ENGENHARIA DE AVALIAÇÕES COM BASE EM MODELOS GAMLSS

LUTEMBERG DE ARAÚJO FLORENCIO

Orientador: Prof. Dr. Francisco Cribari Neto
Co-orientador: Prof. Dr. Raydonal Ospina Martínez

Área de Concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, fevereiro de 2010.

Florencio, Lutemberg de Araújo
Engenharia de avaliações com base em modelos GAMLSS / Lutemberg de Araújo Florencio. - Recife: O Autor, 2010.
xv, 125 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Estatística, 2010.

Inclui bibliografia e apêndice.

1. Estatística aplicada. 2. Estatística aplicada à Engenharia. I. Título.

310 CDD (22. ed) MEI2010 – 033

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

25 de fevereiro de 2010
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Lutemberg de Araújo Florêncio

intitulada

"Engenharia de Avaliações com Base em Modelos GAMLSS"

seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

Coordenador da Pós-Graduação em Estatística

Prof. Francisco Cribari Neto
Coodenador
Pós Graduação em Estatística UFPE

Banca Examinadora:

Raydonal Ospina co-orientador

Mário de Castro Andrade Filho (USP/São Carlos)

Rubens Alves Dantas (UFPE/Economia)

Este documento será anexado à versão final da dissertação.

Dedico este trabalho a meu pai, Laércio, minha mãe, Ester, meu irmão, Laerdes, minha irmã, Jacqueline e minha noiva, Madja.

Agradecimentos

A conclusão de um curso de mestrado em Estatística na Universidade Federal de Pernambuco (UFPE), um dos centros mais avançados em estudos e pesquisas do país, representa para mim muito mais do que uma etapa da vida acadêmica, significa um sonho concretizado e a superação de muitos obstáculos na busca do crescimento pessoal e profissional, sobretudo, um presente divino. Assim, não poderia deixar de agradecer, em primeiro lugar, a Deus por mais uma vez mostrar Sua fidelidade em minha vida e iluminar o meu caminho em todas as minhas escolhas, decisões e projetos. Sinto-me verdadeiramente abençoado por Deus ter-me dado a oportunidade e o prazer de cursar o Mestrado em Estatística na UFPE.

Aos meus pais, Laércio e Ester, as palavras não são suficientes para relatar o profundo sentimento de reconhecimento que trago em meu peito. De meu pai, aprendi que "primeiro vem as obrigações e depois, o lazer" e com minha mãe compreendi que "tudo posso naquele que me fortalece". Vocês são os responsáveis por este momento "primo" que estou vivendo. Muito obrigado pelo apoio incessante e por este amor incondicional!

Aos meus irmãos, agradeço por todo carinho e confiança que sempre depositaram em mim. Não poderia deixar de manifestar gratidão a Laerdes, pelo grande presente (um *notebook*) que chegou na hora mais oportuna desta caminhada, e a Jacqueline pelo encorajamento e pelas silenciosas, porém fortes orações, para o sucesso desta jornada. Aos respectivos cunhados, Edilma e Sérgio, externo minha gratidão pelo contínuo incentivo.

Agradeço a minha noiva, Madja, que além de revisora ortográfica desta dissertação (por livre e espontânea pressão), teve que, por tantas vezes, sacrificar as suas férias e finais de semana. Agradeço-lhe ainda, pela compreensão, força e amor. Estou certo que esta é a mulher da minha vida, afinal "suportar" um mestrando em Estatística é muito mais que uma prova de amor.

Aos professores Francisco Cribari Neto (orientador) e Raydonal Ospina Martínez (co-orientador), agradeço pela orientação firme e incondicional, amizade, paciência (e foi muita) e confiança transmitida no desenvolvimento deste trabalho. O professor Cribari, além de prolífico na comunidade científica, é de exemplar conduta acadêmica e referência de docência na Estatística. Tê-lo como orientador foi um privilégio, afinal de contas o seu desprendimento e sua notável atenção perante as minhas dúvidas — mesmo diante

de tantas outras obrigações e responsabilidades que lhe cabiam —, levam-me a crer que o professor Cribari transcendeu (e muito) as expectativas do que eu esperava de um "orientador" e principalmente, fez-me ter certeza que seu dia não dura apenas 24h. Com o professor Raydonal, pesquisador de notável destaque em modelos de regressão beta e modelos GAMLSS, pude tirar lições de disciplina, ética, independência intelectual, dedicação, entusiasmo e, principalmente, ele fez-me perceber que cientistas não precisam usar linguagem erudita, fumar cachimbo e fazer-se de esquecidos e desligados para serem geniais. Comprometido e sempre disponível, o professor Raydonal mostrou-me o "caminho das pedras" na aplicação dos modelos GAMLSS e também foi um brilhante tutor diante de minhas dúvidas com o ambiente `R`. Gostaria de dividir o mérito desta dissertação com os referidos docentes.

Agradeço também ao professor Rubens Alves Dantas por ter sido um grande incentivador para que eu fizesse este mestrado. O professor Dantas me "apresentou" à Ciência quando lecionou a disciplina de Engenharia de Avaliações e foi o orientador do meu projeto final de conclusão do curso de Engenharia Civil pela Escola Politécnica da Universidade de Pernambuco (POLI-UPE), intitulado **Avaliação de imóveis urbanos: a Engenharia Civil a serviço de uma instituição bancária**, em meados de 2005. Naquela ocasião, o meu interesse pela pesquisa científica emergiu e rapidamente passei a utilizar a Estatística (ainda que de forma elementar) para construção de modelos de avaliação de bens. Agradeço-lhe também por ter me disponibilizado sua literatura e nunca ter poupado seu tempo e paciência para minhas consultas.

Aos colegas do programa de pós-graduação em Estatística, agradeço pelo clima cordial que sempre reinou em nosso convívio social e intelectual, com destaque para os amigos Manoel Neto, Jeremias (Barry Jeremy), Josimar (Jocquinha), Marcelo (Borel), Elton (Scheffé) e Priscila que por muitas vezes dividiram comigo as suas bancas de estudo. Não poderia deixar de agradecer aos demais colegas da pós-graduação (turmas anteriores/posteriores do mestrado e doutorado em Estatística) que de alguma forma contribuíram para o desfecho meritório deste mestrado: Wilton, Olga, Izabel, Tarciana, Tatiene, Fábio Bayer, Laércio, Diego, Silvio, Ivan, Francisco, Marcela e Natasha.

Agradeço a Valéria Bittencourt, secretária da pós-graduação em Estatística, pela competência, carinho e atenção com os alunos do mestrado. Hoje eu entendo o porquê

da frase que ouvi quando iniciei o curso: "qualquer problema, dúvida, desconforto, dificuldade, ansiedade etc., fale com Valéria". De fato, Valéria é sinônimo de presteza e eficiência.

Quero também registrar meus agradecimentos a Leandro Rêgo e Sylvio Santos, meus professores, pelos seus valiosos ensinamentos, que foram muito úteis na elaboração desta dissertação. Em especial, agradeço aos professores Cristiano Ferraz e Audrey Cysneiros por terem confiado em mim desde a minha primeira "visita" ao programa de pós-graduação em Estatística da UFPE, em 2006, e por terem sido verdadeiros "conselheiros" nesta empreitada. Sem dúvida, chegar até aqui sem o apoio de vocês seria impensável.

Agradeço a Suenize Souza, gerente geral da Central de Apoio Operacional de Recife do Banco do Nordeste do Brasil S.A (BNB) — empresa em que trabalho —, por todo o apoio e confiança depositados. Em seu nome, agradeço ao Banco do Nordeste por me conceder uma licença de trabalho por cerca de dois anos para a realização deste mestrado e por permitir a capacitação técnica de sua força de trabalho. Aos colegas do BNB, manifesto minha gratidão pelas palavras de incentivo durante a caminhada. Entre eles, gostaria de destacar aqueles que fazem parte do Departamento de Engenharia em Recife – Ana Emília, Bernardo Vinhas, Efren Girão, Leila Maria e Petronio Rocha – por terem se "desdobrado" ao longo deste período em que estive ausente.

Registro meus agradecimentos a Prefeitura da Cidade de Aracaju, na pessoa do engenheiro civil e diretor do Departamento de Cadastro Imobiliário da Prefeitura de Aracaju, João Freire Prado, pela cessão de dados valiosos relativos a terrenos ofertados e transacionados naquela cidade, sem os quais não seria possível a realização deste trabalho.

A meus amigos que, de uma forma ou de outra, contribuíram com sua amizade e sugestões efetivas para a realização deste trabalho, gostaria de expressar minha profunda gratidão.

Aos participantes da banca examinadora, professores Rubens Alves Dantas e Mário de Castro, agradeço antecipadamente pelos comentários e sugestões.

Agradeço a existência da "dupla" LaTeX e R.

Finalmente, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro a este projeto.

"Buscai antes o reino de Deus, e todas estas coisas vos serão acrescentadas." (Lucas 12:31).

Resumo

A determinação técnica do valor de um bem imóvel (casas, terrenos, entre outros) é de extrema importância para a tomada de decisão em diversos segmentos da sociedade e em muitos órgãos governamentais e privados. Cabe à Engenharia de Avaliações, enquanto ciência do valor, coletar, tratar e analisar dados e estimar modelos que expliquem, de maneira satisfatória, a variabilidade observada nos preços, no mercado em que se estuda. Entretanto, não-normalidade, heteroscedasticidade e heterogeneidade espacial e estrutural são bastante comuns em dados imobiliários, razão pela qual o uso de modelos tradicionais, como o modelo normal de regressão linear clássico (CNLRM) e os modelos lineares generalizados (GLM), pode sofrer limitações. Diante disto e com base numa amostra de 2109 observações de terrenos urbanos situados na cidade de Aracaju-SE, relativas aos anos de 2005, 2006 e 2007, estimamos a função de preços hedônicos mediante uso da classe de modelos de regressão proposta por Rigby & Stasinopoulos (2005), denominada de modelos aditivos generalizados para posição, escala e forma (GAMLSS), a qual permite o ajuste de uma ampla família de distribuições para a variável resposta e possibilita a modelagem direta, utilizando funções paramétricas e/ou não-paramétricas, da estrutura de regressão da variável de interesse. Neste sentido, a presente dissertação descreve e caracteriza os modelos GAMLSS, bem como compara os ajustes realizados entre os modelos estimados via CNLRM, GLM e GAMLSS para o mesmo conjunto de dados. Na análise empírica consideramos como variável resposta o preço unitário do terreno e como variáveis independentes as características estruturais, locacionais e econômicas inerentes ao imóvel. Devido à flexibilidade da estrutura de regressão GAMLSS, modelamos de forma não-paramétrica (utilizando suavizadores *splines*) algumas covariáveis (por exemplo, as coordenadas geográficas referentes à localização do terreno), assim como modelamos os parâmetros de posição (μ) e escala (σ) da variável resposta. Os resultados obtidos mostraram que os modelos GAMLSS forneceram um ajuste superior àqueles obtidos via CNLRM e GLM, segundo as análises gráficas e numéricas dos resíduos e os critérios de Akaike e Schwarz, indicando que a classe de modelos GAMLSS aparenta ser mais apropriada para a estimação dos parâmetros da função de preços hedônicos.

Palavras-chave: Engenharia de Avaliações, função de preços hedônicos, modelos de regressão semiparamétricos, suavizadores não-paramétricos, *splines* cúbicos.

Abstract

The technical determination of the value of real estate (houses, lands, among others) is extremely important for decision making in several professional segments and in many government agencies and private companies. It is a duty to the Engineering Appraisal – while science of value – to collect, analyze and treat data and to estimate models which explain, in a satisfactory way, the variability observed in prices, in a given market of study. Nevertheless, non-normality, heteroskedasticity, and spatial and structural heterogeneity are quite common in real estate data, and that is why the use of traditional models, such as the classical normal linear regression model (CNLRM) and the generalized linear models (GLM), might face limitations. In this context and based on a sample of 2109 observations of urban lands located in the city of Aracaju, Sergipe-Brazil, relative to the years 2005, 2006, and 2007, we estimate a hedonic price function through a class of regression models proposed by Rigby & Stasinopoulos (2005), called generalized additive models for location, scale and shape (GAMLSS), which allows the fit of a broad family of distributions for the response variable and the direct modeling, using either parametric and nonparametric functions, of the regression structure for the variable of interest. From this perspective, the present work describes and characterizes the GAMLSS model, and compares estimated models via CNLRM, GLM, and GAMLSS for the same data set. In the empirical analysis, we considered as the response variable the unit price of the land, and as explanatory variables the structural, locational, and economic characteristics inherent to the real estate. Due to flexibility of the GAMLSS regression framework, we model, in a nonparametric fashion (using smoothing splines) some covariates (for instance, the geographic coordinates concerning the location of the land), as well as the positional (μ) and scale (σ) parameters. The results obtained show that GAMLSS models provided a superior fit when we compared with CNLRM and GLM, according to graphical and numerical analysis of the residuals and the Akaike and Schwarz criteria, thus indicating that the GAMLSS class of models appears to be more appropriate for estimating the hedonic price function than the traditional models (CNLRM and GLM).

Keywords: hedonic price models, engineering appraisal, semiparametric regression models, nonparametric smoothing, cubic splines.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

"De maneira geral, o estatístico deve ser um profissional que, baseado em conhecimentos sólidos e atualizados, seja capaz de abordar com proficiência os problemas usuais de sua área de atuação: coleta, organização e síntese de dados, ajuste de modelos – e ter a capacidade de buscar informação para a solução de problemas novos e, encontrando-as, ser capaz de entendê-las e implementá-las. Além disto, em todas as suas atividades deve estar presente, a curiosidade pelo conhecimento novo e uma postura ética diante dos fatos."

(Diretrizes curriculares para cursos de Estatística
do Ministério da Educação e do Desporto, 1999)

1.1 Preliminares

Há muitos anos não se via no Brasil um período tão promissor para o mercado imobiliário. Antes considerados como péssimas alternativas de aplicação financeira, hoje terrenos, casas, apartamentos e conjuntos comerciais estão, ao lado da Bolsa de Valores, entre as maiores promessas de lucro a longo prazo. Mesmo com a crise financeira global, o mercado imobiliário brasileiro vem batendo recordes de investimento e apresenta-se como um dos setores mais produtivos da economia na geração de negócios, emprego e renda, sendo imprescindível para o crescimento e desenvolvimento do país.[1]

[1]Para mais detalhes sobre o impacto da construção civil e do mercado imobiliário na economia brasileira, vide os sites `http://www.sindusconsp.com.br/` e `http://www.caixa.gov.br/`, respectivamente.

Somente no primeiro semestre de 2009, quase 13 (treze) bilhões de reais em recursos da poupança foram destinados ao financiamento da casa própria e mais de 2 (dois) bilhões de dólares circularam em transações imobiliárias, o que situa o Brasil à frente de outros países da região, como Argentina, México e Chile, segundo informação da Associação das Entidades de Crédito Imobiliário e Poupança (Abecip).

O imóvel constitui, além de um bem de consumo que proporciona conforto e *status* social, um esteio econômico da sociedade capitalista moderna, tornando-se uma forma de reserva e apropriação de capitais, baseadas nas expectativas de valorização, e um meio de ganhos financeiros, por meio de suas rendas locatícias e de vendas.

Em decorrência disto, o valor de mercado do bem imóvel, enquanto produto negociável em função de sua capacidade de aproveitamento e utilização, tornou-se um parâmetro de extrema importância para a tomada de decisão em diversos segmentos da sociedade e em muitos órgãos governamentais ou privados: prefeituras (cobrança do Imposto Predial e Territorial Urbano (IPTU)[2] e do Imposto sobre Transmissão de Bens Imóveis (ITBI),[3] desapropriações e elaboração de plantas de valores genéricos);[4] Serviço de Patrimônio da União (cobrança de laudêmio, foro); Receita Federal (auxílio na determinação da base de cálculo de impostos que envolvam ganhos de capital, identificação de transações que possam prenunciar lavagem de dinheiro); ao Instituto Nacional de Colonização e Reforma Agrária (desapropriações rurais para reforma agrária); Poder Judiciário (avaliações para subsidiar decisões judiciais); agentes financeiros (garantia para financiamento, limite de operações de crédito, leilões) e empresas privadas (operações de compra e venda, análise de viabilidade de empreendimentos), entre outros. Esta demanda gerou a necessidade de se avaliar os bens a partir de análises criteriosas, envolvendo elementos

[2]O IPTU é um imposto cuja incidência se dá sobre a propriedade urbana. Ou seja, o IPTU tem como fato gerador a propriedade, o domínio útil ou a posse de propriedade imóvel localizada em zona urbana ou extensão urbana. A função do IPTU é tipicamente fiscal e sua finalidade principal é a obtenção de recursos financeiros para os municípios, embora ele também possa ser utilizado como instrumento urbanístico de controle do valor da terra.

[3]O ITBI é cobrado pelo município nos casos de transferência – transmissão ou cessão – de propriedade de imóveis como casas, terrenos, apartamentos, salas, lojas e galpões. O pagamento do tributo é condição para o registro em cartório da transferência do imóvel. A função do ITBI é predominantemente fiscal e sua finalidade é a obtenção de recursos financeiros para os municípios.

[4]Também denominada simplesmente de planta de valores, é parte integrante e básica do sistema de informações do cadastro municipal e juntamente com o cadastro imobiliário forma a base de cálculo tanto do IPTU quanto do ITBI e da contribuição de melhoria. Ela deve apresentar valores médios unitários de terrenos para cada face de quadra do município (Liporoni, 2007).

de natureza técnica e científica.

Para suprir esta necessidade, vieram a público diversos trabalhos técnicos sobre o assunto e foram elaboradas normas específicas voltadas à regulamentação das metodologias e procedimentos balizadores para atuar na área denominada de Engenharia de Avaliações de Bens. No Brasil, a primeira norma de avaliação foi editada em 1952 e por muitos anos os profissionais que atuam neste segmento basearam-se exclusivamente em fórmulas determinísticas e ponderações arbitrárias.

Embora existam registros do uso da inferência estatística em trabalhos avaliatórios realizados no Brasil na década de 1950, foi somente em 1974, com o engenheiro Domingos de Saboya Barbosa Filho, que esta ferramenta ganhou grande impulso e possibilitou avanços significativos na área da Engenharia de Avaliações (vide Saboya, 1974). Contudo, foi apenas na década de 1990 que a Engenharia de Avaliações teve o seu maior desenvolvimento, pela consolidação da pesquisa científica[5] como metodologia indispensável ao engenheiro de avaliações.[6]

Entretanto, conforme ficou evidenciado, a Engenharia de Avaliações requer conhecimentos relacionados não apenas à própria Engenharia, mas sobretudo atinentes à Econometria, Teoria das Probabilidades, Amostragem, Álgebra Linear, Matemática Aplicada, Matemática Financeira, Teoria das Decisões, Estatística Aplicada, entre outras áreas.

Em razão disso, a análise de regressão tem desempenhado um papel fundamental na busca de modelos que expliquem, de maneira satisfatória, a variabilidade observada nos preços dos imóveis, com base na variação dos regressores, no mercado que se estuda. Para tanto, são empregadas técnicas da inferência estatística e da teoria econométrica na formulação de modelos hedônicos[7] que representem o mercado imobiliário e sejam, ao mesmo tempo, parcimoniosos e tenham abrangência suficiente para que os principais

[5] Atividade que utiliza a metodologia e os pressupostos científicos (Volpato, 2007).

[6] Deve ser entendido por "engenheiro de avaliações" não só o próprio engenheiro como também o arquiteto, o engenheiro agrônomo ou outro profissional legalmente habilitado e especializado em avaliações.

[7] A abordagem dos preços hedônicos ou preços implícitos deriva da contribuição oferecida por Lancaster (1971), consolidada por Rosen (1974), segundo a qual uma classe de produtos diferenciados pode ser descrita completamente por um vetor de características objetivamente medidas, sendo que os quantitativos das características associadas a cada produto geram os denominados preços hedônicos, definindo decisões locacionais no consumidor. A palavra "hedônica" é proveniente do termo *hedonismo* (do grego *hēdonē* que significa prazer), já que o prazer ou a felicidade que um consumidor apresenta depende do nível de atributos que o "produto" adquirido possui.

fatores intervenientes estejam claramente identificados.

Na análise hedônica para o mercado imobiliário, o imóvel é tratado como um bem heterogêneo composto de um pacote de características e a estimação da função explícita, denominada função de preço hedônico, determina quais são os atributos, ou "pacote" de atributos, mais significativos na composição do preço, quando da avaliação de um bem em particular. Entretanto, a estimação da equação hedônica não é trivial, visto que a teoria não determina sua forma funcional nem as variáveis relevantes para a sua estimação.

Na literatura nacional, as equações de preços hedônicos voltadas para o mercado imobiliário têm sido, em sua maioria, formuladas com base no modelo normal de regressão linear clássico (*Classical Normal Linear Regression Model* — CNLRM) e adotam uma forma linear, log-linear ou fazem uso da transformação de Box-Cox em relação à variável resposta (vide, por exemplo, Aguirre & Macedo, 1996 e Fávero *et al.*, 2003). Contudo, na maioria das vezes, o pesquisador não toma os cuidados necessários na modelagem em relação aos pressupostos básicos do CNLRM. Sobre isto, Dantas (2003) alerta que a não observância destes pressupostos pode ser um dos fatores causadores das distorções encontradas entre os resultados obtidos e os valores reais de mercado, pois questões como falta de normalidade, heteroscedasticidade e autocorrelação são bastante comuns em dados imobiliários. Acrescenta-se que outros trabalhos, em quantidade incipiente, utilizam os modelos lineares generalizados para estimar o valor venal de imóveis urbanos (vide, por exemplo, Dantas & Cordeiro, 1988, 2001) e empregam técnicas de validação cruzada para justificar a escolha da função de distribuição "ideal" para a construção do modelo de regressão, como apresentado em Barbosa & Bidurin (1991), que recomendam as distribuições gama ou lognormal para o conjunto de dados imobiliários analisado. Cumpre registrar que em todos os casos mencionados os modelos resultantes são obtidos a partir do uso estrito da regressão paramétrica.

Em contrapartida, na literatura internacional é possível observar a estimação de funções hedônicas por meio de modelos não-paramétricos e semiparamétricos, como em Hartog & Bierens (1989), Stock (1991), Pace (1993, 1995, 1998), Anglin & Gencay (1996), Gencay & Yang (1996), Iwata *et al.* (2000) e Clapp *et al.* (2002). Além destes, destacamos o estudo desenvolvido por Bin & Martins-Filho (2003), que utiliza dados do mercado imobiliário de Multnomah County, Oregon-USA, para enfatizar a superioridade dos modelos

não-paramétricos em detrimento das estruturas estritamente paramétricas na estimação do valor de comercialização de casas.

De toda forma, as evidências disponíveis, principalmente na literatura nacional, indicam que muito pouco foi realizado em termos de modelos de preços hedônicos que não fazem uso de métodos tradicionais[8] ou que não restrinjam a modelagem da variável resposta às distribuições da família exponencial, razão pela qual se torna imperativa a busca por técnicas estatísticas que conduzam a modelagens mais flexíveis e ao mesmo tempo expliquem, com o máximo de fidelidade, o comportamento do mercado imobiliário.

Esta crescente complexidade de modelização do mundo real, atrelada aos impressionantes ganhos de velocidade e memória dos computadores, têm exigido dos pesquisadores, de forma intensa, o desenvolvimento de métodos estatísticos sofisticados capazes de descrever com maior grau de adequação as inter-relações entre variáveis. A quantidade de dados coletados e a necessidade de análises estatísticas aumentaram significativamente nos últimos anos, permitindo o ajuste de modelos cada vez mais complexos e realistas.

Neste sentido, Rigby & Stasinopoulos (2005) propuseram uma classe de modelos de regressão denominada de modelos aditivos generalizados para posição, escala e forma, em inglês *Generalized Additive Models for Location, Scale and Shape*, GAMLSS. Trata-se de uma técnica de modelagem estatística univariada que permite o ajuste de uma ampla família de distribuições contínuas e discretas para a variável resposta e possibilita a modelagem explícita, utilizando funções paramétricas e/ou não-paramétricas, de todos os parâmetros da distribuição da variável resposta em relação às variáveis explanatórias. Nos modelos GAMLSS, a distribuição da variável resposta não precisa pertencer à família exponencial e diferentes termos aditivos podem ser incluídos no preditor para cada parâmetro da distribuição, a exemplo de *splines* e efeitos aleatórios, o que confere flexibilidade extra ao modelo.

Pode-se afirmar, assim como enfatizou Dantas (2005), que no atual cenário de avaliações imobiliárias há grande probabilidade dos resultados baseados na metodologia tradicional (via CNLRM) serem viesados, ineficientes ou inconsistentes, por negligenciarem ou conflitarem com os pressupostos básicos do modelo clássico de regressão. Além disso, a restrição imposta na abordagem paramétrica para a forma funcional da relação entre a

[8]Usa geralmente como ferramenta estatística o modelo normal de regressão linear clássico e eventualmente, os modelos lineares generalizados.

variável dependente e as variáveis independentes, associada às suposições adicionais sobre a distribuição de probabilidade para os erros aleatórios, constituem limitadores para utilização desta técnica e podem ocasionar possíveis erros de especificação do modelo.

Dado o exposto, acredita-se que o emprego da estrutura de regressão GAMLSS possa acurar o processo de estimação do valor do imóvel e contribuir para a análise e entendimento de quais, e de que forma e com que intensidade, os atributos influenciam na variação dos preços de mercado dos imóveis.

1.2 Objetivos da dissertação

Esta dissertação pretende atingir dois objetivos: um relacionado a aspectos metodológicos e o outro de natureza empírica. O primeiro consiste em apresentar, descrever e caracterizar a classe de modelos estatísticos univariada denominada GAMLSS, destacando aspectos de inferência e diagnóstico inerentes à análise de regressão. O segundo trata da aplicação e incorporação da estrutura de regressão GAMLSS para estimação da equação de preços hedônicos de terrenos urbanos situados na cidade de Aracaju, capital do Estado de Sergipe (SE). Adicionalmente, os resultados obtidos via GAMLSS serão comparados com os ajustes realizados pela metodologia tradicional.

Essencialmente, o que se busca neste trabalho é melhorar a precisão da estimação da equação de preços hedônicos mediante emprego dos modelos GAMLSS, ainda não difundidos na área de Engenharia de Avaliações de Bens no Brasil.

1.3 Estrutura da dissertação

Esta dissertação encontra-se dividida em 6 (seis) capítulos. No Capítulo 1, enfatizamos a evolução e importância do mercado imobiliário e da Engenharia de Avaliações no contexto nacional, estadual e municipal, bem como evidenciamos as técnicas atualmente utilizadas para previsão do valor de mercado dos bens imobiliários. Além disto, mencionamos as principais dificuldades enfrentadas na estimação das equações de preços hedônicos de imóveis e apontamos os modelos GAMLSS como uma possível alternativa para acurar o processo de estimação e superar algumas limitações presentes nas estruturas de regressão tradicionalmente empregadas no ajuste dos modelos. Adicionalmente, são

expostos os objetivos do trabalho.

No Capítulo 2, discorremos acerca da Engenharia de Avaliações e abordamos os métodos e conceitos mais relevantes que dão suporte ao trabalho avaliatório[9] no Brasil. Em seguida, destacamos a incorporação da pesquisa científica nas avaliações imobiliárias e expomos a atual metodologia de estimação da equação de preços hedônicos de imóveis predominante no país.

No Capítulo 3, visando à compreensão prévia de algumas técnicas e termos básicos empregados no ajuste dos modelos GAMLSS, apresentados no Capítulo 4, revisamos alguns conceitos fundamentais de regressão, como a distinção entre modelos paramétricos, não-paramétricos e semiparamétricos, e também apresentamos os principais procedimentos e técnicas não-paramétricas de suavização utilizados na estimação de modelos que envolvem componentes paramétricos e não-paramétricos, como os métodos *kernel* e *spline*. Adicionalmente, discorremos sobre o processo iterativo de ajuste dos modelos semiparamétricos que combina maximização da verossimilhança e o algoritmo *backfitting*.

No Capítulo 4, apresentamos os modelos GAMLSS e mostramos como incorporar nesta estrutura de regressão as modelagens paramétrica, não-paramétrica e de efeitos aleatórios, entre outras. Além disto, detalhamos o processo de estimação e discutimos aspectos técnicos e práticos, incluindo estratégias de ajuste e diagnóstico para estes modelos.

No Capítulo 5, consideramos uma aplicação com dados reais referentes a 2109 observações de terrenos urbanos situados em Aracaju-SE e que estavam à venda (ofertados) ou foram transacionados (negociados) ou constavam nas declarações de ITBI do cadastro da prefeitura. Visando à estimação da equação de preços hedônicos, comparamos os modelos GAMLSS ajustados às equações de preços hedônicos contra alguns modelos ajustados por métodos tradicionais.

Finalmente, no Capítulo 6 são apresentadas conclusões, comentários e sugestões para futuras pesquisas.

[9]Trabalho avaliatório ou processo avaliatório são terminologias próprias da área de Engenharia de Avaliações para referir-se a todas as etapas que envolvem a elaboração de uma avaliação de bens (vide Seção 2.4).

1.4 Suporte computacional

O emprego da metodologia científica e a investigação de modelos explicativos do mercado imobiliário abrangem diversas etapas de análise, razão pela qual se torna imprescindível o uso de computadores e *softwares* adequados à manipulação de dados e à interpretação dos resultados no trabalho avaliatório. Por este motivo, destacamos que todas as apresentações gráficas e a análise de regressão (estimação de parâmetros, testes de hipóteses, intervalos de confiança, entre outras investigações) realizadas ao longo desta dissertação foram produzidas no ambiente de programação `R`, tendo sido utilizada a versão 2.9.2 para a plataforma Windows. O `R` foi criado por Ross Ihaka e Robert Gentleman, na Universidade de Auckland, com o objetivo de produzir um ambiente de programação parecido com o `S`, uma linguagem desenvolvida no AT & T Bell Laboratories, cuja versão comercial é o `S`-Plus, tendo as vantagens de ser de livre distribuição e de possuir código fonte aberto. `R` é um ambiente integrado que possui grandes facilidades para a manipulação de dados, geração de gráficos e modelagem estatística em geral. A linguagem e seus pacotes podem ser obtidos gratuitamente no endereço `http://www.r-project.org`. Mais detalhes podem ser obtidos em Ihaka e Gentleman (1996), Cribari-Neto & Zarkos (1999) e Venables *et al.* (2009).

A presente dissertação foi digitada com auxílio do sistema tipográfico LaTeX, desenvolvido por Leslie Lamport na década de 1980, que consiste em uma série de macros ou rotinas do sistema TeX, criado por Donald Knuth na Universidade de Stanford, que facilitam o desenvolvimento da edição do texto. Uma implementação LaTeX para a plataforma Windows (MikTeX) encontra-se disponível em `http://www.miktex.org`. Detalhes sobre o sistema de tipografia LaTeX podem ser encontrados em Lamport (1994), Mittelbach *et al.* (2004) e em `http://www.tex.ac.uk/CTAN/latex`.

Por fim, registramos que foi utilizado um computador portátil (*notebook*) Compaq Presario CQ50-222BR (2.0GHz Intel Pentium Dual-Core, 3GB de memória RAM, HD de 250GB, clock de 2.0GHz e sistema operacional Windows Vista Basic) para a elaboração desta dissertação.

CAPÍTULO 2

A Engenharia de Avaliações

"O rigor do laudo de avaliação não está na descrição suntuosa dos detalhes do bem avaliando,[1] *nem tampouco na qualidade do papel apresentado, fotografias etc.; o rigor do laudo de avaliação está na fundamentação do valor estimado e no enfoque científico do trabalho avaliatório."*

(Autor desconhecido)

Este capítulo apresenta uma visão global da Engenharia de Avaliações e introduz, de forma sistemática, alguns dos conceitos e diretrizes que norteiam o processo avaliatório no Brasil. Adicionalmente, abordamos, em linhas gerais, a base teórica da estimação empírica da equação de preços hedônicos de imóveis via modelo normal de regressão linear clássico, por ser esta a técnica atualmente predominante nos trabalhos de avaliações imobiliárias em todo o país.

Destacamos que a tipologia do imóvel ao qual evidenciaremos ao longo deste capítulo é o **terreno**, visto que os dados que dão suporte à aplicação realizada no Capítulo 5 são desta natureza.

[1] Terminologia própria da área de Engenharia de Avaliações para se referir ao imóvel que está sendo avaliado.

2.1 Introdução

Desde os primórdios de sua existência o homem mantém uma estreita relação com a terra, pois é por meio dela que adquire seu sustento, produz seus alimentos, cria seus animais, entre outras atividades. No período Neolítico,[2] a humanidade atingiu um notável grau de desenvolvimento, sendo o início da produção agrícola e pecuária um marco na transformação da organização social e econômica dos povos. Cultivando a terra e criando animais, o homem conseguiu diminuir sua dependência em relação à natureza. Com estes avanços, tornou-se possível a sedentarização, que conduziu naturalmente aos primeiros aldeamentos localizados, sobretudo, na proximidade de rios. Nesta etapa da evolução humana, é possível presumir que a necessidade de habitação fixa, atrelada à passagem da economia de sobrevivência para uma economia de produção, originou a demanda para o mercado imobiliário.

Abstraindo um pouco dos relatos históricos, é possível imaginar que, provavelmente, nossos ancestrais estariam dispostos a pagar com alimentos, animais e outras mercadorias aos que se dispusessem a construir suas casas. Já neste momento, fez-se necessária uma primeira avaliação do que seria viável pagar em troca da construção de um abrigo. Surgia o primeiro estudo de avaliação de bens.

A Engenharia de Avaliações é definida em Dantas (2005) como uma especialidade da Engenharia que reúne um conjunto amplo de conhecimentos da área de Engenharia e Arquitetura, bem como de outras áreas das ciências sociais, exatas e da natureza, com o objetivo de determinar tecnicamente o valor de um bem, de seus direitos, frutos[3] e custos de reprodução. Surgiu no Brasil no final da década de 1910, sendo consequência da promulgação da Lei n°. 601, de 1850 (Lei das Terras), que criou a figura da propriedade particular sobre a terra, extinguindo o Sistema de Concessões instituído pelo Estado português desde 1375. Surgiram assim os primeiros proprietários imobiliários, os quais passaram a registrar seus imóveis nos Assentamentos Paroquiais.

A partir deste momento, as avaliações imobiliárias tornaram-se fundamentais para o bom equilíbrio social, político e jurídico das relações humanas. Considerando-se que

[2]Período pré-histórico, também conhecido como Idade da Pedra Polida, que corresponde à época da evolução humana situada entre 10 mil A.C. até 4 mil A.C..

[3]Resultado da exploração econômica de um bem.

o imóvel, em geral, é o bem de maior importância adquirido pelo homem no decorrer da sua vida e, ainda, a relevância de sua avaliação para se aferir o poder econômico de seu detentor e sua capacidade contributiva, é fácil perceber a importância da precisão da avaliação imobiliária para o equilíbrio das diversas relações travadas na sociedade.

Este cenário nos remete a valorar tecnicamente os bens imóveis, objetivando auxiliar as tomadas de decisões a respeito de valores, custos e alternativas de investimentos. Entre os diversos serviços nos quais pode ser aplicada, a Engenharia de Avaliações subsidia operações de garantia, transações de compra e venda, locação, decisões judiciais, incidência de tributos (prediais, territoriais e de transmissão), balanços patrimoniais, operações de seguros, separações ou cisões de empresas, desapropriações, entre outras.

A relevância da avaliação imobiliária atinge não só o cidadão, mas também o próprio processo de globalização, pois mercados imobiliários subavaliados ou hiperavaliados internacionalmente podem causar "bolhas" imprevisíveis, cujos "estouros" podem levar países a situações econômicas preocupantes. Neste sentido, vale lembrar a crise vivenciada pelos Estados Unidos (EUA) em 2005, quando a especulação imobiliária ocasionou a valorização em mais de 85% nos últimos dez anos, favorecendo o crescimento da metade dos empregos desde 2001 e também o endividamento dos proprietários de imóveis, por meio de hipotecas, com consequente aquecimento da economia. A desvalorização imobiliária súbita nos EUA causou danos significativos à economia daquele país e do mundo (Gomide, 2007).

2.2 Conceitos gerais

2.2.1 Normas e legislações

As avaliações de bens estão regulamentadas pela Norma Brasileira Registrada (NBR) 14653, da Associação Brasileira de Normas Técnicas (ABNT), que é constituída pelas seguintes partes, sob o título geral "Avaliação de Bens": Parte 1 – Procedimentos gerais e Parte 2 – Imóveis urbanos.

As demais partes da norma referem-se a outros tipos de avaliações de bens que não serão abordados nesta dissertação, a saber: Parte 3 – Imóveis rurais, Parte 4 – Empreendimentos, Parte 5 – Máquinas, equipamentos, instalações e bens industriais em geral, Parte 6 – Recursos naturais e ambientais e a Parte 7 – Patrimônios históricos.

Conforme destacado, as avaliações de imóveis urbanos são regidas pela "Norma de Avaliação de Bens - Parte 1 (Procedimentos Gerais)" e "Norma de Avaliação de Bens - Parte 2 (Imóveis Urbanos)". A Parte 1 da NBR 14653 fixa as diretrizes básicas para os procedimentos de excelência relativos ao exercício profissional e é exigível em todas as manifestações escritas de trabalhos que caracterizam o valor de imóveis urbanos, de seus frutos ou direitos. A Parte 2 da NBR 14653 detalha os procedimentos gerais da NBR 14653 - Parte 1, bem como complementa os conceitos, métodos e procedimentos gerais para os serviços técnicos de avaliação de imóveis urbanos. Esta parte da norma, portanto, visa a estabelecer os critérios a serem empregados pelos profissionais legalmente habilitados nos Conselhos Regionais de Engenharia, Arquitetura e Agronomia (CREAs).

As avaliações de bens são de competência exclusiva dos engenheiros, arquitetos e agrônomos, de acordo com sua habilitação profissional, conforme preceitua a Lei Federal n°. 5194, de 24 de dezembro de 1966 e as Resoluções n°. 205, n°. 218 e n°. 345 do Conselho Federal de Engenharia e Arquitetura (CONFEA).

2.2.2 Bem

De acordo com a NBR 14653 - Parte 1, bem é coisa que tem valor, suscetível de utilização ou que pode ser objeto de direito, que integra um patrimônio. Os bens são classificados em tangíveis, quando podem ser identificados materialmente, como é o caso dos imóveis e equipamentos, ou intangíveis, quando são imateriais, a exemplos das marcas e patentes.

Para o caso particular de imóveis urbanos, ou seja, aqueles situados dentro do perímetro urbano definido em lei, a NBR 14653 - Parte 2 classifica-os quanto ao uso (residencial, comercial, industrial etc.), quanto ao tipo (terreno, apartamento, casa etc.) e quanto ao agrupamento (loteamento, condomínio de casas, prédio de apartamentos etc.).

2.2.3 Valor

Segundo Ayres (1996), atribui-se valor a tudo aquilo que é útil ou escasso. Cabe à avaliação traduzir essa utilidade ou escassez numa quantia monetária e associá-la a uma necessidade ou desejo de possuir um bem.

