RBEs REVISTA BRASILEIRA DE ESTATÍSTICA
Desde 1940 revelando a Estatística Brasileira
IBGE 90

Avaliação dos preços dos imóveis venda: um estudo realizado em Tangará da Serra – MT

Elton Gean Araújo

egarauj@yahoo.com.br

Universidade Federal de Mato Grosso do Sul

Wilson Tarantin Junior

wilsontarantin@yahoo.com.br

Montvero Consultoria e Treinamento

Resumo: O mercado imobiliário tem grande dinamicidade, o que provoca alta variabilidade nos valores dos imóveis residenciais à venda por períodos economicamente dissemelhantes. O fato é que as técnicas de avaliação dos imóveis precisam ser constantemente atualizadas e revistas do ponto de vista técnico, para fornecerem valores de mercado dos imóveis verossímeis ao horizonte contextual. Para tanto, este trabalho teve como objetivo analisar dados referentes a valores de imóveis à venda localizados no município de Tangará da Serra, no Estado de Mato Grosso (MT). Utilizou-se um modelo de regressão linear múltipla Box Cox e um modelo aditivo generalizado para locação, escala e forma (GAMLSS) considerando-se a distribuição Box-Cox Cole e Green original, tipificando as principais covariáveis que impactaram nos preços desses imóveis, bem como, avaliando-se a capacidade preditiva de cada modelo testado. Os resultados indicaram melhor qualidade de ajuste e menor erro percentual absoluto médio (MAPE) para o GAMLSS, em que o valor do imóvel foi influenciado pela área construída, área do terreno, número de suítes, localização e estado de conservação. O modelo apresentou erro percentual abaixo de 20% e pode ser utilizado como uma alternativa na avaliação de imóveis no município.

Palavras-chave: Modelagem Estatística; Imobiliárias; Heterocedasticidade; Normas.

Abstract: The real estate market exhibits high dynamism, which results in substantial variability in the market values of residential properties for sale across different economic periods. Consequently, property valuation techniques must be continuously updated and technically reviewed to produce market value estimates that are plausible within the relevant contextual horizon. This study aimed to analyze data related to the sale prices of properties located in the municipality of Tangará da Serra, in the state of Mato Grosso (MT). Two modeling approaches were employed: a multiple linear regression model with Box-Cox transformation and a generalized additive model for location, scale, and shape (GAMLSS), considering the original Box-Cox Cole and Green distribution. The models aimed to identify the main covariates influencing property prices and to evaluate the predictive performance of each approach.
Results indicated that the GAMLSS model provided a better fit, with lower mean absolute percentage error (MAPE). The analysis revealed that property value was significantly influenced by the built area, land area, number of suites, location, and condition of conservation. The GAMLSS model achieved an error percentage below 20% and can be considered a viable alternative for property valuation in the municipality.

Keywords: Statistical Modeling; Real Estate; Heteroscedasticity; Standards.

1.INTRODUÇÃO

O mercado imobiliário tem relevância considerável para a economia do país, sendo um setor importante na geração de emprego e renda. Influenciado pelos cenários econômicos, é bastante dinâmico, passando por períodos de crise e expansão; atualmente, o setor passa por um período de recuperação, após atravessar o período pandêmico e alta inflacionária (Matos, 2017; Almeida, 2023).

O fato é que para acompanhar essa dinamicidade na área da construção civil, o mercado precisa de ferramentas precisas e atualizadas no campo da avaliação imobiliária. Metodologias científicas utilizadas na avaliação de imóveis são designadas pela Associação Brasileira de Normas Técnicas (ABNT, 2019), e envolvem técnicas estatísticas, tais como: regressão linear, regressão espacial e redes neurais artificiais.

De acordo com Dodt (2016) e Cordeiro (2022), o relatório final de avaliação de um imóvel deve indicar de forma fidedigna sua valoração, sendo, portanto, a capitalização de mercado de um imóvel o fator mais preponderante no processo decisório nos mais variados segmentos da engenharia de avaliações, como desapropriações, transferência de posse, locação, inventários, entre outros.

O método comparativo direto de dados de mercado tem sido amplamente aplicado na avaliação de imóveis urbanos. Neste caso, o valor de determinado imóvel é obtido comparando-se a outros imóveis com características similares (Oliveira, 2016; Araújo, 2021). Porém, existem muitos fatores intrínsecos e extrínsecos que são heterogêneos e impactam no valor do imóvel, os quais podem ser relacionados por meio de modelos de aprendizado de máquina supervisionados (Fávero et al., 2008).

Os elementos intrínsecos estão associados aos próprios aspectos dos imóveis como, a quantia de dormitórios e sanitários, extensão da área útil, enquanto as características extrínsecas são concernentes à localidade em que esses imóveis estão situados e na composição do entorno destes imóveis, como: hospitais, centros comerciais, espaços recreativos, unidades de ensino particulares, entre outros (Fávero et al., 2008).

Na literatura, pode-se encontrar trabalhos científicos que se utilizaram de conjuntos de dados de determinada região ou cidade brasileira para inferir sobre os preços de imóveis urbanos por meio de modelos de regressão linear múltipla, que estão inseridos no contexto de aprendizado de máquina supervisionado (Izbicki & Dos Santos, 2020). Em Pinto & Fernandes (2019), ajustou-se um modelo hedônico a partir de uma regressão linear múltipla para predizer os preços de imóveis em Lafaiete – MG, por meio de características dessas propriedades e uma amostra de tamanho 150; comprovando-se que covariáveis intrínsecas e a distância dos imóveis até Instituições de ensino e centro da cidade provocaram um aumento no valor de venda dos imóveis.

No estudo de Nunes et al. (2019), foi ajustado um modelo de regressão linear para determinação do valor de mercado de apartamentos residências na cidade de Fortaleza – CE; a técnica regressão ridge foi usada para problemas de multicolinearidade. Características como a quantidade de vagas de garagem e o valor do metro quadrado foram importantes na predição dos preços dos apartamentos, e foi produzido um erro percentual médio de 21,1 % ao serem analisadas as amostras testes.

Kappel & Teodoro (2020) aplicaram técnicas computacionais para a previsão de imóveis em Nova Friburgo – RJ, utilizando-se para o ajuste do modelo dados de 905 imóveis coletados pelo processo de web scraping automatizado a partir do site vivareal.com.br. O modelo apresentou uma capacidade preditiva de aproximadamente 75%, com desvio relativo médio de 25,22%, obtido na base de testes.

Em Duarte et al. (2022), foi proposto um modelo de regressão linear múltipla com objetivo de avaliar imóveis da área central do município de Florianópolis – SC. Para tanto, utilizou-se uma amostra de 208 imóveis com 8 covariáveis para o ajuste do modelo e, utilizando uma segunda amostra de 11 imóveis para avaliar o modelo. Este produziu um R^2 igual 96,32% e valores preditos com erros percentuais menores que 15%, podendo ser enquadrado no grau III de fundamentação da NBR 14653-2.

O trabalho de Dantas (2022) realizou predições dos valores dos imóveis estudados no município de Aracaju - SE por meio de modelagem estatística. Neste estudo, além do modelo de regressão linear, utilizou técnicas geoestatísticas para cálculo da matriz de pesos espaciais, com base em relações de distância ou contiguidade entre os dados de mercado, o que permite captar os efeitos de dependência espacial e incorporá-lo no modelo.

