XVI COBREAP - CONGRESSO BRASILEIRO DE ENGENHARIA DE AVALIAÇÕES E PERÍCIAS - IBAPE/AM - 2011

TRABALHO DE AVALIAÇÃO

AVALIAÇÕES DE IMÓVEIS EM MASSA COM BASE EM MODELOS GAMLSS

Resumo: *O emprego da inferência estatística e da teoria econométrica na formulação de modelos de preços hedônicos voltados para o mercado imobiliário constituiu um marco para a Engenharia de Avaliações no Brasil. Entretanto, características singulares dos imóveis fazem com que a estimação do seu valor por meio do método de preços hedônicos seja uma tarefa complexa, uma vez que a teoria não especifica a forma funcional nem as variáveis relevantes que compõem a equação hedônica. Ademais, questões como falta de normalidade, heteroscedasticidade e autocorrelação são bastante comuns em dados imobiliários, razão pela qual a utilização de modelagens tradicionais de regressão – baseadas em relações estritamente paramétricas e lineares entre as variáveis dependente e independentes – pode sofrer limitações e não ser adequada para explicar o comportamento do mercado. Visando lidar com estas dificuldades, o presente trabalho propõe a indução do modelo explicativo do mercado imobiliário mediante o uso de uma (nova) técnica de modelagem estatística univariada que permite o ajuste de uma ampla família de distribuições para a variável dependente (por exemplo, o preço unitário) e possibilita a modelagem direta, utilizando funções paramétricas e/ou não-paramétricas, de todos os parâmetros da distribuição da variável resposta em relação as variáveis explanatórias. As análises realizadas neste trabalho indicam que os modelos de regressão semiparamétricos ora propostos aparentam ser mais apropriados para a estimação da função de preços hedônicos do que as técnicas clássicas de regressão usualmente empregadas, principalmente no que tange ao poder de explicação do modelo e a ampla flexibilidade funcional.*

Palavras-chave: Avaliação de imóveis, Função de preços hedônicos, Modelos de regressão semiparamétricos.

1 Introdução

1.1 Preliminares

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 o bom equilíbrio social, político e jurídico das relações humanas. Considerando-se que o imóvel, em geral, é o bem de maior importância adquirido pelo homem no decorrer de 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 para os diversos segmentos da sociedade e muitos órgãos governamentais ou privados: prefeituras (cobrança do Imposto Predial e Territorial Urbano (IPTU) e do Imposto sobre Transmissão de Bens Imóveis (ITBI), desapropriações e elaboração de plantas de valores genéricos); 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 de natureza técnica e científica.

Neste sentido, cabe à Engenharia de Avaliações, enquanto ciência do valor, a determinação técnica do valor de um bem, dos seus custos, frutos ou direitos de reprodução. As avaliações de imóveis são realizadas usualmente com base no método comparativo direto de dados de mercado, em que o valor de um bem é obtido por comparação com outros de características similares. Ocorre que, após a coleta dos elementos de referência, o engenheiro de avaliações[1] está geralmente de posse de uma amostra composta de eventos similares entre si mas que 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. De acordo com a NBR 14653-2:2011 (Avaliação de Bens Parte 2: Imóveis Urbanos), no tratamento dos dados podem ser utilizados, alternativamente e em função da qualidade e da quantidade de dados e informações disponíveis:

- tratamento por fatores: homogeneização por fatores e critérios, calculados e fundamentados por metodologia científica, e posterior análise estatística dos resultados homogeneizados;
- tratamento científico: tratamento de evidências empíricas pelo uso de metodologia científica que leve à indução de modelo validado para o comportamento do mercado.

No "tratamento científico" via modelos de regressão – cujas diretrizes do tratamento de dados serão utilizadas no presente trabalho – são empregadas ferramentas da Inferência Estatística, sendo a estimação do valor do imóvel usualmente definida a partir da

[1] 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.

equação de preços hedônicos, conforme metodologia proposta por Rosen (1974). Neste caso, 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 influentes na composição do preço do bem. 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 (ver, 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 (*Generalized Additive Models* — GLM) para estimar o valor venal de imóveis urbanos (ver, 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 Martins-Filho & Bin (2005), 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. Apesar destas referências, algumas limitações estão presentes nas modelagens supracitadas: (i) na abordagem semiparamétrica a distribuição da variável resposta pertence à família exponencial; (ii) no contexto puramente não-paramétrico há o inconveniente da "maldição da dimensionalidade"[2] (em inglês, *curse of dimensionality*).

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[3] ou que não restrinjam a modelagem da variável resposta às distribuições da família exponencial, razão pela qual se torna imperativa a

[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).

[3] Daqui em diante, salvo menção em contrário, sempre que citarmos a expressão "métodos e/ou metodologias tradicionais" estaremos nos referindo ao modelo normal de regressão linear clássico e aos modelos lineares generalizados.

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.

Neste contexto e visando lidar com as dificuldades retromencionadas, o presente trabalho propõe a indução de modelos explicativos do mercado imobiliário – para o caso de avaliações em massa – mediante o uso de uma (nova) 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.

A superioridade do ajuste da classe de modelos GAMLSS comparativamente às metodologias tradicionais é evidenciada neste trabalho a partir da análise empírica com dados de terrenos urbanos situados na cidade de Aracaju-SE, Brasil. Na análise empírica consideramos como variável resposta o preço unitário do terreno e como variáveis independentes as características estruturais, locacionais e econômicas inerentes aos imóveis. 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 (média) e escala (dispersão) da variável resposta. Os resultados obtidos, principalmente no que tange às análises gráficas e numéricas dos resíduos, aos critérios de informação de Akaike (AIC) e Schwarz (SBC), e segundo uma medida do poder explicativo do modelo estimado, denotada adiante de *pseudo*-R^2, indicam que os modelos GAMLSS aparentam ser mais apropriados para a estimação da função de preços hedônicos do que as modelagens via CNLRM e GLM. Este trabalho evidencia o ganho considerável no poder de ajuste ao usar modelos GAMLSS, mesmo sob dados de corte transversal e com elevada variabilidade, como são os terrenos que compõem a amostra de nossa análise empírica.

1.2 Objetivos do trabalho

Este trabalho 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 a estimação da equação de preços hedônicos de terrenos urbanos situados na cidade de Aracaju, capital do Estado de Sergipe (SE). Acredita-se que o emprego da estrutura de regressão GAMLSS possa contribuir para a análise e entendimento de quais, e de que forma e com que intensidade, os atributos influenciam na variabilidade observada nos preços dos imóveis.

Essencialmente, o que se busca neste trabalho é melhorar a precisão e acurar o processo de estimação da equação de preços hedônicos mediante emprego dos modelos GAMLSS, ainda não difundidos na área de Engenharia de Avaliações de Bens.

1.3 Estrutura do trabalho

O presente trabalho está dividido em 5 (cinco) seções. Na Seção 1, enfatizamos a importância da determinação técnica do valor de um bem imóvel para a tomada de decisão em diversos segmentos da sociedade e destacamos o método comparativo direto de dados de mercado como o eletivo nas avaliações de imóveis. Adicionalmente, destacamos as estruturas de regressão atualmente mais utilizadas no ajuste da função de preços hedônicos e mencionamos as principais dificuldades enfrentadas para a sua estimação. Além disto, apontamos os modelos GAMLSS como uma possível alternativa para superar algumas limitações presentes nas estruturas de regressão tradicionalmente empregadas pelo engenheiro de avaliações na estimação do modelo explicativo do mercado imobiliário. Por fim, expusemos os objetivos do trabalho. Na Seção 2, 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. Na Seção 3, descrevemos os dados empregados na análise empírica. Na Seção 4, apresentamos e comparamos os resultados – para o mesmo conjunto de dados – da estimação da equação de preços hedônicos utilizando os modelos GAMLSS contra alguns modelos ajustados por métodos tradicionais (CNLRM e GLM). Finalmente, na Seção 5, são apresentadas as conclusões, comentários e sugestões para futuras pesquisas.

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 deste trabalho 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, e tem 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).

O presente trabalho foi digitado 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 *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 deste trabalho.

2 Modelos GAMLSS

2.1 Introdução

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 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.

