

COMPARAÇÃO ENTRE A COLETA MANUAL E POR RASPAGEM DE DADOS EM UMA AVALIAÇÃO IMOBILIÁRIA

Trabalho de Avaliação

LAURO CAVALCANTI DE SÁ
LEYSER PACHECO PIRES FILHO

João Pessoa - PB
2025

REALIZAÇÃO:



PATROCÍNIO:



COMPARAÇÃO ENTRE A COLETA MANUAL E POR RASPAGEM DE DADOS EM UMA AVALIAÇÃO IMOBILIÁRIA

RESUMO

Este trabalho analisou a aplicação da técnica de raspagem de dados (*web scraping*) na avaliação de imóveis, comparando seus resultados com os obtidos por meio do método tradicional baseado no Método Comparativo Direto de Dados do Mercado (MCDDM), conforme a NBR 14.653. O estudo foi realizado com um apartamento localizado no bairro de Botafogo, no Rio de Janeiro, contando com ampla coleta de dados manuais e automatizados, seguidos por rigoroso tratamento estatístico. A avaliação tradicional utilizou 56 dados amostrais da mesma rua do imóvel, com análise de regressão linear e cumprimento dos critérios de fundamentação e precisão exigidos pela norma. Já a avaliação por raspagem partiu de 4.993 anúncios extraídos automaticamente, que, após limpeza e exclusão de *outliers* e resíduos, resultaram em um modelo de regressão múltipla com bom poder explicativo. Ambos os métodos seguiram princípios éticos e legais quanto à manipulação de dados, e permitiram a comparação entre abordagens distintas de estimativa de valor imobiliário.

Palavras-chave: avaliação imobiliária, raspagem de dados, regressão linear múltipla.

REALIZAÇÃO:



PATROCÍNIO:



1. EXPOSIÇÃO

1.1. Introdução

A estimativa do valor de imóveis insere-se no campo da Engenharia de Avaliações, especificamente na subárea de Avaliações Imobiliárias, voltada à determinação do valor de mercado de bens imóveis. No Brasil, o primeiro registro técnico sobre o tema foi a obra de Berrini (1941), marco inicial da sistematização das avaliações. Atualmente, essa atividade é regulamentada pela Norma Brasileira (NBR) 14.653, da Associação Brasileira de Normas Técnicas (ABNT), cujas partes foram publicadas entre 2011 e 2019.

As técnicas e definições utilizadas evoluíram significativamente ao longo do tempo. Atualmente, as avaliações baseiam-se no conceito de Valor de Mercado, definido na ABNT (2019) como: *“a quantia mais provável pela qual se negociaria voluntária e conscientemente um bem, em uma data de referência, sob as condições do mercado vigente.”*

Para estimar esse valor, aplicam-se diferentes metodologias conforme a natureza do bem e a disponibilidade de dados. As mais utilizadas são: Método Comparativo Direto de Dados do Mercado (MCDDM), Método da Quantificação do Custo, Método Evolutivo e Método Involutivo. Conforme o item 6.6 da NBR 14.653-1, o MCDDM deve ser preferencialmente adotado, com base em transações efetivas ou ofertas verificáveis no mercado.

Entretanto, a escassez de informações confiáveis sobre transações efetivas leva os avaliadores, frequentemente, à coleta manual de dados em sites de anúncios, processo caracterizado por baixa eficiência e alto grau de subjetividade (Steiner et al. (2008) e González (2002)).

Nesse contexto, a Raspagem de Dados (Web Scraping) apresenta-se como uma alternativa promissora. Essa técnica utiliza algoritmos para extrair automaticamente informações estruturadas de páginas web. Para evitar sobrecarga nos servidores, recomenda-se a realização de requisições em horários de baixa demanda e o uso de temporizadores (Silva; Longo; Silva (2020)).

A aplicação dessa tecnologia no campo das avaliações imobiliárias teve início recente. Tavares (2021) identificou que os primeiros trabalhos científicos sobre o tema surgiram em 2018, indicando um campo ainda em desenvolvimento. Em estudo de caso, Cruz e Tamietti (2023) concluíram que o Web Scraping é mais eficaz na produção de conhecimento sobre o mercado do que na aplicação direta em modelos de avaliação.

Porém, eles não compararam os resultados com algum método tradicional para embasar tal afirmação. Os autores sugeriram, ainda, o uso de amostras ampliadas em estudos futuros.

A partir desses estudos, percebe-se que apesar do potencial da raspagem de dados para acelerar e ampliar a coleta de informações, sua aplicabilidade direta na avaliação de imóveis específicos é questionada e que, ao mesmo tempo, não existem estudos que comparem resultados obtidos com raspagem de dados com os métodos tradicionais de avaliação.

A dúvida em relação aos valores obtidos utilizando a raspagem de dados ocorre devido à incapacidade dos respectivos modelos em utilizar dados específicos não estruturados e não automatizados nos portais de anúncios como, por exemplo, o índice fiscal, padrão construtivo e estado de conservação. Assim, a complementação manual dessas informações permaneceria necessária para garantir a precisão nas avaliações.

Diante do que foi apresentado, este trabalho teve como objetivo comparar os resultados obtidos utilizando a técnica de raspagem de dados com o valor obtido por uma avaliação imobiliária tradicional em um estudo de caso, de forma a mostrar objetivamente a diferença entre as duas abordagens. Com isso, espera-se embasar uma possível decisão entre utilizar ou não a raspagem de dados em avaliações imobiliárias.

1.2. Imóvel Avaliando

Todas as etapas previstas na NBR 14.653 foram devidamente cumpridas, inclusive aquelas relativas à vistoria do imóvel e à caracterização da região.

O Imóvel Avaliando localiza-se na Rua Voluntários da Pátria, no bairro de Botafogo, Zona Sul do Rio de Janeiro, região amplamente consolidada do ponto de vista urbano e dotada de completa infraestrutura. A localidade conta com rede de transporte público diversificada, incluindo estação de metrô a poucos metros e linhas regulares de ônibus, garantindo fácil acesso ao Centro e a bairros como Copacabana, Flamengo e Humaitá. A região é atendida por ampla gama de equipamentos urbanos, como o Botafogo Praia Shopping, o Hospital Samaritano, escolas de alto padrão, museus, centros culturais e áreas de lazer na orla da Enseada de Botafogo. A presença de comércio variado, serviços essenciais e vias arteriais de grande fluxo tornam o entorno do imóvel altamente valorizado e atrativo para fins residenciais, comerciais e de investimento.

