

ESTIMATIVA DE VALORES DE IMÓVEIS DA UNIÃO COM TÉCNICAS DE APRENDIZAGEM DE MÁQUINA E COMPONENTES ESPACIAI

AUTORES

José Nilo Alves de Sousa Neto

Juliana Miura

REALIZAÇÃO

ESTIMATIVA DE VALORES DE IMÓVEIS DA UNIÃO COM TÉCNICAS DE APRENDIZAGEM DE MÁQUINA E COMPONENTES ESPACIAIS

RESUMO

A valoração do patrimônio imobilizado de uma instituição representa uma condição necessária a uma gestão eficiente de seus ativos. A execução e a análise dos laudos de avaliação imobiliária são essenciais à consecução de alguns objetivos estratégicos da Força Terrestre do Brasil, mas são, também, bastante onerosas em termos de tempo, mão de obra e recursos financeiros. Ocorre, por vezes, emprego de grande esforço para que as etapas supramencionadas ocorram e o valor de mercado finalmente obtido ser incoerente com o que as autoridades solicitantes imaginavam inicialmente, fazendo com que o estudo técnico realizado não seja efetivamente utilizado em negociações pela organização. Nesse sentido, este trabalho propõe o desenvolvimento de modelos preditivos capazes de construir estimativas de valor de imóveis urbanos. Os modelos têm razoável nível de assertividade e abrangência geográfica nacional ao gerar valores estimativos de mercado dos imóveis da União. Contemplaram-se variáveis intrínsecas e extrínsecas aos imóveis, incluindo testes de agregação de componentes espaciais sobre algumas delas. Como a interpretabilidade da solução proposta é um requisito importante, tanto nas abordagens lineares quanto nas não lineares, com aprendizagem de máquina, adotou-se o valor de Shapley como ferramenta de apoio à garantia de explicabilidade. Modelos de equações estruturais com mínimos quadrados parciais (PLS-SEM) foram construídos para selecionar atributos de maneira melhor fundamentada e visualmente acessível. Essas duas considerações, associadas à precificação estimativa de propriedades a nível nacional, representam uma inovação deste trabalho em relação à literatura científica analisada.

Palavras-chave: aprendizagem de máquina, avaliação, imóveis, componentes espaciais

ABSTRACT

The valuation of an institution's patrimony represents a necessary condition for an efficient management of its assets. The execution and analysis of real estate appraisal reports are essential to the achievement of some strategic objectives of the Brazilian Army, but they are also quite costly in terms of time, labor and financial resources. Sometimes, great effort is required for the aforementioned steps to take place and the market value finally obtained is inconsistent with what was initially imagined by the authorities, causing the technical study carried out to not be effectively used in negotiations by the organization. In this sense, this work proposes the development of predictive models capable of building estimates of urban real estate values. The models have a reasonable level of assertiveness and national geographic coverage when generating estimated market values of Union real estate assets. Intrinsic and extrinsic variables to the properties were considered, including tests of aggregation of spatial components on some of them. As the interpretability of the proposed solution is an important requirement, in both linear and nonlinear approaches, with machine learning algorithms, the Shapley value was adopted as a tool to support the guarantee of explainability. Partial least squares structural equation modeling (PLS-SEM) method was applied in order to select features in a reasoned and visually accessible manner. These two considerations associated with real estate price modeling at a national level represent an innovation of this work in relation to the analyzed scientific literature.

Keywords: machine learning, valuation, real estate, spatial components

1. INTRODUÇÃO

Os bens tangíveis do tipo imobiliário possuem características intrínsecas e extrínsecas que os tornam únicos. Tal unicidade faz com que o processo de mensuração de seus valores mais prováveis e justos de mercado constitua matéria efetivamente complexa.

O Exército Brasileiro (EB) possui mais de vinte mil parcelas imobiliárias sob sua administração e necessita de ferramenta para estimativa, em massa, dos valores desses bens patrimoniais a fim de suportar processos decisórios a nível estratégico visando à arrecadação de recursos para o Tesouro Nacional ou à permuta por outros ativos que melhor atendam suas atuais demandas.

No âmbito do EB, a Diretoria de Patrimônio Imobiliário e Meio Ambiente (DPIMA) representa a instância técnica responsável por analisar e aprovar as avaliações imobiliárias dos imóveis administrados pelo EB e daqueles de interesse da Força Terrestre. Já a nível da União, a Secretaria do Patrimônio da União (SPU) faz a gestão dos ativos imobiliários.

Conforme a norma ABNT NBR 14653, o Método Comparativo Direto de Dados de Mercado deve ser preferencialmente utilizado; ele propõe a inferência dos valores de mercado dos imóveis com base na oferta ou na transação de outros imóveis com determinado nível de similaridade. A valoração de imóveis, ainda que de maneira estimativa, não é uma tarefa trivial. Trabalhos realizados mostram que a localização costuma se comportar como uma variável significativa nos modelos construídos e que os imóveis ofertados ou recentemente transacionados nas adjacências daquele em análise têm influência sobre seu valor.

2. FUNDAMENTAÇÃO TEÓRICA

Segundo [Alves Dantas et al. 2010] e [Barros Antunes Campos and Almeida 2018], considerando modelo tradicional de preços hedônicos definido inicialmente por [Rosen 1974], o valor de mercado das unidades imobiliárias residenciais pode ser explicado por:

$$P = f(E, L, T, \beta) + \varepsilon \quad (\text{eq. 1})$$

sendo o preço da habitação (P) função das suas características estruturais (E), locacionais (L) e da época em que foi demandado (T); f é um operador indicativo da forma funcional, β são parâmetros e ε os erros aleatórios do modelo.

Sob essa ótica, imóveis com outras vocações além da residencial também poderiam ter seus valores prováveis de mercado construídos a partir de características intrínsecas e extrínsecas.

3. TRABALHOS RELACIONADOS

Há diferentes abordagens com foco na valoração de imóveis presentes na literatura. [Alves Dantas et al. 2010] implementam modelos de abrangência geográfica municipal combinados a econometria espacial. A linearidade é considerada como um pressuposto e diversas análises são realizadas com e sem componentes espaciais. Os modelos desenvolvidos chegam a coeficientes de determinação (R^2) da ordem de 90%.

[Park e Bae 2015] aplicam e analisam algumas técnicas de aprendizagem de máquina, inclusive não lineares, no mercado de imóveis residenciais em Fairfax County, Virginia, nos Estados Unidos da América. Algoritmos *C4.5*, *RIPPER*, *Naïve Bayes* e *AdaBoost* são comparados e o *RIPPER* apresenta a melhor performance geral entre eles. São limitações do trabalho a abrangência geográfica restrita e a carência de interpretabilidade das variáveis independentes frente à variável explicada.

[Kiely and Bastian 2020] desenvolveram modelos regressivos de predição de valores de imóveis de diferentes tipologias em Nova York com uso de redes neurais artificiais (ANN) e conceitos de agregação espacial de variáveis. Os R^2 mais elevados obtidos são da ordem de 50%. [Dewan et al. 2019] combinaram ANN a modelos autorregressivos espaciais de caráter global (SAR) e os resultados obtidos indicaram a importância dos componentes espaciais.

[Hagenauer and Helbich 2022] propuseram um modelo de ANN combinado a regressões espaciais locais ponderadas, denominado *Geographically Weighted Artificial Neural Network* (GWANN). Quando aplicado ao domínio real de casas unifamiliares na Áustria, os maiores R^2 médios foram de aproximadamente 45%.