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 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.

Pode-se afirmar, assim como enfatizou Dantas (2005), que no atual cenário de avaliações imobiliárias há grande probabilidade dos resultados baseados no modelo normal de regressão linear 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 variável dependente e as variáveis independentes, associada às suposições adicionais sobre a distribuição de probabilidade para os erros aleatórios, constituem limitadores para a utilização desta técnica.

Para Maddala (2003), algumas vezes métodos mais simplificados foram sugeridos porque outros 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. É sabido que procedimentos de inferência baseados em suposições equivocadas sobre a 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.

Diante da crescente complexidade de modelização do mundo real e da quantidade de dados coletados, 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.

Neste sentido, Rigby & Stasinopoulos (2005) propuseram uma nova classe de modelos estatísticos de regressão (semi)paramétricos,[4] 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.[5]

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 discreta ou contínua (incluindo as distribuições contínuas com acentuada assimetria, positiva ou negativa, e expressiva curtose, leptocúrtica ou platicúrtica). 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.

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 4.3) e permite ajustar mais de 50 distribuições diferentes, entre elas a distribuição exponencial potência de Box-Cox (Rigby & Stasinopoulos, $2004b$) 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 subseções a seguir 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] Os modelos de regressão paramétricos, não-paramétricos e semiparamétricos 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. Mais detalhes sobre os modelos paramétricos, não-paramétricos e semiparamétricos podem ser obtidos em Davidson, 2003; Härdle, 1990; e Härdle *et al*, 2004, respectivamente.

[5] Diversos procedimentos não-paramétricos para estimar a função densidade de probabilidade estão disponíveis na literatura (ver Silverman & Green, 1986; Pagan & Ullah, 1999; Härdle, 1990) e são frequentemente referenciados como métodos de suavização (em inglês, *smoothing methods*). Dois suavizadores têm grande destaque na literatura: *kernel* e *splines* (ver Silverman & Green, 1986 e Silverman, 1984, respectivamente).

2.2 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

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.1}$$

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 (2.1) 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 (2.1) 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, (2.1) se reduz a um modelo linear completamente paramétrico dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k. \tag{2.2}$$

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 (2.1), temos

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{h}_{jk}(\mathtt{x}_{jk}), \tag{2.3}$$

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 (2.3) é denominado de GAMLSS aditivo semiparamétrico linear. O modelo resultante em (2.3) é um caso especial do modelo (2.1) e pode conter termos paramétricos, não-paramétricos e de efeitos aleatórios (Rigby & Stasinopoulos, 2005).

O Modelo (2.3) pode ser estendido para permitir a inclusão de termos não-lineares na modelagem dos k parâmetros da distribuição, na forma

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathtt{x}_{jk}), \tag{2.4}$$

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 (2.4) é designado de GAMLSS aditivo semiparamétrico não-linear. Se $J_k = 0$, então o Modelo (2.4) se reduz a um GAMLSS paramétrico não-linear, expresso por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k). \tag{2.5}$$

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, (2.5) se reduz ao modelo paramétrico linear (2.2). 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 (2.1), 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:

$$\left.\begin{array}{ll} \text{Parâmetros de posição e escala} & \left\{\begin{array}{l} g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ \\ g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \end{array}\right. \\ \\ \text{Parâmetros de forma} & \left\{\begin{array}{l} g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\ \\ g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}. \end{array}\right. \end{array}\right\} \tag{2.6}$$

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.

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 (2.1) é mais geral do que os CNLRM, GLM ou GAM, 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.

2.3 Estimação

Dois aspectos são fundamentais no ajuste de componentes aditivos incorporados na estrutura GAMLSS: o algoritmo *backfitting*[6] 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 (2.1) os termos de efeitos aleatórios $\boldsymbol{\gamma}_{jk}$ sejam independentes e tenham distribuição normal com $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0},\, \boldsymbol{G}_{jk}^{-1})$, 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 deste trabalho, 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 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*; ver 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 $\boldsymbol{\lambda}_{jk}$, temos que $\boldsymbol{\beta}_k$ e $\boldsymbol{\gamma}_{jk}$ são estimados na estrutura de regressão GAMLSS por meio da maximização da função de log-verossimilhança penalizada, ℓ_p , dada por

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\boldsymbol{\gamma}_{jk}^{\top}\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.7}$$

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

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

(ver Lee & Nelder, 1996 e Pawitan, 2001). Acrescenta-se que a maximização de ℓ_p pode ser obtida com a implementação de um algoritmo *backfitting*.

[6] 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*. Para mais detalhes sobre o algoritmo *backfitting* ver Hastie & Tibshirani (1990) e Härdle *et al*. (2004).

2.4 Algoritmos de maximização

No R, dois algoritmos podem ser utilizados para a maximização da função de verossimilhança penalizada dada em (2.7). 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. 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 $k = 1, 2, \ldots, p$.

O objetivo dos algoritmos é maximizar a função de verossimilhança penalizada ℓ_p, dada por (2.7), para hiperparâmetros (λ) fixados. Nos modelos completamente paramétricos, como (2.2) ou (2.4), os algoritmos maximizam a função de verossimilhança ℓ. Mais detalhes sobre os algoritmos CG e RS ver Rigby & Stasinopoulos (2005).

2.5 Preditor linear

2.5.1 Termos paramétricos

No modelo GAMLSS (2.1), 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 e polinômios fracionários para as variáveis exploratórias.

Acrescenta-se ainda que parâmetros não-lineares podem ser incorporados à estrutura GAMLSS (2.1) pelo método perfilado ou pelo método derivado.[7] 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 (ver, por exemplo, Benjamin *et al.*, 2003).

2.5.2 Termos aditivos

Os componentes aditivos $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$ na Equação (2.1) 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.

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

2.5.2.1 *Splines* cúbicos

A utilização de *splines* cúbicos no Modelo (2.3) presume que as funções $h(t)$ são contínuas e duas vezes diferenciáveis e que a maximização da função de verossimilhança penalizada (ver Equação (2.7)) 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 $i = 1, 2, \ldots, n$ (ver de Boor, 1978 e Schumaker, 1993), ou seja, $h(t) = \sum_{i=1}^{n} \delta_i B_i(t)$.

Considere ainda que $\boldsymbol{h} = h(\mathbf{x})$ é um vetor com as avaliações da função $h(t)$ dos valores de **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 **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 **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 (2.1) 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).

Além dos *splines* cúbicos, 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*[8] e os polinômios fracionários.

Acrescenta-se que 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.

2.6 Famílias específicas

2.7 Generalidades

A função densidade de probabilidade populacional $f(y|\boldsymbol{\theta})$ no Modelo (2.1) 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.

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

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 1 e 2 exibem algumas famílias de distribuições contínuas e discretas, respectivamente, que se encontram implementadas no R.

Tabela 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 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 1 e 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,

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

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, 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).

2.8 Seleção do modelo

2.8.1 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.[9] 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 dependente. Aqui, cabe destacarmos que 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.

2.8.2 Seleção do modelo, inferências e diagnósticos

Na estrutura de regressão GAMLSS paramétrica, cada modelo $\mathcal{M}$ da forma (2.2) pode ser avaliado a partir de seu desvio global ajustado (*Global Deviance* — GD), dado

[9] 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.

por $\mathrm{GD} = -2\ell(\hat{\boldsymbol{\theta}})$, 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 $\Lambda = \mathrm{GD}_0 - \mathrm{GD}_1$, 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[10] 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, $\mathrm{GAIC}(\#) = \mathrm{GD} + \#df$, 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 $\mathrm{GAIC}(\#)$ é o selecionado. A sensibilidade do modelo selecionado frente à escolha 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 $\mathrm{GAIC}(\#)$, 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 (ver 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

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

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 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.

3 Análise de dados

Esta seção 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.

3.1 Coleta de dados

O conjunto de dados a ser analisado é composto de 2,109 (duas mil cento e nove) observações de terrenos urbanos nus, ou seja sem construções edificadas, situados na cidade de Aracaju, capital do Estado de Sergipe (SE), Brasil, e são provenientes de duas fontes: (i) coleta pelos autores 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* e tiveram suas posições geográficas (latitude, longitude) projetadas no Sistema Universal Transverso de Mercartor (UTM — *Universal Transversa de Mercator*).