Analogamente, Fiker (1997) definiu "valor" como a relação entre a intensidade das ne-

cessidades econômicas humanas, objetivas ou subjetivas, e a quantidade de bens disponíveis para atendê-las.

Por outro lado, a NBR 14653 - Parte 1 não apresenta a definição de valor, mas conceitua a expressão "valor de mercado" como sendo a quantia mais provável pela qual se negociaria voluntariamente e conscientemente um bem, numa data de referência, dentro das condições do mercado vigente.

Desta forma, o termo valor, quando desprovido de qualquer qualificativo, significará sempre o determinado pela lei da oferta e da procura, sendo também denominado valor de mercado ou valor venal. Ou seja, é o valor pelo qual se realizaria uma transação de compra e venda entre partes, desejosas mas não obrigadas à transação, ambas perfeitas conhecedoras do imóvel e do mercado e admitido um prazo razoável para se encontrarem.

Entretanto, quando a finalidade da avaliação assim exigir, poderão ser identificados outros valores diferentes daquele de mercado, tais como: valor patrimonial, valor em risco, valor de liquidação forçada, valor de desmonte, entre outros.

2.2.4 O mercado imobiliário

Do ponto de vista da economia, o mercado é o local onde agentes econômicos procedem à troca de bens por uma unidade monetária ou por outros bens. A existência de um mercado pressupõe a presença de três componentes: os bens levados ao mercado, as partes interessadas em vendê-los e as partes desejosas em adquiri-los. Quando se trata de bens imóveis, estes três elementos formam o mercado imobiliário.

O mercado imobiliário surge como uma resposta às necessidades de crescimento da cidade e é consequência direta da dinâmica de formação e desenvolvimento dos núcleos urbanos. A formação desses novos espaços urbanos constitui uma atividade econômica imprescindível ao crescimento da cidade e envolve complexa interação entre os agentes de oferta e demanda em torno dos bens imobiliários.

O bem imóvel possui características próprias que o distinguem, em termos econômicos, de outros bens duráveis disponíveis no mercado, fazendo com que o seu tratamento teórico seja diferenciado. Esta singularidade pode ser explicada em função da heterogeneidade, fixação espacial, alto custo de aquisição, longa vida útil e longo período de produção.

O imóvel é também um bem imperfeito, diferente de todos os outros bens econômicos; mesmo que semelhantes, dois ou mais imóveis sempre trarão pelo menos uma peculiaridade que os diferenciará. Consequentemente, o mercado imobiliário não é, pela sua própria natureza, de concorrência perfeita.[4] Neste ponto, convém ressaltar que apenas em um mercado de concorrência perfeita a formação do valor segue a lei da oferta e da procura, com curvas bem definidas. Somente neste caso o preço que se paga por um determinado bem coincide com seu valor de mercado e o equilíbrio entre a quantidade ofertada e a demandada ocorrerá no ponto em que as curvas de oferta e demanda se cruzam, conforme ilustrado na Figura 2.1.

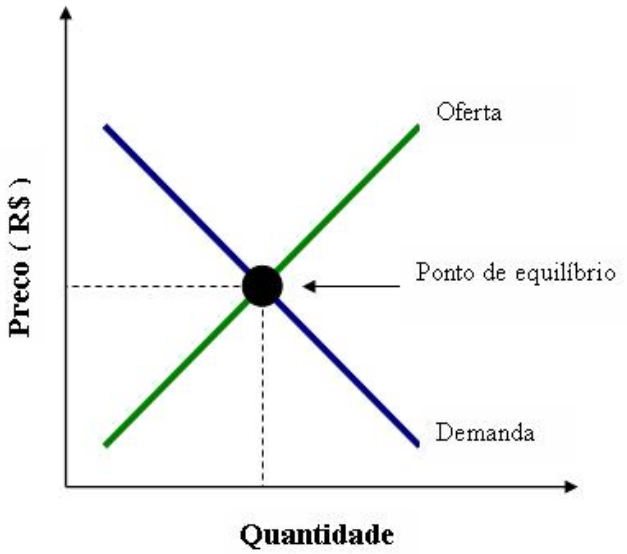

Figura 2.1: Representação do equilíbrio de mercado.

Por este motivo, não necessariamente o valor de mercado coincidirá com o preço do bem. Este último representa a quantidade de dinheiro pago em uma transação, enquanto o primeiro corresponde ao valor médio ou valor mais provável a ser atingido em transações normais, em dado momento.

[4] A concorrência perfeita corresponde a uma situação limite em que nenhum vendedor e nenhum comprador tem poder suficiente para influenciar o preço de mercado de um determinado bem. Para que tal situação se verifique é necessário que sejam atendidos os seguintes pressupostos: (i) Existência de um grande número de vendedores do mesmo produto ou serviço (bem) e com dimensão e estrutura de custos semelhante. (ii) Existência de um grande número de compradores e todos com a mesma informação disponível sobre a oferta existente no mercado. (iii) Existência de homogeneidade nos produtos ou serviços oferecidos no mercado. (iv) Inexistência de barreiras à entrada ou à saída de compradores e vendedores no mercado.

Os preços estão sujeitos às peculiaridades das transações e dos agentes e podem, por exemplo, se diferenciar do valor porque as partes têm liberdade restrita para negociar. Assim, a necessidade de venda ou compra imediata e a inexistência de um livre comércio podem alterar o preço de um bem, tornando-o superior ou inferior ao valor de mercado. Por tal razão, na prática, **estima-se o valor de mercado como a média dos preços**, haja vista as imperfeições do mercado imobiliário.

2.3 Métodos de avaliação

Consoante a NBR 14653 - Parte 1, os principais métodos para identificar o valor de um bem, de seus frutos e direitos são: (i) Método da capitalização da renda; (ii) Método involutivo; (iii) Método evolutivo; (iv) Método comparativo direto de dados de mercado.

Dantas (2005) ressalta que a aplicação da metodologia mais adequada para a realização de um trabalho avaliatório depende fundamentalmente das condições mercadológicas com que se defronta o avaliador, das informações coletadas neste mercado, bem como da natureza do serviço que se pretende desenvolver.

2.3.1 Método da capitalização da renda

O valor do bem é estimado com base na capitalização presente da sua renda líquida prevista e considerando-se cenários viáveis. Neste método, o valor estimado corresponde ao máximo de viabilidade que um investidor estaria disposto a pagar pelo bem, nas condições por ele estabelecidas. Desta forma, o valor obtido é chamado de valor econômico. Entretanto, desde que se utilizem informações advindas do mercado, o método da capitalização da renda pode ser aplicado para expressar o valor venal de um bem. Este método é quase sempre de possível aplicação e muitas vezes constitui-se no único recurso nos casos de imóveis isolados, atípicos ou quando o número de elementos comparáveis disponíveis no mercado é insuficiente para aplicação do método comparativo direto.

2.3.2 Método involutivo

Identifica o valor de mercado do bem, alicerçado no seu aproveitamento eficiente, baseado em modelo de estudo de viabilidade técnico-econômica, mediante empreendimento hipotético compatível com as características do bem e com as condições do mer-

cado no qual está inserido, considerando-se cenários viáveis para a execução e a comercialização do produto. A aplicação do método involutivo considera a receita provável da comercialização das unidades hipotéticas com base em preços obtidos em pesquisas, todas as despesas inerentes à transformação do bem, a margem de lucro do empreendedor, as despesas de comercialização, os prazos viáveis ao projeto, à execução e à comercialização, mediante taxas financeiras operacionais reais. O método involutivo é bastante utilizado na avaliação de glebas urbanizáveis,[5] onde o empreendimento considerado é um loteamento.

2.3.3 Método evolutivo

Consiste em identificar o valor do bem pelo somatório dos seus componentes. A composição do valor total do imóvel *avaliando* pode ser obtida da conjugação de métodos, a partir do valor do terreno, considerados os custos de reprodução das benfeitorias devidamente depreciados e o fator de comercialização,[6] ou seja,

$$VI = (VT + VB) \times FC,$$

em que VI é o valor do imóvel, VT é o valor do terreno, VB é o valor da benfeitoria e FC é o fator de comercialização.

2.3.4 Método comparativo direto de dados de mercado

O valor de mercado do bem é estimado por meio do tratamento técnico dos atributos dos elementos comparáveis, constituintes da amostra. A condição necessária à aplicação deste método é, portanto, a existência de um conjunto de dados comparáveis ao *avaliando*, em quantidade suficiente para uma análise estatística. Quando não for possível reunir elementos similares e em número suficiente, o avaliador deve procurar aplicar outra metodologia, em substituição ou em caráter complementar à avaliação por comparação. De todos os métodos existentes, o método comparativo direto de dados de mercado pode ser considerado o eletivo e é o mais utilizado para a avaliação de imóveis sempre que há dados suficientes para compor uma amostra.

[5] Terreno passível de receber obras de infraestrutura urbana, visando ao seu aproveitamento eficiente, por meio de loteamento, desmembramento ou implantação de empreendimento.

[6] Razão entre o valor de mercado de um bem e o seu custo de reedição ou de substituição, que pode ser maior ou menor que 1 (um).

2.3.4.1. Tratamento de dados

Na avaliação de terrenos urbanos pelo método comparativo direto de dados de mercado, a amostra composta de eventos relativos a lotes[7] similares entre si dificilmente será homogênea o bastante para permitir uma conclusão direta quanto ao valor médio de mercado desses imóveis, tornando-se imprescindível o tratamento dos dados coletados e a homogeneização dos valores.[8]

De acordo com a NBR 14653 - Parte 2, no tratamento dos dados podem ser utilizados, alternativamente e em função da qualidade e da quantidade de dados e informações disponíveis, as seguintes metodologias:

1. Tratamento por fatores (modelos determinísticos): na metodologia clássica, adota-se como denominador comum um terreno ideal, dito "paradigma", em relação ao qual os diferentes atributos dos terrenos observados no mercado são homogeneizados. Embora incontestavelmente cercados de um certo grau de subjetivismo, os fatores a serem aplicados na homogeneização dos diversos aspectos devem ser fruto de observação e aferição do mercado imobiliário. Aqui, o aspecto mais marcante é o uso da Estatística Descritiva;

2. Tratamento científico (modelos probabilísticos): tratamento de evidências empíricas pelo uso da metodologia científica que leve à indução de modelo validado para o comportamento de mercado. Aqui, são utilizadas ferramentas da Inferência Estatística.

Dantas (2005) tece o seguinte comentário acerca dos dois tratamentos acima descritos:

> "Tendo em vista que no tratamento científico a estimativa do valor é realizada utilizando-se modelos elaborados especificamente para avaliação do bem *avaliando*, pela substituição de suas características na equação resultante, pode-se imprimir maior nível de precisão e fundamentação ao trabalho."

[7] Embora do ponto de vista técnico um terreno não seja necessariamente um lote, neste trabalho trataremos as duas palavras como sinônimas.

[8] Entende-se por homogeneização dos valores o tratamento dispensado à amostra coletada, objetivando a retirada das discrepâncias existentes entre as características de cada elemento tomado como referência e o bem *avaliando*.

Dantas (2005) afirma ainda que a utilização generalizada dos fatores de homogeneização pode acarretar numa sensível perda do nível de precisão das avaliações, devido a questões de heterogeneidade espacial e multicolinearidade, principalmente.

Faz-se necessário mencionar que embora o tratamento científico esteja previsto na NBR 14653 - Parte 2, não há quaisquer recomendações ou orientações para análise dos dados e indução do comportamento imobiliário utilizando regressão não-paramétrica ou semiparamétrica, ao contrário do que ocorre para a regressão paramétrica via modelos lineares.

2.4 Metodologia científica no trabalho avaliatório

A introdução da metodologia científica no trabalho avaliatório tem como objetivo orientar o avaliador, desde a escolha das informações de interesse, a forma de coletá-las, analisá-las e tratá-las, na busca de modelos que expliquem, de maneira satisfatória, a variabilidade observada nos preços, no mercado em que se estuda (Dantas, 2005).

O método científico auxilia na compreensão não apenas dos resultados da investigação do valor do imóvel, mas do processo de investigação como um todo, podendo ser resumido nas seguintes etapas: (i) Observação do fenômeno, (ii) Planejamento da pesquisa, (iii) Processamento e edição das informações, (iv) Processamento e análise dos dados, (v) Verificação do ajuste da técnica de análise escolhida e (vi) Redação do relatório de pesquisa.

Acrescenta-se que quaisquer dos métodos apresentados na Seção 2.3 podem ser empregados seguindo as etapas supracitadas. No entanto, neste trabalho será dada maior ênfase à aplicação do método comparativo direto de dados de mercado, visto que este será o método utilizado na modelagem com dados reais do Capítulo 5.

É fundamental para a obtenção de uma avaliação confiável que o método e a técnica estatística empregados sejam compatíveis com o problema e estejam alicerçados na metodologia científica, pois, somente seguindo estes passos a Engenharia de Avaliações pode ser considerada como uma ciência: a ciência do valor.

2.4.1 Observação do fenômeno

Esta etapa, denominada na Engenharia de Avaliações de vistoria, consiste no reconhecimento do imóvel objeto da avaliação e visa à caracterização não apenas do *avaliando*, mas da região e do contexto imobiliário no qual o bem está inserido.

2.4.1.1. Vistoria do terreno

Na vistoria do terreno, contendo ou não construções, devem ser observados os aspectos que possam influenciar na formação do seu valor, a exemplo do formato, dimensões, topografia, consistência do solo, utilização atual, aspectos ligados à infraestrutura urbana, equipamentos comunitários existentes, possibilidades de desenvolvimento local, entre outros.

É nesta fase que se formam as primeiras concepções acerca das possíveis variáveis influenciantes na formação dos preços, embora no decorrer do trabalho outras variáveis possam revelar-se importantes.

2.4.1.2. Seleção e análise de variáveis regressoras

De acordo com o conceito empregado na inferência estatística, as variáveis são características (ou atributos) observáveis na amostra, que, em princípio, devem variar entre os elementos que a compõem. Cabe ao engenheiro de avaliações presumir quais variáveis supostamente são mais relevantes para explicar as variações na variável de interesse, para que se confirme ou não a veracidade dessa suposição.

As variáveis independentes — compreendendo as características físicas (área, frente etc.), de localização (bairro, logradouro etc.) e econômicas (oferta ou transação, época etc.) — devem ser estabelecidas, *a priori*, com base em teorias existentes ou raciocínios precedentes, conhecimentos adquiridos, senso comum, trabalhos anteriores e outros atributos que se revelem importantes no decorrer do processo avaliatório.

2.4.2 Planejamento da pesquisa

O planejamento experimental permite ao investigador direcionar as etapas e prever várias situações da pesquisa para atingir o objetivo do estudo. Embora diversas

conjunturas não previstas possam surgir, muitos problemas podem ser eliminados, ou minimizados, por meio de uma preparação prévia. Essencialmente, busca-se identificar uma estratégia eficiente de medida para cada etapa da pesquisa.

Na fase de planejamento, faz-se a escolha, definição e delimitação do problema em análise, observam-se as teorias e abordagens a serem empregadas, os conceitos e hipóteses que devem ser levados em consideração e os métodos a serem utilizados.

Num trabalho avaliatório em que se opta pelo método comparativo direto de dados de mercado, um possível delineamento da pesquisa pode ser assim estabelecido: (i) Estudo de campo (objetivo: conhecer o *avaliando* e o contexto urbano ao qual o bem está inserido; estratégia: vistoria e visita a região), (ii) Seleção das variáveis que julgamos influenciantes na formação dos preços do imóvel (objetivo: focalizar as características mais importantes da população; estratégia: consulta a trabalhos anteriores), (iii) Planejamento da coleta de dados (objetivo: identificar a técnica mais adequada para obtenção dos preços de mercado; estratégia: entrevista mediante aplicação de questionário diretamente ao proprietário do imóvel), (iv) Definição de um método de avaliação (objetivo: identificar a metodologia avaliatória para estimar o valor do bem; estratégia: lançar mão de todas as evidências disponíveis, tais como, tipologia do *avaliando*, tipo de informação coletada etc), (v) Verificação das hipóteses (objetivo: confrontar com a realidade as suposições sobre o mercado estabelecidas pelo pesquisador, tais como, terrenos de esquina são mais valorizados que terrenos situados no meio da quadra ou ainda, terrenos planos são mais valorizados que terrenos acidentados; estratégia: estudo exploratório dos dados com ênfase na análise dos diagramas de dispersão, testes de hipóteses, entre outros), (vi) Escolha de uma técnica para tratamentos dos dados (objetivo: examinar o comportamento da variação dos preços dos imóveis com base na variação de algumas variáveis explicativas, ou seja, construir modelos elucidativos do mercado imobiliário; estratégia: análise de regressão e emprego do modelo de regressão linear clássico, por exemplo), (vii) Explicação dos resultados (objetivo: interpretar o comportamento do mercado em relação a cada variável; estratégia: análise dos parâmetros quanto aos aspectos de sensibilidade e elasticidade) e (viii) Relatório da pesquisa (objetivo: descrever o raciocínio desenvolvido e os resultados obtidos; estratégia: elaborar o laudo de avaliação baseado na descrição da metodologia utilizada e nos resultados alcançados em cada etapa do processo avaliatório).

Cabe mencionar que o delineamento acima exposto é apenas um exemplo hipotético e que não necessariamente todas as avaliações seguirão a mesma sequência de objetivos e estratégias, visto que cada trabalho de avaliação apresenta um problema diferente.

Um último e importante aspecto do planejamento é a determinação do cronograma de ações. É ele quem indicará o tempo estimado para cada etapa da pesquisa e, consequentemente, a duração total do processo avaliatório.

2.4.3 Processamento e edição das informações

Para Volpato (2007), se considerarmos a natureza empírica da ciência moderna, os enunciados teóricos devem ser confrontados com dados do mundo físico. Nesse sentido, a coleta de dados é parte integrante indispensável de uma pesquisa científica.

Na pesquisa imobiliária, a coleta de dados objetiva a composição de uma amostra formada por imóveis semelhantes entre si e pesquisados na mesma região geoeconômica. Esta amostra é formada pelos preços praticados no mercado imobiliário associados às respectivas características físicas, locacionais e econômicas.

Os preços das transações são, possivelmente, as informações mais cobiçadas nas avaliações de imóveis. Contudo, são também as mais difíceis de serem obtidas, visto que, geralmente, as pessoas podem sentir-se "ofendidas" em sua sensibilidade quando abordadas sobre fatos de sua vida pessoal. Por exemplo, sobre o preço pelo qual comprou ou vendeu um determinado bem.

Entretanto, vale salientar que preços de imóveis colocados em oferta podem também ser considerados como eventos do mercado imobiliário, porém são menos consistentes que os de transação. Nesse caso, a média dos preços de oferta servirá como um indicador de um limite superior de preço que se pagaria pelo bem no mercado.

Normalmente, as informações de imóveis transacionados ou ofertados são obtidas diretamente com o proprietário do bem, por meio de placas no próprio local, por intermédio de agentes imobiliários, via classificados de jornais e/ou internet, em consultas ao Cartório de Registro de Imóveis, declarações de ITBI no arquivo da prefeitura, entre outras fontes. Todavia, é preciso reconhecer que todas as fontes mencionadas têm suas limitações e, por essa razão, devemos ser cautelosos e precavidos na coleta de dados.

Conforme ficou evidenciado, a base da Engenharia de Avaliações não envolve somente

a lógica, mas também a informação (coleta de dados). Por isso, é imprescindível que os dados sejam *fidedignos* e *precisos*,[9] caso contrário toda a avaliação sobre as conclusões elaboradas fica comprometida.

2.4.4 Processamento e análise dos dados

Nesta fase do trabalho avaliatório emprega-se uma grande variedade de técnicas gráficas e analíticas visando a maximizar a descoberta de informações ocultas na estrutura dos dados (como tendências, relações e padrões), a identificação de variáveis importantes, a detecção de comportamentos anômalos do fenômeno observado ou de eventuais observações atípicas e a indicação de possíveis estruturas de regressão para o ajuste dos dados.

Adicionalmente, constroem-se modelos que representem o mercado imobiliário, estudam-se as relações entre as variáveis, estimam-se os parâmetros, realizam-se testes de significância sobre os parâmetros e verificam-se as hipóteses subjacentes ao modelo em uso.

2.4.4.1. Da montagem de modelos que representem o mercado imobiliário

No mercado imobiliário compradores e vendedores praticam atos racionais e emocionais. Ao considerar a prática de atos racionais, os atributos do imóvel são tomados de forma lógica e o seu comportamento pode ser equacionado de forma determinística. Já em relação aos atos emocionais, os participantes do mercado empregam critérios subjetivos, tendo em vista que cada um pondera de forma diferente suas qualidades, seus defeitos, sua utilidade ou sua necessidade, dependendo de usos e costumes locais e das formas diferenciadas ao longo do tempo. Isso explica, em parte, a grande diversidade encontrada nos preços ofertados, geralmente maior que aquela verificada na pesquisa de preços efetivamente transacionados (Grandiski & Oliveira, 2007).

Na prática, é possível verificar certa proporcionalidade entre a variação dos preços dos bens e a variação de suas características ou atributos, o que nos leva a concluir que, bens equivalentes, em mercados também equivalentes, podem ter os mesmos preços. Esse comportamento depende, simultaneamente, de fatores endógenos (próprios do mercado

[9] Fidedignidade significa que os dados coletados correspondem ao que imaginamos que correspondam (Volpato, 2007). O conceito de precisão refere-se ao rigor na determinação de uma medida.

e específico dos bens) e exógenos (decorrentes de crises econômicas, políticas, guerras, legislação que confere incentivos, entre outros).

Em Engenharia de Avaliações o que se busca é uma relação entre os preços que são praticados no mercado, que formam a imagem da função, e as diversas características que influem decisivamente na formação dos mesmos, as quais compõem o domínio. Como dados de mesmas características não são negociados necessariamente pelo mesmo preço, devido a várias razões, entre elas a própria aleatoriedade do comportamento humano, para um mesmo elemento do domínio poderiam estar associados vários elementos da imagem. Isto torna a relação entre os preços e as características que os influenciam de caráter aleatório. Desta forma, o valor de mercado pode ser considerado como uma variável aleatória, cuja estimação pontual é feita pela média dos preços nele praticados, atendidos os pressupostos e condições da técnica utilizada na modelagem (Dantas, 2003).

2.4.4.2. A modelagem tradicional

O preço do bem é definido em função de seus diversos atributos, os quais são de difícil quantificação e qualificação. Daí a importância do uso de modelos probabilísticos, que podem retratar bem essa impossibilidade de abranger todos os aspectos que determinam ou condicionam o preço do imóvel.

Isto significa dizer que o valor de um bem imóvel é explicado por uma parcela sistemática ou determinística e por um componente aleatório imprevisível ou não sistemático, cujo modelo probabilístico, derivado de modelos econômicos baseados na teoria dos preços hedônicos, passa a ser descrito por uma relação estocástica entre k variáveis independentes $(X_1, X_2, \ldots, X_k)$ conhecidas e uma variável dependente Y definida pela equação

$$Y = \underbrace{f(X_1, X_2, \ldots, X_k)}_{\text{componente sistemático}} + \underbrace{\epsilon}_{\text{componente aleatório}},$$

em que ϵ é o distúrbio estocástico.[10]

Tem-se observado na literatura uma intensa utilização do modelo clássico de regressão linear na estimação das equações de preços hedônicos de imóveis, constituindo-se quase uma metodologia única. Porém, falta de normalidade e presença de heterosce-

[10] Também denominado de termo de erro estocástico ou termo aleatório não-observável.

dasticidade[11] são bastante comuns em dados imobiliários e a negligência destes efeitos pelo pesquisador pode resultar em inferências enganosas sobre os parâmetros do modelo (Dantas, 2003).

2.4.5 Verificação do ajuste da técnica de análise escolhida

É importante observar que o emprego da metodologia científica mediante o uso da Estatística indutiva minimiza a subjetividade do trabalho avaliatório. Contudo, deve-se registrar que os procedimentos e técnicas estatísticas constituem apenas instrumentos que facilitam a interpretação dos resultados, sendo necessário que o pesquisador trace um paralelo entre os resultados obtidos empiricamente e as teorias já existentes, bem como argumente sobre a corroboração ou refutação das hipóteses e indique, se for o caso, a construção de novas teorias e paradigmas.[12]

Neste sentido, é recomendável que o pesquisador explicite os resultados do trabalho avaliatório por intermédio de análise quantitativa e qualitativa do comportamento do mercado em relação a cada variável, bem como interprete os parâmetros quanto aos aspectos de sensibilidade e elasticidade, a fim de constatar a adequação teórica e lógica do modelo.

Nesta etapa, infere-se o valor médio de mercado e o intervalo de confiança, no qual se afirma com determinada probabilidade que o valor de mercado está nele contido.

2.4.6 Redação do relatório da pesquisa

O relatório de pesquisa, no âmbito da Engenharia de Avaliações, corresponde ao laudo de avaliação e constitui a etapa final do processo avaliatório. Deve conter todos os elementos relevantes considerados no seu desenvolvimento: a identificação do objeto da avaliação, a técnica de coleta de dados, a metodologia de regressão adotada e as devidas interpretações e conclusões fundamentadas. A redação do conteúdo deve ser feita de forma clara, objetiva e sequenciada. Mais detalhes sobre os requisitos exigidos e a forma de apresentação do laudo de avaliação podem ser obtidos na NBR 14653 - Parte 2.

[11] Sob heteroscedasticidade, os estimadores de mínimos quadrados ordinários permanecem não-viesados e consistentes, mas deixam de ser eficientes (variância mínima). Adicionalmente, o estimador usual de sua matriz de covariâncias não é válido. Em consequência, os testes t e F convencionais podem conduzir a inferências errôneas.

[12] A idéia geral do vocábulo "paradigma" é de uma grande noção que norteia as pessoas em suas decisões (Volpato, 2007).

CAPÍTULO 3

Regressão: alguns conceitos básicos

"... geralmente dirigimos nas pontes sem nos preocupar com a solidez de sua construção porque estamos razoavelmente certos de que alguém conferiu rigorosamente os princípios e a prática de sua engenharia. Os estatísticos devem fazer o mesmo com os modelos de regressão ou, então, incluir a advertência 'não nos responsabilizamos pelo colapso'..."

(Texto extraído e adaptado de Hendry, D., *Dynamic Econometrics*, 1998)

O objetivo deste capítulo é sumarizar alguns conceitos básicos de regressão e fazer uma revisão acerca dos principais procedimentos e técnicas não-paramétricas de suavização (também denotadas na literatura de técnicas não-paramétricas de *alisamento*) utilizados na estimação de modelos que combinam componentes não-paramétricos e paramétricos.

Embora a teoria exposta neste capítulo seja diminuta comparada ao que se tem disponível na literatura sobre o assunto, descreveremos de forma concisa, porém sem prejuízo do rigor matemático, os tópicos necessários para compreensão do desenvolvimento deste trabalho, particularmente no que se refere ao ajuste de modelos aditivos generalizados para posição, escala e forma, que serão apresentados no Capítulo 4.

3.1 Regressão paramétrica e não-paramétrica

Uma das abordagens mais utilizadas em análise de dados experimentais ou observacionais é o estudo e análise da relação funcional entre uma ou mais variáveis explicativas e uma variável de interesse mediante ajuste de modelos de regressão.

Os modelos de regressão paramétricos e não-paramétricos representam distintas formas para a análise de regressão e constituem, essencialmente, técnicas estatísticas que buscam estabelecer uma relação matemática entre as variáveis dependentes e independentes que caracterizam um fenômeno aleatório de interesse.

Para tanto, suponha que as observações são coletadas de uma variável contínua Y em n valores da variável independente X. Seja (x_i, y_i), $i = 1, 2, \ldots, n$, tal que o seguinte modelo de regressão pode ser formulado:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

em que as variáveis aleatórias ϵ_i têm média zero, são descorrelacionadas e possuem variância comum σ^2.Mais ainda, $f(x_i)$ são valores obtidos de alguma função f, desconhecida, calculada nos pontos $x_i, \ldots, x_n$. A função f é geralmente chamada de função de regressão ou curva de regressão (Dias, 2001b).

Na regressão paramétrica, as distribuições condicionais da relação entre as variáveis preditoras e a variável resposta são modeladas por meio de famílias de distribuições paramétricas, cuja forma da função distribucional é conhecida, sendo desconhecidos apenas os valores dos seus parâmetros. Além disto, a forma funcional (matemática) desconhecida da relação entre regressando e regressores pode ser determinada mediante o ajuste de curvas aos dados e estimação dos coeficientes da equação de regressão. Isto é, um modelo de regressão paramétrico assume que a forma de f é conhecida, exceto por um número finito de parâmetros, e pode ser descrito por

$$y_i = f(x_i, \beta_1, \ldots, \beta_k) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3.1}$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top \in I\!R^k$. Note que determinar, a partir dos dados, a curva f é equivalente a determinar o vetor $\boldsymbol{\beta}$ de parâmetros. Observe que se f tem forma linear, ou seja, $f(x, \boldsymbol{\beta}) = \sum_{i=1}^{k} \beta_i \vartheta_i(x)$, para algum conjunto de funções $\vartheta_1, \ldots, \vartheta_k$, estamos na situação do modelo de regressão linear paramétrico. Neste caso, são estimados os

coeficientes de uma forma funcional determinada previamente e o pesquisador verifica quão bem as estimativas se aproximam dos coeficientes reais (populacionais) por meio de testes de hipóteses. Não há maior preocupação com a curva estimada.

Na regressão não-paramétrica ocorre uma mudança de perspectiva. Como a forma funcional de f não é conhecida, assume-se que a curva de regressão pertence a uma coleção infinito-dimensional de funções e sua estimação passa a ocupar o papel central na análise. Sendo assim, mesmo que a função continue a ser estimada a partir do ajuste de parâmetros livres, o conjunto de "formas" que a função pode assumir (classe de funções que o modelo do estimador pode prever) é muito amplo. Note que o caso paramétrico é mais restrito, pois nele presume-se que f pertence a uma família específica de curvas.

Na análise não-paramétrica, em contraste com o método paramétrico, estima-se uma função média sem referência a uma forma funcional previamente estabelecida e o experimentador precisa apenas escolher o espaço de funções apropriado, ao qual se acredita que f pertença, permitindo que os "dados falem por si mesmos". Tal característica é de grande interesse, pois muitas vezes a análise teórica não estabelece a forma estrutural entre as variáveis ou estabelece formas estruturais competitivas. Neste caso, um teste de hipóteses pode ser empregado para verificar se a curva estimada reproduz a verdadeira função média.

Para Härdle (1990), a abordagem não-paramétrica apresenta pelo menos quatro pontos fortes em relação ao método paramétrico: (1) possibilita uma forma versátil para explorar a relação entre as variáveis do modelo; (2) fornece um modelo preditivo sem a necessidade de referências a um modelo paramétrico fixo; (3) é uma boa ferramenta para a detecção de pontos espúrios[1] a partir do estudo da influência que cada ponto exerce no ajuste; (4) trata-se de um método com moderada flexibilidade para substituição de observações desconhecidas (*missing*) mediante interpolação de observações adjacentes.

No entanto, conforme apropriadamente ressaltou Dias (2001b), a flexibilidade e a versatilidade dos modelos de regressão não-paramétricos não significam que este método é superior à abordagem paramétrica. Na verdade, técnicas de regressão não-paramétricas podem ser usadas para verificar a validade de um certo modelo paramétrico proposto. Reciprocamente, a forma da curva de regressão obtida por técnicas não-paramétricas pode

[1] Isto é, pontos de alavancagem e *outliers*.

sugerir um modelo paramétrico. Assim, procedimentos de regressão não-paramétricos poderão ser o estágio final de uma análise de dados ou meramente um passo confirmatório ou exploratório do processo de modelagem.

3.2 Regressão semiparamétrica

Infelizmente, na prática, raramente conhecemos a forma funcional verdadeira do modelo ou as suposições probabilísticas a serem postuladas. Contudo, é possível que num rol de k variáveis integrantes de um determinado modelo, tenhamos conhecimento *a priori* das hipóteses da distribuição real e/ou do formato das funções de exatamente g variáveis, com $g < k$, e para as demais variáveis, $k - g$, não dispomos de qualquer informação sobre a forma funcional entre as variáveis.

A regressão semiparamétrica surge como uma opção prática, consistente e robusta para este tipo de análise ao permitir a modelagem de formas funcionais complexas que contemplam simultaneamente componentes paramétricos e não-paramétricos. Trata-se de uma alternativa mais flexível do que a abordagem clássica e menos restritiva para a estimação de uma curva desconhecida.

Os modelos semiparamétricos estão intrinsecamente relacionados ao conceito de modelos aditivos, razão pela qual faremos uma breve exposição acerca desta classe de modelos na subseção a seguir.

3.2.1 Modelos aditivos

Os modelos aditivos são caracterizados pela habilidade de prover bons ajustes a um número variado de situações e o seu emprego pode ser observado tanto no desenvolvimento e aprimoramento de métodos estatísticos quanto em aplicações nas mais diversas áreas do conhecimento.

O modelo aditivo é uma generalização de um modelo linear e, sendo assim, possui uma característica importante: o efeito de uma variável em uma superfície de resposta é aditivo em relação aos efeitos de qualquer outra variável. Em outras palavras, podemos verificar a contribuição individual de cada variável na explicação da variabilidade da variável de interesse.

Nesta classe de modelos há três tipos de abordagens: apenas paramétrica, estritamente não-paramétrica e uma mistura das duas abordagens supracitadas de forma simultânea, conforme apresentaremos a seguir.

Semelhantemente ao que foi estabelecido no modelo de regressão da Equação (3.1), consideremos n pares de observações (x_i, y_i), $i = 1, 2, \ldots, n$, e tomemos uma função f que estabelece a relação entre as variáveis X e Y da foma

$$y_i = f(x_i) + \epsilon_i, \tag{3.2}$$

sendo f uma função a ser estimada e supondo que, para os erros aleatórios ϵ_i, tenhamos $\mathrm{E}(\epsilon_i) = 0$ e $\mathrm{Var}(\epsilon_i) = \sigma^2$. Consequentemente, se considerarmos um conjunto de k variáveis explicativas representadas em uma matriz $\boldsymbol{X}$, de dimensão $n \times k$ e posto completo, com a i-ésima linha dada por $\boldsymbol{X}_i = x_{i1}, x_{i2}, \ldots, x_{ik}$, teremos uma função f de modo que $y_i = f(x_{i1}, x_{i2}, \ldots, x_{ik}) + \epsilon_i$.

Objetivando tornar o modelo linear no efeito das variáveis regressoras, Buja *et al.* (1989) e Hastie & Tibshirani (1990) consideraram a função f como uma soma de funções f_j, $j = 1, \ldots, k$, para cada uma das colunas de $\boldsymbol{X}$, sendo que para a i-ésima linha de $\boldsymbol{X}$ temos

$$f(\boldsymbol{X}_i) = f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik})$$

e, desta forma, o modelo passa a ser

$$\boldsymbol{y}_i = f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik}) + \epsilon_i, \tag{3.3}$$

em que $\mathrm{y}^\top = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ é um vetor $n \times 1$ de respostas e a i-ésima linha da matriz $\boldsymbol{X}$ é $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ik})$ é um vetor $n \times 1$ de variáveis explicativas, com $\mathrm{E}(\epsilon_i) = 0$ e $\mathrm{Var}(\epsilon_i) = \sigma^2$ (Bidurin & Milan, 2001). O Modelo (3.3) é chamado por Buja *et al.* (1989) de regressão aditiva ou modelo aditivo.

O preditor aditivo obtido em (3.3) corresponde a um modelo totalmente não-paramétrico e cada função f_j, para $j = 1, \ldots, k$, é uma função univariada arbitrária a ser estimada por meio de algum procedimento não-paramétrico (mecanismo de suavização). Aqui, o problema conhecido na literatura como "maldição da dimensionalidade"[2] (em

[2] À medida em que o número de variáveis independentes cresce o estimador não-paramétrico deve ponderar sobre regiões muito grandes do espaço, aumentando rapidamente o número de observações necessário para produzir uma estimativa de qualidade (Hastie *et al.*, 2001).

inglês, *curse of dimensionality*) é superado, uma vez que o procedimento de estimação é construído com suavizadores univariados, isto é, cada função f_j é estimada de modo univariado. Logo, o modelo aditivo generaliza o modelo de regressão linear múltipla com a função linear substituída por uma quantidade aditiva de funções univariadas.

Modelos cujos preditores combinam formas paramétricas de algumas (g) variáveis preditoras com termos não-paramétricos de outras $(k-g)$ variáveis também fazem parte dessa classe. Neste caso, o preditor pode ser escrito como

$$y_i = \beta_1 x_{i1} + \cdots + \beta_g x_{ig} + f_1(x_{i,g+1}) + \cdots + f_{k-g}(x_{ik}) + \epsilon_i.$$

Esses modelos são denominados de semiparamétricos.[3]

Acrescenta-se que os resultados assintóticos sobre a eficiência e consistência dos estimadores obtidos por Schick (1986, 1993 e 1996) e Bhattacharya & Zao (1997) têm ajudado a consolidar a aplicação dos modelos semiparamétricos, os quais têm sido amplamente discutidos na literatura (vide Stone, 1985; Hastie & Tibshirani, 1990; Lee, 1990).

3.2.2 Função densidade de probabilidade

A função densidade de probabilidade caracteriza completamente um espaço amostral composto por uma determinada variável aleatória e é definida como uma função real mensurável e não-negativa satisfazendo

$$\int_{-\infty}^{\infty} f(y)dy = 1.$$

Seja Y uma variável aleatória com função densidade de probabilidade f. A especificação da função f fornece uma descrição natural da distribuição da variável e permite que probabilidades associadas a Y sejam determinadas mediante a relação

$$P(a < Y < b) = \int_a^b f(y)dy \quad \text{para todo } a < b.$$

Entende-se por estimação de densidade o processo de construção de uma estimativa da função densidade de probabilidade que representa um determinado conjunto de dados. A sua identificação fornece valiosas indicações na análise exploratória, em que descrevem-se aspectos como multimodalidade, assimetria, tipo de cauda (longa ou pesada), curtose,

[3] Também referenciados na literatura por modelos parcialmente lineares (Speckman, 1988).

entre outras características, e também na análise confirmatória, como instrumento indicativo para utilização de diferentes métodos (análise discriminante, análise de agrupamentos, testes para a moda etc.).

Silverman (1986) destaca ainda que a função densidade estimada tem um efeito visual impactante que facilita a compreensão do problema até por pessoas não familiarizadas com a Estatística. Neste sentido e adaptando o exemplo apresentado em Silverman (1986), expomos na Figura 3.1 três maneiras de descrever a distribuição gama com parâmetros $\theta = 2.0$ e $\alpha = 2.0$: (a) gráfico da função densidade; (b) gráfico da função de distribuição acumulada; (c) a expressão matemática da função densidade. Possivelmente, o gráfico da função densidade de probabilidade seja o mais propenso a ser escolhido para explicação da distribuição gama, haja vista a interpretação intuitiva inerente à sua representação e possibilidade de expor considerações acerca de (b) e (c) sem que fossem necessárias às suas visualizações.