[Tchunte and Nyawa 2022] compararam o desempenho de 7 algoritmos de aprendizagem de máquina, lineares e não lineares, aplicados com e sem geocodificação a uma base de dados disponibilizada pelo Governo da França de valores de imóveis. Os modelos não lineares com geocodificação apresentaram maiores coeficientes de determinação, aproximadamente 74%.

A interpretabilidade utilizando o valor de Shapley pode ser garantida por meio de definições colaborativas da Teoria dos Jogos.

Conforme [Lundberg et al. 2020] e [Chakraborty et al. 2020], o valor de Shapley é a contribuição marginal média de cada valor de atributo em todas as combinações possíveis de características. Os recursos com grandes valores de Shapley absolutos são considerados importantes. Para obter a importância relativa global da variável, calcula-se a média de seus valores absolutos de Shapley em todas as ocorrências; normalmente, as medidas de tendência central são ordenadas em importância decrescente e plotadas.

Fundamentado em pesquisa bibliográfica realizada, a Tabela 1 apresenta um resumo comparativo entre a solução proposta e os trabalhos da literatura.

Tabela 1 – Comparação entre trabalhos relacionados.

Trabalho	$R^2 > 57\%$	Abrangência nível País	Modelagem linear	Modelagem não linear	Ajuste de hiperparâmetros	Componentes espaciais	Mais de uma vocação de imóvel	Interpretabilidade
[Alves Dantas et al. 2010]	✓		✓			✓		✓
[Park and Bae 2015]	✓		✓	✓	✓			
[Kiely and Bastian 2020]			✓	✓	✓	✓	✓	
[Hagenauer and Helbich 2022]		✓	✓	✓		✓		
[Tchuente and Nyawa 2022]	✓	✓	✓	✓	✓	✓		
Solução proposta	✓	✓	✓	✓	✓	✓	✓	✓

Fonte: Elaborado pelos Autores.

4. METODOLOGIA

Nesta seção, discorre-se acerca da extração de dados, de sua análise exploratória multivariada e da concepção inicial dos modelos associados à solução proposta.

4.1. Extração de Dados

Os dados diretamente associados a cada um dos imóveis foram extraídos de bases de dados de acesso restrito do EB, com 258 imóveis anotados, e da SPU, com 4981 imóveis anotados. As ocorrências correspondem a imóveis avaliados e com laudos devidamente homologados pelas referidas instituições.

Todos os imóveis do EB foram georreferenciados em formato *shapefile*. Como as instâncias da SPU não continham polígonos georreferenciados, optou-se por geocodificar o campo textual *Endereço*. No processo de geocodificação, 925 instâncias, 17,6% do total coletado, não puderam ter suas coordenadas espaciais extraídas no *Google Earth Pro*. No intuito de amenizar os efeitos de tal perda, 284 ocorrências foram recuperadas manualmente, resultando nos quantitativos resumidos na Tabela 2.

Essas bases foram enriquecidas com atributos socioeconômicos do Censo 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE) tratados por área de ponderação (AP). Para associar os imóveis às AP, realizaram-se operações de união espacial no *QGIS 3.16 Hannover* entre os pontos relativos aos centroides das

propriedades e os polígonos das AP, extraídos de [Furtado 2020]. As 4598 instâncias resultantes e as AP encontram-se representadas na Figura 2. Nos casos em que o centroide do imóvel ficou em área descoberta pelas AP, utilizou-se a média dos atributos das AP adjacentes.

Tabela 2 – Informações sobre as instâncias urbanas com valor conhecido coletadas após processo de recuperação manual.

Fonte	Qtde de instâncias urbanas com valor conhecido	Participação relativa
EB	258	5,6%
SPU	4340	94,4%
EB e SPU	4598	100,0%

Fonte: Elaborado pelos Autores.

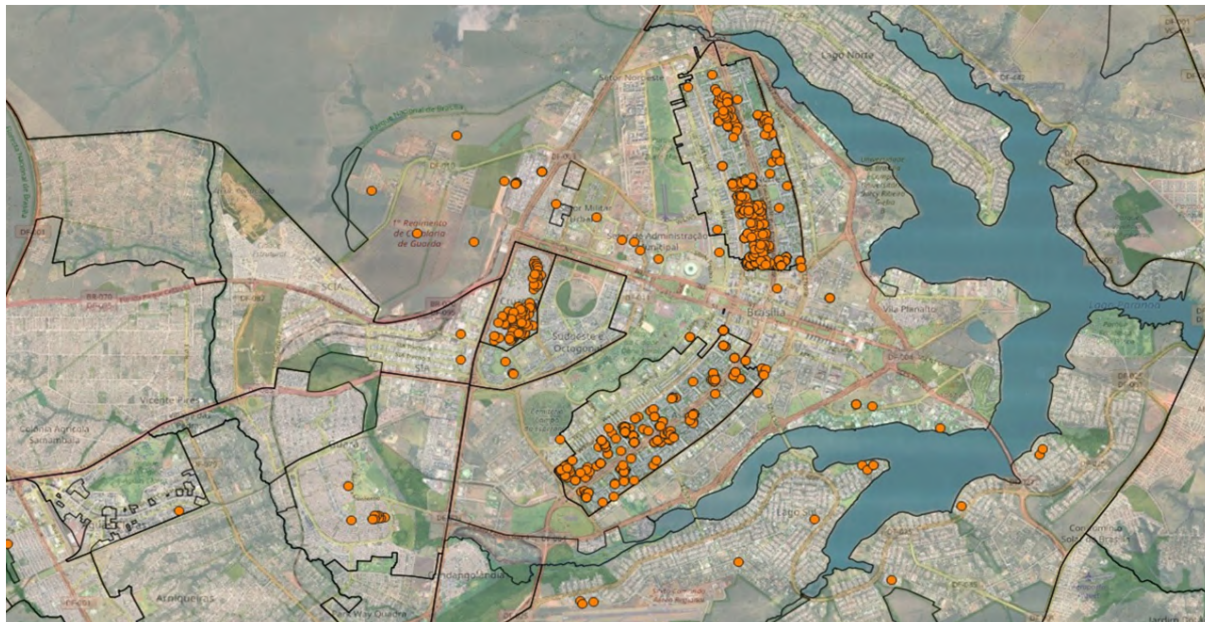
Figura 2 – AP do IBGE (em azul) sobrepostas pelos centroides de instâncias do EB e da SPU.



Fonte: Elaborado pelos Autores.

A Figura 3, por sua vez, ilustra alguns imóveis da base híbrida composta por elementos do EB e da SPU na região do Distrito Federal.

Figura 3 – AP do IBGE (transparentes com contorno em preto) sobrepostas pelos centroides de instâncias do EB e da SPU na região do Distrito Federal.



Fonte: Elaborado pelos Autores.

Por fim, foram coletadas as quantidades de diversas tipologias de pontos de interesse, tais como estações de metrô, parques, *shopping centers* e hospitais, em um raio preferencial de 400 metros de cada propriedade, correspondente a aproximadamente cinco (5) minutos de caminhada a pé, tomando as coordenadas dos centroides dos imóveis como referência. Para tal, consumiu-se a *Application Programming Interface (API) Google Places Nearby Search*.