3.2 Análise exploratória de dados

A amostra utilizada para a estimação da equação de preços hedônicos[11] contém, além do ano de referência, 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 observação, renda média do chefe de família do setor censitário[12] onde se situa o imóvel e valor do terreno). A seguir, definimos as variáveis observadas na amostra e detalhamos o tratamento dispensado para cada atributo visando à estimação da equação de preços hedônicos. Neste sentido, temos:

- ANO (ANO): variável qualitativa ordinal que identifica o ano em que a informação foi obtida. Pode assumir os valores 2005 ou 2006 (ANO06) ou 2007 (ANO07). Tratada na modelagem como variáveis *dummys*;

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

[11] Para simplificação da linguagem empregada ao longo deste trabalho, 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".

[12] Os setores censitários são unidades territoriais definidas pelo Instituto Brasileiro de Geografia e Estatística (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).

- FRENTE (FR): variável quantitativa contínua, também denominada de “testada” e medida em m (metros), 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. Tratada na modelagem como uma variável *dummy*, em que atribuiu-se 1 para terrenos planos e 0 para acidentados;

- 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. Tratada na modelagem como uma variável *dummy*, em que atribuiu-se 1 para terrenos situados em vias pavimentadas e 0 caso contrário;

- SITUAÇÃO (SI): variável qualitativa nominal empregada para discernir a disposição do terreno na quadra. Classifica-se em lote de “esquina” ou “meio”. Tratada na modelagem como uma variável *dummy*, em que atribuiu-se 1 para terrenos de esquina e 0 caso contrário;

- BAIRRO (BAIRRO): variável qualitativa nominal referente ao nome do bairro onde o terreno observado está situado. Na modelagem foi subclassificada em bairros supostamente valorizados e não-valorizados, sendo esta variável denotada por BV e tratada como *dummy*, em que atribuiu-se 1 para bairros admitidos como valorizados e 0 caso contrário. Os bairros foram ainda agrupados em pertencentes à Zona Sul da cidade ou não, sendo esta variável tratada como *dummy* e representada na modelagem por DZSU. Aqui, atribuiu-se 1 para terrenos localizados na Zona Sul e 0 caso contrário;

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

- 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. O CA é definido a partir do plano diretor de desenvolvimento urbano de Aracaju. Pode assumir os valores 3.0, 3.5,. . . ,5.5 ou 6.0.

- 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” (VIAP), “via secundária” (VIAS) ou “via terciária/superior”, conforme importância da via pública no contexto da região. Tratada na modelagem como variáveis *dummys*;

- NATUREZA DA INFORMAÇÃO (NI): variável qualitativa nominal que define se o dado coletado é oriundo de “oferta”(NIO), “transação” (NIT) ou do registro da prefeitura acerca do Imposto sobre Transmissão de Bens Imóveis (ITBI). Tratada na modelagem como variáveis *dummys*;

- SETOR (ST): variável *proxy*[13] quantitativa discreta de macrolocalização para distinguir o nível socioeconômico dos diversos bairros da cidade, representada pela renda 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.[14]Pode assumir os valores 1,2,. . . ,17 ou 18;

- FRENTE EM BAIRROS VALORIZADOS (FRBV): variável quantitativa contínua que assume valores estritamente positivos e que corresponde a interação entre as variáveis FR e BV. Incluída na modelagem 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";

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

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, BV, DZSU, LAT, LONG, ST, CA e VIA), físicas (AR, FR, TO SI e FRBV) e econômicas (NI), além do ano em que a observação foi coletada.

Na Figura 1 apresentamos os gráficos box-plot (também denotados na literatura de *gráficos de caixa*) das variáveis PU, AR e FR, enquanto que na Tabela 3 mostramos um resumo de algumas medidas de posição e dispersão destas variáveis.

Verificamos por meio dos gráfico box-plot que PU aparenta ter distribuição assimétrica à direita e que há uma quantidade de observações atípicas considerável. Estas características da variável PU também podem ser ratificadas mediante inspeção de seu histograma constante na Figura 2. Já na Tabela 3 observamos que PU abrange um expressivo intervalo de valores (entre $\mathrm{R\$2.36/m^2}$ e $\mathrm{R\$\ 800.00/m^2}$), bem como evidencia que cerca de 75$\%$ dos terrenos observados têm preços unitários inferiores a $\mathrm{R\$\ 82.82/m^2}$.

Embora tenham sido identificadas 263 observações atípicas mediante inspeção do gráfico box-plot de AR (ver Figura 1), constatamos que as discrepâncias não estão relacionadas a erros de mensuração, mas à elevada amplitude e dispersão da própria variável. Adicionalmente, percebemos que AR varia de $41\ \mathrm{m^2}$ a $91,780\ \mathrm{m^2}$, isto é, o maior terreno é 1,912 vezes superior ao menor, em área. Em se tratando da variável FR, notamos pelo gráfico de box-plot (ver Figura 1) que há uma acentuada variabilidade entre os dados, revelada também pela amplitude total ($= 513.40\ \mathrm{m}$) registrada na Tabela 3. Ou seja, o menor terreno é cerca de 198 vezes menor que o maior terreno observado (em relação à frente).

[13] *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.

[14] 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).

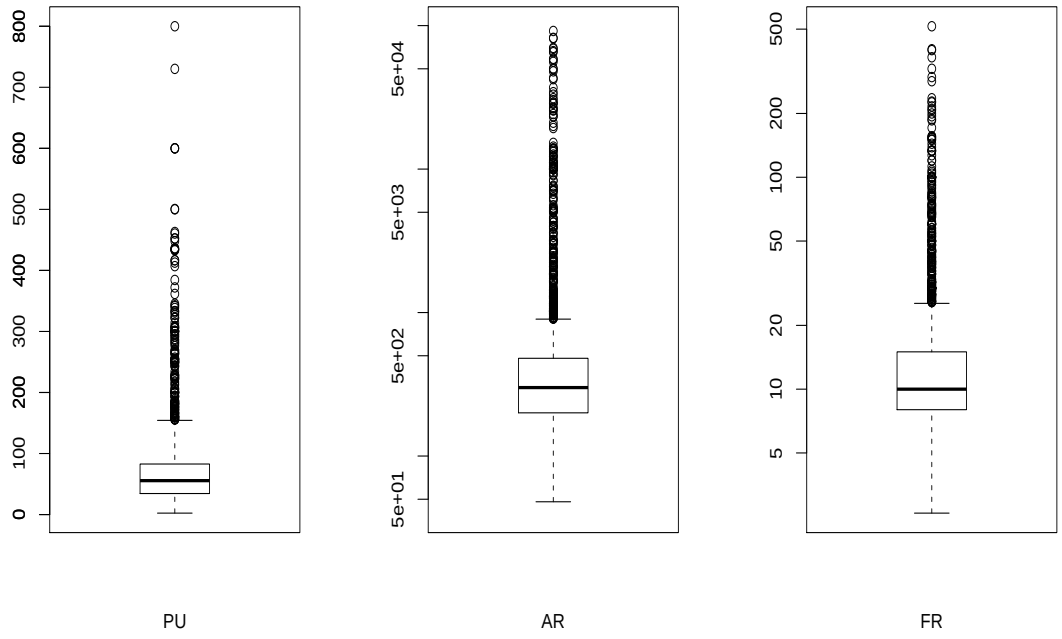

Figura 1: Gráficos box-plot das variáveis PU, AR e FR.

Tabela 3: 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

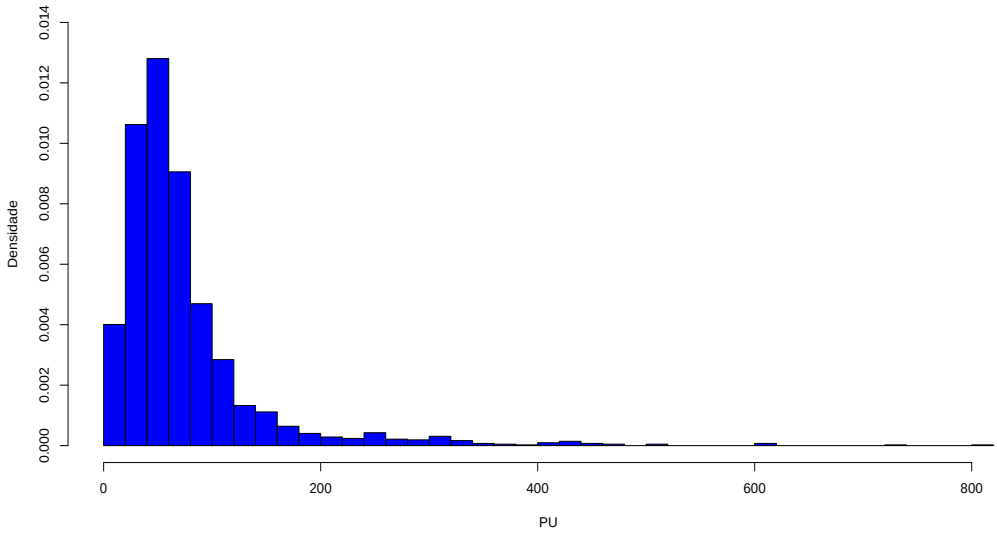

Figura 2: Histograma de PU.

Para analisar o comportamento de PU em relação a algumas variáveis explicativas usamos diagramas de dispersão. Neste sentido, apresentamos na Figura 3 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.

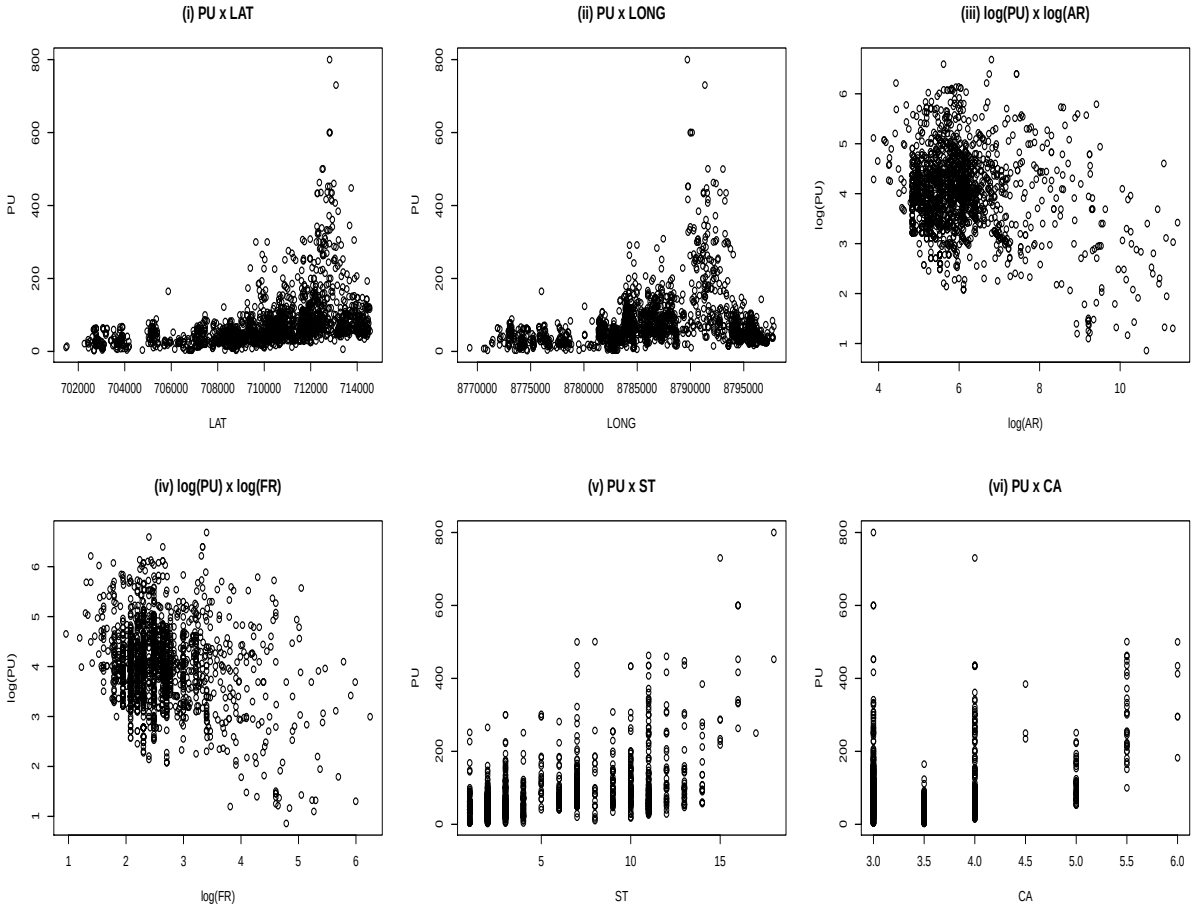

Figura 3: Gráficos de dispersão entre PU e as variáveis quantitativas explicativas.

Conforme podemos observar na Figura 3, 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 à medida em 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 destacar que a expectativa que tínhamos *a priori* do mercado somente não foi ratificada em (iv), visto que esperávamos o aumento de PU quando FR crescesse. Esta aparente "incoerência" motivou a inclusão da variável FRBV, já definida anteriormente, na modelagem empírica a ser apresentada na Seção 4.

Outro aspecto importante que podemos mencionar acerca da Figura 3 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. Adicionalmente, sustentar as hipóteses de homoscedasticidade e normalidade da distribuição condicional de PU dadas as variáveis explicativas (analisadas individualmente ou 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 e incoerente com a teoria subjacente, além de tipicamente resultar em expressões de difícil interpretação.

Visando à identificação de alguma tendência entre as variáveis qualitativas e o preço unitário, construímos na Figura 4 os gráficos box-plot entre: (i) PU $\times$ SI; (ii) PU $\times$ PA; (iii) PU $\times$ TO; (iv) PU $\times$ NI; (v) PU $\times$ VIA; (vi) PU $\times$ 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.

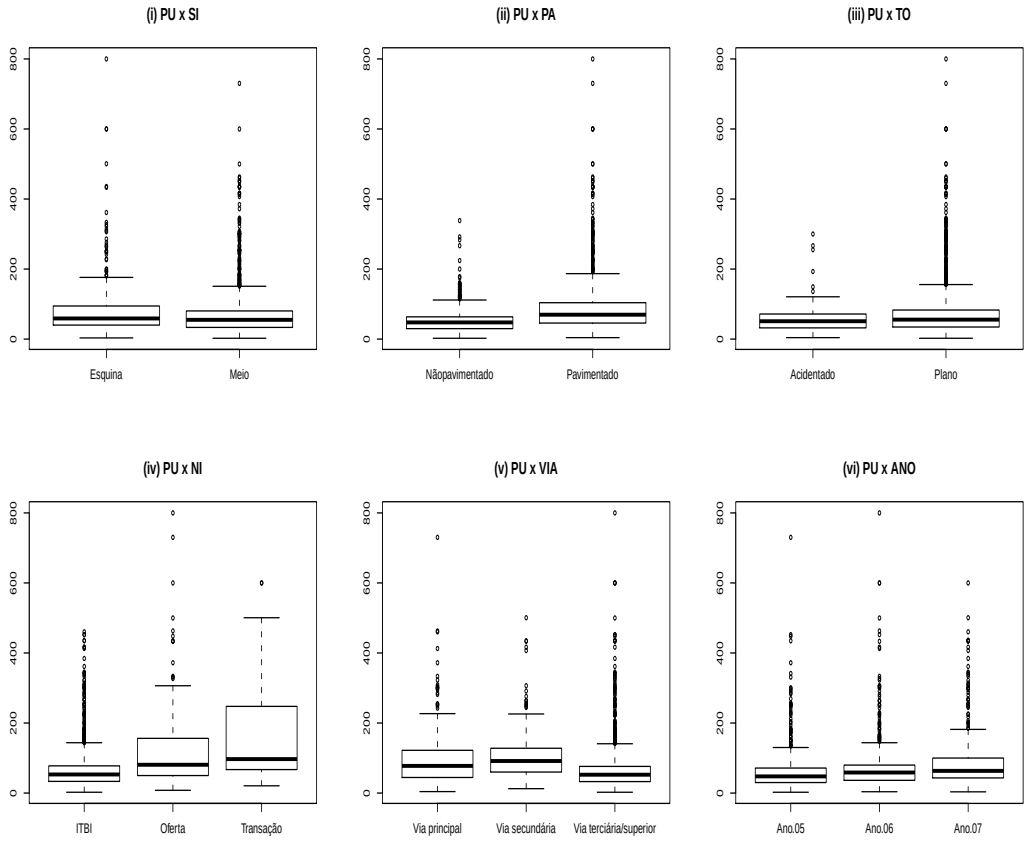

Figura 4: Gráficos box-plot entre PU e as variáveis qualitativas explicativas.

3.3 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 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.

Finalmente, chamamos a atenção para a forte correlação positiva entre AR $\times$ FR $(= 0.77)$ e log(AR) $\times$ log(FR) $(= 0.93)$, 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.

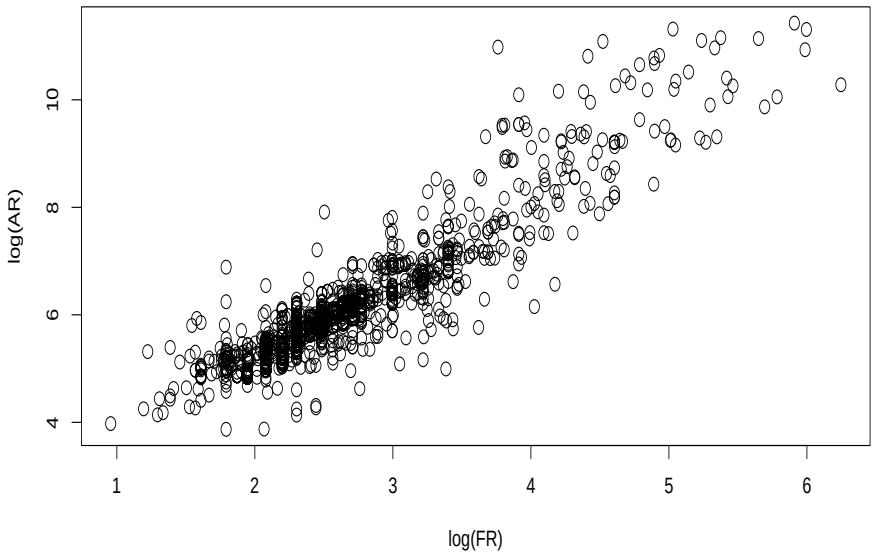

Figura 5: Gráfico de dispersão entre as variáveis FR e AR.

Tabela 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.

4 Modelagem empírica

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.

Esclarecemos ainda que a variável FR mostrou-se altamente correlacionada com AR (ver Subseção 3.3) e em todos os modelos ajustados apresentou o sinal do coeficiente estimado negativo, ou seja, contrário à expectativa do mercado imobiliário, motivos pelos quais foi excluída durante a modelagem via CNLRM, GLM e GAMLSS.

4.1 A modelagem via CNLRM

No modelo normal de regressão linear clássico o modelo adotado para inferir o comportamento do mercado imobiliário é dado por

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.1}$$

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).

Tradicionalmente, a estimação dos parâmetros é realizada com base no método de mínimos quadrados ordinários (*Ordinary Least Squares* — OLS),[15] e alguns pressupostos devem ser atendidos se o objetivo é fazer testes de hipóteses, estimação intervalar

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

e garantir que os parâmetros inferidos no mercado sejam não-tendenciosos, eficientes e consistentes, a saber: (i) o modelo $\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\epsilon}$ 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) $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, em que $\mathbf{0}$ 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) $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2)$, ou seja, os erros têm distribuição normal[16] — com média 0 e variância σ^2 — e são independentes.

Na Tabela 5 resumimos os principais ajustes realizados via CNLRM e as observações relevantes acerca dos modelos concorrentes à predição da equação de preços hedônicos.

Tabela 5: Modelos ajustados via CNLRM

Modelos	Forma Funcional	Considerações
1.1	$\mathsf{PU} = \beta_0 + \beta_1\mathsf{LAT} + \beta_2\mathsf{LONG} + \beta_3\mathsf{AR} + \beta_4\mathsf{CA} + \beta_5\mathsf{ST} + \beta_6\mathsf{VIAP} + \beta_7\mathsf{VIAS} + \beta_8\mathsf{SI} + \beta_9\mathsf{PA} + \beta_{10}\mathsf{TO} + \beta_{11}\mathsf{NIO} + \beta_{12}\mathsf{NIT} + \beta_{13}\mathsf{ANO06} + \beta_{14}\mathsf{ANO07} + \beta_{15}\mathsf{DZSU} + \beta_{16}\mathsf{FRBV} + \epsilon$	As hipóteses nulas de que os resíduos são homocedásticos e normais foram rejeitadas ao nível nominal de 1% pelos teste de Breusch-Pagan e Jarque-Bera, respectivamente. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível nominal de 1% quando utilizado o teste z. $\overline{R}^2$=0.539, AIC=22304 e BIC=22406.
1.2	$\log(\mathsf{PU}) = \beta_0 + \beta_1\mathsf{LAT} + \beta_2\mathsf{LONG} + \beta_3\mathsf{AR} + \beta_4\mathsf{CA} + \beta_5\mathsf{ST} + \beta_6\mathsf{VIAP} + \beta_7\mathsf{VIAS} + \beta_8\mathsf{SI} + \beta_9\mathsf{PA} + \beta_{10}\mathsf{TO} + \beta_{11}\mathsf{NIO} + \beta_{12}\mathsf{NIT} + \beta_{13}\mathsf{ANO06} + \beta_{14}\mathsf{ANO07} + \beta_{15}\mathsf{DZSU} + \beta_{16}\mathsf{FRBV} + \epsilon$	As hipóteses nulas de que os resíduos são homocedásticos e normais foram rejeitadas ao nível nominal de 1% pelos teste de Breusch-Pagan e Jarque-Bera, respectivamente. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível nominal de 1% quando utilizado o teste z. $\overline{R}^2$=0.599, AIC=2912 e BIC=3014.
1.3	$\log(\mathsf{PU}) = \beta_0 + \beta_1\mathsf{LAT} + \beta_2\mathsf{LONG} + \beta_3\log(\mathsf{AR}) + \beta_4\mathsf{CA} + \beta_5\log(\mathsf{ST}) + \beta_6\mathsf{VIAP} + \beta_7\mathsf{VIAS} + \beta_8\mathsf{SI} + \beta_9\mathsf{PA} + \beta_{10}\mathsf{TO} + \beta_{11}\mathsf{NIO} + \beta_{12}\mathsf{NIT} + \beta_{13}\mathsf{ANO06} + \beta_{14}\mathsf{ANO07} + \beta_{15}\mathsf{DZSU} + \beta_{16}\log(\mathsf{FRBV}) + \epsilon$	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 nominal de 1%. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível nominal de 1%, exceto para a variável LAT (valor-$p = 0.0190$). $\overline{R}^2$=0.651, AIC=2619 e BIC=2721.
1.4	$\frac{\mathsf{PU}^{\lambda}-1}{\lambda} = \beta_0 + \beta_1\mathsf{LAT} + \beta_2\mathsf{LONG} + \beta_3\log(\mathsf{AR}) + \beta_4\mathsf{CA} + \beta_5\log(\mathsf{ST}) + \beta_6\mathsf{VIAP} + \beta_7\mathsf{VIAS} + \beta_8\mathsf{PA} + \beta_9\mathsf{TO} + \beta_{10}\mathsf{NIO} + \beta_{11}\mathsf{NIT} + \beta_{12}\mathsf{ANO06} + \beta_{13}\mathsf{ANO07} + \beta_{14}\log(\mathsf{FRBV}) + \epsilon$	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 nominal de 1%. Os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível nominal 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, verificamos que o Modelo (1.4), referente à transformação de Box-Cox (com $\hat{\lambda} = 0.1010$), apresentou os "melho-

[16]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.

res" resultados no que tange ao coeficiente de determinação ajustado $\overline{R}^2$, AIC e BIC. Porém, não foi capaz de estabilizar a variância dos resíduos, conforme verificado pelo 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 adotado 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, 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 6 o ajuste realizado para o Modelo (1.4) utilizando erros-padrão robustos HC3 (Davidson & MacKinnon, 1993) para corrigir o efeito da heteroscedasticidade.