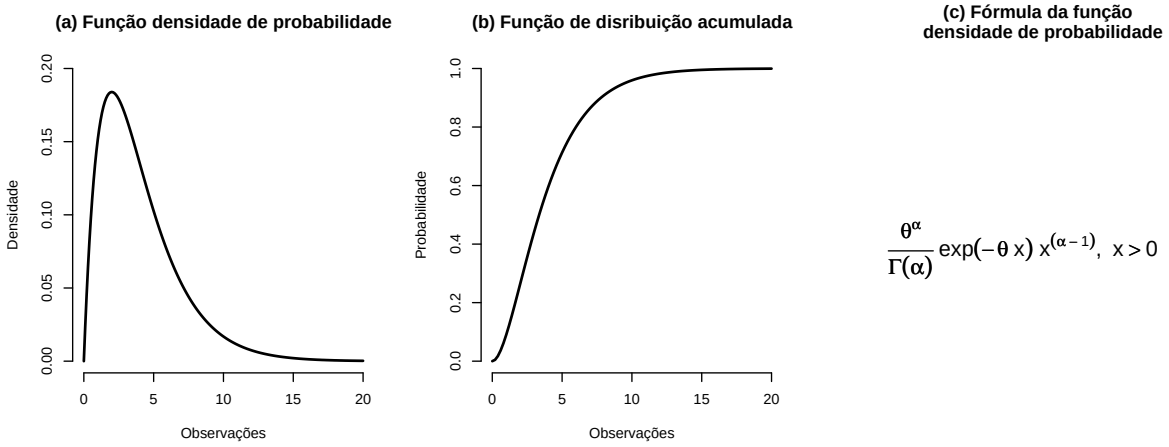

Figura 3.1: Três maneiras de descrever a distribuição gama.

Diversos procedimentos não-paramétricos para estimar a função densidade de probabilidade estão disponíveis na literatura (vide Silverman, 1986; Pagan & Ullah, 1999; Härdle, 1990) e são frequentemente referenciados como métodos de suavização (em inglês, *smoothing methods*).

3.3 Métodos de suavização

Para Lima *et al.* (2001), um suavizador (também denotado na literatura por *alisador*) é uma ferramenta que descreve a variação da média de uma variável Y como função de uma ou mais variáveis não-estocásticas[4] $X_1, \ldots, X_k$. Quando a variação da média de Y é descrita em função de apenas uma variável X, o suavizador é denominado unidimensional. Quando k variáveis, $X_1, \ldots, X_k$, são consideradas, diz-se que o suavizador é multidimensional.

Na maioria das vezes, um suavizador é utilizado com o objetivo de ajustar o modelo

$$y_i = f(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.4}$$

em que f é uma função não especificada e os ϵ_i são erros aleatórios distribuídos independentemente com média zero e variância σ^2.

Em boa parte das técnicas de suavização, o valor suavizado $\hat{y}_i$ é obtido com base em uma "média" de t observações na vizinhança de uma dado valor x_i. Diferentes formas de cálculo dessa média em uma vizinhança de x_i definem diferentes métodos de suavização.

Dois suavizadores serão destacados e descritos neste trabalho: *kernel* e *splines*. Enquanto o primeiro distingue-se pela robustez da teoria e aplicabilidade nas mais diversas situações práticas, primariamente na modelagem de dados econométricos, o último constitui uma alternativa à suavização por *kernel* baseada na penalização da curvatura da função a ser estimada e caracteriza-se pela eficiência e rapidez sob o ponto de vista computacional. Em ambos os casos, a curva suavizada ou curva ajustada é construída com base nos pontos $(x_i, \hat{y}_i)$, em que $\hat{y}_i$ é o valor previsto (pela suavização) de Y para $X = x_i$. Esses valores são obtidos sem a adoção de um modelo paramétrico relacionando Y e X.

3.3.1 Suavização por *kernel*

O método *kernel* ou função núcleo, desenvolvido a partir dos trabalhos de Rosenblatt (1956) e Parzen (1962), é uma técnica não-paramétrica para estimação de curvas de densidade baseado na ponderação local.[5] A teoria que embasa este método é bem estabelecida e auxilia a compreensão de diversos aspectos no campo da regressão não-paramétrica.

[4] Embora esta suposição possa ser relaxada, admitiremos neste trabalho que os valores assumidos pelos regressores X_i, com $i = 1, \ldots, k$, são fixados em amostras repetidas.

[5] A estimativa do valor da função em um ponto x_i é influenciado pelas observações próximas de x_i.

Antes de apresentarmos o método *kernel* propriamente dito, discorreremos sobre o histograma e o estimador "ingênuo", que são aproximações "grosseiras" da função densidade de probabilidade e cujas generalizações auxiliarão na construção dos estimadores *kernel*.

3.3.1.1. Histograma

Histograma é o método não-paramétrico mais antigo e utilizado para estimação de densidades, porém é pouco rigoroso e de aplicabilidade complexa quando não estamos no caso univariado. A ideia por trás desta técnica é dividir o intervalo de variação dos dados em subintervalos de comprimento h (em inglês, denominados de *bins*) e quantificar o número de observações que pertence a cada intervalo.

Inicialmente, consideremos uma variável aleatória discreta X, em que x seja um dos valores que a variável pode assumir, e o interesse seja a estimação de $f(x)$ a partir das observações $x_i,\ i = 1, \ldots, n$. O histograma é, então, definido por

$$\hat{f}_1(x) = \frac{1}{nh} \times (\text{número de } x_1, \ldots, x_n \text{ iguais a } x).$$

Cabe aqui ressaltar que o gráfico de pontos (em inglês, *dot plot*) é um tipo particular de histograma para $h \to 0$.

O histograma constitui uma ferramenta bastante útil para representação dos dados mas apresenta uma série de inconvenientes matemáticos (por exemplo, a dependência do comprimento do intervalo e a descontinuidade da função) que o torna pouco atrativo e limitado quando comparado aos demais métodos não-paramétricos disponíveis para estimação de densidades (vide Silverman, 1986).

3.3.1.2. Estimador "ingênuo"

Agora, admitamos que X é uma variável aleatória contínua. Neste caso, a probabilidade de X assumir um valor específico x é igual a 0, e $f(x)$ será estimada a partir da média dos valores x_i que estão localizados próximos de x, digamos no intervalo $x \pm \frac{h}{2}$, sendo que h, como já mencionamos, é o comprimento do intervalo.

Dado o exposto, um estimador da função densidade $f(x)$ pode ser dado por

$$\hat{f}_2(x) = (nh)^{-1} \sum_{i=1}^{n} I(x - \frac{h}{2} \leq x_i \leq x + \frac{h}{2}),$$

em que $I(\mathcal{A}) = 1$, se $\mathcal{A}$ for verdadeiro e 0, caso contrário. Alternativamente, podemos escrever

$$\begin{aligned} \hat{f}_2(x) &= \frac{1}{nh}\sum_{i=1}^{n} I\left(-\frac{1}{2} \leq \frac{x_i - x}{h} \leq \frac{1}{2}\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} I\left(|\psi_i| \leq \frac{1}{2}\right), \end{aligned} \tag{3.5}$$

em que $\psi_i = (x_i - x)/h$.

Note que na Expressão (3.5), $\hat{f}_2(x)$ é a frequência relativa por unidade no intervalo $(x - h/2,\ x + h/2)$, sendo que x é o ponto central. Em um sentido mais estrito, $\hat{f}_2(x)$ é exatamente a ordenada do histograma em x. Assim, o estimador em (3.5) pode ser visto como uma tentativa de construção de um histograma que se baseia nas observações "locais" para x, em que cada ponto x é o centro de um intervalo amostral. Fix & Hodges (1951) denominaram $\hat{f}_2(x)$ de estimador "ingênuo".

Claramente, a indicatriz ou função peso $I(-1/2 \leq \psi \leq 1/2)$, em (3.5), depende da distância entre x_i e x. Se esta distância, em valor absoluto, for menor ou igual a 1/2 o peso será 1, caso contrário será 0 (zero).

Perceba que a estimativa $\hat{f}_2(x)$ depende fortemente da escolha de h. Quando variamos o tamanho do intervalo h obtemos diferentes formas de $\hat{f}_2(x)$. Por exemplo, quanto menor for o tamanho de h, menos suave será a estimativa. De maneira oposta, quanto maior for h, mais suave será a estimativa final. Mais detalhes sobre o processo de escolha de h para representação dos dados via histograma e estimador "ingênuo" podem ser obtidos em Härdle (1990) e Silverman (1986).

3.3.1.3. Estimação de densidades por *kernel*

O estimador da função densidade dado em (3.5) apresenta a inconveniente característica de não-suavidade (Silverman, 1986). Além disso, $\hat{f}_2(x)$ não é uma função contínua e tem derivada nula em todos os pontos, exceto nos pontos de salto $x_i \pm h/2$. Rosenblatt (1956) tratou este problema substituindo a função indicadora em (3.5) por uma função *kernel* K, real e positiva, satisfazendo

$$\int_{-\infty}^{\infty} K(\psi)\, d\psi = 1.$$

Geralmente K é uma função densidade de probabilidade simétrica, a exemplo da densidade normal, ou uma função indicatriz $I(\psi)$, como definida para o estimador "ingênuo". Desta forma, generalizando (3.5), o estimador *kernel* com *kernel* K é dado por

$$\hat{f}_3(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n} K(\psi_i), \tag{3.6}$$

em que $\psi_i = (x_i - x)/h$ e h, como definido para o estimador ingênuo em (3.5), é o parâmetro de suavização (*smoothing parameter*), também denotado na literatura de janela (em inglês, *bandwidth)*, que controla o tamanho da vizinhança no entorno de x no qual a função núcleo será aplicada.

Perceba que valores pequenos de h implicam que somente observações perto de x recebem algum peso, enquanto que h grande significa que mesmo valores a uma distância considerável de x serão utilizados. Note que se o *kernel* é uma função densidade de probabilidade, então h é o parâmetro de escala no sentido estatístico do termo (Souza, 2008).

A estimativa gerada pelo processo *kernel* é definida como sendo uma soma de "protuberâncias" (em inglês, *bumps*) suaves postas nas observações e que resulta numa aproximação bastante razoável da verdadeira densidade, desde que $K(\psi)$ seja contínua. Alterada a forma funcional de $K(\psi)$, obtemos da Expressão (3.6) uma grande variedade de estimadores. Várias são as funções que podem servir como núcleo, dentre estas estão: gaussiana, Epanechnikov, retangular (uniforme), triangular, *biweight*, *cosine* e *optcosine*. Mostramos na Tabela 3.1 e na Figura 3.2 as expressões analíticas e as formas funcionais (para $h = 1.2$), respectivamente, das sete funções supracitadas.

Segundo Bidurin & Milan (2001), a aplicação do estimador *kernel* depende essencialmente da escolha da função *kernel* e da definição do valor de h. Ainda de acordo com os autores, o primeiro aspecto parece pouco relevante, visto que a mudança de $K(\psi)$ não causa diferenças significativas no ajuste (vide Figura 3.2). No entanto, o segundo aspecto é de grande importância, visto que, se tomarmos o valor de h muito baixo em relação à variação de x, poderemos estar deixando de suavizar a função f e, se tomarmos h muito alto, poderemos estar suavizando em excesso. Em outras palavras, na medida que h se aproxima de 0, a estimativa tende a interpolar as observações e, quando h au-

menta, a curva estimada aproxima-se de uma regressão linear de grau d, o grau da função polinomial utilizada.

Tabela 3.1: Expressões analíticas de funções kernel comumente utilizadas.

Função Kernel	Forma analítica, $K(\psi)$
retangular	$\frac{1}{2}$ para $\|\psi\| < 1$, 0 caso contrário
gaussiana	$2\pi^{-1/2}\exp\left[-\frac{1}{2}(\psi)^2\right]$
triangular	$1 - \|\psi\|$ para $\|\psi\| < 1$, 0 caso contrário
biweight	$\frac{15}{16}(1-\psi^2)^2$ para $\|\psi\| < 1$, 0 caso contrário
Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}\psi^2)/\sqrt{5}$ para $\|\psi\| < \sqrt{5}$, 0 caso contrário
optcosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}\psi\right)$ para $\|\psi\| < 1$, 0 caso contrário
cosine	$1 + \cos(\frac{\pi\psi}{2})$ para $\|\psi\| < 1$, 0 caso contrário

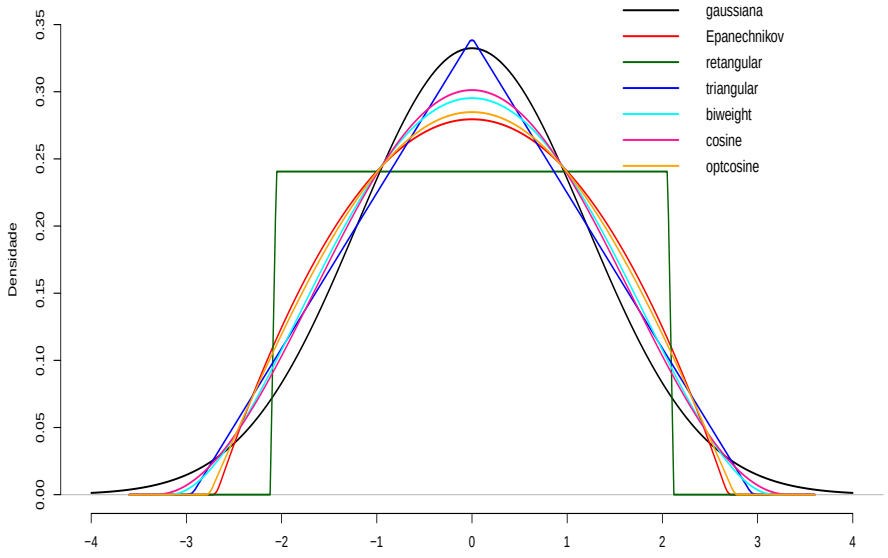

Figura 3.2: Densidades de funções kernel comumente utilizadas com $h = 1.2$.

Pinto (2003) observa que a função de densidade *kernel* estimada é assintoticamente não-viesada quando $h \to 0$. Mas, como geralmente trabalha-se com uma janela diferente

de 0, tem-se um estimador viesado. O viés do estimador é uma função da janela escolhida, sendo menor quando a janela escolhida é pequena. A variância da função de densidade *kernel* também dependerá do tamanho da janela escolhida. Quanto maior for h, menor será a variância. Logo, quando se aumenta o tamanho da janela diminui-se a variância, mas aumenta-se o viés. A escolha do valor "ótimo" para o parâmetro de suavização h é o que discutiremos a seguir.

Escolha da janela h

A escolha do valor ótimo para o parâmetro de suavização é crucial para a obtenção de uma boa estimativa. Podemos dizer que a escolha de h determina a escolha do estimador, pois as estimativas variam consideravelmente em função de h. É comum a utilização de métodos subjetivos, tais como análises gráficas ou aplicação de critérios que estabelecem uma relação entre o parâmetro de suavização e o tamanho da amostra.

Souza (2008) destaca que há uma variedade de técnicas automáticas de seleção de h, geralmente baseadas na minimização do erro quadrático médio (*Mean Quadratic Error* — MQE) da estimação de $f(x)$, dado por

$$\text{MQE}[\hat{f}(x)] = \text{E}[\hat{f}(x) - f(x)]^2, \tag{3.7}$$

também chamado de função perda $\mathcal{L}^2$. Porém, na maioria das aplicações não é possível minimizar (3.7) diretamente, pois $f(x)$ não é conhecida. Note que (3.7) pode ser reescrito de modo a relacionar o vício e a variância de $\hat{f}(x)$:

$$\text{MQE}[\hat{f}(x)] = \{\text{B}[\hat{f}(x)]\}^2 + \text{Var}[\hat{f}(x)], \tag{3.8}$$

em que $\text{B}[\hat{f}(x)] = f(x) - \text{E}[\hat{f}(x)]$ e $\text{Var}[\hat{f}(x)] = \text{E}\{[\hat{f}(x) - \text{E}(\hat{f}(x))]^2\}$ correspondem ao viés e à variância do estimador de $f(x)$, respectivamente. Essa relação mostra que seria válido tolerar um pouco de vício se o resultado for uma grande redução na variância amostral. Na verdade, isso é o que quase sempre acontece, e é a razão fundamental de suavizar os dados para estimar funções.

De acordo com Dias (2001b), o método da validação cruzada generalizada (*Generalized Cross Validation — GCV*) é o mais utilizado na estimação dos parâmetros de suavização seja qual for o método de estimação da função de regressão. A ideia do critério GCV consiste em retirar sucessivamente elementos da amostra e fazer uma estimativa

do ponto retirado, obtendo-se um erro de predição. Procura-se então o conjunto de parâmetros que minimiza esse erro. O critério GCV apresenta a seguinte forma quando a estimação de f é feita pelo método *kernel*:

$$\mathrm{GCV}(h) = \frac{\mathrm{MQE}(h)}{n^{-1}\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}_h)} = \frac{\frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{H}_h)\mathrm{y}\|^2}{\left[\frac{1}{n}\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}_h)^2\right]}, \tag{3.9}$$

em que $\boldsymbol{H}_h = (\boldsymbol{X}_h^\top \boldsymbol{X}_h)^{-1}\boldsymbol{X}_h^\top$. Especificamente, procura-se o valor de h que minimize a função objetivo $\mathrm{GCV}(h)$.

Para Souza (2008), não há um método ótimo para a escolha de h e que seja sempre confiável. Algoritmos de seleção ótima de h continuam sendo objeto de muitos estudos. Em muitos casos práticos o que se faz é escolher dentre uma variedade de valores de h aquele que produz graficamente o melhor resultado.

3.3.2 Suavização por *splines*

O termo *splines* é emprestado de um dispositivo mecânico muito utilizado (antes da computação gráfica) para desenhar secções de cascos de navios e linhas férreas. Os *splines* mecânicos constituíam tiras flexíveis de madeira que eram forçadas a passar por determinados pontos fixos, de forma a atingir a posição desejada (Bowman & Azzalini, 1997). A ideia por trás deste dispositivo era possibilitar a confluência de duas ou mais curvas de maneira suave, passando por pontos pré-fixados e se moldando às diferentes curvaturas do projeto de construção da embarcação ou da ferrovia.

A transposição desta ideia ao ajuste de modelos é feita de forma a termos uma regressão polinomial por "pedaços". Para se trabalhar com o ajustamento e interpolação de funções, a aproximação por polinômios é muito conveniente, uma vez que os polinômios têm muitas propriedades interessantes, dentre estas a de serem funções analíticas,[6] que torna possível calcular as derivadas de qualquer ordem dos polinômios. Entretanto, a necessidade de muitas derivadas da função que está sendo aproximada por polinômios pode ser muito restritiva. Uma maneira de contornar este problema é a utilização de polinômios

[6]Uma função $f :]a, b[\to \mathbb{R}$ é dita ser analítica em $x_0 \in]a, b[$ se existe uma série de potências $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ tal que $f(x)$ seja a soma dessa série para todo x numa vizinhança de x_0, isto é, para todo $x \in]x_0 - \epsilon, x_0 + \epsilon[\subset]a, b[$, com $\epsilon > 0$. Logo, se f é analítica em x_0, f tem derivadas de qualquer ordem numa vizinhança desse ponto e todas as suas derivadas são funções analíticas. Além disto, se $f :]a, b[\to \mathbb{R}$ for analítica em $x_0 \in]a, b[$, então f é a soma da sua série de Taylor numa vizinhança de x_0, ou seja, $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$, para todo $x \in]x_0 - \epsilon, x_0 + \epsilon[\subset]a, b[$.

por partes, pois desta forma pode-se escapar da analiticidade no intervalo inteiro, permitindo descontinuidades das derivadas de ordem mais elevadas em alguns pontos. Essa característica de "pseudo-analiticidade" confere às funções polinomiais por partes, denominadas de *splines*, boas propriedades de aproximação, convergência e estabilidade (Cunha, 2000).

A suavização por *splines* (*smoothing splines*) teve origem com Whittaker (1923). Porém, foi Schoenberg (1964) que, empregando os métodos desenvolvidos por Whittaker, obteve um estimador de suavização por *splines* (Dias, 2001c).

As funções *splines* estão associadas à partição de um intervalo $[a, b]$ do domínio $\mathcal{D}$ de f em que se pretende trabalhar. Uma partição I será definida pelos pontos $x_0, \ldots, x_k$ tais que

$$a = x_0 < x_1 < \cdots < x_{k-1} < x_k = b.$$

Em cada subintervalo $[x_i, x_{i+1}],\ i = 0, \ldots, k$, as *splines* são polinômios de um determinado grau d. Estes "pedaços" de polinômios são colocados convenientemente para que algumas derivadas, de ordem ditada pelo problema, existam em todo o intervalo $[a, b]$. Existe uma relação entre o grau dos "pedaços" dos polinômios e a ordem das derivadas exigidas nos pontos da partição. Assim, algumas restrições devem ser impostas na definição geral das *splines* (Cunha, 2000).

Definição 3.3.1. *A função $s(x)$ é chamada de spline de grau d, com nós (knots) em $\{x_i\}_{i=1}^{k}$, associada a uma partição de $[a, b]$, se:*

- *$s(x)$ é um polinômio de grau d em cada intervalo de subíndices consecutivos $[x_i, x_{i+1}]$;*
- *$s(x)$ tem $d-1$ derivadas contínuas em cada x_i e, portanto, em $[a, b]$.*

O conjunto das funções $\mathcal{S}_d(x_0, \ldots, x_k)$ é um espaço linear e recebe o nome de espaço *spline* (em inglês, *spline space*), em que seus elementos são funções *splines*. Embora diversas configurações sejam possíveis, uma escolha bastante popular são as *splines* cúbicas ($d = 3$), que consistem em uma coleção de polinômios cúbicos com primeira e segunda derivadas contínuas nos nós.

Então, se $b_j(x)$ é a j-ésima base da função[7] que define o seu espaço, $s(x)$ pode ser

[7]Em análise de funções e nas suas aplicações, um espaço funcional pode ser visto como um espaço vetorial de dimensão infinita cujos vetores-base são funções e não vetores. Isto significa que cada função no espaço funcional pode ser representada como uma combinação linear das funções de base.

representada pela expressão

$$s(x) = \sum_{j=1}^{q} b_j(x)\beta_j, \tag{3.10}$$

para algum valor do parâmetro desconhecido β_j. Por exemplo, se s for uma função polinomial de grau 5, sua base é dada por $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$, $b_4(x) = x^3$, $b_5(x) = x^4$ e $b_6(x) = x^5$. Com isto, (3.10) pode ser expressa por

$$s(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5 + x^5\beta_6.$$

3.3.2.1. Penalização pela não suavidade

Um bom ajuste aos dados não é o único objetivo ao se estimar uma curva. Existe um outro objetivo, geralmente conflitante, que é obter uma estimativa que não oscile muito rapidamente. Para dirimir este dilema e representar o problema de estimação de maneira que o compromisso entre esses dois objetivos seja explícito, utiliza-se uma penalização pela não suavidade do ajuste que quantifica quão rapidamente uma curva oscila.

Considere o modelo de regressão proposto em (3.4) e suponha que $\hat{f}(x)$ estime a função $f(x)$. Um critério de bondade de ajuste poderia ser dado pela soma dos quadrados dos resíduos

$$\sum_{i=1}^{n}[y_i - \hat{f}(x_i)]^2. \tag{3.11}$$

Se f é assumida com uma forma irrestrita (conceito não-paramétrico), então o valor da soma acima pode ser reduzido a 0 para alguma curva em particular, e para algum comportamento específico dos dados.

Conforme Rosa & Soler (2004) destacam, somente a soma dos quadrados dos resíduos não é, isoladamente, um bom critério de ajuste, uma vez que estamos tratando com modelos numéricos inicialmente criados para interpolação, e no caso de uma interpolação dos dados o ajuste teria uma bondade de ajuste "perfeita", mas seria pouquíssimo suave.

Acrescentamos, então, um critério de penalização para a falta de suavidade do ajuste a partir das derivadas da função f, proposto por Handscomb (1966), da forma

$$\int_a^b [f^m(x)]^2 dx, \tag{3.12}$$

para a e b tais que $a \leq x_0 \leq \cdots \leq x_k \leq b$ e m é um parâmetro de ordem da derivada da função f que está relacionado com o grau d dos polinômios a serem ajustados, ou seja, f^m denota a m-ésima derivada. No caso dos *splines* cúbicos ($d = 3$ e $m = 2$), penaliza-se a segunda derivada, ou aceleração da curva.

Compondo os critérios (3.11) e (3.12) em uma única equação, temos a forma tradicional de *smoothing splines*, que na verdade é a procura por uma função $f(x)$ com m derivadas contínuas que minimiza a soma de quadrados de resíduos penalizada

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_a^b [f^m(x)]^2 dx, \tag{3.13}$$

sendo que a primeira parcela penaliza a falta de ajuste da função de regressão aos dados e a segunda parcela penaliza a falta de suavidade da função $\hat{f}(x)$ e λ é o parâmetro de suavização que determina o grau de suavidade da estimativa, controlando o quanto andamos na direção da interpolação dos dados ou na direção da suavização excessiva.

Destaca-se que os *splines* gozam de boas propriedades tais como existência, unicidade e flexibilidade, além de serem fáceis de calcular computacionalmente quando comparados a outros métodos de suavização. Em particular, os *splines* cúbicos (vide Eubank, 1994) têm sido amplamente utilizados na estimação não-paramétrica e são bastante empregados para solução do Critério (3.13).

Estudos sobre o comportamento assintótico do estimador obtido utilizando o método de suavização por *splines* podem ser encontrados em Silverman (1984) e Eubank (1988). O primeiro autor mostra ainda que, sob certas condições, a suavização *spline* corresponde aproximadamente à suavização por *kernel* com a janela h dependendo da densidade local dos pontos de observação. Para mais detalhes vide Wegman & Wright (1983) e Härdle (1990).

3.3.2.2. Estimação do parâmetro suavizador λ

De acordo com Souza (2008), existem duas abordagens distintas com relação à escolha do parâmetro de suavização. A primeira abordagem considera a livre escolha do parâmetro de suavização como uma característica importante do procedimento. O que se faz é utilizar diferentes valores e, assim, escolher aquele que, de certa forma, produz a estimativa que causa o "melhor" ajuste visual da curva suavizada aos dados. Isso faz

com que esse método seja subjetivo, porém, muito utilizado na prática. Ele constitui uma ótima opção quando se deseja ajustar uma única curva.

A outra abordagem lida com a necessidade de se ter um procedimento automático para a escolha de λ com base nos dados. Pode-se dizer que, condicionado na escolha do método automático a ser usado, essa é uma forma objetiva de escolha de λ.

Dentre os procedimentos automáticos de escolha do parâmetro de suavização λ, o mais conhecido de todos, semelhantemente ao que se utiliza na estimação da janela h do estimador *kernel* (vide Seção 3.3.1.3), é o critério GCV, que apresenta a seguinte forma quando a estimação de f é feita pelo método *splines*:

$$\text{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}(x_i)]^2}{[1 - \overline{h}_i(\lambda)]^2} = \frac{\frac{1}{n}\|(\boldsymbol{I} - \boldsymbol{H}_\lambda)\mathrm{y}\|^2}{\left[\frac{1}{n}\text{tr}(\boldsymbol{I} - \boldsymbol{H}_\lambda)^2\right]}, \tag{3.14}$$

sendo $||\cdot||$ a norma euclidiana, $\overline{h}(\lambda) = (1/n)\text{tr}(\boldsymbol{H}_\lambda)$ e $\text{tr}(\boldsymbol{H}_\lambda)$ é o traço da matriz $\boldsymbol{H}_\lambda = (\boldsymbol{X}_\lambda^\top \boldsymbol{X}_\lambda)^{-1}\boldsymbol{X}_\lambda^\top$. A escolha do parâmetro de suavização é feita encontrando o valor de λ que minimiza o critério GCV(λ). É importante ainda notar que os métodos de validação cruzada estão bem definidos para um conjunto de pontos $\{x_i\}_{i=1}^n$ distintos e que, portanto, deve-se tomar cuidado na sua implementação para eliminar grupos de pontos não distintos antes de realizar o procedimento de otimização. Para mais detalhes, vide Ramsay & Silverman (2006).

CAPÍTULO 4

Modelos GAMLSS

"As teorias científicas lidam com conceitos, não com a realidade. Embora elas sejam formuladas para corresponder à realidade, esta correspondência é aproximada e a justificativa para todas as conclusões teóricas é baseada em alguma forma de raciocínio indutivo."

(Papoulis, A., em *Probability, Random Variables, and Stochastic Processes*, 1965)

4.1 Introdução

Procedimentos de inferência baseados em suposições equivocadas da distribuição de probabilidade do termo de erro estocástico associadas à adoção de formas funcionais incorretas entre regressando e regressores podem gerar resultados duvidosos e irrealistas, frutos do erro de especificação do modelo. Por este motivo, pesquisadores têm dedicado especial atenção ao desenvolvimento de técnicas estatísticas de modelagem mais flexíveis e menos restritivas como forma de minimizar possíveis fontes de erros de especificação do modelo e aumentar a acurácia das estimativas de quantidades de interesse.

Contudo, esta busca incessante por procedimentos estatísticos "inovadores" não garante a construção de conclusões indubitavelmente certas; ao contrário, conseguem-se apenas explanações coerentes com o conhecimento da época, o qual pode ser modificado no futuro, a exemplo do que se observa com o desenvolvimento histórico dos modelos estatísticos.

Segundo Paula (2004), por muitos anos os modelos normais lineares foram utilizados para descrever a maioria dos fenômenos aleatórios. Mesmo quando o fenômeno sob estudo não apresentava uma resposta para a qual fosse razoável a suposição da normalidade, tentava-se algum tipo de transformação no sentido de alcançar a normalidade procurada. Provavelmente a transformação mais conhecida foi proposta por Box & Cox (1964), a qual transforma o valor observado y positivo em

$$z = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{se} \quad \lambda \neq 0, \\ \log y, & \text{se} \quad \lambda = 0, \end{cases}$$

sendo λ[1] uma constante desconhecida. Acreditava-se que para um único valor de λ a transformação de Box-Cox, quando aplicada a um conjunto de valores observados, produzia normalidade aproximada, constância de variância e também linearidade $\mathrm{E}(Z) = \eta$, em que $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, sendo que $\beta_0, \ldots, \beta_k$ são os parâmetros (coeficientes no modelo de regressão) a serem estimados e $X_1, \ldots, X_k$ são variáveis preditoras conhecidas. No entanto, isso raramente acontece.

Algumas vezes, métodos mais simplificados, como o de mínimos quadrados em dois estágios, foram sugeridos porque outros, como o de máxima verossimilhança com informação limitada,[2] envolviam cálculos complicados e difíceis de serem manipulados. Com os recentes avanços computacionais, tal busca por modelos simplistas não mais se justifica, pelo menos para a maior parte dos problemas (Maddala, 2003).

Em decorrência disto, alguns modelos que exigiam a utilização de esquemas iterativos para a estimação dos parâmetros começaram a ser mais empregados, como o modelo normal não-linear, que assume uma estrutura não-linear para os parâmetros em η, e os modelos não-lineares da família exponencial (Cordeiro & Paula, 1989a e Wei, 1998), que admitem preditores não-lineares nos parâmetros.

Dentre as técnicas de modelagem de regressão univariada, os modelos lineares generalizados (*Generalized Linear Models* — GLM) e os modelos aditivos generalizados (*Generalized Additive Models* — GAM) ocupam lugar de destaque na literatura (Nelder & Wedderburn, 1972 e Hastie & Tibshirani, 1990, respectivamente). Ambos os modelos

[1] O parâmetro λ da transformação de Box-Cox é um parâmetro a ser estimado a partir dos dados da amostra e não deve ser confundido com o parâmetro λ de suavização da Equação (3.13).

[2] Uma abordagem detalhada sobre o método dos mínimos quadrados em dois estágios e método da máxima verossimilhança com informação limitada é apresentada em Anderson (2005).

assumem que a distribuição da variável resposta pertence à família exponencial e sua média μ é modelada a partir das variáveis explanatórias. Adicionalmente, $\mathrm{Var}(y) = \phi v(\mu)$, em que $v(\mu)$ é a "função de variância" que depende de μ e ϕ é um parâmetro de dispersão, que na maioria das vezes é suposto ser constante para todas as observações. Note que numa distribuição da família exponencial a simetria e curtose de y são, em geral, funções de μ e ϕ. Assim, nos GLM e GAM a variância, simetria e curtose não são modeladas explicitamente em termos das variáveis explanatórias, mas implicitamente através da dependência com o parâmetro μ. Uma das extensões mais importantes dos GLM foi apresentada por Wedderburn (1974), os modelos de quase-verossimilhança, que expandem a ideia dos GLM para situações mais gerais incluindo dados correlacionados. Liang e Zeger (1986) estendem os modelos de quase-verossimilhança propondo as equações de estimação generalizadas que permitem o estudo de variáveis aleatórias correlacionadas não-gaussianas.

Outra importante classe são os modelos lineares mistos de efeitos aleatórios, que fornecem uma estrutura bastante geral para a modelagem de dados dependentes derivados de estudos longitudinais, espaciais ou hierárquicos. Aqui, presume-se normalidade para a distribuição condicional de y dados os efeitos aleatórios e não é possível modelar explicitamente a simetria e a curtose. Muitos desses resultados são discutidos no livro do McCulloch & Searle (2001).

O modelo linear generalizado misto(*Generalized Linear Mixed Model* — GLMM) é uma combinação do GLM com o modelo linear misto mediante a inserção de um termo (quase sempre normal) de efeito aleatório no preditor linear para a média de um GLM. Procedimentos bayesianos para ajuste de modelos GLMM empregando o algoritmo de maximização da esperança (*Expectation Maximization* — EM) e o método de Monte Carlo baseado em cadeias de Markov estão detalhados em McCulloch (1997) e Zerger & Karin (1991). Lin e Zhang (1999) exemplificam o uso de modelos aditivos generalizados mistos (*Generalized Additive Mixed Model* — GAMM), enquanto Fahrmeir & Lang (2001) abordam a modelagem GAMM utilizando inferência bayesiana e Fahrmeir & Tutz (2001) discutem procedimentos alternativos de estimação via GLMM e GAMM. Cumpre registrar que os GLMM e GAMM, embora mais flexíveis do que os GLM e GAM, também pressupõem uma distribuição condicional da família exponencial para y e, à exceção da

média, raramente permitem a modelagem de outros parâmetros da distribuição da variável resposta em função das covariáveis. Aqui, o ajuste do modelo geralmente depende da utilização do método de Monte Carlo baseado em cadeias de Markov e da verossimilhança (por exemplo, quadratura gaussiana) integrada (distribuição marginal), resultando em procedimentos computacionalmente intensivos, principalmente quando se trabalha com conjuntos de dados extensos e se faz necessária a análise comparativa de diversos modelos alternativos. Vários estudos foram desenvolvidos e propostos visando ao ajustamento de GLMM (vide Breslow & Clayton, 1993; Breslow & Lin, 1995 e Lee & Nelder, 1996, 2001a,b). Uma outra alternativa de aproximação é utilizar máxima verossimilhança não-paramétrica baseada em misturas finitas; vide Aitkin (1999).

Objetivando superar algumas das limitações associadas aos modelos acima descritos, Rigby & Stasinopoulos (2005) propuseram uma nova classe de modelos estatísticos de regressão (semi)paramétricos, denominada de modelos aditivos generalizados para posição, escala e forma (GAMLSS). São paramétricos no sentido de que uma distribuição paramétrica é requerida para a variável resposta e ao mesmo tempo semiparamétricos por permitirem que a modelagem dos parâmetros da distribuição e das funções das variáveis explanatórias possa envolver o uso de funções de suavização não-paramétricas.

Nos modelos GAMLSS, a premissa de que a variável resposta pertence à família exponencial é relaxada e substituída por uma família de distribuições mais geral $\mathcal{D}$. A variável resposta y tem distribuição $D(y|\mu, \sigma, \nu, \tau)$, em que $D \in \mathcal{D}$ pode ser qualquer distribuição (incluindo distribuições contínuas com assimetria ou curtose acentuadas e distribuições discretas). Além disso, a parte sistemática do modelo é amplificada para permitir a modelagem não apenas da média (ou posição), mas de todos os parâmetros da distribuição condicional de y, sejam através de funções paramétricas ou não-paramétricas (de suavização) das variáveis explanatórias e/ou termos de efeitos aleatórios.

Os modelos GAMLSS são adequados, sobretudo, para modelagem da variável resposta que não segue uma distribuição da família exponencial (por exemplo, leptocúrtica ou platicúrtica e/ou com assimetria positiva ou negativa) e nos casos em que o regressando exibe heterogeneidade (por exemplo, quando a escala ou a forma da distribuição da variável resposta mudam com as variáveis explanatórias) ou esteja relacionado a dados de contagem com sobredispersão.

Um aspecto relevante e que deve ser considerado como uma vantagem dessa abordagem diz respeito à facilidade de acesso a programas de livre distribuição, como o ambiente de programação R. A estrutura de modelagem GAMLSS está implementada em uma série de pacotes no R (ver Seção 5.4.3) e permite ajustar mais de 50 distribuições diferentes, entre elas a distribuição exponencial potência de Box-Cox (Rigby & Stasinopoulos, 2004) utilizada pela Organização Mundial de Saúde para a construção das curvas de crescimento padrão mundial (*WHO Multicentre Growth Reference Study Group*). Os modelos GAMLSS também possibilitam o ajuste de versões truncadas, censuradas ou de misturas finitas das distribuições e sua aplicação já pode ser observada em diversas áreas do conhecimento, como na medicina (ver Beyerlein *et al.*, 2008) e economia (ver Ferreira, 2008), entre outras.

Nas seções subsequentes deste capítulo iremos descrever detalhadamente os modelos GAMLSS no que tange aos aspectos de estimação, inferência e diagnóstico. Acrescenta-se que os resultados e teoria aqui expostos estão fortemente embasados em Rigby & Stasinopoulos (2001, 2005, 2006 e 2007) e Akantziliotou *et al.* (2002, 2006).

4.2 Modelos aditivos generalizados para posição, escala e forma (GAMLSS)

4.2.1 Definição

Na estrutura de regressão GAMLSS os p parâmetros $\boldsymbol{\theta}^\top = (\theta_1, \theta_2, \ldots, \theta_p)$ de uma função densidade de probabilidade $f(y|\boldsymbol{\theta})$ são modelados utilizando termos aditivos. Aqui, presume-se que para $i = 1, 2, \ldots, n$ as observações y_i são independentes e condicionais a $\boldsymbol{\theta^i}$, com função densidade de probabilidade $f(y_i|\boldsymbol{\theta^i})$, onde $\boldsymbol{\theta^{i\top}} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})$ é um vetor de p parâmetros relacionado às variáveis explanatórias e efeitos aleatórios. Destaca-se que quando os valores assumidos pelas covariáveis são estocásticos ou as observações y_i dependem de seus valores passados, então $f(y_i|\boldsymbol{\theta^i})$ é interpretada como sendo condicional a estes valores.