Avançando para uma modelagem mais complexa e possivelmente mais robusta no que tange à avaliação de imóveis, Cajias (2018) utilizaram modelos aditivos generalizados (GAM) e GAMLSS para predição de aluguéis em Munique, Alemanha; em que obtiveram resultados mais precisos na abordagem GAMLSS do que em relação ao GAM. Lee e Park (2019) utilizaram GAMLSS para obter estimativas de preços de imóveis em Seoul, Coréia do Sul, pois, segundo os autores, este tipo de modelo acomodava limitações da modelagem clássica (modelos hedônicos), tais como suposição de uma relação linear entre a variável resposta e as covariáveis.

Por fim, Costa et al. (2025), com intuito de realizar uma análise espaço-temporal dos preços de terrenos urbanos em Londrina, estado do Paraná, Brasil, propuseram um GAMLSS, gerando preço de superfície para a área de estudo, abrangendo qualquer período desejado. Os autores concluem que a metodologia de modelagem se mostrou robusta e apropriada para a situação estudada, levando vantagem em relação às abordagens convencionais.

Diante deste cenário, o intuito deste trabalho é analisar dados referentes aos preços de venda de uma amostra de imóveis localizados no município de Tangará da Serra, no Estado de Mato Grosso (MT), por meio de modelagem estatística, identificando as principais covariáveis que influenciam nos valores desses imóveis e com bom potencial de predição.

2. MATERIAL E MÉTODOS

A pesquisa foi realizada no município de Tangará da Serra – MT, município com população de 106434 habitantes; quinto maior Produto Interno Bruto per capita (PIB per capita) na região geográfica imediata (IBGE, 2022), apontando um crescimento de 34,5% em relação ao ano anterior, sendo os setores de Serviços e Agropecuária os principais responsáveis pelo incremento.

Os dados deste estudo foram obtidos a partir de pesquisa documental por meio da busca e tabulação das informações referentes às características dos imóveis (casas residenciais) junto às imobiliárias do município de Tangará da Serra – MT, entre os meses de janeiro e fevereiro de 2023; no total, foi coletada uma amostra sobre 340 imóveis.

As covariáveis consideradas nesse estudo são: número de quartos (N.Quartos), vagas de garagem (VG), número de suítes (N.Suites), número de banheiros (N.Banheiros), número de cozinhas (N.Cozinhas), estado de conservação do imóvel (E.conservação; baixa, média, alta), localização do imóvel (localização; Fora.Centro – fora do centro da cidade, Centro_AltoValor - no centro da cidade e regiões de alta valorização), área do terreno (A.Terreno;

em m²), área construída do imóvel (A.Construida;
em m²) e a variável resposta preço de venda dos imóveis nas imobiliárias no ato da consulta (Y, em Reais (R$)). Para a entrada no modelo da covariável categórica estado de conservação, foram criadas covariáveis dummies, ficando da seguinte forma: estado de conservação do imóvel (imóveis com baixa conservação: E.Conservação_Baixa; imóveis com média conservação: E.Conservação_Média; categoria de referência: imóveis com alta conservação). Os níveis da variável localização do imóvel para entrada no modelo foram: localização do imóvel (imóveis no centro da cidade ou região de alta valorização: Localização_Centro.AltoValor; categoria de referência: imóveis fora do centro).

Foi utilizado neste trabalho a abordagem quantitativa por meio de modelos estatísticos para analisar os dados de imóveis à venda, pois objetivou-se ajustar e testar um modelo que pudesse descrever adequadamente a relação entre a variável resposta preço do imóvel (*Y*) e as demais covariáveis (relacionadas às características do imóvel), e com boa capacidade preditiva.

Considerando-se n observações independentes e $X_j (j = 1, \dots k)$ covariáveis, o modelo de regressão linear múltipla pode ser expresso pela Equação 1 (Montgomery & Runger, 2008):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (1)$$

em que β_j , $j = 1, \dots k$ são os coeficientes de regressão, sendo parâmetros que representam a variação esperada em Y por unidade de variação em X_j, quando todos os outros regressores são mantidos constantes; β_0 é o intercepto do modelo e $\varepsilon \sim N(0, \sigma^2)$ é uma variável aleatória que representa o termo de erro, sendo a variabilidade em y que não pode ser explicada pelo efeito linear das k covariáveis.

A seleção de covariáveis para fazerem parte do modelo foi feita através do procedimento Stepwise (passo a passo), na direção "*backward*" e avaliadas em cada etapa pelo teste F (PEREIRA et. al, 2012). O método de estimação dos parâmetros do modelo de regressão linear múltipla utilizado neste trabalho foi o método dos mínimos quadrados ordinários.

Para a avaliação da qualidade de ajuste do modelo, foram analisados os testes de significância F e t para inferir sobre a significância global e individual das covariáveis, respectivamente; o coeficiente de determinação R^2 que representa qual a porcentagem da variável Y está sendo explicada pelo modelo; análise de resíduo para verificar suposições do modelo tais como normalidade e homogeneidade dos erros; multicolinearidade (definida como a presença de

fortes relações lineares entre as covariáveis, o que
acarreta na produção de coeficientes de regressão instáveis e com grande erro padrão, levando às inferências errôneas (Mourão et al. (2008)) e possível presença de pontos atípicos, isto é, outliers (Sweeney et al., 2013; Devore & Cordeiro, 2014). Ademais, o erro médio absoluto (MAE), a raiz do erro quadrático médio (RMSE) e MAPE também foram consideradas como medida de qualidade de ajuste no que tange à capacidade preditiva do modelo. As métricas MAE, RMSE e o MAPE pode sem ser obtidas pelas Equações 2, 3 e 4 (Ji et al., 2025):

$$MAE = N^{-1}\sum_{i=1}^{N}(|\hat{y}_i - y_i|) \tag{2}$$

$$RMSE = \sqrt{N^{-1}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{3}$$

$$MAPE = N^{-1}\sum_{1=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{4}$$

em que N representa o número de observações, $\hat{y}_i$ o valor predito para a i-ésima observação e y_i o valor observado para a i-ésima observação.

O modelo foi ajustado após uma transformação Box-Cox na variável dependente preço de vendas dos imóveis, com intuito de melhorar sua precisão (Blum et al., 2022), bem como reparar desvios de normalidade e homogeneidade dos resíduos. Tal transformação foi proposta por Box & Cox (1964), consiste em determinar o parâmetro $\lambda \in \mathbb{R}$ a ser aplicado a um vetor $Y \in \mathbb{R}^n$ de acordo com a Equação 5:

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda} & se\ \lambda \neq 0 \\ \ln Y_i & se\ \lambda = 0 \end{cases} \tag{5}$$

De acordo com Blum et. al (2022), muitos critérios têm sido abordados para obter o λ ótimo, sendo o estimador de máxima verossimilhança (MLE), introduzido por Box & Cox (1964) o mais utilizado.