A Figura 1 apresenta registros fotográficos do interior do Imóvel Avaliando, que se trata de um apartamento padrão de 100 m², 3 (três) dormitórios, sem vaga de garagem,

situado em edifício residencial que dispõe de portaria 24 horas, elevadores social e de serviço. Esse apartamento apresenta acabamentos compatíveis com padrão médio e encontra-se em bom estado de conservação, sem sinais visíveis de deterioração ou necessidade de manutenção imediata, aparentando cuidados regulares com limpeza, pintura e conservação geral.



Figura 1. Registros fotográficos do Imóvel Avaliando (autoria própria).

1.3. Considerações Iniciais da Avaliação Imobiliária

Este estudo seguiu integralmente as diretrizes da Lei Geral de Proteção de Dados Pessoais (LGPD). Os dados utilizados foram tornados publicamente acessíveis pelos próprios titulares em plataformas de anúncios, sem exigência de login ou autenticação, e passíveis de indexação por ferramentas de busca. Nessas condições, o consentimento não é exigido, conforme §4º do Art. 7º da LGPD. A finalidade da coleta está amparada no inciso IV do mesmo artigo, voltada exclusivamente à pesquisa. Para assegurar a anonimização, conforme exigido pela legislação, foram adotados os meios técnicos disponíveis, de forma a eliminar a possibilidade de associação direta ou indireta aos indivíduos. Assim, os campos “Endereço”, “Link para o Anúncio”, nomes dos informantes e seus respectivos contatos foram removidos da base de dados analisada.

Com o objetivo de possibilitar as análises da precisão das ferramentas de predição, procedeu-se a avaliação imobiliária do Imóvel Avaliando por meio do MCDDM, com aplicação de tratamento científico (inferência estatística), conforme preconizado no subitem 8.2.1 da Parte 2 da NBR 14.653.

E com a finalidade de conferir maior robustez e fundamentação técnica ao trabalho, adotou-se como premissa a obtenção, no mínimo, do Grau de Fundamentação II, conforme os critérios estabelecidos no subitem 9.2.1 da Parte 2 da NBR 14.653.

1.4. Avaliação Imobiliária utilizando a Coleta Manual

Os dados amostrais foram coletados manualmente no sítio eletrônico *ZapImoveis*, especializado na comercialização de imóveis. Em virtude da ampla oferta de dados na região, todos os 60 dados coletados encontram-se localizados na mesma rua do Imóvel Avaliando. A Tabela 1 apresenta as variáveis consideradas na avaliação, incluindo variáveis textuais utilizadas na caracterização dos elementos. A coincidência do logradouro entre os dados amostrais e o Imóvel Avaliando tornou dispensável o tratamento da variável “Índice Fiscal”, a qual, tradicionalmente, é empregada para tratamento do atributo de localização.

A variável “Dormitórios” não atendeu quanto ao nível máximo de significância exigido pela norma para rejeição da hipótese nula, de modo que sua inclusão no modelo não contribuiu para a explicação das variações de preços. Por esse motivo, foi retirada do modelo final. Assim, foram utilizadas seis variáveis numéricas independentes: “Área

privativa”, “Vaga de garagem”, “Elevador”, “Equipamentos”, “Padrão construtivo” e “Estado de conservação”, sendo considerada como variável dependente o “Valor unitário”.

Foram coletados 60 dados amostrais, com a identificação de todos os seus atributos, em conformidade com as variáveis descritas na Tabela 1. Os dados foram processados e analisados utilizando-se o software de Engenharia de Avaliações SisDEA, desenvolvido pela Pelli Sistemas Engenharia.

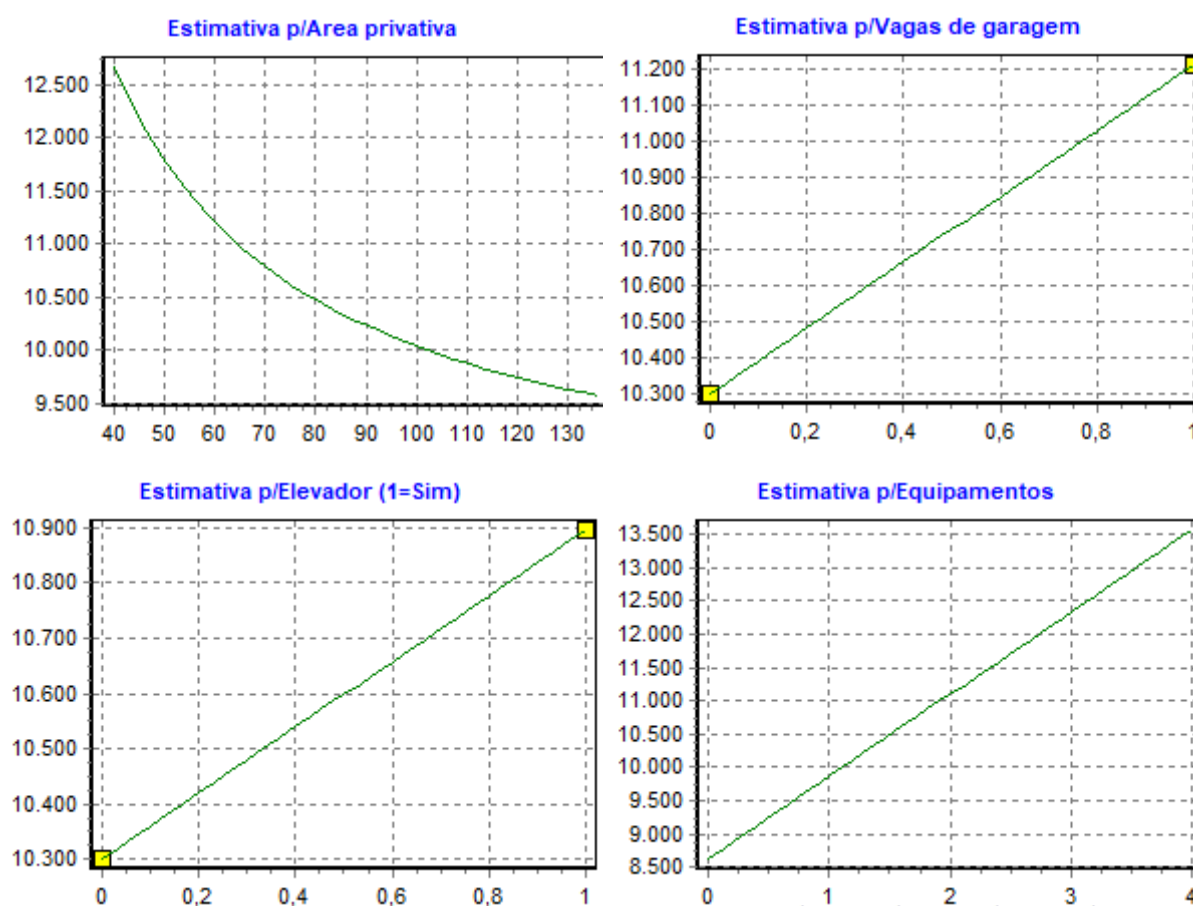
Tabela 1. Variáveis consideradas na avaliação tradicional.

Nome	Tipo	Classificação	Descrição da varável	Habilitada
Endereço	Texto	Texto	Endereço completo do imóvel	sim
Bairro	Texto	Texto	Bairro onde o imóvel se localiza	sim
Informante	Texto	Texto	Nome ou identificação do informante	sim
Contato	Texto	Texto	Telefone de contato do informante	sim
Área privativa	Numérica	Quantitativa	Área privativa da unidade (m ²)	sim
Dormitórios	Numérica	Quantitativa	Quantidade de dormitórios da unidade	não
Vaga de garagem	Numérica	Dicotômica	Indica a presença ou ausência de vaga de garagem (0-não, 1-sim)	sim
Elevador	Numérica	Dicotômica	Indica a presença ou ausência de elevadores no prédio (0-não, 1-sim)	sim
Equipamentos	Numérica	Quantitativa	Número de equipamentos disponíveis no condomínio (Portaria 24hrs, academia, piscina, área de lazer/playground)	sim
Padrão construtivo	Numérica	Qualitativa (Códigos Alocados)	Padrão construtivo do imóvel (1-Baixo, 2-Normal, 3-Alto)	sim
Estado de conservação	Numérica	Qualitativa (Códigos Alocados)	Estado de conservação do imóvel (1-Reparos Simples, 2-Bom, 3-Novo)	sim
Índice Fiscal	Numérica	Proxy	Variável de macrolocalização construída com base nos índices fiscais da PGV do município	não
Valor total	Numérica	Dependente	Valor total do imóvel (R\$)	não
Valor unitário	Numérica	Dependente	Valor total do imóvel dividido pela área total (R\$/m ²)	sim

Foram excluídos 4 (quatro) dados do modelo final com o objetivo de viabilizar sua aceitação, conforme os parâmetros exigidos pela norma para o atingimento do Grau de Fundamentação II. Dessa forma, foram efetivamente utilizados 56 dados, quantidade que

atende ao número mínimo de dados de mercado exigido para o grau de fundamentação pretendido. As variáveis qualitativas codificadas (“Padrão construtivo” e “Estado de conservação”) atenderam ao critério de micronumerosidade, uma vez que há, no mínimo, três elementos amostrais associados a cada código alocado.

A Figura 2 apresenta os relacionamentos gráficos obtidos para o valor unitário em função de cada variável considerada. Observa-se que, conforme esperado, todas as variáveis apresentam relação positiva e crescente com o valor unitário, exceto a variável “Área privativa”, para a qual maiores valores correspondem a valores unitários menos elevados.



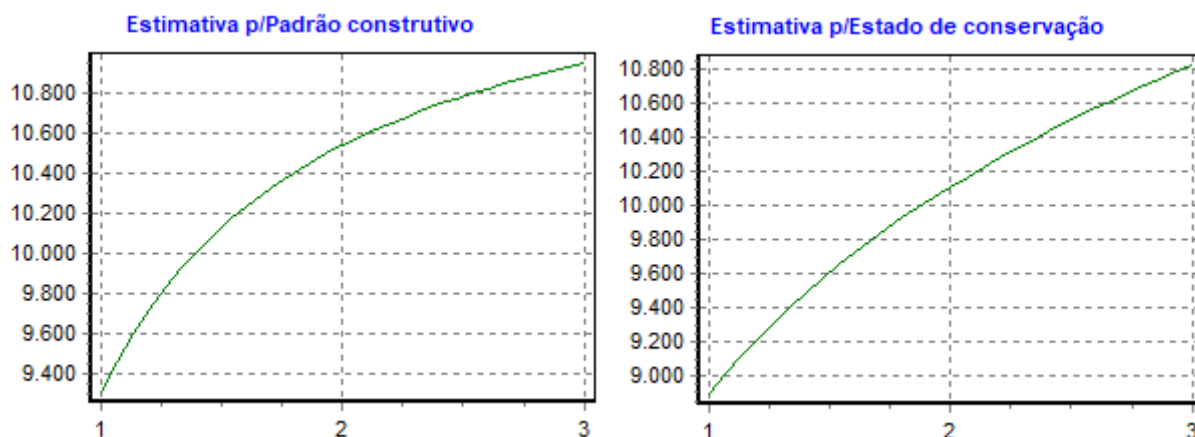


Figura 2. Relacionamento das variáveis numéricas.

Quanto às estatísticas da avaliação, o coeficiente de correlação (R) do modelo foi de 0,85, indicando forte correlação entre as variáveis independentes e a variável dependente. O coeficiente de determinação (R^2) foi de 72%, o que significa que 72% da variabilidade nos valores unitários é explicada pelas variáveis incluídas no modelo. A Tabela 2 apresenta a análise de variância (ANOVA) do modelo. A estatística F , calculada como o quociente entre as variâncias explicada e não explicada, foi de 21,33, o que corresponde a um nível de significância inferior a 1% ($p < 0,01$), indicando que o modelo é estatisticamente significativo.

Tabela 2. Análise de Variância (Estatística F).

	Graus de liberdade	Soma dos quadrados	Quadrado médio	F	Nível de significância
Regressão	6	258.268.913	43.044.819	21,33	< 0,01
Residual	49	98.888.615	2.018.135		
Total	55	357.157.528			

A Tabela 3 apresenta os resultados de significância dos regressores, nos quais se observa que a maior estatística t foi obtida para a variável “Equipamentos” ($t = 7,90$), resultando em um nível de significância de 0,01. Esse resultado indica que se trata da variável mais relevante do modelo, por ser a que mais contribui para explicar a variação dos preços, conforme evidenciado pelo seu baixo nível de significância. Como todos os níveis de significância são menores ou iguais a 20%, o modelo atende a estatística t quanto ao mínimo requerido para o Grau de Fundamentação II.

Tabela 3. Significância dos regressores.

	Coeficientes	Estatística t	Nível de significância (%)
Área privativa	175.563,28	2,58	1,29
Vaga de garagem	911,33	2,05	4,62
Elevador	594,91	1,32	19,19
Equipamentos	1.233,55	7,90	0,01
Padrão construtivo	-2.464,06	-3,05	0,37
Estado de conservação	1.751,65	2,97	0,46
Valor unitário	6.664,58	5,09	0,01

As Figuras 3 e 4 apresentam, respectivamente, o gráfico de resíduos e o gráfico de aderência. Conforme observado, nenhum ponto ultrapassou os limites de ± 2 desvios padrão (indicados pelas retas horizontais vermelhas na Figura 3), não havendo, portanto, a presença de *ouliers* no modelo. A normalidade dos resíduos amostrais foi verificada pelas frequências relativas dos resíduos padronizados nos intervalos de $[-1,+1]$, $[-1,64;+1,64]$ e $[-1,96;+1,96]$ do desvio-padrão, que apresentam, respectivamente, probabilidades de 75%, 87% e 100%. Cabe destacar, ainda, que as variáveis independentes não apresentaram correlações parciais entre si superiores ao limite estabelecido pela norma, de 0,80.

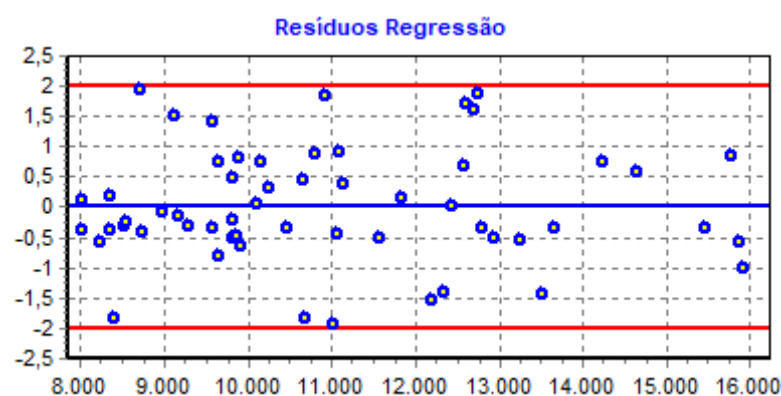


Figura 3. Distribuição dos resíduos.

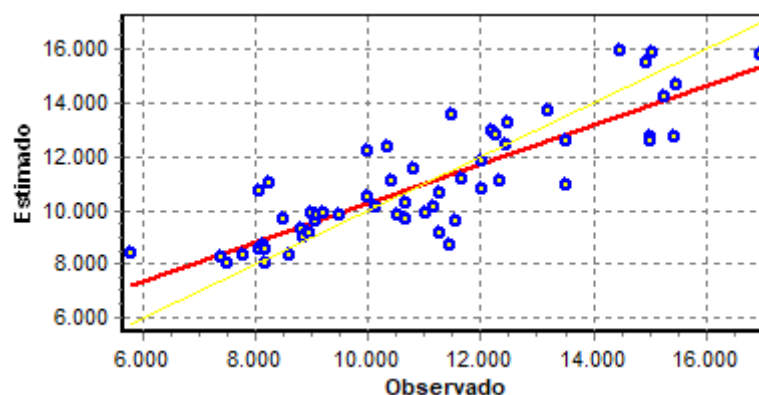


Figura 4. Aderência.

Por fim, tem-se a equação da função estimativa para o modelo final:

$$\begin{aligned} \text{Valor unitário} = & 6664,580441 + 175563,282 / \text{Área privativa} \\ & + 911,3309479 * \text{Vaga de garagem} + 594,9066115 * \text{Elevador} \\ & + 1233,546663 * \text{Equipamentos} - 2464,062833 / \text{Padrão construtivo} \\ & + 1751,654957 * \ln (\text{Estado de conservação}) \end{aligned}$$

A Tabela 4 apresenta a projeção de valores para o bem avaliando, considerando seus atributos e o intervalo de confiança do modelo. O apartamento objeto do estudo de caso foi avaliado pelo valor médio de **R\$ 10.941,02/m²**, totalizando R\$ 1.094.102,48. Ressalta-se que todos os requisitos para o atingimento do Grau de Fundamentação II foram devidamente atendidos e que o intervalo de confiança obtido foi de 10,31%, o que corresponde ao Grau de Precisão III, conforme os critérios estabelecidos na NBR 14653-2.

Tabela 4. Projeção de valores para o avaliando.

Área (m²)	Dorm.	Vaga de garagem	Elevador	Equip.	Padrão construtivo	Estado de Conservação	Valor Unitário Mínimo (R\$/m²)	Valor Unitário Médio (R\$/m²)	Valor Unitário Máximo (R\$/m²)
100	3	0	1	1	2	3	10.377,00	10.941,02	11505,05

1.5. Estimativa de Valor utilizando a Raspagem de Dados

Para a implementação da raspagem de dados, definiu-se o escopo da coleta com base em três critérios: localidade, tipo de imóvel e portal de anúncios. A seleção do portal considerou a quantidade de ofertas disponíveis, a padronização das informações e a estrutura do site compatível com a automação da coleta. Portais com restrições expressas

em seus Termos de Uso foram desconsiderados. Optou-se por imóveis do tipo apartamento, localizados no bairro de Botafogo, Rio de Janeiro – RJ, conforme o Imóvel Avaliando. O portal *Imove/web* foi selecionado por apresentar 5.861 anúncios compatíveis com os critérios estabelecidos e com dados estruturados adequadamente.

Utilizou-se o sítio da internet *Webscraper.io* para configurar a extração automatizada de dados do portal imobiliário, abrangendo os campos: endereço, área, número de quartos, banheiros, vagas, preço de compra/venda, link do anúncio e tempo de publicação.

Foram coletados 4.993 registros, gerando um arquivo em formato .csv, com campos separados por vírgula, onde cada linha representa um imóvel e cada coluna, uma variável. Os dados foram extraídos em ordem decrescente de publicação, permitindo a verificação da contemporaneidade (até dois anos), conforme o Art. 3º, inciso XIII, da SPU (2025), sendo selecionadas 153 publicações dos últimos 30 dias na mesma rua do Imóvel Avaliando. Ressalta-se que a quantidade efetivamente coletada pode divergir do número inicialmente informado pelo portal, o que constitui uma limitação inerente ao portal de anúncios.

Utilizou-se o software LibreOffice 7 para o cálculo do preço unitário e, nos casos em que não havia informação sobre vagas de garagem, foi atribuído o valor zero ao campo correspondente.

Em um primeiro momento, os dados imobiliários obtidos foram sanitizados utilizando a biblioteca *Pandas* do programa *Python*, que:

- importou e definiu corretamente os tipos dos dados, possibilitando que os valores não numéricos fossem convertidos para “N/A” (valores ausentes) nos campos “Área”, “Quantidade de Quartos”, “Quantidade de Banheiros”, “Quantidade de Vagas”, ou “Preço de Compra/Venda”;
- excluiu os dados com valores “N/A”, já que não seria possível aproveitá-los no modelo estatístico; e
- verificou e confirmou que não havia dados repetidos.

Após a sanitização dos campos “Área”, “Quartos”, “Banheiros”, “Vagas” e “Preço”, com a remoção de valores não numéricos e ausentes, a base foi reduzida para 142 registros. Essa perda de dados decorre da incompletude dos anúncios, evidenciando uma limitação do uso de portais como fonte de dados. Em seguida, foram eliminados 19 registros duplicados, resultando em uma amostra final de 123 imóveis. Ressalta-se que este trabalho não abordou a possibilidade de duplicidade de imóveis anunciados por diferentes usuários.

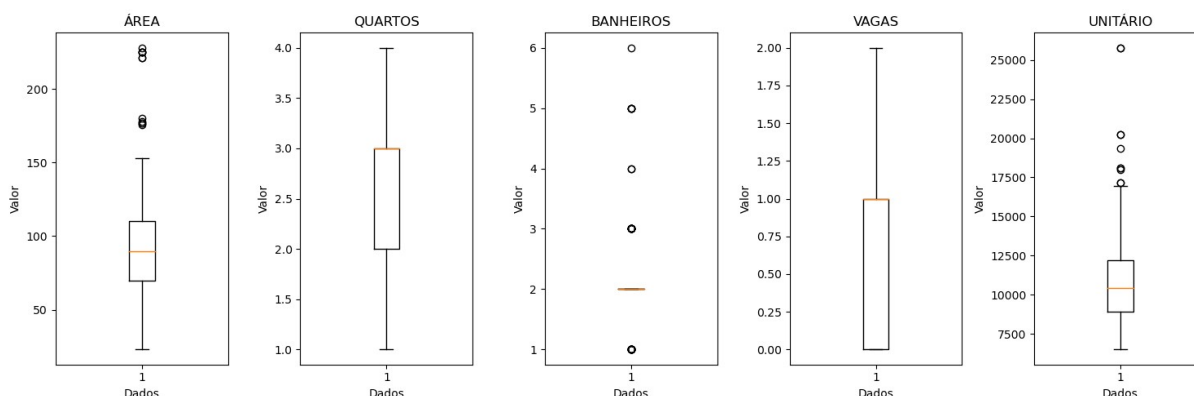


Figura 5. Distribuição dos dados imobiliários iniciais por variáveis.

Dessa maneira, obteve-se a distribuição dos 123 dados mostrados na Figura 5 das quatro variáveis numéricas independentes (“Área”, “Quartos”, “Banheiros” e “Vagas”) e da variável numérica dependente “Preço Unitário”, resultando no resumo estatístico da Tabela 5.

Tabela 5. Resumo dos dados imobiliários iniciais.

	Área	Quartos	Banheiros	Vagas	Unitário
Média	99,03	2,50	2,09	0,52	11.221,57
Desvio Padrão	42,66	0,66	0,89	0,53	3.490,19
Mínimo	23	1	1	0	6.500,00
25%	70	2	2	0	8.929,74
50%	90	3	2	1	10.441,18
75%	110	3	2	1	12.218,56
Máximo	228	4	6	2	25.775,86

De acordo com o painel de distribuições entre variáveis mostrado na Figura 6, pôde-se constatar os seguintes comportamentos entre elas, os quais eram esperados em virtude da natureza das mesmas perante aos imóveis:

- “Área” varia positivamente com “Banheiros”, “Quartos” e “Vagas”;
- “Quartos”, “Banheiros” e “Vagas” variam positivamente entre si;
- “Valor Unitário” varia positivamente com “Vagas”, “Área” e “Banheiros”; e
- “Valor Unitário” varia inversamente com a área.

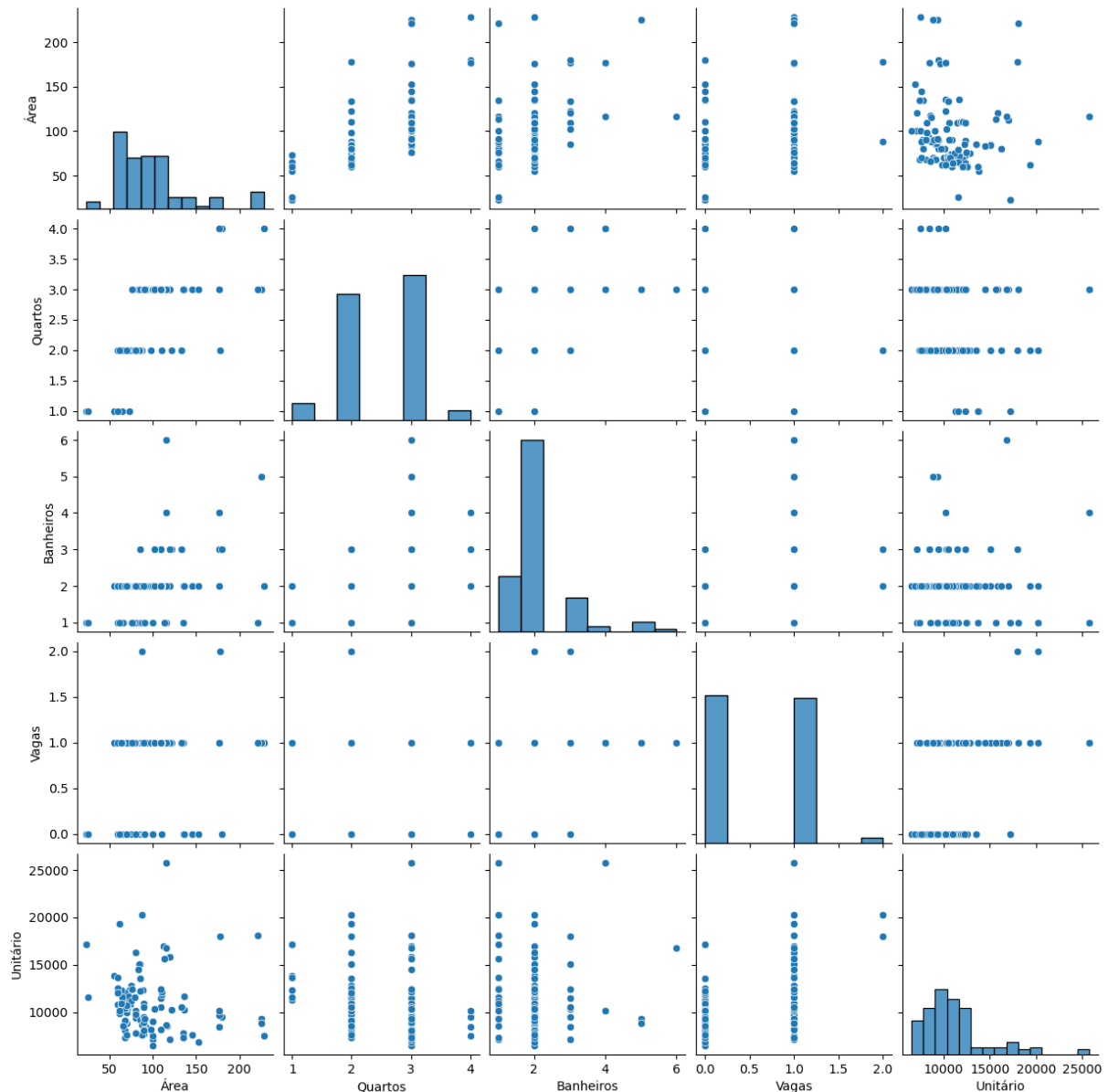


Figura 6. Painel de distribuição dos dados iniciais entre variáveis.

Após o tratamento inicial dos dados, foram realizadas análises estatísticas dos dados por regressão linear múltipla utilizando as bibliotecas *itertools*, *matplotlib*, *pickle*, *scipy*, *seaborn*, *statsmodels*, *math*, *numpy* e *pandas* no programa *Python*. Para a rejeição da hipótese nula e para explicar o comportamento dos preços unitários em função das demais variáveis, foi considerado um nível de significância (*p-valor*) igual a 0,05 para o teste “t” dos parâmetros e “F” para o modelo.

Dessa forma, após a visualização dos dados iniciais, os mesmos foram testados em um modelo de regressão linear múltipla sem transformações, que apresentou um

Coefficiente de Determinação igual a 0,256 com valores de significância todos menores do que 5% (valor $p < 0,05$), indicando a representatividade das variáveis “Área”, “Quartos” e “Vagas” para explicar o “Preço Unitário”, porém, com valor maior para a variável “Banheiros”. Por esse motivo, a mesma foi excluída do modelo e os passos foram refeitos.

Com a finalidade de confirmar a representatividade das variáveis independentes, foi executado uma análise *Step Wise*, que atestou o anteriormente descrito, alterando as características do modelo inicial.

Sendo assim, cada uma das variáveis numéricas independentes “Área”, “Quartos”, “Banheiros” e “Vagas”, assim como a variável dependente, foram testadas com as transformações “1/x”, “ln(x)”, “x²”, “√x”, “1/x²” e “√1/x”, além das combinações sem transformação, totalizando 2.401 análises.

Após isso, a melhor combinação de transformações que resultou em um Coeficiente de Determinação igual a 0,313 foi a seguinte equação:

$$1/\text{UNITARIO} \sim \text{const.} + a \cdot 1/\text{AREA}^2 + b \cdot \text{QUARTOS}^2 + d \cdot \text{VAGAS}$$

Com a finalidade de aumentar o grau de explicação do modelo, optou-se pela análise dos resíduos e exclusão de 60 dados com os resíduos padronizados acima de dois em módulo. Após cada exclusão, foi realizada uma nova análise do modelo, resultando em um Coeficiente de Determinação igual a 0,785, resultando na seguinte equação e na Tabela 6:

$$1/\text{UNITARIO} = 0,000084 - 0,016135/\text{AREA}^2 + 0,000003 \cdot \text{QUARTOS}^2 - 0,000007 \cdot \text{VAGAS}$$

Dessa forma, optou-se pela equação encontrada como a que melhor explicaria o comportamento da amostra imobiliária, podendo estimar o valor de outros imóveis.

Utilizando-se a ferramenta de predição para o apartamento objeto do estudo de caso, com 100 m², 3 quartos, 2 banheiros e sem vaga de garagem, estimou-se um valor unitário médio igual a **R\$ 9.015,44/m²**.

Tabela 6. Resultados do modelo de regressão múltipla linear final.

Dep. Variable:	Unitário	R-squared:	0.785			
Model:	OLS	Adj. R-squared:	0.775			
Method:	Least Squares	F-statistic:	80.19			
Date:	Mon, 07 Jul 2025	Prob (F-statistic):	5.70e-22			
Time:	21:47:41	Log-Likelihood:	733.12			
No. Observations:	70	AIC:	-1458.			
Df Residuals:	66	BIC:	-1449.			
Df Model:	3					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	8.399e-05	2.28e-06	36.842	0.000	7.94e-05	8.85e-05
Área	-0.0161	0.003	-5.089	0.000	-0.022	-0.010
Quartos	3.172e-06	2.79e-07	11.350	0.000	2.61e-06	3.73e-06
Vagas	-6.753e-06	1.8e-06	-3.760	0.000	-1.03e-05	-3.17e-06
=====						
Omnibus:	8.831	Durbin-Watson:	2.063			
Prob(Omnibus):	0.012	Jarque-Bera (JB):	2.964			
Skew:	-0.022	Prob(JB):	0.227			
Kurtosis:	1.993	Cond. No.	2.66e+04			
=====						

De forma a validar o modelo, também foram verificados: a distribuição dos dados, os resíduos padronizados, a normalidade dos resíduos, os valores da variável dependente, a heterocedasticidade, a autocorrelação e a multicolinearidade. Todos os resíduos padronizados resultantes se encontravam dentro do intervalo de -2 a +2 desvios padrão, conforme mostrado na Figura 7. Os resíduos apresentaram uma distribuição igual a 58%, 88% e 98% nos intervalos de ± 1 , $\pm 1,64$ e $\pm 1,96$ desvios padrão. Os valores da variável dependente se encontravam bem distribuídos em relação aos valores preditos, conforme mostra a Figura 8. O teste de Breusch-Pagan não confirmou a heterocedasticidade, apresentando um p-value igual a 0,346, validando o modelo proposto. O teste de Durbin-Watson apresentou um valor igual a 2,06, afastando a hipótese de autocorrelação entre as variáveis independentes, corroborando o modelo proposto. E, por fim, a hipótese de multicolinearidade entre variáveis foi afastada com a verificação dos valores do Fator de Inflação da Variância (em inglês, *Variance Inflation Factor – VIF*) menores que 5, conforme a Tabela 7.

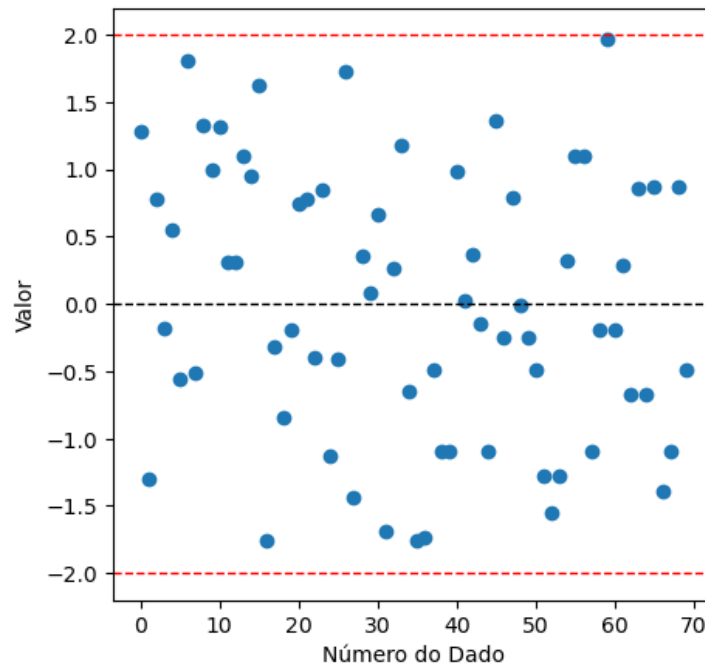


Figura 7. Resíduos padronizados.

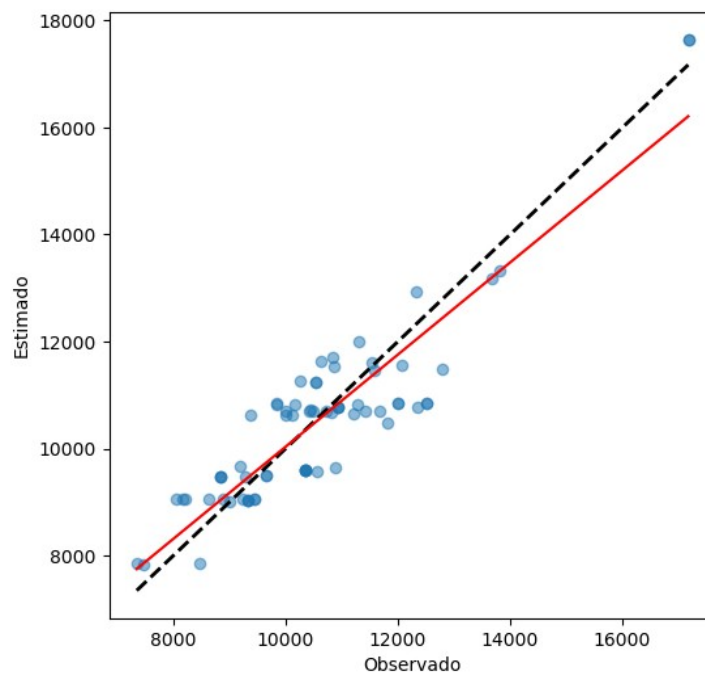


Figura 8. Valores da variável dependente x Valores preditos.

Tabela 7. Fator de Inflação da Variância.

	Área
const	-----
Área	1,39
Quartos	1,43
Vagas	1,43
Unitário	1,46

1.6. Comparação entre as Metodologias e Valores

De acordo com as análises realizadas, constatou-se que o valor obtido utilizando a raspagem de dados (R\$ 9.015,44 m²) foi **17,6%** inferior ao obtido pela coleta manual (R\$ 10.941,02 m²). Essa diferença pode ser atribuída a diversos fatores.