Seja $\mathtt{y}^\top = (y_1, y_2, \ldots, y_n)$ o vetor de observações da variável resposta. Considere ainda, para $k = 1, 2, \ldots, p$, uma função de ligação monótona $g_k(\cdot)$ relacionando o k-ésimo parâmetro $\boldsymbol{\theta}_k$ às variáveis explanatórias e efeitos aleatórios por meio de um modelo aditivo

dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{4.1}$$

em que $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores $n \times 1$, por exemplo $\boldsymbol{\theta}_k^\top = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})$, $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'_k k})$ é um vetor de parâmetros de tamanho J'_k e $\boldsymbol{X}_k$ e $\boldsymbol{Z}_{jk}$ são matrizes de planejamento (covariáveis) fixas, conhecidas e de ordens $n \times J'_k$ e $n \times q_{jk}$, respectivamente. Já $\boldsymbol{\gamma}_{jk}$ é uma variável aleatória q_{jk}-dimensional. O Modelo (4.1) é denominado de GAMLSS (Rigby & Stasinopoulos, 2005).

Os vetores $\boldsymbol{\gamma}_{jk}$, para $j = 1, 2, \ldots, J_k$, podem ser manipulados e combinados em um único vetor $\boldsymbol{\gamma}_k$ e numa única matriz de covariáveis $\boldsymbol{Z}_k$. Entretanto, a formulação proposta em (4.1) é mais apropriada por dois motivos: facilita o uso dos algoritmos de retroajuste (*backfitting*) e permite que combinações de diferentes tipos de termos aditivos e/ou de efeitos aleatórios sejam facilmente incorporadas no modelo (Rigby & Stasinopoulos, 2005).

No caso em que $J_k = 0$, não há termos aditivos associados aos parâmetros da distribuição. Então, (4.1) se reduz a um modelo linear completamente paramétrico dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k. \tag{4.2}$$

Se $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, em que $\boldsymbol{I}_n$ é uma matriz identidade de ordem $n \times n$, e $\boldsymbol{\gamma}_{jk} = \boldsymbol{h}_{jk} = h_{jk}(\mathtt{x}_{jk})$ para todas as combinações de j e k no Modelo (4.1), temos

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{h}_{jk}(\mathtt{x}_{jk}), \tag{4.3}$$

em que $\mathtt{x}_{jk}$, para $j = 1, 2, \ldots, J_k$ e $k = 1, 2, \ldots, p$, são vetores de tamanho n. A função h_{jk} é uma função desconhecida da variável explanatória X_{jk} e $\boldsymbol{h}_{jk} = h_{jk}(\mathtt{x}_{jk})$ é um vetor que avalia a função h_{jk} em $\mathtt{x}_{jk}$. Neste caso, assume-se que os vetores $\mathtt{x}_{jk}$ são conhecidos e o modelo apresentado na Equação (4.3) é denominado de GAMLSS aditivo semiparamétrico linear. O modelo resultante em (4.3) é um caso especial do modelo (4.1) e pode conter termos paramétricos, não-paramétricos e de efeitos aleatórios (Rigby & Stasinopoulos, 2005).

O Modelo (4.3) pode ser estendido para permitir a inclusão de termos não-lineares na modelagem dos k parâmetros da distribuição, na forma

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathtt{x}_{jk}), \tag{4.4}$$

em que h_k para $k = 1, 2, \ldots, p$ são funções não-lineares e $\boldsymbol{X}_k$ é uma matriz de covariáveis conhecida de ordem $n \times J_k''$. O Modelo (4.4) é designado de GAMLSS aditivo semiparamétrico não-linear. Se $J_k = 0$, então o Modelo (4.4) se reduz a um GAMLSS paramétrico não-linear, expresso por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k). \tag{4.5}$$

Finalmente, se $h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) = \boldsymbol{X}_k^\top \boldsymbol{\beta}_k$, para $i = 1, 2, \ldots, n$ e $k = 1, 2, \ldots, p$, então, (4.5) se reduz ao modelo paramétrico linear (4.2). Note que alguns termos de $h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k)$ podem ser lineares, o que resulta num modelo GAMLSS com a combinação de termos paramétricos lineares e não-lineares.

Em muitas situações práticas são requeridos no máximo quatro parâmetros ($p = 4$), usualmente caracterizados pela posição (μ), escala (σ), assimetria (ν) e curtose (τ). Enquanto os dois primeiros parâmetros populacionais θ_1 e θ_2 no Modelo (4.1), aqui denotados por μ e σ, são referidos na literatura por parâmetros de posição (ou locação) e escala, respectivamente, os dois últimos $\nu = \theta_3$ e $\tau = \theta_4$ são denominados de parâmetros de forma. Com isto, temos os seguintes modelos:

$$\begin{array}{ll} \text{Parâmetros de posição e escala} & \left\{ \begin{array}{l} g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ \\ g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \end{array} \right. \\ \\ \text{Parâmetros de forma} & \left\{ \begin{array}{l} g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\ \\ g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}. \end{array} \right. \end{array} \tag{4.6}$$

Acrescenta-se que os *pacotes* disponíveis e implementados no `R` referentes à estrutura GAMLSS permitem que as funções aditivas h_{jk} admitam *splines* cúbicos, *splines* penalizados, polinômios fracionários, polinômios potência não-lineares em que o parâmetro potência assume qualquer valor real (por exemplo, $b_0 + b_1 x^{p_1} + b_2 x^{p_2}$), curvas *loess*, termos

de coeficientes variáveis, entre outras. Desta forma, qualquer combinação destas funções pode ser incluída no modelo para cada μ, σ, ν ou τ (ver Seções 4.5.2 e 4.5.3).

Conforme destacam Akantziliotou *et al.* (2002), a estrutura GAMLSS pode ser aplicada aos parâmetros de qualquer distribuição populacional e generalizada para modelagem de mais de quatro parâmetros da distribuição. Além disto, Rigby & Stasinopoulos (2005) salientam que a classe de modelos GAMLSS (4.1) é mais geral do que os GLM, GAM, GLMM ou GAMM, no sentido de que a distribuição da variável resposta não se restringe à família exponencial e todos os parâmetros (não apenas a média) são modelados em termos de efeitos fixos e aleatórios.

4.3 Estimação

Dois aspectos são fundamentais no ajuste de componentes aditivos incorporados na estrutura GAMLSS: o algoritmo *backfitting*[3] (ver Seção **??**) e o fato de que as penalidades quadráticas na função de verossimilhança resultam da premissa de que os efeitos aleatórios no preditor linear seguem distribuição normal. Com isto, o processo de estimação do modelo utilizará, basicamente, matrizes de encolhimento (alisamento) associadas à estrutura do algoritmo *backfitting*, conforme apresentaremos a seguir.

Admitamos que no Modelo (4.1) os termos de efeitos aleatórios $\boldsymbol{\gamma}_{jk}$ sejam independentes e tenham distribuição normal com $\gamma_{jk} \sim N_{q_{jk}}(\mathbf{0},\, \boldsymbol{G}_{jk}^{-1})$, em que $\boldsymbol{G}_{jk}^{-1}$ é a inversa (generalizada) de ordem $q_{jk} \times q_{jk}$ da matriz simétrica $\boldsymbol{G}_{jk} = \boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$. Esta matriz pode depender de um vetor de hiperparâmetros $\boldsymbol{\lambda}_{jk}$ e, sendo $\boldsymbol{G}_{jk}$ singular, $\boldsymbol{\gamma}_{jk}$ especifica uma função de densidade imprópria proporcional a $\exp\left(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^{\top}\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}\right)$. A fim de simplificar a notação ao longo desta dissertação, iremos nos referir a $\boldsymbol{G}_{jk}$ ao invés de $\boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$, embora a dependência de $\boldsymbol{G}_{jk}$ aos hiperparâmetros $\boldsymbol{\lambda}_{jk}$ continue existindo.

A premissa de independência entre diferentes vetores $\boldsymbol{\gamma}_{jk}$ de efeitos aleatórios é fundamental no contexto da estrutura GAMLSS. Se para um particular k, dois ou mais vetores de efeitos aleatórios não forem independentes, pode-se combiná-los em um único

[3] A ideia central do algoritmo *backfitting* é de um processo de ajuste iterativo que busca minimizar uma função de perda (normalmente um erro quadrático) em relação à cada uma das funções (uma das variáveis preditoras de cada vez) até a convergência. Hastie & Tibshirani (1990) provaram que este algoritmo atinge uma solução única independente de valores iniciais para funções de ajuste simétricas, como as funções *splines*, discutidas na Seção 3.3. Para mais detalhes sobre o algoritmo *backfitting* ver Hastie & Tibshirani (1990) e Härdle *et al.* (2004).

vetor de efeitos aleatórios. Analogamente, as correspondentes matrizes de covariáveis $\boldsymbol{Z}_{jk}$ também podem ser transformadas numa matriz única, satisfazendo a condição de independência (Rigby & Stasinopoulos, 2005).

Rigby & Stasinopoulos (2005) mostraram, utilizando argumentos bayesianos empíricos, que o método da estimação máximo *a posteriori* (*Maximum a Posteriori* (MAP) *Estimation*; vide Berger, 1985) para o vetor de parâmetros $\boldsymbol{\beta}_k$ e termos de efeitos aleatórios $\boldsymbol{\gamma}_{jk}$ (com valores fixos do parâmetro de suavização ou hiperparâmetros $\boldsymbol{\lambda}_{jk}$), para $j = 1, 2, \ldots, J_k$ e $k = 1, 2, \ldots, p$, é equivalente à estimação por máxima verossimilhança penalizada.

Desta forma, para valores fixados de $\boldsymbol{\lambda}_{jk}, \boldsymbol{\beta}_k$ e $\boldsymbol{\gamma}_{jk}$ são estimados na estrutura de regressão GAMLSS por meio da maximização da função de verossimilhança penalizada, ℓ_p , dada por

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\boldsymbol{\gamma}_{jk}^{\top}\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}, \tag{4.7}$$

em que $\ell = \sum_{i=1}^{n}\log\{f(y_i|\boldsymbol{\theta}^i)\}$ é a função de log-verossimilhança dos dados condicionais a $\boldsymbol{\theta}^i$, para $i = 1, 2, \ldots, n$. Isto é equivalente a maximizar a verossimilhança estendida ou hierárquica definida por

$$\ell_h = \ell_p + \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\{\log|\boldsymbol{G}_{jk}| - q_{jk}\log(2\pi)\}$$

(vide Lee & Nelder, 1996 e Pawitan, 2001).

Rigby & Stasinopoulos (2005) ressaltam que a maximização de ℓ_p pode ser obtida com a implementação de um algoritmo *backfitting* e demonstram que a maximização de (4.7), aplicada aos resíduos parciais $\boldsymbol{\epsilon}_{jk}$ (vide nota de rodapé[4]) para atualizar a estimativa do preditor aditivo $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$, conduz à matriz de encolhimento (alisamento) $\boldsymbol{S}_{jk}$, dada por

$$\boldsymbol{S}_{jk} = \boldsymbol{Z}_{jk}(\boldsymbol{Z}_{jk}^{\top}\boldsymbol{W}_{kk}\boldsymbol{Z}_{jk} + \boldsymbol{G}_{jk})^{-1}\boldsymbol{Z}_{jk}^{\top}\boldsymbol{W}_{kk}, \tag{4.8}$$

para $j = 1, 2, \ldots, J_k$ e $k = 1, 2, \ldots, p$, em que $\boldsymbol{W}_{kk}$ é uma matriz diagonal de pesos iterativos.

[4] $\boldsymbol{\epsilon}_{jk}^{(r)} = \boldsymbol{Z}_k^{(r)} - \boldsymbol{X}_k\boldsymbol{\beta}_k^{(r+1)} - \sum_{t=1,t\neq j}^{J_k}\boldsymbol{Z}_{tk}\boldsymbol{\gamma}_{tk}^{(r+1)} - \boldsymbol{W}_{kk}^{(r)^{-1}}\sum_{s=1,s\neq k}^{p}\boldsymbol{W}_{ks}^{(r)}(\boldsymbol{\eta}_s^{(r+1)} - \boldsymbol{\eta}_s^{(r)})$, em que r é o ciclo da iteração, isto é, $r = 1, 2\ldots$ até a convergência; $\boldsymbol{W}_{ks} = -\frac{\partial^2\ell}{\partial\boldsymbol{\eta}_k\boldsymbol{\eta}_s^{\top}}$, $\boldsymbol{Z}_k^{(r)} = \boldsymbol{\eta}_k^{(r)} + \boldsymbol{W}_{kk}^{(r)^{-1}}\boldsymbol{u}_k^{(r)}$ e $u_k = \frac{\partial\ell}{\partial\boldsymbol{\eta}_k}$ é a variável dependente ajustada. Para mais detalhes vide Rigby & Stasinopoulos (2005).

Diferentes formas de $\boldsymbol{Z}_{jk}$ e $\boldsymbol{G}_{jk}$ correspondem a diferentes tipos de termos aditivos no preditor linear $\boldsymbol{\eta}_k$ para $k = 1, 2, \ldots, p$. Em relação aos termos de efeitos aleatórios, $\boldsymbol{G}_{jk}$ é geralmente uma matriz de ordem pequena, considerando que para um termo de suavização *spline* cúbico temos $\boldsymbol{\gamma}_{jk} = \boldsymbol{h}_{jk}$, $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$ e $\boldsymbol{G}_{jk} = \lambda_{jk}\boldsymbol{K}_{jk}$, em que $\boldsymbol{K}_{jk}$ é uma matriz estruturada. Em qualquer um dos casos fica fácil a atualização de $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$.

4.4 Algoritmos de maximização

No `R`, dois algoritmos podem ser utilizados para a maximização da função de verossimilhança penalizada dada em (4.7). O primeiro, algoritmo `CG`, é uma generalização do algoritmo de Cole & Green (1992) e usa a primeira derivada — e o valor esperado ou aproximado das derivadas de segunda ordem e das derivadas cruzadas — da função de log-verossimilhança em relação aos parâmetros da distribuição (por exemplo, $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$ para uma distribuição com quatro parâmetros). Entretanto, para muitas funções de densidade de probabilidade, $f(y|\boldsymbol{\theta})$, os parâmetros $\boldsymbol{\theta}$ são ortogonais, ou seja, os valores esperados das derivadas cruzadas da função de log-verossimilhança são iguais a 0 (por exemplo, modelos de posição e escala e modelos da família de dispersão). Neste caso, é utilizado um algoritmo mais simples e que não utiliza o valor esperado das derivadas cruzadas, conhecido como `RS`, que é uma generalização do algoritmo usado por Rigby & Stasinopoulos (1996a, b) no ajuste da média e da dispersão de modelos aditivos. Destaca-se que o algoritmo `RS` não é um caso especial do algoritmo `CG`, uma vez que no algoritmo `RS` a matriz diagonal de pesos $\boldsymbol{W}_{kk}$ é avaliada (isto é, atualizada) "dentro" de cada ajuste do parâmetro $\boldsymbol{\theta}_k$, enquanto que no `CG` todas as matrizes de pesos $\boldsymbol{W}_{ks}$, para $k = 1, 2, \ldots, p$ e $s = 1, 2, \ldots, p$, são avaliadas depois do ajuste de todos os parâmetros $\boldsymbol{\theta}_k$, para $k = 1, 2, \ldots, p$. Acrescenta-se que no caso totalmente paramétrico, o algoritmo `CG` corresponde ao método escore de Fisher.

O objetivo dos algoritmos é maximizar a função de verossimilhança penalizada ℓ_p, dada por (4.7), para hiperparâmetros (λ) fixados. Nos modelos completamente paramétricos, como (4.2) ou (4.4), os algoritmos maximizam a função de verossimilhança ℓ. A escolha dos algoritmos é efetuada com a opção `method` da função `gamlss( )`, em que uma combinação dos dois algoritmos também é permitida. Mais detalhes sobre os algoritmos `CG` e `RS` podem ser obtidos em Rigby & Stasinopoulos (2005).

4.5 Preditor linear

4.5.1 Termos paramétricos

No modelo GAMLSS (4.1), os preditores lineares $\boldsymbol{\eta}_k$, para $k = 1, 2, \ldots, p$, incluem componentes paramétricos, $\boldsymbol{X}_k\boldsymbol{\beta}_k$, e aditivos, $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$, para $j = 1, 2, \ldots, J_k$. O componente paramétrico pode conter termos lineares e de interação, bem como fatores, polinômios, polinômios fracionários (Royston & Altman, 1994) e polinômios segmentados (com nós fixados) para as variáveis exploratórias.

Acrescenta-se ainda que parâmetros não-lineares podem ser incorporados à estrutura GAMLSS (4.1) pelo método perfilado ou pelo método derivado.[5] No primeiro método, a estimação dos parâmetros é realizada mediante a maximização da função de verossimilhança perfilada. No último método, as derivadas do preditor $\boldsymbol{\eta}_k$ em relação aos parâmetros não-lineares são incluídas na matriz de covariáveis $\boldsymbol{X}_k$ do algoritmo de ajustamento (vide, por exemplo, Benjamin *et al.*, 2003).

4.5.2 Termos aditivos

Os componentes aditivos $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$ na Equação (4.1) podem modelar uma variedade de termos, tais como de suavização e efeitos aleatórios, bem como termos que são úteis na análise de séries temporais, como passeios aleatórios. Diferentes termos aditivos podem ser integrados à estrutura GAMLSS, conforme apresentaremos a seguir. Antes, porém, esclarecemos que, no intuito de simplificar a exposição e notação dos tópicos adiante, iremos omitir (onde for apropriado) os subscritos j e k nos vetores e matrizes.

4.5.2.1. *Splines* cúbicos

A utilização de *splines* cúbicos no Modelo (4.3) presume que as funções $h(t)$ são contínuas e duas vezes diferenciáveis e que a maximização da função de verossimilhança penalizada (vide Equação (4.7)) está sujeita aos termos de penalização da forma $\lambda \int_{-\infty}^{\infty} h''(t)^2 dt$. De acordo com Reinsch (1967), as funções de maximização $h(t)$ são todas *splines* cúbicas e por isso podem ser expressas como combinações lineares de suas funções bases *splines* cúbicas $B_i(t)$, para $i = 1, 2, \ldots, n$ (vide de Boor, 1978 e Schumaker, 1993), ou seja, $h(t) = \sum_{i=1}^{n} \delta_i B_i(t)$.

[5] Mais detalhes sobre os métodos derivado e perfilado podem ser obtidos em Bates & Watts (1988).

Considere ainda que $\boldsymbol{h} = h(\mathbf{x})$ é um vetor com as avaliações da função $h(t)$ dos valores de $\mathbf{x}$ que a variável explanatória X assume (os quais admitimos serem distintos para simplificação da exposição). Seja $\boldsymbol{N}$ uma matriz não-singular de ordem $n \times n$, em que as colunas contêm os vetores de avaliação das funções $B_i(t)$, para $i = 1, 2, \ldots, n$, em $\mathbf{x}$. Assim, $\boldsymbol{h}$ pode ser expresso por meio de um vetor (coeficiente) $\boldsymbol{\delta}$, resultado da combinação linear das colunas de $\boldsymbol{N}$, por $\boldsymbol{h} = \boldsymbol{N}\boldsymbol{\delta}$.

Seja $\boldsymbol{\Omega}$ uma matriz $n \times n$ dos produtos internos das segundas derivadas das funções bases *splines* cúbicas para os (r, s)-ésimos registros, dada por

$$\Omega_{rs} = \int B_r^{''}(t) B_s^{''}(t) dt.$$

A penalidade é dada pela forma quadrática

$$Q(\boldsymbol{h}) = \lambda \int_{-\infty}^{\infty} h^{''}(t)^2 dt = \lambda \delta^{\top} \boldsymbol{\Omega} \delta = \lambda \boldsymbol{h}^{\top} \boldsymbol{N}^{-\top} \boldsymbol{\Omega} \boldsymbol{N}^{-1} \boldsymbol{h} = \lambda \boldsymbol{h}^{\top} \boldsymbol{K} \boldsymbol{h},$$

em que $\boldsymbol{K} = \boldsymbol{N}^{-\top} \boldsymbol{\Omega} \boldsymbol{N}^{-1}$ é uma matriz de penalidade conhecida que depende apenas dos valores do vetor explanatório $\mathbf{x}$ (Hastie & Tibshirani, 1990). A forma precisa da matriz $\boldsymbol{K}$ pode ser obtida em Green & Silverman (1994).

Para que a estrutura de regressão seja formulada segundo um modelo GAMLSS (4.1) de efeitos aleatórios é necessário que $\boldsymbol{\gamma} = \boldsymbol{h}$, $\boldsymbol{Z} = \boldsymbol{I}_n, \boldsymbol{K} = \boldsymbol{N}^{-\top} \boldsymbol{\Omega} \boldsymbol{N}^{-1}$ e $\boldsymbol{G} = \lambda \boldsymbol{K}$, de forma que $\boldsymbol{h} \sim N_n(0, \lambda^{-1} \boldsymbol{K}^{-})$, em que $\boldsymbol{K}^{-}$ é uma inversa generalizada de $\boldsymbol{K}$, resulte numa densidade parcialmente imprópria (Silverman, 1985). Ou seja, assume-se completa indeterminação *a priori* sobre a constante e as funções lineares, assim como reduz-se a incerteza acerca das funções de ordem superiores (Verbyla *et al.*, 1999).

Acrescenta-se ainda que suavizadores *splines* cúbicos podem ser combinados em diferentes covariáveis e resultar no modelo aditivo (Hastie & Tibshirani, 1990).

4.5.2.2. *Splines* de penalização

Suavizadores em que o número de funções bases é menor que o número de observações mas seus coeficientes de regressão são penalizados são chamados de *splines* penalizados ou *P-splines*.

Eilers & Marx (1996) utilizaram um conjunto de q funções bases *B-splines* na variável explanatória X (cujas avaliações nos valores $\mathbf{x}$ de X são as colunas da matriz de covariáveis

$\boldsymbol{Z}$, de ordem $n \times q$, na Equação (4.1) e sugeriram o uso de um número razoável (algo entre 20 e 40) de nós igualmente espaçados, em que os segmentos *splines* se unem e garantem a flexibilidade no ajustamento das curvas. Contudo, estes autores também impuseram penalidades nos parâmetros γ das funções bases B-*splines* a fim de assegurar a suavidade do ajuste. Em essência, Eilers & Marx (1996) assumiram que $\boldsymbol{\gamma}\boldsymbol{D}_r \sim N_{n-r}(0, \lambda^{-1}\boldsymbol{I})$, em que $\boldsymbol{D}_r$ é uma matriz $(q-r) \times q$ que fornece r-ésimas diferenças do vetor q-dimensional $\boldsymbol{\gamma}$.

Uma aproximação semelhante foi proposta por Wood (2001), que utilizou uma base polinomial Hermite cúbica ao invés de *B-spline*. Wood (2000) forneceu uma maneira de estimar os hiperparâmetros empregando validação cruzada, o que corresponde, na estrutura de regressão GAMLSS (4.1), a admitir $\boldsymbol{G} = \lambda\boldsymbol{K}$, de modo que $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \lambda^{-1}\boldsymbol{K}^{-})$, em que $\boldsymbol{K} = \boldsymbol{D}_r^{\top}\boldsymbol{D}_r$ (Rigby & Stasinopoulos, 2005).

4.5.2.3. Outras suavizações

Além dos *splines* cúbicos e dos *splines* penalizados, outros suavizadores podem ser usados como termos aditivos, por exemplo, a implementação no `R` da estrutura GAMLSS permite incorporar suavizadores de regressão local, como o *loess*[6] e os polinômios fracionários.

4.5.2.4. Termos de coeficientes variáveis

Os modelos de coeficientes variáveis (Hastie & Tibshirani, 1993) permitem um tipo particular de interação entre suavizadores aditivos e variáveis contínuas ou fatores. Estes modelos são da forma $\boldsymbol{s}h(\mathbf{x})$, em que $\boldsymbol{s}$ e $\mathbf{x}$ são vetores de valores fixos das variáveis explanatórias S e X.

É possível mostrar que os termos de coeficientes variáveis podem ser incorporados facilmente ao algoritmo de ajuste dos modelos GAMLSS mediante o uso da matriz de alisamento na forma da Equação (4.8), com $\boldsymbol{Z} = \boldsymbol{I}_n$, $\boldsymbol{K} = \boldsymbol{N}^{-\top}\boldsymbol{\Omega}\boldsymbol{N}^{-1}$ e $\boldsymbol{G} = \lambda\boldsymbol{K}$. Entretanto, é necessário assumir que os valores de S são distintos, com uma matriz diagonal de pesos iterativos $\boldsymbol{W}$ multiplicada pela matriz diagonal com elementos $s_1^2, s_2^2, \ldots, s_n^2$ e os resíduos parciais ϵ_i divididos por s_i, para $i = 1, 2, \ldots, n$.

[6] Uma referência sobre o suavizador *loess* é Cleveland *et al.* (1993).

4.5.2.5. Termos de efeitos aleatórios específicos

Lee & Nelder (2001b) consideraram vários termos de efeitos aleatórios no preditor da média nos modelos GLMM. Dentre os termos de efeitos aleatórios específicos que podem ser incorporados na estrutura GAMLSS (4.1) destacam-se os seguintes:

1. Termo de sobredispersão: no Modelo (4.1) considere $\boldsymbol{Z} = \boldsymbol{I}_n$ e $\boldsymbol{\gamma} \sim N_n(\boldsymbol{0}, \lambda^{-1}\boldsymbol{I}_n)$, o que fornece um termo de sobredispersão para cada observação no preditor.

2. Termo de efeito aleatório de um fator: no Modelo (4.1) considere que $\boldsymbol{Z}$ é uma matriz de incidência das covariáveis, de ordem $n \times q$ (para um fator de nível q), definida pelos elementos $z_{it} = 1$, se a i-ésima observação pertence ao t-ésimo nível do fator, e $z_{it} = 0$ caso contrário, com $\boldsymbol{\gamma} \sim N_q(\boldsymbol{0}, \lambda^{-1}\boldsymbol{I}_n)$, o que resulta no modelo de efeitos aleatórios de um fator.

3. Termos de efeitos aleatórios correlacionados: no Modelo (4.1), desde que $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{G}^-)$, em que $\boldsymbol{G}^-$ é a inversa generalizada de $\boldsymbol{G}$, estruturas correlacionadas podem ser aplicadas aos efeitos aleatórios mediante escolha adequada da matriz $\boldsymbol{G}$, por exemplo, passeios aleatórios de primeira ou segunda ordem, autorregressivos de primeira ou segunda ordem, modelos de decaimento exponencial (dependência temporal) e de correlação simétrica composta.

Acrescente-se ainda que existem diversas combinações úteis entre efeitos aleatórios e *splines*, como o agrupamento dos coeficientes (de covariáveis) na mesma covariável.

4.5.3 Combinações de termos

Quaisquer combinações de termos aditivos e paramétricos podem ser aplicadas (em um ou mais preditores dos parâmetros de posição, escala ou forma) para gerar modelos e termos ainda mais complexos.

4.5.3.1. Desenho de medidas repetidas longitudinal de dois níveis

Considere um planejamento experimental em dois níveis com indivíduos no primeiro nível, em que y_{ij} para $i = 1, 2, \ldots, n_j$ são medidas repetidas do segundo nível no indivíduo j, para $j = 1, 2, \ldots, J$. Seja $\boldsymbol{\eta}$ um vetor de valores preditos, associado aos valores de cada

indivíduo, ou seja, $\boldsymbol{\eta}^\top = (\boldsymbol{\eta}_1^\top, \boldsymbol{\eta}_2^\top, \ldots, \boldsymbol{\eta}_j^\top)$ de dimensão $n = \sum_{j=1}^{J} n_j$. Seja $\boldsymbol{Z}_j$ uma matriz de covariáveis $n \times q_j$ (para efeitos aleatórios γ_j e indivíduo j) que não possui zeros nas n_j linhas correspondentes ao indivíduo j, sendo que os $\boldsymbol{\gamma}_j$ são independentes e $\boldsymbol{\gamma}_j \sim N_{qj}(0, \boldsymbol{G}_j^{-1})$, para $j = 1, 2, \ldots, J$. Acrescenta-se que as matrizes $\boldsymbol{Z}_j$ e os efeitos aleatórios, para $j = 1, 2, \ldots, J$, podem ser combinados em uma única matriz de covariáveis $\boldsymbol{Z}$ e em apenas um vetor aleatório $\boldsymbol{\gamma}$.

4.5.3.2. Termos de efeitos aleatórios correlacionados com medidas repetidas

Na Seção 4.5.3.1, considere $q_j = n_j$ e que a submatriz (de elementos não nulos) $\boldsymbol{Z}_j$ seja uma matriz identidade I_{nj}, para $j = 1, 2, \ldots, J$. Com isto, obtêm-se diversas estruturas de covariância ou correlação nos efeitos aleatórios das medidas repetidas e que podem ser especificadas mediante escolha adequada de matrizes $\boldsymbol{G}_j$, como destacado no item (3) da Seção 4.5.2.5.

4.6 Famílias específicas

4.6.1 Generalidades

A função densidade de probabilidade populacional $f(y|\boldsymbol{\theta})$ no Modelo (4.1) pode pertencer a uma família de distribuições bastante geral sem que haja a obrigatoriedade de uma forma explícita para a distribuição condicional da variável resposta y.

No `R`, a única restrição que a implementação do modelo GAMLSS exige na especificação da distribuição de y é que a função $f(y|\theta)$ e sua primeira derivada (e opcionalmente o valor esperado das derivadas de segunda ordem e as derivadas cruzadas) com relação a cada um dos parâmetros de $\boldsymbol{\theta}$ sejam calculáveis. Embora as expressões das derivadas sejam preferíveis, derivadas numéricas também podem ser obtidas e usadas, ainda que neste último caso ocorra uma redução na velocidade de processamento dos dados.

As Tabelas 4.1 e 4.2 exibem algumas famílias de distribuições contínuas e discretas, respectivamente, que se encontram implementadas no `R`.

Tabela 4.1: Exemplos de distribuições contínuas implementadas à estrutura GAMLSS e disponíveis no R.

		Função de ligação			
Distribuição	Nomenclatura	μ	σ	ν	τ
beta	BE()	logit	logit	–	–
beta inflacionada (em zero)	BEOI()	logit	log	logit	–
beta inflacionada (em um)	BEZI()	logit	log	logit	–
beta inflacionada (em 0 e 1)	BEINF()	logit	logit	log	log
Box-Cox (Cole & Green)	BCCG()	identidade	log	identidade	–
Box-Cox exponencial potência	BCPE()	identidade	log	identidade	log
Box-Cox-t	BCT()	identidade	log	identidade	log
exponencial	EXP()	log	–	–	–
exponencial gaussiana	exGAUS()	identidade	log	log	–
exponencial poder	PE()	identidade	log	log	–
família t	TF()	identidade	log	log	–
gama	GA()	log	log	–	–
gama generalizada	GG()	log	log	identidade	–
gaussiana inversa	IG()	log	log	–	–
gaussiana inversa ajustada a zero	ZAIG()	log	log	logit	–
gaussiana inversa generalizada	GIG()	log	log	identidade	–
Gumbel	GU()	identidade	log	–	–
Gumbel reversa	RG()	identidade	log	–	–
log normal	LOGNO()	log	log	–	–
log normal (Box-Cox)	LNO()	log	log	fixed	–
logística	LO()	identidade	log	–	–
normal	NO()	identidade	log	–	–
shash	SHASH()	identidade	log	log	log
Weibull	WEI()	log	log	–	–
Weibull (reparametrizada)	WEI3()	log	log	–	–

Tabela 4.2: Exemplos de distribuições discretas implementadas à estrutura GAMLSS e disponíveis no R.

		Função de ligação		
Distribuição	Nomenclatura	μ	σ	ν
beta binomial	BB()	logit	log	–
binomial	BI()	logit	–	–
binomial negativa tipo I	NBI()	log	log	–
binomial negativa tipo II	NBII()	log	log	–
Delaporte	DEL()	log	log	logit
Gaussiana inversa Poisson	PIG()	log	–	–
Poisson	PO()	log	–	–
Poisson inflacionada de zeros	ZIP()	log	logit	–
Sichel	SI()	log	log	identidade
Sichel (reparametrizada)	SICHEL()	log	log	identidade

Nas seções seguintes utilizaremos a notação

$$y \sim \mathcal{D}\{g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, \ldots, g_p(\theta_p) = t_p\}$$

para identificar exclusivamente um modelo GAMLSS, em que $\mathcal{D}$ é a distribuição da variável resposta, $\theta_1, \ldots, \theta_p$ são os parâmetros de $\mathcal{D}$ (conforme abreviado nas Tabelas 4.1

e 4.2), $g_1, \ldots, g_p$ são as funções de ligação e $t_1, \ldots, t_p$ são as fórmulas dos modelos para os termos explanatórios e/ou efeitos aleatórios nos preditores $\eta_1, \ldots, \eta_p$, respectivamente. Por exemplo,

$$y \sim \text{PE}\{\mu = cs(x, 5),\ \log(\sigma) = x,\ \log(\nu) = 1\}$$

é um modelo GAMLSS em que a variável resposta y tem distribuição exponencial potência (PE); o parâmetro de posição μ é modelado usando uma função de ligação identidade e suavizadores *splines* cúbicos com cinco graus de liberdade efetivos em x, ou seja, $\text{cs}(x, 5)$; o parâmetro de escala σ é modelado a partir de um modelo log-linear em x e o parâmetro ν admitido como constante e igual a 1 (mas na escala logarítmica).

4.6.2 Distribuições específicas

Muitas famílias de distribuições contínuas podem ser definidas assumindo uma variável transformada z, obtida a partir de y, cuja distribuição resultante seja simples e bem conhecida.

A família **Box-Cox normal** para $y > 0$, reparametrizada de Box & Cox (1964), denotada por $\text{BCN}(\mu, \sigma, \nu)$ e utilizada por Cole & Green (1992), assume que z tem uma distribuição normal padrão $N(0, 1)$, com média 0 e variância 1, em que

$$z = \begin{cases} \frac{1}{\sigma\nu}\left\{\left(\frac{y}{u}\right)^{\nu} - 1\right\}, & \text{se } \nu \neq 0, \\ \frac{1}{\sigma}\log(\frac{y}{\mu}), & \text{se } \nu = 0. \end{cases} \tag{4.9}$$

Cole & Green (1992) foram os primeiros a modelar todos os três parâmetros de uma distribuição como funções de suavização não-paramétricas de uma única variável explanatória.

A família **gama generalizada** para $y > 0$, parametrizada por Lopatatzidis & Green (2000) e denotada por $\text{GG}(\mu, \sigma, \nu)$, assume que z tem distribuição gama $\text{GA}(1, \sigma^2\nu^2)$ com média 1 e variância $\sigma^2\nu^2$, com $z = (\frac{y}{\mu})^{\nu}$, para $\nu > 0$.

A família **exponencial potência** para $-\infty < y < \infty$, utilizada por Nelson (1991), denotada por $\text{PE}(\mu, \sigma, \nu)$, é uma reparametrização daquela desenvolvida por Box & Tiao (1973) e assume que z tem distribuição gama $\text{GA}(1, \nu)$ com média 1 e variância ν, sendo que

$$z = \frac{\nu}{2}\left|\frac{y - \mu}{\sigma c(\nu)}\right|^{\nu}$$

e

$$c(\nu) = \left\{ 2^{-2/\nu} \frac{\Gamma(1/\nu)}{\Gamma(3/\nu)} \right\}^{1/2},$$

$\nu > 0$. Nesta reparametrização, μ e σ são a média e o desvio-padrão de y, respectivamente.

A família ***t* de Student** para $-\infty < y < \infty$, denotada por $\text{TF}(\mu, \sigma, \nu)$, assume que z tem distribuição t padrão com ν graus de liberdade, em que $z = (y - \mu)/\sigma$.

Os quatro parâmetros da família ***t* de Box-Cox** para $y > 0$, denotada por BCT (μ, σ, ν, τ), são definidos assumindo que a variável z dada na Expressão (4.9) tem distribuição t padrão com τ graus de liberdade; vide Rigby & Stasinopoulos (2004a).

A família **exponencial potência de Box-Cox** para $y > 0$, denotada por BCPE (μ, σ, ν, τ), é definida assumindo que a variável z dada na Expressão (4.9) tem distribuição exponencial potência padrão; vide Rigby & Stasinopoulos (2004b). Essa distribuição é útil para a modelagem de dados contínuos sujeitos a assimetria (positiva ou negativa) combinada com (lepto ou plati) curtose.

4.7 Seleção do modelo

4.7.1 Modelagem estatística

Considere que $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ representa um modelo GAMLSS, em que $\mathcal{D}$ especifica a distribuição da variável resposta, $\mathcal{G}$ o conjunto das funções de ligação $(g_1, \ldots, g_p)$ para os parâmetros $(\theta_1, \ldots, \theta_p)$, $\mathcal{T}$ define o conjunto de termos preditores $(t_1, \ldots, t_p)$ para os preditores $(\eta_1, \ldots, \eta_p)$ e $\boldsymbol{\lambda}$ explicita o conjunto de hiperparâmetros.

Para um conjunto de dados específico, o processo de construção de um modelo GAMLSS consiste em comparar diversos modelos concorrentes onde diferentes combinações dos componentes $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ foram utilizadas. Como podemos perceber, há uma grande quantidade de possibilidades a serem avaliadas e testadas, o que sugere, em certa medida, um mecanismo de tentativa e erro envolvido na escolha do modelo "certo" para a análise empírica.[7] Parece bastante razoável procurar por um modelo que capte a essência do fenômeno estudado e que ratifique a relevância lógica ou teórica das variáveis explanatórias em relação à variável independente. Aqui, cabe destacarmos que

[7] No entanto, deve-se evitar o que é conhecido como "garimpagem de dados", isto é, a procura indiscriminada e arbitrária por modelos que se ajustem bem aos dados.

um grande número de covariáveis significa um alto grau de complexidade na interpretação do modelo. Por outro lado, um modelo com um pequeno número de covariáveis pode ter uma interpretação fácil mas pode se ajustar "pobremente" aos dados. Neste sentido, devemos procurar um modelo intermediário entre o minimal, que possui o menor número de termos necessários para o ajustamento, e o maximal, ou seja, aquele com o maior número de variáveis independentes que se pretende trabalhar.

Assim como todas as inferências científicas, a determinação da adequabilidade de qualquer modelo depende substancialmente do problema de interesse e requer conhecimentos específicos do pesquisador.

4.7.2 Seleção do modelo, inferências e diagnósticos

Na estrutura de regressão GAMLSS paramétrica, cada modelo $\mathcal{M}$ da forma (4.2) pode ser avaliado a partir de seu desvio global ajustado (*Global Deviance* — GD), dado por $\text{GD} = -2\ell(\hat{\boldsymbol{\theta}})$, em que $\ell(\hat{\theta}) = \sum_{i=1}^{n} \ell(\hat{\boldsymbol{\theta}}^i)$. Dois modelos GAMLSS paramétricos encaixados e concorrentes à predição, $\mathcal{M}_0$ e $\mathcal{M}_1$, com desvios globais ajustados, GD_0 e GD_1, e graus de liberdade dos erros, df_{e0} e df_{e1}, respectivamente, podem ser comparados usando o teste da razão de verossimilhanças generalizado com estatística de teste $\Lambda = \text{GD}_0 - \text{GD}_1$, que tem distribuição assintótica χ^2 sob $\mathcal{M}_0$ com $d = df_{e0} - df_{e1}$ graus de liberdade (dado que as condições de regularidade[8] sejam satisfeitas). Para cada modelo $\mathcal{M}$ o número de graus de liberdade dos erros para os parâmetros df_e é definido por $df_e = n - \sum_{k=1}^{p} df_{\theta k}$, em que $df_{\theta k}$ são os graus de liberdade utilizados no modelo preditor para o parâmetro θ_k, para $k = 1, \ldots, p$.

