

Estimativa de Valores de Imóveis da União com Técnicas de Aprendizagem de Máquina e Componentes Espaciais

Autores:

José Nilo Alves de Sousa Neto e Juliana Miura

Orientador:

Prof. Dr. Marcelo Ladeira

(Departamento de Computação da UnB)



11 A 15

SETEMBRO | 2023

SÃO PAULO - SP

REALIZAÇÃO



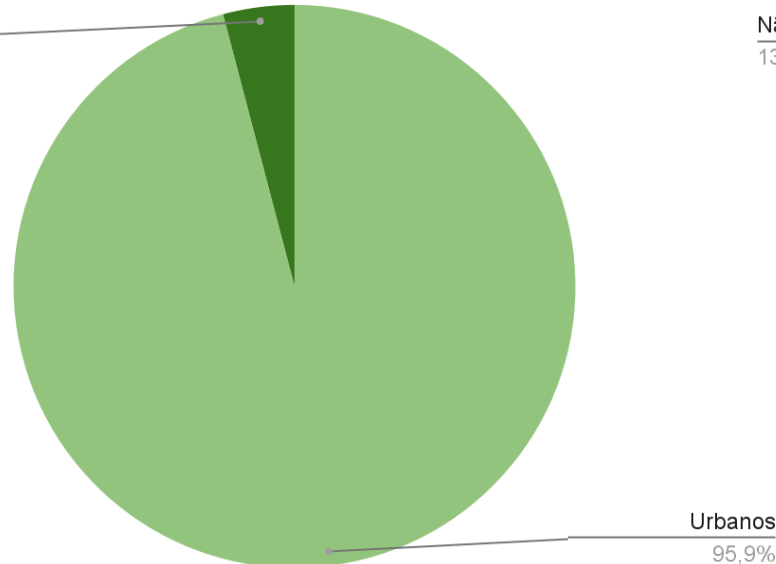
IBAPE NACIONAL
Instituto Brasileiro de Avaliações
e Perícias de Engenharia

PROBLEMA E CONTEXTO

- **Problema:** a mensuração de valores de imóveis no Exército Brasileiro é difícil, onerosa e demorada. Apenas 1,2% valorados.
- O EB possui **mais de 20 mil** parcelas imobiliárias da União sob sua administração. A União, **mais de 700 mil**.

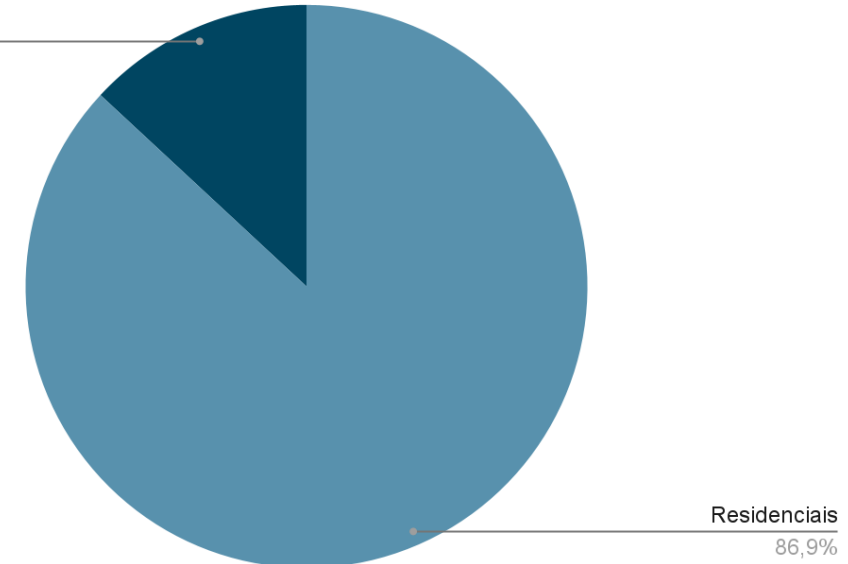
População (aprox. 21500 unidades imobiliárias)

Rurais
4,1%



População (aprox. 21500 unidades imobiliárias)

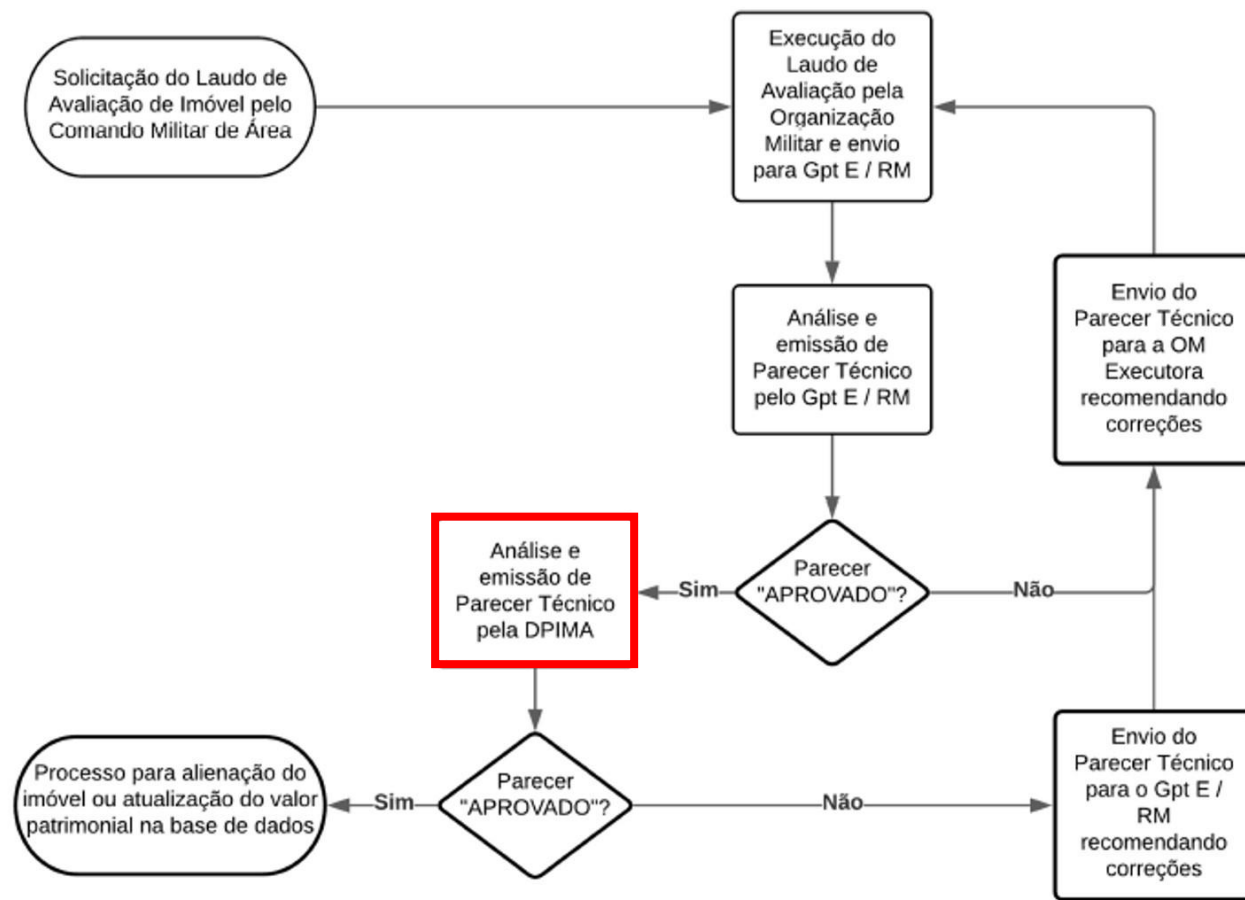
Não Residenciais
13,1%



CONTEXTUALIZAÇÃO



Fonte: DPIMA.



QUESTÃO DE PESQUISA

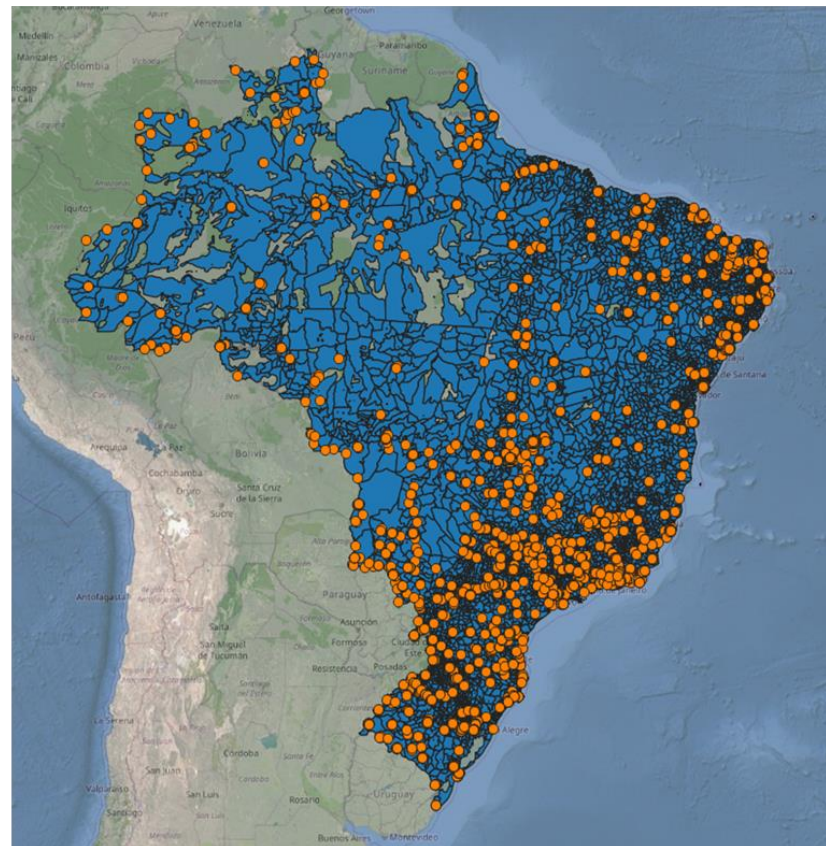
- Como propor modelos capazes de construir **estimativas de valor** dos imóveis administrados pelo EB de forma que a **solicitação formal dos laudos de avaliação** possa ser feita já com essa informação disponível?

Trabalho	$R^2 > 57\%$	Abrangência nível País	Modelagem linear	Modelagem não linear	Ajuste de hiperparâmetros	Componentes espaciais	Mais de uma vocalção de imóvel	Interpretabilidade
Dantas et al. 2010 [3]	✓		✓			✓		✓
Park e Bae 2015 [4]	✓		✓	✓	✓			
Kiely e Bastian 2020 [1]			✓	✓	✓	✓	✓	
Hagenauer e Helbich 2022 [2]		✓	✓	✓		✓		
Tchunte e Nyawa 2022 [5]	✓	✓	✓	✓	✓	✓		
Solução proposta	✓	✓	✓	✓	✓	✓	✓	✓

11^A15
SETEMBRO | 2023
SÃO PAULO - SP

REALIZAÇÃO

BASES DE DADOS

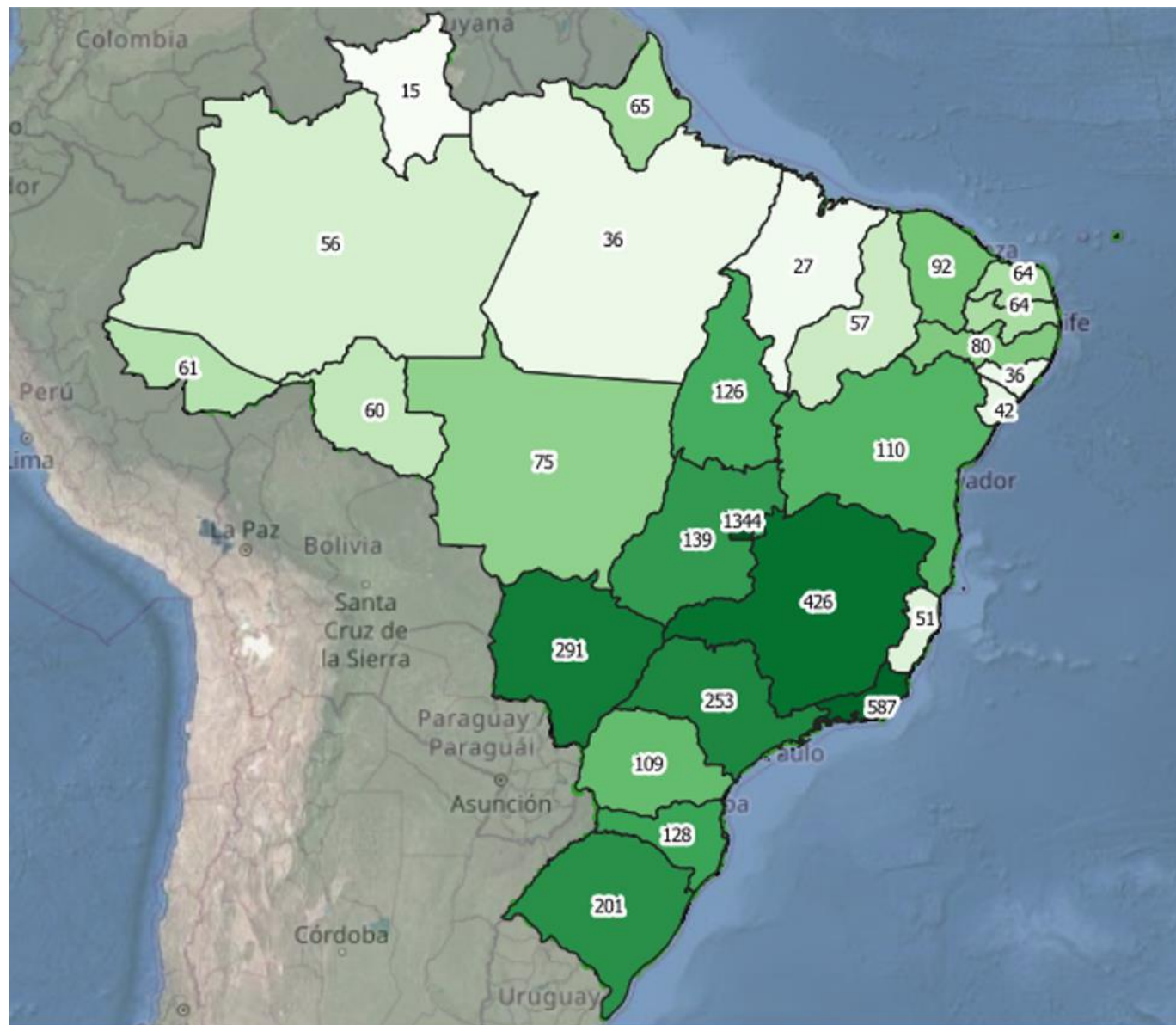
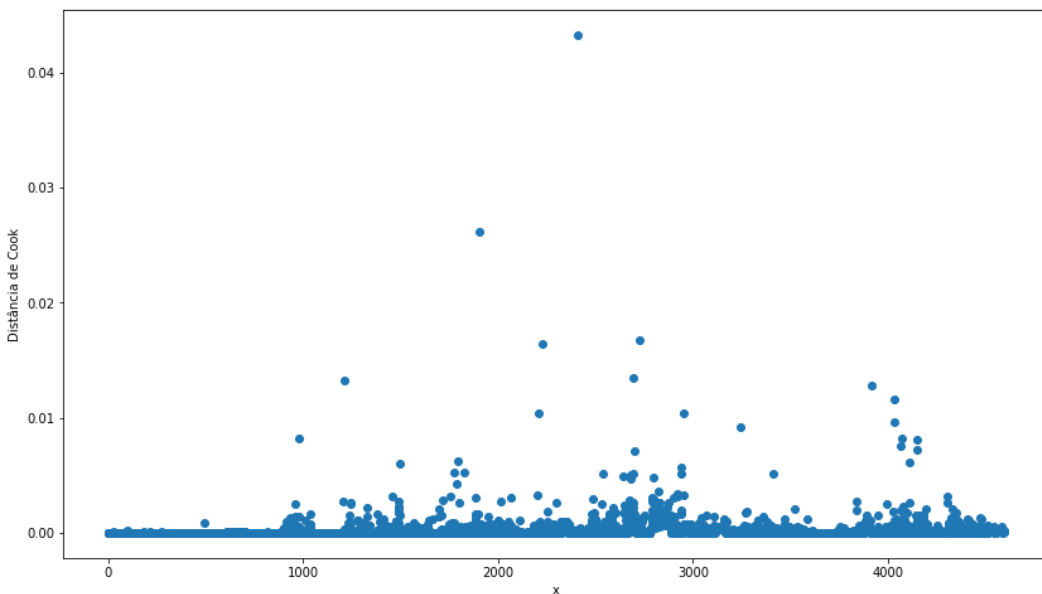
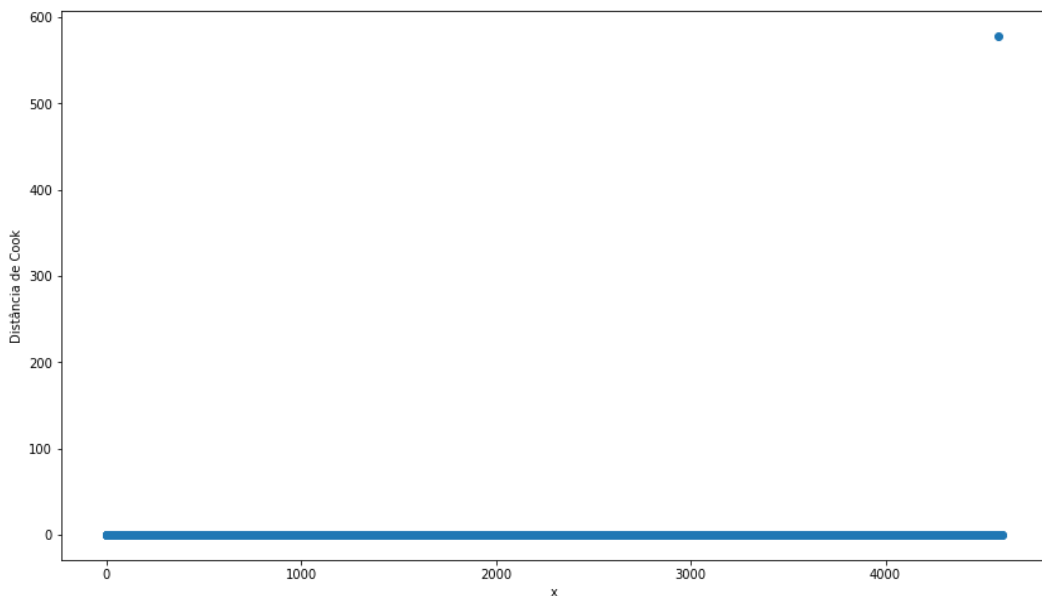


Fonte: EB /
SPU / IBGE /
GitHub Prof.
Dr. Bernardo
Furtado.

Fonte	Qtde instâncias urbanas com valor conhecido	Perda geocodificação	Qtde pós-geocodificação	Participação relativa
EB	258	0 (0,0%)	258	6,0%
SPU	4981	925 (18,5%)	4056	94,0%
EB e SPU	5239	925 (17,6%)	4313	100,0%

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

BASES DE DADOS



METODOLOGIA

- **Pré-processamento:** união dos arquivos gerados (incluindo elementos geográficos IBGE e Google), **codificação binária** de variáveis categóricas nominais.
- **Análise exploratória:** verificação gráfica e seleção de atributos (com SEM).
- Aplicação de **testes de normalidade** e de **transformações matemáticas** sobre os atributos.
- Validação cruzada: **10 folds**.
- **Testes intradomínio:** treinamento e teste em base **EB/SPU** (**Cenário A**).
- **Testes intradomínio:** treinamento em base **EB/SPU** e teste em base **EB** (**Cenário B**).
- **Comparação** dos modelos entre si (lineares e não lineares).
- Métricas de avaliação utilizadas: **coeficiente de determinação (R^2)** e **raiz do erro quadrático médio ($RMSE$)**.
- Interpretabilidade com **valor de Shapley** (Teoria dos Jogos).

EXPERIMENTOS REALIZADOS



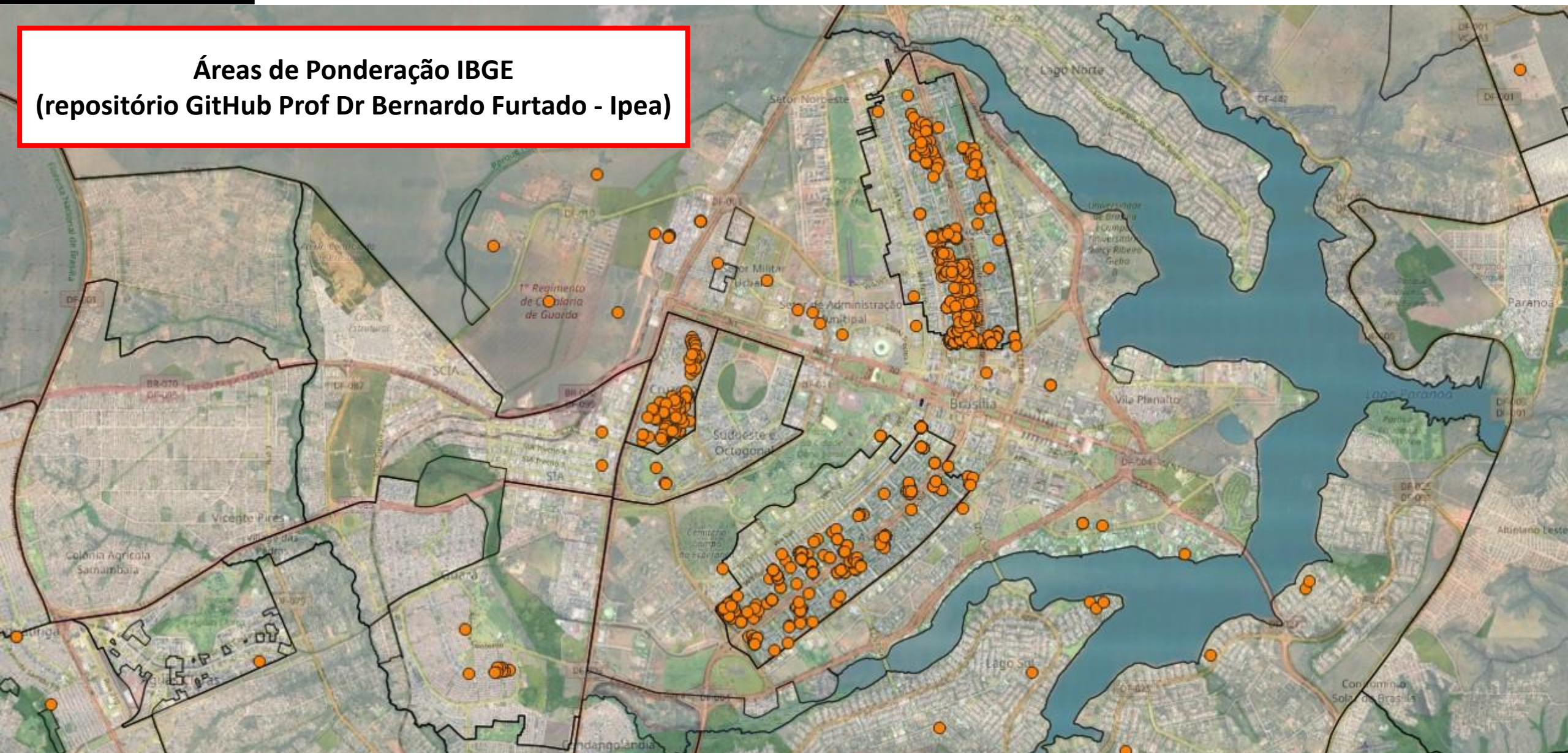
Geocodificação da base da SPU



Tratamentos e uniões espaciais

EXPERIMENTOS REALIZADOS

Áreas de Ponderação IBGE
(repositório GitHub Prof Dr Bernardo Furtado - Ipea)



EXPERIMENTOS REALIZADOS

Pontos de interesse
em raio de 400 m de cada imóvel:

IMÓVEIS URBANOS

data zap⁺



Google Places API

- **agregado equipamentos**
- atrações turísticas
- delegacias
- escolas
- hospitais
- parques
- universidades
- aeroportos
- estações de metrô
- estações de trem
- estações de VLT
- paradas de ônibus
- **agregado estabelecimentos**
- cafeterias
- lojas
- padarias
- restaurantes
- ***shopping centers***
- supermercados

EXPERIMENTOS REALIZADOS

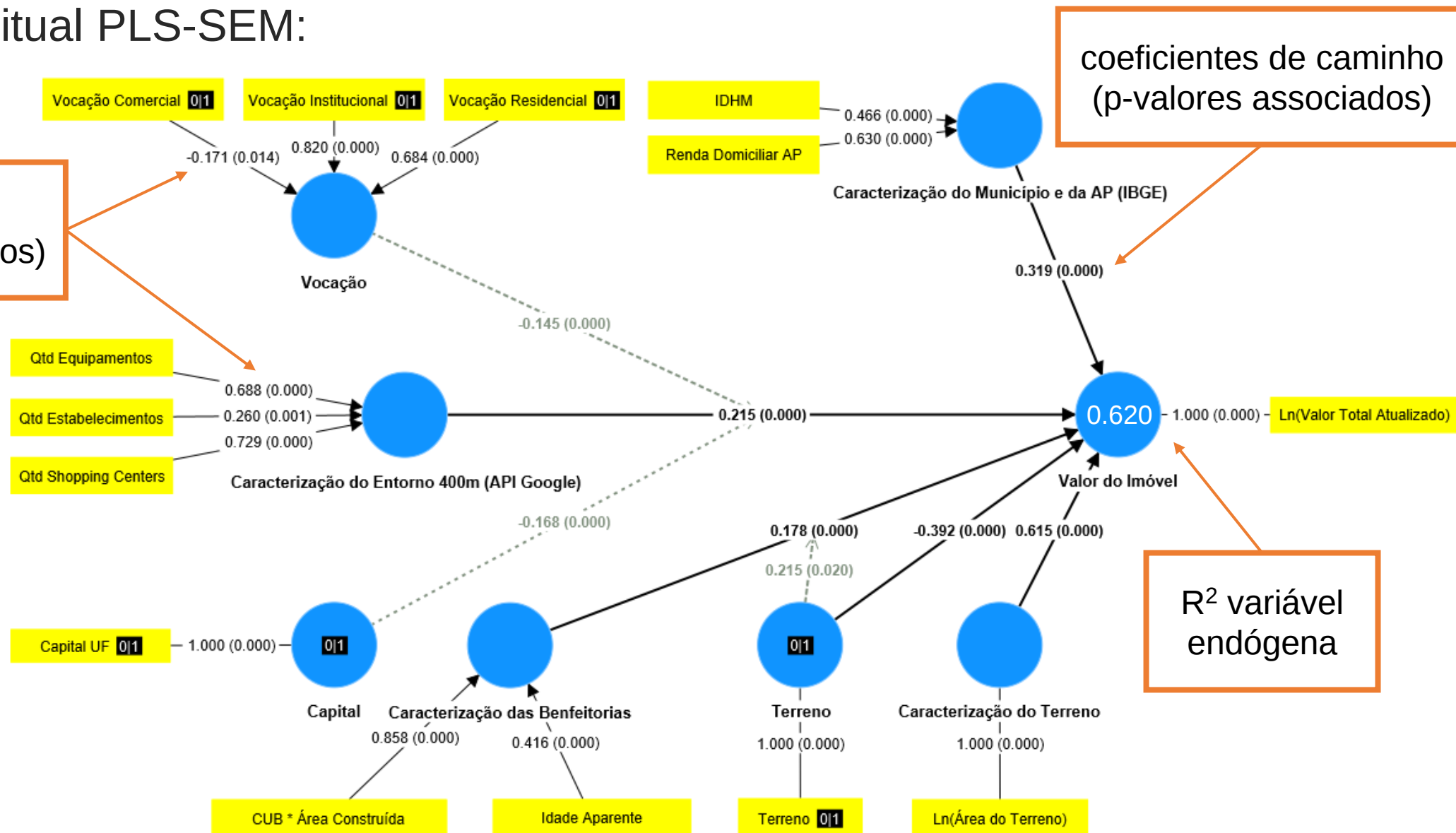
- Os seguintes atributos **urbanos** foram analisados:

Variável	Tipo	Unidade	Faixa de Valores	Fonte	Critério de Exclusão
<i>Capital UF</i>	binária	categ.	0 ou 1	IBGE	-
<i>Tipologia Municipal</i>	cód. binários	categ.	0 ou 1	IBGE	C ₂
<i>Grau de Urbanização</i>	numérica	adim.	0 a 0,986	IBGE	C ₄
<i>Índice de Desenvolvimento Humano Municipal</i>	numérica	adim.	0,522 a 0,847	IBGE	-
<i>Índice de Vulnerabilidade Social</i>	numérica	adim.	0,117 a 0,588	Ipea	C ₄
<i>Taxa de Homicídios</i>	numérica	%habitantes	0,0 a 138,76	Ipea	C ₄
<i>Vocação do Imóvel</i>	cód. binários	categ.	0 ou 1	EB/SPU	-
<i>Terreno</i>	binária	categ.	0 ou 1	EB/SPU	-
<i>Ln(Área do Terreno)</i>	numérica	ln(m ²)	2,0 a 11,2	EB/SPU	-
<i>CUB * Área Construída</i>	numérica	R\$	0 a 129 983 113,42	Sinduscon/EB/SPU	-
<i>Idade Aparente</i>	cód. alocados	categ.	0 a 4	EB/SPU	-
<i>Vida Útil</i>	numérica	anos	45 a 75	BIR	C ₄
<i>% Residentes com Ensino Superior Completo AP</i>	numérica	%	0,3 a 62,4	IBGE	C ₄
<i>% Domicílios com Rede Geral de Distribuição de Água AP</i>	numérica	%	52,2 a 100,0	IBGE	C ₄
<i>% Domicílios com Microcomputador com Acesso à Internet AP</i>	numérica	%	2,3 a 93,4	IBGE	C ₁
<i>Renda Domiciliar AP</i>	numérica	R\$	905,23 a 22 047,71	IBGE	-
<i>Pontos de Interesse API Google Places</i>	numérica	adim.	0 a 60	API Google	-
<i>Pontos de Interesse Excluídos API Google Places</i>	numérica	adim.	0 a 60	API Google	C ₄
<i>Google Trends</i>	numérica	adim.	48 a 100	Google Trends	C ₄
<i>Coordenadas Geográficas</i>	numérica	graus	todas UF	EB/SPU	C ₃
<i>Ln(Valor Total Atualizado)</i>	numérica	ln(R\$)	8,1 a 18,5	EB/SPU	-

*C₁ - correlações bivariadas / C₂ - regressão *forward-backward stepwise* / C₃ - eliminação recursiva com estimador *RF* / C₄ - PLS-SEM

EXPERIMENTOS REALIZADOS

- Modelo conceitual PLS-SEM:

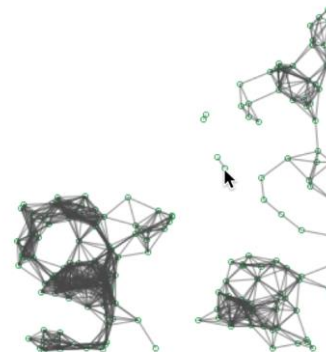


EXPERIMENTOS REALIZADOS

- Atributos **urbanos** efetivamente utilizados em cada nível de modelagem:

Variável	Nível de Modelagem
<i>Capital UF</i>	U ₁ , U ₂ , U ₃
<i>Vocação do Imóvel</i>	U ₁ , U ₂ , U ₃
<i>Índice de Desenvolvimento Humano Municipal</i>	U ₁ , U ₂ , U ₃
<i>Ln(Área do Terreno)</i>	U ₁ , U ₂ , U ₃
<i>Custo Unitário Básico * Área Construída</i>	U ₁ , U ₂ , U ₃
<i>Idade Aparente</i>	U ₂ , U ₃
<i>Renda Domiciliar AP</i>	U ₃
<i>Pontos de Interesse API Google Places</i>	U ₃
<i>Ln(Valor Total Atualizado)</i>	U ₁ , U ₂ , U ₃

- Regressão OLS
- Regressão Espacial S2SLS
- SGDRegressor linear
- ANN MLPRegressor não linear
- XGBRegressor não linear



Fonte: The Center For Spatial Data Science, Universidade de Chicago.

EXPERIMENTOS REALIZADOS

IMÓVEIS URBANOS

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	$L2$	$L2$ (multiplicador $\alpha = 0,1$)	$L1$ ($\alpha = 5$) e $L2$ ($\lambda = 1$)
Arquitetura básica	não se aplica	(5, 5) ^a em U_1 , (6, 6) em U_2 e (8, 8) em U_3	$max_depth = 5$

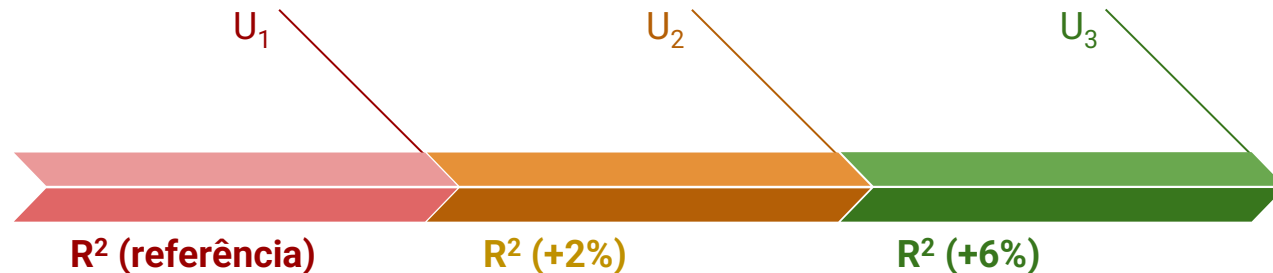
^a camadas ocultas da rede neural artificial



Treinamento - Validação - Teste

Amostragem estratificada com base
na variável "Vocação do Imóvel"

RESULTADOS OBTIDOS



Modelos U ₃ (Cenário)	R^2 médio 10 <i>folds</i>	dp R^2 10 <i>folds</i>	R^2 <i>train</i>	R^2 <i>test</i>	$RMSE$ <i>train</i>	$RMSE$ <i>test</i>
OLS (A)	NA ^a	NA ^a	61,9%	52,3%	1,28	1,73
Regressão Espacial (A)	NA ^a	NA ^a	63,4%	55,7%	1,26	1,39
<i>SGDRegressor</i> linear (A)	61,2%	0,047	61,9%	61,1%	1,27	1,31
<i>SGDRegressor</i> linear (B)	60,7%	0,044	61,7%	50,3%	1,29	1,35
ANN não linear (A)	73,8%	0,048	74,8%	69,3%	1,04	1,17
ANN não linear (B)	71,8%	0,041	73,8%	71,3%	1,06	1,02
<i>XGBRegressor</i> não linear (A)	86,2%	0,019	95,7%	83,6%	0,43	0,85
<i>XGBRegressor</i> não linear (B)	86,7%	0,035	94,8%	80,6%	0,47	0,84

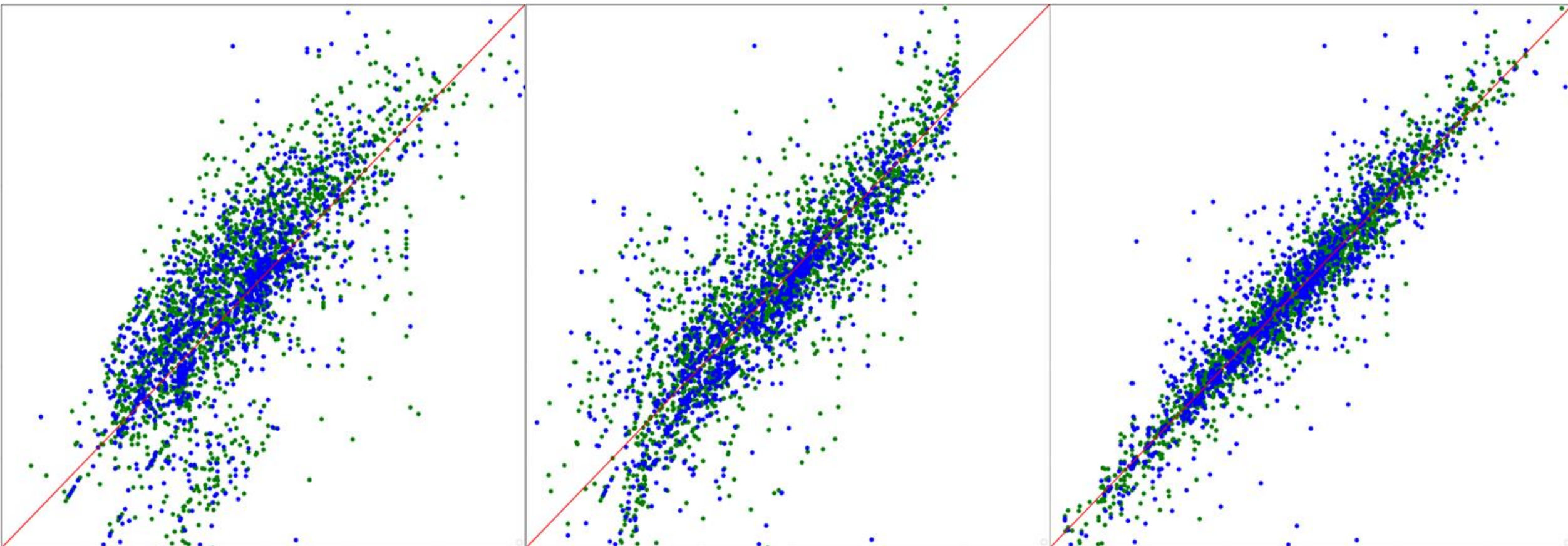
^a não se aplica

RESULTADOS OBTIDOS

U₃ *SGDRegressor* (Cenário A)

U₃ *ANN MLPRegressor* (Cenário A)

U₃ *XGBRegressor* (Cenário A)



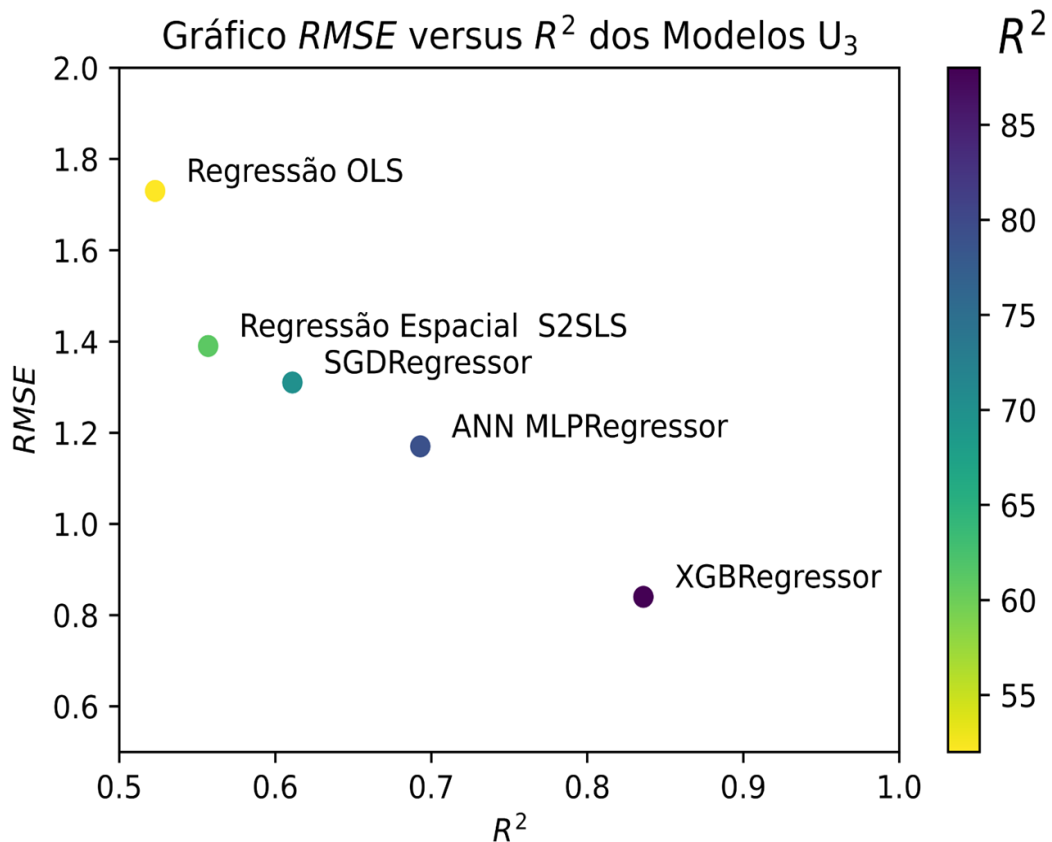
- dados de treinamento
- dados de teste

eixo das ordenadas
Ln(Valores Observados)

(8,8)

eixo das abscissas
Ln(Valores Estimados pelo Modelo)

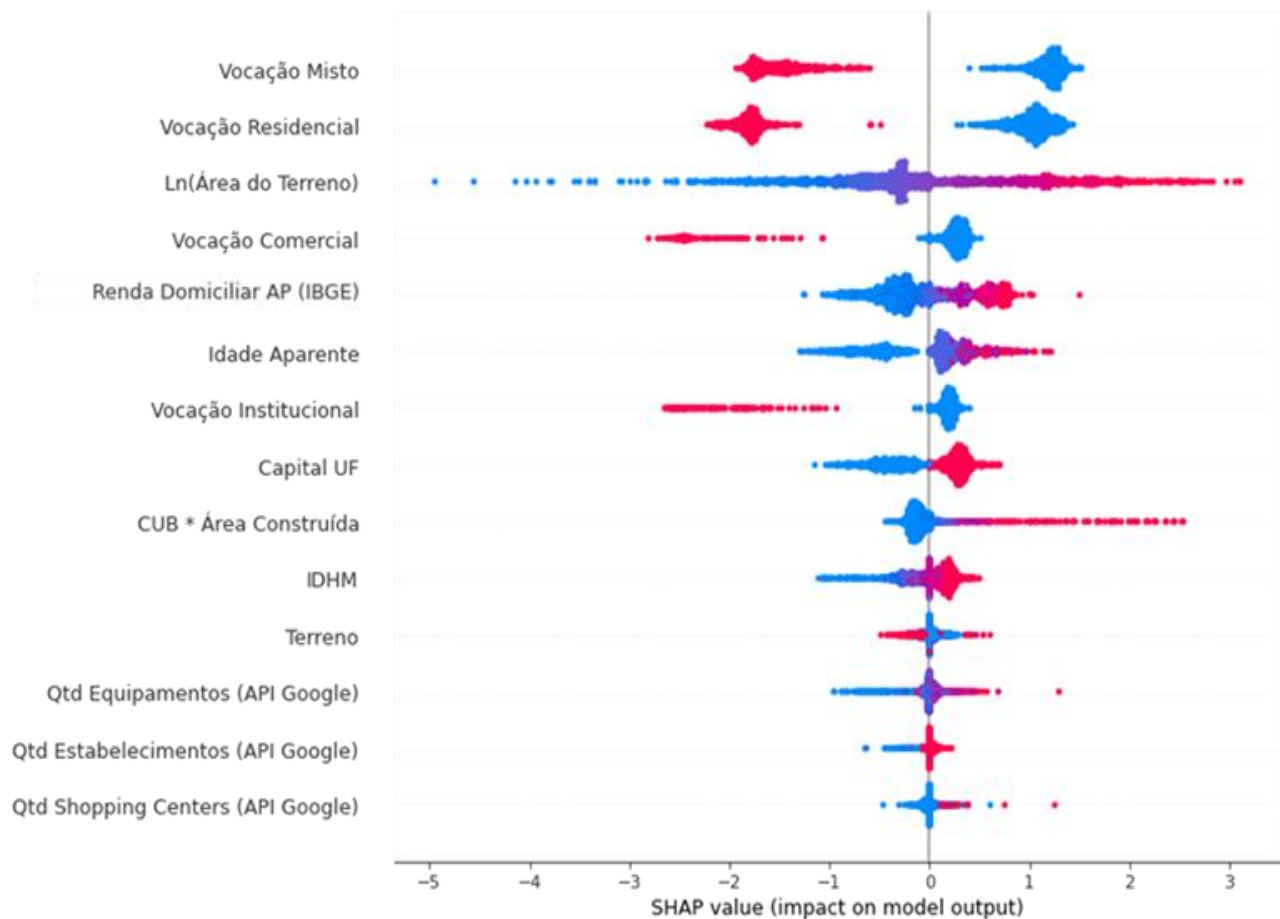
RESULTADOS OBTIDOS



Modelo	Limite IC Inferior	Limite IC Superior	Amplitude IC
<i>ANN MPLRegressor</i> U_1	-51,0%	+109,0%	160,0%
<i>ANN MPLRegressor</i> U_2	-50,9%	+107,9%	158,8%
<i>ANN MPLRegressor</i> U_3	-51,1%	+98,8%	149,9%
<i>XGBRegressor</i> U_1	-43,9%	+79,8%	123,7%
<i>XGBRegressor</i> U_2	-43,8%	+79,6%	123,4%
<i>XGBRegressor</i> U_3	-41,9%	+73,3%	115,2%

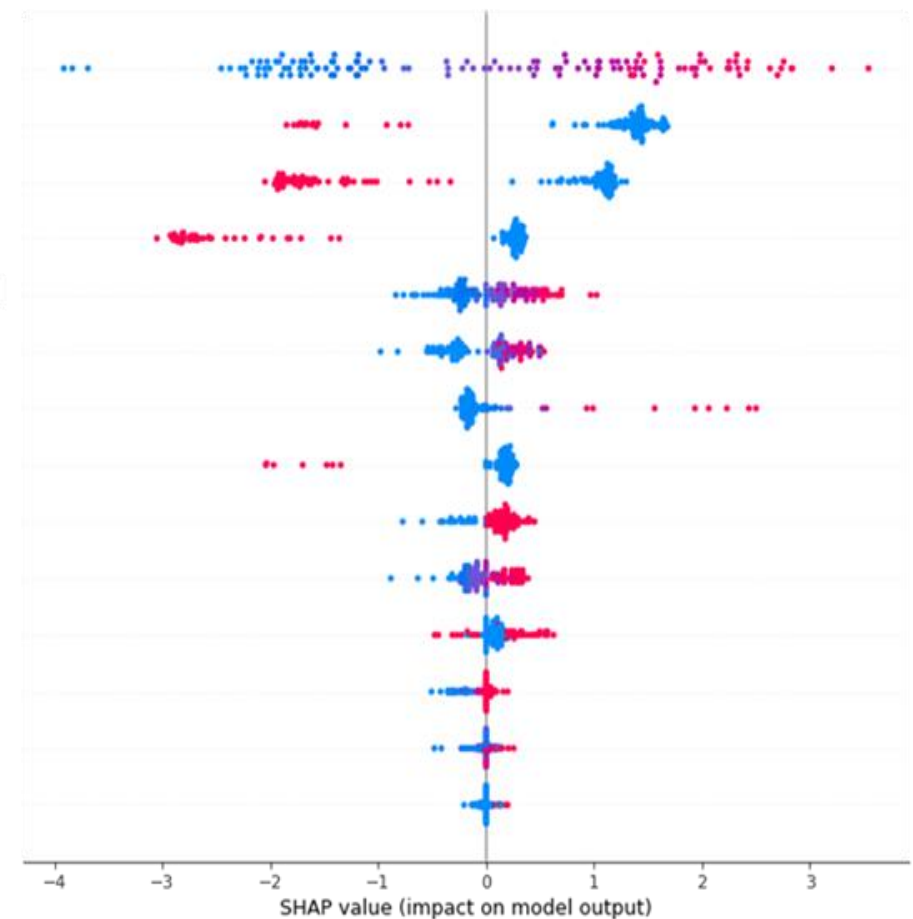
RESULTADOS OBTIDOS

SHAP values (*ANN MLPRegressor* não linear)



Cenário A

menor

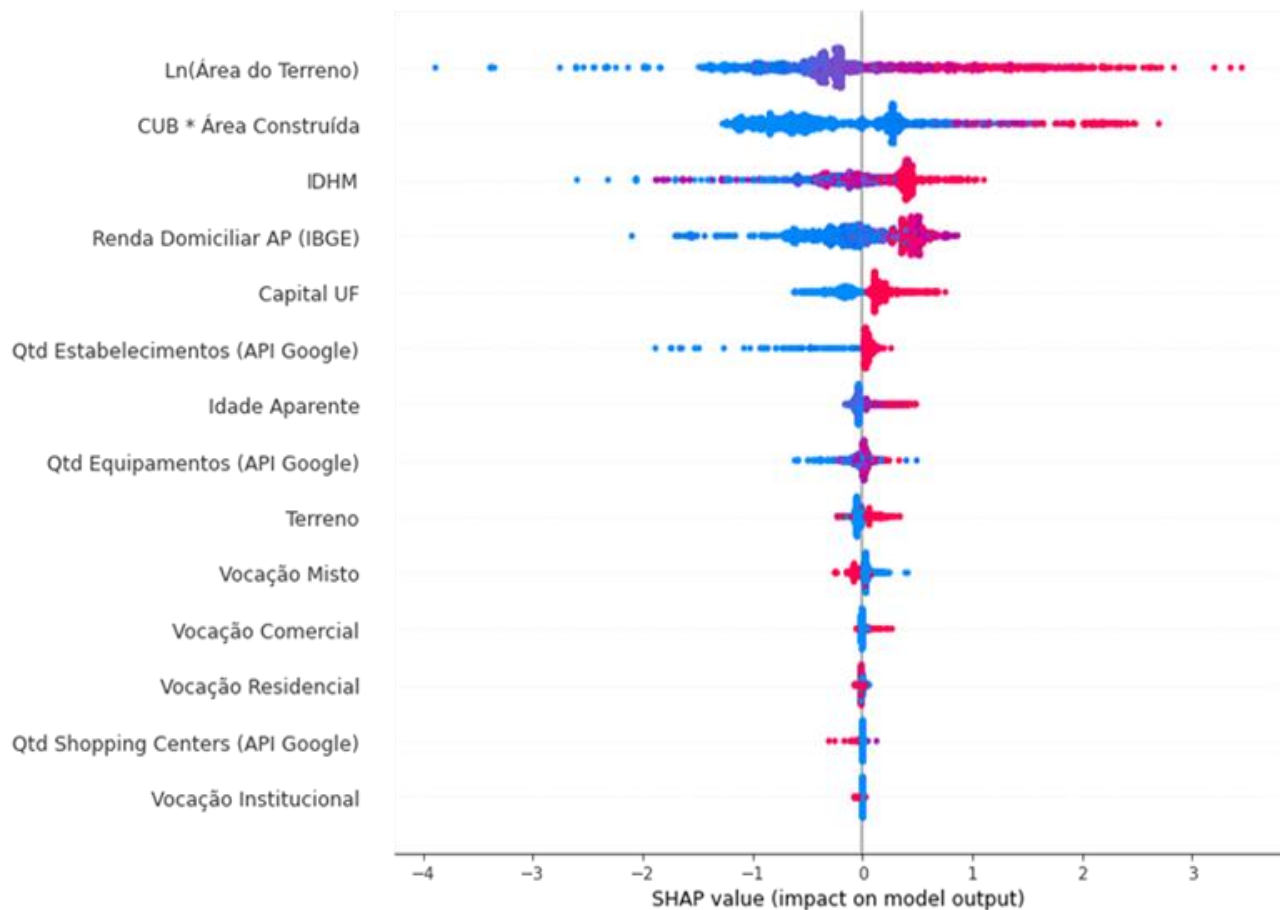


Cenário B

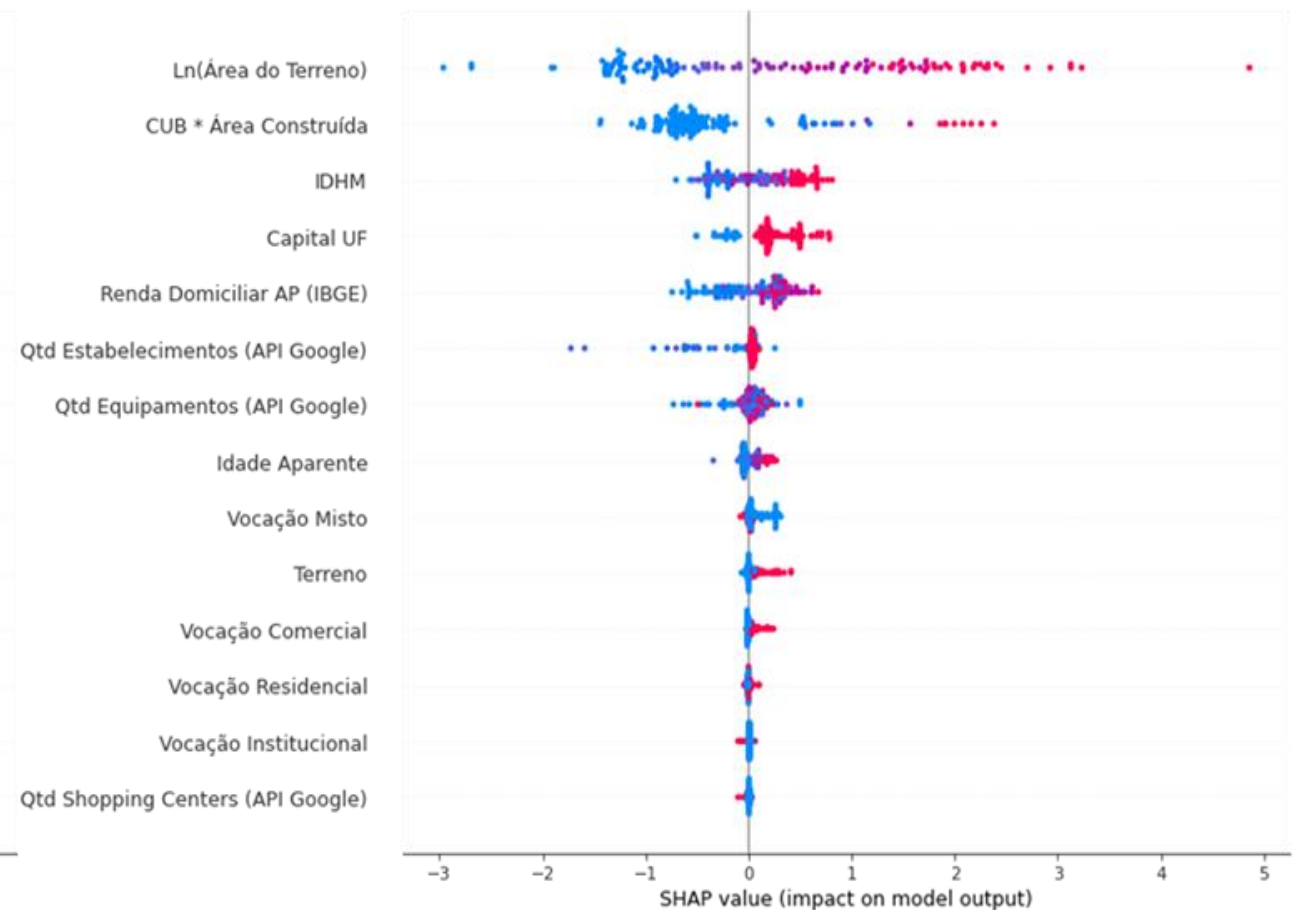
maior

RESULTADOS OBTIDOS

SHAP values (*XGBRegressor* não linear)



Cenário A



Cenário B

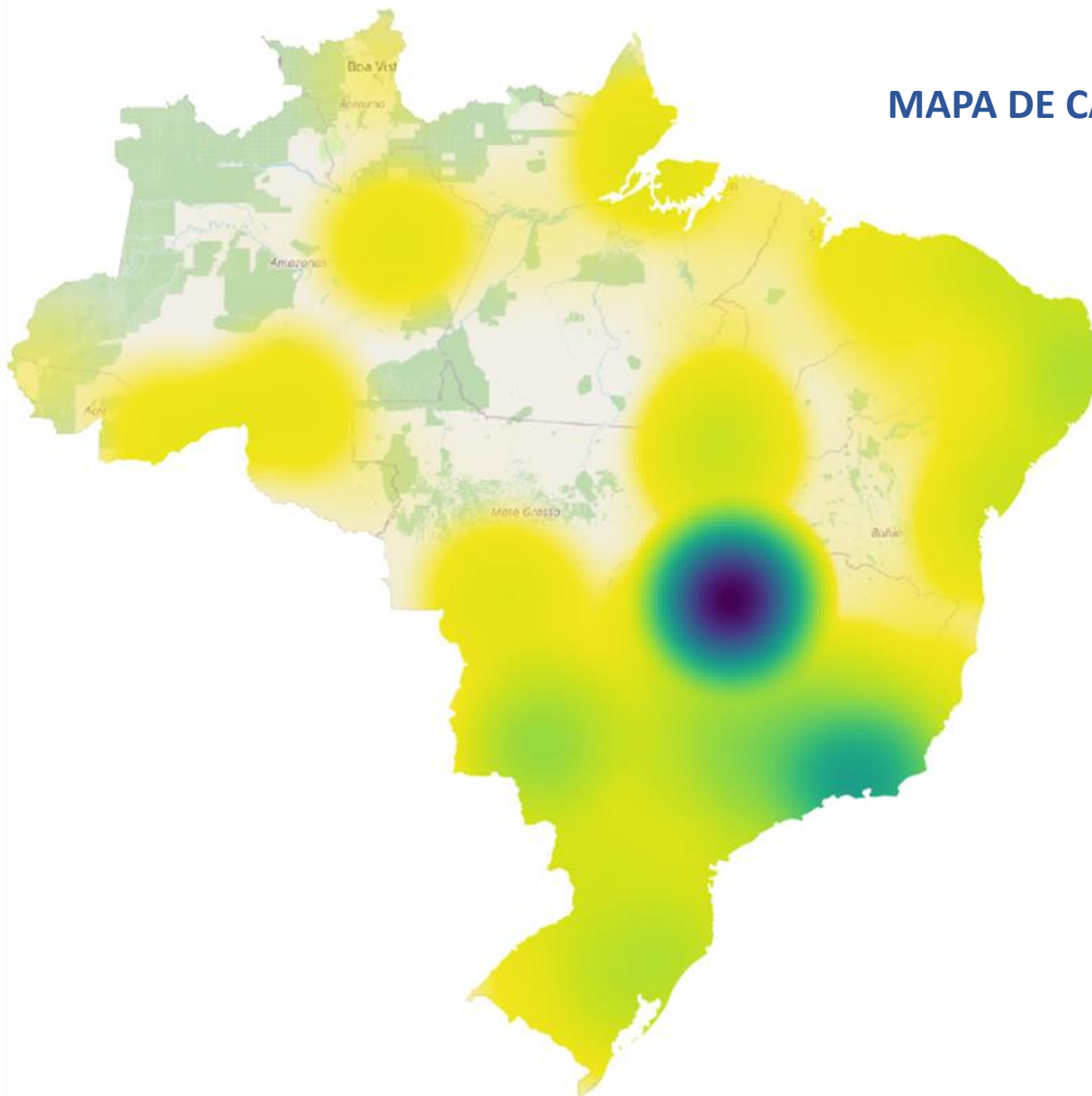
menor

maior

CONSIDERAÇÕES FINAIS

- Alguns modelos com **coeficiente de determinação superior a 57%**.
- **Os coeficientes de determinação** obtidos com os modelos não lineares foram superiores aos lineares.
- Há **limitações** de performance regional. As estimativas urbanas são mais robustas para regiões com maior densidade de dados anotados, como Brasília, Rio de Janeiro e São Paulo.
- Variáveis mais importantes acrescentadas nível a nível: U_1 (*Área do Terreno*), U_2 (*Idade Aparente*) e U_3 (*Renda Domiciliar AP*)

CONSIDERAÇÕES FINAIS



MAPA DE CALOR DE IMÓVEIS URBANOS

CONTRIBUIÇÕES

- Contribuição em vertente de **inovação** ao EB e à SPU: possibilidade de automatização de estimadores iniciais de valor dos imóveis antes de iniciar procedimentos técnicos segundo a NBR 14653.
- Redução do número de imóveis do EB e da SPU sem valor estimado nos sistemas de gerenciamento (EB: somente 1,2% dos imóveis com valor cadastrado conforme NBR 14653; SPU: 0,7%). Isso pode ajudar a projetar lastros para créditos internacionais à União.
- Utilização de **PLS-SEM** para selecionar atributos e de **API Google** para melhor entendimento do entorno dos imóveis.

TRABALHOS FUTUROS

Sugestões de trabalhos futuros:

- Utilização de informações do **Censo IBGE 2022**, quando disponíveis.
- Considerar variáveis de **demanda**.
- Utilização de **redes convolucionais** para identificação de padrões em imagens de imóveis e criação de novos atributos.

AGRADECIMENTOS

GRATOS!



José Nilo Alves de Sousa Neto



Juliana Miura



<https://github.com/joseniloneto/imoveismi>

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Kiely, Timothy e Nathaniel Bastian: *The spatially conscious machine learning model*. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13, 2019.
- [2] Hagenauer, Julian e Marco Helbich: *A geographically weighted artificial neural network*. International Journal of Geographical Information Science, 2021.
- [3] Alves Dantas, Rubens, André Matos Magalhães e José Raimundo de Oliveira Vergolino: Um modelo espacial de demanda habitacional para a cidade do Recife. Estudos Econômicos (São Paulo), 40(4), 2010.
- [4] Park, Byeonghwa e Jae Bae: *Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data*. Expert Systems with Applications, 42, 2015.
- [5] Tchunte, Dieudonne e Serge Nyawa: *Real estate price estimation in French cities using geocoding and machine learning*. Annals of Operations Research, 308(1-2,SI):571–608, 2022.
- [6] Furtado, Bernardo Alves: *Polycspace2: modeling markets and endogenous public policies*. Journal of Artificial Societies and Social Simulation, 25(1), 2022.

REFERÊNCIAS BIBLIOGRÁFICAS

- [7] Barros Antunes Campos, Rodger e Eduardo Almeida: Decomposição espacial nos preços residenciais no município de São Paulo. Estudos Econômicos (São Paulo), 48, 2018.
- [8] Anselin, Luc: *A local indicator of multivariate spatial association: Extending Geary's C*. Geographical Analysis, 51, 2019.
- [9] Lundberg, SM, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, J Himmelfarb, N Bansal e SI Lee: *From local explanations to global understanding with explainable AI for trees*. Nature Machine Intelligence, 2, 2020.
- [10] Chakraborty, Debaditya e Hazem Elzarka: *A novel construction cost prediction model using hybrid natural and light gradient boosting*. Advanced Engineering Informatics, 2020.
- [11] Dewan, Pranita, Raghu Ganti, Mudhakar Srivatsa e Sebastian Stein: *NN-SAR: A neural network approach for spatial autoregression*. Em 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (Per-Com Workshops), 2019.
- [12] Alves de Sousa Neto, José Nilo e Marcelo Ladeira: *Value Estimation of Properties Administered by the Brazilian Army Using Machine Learning and Spatial Components*. Em Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), Campinas/SP, 2022.