Por fim, na possibilidade de melhorar ainda mais a capacidade preditiva do modelo, ajustou-se aos dados o GAMLSS, que fornece uma estrutura de modelagem mais flexível, em que a variável de interesse pode seguir uma distribuição que não seja necessariamente gaussiana, como distribuições altamente assimétricas. Permite-se ainda, de acordo com Rigby & Stasinopoulos (2005), que todos os parâmetros sejam modelados como funções paramétricas e/ou não paramétricas (funções de suavização) aditivas de covariáveis de interesse

Considere a variável resposta $y_i (i = 1, \ldots, n)$, com função densidade de probabilidade $f(y_i|\theta_i)$, em que $\theta_i = (\mu_i, \sigma_i, \nu_i, \tau_i)^T$ corresponde ao vetor de parâmetros. Seja $y' = (y_1, y_2, \cdots, y_n)$ o vetor variável resposta com n observações independentes. Para *k = 1, 2, ..., 4* parâmetros, sejam $g_k(.)$ funções de ligação monótonas conhecidas relacionando os parâmetros e as covariáveis da seguinte forma (Rigby & Stasinopoulos, 2005):

$$Y \overset{ind}{\sim} D(\mu_i, \sigma_i, \nu_i, \tau_i)$$

$$g_1(\mu_i) = \eta_1 = X_1\beta_1 + \sum_{i=1}^{J_1} Z_{j1}\gamma_{j1} \quad (6)$$

$$g_2(\sigma_i) = \eta_2 = X_2\beta_2 + \sum_{i=1}^{J_2} Z_{j2}\gamma_{j2} \quad (7)$$

$$g_3(\nu_i) = \eta_3 = X_3\beta_3 + \sum_{i=1}^{J_3} Z_{j3}\gamma_{j3} \quad (8)$$

$$g_4(\tau_i) = \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}\gamma_{j4} \quad (9)$$

em que $\mu_i, \sigma_i, \nu_i, \tau_i$ e η_k são vetores de dimensão n, X_k são matrizes de ordem $n \times J'_k$ associados com os efeitos fixos β_k, de dimensão $J'_k \times 1$; Z_{jk} são matrizes conhecidas de dimensão $n \times q_{jk}$ associados com os efeitos aleatórios γ_{jk} de dimensão $q_{jk} \times 1$ com distribuição normal multivariada; J'_k representa o número de covariáveis usadas nos efeitos fixos (que está relacionado às covariáveis de efeito linear, em que não foram aplicados suavizadores splines, parte paramétrica do modelo) e J_k representa o número de efeitos aleatórios em η_k (que está relacionado às covariáveis de efeito não linear, em que foram aplicados suavizadores splines, parte não paramétrica do modelo).

Para definição da distribuição de probabilidade com melhor ajuste à variável resposta preço dos imóveis, utilizou-se a função fitDist considerando todas as distribuições de probabilidade para valores positivos disponíveis no pacote gamlss (Rigby & Stasinopoulos, 2005), comparadas pelo critério de informação generalizado de Akaike (GAIC); sem ajuste dos efeitos das covariáveis, nesta etapa. A distribuição com melhor ajuste aos valores dos imóveis foi a Box-Cox Cole e Green original (BCCGo). Seja $Y \sim D(\theta)$, sendo D a distribuição da variável resposta e θ o vetor de parâmetros da distribuição, a

distribuição de Box-Cox Cole e Green original (BCCGo) (Cole & Green, 1992) é definida da transformação:

$$\mathrm{Z} = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right] & se\ \nu \neq 0 \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right) & se\ \nu = 0 \end{cases} \tag{10}$$

onde se supõe que Z tenha uma distribuição normal truncada; μ é o parâmetro de posição, $\sigma > 0$ é o parâmetro de dispersão e $-\infty < \nu < +\infty$ é o parâmetro de assimetria, modelados por funções de ligação logarítmica, logarítmica e identidade, respectivamente (Stasinopoulos et al., 2017). Os efeitos não lineares foram modelados por funções de suavização spline cúbica.

De acordo com Rigby & Stasinopoulos (2005), os parâmetros do GAMLSS podem ser estimados pelo algoritmo RS, usando a máxima verossimilhança penalizada; método que foi utilizado neste trabalho. Para a seleção das covariáveis e apontamento do melhor GAMLSS, foi considerado o GAIC com penalização igual a 2, através da função stepGAIC do pacote gamlss (Rigby & Stasinopoulos, 2005) , na direção "*backward*", de acordo com Venables and Ripley (2002); e quanto à verificação da qualidade do ajuste, utilizou-se, de acordo com Buuren & Fredriks (2001), o gráfico worm plot (gráfico de minhoca), e, caso os resíduos estejam situados no interior da banda de confiança de 95%, o modelo está bem ajustado; bem como as métricas MAE, RMSE e o MAPE. Todas as análises foram realizadas pelo software de código fonte aberto R (R CORE TEAM, 2025).

3.RESULTADOS E DISCUSSÃO

Na Tabela 1 tem-se um resumo estatístico das covariáveis qualitativas e quantitativas envolvidas no estudo, bem como da variável resposta, o que é importante para entender o comportamento dos dados e evitar extrapolação na predição.

Tabela 1 - Resumo Estatístico das variáveis da pesquisa

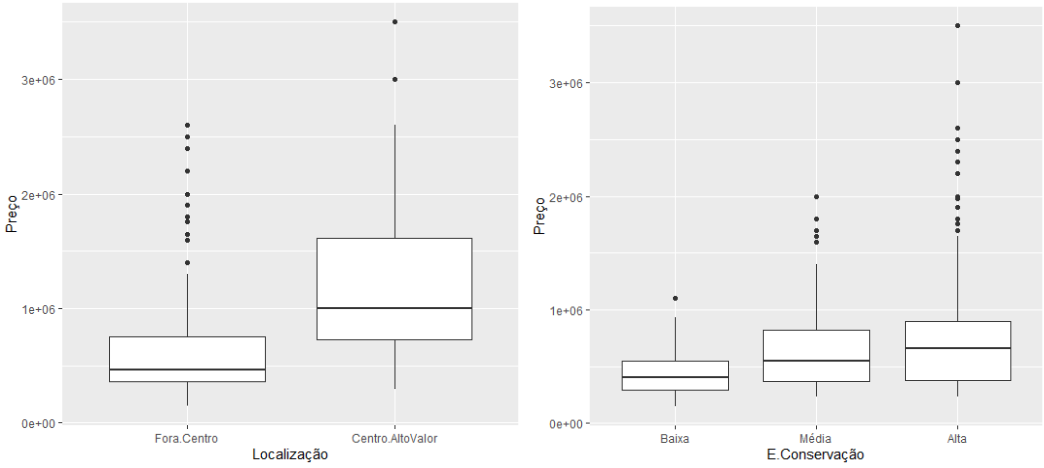
Variáveis Quantitativas	Mínimo	Média	Mediana	Desvio Padrão	Máximo	Assimetria	Curtose
A.Construida	52,00	164,87	156,00	83,47	504,74	1,02	1,00
A.Terreno	101,50	352,12	275,00	198,38	2.700,00	5,40	56,13
N.Suites	0,00	1,10	1,00	0,88	4,00	1,01	0,78
N.Quartos	1,00	1,84	2,00	0,69	5,00	0,64	1,03
N.Banheiros	1,00	1,53	1,00	0,83	6,00	1,99	4,60
VG	0,00	1,82	2,00	1,01	7,00	1,01	2,37
N.Cozinhas	1,00	1,13	1,00	0,34	3,00	2,48	4,93
Preço	145.000,00	72.9847,06	580.000,00	533.888,07	3.500.000,00	2,05	4,60
Variáveis Qualitativas	Frequência Absoluta	Percentual					
Localização (Fora.Centro)	268	78,80					

Localização (Centro.AltoValor)	72	21,20
E.Conservação (Baixa)	44	12,90
E.Conservação (Média)	118	34,70
E.Conservação (Alta)	178	52,40

Fonte - Dados originais da pesquisa.