Na comparação de modelos GAMLSS não-encaixados (incluindo modelos com termos de suavização), o critério de informação de Akaike generalizado (Generalized Akaike Information Criterion — GAIC; Akaike, 1983) pode ser utilizado para penalizar sobreajustes (em inglês, *overfitting*). Isto é obtido adicionando aos desvios globais ajustados uma penalidade fixa # para cada grau de liberdade efetivo que é usado no modelo, ou seja, $\text{GAIC}(\#) = \text{GD} + \#df$, onde df denota o total de graus de liberdade efetivos utilizados no modelo e GD é o desvio global ajustado. O modelo com o menor valor do critério GAIC(#) é o selecionado. A sensibilidade do modelo selecionado frente à escolha

[8] Para uma listagem das condições de regularidade vide, por exemplo, Sen & Singer (1993).

da penalidade # também pode ser investigada.

O critério de informação de Akaike (*Akaike Information Criterion* — AIC; Akaike, 1974) e o critério bayesiano de Schwarz (*Schwarz Bayesian Criterion* — SBC; Schwarz, 1978) são casos especiais do critério GAIC(#), e correspondem a $\# = 2$ e $\# = \log(n)$, respectivamente. Acrescenta-se que os dois critérios, AIC e SBC, permitem comparar modelos não-encaixados e penalizam aqueles com maiores números de parâmetros. Embora no critério SBC esta penalidade seja mais rigorosa e favoreça modelos mais parcimoniosos, ambos os critérios possuem fundamentação assintótica.

Os parâmetros dos modelos GAMLSS com hiperparâmetros $\boldsymbol{\lambda}$ podem ser estimados a partir dos seguintes métodos: (i) minimização do critério GAIC perfilado sobre $\boldsymbol{\lambda}$; (ii) minimização do critério de validação cruzada generalizado perfilado sobre $\boldsymbol{\lambda}$; (iii) maximização da função densidade marginal aproximada (ou verossimilhança marginal perfilada) para $\boldsymbol{\lambda}$ mediante o uso da aproximação de Laplace ou (iv) maximização da verossimilhança marginal para $\boldsymbol{\lambda}$ por meio do uso de um algoritmo EM aproximado. Fixados os hiperparâmetros $\boldsymbol{\lambda}$, utiliza-se um algoritmo *backfitting* para se proceder à estimação máximo *a posteriori* (MAP) de $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Mais detalhes sobre os métodos apresentados podem ser obtidos em Rigby & Stasinopoulos (2005).

Para testar se um parâmetro específico do preditor de efeito fixo é diferente de 0, um teste χ^2 é empregado, comparando a mudança no desvio global Λ para modelos paramétricos (ou a mudança no desvio da aproximação marginal, eliminando os efeitos aleatórios, para os modelos de efeitos aleatórios) quando o parâmetro é atribuído 0 com um valor crítico χ^2. A função de verossimilhança perfilada (marginal) para parâmetros em modelos de efeitos fixos pode ser utilizada para a construção de intervalos de confiança. Os testes mencionados acima e os intervalos de confiança são para quaisquer hiperparâmetros fixados em valores selecionados.

Uma aproximação alternativa, que é apropriada para conjunto de dados extensos, é "dividir" a análise em três etapas: treinamento, validação e teste do conjunto de dados (vide Ripley, 1996 e Hastie *et al.*, 2001). No treinamento, os dados são utilizados para o ajuste do modelo a partir da minimização do GD, na validação, os dados servem para seleção do modelo também via minimização do GD e na fase de teste do conjunto de dados são feitas avaliações do poder preditivo do modelo escolhido (mais uma vez com

base no GD).

Os resíduos (dos quantis aleatórios normalizados) de Dunn & Smyth (1996) são usados para checar a adequabilidade de cada $\mathcal{M}$ e, em particular, a distribuição do componente $\mathcal{D}$. Estes resíduos são dados por $\hat{r_i} = \mathbf{\Phi}^{-1}(u_i)$, em que $\mathbf{\Phi}^{-1}$ é a inversa da função de distribuição acumulada (*Cumulative Distribution Function* — CDF) de uma normal padrão e $u_i = F(y_i|\hat{\boldsymbol{\theta}}^i)$ se y_i é uma observação de uma resposta contínua. Considera-se ainda u_i um valor aleatório de uma distribuição uniforme no intervalo $[F(y_i - 1|\hat{\boldsymbol{\theta}}^i), [F(y_i|\hat{\boldsymbol{\theta}}^i)]$ se y_i é uma observação de uma resposta inteira discreta, em que $F(y|\boldsymbol{\theta})$ é a função de distribuição de $\mathcal{D}$. Para respostas contínuas censuradas a direita, u_i é definido como um valor aleatório de uma distribuição uniforme no intervalo $[F(y_i|\hat{\boldsymbol{\theta}}^i), 1]$. Note que, quando a aleatorização é utilizada, muitos conjuntos aleatórios de resíduos devem ser estudados antes de uma decisão acerca da adequabilidade do modelo $\mathcal{M}$ adotado. Para as distribuições contínuas, os verdadeiros resíduos r_i seguem distribuição normal padrão quando o modelo está corretamente especificado.

Outro aspecto importante dos modelos GAMLSS diz respeito à estimação centílica. Conforme destacado, os resíduos quantílicos são computados facilmente quando é fornecida a CDF de y e, neste caso, a estimação centílica pode ser feita sempre que a inversa da CDF pode ser obtida. Isto se aplica às distribuições contínuas da Tabela 4.1 que podem ser transformadas em distribuições-padrão simples, enquanto que para as distribuições discretas, a CDF e a inversa da CDF podem ser computadas numericamente, se necessário.

CAPÍTULO 5

Análise de dados: modelos GAMLSS a serviço da Engenharia de Avaliações

"Quando o Senhor criou o mundo, Ele deve ter pensado consigo mesmo: "Se Eu fizer tudo previsível, os seres humanos, que Eu dotei de bons cérebros, sem dúvida irão aprender a predizer tudo, e, por causa disso, não terão motivos para fazer nada, porque eles reconhecerão que o futuro é totalmente determinado e não pode ser influenciado por nenhuma ação humana. Por outro lado, se Eu fizer tudo imprevisível, eles irão gradualmente descobrir que não há nenhuma base racional para qualquer decisão e, como no primeiro caso, eles não terão motivo para fazer nada. Nenhum destes esquemas teria sentido. Eu preciso, portanto, criar uma mistura dos dois: deixar algumas coisas serem previsíveis, e outras imprevisíveis. Eles terão, então, entre muitas outras coisas a tarefa de descobrir o que é o quê."

(Schumacher, E. F., em *Small is Beautiful: Economics as if People Mattered*, 1973)

Este capítulo objetiva ilustrar as técnicas descritas no ajuste de modelos GAMLSS a partir da estimação empírica da equação de preços hedônicos para terrenos urbanos situados em Aracaju, Sergipe. Acrescenta-se que, para o mesmo conjunto de dados, os resultados são comparados com aqueles obtidos mediante aplicação do modelo normal de regressão linear clássico e dos modelos lineares generalizados.

Neste estudo foram percorridas 4 (quatro) fases interrelacionadas, a saber: (i) Coleta de dados; (ii) Análise exploratória de dados; (iii) Especificação e estimação dos modelos; (iv) Seleção do modelo.

5.1 Coleta de dados

O conjunto de dados utilizado é composto de 2109 (duas mil cento e nove) observações de terrenos urbanos nus (sem benfeitorias edificadas)[1] situados na cidade de Aracaju-SE e são provenientes de duas fontes: (i) coleta pelo autor deste trabalho junto a empresas imobiliárias, corretores autônomos, anúncios em jornais e percorrendo a região em busca de informações sobre terrenos em oferta ou negociados; (ii) cessão do Departamento de Cadastro Imobiliário da Prefeitura de Aracaju. Acrescenta-se que os dados são relativos aos anos de 2005, 2006 e 2007, porém, não são dados de séries temporais, visto que cada terreno i, $i = 1, \ldots, n$, foi observado em apenas um dos anos j, $j = 2005, 2006, 2007$. Destaca-se que todos os terrenos que compõem a amostra foram georeferenciados em relação ao *South American Datum*[2] e tiveram suas posições geográficas (latitude, longitude) projetadas no Sistema Universal Transverso de Mercartor (UTM — *Universal Transversa de Mercator*).[3]

5.2 Análise exploratória de dados

5.2.1 A cidade de Aracaju

Aracaju é um município brasileiro e capital do estado de Sergipe. Localiza-se no litoral sergipano e limita-se com os municípios de São Cristóvão, Barra dos Coqueiros, Nossa Senhora do Socorro e Itaporanga d'Ajuda. O topônimo "Aracaju" deriva da expressão indígena *ará acaiú*, que em tupi-guarani significa "cajueiro dos papagaios".

Fundada em 1855 para abrigar a capital da Província, até então localizada em São Cristóvão, Aracaju foi a segunda capital brasileira planejada. O centro do poder político-administrativo, atual Praça Fausto Cardoso, foi o ponto de partida para o crescimento da cidade e todas as ruas foram construídas como um tabuleiro de xadrez — ruas retas e quarteirões quadrados — desembocando no Rio Sergipe. Sua construção foi um

[1] Terminologia própria da área de Engenharia Civil para se referir a obras ou serviços realizados num bem e que não podem ser retirados sem destruição, fratura ou dano.

[2] Datum, do latim *dado*, detalhe, pormenor. Em cartografia, refere-se ao modelo matemático teórico da representação da superfície da Terra ao nível do mar para uso na geodésia e navegação. O *South American Datum* (SAD) é o sistema geodésico regional para a América do Sul.

[3] Projeção cartográfica cilíndrica do esferóide terrestre em 60 cilindros secantes à supefície da Terra ao longo de meridianos em zonas múltiplas de 6 graus de longitude e estendendo-se de 80 graus de latitude sul a 84 graus de latitude norte.

desafio à engenharia, face à sua localização numa área dominada por pântanos e charcos. Até então, as cidades existentes antes do século XVII adaptavam-se às respectivas condições topográficas naturais, estabelecendo uma irregularidade no panorama urbano. O engenheiro Sebastião Basílio Pirro contrapôs a essa irregularidade e Aracaju foi uma das primeiras cidades no Brasil a ter essa tendência geométrica. Uma visão parcial da cidade pode ser observada na Figura 5.1, onde percebe-se grande parte da região centro-norte da cidade banhada pelo estuário do Rio Sergipe.

Figura 5.1: Vista aérea da cidade de Aracaju.

Aracaju é bastante quente durante a maior parte do ano e a temperatura média é de 26°C. As chuvas se concentram entre os meses de março e agosto e a precipitação média anual é de 1.590 mm. No que diz respeito à pedologia, é constituída por depósitos marinhos de areia quartzosa e podzólico vermelho e amarelo. No que se refere à hidrografia, a cidade é banhada pelos rios Sergipe, Vaza Barris, Rio do Sal, Poxim, Rio Pitanga e Canal de Santa Maria. Quanto à vegetação, é predominantemente composta de higrófilos (campos de várzeas e manguezais).

Nas zonas mais próximas ao rio Sergipe (bairros Salgado Filho, Grageru, 13 de Julho,

São José, entre outros) existia uma área de manguezal constantemente inundada. Hoje, a área de manguezal está coberta por concreto e é onde localiza-se a área mais nobre da cidade, com enorme concentração de prédios, que por muitos anos possuíam gabarito[4] de 12 andares. Com a aprovação do "novo plano diretor", essa limitação subiu para 23 andares. A vegetação original e o mangue, que ficavam principalmente às margens do rio Sergipe, foram quase que completamente soterrados.

A orla de Aracaju possui aproximadamente 35 km de extensão e oferece belíssimas praias que chamam a atenção pelo mar limpo, dunas de areias bancas, coqueirais, lagos, pela temperatura da água, sempre morna, e pela pouca profundidade. As praias mais frequentadas são Atalaia, Aruana, Robalo, Náufragos e Mosqueiro (na rodovia Airton Sena), Hawaizinho e Praia dos Artistas. À beira-mar, estão os hotéis e as casas de veraneio, com exceção de bairros como Atalaia e Coroa do Meio, que possuem uma grande densidade demográfica. Os prédios baixos facilitam a circulação de ar pela cidade e, ao contrário do que acontece nas capitais litorâneas, a zona mais rica da capital está às margens do rio Sergipe, assim como o Centro.

O relevo plano é propício à prática do ciclismo, sendo o uso da bicicleta como meio de transporte bastante incentivado pela Prefeitura, que nos últimos anos construiu mais de 50 km de ciclovias. A política de ampliação da rede cicloviária tem ajudado a diminuir os congestionamentos, além de evitar a sobrecarga do sistema de transporte público. Existem algumas grandes ciclovias na cidade. As mais antigas são da avenida Augusto Franco, avenida Beira Mar e, mais recentemente, avenida São Paulo (em direção aos bairros mais periféricos), e da praia de Atalaia.

Aracaju faz parte da região de Produção Nordeste da Petrobrás, possuindo indústrias no setor têxtil e confecções, artigos de couro, sabão, cal, extração de petróleo, além de um grande potencial turístico. Os serviços, a indústria e o comércio são a base da economia aracajuana. Em 2005, o Produto Interno Bruto (PIB) do município chegou a R$ 5.021 bilhões e o PIB per capita a R$ 10.071,00, o que rendeu a 13ª colocação entre todas as capitais do país e o segundo lugar na região Nordeste, conforme levantamento do Instituto Brasileiro de Geografia e Estatística (IBGE).

A população da cidade cresceu muito desde que foi fundada, em 1855. O primeiro

[4]Número máximo de pisos (pavimentos) numa edificação permitidos pela legislação.

levantamento de que se tem notícia data de 1872, quando foram contabilizados 9.559 moradores. De lá para cá, os números evoluíram da seguinte forma: 16.336 (1890); 21.132 (1900); 37.440 (1920); 59.031 (1940); 78.364 (1950); 115.713 (1960); 183.670 (1970); 293.100 (1980); 402.341 (1991); 425.726 (1996) e 461.534 (2000). Mais recentemente, em 2007, o IBGE contabilizou 520.303 habitantes, distribuídos em 37 bairros e uma zona de expansão urbana, perfazendo uma área total de 174 km^2. De acordo com o cadastro imobiliário da Prefeitura Municipal de Aracaju, a cidade possuía, em 2006, aproximadamente 180.000 unidades imobiliárias, 5.000 quadras e 115.000 lotes.

O crescimento da cidade de Aracaju tem tomado todas as direções, tanto em expansão física horizontal com a formação de novas periferias, como a verticalização, que constitui símbolo de *status* para a classe mais abastada, que originalmente vivia nas áreas familiares do centro. Depois de décadas de dominação e saturação da região central, agora a forte tendência é a invasão da zona sul pelas classes média, média alta e altíssima. No outro lado da cidade, os bairros situados nos extremos sul, sudoeste, oeste e noroeste (Mosqueiro, Santa Maria, Capucho, Lamarão, Soledade, entre outros) têm sido ocupados pela parcela da população de menor poder aquisitivo.

Nos últimos dez anos, Aracaju tem vivenciado um aumento na procura de residências e uma supervalorização nos terrenos e imóveis já construídos. Essa é uma das razões para a elevação do custo médio de construção na capital. Em seis anos, o preço do metro quadrado dos imóveis à venda na capital aumentou cerca de 300% em média e se tornou um dos mais caros do Nordeste, conforme dados da Associação de Dirigentes de Empresas do Mercado Imobiliário (ADEMI).

5.2.2 Descrição da amostra

A amostra utilizada para a estimação da equação de preços hedônicos[5] contém, além do período, informações sobre as características físicas dos terrenos (área, frente, topografia, infraestrutura (pavimentação) e posição na quadra), locacionais (bairro, coordenadas geográficas (latitude, longitude), coeficiente de aproveitamento e tipo de via na qual está localizado o imóvel) e econômicas (natureza da informação que gerou a ob-

[5]Para simplificação da linguagem empregada ao longo desta dissertação, daqui em diante, salvo menção em contrário, sempre que citarmos a expressão "equação de preços hedônicos" estaremos nos referindo à "equação de preços hedônicos de terrenos urbanos em Aracaju-SE".

servação, renda média do chefe de família do setor censitário[6] onde situa-se o imóvel e valor do terreno). A seguir, discriminamos as características de cada variável e que tipo de informação foi registrada. Neste sentido, temos:

- ANO (ANO): variável qualitativa ordinal que identifica o ano em que a informação foi obtida;

- ÁREA (AR): variável quantitativa contínua, medida em m^2 (metro quadrado), que concerne à projeção num plano horizontal da superfície do terreno examinado;

- FRENTE (FR): variável quantitativa contínua, também denominada de "testada" e medida em m (metro), que diz respeito à projeção da frente real sobre a perpendicular a uma das divisas do lote, quando ambas são oblíquas no mesmo sentido, ou à corda no caso de frentes curvas;

- TOPOGRAFIA (TO): variável qualitativa nominal que denota as conformações topográficas do imóvel. Classifica-se em "plano" se o terreno possui aclive inferior a 10% ou declive inferior a 5%, e em "acidentado", caso contrário;

- PAVIMENTAÇÃO (PA): variável qualitativa nominal que indica a presença ou ausência de pavimentação (em concreto, asfáltica ou granítica) na via principal em que se localiza a frente preponderante do terreno;

- SITUAÇÃO (SI): variável qualitativa nominal empregada para discernir a disposição do terreno na quadra. Classifica-se em lote de "esquina" ou "meio";

- BAIRRO (BAIRRO): variável qualitativa nominal referente ao nome do bairro onde o terreno observado está situado;

- LATITUDE (LAT) e LONGITUDE (LONG): variáveis quantitativas contínuas correspondentes à posição geográfica do imóvel no ponto $z = (\text{LAT}, \text{LONG})$, em que LAT e LONG são as coordenadas medidas em UTM;

[6] Os setores censitários são unidades territoriais definidas pelo IBGE para orientar a distribuição espacial da população, sendo mais de 200.000 em todo o Brasil. Obedecem a critérios de operacionalização da coleta de dados, de tal maneira que abranjam uma área que possa ser percorrida por um único recenseador em um mês e que possua em torno de 250 a 350 domicílios (em áreas urbanas).

- COEFICIENTE DE APROVEITAMENTO (CA): variável quantitativa discreta referente a um número que, multiplicado pela área do terreno, indica a quantidade máxima de metros quadrados que podem ser construídos em um lote, somando-se as áreas de todos os pavimentos. Por exemplo, se dispomos de um lote retangular medindo 24×30 m (área total = 720 m^2) e CA = 2, então podemos construir 1440 m^2 ($720 \times 2 = 1440$). Se a taxa de ocupação[7] do terreno for de 50%, necessitaríamos de 4 pavimentos (cada um com 360 m^2) para distribuir a área edificada (vide Figura 5.2). O CA é definido a partir do plano diretor de desenvolvimento urbano de Aracaju.

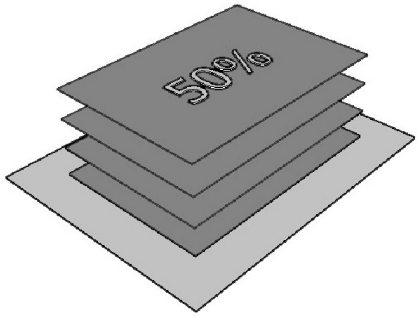

Figura 5.2: Exemplo de distribuição da área edificada em um lote de 24×30 m com CA=2 e taxa de ocupação de 50%.

- VIA (VIA): variável qualitativa ordinal utilizada para diferenciar a posição do imóvel em relação ao logradouro em que se situa. Classifica-se em "via principal", "via secundária" ou "via terciária/superior", conforme importância da via pública no contexto da região;

- NATUREZA DA INFORMAÇÃO (NI): variável qualitativa nominal que define se o dado coletado é oriundo de "oferta", "transação" ou "ITBI";

- SETOR (ST): variável *proxy*[8] quantitativa discreta de macrolocalização para distinguir o nível socioeconômico dos diversos bairros da cidade, representada pela renda

[7] A taxa de ocupação é a relação percentual entre a projeção da edificação e a área do terreno. Ou seja, ela representa a porcentagem do terreno sobre o qual há edificação.

[8] *Proxy* é uma variável tomada como medida aproximada de uma outra variável para a qual não se tem informações. Ou ainda, variável utilizada para substituir outra de difícil mensuração e que se presume guardar com ela relação de pertinência.

média do chefe da família, em salários mínimos, divulgada pelo censo do IBGE (2000). Neste caso, a renda do bairro servirá como *proxy* para outras características, tais como as amenidades urbanas;[9]

- PREÇO UNITÁRIO (PU): variável quantitativa contínua que assume valores estritamente positivos e corresponde ao valor do terreno dividido pela sua área, medida em R\$/m^2 (reais por metro quadrado).

A fim de facilitar a análise exploratória dos dados, classificamos as variáveis em quatro grupos: (i)variáveis quantitativas contínuas; (ii) variáveis quantitativas discretas; (iii) variáveis qualitativas nominais; (iv) variáveis qualitativas ordinais.

5.2.2.1. Variáveis quantitativas contínuas

Na Figura 5.3 apresentamos os gráficos box-plot (também denotados na literatura de *gráficos de caixa*) das variáveis PU, AR e FR, enquanto que na Tabela 5.1 mostramos um resumo de algumas medidas de posição e dispersão destas variáveis. Verificamos por meio dos gráfico box-plot que PU se distribui de forma assimétrica à direita e que há uma considerável quantidade de observações atípicas associada a uma alta dispersão dos dados. Estas características da variável PU podem ser ratificadas mediante inspeção de seu histograma constante na Figura 5.4. Já na Tabela 5.1 observamos que PU abrange um expressivo intervalo de valores (entre R\$ 2.36/m^2 e R\$ 800.00/m^2), bem como evidencia que cerca de 75% dos terrenos observados têm preços unitários inferiores a R\$ 82.82/m^2.

Embora tenham sido identificadas 263 observações atípicas mediante inspeção do gráfico box-plot de AR (vide Figura 5.3), constatamos que as discrepâncias não estão relacionadas a erros de mensuração, mas à elevada magnitude e dispersão da própria variável. Além disso, percebemos que AR varia de 41.00 m^2 a 91.780 m^2, isto é, o maior terreno é 1912 vezes superior ao menor, em área. Em se tratando da variável FR, notamos pelo gráfico de box-plot (vide Figura 5.3) que há uma acentuada variabilidade entre os dados, revelada também pela amplitude total (= 513.40 m) registrada na Tabela 5.1. Ou seja, o menor terreno é cerca de 198 vezes menor que o maior terreno observado (em relação à frente).

[9]Entende-se por *amenidades urbanas* um conjunto de características específicas de uma localidade com contribuição positiva ou negativa para a satisfação dos indivíduos (por exemplo, oferta de entretenimento, segurança, área verde, entre outras).

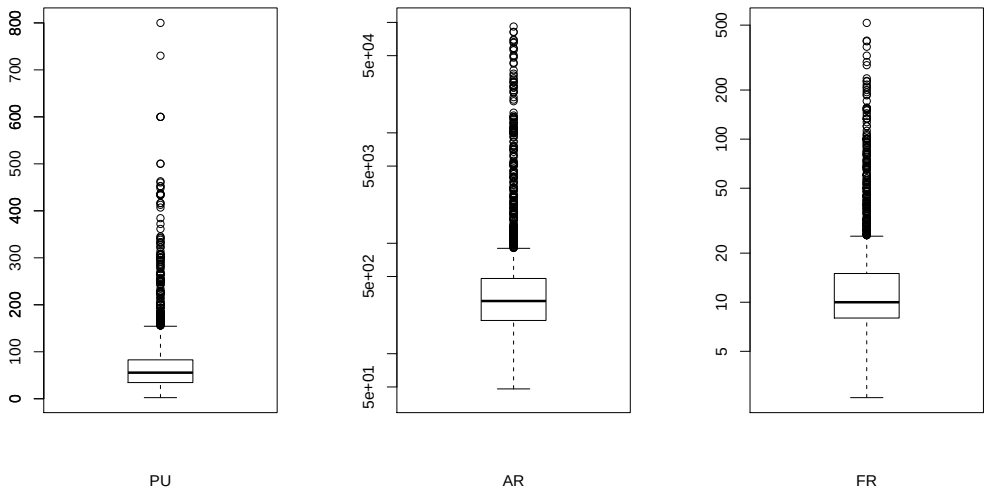

Figura 5.3: Gráficos box-plot das variáveis PU, AR e FR.

Tabela 5.1: Medidas de posição e dispersão.

Variável	Sigla	Média	Mediana	Desvio-padrão	Mínimo	Máximo	Amplitude
Preço unitário	PU	72.82	55.56	70.28	2.36	800.00	797.64
Latitute	LAT	710100.00	710300.00	2722.34	701500.00	714600.00	13100.00
Longitude	LONG	8787000.00	8786000.00	6638.77	8769000.00	8798000.00	29000.00
Área	AR	1355.00	300.00	6063.53	48.00	91780.00	91732.00
Frente	FR	18.13	10.00	30.54	2.60	516.00	513.40

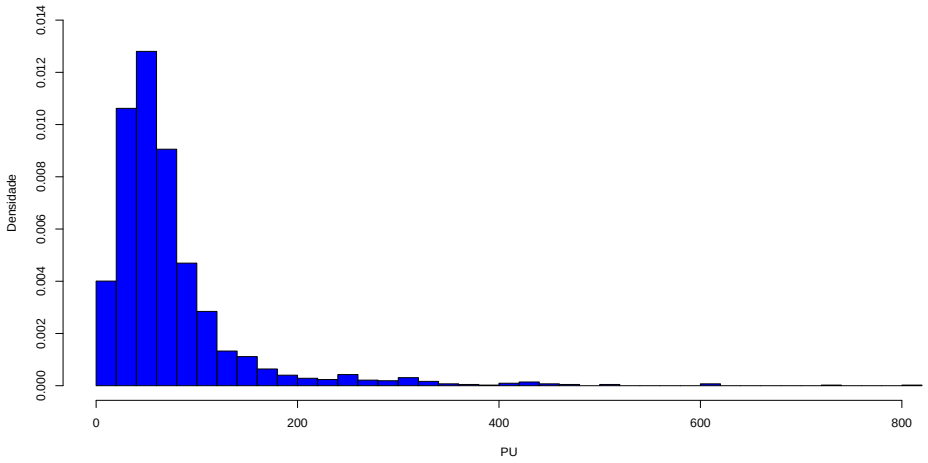

Figura 5.4: Histograma de PU.

5.2.2.2. Variáveis quantitativas discretas

Conforme podemos observar no gráfico de barras da Figura 5.5, a variável CA assume apenas 7 (sete) valores: 3.0, 3.5, 4.0, 4.5, 5.0, 5.5 e 6.0, sendo 3.0 o de maior frequência relativa (= 45%) e 4.5 o de menor frequência (= 1%), indicando que a maior parte das observações está situada em bairros que não permitem a construção de edificações muito altas. Por outro lado, a variável ST assume valores discretos e consecutivos de 1 até 18, sendo cerca de 65% dos dados localizados em zonas censitárias com renda média dos chefes de família inferiores a 4 salários mínimos, denotando a concentração das observações nos bairros de menor poder aquisitivo.

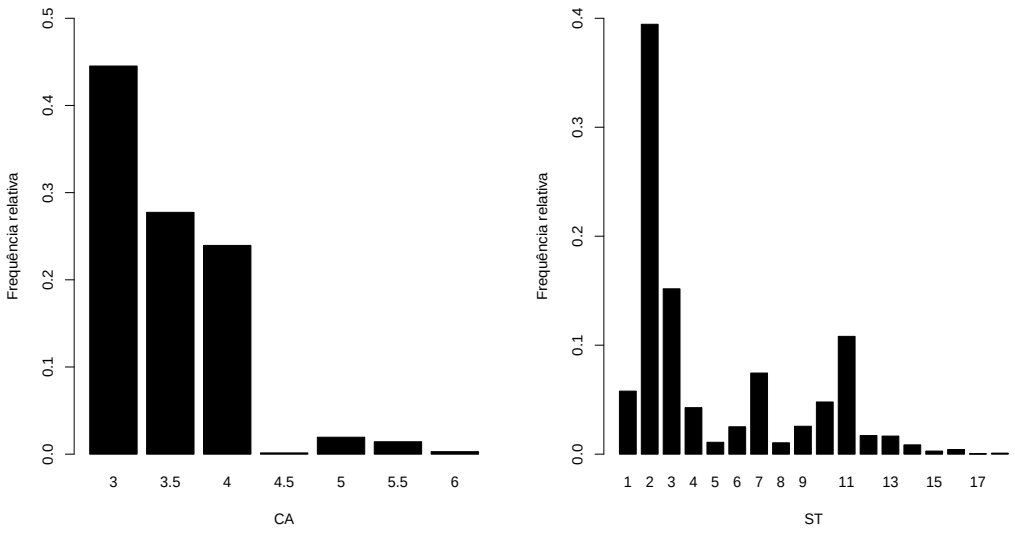

Figura 5.5: Gráficos de barras das variáveis CA e ST.

Na Figura 5.6 apresentamos os gráficos box-plot das variáveis CA e ST, sobre os quais enfatizamos que, embora tenham sido registradas poucas observações discrepantes tanto na variável CA como na ST, há uma evidente assimetria negativa na distribuição de ambas as variáveis.

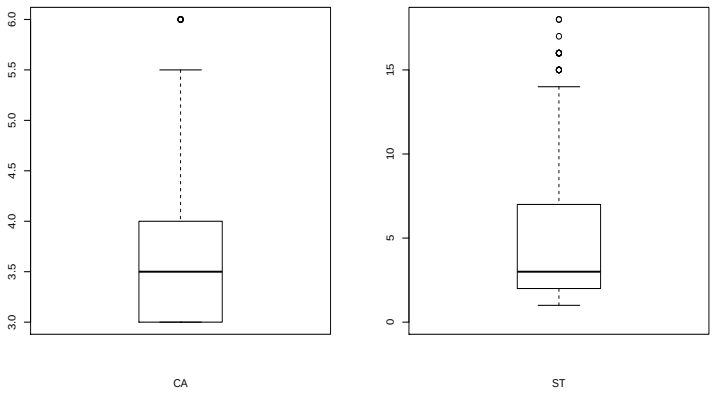

Figura 5.6: Gráficos box-plot das variáveis CA e ST.

5.2.2.3. Variáveis qualitativas nominais

A partir da Figura 5.7, referente ao gráfico de barras da variável BAIRRO, é possível listar todos os bairros que têm dados contidos na amostra, a saber: Aeroporto, América, Atalaia, Bugio, Capucho, Centro, Cidade Nova, Cirurgia, Coroa do Meio, Dezoito do Forte, Farolândia, Getúlio Vargas, Grageru, Inácio Barbosa, Industrial, Jabotiana, Jardim Centenário, Jardins, José Conrado de Araújo, Lamarão, Luzia, Mosqueiro, Novo Paraíso, Olaria, Palestina, Pereira Lobo, Ponto Novo, Porto Dantas, Salgado Filho, Santa Maria, Santo Antônio, Santos Dumont, São Conrado, São José, Siqueira Campos, Soledade, Suíça e Treze de Julho. Além disso, verificamos que os bairros do Mosqueiro, Atalaia, Coroa do Meio e Cidade Nova são os que possuem o maior número de terrenos observados na amostra (659, 225, 180 e 180, respectivamente), enquanto os bairros Bugio, Palestina e Novo Paraíso são os menos representados na amostra (2, 2 e 7, respectivamente).

Na Figura 5.8, exibimos os gráficos de setores das variáveis SI, PA, TO e NI, onde são evidenciados que os terrenos situados no "meio" da quadra, localizados em logradouros "não-pavimentados", de conformidade topográfica "plana" e preços unitários oriundos do "ITBI", respectivamente, são majoritários na amostra para este grupo de variáveis. Destacamos, entretanto, a discrepância entre a quantidade de observações coletadas nas variáveis TO, para as situações de "plano" (= 2022) e "acidentado" (= 87), e na variável NI, para os casos de "ITBI" (= 1852), oferta (= 204) e "transação"(= 53).

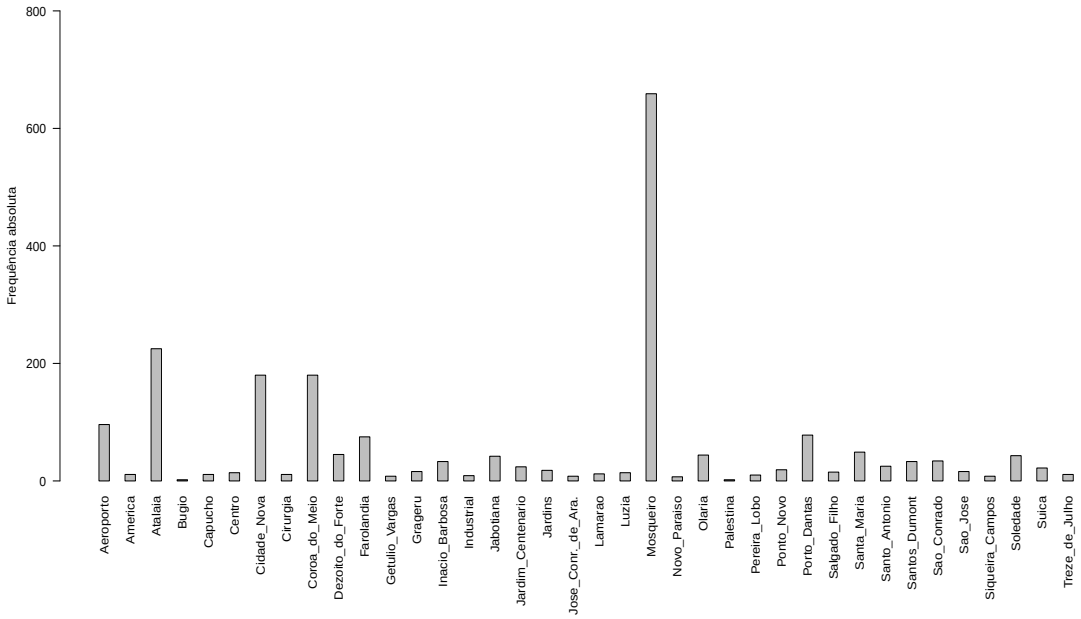

Figura 5.7: Gráfico de barras da variável BAIRRO.

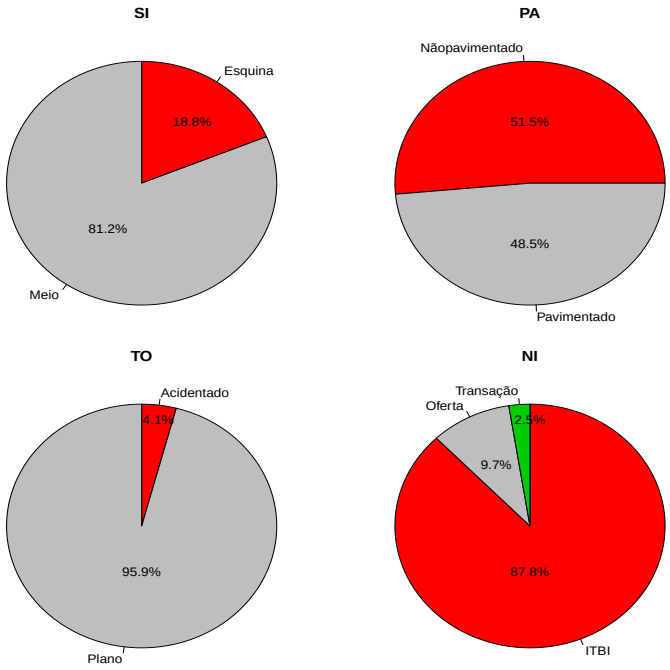

Figura 5.8: Gráfico de setores das variáveis SI, PA, TO e NI.

5.2.2.4. Variáveis qualitativas ordinais

Notamos da Figura 5.9, relativa aos gráficos de setores das variáveis VIA e ANO, que os terrenos observados estão preponderantemente situados em "vias terciárias/superior", seguidos pelas "vias principais" e "vias secundárias". É possível ainda identificar que a maioria dos dados foi coletada no ano de 2005, sendo o ano de 2006 aquele com o menor número de registros. Aqui, novamente, chamamos atenção para a notável diferença entre o número de dados localizados entre "vias terciárias/superior" (= 1799), "vias principais" (= 163) e "vias secundárias" (= 147).

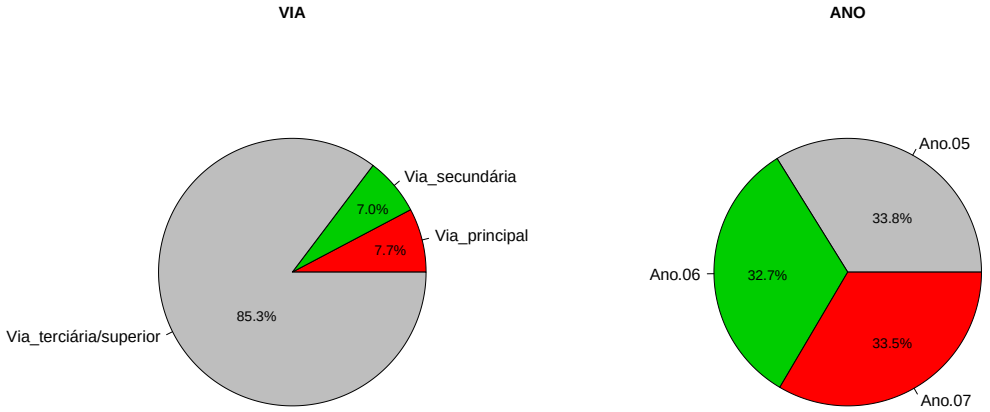

Figura 5.9: Gráfico de setores das variáveis VIA e ANO.

5.2.3 Análise de associação entre variáveis

Na Engenharia de Avaliações e para o caso de terrenos, o interesse recai, geralmente, na modelagem do preço unitário, com base na área do terreno, em função das características estruturais, locacionais e econômicas que o bem pode assumir. Sendo assim, adotaremos neste trabalho como variável dependente PU e como variáveis independentes as respectivas características locacionais (BAIRRO, LAT, LONG, ST, CA e VIA), físicas (AR, FR, TO e SI) e econômicas (NI), além do ANO em que a observação foi coletada.

Novamente e a fim de tornar o estudo sequenciado, subdividiremos a análise em dois grupos de variáveis: (i) variáveis quantitativas e (ii) variáveis qualitativas. Evidentemente, a variável dependente PU estará presente nos dois grupos supracitados para que também se examine o comportamento desta variável frente às demais.