O principal deles refere-se à capacidade da coleta manual de incorporar variáveis adicionais, como o estado de conservação e o padrão construtivo dos imóveis. A coleta por raspagem, por sua vez, não permite a inclusão dessas variáveis, uma vez que tais informações não se apresentam de forma padronizada nos sítios de anúncios consultados.

Além desse fator, destaca-se a capacidade da raspagem de dados de abranger uma faixa temporal mais ampla de anúncios, permitindo a coleta de ofertas desde as mais antigas até as mais recentes. No entanto, essa característica pode representar uma limitação, caso sejam incluídos anúncios antigos com preços desatualizados. Por esse motivo, somente foram selecionados anúncios com até 30 dias de postagem.

Neste estudo buscou-se a semelhança no critério de amostragem. Tanto na coleta manual quanto na raspagem, foram selecionados somente os anúncios de imóveis localizados na mesma rua do Imóvel Avaliando. Dessa forma, não foi necessário considerar outro fator de localização em ambos os métodos de coleta.

Observou-se também que a duplicidade e a incompletude dos anúncios dificultam a extração eficiente de dados por meio da raspagem, além da possibilidade de um mesmo imóvel ser anunciado por diferentes corretores ou plataformas, situação que não foi investigada no presente estudo.

Outra dificuldade associada à raspagem de dados decorre da própria área de Ciência de Dados: o tratamento de grandes volumes amostrais, tanto na etapa de sanitização quanto no processamento das informações. Como exemplo, foi necessário desenvolver um método automatizado para a remoção de *outliers*, uma vez que a execução manual dessa tarefa tornaria o processo excessivamente demorado e pouco viável.

Foram estudados os valores por m² por bairros do índice *FipeZap+*, da Fundação Instituto de Pesquisas Econômicas (FIPE), que apresentou o preço unitário igual a R\$ 12.888,00/m² para um imóvel residencial independente da quantidade de dormitórios no mesmo mês. Destaca-se que esses índices não consideram a variação das variáveis como, por exemplo, a quantidade de quartos e vagas, inviabilizando a comparação com os resultados obtidos.

Diante dos resultados obtidos nas duas análises e das considerações previamente apresentadas, conclui-se que a aplicação da raspagem de dados em avaliações imobiliárias ainda demanda aprofundamento em termos de estudos, sistematização e metodologia, de modo a validá-la e assegurar que os resultados produzidos sejam compatíveis com aqueles obtidos por meio da coleta manual de dados.

Nesse sentido, recomenda-se a realização de novos estudos comparativos que empreguem a raspagem de dados, a fim de contribuir para o desenvolvimento e consolidação dessa abordagem no campo das avaliações.

2. CONCLUSÃO

O estudo teve como objetivo comparar os valores estimados para um imóvel por meio de duas abordagens distintas: a avaliação tradicional conforme a NBR 14.653, baseada em coleta manual de dados e tratamento estatístico rigoroso, e a estimativa via raspagem de dados (*web scraping*), com utilização de modelos preditivos automatizados. A análise revelou que, embora a raspagem de dados permita ampla coleta e agilidade na obtenção de informações, sua aplicação direta na avaliação de um imóvel específico ainda enfrenta limitações importantes.

Enquanto a avaliação tradicional alcançou um valor estimado de R\$ 10.941,02/m², com Grau de Fundamentação II e Precisão III, a avaliação com a raspagem de dados, mesmo após refinamentos estatísticos e exclusão de *outliers*, resultou em um valor inferior, de R\$ 9.015,44/m², com poder explicativo similar ($R^2 = 0,785$). A diferença significativa entre os resultados evidencia que, apesar do potencial da raspagem de dados como ferramenta complementar para análise de mercado, ela ainda não substitui a precisão e a contextualização técnica proporcionadas pela avaliação convencional, especialmente quando se trata de imóveis específicos com características singulares não capturadas automaticamente (como padrão construtivo, conservação e índice fiscal).

Conclui-se, portanto, que a raspagem de dados pode ser útil para estudos exploratórios e levantamento de tendências, mas não deve ser utilizada isoladamente como substituto das avaliações técnicas normatizadas, sem as devidas complementações manuais e criteriosas exigidas pela prática profissional da Engenharia de Avaliações.

Referências

ABNT. 2019. Avaliação de bens. Parte 1: Procedimentos gerais. Rio de Janeiro, RJ, Brasil. ISBN 978-85-07-08096-1.

ABNT. 2011. Avaliação de bens. Parte 2: Imóveis Urbanos. Rio de Janeiro, RJ, Brasil. ISBN 978-85-07-02596-2.

Berrini, L.C. 1941. Avaliações de Terrenos. 1ed. São Paulo, SP, Brasil.

Cruz, B. H. S. G.; Tamietti, L. F. de M.. 2023. Raspagem de Dados para Avaliação Imobiliária um Estudo de Caso para Mapeamento de Riscos e Oportunidades. (): .

González (2002): González, M. A. S, Aplicação de Técnicas de Descobrimento de Conhecimento em Bases de Dados e de Inteligência Artificial em Avaliação de Imóveis, Porto Alegre, RS, Brasil.

Silva, C. B.; Longo, E. S.; Silva, E. da. 2020. Coleta de Dados de Imóveis de Forma Automatizada para Fins de Políticas Públicas. Anais do 14º Congresso de Cadastro Multifinalitário e Gestão Territorial p. 1-8.

SPU. 06/03/2025. Instrução Normativa SPU/MGI nº 98. (): .

Steiner, M. T. A. et al. 2008. Métodos estatísticos multivariados aplicados à engenharia de avaliações. Revista Gestão e Produção 1 (15): 23-32.

Tavares, L. da S. 2021 . Framework Automatizado para Avaliação de Imóveis Urbanos utilizando Técnicas de Mineração de Dados. Dissertação. Instituto Federal de Educação, Ciência e Tecnologia Fluminense, Campos dos Goytacazes, RJ, Brasil.