De acordo com a Tabela 6, todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível de 2%, exceto para a variável LAT (valor-$p = 0.1263$), indicando que as maiores variações dos preços, a grande escala espacial, ocorrem no sentido norte-sul.

Os resultados obtidos neste trabalho mediante uso dos modelos CNLRM ratificam, conforme observado por Dantas & Cordeiro (2000),[17] que a falta de normalidade é indubitável nos preços de compra e venda 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.

Tabela 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
IFRBV	1.2041	0.0137	8.797	0.0000

[17] 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.

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:

$$g(\boldsymbol{PU^*}) \;\; = \;\; f(F, L, E, \boldsymbol{\beta}), \tag{4.2}$$

em que $\boldsymbol{PU^*} = \mathsf{E}(\boldsymbol{PU}) = \boldsymbol{\mu}$ e $f(F, L, E, \boldsymbol{\beta}) = \boldsymbol{X\beta} = \boldsymbol{\eta}$, ou seja, a estimação empírica da Equação (4.2) 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.[18]

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 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 (4.2) 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.

Exibimos na Tabela 7 o ajuste realizado via GLM do modelo preditor da equação de preços hedônicos, dado por

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

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

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.

Destaca-se também que os coeficientes das variáveis explicativas mostraram-se estatisticamente significativos ao nível nominal de 1% quando utilizado o teste z, 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 usando GLM. Entretanto, o uso da distribuição gama, ao invés da normal, resultou numa leve melhora no ajuste dos dados no que tange à relação entre os valores observados e os valores preditos.

[18] 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.

Tabela 7: Ajuste do modelo de preços hedônicos via GLM - Modelo (2.1).

	Estimativa	Erro-padrão	Estatística z	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

4.3 A modelagem via GAMLSS

Conforme salientado na Seção 2, 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 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 (ver Seção 2.8). 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; ver Rigby & Stasinopoulos, 2008) e na *biblioteca* `MASS` (como `addterm()` e `dropterm()`; ver 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.

4.3.1 Modelagem do parâmetro de posição (μ)

Visto que a variável PU assume apenas valores positivos, elegemos as distribuições log-normal (LOGNO), gaussiana inversa (IG), Weibull (WEI) e gama (GA)[19] como potenciais candidatas ao ajuste da variável resposta. Apresentamos na Tabela 8 os principais modelos considerados com o objetivo de modelar o parâmetro μ e os respectivos comentários acerca dos ajustes.

Com base na Tabela 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 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 9 o ajuste referente a este modelo e relativo à estimação da equação de preços hedônicos.

Tabela 8: Modelos ajustados via GAMLSS

Modelos	$\mathcal{D}$	$\mathcal{G}$	Forma funcional	Considerações
3.1	LOGNO	logarítmica	PU $=\beta_0+$ cs(LAT) $+$ cs(LONG) $+$ cs(log(AR)) $+$ cs(CA) $+$ cs(ST) $+\beta_1$VIAP $+$ β_2VIAS $+\beta_3$SI $+\beta_4$PA $+\beta_5$TO $+\beta_6$NIO $+$ β_7NIT$+\beta_8$ANO06$+\beta_9$ANO07$+\beta_{10}$DZSU$+$ cs(log(FRBV))	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível nominal de 1% quando utilizado o teste z. AIC=19155, SBC=19359 e GD=19083.
3.2	IG	logarítmica	PU $=\beta_0+$ cs(LAT) $+$ cs(LONG) $+$ cs(log(AR)) $+$ cs(CA) $+$ cs(ST) $+\beta_1$VIAP $+$ β_2VIAS $+\beta_3$SI $+\beta_4$PA $+\beta_5$TO $+\beta_6$NIO $+$ β_7NIT$+\beta_8$ANO06$+\beta_9$ANO07$+\beta_{10}$DZSU$+$ cs(log(FRBV))	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível nominal de 1% quando utilizado o teste z. AIC=19845, SBC=20048 e GD=19773.
3.3	WEI	logarítmica	PU $=\beta_0+$ cs(LAT) $+$ cs(LONG) $+$ cs(log(AR)) $+$ cs(CA) $+$ cs(ST) $+\beta_1$VIAP $+$ β_2VIAS $+\beta_3$SI $+\beta_4$PA $+\beta_5$TO $+\beta_6$NIO $+$ β_7NIT$+\beta_8$ANO06$+\beta_9$ANO07$+\beta_{10}$DZSU$+$ cs(log(FRBV))	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível nominal de 1% quando utilizado o teste z. AIC=19260, SBC=19463 e GD=19188.
3.4	GA	logarítmica	PU $=\beta_0+$ cs(LAT) $+$ cs(LONG) $+$ cs(log(AR)) $+$ cs(CA) $+$ cs(ST) $+\beta_1$VIAP $+$ β_2VIAS $+\beta_3$SI $+\beta_4$PA $+\beta_5$TO $+\beta_6$NIO $+$ β_7NIT$+\beta_8$ANO06$+\beta_9$ANO07$+\beta_{10}$DZSU$+$ cs(log(FRBV))	Todos os coeficientes das variáveis explicativas mostraram-se significativos ao nível nominal de 1% quando utilizado o teste z. AIC=19062, SBC=19337 e GD=19062.

Embora as funções estimadas não-parametricamente utilizando 3 (três) graus de liberdade efetivos em todas as funções suavizadoras tenham conduzido a um ajuste razoável da equação de preços hedônicos, é possível obter o número de graus de liberdade "ótimo" para os suavizadores. Neste sentido, reestimamos o Modelo (3.4) levando em consideração dois aspectos: a escolha via AIC e a inspeção visual das curvas suavizadas — este último aspecto teve por objetivo evitar "sobreajustamentos" (*overfitting*). O "novo" modelo estimado (Modelo (3.5)) também fez uso dos suavizadores *splines* cúbicos (cs), porém com diferentes graus de liberdade (df) efetivos nas funções

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

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

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).

alisadoras, conforme destacado na Tabela 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 entre os valores observados contra os valores preditos.

Tabela 9: Ajuste do modelo de preços hedônicos via GAMLSS - Modelo (3.4).

	Estimativa	Erro-padrão	Estatística z	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(IAR)	-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(IFRBV)	0.6513	0.0569	11.443	0.0000
DZSU	0.3875	0.0299	12.935	0.0000

Tabela 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(IAR, 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(IFRBV, df=10)	0.6599	0.0522	12.630	0.0000
DZSU	0.5119	0.0275	18.613	0.0000

Exibimos na Figura 6 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 verde corresponde aos erros-padrão pontuais *(pointwise standard errors)*.

