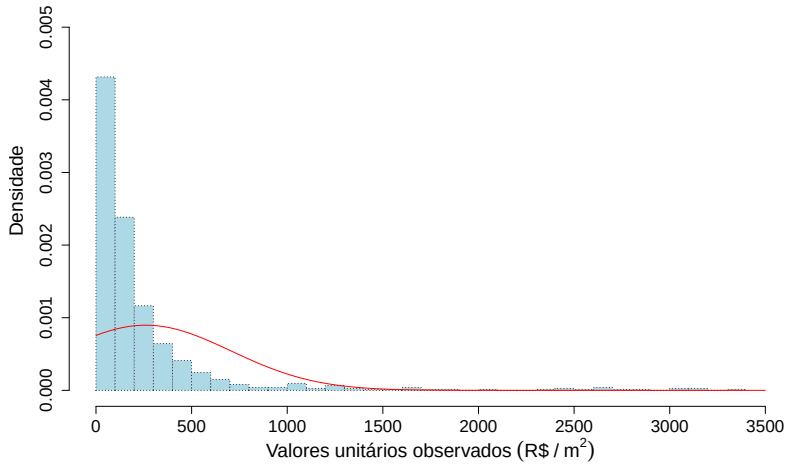
Por fim destaca-se que, com o propósito único de facilitar a visualização das variações do valor unitário mediano entre imagens geradas em períodos distintos, a resposta foi transformada pela aplicação da função logaritmo, a qual tornou a variação cromática observada nas imagens mais pronunciada, quando diversos períodos temporais são expostos simultaneamente.

5 RESULTADOS OBTIDOS

5.1 ANÁLISE DESCRITIVA DA AMOSTRA

Depreende-se da informação trazida pela Tabela 5.1 que a variável resposta "UNIT" apresenta uma distribuição com forte assimetria à direita e seu histograma, ilustrado na Figura 5.1, revela também uma forma leptocúrtica tendo como referência a distribuição Normal (simétrica e mesocúrtica).

Figura 5.1: Distribuição marginal da resposta sob a distribuição Normal (NO)



Fonte: Próprio autor

Tabela 5.1: Medidas resumo das variáveis quantitativas incorporadas como termos nos preditores do modelo

Variável (sigla)	Valor mínimo	1^o Quartil	Mediana	Valor médio	3^o Quartil	Valor máximo	Coef. assimetria	Coef. curtose
UNIT	5,00	68,13	119,60	260,29	263,79	3.348,21	4,28	23,89
AT	157,00	289,20	377,00	1.621,60	607,00	43.088,00	5,06	31,85
DATA[(1)]	9.279	12.053	12.889	13.341	14.722	18.697	0,79	2,87
UTM_X	477.200	481.292	483.240	483.218	485.020	489.853	0,05	2,26
UTM_Y	7.410.324	7.419.294	7.421.505	7.421.978	7.424.869	7.429.181	0,08	2,35

[(1)]: 29/05/1995; 01/01/2003; 16/04/2005; 28/07/2006; 10/06/2010; 11/03/0021

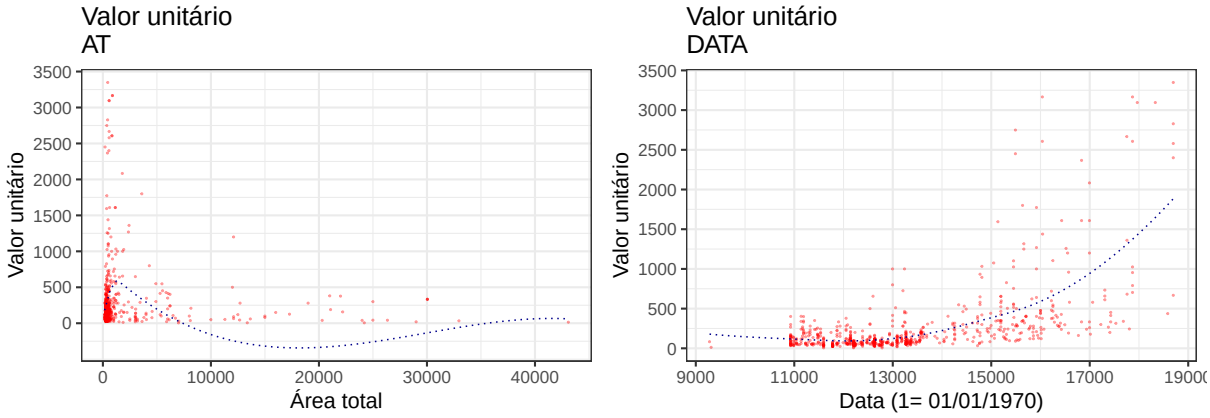
Fonte: Próprio autor

De forma complementar, os gráficos de dispersão mostrados nas Figuras 5.2 tornam claro que a relação da resposta "UNIT" condicionada às variáveis "AT" e "DATA" não

é linear. Observa-se na Figura 5.2 (b) que o padrão de crescimento do valor unitário observado intensifica-se com o progredir do tempo.

Com o propósito de aclarar a relação da variável resposta "UNIT" condicionada à variável "AT", não muito definido na Figura 5.2 (a), na Figura 5.3 investiga-se o padrão das mudanças sob duas transformações diferentes da variável explanatória, igualmente não lineares.

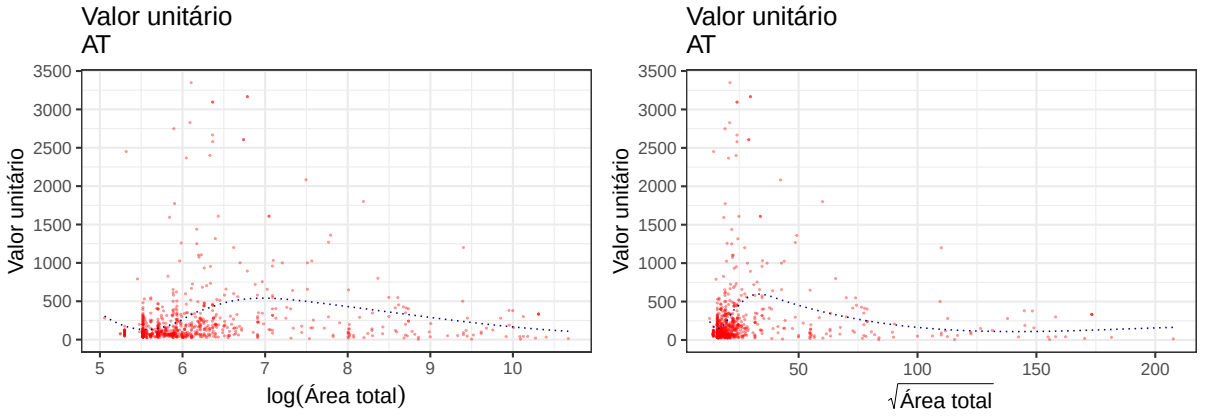
Figura 5.2: Dispersão da variável resposta



(a) Variável: AT (área total) (b) Variável: DATA

Fonte: Próprio autor

Figura 5.3: Dispersão da variável resposta

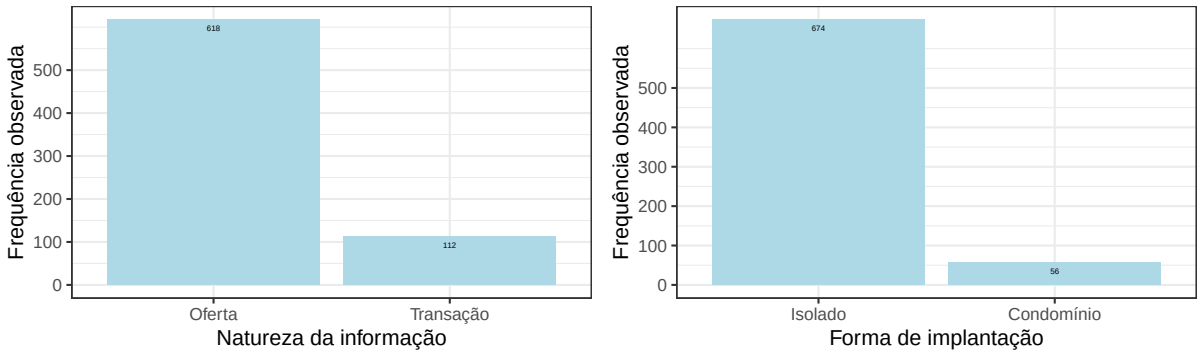


(a) Variável transformada: AT (área total) (b) Variável transformada: AT (área total)

Fonte: Próprio autor

As frequências observadas nos dados amostrais relativas aos níveis assumidos pelos fatores "NATUREZA" e "IMPLANTACAO" são mostradas na Figura 5.4 e "PAVIMENTACAO" e "RELEVO" na Figura 5.5. Tais verificações visam assegurar um equilíbrio amostral para os variados níveis assumidos por esses fatores.

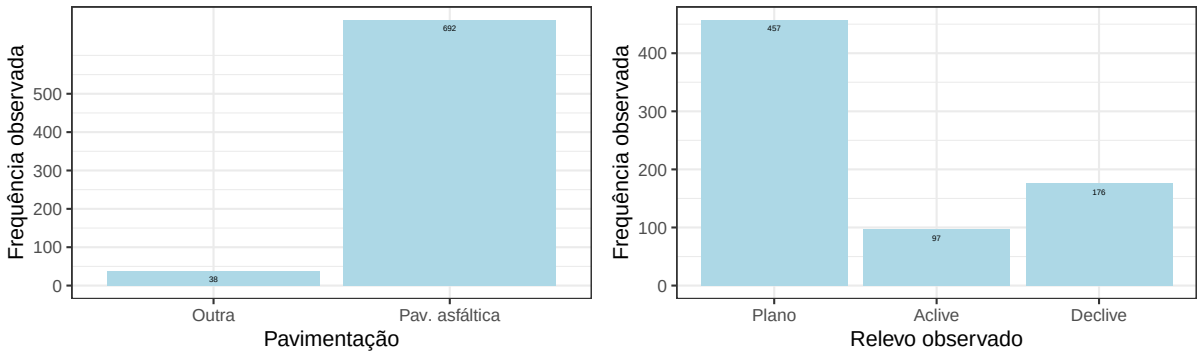
Figura 5.4: Frequências observadas na amostra dos níveis dos fatores “NATUREZA” e “IMPLANTACAO”



(a) Fator: “NATUREZA” (b) Fator: “IMPLANTACAO” (implantação)

Fonte: Próprio autor

Figura 5.5: Frequências observadas na amostra dos níveis dos faores “PAVIMENTACAO” e “RELEVO”

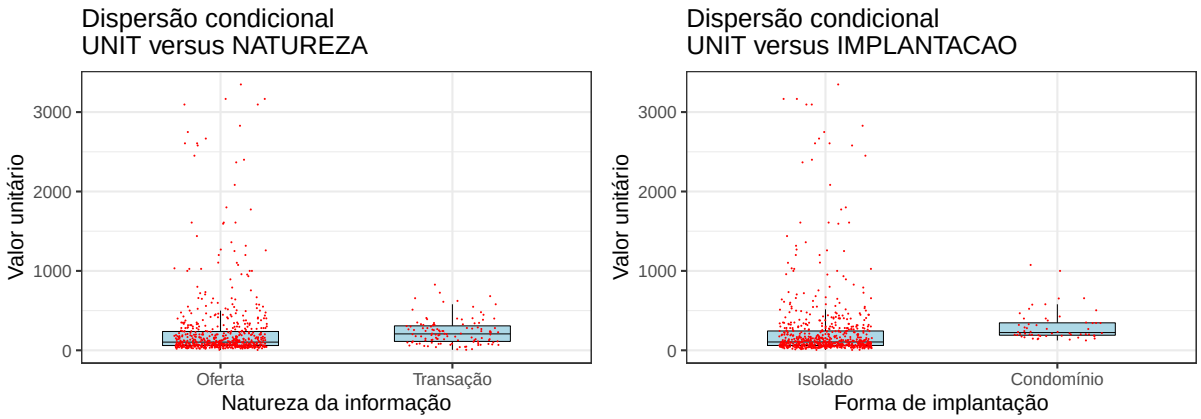


(a) Fator: “PAVIMENTACAO” (pavimentação) (b) Fator: “RELEVO”

Fonte: Próprio autor

A variabilidade observada na Figura 5.6 da variável resposta “UNIT” relacionada aos níveis dos fatores “NATUREZA” e “IMPLANTACAO” expõe, na primeira qualificação (a), que os preços (anunciados) são propostos de modo muito subjetivo quando comparado aos valores efetivamente transacionados. Na segunda qualificação (b) a menor variabilidade da resposta observada para a implantação em condomínio resulta de uma certa homogeneidade relacionada às amenidades existentes sob esse tipo de implantação e também evidencia uma certa independência da localização espacial do elemento. Assim pode-se entender que, para o adquirente, se o terreno está inserido em uma estrutura de condomínio fechado, então a localização desse empreendimento já não lhe é tão relevante.

Figura 5.6: *Box-plot* da variável resposta em relação aos níveis dos fatores “NATUREZA” e “IMPLANTACAO”

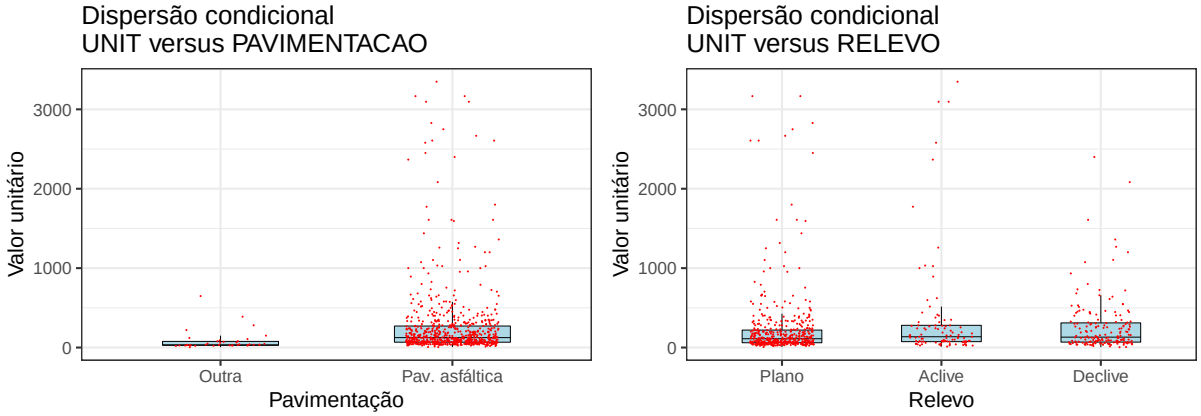


(a) Fator: “NATUREZA” (b) Fator: “IMPLANTACAO” (implantação)

Fonte: Próprio autor

A variabilidade da variável resposta “UNIT” relacionada aos níveis dos fatores “PAVIMENTACAO” e “RELEVO” é exposta na Figura 5.7. A menor variabilidade na primeira situação (a) resulta da reduzida quantidade de elementos ainda não dotados de pavimentação asfáltica e, provavelmente situados em locais muito assemelhados entre si, quando comparados à outra situação.

Figura 5.7: *Box-plot* da variável resposta em relação aos níveis dos fatores “PAVIMENTACAO” e “RELEVO”



(a) Fator: “PAVIMENTACAO” (pavimentação) (b) Fator: “RELEVO”

Fonte: Próprio autor

5.2 MODELO PROPOSTO

O *GAMLSS* proposto assumiu a estrutura mostrada em (5.1).

$$
\begin{aligned}
\log(\hat{\mu}) =& 4,53 + s_{11}(AT) + s_{12}(DATA) + s_{13}(UTM\_X, UTM\_Y) \\
&+ s_{14}(UTM\_X, UTM\_Y; DATA) - 0,11(\text{se “RELEVO"=}\textit{Aclive}) \\
&- 0,06(\text{se RELEVO=}\textit{“Declive”}) - 0,21(\text{se NATUREZA=}\textit{“Transação”}) \\
&+ 0,37(\text{se PAVIMENTACAO=}\textit{“Pav. asfáltica”}) + 0,75(\text{se IMPLANTACAO=}\textit{“Condomínio”}) \\
\log(\hat{\sigma}) =& -1,23 + s_{21}(AT) + s_{22}(DATA) + s_{23}(UTM\_X, UTM\_Y) \\
\hat{\nu} =& -0,94 + s_{31}(AT) + s_{32}(DATA) + s_{33}(UTM\_X, UTM\_Y) \\
\log(\hat{\tau}) =& 0,65 + s_{41}(AT)
\end{aligned}
\tag{5.1}
$$

O Quadro 5.1 expõe a parametrização das funções computacionais usadas para os variados suavizadores e a Tabela 5.2 apresenta algumas métricas associadas ao modelo.