4.2. Tratamento de Instâncias

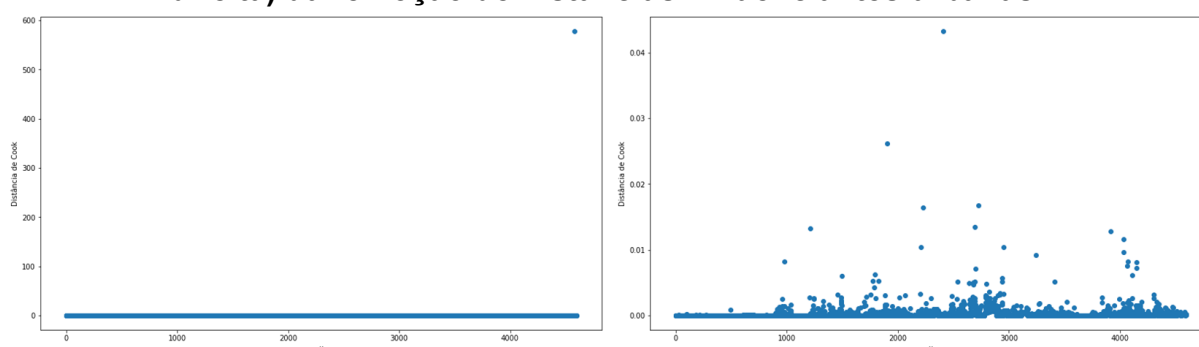
Foram desenvolvidas três abordagens para tratamento das instâncias coletadas, em decorrência da elevada dispersão observada das variáveis *Área do Terreno*, *Valor Total Atualizado* e *Valor Unitário Atualizado*, estas duas últimas atualizadas a janeiro de 2022 com uso do índice FipeZap Brasil, respeitando-se o tipo de uso do bem em questão. Segue uma síntese das abordagens abaixo:

- (1) **Abordagem 1 (A_1)** Remoção de *outliers* pelo método de Tukey baseada na variável *Valor Unitário Atualizado* (R\$/m²) e utilização deste atributo como variável explicada.
- (2) **Abordagem 2 (A_2)** Transformação das variáveis *Área do Terreno* com logaritmo neperiano (ln), sem remover outliers, e ln de *Valor Total Atualizado* como variável explicada.

(3) **Abordagem 3 (A_3)** Transformação do atributo *Valor Unitário Atualizado* com $\ln$, remoção da variável *Área do Terreno* e $\ln$ de *Valor Unitário Atualizado* como variável explicada.

A abordagem A_2 apresentou melhores resultados, indicando necessidade de eliminação de apenas três instâncias, todas da SPU, das 4598 provenientes do processo de geocodificação. Na modelagem mais simples, linear, as 4595 ocorrências resultantes se mostraram não influenciantes sob avaliação da distância de Cook, conforme Figura 4.

Figura 4 – Distâncias de Cook calculadas antes (à esquerda) e depois (à direita) da remoção de instâncias influenciantes urbanas.



Fonte: Elaborado pelos Autores.

4.3. Modelagem

Os modelos urbanos propostos utilizaram os atributos constantes na Tabela 3 após rigoroso processo de seleção de variáveis, sobre o qual se discorrerá na Subseção 4.4, e se valeram das transformações matemáticas da abordagem A_2 . Eles foram implementados com uso dos seguintes algoritmos e bibliotecas:

- Regressão Linear Múltipla (OLS), biblioteca *statsmodels* versão 0.14.0 em *Python*;
- Regressão Espacial *Spatial Two Stage Least Squares* (S2SLS), conforme proposto por [Anselin 1988], pacote *sprege* versão 1.3.0 da biblioteca *PySAL* em *Python*;
- *SGDRegressor* linear, biblioteca *scikit-learn* versão 1.2.2 em *Python*;
- Rede Neural Artificial *ANN MLPRegressor* não linear, biblioteca *scikit-learn* versão 1.2.2 em *Python*; e
- *XGBRegressor* não linear, biblioteca *XGBoost* versão 1.7.4 em *Python*.

Para o modelo de regressão espacial, foram criadas matrizes de pesos espaciais W fundamentadas no livro de [Anselin et al. 2014]. Testamos algumas distâncias como largura de banda δ a fim de definir os elementos diferentes de zero na matriz apenas para os casos $d_{ij} < \delta$, em que d_{ij} representa a distância euclidiana entre os elementos i e j considerando suas coordenadas geográficas e o raio de curvatura terrestre.

Para a amostra de imóveis urbanos, a distância de 1 km se mostrou um bom limite (após testes realizados entre 100 metros e 20 km), implicando boa significância fundamentada em t de Student para a matriz de vizinhança e melhora do R^2 em relação ao modelo OLS, o qual não considera a autocorrelação espacial, a influência que o valor de um imóvel exerce sobre os valores de seus vizinhos matriciais e vice-versa.

Os pesos das matrizes espaciais foram calculados com função de decaimento igual ao inverso da distância entre os imóveis.

4.4. Análise e Seleção de Atributos

O critério para consideração inicial de um atributo foi baseado na sua disponibilidade sob forma estruturada ou passível de estruturação e pela sua presença nos bancos de dados do EB e da SPU.

No caso de imóveis urbanos, a identificação dos subconjuntos de preditores úteis e alinhados ao pressuposto de não multicolinearidade exigido à aplicação de modelos de regressão linear múltipla foi realizada por meio das técnicas: análise de correlações bivariadas (máximo de 80% de correlação de Pearson, conforme indicado na NBR 14653), denominada critério C_1 ; regressão *forward-backward stepwise* (limites p -valor de entrada e de saída de 30%, fundamentados na NBR 14653), denominada critério C_2 ; eliminação recursiva de atributos com estimador regressivo *Random Forest*, denominada critério C_3 ; e análise de significâncias com procedimento não paramétrico *bootstrapping* bicaudal de 10 mil subamostras aleatórias por meio da construção de modelo conceitual de equações estruturais com mínimos quadrados parciais (PLS-SEM), denominada critério C_4 . À exceção do PLS-SEM, implementado com uso do *software SmartPLS 4*, as demais técnicas foram desenvolvidas com as bibliotecas *pandas* e *scikit-learn* em linguagem *Python*.

A aplicação dos algoritmos de seleção resultou nos atributos apresentados na Tabela 3 e no modelo conceitual PLS-SEM representados nas Figuras 5, na qual constam os modelos de medida (externos), caracterizados pelo relacionamento entre os indicadores (em amarelo) e as variáveis latentes exógenas (em azul), e o modelo estrutural (interno), marcado pelo relacionamento entre as variáveis latentes exógenas, de controle, as moderadoras e a endógena. Os indicadores destacados com contorno em cor laranja são aqueles com nível de granularidade espacial intramunicipal.

Nos modelos de medida (externos), constam os pesos externos e os p -valores associados aos testes de hipótese de nulidade de regressores com base em distribuição t de Student, entre parênteses; já para os estruturais (internos), constam os coeficientes de caminho, os p -valores de seus testes de hipótese, entre parênteses, e o coeficiente de determinação para a variável latente endógena $Ln(\text{Valor Total Atualizado})$.

Analisando-se a relação entre o construto formativo *Caracterização do Terreno* e a variável latente endógena *Valor do Imóvel*, ilustrados na Figura 5, percebe-se que se trata da conexão mais forte do modelo estrutural.