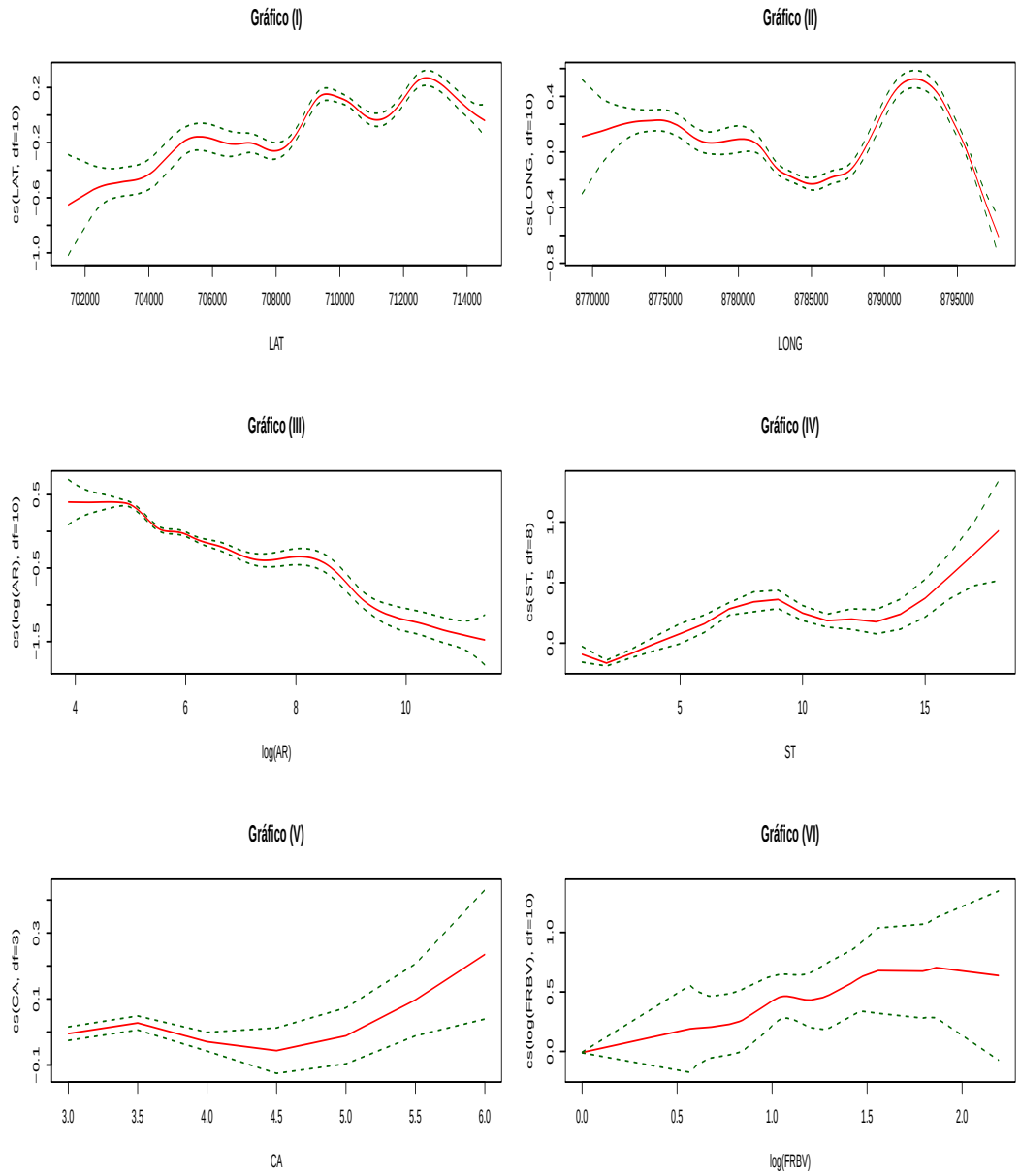

Figura 6: Gráficos dos termos aditivos suavizados - Modelo (3.5).

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,[20] decrescentes,

[20] 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.

crescentes, crescentes e crescentes, respectivamente, com os aumentos da latitude, longitude, logaritmo da área, setor socioeconômico, coeficiente de aproveitamento e logaritmo da frente do terreno nos bairros valorizados, respectivamente. Contudo, estas mesmas informações também foram fornecidas anteriormente pelos modelos CNLRM e GLM 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), percebe-se 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. Adicionalmente, 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 socioeconômicos 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), nota-se 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, quase exclusivamente, observações do bairro do Mosqueiro. A partir da posição 8785000 há uma notável 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 tornar-se negativo.

No Gráfico (III), percebe-se 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), nota-se que à medida que o setor socioeconômico 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 socioeconômico 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), percebe-se 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 a 5.0 é que se verifica efeito positivo crescente.

No Gráfico (VI), nota-se que à medida que o logaritmo da frente dos terrenos aumenta nos bairros valorizados, a "contribuição" do termo aditivo ajustado à covariável log(FRBV) é preponderantemente crescente e 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 6, 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 a variável dependente (PU) e as covariáveis 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 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.

4.3.2 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.[21] Adicionalmente, os modelos serão confrontados por meio de um "*pseudo* coeficiente de determinação" (*pseudo*-R^2), o qual será calculado pela expressão

$$\textit{pseudo} - R^2 = (\text{correlação (valores observados de } PU, \text{ valores preditos de } PU))^2. \quad (4.3)$$

Com base nas considerações anteriores, apresentamos na Tabela 11 um resumo comparativo entre os modelos supracitados e claramente percebemos a superioridade 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. Para o modelo GAMLSS (3.5), o *pseudo*-R^2 supera 0.80.

Tabela 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

4.3.3 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)[22] para investigar o comportamento — se homoscedástico ou heteroscedástico — do parâmetro de escala σ. Tendo em vista

[21] 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)).

[22] 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 logaritmo da razão entre as duas verossimilhanças (L($\widetilde{\boldsymbol{\theta}}$) e L($\widehat{\boldsymbol{\theta}}$)), isto é, na diferença entre logL($\widetilde{\boldsymbol{\theta}}$) e logL($\widehat{\boldsymbol{\theta}}$). Se H_0 é verdadeira, então $\mathrm{LR} = -2[\log\mathrm{L}(\widetilde{\boldsymbol{\theta}}) - \log\mathrm{L}(\widehat{\boldsymbol{\theta}})] \xrightarrow{d} \chi^2_m$, em que m é o número de restrições, quando $n \to \infty$.

que a hipótese nula de dispersão constante foi rejeitada, segundo o teste LR, modelamos a dispersão (σ) tomando por base o Modelo (3.5), uma vez que esse foi o ajuste que melhor "representou" os dados. Para modelar o parâmetro de dispersão adotamos procedimento semelhante ao utilizado anteriormente na modelagem do parâmetro de posição, ou seja, aplicamos a técnica *stepwise* de seleção de covariáveis, 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. Acrescenta-se que os procedimentos citados não foram novamente aplicados ao parâmetro de posição, mas apenas impostos à modelagem do parâmetro de dispersão. Destaca-se ainda que nesta etapa também utilizamos o critério AIC na escolha dos suavizadores e definição dos graus de liberdade das funções alisadoras, bem como fizemos a inspeção visual das curvas suavizadas na busca do "melhor" modelo.

Apresentamos na Tabela 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 logarítmicas. Note que o Modelo (3.6) contém termos paramétricos e não-paramétricos, motivo pelo qual é denominado de GAMLSS aditivo semiparamétrico linear.

Tabela 12: Ajuste do modelo de preços hedônicos via GAMLSS - Modelo (3.6).

Coeficientes de μ				
	Estimativa	Erro-padrão	Estatística z	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(IAR, 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(IFRBV, 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(IAR, df=10)	0.1370	0.0143	9.593	0.0000
ST	-0.0391	0.0040	-9.632	0.0000

A partir dos resultados da Tabela 12, verificamos que as estimativas dos coeficientes do submodelo da média não sofreram grandes alterações em relação àquelas obtidas para o Modelo (3.5) (ver Tabela 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*[23] em relação ao Modelo (3.5) (ver Figuras 7 e 8).

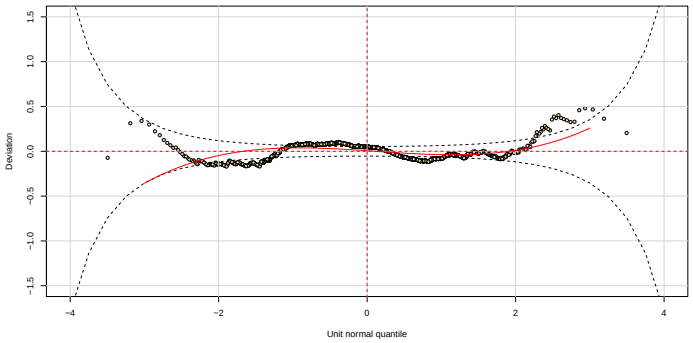

Figura 7: Gráfico *worm-plot* - Modelo (3.5).