Quadro 5.1: Parametrização das funções computacionais

Suavizador	**Parametrização da função computacional**
$s_{11}(AT), s_{21}(AT), s_{31}(AT), s_{41}(AT)$	ti(AT, bs="cr")
$s_{12}(DATA), s_{22}(DATA), s_{32}(DATA)$	ti(DATA, bs="cr")
$s_{13}(UTM\_X, UTM\_Y), s_{23}(UTM\_X, UTM\_Y)$	ti(UTM_X, UTM_Y, bs= "tp", d=2)
$s_{14}(UTM\_X, UTM\_Y; DATA)$	ti(UTM_X, UTM_Y, DATA, bs= c("tp", "cr"), d = c(2, 1))

Fonte: Próprio autor

Tabela 5.2: Informações sobre algumas métricas do modelo proposto

Métrica	Valor
Deviance global	7618,53
AIC	7869,96
BIC	8447,38
Pseudo R^2 (Cox&Snell)	0,894
Graus de liberdade do ajuste	125,72
Graus de liberdade dos resíduos	604,28
Número e observações	730

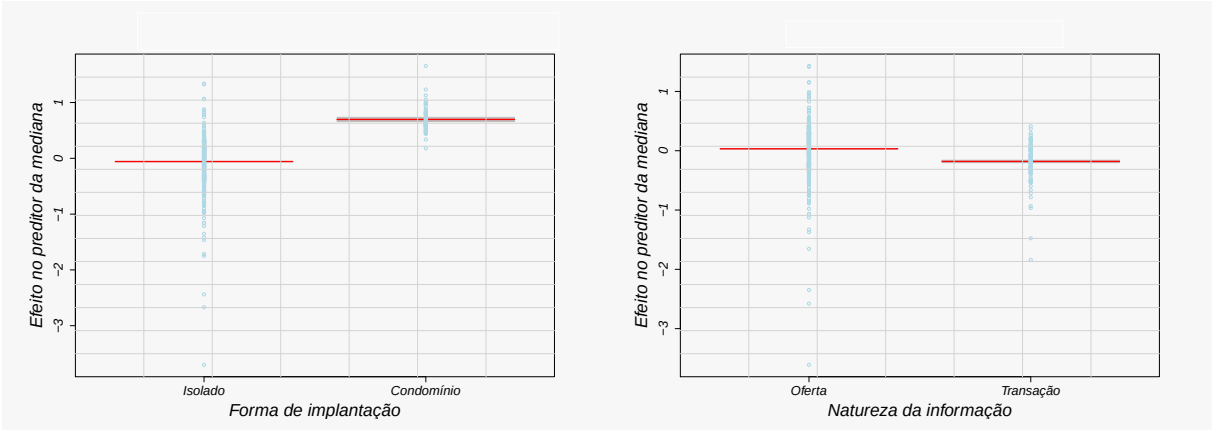
Fonte: Próprio autor

5.2.1 Efeitos dos termos incorporados ao preditor do primeiro parâmetro da distribuição (θ_1: mediana)

Os efeitos exercidos pelos fatores "IMPLANTACAO" e "NATUREZA" sobre o preditor linear da mediana distribucional ($\eta_1 = \log(\hat{\mu})$), exibidos na Figura 5.8 (a) , confirmam que terrenos implantados dentro de uma instalação condominial (b), dotada de alguma infra-estrutura de amenidades, apresentam um valor unitário mediano superior àquele que se situam isolados nas ruas. Também se valida a prática comum de se anunciar terrenos por valores

unitários artificialmente superiores comparados aos valores pelos quais serão efetivamente transacionados conforme exibido na Figura 5.8 (b). Ambos os resultados são coerentes com o que se observa no mercado imobiliário.

Figura 5.8: Efeitos dos fatores “IMPLANTACAO” e “NATUREZA” sobre a mediana distribucional (incorporados parametricamente ao seu preditor linear)



(a) Fator: “IMPLANTACAO” (implantação) (b) Fator: “NATUREZA”

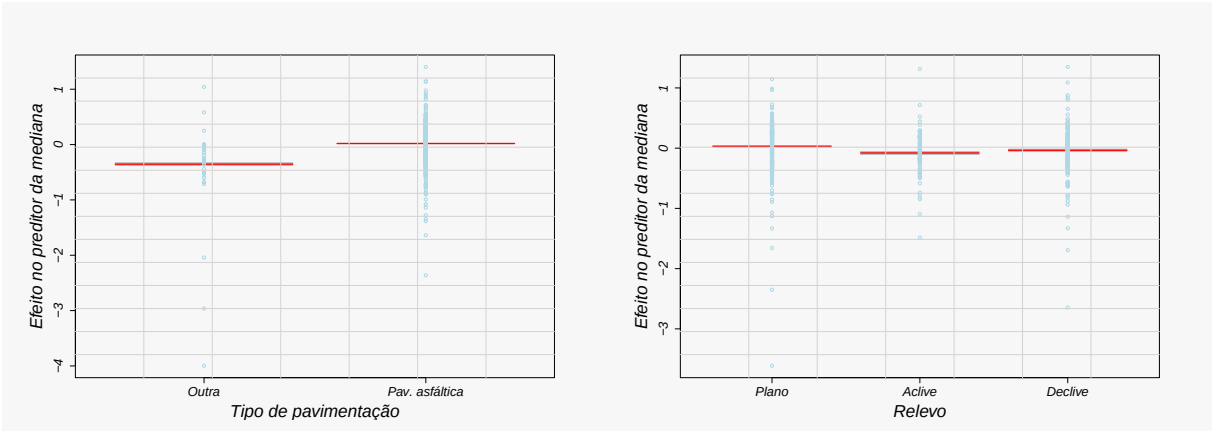
Fonte: Próprio autor

Os efeitos exercidos pelos fatores “PAVIMENTACAO” e “RELEVO” sobre o preditor linear da mediana distribucional ($\eta_1 = \log(\hat{\mu})$), exibidos na Figura 5.9 (a), confirmam que terrenos com sua frente principal voltada para um logradouro com pavimentação asfáltica possuem um valor unitário mediano superior e que os terrenos planos são mais valorizados. Ambos os resultados são coerentes com o que se observa no mercado.

À guisa de informação complementar, acreditamos que a pequena variação dos efeitos exercidos pelos níveis “DECLIVE” e “ACLIVE” no gráfico da Figura 5.9 (b) decorre da situação prática, algumas vezes até favorável, para terrenos em declive quando suas dimensões favorecem seu aproveitamento para incorporação imobiliária. Nesses casos tal relevo resulta em uma economia construtiva como se verifica, por exemplo, na construção de edifícios, pela redução de custos associados à escavação e proteção de encostas. Por outro lado, terrenos em aclive com dimensões coerentes à construção isolada facilitam a instalação da garagem no nível inferior e da parte residencial em um plano mais elevado.

Sob um ponto de vista macroscópico, em geral, o valor unitário mediano de um terreno é inversamente relacionado à sua área: terrenos com reduzidas dimensões apresentam um valor unitário superior aos de terrenos com grandes áreas. Tal fato decorre do fenômeno de natureza econômica observado quando qualquer produto é oferecido sob diferentes quantidades e, por tal variação, atingem diferentes mercados consumidores. Isso pode ser ilustrado valendo-se de uma comparação simples e, para tanto, considere uma região qualquer dentro de uma cidade. Admita nessa região a existência de dois terrenos contíguos, mas, todavia, com áreas bastante distintas: um deles com uma área total muito superior ao outro.

Figura 5.9: Efeitos dos fatores “PAVIMENTACAO” e “RELEVO” sobre a mediana distribucional (incorporados parametricamente ao seu preditor linear)



(a) Fator: “PAVIMENTACAO” (pavimentação) (b) Fator: “RELEVO”

Fonte: Próprio autor

Quando oferecidos ao mercado, o que irá se verificar é que os valores unitários de cada um desses dois imóveis, calculados após sua negociação efetiva, serão diferentes. Mesmo tendo as mesmas características intrínsecas e extrínsecas e tendo sido negociados em uma mesma época o valor unitário mediano do terreno com área menor será superior ao do outro.

O efeito exercido pela variável “AT”, por meio da função de suavização estimada, sobre o preditor linear da mediana distribucional ($\eta_1 = \log(\hat{\mu})$), exibido na Figura 5.10 (a), revela algumas sutilezas que merecem considerações de ordem microscópica, por assim dizer.

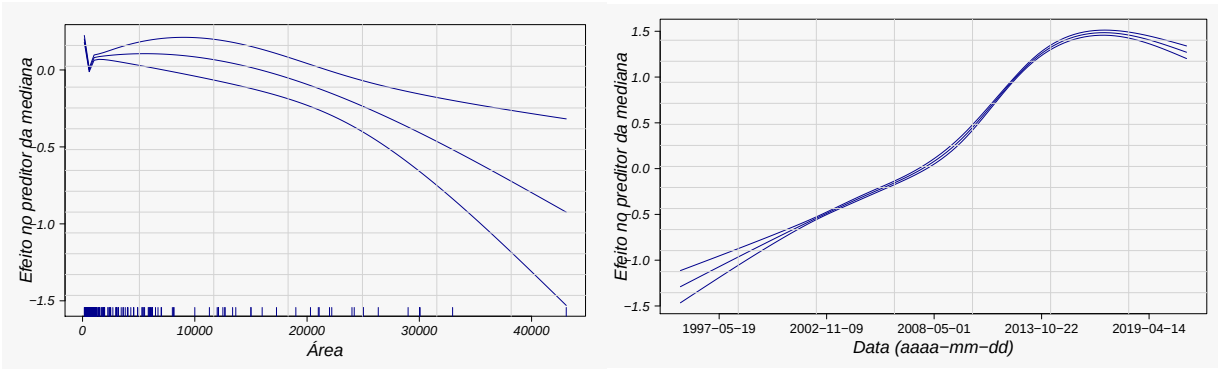
Observa-se que o efeito exercido pela dimensão do imóvel no valor unitário mediano tenha um primeiro valor máximo para a área mínima (125,00 m^2) permitida pela legislação (Lei n^o 6.766 de 19/12/1979) para um lote urbano resultante do parcelamento de outro, com área maior. Esse efeito progressivamente se reduz até uma área aproximada de 250 m^2 (sem efeito). Verifica-se a partir dessa dimensão, até uma área aproximada de 400 m^2, que o efeito passa a ser de redução no valor unitário mediano. A partir desse valor de área total observa-se que o efeito da dimensão de um terreno volta a ser de incremento no seu valor unitário mediano, atingindo um novo valor máximo para terrenos com uma área total aproximada de 3.000 m^2.

Depreende-se, em linhas gerais, que o efeito da área total no valor unitário mediano seja máximo tanto para terrenos com a área total mínima permitida pela legislação, invariavelmente destinados a edificações de padrão construtivo bastante modesto, quanto para terrenos com área total aproximada de 3.000 m^2, aparentemente, a dimensão mais procurada para a realização de incorporações imobiliárias.

O efeito exercido pela variável “DATA”, por meio da função de suavização estimada, sobre o preditor linear da mediana distribucional ($\eta_1 = \log(\hat{\mu})$), exibido na Figura 5.10 (b), revela um comportamento bastante interessante ao longo do período estudado.

Em grande parte desse período observa-se que o efeito sobre a resposta é sempre crescente e também que - em maio de 2008 - houve uma intensificação pela alteração em sua taxa. Já a partir de novembro de 2013, verifica-se uma desaceleração desse efeito - ainda incremental - com o progredir do tempo, muito provavelmente associada a fatos relevantes ocorridos na economia nacional naquela data.

Figura 5.10: Efeitos das variáveis “AT” e “DATA” sobre a mediana distribucional (incorporadas ao seu preditor linear sob a forma de funções suavizadoras unidimensionais)



(a) Variável: “AT” (área total), edf=4 (b) Variável: “DATA”, edf=3,99

Fonte: Próprio autor

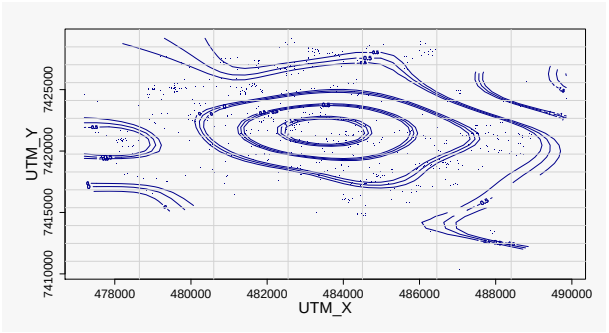
As Figuras 5.11 (a) e (b) ilustram o efeito exercido pela localização, por meio da função de suavização estimada, sobre o preditor linear da mediana distribucional ($\eta_1 = \log(\hat{\mu})$).

A Figura 5.12 ilustra o efeito exercido pela interação da localização do elemento e a data de referência, por meio da função de suavização estimada, sobre o preditor linear da mediana distribucional ($\eta_1 = \log(\hat{\mu})$).

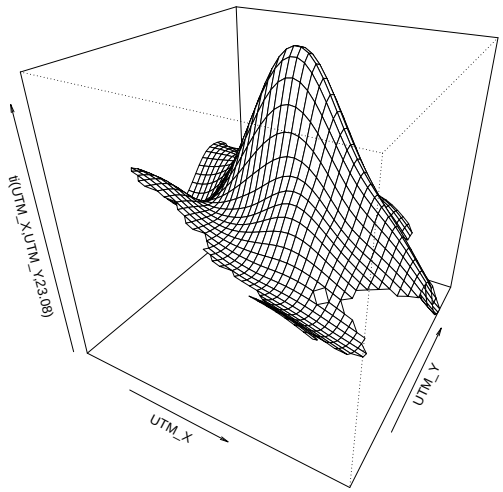
5.2.2 Efeitos dos termos incorporados ao preditor do segundo parâmetro da distribuição (θ_2: coeficiente de variação)

Os efeitos exercidos pelas variáveis “AT” e “DATA” incorporadas ao preditor do segundo parâmetro distribucional ($\eta_2 = \log(\hat{\sigma})$) sob a forma de funções suavizadoras podem ser visualmente verificados nas Figuras 5.13 e 5.14.

Figura 5.11: Efeito conjunto das variáveis “UTM_X” e “UTM_Y” na mediana distribucional (incorporadas ao seu preditor linear sob a forma de uma função suavizadora bidimensional): efeito puramente espacial da localização



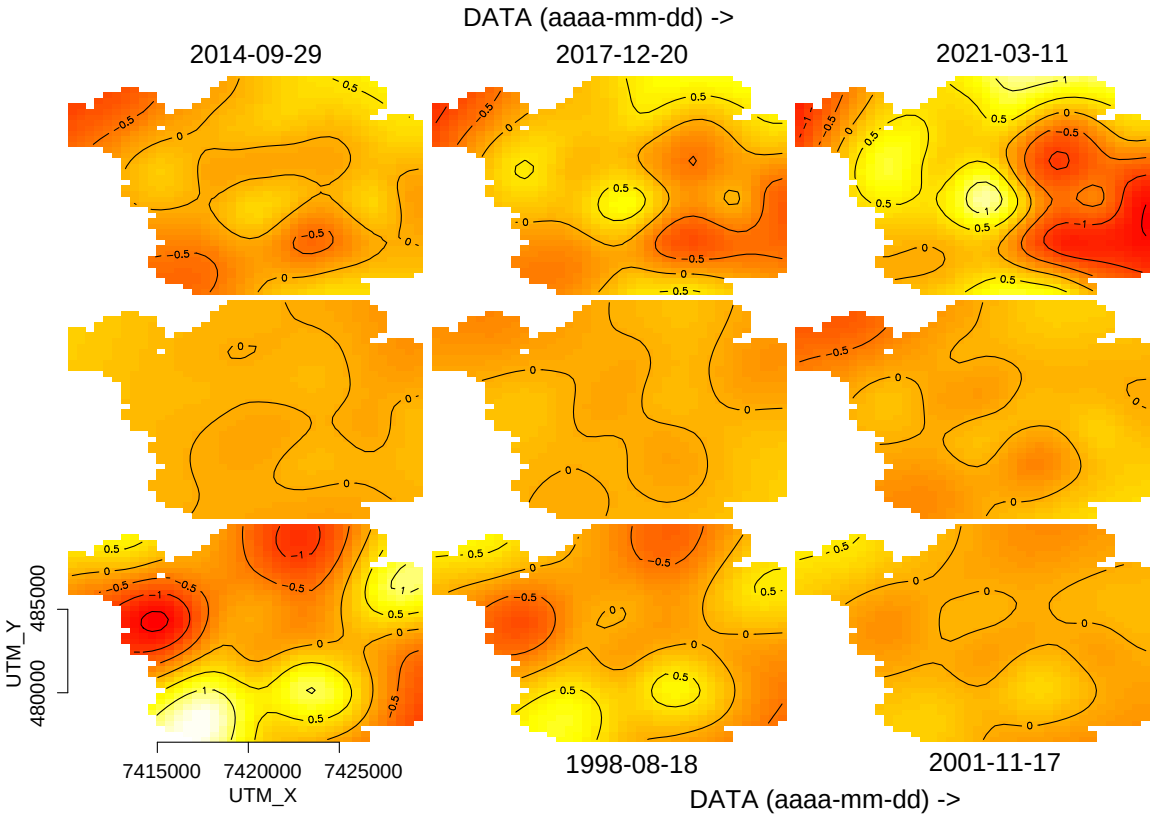
(a) Gráfico de contornos (isolinhas), edf=23,08



(b) Superfície de valores, edf=23,08

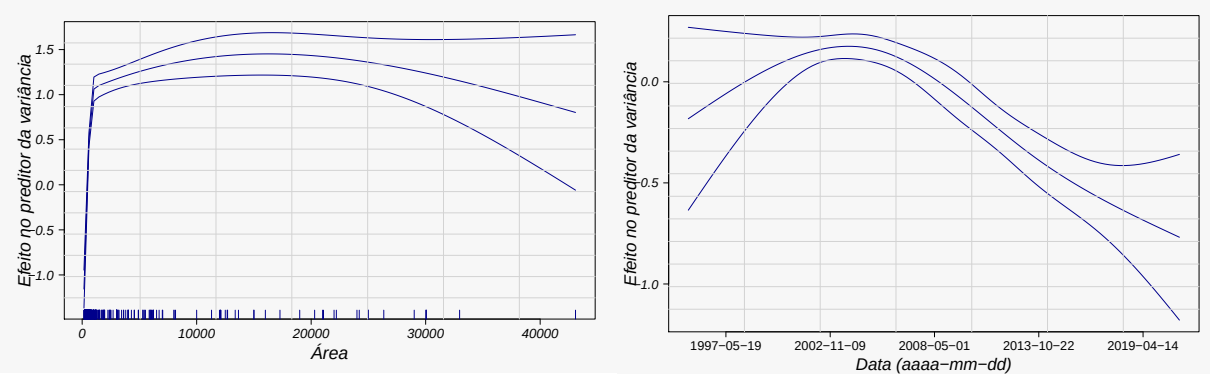
Fonte: Próprio autor

Figura 5.12: Efeito conjunto das variáveis “UTM_X”, “UTM_Y” e “DATA” na mediana distribucional (incorporadas ao seu preditor linear sob a forma de uma função suavizadora multidimensional): efeito espaçotemporal das coordenadas mostradas (UTM) e a data da informação, edf=50,95



Fonte: Próprio autor

Figura 5.13: Efeitos das variáveis “AT” e “DATA” sobre o coeficiente de variação da distribuição (incorporadas ao seu preditor linear sob a forma de funções suavizadoras unidimensionais)

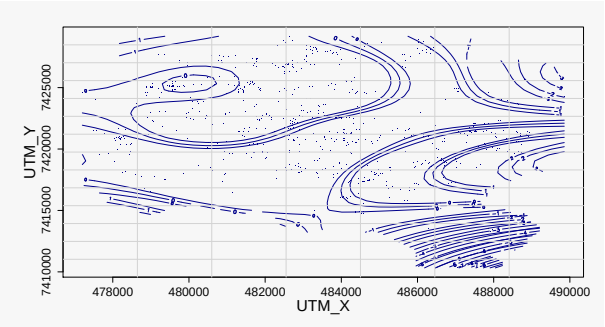


(a) Variável: “AT” (área total), edf=3,2

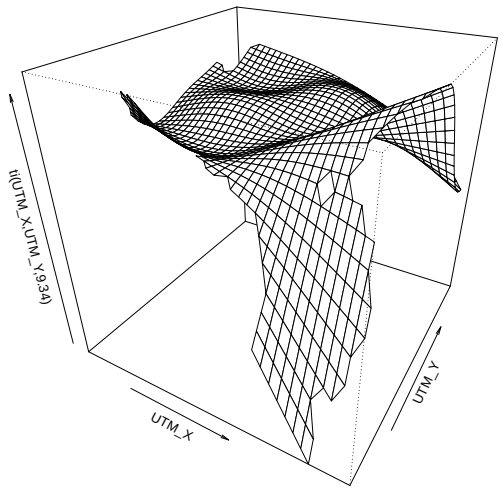
(b) Variável: “DATA”, edf=2,87

Fonte: Próprio autor

Figura 5.14: Efeito conjunto das variáveis “UTM_X” e “UTM_Y” no coeficiente de variação da distribuição (incorporadas ao seu preditor linear sob a forma de uma função suavizadora bidimensional): efeito puramente espacial da localização



(a) Gráfico de contornos (isolinhas), edf=9,34



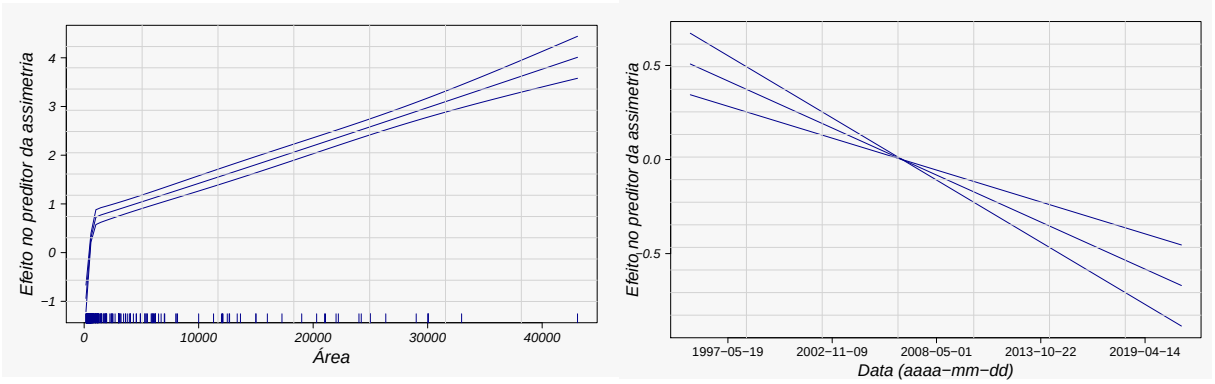
(b) Superfície de valores, edf=9,34

Fonte: Próprio autor

5.2.3 Efeitos dos termos incorporados ao preditor do terceiro parâmetro da distribuição (θ_3: coeficiente *proxy* da assimetria)

Os efeitos exercidos pelas variáveis “AT” e “DATA” incorporadas ao preditor do terceiro parâmetro distribucional ($\eta_3 = \log(\hat{\nu})$) sob a forma de funções suavizadoras podem ser visualmente verificados nas Figuras 5.15 e 5.16. Na Figura 5.15 (b) observa-se que a penalização aplicada ao suavizador estimado com a variável “DATA” resultou em uma relação linear (edf=1). As restrições de identificabilidade de soma zero (3.15) impostas a um suavizador implicam que, se a função estimada for linear, então o intervalo de confiança deve desaparecer no ponto zero (WOOD, 2017).

Figura 5.15: Efeitos das variáveis “AT” e “DATA” sobre o coeficiente *proxy* da assimetria da distribuição (incorporadas ao seu preditor linear sob a forma de funções suavizadoras unidimensionais)



(a) Variável: “AT” (área total), edf=3,09

(b) Variável: “DATA”, edf=1

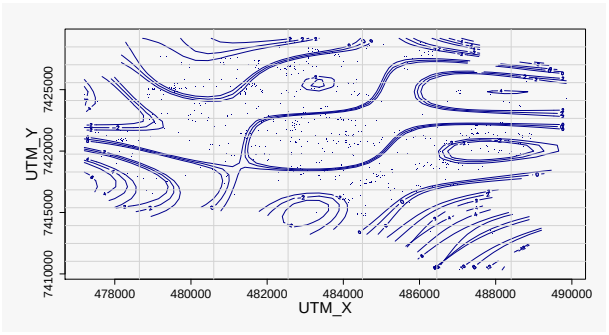
Fonte: Próprio autor

5.2.4 Efeitos dos termos incorporados ao preditor do quarto parâmetro da distribuição (θ_4: coeficiente *proxy* da curtose)

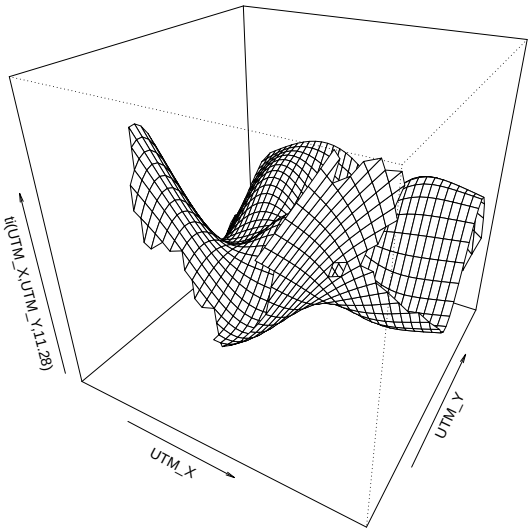
O efeito exercido pela variável“AT” incorporada ao preditor do quarto parâmetro distribucional ($\eta_4 = \log(\hat{\tau})$) sob a forma de uma função suavizadora pode ser visualmente verificado na Figura 5.17.

Os graus de liberdade efetivos indicados nas Figuras 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16 e 5.17 confirmam a não linearidade das relações pois, em seu processo de estimação, a penalização imposta nunca foi de tal ordem que o perfil final assumido pela função suavizada fosse linear e, consequentemente, seu efeito constante para qualquer valor da variável. À guisa de informação, tais efeitos podem ser extraídos da função *predictAll()* da biblioteca principal da implementação dos *GAMLSS* (predictAll()) com o argumento *type="terms"*, que retorna o efeito aditivo de cada um dos termos de cada um dos preditores.

Figura 5.16: Efeito conjunto das variáveis “UTM_X” e “UTM_Y” no coeficiente *proxy* da assimetria da distribuição (incorporadas ao seu preditor linear sob a forma de uma função suavizadora bidimensional): efeito puramente espacial da localização



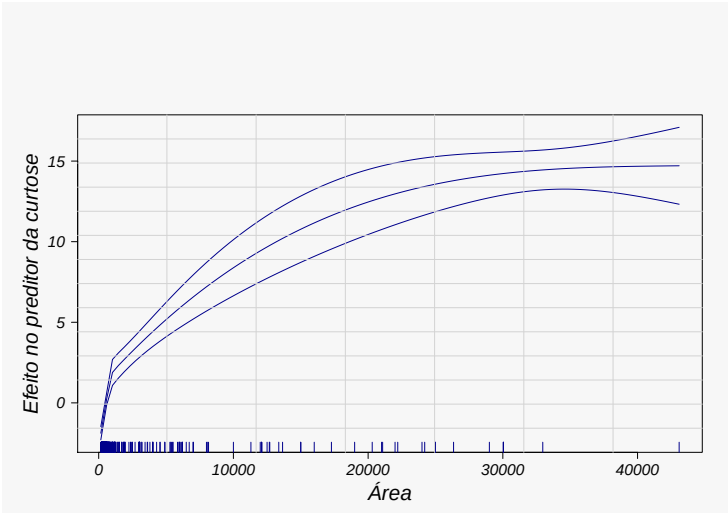
(a) Gráfico de contornos (isolinhas), edf=11,28



(b) Superfície de valores, edf=11,28

Fonte: Próprio autor

Figura 5.17: Efeito da variável “AT” sobre o coeficiente *proxy* da curtose da distribuição (incorporada ao seu preditor linear sob a forma de uma função suavizadora unidimensional) edf=3,92



Fonte: Próprio autor

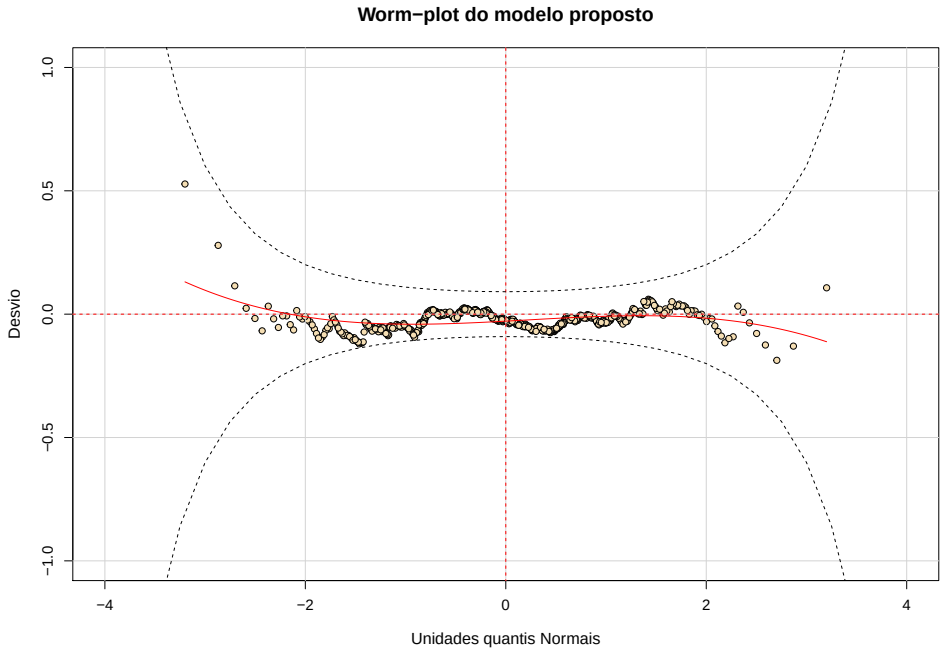
5.2.5 Análise diagnóstica do modelo proposto

5.2.5.1 *Worm plot* dos resíduos quantílicos

O padrão exibido na Figura 5.18 do *worm plot* dos resíduos quantílicos sugere um bom ajustamento do modelo proposto e os coeficientes estimados do polinômio cúbico ajustado: $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3$ podem ser observados na Tabela 5.3 e cotejados aos valores limite (em módulo) sugeridos por Van Buuren e Fredriks (2001): 0,10; 0,10; 0,05 e 0,03.

De modo complementar, são apresentados nas Tabelas 5.4, 5.5, 5.6, 5.7, 5.8, 5.9 e 5.10 os coeficientes estimados dos polinômios cúbicos ajustados para quatro intervalos de valores assumidos pelas variáveis “AT”, “DATA”, pelos níveis assumidos pelos fatores “RELEVO”, “PAVIMENTACAO”, “NATUREZA”, “IMPLANTACAO” e, por fim, pelos níveis assumidos pelo fator “NATUREZA” conjuntamente aos quatro intervalos de valores assumidos pela variável “AT”, com o propósito e se investigar o ajustamento conjunto do fator que reflete o deságio aplicado sobre um valor relativo a um imóvel meramente ofertado ao mercado sob várias dimensões de sua área total.