A Figura 1 mostra o comportamento da variável resposta (valor dos imóveis) para cada categoria das covariáveis qualitativas.

Figura 1 - Boxplot da variável resposta preço dos imóveis em função das covariáveis categóricas. (a) Localização; (b) Estado de Conservação



(a) (b)

Fonte - Resultados originais da pesquisa.

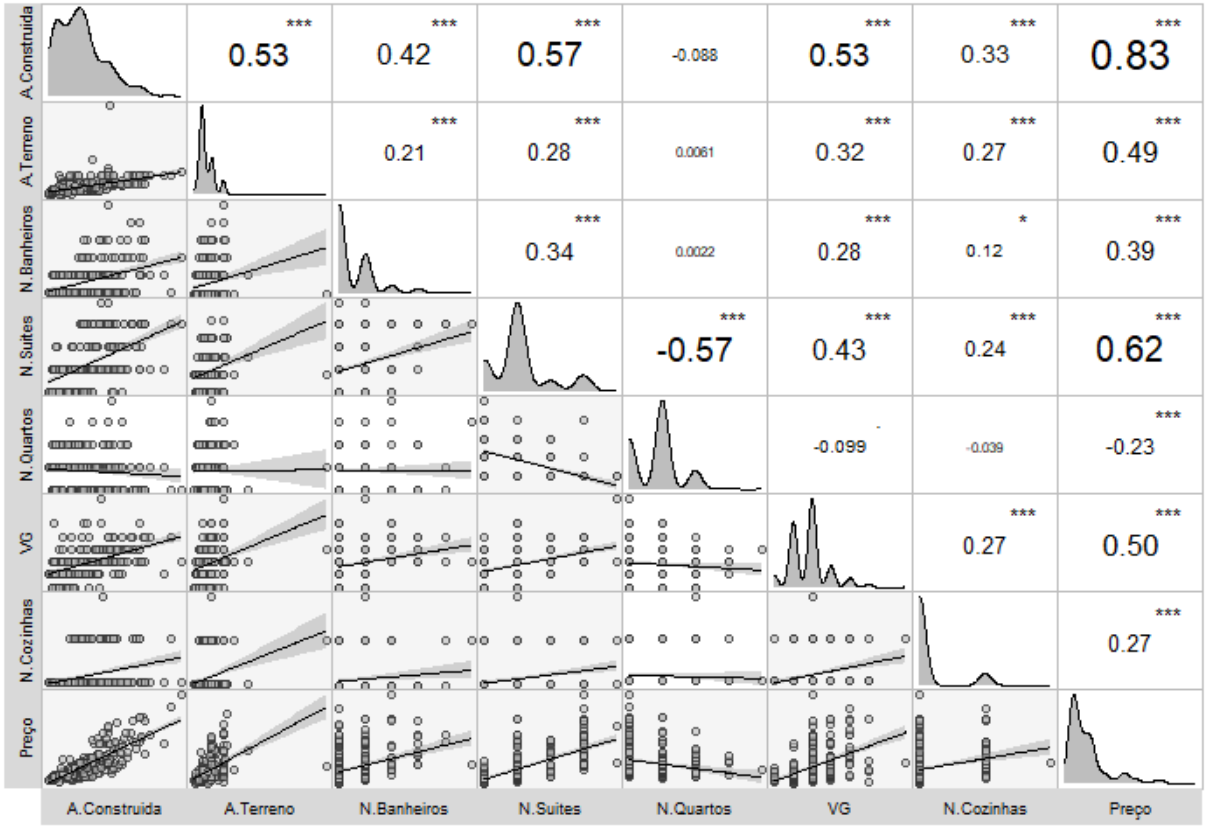
O gráfico da Figura 1, parte (a), evidencia que os imóveis localizados no centro da cidade e em regiões valorizadas tem valor mediano (R$ 1.000.000,00) superior ao valor mediano dos imóveis localizados fora do centro da cidade (R$ 460.000,00). Fato também observado em relação ao estado de conservação dos imóveis (Figura 1, parte (b)), sendo que, quanto mais conservado o imóvel, maior seu preço de venda, mais especificamente, os imóveis com baixa conservação apresentaram valor mediano igual a R$ 400.000,00; os de média conservação R$ 545.000,00; e, alta conservação R$ 655.000,00. Kappel & Teodoro (2020), ao construírem um modelo para predição de preços de imóveis à venda no município de Nova Friburgo, RJ destacaram forte valorização dos imóveis na região central da cidade, se comparado aos imóveis localizados fora do centro da cidade. Os autores justificam tal fato pelo centro da cidade concentrar o maior número de atividades comerciais e financeiras.

Nota-se ainda, pela Tabela 1, maiores assimetrias e curtoses para as covariáveis área do terreno, número de cozinhas, número de banheiros e preço

dos imóveis. Em Costa et al. (2025), ao realizarem
uma análise espaço-temporal dos preços da terra urbana encontraram valores expressivos na magnitude de 4,28 e 23,89 para assimetria e curtose, respectivamente, para a variável valores unitários (R$/m²), evidenciando que essas características são essenciais para selecionar uma distribuição apropriada para modelar a variável resposta de forma eficaz.

No que se refere à associação linear (correlação de Pearson e diagrama de dispersão) da variável preço dos imóveis em relação às covariáveis independentes contínuas, os resultados são mostrados na Figura 2.

Figura 2 - Correlações e diagrama de dispersão entre a variável resposta preço dos imóveis e as covariáveis independentes contínuas



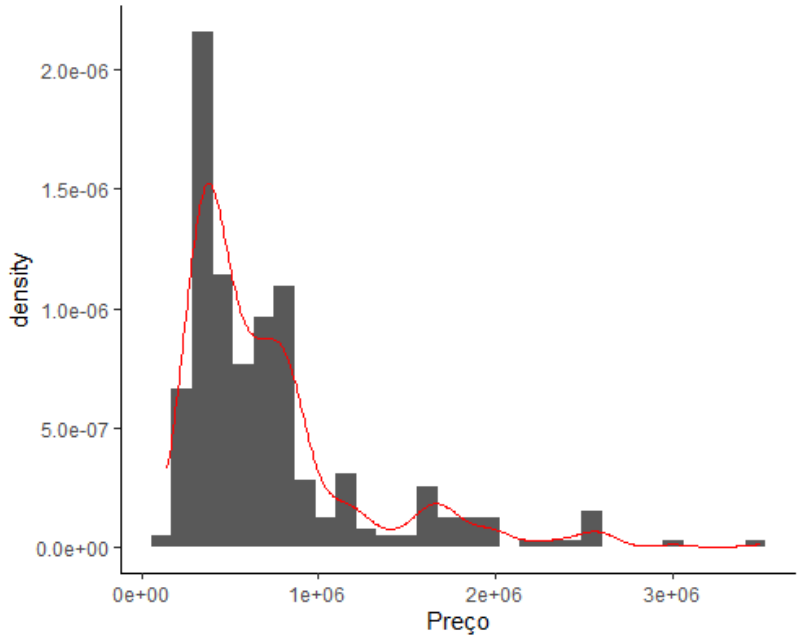
Fonte - Resultados originais da pesquisa.

Os resultados da Figura 2 mostram correlações lineares de Pearson significantes entre a variável preço e todas as covariáveis independentes contínuas, com destaque para número de suítes e área construída, com valores acima 0,6. Fato curioso foi a correlação negativa entre número de quartos e preço do imóvel, o que pode ser justificado por imóveis maiores e antigos (com maior número de quartos) fora do centro da cidade e de regiões valorizadas. Em

Kappel & Teodoro (2020), a área construída (denominada área total do imóvel) também foi a variável mais importante no valor de um imóvel. Resultado similar encontrado por Nunes et al. (2019), em que a área construída (denominada área útil) foi a variável que apresentou a maior correlação com o preço do imóvel.

Uma característica importante, que não poderia deixar de ser mencionada na análise descritiva, é o comportamento da distribuição da variável resposta preço dos imóveis. Como pode ser verificado na Figura 3, é notada forte assimetria positiva nos valores desta variável; Zilli et al. (2022), encontraram leve assimetria a direita na variável valor unitário dos imóveis, em estudo realizado em Florianópolis, SC.

Figura 3 - Histograma variável resposta preço dos imóveis com função densidade



Fonte - Resultados originais da pesquisa.

A próxima etapa foi especificação e ajuste do modelo de regressão linear múltipla. Mesmo com a remoção de observações atípicas, que apresentaram maiores desvios de normalidade, os resíduos não seguiram distribuição normal; e, somando-se ao fato de ainda não terem variância constante, o modelo de regressão deixa de ser válido (Marôco, 2021). Ademais, observou-se, por exemplo, que o MAPE foi igual a 27,12 %. De acordo com a NBR 14653-3 (ABNT, 2019), erros percentuais médios até 20,00% são considerados aceitáveis.

Portanto, utilizou-se da transformação Box Cox para corrigir os problemas referentes à violação dos pressupostos do modelo (Morais et al., 2020). A estimativa do λ ótimo via MLE foi -0,3998. Após a transformação, o

modelo foi reajustado e denominado modelo Box-Cox. A Tabela 2 exibe o resumo do resultado do ajuste do modelo Box-Cox.

Tabela 2 - Resumo do resultado do ajuste do modelo Box-Cox

Efeito	Estimativa	Erro Padrão	valor-p
Intercepto	2,48	$1{,}95x10^{-4}$	<0,01
A.Construida	$2{,}03x10^{-5}$	$1{,}30x10^{-6}$	<0,01
A.Terreno	$1{,}41x10^{-6}$	$4{,}45x10^{-7}$	0,01
N.Suites	$7{,}19x10^{-4}$	$1{,}07x10^{-4}$	<0,01
Localização_Centro.AltoValor	$5{,}27x10^{-4}$	$2{,}07x10^{-4}$	0,029
E.Conservação_Baixa	$-1{,}70x10^{-3}$	$2{,}44x10^{-4}$	<0,01
E.Conservação_Média	$-6{,}99x10^{-4}$	$1{,}65x10^{-4}$	<0,01
R^2	0,7770		

Fonte - Dados originais da pesquisa.

Observou-se significância estatística ao nível de 5% para todas as covariáveis independentes do modelo Box-Cox, houve pequeno acréscimo do R^2 do modelo anterior para este (0,0071). Não se observou presença de multicolinearidade (todos os *Variance Inflation Factor (VIF)* menores que 2,2*).* Os pressupostos do modelo Box-Cox foram todos atendidos (Tabela 3).

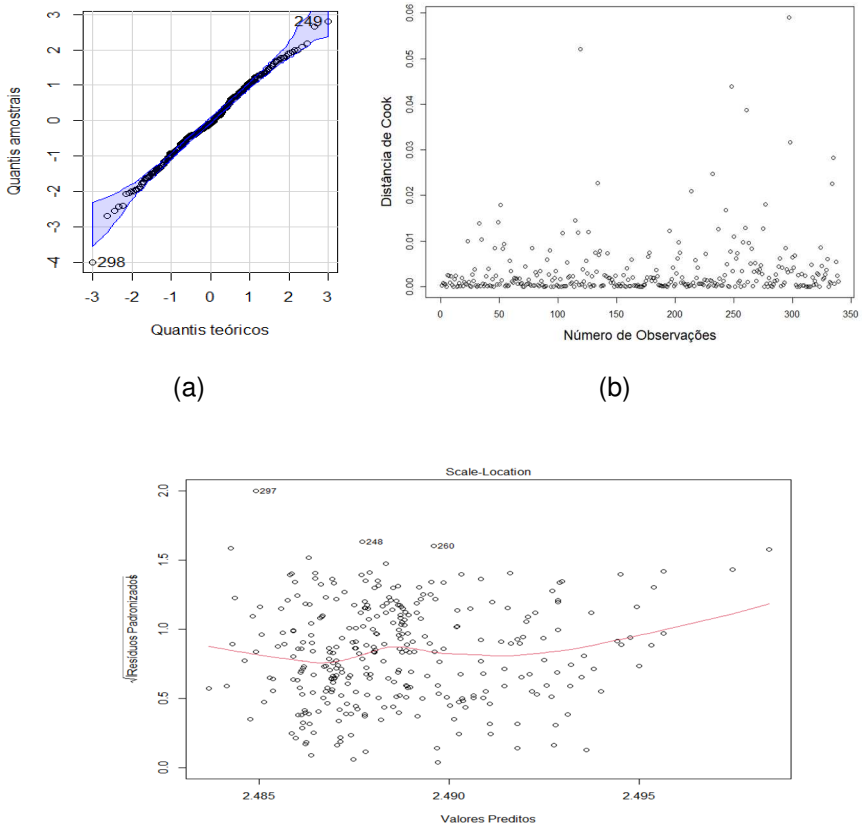
Tabela 3 - Resultado dos Testes Diagnósticos do modelo Box-Cox

Teste Estatístico	Estatística do Teste	valor-p
Shapiro - Francia	W = 0,9935	0,1365
Breusch Pagan	χ^2 = 0,7924	0,3734

Fonte - Dados originais da pesquisa.

A acomodação dos pressupostos no modelo Box-Cox pode ser verificada, em relação à normalidade dos resíduos, também pelo gráfico Quantil – Quantil (Figura 4(a)), onde se percebe uma melhora significativa, em que apenas a observação 298 apresenta desvio de normalidade; homogeneidade de variância (Figura 4(c)), bem como nenhuma distância de Cook acima de 1, o que indica que não há pontos de alta influência (Figura 4(b)) (Cook & Weisberg,1982). Porém, ainda assim, como no modelo de regressão linear múltipla ajustado anteriormente (sem a transformação), o modelo Box-Cox apresenta MAPE igual a 24,77%, portanto, acima dos 20,00% considerados limite pela NBR 14653-3 (ABNT, 2019). Ademais, as métricas RMSE e MAE foram 1192470 e 270920, respectivamente.

Figura 4 – Gráficos para análise de resíduos do modelo Box Cox.



Fonte - Resultados originais da pesquisa.

Desta forma, a próxima etapa foi o ajuste do GAMLSS, visto que a distribuição da variável resposta (preço do imóvel) apresenta comportamento assimétrico (Figura 3), e de acordo com Rigby & Stasinopoulos (2005), distribuições com este tipo de comportamento podem ser utilizadas por essa classe de modelos.

Levando-se em consideração o worm plot e as métricas MAE, RMSE e o MAPE, o melhor ajuste foi para o GAMLSS com suavizadores *splines* cúbicos (cs) para as covariáveis “A.Construida” e “A.Terreno”.

Quanto ao efeito linear das covariáveis na variável resposta valor dos imóveis, apenas a variável "E.Conservação_Baixa" para predição de ν não foi significante. Pode-se notar ainda, que, para a predição do parâmetro de posição (μ), todas exercem relação direta, com exceção de "E.Conservação_Baixa" e "E.Conservação_Média". Tal resultado mostra que os imóveis com estado de conservação "baixo" e "médio", apresentam valores menores que os imóveis com estado de conservação "alto", sendo este último categoria de referência. O GAMLSS é descrito pelas equações 11, 12 e 13:

$$log(\mu) = 0{,}000124 + cs(\text{A. Construida}, \text{df} = 8) + cs(\text{A. Terreno}, \text{df} = 3) + 0{,}133 * \text{N. Suites} + 0{,}214 * \text{Localização_Centro. AltoValor} + -0{,}335 * \text{E. Conservação_Baixa} + -0{,}124 * \text{E. Conservação_Média} \quad (11)$$

$$log(\sigma) = -1{,}9431 + cs(\text{A. Construida}, \text{df} = 8) + 0{,}4001 * \text{E. Conservação_Baixa} \quad (12)$$

$$\nu = 0{,}1782 + cs(\text{A. Construida}, \text{df} = 8) + 0{,}0977 * \text{E. Conservação_Baixa} \quad (13)$$

Mais especificamente, pode-se notar que, a cada suíte adicional, o valor mediano de imóvel tende a ser 1,14 vezes maior; imóveis que estão no centro da cidade e regiões de alta valorização aumentam seu preço mediano em 1,24 vezes, em relação aos que estão nessas regiões. Em Zhang (2021), trabalho que abordou o ajuste de um modelo estatístico para predizer imóveis na cidade de Boston, nos Estados Unidos, apontou que, o preço dos imóveis tem tendência de queda para casas que estão localizadas em regiões suburbanas.

Também se observa que, imóveis que tem estado de conservação baixa e média, têm valores medianos 1, 4 e 1,13 menores, respectivamente, em relação a imóveis com estado de conservação alta. Lee e Park (2019) mostram que, ao analisar o efeito não linear da covariável idade do imóvel, há uma depreciação acelerada no preço de venda ao longo dos primeiros 10 anos. Fato que pode estar relacionado ao estado de conservação do imóvel.

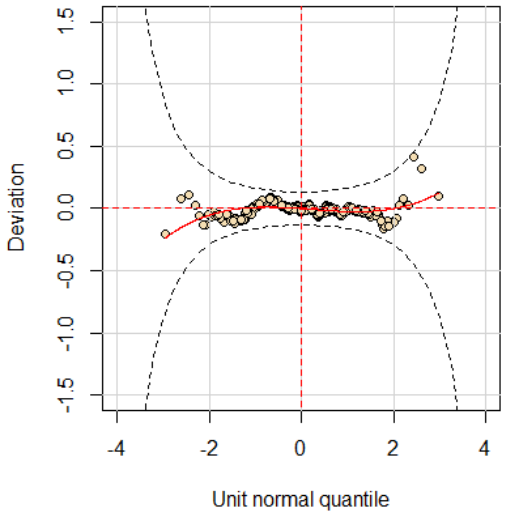
O comportamento dos resíduos do GAMLSS pode ser visualizado na Figura 5, por meio do Gráfico Worm plot. Como observado, os resíduos estão situados no interior da banda de confiança de 95%, logo o GAMLSS está bem ajustado, conforme Buuren & Fredriks (2001).

Também foi calculado o MAPE, assim como feito para os modelos ajustados anteriormente. O MAPE do GAMLSS foi 18,85%, dentro do limite aceitável de 20,00%, de acordo com NBR 14653-3 (ABNT, 2019), mostrando que este modelo é aceitável, com boa qualidade de ajuste e capacidade

preditiva. Ademais, as métricas RMSE e MAE foram 246547 e 147822, respectivamente.

Na Figura 6, pode-se observar os gráficos relativos às curvas de suavização dos termos aditivos do GAMLSS (não lineares) em relação ao parâmetro de posição (μ). A faixa cinza em torno das curvas são os erros padrões pontuais. Nota-se que a contribuição do termo aditivo ajustado para a variável "A.Construida" (parte (a)) é basicamente crescente, com inclinações maiores para intervalos de imóveis com até 200 m², entre aproximadamente 250 m² e 350 m² e maiores que 400 m² de área construída.

Figura 5 - Gráfico worm-plot – GAMLSS



Fonte - Resultados originais da pesquisa.

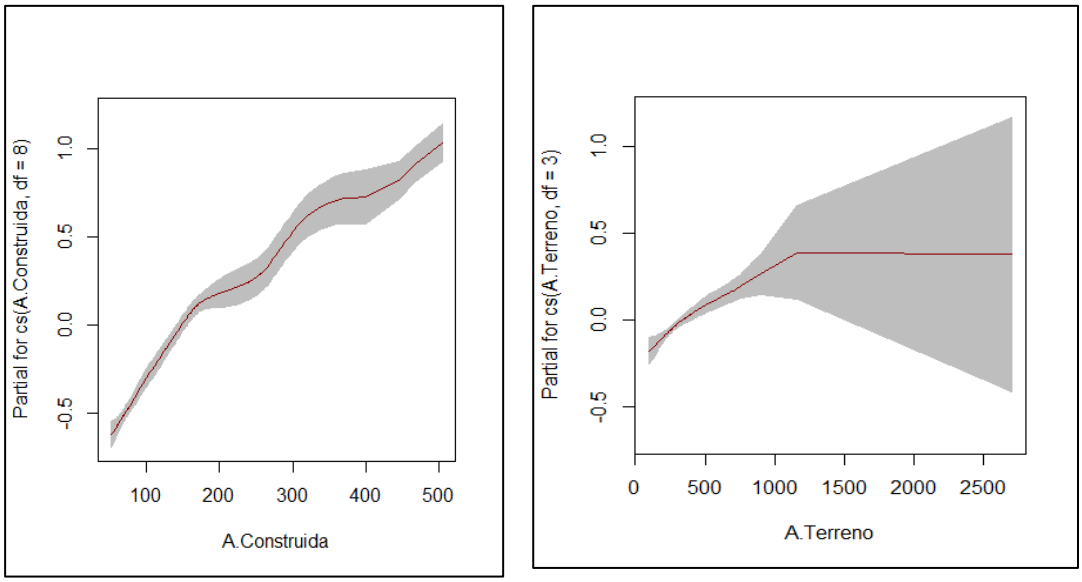
No trabalho de Lee e Park (2019), a pesquisa mostrou que os preços dos imóveis tiveram uma tendência de queda para imóveis com área 0m² e 200m², mas aumenta abruptamente para imóveis com área acima 200 m².