5.2.3.1. Variáveis quantitativas

O primeiro passo para se analisar o comportamento de PU em relação às variáveis explicativas é analisar o gráfico de dispersão. Neste sentido, apresentamos na Figura 5.10 os seguintes gráficos de dispersão: (i) PU $\times$ LAT; (ii) PU $\times$ LONG; (iii) log(PU) $\times$ log(AR); (iv) log(PU) $\times$ log(FR); (v) PU $\times$ ST; (vi) PU$\times$ CA. Note que em (iii) e (iv) foi necessário aplicar uma transformação logarítmica em PU, AR e FR para uma melhor visualização gráfica da relação entre as variáveis, visto que a grande amplitude e a alta variabilidade observadas em AR e FR dificultam a análise em suas respectivas escalas de medidas originais.

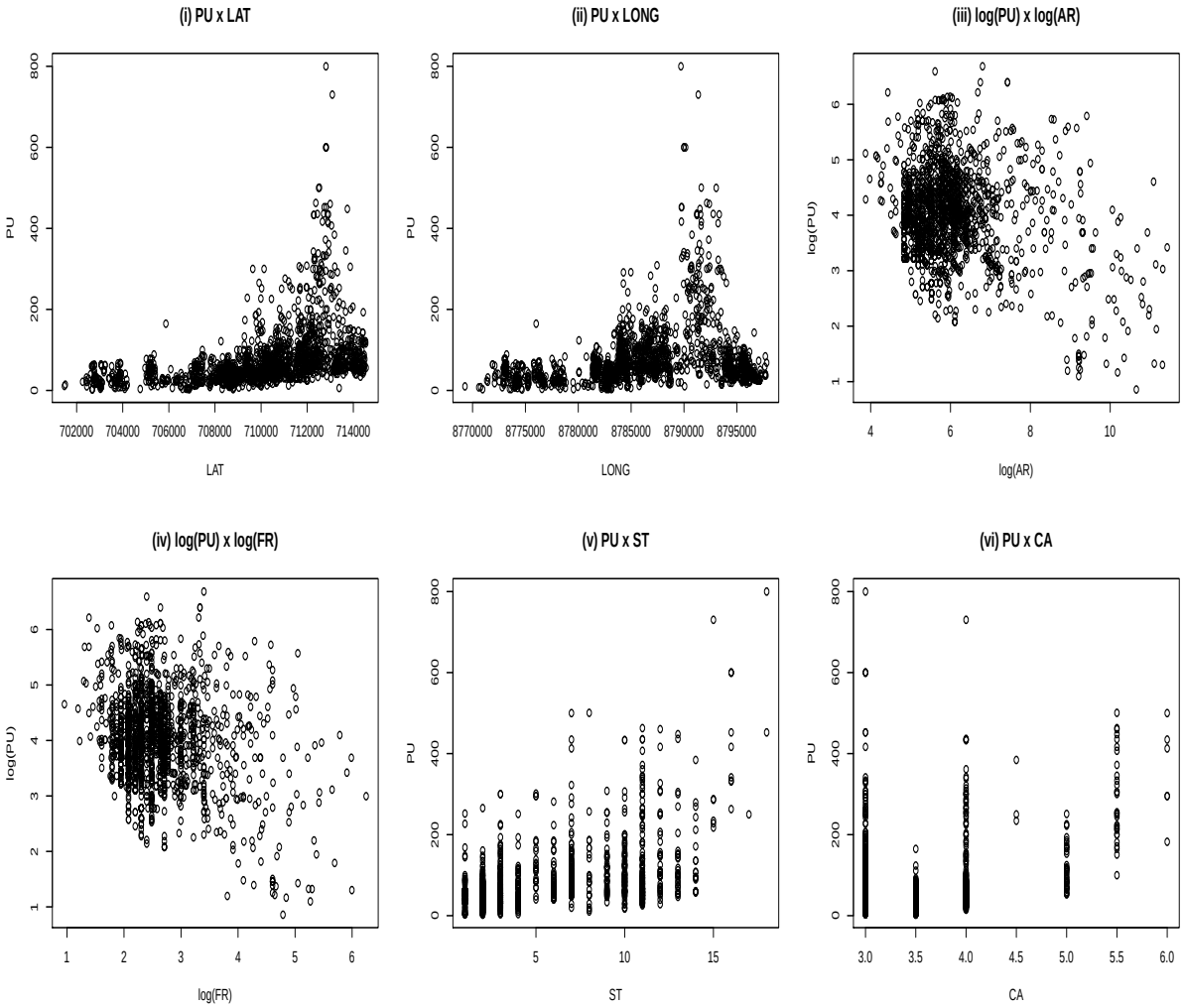

Figura 5.10: Gráficos de dispersão entre PU e as variáveis quantitativas explicativas.

Conforme podemos observar na Figura 5.10, aparentemente há uma relação diretamente proporcional — embora a intensidade desta relação não seja tão acentuada — entre PU e as variáveis explicativas em (i), (ii), (v) e (vi), enquanto que em (iii) e (iv) percebe-se uma relação inversamente proporcional. A partir disto e em princípio, podemos constatar que existe uma tendência de acréscimo do valor unitário na medida que a latitude, longitude, setor e coeficiente de aproveitamento aumentam. Contudo, em (iii) e (iv) há uma tendência de decréscimo do preço unitário quando a área e a frente crescem. Aqui, cabe destacarmos que a expectativa, *a priori*, que tínhamos do mercado somente não foi ratificada em (iv), visto que esperávamos o aumento de PU quando FR crescesse. De fato, é de se esperar que FR tenha uma influência positiva sobre PU, fundamentalmente em terrenos situados em áreas comerciais e nobres da cidade de Aracaju, porém, é provável que este efeito não tenha sido evidenciado quando considerados na amostra os terrenos situados em bairros da periferia, onde a influência isolada de FR sobre PU não segue a mesma intensidade das áreas nobres.

Outro aspecto importante que podemos mencionar acerca da Figura 5.10 diz respeito à forma funcional da curva que melhor se ajustaria aos dados. Note que é difícil afirmar com segurança se a interdependência observada entre PU e as demais variáveis é linear ou não. Além disso, sustentar as hipóteses de homoscedasticidade e normalidade da distribuição condicional de PU dadas as variáveis explicativas (analisadas individualmente e conjuntamente) pode não ser razoável. Para situações desta natureza, Rigby & Stasinopoulos (2007) ressaltam que costumeiramente são realizadas transformações na variável resposta e/ou nas variáveis explanatórias, como em (iii) e (iv), a fim de tentar "corrigir" algum ou todos os problemas mencionados anteriormente. Contudo, este artifício nem sempre é exitoso e a tarefa de obter as transformações nas variáveis que minimizam os efeitos da não-linearidade, heteroscedasticidade e ausência de normalidade pode ser laboriosa, incoerente com a teoria subjacente e resultar em expressões de difícil interpretação.

Mediante análise da matriz de correlações dois a dois (variáveis analisadas em suas respectivas escalas de medidas originais) apresentada na Tabela 5.2, podemos complementar as observações mencionadas nos dois parágrafos anteriores, uma vez que é possível constatar as relações diretas entre PU $\times$ LAT, PU $\times$ LONG, PU $\times$ CA e PU $\times$ ST, dadas,

respectivamente, por 0.40, 0.21, 0.19 e 0.50. Ademais, ratificam-se as influências inversas de FR sobre PU e de AR sobre PU por meio das medidas de correlação (-0.07) e (-0.09), respectivamente. Note que, linearmente e sem quaisquer transformações nas variáveis PU $\times$ AR e PU $\times$ FR, a relação de dependência é quase nula, ao passo que procedendo à transformação logarítmica em PU, AR e FR, há um aumento (em valor absoluto) nas medidas de correlação (vide Tabela 5.3) entre log(PU) $\times$ log(AR) e log(PU) $\times$ log(FR), embora ainda estejam longe de refletir a real importância das variáveis AR e FR na determinação do preço unitário de um terreno.

Finalmente, chamamos a atenção para a forte correlação positiva entre AR $\times$ FR $(= 0.77$, vide Tabela 5.2) e log(AR) $\times$ log(FR) $(= 0.93$, vide Tabela 5.3), indicando que podemos ter multicolinearidade no modelo de regressão se estas variáveis forem incluídas conjuntamente. Este fato é esperado, haja vista que terrenos com frentes grandes tendem a ter áreas grandes e vice-versa, conforme ilustrado no gráfico de dispersão log(FR) $\times$ log(AR) da Figura 5.11.

Tabela 5.2: Matriz de correlações dois a dois - variáveis nas escalas de medidas originais.

	LAT	LONG	AR	FR	CA	ST	PU
LAT	1.00	0.53	-0.08	-0.07	-0.06	0.58	0.40
LONG	0.53	1.00	-0.13	-0.18	0.41	-0.04	0.21
AR	-0.08	-0.13	1.00	0.77	-0.06	-0.00	-0.09
FR	-0.07	-0.18	0.77	1.00	-0.12	0.07	-0.07
CA	-0.06	0.41	-0.06	-0.12	1.00	-0.26	0.19
ST	0.58	-0.04	-0.00	0.07	-0.26	1.00	0.50
PU	0.40	0.21	-0.09	-0.07	0.19	0.50	1.00

Tabela 5.3: Matriz de correlações dois a dois - variáveis PU, AR e FR transformadas.

	log(AR)	log(FR)	log(PU)
log(AR)	1.00	0.93	-0.21
log(FR)	0.93	1.00	-0.21
log(PU)	-0.21	-0.21	1.00

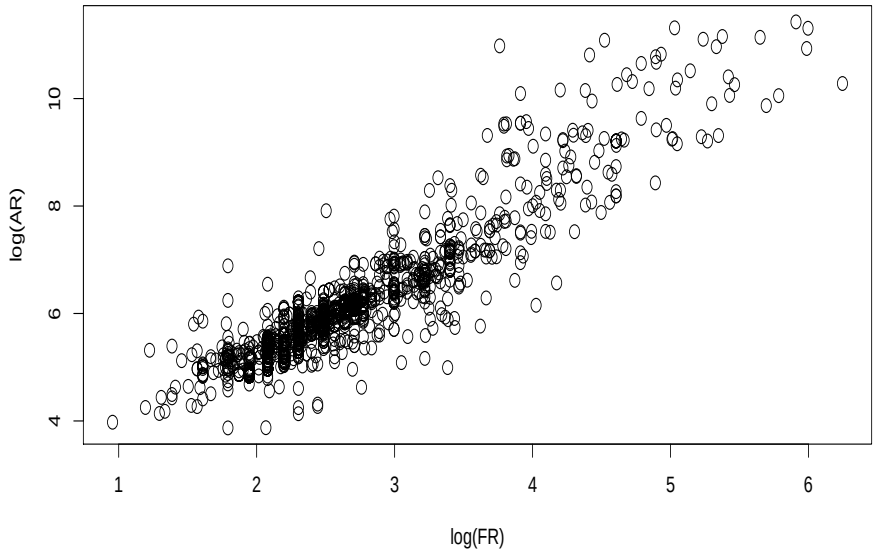

Figura 5.11: Gráfico de dispersão entre as variáveis FR e AR.

5.2.3.2. Variáveis qualitativas

Visando à identificação de alguma tendência entre as variáveis qualitativas e o preço unitário, construímos na Figura 5.12 os gráficos box-plot entre: (i) PU × SI; (ii) PU × PA; (iii) PU × TO; (iv) PU × NI; (v) PU × VIA; (vi) PU × ANO. É possível destacar que no gráfico (i) há uma leve tendência de terrenos de "esquina" serem mais valorizados do que os de "meio" de quadra; no gráfico (ii) terrenos situados em vias "pavimentadas" aparentam ser mais caros que aqueles localizados em vias "não-pavimentadas"; no gráfico (iii) há uma suave valorização de terrenos "planos" em detrimento de terrenos "acidentados"; no gráfico (iv) existe uma clara tendência de preços unitários oriundos de "ITBI" serem inferiores àqueles oriundos de "oferta" ou "transação"; no gráfico (v) é perceptível a desvalorização de terrenos localizados em "vias terciárias/superiores" frente àqueles situados em vias "principais" ou "secundárias" e no gráfico (vi) notamos uma tendência de aumento do preço unitário no mesmo sentido de crescimento da ordem cronológica dos anos.

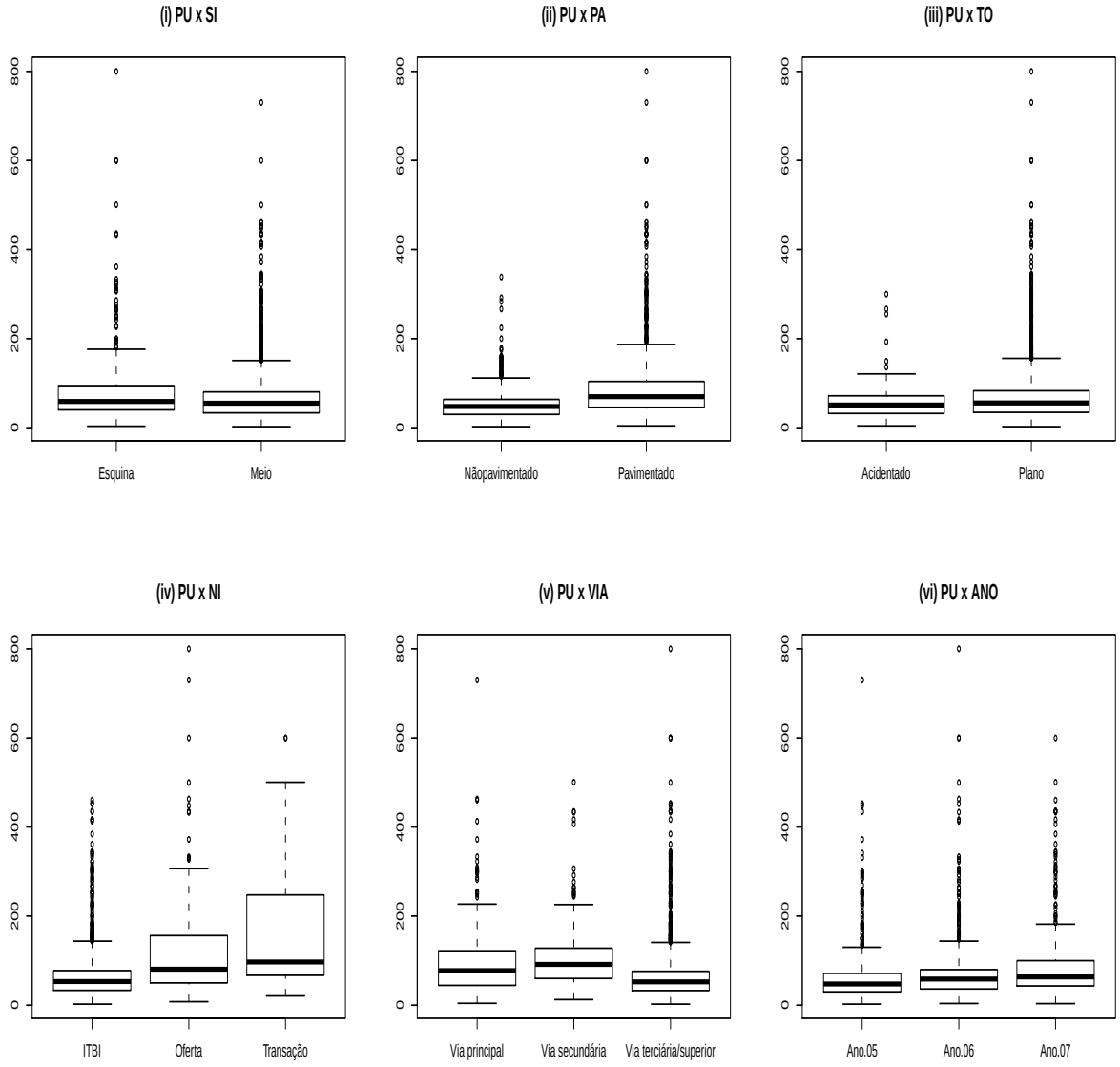

Figura 5.12: Gráficos box-plot entre PU e as variáveis qualitativas explicativas.

No que tange à variável BAIRRO, apresentamos na Figura 5.13 o gráfico box-plot desta variável em função de PU. Claramente, é possível identificar que os bairros Jardins, Centro e Salgado Filho apresentam a maior variabilidade entre os preços unitários, enquanto os bairros de Jardins, Treze de Julho e Centro têm os preços unitários medianos mais altos e os bairros de Capucho, Santa Maria e Soledade, os mais baixos.

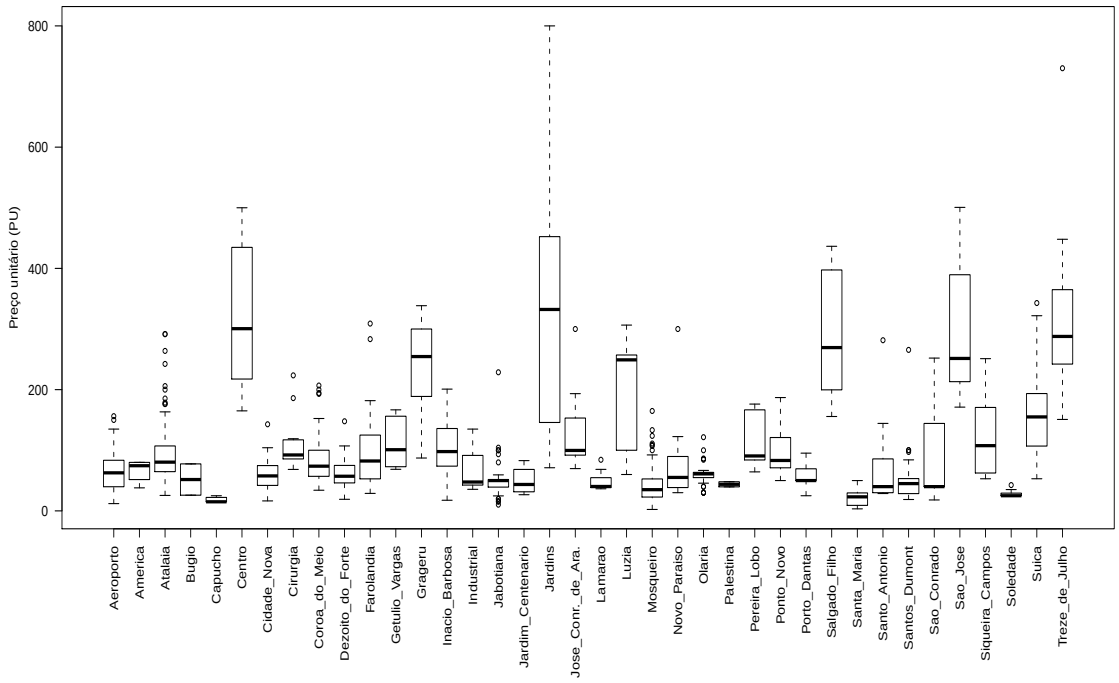

Figura 5.13: Gráfico box-plot da variável BAIRRO em função de PU.

5.3 Informações adicionais sobre as variáveis

Tendo em vista que na seção seguinte estimaremos a equação de preços hedônicos para terrenos situados em Aracaju-SE, faz-se necessário definir de que "forma" as variáveis serão avaliadas e incorporadas no modelo de regressão. Para tanto, apresentamos na Tabela 5.4 um quadro-resumo com as principais características e tratamentos considerados para cada variável.

Cumpre registrar que a variável de interação denominada de FRBV foi incluída para verificar se a influência da dimensão da frente dos terrenos localizados nos bairros admitidos como "valorizados" é significativa em relação àqueles situados nos bairros supostamente "menos valorizados", haja vista que a expectativa *a priori* é de que os bairros comerciais e residenciais nobres (por exemplo, Centro, Jardins e Treze de Julho) tenham os preços unitários dos terrenos fortemente impactados e acrescidos com o aumento do tamanho da testada, ao passo que nos demais bairros este efeito pode não ser tão significativo, conforme conjecturamos quando analisamos o gráfico de dispersão entre as variáveis PU $\times$ FR (vide Seção 5.2.3.1).

Tabela 5.4: Quadro-resumo das variáveis utilizadas nos modelos de regressão.

Variável	**Sigla**	**Classificação I**	**Classificação II**	**Classificação III**	**Domínio**
Preço unitário	PU	Dependente	Contínua	————	$\mathbb{R}_+^*$
Latitute	LAT	Independente	Contínua	————	$\mathbb{R}$
Longitude	LONG	Independente	Contínua	————	$\mathbb{R}$
Área	AR	Independente	Contínua	————	$\mathbb{R}_+^*$
Frente	FR	Independente	Contínua	————	$\mathbb{R}_+^*$
Coef. de aproveitamento	CA	Independente	Discreta	————	3, 3.5,..., 5.5, 6.0
Setor	ST	Independente	Discreta	*Proxy*	1,2,..., 17, 18
Topografia	TO	Independente	Nominal	*Dummy*	0 se não for plano 1 se for plano
Pavimentação	PA	Independente	Nominal	*Dummy*	0 se não for pavimentado 1 se for pavimentado
Situação	SI	Independente	Nominal	*Dummy*	0 se for de meio 1 se for de esquina
Bairros valorizados*	BV	Independente	Nominal	*Dummy*	0 se não for bairro valorizado 1 se for bairro valorizado
Via Via principal Via secundária Via terciária/superior	VIA VIAP VIAS VIAT	Independente	Nominal	*Dummy*	Vias: principal, secundária, ou terciária/superior VIAP=1 e VIAS=0 VIAP=0 e VIAS=1 VIAP=0 e VIAS=0
Natureza da informação Oferta Transação ITBI	NI NIO NIT NIBI	Independente	Nominal	*Dummy*	Oferta, transação, ou ITBI Oferta=1 e transação=0 Oferta=0 e transação=1 Oferta=0 e transação=0
Ano 2007 2006 2005	ANO ANO.07 ANO.06 ANO.05	Independente	Ordinal	*Dummy*	2005, 2006, ou 2007 AN0.06=0 e ANO.07=1 AN0.06=1 e ANO.07=0 AN0.06=0 e ANO.07=0
Frente em bairros valorizados**	FRBV	Independente	Contínua	Interação	$\mathbb{R}_+$

* Foram considerados como bairros supostamente valorizados: Jardins, Treze de Julho e Centro.

** Variável correspondente à interação entre as variáveis FR e BV.

5.4 Estimação de modelos

A especificação de modelos que visam à estimação empírica da equação de preços hedônicos não pode ser feita mecanicamente; precisa de compreensão, intuição e habilidade. Embora o senso comum, a lógica e a experiência de outros pesquisadores proporcionem guias para a escolha do "melhor" método para explicar a formação dos preços, essas são teorias que devem ser comprovadas com a realidade, a partir dos dados de mercado.

Conforme já destacado, na literatura nacional as equações de preços hedônicos voltadas para o mercado imobiliário têm sido, em sua maioria, formuladas com base no modelo normal de regressão linear clássico e adotam uma forma linear, log-linear ou fazem uso da transformação de Box-Cox na variável resposta. Uma outra alternativa tem sido a utilização dos modelos lineares generalizados com emprego das distribuições gama e lognormal.

Contudo, a heterogeneidade intrínseca presente nos dados imobiliários e a inexistência de uma teoria que determine a forma funcional da equação hedônica dificultam a aplicação de metodologias econométricas que resultem em modelos simultaneamente parcimoniosos, abrangentes e fidedignos ao mercado. É necessário que a estrutura de regressão utilizada seja flexível, a ponto de "acomodar" as peculiaridades do bem imóvel e as limitações da própria teoria.

Em virtude do exposto e considerando que o ponto central de nossa análise é conferir flexibilidade ao ajuste, estimaremos a função de preços hedônicos para terrenos urbanos situados em Aracaju-SE utilizando a classe de modelos GAMLSS. Antes, porém, ajustaremos os modelos CNLRM e GLM para comparações com os modelos GAMLSS.

5.4.1 A modelagem via CNLRM

No modelo normal de regressão linear clássico o preço unitário do terreno (PU) é função das suas características físicas (F) – área, frente, topografia etc. –, locacionais (L) – bairro onde se situa o imóvel, distância a pólos de influência, amenidades[10] do entorno etc. – e econômicas(E) – época da transação, condições de pagamento, natureza do evento:

[10]Entende-se por amenidades urbanas um conjunto de características específicas de uma localidade com contribuição positiva ou negativa para a satisfação dos indivíduos.

em oferta ou efetivamente vendido etc. –, conforme definido na Equação (5.1):

$$PU = f(F, L, E, \beta) + \epsilon, \tag{5.1}$$

em que f é um operador indicativo da forma funcional linear, $\boldsymbol{\beta}$ é um vetor de parâmetros e ϵ é um erro aleatório do modelo, respectivamente. Considera-se o conjunto de erros para várias observações como sendo *i.i.d*, ou seja, admite-se que os erros aleatórios são independentes e identicamente distribuídos (normais, homoscedásticos e não-autocorrelacionados).

Nesse caso, o modelo adotado para inferir o comportamento do mercado imobiliário é dado por

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad i = 1, \ldots, n, \tag{5.2}$$

em que Y expressa a variável dependente, retratada pelo preço do imóvel observado no mercado; $X_{i1}, \ldots, X_{ik}$ são as variáveis independentes; $\beta_0, \ldots, \beta_k$ são parâmetros desconhecidos da regressão a serem estimados e $\epsilon_1, \ldots \epsilon_n$ são termos de perturbação estocástica que causam a "natural flutuação" dos preços de mercado e são provenientes da imprevisibilidade do comportamento humano, da não inclusão de variáveis independentes que contribuem muito pouco para a formação dos preços de mercado e de erros amostrais e não amostrais (erros de mensuração, especificação, processamento, entre outros). Acrescenta-se que o i minúsculo na base do X refere-se à i-ésima observação, enquanto o segundo subíndice em X identifica o número da variável em questão e n corresponde à quantidade total de unidades observadas. O número de variáveis independentes é k, de modo que para $k = 1$ a equação de regressão linear múltipla (5.2) se reduz a um modelo de regressão linear simples.

Em forma matricial o modelo (5.2) é dado por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.3}$$

em que

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ e } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

sendo $\boldsymbol{Y}$, $\boldsymbol{\beta}$ e $\boldsymbol{\epsilon}$ os vetores de preços observados, de parâmetros e de erros aleatórios do modelo de regressão, respectivamente, e $\boldsymbol{X}$ a matriz das observações das variáveis independentes conhecidas.

Tradicionalmente, a estimação dos parâmetros é realizada com base no método de mínimos quadrados ordinários (*Ordinary Least Squares* — OLS),[11] de forma que o sistema de equações normais e os estimadores OLS para o Modelo (5.3) são dados, respectivamente, por

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y} \quad \text{e} \quad \boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

em que $\boldsymbol{b}$ corresponde ao vetor de coeficientes estimados do modelo, também denotado de vetor de preços implícitos.

Assim, a estimativa do valor de mercado de um imóvel é dada por

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_k X_{ik}, \ i = 1, \ldots, n, \tag{5.4}$$

em que $b_0, \ldots, b_k$ são as estimativas correspondentes a $\beta_0, \ldots, \beta_k$, respectivamente.

Tendo em vista que a base teórica da estimação empírica tradicional utiliza os estimadores OLS, alguns pressupostos devem ser atendidos se o objetivo é fazer testes de hipóteses, estimação intervalar e garantir que os parâmetros inferidos no mercado sejam não-tendenciosos, eficientes e consistentes, a saber: (i) o modelo $\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\epsilon}$ está corretamente especificado, ou seja, a forma funcional está correta, na sua composição estão incluídas apenas variáveis explicativas relevantes, o termo de erro estocástico está corretamente definido e não há erros de medição nas covariáveis, (ii) $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$, em que $\boldsymbol{0}$ é um vetor $n \times 1$ de zeros, ou seja, fatores não incluídos explicitamente no modelo e, portanto, agrupados em $\boldsymbol{\epsilon}$, não afetam sistematicamente o valor médio de $\boldsymbol{Y}$, (iii) $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$, em que $\boldsymbol{I}$ é a matriz identidade de dimensão $n \times n$ e $0 < \sigma^2 < \infty$, ou seja, os termos de erro são descorrelacionados e possuem variância constante (modelo homoscedástico), (iv) $\boldsymbol{X}$ possui posto coluna completo, ou seja, as colunas de $\boldsymbol{X}$ são linearmente independentes e (v) $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$, ou seja, os erros têm distribuição normal[12] — com média 0 e variância σ^2 — e são independentes.

[11] Uma referência sobre o assunto é Davidson & MacKinnon (2004, Capítulo 15).

[12] Embora a suposição de normalidade para a distribuição de probabilidade do termo de erro estocástico não seja necessária para que os estimadores OLS sejam não-viesados, consistentes e eficientes, ela é tipicamente usada para estimação intervalar e para a realização de testes de hipóteses sobre os parâmetros da regressão. Assim, inferências realizadas sobre preços hedônicos em regressões lineares não-normais baseadas na suposição de normalidade podem ser imprecisas.

Na Tabela 5.5 resumimos os principais ajustes realizados via CNLRM e as observações relevantes acerca dos modelos concorrentes à predição da equação de preços hedônicos. Contudo, antes de analisarmos cada modelo estimado, esclarecemos que a variável FR mostrou-se altamente correlacionada com AR (vide Seção 5.2.3.1) e em todos os modelos ajustados apresentou o sinal do coeficiente estimado negativo, ou seja, contrário à expectativa do mercado imobiliário, motivo pelo qual foi excluída durante a modelagem.

Tabela 5.5: Modelos ajustados via CNLRM

Modelos	Forma Funcional	Considerações
1.1	$\text{PU} = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LONG} + \beta_3\text{AR} + \beta_4\text{CA} + \beta_5\text{ST} + \beta_6\text{VIAP} + \beta_7\text{VIAS} + \beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{ANO06} + \beta_{14}\text{ANO07} + \beta_{15}\text{DZSU} + \beta_{16}\text{FRBV} + \epsilon$	As hipóteses nulas de que os resíduos são homocedásticos e normais foram rejeitadas ao nível de 1% pelos teste de Breusch-Pagan e Jarque-Bera, respectivamente. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível de 1% quando utilizado o teste t. $\overline{R}^2$=0.539, AIC=22304 e BIC=22406.
1.2	$\log(\text{PU}) = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LONG} + \beta_3\text{AR} + \beta_4\text{CA} + \beta_5\text{ST} + \beta_6\text{VIAP} + \beta_7\text{VIAS} + \beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{ANO06} + \beta_{14}\text{ANO07} + \beta_{15}\text{DZSU} + \beta_{16}\text{FRBV} + \epsilon$	As hipóteses nulas de que os resíduos são homoscedásticos e normais foram rejeitadas ao nível de 1% pelos teste de Breusch-Pagan e Jarque-Bera, respectivamente. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível de 1% quando utilizado o teste t. $\overline{R}^2$=0.599, AIC=2912 e BIC=3014.
1.3	$\log(\text{PU}) = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LONG} + \beta_3\log(\text{AR}) + \beta_4\text{CA} + \beta_5\log(\text{ST}) + \beta_6\text{VIAP} + \beta_7\text{VIAS} + \beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{ANO06} + \beta_{14}\text{ANO07} + \beta_{15}\text{DZSU} + \beta_{16}\log(\text{FRBV}) + \epsilon$	A estatística Jarque-Bera indicou a não rejeição da hipótese nula de uma distribuição normal dos resíduos, mas o teste de Breusch-Pagan rejeitou a hipóteses nula de homoscedasticidade ao nível de 1%. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível de 1%, exceto para a variável LAT (valor-p = 0.0190). $\overline{R}^2$=0.651, AIC=2619 e BIC=2721.
1.4	$\frac{\text{PU}^\lambda - 1}{\lambda} = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LONG} + \beta_3\log(\text{AR}) + \beta_4\text{CA} + \beta_5\log(\text{ST}) + \beta_6\text{VIAP} + \beta_7\text{VIAS} + \beta_8\text{PA} + \beta_9\text{TO} + \beta_{10}\text{NIO} + \beta_{11}\text{NIT} + \beta_{12}\text{ANO06} + \beta_{13}\text{ANO07} + \beta_{14}\log(\text{FRBV}) + \epsilon$	A estatística Jarque-Bera indicou a não rejeição da hipótese nula de uma distribuição normal dos resíduos, mas o teste de Breusch-Pagan rejeitou a hipóteses nula de homoscedasticidade ao nível de 1%. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível de 1%, exceto para a variável LAT (valor-p = 0.0881). $\overline{R}^2$ = 0.657, AIC=4290 e BIC=4392.

De acordo com os resultados apresentados na Tabela 5.5, verificamos que o Modelo (1.1) teve as hipóteses básicas de normalidade e homoscedasticidade dos erros aleatórios fortemente rejeitadas quando utilizados os testes de Jarque-Bera e Breusch-Pagan, respectivamente, indicando que esta forma funcional parece não ser a mais adequada.

O segundo modelo proposto (Modelo (1.2)) considera a forma funcional log-linear,[13]

[13]De acordo com Halvorsen & Palmquist (1980), a forma funcional log-linear (semilogarítmica) é a mais comum na literatura hedônica devido ao "razoável" ajuste do modelo aos dados e a interpretabilidade

onde o preço unitário dos terrenos é tomado na escala logarítmica e as demais variáveis na escala original. Aqui, novamente, as hipóteses básicas de normalidade e homoscedasticidade dos erros aleatórios foram rejeitadas.

A construção do Modelo (1.3) baseia-se na inspeção visual do diagrama de dispersão entre a variável resposta (na escala logarítmica) *versus* as variáveis independentes (na escala original e logarítmica). Caso as transformações realizadas evidenciem ou acentuem uma relação linear, a covariável é incluída no modelo na forma transformada (logarítmica), caso contrário é incorporada ao modelo na escala original. Assim, considerou-se a transformação logarítmica em PU e nas covariáveis AR, ST e FRBV. Todavia, embora o teste de Jarque-Bera não tenha rejeitado a hipótese nula de normalidade dos resíduos, o teste de Breusch-Pagan rejeitou a hipótese de homoscedasticidade dos erros aleatórios.

Embora o Modelo (1.4), referente à transformação de Box-Cox (com $\lambda = 0.1010$), tenha apresentado os "melhores" resultados no que tange ao coeficiente de determinação ajustado $\overline{R}^2$, AIC, BIC e o gráfico dos *valores observados* $\times$ *valores preditos* (vide Figura 5.14),[14] não foi capaz de estabilizar a variância dos resíduos, conforme teste de Breusch-Pagan. Apesar da estatística Jarque-Bera não ter rejeitado a hipótese nula de normalidade dos resíduos e a hipótese nula de que o conjunto de variáveis explicativas adotadas não é importante para explicar a variabilidade observada nos preços dos terrenos ter sido rejeitada – quando utilizado o teste F (valor-$p \cong 0.00$) –, inferências baseadas nas estimativas dos parâmetros β's podem ser enganosas (vide Davidson & MacKinnon, 1993), visto que os estimadores de mínimos quadrados ordinários, embora ainda não tendenciosos e consistentes, deixam de ser eficientes (mesmo assintoticamente) sob heteroscedasticidade. Diante disto, apresentamos na Tabela 5.6 o ajuste realizado para o Modelo (1.4) utilizando o estimador HC3 (Davidson & Mackinnon, 1993) para corrigir o efeito da heteroscedasticidade. Para mais detalhes sobre as técnicas de detecção e correção de heteroscedasticidade, vide Mackinnon & White (1985, 1993) e Godfrey (2006).

De acordo com a Tabela 5.6, todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível de 2%, exceto para a variável LAT (valor-$p = 0.1263$),

direta dos coeficientes estimados – o coeficiente angular mede a variação proporcional constante em Y para dada variação absoluta no valor do regressor.

[14]Neste gráfico, a reta vermelha traçada forma um ângulo de 45° com o eixo horizontal e representa o "ajuste ideal" sobre a qual se deseja que os valores estimados recaiam.

indicando que as maiores variações dos preços, a grande escala espacial, ocorrem no sentido norte-sul. Realmente isto pode ser verificado em função do bairro do Mosqueiro, que se situa na parte sul da cidade e abrange quase 40% da sua extensão, ser um dos trechos de menor preço unitário.

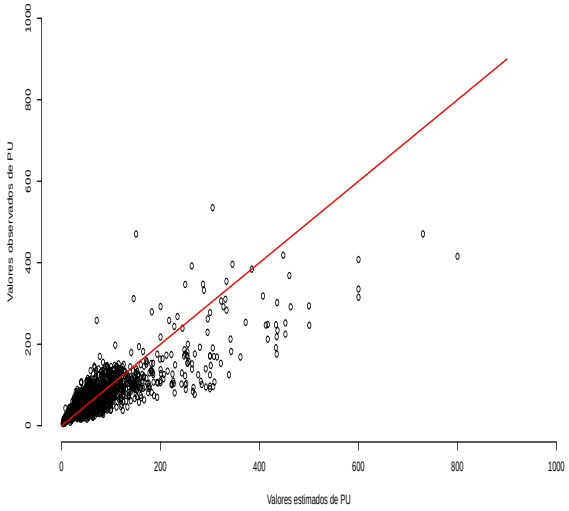

Figura 5.14: Gráfico dos *valores observados* × *valores preditos* de PU - Modelo (1.4).

Tabela 5.6: Ajuste do modelo de preços hedônicos via CNLRM - Modelo (1.4).

	Estimativa	Erro-padrão	Estatística t	valor-p
(Intercepto)	-162.6307	34.1920	-4.756	0.0000
LAT	1.85e-05	1.21e-05	1.529	0.1263
LONG	1.74e-05	4.60e-06	3.798	0.0001
log(AR)	-0.3507	0.0192	-18.236	0.0000
log(ST)	0.4423	0.0332	13.297	0.0000
CA	0.2651	0.0412	6.429	0.0000
VIAP	0.4874	0.0717	6.789	0.0000
VIAS	0.1678	0.0675	2.485	0.0130
SI	0.1119	0.0405	2.757	0.0058
PA	0.3853	0.0302	12.767	0.0000
TO	0.4905	0.0798	6.145	0.0000
NIO	0.5994	0.0592	10.131	0.0000
NIT	0.5111	0.0131	3.886	0.0000
ANO06	0.2560	0.0351	7.289	0.0000
ANO07	0.6450	0.0345	18.645	0.0000
DZSU	0.7221	0.0474	15.239	0.0000
lFRBV	1.2041	0.0137	8.797	0.0000

Os resultados obtidos neste trabalho mediante uso dos modelos CNLRM ratificam, conforme observado por Dantas & Cordeiro (2000),[15] que a falta de normalidade é indubitável nos preços de compra de imóveis, pois estes se situam no campo dos reais positivos, enquanto que a distribuição normal abrange todo o campo dos reais. Também é natural que a heteroscedasticidade esteja presente nos dados imobiliários, uma vez que nas negociações estão presentes classes de consumidores com rendas variadas, que adquirem bens imóveis proporcionalmente às suas rendas.