Figura 5.18: *Worm plot* do modelo proposto



Fonte: Próprio autor

Tabela 5.3: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots*

Coeficientes	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Valores ajustados	-0,0283	0,02198	0,00369	-0,00585

Fonte: Próprio autor

Tabela 5.4: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada intervalo de valores da variável "AT"

Intervalo considerado (m^2)	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
(156,995;289,005)	-0,1203	-0,0543	0,0216	0,0100
(289,995;377,005)	0,1154	0,0263	0,0027	-0,0078
(376,995;608,005)	-0,1144	-0,0124	-0,0091	0,0009
(607,995;43.088,025)	0,0182	0,0914	-0,0064	-0,0195

Fonte: Próprio autor

Tabela 5.5: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada intervalo de valores da variável "DATA"

Intervalo considerado	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
28/05/1995 a 01/01/2003	-0,0307	0,0271	-0,0002	-0,0148
31/12/2002 a 01/04/2005	0,0654	-0,0473	-0,0451	-0,0067
30/04/2005 a 23/04/2010	-0,0794	0,0012	0,0071	-0,0149
22/04/2010 a 11/03/2021	-0,0281	0,1100	0,0275	0,0027

Fonte: Próprio autor

Tabela 5.6: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "RELEVO"

Nível considerado	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Plano	0,0132	-0,0072	0,0197	-0,0020
Aclive	-0,1159	0,1039	0,0176	-0,0199
Declive	-0,0932	0,0525	-0,0400	-0,0125

Fonte: Próprio autor

Tabela 5.7: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "PAVIMENTACAO"

Nível considerado	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Outra	-0,0524	-0,3161	0,0053	0,1291
Pav. asfáltica	0,0108	0,0225	-0,0035	-0,0072

Fonte: Próprio autor

Tabela 5.8: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "NATUREZA"

Nível considerado	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Oferta	-0,0254	0,0086	0,0080	-0,0063
Transação	-0,0516	0,0893	-0,0130	-0,0025

Fonte: Próprio autor

Tabela 5.9: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* parciais em cada nível do fator "IMPLANTACAO"

Nível considerado	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Isolada	-0,0187	0,0284	0,0006	-0,0058
Condomínio	-0,1540	-0,0517	0,0528	-0,0095

Fonte: Próprio autor

Tabela 5.10: Coeficientes dos polinômios cúbicos ajustados sobre os *worm plots* conjuntos de cada intervalo de valores da variável "AT" e nível do fator "NATUREZA"

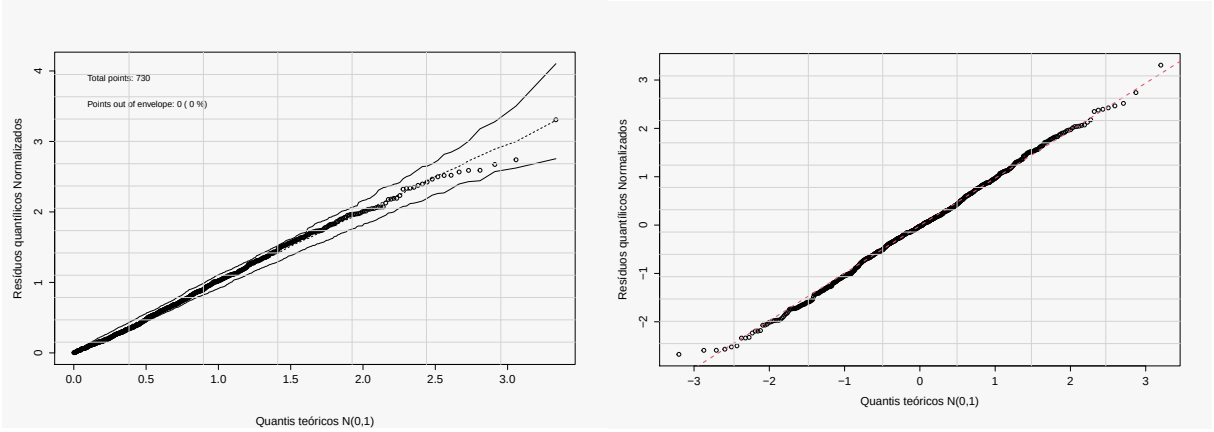
Coeficientes	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Nível considerado	Intervalo considerado (m^2)			
	(156,995;289,005)	(289,995;377,005)	(376,995;608,005)	(607,995;43.088,025)
Oferta	-0,1523	-0,0593	0,0640	0,0041
	0,1213	0,0113	-0,0118	-0,0116
	-0,1042	-0,0248	-0,0006	0,0006
	0,0252	0,0802	-0,0093	-0,0156
Transação	-0,0418	0,0104	-0,0923	0,0087
	0,1024	0,1630	0,0661	-0,0083
	-0,2219	0,0630	-0,0587	0,0144
	-0,0541	0,1988	0,0371	-0,0610

Fonte: Próprio autor

5.2.5.2 Gráficos de probabilidade dos resíduos quantílicos

Na Figura 5.19 pode-se investigar a Normalidade dos resíduos quantílicos por meio dos gráficos *half Normal plot* com envelope simulado e de probabilidade quantil-quantil e, na Tabela 5.11, algumas medidas descritivas são apresentadas. Os padrões visuais de distribuição e os valores descritivos apresentados para os resíduos não nos permitem afirmar que destoem do padrão assumido por resíduos com uma distribuição Normal padrão.

Figura 5.19: Gráficos de probabilidade Normal dos resíduos quantílicos



(a) *Half Normal plot* com envelope simulado (100 simulações)

(b) Probabilidade quantil-quantil

Fonte: Próprio autor

Tabela 5.11: Medidas descritivas dos resíduos quantílicos do modelo proposto

Medida	Valor
Média	-0,0246
Variância	1,011
Coeficiente de assimetria	0,0167
Coeficiente de curtose	2,8664
Coeficiente de correlação de Filliben	0,9991

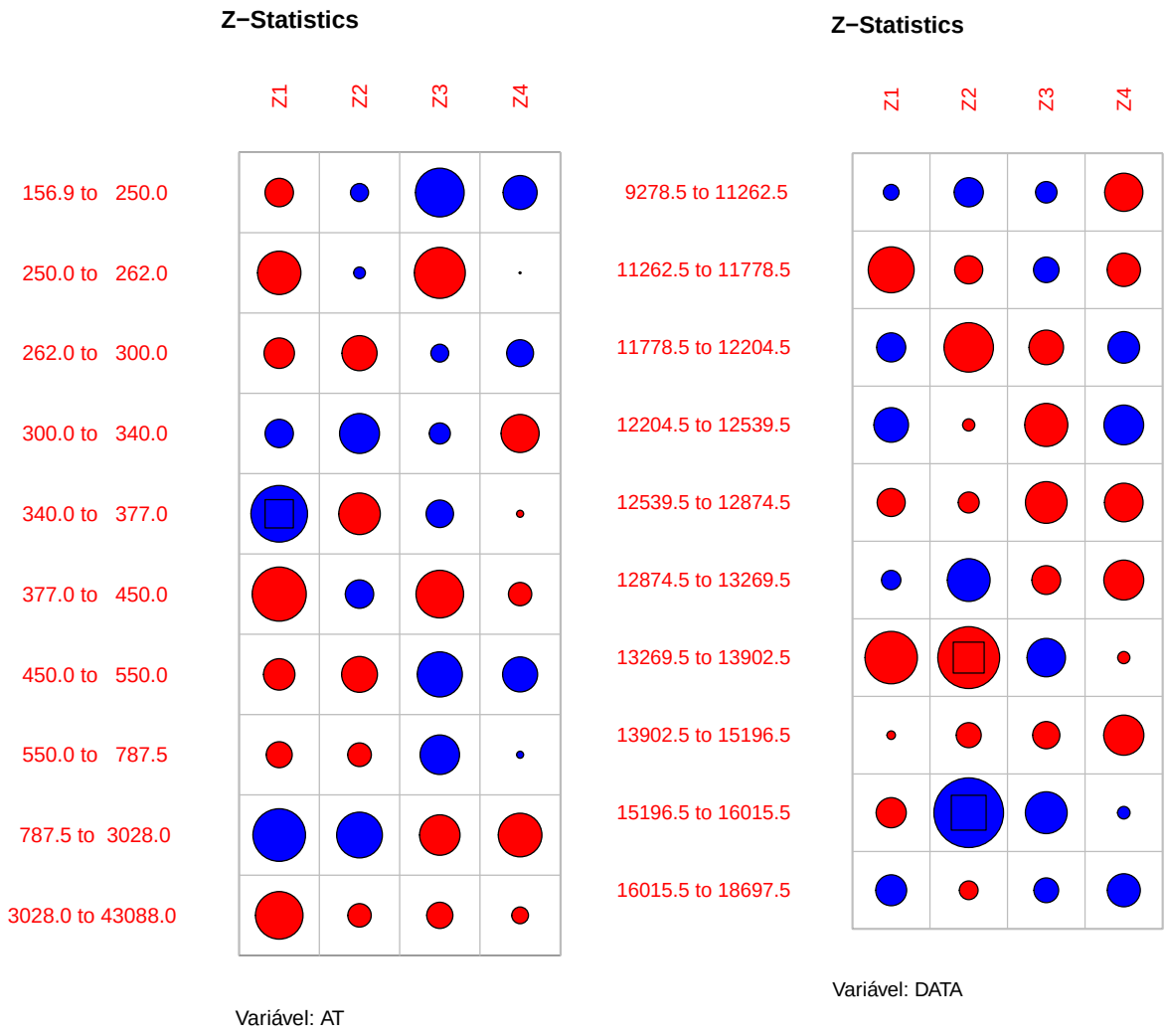
Fonte: Próprio autor

5.2.5.3 Estatísticas Z

Os resultados de uma análise complementar da Normalidade dos resíduos quantílicos pelos valores da estatística Z, para cada uma das variáveis contínuas, estão expostos nas representações gráficas das Figuras 5.20 e 5.21. Nelas a magnitude dos círculos mostrados é proporcional ao valor calculado da estatística Z_j em um dos grupos g_i e as cores indicam seu sinal (vermelho para negativo e azul para positivo). Valores superiores a $|2|$ recebem um destaque pela inserção de um quadrado como meio de destaque.

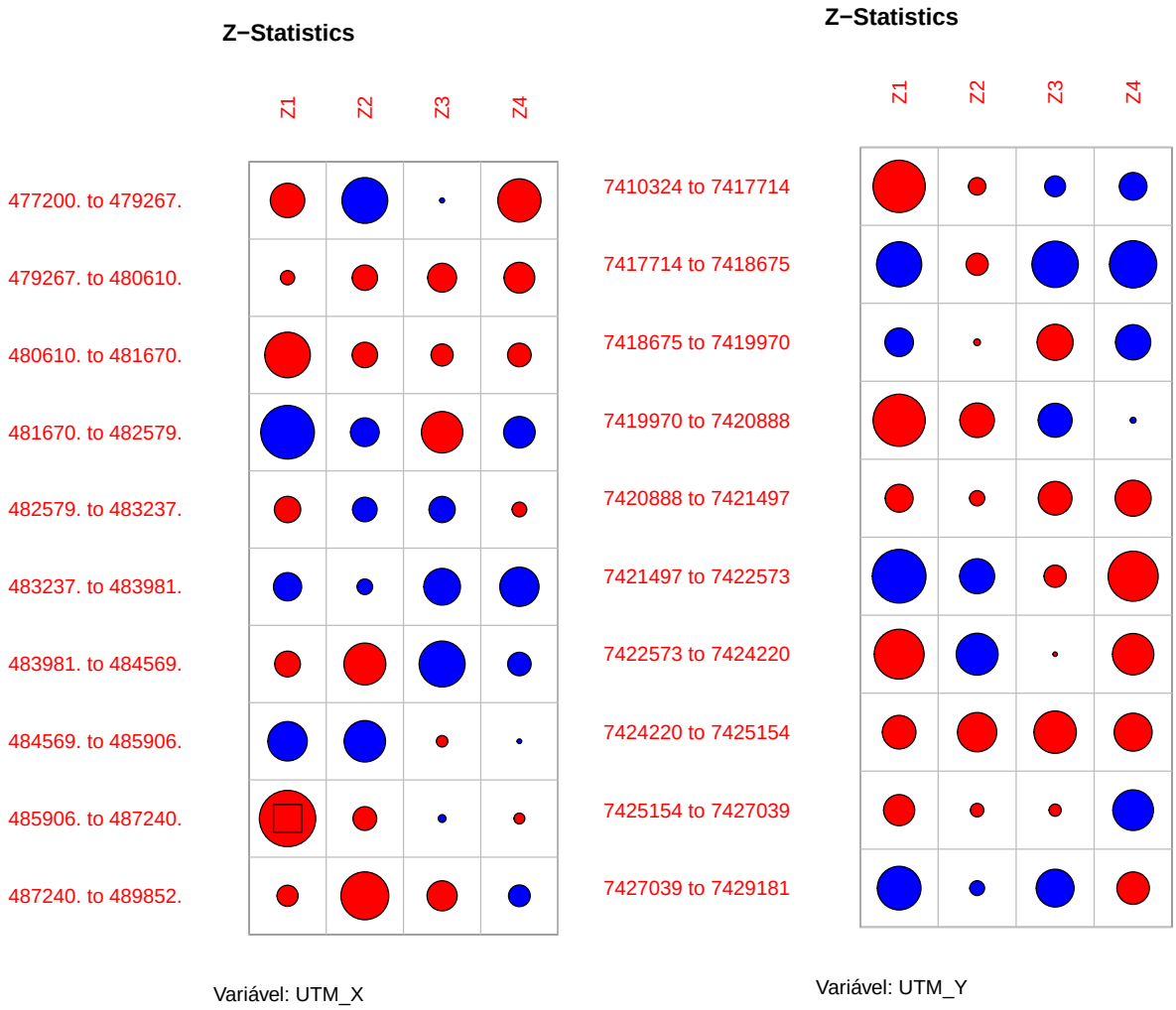
O aspecto visual indicado pela alternância das cores, bem como a magnitude dos círculos em cada um dos 10 grupos das estatísticas Z calculados para cada variável, não nos permitem dizer que destoem em muito do padrão assumido por resíduos com uma distribuição Normal padrão.

Figura 5.20: Representação gráfica dos valores da estatística Z em cada faixa definida para as variáveis “AT” e “DATA”



Fonte: Próprio autor

Figura 5.21: Representação gráfica dos valores da estatística Z em cada faixa definida para as variáveis “UTM_X” e “UTM_Y”



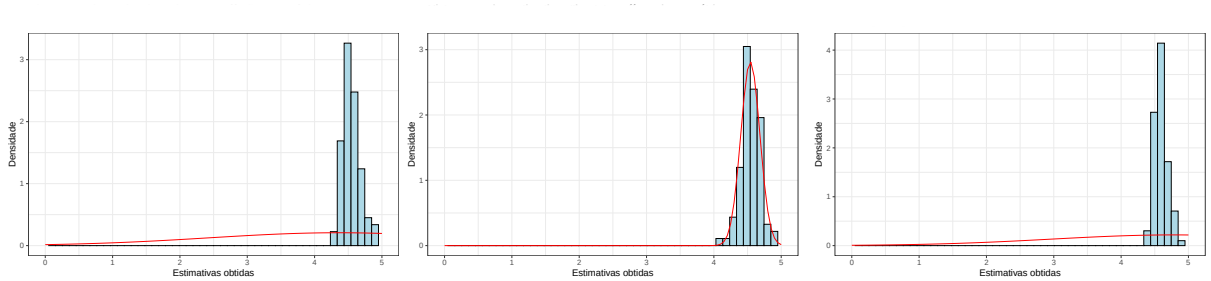
Fonte: Próprio autor

5.2.6 Inferências sobre coeficientes estimados

Os intervalos de confiança de Wald, baseados nos erros-padrão robustos com 95 % confiança, para os coeficientes dos termos considerados de forma paramétrica no preditor da mediana da distribuição da resposta ($\eta_1 = \log(\hat{\mu})$) do modelo proposto (os fatores "se NATUREZA: Transação", "se IMPLANTAÇÃO: Condomínio", "se RELEVO: Aclive", "se RELEVO: Declive" e "se PAVIMENTACAO: Pav. asfáltica"), assim como intercepto estimado, são apresentados e comparados aos intervalos de confiança construídos segundo as três propostas de *bootstrap* apresentadas por Stasinopoulos et al. (2017) (paramétrico, semiparamétrico e não paramétrico, sob 100 replicações).

As Figuras 5.22, 5.23, 5.24, 5.25, 5.26 e 5.27 ilustram a distribuição das estimativas *bootstrap* e as Tabelas 5.12, 5.13, 5.14, 5.15, 5.16 e 5.17 apresentam as estimativas ($\hat{\beta}$) e as estimativas *bootstrap* ($\bar{\beta}^*$) obtidas. Os limites para o intervalo de confiança de 95 % foram estabelecidos a partir dos quantis 0,025 e 0,975 dos afastamentos δ_b^* medidos entre a estimativa do parâmetro sob análise ($\hat{\beta}$) e cada uma das 100 estimativas *bootstrap* ($\hat{\beta}^*$) tais que $\delta_b^* = \hat{\beta} - \beta_b^*,\ b = 1, \ldots, B$ (ORLOFF; BLOOM, 2014). Assim, os limites inferior e superior do intervalo de confiança assumem os valores $[\hat{\beta} - \delta^*_{(1-\alpha/2=0{,}025)}; \hat{\beta} + \delta^*_{(\alpha/2=0{,}975)}]$.

Figura 5.22: Histogramas das estimativas *bootstrap* do intercepto do modelo proposto



(a) *Bootstrap* não paramétrico (b) *Bootstrap* semiparamétrico (c) *Bootstrap* paramétrico

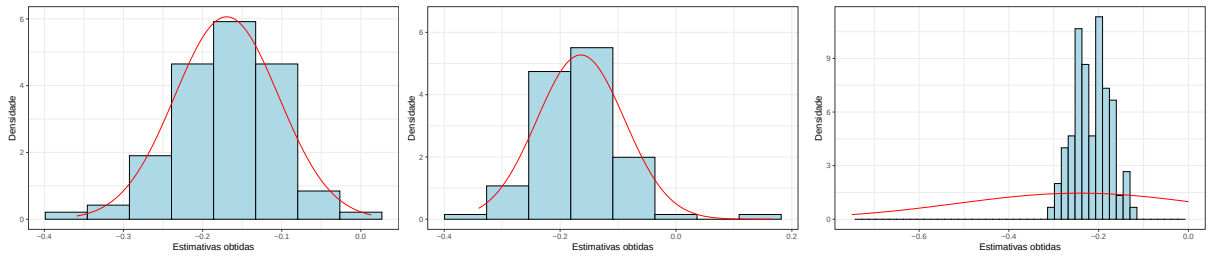
Fonte: Próprio autor

Tabela 5.12: Limites do intervalo de confiança para as estimativas do intercepto do modelo proposto

Tipo	Limite inferior	Estimativas	Limite superior
Intervalos de Wald sob erros-padrão robustos	$\hat{\beta}_{(\alpha/2=0,025)} = 4,51$	$\hat{\beta} = 4,53$	$\hat{\beta}_{(1-\alpha/2=0,975)} = 4,55$
bootstrap não paramétrico $\delta^*_{(\alpha/2=0,025)} = -0,20$; $\delta^*_{(1-\alpha/2=0,975)} = 0,40$	$\hat{\beta}^*_{(\alpha/2=0,025)} = 4,33$	$\bar{\beta}^* = 4,36$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 4,93$
bootstrap semiparamétrico $\delta^*_{(\alpha/2=0,025)} = -0,24$; $\delta^*_{(1-\alpha/2=0,975)} = 0,30$	$\hat{\beta}^*_{(\alpha/2=0,025)} = 4,29$	$\bar{\beta}^* = 4,54$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 4,83$
bootstrap paramétrico $\delta^*_{(\alpha/2=0,025)} = -0,09x$; $\delta^*_{(1-\alpha/2=0,975)} = 0,41$	$\hat{\beta}^*_{(\alpha/2=0,025)} = 4,44$	$\bar{\beta}^* = 4,79$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 4,94$

Fonte: Próprio autor

Figura 5.23: Histogramas das estimativas *bootstrap* do coeficiente do fator: “NATUREZA” (se transação)



(a) *Bootstrap* não paramétrico (b) *Bootstrap* semiparamétrico (c) *Bootstrap* paramétrico

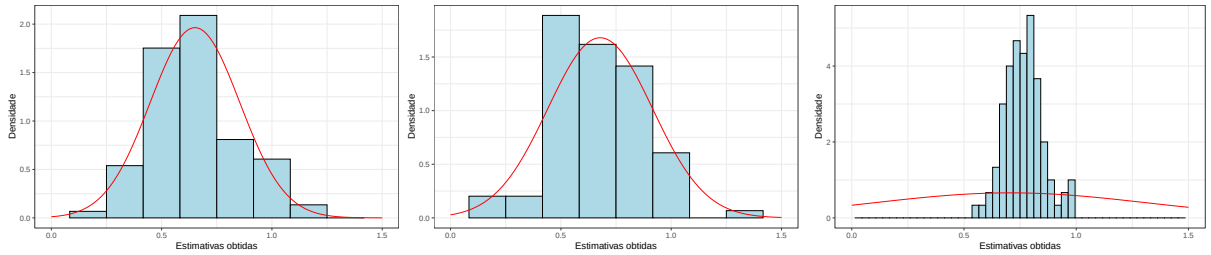
Fonte: Próprio autor

Tabela 5.13: Limites do intervalo de confiança para as estimativas do coeficiente do fator: “NATUREZA” (se transação)

Tipo	Limite inferior	Estimativas	Limite superior
Intervalos de Wald sob erros-padrão robustos	$\hat{\beta}_{(\alpha/2=0,025)} = -0,23$	$\hat{\beta} = -0,21$	$\hat{\beta}_{(1-\alpha/2=0,975)} = -0,19$
bootstrap não paramétrico ($\delta^*_{(\alpha/2=0,025)} = -0,08$, $\delta^*_{(1-\alpha/2=0,975)} = 0,16$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = -0,29$	$\bar{\beta}^* = -0,17$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = -0,05$
bootstrap semiparamétrico ($\delta^*_{(\alpha/2=0,0250} = -0,09$, $\delta^*_{(1-\alpha/2=0,975)} = 017$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = -0,30$	$\bar{\beta}^* = -0,16$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = -0,04$
bootstrap paramétrico ($\delta^*_{(\alpha/2=0,025)} = -0,08$, $\delta^*_{(1-\alpha/2=0,975)} = 0,08$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = -0,29$	$\bar{\beta}^* = -0,24$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = -0,13$

Fonte: Próprio autor

Figura 5.24: Histogramas das estimativas *bootstrap* do coeficiente do fator: "IMPLANTACAO" (se condomínio)



(a) *Bootstrap* não paramétrico (b) *Bootstrap* semiparamétrico (c) *Bootstrap* paramétrico

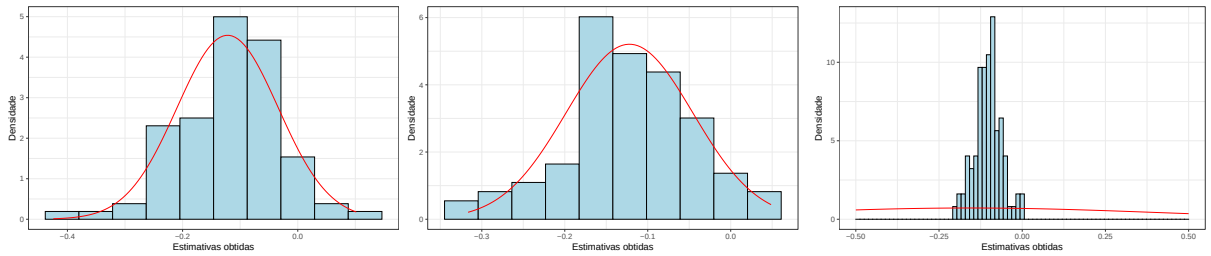
Fonte: Próprio autor

Tabela 5.14: Limites do intervalo de confiança para as estimativas do coeficiente do fator: "IMPLANTACAO" (se condomínio)

Tipo	Limite inferior	Estimativas	Limite superior
Intervalos de Wald sob erros-padrão robustos	$\hat{\beta}_{(\alpha/2=0,025)} = 0,72$	$\hat{\beta} = 0,75$	$\hat{\beta}_{(1-\alpha/2=0,975)} = 0,79$
bootstrap não paramétrico ($\delta^*_{(\alpha/2=0,025)} = -0,46$, $\delta^*_{(1-\alpha/2=0,975)} = 0,27$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = 0,29$	$\bar{\beta}^* = 0,65$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 1,03$
bootstrap semiparamétrico ($\delta^*_{(\alpha/2=0,025)} = -0,49$, $\delta^*_{(1-\alpha/2=0,975)} = 0,31$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = 0,26$	$\hat{\beta}^* = 0,68$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 1,06$
bootstrap paramétrico ($\delta^*_{(\alpha/2=0,025)} = -0,16$, $\delta^*_{(1-\alpha/2=0,975)} = 0,21$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = 0,60$	$\bar{\beta}^* = 0,71$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 0,96$

Fonte: Próprio autor

Figura 5.25: Histogramas das estimativas *bootstrap* do coeficiente do fator: “RELEVO” (se aclive)



(a) *Bootstrap* não paramétrico (b) *Bootstrap* semiparamétrico (c) *Bootstrap* paramétrico

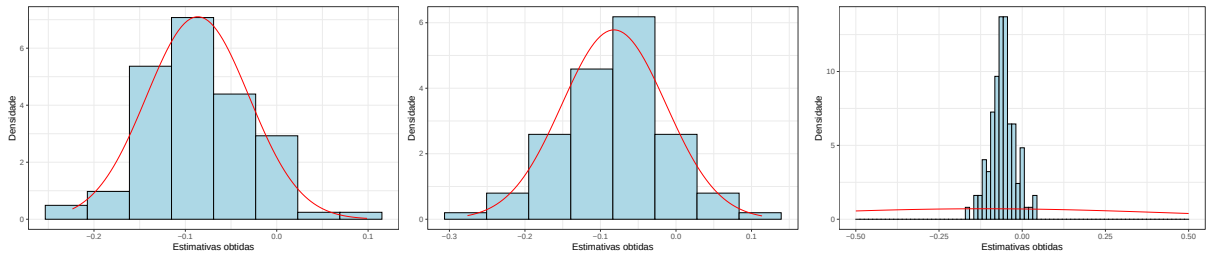
Fonte: Próprio autor

Tabela 5.15: Limites do intervalo de confiança para as estimativas do coeficiente do fator: “RELEVO” (se aclive)

Tipo	Limite inferior	Estimativas	Limite superior
Intervalos de Wald sob erros-padrão robustos	$\hat{\beta}_{(\alpha/2=0{,}025)} = -0,14$	$\hat{\beta} = -0,11$	$\hat{\beta}_{(1-\alpha/2=0{,}975)} = -0,08$
bootstrap não paramétrico ($\delta^*_{(\alpha/2=0{,}025)} = -0,19$, $\delta^*_{(1-\alpha/2=0{,}975)} = 0,14$)	$\hat{\beta}^*_{(\alpha/2=0{,}025)} = -0,30$	$\bar{\beta}^* = -0,12$	$\hat{\beta}^*_{(1-\alpha/2=0{,}975)} = 0,03$
bootstrap semiparamétrico ($\delta^*_{(\alpha/2=0{,}025)} = -0,18$, $\delta^*_{(1-\alpha/2=0{,}975)} = 0,13$)	$\hat{\beta}^*_{(\alpha/2=0{,}025)} = -0,29$	$\bar{\beta}^* = -0,12$	$\hat{\beta}^*_{(1-\alpha/2=0{,}975)} = 0,02$
bootstrap paramétrico ($\delta^*_{(\alpha/2=0{,}025)} = 0,08$, $\delta^*_{(1-\alpha/2=0{,}975)} = 0,10$)	$\hat{\beta}^*_{(\alpha/2=0{,}025)} = -0,19$	$\bar{\beta}^* = -0,16$	$\hat{\beta}^*_{(1-\alpha/2=0{,}975)} = -0,01$

Fonte: Próprio autor

Figura 5.26: Histogramas das estimativas *bootstrap* do coeficiente do fator: “RELEVO” (se declive)



(a) *Bootstrap* não paramétrico (b) *Bootstrap* semiparamétrico (c) *Bootstrap* paramétrico

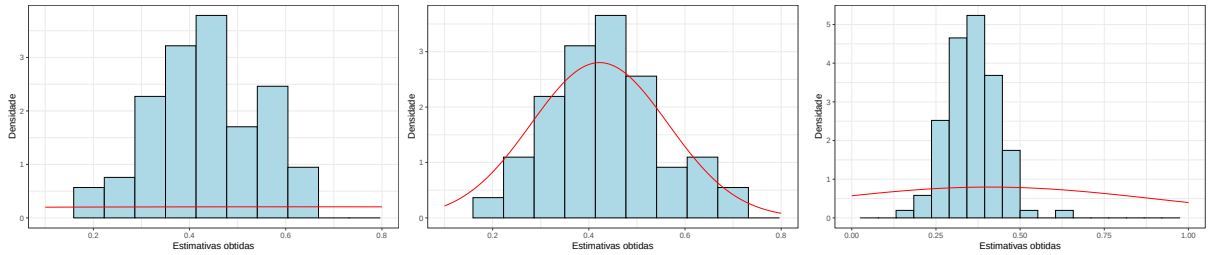
Fonte: Próprio autor

Tabela 5.16: Limites do intervalo de confiança para as estimativas do coeficiente do fator: “RELEVO” (se declive)

Tipo	Limite inferior	Estimativas	Limite superior
Intervalos de Wald sob erros-padrão robustos	$\hat{\beta}_{(\alpha/2=0{,}025)} = -0,08$	$\hat{\beta} = -0,06$	$\hat{\beta}_{(1-\alpha/2=0{,}975)} = -0,04$
bootstrap não paramétrico ($\delta^*_{(\alpha/2=0{,}025)} = -0,14$, $\delta^*_{(1-\alpha/2=0{,}975)} = 0,07$)	$\hat{\beta}^*_{(\alpha/2=0{,}025)} = -0,20$	$\bar{\beta}^* = -0,09$	$\hat{\beta}^*_{(1-\alpha/2=0{,}975)} = 0,01$
bootstrap semiparamétrico ($\delta^*_{(\alpha/2=0{,}025)} = -0,16$, $\delta^*_{(1-\alpha/2=0{,}975)} = 0,13$)	$\hat{\beta}^*_{(\alpha/2=0{,}025)} = -0,22$	$\bar{\beta}^* = -0,08$	$\hat{\beta}^*_{(1-\alpha/2=0{,}975)} = 0,07$
bootstrap paramétrico ($\delta^*_{(\alpha/2=0{,}025)} = -0,08$, $\delta^*_{(1-\alpha/2=0{,}975)} = 0,08$)	$\hat{\beta}^*_{(\alpha/2=0{,}025)} = -0,14$	$\bar{\beta}^* = -0,11$	$\hat{\beta}^*_{(1-\alpha/2=0{,}975)} = 0,02$

Fonte: Próprio autor

Figura 5.27: Histogramas das estimativas *bootstrap* do coeficiente do fator: "PAVIMENTACAO" (se asfáltica)



(a) *Bootstrap* não paramétrico (b) *Bootstrap* semiparamétrico (c) *Bootstrap* paramétrico

Fonte: Próprio autor

Tabela 5.17: Limites do intervalo de confiança para as estimativas do coeficiente do fator: "PAVIMENTACAO" (se asfáltica)

Tipo	Limite inferior	Estimativas	Limite superior
Intervalos de Wald sob erros-padrão robustos	$\hat{\beta}_{(\alpha/2=0,025)} = 0,35$	$\hat{\beta} = 0,37$	$\hat{\beta}_{(1-\alpha/2=0,975)} = 0,39$
bootstrap não paramétrico ($\delta^*_{(\alpha/2=0,025)} = -0,37$, $\delta^*_{(1-\alpha/2=0,975)} = 0,27$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = 0$	$\bar{\beta}^* = 0,61$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 0,64$
bootstrap semiparamétrico ($\delta^*_{(\alpha/2=0,025)} = -0,27$, $\delta^*_{(1-\alpha/2=0,975)} = 0,31$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = 0,10$	$\bar{\beta}^* = 0,42$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 0,68$
bootstrap paramétrico ($\delta^*_{(\alpha/2=0,025)} = -0,14$, $\delta^*_{(1-\alpha/2=0,975)} = 0,14$)	$\hat{\beta}^*_{(\alpha/2=0,025)} = 0,23$	$\bar{\beta}^* = 0,41$	$\hat{\beta}^*_{(1-\alpha/2=0,975)} = 0,51$

Fonte: Próprio autor

5.2.7 Autocorrelação espaçotemporal nos resíduos simples

Admitindo-se que as estruturas paramétrica e não paramétrica consideradas para modelar a variabilidade espacial e temporal no modelo proposto tenham sido adequadamente incorporadas aos preditores dos parâmetros distribucionais e bem ajustadas, espera-se que os resíduos simples desse modelo não exibam nenhum padrão de autocorrelação espacial e, ou, temporal. Tal verificação pode ser realizada pela construção do semivariograma empírico. A ausência de padrões que sugiram qualquer associação dos valores da semivariância com a localização espacial e o tempo assegura que os resíduos são independentes. Winkle, Zammit-Mangion e Cressie (2019) expõem, de modo didático, a construção dos objetos computacionais necessários à estimação do variograma experimental por meio das bibliotecas spacetime e sp, implementadas em R.

Ao contrário da correlação e da covariância, a semivariância é uma medida de dissimilaridade dos valores W_i e W_j de duas observações espacialmente distantes h m e temporalmente afastadas u dias, dada por (5.2).

$$\hat{\varphi}_{(h,u)} = \frac{1}{2N_{(h,u)}} \sum_{N(h,u)} \{[W(s_i,t_i) - W(s_j,t_j)]\}^2, \tag{5.2}$$

em que $N_{(h,u)}$ é o número de pares de observações distantes espacialmente h m e temporalmente u dias.

Resíduos autocorrelacionados espaçotemporalmente exibem valores de semivariância crescentes com o aumento da distância espacial ou o afastamento temporal entre eles, diferentemente do padrão observado nas Figuras 5.28 e 5.29, que expõem a semivariância calculada entre pares de observações espacial ou temporalmente adjacentes seguindo os seguintes parâmetros:

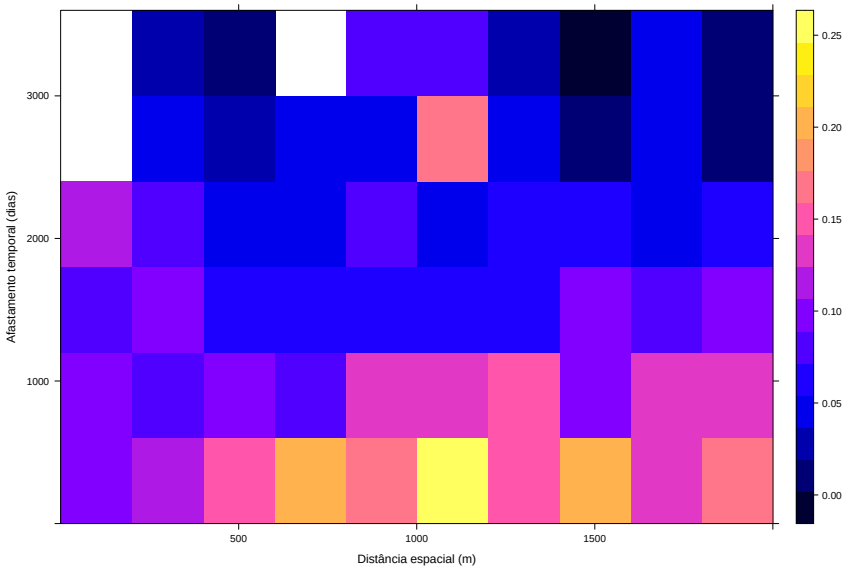
- a máxima distância espacial entre duas observações consideradas para o cálculo da semivariância é de 2000 m;
- o máximo afastamento temporal entre duas observações para o cálculo da semivariância é de 120 dias.

5.2.8 Coeficiente de determinação generalizado

A Figura 5.30 ilustra a relação entre os valores estimados pelo modelo e os observados. O coeficiente de determinação generalizado R^2 do modelo proposto, tal como proposto por Nagelkerke (1991)[1] é de 0,894.

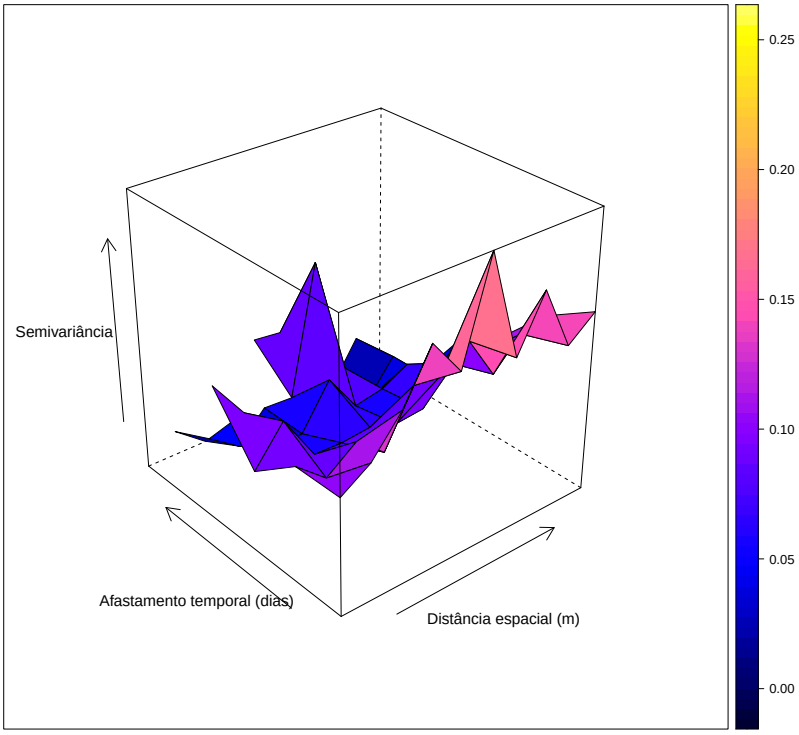
[1] O coeficiente de determinação assim especificado é ligeiramente diferente do proposto por Cook e Weisberg (1982).

Figura 5.28: Variograma espaçotemporal dos resíduos simples do modelo proposto



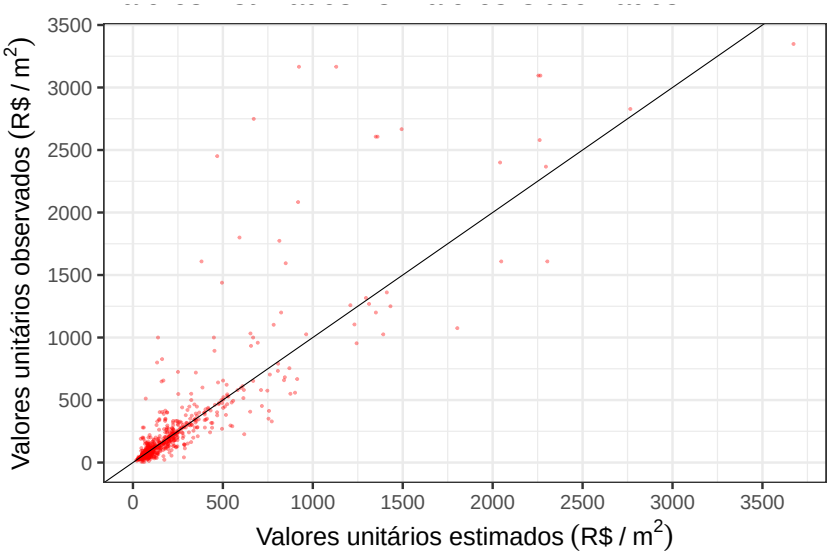
Fonte: Próprio autor

Figura 5.29: Variograma espaçotemporal dos resíduos simples do modelo proposto



Fonte: Próprio autor

Figura 5.30: Valores estimados *versus* observados



Fonte: Próprio autor

5.2.9 Superfícies de valor

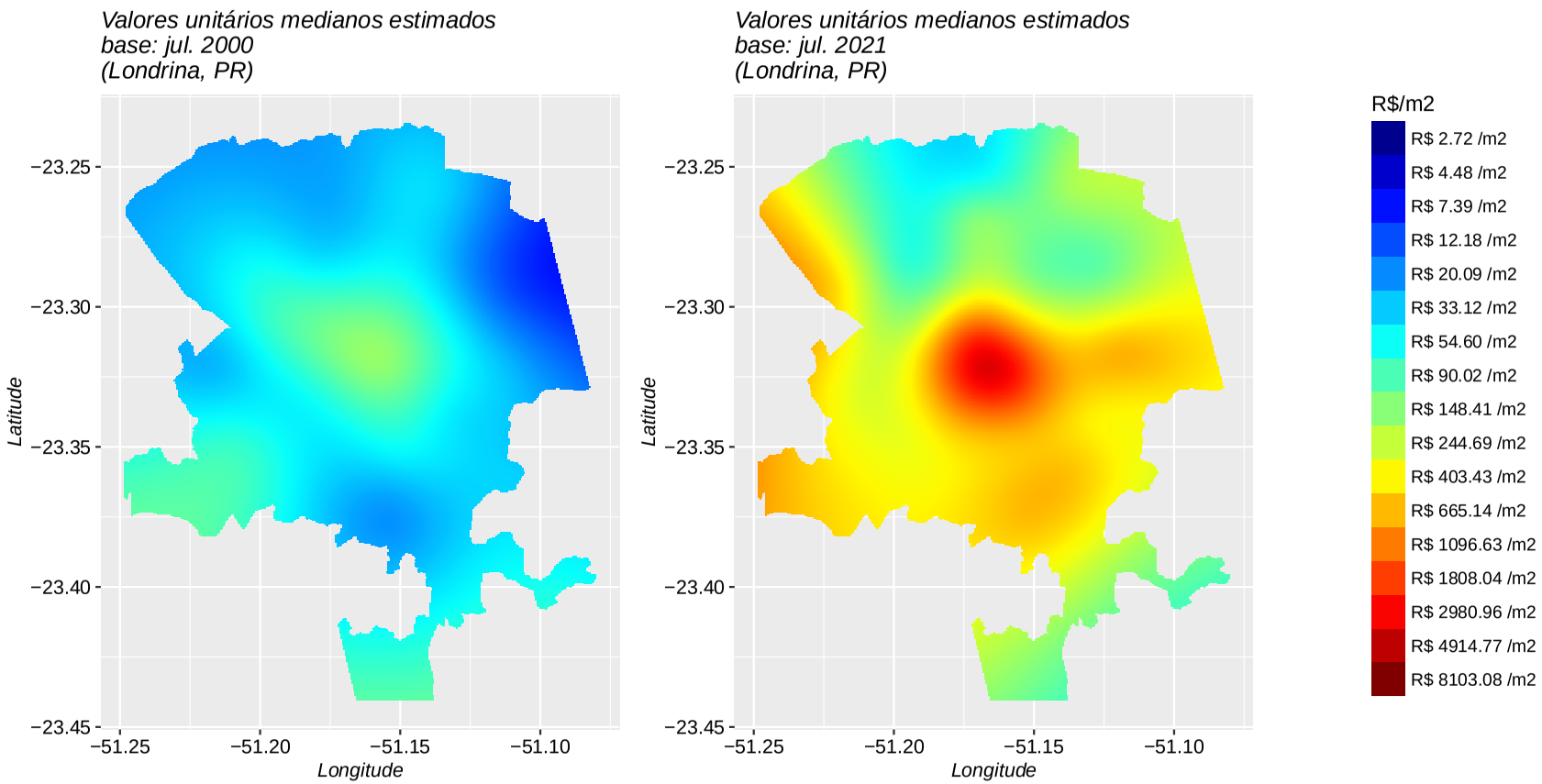
A situação paradigma do imóvel a ser estimado, de modo reiterado, em cada localização espacial definida no *grid* criado, necessária a geração de superfícies de valor sob diferentes tempos de referência foi assim estabelecida:

- área total de 377 m^2 (a mediana da variável);
- relevo plano;
- que a resposta média se refira a uma transação efetivamente concluída;
- esteja situado defronte a uma via com pavimentação asfáltica;
- esteja implantado de modo isolado;
- datas de referência: jul. 2000; jul. 2021 e também de jan. 2000 a jul.2021 em intervalos temporais de 6 meses.

As superfícies geradas e apresentadas na Figura 5.31 expõem a variabilidade do valor unitário mediano para um terreno, estimado sob a situação paradigma antes estabelecida, em dois períodos temporais distintos: jul. 2000 e jul. 2021.

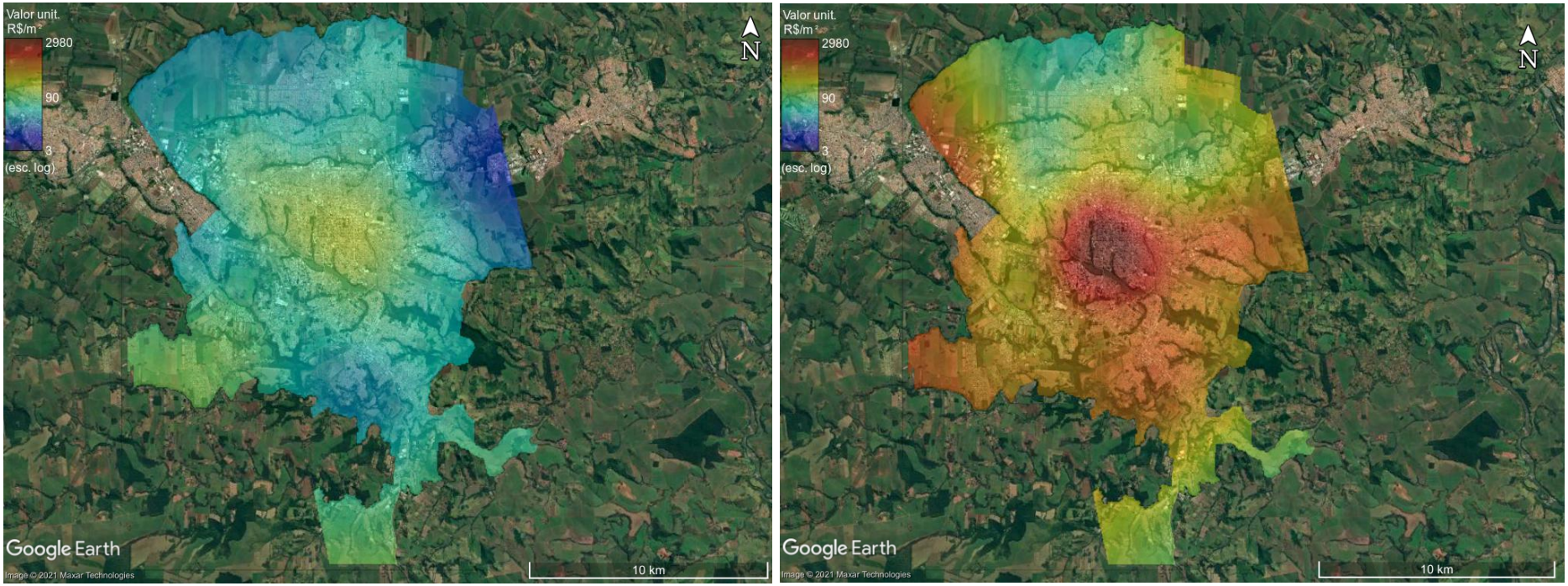
Uma representação alternativa pode ser observada na Figura 5.32, pela justaposição das mesmas superfícies a uma imagem de satélite da cidade de Londrina, visando facilitar a identificação das regiões geográficas da área do estudo.

Figura 5.31: Alterações do valor unitário mediano na área delimitada pelo perímetro urbano de Londrina com recortes temporais definidos para jun. 2000 e jun. 2021 (sob a situação definida como paradigma)



Fonte: Próprio autor

Figura 5.32: Superfície de valores unitários medianos estimados (sob a situação definida como paradigma)

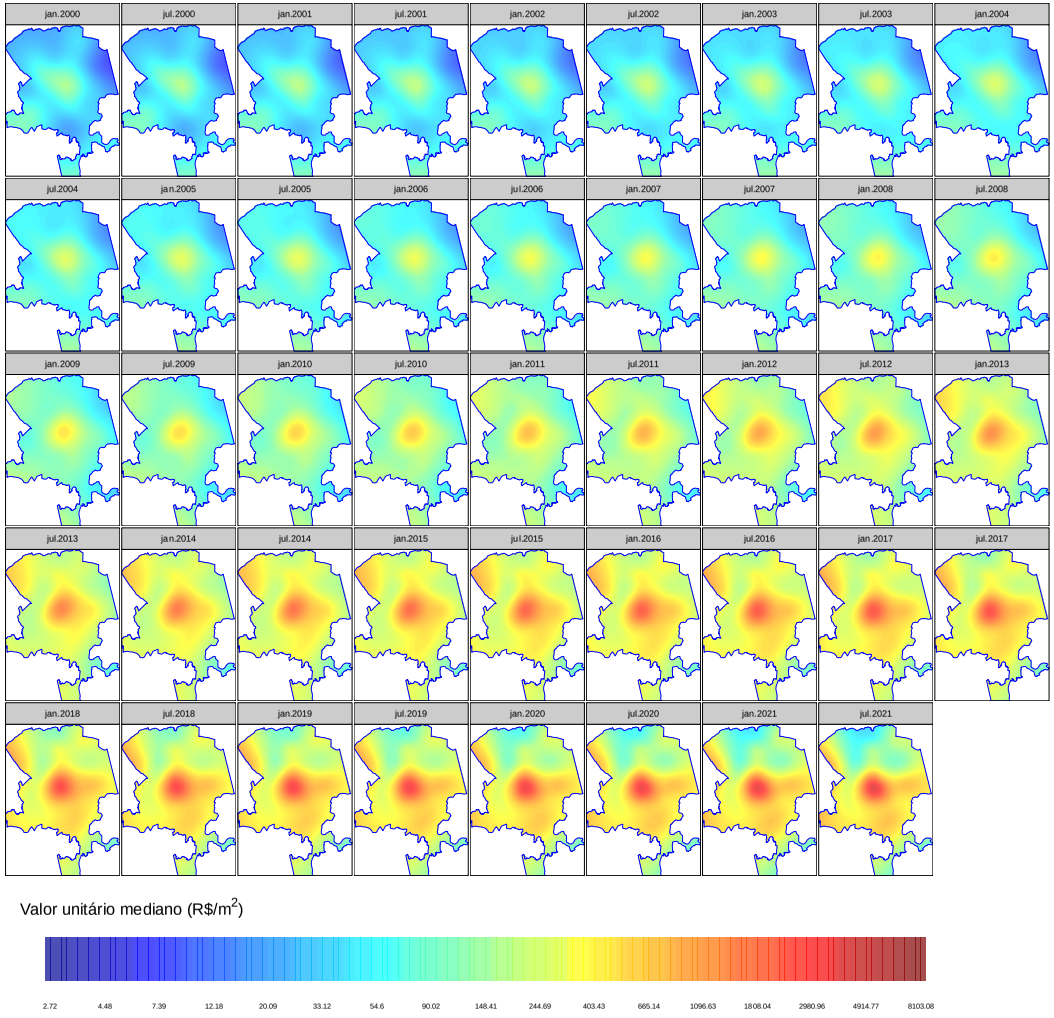


(a) Data de referência jul. 2000

(b) Data de referência jul. 2021

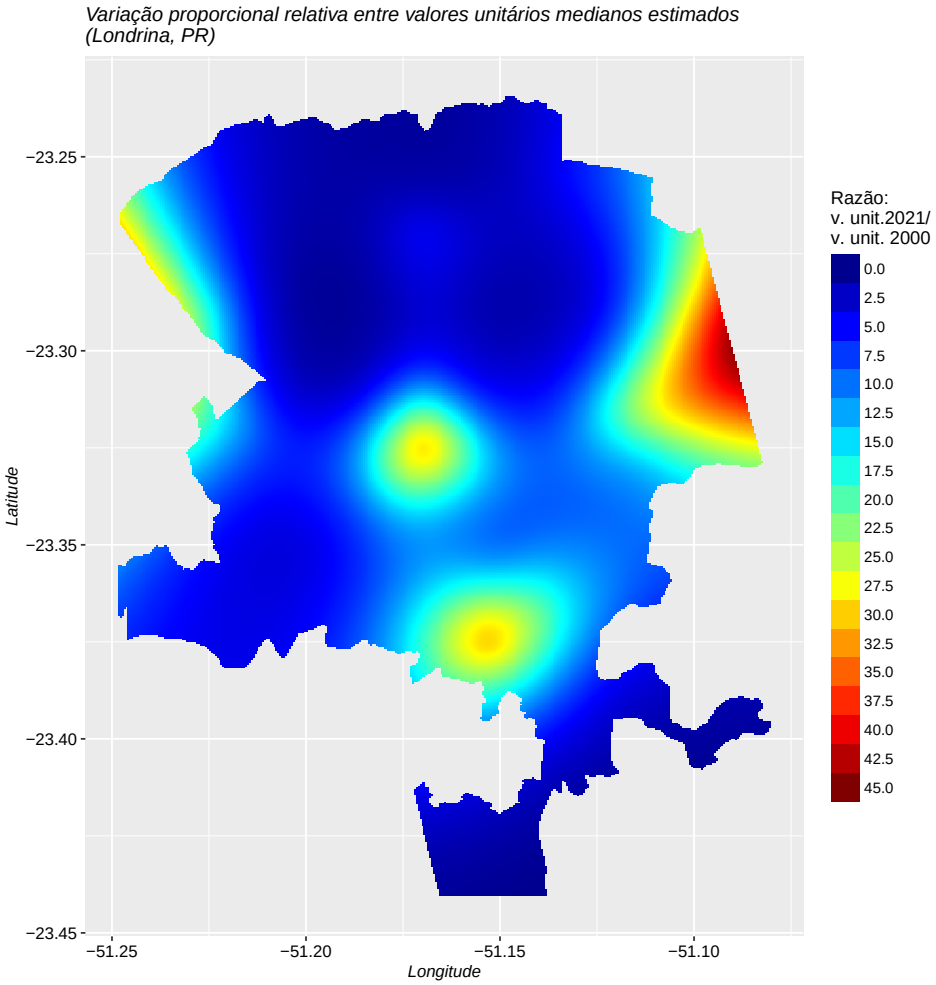
Fonte: Próprio autor (imagem de fundo: Google Earth Pro, 2021 Londrina, PR nas coordenadas 23,3197S;51,1662W)

Figura 5.33: Alterações do valor unitário mediano na área delimitada pelo perímetro urbano de Londrina com recortes temporais definidos semestralmente ao longo do período de jan. 2000 a jul. 2021 (sob a situação definida como paradigma)



Fonte: Próprio autor

Figura 5.34: Variação proporcional relativa entre o valor unitário mediano na área delimitada pelo perímetro urbano de Londrina entre as datas de jul. 2000 e jul. 2021 (sob a situação definida como paradigma)



Fonte: Próprio autor

5.2.10 Variabilidade espaçotemporal

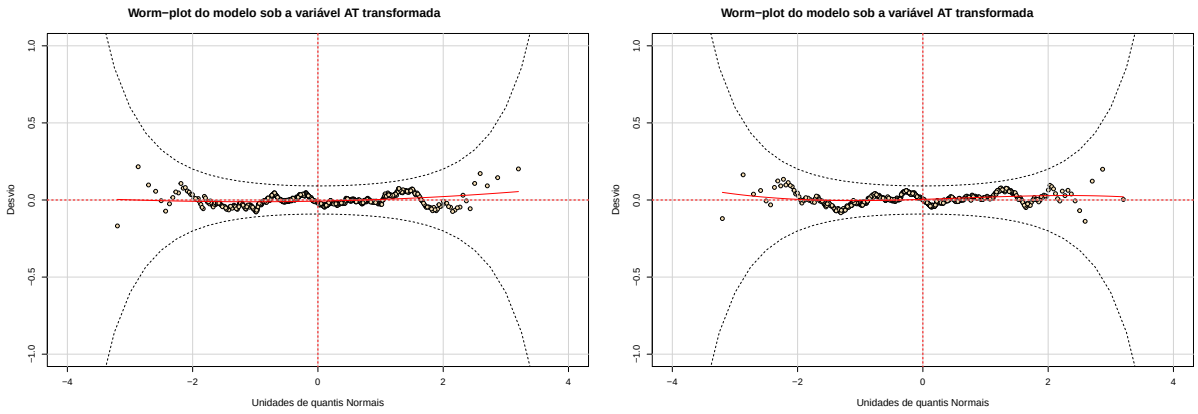
A Figura 5.34 ilustra a variação relativa entre os valores unitários medianos estimados para jul. 2021 e os de jul. 2000 na área delimitada pelo perímetro urbano de Londrina, corroborando a afirmação feita na parte inicial desse trabalho que diferentes regiões de uma cidade se valorizam (ou desvalorizam) de modos distintos ao longo do tempo .

5.2.11 Transformação da variável "AT"

Com o propósito de melhor investigar o perfil do efeito da variável "AT" exibido na Figura 5.10, ajustam-se dois outros modelos que, embora possuam idêntica estrutura que a do modelo proposto, a variável "AT" está transformada pela aplicação da função logaritmo e de uma transformação de potência ao ser incorporada na estrutura do preditor do primeiro parâmetro da distribuição.

Os *worm plots* exibidos na Figura 5.35 exibem um perfil tão semelhante, quanto indicativo de bom ajuste, ao do modelo proposto.

Figura 5.35: *Worm plots* de modelos com estrutura idêntica à do modelo proposto, mas sob variadas transformações da variável "AT"

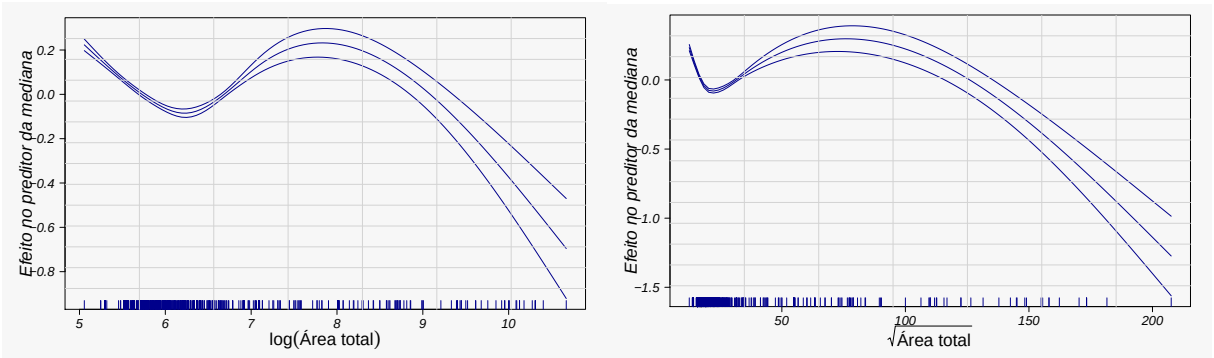


(a) Variável transformada: $\log(AT)$ (b) Variável transformada: $\sqrt{AT}$

Próprio autor

Os perfis do efeito da variável "AT" exibidos na Figura 5.36, considerada sob essas duas transformações, no preditor da mediana distribucional ainda mostram o mesmo comportamento que o perfil exibido na Figura 5.10, um ponto de inflexão para uma área total aproximada de 400 m^2 e dois máximos (um para as menores áreas totais da amostra e outro para uma área aproximada de 3.000 m^2).

Figura 5.36: Efeitos da variável "AT" sob variadas transformações, sobre a mediana distribucional (incorporadas ao seu preditor linear sob a forma de função suavizadora unidimensional)



(a) Variável transformada: $\log(AT)$, edf=3,96 (b) Variável transformada: $\sqrt{AT}$, edf=3,99

Próprio autor

Algumas métricas levantadas para os dois novos modelos e exibidas na Tabela 5.18 mostram valores ligeiramente mais favoráveis em relação ao modelo proposto, como

exposto na Tabela 5.2. Todavia, tais diferenças não são estatisticamente significativas para se estabelecer distinção entre os três modelos como apresentado nos Quadros 5.2 e 5.3. De modo complementar, no Teste de Clarke entre o modelo proposto e aquele que apresentava a variável "AT" sob uma transformação de potência, o resultado pendeu favoravelmente para o modelo proposto razão pela qual o mantivemos.

Tabela 5.18: Informações sobre algumas métricas dos modelos sob variadas transformações da variável "AT"

Métrica	Modelo sob $\log(AT)$	Modelo sob $\sqrt{AT}$
Deviance global	7566,32	7556,74
AIC	7829,76	7828,83
Pseudo R^2 (Cox&Snell)	0,901	0,903

Fonte: Próprio autor

Quadro 5.2: Diferenças estatisticamente significativas pelo Teste de Voung ($\alpha = 0,05$)

Modelos	Modelo proposto (1)	Modelo sob $\log(AT)$ (2)	Modelo sob $\sqrt(AT)$ (3)
Modelo proposto (1)	-	indiferente	indiferente
Modelo sob $\log(AT)$ (2)	indiferente	-	indiferente
Modelo sob $\sqrt(AT)$ (3)	indiferente	indiferente	-

Fonte: Próprio autor

Quadro 5.3: Diferenças estatisticamente significativas pelo Teste de Clarke ($\alpha = 0,05$)

Modelos	Modelo proposto (1)	Modelo sob $\log(AT)$ (2)	Modelo sob $\sqrt(AT)$ (3)
Modelo proposto (1)	-	indiferente	preferível (1)
Modelo sob $\log(AT)$ (2)	indiferente	-	preferível (2)
Modelo sob $\sqrt(AT)$ (3)	preferível (1)	preferível (2)	-

Fonte: Próprio autor

6 CONCLUSÃO

A metodologia empregada na modelagem mostrou-se adequada para o propósito que se desejava, estimar os valores unitários medianos de terrenos urbanos, sob uma situação paradigma pré-estabelecida, para qualquer localização na área estudada e data dentro dos horizontes temporais da amostra.

As estruturas empregadas para a modelagem da variabilidade espacial, temporal e espaçotemporal propiciaram estimativas cujos resíduos brutos não apresentaram padrões sistemáticos de autocorrelação espacial e temporal.

As variadas representações gráficas e animações visuais geradas com as estimativas produzidas com o modelo ilustram que as alterações, sejam valorização ou desvalorização, não podem ser admitidas como espacialmente uniformes ao longo da escala temporal, tal como foi afirmado na parte inicial desse trabalho.

Os valores unitários medianos estimados pelo modelo podem também ser incorporados na forma de uma variável que busque explicar a variabilidade espaçotemporal em Modelos Hedônicos de Regressão de outras tipologias de imóveis tais como casas, apartamentos, lojas, galpões, uma vez que a localização de um imóvel é uma variável comum a todas elas.

REFERÊNCIAS

ABNT: Associação Brasileira de Normas Técnicas. *NBR 14.653-2:2011*: Avaliação de bens parte 2: imóveis urbanos. Rio de Janeiro, RJ, 2011. 54 p.

ABNT: Associação Brasileira de Normas Técnicas. *NBR 14.653-1:2019*: Avaliação de bens parte 1: procedimentos gerais. Rio de Janeiro, RJ, 2019. 19 p.

ALONSO, N. R. P. (org.). *Engenharia de avaliações*. São Paulo, SP: Editora Pini, 2007. 987 p.

APPELHANS, T. et al. *mapview: Interactive Viewing of Spatial Data in R*. [S.l.], 2021. Versão 2.10.0, disponível em: .

BARBOSA, A. G.; COSTA, A. A. O solo urbano e a apropriação da natureza na cidade. *Sociedade & Natureza*, v. 24, n. 3, p. 477–488, 2012. Disponível em: . Acesso em: abr. 2021.

BECKER, R. A. et al. *maps: Draw Geographical Maps*. [S.l.], 2021. Versão 3.4.0, disponível em: .

BERRINI, L. C. *Avaliações de imóveis*. São Paulo, SP: (publicado e editado pelo próprio autor), 1949. 390 p. Disponível em: . Acesso em: abr. 2021.

BIVAND, R. et al. *maptools: Tools for Handling Spatial Objects*. [S.l.], 2022. Versão 1.1-3, disponível em: .

BIVAND, R. et al. *spData: Datasets for Spatial Analysis*. [S.l.], 2021. Versão 2.0.1, disponível em: .

BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society*, v. 26, n. 2, p. 211–252, 1964. Disponível em: . Acesso em: abr. 2021.

BROWN, L. D. Fundamentals of statistical exponential families: with applications in statistical decision theory. *Lecture Notes-Monograph Series*, v. 9, p. 279, 1986. Disponível em: . Acesso em: abr. 2021.

BRUNSDON, C.; CHEN, H. *GISTools: Some further GIS capabilities for R*. [S.l.], 2014. Versão 0.7-4, disponível em: .

BURDEN, R. L.; FAIRES, D. J.; BURDEN, A. M. *Análise Numérica*. 3. ed. São Paulo, SP: Cengage, 2017. 879 p.

CAN, A. Gis and spatial analisys of housing mortgage markets. *Journal of housing research*, v. 9, n. 1, p. 61–86, 1998. Disponível em: . Acesso em: abr. 2021.

CHICA-OLMO, J. *Teoría de las variables regionalizadas. Aplicación en economia espacial y valoración inmobiliária*. Granada, ES: Servicio de publicaciones de la Universidad de Granada, 1994. 147 p.

CHICA-OLMO, J. Prediction of housing location price by a multivariate spatial method: cokriging. *Journal of Real Estate Research*, v. 29, n. 1, p. 91–114, 2007. Disponível em: . Acesso em: abr. 2021.

CHICA-OLMO, J.; CANO-GUERVOS, R. Does my house have a premium or discount in relation to my neighbors? a regression-kriging approach. *Socio-Economic Planning Sciences*, v. 72, 07 2020. Disponível em: . Acesso em: jan. 2022.

CLARKE, K. A. A simple distribution-free test for nonnested model selection. *Political Analysis*, v. 15, p. 347–363, 2007. Disponível em: . Acesso em: jul. 2021.

CLEVELAND, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, v. 74, p. 829–836, 1979. Disponível em: . Acesso em: mai. 2021.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: The ims method and penalized likelihood. *Statistics in Medicine*, v. 11, p. 1305–1319, 1992. Disponível em: . Acesso em: jun. 2021.

COOK, R. D.; WEISBERG, S. *Residuals and Influence in Regression*. London, EN: Chapman and Hall, 1982. 240 p. Disponível em: . Acesso em: abr. 2021.

COOLEY, D. *mapdeck: Interactive Maps Using "Mapbox GL JS" and "Deck.gl"*. [S.l.], 2020. Versão 0.3.4, disponível em: .

COURT, A. T. Hedonic price indexes with automotive examples. *Automobile manufactures association*, p. 99–119, 1939. Disponível em: . (Acesso em: abr/2021).

COX, D. R.; HINKLEY, D. V. *Theoretical Statistics*. London, EN: Chapman & Hall/CRC, 1974. 522 p. Disponível em: . Acesso em: jul. 2021.

COX, D. R.; SNELL, E. J. *Analysis of Binary Data*. Boca Raton, FL: Chapman & Hall/CRC, 1989. 253 p.

COX, M. G. The numerical evaluation of b-splines. *Journal of Applied Mathematics*, v. 10, p. 134–149, 1972. Disponível em: . Acesso em: mai. 2021.

CURRIE, I.; REGUERA, M. L. D.; EILERS, P. H. C. Smoothing abd forecasting mortality rates. *Statistical Modelling*, v. 4, n. 4, p. 279–298, 2004. Disponível em: . Acesso em: nov 2021.

D'AGOSTINO, R. B.; BELANGER, A.; D'AGOSTINO JR., R. B. A suggestion for using powerful and informative tests of normality. *The American Statistician*, v. 44, p. 316–321, 1990. Disponível em: . Acesso em: jul. 2021.

DANTAS, R. A. Métodos matemáticos e estatísticos na engenharia de avaliações. In: *II Simpósio Mineiro de Engenharia de Avaliações e Perícias*. Belo Horizonte, MG: IBAPE-MG, 1988. p. 35–129.

DANTAS, R. A. *Engenharia de avaliações: uma introdução à metodologia científica*. São Paulo, SP: Editora Pini, 1998. 251 p.

DANTAS, R. A. *Modelos Espaciais Aplicados ao Mercado Habitacional um Estudo de Caso para a Cidade do Recife*. Tese (Doutorado em Economia) — Universidade Federal de Pernambuco, Recife, PE, 2003. Disponível em: . Acesso em: abr. 2021.

DANTAS, R. A.; CORDEIRO, G. M. A avaliação de imóveis através da metodologia de pesquisa científica. *Caderno Brasileiro de Avaliações e Perícias*, n. 26, p. 35–50, 1991.

DATAS, R. A. et al. Uma nova metodologia para a avaliação de imóveis utilizando regressão espacial. In: *XI Congresso Brasileiro de Engenharia de Avaliações e Perícias*. Guarapari, ES: IBAPE, 2001. Disponível em: . Acesso em: abr. 2021.

DE BASTIANI, F. *Inference and Diagnostics in Spatial Models*. Tese (Doutorado em Estatística) — Universidade Federal de Pernambuco, Recife, PE, 2016. Disponível em: . Acesso em: jul. 2021.

DE BOOR, C. W. R. *A Practical Guide to Splines*: Revised edition. New York, NY: Springer, 2001. Disponível em: . Acesso em: mai. 2021.

DUBIN, R. A. Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *The Review of Economics and Statistics*, v. 70, n. 3, p. 466–474, 1988. Disponível em: . Acesso em: abr. 2021.

DUBIN, R. A. Spatial autocorrelation and neihborhood quality. *Regional Science and Urban Economics*, v. 22, n. 3, p. 433–452, 1992. Disponível em: . Acesso em: abr. 2021.

DUCHON, J. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: *Constructive Theory of Functions of Several Variables*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1977. p. 85–100. Disponível em: . Acesso em: mai. 2021.

DUNN, P. K.; SMITH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, v. 5, p. 236–244, 1996. Disponível em: . Acesso em: abr. 2021.

EFRON, B.; TIBSHIRANI, R. J. *An Introduction to Bootstrap*. Berlin, DE: Springer Science Business Media, 1993. 452 p.

EILERS, P. H.; MARX, B. D. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, v. 66, n. 2, p. 159–174, 2003. Disponível em: . Acesso em: nov. 2021.

EILERS, P. H. C.; MARX, B. D. Flexible smoothing with b-splines and penalties. *Statistical Science*, v. 11, n. 2, p. 89–121, 1996. Disponível em: . Acesso em: mai. 2021.

EILERS, P. H. C.; MARX, B. D.; REGUERA, M. L. D. Twenty years of p-splines. *SORT-Statistics and Operations Research Transactions*, Catalunha, SP, v. 39, n. 2, p. 149–186, 2015. Disponível em:. Acesso em: jun. 2021.

EVANS, J. S.; MURPHY, M. A.; RAM, K. *spatialEco: Spatial Analysis and Modelling Utilities*. [S.l.], 2021. Versão 1.3-7, disponível em: .

FLORÊNCIO, L. de A. *Estruturação de um sistema de rating para a classificação de risco de vinculação de empreendimentos de base imobiliária em garantia de operações de crédito de longo prazo*. Tese (Doutorado em Ciências) — Universidade de São Paulo, São Paulo, SP, 2018. Disponível em: . Acesso em: abr. 2021.

FLORÊNCIO, L. de A.; CRIBARI-NETO, F.; OSPINA, R. Real estate appraisal of land lots using gamlss models. *Chilean Journal of Statistics*, v. 3, p. 75–91, 2012. Disponível em: . Acesso em: fev. 2022.

GALTON, F.; DICKSON, J. D. H. Family likeness in stature. *The Royal Society*, v. 40, p. 42–73, 1886. Disponível em: . Acesso em: abr. 2021.

GALTON, F.; DICKSON, J. D. H. Co-relations and their measurement, chiefly from anthropometric data. *The Royal Society*, v. 45, p. 135–145, 1889. Disponível em: . Acesso em: abr. 2021.

GAUSS, C. F. *Theoria Motus Corporum Coelestium*. Boston, MA: Little, Brown and Company, 1857. 508 p. Disponível em: . Acesso em: abr. 2021.

GOLUB, G. H.; HEATH, M.; WAHBA, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, v. 21, p. 215–223, 1979. Disponível em: . Acesso em: jun. 2021.

GOMES, M. S.; MONTEIRO, M. M. Estudo da contribuição marginal de variáveis relevantes na formação de preços imobiliários no brasil. In: *Anais da 9ª Conferência Internacional da Latin America Real Estate Society*. São Paulo, SP: Latin America Real Estate Society, 2009. p. 206–257. Disponível em: . Acesso em: ago. 2014.

GOODMAN, A. C. Hedonic prices, price indices and housing markets. *Journal of urban economics*, v. 5, n. 4, p. 471–484, 1978. Disponível em: . Acesso em: abr. 2021.

GRANELLE, J. *Espace urbain et prix du sol*. Paris, FR: SIREY Recherches économiques et financières, 1970. 296 p.

GREEN, P. J.; SILVERMAN, B. W. *Nonparametric Regression and Generalized Linear Models a Roughness Penalty Approach*. Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science, 1994. 194 p.

GUJARATI, D. N.; PORTER, D. C. *Econometria Básica*. São Paulo, SP: AMGH Editora, 2011. Disponível em: . Acesso em: abr. 2021.

HAMILTON, W. *Discussions on Philosophy and Literature, Education and University Reform*. Edimburgh, SC: Longman, Brown, Green and Longmans, 1853. 875 p. Disponível em: . Acesso em: abr. 2021.

HANDSCOMB, D. C. *Methods of Numerical Approximation*: Lectures delivered at a summer school held at oxford university, 1965. London, EN: Pergamon Press, 1966.

HASTIE, T. J.; TIBSHIRANI, R. J. Generalized additive models. *Statistical Science*, v. 1, n. 3, p. 297–318, 1986. Disponível em: Acesso em: abr. 2021.

HASTIE, T. J.; TIBSHIRANI, R. J.; FRIEDMAN, J. *The Elements of Statistical Learning*: Data mining, inference and prediction. 2nd. ed. New York, NY: Springer, 2009. Disponível em: . Acesso em: abr. 2021.

HENGL, T. et al. *plotKML: Visualization of Spatial and Spatio-Temporal Objects in Google Earth*. [S.l.], 2021. Versão 0.8-2, disponível em: .

HIJMANS, R. J. et al. *raster: Geographic Data Analysis and Modeling*. [S.l.], 2022. Versão 3.5-15, disponível em: .

HUBER, P. J. The behavior of maximum likelihood estimates under nonstandard conditions. In: *Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California, 1967. v. 1, p. 221–233. Disponível em: . Acesso em: mai. 2021.

HURD, R. M. *Principles of city land values*. New York, NY: The Record and Guide, 1924. 176 p. Disponível em: . (Acesso em: abr. 2021.

JR, P. J. R. et al. *geoR: Analysis of Geostatistical Data*. [S.l.], 2020. Versão 1.8-1, disponível em: .

KILIBARDA, M.; SEKULIC, A. *plotGoogleMaps: Plot Spatial or Spatio-Temporal Data Over Google Maps*. [S.l.], 2015. Versão 2.6.21, disponível em: .

KING, A. T.; MIESZKOWSKI, P. Racial discrimination, segregation, and the price of housing. *Journal of Political Economy*, v. 81, p. 590–606, 1973. Disponível em: . Acesso em: set. 2021.

KLEIN, N. et al. Bayesian structured additive distributional regression with an application to regional income inequaty in germany. *The Annals of Applied Statistics*, v. 9, p. 1024–1052, 2015. Disponível em: . Acesso em: jul. 2021.

LINDGREN, F. et al. *inlabru: Bayesian Latent Gaussian Modelling using INLA and Extensions*. [S.l.], 2022. Versão 2.5.2, disponível em: .

LINDSEY, J. K. *Applying Generalized Linear Models*. New York, NY: Springer, 1997. 265 p. Disponível em: . Acesso em: abr. 2021.

MAIA NETO, Francisco. *Introdução à engenharia de avaliações e perícias judiciais*. Belo Horizonte, MG: Editora Del Rey, 1992. 166 p.

MICHAEL, R. *Avaliação em Massa de Imóveis com uso de Inferência Estatística e Análise Espacial de Superfícies de Tendência*. Dissertação (Mestrado em Engenharia Civil) — Universidade Federal de Santa Catarina, Florianópolis, SC, 2004. Disponível em: . Acesso em: abr. 2021.

MILLS, E. S. Urban density functions. *Journal of the Royal Statistical Society*, v. 7, n. 1, p. 2–20, 1970. Disponível em: . Acesso em: abr. 2021.

MUTH, R. F. The spatial structure of the house market. *Papers in Regional Science*, v. 7, n. 1, p. 207–220, 1961. Disponível em: . Acesso em: abr. 2021.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika*, v. 78, p. 691–692, 1991. Disponível em: . Acesso em: set. 2021.

NELDER, J. A.; MCCULLAGH, P. *Generalized Linear Models*. 2nd. ed. Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science, 1989. 532 p.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society*, v. 135, n. 3, p. 370–384, 1972. Disponível em: . Acesso em: abr. 2021.

ORCID, H. W. et al. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. [S.l.], 2021. Versão 3.3.5, disponível em: .

ORLOFF, J.; BLOOM, J. *Bootstrap confidence intervals*. Cambridge, MA: Massachusets Institute os Technology, 2014. Disponível em: . Acesso em: set. 2021.

PAGOURTZI, E. et al. Real estate appraisal: A review of valuation methods. *Journal of Property Investment and Finance*, Emerald Group Publishing Limited, v. 21, n. 4, p. 383–401, 2003. Disponível em: . Acesso em: jan. 2022.

PAIXÃO, L. A. R. Índice de preços hedônicos para imóveis: uma análise para o município de Belo Horizonte. *Economia Aplicada*, v. 19, n. 1, p. 5–29, 2015. Disponível em: . Acesso em: nov. 2021.

PEARSON, K. *The Grammar of Science*. 1900. ed. London, EN: Adam and Charles Black, 1892. 574 p. Disponível em: . Acesso em: abr. 2021.

PEARSON, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, v. 186, p. 343–414, 1895. Disponível em: . Acesso em: out 2021.

PEARSON, K. Notes on the history of correlation. *Biometrika*, v. 13, n. 1, p. 25–45, 1920. Disponível em: . Acesso em: mai. 2021.

PEBESMA, E. et al. *sf: Simple Features for R*. [S.l.], 2022. Versão 1.0-7, disponível em: .

PEBESMA, E. et al. *sp: Classes and Methods for Spatial Data*. [S.l.], 2021. Versão 1.4-6, disponível em: .

PEBESMA, E. et al. *spacetime: Classes and Methods for Spatio-Temporal Data*. [S.l.], 2022. Versão 1.2-6, disponível em: .

PEREIRA, R. H. M.; GONÇALVES, C. N. *geobr: Download Official Spatial Data Sets of Brazil*. [S.l.], 2022. Versão 1.6.5, disponível em: .

PRICE, M. J. *Penalized b-splines and their application with an in depth look at the bivariate tensor product penalized b-spline*. Tese (Doctor of Philosophy) — Iowa State University, Ames, IA, 2018. Disponível em: . Acesso em: abr. 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: .

RACINE, J. S. *Appendix E: A Primer on Regression Splines*. Oxford, EN: Oxford University Press, 2019. Disponível em: . Acesso em: nov. 2021.

REGUERA, M. L. D. *Modelos Aditivos Generalizados con Psplines*. Madrid, SP: Universidade Carlos III, 2021. Disponível em: . Acesso em: mai. 2021.

REINSCH, C. H. Smoothing by spline functions. *Numerische Mathematik*, v. 10, p. 177–183, 1967. Disponível em: . Acesso em: mai. 2021.

RICHARDSON, H. W. *Economía del Urbanismo*. Madrid, ES: Alianza Editorial, 1975. 211 p.

RIDKER, R. G.; HENNING, J. A. The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics*, v. 49, p. 246–257, 1967. Disponível em: . Acesso em: set. 2021.

RIGBY, R. A.; STASINOPOULOS, D. M. Mean and dispersion additive models. In: *Statistical Theory and Computacional Aspects of Smoothing*. Semmering, AU: Springer, 1996. p. 213–230. Disponível em: . Acesso em: dez. 2021.

RIGBY, R. A.; STASINOPOULOS, D. M. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, v. 6, p. 57–65, 1996. Disponível em: . Acesso em: dez. 2021.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistics Society*, v. 54, n. 3, p. 507–554, 2005. Disponível em:. Acesso em: jun. 2021.

RIGBY, R. A.; STASINOPOULOS, D. M. Using the box-cox t distribution in gamlss to moel skewness and kurtosis. *Statistical Modelling*, v. 6, p. 209–229, 2006. Disponível em: . Acesso em: jun. 2021.

RIGBY, R. A. et al. *Distributions for Modeling Location, Scale and Shape using GAMLSS in R*. London, EN: Chapman & Hall/CRC, 2020. 588 p.

RIPLEY, B.; VENABLES, W. *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. [S.l.], 2022. Versão 7.3-17, disponível em: . Acesso em: abr. 2021.

ROYSTON, P.; WRIGHT, E. M. Goodness-of-fit statistics for age-specific reference intervals. *Statistics in Medicine*, v. 19, p. 2943–2962, 2000. Disponível em: . Acesso em: jul. 2021.

SCHOENBERG, I. J. Contributions to the problem of approximation of equidistant data bu analytic functions: Part a: On the problem of smoothing or graduation. a fisrt class of analytic formulae. *Quarterly of Applied Mathematics*, v. 4, n. 1, p. 45–99, 1945. Disponível em: . Acesso em: mai. 2021.

STASINOPOULOS, D. M. et al. *Flexible Regression and Smoothing Using GAMLSS in R*. Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science, 2017. 549 p.

STASINOPOULOS, M. et al. *gamlss.add: Extra Additive Terms for Generalized Additive Models for Location Scale and Shape*. [S.l.], 2020. Versão 5.1-6, disponível em: . Acesso em: abr. 2021.

STASINOPOULOS, M. et al. *gamlss: Generalised Additive Models for Location Scale and Shape*. [S.l.], 2022. Versão 5.4-1, disponível em: .

TENNEKES, M. et al. *tmap: Thematic Maps*. [S.l.], 2022. Versão 3.3-3, disponível em: .

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. *rpart: Recursive Partitioning and Regression Trees*. [S.l.], 2022. Versão 4.1.16, disponível em: .

THOMAS, G. *GAMLSS with applications to zero inflated and hierarchical data*. Dissertação (Mestrado em Estatística e Experimentação Agronômica) — Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, SP, 2017. Disponível em: . Acesso em: jul. 2021.

THÜNEN, J. H. von. *Der Isoliert Staat*. Hamburg, GE: Wirtschaft & Finan, 1826. 319 p. Disponível em: . Acesso em: nov. 2021.

TRIVELLONI, C. A. P. *Método para a determinação do valor da localização com o uso de técnicas inferenciais e geoestatísticas na avaliação em massa de imóveis*. Tese (Doutorado em Engenharia Civil) — Universidade Federal de Santa Catarina, Florianópolis, SC, 2005. Disponível em: . Acesso em: jul. 2014.

TURKMAN, M. A. da C. A.; SILVA, G. L. Modelos lineares generalizados - da teoria à prática. In: *VIII Congresso Anual da Sociedade Portuguesa de Estatística*. Peniche, PT: SPE - Sociedade Portuguesa de Estatística, 2000. Disponível em: . Acesso em: abr. 2021.

VAN BUUREN, S.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference tools. *Statistics in Medicine*, v. 20, p. 1259–1277, 2001. Disponível em: . Acesso em: abr. 2021.

VANEGAS, L. H.; PAULA, G. A. An extension of log-symmetric regression models: R codes and applications. *Journal of Statistical Computation and Simulation*, v. 86, p. 1709–1735, 2015. Disponível em: . Acesso em: mai. 2021.

VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, v. 57, p. 307–333, 1989. Disponível em: . Acesso em: jul. 2021.

WAN, C.; ZHONG, W. *MultiKink: Estimation and Inference for Multi-Kink Quantile Regression*. [S.l.], 2020. Versão 0.1.0, disponível em: .

WHITE, H. A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, v. 48, n. 4, p. 817–838, 1980. Disponível em: . Acesso em: abr. 2021.

WINKLE, C. K.; ZAMMIT-MANGION, A.; CRESSIE, N. *Spatial-temporal Statistics with R*. Boca Raton, FL: CRC Press Taylor &Francis Group, 2019. 396 p.

WOLVERTON, M. L. Empirical study of the relationship between residential lot price, size and view. *Journal of Property Valuation and Investment*, Emerald Group Publishing Limited, v. 15, n. 1, p. 48–57, 1997. Disponível em: . Acesso em: jan. 2022.

WOOD, S. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. [S.l.], 2022. Versão 1.8-40, disponível em: .

WOOD, S. N. Thin plate regression splines. *Royal Statistical Society*, v. 65, n. 1, p. 95–114, 2003. Disponível em: . Acesso em: abr. 2021.

WOOD, S. N. Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistics Society*, v. 70, n. 3, p. 495–518, 2008. Disponível em: . Acesso em: jun. 2021.

WOOD, S. N. P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing*, v. 27, p. 985–989, 2016. Disponível em: . Acesso em: abr. 2021.

WOOD, S. N. *Generalized Additive Models: an introduction with R*: Second edition. Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science, 2017. 476 p.

YU, Y.; RUPPERT, D. Penalized spline estimation for partially linear single-index models. *American Statistical Association*, v. 97, p. 1042–1054, 2002. Disponível em: . Acesso em: mai. 2021.

ZYGA, J. Evaluation of usefulness of real estate data contained in the register of prices and values of real estates. III, 03 2017. Disponível em: . Acesso em: jan. 2022.