Ainda com relação à Figura 6 (parte (b)), tem-se a contribuição do termo aditivo ajustado para a variável "A.Terreno", com crescimento para os imóveis com terrenos até aproximadamente 1200 m², e, uma estabilização para imóveis com área de terreno acima dessa metragem.

Analisando o gráfico relativo à curva de suavização do termo aditivo do GAMLSS (efeito não linear) em relação ao parâmetro de dispersão (σ), Figura 7 (parte (a)), a contribuição do termo aditivo ajustado para a variável "A.Construida" é estável para imóveis com área construída de até

aproximadamente 400 m², e, decrescente após este tamanho. Já em relação ao gráfico relativo à curva de suavização do termo aditivo do GAMLSS (efeitos não linear) em relação ao parâmetro de forma (ν)), Figura 7 (parte (b)), a contribuição do termo aditivo para a variável "A.Construida" é constante até imóveis com aproximadamente 360 m² de área construída, e, a partir desta metragem a contribuição aumenta. Liao et al. (2022), cujo intui foi ajustar um modelo de regressão para predição de imóveis em Pequim e Chongqing, na China, observaram que a área do imóvel e a presença de elevador tiveram impacto significativo no preço dos imóveis, concluindo que casas com maiores áreas, tem maiores preços.

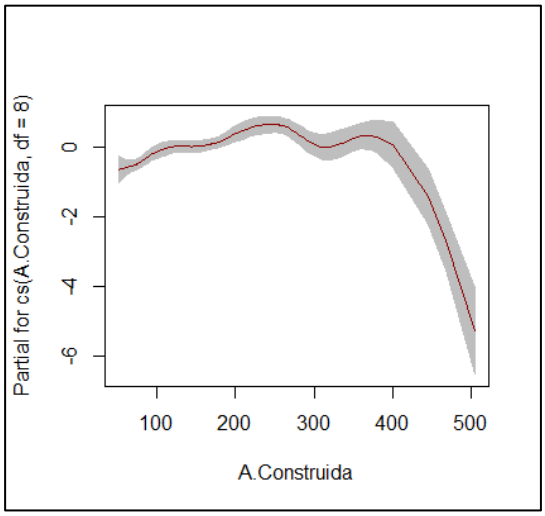
Figura 6 - Gráficos dos termos aditivos suavizados - GAMLSS - parâmetro de posição (μ).

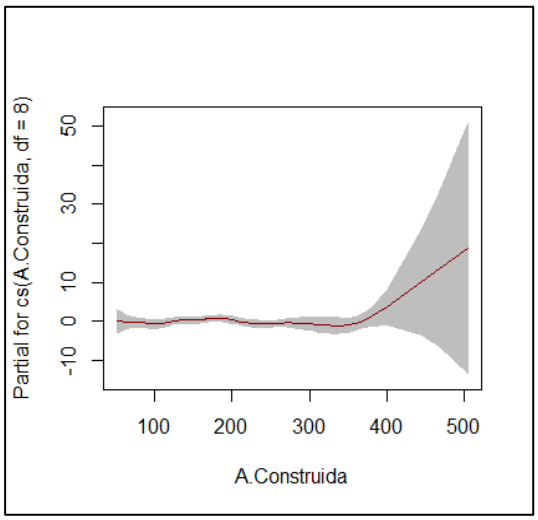


(a) (b)

Fonte - Resultados originais da pesquisa.

Figura 7 - *Gráficos dos termos aditivos suavizados – GAMLSS* - σ (a) e ν (b)





(a) (b)

Fonte - Resultados originais da pesquisa.

Em Lee e Park (2019) , o GAMLSS espacial ajustado para predição de imóveis à venda, com suavizadores *splines* penalizados (pb) para as covariáveis "área da sala", "andar em que a casa está localizada" e "idade do imóvel", com a distribuição exponencial gaussiana, com três parâmetros: de posição (μ), de dispersão (σ) e forma (assimetria, ν), modelados por funções de ligação identidade, logarítmica e logarítmica, respectivamente, foi o que obteve a melhor performance dentre os testados no estudo. Neste trabalho, O GAMLSS, também foi o que obteve a melhor capacidade preditiva (Tabela 4).

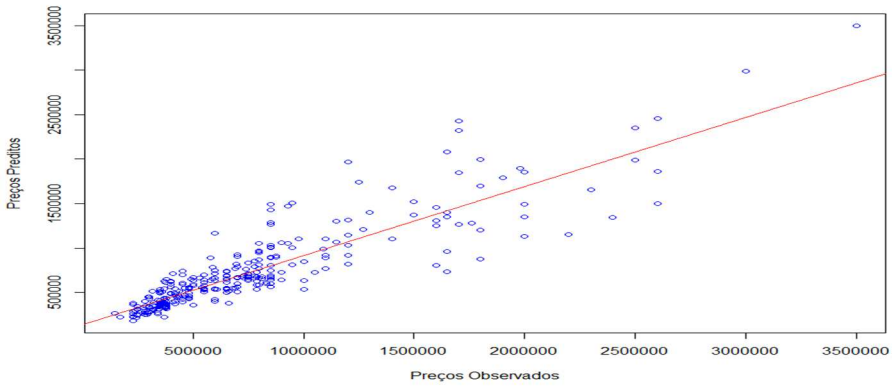
Tabela 4 - Comparação da Precisão dos Modelos avaliados

Modelos Testados	RMSE	MAE	MAPE (%)
Modelo Regressão Box Cox	1192470	270920	24,77
(GAMLSS)	246547	147822	18,85

Fonte - Resultados originais da pesquisa.

O gráfico dos valores observados *vs* ajustados do GAMLSS é apresentado na Figura 8. A correlação entre os valores observados e preditos é 0,89, evidenciando que o GAMLSS apresentou performance satisfatória, sendo melhor que o modelo regressão Box Cox (Tabela 4).

Figura 8 - Gráfico dos valores observados *vs* ajustados – *GAMLSS*



Fonte - Resultados originais da pesquisa.

4.CONCLUSÕES

O GAMLSS ajustado para locação, escala e forma, considerando-se a distribuição Box-Cox Cole e Green original, apresentou boa performance preditiva com erro percentual abaixo de 20%, e então, pode ser utilizado como uma técnica alternativa para o mercado imobiliário na avaliação de imóveis residenciais à venda no município de Tangará da Serra - MT. As covariáveis área construída, área do terreno, número de suítes, localização e estado de conservação do imóvel foram importantes e significantes para o modelo, sendo então os principais fatores na explicação/formação dos preços dos imóveis à venda no município. Este trabalho pôde fornecer uma metodologia mais robusta no que tange à avaliação dos preços de imóveis à venda, se comparado aos modelos de regressão clássicos. Em estudos futuros, pretende-se a inserção de alguma componente espacial neste modelo, bem como utilização de árvores de decisão e redes neurais com a finalidade de aferir superioridade preditiva.

5.REFERÊNCIAS BIBLIOGRÁFICAS

ABNT. (2019). NBR14653-3: Avaliação de Bens, Parte 3: Imóveis Rurais. Rio de Janeiro: ABNT.

Almeida, G. R. (2023). Avaliação do comportamento do mercado imobiliário residencial vertical de padrão econômico, diante dos impactos da pandemia da COVID-19, nas cidades de Goiânia/GO e Aparecida de Goiânia/GO. Monografia, Instituto Federal de

Educação, Ciência e Tecnologia de Goiás.
Departamento de áreas acadêmicas III, Engenharia Civil, Goiânia.

Araújo, M. M. (2021). Avaliação de imóvel residencial urbano pelo método comparativo direto de dados de mercado. Monografia, Universidade Federal do Ceará, Centro de Tecnologia, João Pessoa.

Blum, L., Elgendi, M., & Menon, C. (2022). Impact of Box-Cox Transformation on Machine-Learning Algorithms. Frontiers in Artificial Intelligence, 5.

Buuren, S. V., & Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. Statistics in Medicine, 20(8), 1259-1277.

Cajias, M. (2018), "Is there room for another hedonic model? The advantages of the GAMLSS approach in real estate research", Journal of European Real Estate Research, Vol. 11 No. 2, pp. 204-245. https://doi.org/10.1108/JERER-07-2017-0025.

Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. Statistics in medicine, 11(10), 1305-1319.

Cook, R. D. W. S. (1982). Residuals and influence in regression.

Cordeiro, C. C. S. (2022). Um breve estudo sobre avaliações de imóveis urbanos: uma revisão de literatura (Trabalho de conclusão de curso). Universidade Federal do Ceará, Fortaleza.

Costa, F. J., Pescim, R. R., Nakamura, L. R., Urbano, M. R., Silva, V. C., & Ramires, T. G. (2025). Spatiotemporal analysis of urban land prices: a distributional regression approach. Spatial Economic Analysis, 1–13. https://doi.org/10.1080/17421772.2025.2486144.

Dantas, R. A. (2022). Modelos geoespaciais aplicados na avaliação de imóveis: um estudo de caso em Aracaju, Brasil. Revista Valorem, 1(1), 42-51.

De Morais, E. O., et al. (2020). Aplicação de modelos de regressão linear e não linear para estimativa de volume de biomassa e estoque de carbono. Brazilian Journal of Development, 6(7), 45621-45632.

Devore, J. L., & Cordeiro, M. T. A. (2014). Probabilidade e estatística: para engenharia e ciências. Cengage Learning Edições Ltda.

Dodt, E. F. (2016). Avaliação de imóvel: elaboração do laudo de avaliação pelo método comparativo direto. In Anais do XX Congresso Brasileiro de Comunidades (pp. 12-21). Salvador: CBC.

Duarte, L., Hochheim, N., & Reginato, V. (2022). Regressão linear múltipla aplicada em avaliação imobiliária: estudo de caso na área central de Florianópolis. Concilium, 22(6), 880-899.

Fávero, L. P. L., Belfiore, P. P., & Lima, G. A. S. (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 (São Paulo), 38, 73-96.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. *Produto Interno Bruto dos Municípios 2022*. Rio de Janeiro, 2022. Disponível em: https://cidades.ibge.gov.br/brasil/mt/tangara-da-serra/panorama. Acesso em: 22/03/2023.

Izbicki, R., & Dos Santos, T. M. (2020). Aprendizado de máquina: uma abordagem estatística. Rafael Izbicki.

Ji, H., Xu, Y., & Teng, G. (2025). Predicting egg production rate and egg weight of broiler breeders based on machine learning and Shapley additive explanations. Poultry Science, 104(1), 104458.

Kappel, M. A. A., & De Almeida Teodoro, L. (2020). Modelo Hedônico para Estimação do Valor de Imóveis: Aplicação em Nova Friburgo-RJ. VETOR - Revista de Ciências Exatas e Engenharias, 30(1), 28-37.

Lee, C., & Park, K. (2019). Estimating Housing Prices through a Spatial GAMLSS Modeling Approach. Journal of the Korean Geographical Society, 54(2), 271-284.

Liao, X., Deng, M., & Huang, H. (2022). Analisando relações espaciais multiescalares entre o preço da habitação e os fatores do ambiente visual. Ciências Aplicadas , 12 (1), 213. https://doi.org/10.3390/app12010213.

Marôco, J. (2021). Análise Estatística com o SPSS Statistics v18–v27. Lisboa: Report Number.

Matos, T. O. (2017). A expansão do mercado imobiliário no Brasil: Um paralelo entre a evolução dos preços no mercado brasileiro e a bolha imobiliária norte-americana. Monografia, Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Ciências Econômicas, Rio de Janeiro.

Montgomery, D. C., & Runger, G. C. (2008). Estatística Aplicada e Probabilidade para Engenheiros (2a ed.). Rio de Janeiro: LTC Editora.

Mourão, G. B., Petrini, J., Dias, R. A. P., Eler, J. P., Ferraz, J. B. S., Balieiro, J. C. D. C., & Mattos, E. C. (2008). Diagnóstico de multicolinearidade em modelos aditivo-dominante em uma população de bovinos de corte compostos (Bos taurus x Bos indicus). 45ª Reunião Anual da Sociedade Brasileira de Zootecnia.

Nunes, D. B., Barros, J. P., & Freitas, S. M. (2019). Modelo de regressão linear múltipla para avaliação do valor de mercado de apartamentos residenciais em Fortaleza, CE. Ambiente Construído, 19, 89-104.

OLIVEIRA, H. (2016). Avaliação de um imóvel residencial multifamiliar urbano usando o método comparativo direto de dados de mercado. Monografia, Universidade Federal da Paraíba, Centro de Tecnologia, João Pessoa.

Pereira, J. C., Garson, S., & Araújo, E. G. (2012).
Construção de um modelo para o preço de venda de casas residenciais na cidade de Sorocaba-SP. Gestão da Produção, Operações e Sistemas, 7(4), 153-167.

Pinto, V. H. L., & Fernandes, R. A. S. (2019). Análise de preços hedônicos no mercado imobiliário residencial de Conselheiro Lafaiete, MG. Interações (Campo Grande), 20, 627-643.

R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rigby R.A. and Stasinopoulos D.M. (2005). Generalized additive models for location, scale and shape,(with discussion), Appl. Statist., 54, part 3, pp 507-554.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3), 507-554.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). Flexible regression and smoothing: using GAMLSS in R (p. 549). CRC Press, Taylor & Francis Group.

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. Scientific Programming, 2021(1), 7678931.

Zilli, C. A., Droubi, L. F. P., & Hochheim, N. (2022). Avaliação em massa de apartamentos com uso de regressão geograficamente ponderada. Revista Valorem, 1(1), 16-31.