5.4.2 A modelagem via GLM

Nos modelos lineares generalizados os pressupostos de variância constante e distribuição normal para o erro não são mais exigidos, sendo requeridos agora uma distribuição de probabilidades (membro da família exponencial de distribuições) para a variável resposta (componente aleatória), um conjunto de variáveis independentes descrevendo a estrutura linear do modelo (componente sistemática) e uma função de ligação ($g(\cdot)$) entre a média da variável de resposta (μ) e a estrutura linear (η). Aqui, a média do preço unitário do terreno (PU^*) é função das suas características físicas (F), locacionais (L) e econômicas (E), ou seja, nos GLMs modela-se o valor esperado dos dados ao invés de transformar as observações como nos modelos Box-Cox:

$$g(\boldsymbol{PU^*}) \;\; = \;\; f(F, L, E, \boldsymbol{\beta}), \tag{5.5}$$

em que $\boldsymbol{PU^*} = \mathrm{E}(\boldsymbol{PU}) = \boldsymbol{\mu}$ e $f(F, L, E, \boldsymbol{\beta}) = \boldsymbol{X\beta} = \boldsymbol{\eta}$, ou seja, a estimação empírica da Equação (5.5) via GLM admite que a componente sistemática é uma função linear dos parâmetros desconhecidos $(\beta_1, \ldots, \beta_p)$, em que p é o número de variáveis explicativas. O método tradicionalmente usado na estimação do vetor de parâmetros $\boldsymbol{\beta}$ de um GLM é o da máxima verossimilhança.[16]

Perceba que a análise de dados a partir dos modelos GLMs é bem mais flexível do que via CNLRM, pois para uma mesma estrutura linear pode-se obter vários modelos

[15] Em uma avaliação do mercado de apartamentos na região metropolitana do Recife, os autores verificaram que ao considerar a distribuição normal para os dados, alguns preços ajustados foram negativos, uma situação impossível de acontecer.

[16] O algoritmo de cálculo das estimativas de máxima verossimilhança foi desenvolvido por Nelder e Wedderburn (1972) e baseia-se em um método semelhante ao de Newton-Raphson, conhecido como método escore de Fisher.

dependendo da distribuição proposta para o erro e da função de ligação escolhida. Note também que quando o erro é normal e a função de ligação é a identidade, tem-se o modelo normal clássico de regressão linear como um caso particular de um GLM e a Expressão (5.5) é resolvida por um processo direto de diferenciação envolvendo equações lineares. Nos demais casos, tem-se um sistema de equações não-lineares e métodos numéricos iterativos são necessários para estimar os $\boldsymbol{\beta}$'s.

Com base nas considerações acima mencionadas e motivado pela própria natureza dos dados, todos situados no campo dos reais positivos, exibimos na Tabela 5.7 o ajuste realizado via GLM do modelo preditor da equação de preços hedônicos, dado por

$$\begin{aligned} g(PU^*) \;=\; & \beta_0 + \beta_2\text{LONG} + \beta_3\log(\text{AR}) + \beta_4\text{CA} + \beta_5\log(\text{ST}) + \beta_6\text{VIAP} + \beta_7\text{VIAS} + \\ & \beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{ANO06} + \beta_{14}\text{ANO07} + \\ & \beta_{15} + \text{DZSU} + \beta_{16}\log(\text{FRBV}), \qquad \text{(Modelo 2.1)} \end{aligned}$$

em que $PU \sim$ gama (μ, σ) e $\eta = \log(\mu)$.

Tabela 5.7: Ajuste do modelo de preços hedônicos via GLM - Modelo (2.1).

	Estimativa	Erro-padrão	Estatística t	valor-p
(Intercepto)	-151.8019	15.7792	-9.620	0.0000
LONG	1.77e-05	1.80e-06	9.851	0.0000
log(AR)	-0.2276	0.0108	-21.120	0.0000
CA	0.1272	0.0231	5.515	0.0000
log(ST)	0.2880	0.0193	14.954	0.0000
VIAP	0.3562	0.0395	9.021	0.0000
VIAS	0.1419	0.0408	3.482	0.0005
SI	0.0945	0.0255	3.707	0.0002
PA	0.2324	0.0220	10.556	0.0000
TO	0.3139	0.0503	6.236	0.0000
NIO	0.4208	0.0348	12.087	0.0000
NIT	0.3779	0.0642	5.884	0.0000
ANO06	0.1947	0.0242	8.035	0.0000
ANO07	0.4551	0.0242	18.780	0.0000
DZSU	0.4716	0.0310	15.220	0.0000
lFRBV	0.7467	0.0622	11.997	0.0000

Note que consideramos a distribuição gama para a variável resposta e função de ligação logarítmica, visto que esta combinação apresentou os melhores resultados dentre

as possibilidades oferecidas pela classe de modelos lineares generalizados.[17]

Destaca-se também que os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível de 1% quando utilizado o teste t, exceto para LAT (valor-p = 0.5295) – razão pela qual esta variável foi excluída do modelo. Acrescenta-se ainda que os mesmos sinais das estimativas para os coeficientes do Modelo (1.4) (via CNLRM) também foram obtidos pelo GLM. Entretanto, o uso da distribuição gama, ao invés da normal, resultou numa leve melhora no ajuste dos dados (vide Figura 5.15).

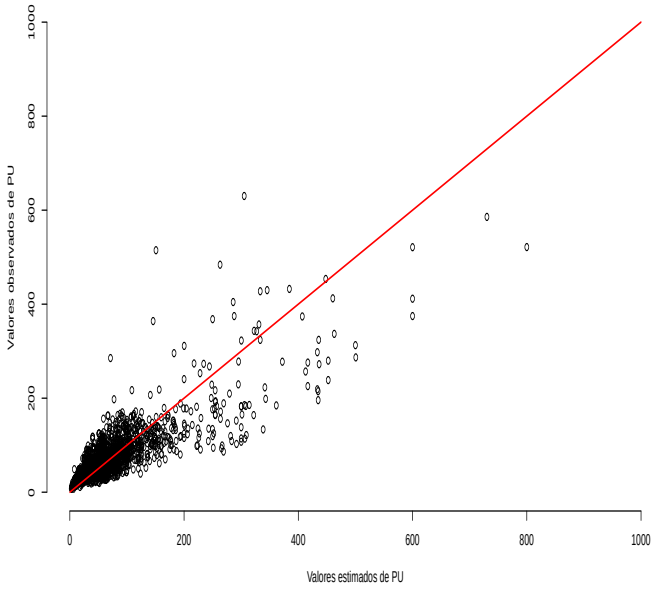

Figura 5.15: Gráfico dos *valores observados* × *valores preditos* de PU - Modelo (2.1).

5.4.3 A modelagem via GAMLSS

Conforme salientado no Capítulo 4, na classe de modelos GAMLSS a premissa de que a variável resposta pertence à família exponencial é relaxada e substituída por uma família de distribuições mais geral $\mathcal{D}$. Além disso, a parte sistemática do modelo é amplificada para permitir a modelagem não apenas da média (ou posição), mas de todos os parâmetros da distribuição condicional de y, por meio de funções paramétricas ou não-paramétricas

[17] Resultados semelhantes foram obtidos por Dantas & Cordeiro (1988) ao analisar uma amostra composta de 50 dados de lotes urbanos situados em Recife-PE.

das variáveis explanatórias e/ou termos de efeitos aleatórios, o que confere flexibilidade extra ao modelo. Note que a classe de modelos GLM é um caso particular da estrutura de regressão GAMLSS.

O processo de construção e seleção de um modelo GAMLSS consiste em comparar diversos modelos concorrentes em que diferentes combinações dos componentes $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ são utilizadas (vide Seção 4.7). Entretanto, a tarefa de escolha dos componentes acima mencionados em busca do modelo mais adequado aos dados não é trivial e requer, além de experiência e familiaridade do pesquisador com o problema, um *software* confiável e que forneça resultados em curto espaço de tempo. Neste sentido, utilizamos o *software* livre `R` e lançamos mão de poderosas funções disponíveis no pacote `gamlss` (por exemplo, `stepGAIC()`, `stepGAIC.VR()`, `stepGAIC.CH()`, `find.hyper()`, `histDist()`, entre outras; vide Rigby & Stasinopoulos, 2008) e na *biblioteca* `MASS` (como `addterm()` e `dropterm()`; vide Venables & Ripley, 2002).

A construção dos modelos consistiu das seguintes etapas: (i) identificação das distribuições plausíveis para a variável resposta; (ii) escolha da função de ligação para modelar o parâmetro de posição (μ); (iii) aplicação da técnica *stepwise* de seleção de covariáveis para modelar μ; (iv) inclusão de termos aditivos não-paramétricos, a exemplo de *splines*; (v) escolha da função de ligação para modelar o parâmetro de escala (σ); (vi) aplicação da técnica *stepwise* de seleção de covariáveis para modelar σ.

5.4.3.1. Modelagem do parâmetro de posição (μ)

A identificação das distribuições plausíveis para a variável resposta pode ser facilitada com o auxílio da função `histDist()` do pacote `gamlss` do `R`, que com base no histograma de frequências da variável dependente estima a função densidade de probabilidade de forma não-paramétrica e ajusta a distribuição (paramétrica) que suspeitamos ser mais adequada aos dados. Assim, visto que a variável PU assume apenas valores positivos, elegemos as distribuições log-normal (LOGNO), gaussiana inversa (IG), Weibull (WEI) e gama (GA)[18] como potenciais candidatas ao ajuste da variável resposta (vide

[18] Aqui, a função densidade de probabilidade da distribuição gama, denotada por GA (μ, σ), é defnida por

$$f_Y(y|\mu,\sigma) = \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{\frac{1}{\sigma^2}-1} e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}$$

para $y > 0$, em que $\mu > 0$ e $\sigma > 0$. Temos que $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \mu^2\sigma^2$ (Johnson *et al.*, 1994).

Figura 5.16). Conforme podemos observar na referida figura, as distribuições gama, log-normal e gaussiana inversa, traçadas na cor vermelha, parecem estar mais próximas da função densidade de probabilidade estimada não-parametricamente (traçada na cor azul), indicando que estas distribuições apresentam uma maior aderência aos dados.

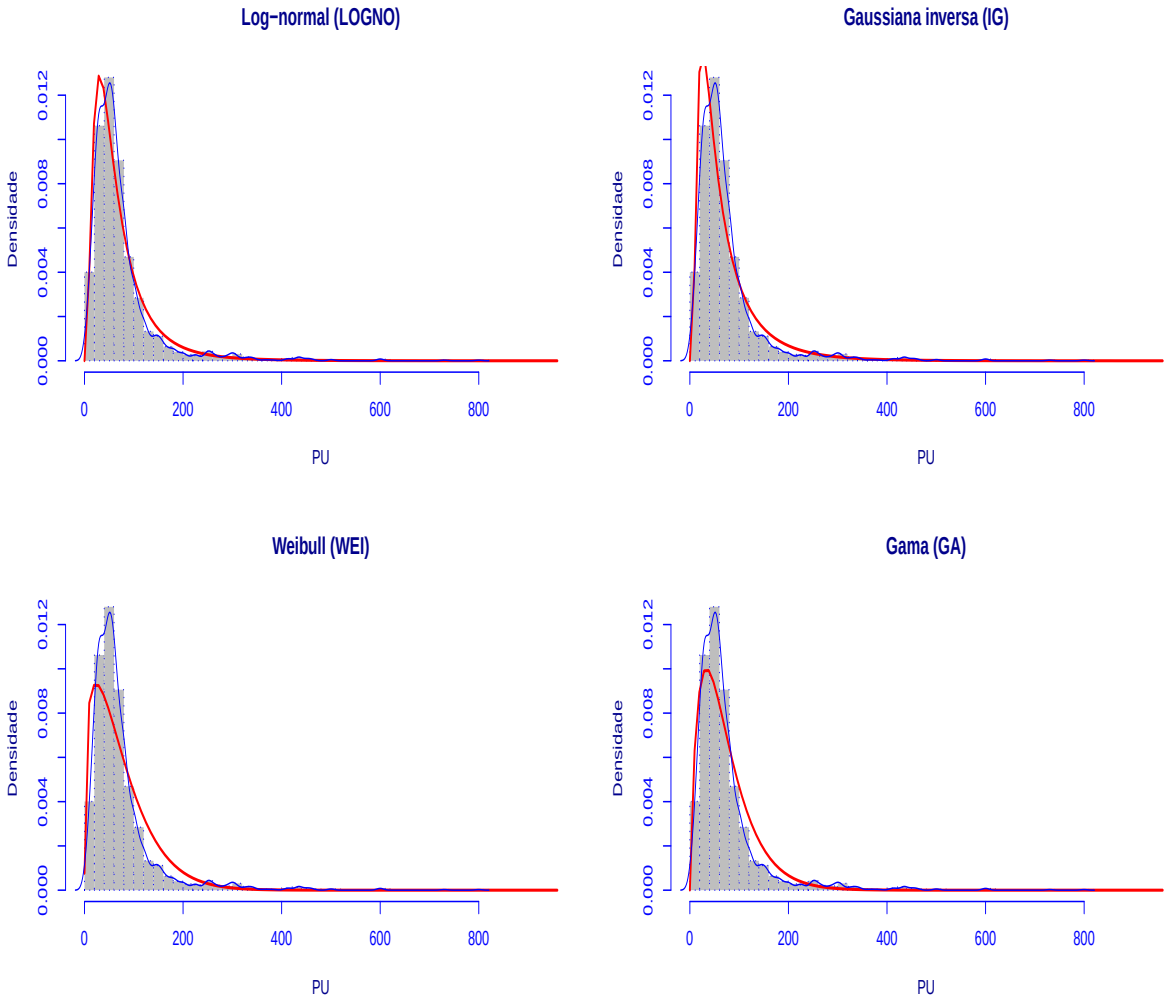

Figura 5.16: Ajustes das distribuições LOGNO, IG, WEI e GA à variável resposta (PU).

A inspeção visual resultante da aplicação da função `histDist()` serve, todavia, apenas como um "ponto de partida", visto que este procedimento não leva em consideração a existência de variáveis explanatórias. Por isto, faz-se necessário realizar as 6 (seis) etapas mencionadas no terceiro parágrafo desta seção para cada possível distribuição assumida pela variável resposta, a fim de comparar os modelos resultantes segundo critérios obje-

tivos (por exemplo, *desvio global* (GD), AIC e SBC) e análises gráficas (por exemplo, o gráfico dos quantis normais dos resíduos). Apresentamos na Tabela 5.8 os principais modelos considerados com o objetivo de modelar o parâmetro μ e os respectivos comentários acerca dos ajustes.

Tabela 5.8: Modelos ajustados via GAMLSS

Modelos	$\mathcal{D}$	$\mathcal{G}$	Forma funcional	Considerações
3.1	LOGNO	logarítmica	$\text{PU} = \beta_0 + \text{cs(LAT)} + \text{cs(LONG)} + \text{cs(log(AR))} + \text{cs(CA)} + \text{cs(ST)} + \beta_1\text{VIAP} + \beta_2\text{VIAS} + \beta_3\text{SI} + \beta_4\text{PA} + \beta_5\text{TO} + \beta_6\text{NIO} + \beta_7\text{NIT} + \beta_8\text{ANO06} + \beta_9\text{ANO07} + \beta_{10}\text{DZSU} + \text{cs(log(FRBV))}$	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível de 1% quando utilizado o teste t. AIC=19155, SBC=19359 e GD=19083.
3.2	IG	logarítmica	$\text{PU} = \beta_0 + \text{cs(LAT)} + \text{cs(LONG)} + \text{cs(log(AR))} + \text{cs(CA)} + \text{cs(ST)} + \beta_1\text{VIAP} + \beta_2\text{VIAS} + \beta_3\text{SI} + \beta_4\text{PA} + \beta_5\text{TO} + \beta_6\text{NIO} + \beta_7\text{NIT} + \beta_8\text{ANO06} + \beta_9\text{ANO07} + \beta_{10}\text{DZSU} + \text{cs(log(FRBV))}$	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível de 1% quando utilizado o teste t. AIC=19845, SBC=20048 e GD=19773.
3.3	WEI	logarítmica	$\text{PU} = \beta_0 + \text{cs(LAT)} + \text{cs(LONG)} + \text{cs(log(AR))} + \text{cs(CA)} + \text{cs(ST)} + \beta_1\text{VIAP} + \beta_2\text{VIAS} + \beta_3\text{SI} + \beta_4\text{PA} + \beta_5\text{TO} + \beta_6\text{NIO} + \beta_7\text{NIT} + \beta_8\text{ANO06} + \beta_9\text{ANO07} + \beta_{10}\text{DZSU} + \text{cs(log(FRBV))}$	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível de 1% quando utilizado o teste t. AIC=19260, SBC=19463 e GD=19188.
3.4	GA	logarítmica	$\text{PU} = \beta_0 + \text{cs(LAT)} + \text{cs(LONG)} + \text{cs(log(AR))} + \text{cs(CA)} + \text{cs(ST)} + \beta_1\text{VIAP} + \beta_2\text{VIAS} + \beta_3\text{SI} + \beta_4\text{PA} + \beta_5\text{TO} + \beta_6\text{NIO} + \beta_7\text{NIT} + \beta_8\text{ANO06} + \beta_9\text{ANO07} + \beta_{10}\text{DZSU} + \text{cs(log(FRBV))}$	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível de 1% quando utilizado o teste t. AIC=19062, SBC=19337 e GD=19062.

Com base na Tabela 5.8 esclarecemos que os modelos ajustados utilizaram suavizadores *splines* cúbicos (cs) com 3 (três) graus de liberdade efetivos nas covariáveis LAT, LONG, logAR, CA, ST e logFRBV. Acrescenta-se ainda que outros suavizadores (por exemplo, *loess* e *splines* penalizados), bem como diferentes combinações de $\mathcal{D}$ (por exemplo, BCPE, BCCG, LNO, BCT, exGAUSS, entre outras) e de $\mathcal{G}$ (por exemplo, identidade, inversa, recíproca, entre outras), foram avaliados, mas não apresentaram resultados superiores àqueles exibidos na Tabela 5.8. Ainda com base nesta tabela, observamos que o Modelo (3.4) apresentou os melhores resultados no que tange aos critérios GD, AIC e SBC. Diante disto, exibimos na Tabela 5.9 o ajuste referente a este modelo e relativo à estimação da equação de preços hedônicos.

Embora as funções estimadas não-parametricamente utilizando 3 (três) graus de liberdade (*degrees of freedom* — df) efetivos em todas as funções suavizadoras tenham conduzido a um ajuste razoável da equação de preços hedônicos, é possível, com o auxílio da função `find.hyper`, implementada no pacote `gamlss` do `R`, obter o número de graus de

liberdade "ótimo" para os suavizadores. Esta seleção baseia-se na minimização do critério AIC e a convergência do algoritmo pode ser bastante lenta dependendo do tamanho do conjunto de dados e do número de parâmetros a otimizar. Neste sentido, reestimamos o Modelo (3.4) levando em consideração dois aspectos: o emprego da função `find.hyper` e a inspeção visual das curvas suavizadas — este último aspecto teve por objetivo evitar "sobreajustamentos" (*overfitting*). O "novo" modelo estimado (Modelo (3.5)) também lançou mão dos suavizadores *splines* cúbicos (cs), porém com diferentes graus de liberdade (df) efetivos nas funções alisadoras, conforme sugerido pela função `find.hyper` e destacado na Tabela 5.10. Salienta-se que houve uma considerável redução — em relação ao Modelo (3.4) — nos valores do AIC, SBC e GD (18822, 19212 e 18684, respectivamente) e uma significativa melhora no ajuste do gráfico entre os *valores observados* $\times$ *valores preditos* (vide Figura 5.17).

Tabela 5.9: Ajuste do modelo de preços hedônicos via GAMLSS - Modelo (3.4).

	Estimativa	Erro-padrão	Estatística t	valor-p
(Intercepto)	−165.4000	16.1300	−10.251	0.0000
cs(LAT)	5.17e-05	6.22e-06	8.307	0.0000
cs(LONG)	1.51e-05	2.13e-06	7.071	0.0000
cs(lAR)	−0.2317	0.0096	−24.074	0.0000
cs(ST)	0.0465	0.0037	12.416	0.0000
cs(CA)	0.1223	0.0206	5.947	0.0000
VIAP	0.3133	0.0349	8.963	0.0000
VIAS	0.0926	0.0364	2.545	0.0100
SI	0.0920	0.0227	4.054	0.0000
PA	0.1891	0.0195	9.670	0.0000
TO	0.2662	0.0474	5.951	0.0000
NIO	0.4135	0.0395	13.362	0.0000
NIT	0.3485	0.0571	6.102	0.0000
ANO06	0.1645	0.0215	7.632	0.0000
ANO07	0.4358	0.0215	20.235	0.0000
cs(lFRBV)	0.6513	0.0569	11.443	0.0000
DZSU	0.3875	0.0299	12.935	0.0000

Tabela 5.10: Ajuste do modelo de preços hedônicos via GAMLSS - Modelo (3.5).

	Estimativa	Erro-padrão	Estatística t	valor-p
(Intercepto)	−130.1000	14.8100	−8.787	0.0000
cs(LAT, df=10)	5.92e-05	5.71e-06	10.354	0.0000
cs(LONG, df=10)	1.05e-05	1.96e-06	5.352	0.0000
cs(lAR, df=10)	−0.2559	8.83e-03	−28.963	0.0000
cs(ST, df=8)	0.0373	3.44e-03	10.831	0.0000
cs(CA, df=3)	0.1769	0.0188	9.370	0.0000
VIAP	0.2571	0.0320	8.012	0.0000
VIAS	0.0728	0.0334	2.180	0.0293
SI	0.1029	0.0208	4.940	0.0000
PA	0.1436	0.0179	7.999	0.0000
TO	0.1822	0.0410	4.436	0.0000
NIO	0.4173	0.0284	14.690	0.0000
NIT	0.3388	0.0524	6.462	0.0000
ANO06	0.1373	0.0198	6.941	0.0000
ANO07	0.4190	0.0197	21.190	0.0000
cs(lFRBV, df=10)	0.6599	0.0522	12.630	0.0000
DZSU	0.5119	0.0275	18.613	0.0000

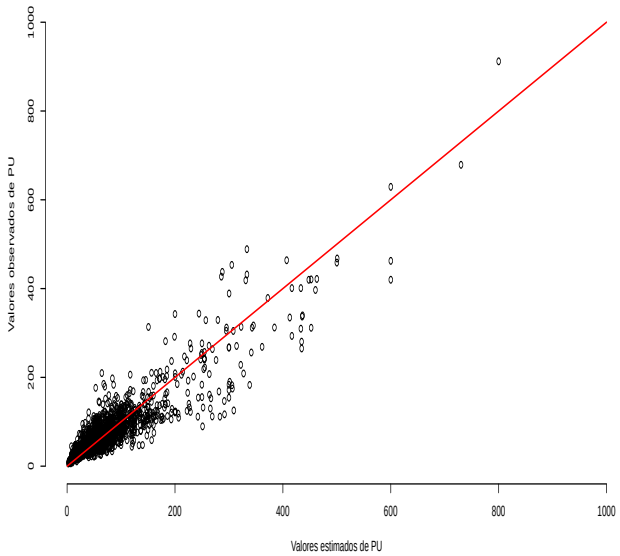

Figura 5.17: Gráfico dos *valores observados* × *valores preditos* de PU - Modelo (3.5).

5.4.3.2. Interpretação dos coeficientes estimados em relação à posição (μ)

Embora os coeficientes estimados correspondentes às curvas de suavização do Modelo (3.5) não sejam passíveis de interpretação direta, é possível examinar, em termos bastante gerais, os sinais destes coeficientes e dos parâmetros estimados ($\widehat{\beta}$'s referentes às funções paramétricas do modelo) a fim de confrontar com as expectativas *à priori* e tecer algumas considerações acerca do comportamento do mercado imobiliário em estudo. Neste sentido, fazemos as seguintes análises com base na Tabela 5.10:

- O sinal positivo do coeficiente da variável LAT indica que o preço unitário médio dos terrenos aumenta à medida em que a localização dos lotes se desloca para o leste. Isto é explicado pela influência da proximidade da praia nessa direção;

- O sinal positivo do coeficiente da variável LONG indica que o preço unitário médio dos terrenos aumenta à medida em que a localização dos lotes se desloca para o norte. Isto ocorre porque o bairro do Mosqueiro, vizinho sul de Atalaia, abrange quase 40% da cidade e é o local onde os preços unitários observados estão entre os menores;

- O sinal negativo da variável log(AR) indica que os terrenos com grandes áreas tendem a ter preços unitários menores do que os de pequenas áreas, o que é esperado em condições normais de mercado;

- O sinal positivo do coeficiente da variável CA indica que quanto maior o potencial construtivo do terreno, maior será seu preço unitário médio;

- O sinal positivo do coeficiente da variável ST indica que há expectativa de elevação do preço unitário médio dos terrenos com o aumento do padrão sócioeconômico do setor censitário onde o imóvel está localizado;

- O coeficiente maior da variável VIAP relativamente ao da variável VIAS também é esperado em condições normais de mercado. Isto ocorre porque terrenos situados em vias principais tendem a ser mais valorizados do que aqueles situados em vias secundárias e estes, por sua vez, são mais valorizados do que aqueles localizados em vias terciárias/superiores;

- O sinal positivo dos coeficientes das variáveis PA, SI, TO e DZSU indicam, respectivamente, que os terrenos localizados em ruas pavimentadas, de esquina, com relevo plano e na zona sul da cidade são em média mais valorizados;

- O coeficiente maior da variável NIO relativamente ao da variável NIT também é esperado em condições normais de mercado. Isto ocorre porque os preços unitários de terrenos advindos de oferta tendem a ser maiores do que aqueles provenientes de transação e estes, por sua vez, são superiores aqueles oriundos do ITBI;

- O maior coeficiente da variável ANO07 relativamente ao da variável ANO06 indica que os preços unitários dos terrenos coletados no ano de 2007 são superiores em média àqueles observados em 2006 e estes, por sua vez, são maiores do que aqueles coletados em 2005;

- O sinal positivo do coeficiente da variável log(FRBV) indica que o aumento na dimensão da frente do lote provoca um aumento no preço unitário médio dos terrenos situados nos bairros valorizados (BV) de Aracaju. Isto se deve ao fato destas áreas valorizadas abrangerem uma grande concentração de prédios residenciais e empreendimentos comerciais, onde o "fator" frente pode ser decisivo na implantação de um projeto imobiliário.

Note que embora tenhamos interpretado os sinais dos coeficientes estimados apenas para o Modelo (3.5), estas mesmas conclusões também podem ser estendidas para os Modelos (1.4) e (2.1), logicamente com as restrições de que no Modelo (1.4) a variável LAT não se mostrou significativa ao nível de 10% e de que no Modelo (2.1) a variável LAT não foi considerada – excluída durante a modelagem por não se mostrar estatisticamente significativa.

Adicionalmente, exibimos na Figura 5.18 os gráficos referentes às curvas de suavização dos termos aditivos do Modelo (3.5). É possível verificar por meio destes gráficos os comportamentos e as contribuições aditivas dos termos ajustados de forma não-paramétrica – em relação ao parâmetro de posição (μ) – ao longo dos possíveis valores assumidos pelas variáveis explanatórias. A linha tracejada em azul corresponde aos erros-padrão pontuais *(pointwise standard errors)*.

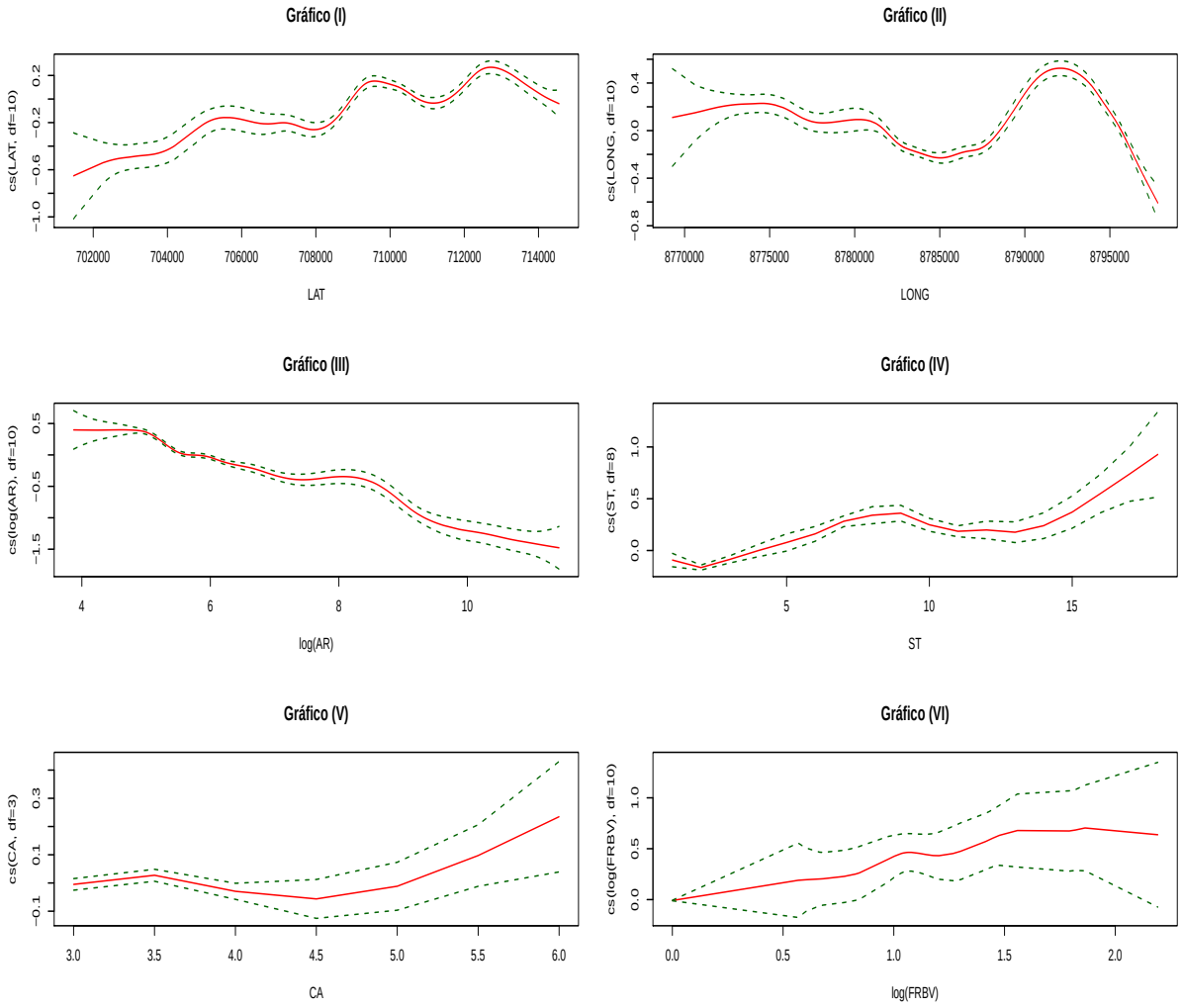

Figura 5.18: Gráficos dos termos aditivos suavizados - Modelo (3.5).

Note que nos Gráficos (I), (II), (III), (IV), (V) e (VI) as funções estimadas indicam que as "contribuições" dos termos aditivos ajustados às covariáveis LAT, LONG, log(AR), ST, CA e log(FRBV) são, em geral, crescentes, crescentes/decrescentes,[19] decrescentes, crescentes, crescentes e crescentes, respectivamente, com os aumentos da latitude, longitude, logaritmo da área, setor sócioeconômico, coeficiente de aproveitamento e logaritmo da frente do terreno nos bairros valorizados, respectivamente. Contudo, perceba que estas mesmas informações também foram fornecidas anteriormente pelos modelos CNLRM,

[19] O Gráfico (II) apresenta alternadamente tendências de crescimento e decrescimento acentuadas, razão pela qual é inapropriado fazer qualquer afirmação sobre a contribuição, ainda que em termos gerais, do termo aditivo ajustado à covariável LONG baseando-se apenas na análise gráfica.

GLM e GAMLSS mediante a verificação dos sinais dos coeficientes estimados para cada regressor, razão pela qual enfatizaremos uma outra abordagem na descrição destes gráficos e que constitui uma importante vantagem dos modelos semiparamétricos em detrimento dos paramétricos: a análise parcial dos termos aditivos suavizados.

No Gráfico (I), perceba que à medida que a latitude aumenta a "contribuição" do termo aditivo ajustado à covariável LAT entre as latitudes 702000 e 709000 (aproximadamente) – onde estão localizados os bairros pertencentes à zona de expansão da cidade –, é negativa, enquanto que a partir da posição 709000 (aproximadamente) – onde estão localizados a Zona Sul e o Centro da cidade de Aracaju – o efeito ocorre de maneira positiva. Além disto, podemos destacar que em alguns intervalos o aumento da latitude provoca uma acentuada mudança na "inclinação" da curva ajustada, como podemos observar entre as posições 708000 e 710000 – correspondente à divisa entre regiões/bairros de padrões sócioeconômicos distintos –, enquanto que em outras zonas, como podemos verificar entre as latitudes 706000 e 708000 – onde se concentram, praticamente, observações de um único bairro, o Mosqueiro –, o aumento da latitude provoca um efeito negativo uniforme ao longo deste intervalo.

No Gráfico (II), note que a "contribuição" do termo aditivo ajustado à covariável LONG, à medida que a longitude aumenta até a posição 8780000, é positiva e praticamente uniforme, uma vez que neste intervalo estão inseridas, praticamente, apenas observações do bairro do Mosqueiro. A partir da posição 8785000 há uma notória mudança de tendência na "inclinação" da curva ajustada – provocada pela localização dos bairros mais nobres da cidade entre as longitudes 8785000 e 8794000 (aproximadamente). Após a posição 8794000, o efeito permanece positivo mas decresce até se tornar negativo.

No Gráfico (III), perceba que à medida que o logaritmo da área aumenta a "contribuição" do termo aditivo ajustado à covariável log(AR), entre os terrenos com áreas (em escala logarítmica) 4 e 5 (aproximadamente), sofre um efeito positivo. Para terrenos com áreas (em escala logarítmica) superior a 5, o efeito é negativo.

No Gráfico (IV), note que à medida que o setor sócioeconômico aumenta a "contribuição" do termo aditivo ajustado à covariável ST, entre o intervalo de 1 a 4 salários mínimos, é negativa, embora a tendência seja crescente. Para terrenos situados em bairros de setor sócioeconômico superior a 4 salários mínimos, o efeito é sempre positivo, apesar

de entre 10 e 15 salários mínimos o efeito ser praticamente uniforme.

No Gráfico (V), perceba que à medida que o coeficiente de aproveitamento aumenta a "contribuição" do termo aditivo ajustado à covariável CA, ao contrário da expectativa *a priori*, não evidenciou efeito positivo sempre crescente. Note que no intervalo de 3.0 a 5.0, a curva ajustada é bastante suave e oscila muito pouco, de forma que há uma alternância entre efeitos positivos e negativos. Somente para coeficientes de aproveitamento superiores à 5.0 verifica-se um efeito positivo crescente.

No Gráfico (VI), note que à medida que o logaritmo da frente dos terreno aumenta nos bairros valorizados a "contribuição" do termo aditivo ajustado à covariável log(FRBV) é sempre positiva. Entretanto, no intervalo de 1.5 a 2.0 este efeito positivo é aproximadamente uniforme.

De acordo com o que foi descrito nas análises dos Gráficos (I), (II), (III), (IV), (V) e (VI) da Figura 5.18, fica evidente o poder do Modelo (3.5) na detecção de efeitos significativos nas relações não-lineares — que não apresentam uma forma definida — presentes nas associações entre o preço unitário (PU) e as variáveis explicativas. Conforme destacado, as associações entre as diversas variáveis intervenientes não apresentaram o mesmo comportamento e sofreram alterações de intensidade e forma ao longo de todos os seus valores do domínio. Dada a complexidade desta interdependência, é razoável imaginar que o uso de modelos estritamente paramétricos — como os Modelos (1.4) e (2.1) — dificilmente corresponderão à realidade, uma vez que apenas as associações lineares entre as variáveis serão avaliadas, o que nem sempre é adequado em estudos de avaliações de bens.

5.5 Escolha do modelo

A fim de compararmos os "melhores" modelos estimados via CNLRM (Modelo (1.4)), GLM (Modelo (2.1)) e GAMLSS (Modelo (3.5)) utilizaremos os critérios AIC e SBC.[20] Adicionalmente, os modelos serão confrontados por meio de um "*pseudo* coeficiente de determinação" (*pseudo*-R^2), o qual será calculado pela expressão

$$pseudo - R^2 = (\text{correlação (valores observados de } PU, \text{valores preditos de } PU))^2. \quad (5.6)$$

[20]Somente será possível a comparação utilizando AIC e SBC entre os modelos que apresentam a variável resposta (PU) na mesma escala de medida, como é o caso dos Modelos (2.1) e (3.5)).

Com base nas considerações anteriores, apresentamos na Tabela 5.11 um resumo comparativo entre os modelos supracitados e claramente percebemos a preponderância do Modelo (3.5) frente aos demais, não apenas pelos menores valores obtidos de AIC e SBC (comparativamente ao Modelo (2.1)), mas pela superioridade expressiva no valor do *pseudo-*R^2.

Tabela 5.11: Tabela-resumo comparativa entre os modelos estimados via CNLRM, GLM e GAMLSS.

Modelo	Classe	AIC	SBC	Pseudo-R^2
1.4	(CNLRM)	4290	4392	0.667
2.1	(GLM)	19486	19581	0.672
3.5	(GAMLSS)	18822	19212	0.811

5.5.1 Modelagem do parâmetro de dispersão (σ)

Uma vez estabelecido um bom modelo para predição de μ, realizamos o teste da razão de verossimilhanças (*likelihood ratio* - LR)[21] para investigar o comportamento — se homoscedástico ou heteroscedástico — do parâmetro de escala σ. Tendo em vista que a hipótese nula de dispersão constante foi rejeitada, segundo o teste LR, modelamos a dispersão (σ) tomando por base o Modelo (3.5), haja vista ter sido este o ajuste que melhor "representou" os dados. Aqui, cumpre registrar que para modelarmos o parâmetro de dispersão adotamos procedimento semelhante ao utilizado anteriormente na modelagem do parâmetro de posição, ou seja, aplicamos a técnica *stepwise* de seleção das covariáveis significativas, testamos possíveis funções de ligação (por exemplo, identidade, inversa, recíproca, entre outras) e incluímos funções de suavização (por exemplo, *splines* cúbicos, *loess* e *splines* penalizados) no termo preditor do parâmetro de dispersão do modelo. Note que os procedimentos citados não foram novamente aplicados ao parâmetro de posição, mas apenas impostos à modelagem do parâmetro de dispersão, conforme sugerido em

[21] O teste LR requer a estimação do modelo restrito (cujo vetor de parâmetros restrito denominamos por $\widetilde{\boldsymbol{\theta}}$) e sem restrição (cujo vetor de parâmetros não-restrito denominamos por $\widehat{\boldsymbol{\theta}}$). O teste LR é baseado no *log* da razão entre as duas verossimilhanças ($\mathrm{L}(\widetilde{\boldsymbol{\theta}})$ e $\mathrm{L}(\widehat{\boldsymbol{\theta}})$), isto é, na diferença entre $\mathrm{logL}(\widetilde{\boldsymbol{\theta}})$ e $\mathrm{logL}(\widehat{\boldsymbol{\theta}})$. Se H_0 é verdadeira, então $\mathrm{LR} = -2[\mathrm{logL}(\widetilde{\boldsymbol{\theta}}) - \mathrm{logL}(\widehat{\boldsymbol{\theta}})] \xrightarrow{d} \chi^2_g$, em que g é o número de restrições, quando $n \to \infty$.