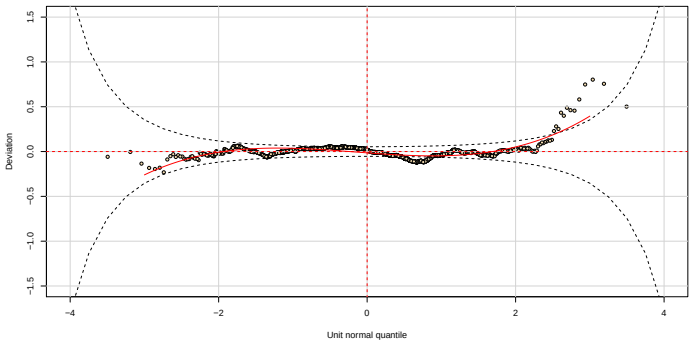

Figura 8: Gráfico *worm-plot* - Modelo (3.6).

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. 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 socioeconômico do setor censitário onde o imóvel está localizado. Aqui, cabe ressaltar que o comportamento observado da variância em função da covariável ST aparenta refletir

[23]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), então o modelo fornece um bom ajuste.

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 socioeconômicos do setor censitário.

Cumpre registrar ainda que o valor 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 nominal de 1%. Note que estes resultados, especialmente no que tange ao elevado valor do *pseudo*-R^2, em geral são raros de serem atingidos quando se trabalha com dados de corte transversal, 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 acentuada variabilidade entre as características físicas, estruturais e locacionais dos imóveis observados, é apreciável a superioridade da qualidade (ver Figura 9 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.

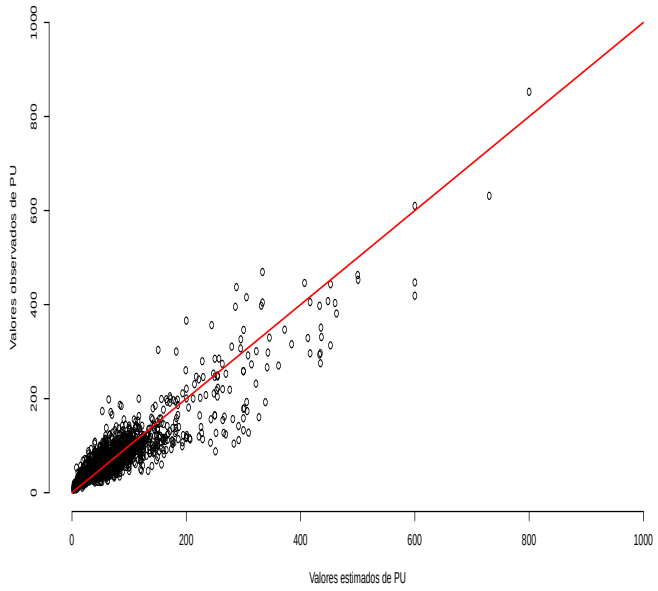

Figura 9: Gráfico dos *valores observados* $\times$ *valores preditos* de PU – Modelo (3.6).

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

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

$$\log(\sigma) \;=\; \gamma_0 + \gamma_1\text{ST} + cs(\log(\text{AR}), \text{df}=10),$$

em que a variável resposta resposta (PU) segue distribuição gama (GA) com parâmetros de posição μ e escala σ, aparenta ser o mais indicado para a estimação da equação de preços hedônicos para terrenos urbanos situados na cidade de Aracaju-SE.

5 Considerações finais

5.1 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.

O enfoque central deste trabalho 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 (CNLRM) e dos modelos lineares generalizados (GLM). 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. Ademais, a indução do modelo explicativo do mercado imobiliário mediante o emprego dos modelos GAMLSS possibilitou um forma versátil para explorar a relação entre as variáveis do modelo, bem como mostrou-se uma boa ferramente para a detecção e a acomodação de pontos espúrios (ou seja, pontos de alavancagem e *outliers*) a partir do estudo da influência que cada ponto exerce no ajuste.

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 deste trabalho. Embora a natureza dos dados analisados neste estudo 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" (ou seja, nesta abordagem busca-se o ajuste do "modelo aos dados"), o que minimiza quaisquer tentativas "forçosas" de ajustar os "dados ao modelo".

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 — ex-

cluída 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 a 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.

Importante ressaltar, todavia, que os modelos GAMLSS não são uma "fórmula" mágica e perfeita para se avaliar bens, ao contrário, ainda que aparentem espelhar a "realidade" e sejam parcimoniosos, plausíveis e informativos, mesmo assim serão "representações" simplificadas e aproximadas do mercado, uma vez que sempre compreenderão uma parte não explicada que incorpora erros.

5.2 Utilidade do estudo

O emprego de métodos estatísticos mais flexíveis e que são capazes de descrever com um 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 quatro contribuições deste trabalho para os profissionais que militam na área e para a 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, razão pela qual esperamos despertar e instigar entre os pesquisadores e avaliadores 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;

2. O emprego da classe de modelos GAMLSS possibilita ao engenheiro de avaliações uma interação abrangente com o mercado imobiliário e não limita-se apenas a atribuir valor a um bem. Além de permitir a modelagem direta de todos os parâmetros – e não apenas a média (μ) – da distribuição condicional do preço unitário, os modelos GAMLSS proporcionam a análise parcial dos termos aditivos suavizados, o que contribui para um diagnóstico mais verossímil do mercado;

3. 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 semi-paramétrica, bem como não prevê a estimação de outros parâmetros da variável resposta, como a dispersão (σ), que foi estimada neste trabalho e constitui uma importante informação para o entendimento do comportamento do mercado imobiliário. 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 os engenheiros e arquitetos especialistas em avaliações. Desta forma, esperamos contribuir com o crescimento técnico-científico da Engenharia de Avaliações no país;

4. A metodologia GAMLSS exposta neste trabalho pode ser de grande utilidade para a elaboração de plantas genéricas de valores pelas prefeituras para fins de cobrança do IPTU e ITBI, favorecendo uma política fiscal mais justa para os municípios e contribuintes. Aqui, o desafio é promover mais equidade (maior uniformidade dos níveis de avaliação entre imóveis distintos).

5.3 Sugestões para novas pesquisas

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

- Devido à existência de pesquisas recentes que sugerem a presença de correlação espacial em dados imobiliários (ver, 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[24] (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;

- Promover avaliações de bens com base em técnicas de estimação centílica via modelos GAMLSS;

- Promover avaliações de bens via modelos GAMLSS que incluam simultaneamente funções lineares e não-lineares (nos parâmetros) no mesmo modelo.

[24] 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

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

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

[4] 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.

[5] 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.

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

[7] 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.

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

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

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

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] Cribari-Neto, F. & Zarkos, S.G. (1999). R: yet another econometric programming environment. *Journal of Applied Econometrics* 14, 319-329.

[18] 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.

[19] 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.

[20] 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.

[21] Dantas, R.A. (2005). *Engenharia de Avaliações: Uma Introdução à Metodologia Científica*, $2^{\underline{a}}$ ed. São Paulo: Pini.

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

[23] Davidson, A.C. (2003). *Statistical Models*. Cambridge University Press.

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

[25] Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing*. 2nd ed. Marcel Dekker: New York.

[26] 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.

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

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

[29] 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.

[30] 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.

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

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

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

[34] 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.

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

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

[37] Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational Graphics and Statistics* 5, 299-314.

[38] 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.

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

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

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

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

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

[44] Mittelbach, F.; Goossens, M.; Braams, J.; Carlisl, D. & Rowley, C. (2004). *The LATEX Companion: Tools and Techniques for Computer Typesetting*. Addison Wesley, Boston.

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

[46] Pace, R.K. (1993). Non-Parametric Methods with Applications to Hedonic Models. *Journal of Real Estate Finance and Economics* 7, 185-204.

[47] 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.

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

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

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

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

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

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

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

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

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

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

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

[67] Silverman, B.W. 1984. Spline Smoothing: The Equivalent Kernel Method. *Annals of Statistics* 12, 896-916.

[68] 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.

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

[70] Stock, J. (1991). Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. In *Nonparametric and Semiparametric Methods in Econometrias 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.

[71] Thorsnes, P. & McMillen, D.P. (1998). Land Value and Parcel Size: A Semiparametric Analysis. *Journal of Real Estate Finance and Economics* 17, 233-244.

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

[73] 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.

[74] 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.