Ainda a respeito do PLS-SEM, já que os construtos se mostraram de natureza formativa, analisamos os fatores de inflação da variância (VIF) e tanto os relativos ao modelo interno quanto ao externo foram inferiores a 3,3. Apesar de nem todas as magnitudes dos coeficientes de caminho serem superiores a 0,2 em módulo, o que representa efeitos menores segundo as regras heurísticas de [Cohen 1988], todos os *p-valores* associados a seus testes de hipóteses e àqueles dos pesos e cargas externas do modelo externo foram inferiores a 5%.

O coeficiente de determinação do modelo conceitual PLS-SEM de 62,0% sugere poder explicativo moderado, de acordo com [Hair 2010]. Pontua-se a fragilidade dos modelos conceituais construídos em decorrência de a variável latente endógena *Valor do Imóvel* ter apenas um indicador a ela associado.

Tabela 3 – Ocorrência de atributos nas modelagens finais de imóveis urbanos.

Variável	Tipo	Fonte	Critério de Exclusão
<i>Capital UF</i>	binária	IBGE	-
<i>Tipologia Municipal</i>	cód. binários	IBGE	C ₂
<i>Grau de Urbanização</i>	numérica	IBGE	C ₄
<i>IDHM</i>	numérica	IBGE	-
<i>IVS</i>	numérica	Ipea	C ₄
<i>Taxa de Homicídios</i>	numérica	Ipea	C ₄
<i>Vocação do Imóvel</i>	cód. binários	EB/SPU	-
<i>Terreno</i>	binária	EB/SPU	-
<i>Ln(Área do Terreno)</i>	numérica	EB/SPU	-
<i>CUB * Área Construída</i>	numérica	Sinduscon ^a /EB/SPU	-
<i>Idade Aparente</i>	cód. alocados	EB/SPU	-
<i>Vida Útil</i>	numérica	BIR ^b	C ₄
<i>% Residentes com Ensino Superior Completo AP</i>	numérica	IBGE	C ₄
<i>% Domicílios com Rede Geral de Distribuição de Água AP</i>	numérica	IBGE	C ₄
<i>% Domicílios com Microcomputador com Acesso à Internet AP</i>	numérica	IBGE	C ₁
<i>Renda Domiciliar AP</i>	numérica	IBGE	-
<i>Pontos de Interesse API Google Places^c</i>	numérica	API Google	-
<i>Pontos de Interesse Excluídos API Google Places^d</i>	numérica	API Google	C ₄
<i>Google Trends</i>	numérica	Google Trends	C ₄
<i>Coordenadas Geográficas^e</i>	numérica	EB/SPU	C ₃
<i>Ln(Valor Total Atualizado)</i>	numérica	EB/SPU	-

^a <https://www.cub.org.br/>

^b Bureau of Internal Revenue

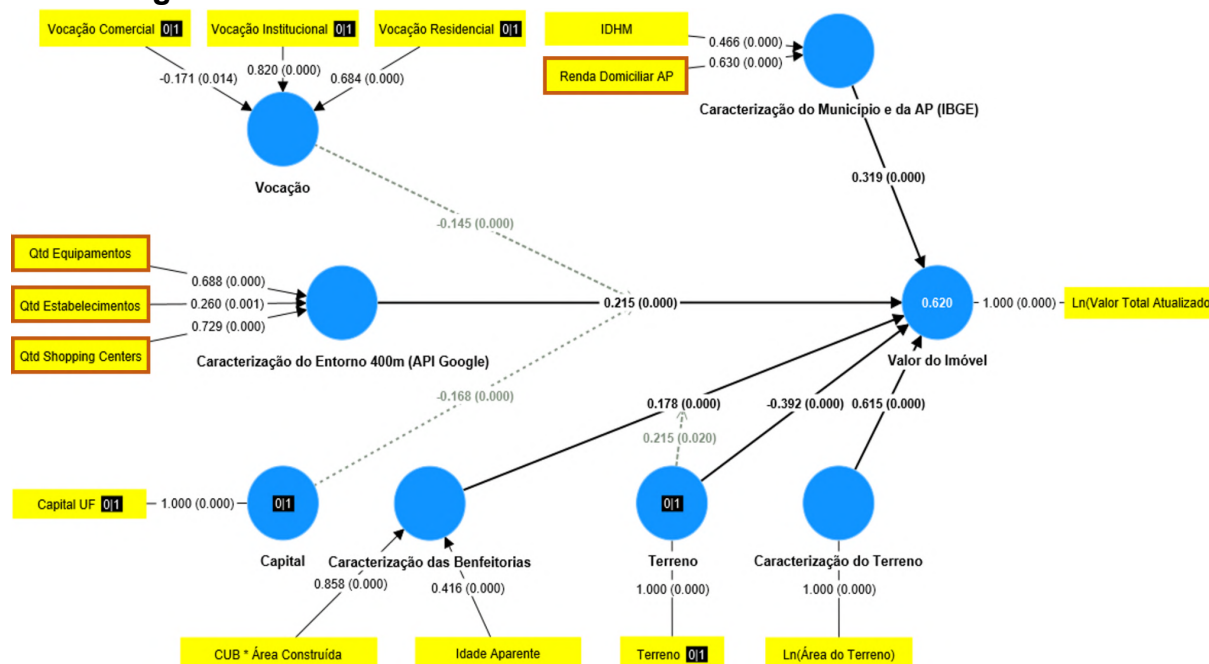
^c equipamentos (somatório de equipamentos urbanos, tais como hospitais, parques, delegacias, escolas, universidades, atrações turísticas e infraestrutura de transporte), estabelecimentos (quantidade de estabelecimentos comerciais, exceto *shopping centers*) e *shopping centers*

^d hospitais, parques, delegacias, escolas, universidades, atrações turísticas, supermercados, restaurantes, padarias, cafeterias, lojas, paradas de ônibus, estações de VLT, estações de metrô, estações de trem e aeroportos (apresentaram efeito difuso no PLS-SEM)

^e utilizadas para uniões espaciais, consultas API Google e construção de matriz de pesos espaciais

Fonte: Elaborado pelos Autores.

Figura 5 – Modelo conceitual PLS-SEM no software SmartPLS 4.



Fonte: Elaborado pelos Autores.

4.5. Tratamento de Vari veis

Foram aplicados testes de normalidade num ricos aos atributos explicativos selecionados. Tanto o teste de Shapiro-Wilk quanto o teste de Jarque-Bera indicaram, pelos *p-valores* obtidos, que h  evid ncias de que as vari veis t m assimetria e curtose significativamente diferentes de uma distribui  o normal.

Optou-se, portanto, pela aplica  o de transforma  o do tipo *MinMaxScaler*, utilizando a biblioteca *scikit-learn* vers o 1.2.2 na linguagem *Python*  s vari veis explicativas.

Na fase de pr -processamento, portanto, os atributos foram reescalados entre 0 e 1. Tal etapa evita que vari veis com maior amplitude de varia  o influenciem demasiadamente o modelo.

5. EXPERIMENTOS E RESULTADOS

Nesta se  o, s o apresentados os dados efetivamente utilizados, a configura  o dos experimentos, os cen rios de aplica  o, os resultados dos modelos e uma an lise das sa idas   luz das principais m tricas de avalia  o adotadas, o coeficiente de determina  o (R^2) e a raiz do erro quadr tico m dio (RMSE).

Vale ressaltar que a avaliação dos modelos de aprendizagem de máquina foram realizadas com validação cruzada com 10 *folds*, após testes de 2 a 15 *folds*, programados na biblioteca *scikit-learn* versão 1.2.2 em *Python*. As médias e os desvios padrões das métricas foram calculados para o conjunto de treinamento e para o de teste.

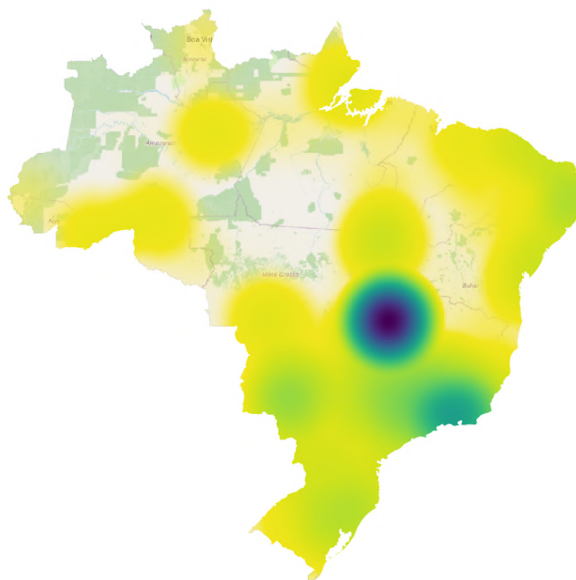
Os modelos OLS e de regressão espacial também foram avaliados quanto à probabilidade de nulidade simultânea de todos os seus regressores associados, o que representa uma hipótese grave, por meio da distribuição *F* de Fisher–Snedecor.

5.1. Bases de Dados e Cenários de Aplicação

Conforme etapas de pré-processamento discutidas Seção 4, a submissão de dados dos bancos do EB e da SPU, de maneira concatenada, a processos de geocodificação, enriquecimento espacial com informações de AP do IBGE (no caso de imóveis urbanos) e detecção e remoção de elementos influenciadores da amostra levou a reduções quantitativas das bases originais.

A amostra urbana efetivamente utilizada nos experimentos descritos nas subseções a seguir contemplou 4595 instâncias. A Figura 6 ilustra a distribuição geográfica dos elementos da amostra urbana, com destaque para o Distrito Federal, que concentra aproximadamente 29% dos dados.

Figura 6 – Mapa de densidade de imóveis que compõem a amostra urbana.



Fonte: Elaborado pelos Autores.

As correlações de Pearson entre as variáveis explicativas utilizadas nos experimentos desta seção foram calculadas e não se observaram correlações

parciais superiores a 80%, o que representa atendimento a um requisito normativo da NBR 14653.

5.2. Configuração dos Experimentos

A Tabela 4 contém a configuração dos principais parâmetros para a modelagem de imóveis urbanos com uso de aprendizagem de máquina após testes realizados com a configuração padrão dos algoritmos e diversas outras com o *GridSearchCV* da biblioteca *scikit-learn* versão 1.2.2. A ANN foi configurada com função de ativação tangente hiperbólica e *solver adam*, baseado em gradiente estocástico, como otimizador de pesos.

Tabela 4 – Parâmetros utilizados nos modelos de aprendizagem de máquina.

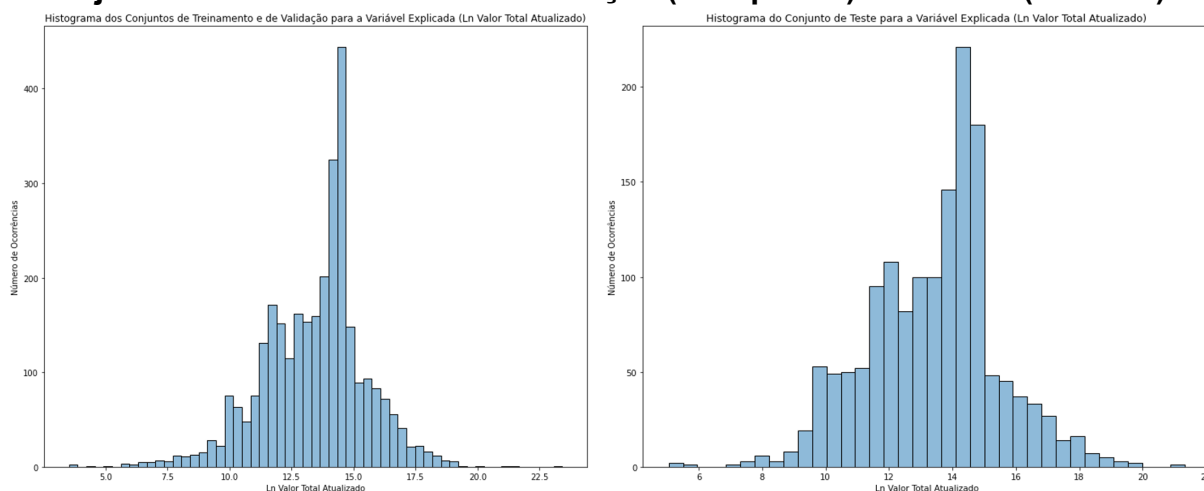
Parâmetro	<i>SGDRegressor</i> linear	<i>ANN MLPRegressor</i> não linear	<i>XGBRegressor</i> não linear
Fração de Treinamento	53,3%	53,3%	53,3%
Fração de Validação	13,3%	13,3	13,3%
Fração de Teste	33,3%	33,3	33,3%
Tipo de taxa de aprendizagem	constante	constante	constante
Valor da taxa de aprendizagem	0,0001	0,01	0,3
Função de custo	erro quadrático	erro quadrático	erro quadrático
Termo de regularização	<i>L2</i>	<i>L2</i> (multiplicador $\alpha = 0,1$)	<i>L1</i> ($\alpha = 5$) e <i>L2</i> ($\lambda = 1$)
Arquitetura básica	não se aplica	(5,5) ^a em U_1 , (6,6) em U_2 e (8,8) em U_3	<i>max_depth</i> = 5

^a camadas ocultas da rede neural artificial

Fonte: Elaborado pelos Autores.

A Figura 7 mostra a distribuição da variável dependente quando de sua utilização nos modelos desenvolvidos. A partição entre os conjuntos de treinamento, validação e teste foi feita de forma estratificada, com base na vocação dos imóveis.

Figura 7 – Histograma da variável dependente *Ln(Valor Total Atualizado)* nos conjuntos de treinamento e de validação (à esquerda) e de teste (à direita).



Fonte: Elaborado pelos Autores.

5.3. Resultados Obtidos

Foram obtidos os coeficientes de determinação (R^2) e as raízes dos erros quadráticos médios (RMSE) especificados na Tabela 5.

Tabela 5 – Resultados das métricas de avaliação dos modelos urbanos para os dois cenários considerados.

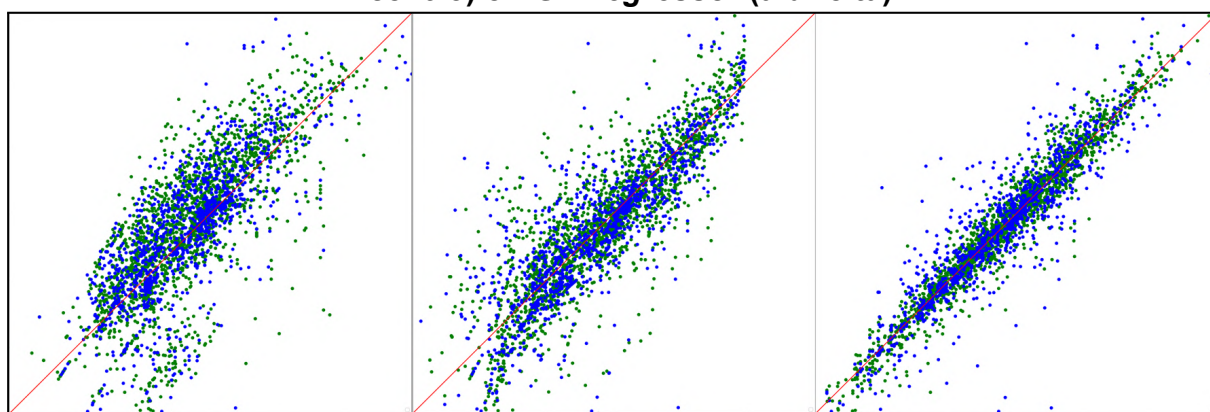
Modelo (Cenário)	R^2 médio 10 folds	dp R^2 10 folds	R^2 train	R^2 test	RMSE train	RMSE test
OLS (A)	NA	NA	61,9%	52,3%	1,28	1,73
Regressão Espacial (A)	NA	NA	63,4%	55,7%	1,26	1,39
<i>SGDRegressor</i> linear (A)	61,2%	4,7%	61,9%	61,1%	1,27	1,31
<i>SGDRegressor</i> linear (B)	60,7%	4,4%	61,7%	50,3%	1,29	1,35
ANN não linear (A)	73,8%	4,8%	74,8%	69,3%	1,04	1,17
ANN não linear (B)	71,8%	4,1%	73,8%	71,3%	1,06	1,02
<i>XGBRegressor</i> não linear (A)	86,2%	1,9%	95,7%	83,6%	0,43	0,85
<i>XGBRegressor</i> não linear (B)	86,7%	3,5%	94,8%	80,6%	0,47	0,84

NA: não se aplica

Fonte: Elaborado pelos Autores.

A maior concentração de pontos próximo às bissetrizes, destacadas em vermelho nos gráficos constantes na Figura 8, aliada às medidas de dispersão expostas na Tabela 5, são indícios de razoável assertividade das predições dos modelos não lineares. Na mesma figura, os dados de treinamento estão representados por pontos verdes e os de teste, por pontos azuis. Os eixos das ordenadas e das abscissas estão na mesma escala, na faixa de valores de 8 a 20.

Figura 8 – Gráficos de dispersão de ln de valores observados (ordenadas) versus ln de valores projetados pelos modelos (abscissas). Implementado com os algoritmos *SGDRegressor* (à esquerda), *ANN MLPRegressor* (ao centro) e *XGBRegressor* (à direita).



Fonte: Elaborado pelos Autores.

Para o modelo de regressão espacial, analisou-se a dependência espacial, autocorrelação global, a partir do índice de Moran e do teste de Anselin-Kelejian.

Ambos indicaram que as matrizes de pesos espaciais construídas com função de decaimento do inverso da distância entre os imóveis até o limite de 1 km, conforme detalhado na Subseção 4.3, são significantes.

Como os algoritmos *MLPRegressor* e *XGBRegressor* foram os únicos que atingiram consistentemente, nos dois cenários de aplicação, o objetivo específico de se obter coeficiente de determinação maior que 57%, um requisito normativo da NBR 14653, calculamos seus intervalos de confiança (IC) com 90% de grau de confiança. As semi-amplitudes inferiores e superiores dos intervalos calculados encontram-se na Tabela 6.

Tabela 6 – Semi-amplitudes dos intervalos de confiança 90% calculados para os modelos urbanos de melhor performance.

Modelo	Limite IC Inferior	Limite IC Superior	Amplitude IC
<i>ANN MLPRegressor</i>	-51,1%	+98,8%	149,9%
<i>XGBRegressor</i>	-41,9%	+73,3%	115,2%

Fonte: Elaborado pelos Autores.

5.4. Interpretabilidade dos Resultados

Calcularam-se os coeficientes associados a cada variável e sua significância nas regressões lineares, com *p-valores* baseados na distribuição *t* de Student. Com as informações da Tabela 7, é possível se identificar as variáveis mais relevantes aos modelos específicos mais convencionais, regressão linear múltipla (OLS) e regressão espacial; destaca-se o atributo *Ln(Área do Terreno)*.

Tabela 7 – Coeficientes e *p-valores* associados calculados para os modelos mais convencionais. RE como abreviação de Regressão Espacial.

Atributo	Coef. OLS	<i>p-valor</i> OLS	Coef. RE	<i>p-valor</i> RE
<i>Constante / Intercepto</i>	5,80	0,00	5,95	0,00
<i>Capital UF</i>	0,73	0,00	0,53	0,00
<i>IDHM</i>	1,99	0,00	1,65	0,00
<i>Terreno</i>	-0,78	0,00	-0,66	0,00
<i>Idade Aparente</i>	1,49	0,00	1,60	0,00
<i>Renda Domiciliar AP (IBGE)</i>	1,69	0,00	1,43	0,00
<i>Equipamento (API Google)</i>	1,84	0,00	1,77	0,00
<i>Shopping Center (API Google)</i>	2,10	0,00	1,86	0,00
<i>Estabelecimento (API Google)</i>	0,14	0,33	0,13	0,36
<i>CUB * Área Construída</i>	4,26	0,00	4,31	0,00
<i>Ln(Área do Terreno)</i>	12,02	0,00	12,20	0,00
<i>Vocação Comercial</i>	-0,16	0,03	-0,10	0,18
<i>Vocação Institucional</i>	0,24	0,00	0,25	0,00
<i>Vocação Residencial</i>	-0,17	0,01	-0,13	0,03
<i>Matriz de Vizinhaça W</i>	NA	NA	0,05	0,00

NA: não se aplica

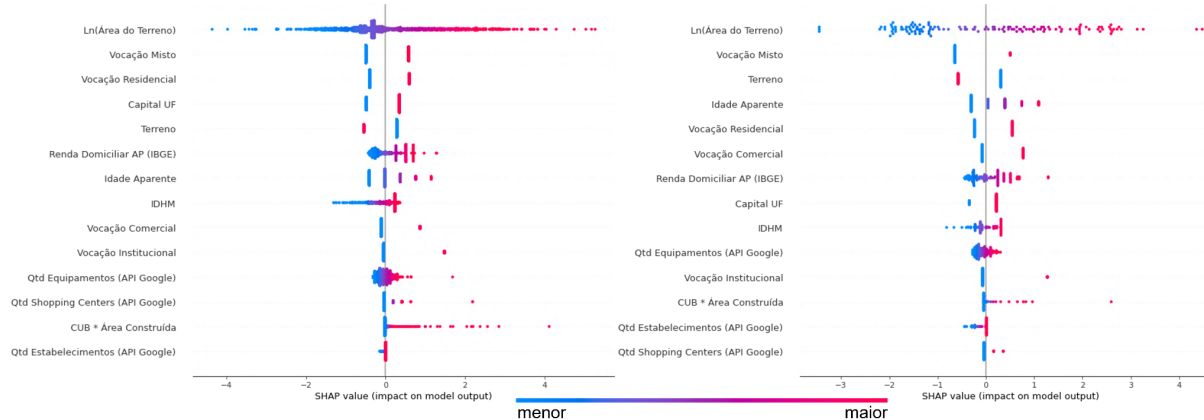
Fonte: Elaborado pelos Autores.

Cada ponto nos gráficos das Figuras 9, 10 e 11 representa um valor de Shapley para atributos individuais e instâncias específicas. As variáveis mais importantes à formação do valor dos imóveis em cada modelo estão posicionadas na porção superior das imagens.

Verifica-se que a relação das variáveis explicativas com a variável dependente se mostra coerente com a realidade observada no domínio em estudo para os modelos de aprendizagem *SGDRegressor* e *ANN MLPRegressor*. Para o *XGBRegressor*, atributos como *Terreno* e *Qtde Shopping Centers (API Google)* apresentam relação direta e inversa com o valor do imóvel, respectivamente, o que não faz sentido sob a ótica do senso comum. Tentou-se aumentar a força dos termos de regularização *L1* e *L2*, mas o referido comportamento não se alterou.

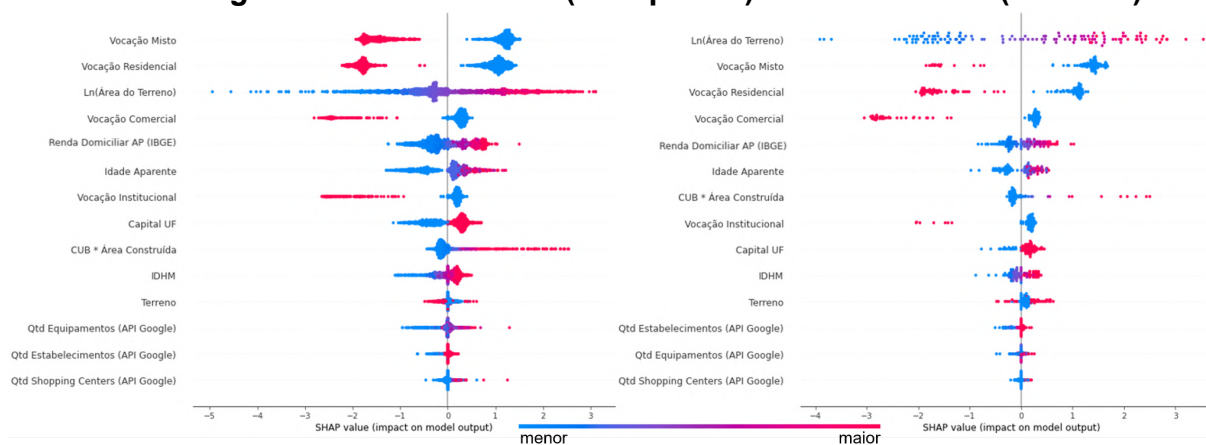
O que talvez possa explicar a mencionada inversão do *XGBRegressor* quanto à variável *Terreno* é a assimetria negativa acentuada da variável *CUB * Área Construída* em relação ao ponto de equilíbrio do valor de Shapley, observável na Figura 11. Ou seja, o simples fato de um imóvel não possuir benfeitorias construtivas já é tão penalizado pelo valor nulo de *CUB * Área Construída*, que *Terreno* entra com efeito compensatório. Tal comentário carece de análise de causalidade cuidadosa e detalhada.

Figura 9 – Valores de Shapley calculados para o algoritmo *SGDRegressor* no Cenário A (à esquerda) e no Cenário B (à direita).



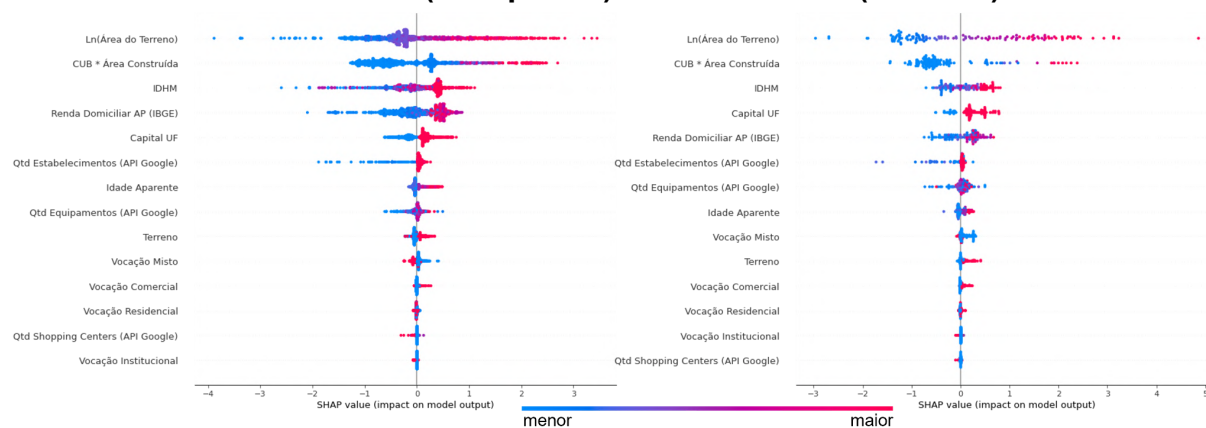
Fonte: Elaborado pelos Autores.

Figura 10 – Valores de Shapley calculados para o algoritmo ANN MLPRegressor no Cenário A (à esquerda) e no Cenário B (à direita).



Fonte: Elaborado pelos Autores.

Figura 11 – Valores de Shapley calculados para o algoritmo XGBRegressor no Cenário A (à esquerda) e no Cenário B (à direita).



Fonte: Elaborado pelos Autores.

5.5. Discussão dos Resultados

O PLS-SEM se mostrou um critério relevante e visualmente acessível no sentido de selecionar variáveis explicativas urbanas, contribuindo para a interpretabilidade da metodologia.

O atributo *Ln(Área do Terreno)* se mostra como a variável quantitativa mais importante na formação de valor dos imóveis urbanos, em modelos lineares e não lineares.

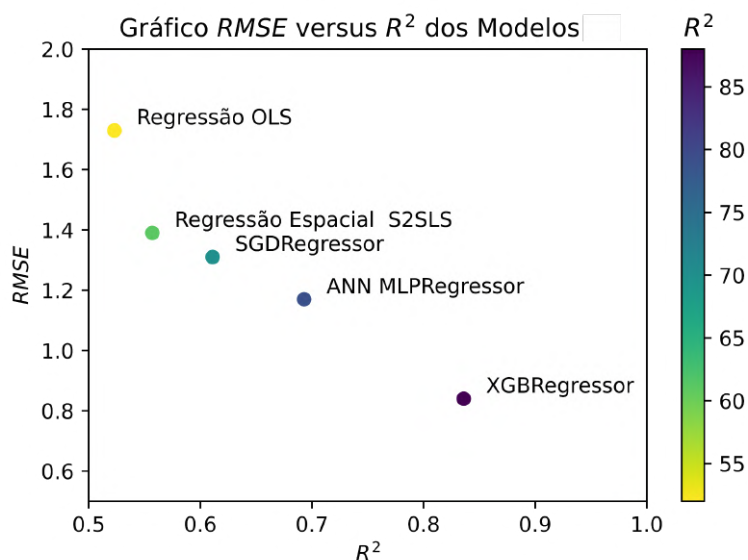
A partir da inclusão de variáveis com granularidade espacial nível AP (IBGE) e relativas ao entorno de 400 metros (API Google) dos imóveis urbanos, obteve-se incremento de poder de diferenciação entre ocorrências dentro de um mesmo

município. Os referidos atributos possibilitaram a obtenção de R^2 mais elevados, RMSE menores e intervalos de confiança de menor amplitude para um mesmo grau de risco.

A despeito de o algoritmo *XGBRegressor* ter performance melhor em todos os cenários urbanos, a exemplo do comparativo entre os modelos no Cenário A representado na Figura 12, as *ANN MLPRegressor* desenvolvidas parecem mais adequadas à valoração dos imóveis administrados pelo EB. Os pontos levantados nas Subseções 5.3 e 5.4 indicam que as redes neurais treinadas têm uma performance relativamente equilibrada e que a análise de sua interpretabilidade à luz do valor de Shapley se mostra mais aderente à realidade; a valorização e a depreciação da variável dependente a partir da variação de cada atributo explicativo ocorrem de forma mais alinhada ao senso comum.

Com a análise do *p-valor* associado à matriz de vizinhança *W* e com a autocorrelação espacial dos valores dos imóveis urbanos atestada pelo *i* de Moran, fica ainda mais evidente a importância da localização na formação das valorações imobiliárias.

Figura 12 – Gráfico comparativo entre os modelos urbanos no Cenário A considerando as métricas de avaliação R^2 e RMSE.



Fonte: Elaborado pelos Autores.

Em decorrência da importância de fatores locacionais, os modelos destinados a imóveis urbanos tendem a performar de maneira superior nas regiões de maior densidade de bens anotados, ilustradas na Figura 6.

Para os modelos urbanos, as estimativas de valor são mais robustas, com menor grau de incerteza, para locais como Brasília e Rio de Janeiro. As projeções para

ativos situados em lugares ermos, no interior da região Norte, por exemplo, distantes de outros imóveis, pela sua própria natureza, envolvem grau maior de incerteza.

Imóveis com valor histórico-cultural associado, tais como fortalezas históricas, podem ter componente intangível e valor de bens integrados, como pinturas e monumentos, acrescidos às estimativas deste trabalho.

Os *scripts* relativos aos experimentos encontram-se parcialmente disponíveis no seguinte repositório GitHub: *imoveisml*. Os modelos treinados e as bases de dados são de acesso restrito.

6. CONCLUSÕES E RECOMENDAÇÕES

Este trabalho é um subproduto de uma dissertação de mestrado apresentada no Programa de Pós-Graduação em Computação Aplicada (PPCA) da Universidade de Brasília (UnB). Ele explorou o problema de se estimar, em níveis iniciais, valores de imóveis urbanos da União, em especial daqueles administrados pelo EB, de forma a permitir que as autoridades decisoras tenham acesso a tal informação antes de solicitar formalmente laudos de avaliação imobiliária.

Os modelos implementados com os algoritmos *ANN MLPRegressor* e *XGBRegressor* tiveram coeficiente de determinação consistentemente superior a 57%, em todos os cenários desenhados. Ressalta-se, entretanto, que se tratam de estimativas com grau de incerteza associado, mais acentuado na região Norte e menos acentuado em zonas de maior densidade de dados anotados, tais como Brasília, Rio de Janeiro e São Paulo.

A partir dos experimentos realizados e dos resultados alcançados nesta pesquisa, enxerga-se como promissor realizar novos experimentos incluindo informações do Censo IBGE 2022, quando disponíveis. O valor venal referencial referencial à cobrança de Imposto Predial e Territorial Urbano (IPTU) extraído das plantas de valores genéricos dos municípios talvez constitua um atributo significativo e, conseqüentemente, ajude a aperfeiçoar os modelos; ressalta-se que, muitas vezes, ele não se encontra disponível de maneira estruturada e que os municípios não o calculam de uma única forma, o que exigiria intensa etapa de tratamento sobre ele.

Complementarmente, fotografias dos imóveis poderiam ser analisadas e ter padrões próprios reconhecidos por meio de redes convolucionais, com potencial de aprimoramento dos modelos preditivos construídos neste trabalho, possivelmente com a agregação de uma variável indicativa do estado de conservação dos imóveis.

Abordagens sob a ótica de demanda mercadológica por imóveis também seria um ponto interessante. Tentou-se incluir a variável *Google Trends* com granularidade espacial estadual na etapa de seleção de atributos, mas ela não se mostrou significativa. O desemprego local pode exercer influência sobre este aspecto.

REFERÊNCIAS BIBLIOGRÁFICAS

ANSELIN, L. *Spatial Econometrics: Methods and Models*. Springer Dordrecht, ISBN 978-94-015-7799-1, 1988.

ANSELIN, L.; REY, S. *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. Geoda Press LLC, ISBN 978-0-986-34210-3, 2014.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **14653-1**: Avaliação de bens — Procedimentos gerais — Apresentação. Rio de Janeiro: ABNT, 2019.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **14653-2**: Avaliação de bens — Imóveis urbanos — Apresentação. Rio de Janeiro: ABNT, 2011.

CAMPOS, R. B. A.; ALMEIDA, E. *Decomposição espacial nos preços residenciais no município de São Paulo*. **Estudos Econômicos**, v. 48(1), p. 5-38, 2018.

COHEN, J. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, ISBN 978-0-203-77158-7, 1988.

CHAKRABORTY, D.; ELZARKA, H. *A novel construction cost prediction model using hybrid natural and light gradient boosting*. **Advanced Engineering Informatics**, v. 46, 2020.

DANTAS, R. A.; MAGALHÃES, A. M.; VERGOLINO, J. R. O. *Um Modelo Espacial de Demanda Habitacional para a Cidade do Recife*. **Estudos Econômicos**, v. 40(4), p. 891-916, 2010.

DEWAN, P.; GANTI, R.; SRIVATSA, M.; STEIN, S. *NN-SAR: A neural network approach for spatial autoregression*. **2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)**, Kyoto, p. 783-789, 2019.

FURTADO, B. A. T. *DISET 78 - Gerando Famílias Artificiais Intraurbanas: Censo 2010*. Ipea, 2020.

HAGENAUER, J.; HELBICH, M. *A geographically weighted artificial neural network*. **International Journal of Geographical Information Science**, v. 36(2), p. 215-235, 2022.

HAIR, J.; BLACK, W.; BABIN, B.; ANDERSON, R. *Multivariate Data Analysis: A Global Perspective*. Pearson Education, ISBN 978-0-135-15309-3, 2010.

HAIR, J.; HULT, G. T. M.; RINGLE, C.; SARSTEDT, M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Sage Publishing, ISBN 978-1-5443-9640-8, 2022.

KIELY, T.; BASTIAN, N. *The spatially conscious machine learning model*. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, v. 13(1), p. 31-49, 2020.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DEGRAVE, A.; PRUTKIN, J. M.; NAIR, B.; KATZ, R.; HIMMELFARB, N. B.; LEE, S. I. *From local explanations to global understanding with explainable AI for trees*. **Nature Machine Intelligence**, v. 2, p. 56-67, 2020.

PARK, B.; BAE, J. *Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data*. **Expert Systems with Applications**, v. 42(6), p. 2928-2934, 2015.

ROSEN, S. *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition*. **Journal Of Political Economy**, v. 82(1), p. 34-55, 1974.

TCHUENTE, D.; NYAWA, S. *Real estate price estimation in French cities using geocoding and machine learning*. **Annals of Operations Research**, v. 308(1-2), p. 571-608, 2022.

REALIZAÇÃO