Rigby & Stasinopoulos (2008). Destaca-se ainda que nesta etapa também utilizamos a função `find.hyper` e fizemos a inspeção visual das curvas suavizadas na busca do "melhor" modelo.

Neste sentido, apresentamos na Tabela 5.12 os resultados do ajuste referente ao modelo GAMLSS (Modelo (3.6)) que contempla a modelagem explícita dos parâmetros de posição (μ) e dispersão (σ). Sobre este modelo, salientamos que a variável resposta (PU) segue distribuição gama e as funções de ligação utilizadas para modelar μ e σ são as logarítmicas. Note que o Modelo (3.6) contém termos paramétricos e não-paramétricos, motivo pelo qual é denominado de GAMLSS aditivo semiparamétrico linear.

Tabela 5.12: Ajuste do modelo de preços hedônicos via GAMLSS - Modelo (3.6).

Coeficientes de μ				
	Estimativa	Erro-padrão	Estatística t	valor-p
(Intercepto)	−95.1300	14.2700	−6.665	0.0000
cs(LAT, df=10)	5.94e-05	5.37e-06	11.053	0.0000
cs(LONG, df=10)	6.45e-06	1.86e-06	3.460	0.0000
cs(lAR, df=10)	−0.2087	0.0104	−20.138	0.0000
cs(ST, df=8)	0.0321	0.0030	10.666	0.0000
cs(CA, df=3)	0.2095	0.0161	13.006	0.0000
VIAP1	0.2039	0.0298	6.838	0.0000
VIAS1	0.0729	0.0276	2.635	0.0084
SI1	0.7136	0.0192	3.705	0.0000
PA1	0.1653	0.0157	10.465	0.0000
TO1	0.1778	0.0370	4.799	0.0000
NIO1	0.3722	0.0251	14.799	0.0000
NIT1	0.2790	0.0468	5.957	0.0000
ANO061	0.1255	0.0175	7.144	0.0000
ANO071	0.4195	0.0177	23.622	0.00
cs(lFRBV, df=10)	0.6809	0.0403	16.88	0.0000
DZSU1	0.4824	0.0241	20.001	0.0000
Coeficientes de σ				
(Intercepto)	−1.6838	0.0839	−20.072	0.0000
cs(lAR, df=10)	0.1370	0.0143	9.593	0.0000
ST	−0.0391	0.0040	−9.632	0.0000

Pelos resultados da Tabela 5.12, verificamos que os valores das estimativas dos coeficientes do submodelo da média não sofreram grandes alterações em relação àqueles

obtidos para o Modelo (3.5) (vide Tabela 5.10). Todavia, destacamos que houve uma expressiva redução do GD, AIC e SBC (18445, 18607 e 19065, respectivamente) e, também, uma melhora no comportamento dos resíduos apresentados no gráfico *worm plot*[22] em relação ao Modelo (3.5) (vide Figuras 5.19 e 5.20).

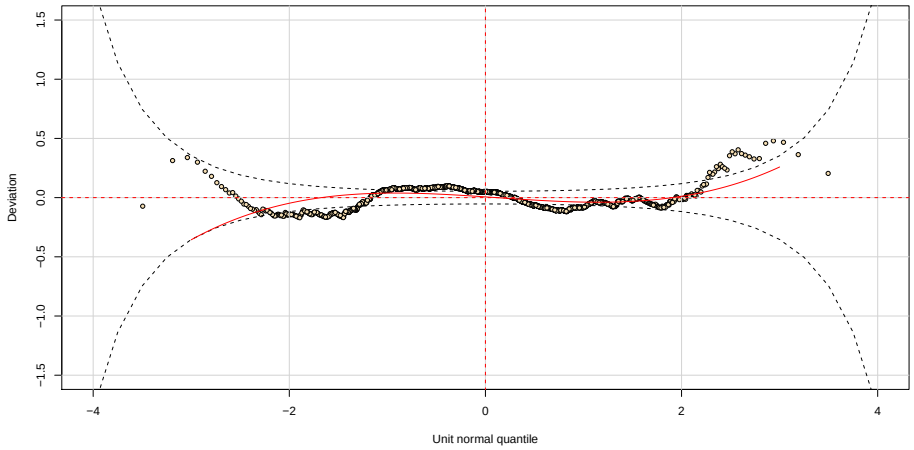

Figura 5.19: Gráfico *worm-plot* - Modelo (3.5).

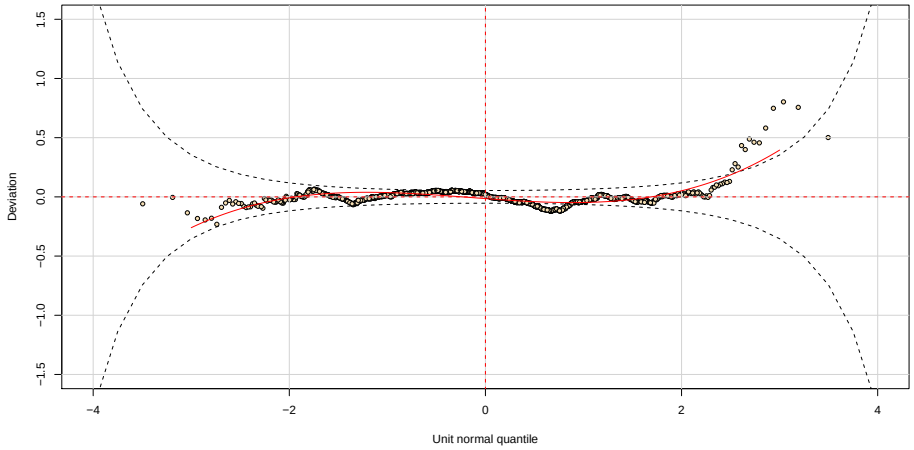

Figura 5.20: Gráfico *worm-plot* - Modelo (3.6).

[22]Gráficos *worm plots* foram introduzidos por van Buuren & Fredriks (2001) e consistem em ferramentas de diagnóstico para análise dos resíduos em diferentes regiões (intervalos) da variável explanatória. Se nenhuma variável explanatória é especificada, o gráfico *worm plot* funciona como o gráfico dos quantis normais dos resíduos sem a tendência. Se os pontos estão situados no interior da região de "aceitação" (entre as duas curvas elípticas), o modelo fornece um bom ajuste.

Acrescenta-se ainda que para o Modelo (3.6) as interpretações – em relação ao parâmetro de posição (μ) – dos sinais dos coeficientes estimados correspondentes às curvas de suavização e aos $\widehat{\beta}$'s referentes às funções paramétricas do modelo são análogas àquelas descritas para o Modelo (3.5) (vide Seção 5.4.3.2). Em se tratando do parâmetro de dispersão (σ), verificamos que apenas 2 (duas) variáveis foram efetivamente consideradas no Modelo (3.6): ST e AR, sendo ST tratada de forma paramétrica e AR ajustada de forma não-paramétrica por meio de uma função suavizadora *spline* cúbica com 10 (dez) graus de liberdade efetivos, ou seja, cs(AR, df10). Acrescentamos, em termos bastante gerais, que o sinal positivo do coeficiente estimado em AR indica que a dispersão de PU é maior entre os terrenos que possuem grandes áreas – pertencentes, em geral, à classe mais abastada e com maior poder aquisitivo –, enquanto que o sinal negativo em ST indica que a variabilidade de PU diminui com o aumento do padrão sócioeconômico do setor censitário onde o imóvel está localizado. Aqui, cabe ressaltar que o comportamento observado da variância em função da covariável ST aparenta refletir mais uma característica intrínseca da amostra coletada do que propriamente do mercado imobiliário de terrenos. Isto pode ser devido ao desequilíbrio observado na amostra no que tange à discrepância da quantidade de terrenos que estão localizados em setores de baixo e alto padrão sócio-econômico do setor censitário, conforme evidenciado na Seção 5.2.

Cumpre registrar ainda que o valor obtido do *pseudo-*R^2 para o Modelo (3.6) foi de 0.817 e que todas as variáveis explicativas mostraram-se estatisticamente significativas ao nível de 1%. Note que estes resultados, substancialmente em relação ao valor alcançado pelo *pseudo-*R^2, em geral, são raros de serem atingidos quando se trabalha com dados de corte transversal e especialmente, nas avaliações imobiliárias em massa. No presente estudo, em que a amostra coletada contempla observações de terrenos situados ao longo de **toda** a cidade de Aracaju e cuja análise exploratória de dados indicou uma acentuada variabilidade entre as características físicas, estruturais e locacionais dos imóveis observados, é apreciável a superioridade da qualidade (vide Figura 5.21 referente ao gráfico dos *valores observados* $\times$ *valores preditos* de PU para o Modelo (3.6)) e do poder de ajuste (*pseudo-*$R^2 = 0.817$) do Modelo GAMLSS (3.6) frente aos métodos tradicionais.

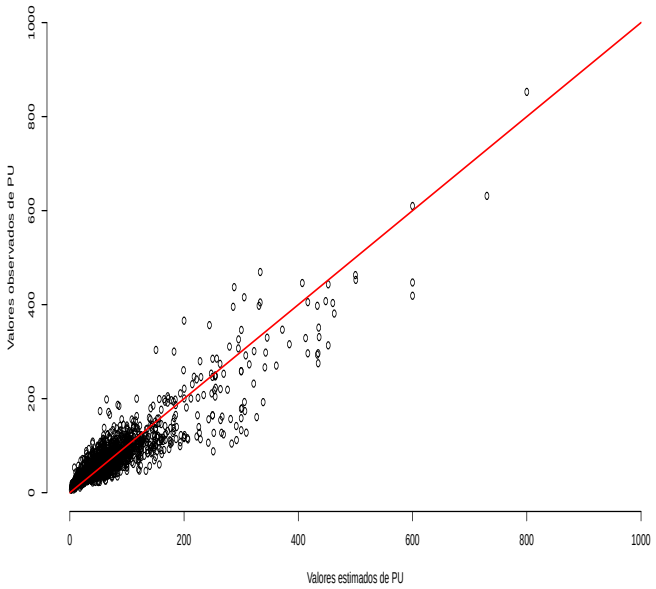

Figura 5.21: Gráfico dos *valores observados* × *valores preditos* de PU – Modelo (3.6).

Em virtude do exposto, o Modelo (3.6) dado por

$$\begin{aligned}
\log(\mu) \;=\; & \beta_0 \;+\; cs(\text{LAT},\ \text{df}=10) \;+\; cs(\text{LONG},\ \text{df}=10) \;+\; cs(\log(\text{AR}),\ \text{df}=10) \;+ \\
& cs(\text{CA},\ \text{df}=3) \;+\; cs(\text{ST},\ \text{df}=8) \;+\; \beta_1\text{VIAP} \;+\; \beta_2\text{VIAS} + \beta_3\text{SI} + \beta_4\text{PA} + \\
& \beta_5\text{TO} + \beta_6\text{NIO} + \beta_7\text{NIT} + \beta_8\text{ANO06} + \beta_9\text{ANO07} + \beta_{10}\text{DZSU} + \\
& cs(\log(\text{FRBV}),\ \text{df}=10), \\
\log(\sigma) \;=\; & \gamma_0 + \gamma_1\text{ST} + cs(\log(\text{AR}), \text{df}=10),
\end{aligned}$$

em que a variável resposta resposta (PU) segue uma distribuição gama (GA) com parâmetro de posição (μ) e de escala (σ), aparenta ser o mais indicado para a estimação da equação de preços hedônicos para terrenos urbanos situados na cidade de Aracaju-SE, pois além de captar a essência do fenômeno estudado e descrever bem os dados, apresentou os melhores resultados numéricos e gráficos.

CAPÍTULO 6

Considerações finais

"... Que tenhamos claro que do conhecimento produzido podemos resolver problemas concretos por meio da tecnologia. Que uma tecnologia pode ser validada, constituindo essa validação um conhecimento científico. Seja como for, os preconceitos só atrapalham e a função mais nobre do cientista é buscar novos conhecimentos, divulgá-los e tentar, com isso, construir uma sociedade melhor. Não se constrói um país livre e independente apenas com ciência "básica", e nem tampouco só com ciência "aplicada". Mas com ciência de bom nível pode-se construir o país ideal."

(Volpato, G., em *Ciência: Da Filosofia à Publicação*, 2007)

6.1 Conclusões

No desenvolvimento deste trabalho foram apresentadas as características e propriedades da classe de modelos de regressão proposta por Rigby & Stasinopoulos (2005), denominada de modelos aditivos generalizados para posição, escala e forma (GAMLSS). Além dos aspectos de inferência e diagnóstico, enfatizou-se a flexibilidade inerente à análise de regressão via GAMLSS, que permite o ajuste de uma ampla família de distribuições para a variável resposta e possibilita a modelagem direta, utilizando funções paramétricas

e/ou não-paramétricas, de todos os parâmetros da distribuição da variável resposta em relação às variáveis explanatórias. Adicionalmente, fizemos uma breve revisão de alguns conceitos fundamentais de regressão, como a distinção entre modelos paramétricos, não-paramétricos e semiparamétricos, e também apresentamos os principais procedimentos e técnicas não-paramétricas de suavização utilizados na estimação de modelos que combinam componentes paramétricos e não-paramétricos, como os métodos *kernel* e *spline*.

O enfoque central desta dissertação consistiu na estimação empírica da equação de preços hedônicos para terrenos urbanos situados em Aracaju-SE com base em modelos GAMLSS. Acrescenta-se que, para o mesmo conjunto de dados, os resultados foram comparados com aqueles obtidos pela aplicação do modelo normal de regressão linear clássico e dos modelos lineares generalizados. As análises realizadas mostraram que os modelos estimados via GAMLSS forneceram um ajuste superior àqueles obtidos via CNLRM e GLM, segundo os critérios de Akaike e Schwarz e as análises dos resíduos (gráficos *worm plot*), indicando que a classe de modelos GAMLSS aparenta ser mais apropriada para a estimação da função de preços hedônicos do que as tradicionais modelagens via CNLRM e GLM.

Outro aspecto que evidenciou a preponderância do modelo GAMLSS foi o valor obtido do *pseudo*-R^2 (=0.817) comparativamente àqueles obtidos via CNLRM (=0.667) e GLM (=0.672). Aqui, cabe destacar além desta superioridade de magnitude "numérica" do *pseudo*-R^2, o considerável poder de ajuste desta classe de modelos mesmo sob dados de corte transversal e com excessiva variabilidade, como são os terrenos que compõem a amostra da análise de dados desta dissertação. Embora a natureza dos dados analisados neste trabalho tenha sugerido a distribuição gama para modelagem da variável resposta – motivo pelo qual modelamos apenas os parâmetros de posição e escala – os modelos GAMLSS possibilitam o ajuste de uma ampla família de distribuições que podem fornecer informações adicionais sobre a assimetria e a curtose, o que não é permitido na modelagem via GLM.

Cumpre registrar ainda que o emprego dos modelos GAMLSS conduziu a ajustes mais realistas (ratificados pelo cálculo do pseudo-R^2) e menos sujeitos à influência e subjetividade do pesquisador, haja vista que ao tratarmos algumas variáveis explanatórias de forma não-paramétrica deixamos que os "dados falassem por si mesmos". Acrescenta-

se que no modelo GAMLSS final adotado (Modelo (3.6)) todas as variáveis explicativas mostraram-se estatisticamente significativas ao nível de 1%, enquanto que no modelo CNLRM a variável latitude (LAT) não se mostrou significativa ao nível de 10% e no modelo GLM a mesma variável latitude (LAT) não foi considerada — excluída durante a modelagem por não se mostrar estatisticamente significativa. Embora os modelos estimados via CNLRM e GLM tenham produzido resultados "coerentes" — no sentido da ratificação das expectativas *a priori* sobre os sinais dos coeficientes estimados —, nestas análises as associações avaliadas entre a variável dependente (PU) e os regressores são estritamente paramétricas e lineares, as quais podem não ser adequadas para o fenômeno estudado, conforme resultados apresentados ao longo deste trabalho. É fato conhecido da teoria que a adoção de formas funcionais equivocadas ou omissão de variáveis independentes importantes resultam em erros de especificação do modelo, sobre o qual a validade das interpretações e estimativas dos parâmetros são altamente questionáveis.

Vale salientar que o uso da classe de modelos GAMLSS na Engenharia de Avaliações não deve ser confundido com "refinamento", "preciosismo" ou "sofisticação" da análise de regressão e da valoração de bens, mas método eficiente de modelagem fruto de técnicas avançadas da pesquisa científica que aumentam a acurácia do trabalho avaliatório. Os modelos GAMLSS constituem atualmente uma das ferramentas estatísticas mais poderosas para análise de dados univariados com estrutura de regressão e parecem ser bastante promissores para o mercado imobiliário. A relevância dos modelos GAMLSS não é apenas de índole prática. Do ponto de vista teórico, a sua importância advém, essencialmente, do fato de a metodologia destes modelos constituir uma abordagem unificada de muitos procedimentos estatísticos.

6.2 Utilidade do estudo

O emprego de métodos estatísticos mais flexíveis e que são capazes de descrever com maior grau de adequação as inter-relações entre variáveis tem sido cada vez mais "exigido" pelo mercado imobiliário. Por isto e conforme demonstrado neste trabalho, a classe de modelos GAMLSS surge como uma ferramenta poderosa para lidar com as peculiaridades intrínsecas do bem imóvel e com as limitações presentes nos modelos tradicionais (CNLRM

e GLM). De imediato, elencamos três contribuições deste trabalho para a comunidade acadêmica e para sociedade:

1. Trata-se de trabalho inovador no Brasil (e também no exterior) em que se estuda o uso dos modelos GAMLSS na Engenharia de Avaliações – constitui, portanto, um dos primeiros textos em português sobre o assunto. Diante disto, esperamos despertar e instigar entre os pesquisadores e profissionais atuantes no mercado imobiliário as potencialidades e benefícios dos modelos GAMLSS no que tange aos ganhos de precisão e melhoria na qualidade do ajuste de funções de preços hedônicos, bem como evidenciar a aplicabilidade da Estatística nas avaliações imobiliárias – ainda carente de capital humano especializado. Salienta-se que no Brasil não é prática publicar os estudos referentes a avaliações imobiliárias, o que justifica a quantidade ínfima de discussões sobre o tema na comunidade científica nacional;

2. A atual norma de avaliação de bens para imóveis urbanos (NBR 14653 - Parte 2) não aborda a análise de dados utilizando regressão não-paramétrica ou semiparamétrica, ao contrário do que ocorre para a regressão paramétrica via modelos lineares. Almejamos com este trabalho incluir os modelos GAMLSS nas próximas discussões de revisão da norma e, a partir disto, torná-los ainda mais difundidos entre engenheiros e arquitetos especialistas em avaliações. Desta forma, esperamos contribuir com o crescimento técnico-científico da Engenharia de Avaliações no país;

3. A metodologia GAMLSS exposta neste trabalho pode ser de grande utilidade para os diversos órgãos públicos e privados já citados, principalmente para instituições financeiras responsáveis pela execução de políticas públicas do governo federal, como o Banco do Nordeste do Brasil S.A.[1] e a Caixa Econômica Federal,[2] na estimação de modelos de regressão que subsidiem a tomada de decisão nas operações creditícias que envolvem avaliação de imóveis (por exemplo, garantia hipotecária das operações

[1] O Banco do Nordeste do Brasil S.A. (BNB) é o maior banco de desenvolvimento regional da América Latina e diferencia-se das demais instituições financeiras pela missão que tem a cumprir: atuar, na capacidade de instituição financeira pública, como agente catalisador do desenvolvimento sustentável do Nordeste, integrando-o na dinâmica da economia nacional.

[2] A Caixa Econômica Federal (CEF) atua em todo o território nacional e prioriza os setores como habitação, saneamento básico, infra-estrutura e prestação de serviços. A CEF exerce um papel fundamental na promoção do desenvolvimento urbano e da justiça social no país, contribuindo para melhorar a qualidade de vida da população, especialmente a de baixa renda.

de financiamento). Uma outra aplicação interessante diz respeito à elaboração de plantas genéricas de valores pela prefeitura para fins de cobranças do IPTU e ITBI, favorecendo uma política fiscal mais justa para o município e contribuintes. Aqui, o desafio é promover mais equidade (maior uniformidade dos níveis de avaliação entre imóveis distintos).

6.3 Sugestões para novas pesquisas

Evidentemente este trabalho não esgotou a teoria e multiplicidade de aplicações dos modelos GAMLSS, razão pela qual sugerimos para o desenvolvimento de trabalhos futuros:

- Análise dos dados com base em técnicas de estimação centílica via modelos GAMLSS;
- Avaliação comparativa entre modelos GAMLSS que incluem simultaneamente funções lineares e não-lineares (nos parâmetros) no mesmo modelo;
- Devido à existência de pesquisas recentes que sugerem a presença de correlação espacial em dados imobiliários (vide, por exemplo, Dantas, 2003), recomendamos que seja investigada a incorporação dos efeitos da dependência espacial utilizando modelos GAMLSS. Esta é uma combinação (modelos espacias + modelos GAMLSS) que aparenta ser bastante promissora, visto que a flexibilidade característica dos modelos GAMLSS pode auxiliar na especificação da matriz de pesos espaciais[3] (geralmente construída de maneira *ad hoc*) e na captação de efeitos de anisotropia (caso em que a estrutura espacial do fenômeno varia conforme a direção), possibilitando ajustes ainda mais fidedignos ao comportamento do mercado imobiliário.

[3]Também denominada de matriz de proximidade espacial ou matriz de vizinhanças (W). Corresponde a uma matriz quadrada que estima a variabilidade espacial de dados de área, em que cada elemento w_{ij} representa uma medida de proximidade entre A_i e A_j, sendo A_i e A_j as zonas que estão sendo analisadas.

Referências Bibliográficas

[1] Aguirre, A. & Macedo, P.B.R. (1996). Estimativas de Preços Hedônicos para o Mercado Imobiliário de Belo Horizonte. *Anais do XVIII Encontro Brasileiro de Econometria* 1, 1–16. Águas de Lindóia-SP.

[2] Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 55, 117–128.

[3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716–723.

[4] Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute* 50, 277–290.

[5] Akantziliotou, C.; Rigby, R.A. & Stasinopoulos, D.M. (2002). The `R` implementation of generalized additive models for location scale and shape. In *Statistical modelling in Society: Proceedings of the 17th International Workshop on Statistical Modelling.* Eds: Stasinopoulos, M. and Touloumi, G., 75–83. Chania, Greece.

[6] Akantziliotou C.; Rigby, R.A. & Stasinopoulos, D.M. (2006). Instructions on how to use the GAMLSS package in `R`. *Technical Report 01/06.* STORM Research Centre, London Metropolitan University, London.

[7] Anderson, T.W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics* 127, 1–16.

[8] Anglin, P. & Gencay, R. (1996). Semiparametric estimation of hedonic price functions. *Journal of Applied Econometrics* 11, 633–648.

[9] Ayres, A. (1996). *Como Avaliar Imóveis*. São Paulo: Editora Imobiliária S/C Ltda.

[10] Barbosa, E.P. & Bidurin, C.P. (1991). Seleção de modelos de regressão para predição via validação cruzada: uma aplicação na avaliação de imóveis. *Revista Brasileira de Estatística* 52, 105–120.

[11] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

[12] Benjamin, M.; Rigby, R.A. & Stasinopoulos, D.M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association* 98, 214–223.

[13] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

[14] Beyerlein, A.; Fahrmeir, L.; Mansmann, U. & Toschke, M.A. (2008). Alternative regression models to assess increase in childhood BMI. *BMC Medical Research Methodology*, 8:59.

[15] Bhattacharya, P.K & Zao, P.L. (1997). Semiparametric inference in a partial linear model. *Annals of Statistics* 25, 244–262.

[16] Bidurin, C.P & André, L.A. (2001). Modelos semiparamétricos: uma revisão. *Revista Brasileira de Estatística* 62, 71–90.

[17] de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.

[18] Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. New York: Oxford University Press.

[19] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* 26, 211–252.

[20] Box, G.E.P. & Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. New York: Wiley.

[21] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88, 9–25.

[22] Breslow, N.E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* 82, 81–91.

[23] Buja, A.; Hastie, T. & Tibshirani, R. (1989). Linear smoothers and additive models. *Annals of Statistics* 17, 453–510.

[24] Clapp, J.M.; Kim, H.J. & Gelfand, A. (2002). Predicting spatial patterns of house prices using LPR and bayesian smoothing. *Real Estate Economics* 30, 505–532.

[25] Cleveland, W.S.; Grosse, E. & Shyu, M.J. (1992). Local regression models. *In Statistical Modelling in S*. Eds: Chambers, J.M. and Hastie, T.J., 309–376. New York: Chapman and Hall.

[26] Cole, T.J. & Green, P.J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* 11, 1305–1319.

[27] Cunha, M.C. (2000). *Métodos Numéricos*, 2ª ed. São Paulo: Unicamp.

[28] Dantas, R.A. & Cordeiro G.M. (1988). Uma nova metodologia para avaliação de imóveis utilizando modelos lineares generalizados. *Revista Brasileira de Estatística* 191, 27–46.

[29] Dantas, R.A. & Cordeiro, G.M. (2000). Uma avaliação do mercado de apartamentos do Recife utilizando modelos lineares generalizados. *XIX Congresso Panamericano de Avaliações*, Margarita, Venezuela.

[30] Dantas, R.A. & Cordeiro G.M. (2001). Evaluation of the Brazilian city of Recife's condominium market using generalized linear models. *The Appraisal Journal* 69, 247–257.

[31] Dantas, R.A. (2003). *Modelos Espaciais Aplicados ao Mercado Habitacional: Um Estudo de Caso Para a Cidade do Recife*. Tese (Doutorado em Economia - Área de concentração: Métodos quantitativos) - Universidade Federal de Pernambuco (UFPE), Recife.

[32] Dantas, R.A. (2005). *Engenharia de Avaliações: Uma Introdução à Metodologia Científica*, 2ª ed. São Paulo: Pini.

[33] Davidson, R. & MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. New-York: Oxford University Press.

[34] Davidson, R. & MacKinnon, J.G. (2004). *Econometric Theory and Methods*. New-York: Oxford University Press.

[35] Dias, R. (2001a). *Tutorial em Métodos Paramétricos para Estimação de Curvas*. Disponível na internet em: `http://www.ime.unicamp.br/~dias/np.html`. Arquivo obtido em 18 de agosto de 2009.

[36] Dias, R. (2001b). *Regressão Não-Paramétrica*. Disponível na internet em `http://www.ime.unicamp.br/~dias/np.html`. Arquivo obtido em 18 de agosto de 2009.

[37] Dias, R. (2001c). *O uso de Splines em Regressão Não-Paramétrica*. Disponível na internet em: `http://www.ime.unicamp.br/~dias/np.html`. Arquivo obtido em 18 de agosto de 2009.

[38] Dunn, P.K. & Smyth, G.K. (1996). Randomised quantile residuals. *Journal of Computational and Graphical Statistics* 5, 236–244.

[39] Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society* B 57, 45–97.

[40] Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science* 11, 89–121.

[41] Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

[42] Eubank, R.L. (1994). A simple smoothing spline. *American Statistician* 2, 103–106.

[43] Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed. New York: Springer.

[44] Fávero, L.P.L.; Belfiore, P.P. & Lima, G.A.S.F. (2008). Modelos de precificação hedônica de imóveis residenciais na Região Metropolitana de São Paulo: uma abordagem sob as perspectivas da demanda e da oferta. *Estudos Econômicos* 38, 73–96.

[45] Ferreira, J. (2008). *Modelos de Previsão de Perdas para Crédito Massificado*. Dissertação (Mestrado em Economia - Área de concentração: Finanças) - Faculdade IBMEC São Paulo.

[46] Fiker, J. (1997). *Avaliação de Imóveis Urbanos*, 5ª ed. São Paulo: Pini.

[47] Fix, E. & Hodges Jr., J. (1951). Discriminatory analysis: nonparametric discrimination: consistency properties. *Report n°. 4*, USAF School of Aviation Medicine, Randolph Field, TX.

[48] Friedman, J.H. & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association* 76, 817–823.

[49] Florencio, L.A (2006). *Avaliação de Imóveis Urbanos: a Engenharia Civil a Serviço de uma Instituição Bancária*. Projeto final (graduação em Engenharia Civil) - Escola Politécnica da Universidade de Pernambuco (POLI-UPE).

[50] Gencay, R. & Yang, X. (1996). A forecast comparison of residential housing prices by parametric and semiparametric conditional mean estimators. *Economic Letters* 52, 129–135.

[51] Godfrey, L.G. Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics and Data Analysis* 50, 2715–2733.

[52] Gomide, T.L.F. (2007). Panorama geral e importância jurídica. In: Instituto Brasileiro de Avaliações e Perícias de Engenharia de São Paulo. *Engenharia de Avaliações*, São Paulo: Pini.

[53] Grandiski, P. & Oliveira A.M.B.D. (2007). Engenharia de Avaliações. In: Instituto Brasileiro de Avaliações e Perícias de Engenharia de São Paulo. *Engenharia de Avaliações*, São Paulo: Pini.

[54] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

[55] Gujarati, D.N. (2006). *Basic Econometrics*, 4th ed. Nova York: McGraw-Hill.

[56] Halvorsen, R. & Palmquist, R. (1980). The interpretation of dummy variables in semilogarithmic equations. *American Economic Review* 70, 474–475.

[57] Handscomb, D.C. (1966). Spline functions. *In Methods of Numerical Approximation*. Oxford: Pergamon Press.

[58] Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

[59] Härdle, W.; Müller, M.; Sperlich, S. & Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Berlin: Springer-Verlag.

[60] Hartog, J. & Bierens, H. (1991). Estimating a hedonic earnings function with a nonparametric method. In *Semiparametric and Nonparametric Econometrics: Studies in Empirical Economics*. Ed: Ullah, A., New York: Springer.

[61] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

[62] Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society* B 55, 757–796.

[63] Hastie, T.; Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

[64] Hjort, N.L. & Claeskens, G. (2003). Frequentist model average estimation. *Journal of the American Statistical Association* 98, 879–899.

[65] Iwata, S.; Murao, H. & Wang, Q. (2000). Nonparametric assessment of the effects of neighborhood land uses on the residential house values. In: *Advances in Econometrics: Applying Kernel and Nonparametric Estimation to Economic Topics*. Eds: Fomby, T. and Carter, H.R. New York: JAI Press.

[66] Johnson, N.L.; Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, volume I, 2nd ed. Wiley, New York.

[67] Lamport, L. (1994). *A Document Preparation System LATEX, User's Guide and Reference Manual*, 2nd ed. Massachusett: Addison-Wesley.

[68] Lancaster, K.J. (1966). A new approach to consumer theory. *Journal of Political Economy* 74, 132–157.

[69] Lee, D.K.C. (1990). Cross-validation in semiparametric models: some Monte Carlo results. *Journal of Statistical Computation and Simulation* 37, 171–187.

[70] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society B* 58, 619–678.

[71] Lee, Y. & Nelder, J.A. (2001a). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika* 88, 987–1006.

[72] Lee, Y. & Nelder, J.A. (2001b). Modelling and analysing correlated non-normal data. *Statistical Modelling* 1, 3–16.

[73] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

[74] Lima, L.P; André, C.D.S & Singer, J.M. (2001). Modelos aditivos generalizados: metodologia e prática. *Revista Brasileira de Estatística* 62, 37–69.

[75] Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society* B 61, 381–400.

[76] Liporoni, A.S. (2007). Avaliação em massa com ênfase em planta de valores. In: Instituto Brasileiro de Avaliações e Perícias de Engenharia de São Paulo. *Engenharia de Avaliações*, São Paulo: Pini.

[77] Lopatatzidis, A. & Green, P.J. (2000). Nonparametric quantile regression using the gamma distribution. *Submetido para publicação.*

[78] MacKinnon, J.G & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics* 29, 305-325.

[79] Maddala, G.S. (2003). *Introdução à Econometria.* Rio de Janeiro: LTC.

[80] Madigan, D. & Raftery, A.E. (1994). Model selection and accounting formodel uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89, 1535–1546.

[81] Marquetti, A. & Vialli, L. (2004). Princípios e aplicações da regressão local. *Análise Econômica* 22, 253–277.

[82] Martins-Filho, C. & Bin, O. (2005). Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics* 30, 93–114.

[83] Nelder, J.A & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society* A 135, 370–384.

[84] Pace, R.K. (1993). Nonparametric methods with applications to hedonic models. *Journal of Real Estate Finance and Economics* 7, 185–204.

[85] Pace, R.K. (1995). Parametric, semiparametric, and nonparametric estimation of characteristics values within mass assessment and hedonic pricing models. *Journal of Real Estate Finance and Economics* 11, 195–217.

[86] Pace, R.K. (1998). Appraisal using generalized additive models. *Journal of Real Estate Research* 15, 77–99.

[87] Pagan, A. & Ulah, A. (1999). *Nonparametric Econometrics.* Cambridge: Cambridge University.

[88] Papoulis, A. (1965). *Probability, Random Variables, and Stochastic Processes*. New York: McGraw Hill.

[89] Parzen, E. (1962). On-estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33, 1065–1076.

[90] Paula, G.A. (2004). *Modelos de Regressão com Apoio Computacional*. São Paulo: IME/USP.

[91] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

[92] Pinto, C.C.X. (2003). *Diversidade do Lucro Entre as Pequenas Empresas Brasileiras: O Mercado de Crédito Como Um de Seus Possíveis Determinantes*. Dissertação (Mestrado em Economia) - Pontifícia Universidade Católica do Rio de Janeiro.

[93] Ramsay, J.O. & Silverman, B.W. (2006). *Functional Data Analysis*, 2nd ed. New York: Springer.

[94] Reinsch, C. (1967). Smoothing by spline functions. *Numerical Mathematics* 10, 177–183.

[95] Rigby, R. A. & Stasinopoulos, D.M. (1996a). A semi-parametric additive model for variance heterogeneity. *Statistical Computing* 6, 57–65.

[96] Rigby, R. A. & Stasinopoulos, D.M. (1996b). Mean and dispersion additive models. In *Statistical Theory and Computational Aspects of Smoothing*. Eds: Härdle, W. and Schimek, M.G., 215–230. Heidelberg: Physica.

[97] Rigby, R.A. & Stasinopoulos, D.M. (2001). The GAMLSS project: a flexible approach to statistical modelling. In *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*. Eds: Klein, B. and Korsholm, L., 337–345. Odense: Denmark.

[98] Rigby, R.A. & Stasinopoulos, D.M. (2004a) Box Cox t distribution for modelling skew and leptokurtotic data. *Technical Report 01/04*. STORM Research Centre, London Metropolitan University, London.

[99] Rigby R.A. & Stasinopoulos D.M. (2004b). Smooth centile curves for skew and kurtotic data modelled using the Box Cox power exponential distribution. *Statistics in Medicine* 23, 3053–3076.

[100] Rigby, R.A. & Stasinopoulos D.M. (2005). Generalized additive models for location, scale and shape (with discussion), *Applied Statistics* 54, 507–554.

[101] Rigby, R.A. & Stasinopoulos D.M. (2006). Using the Box Cox `t` distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling* 6, 209–229.

[102] Rigby, R.A. & Stasinopoulos D.M. (2007). Generalized additive models for location scale and shape (GAMLSS) in `R`. *Journal of Statistical Software*, vol. 23, Issue 7.

[103] Rigby, R.A. & Stasinopoulos, D.M. (2008). Instructions on How to Use the Gamlss Package in `R`. Disponível na internet em `http://www.londonmet.ac.uk/gamlss/`. Arquivo obtido em 10 de junho de 2009.

[104] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

[105] Rosa, F.H.F.P. & Soler, J.M. (2004). Avaliando técnicas de normalização para Microarrays de cDNA. In: 16° Sinape, Caxambu - MG. *Anais do 16° Sinape*.

[106] Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation perfect competition. *Journal of Political Economy* 82, 34–55.

[107] Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function. *Annals of Mathematical Statistics* 27, 832–837.

[108] Royston, P. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics* 43, 429–467.

[109] Ruppert, D. & Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics* 22, 1346–1370.

[110] Saboya, B.F.D. (1974). Avaliação de terras conflagradas pelas fraldas urbanas. *Anais do I Congresso Brasileiro de Avaliações e Perícias de Engenharia*. São Paulo: Pini.

[111] Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Annals of Statistics* 14, 1139–1151.

[112] Schick, A. (1993). On efficient estimation in regression models. *Annals of Statistics* 21, 1486–1521.

[113] Schick, A. (1996). Root-n-consistent and efficient estimation in semiparametric additive regression models. *Statistics & Probability Letters* 30, 45–51.

[114] Schoenberg, I.J. (1964). Spline interpolation and best quadrature formulae. *Bulletin of the American Mathematical Society* 70, 143–148.

[115] Schumaker, L.L. (1993). *Spline Functions: Basic Theory*. Melbourne: Krieger.

[116] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

[117] Sen, P.K. & Singer, J.M. (1993). *Large Sample Methods in Statistics. An Introduction with Applications*. New York: Chapman and Hall.

[118] Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *Annals of Statistics* 12, 898–916.

[119] Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society* B 47, 1–52.

[120] Silverman, B.W & Green, P.J. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[121] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society* 50, 413–436.

[122] Souza, C.P.E. (2008). *Testes de Hipóteses para Dados Funcionais Baseados em Distâncias: Um Estudo Usando Splines*. Dissertação (Mestrado em Estatística) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica (UNICAMP/IMECC), Campinas-SP.

[123] Stock, J. (1991). Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the 5th International Symposium in Economic Theory and Econometrics*. Eds: Barnett, W., Powell, J. and Tauchen, G. New York: Cambridge University Press.

[124] Stone, C.J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* 13, 689–705.

[125] Tukey, J.W. (1962). The future of data analysis. *Annals of Mathematical Statistics* 33, 1–67.

[126] van Buuren, S. & Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20, 1259–1277.

[127] Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*. 4th ed. Springer.

[128] Venables, W.N; Smith, D.M. & R Development Core Team. (2009). An introduction to R. Disponível em: `http://cran.r-project.org/doc/manuals/R-intro.pdf`. Arquivo obtido em 17 de setembro de 2009.

[129] Verbyla, A.P.; Cullis, B.R.; Kenward, M.G. & Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistical* 48, 269–311.

[130] Volpato, G.L. (2007). *Ciência: Da Filosofia à Publicação*, 5ª ed. São Paulo & Vinhedo: Cultura Acadêmica Editora & Scripta.

[131] Wahba, G. (1990). *Spline Models for Observation Data*. Philadelphia: Society for Industrial and Applied Mathematics.

[132] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* 61, 439–447.

[133] Whittaker, E.T. (1923). On new method of graduation. *Proceedings of the Edinburgh Mathematical Society* 41, 63–75.

[134] WHO Multicentre Growth Reference Study Group (2006). *WHO child growth standards: methods and development*. World Health Organization, Geneva, Switzerland.

[135] WHO Multicentre Growth Reference Study Group (2007). *WHO child growth standards: methods and development*. World Health Organization, Geneva, Switzerland.

[136] Wood, S.N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society* B 62, 413–428.

[137] Wood, S.N. (2001). Mgcv: GAMs and generalized Ridge regression for R. *R News* 1, 20–25.

[138] Zeni, A.M. (1996). *Curso básico de Engenharia de Avaliações - metodologia científica*. ABDE.

[139] Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 86, 79–95.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo