Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Estimativa de Valores de Imóveis da União Administrados pelo Exército Brasileiro com Técnicas de Aprendizagem de Máquina e Componentes Espaciais

José Nilo Alves de Sousa Neto

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcelo Ladeira

Brasília
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

AA474e

Alves de Sousa Neto, José Nilo
Estimativa de Valores de Imóveis da União Administrados pelo Exército Brasileiro com Técnicas de Aprendizagem de Máquina e Componentes Espaciais / José Nilo Alves de Sousa Neto; orientador Marcelo Ladeira. -- Brasília, 2023.
93 p.

Dissertação(Mestrado Profissional em Computação Aplicada) - Universidade de Brasília, 2023.

1. aprendizagem de máquina. 2. imóveis. 3. componentes espaciais. I. Ladeira, Marcelo, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Estimativa de Valores de Imóveis da União Administrados pelo Exército Brasileiro com Técnicas de Aprendizagem de Máquina e Componentes Espaciais

José Nilo Alves de Sousa Neto

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcelo Ladeira (Orientador)
Presidente

Prof. Dr. Edison Ishikawa
Membro Interno

Prof. Dr. Bernardo Alves Furtado
Membro Externo

Prof. Dr. Ari Melo Mariano
Suplente

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 18 de abril de 2023

Dedicatória

Dedico este trabalho, com carinho e saudade, à minha avó Isete, por todo seu amor, pela sua amizade e por me fazer acreditar que todos os desafios são transponíveis.

Agradecimentos

Agradeço, primeiramente, a Deus, por me proteger e me dar saúde.

Agradeço imensamente à minha família, em nome dos meus pais, Nilo e Jevânia, exemplos de garra, amor e incentivo constantes.

Gratidão ao meu orientador, Prof. Dr. Marcelo Ladeira, pelos conhecimentos compartilhados, pela paciência e pelo valioso tempo a mim dispensado.

Destaco a importância das instituições Exército Brasileiro, Universidade de Brasília, Secretaria do Patrimônio da União e DataZAP+ na realização desta pesquisa. Este trabalho certamente não teria sido possível sem seu suporte, de capacitados profissionais delas integrantes.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

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 multiníveis capazes de construir estimativas do valor dos imóveis urbanos e rurais. Os modelos têm razoável nível de assertividade e abrangência geográfica nacional ao gerar predições de valores de mercado, associadas a graus de incerteza, 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, 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 valoração estimativa de propriedades a nível nacional, representam uma inovação deste trabalho em relação à literatura científica analisada.

Palavras-chave: aprendizagem de máquina, 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 multilevel predictive models capable of building estimates of urban and rural real estate values. The models have a reasonable level of assertiveness and national geographic coverage when generating estimated market values, associated with degrees of uncertainty, 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, 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 value modeling at a national level represent an innovation of this work in relation to the analyzed scientific literature.

Keywords: machine learning, real estate, spatial components

Sumário

1 Introdução **1**

1.1 Contextualização e Problema 1
1.2 Objetivos 4
1.3 Contribuição Esperada da Pesquisa 6
1.4 Estrutura do Documento 6

2 Fundamentação Teórica **7**

2.1 Valoração de Imóveis 7
2.2 Imóveis Urbanos 7
2.2.1 Índices da Construção Civil 8
2.2.2 Censo IBGE 2010 9
2.2.3 API Google Places 10
2.3 Imóveis Rurais 10
2.3.1 Classes de Capacidade de Uso dos Solos e Nota Agronômica 11
2.3.2 Potencialidade Agrícola Natural dos Solos 12
2.4 Modelos de Regressão Múltipla Linear e Espacial 16
2.5 Modelos de Aprendizagem de Máquina 16
2.5.1 SGDRegressor 17
2.5.2 MLPRegressor 18
2.5.3 XGBoost 20
2.6 PLS-SEM 22
2.7 Interpretabilidade com valor de Shapley 23
2.8 Considerações Finais 24

3 Trabalhos Relacionados **25**

3.1 Discussão dos Trabalhos Relacionados 27

4 Solução Proposta **29**

4.1 Modelos para Construção de Estimativas de Valor de Mercado de Imóveis da União 29

4.1.1 Modelos Imóveis Urbanos 29
4.1.2 Modelos Imóveis Rurais 30
4.2 Metodologia 30
4.2.1 Extração de Dados 30
4.2.2 Atributos Analisados 35
4.2.3 Análise Gráfica de Atributos 41
4.2.4 Abordagens 43
4.2.5 Modelagens 43
4.2.6 Seleção de Atributos 45
4.2.7 Tratamento de Variáveis 50
4.2.8 Tratamento de Instâncias Coletadas 50
4.2.9 Métricas de Avaliação 51
4.2.10 Intervalo de Confiança 52

5 Experimentos e Resultados **53**
5.1 Bases de Dados e Cenários de Aplicação 53
5.2 Configuração dos Experimentos 55
5.3 Resultados Obtidos 55
5.3.1 Imóveis Urbanos 55
5.3.2 Imóveis Rurais 59
5.4 Interpretabilidade dos Resultados 62
5.5 Discussão dos Resultados 64

6 Conclusões e Trabalhos Futuros **70**
6.1 Contribuições 70
6.2 Trabalhos Futuros 71

Referências **73**

Lista de Figuras

1.1 Diagrama de processo CRISP-DM. 2
1.2 Fluxograma de execução e análise de laudos de avaliação no Exército Brasileiro (EB). 3

2.1 Mapa de classes de declividade na região do município de Maranguape-CE produzido no *software QGIS 3.16.3 Hannover*. 12
2.2 Imagem satelital sobreposta por mapa de classes de declividade também na região de Maranguape-CE produzido no *software QGIS 3.16.3 Hannover*. 13
2.3 Diagrama das etapas de análise automatizada, validação e classificação da potencialidade agrícola natural das terras. 14
2.4 Mapa de potencialidade agrícola natural das terras. 15

3.1 Coeficientes de determinação obtidos na predição de preços de casas na Áustria. 27

4.1 Regiões rurais publicadas pelo IBGE em 2015 e para as quais há VTN associado à cobrança de ITR pela RFB. 31
4.2 AP do IBGE (em azul) sobrepostas pelos centroides de instâncias do EB e da SPU. 33
4.3 AP do IBGE (transparentes com contorno em preto) sobrepostas pelos centroides de instâncias do EB e da SPU na região do Distrito Federal. . . 33
4.4 Polígono de imóvel rural da base de dados avaliada com suas parcelas de potencialidade agrícola natural de solo (em cores verde, amarelo e laranja), cursos d'água da ANA (linha contínua azul) e rodovia do DNIT (linha tracejada preta). 34
4.5 *Boxplot* para *Área do Terreno* (à esquerda) e *Ln(Área do Terreno)* (à direita), após eliminação de pontos influenciantes. 41
4.6 *Boxplot* para *Valor Total Atualizado* (à esquerda) e *Ln(Valor Total Atualizado)* (à direita), após eliminação de pontos influenciantes. 42

4.7 Gráficos de dispersão bivariados. Os pontos em cor laranja representam os imóveis sem construções; já os azuis, representam as propriedades com benfeitorias construtivas. 42
4.8 Grafo de conectividades gerado a partir de definição de largura de banda. . 45
4.9 Modelo conceitual U_1 PLS-SEM. 47
4.10 Modelo conceitual U_2 PLS-SEM. 47
4.11 Modelo conceitual U_3 PLS-SEM. 48
4.12 Distâncias de Cook calculadas antes (à esquerda) e depois (à direita) da remoção de instâncias influenciantes urbanas. 51
4.13 Distâncias de Cook calculadas antes (à esquerda) e depois (à direita) da remoção de instâncias influenciantes rurais. 51

5.1 Distribuição das 4595 instâncias urbanas por UF com suas respectivas quantidades absolutas. 54
5.2 Mapa de densidade dos imóveis que compõem a amostra urbana (à esquerda) e rural (à direita). 54
5.3 Gráfico de calor de correlações entre as variáveis explicativas quantitativas urbanas já tratadas e selecionadas. 56
5.4 Gráfico de calor correlações entre as variáveis explicativas rurais já tratadas e selecionadas. 57
5.5 Histograma da variável dependente *Ln(Valor Total Atualizado)* nos conjuntos de treinamento e de validação (à esquerda) e de teste (à direita). . . 58
5.6 Gráficos de dispersão de *ln* de valores observados (ordenadas) versus *ln* de valores projetados pelos modelos U_3 (abscissas). Implementado com os algoritmos *SGDRegressor* (à esquerda), *ANN MLPRegressor* (ao centro) e *XGBRegressor* (à direita). 60
5.7 Valores de Shapley calculados para o modelo U_3 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *SGDRegressor*. . 65
5.8 Valores de Shapley calculados para o modelo U_3 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *ANN MLPRegressor*. 65
5.9 Valores de Shapley calculados para o modelo U_3 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *XGBRegressor*. . 66
5.10 Valores de Shapley calculados para o modelo U_1 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *ANN MLPRegressor*. 66
5.11 Valores de Shapley calculados para o modelo U_2 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *ANN MLPRegressor*. 67
5.12 Gráfico comparativo entre os modelos urbanos U_3 no Cenário A considerando as métricas de avaliação R^2 e $RMSE$. 68

5.13 Zonas de amortecimento criadas para os modelos específicos rurais na região de Sinop-MT. 69

Lista de Tabelas

2.1 Comparação entre os índices de construção da construção civil nacional. . . 9
2.2 Classificação do relevo conforme Manual de Obtenção de Terras do INCRA de 2006. 12

3.1 Comparação com os trabalhos relacionados. 28

4.1 Informações sobre as instâncias com valor conhecido. 32
4.2 Informações sobre as instâncias urbanas com valor conhecido coletadas após processo de recuperação manual. 32
4.3 Informações sobre as instâncias rurais avaliadas coletadas e com polígono georreferenciado disponível. 34
4.4 Vida útil dos imóveis. 37
4.5 Intensidade de procura relativa pelo termo "imóvel" em ferramentas de busca *Google* por UF e para o ano de 2022. 39
4.6 Ocorrência de atributos nas modelagens finais de imóveis urbanos. 48
4.7 Ocorrência dos atributos selecionados nos diferentes níveis de modelos urbanos. 49
4.8 Ocorrência de atributos nas modelagens finais de imóveis rurais. 49
4.9 Ocorrência dos atributos selecionados nos diferentes níveis de modelos rurais. 49

5.1 Parâmetros utilizados nos modelos de aprendizagem de máquina. 55
5.2 Resultados das métricas de avaliação dos modelos urbanos básicos para os dois cenários considerados. 58
5.3 Resultados das métricas de avaliação dos modelos urbanos intermediários para os dois cenários considerados. 59
5.4 Resultados das métricas de avaliação dos modelos urbanos específicos para os dois cenários considerados. 59
5.5 Semiamplitudes dos intervalos de confiança 90% calculados para os modelos urbanos multiníveis. 60

5.6 Diferenças relativas entre entre os valores observados dos imóveis y_i e valores $\hat{y}_i$ projetados pelos modelos, calculados como alíquotas de $\hat{y}_i$. Valores positivos indicam valores observados maiores que predições. 61
5.7 Valores observados dos imóveis y_i e valores $\hat{y}_i$ estimados pelo modelo OLS R_2. Encontram-se tabulados as medidas de tendência central dos valores projetados, os mínimos e os máximos calculados. 62
5.8 Coeficientes de determinação (R^2) e raízes dos erros quadráticos médios ($RMSE$) calculados para os modelos lineares rurais. 62
5.9 Semiamplitudes dos intervalos de confiança 90% calculados para os modelos rurais multiníveis. 63
5.10 Coeficientes e *p-valores* associados calculados para os modelos específicos convencionais U_3. RE como abreviação de Regressão Espacial. 63
5.11 Coeficientes e *p-valores* associados calculados para os modelos específicos R_2. RE como abreviação de Regressão Espacial. 64

Lista de Abreviaturas e Siglas

ANA Agência Nacional de Águas e Saneamento Básico.

ANN *Artificial Neural Network*.

AP Área de Ponderação (IBGE).

API *Application Programming Interface*.

BDI Benefícios e Despesas Indiretas.

CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

Cmdo Mil A *Comando Militar de Área*.

CRISP-DM *CRoss-Industry Standard Process for Data Mining*.

CUB Custo Unitário Básico.

DEC Departamento de Engenharia e Construção.

DNIT Departamento Nacional de Infraestrutura de Transportes.

DPIMA Diretoria de Patrimônio Imobiliário e Meio Ambiente.

EB Exército Brasileiro.

Embrapa Empresa Brasileira de Pesquisa Agropecuária.

FIPE Fundação Instituto de Pesquisas Econômicas.

GCP Google Cloud Platform.

Gpt E Grupamento de Engenharia.

IBGE Instituto Brasileiro de Geografia e Estatística.

IC Intervalo de Confiança.

IDHM Índice de Desenvolvimento Humano Municipal.

IN Instrução Normativa.

INCRA Instituto Nacional de Colonização e Reforma Agrária.

Ipea Instituto de Pesquisa Econômica Aplicada.

ITR Imposto sobre a Propriedade Territorial Rural.

IVS índice de Vulnerabilidade Social.

ln Logaritmo Neperiano.

MSE Erro Quadrático Médio.

OLS *Ordinary Least Squares*.

OM Organização Militar.

OPUS Sistema Unificado do Processo de Obras.

PIB Produto Interno Bruto.

PLS-SEM *Partial Least Squares Structural Equation Modeling*.

RFB *Receita Federal do Brasil*.

RM Região Militar.

RMSE Raiz do Erro Quadrático Médio.

SGD *Stochastic Gradient Descent*.

SHAP *SHapley Additive exPlanations*.

SIGPIMA Sistema Informatizado de Gestão do Patrimônio Imobiliário e Meio Ambiente.

Sinduscon Sindicato da Indústria da Construção Civil.

SPIUnet Sistema de Gerenciamento dos Imóveis de Uso Especial da União.

SPU Secretaria do Patrimônio da União.

UF Unidade da Federação.

VLT Veículo Leve sobre Trilhos.

VTN Valor de Terra Nua.

Capítulo 1

Introdução

Neste capítulo, o problema é definido, o tema é justificado e a contribuição esperada é descrita.

1.1 Contextualização e Problema

Diferentemente de ativos financeiros de precificação mais simples e em tempo quase real, 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.

Além da importância social de se valorar de maneira justa os ativos mencionados, no Brasil, há recomendação do Governo Federal quanto à necessidade de mensuração do valor do patrimônio imobiliário da União, no intuito de discriminar garantias, aumentar as margens brasileiras para obtenção de créditos internacionais e promover redução do risco-país. Muitos imóveis, dos mais de 700 mil de propriedade do Estado, estão cadastrados no Sistema de Gerenciamento dos Imóveis de Uso Especial da União (SPIUnet) com valores que não traduzem sua realidade atual.

O Exército Brasileiro (EB), por sua vez, possui mais de 21 mil parcelas imobiliárias sob sua jurisdição 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. Hoje, menos de 2% dos imóveis têm valor de avaliação devidamente cadastrado no sistema de gerenciamento.

No âmbito do EB, a Diretoria de Patrimônio Imobiliário e Meio Ambiente (DPIMA) [1] representa a instância técnico-normativa responsável por analisar e aprovar as avaliações

[1] `http://www.dpima.eb.mil.br`

imobiliárias dos imóveis da União administrados pelo EB e daqueles de interesse da Força Terrestre.

Conforme a norma ABNT NBR 14653, o Método Comparativo Direto de Dados de Mercado deve ser preferencialmente utilizado. Segundo o referido método, inferem-se os valores de mercado dos imóveis avaliandos 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 significante nos modelos construídos e que os imóveis ofertados ou recentemente transacionados nas adjacências daquele em análise têm influência sobre seu valor.

Na realização deste projeto de pesquisa, optou-se por seguir o processo *CRoss-Industry Standard Process for Data Mining*, conhecido como CRISP-DM e detalhado na Figura 1.1. Foram percorridas as fases de entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem e avaliação. A etapa de produção e implantação não foi abordada neste trabalho.

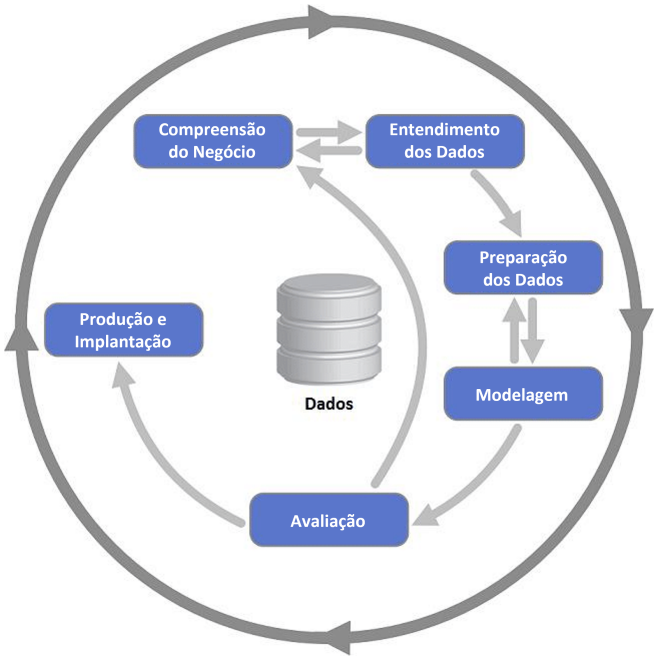

Figura 1.1: Diagrama de processo CRISP-DM.

Na fase de compreensão do negócio, identificou-se a carência em estimar os valores dos imóveis da União antes de iniciar o oneroso processo de execução e análise de laudos de avaliação e a necessidade de se garantir interpretabilidade aos modelos construídos.

No âmbito do EB, as avaliações imobiliárias costumam ser solicitadas pelos Comandos Militares de Área (Cmdo Mil A), motivados por algum interesse institucional, aos Grupamentos de Engenharia (Gpt E) ou às Regiões Militares (RM), os quais designam

arquitetos ou engenheiros avaliadores para executá-las. Após sua execução, são analisadas pelo corpo técnico dos Gpt E ou das RM. Após a devida aprovação do respectivo Gpt E ou RM, a peça técnica é enviada para a DPIMA para análise, podendo ser homologada ou reprovada. Se for constatada a necessidade de correções em qualquer um dos processos de verificação técnica, o laudo retorna para a Organização Militar (OM) onde foi executado. A análise na DPIMA é de caráter puramente técnico e verifica se o trabalho respeita as diretrizes da NBR 14653 e da Instrução Normativa (IN) nº 67/2022 da Secretaria do Patrimônio da União (SPU), que regula a avaliação de imóveis da União. Esse processo está representado na Figura 1.2.

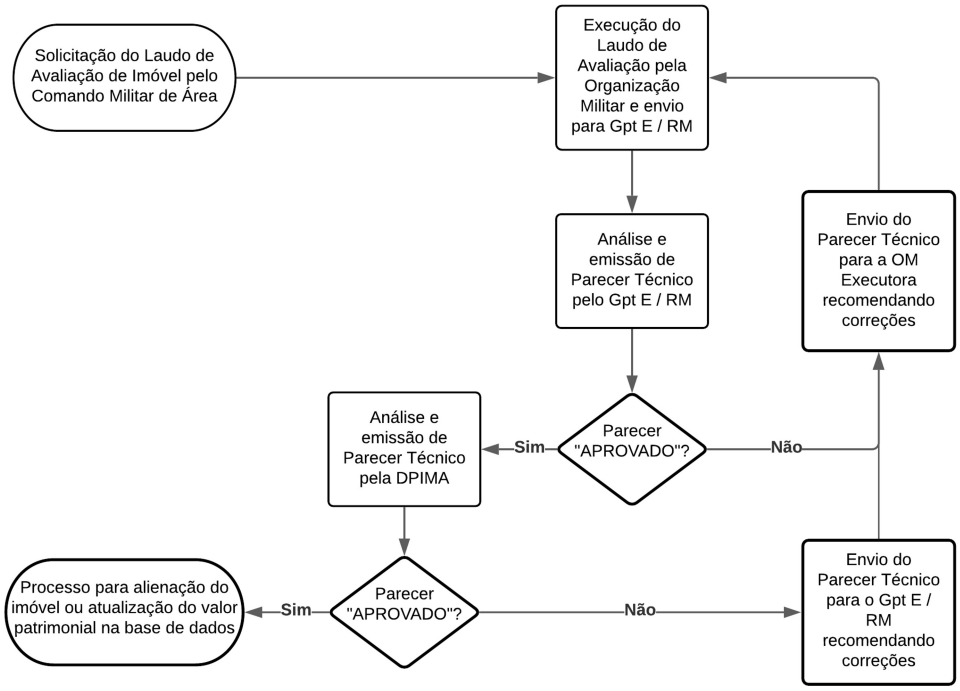

Figura 1.2: Fluxograma de execução e análise de laudos de avaliação no Exército Brasileiro (EB).

Esse trâmite é bastaste oneroso em termos de utilização de mão de obra especializada, tempo de execução e correspondência e recursos financeiros e pode ser iniciado sem se ter uma estimativa razoável do valor de mercado do imóvel de interesse. Por vezes, o processo ocorre e, pelo fato de o valor calculado pelo avaliador e homologado pela DPIMA estar desalinhado àquele inicialmente imaginado pelo Cmdo Mil A, a linha de ação de alienar ou incorporar determinado bem imóvel acaba sendo abandonada.

Em outubro de 2022, o EB administrava aproximadamente 21420 unidades ou parcelas imobiliárias, das quais 95,9% eram urbanas e 4,1% eram rurais.

O valor de mercado dos imóveis pode ser estimado de forma endógena, com a construção de modelos multiagentes, como proposto por Williams [1], ou de forma hedônica,

conforme percepção da utilidade de atributos implícitos de Rosen [2], havendo entendimento de que a variância do valor imobiliário pode ser explicada por meio de variáveis intrínsecas ou extrínsecas aos bens de caráter imobilizado.

Diferentes modelos preditivos foram propostos para valorar imóveis com relativo grau de assertividade. Em [3], Dantas constrói algoritmos lineares, com e sem componentes espaciais, de demanda habitacional aplicados à cidade do Recife e verifica-se que a autocorrelação espacial exerce um importante papel na identificação do equilíbrio natural existente no mercado imobiliário.

Já em [4] e em [5], são aplicadas técnicas de aprendizagem de máquina a mercados imobiliários nos Estados Unidos da América e na França visando a predizer valores transacionais. O aumento de complexidade dos modelos eleva seu poder de explicação, mas acaba por reduzir sua interpretabilidade.

No trabalho [6], Furtado constrói um modelo multiagentes a fim de avaliar, de maneira endógena, políticas públicas e ressalta que o mercado imobiliário, particularmente o residencial, sofre influência de ciclos econômicos, taxas de juros e liquidez, decisões e mudanças de famílias, interesse de investidores locais e estrangeiros, regulamentação quanto a uso e ocupação do solo, dinâmicas construtivas, localização, incluindo vizinhança, e características intrínsecas aos próprios imóveis.

Na pesquisa bibliográfica realizada, foi observada a inexistência de um modelo hedônico de predição de valores de imóveis de abrangência nacional, nível geográfico Brasil, com razoável poder de explicação, interpretável e que contemple, dentro de seu domínio de análise, instâncias de bens avaliados à luz da NBR 14653 e da IN SPU em que constam as diretrizes de avaliações imobiliárias.

1.2 Objetivos

O principal objetivo é desenvolver modelos capazes de construir estimativas de valores de mercado dos imóveis da União administrados pelo Exército Brasileiro com três níveis de especificidade para imóveis urbanos e com dois níveis de especificidade para imóveis rurais a fim de permitir que o processo ilustrado na Figura 1.2, se acionado, seja iniciado com essa informação disponível previamente. Dessa forma, a autoridade solicitante somente formalizaria seu pedido de laudo de avaliação se a estimativa fosse coerente com o plano que pretende executar.

Para alcançar tal objetivo, foram construídos modelos com a aplicação de técnicas de aprendizagem de máquina de caráter preditivo com e sem adição de componentes espaciais extraídas de bases de dados do Instituto Brasileiro de Geografia e Estatistica (IBGE) e

com uso de *Application Programming Interface* (API) *Google Places Nearby Search*, no caso de imóveis urbanos.

Os resultados obtidos com uso de aprendizagem de máquina foram comparados àqueles alcançados por meio de regressões mais simples, tais como a linear múltipla e a espacial. Um ponto que distinguiu as técnicas foi a necessidade de adaptabilidade dos algoritmos a imóveis inseridos em cenários por vezes não constantes nas instâncias coletadas sem perda de poder de predição.

Já para os imóveis rurais, em decorrência do reduzido número de parcelas avaliadas, cujo valor fora calculado e aprovado pelo EB ou pela SPU, e com polígonos georreferenciados associados, o problema foi abordado com uso de tratamento espacial, regressão linear múltipla e estatística espacial. A discussão de resultados se deu mais qualitativamente que quantitativamente.

Os três níveis de modelo destinados à valoração estimativa de imóveis urbanos da União podem ser descritos sucintamente da seguinte forma:

- Nível 1 Urbano (U_1): básico - baseado em variáveis genéricas (tipo de uso dos imóveis, área de terreno, área construída, Custo Unitário Básico da Construção associado, IDHM, etc.);
- Nível 2 Urbano (U_2): intermediário - baseado nas variáveis genéricas de U_1 e em variáveis intrínsecas mais próprias aos imóveis (como idade aparente e vida útil); e
- Nível 3 Urbano (U_3): específico - baseado nas variáveis genéricas, nas variáveis intrínsecas aos imóveis, em componentes espaciais, em indicadores socioeconômicos associados a áreas de ponderação (AP) de setores censitários do IBGE e em quantitativos de equipamentos urbanos e de estabelecimentos existentes no entorno imediato do imóvel obtidos por meio de API *Google Places*.

Já os dois níveis de modelo destinados à valoração estimativa de imóveis rurais da União podem ser resumidos da maneira a seguir:

- Nível 1 Rural (R_1): básico - baseado em variáveis de caráter físico (área, presença de cursos d'água, acesso ao imóvel, potencialidade agrícola natural da terra); e
- Nível 2 Rural (R_2): específico - baseado nas variáveis de R_1 acrescidas de atributos indicativos do tipo de uso que se dá à terra, como agricultura, pecuária, silvicultura, preservação da vegetação nativa, dentre outros, e de referência de qualidade de vida no município que contém o imóvel rural dentro de seus limites geográficos.

São objetivos específicos: obter coeficiente de determinação igual ou superior a 57%, conforme NBR 14653, garantir interpretabilidade aos modelos desenvolvidos e testar seu desempenho em cenários envolvendo instâncias do EB e da SPU.

1.3 Contribuição Esperada da Pesquisa

Prover o EB de modelos para construir estimativas dos valores dos imóveis por ele administrados com abrangência nacional, produto que tal instituição não possui e com potencial de gerar economia de recursos laborais e financeiros. Sob a ótica do EB, portanto, verifica-se que a contribuição principal deste trabalho de pesquisa deverá se dar no âmbito da inovação.

1.4 Estrutura do Documento

Figuras e tabelas de terceiros possuem indicação dos autores no texto. Aquelas sem tal indicação foram elaboradas durante esta pesquisa e são de nossa autoria.

O restante deste documento está organizado de maneira construtiva e linear, a fim de proporcionar fluidez de entendimento da pesquisa aos leitores.

No Capítulo 2, é apresentada a fundamentação teórica, visando a oferecer a base necessária para a compreensão do assunto abordado e seus desafios.

Posteriormente, no Capítulo 3, citam-se trabalhos similares, com aplicações de algoritmos a domínios de valoração imobiliária, e faz-se uma breve comparação entre eles.

Já no Capítulo 4, apresenta-se a solução proposta e a metodologia a ela associada, incluindo diferentes abordagens e critérios de avaliação dos modelos.

Complementarmente, no Capítulo 5, são apresentados os experimentos realizados, os resultados obtidos e análises acerca deles.

Por fim, no Capítulo 6, destacam-se as conclusões, as principais contribuições alcançadas e algumas sugestões de trabalhos futuros, bem como sucintas considerações finais.

Capítulo 2

Fundamentação Teórica

Inicialmente, é apresentado um embasamento téorico sobre avaliações imobiliárias, imóveis urbanos e rurais, estatística espacial, modelos de aprendizagem de máquina lineares e não lineares e interpretabilidade à luz do valor de Shapley (Teoria dos Jogos). Subsequentemente, propõe-se a construção de uma modelagem multiníveis visando à estimativa de valores de imóveis da União. Por fim, discorre-se acerca da integração entre os diferentes conceitos.

2.1 Valoração de Imóveis

Segundo Dantas [3] e Barros Antunes Campos e Almeida [7], considerando modelo tradicional de preços hedônicos definido inicialmente por Rosen [2], o valor de mercado das unidades imobiliárias residenciais pode ser explicado pela Equação 2.1.

$$P = f(E, L, T, \beta) + \varepsilon \tag{2.1}$$

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 são 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.

2.2 Imóveis Urbanos

Nas Subseções 2.2.1, 2.2.2 e 2.2.3 a seguir, serão abordados conceitos que auxiliam na escolha de variáveis explicativas com potencial de explicação da variância do valor de

imóveis urbanos.

2.2.1 Índices da Construção Civil

Os bens imóveis podem ser, via de regra, desmembrados em terreno e benfeitorias. O terreno seria o correspondente à terra nua com limites geográficos bem definidos. A benfeitoria pode ser entendida como resultado de obra ou serviço realizado em um bem e que não pode ser retirada sem destruição, fratura ou dano.

2.2.1.1 Custo Unitário Básico (CUB)

O CUB é divulgado mensalmente e para cada projeto-padrão disponível na ABNT NBR 12.721:2006 pelos Sindicatos da Construção Civil (Sinduscon) de cada unidade da federação (de cada UF). Duas UF atingem menor granularidade de divulgação dos resultados: PR e MG.

De acordo com o item 3.9 da NBR 12.721:2006, o conceito de Custo Unitário Básico é o seguinte:

"*Custo por metro quadrado de construção do projeto-padrão considerado, calculado de acordo com a metodologia estabelecida em 8.3, pelos Sindicatos da Indústria da Construção Civil, em atendimento ao disposto no artigo 54 da Lei nº 4.591/64 e que serve de base para avaliação de parte dos custos de construção das edificações.*"

O CUB/m² representa o custo parcial da obra, isto é, não leva em conta os demais custos adicionais, que, por sua vez, podem ser parcialmente estimados por uma alíquota de benefícios e despesas indiretas (BDI), que contempla lucro, riscos, seguros, garantias, despesas financeiras, dentre outros itens.

2.2.1.2 Sistema Nacional de Pesquisa de Custos e Índices da Construção Civil (SINAPI)

Conforme o IBGE [1], "o SINAPI tem por objetivo a produção de séries mensais de custos e índices para o setor habitacional, e de séries mensais de salários medianos de mão de obra e preços medianos de materiais, máquinas e equipamentos e serviços da construção para os setores de saneamento básico, infraestrutura e habitação. O Sistema é uma produção conjunta do IBGE e da Caixa Econômica Federal (CEF)".

[1] `https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9270-sistema-nacional-de-pesquisa-de-custos-e-indices-da-construcao-civil`

2.2.1.3 Custo Unitário Pini de Edificações (CUPE)

De acordo com o sítio eletrônico PINI[2], "o cálculo mensal do CUPE ocorre por meio da atualização do orçamento global do projeto-padrão de cada tipo de obra. Ou seja, mensalmente são atualizados os preços de todos os insumos que participam do cálculo, entre materiais, mão de obra, equipamentos. As estimativas são levantadas com base nos orçamentos de cada obra nas principais capitais brasileiras".

2.2.1.4 Comparativo de Índices da Construção Civil

A Tabela 2.1 contém uma comparação bastante sumária entre os índices analisados. Optou-se pela utilização do CUB nos modelos pelo fato de ele ser o índice preferencial nas estimativas dos custos construtivos, segundo a NBR 14653, de ele ser disponibilizado mensalmente pelos Sinduscon regionais e de o nível de granularidade espacial dos três índices analisados ser praticamente o mesmo.

Tabela 2.1: Comparação entre os índices de construção da construção civil nacional.

Índice	**Granularidade Espacial**	**Valor Médio (R$/m²)**	**Nível de Detalhamento**	**Periodicidade de Divulgação**	**Responsável pela Divulgação**
CUB	estadual	intermediário	médio	mensal	SINDUSCON
SINAPI	estadual	mais baixo	alto	mensal	IBGE / CEF
CUPE	capitais	mais alto	médio	mensal	PINI

2.2.2 Censo IBGE 2010

Os dados do IBGE utilizados são os do último censo realizado, ou seja, o Censo 2010. As informações do questionário da amostra do Censo 2010 são disponibilizadas tendo como unidade territorial as áreas de ponderação (AP), definidas pelo IBGE como o agrupamento de vários setores censitários.

Conforme endereço eletrônico do Centro de Estudos da Metrópole (CEM)[3], sediado na Universidade de São Paulo (USP), as AP são unidades geográficas, definidas apenas para os Censos de 2000 e 2010, constituídas do agrupamento mutuamente exclusivo de setores censitários contíguos. São construídas pelo IBGE para que seja possível aplicar procedimentos de calibração dos pesos amostrais.

As AP são a menor unidade geográfica com possibilidade de obter representatividade estatística a partir das amostras dos censos demográficos.

[2] `https://tcpoweb.pini.com.br`

[3] `https://centrodametropole.fflch.usp.br/pt-br`

O número de domicílios e de indivíduos habitando em uma AP, consequentemente, não pode ser muito reduzido, sob pena de perda de precisão das estimativas de suas características. Por essa razão, principalmente nas regiões menos povoadas, as AP acabam ocupando uma larga extensão territorial. Para o Censo de 2010, o IBGE estabeleceu que uma AP deveria ter, no mínimo, 400 domicílios ocupados na amostra. Em geral, AP são regiões dentro de municípios e, portanto, permitem fazer análises intramunicipais. No entanto, quando os municípios não possuem o mínimo de domicílios estabelecido, o próprio município é considerado, por inteiro, como uma única AP.

Os resultados da amostra do Censo 2010 por AP foram extraídos do endereço eletrônico de *downloads* do IBGE[4].

2.2.3 API Google Places

A *Google Places Nearby Search* é uma API da plataforma de serviços em nuvem da Google (GCP) que permite a desenvolvedores obter informações sobre lugares pré-cadastrados próximos a pares de coordenadas especificados por eles. A API retorna os resultados da consulta em formato JSON.

Neste trabalho, os retornos em formato JSON foram convertidos no tipo dicionário de estrutura de dados, programado em linguagem *Python* em ambiente *Google Colaboratory*. O entorno dos imóveis urbanos foi melhor compreendido por meio de consultas sobre a existência de diversas tipologias de equipamentos urbanos e de estabelecimentos comerciais.

Com a criação de uma chave API na *Google Cloud Platform*, são creditados 200 dólares americanos mensalmente à conta do usuário. Cada requisição da API *Google Places Nearby Search* consome US$ 0,040 e retorna informações divididas sobre estabelecimentos em três *stock keeping units* (SKU): básica (inclui tipologia), ambiência e contato. Respeitando o limite de gratuidade, é possível realizar 5 mil consultas mensais.

2.3 Imóveis Rurais

Na avaliação de imóveis rurais, fatores como a situação do imóvel (localização e acesso), suas dimensões (área e perímetro), sua forma (conformação do polígono), a declividade do terreno (relevo), os recursos naturais disponíveis (principalmente, água), os tipos de solo presentes e seu uso (lavoura, pastagem, silvicultura, preservação da fauna e da flora, dentre outros) são de suma importância à sua valoração e se relacionam diretamente com a determinação da nota agronômica (NA) a ele associada.

[4] `https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html`

Para obtenção desses parâmetros de maneira fiel à realidade e com rigor técnico, faz-se necessária a participação de profissionais capacitados e habilitados, tais como engenheiros agrônomos e florestais. Em apoio a tais profissionais, a geoinformação consiste em importante ferramenta. Tentou-se, neste trabalho, utilizar dados disponíveis publicamente e nas bases do EB e da SPU a fim de se estimar em primeiro nível os valores de imóveis rurais.

Nas Subseções 2.3.1 e 2.3.2 a seguir, serão abordados alguns conceitos e fontes de dados importantes às abordagens estimativas propostas no Capítulo 4.

2.3.1 Classes de Capacidade de Uso dos Solos e Nota Agronômica

O sistema de capacidade de uso constitui uma classificação técnica que envolve um grupamento qualitativo de condições ligadas aos atributos das terras sem priorizar localização e características econômicas. Ele ajuda a avaliar a capacidade do solo para suportar diferentes usos, como agricultura, pastagem e silvicultura.

Nesse sistema, características físicas (tipo, profundidade, textura, drenagem e erosão) e químicas (composição) de solo e topográficas são sintetizadas visando a obter grupamentos de terras similares, fundamentados na máxima capacidade de uso para agricultura sem risco de degradação.

Informações de tipo de solo, alinhadas ao sistema brasileiro de classificação, podem ser extraídas do sítio eletrônico da Empresa Brasileira de Pesquisa Agropecuária (Embrapa)[5] em formato georreferenciado *shapefile*. Para se aferir outros aspectos, como profundidade e erosão, um profissional habilitado teria que visitar cada imóvel; o que fugiria à proposta deste trabalho.

Já informações de relevo podem ser obtidas do projeto Topodata do Instituto Nacional de Pesquisas Espaciais (INPE). No referido projeto, um modelo digital de elevação (MDE) em definição 30 por 30 metros e de cobertura nacional foi gerado a partir de processamento e refinamento de dados da missão *Shuttle Radar Topography Mission* (SRTM) da *National Aeronautics and Space Administration* (NASA).

Com as imagens em formato GeoTiff (.tif) do MDE disponíveis no endereço eletrônico do Topodata[6], é possível calcular as declividades de interesse e dividi-las nas classes utilizadas pelo Instituto Nacional de Colonização e Reforma Agrária (INCRA), conforme constante na Tabela 2.2. Em *softwares* de geoprocessamento, tais como o *QGIS*, é possível plotar camada de categorias de declividade, Figura 2.1, sobre imagens satelitais,

[5] `https://www.embrapa.br/solos/sibcs/bases-de-dados-de-solos`
[6] `http://www.dsr.inpe.br/topodata/dados.php`

Figura 2.2, e fazer a interseção espacial da camada gerada com os polígonos dos imóveis de interesse.

A nota agronômica (NA) de um imóvel, por sua vez, é atribuída por meio de uma combinação ponderada entre as classes de capacidade de uso de solo (pedologia e relevo) de suas parcelas e as condições de acesso e de localização da propriedade.

Tabela 2.2: Classificação do relevo conforme Manual de Obtenção de Terras do INCRA de 2006.

Classe de Relevo	**Intervalo de Declividade**
Plano	0-2%
Suave Ondulado	2-5%
Moderado Ondulado	5-10%
Ondulado	10-15%
Forte Ondulado	15-45%
Montanhoso	45-70%
Escarpado	$> 70\%$

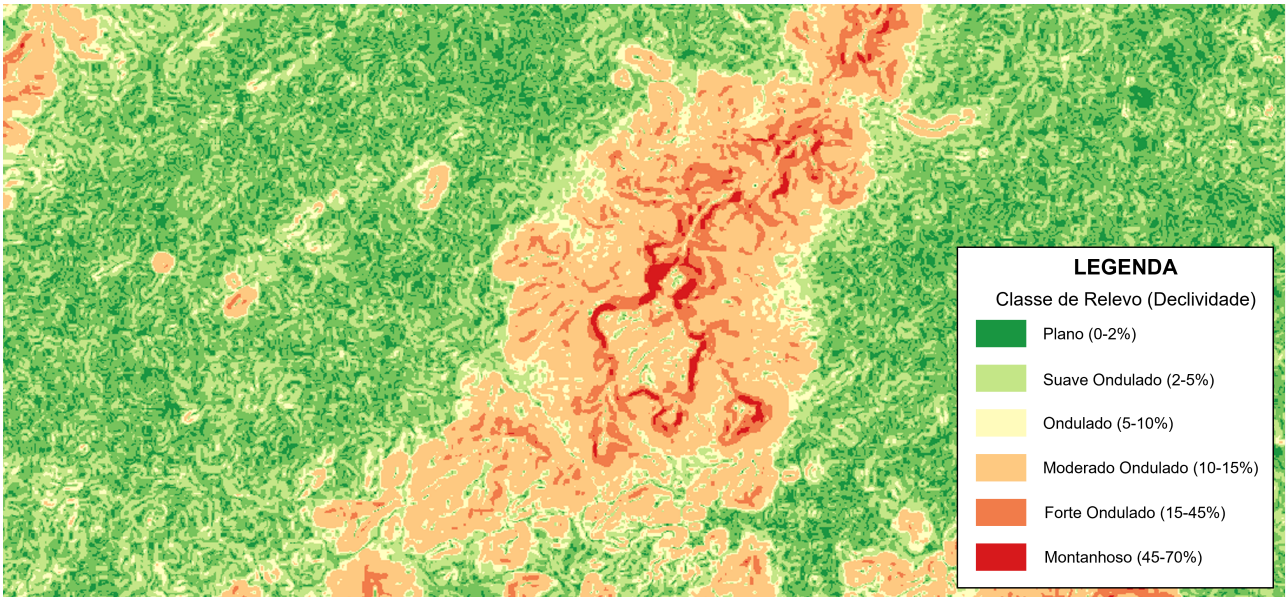

Figura 2.1: Mapa de classes de declividade na região do município de Maranguape-CE produzido no *software QGIS 3.16.3 Hannover*.

2.3.2 Potencialidade Agrícola Natural dos Solos

Conforme macrocaracterização dos recursos naturais do Brasil, divulgada pelo IBGE, a análise das correlações entre as classes de solos, considerando características como tipo, profundidade, textura, fertilidade e pedregosidade, e de relevo gera interpretações de suas

Figura 2.2: Imagem satelital sobreposta por mapa de classes de declividade também na região de Maranguape-CE produzido no *software QGIS 3.16.3 Hannover*.

potencialidades e limitações sob a perspectiva do desenvolvimento agrícola. A Figura 2.3 contém o diagrama de tratamento de informações publicado pelo IBGE.

Um grupo técnico da Diretoria de Geociências do IBGE identificou cinco categorias de terras, hierarquizadas em função de suas potencialidades e limitações quanto ao uso agrícola, enumeradas abaixo:

1. **Classe A1** Muito boa: terras com muito boa potencialidade ao desenvolvimento agrícola - compreende solos com muito boas condições para o desenvolvimento da agricultura, situados em relevo aplainado, com boa fertilidade, profundidade e permeabilidade.

2. **Classe A2** Boa: terras com boa potencialidade ao desenvolvimento agrícola; compreende solos com condições propícias para o desenvolvimento da agricultura, em sua maioria localizados em relevo aplainado, podendo ocorrer pequenas restrições quanto à presença de íons indesejáveis/prejudiciais, mas facilmente corrigíveis e, por vezes, com limitações suaves pela pouca profundidade.

3. **Classe B** Moderada: terras com moderada potencialidade ao desenvolvimento agrícola; compreende solos com condições moderadas para o uso agrícola, presentes dominantemente em relevos ligeiramente acidentados, que podem precisar de ações de manejo adequadas a fim de desenvolver a agricultura, podendo ocorrer moderadas restrições quanto à fertilidade, argilas expansíveis, e presença de íons indesejáveis/prejudiciais, mas relativamente fáceis de serem corrigidas.

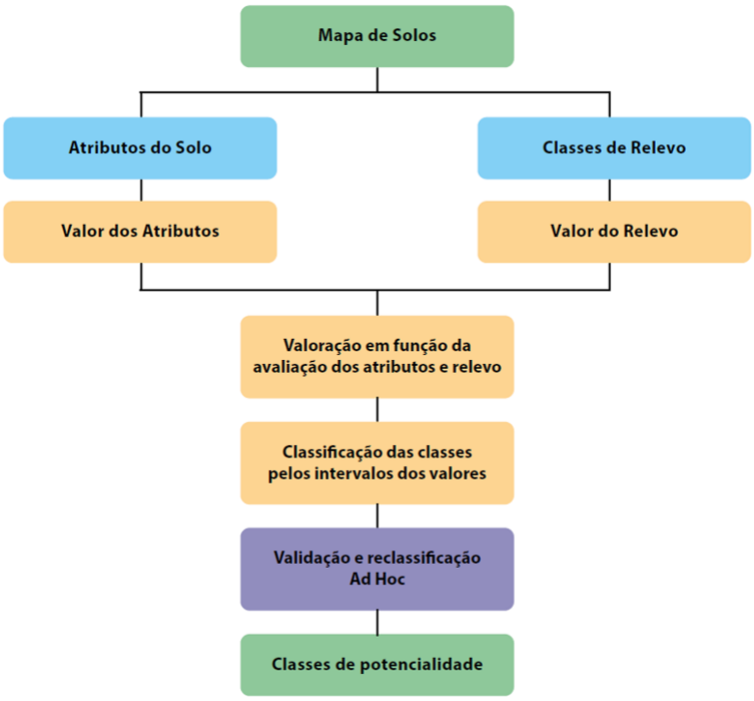

Figura 2.3: Diagrama das etapas de análise automatizada, validação e classificação da potencialidade agrícola natural das terras.

4. **Classe C** Restrita: terras com restrita potencialidade ao desenvolvimento agrícola; compreende solos com condições restritivas para uso agrícola, localizados dominantemente em relevos mais acidentados, que precisam de ações relativamente mais complexas de manejo para o desenvolvimento da agricultura, pela presença de íons indesejáveis/prejudiciais, argilas expansíveis e restrições importantes quanto à profundidade. Também podem ocorrer em áreas aplainadas com restrições pela presença de hidromorfismo, devido às oscilações ou elevações significativas do lençol freático. Para utilização agrícola, necessitaria de ações de manejo significativas e intensivas e sua utilização se daria por uma agricultura especializada adaptada a esses tipos de ambiente.

5. **Classe D** Fortemente restrita: terras com potencialidade fortemente restrita ao desenvolvimento agrícola e terras para proteção, preservação e conservação da vegetação nativa; compreende solos com restrições muito fortes ao uso agrícola, principalmente em superfícies com declividade muito acentuada, presença de sais solúveis indesejáveis e restrições importantes quanto à profundidade. Podem ocorrer em áreas aplainadas com restrições pela forte presença de hidromorfismo e significativa elevação ou oscilação do lençol freático. Para utilização agrícola, necessitariam de ações de manejo significativas e intensivas e sua utilização se daria por uma agricultura especializada adaptada a esses tipos de ambiente. Em alguns locais, essas

terras seriam indicadas como áreas de preservação ambiental, ora pela fragilidade do ambiente, ora pela legislação a qual estão submetidas.

Foi realizado um mapeamento das classes supramencionadas a nível nacional em 2019, o qual foi divulgado em 2022 em formato *shapefile* no sítio eletrônico do IBGE[7]. O mapa das ocorrências de classes de potencialidade agrícola natural das terras no Brasil encontra-se representado na Figura 2.4.

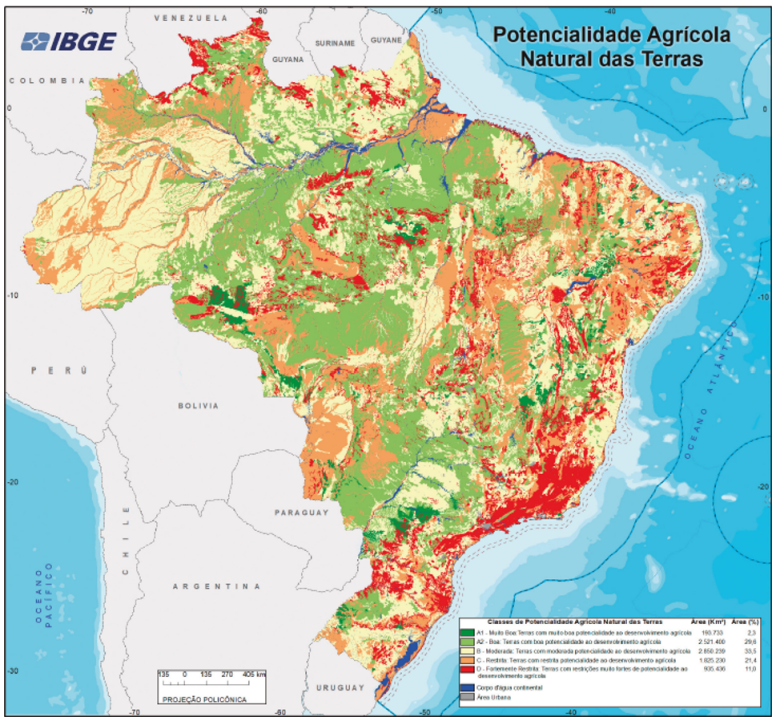

Figura 2.4: Mapa de potencialidade agrícola natural das terras.

Neste trabalho, optou-se por utilizar as informações de potencialidade agrícola natural das terras do IBGE combinadas à disponibilidade de alguns recursos adicionais, conforme indicado por Lima [8], como cursos d'água, da Agência Nacional de Águas e Saneamento Básico (ANA)[8], e como vias pavimentadas, do Departamento Nacional de Infraestrutura de Transportes (DNIT)[9]. Este nível informacional se mostrou mais adequado ao propósito deste trabalho em comparação àquele de classes de capacidade de uso e de nota agronômica, que estaria associado a enorme grau de incerteza sem a visita de profissionais habilitados a todos os imóveis rurais utilizados para construir os modelos matemáticos e estatísticos de precificação.

[7] https://www.ibge.gov.br/geociencias/informacoes-ambientais/estudos-ambientais/24252-macrocaracterizacao-dos-recursos-naturais-do-brasil.html?edicao=35547&t=acesso-ao-produto

[8] https://metadados.snirh.gov.br/geonetwork/srv/api/records/5dd8982f-afe3-4bf0-88d1-73fd53bc196c

[9] https://metadados.snirh.gov.br/geonetwork/srv/api/records/ff37f924-e88d-4ee4-82e7-14a3e5efe0fd

2.4 Modelos de Regressão Múltipla Linear e Espacial

As regressões lineares múltiplas com estimação de regressores ou coeficientes pelo método dos mínimos quadrados (OLS) são uma técnica comumente utilizada na construção de modelos locais, de pequena cobertura geográfica, com finalidade de avaliar imóveis.

Alguns pressupostos precisam ser observados a fim de que os modelos OLS sejam efetivamente válidos, tais como: linearidade, não multicolinearidade, homocedasticidade (variância dos erros aproximadamente constante), não autocorrelação dos erros, normalidade dos resíduos e a ausência de pontos muito influenciantes na amostra coletada e avaliada.

Neste trabalho, o modelo OLS foi calculado e seus resultados foram utilizados como referência na comparação entre modelos acrescidos de componentes espaciais e entre modelos de aprendizagem de máquina.

Como os valores de mercado dos imóveis costumam ser função de relações heterogêneas, não deriváveis e não estacionárias no espaço, os componentes e agregações espaciais utilizados para enriquecimento dos modelos preditivos costumam ser objeto de estudo da estatística espacial e não da geoestatística.

De acordo com Getis e Aldstadt [9], a definição de larguras de banda e de funções de decaimento adequadas na construção de matrizes de vizinhança, normalmente designadas como W, correspondentes às observações são etapas bastante importantes na implementação de modelos de predição com componentes espaciais.

Como abordagens regressivas, há modelagens globais, como a *Spatial AutoRegression* (SAR), e locais, como a *Geographically Weighted Regression* (GWR). Há, ainda, técnicas híbridas que combinam ferramentas espaciais diversas.

Conforme especificado por Anselin [10], a análise da autocorrelação espacial dos atributos é essencial para verificar a importância das variáveis tratadas espacialmente e pode ser feita ainda na etapa de análise exploratória dos dados. Algumas métricas passíveis de uso são: o índice de Moran (I de Moran) para contextos globais, o *Local Indicators of Spatial Association* (LISA) para contextos locais univariados e o *Local Geary c* em cenários locais multivariados.

2.5 Modelos de Aprendizagem de Máquina

Diferentemente da programação computacional tradicional, na qual dados de entrada são passados a um algoritmo e ele retorna dados de saída processados; em aprendizagem de máquina, extraem-se padrões a partir dos dados e constroem-se modelos.

Aplicando-se o conceito a este projeto de pesquisa, busca-se a construção de modelos capazes de estimar valores de mercado de imóveis.

Para tanto, assume-se, inicialmente, a existência de uma função de custo convexa cujos pontos de mínimo local podem ser atingidos a partir de uma configuração adequada de hiperparâmetros de algoritmos, tanto em modelos regressivos lineares quanto em modelos regressivos não lineares, conforme Ben-David e Shalev-Shwartz [11].

Alguns algoritmos utilizados para comparação de resultados são melhor descritos nas subseções abaixo. Eles foram escolhidos com base na literatura e considerando os resultados retornados do método *LazyRegressor* da biblioteca *PyPI Lazy Predict* [10], aplicado aos algoritmos em suas configurações originais. Eles foram avaliados com uso da técnica de validação cruzada, melhor detalhada no Capítulo 4.

2.5.1 SGDRegressor

Optou-se por utilizar o método de aproximação do gradiente descendente estocástico aplicado a minilotes do conjunto de treinamento como otimizador, em busca dos mínimos locais da função de custo, por ser mais simples e ter convergido a mínimos locais similares àqueles obtidos com o método quasi-Newton *Limited-memory Broyden–Fletcher–Goldfarb –Shanno* (L-BFGS) [12].

O algoritmo *SGDRegressor* da biblioteca *scikit-learn* versão 1.2.2 é um modelo de regressão linear que utiliza o método de gradiente descendente estocástico (SGD), definido por Bottou et al. [13] e Zhang [14], para otimizar a função de perda. O *SGDRegressor* é adequado para lidar com grandes conjuntos de dados, pois atualiza os pesos do modelo em pequenos lotes em vez de usar todo o conjunto de dados de uma só vez.

O *SGDRegressor* aceita uma variedade de funções de perda (*loss*) convexas e de penalização (*penalty*), que podem ser especificadas pelo usuário. As funções de perda incluem a função de erro quadrático médio, *mean squared error* (MSE), a função de erro absoluto médio (mean absolute error) e outras. As penalizações incluem a penalização *L1* (*LASSO*) e a penalização *L2* (*Ridge*).

Conforme Tibshirani [15], a penalização *L1* (*LASSO*) adiciona a soma absoluta dos coeficientes à função de custo, podendo levar a zero aqueles relativos às variáveis menos importantes, já a *L2* (*Ridge*), proposta por Hoerl e Kennard [16], adiciona o quadrado da magnitude dos coeficientes à perda convexa, penalizando os modelos mais complexos, mas não chegando a zerar os regressores associados aos atributos menos relevantes. Complementarmente, Zou e Hastie [17] desenvolveram um método de regularização denominado *Elastic Net*, o qual corresponde a uma combinação ponderada entre as duas outras penalizações.

[10] `https://pypi.org/project/lazypredict/`

Neste trabalho, a função de custo regularizada teve por base o erro quadrático médio e encontra-se representada pela Equação 2.2, disponível em detalhes no sítio eletrônico da biblioteca *scikit-learn* [11].

$$L(w,b) = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - f(x_i))^2}{2} + \alpha R(w) \tag{2.2}$$

na qual, L é a função de perda regularizada, n é o número de instâncias do conjunto de treinamento, Y_i é a saída desejada para a instância i, $f(x_i) = w^T * x_i + b$ é a função linear cujos regressores w e intercepto b se deseja determinar a fim de minimizar L, α é um hiperparâmetro não negativo que controla a força do termo de regularização $R(w)$, o qual penaliza a complexidade do modelo de acordo com os tipos *L1*, *L2* ou *Elastic Net*.

O algoritmo *SGDRegressor* também permite ajustar outros hiperparâmetros, tais como a taxa de aprendizado (*learning rate*), o tamanho do lote (*batch size*) e o número de épocas (*number of epochs*).

A referida rotina ajusta os coeficientes do modelo linear de forma iterativa, atualizando os valores a cada passo com base no gradiente da função de perda regularizada em relação aos coeficientes, o qual é calculado para cada instância de treinamento de forma estocástica, ou seja, uma amostra aleatória dos dados é utilizada em cada iteração.

Em resumo, o algoritmo *SGDRegressor* da biblioteca *scikit-learn* é um modelo de regressão linear flexível e escalável que permite aos usuários especificar diferentes funções de perda, penalizações e hiperparâmetros visando à obtenção de uma boa capacidade preditiva.

São vantagens do SGD, quando aplicado ao contexto de mineração de dados, sua eficiência e a facilidade de sua implementação com muitas possibilidades de ajuste de código. Como desvantagens, pode-se citar a quantidade relativamente grande de hiperparâmetros, tais como tipo de regularização e número de iterações, requeridos em sua implementação e sua sensibilidade ao reescalamento de atributos.

2.5.2 MLPRegressor

Uma rede neural artificial (ANN) *multilayer perceptron* (MLP) regressiva é um modelo de aprendizado de máquina que é frequentemente utilizado para problemas de regressão. É uma classe de redes neurais artificiais que consiste em camadas de neurônios, cada qual composta por conjuntos de neurônios interconectados.

Conforme Hinton [18], a arquitetura de uma rede neural MLP regressiva conta normalmente com uma camada de entrada, uma ou mais camadas intermediárias, também conhecidas como camadas ocultas, e uma camada de saída. Os neurônios em cada ca-

[11] `https://scikit-learn.org/stable/modules/sgd.html`

mada estão conectados aos neurônios na camada seguinte por meio de pesos sinápticos, geralmente designados por w_n, que são ajustados durante o processo de treinamento.

Durante o treinamento, a ANN tenta aprender uma função que mapeia as entradas para as saídas desejadas. Isso é feito ajustando os pesos entre os neurônios de modo a minimizar uma medida de erro entre a saída produzida pela rede e a saída observada em dados coletados que contêm valores referenciais.

Uma das vantagens da rede neural MLP regressiva é que ela pode ser usada para aprender funções não lineares complexas. As camadas intermediárias permitem que a rede aprenda representações mais abstratas dos dados, enquanto a camada de saída produz a saída desejada.

No entanto, conforme demonstrado por Pedregosa et al em [19], a rede neural artificial MLP pode ser sensível à inicialização dos pesos e ao tamanho das camadas ocultas, o que pode afetar a capacidade da rede de generalizar para novos dados. Além disso, o treinamento de uma MLP pode ser computacionalmente caro, especialmente para grandes conjuntos de dados.

Em resumo, a MLP regressiva é uma técnica de aprendizado de máquina poderosa e amplamente utilizada para problemas de regressão. É capaz de aprender funções complexas e não lineares, mas pode ser sensível à configuração de hiperparâmetros e computacionalmente cara.

No caso deste trabalho, o número de neurônios da camada de entrada pode ser compreendido como a quantidade de variáveis explicativas do valor do imóvel, o qual pode ser projetado e acessado no único neurônio da camada visível ou de saída.

Os ajustes dos pesos ao longo da etapa de treinamento ocorrem de acordo com algumas equações. A entrada de um neurônio ocorre de acordo com a Equação 2.3.

$$z = w \cdot x + b \tag{2.3}$$

na qual z é a entrada do neurônio, w são os pesos das conexões, x são os valores emitidos pela camada anterior e associada ao neurônio em questão e b é o viés (*bias*) do neurônio, átomo da rede.

Já a saída de um neurônio pode ser representada pela Equação 2.4.

$$a = f(z) \tag{2.4}$$

na qual a é a saída do neurônio e f é uma função de ativação (identidade, sigmoide logística, tangente hiperbólica, *relu*, dentre outras) que pode introduzir não linearidade à rede.

Na versão 1.2.2 da biblioteca *scikit-learn*, utilizada neste trabalho, o algoritmo de *backpropagation* das Equações 2.5, 2.6 e 2.7 é utilizado para ajustar iterativamente os pesos da rede neural *MLPRegressor*.

$$\delta = (Y_i - \hat{Y}_i) \cdot f'(z) \tag{2.5}$$

em que δ é o erro do neurônio da camada de saída, Y_i é a saída desejada para a instância i, $\hat{Y}_i$ é a saída produzida pela rede para a instância i e $f'(z)$ é a derivada primeira da função de ativação.

$$\delta = f'(z) \cdot \sum_j w_{ji} \cdot \delta_j \tag{2.6}$$

em que δ é o erro de cada neurônio das camadas ocultas e $\sum(w \cdot \delta)$ é a soma dos erros ponderados pelos pesos das associações com a camada seguinte.

A atualização dos pesos da ANN ocorre de acordo com a Equação 2.7.

$$w = w - \eta \cdot \delta \cdot x \tag{2.7}$$

na qual η é uma taxa de aprendizagem que controla a magnitude das atualizações dos pesos.

A rede é treinada exaustivamente visando à minimização da função de custo, normalmente definida como o erro quadrático médio (MSE), representado na Equação 2.8.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.8}$$

Conforme documentação da biblioteca *scikit-learn*, disponível em sítio eletrônico[12], no algoritmo *MLPRegressor*, além de as derivadas parciais da função de perda em relação aos parâmetros do modelo serem calculadas para atualizar os pesos, também é possível se adicionar um termo de regularização à função de perda, o qual diminui parâmetros do modelo para simplificá-lo e evitar que ocorra *overfitting*.

2.5.3 XGBoost

O *XGBoost* é um algoritmo de aprendizagem de máquina para problemas de regressão e de classificação que utiliza a técnica de *boosting* para melhorar o desempenho dos modelos.

Ele foi criado por Tianqi Chen em 2014 e, conforme publicado em [20], pode ser considerado um algoritmo de *ensemble*, pois combina múltiplos modelos de árvores de

[12] `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html`

decisão em um modelo mais forte e robusto. O algoritmo utiliza a técnica de *boosting* para criar uma sequência de modelos, em que cada um deles é treinado para corrigir os erros do modelo anterior.

No *XGBoost*, cada árvore é construída de forma a minimizar a função de perda regularizada, que é definida como a soma da função de perda e de um termo de regularização que penaliza a complexidade do modelo. A ideia é que cada árvore adicione informações complementares às árvores anteriores e que o modelo final seja capaz de capturar padrões complexos nos dados.

Durante o treinamento do *XGBoost*, cada árvore, representação de regras de decisão sob forma de diagrama que remete a corpo arbóreo com ramificações, segundo Breiman et al., [21], é treinada em uma amostra aleatória dos dados de treinamento, o que ajuda a evitar *overfitting* e a melhorar a generalização do modelo. Além disso, o algoritmo utiliza tanto técnicas de regularização *L1* quanto *L2* para melhorar a performance do modelo.

Em problemas de regressão, caso deste trabalho, o *XGBoost* utiliza a função de perda de erro regularizada, representada pela Equação 2.9, para calcular o erro do modelo em cada iteração e ajustar os pesos dos exemplos cujos resíduos, diferenças entre os valores observados e aqueles calculados pelo modelo, são superiores a determinado limiar. O modelo pode ser entendido como um conjunto de k árvores de decisão.

$$L(w) = \sum_i l(Y_i, \hat{Y}_i) + \sum_k \Omega(f_k) \tag{2.9}$$

sendo a função objetivo regularizada $L(w)$, que penaliza a complexidade do modelo e evita *overfitting*, dependente da função de perda convexa diferenciável $l(Y_i, \hat{Y}_i)$, que mede a diferença entre o valor observado Y_i e o valor estimado $\hat{Y}_i$, acrescida do termo de regularização $\Omega(w)$, que penaliza a complexidade do modelo em função do vetor de pesos w dos nós de cada estrutura de árvore independente.

O termo de regularização Ω, por sua vez, encontra-se melhor detalhado na Equação 2.10.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{2.10}$$

em que γ é o hiperparâmetro de complexidade da árvore, T é o número de folhas da árvore e λ é o hiperparâmetro de regularização $L2$.

O algoritmo *XGBRegressor*, da biblioteca *XGBoost* versão 1.7.4, utilizado neste trabalho permite, ainda, definição de hiperparâmetros personalizados para ajustar o modelo às necessidades específicas do problema, o que consta detalhadamente no sítio eletrônico da biblioteca[13], juntamente a outros itens da documentação.

[13] `https://xgboost.readthedocs.io/en/stable/parameter.html`

O *XGBRegressor* conta com uma técnica de aproximação de segundo grau utilizada para melhorar a eficiência computacional e a precisão das predições do modelo. Em vez de usar apenas o gradiente da função objetivo, a aproximação de segundo grau leva em conta também a curvatura da função objetivo, calculando a derivada de segundo grau, o que exige que a função de custo seja duplamente diferenciável.

A técnica de aproximação de segundo grau envolve a expansão da função objetivo em uma série de Taylor até segunda ordem. A primeira derivada da série é o gradiente (g), Equação 2.11, e a segunda derivada é a curvatura da função objetivo (h), Equação 2.12. A expansão da série é então usada para aproximar a função objetivo em torno dos pontos de corte das árvores de decisão, a fim de que se obtenha uma melhor estimativa do valor da referida função em cada ponto.

$$g_i = \frac{\partial}{\partial \hat{Y}_i} l(Y_i, \hat{Y} i) \tag{2.11}$$

$$h_i = \frac{\partial^2}{\partial \hat{Y}_i^2} l(Y_i, \hat{Y} i) \tag{2.12}$$

Essa técnica de aproximação de segundo grau é usada para ajustar os pesos w da folha j de cada árvore de decisão k com base em seu conjunto de instâncias i, conforme Equação 2.13, em vez de simplesmente atualizá-los com base no gradiente. O objetivo é melhorar a precisão do modelo ao torná-lo mais sensível às pequenas variações na função objetivo em torno dos pontos de corte das árvores de decisão.

$$w_j = -\frac{\sum_i g_i}{\sum_i h_i + \lambda} \tag{2.13}$$

O algoritmo *XGBRegressor* é, portanto, altamente escalável e, por penalizar modelos demasiadamente complexos, com podas de suas árvores constituintes bem definidas e com função de custo regularizada, costuma apresentar boa generalização e bom poder preditivo a depender da qualidade dos dados coletados e tratados.

2.6 PLS-SEM

De acordo com Hair et al. [22], os modelos de equações estruturais (SEM, do inglês *structural equation modeling*) com mínimos quadrados parciais (PLS, do inglês *partial least squares*) são um método estatístico de segunda geração que permite a pesquisadores modelar e estimar relacionamentos complexos entre múltiplas variáveis, dependentes ou independentes.

Conforme Basco [23], diferentemente dos modelos econométricos, mais fechados, os modelos de equações estruturais admitem a existência de variáveis não consideradas na representação simulada da realidade e contemplam parcelas de erro por essa consideração.

O PLS-SEM normalmente é composto por um modelo de medida, mais externo e constituído pelos relacionamentos entre indicadores, no caso deste trabalho, dados secundários observados diretamente, e construtos ou variáveis latentes, cada qual composto por um ou mais indicadores, e por um modelo estrutural, mais interno e constituído pelo relacionamento entre os construtos. Os relacionamentos são testados sob testes de hipótese fundamentados em conceito de variância total. Os construtos endógenos, explicados por outros construtos, têm variância não explicada traduzida em forma de termos de erro.

Teoria e lógica devem determinar a sequência de construtos em um modelo de equação estrutural e, no caso de não haver literatura com arquitetura clara do problema, os pesquisadores devem utilizar senso comum aos especialistas da área de conhecimento para determinar a sequência, avaliando o modelo construído posteriormente, de acordo com Hair et al. [22].

O PLS-SEM lida bem com construtos formativos e reflexivos e permite trabalhar com o conceito estatístico de moderação. Segundo Becker et al. [24], em trabalho publicado em 2018, uma variável ou um construto moderador muda a força ou até mesmo o sentido do relacionamento entre dois outros construtos no modelo.

Por ser um método estatístico não paramétrico, o PLS-SEM não requer que os atributos sigam determinada distribuição estatística.

Segundo Hair et al. [22], resumido no *site* de apresentação do *software SmartPLS*[14], "o PLS-SEM conta com um *bootstrapping* para testar a significância dos coeficientes de caminho estimados, semelhantes a regressores. No procedimento, subamostras são criadas com observações extraídas aleatoriamente do conjunto original de dados (com reposição). A subamostra é então usada para estimar o modelo de caminho PLS. Esse processo é repetido até que um grande número de subamostras aleatórias tenha sido criado, normalmente cerca de 10.000."

O PLS-SEM foi utilizado como parte da solução proposta neste trabalho a fim de visualizar os relacionamentos entre as variáveis, contribuindo à interpretabilidade, e para selecioná-las.

2.7 Interpretabilidade com valor de Shapley

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

[14] `https://www.smartpls.com/documentation/algorithms-and-techniques/bootstrapping/`

Conforme Lundberg et al. [25] e Chakraborty et al. [26], 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.

Em [27], Lundberg e Lee afirmam que a obtenção dos valores de Shapley requer retreinar o modelo de aprendizagem de máquina para todos os subconjuntos $S \subseteq F$, sendo F o conjunto de todas as variáveis explicativas. Para calcular um valor de importância para cada variável baseado no efeito de sua inclusão sobre as predições, um modelo $f_{S\cup\{i\}}$ é treinado com o atributo e um outro modelo f_S é treinado sem o atributo. Os resultados das predições são comparados e ponderados para todos os arranjos possíveis de variáveis $S \subseteq F\backslash\{i\}$ utilizando a Equação 2.14.

$$\phi_i(f) = \sum_{S\subseteq F\backslash\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[\, f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{2.14}$$

em que x_S representa o vetor de valores de entrada das variáveis de um subconjunto S.

Esta técnica de interpretabilidade pode ser aplicada tanto a modelos lineares mais simples quanto a modelos não lineares mais complexos e sua fundamentação matemática pode ser encontrada no site documentação da biblioteca SHAP (*SHapley Additive exPlanations*)[15] em linguagem *Python*.

2.8 Considerações Finais

Neste capítulo, foram abordados os principais conceitos utilizados nesta pesquisa. Em razão disso, foram registradas definições relativas a avaliações imobiliárias, em considerações acerca da seleção de atributos, em modelos de aprendizagem de máquina diretamente relacionados a esta pesquisa e em interpretabilidade. Convém salientar que algumas características particulares de cada tema, tais como parâmetros, métodos e pressupostos, serão apresentadas e justificadas no Capítulo 4, no qual se propõe uma solução integrada em linguagem *Python*. No próximo capítulo, serão apresentados os trabalhos relacionados a esta pesquisa visando a identificar as lacunas deste tema na literatura.

[15] `https://shap.readthedocs.io/en/latest/index.html`

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são apresentadas as abordagens que exploram o problema de valoração imobiliária por intermédio de técnicas científicas e de maneira hedônica [2].

A literatura discutida nesta parte do trabalho, que se comunica diretamente com o Capítulo 2, está relacionada, mais especificamente, a domínios de valoração de imóveis.

A revisão bibliográfica com método científico permitiu a identificação das tendências na linha de pesquisa pretendida. A coleta de informações foi majoritariamente concentrada nas plataformas *Web of Science* e *Google Scholar*.

Há diferentes abordagens com foco na valoração de imóveis presentes na literatura. Alves Dantas et al. [3] 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 da ordem de 90%.

Park e Bae [4] 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 e Bastian [28] desenvolvem modelos regressivos de predição de valores de imóveis de diferentes tipologias em Nova York e um outro protótipo de classificação de bens quanto à sua probabilidade de venda, ou seja, de chance de liquidação. Constata-se a heterogeneidade espacial de relações lineares e não lineares. O valor dos bens imobiliários é desmembrado em componentes intrínsecos e extrínsecos: fatores ambientais, características estruturais, variáveis agregadas de vizinhança e geolocalização. As funções de agregação espacial criam novas variáveis que tendem a aumentar o poder preditivo dos

modelos de aprendizagem de máquina para determinados tipos de imóveis. Os coeficientes de determinação (R^2) mais elevados obtidos são da ordem de 50%.

Dewan et al. [29] combinam redes neurais artificiais a modelos autorregressivos espaciais de caráter global (SAR), nos quais as autocorrelações espaciais podem ser analisadas à luz do índice de Moran. Demonstra-se, preliminarmente, que a combinação proposta pelos autores chega a ser 20% superior em poder de explicação se comparada a modelos SAR convencionais para o domínio de funções contínuas e deriváveis. A dificuldade em se definir de forma precisa a matriz de vizinhança W faz com que ela seja trabalhada com utilização de séries numéricas, mais especificamente, de potências. O domínio explorado no trabalho foi diferente daquele que representa o foco deste trabalho, entretanto, a arquitetura construída pôde ser conceitualmente aproveitada.

Lundberg et al. [25] apresentam uma estrutura unificada denominada *SHapley Additive exPlanations* (SHAP), a qual utiliza o conceito de contribuições marginais das variáveis nos modelos preditivos visando a garantir interpretabilidade aos mais simples assim como àqueles mais complexos. Fazendo uso de seus recursos visuais, se mostra possível realizar ricas análises locais e globais.

Já em [26], Chakraborty et al. aplicam o algoritmo SHAP para garantir interpretabilidade a um modelo *ensemble* que combina *light gradient boosting* (*LGBoost*) e *natural gradient boosting* (*NGBoost*) aplicados ao domínio de custos de construção nos Estados Unidos da América. Tal modelo híbrido apresenta resultados consistentes em termos de poder de explicação e com compreensível participação e influência dos atributos selecionados.

Considerando a ausência de estacionariedade, comportamento não contínuo e não derivável, no espaço do atributo *valor do imóvel*, Hagenauer e Helbich [30] propõem um modelo de rede neural combinado a regressões espaciais locais ponderadas, denominado *Geographically Weighted Artificial Neural Network* (GWANN). É realizada uma análise comparativa entre o modelo *Geographically Weighted Regression* (GWR) tradicional e o GWANN. Cada neurônio da camada de saída da arquitetura do GWANN é associado a uma coordenada geográfica espacial. Os pesos das conexões entre os neurônios da última camada oculta e os neurônios da camada de saída da rede são estimados utilizando uma função de erro de base espacial local.

Nesse artigo, o decaimento da influência espacial é definido por uma função kernel gaussiana com base em distância euclidiana (ED) e em distância de viagem pelo *software* OpenStreetMap[1] (TTD). A definição da largura de banda mais adequada para conjunção das instâncias vizinhas tanto de forma fixa quanto de forma adaptativa se mostra bastante desafiadora. Quando aplicado ao domínio real de casas unifamiliares na Áustria, os mai-

[1] `https://www.openstreetmap.org/`

ores coeficientes médios de determinação (R^2) são de aproximadamente 45%, conforme pode ser visto na Figura 3.1.

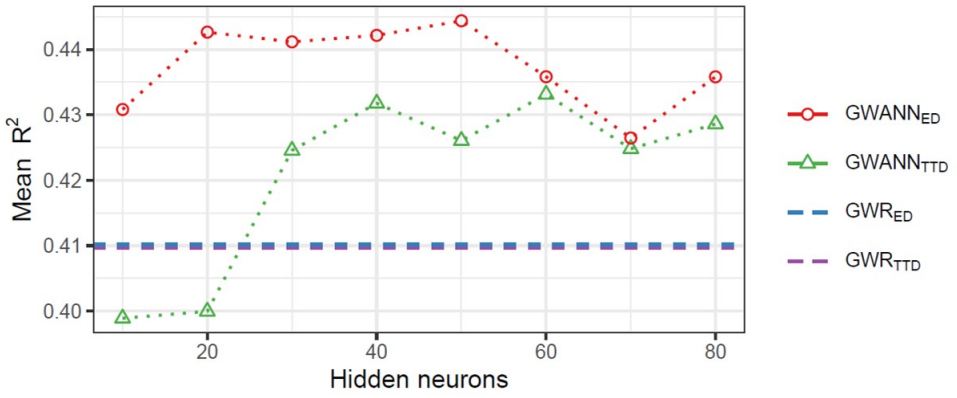

Figura 3.1: Coeficientes de determinação obtidos na predição de preços de casas na Áustria.

Tchuente e Nyawa [5] comparam o desempenho de sete algoritmos de aprendizagem de máquina, tanto lineares quanto não lineares, aplicados com e sem geocodificação a uma base de dados disponibilizada pelo Governo da França de valores de imóveis em cidades francesas. Sem a utilização de latitude e longitude, o modelo de rede neural artifical *multilayer perceptron* (MLP) apresenta o melhor coeficiente de determinação médio para as cidades, de aproximadamente 61%; já com o uso de geocodificação, o algoritmo de *random forest* (RF) gera o melhor R^2, de aproximadamente 74%. Foram testados diferentes níveis de granularidade espacial.

3.1 Discussão dos Trabalhos Relacionados

Fundamentado na pesquisa bibliográfica realizada, a Tabela 3.1 apresenta um resumo comparativo entre a solução proposta (Capítulo 4) e os trabalhos da literatura.

As lacunas dos trabalhos anteriormente apresentadas e resumidas na Tabela 3.1 indicam que este projeto de pesquisa tem potencial de gerar modelo de aprendizagem de máquina com as seguintes contribuições:

- elevado poder de explicação da variância do atributo dependente (valor dos imóveis);
- ampla abrangência geográfica;
- inclusão de mais de um tipo de vocação imobiliária (usos residencial, comercial, institucional e misto) nas análises realizadas;
- interpretabilidade; e

- conjunto de dados de avaliações imobiliárias realizadas conforme NBR 14653 e IN SPU que rege o assunto no âmbito da União.

Tabela 3.1: Comparação com os 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
Dantas et al. 2010 [3]	√		√			√		√
Park e Bae 2015 [4]	√		√	√	√			
Kiely e Bastian 2020 [28]			√	√	√	√	√	
Hagenauer e Helbich 2022 [30]		√	√	√		√		
Tchuente e Nyawa 2022 [5]	√	√	√	√	√	√		
Solução proposta	√	√	√	√	√	√	√	√

Capítulo 4

Solução Proposta

Neste capítulo, são apresentados os modelos construídos como parte da solução proposta e os conjuntos de dados utilizados para treinamento, validação e teste dos algoritmos de aprendizagem de máquina lineares e não lineares. Na sequência, a metodologia é descrita e discorre-se sobre a forma de avaliação dos resultados.

4.1 Modelos para Construção de Estimativas de Valor de Mercado de Imóveis da União

Este trabalho propõe uma ferramenta automatizada em três níveis para estimativa de valores de imóveis urbanos e em dois níveis para o caso de imóveis rurais.

4.1.1 Modelos Imóveis Urbanos

No primeiro nível de modelo (U_1), tido como aquele mais básico, busca-se estimar o valor de mercado de imóveis urbanos da União, em especial aqueles administrados pelo Exército Brasileiro, por meio de variáveis de área intrínsecas aos imóveis e de atributos genéricos extrínsecos aos bens relativos aos municípios onde eles se encontram, tais como grau de urbanização e Índice de Desenvolvimento Humano. Ressalta-se que a utilização da variável Produto Interno Bruto (PIB) *per capita* foi explorada, mas que seu comportamento nos modelos pouco traduziu as diferenças municipais entre os valores de imóveis.

No segundo nível de modelagem (U_2), intermediário, adicionam-se variáveis ligadas às condições próprias dos imóveis, tais como a vida útil por tipo construtivo e de uso e sua idade aparente, referenciada como o tempo decorrido desde a última grande reforma e o mês de janeiro de 2022, mês a que foram atualizados os valores das avaliações imobiliárias quando da execução dos experimentos.

Já no terceiro nível de modelo (U_3), específico, adicionam-se atributos relativos às áreas de ponderação (AP) definidas pelo IBGE a nível de granularidade intramunicipal que contêm os centroides do imóveis ora avaliados. E, por meio da API *Google Places Nearby Search*, coletaram-se informações sobre o entorno imediato dos imóveis, em um raio prioritário de 400 metros, correspondente a 5 minutos de caminhada, aproximadamente. Tal escolha foi fundamentada em publicação de El-Geneidy et al. [31].

4.1.2 Modelos Imóveis Rurais

No primeiro nível de modelo (R_1), tido como básico, procura-se estimar o valor de mercado de imóveis rurais da União, em especial aqueles administrados pelo Exército Brasileiro, com o uso de variáveis disponíveis e de caráter físico, intrínsecas às parcelas de terra, como suas áreas, a presença de cursos d'água e de infraestrutura para escoamento de produção, e às suas condições naturais, como seus potenciais agrícolas naturais.

Já no segundo nível (R_2), atributos associados à localização e ao uso das propriedades são adicionados, tais como o valor de terra nua (VTN) com fins de titulação, indicado para cada município pelo Instituto Nacional de Colonização e Reforma Agrária (INCRA), e o VTN utilizado como indexador à declaração do Imposto sobre a Propriedade Territorial Rural (ITR) por região rural definida pelo IBGE e para cada tipo de uso (agricultura, pecuária, silvicultura, preservação, dentre outros) das parcelas rurais, informações publicadas pela Receita Federal do Brasil (RFB). Uma outra variável considerada é o Índice de Desenvolvimento Humano Municipal (IBGE 2010) dos municípios que contêm os centroides dos imóveis dentro de sus limites geográficos, a fim de traduzir um pouco da qualidade de vida local.

Os polígonos georreferenciados das regiões rurais publicados pelo IBGE no ano de 2015 encontram-se ilustradas na Figura 4.1.

4.2 Metodologia

Nesta seção, discorre-se acerca da extração de dados, de sua análise exploratória univariada e bivariada com gráficos e da concepção dos modelos associados à solução proposta.

4.2.1 Extração de Dados

Os dados diretamente associados a cada um dos imóveis foram extraídos de bases de acesso restrito do EB e da SPU.

Quanto ao EB, foram utilizados o Sistema Informatizado de Gestão do Patrimônio Imobiliário e Meio Ambiente (SIGPIMA) e o Sistema Unificado do Processo de Obras

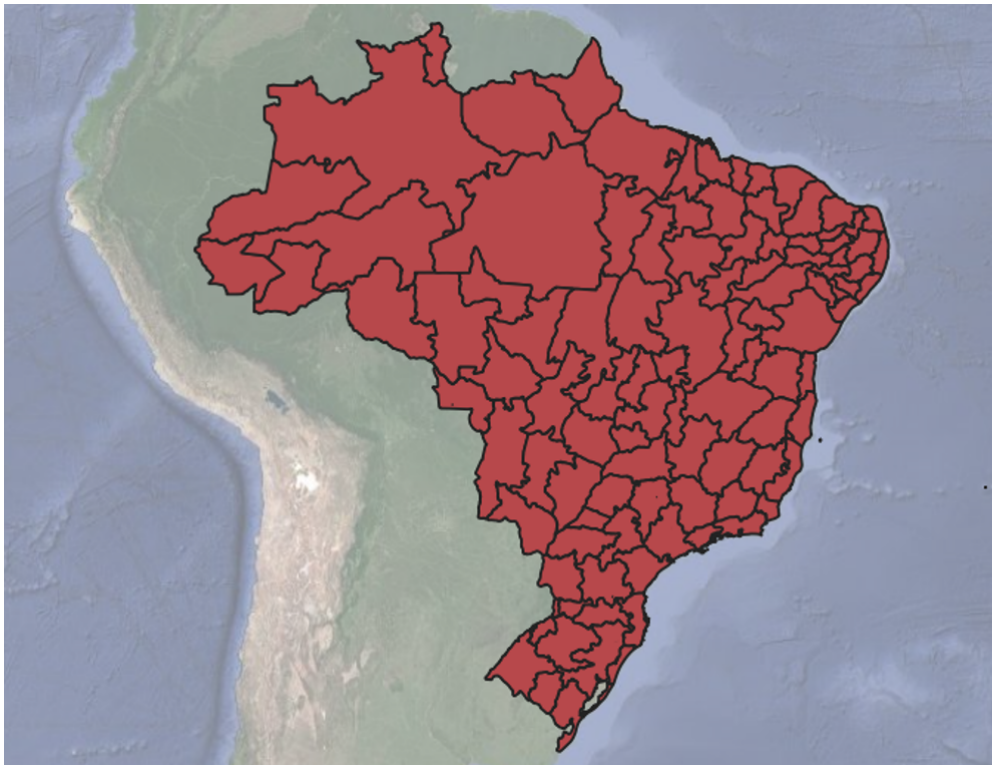

Figura 4.1: Regiões rurais publicadas pelo IBGE em 2015 e para as quais há VTN associado à cobrança de ITR pela RFB.

(OPUS) como fontes principais, ambos geridos pelo Departamento de Engenharia e Construção (DEC) do EB. Todos os imóveis do EB encontram-se georreferenciados em formato *shapefile* e com atributos em extensão *dbf*, tanto os de valor calculado, denominados avaliados neste trabalho, quanto aqueles de valor desconhecido.

Já em relação à SPU, foram extraídos dados do Sistema de Gerenciamento dos Imóveis de Uso Especial da União SPIUnet. Como as instâncias da SPU não continham polígonos georreferenciados, optou-se por geocodificar o campo textual *Endereço* no *Google Earth Pro*.

4.2.1.1 Extração de Dados Urbanos

No processo de geocodificação de imóveis urbanos, 925 instâncias não puderam ter suas coordenadas espaciais extraídas no *Google Earth Pro*, conforme resumo constante na Tabela 4.1.

As ocorrências com campo *valor* conhecido correspondem a avaliações imobiliárias executadas por engenheiros ou arquitetos entre os anos de 2016 e 2022, inclusive, por serem aquelas disponíveis nos bancos de dados, e homologadas pelo corpo técnico do EB, mais especificamente da DPIMA, ou da SPU à luz da IN SPU nº 1, de 02 de dezembro de 2014, da IN SPU nº 2, de 02 de maio de 2017, da IN SPU nº 5, de 28 de novembro

Tabela 4.1: Informações sobre as instâncias com valor conhecido.

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%

de 2018, ou da IN SPU nº 67, de 20 de setembro de 2022, a depender de sua data de referência, e em respeitos à norma ABNT NBR 14653.

Parte das 925 instâncias que não puderam ser geocodificadas, mais especificamente 284, foram recuperadas manualmente, resultando nos quantitativos finais constantes na Tabela 4.2.

Tabela 4.2: 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%

Atributos socioeconômicos são disponibilizados em diferentes granularidades espaciais, desde nacional até AP, pelo IBGE.

Os dados diretamente extraídos das bases do EB e da SPU e geocodificados foram enriquecidos com atributos socioeconômicos do Censo 2010 do IBGE tratados por área de ponderação (AP). Para associar os imóveis às AP, realizamos operações de união espacial no *QGIS 3.16 Hannover* entre os pontos relativos aos centroides das propriedades e os polígonos das AP, extraídos do repositório de Furtado no GitHub[1], associado a trabalho técnico do Ipea [32]. As instâncias do EB e da SPU sobrepõem as AP do IBGE nas Figuras 4.2 e 4.3. Nos casos em que o centroide do imóvel ficou em área descoberta pelas AP, utilizou-se a média dos atributos das AP adjacentes.

Por fim, foram coletadas as quantidades de diversas tipologias de pontos de interesse, tais como estações de metrô, parques 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 API *Google Places Nearby Search*.

4.2.1.2 Extração de Dados Rurais

Para os imóveis rurais, a geocodificação não se mostrou suficiente, tendo em vista a necessidade de se calcular o percentual de incidência espacial de cada potencialidade

[1] `https://github.com/BAFurtado/censo2010`

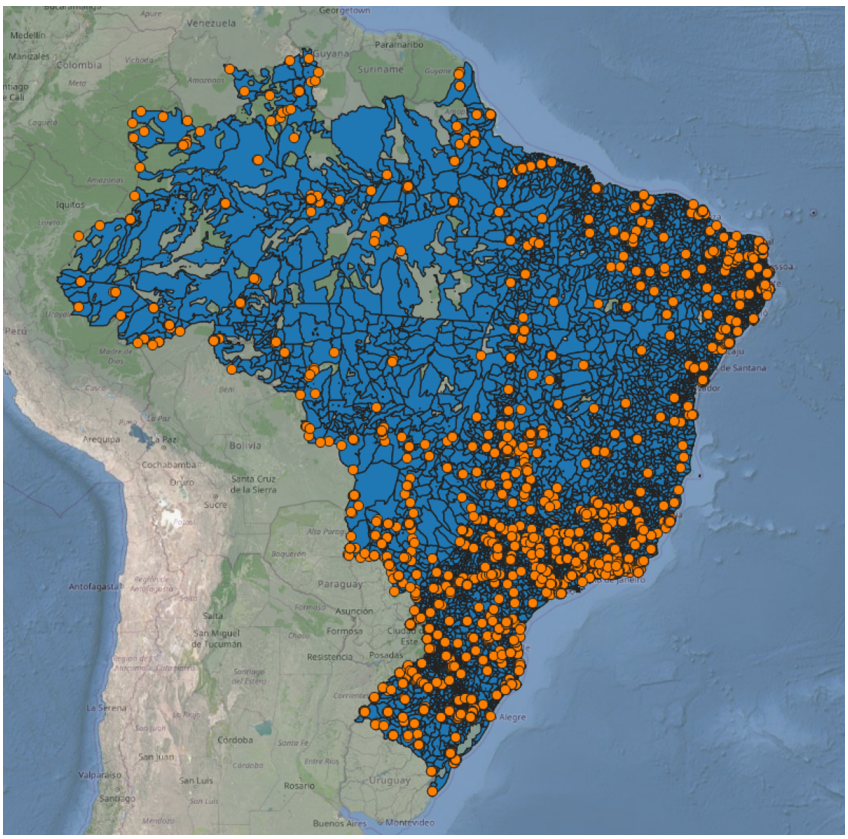

Figura 4.2: AP do IBGE (em azul) sobrepostas pelos centroides de instâncias do EB e da SPU.

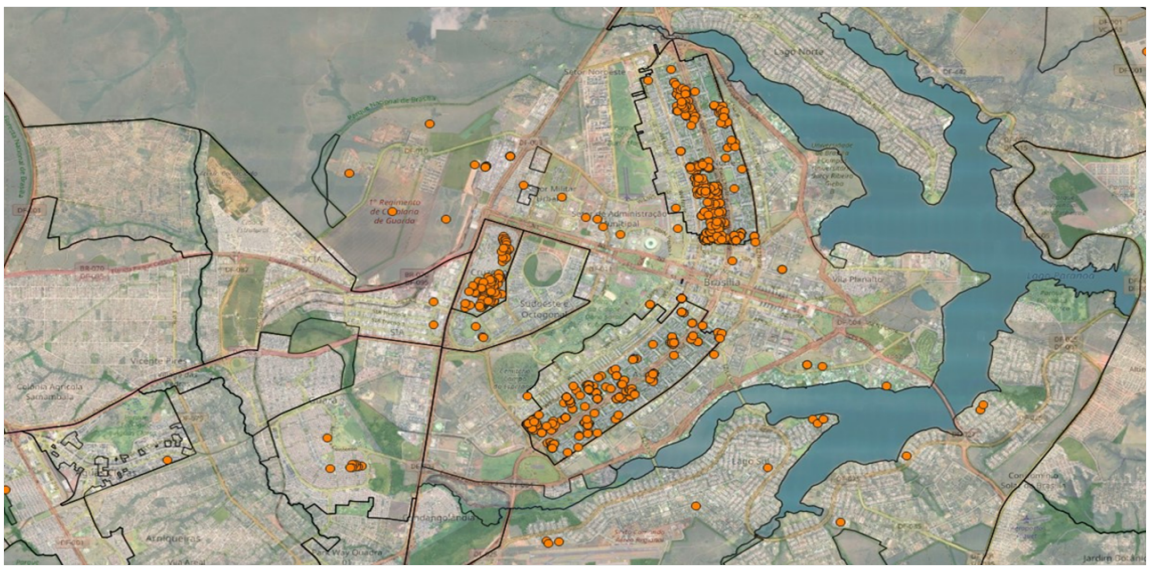

Figura 4.3: AP do IBGE (transparentes com contorno em preto) sobrepostas pelos centroides de instâncias do EB e da SPU na região do Distrito Federal.

agrícola natural dos solos dos imóveis.

Optou-se, portanto, por extrair os arquivos em formato *shapefile* relativos às poligonais dos imóveis rurais avaliados da SPU, constantes no portal do Sistema de Gestão Fundiária (SIGEF)[2] do INCRA. Observaram-se apenas 17 ocorrências da SPU, com valor conhecido, no referido sistema, as quais foram acrescidas ao conjunto de sete (7) avaliações do EB com polígonos georreferenciados, conforme Tabela 4.3.

Tabela 4.3: Informações sobre as instâncias rurais avaliadas coletadas e com polígono georreferenciado disponível.

Fonte	**Qtde de instâncias rurais com valor conhecido**	**Participação relativa**
EB	7	29,2%
SPU	17	70,8%
EB e SPU	24	100,0%

Os polígonos dos imóveis administrados pelo EB e pela SPU foram enriquecidos com informações de potencialidade agrícola natural das terras do IBGE, de cursos d'água da ANA e de rodovias do DNIT. Para tanto, operações de união e de interseção espacial foram realizadas no *QGIS 3.16 Hannover*. A Figura 4.4 ilustra as camadas trabalhadas na ferramenta mencionada.

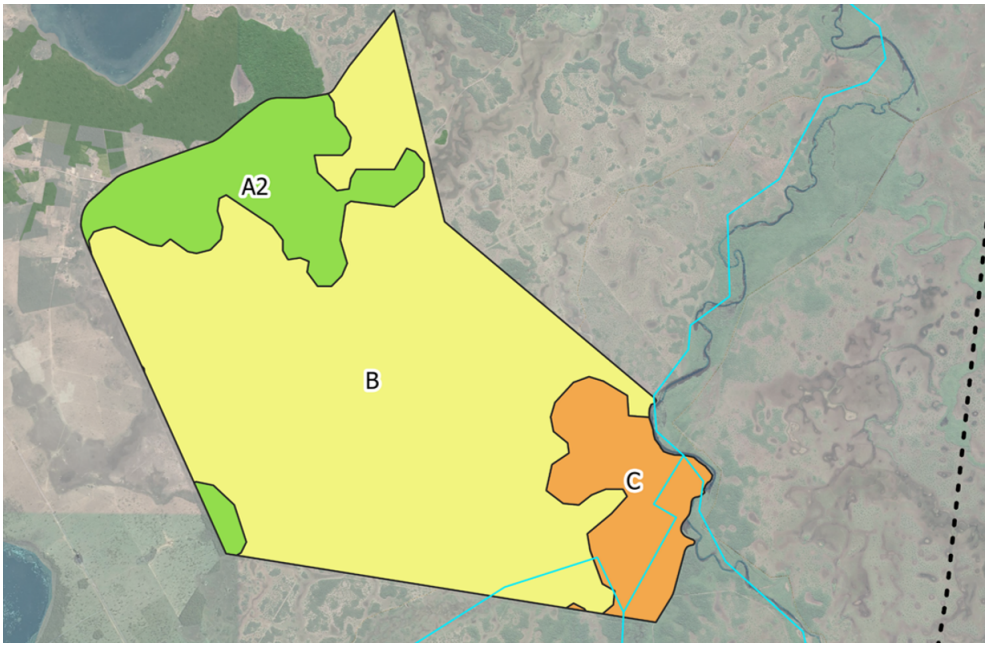

Figura 4.4: Polígono de imóvel rural da base de dados avaliada com suas parcelas de potencialidade agrícola natural de solo (em cores verde, amarelo e laranja), cursos d'água da ANA (linha contínua azul) e rodovia do DNIT (linha tracejada preta).

[2] `https://sigef.incra.gov.br/`

4.2.2 Atributos Analisados

As seguintes variáveis foram analisadas preliminarmente nos modelos experimentais de imóveis urbanos:

1. **Capital UF** Variável dicotômica que recebe valor "1" quando o imóvel se situa em capital de Unidade da Federação e valor "0" caso contrário. Fonte: IBGE.

2. **Tipologia Municipal** Variável qualitativa nominal que indica o tipo do município cujos limites geográficos incluem o centroide do imóvel em questão, segundo a classificação definida pelo IBGE: *município predominantemente urbano*, *município intermediário adjacente*, *município intermediário remoto*, *município rural adjacente* e *município rural remoto*. Unidade: adimensional (codificação binária). Granularidade espacial: município. Fonte: IBGE.

3. **Grau de Urbanização** Variável numérica que indica o grau de urbanização do município cujos limites geográficos contêm o centroide do imóvel em questão. Escala numérica contínua de 0 (município sem indícios de urbanização) a 1 (município integralmente urbanizado). Referente ao ano de 2010. Unidade: adimensional. Granularidade espacial: município. Fonte: IBGE.

4. **IDHM** O Índice de Desenvolvimento Humano Municipal (IDHM) é uma medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda. O índice numérico contínuo varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento humano. Calculado com base nas informações do Censo 2010. Unidade: adimensional. Granularidade espacial: município. Fonte: IBGE.

5. **IDHM-R** Parcela de renda do IDHM 2010. O índice numérico contínuo varia de 0 a 1. Unidade: adimensional. Granularidade espacial: município. Fonte: IBGE.

6. **IVS** O Índice de Vulnerabilidade Social (IVS) 2010 municipal congrega três dimensões da vulnerabilidade social: infraestrutura urbana do município, capital humano dos domicílios nele contidos e renda (acesso a trabalho e forma de inserção, formal ou não, dos residentes nos referidos domicílios). Cada uma dessas dimensões reúne, por sua vez, um conjunto de variáveis obtidas nas bases dos censos demográficos do IBGE, as quais refletem diferentes aspectos das condições de vida. O índice numérico contínuo varia de 0 a 1. Quanto mais próximo de 1, pior a situação. Unidade: adimensional. Granularidade espacial: município. Fonte: Ipea.

7. **Taxa de Homicídios** Variável numérica contínua que representa uma métrica de violência calculada por 100 mil habitantes em 2019. Granularidade espacial: município. Fonte: Ipea.

8. **PIB *per capita*** Variável numérica contínua calculada como a divisão do PIB (soma de todos os bens e serviços finais produzidos) pelo número de habitantes do município referente ao ano de 2018. Unidade: R$/habitante. Granularidade espacial: município. Fonte: IBGE.

9. **Vocação do Imóvel** Variável qualitativa nominal que indica o tipo de uso do imóvel em questão, nas seguintes categorias: *residencial*, *comercial*, *institucional* e *misto* (quando o imóvel atende a mais de um tipo de uso). Unidade: adimensional (codificação binária). Fonte: EB/SPU.

10. **Terreno** Variável dicotômica que recebe valor "1" quando o imóvel não contém benfeitorias construtivas e valor "0" caso contrário. Fonte: EB/SPU.

11. **Área do Terreno** Variável numérica contínua que indica a área do terreno do imóvel avaliado. Quando se tratam de unidades imobiliárias, assume o valor de área privativa total. Unidade: m^2. Fonte: SIGPIMA (EB) / SPIUnet (SPU).

12. **Área Construída** Variável numérica contínua. Somatório das áreas das benfeitorias construtivas do imóvel avaliado. Unidade: m^2. Fonte: OPUS (EB)/SPU.

13. **CUB** Variável numérica discreta. O CUB analisado foi aquele correspondente ao projeto-padrão, ao padrão construtivo (baixo, normal ou alto) e ao tipo de uso do imóvel avaliado referente ao mês de janeiro de 2022, sem desoneração. Em razão da pandemia, alguns Sinduscon estaduais não tinham publicado resultados mensais mais recentes. Unidade: R$/$m^2$. Fonte: EB/SPU e Sinduscon[3].

14. **Idade Aparente** Variável categórica ordinal discreta que representa a diferença temporal ponderada por área entre a data da última grande reforma das benfeitorias construtivas do imóvel e janeiro de 2022. Os códigos foram baseados na forma de apresentação dos dados da SPU e podem ser assim resumidos: "0" se não há construções, "1" se acima de 20 anos, "2" se entre 10 e 20 anos, "3" se entre 5 e 10 anos e "4" se menor que 5 anos. Unidade: adimensional (categorias baseadas em anos). Fonte: EB/SPU.

15. **Vida Útil** Variável numérica discreta que representa a vida útil total do imóvel avaliado considerando o tipo de edificação e o tipo de uso do bem. Os valores

[3] `https://www.cub.org.br/`

assumidos por este atributo estão ilustrados na Tabela 4.4. Unidade: anos. Fonte: EB/SPU e *Bureau of Internal Revenue*.

Tabela 4.4: Vida útil dos imóveis.

Tipo de Edificação	**Vida Útil**
Apartamentos	60 anos
Bancos	70 anos
Casas de Alvenaria	65 anos
Casas de Madeira	45 anos
Hotéis	50 anos
Lojas	70 anos
Teatros	50 anos
Armazéns	75 anos
Fábricas	50 anos
Construções Rurais	60 anos
Garagens	60 anos
Edifícios de escritórios	70 anos
Galpões (Depósitos)	70 anos
Silos	75 anos

Fonte: *Bureau of Internal Revenue*.

16. **% Residentes com Ensino Superior Completo AP** Variável numérica contínua que representa o percentual de residentes com ensino superior completo na AP do IBGE que contém o centroide do imóvel dentro de seus limites geográficos de acordo com informações do Censo IBGE 2010. Unidade: adimensional. Fonte: IBGE.

17. **% Domicílios com Rede Geral de Distribuição de Água AP** Variável numérica contínua que representa o percentual de domicílios com acesso à rede geral de distribuição de água na AP do IBGE que contém o centroide do imóvel dentro de seus limites geográficos de acordo com informações do Censo IBGE 2010. Unidade: adimensional. Fonte: IBGE.

18. **% Domicílios com Microcomputador com Acesso à Internet AP** Variável numérica contínua que representa o percentual de domicílios com microcomputador com acesso à internet na AP do IBGE que contém o centroide do imóvel dentro de seus limites geográficos de acordo com informações do Censo IBGE 2010. Unidade: adimensional. Fonte: IBGE.

19. **Renda Domiciliar AP** Variável numérica contínua que representa a renda domiciliar mediana da AP do IBGE que contém o imóvel dentro de seus limites geográficos de acordo com informações do Censo IBGE 2010. Unidade: R$/domicílio. Fonte: IBGE.

20. **Pontos de Interesse API Google Places** Variável numérica discreta que representa a quantidade de pontos de interesse situados em um raio preferencial de 400 metros, correspondente a aproximadamente 5 minutos de caminhada, do centroide do imóvel. Tais pontos foram mapeados por meio de API *Google Places Nearby Search*, com limite de rastreamento de 60 pontos de interesse por par de coordenadas de centroide de imóvel. Os lugares pesquisados foram divididos nas seguintes tipologias: delegacias, hospitais, parques, escolas, *shopping centers*, universidades, atrações turísticas, supermercados, restaurantes, padarias, cafeterias, lojas, paradas de ônibus, estações de veículo leve sobre trilhos (VLT), estações de metrô, estações de trem e aeroportos. Unidade: adimensional (quantidade absoluta). Foram analisadas isoladamente e dentro de dois (2) grupos mutuamente exclusivos: equipamentos urbanos e estabelecimentos comerciais. Fonte: API *Google Places Nearby Search*.

21. **Google Trends** Variável numérica contínua que representa, em uma escala de 0 a 100, a intensidade de busca pelo termo "imóvel" em ferramentas Google no ano de 2022. Analisou-se este atributo no intuito de considerar a demanda por imóveis regionalmente. Os valores extraídos para este atributo podem ser verificados na Tabela 4.5. Unidade: adimensional. Granularidade espacial: Unidade da Federação (UF). Fonte: *Google Trends*.

22. **Valor Total** Variável numérica contínua que representa o valor total do imóvel avaliado em determinada data de referência. Unidade: R$. Fonte: EB/SPU.

23. **Valor Total Atualizado** Variável numérica contínua que representa o valor total do imóvel avaliado atualizado ao mês de janeiro de 2022 pelo índice FipeZap Brasil, respeitando o tipo de uso do bem em questão. Janeiro de 2022 foi definido como a referência temporal em decorrência de todos os atributos analisados terem sido devidamente publicados para o mês. Unidade: R$. Fonte: EB/SPU e FIPE[4].

24. **Valor Unitário** Variável numérica contínua que representa o valor total do imóvel avaliado dividido pela área do terreno. Unidade: R$/m^2. Fonte: EB/SPU.

25. **Valor Unitário Atualizado** Variável numérica contínua que representa o valor unitário do imóvel avaliado atualizado ao mês de janeiro de 2022 pelo índice FipeZap Brasil, respeitando o tipo de uso do bem em questão. Unidade: R$/m^2. Fonte: EB/SPU e FIPE[4].

Já no caso dos imóveis rurais, as seguintes variáveis foram analisadas preliminarmente nos modelos experimentais:

[4] `https://www.fipe.org.br/pt-br/indices/fipezap/`

Tabela 4.5: Intensidade de procura relativa pelo termo "imóvel" em ferramentas de busca *Google* por UF e para o ano de 2022.

UF	**Grau de Interesse Relativo 2022**
Distrito Federal	100
Paraná	94
São Paulo	93
Santa Catarina	91
Minas Gerais	86
Rio de Janeiro	82
Roraima	82
Tocantins	79
Goiás	78
Rio Grande do Sul	77
Mato Grosso do Sul	74
Mato Grosso	71
Sergipe	71
Rondônia	69
Alagoas	64
Espírito Santo	64
Paraíba	62
Bahia	60
Acre	60
Piauí	57
Rio Grande do Norte	56
Amazonas	56
Maranhão	55
Pará	52
Amapá	51
Pernambuco	49
Ceará	48

Fonte: *Google Trends*.

1. **Área do Terreno** Variável numérica contínua que indica a área do terreno do imóvel avaliado. Unidade: hectare (ha). Fonte: SIGPIMA (EB) e SPIUnet (SPU).

2. **Cursos d'Água** Variável dicotômica que recebe valor "1" quando o imóvel contém curso d'água dentro de seus limites e valor "0" caso contrário. Unidade: adimensional. Fonte: Agência Nacional de Águas e Saneamento Básico (ANA).

3. **Acesso Pavimentado** Variável dicotômica que recebe valor "1" quando o imóvel pode ser acessado por via pavimentada e valor "0" caso contrário. Unidade: adimensional. Fonte: Departamento Nacional de Infraestrutura de Transportes (DNIT).

4. **Classe de Potencialidade Agrícola D** Variável numérica contínua que indica a

fração relativa do imóvel com potencial agrícola natural de classe D, que corresponde a terras com restrições muito fortes de potencialidade ao desenvolvimento agrícola e áreas para proteção, preservação e conservação da vegetação. Havendo potencial de exploração mineral, é interessante discutir formas de rentabilização com a Agência Nacional de Mineração (ANM) e uma possível migração da parcela de terra para uma classe de potencialidade agrícola equivalente à mineral em termos econômicos. Unidade: %. Fonte: IBGE.

5. **Classe de Potencialidade Agrícola C** Variável numérica contínua que indica a fração relativa do imóvel com potencial agrícola natural de classe C, que corresponde a terras com restrita potencialidade ao desenvolvimento agrícola. Unidade: %. Fonte: IBGE.

6. **Classe de Potencialidade Agrícola B** Variável numérica contínua que indica a fração relativa do imóvel com potencial agrícola natural de classe B, que corresponde a terras com moderada potencialidade ao desenvolvimento agrícola. Unidade: %. Fonte: IBGE.

7. **Classe de Potencialidade Agrícola A2** Variável numérica contínua que indica a fração relativa do imóvel com potencial agrícola natural de classe A2, que corresponde a terras com boa potencialidade ao desenvolvimento agrícola. Unidade: %. Fonte: IBGE.

8. **Classe de Potencialidade Agrícola A1** Variável numérica contínua que indica a fração relativa do imóvel com potencial agrícola natural de classe A1, que corresponde a terras com muito boa potencialidade ao desenvolvimento agrícola. Unidade: %. Fonte: IBGE.

9. **Distância Zona Urbana** Variável numérica contínua que indica a distância euclidiana entre o centroide do imóvel e a zona urbana mais próxima a ele. Unidade: km. Fonte: EB/SPU e *Google Earth Pro*.

10. **VTN (INCRA)** Variável numérica contínua que indica a pauta de valores de terra nua para fins de titulação de terras em 2022 e é utilizada pelo INCRA para calcular o valor a ser cobrado das parcelas em assentamentos da reforma agrária. Granularidade espacial: município. Unidade: R$/hectare. Fonte: INCRA.

11. **VTN (RFB)** Variável numérica contínua que indica a pauta de valores de terra nua utilizados como parâmetro para declaração de ITR. Granularidade espacial: região rural. Unidade: R$/hectare. Fonte: RFB.

12. **IDHM** O Índice de Desenvolvimento Humano Municipal (IDHM) é uma medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda. O índice numérico contínuo varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento humano. Calculado com base nas informações do Censo 2010. Unidade: adimensional. Granularidade espacial: município. Fonte: IBGE.

13. **Valor Total Atualizado** Variável numérica contínua que representa o valor total do imóvel avaliado atualizado ao mês de janeiro de 2022 pelo índice FipeZap Brasil, por não haver índice rural específico e contemporâneo. Unidade: R$. Fonte: EB/SPU e FIPE[5].

4.2.3 Análise Gráfica de Atributos

Da análise de gráficos *boxplot* univariados, como os constantes nas imagens à esquerda das Figuras 4.5 e 4.6, foi possível perceber que as variáveis *Área do Terreno*, *Área Construída*, *Valor Total Atualizado* e *Valor Unitário Atualizado* são as que apresentam maior grau de dispersão na amostra de dados urbanos, sugerindo necessidade de remoção de *outliers* ou de incidência de transformações matemáticas sobre elas, por serem menos dispersas. Tais transformações serão melhor abordadas na Subseção 4.2.4; visualmente, seus efeitos podem ser constatados nas imagens à direita das Figuras 4.5 e 4.6.

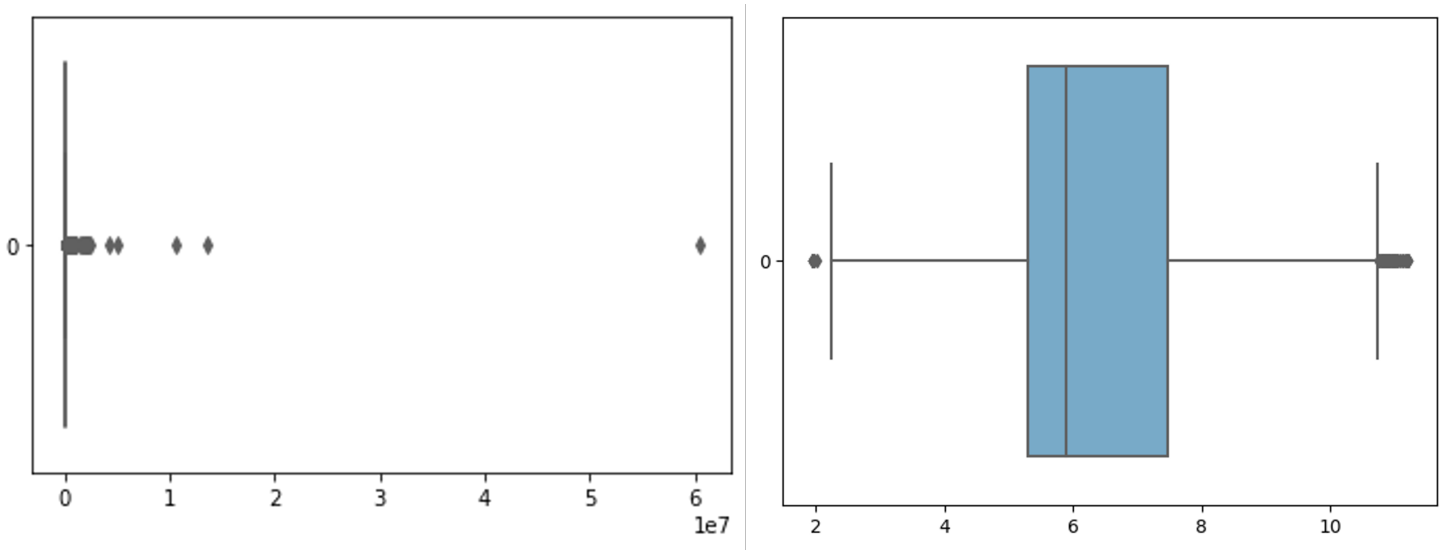

Figura 4.5: *Boxplot* para *Área do Terreno* (à esquerda) e *Ln(Área do Terreno)* (à direita), após eliminação de pontos influenciantes.

[5] https://www.fipe.org.br/pt-br/indices/fipezap/

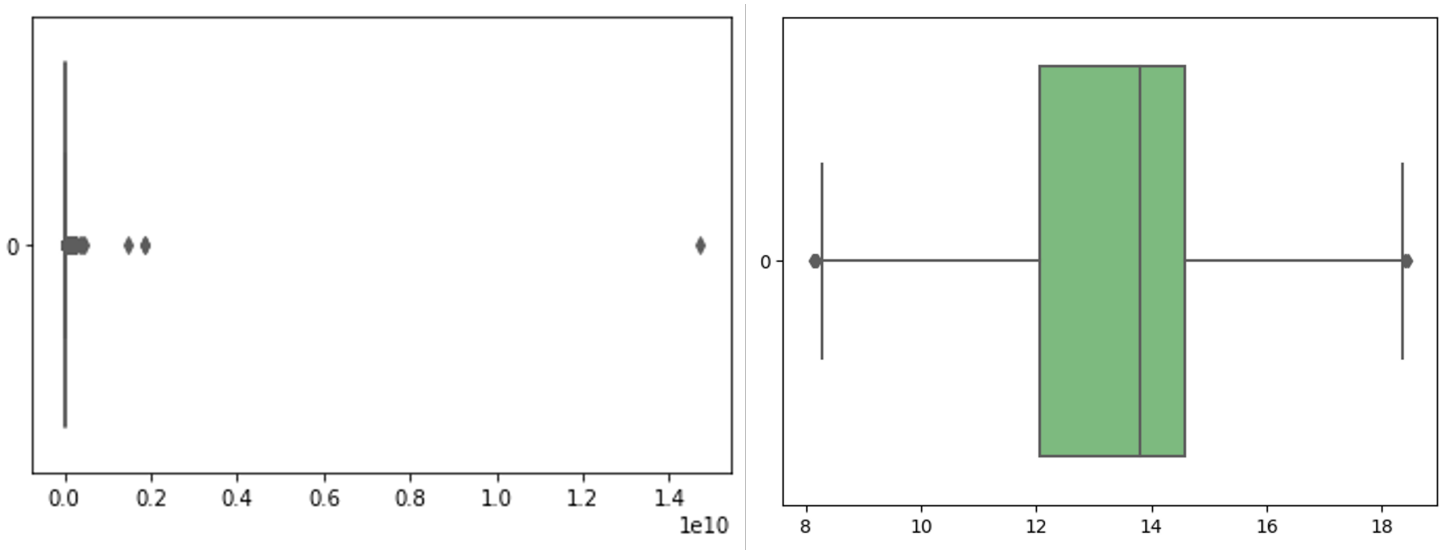

Figura 4.6: *Boxplot* para *Valor Total Atualizado* (à esquerda) e *Ln(Valor Total Atualizado)* (à direita), após eliminação de pontos influenciantes.

Como a variável *Área Construída* apresenta algumas ocorrências de valor nulo (igual a zero), algumas transformações não podem ser a ela aplicadas. Entretanto, verificou-se que a maior parte de sua dispersão se dá em decorrência de 1396 (30% do total) imóveis urbanos avaliados serem do tipo terreno, sem benfeitorias construtivas. A variável dicotômica *Terreno* auxilia na distinção entre os dois grupos na construção dos modelos.

Na Figura 4.7, gráficos de dispersão traçados par a par entre algumas variáveis explicativas e a variáveis dependente *Ln(Valor Total Atualizado)* podem ser observados. Verifica-se, a princípio, que as relações entre as variáveis explicativas presentes e a explicada são positivas, considerando os grupos laranja (imóveis sem construção) e azul (imóveis com benfeitorias construtivas). O comentário se dá a nível preliminar, já que não há clareza de proporcionalidade.

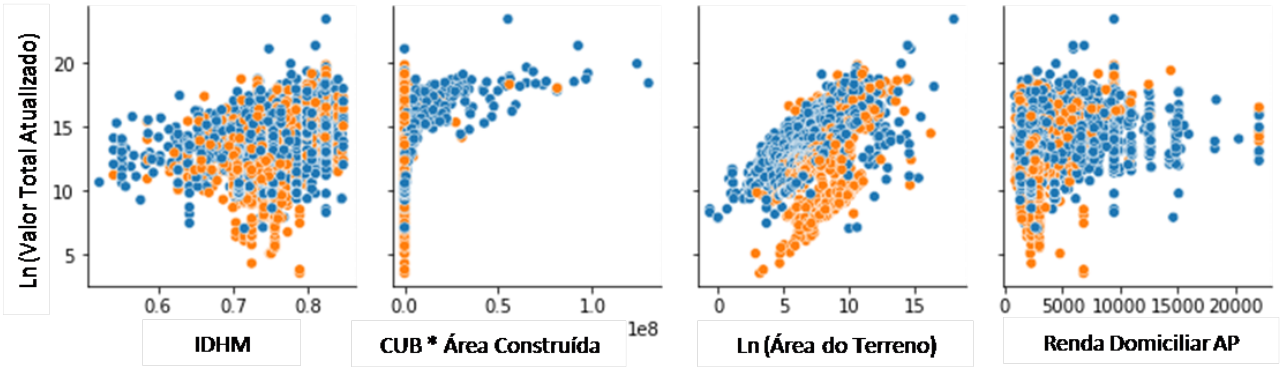

Figura 4.7: Gráficos de dispersão bivariados. Os pontos em cor laranja representam os imóveis sem construções; já os azuis, representam as propriedades com benfeitorias construtivas.

4.2.4 Abordagens

Em decorrência da elevada dispersão observada das variáveis *Área do Terreno*, *Valor Total Atualizado* e *Valor Unitário Atualizado*, foram desenvolvidas três (3) abordagens:

1. **Abordagem 1 ($\mathbf{A_1}$)** Remoção de *outliers* baseada na variável *Valor Unitário Atualizado* (R\$/m^2) e utilização deste atributo como variável dependente.
2. **Abordagem 2 ($\mathbf{A_2}$)** Transformação da variável *Área do Terreno* com logaritmo neperiano (*ln*), sem remoção de *outliers* e *ln* de *Valor Total Atualizado* como variável dependente.
3. **Abordagem 3 ($\mathbf{A_3}$)** Transformação do atributo *Valor Unitário Atualizado* com *ln*, remoção da variável *Área do Terreno* e *ln* de *Valor Unitário Atualizado* como variável dependente.

As três abordagens foram preliminarmente implementadas com a técnica de aprendizagem *SGDRegressor* da biblioteca *scikit-learn* versão 1.2.2 em linguagem *Python* e considerando apenas variáveis genéricas aos imóveis.

Considerando os objetivos de (i) maximizar a precisão das estimativas e de (ii) ser o mais abrangente possível a nível geográfico de país (Brasil), a abordagem matemática A_2 apresentou raiz do erro quadrático médio ($RMSE$) inferior e coeficiente de determinação (R^2) superior àqueles obtidos com as abordagens A_1 e A_3.

Decidiu-se, portanto, realizar os experimentos relativos aos níveis mais específicos, linearmente e não linearmente, utilizando as transformações propostas em A_2.

4.2.5 Modelagens

Os três níveis de modelagem urbana propostos utilizaram os atributos constantes na Tabela 4.7 após rigoroso processo de seleção de variáveis, sobre o qual se discorrerá na Subseção 4.2.6, e utilizaram as transformações matemáticas da abordagem A_2.

Os modelos urbanos básico (U_1), intermediário (U_2) e específico (U_3) 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 [33], pacote *spreg* versão 1.3.0 da biblioteca *PySAL* em *Python*;
- *SGDRegressor*, biblioteca *scikit-learn* versão 1.2.2 em *Python*;
- *ANN MLPRegressor*, biblioteca *scikit-learn* versão 1.2.2 em *Python*; e

- *XGBRegressor*, biblioteca *XGBoost* versão 1.7.4 em *Python*.

Já os modelos rurais básico e específico (R_1, R_2) foram implementados considerando os atributos listados na Tabela 4.9 com uso dos seguintes algoritmos:

- Regressão Linear Múltipla (OLS), biblioteca *statsmodels* versão 0.14.0 em *Python*; e
- Regressão Espacial *Spatial Two Stage Least Squares* (S2SLS), conforme proposto por Anselin [33], pacote *spreg* versão 1.3.0 da biblioteca *PySAL* em *Python*.

Tendo em vista o reduzido número de instâncias rurais coletadas, não haveria coerência na aplicação de modelos de aprendizagem de máquina à amostra disponível. Em se tratando de imóveis rurais, a discussão de resultados no Capítulo 5 se dá em teor mais qualitativo que quantitativo.

Para os modelos de regressão espacial, foram criadas matrizes de pesos espaciais W fundamentadas no livro de Anselin et al. [34]. 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 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 autocorrrelação espacial, a influência que o valor de um imóvel exerce sobre os valores de seus vizinhos matriciais e vice-versa. A Figura 4.8 ilustra um grafo de conectividades gerado a partir da definição de δ no *software* GeoDa[6], disponibilizado gratuitamente pelo *The Center for Spatial Data Science*, da Universidade de Chicago.

Já para a amostra de imóveis rurais, a distância de 21 km se mostrou uma boa largura de banda (após testes realizados entre 1 km e 40 km). Os pesos w_{ij} da matriz espacial foram calculados com função de decaimento igual ao inverso da distância entre os imóveis, conforme Equações 4.1 e 4.2. As matrizes W foram implementadas com uso da biblioteca *PySAL* em linguagem *Python* no ambiente *Google Colaboratory* (Colab).

$$w_{ij} = \frac{1}{d_{ij}} \quad se \quad d_{ij} < \delta \tag{4.1}$$

$$w_{ij} = 0 \quad se \quad d_{ij} \geq \delta \tag{4.2}$$

[6] `https://spatial.uchicago.edu/geoda`

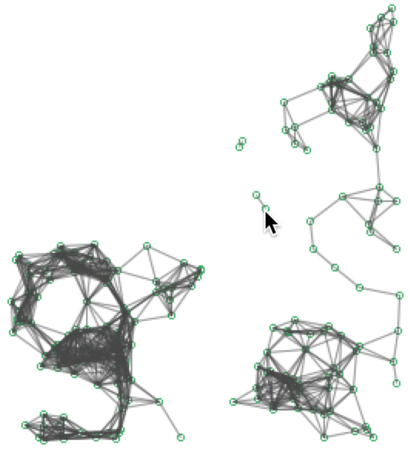

Figura 4.8: Grafo de conectividades gerado a partir de definição de largura de banda.

4.2.6 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.

4.2.6.1 Seleção de Atributos Urbanos

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 PLS-SEM, denominada critério C_4. À exceção do PLS-SEM, implementado com uso do *software SmartPLS 4*[7], as demais técnicas foram implementadas com as bibliotecas *pandas* e *scikit-learn* em linguagem *Python*.

A aplicação dos algoritmos de seleção resultou nos atributos apresentados na Tabela 4.6 e nos modelos conceituais PLS-SEM representados nas Figuras 4.9, 4.10 e 4.11, nas quais 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,

[7] `https://www.smartpls.com/`

as moderadoras e a endógena. Os indicadores destacados com contorno em cor laranja são aqueles adicionados em relação ao nível anterior.

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(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 4.11, percebe-se que se trata da conexão mais forte do modelo estrutural U_3.

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 [35], todos os $p - valores$ associados a seus testes de hipóteses e àqueles dos pesos e cargas externas do modelo externo foram inferiores a 5%.

Os coeficientes de determinação dos modelos conceituais PLS-SEM U_1, U_2 e U_3 de 57,1%, 59,0% e 62,0%, respectivamente, sugerem poder explicativo moderado, de acordo com Hair [36]. 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.

A Tabela 4.7 contém a ocorrência de atributos relacionados a imóveis urbanos em cada um dos níveis de modelagem propostos.

4.2.6.2 Seleção de Atributos de Imóveis Rurais

Como a quantidade de instâncias de imóveis rurais com valor de avaliação é pequena, optou-se por identificar os subconjuntos significantes por meio das seguintes técnicas: análise de correlações bivariadas (máximo de 80% de correlação de Pearson, conforme indicado na NBR 14653), denominada critério C_1; e regressão *forward-backward stepwise* (limites *p-valor* de entrada e de saída de 30%, fundamentados na NBR 14653), denominada critério C_2. A aplicação dos algoritmos de seleção resultou nos atributos da Tabela 4.8.

Na Tabela 4.9, pode ser verificada a ocorrência de variáveis em cada um dos dois níveis de modelagem rural.

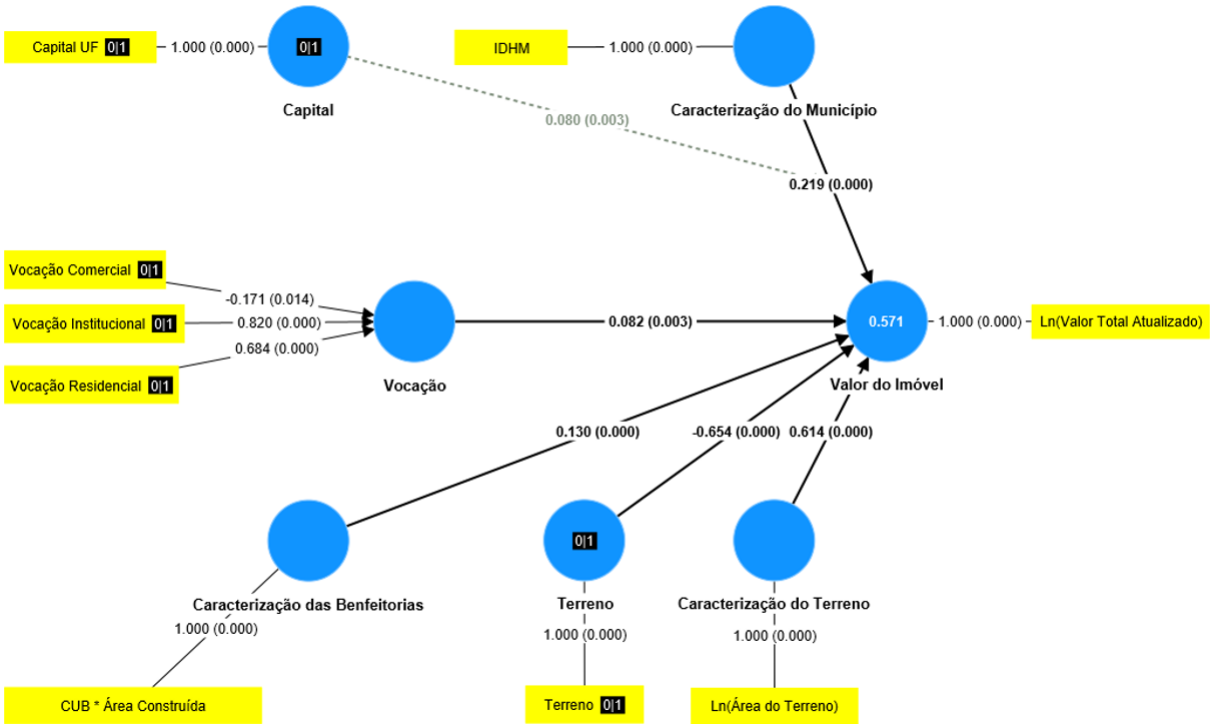

Figura 4.9: Modelo conceitual U_1 PLS-SEM.

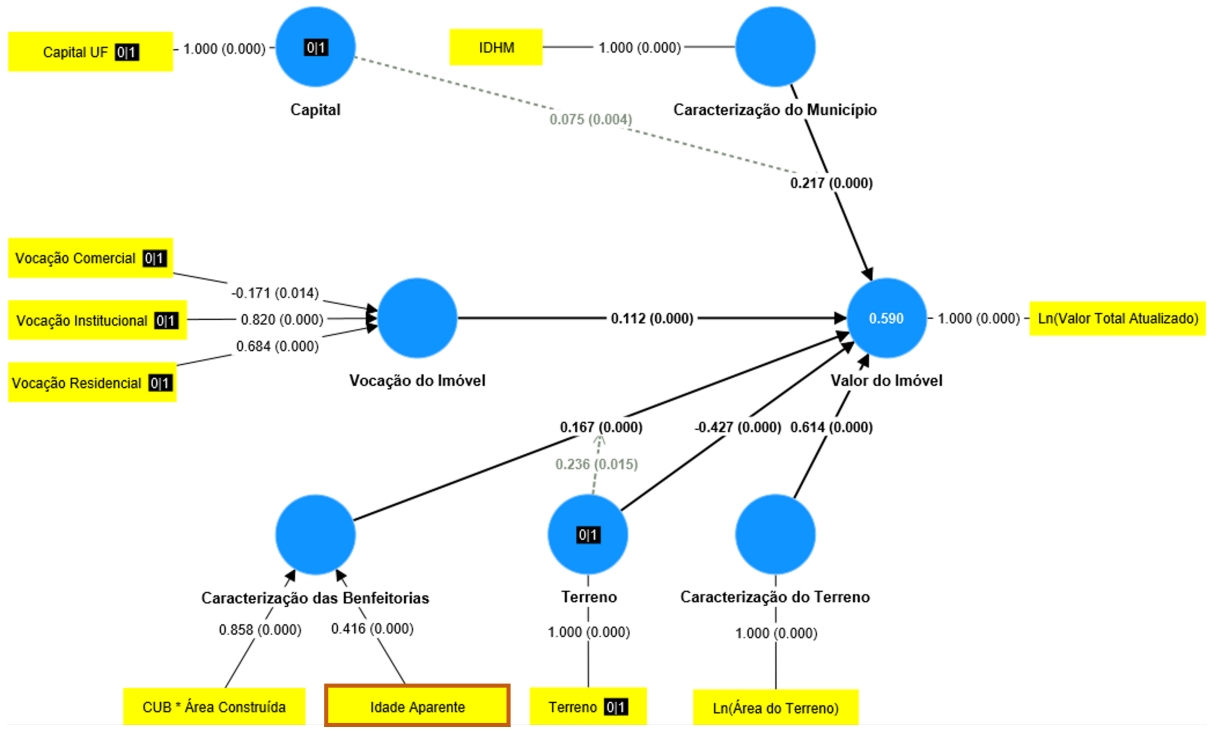

Figura 4.10: Modelo conceitual U_2 PLS-SEM.

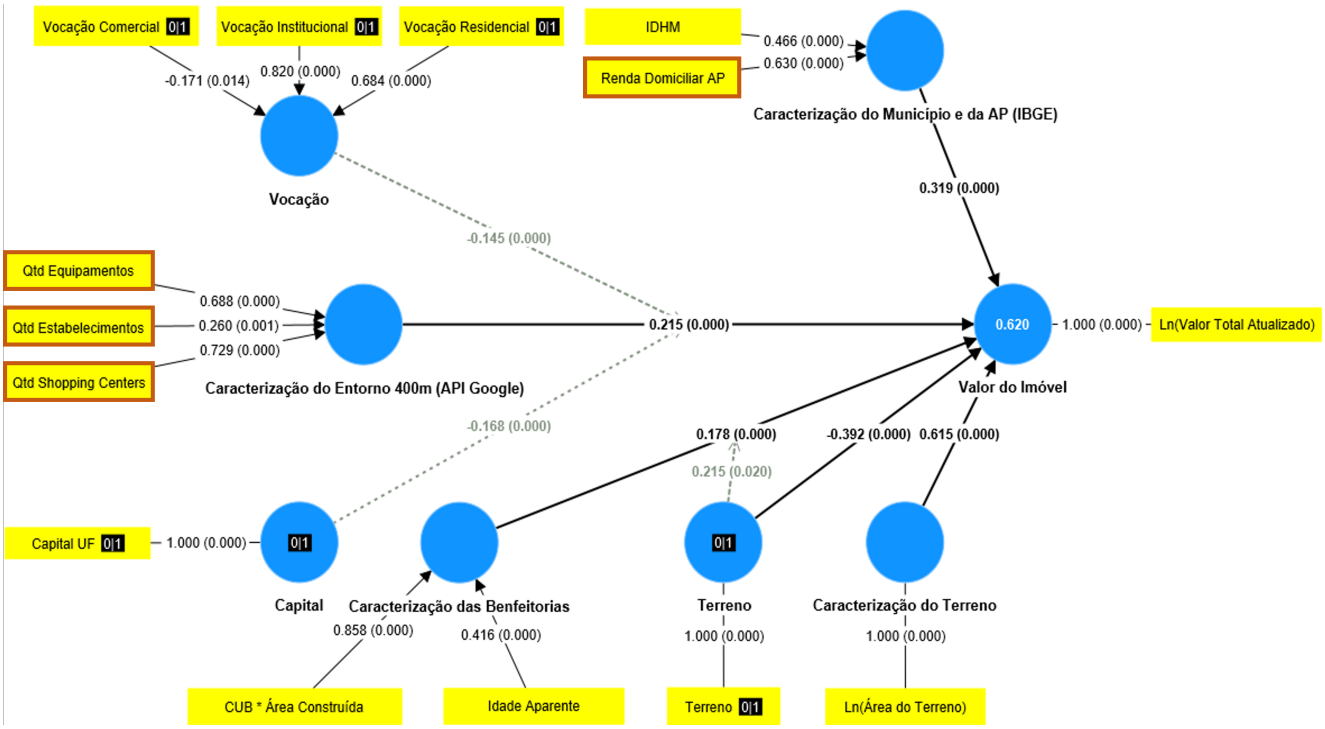

Figura 4.11: Modelo conceitual U_3 PLS-SEM.

Tabela 4.6: Ocorrência de atributos nas modelagens finais de imóveis urbanos.

Variável	**Tipo**	**Fonte**	**Critério de Exclusão**
Capital UF	binária	IBGE	-
Tipologia Municipal	cód. binários	IBGE	C_2
Grau de Urbanização	numérica	IBGE	C_4
IDHM	numérica	IBGE	-
IVS	numérica	Ipea	C_4
Taxa de Homicídios	numérica	Ipea	C_4
Vocação do Imóvel	cód. binários	EB/SPU	-
Terreno	binária	EB/SPU	-
Ln(Área do Terreno)	numérica	EB/SPU	-
*CUB * Área Construída*	numérica	Sinduscon[a]/EB/SPU	-
Idade Aparente	cód. alocados	EB/SPU	-
Vida Útil	numérica	BIR[b]	C_4
% Residentes com Ensino Superior Completo AP	numérica	IBGE	C_4
% Domicílios com Rede Geral de Distribuição de Água AP	numérica	IBGE	C_4
% Domicílios com Microcomputador com Acesso à Internet AP	numérica	IBGE	C_1
Renda Domiciliar AP	numérica	IBGE	-
Pontos de Interesse API Google Places[c]	numérica	API *Google*	-
Pontos de Interesse Excluídos API Google Places[d]	numérica	API *Google*	C_4
Google Trends	numérica	*Google Trends*	C_4
Coordenadas Geográficas[e]	numérica	EB/SPU	C_3
Ln(Valor Total Atualizado)	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

Tabela 4.7: Ocorrência dos atributos selecionados nos diferentes níveis de modelos urbanos.

Variável	**Nível de Modelagem**
Capital UF	U_1, U_2, U_3
Vocação do Imóvel	U_1, U_2, U_3
IDHM	U_1, U_2, U_3
Ln(Área do Terreno)	U_1, U_2, U_3
*CUB * Área Construída*	U_1, U_2, U_3
Idade Aparente	U_2, U_3
Renda Domiciliar AP	U_3
Pontos de Interesse API Google Places	U_3
Ln(Valor Total Atualizado)	U_1, U_2, U_3

Tabela 4.8: Ocorrência de atributos nas modelagens finais de imóveis rurais.

Variável	**Tipo**	**Fonte**	**Critério de Exclusão**
Ln(Área do Terreno)	numérica	EB/SPU	-
Cursos d'Água	cód. binários	ANA	-
Acesso Pavimentado	cód. binários	DNIT	-
Classe de Potencialidade Agrícola D	numérica	IBGE	C_1
Classe de Potencialidade Agrícola C	numérica	IBGE	C_2
Classe de Potencialidade Agrícola B	numérica	IBGE	C_2
Classe de Potencialidade Agrícola A2	numérica	IBGE	-
Classe de Potencialidade Agrícola A1	numérica	IBGE	C_1
Distância Zona Urbana	numérica	EB/SPU	-
VTN (INCRA)	numérica	INCRA	C_2
VTN (RFB)	numérica	RFB	-
IDHM	numérica	IBGE	-
Ln(Valor Total Atualizado)	numérica	EB/SPU	-

Tabela 4.9: Ocorrência dos atributos selecionados nos diferentes níveis de modelos rurais.

Variável	**Nível de Modelagem**
Ln(Área do Terreno)	R_1, R_2
Cursos d'Água	R_1, R_2
Acesso Pavimentado	R_1, R_2
Classe de Potencialidade Agrícola A2	R_1, R_2
Distância Zona Urbana	R_1, R_2
VTN (RFB)	R_2
IDHM	R_2
Ln(Valor Total Atualizado)	R_1, R_2

4.2.7 Tratamento de Variáveis

Foram aplicados testes de normalidade numéricos aos atributos explicativos selecionados. Tanto o teste de Shapiro-Wilk quanto o teste de Jarque-Bera indicaram, pelos *p-valores* obtidos, que há evidências de que as variáveis têm assimetria e curtose significativamente diferentes de uma distribuição normal.

Optou-se, portanto, pela aplicação de transformação do tipo *MinMaxScaler*, utilizando a biblioteca *scikit-learn* versão 1.2.2 na linguagem *Python* às variáveis explicativas.

Na fase de pré-processamento, portanto, os atributos foram reescalados entre 0 e 1. Tal etapa evita que variáveis com maior amplitude de variação influenciem demasiadamente o modelo e não prejudica a interpretabilidade, por se tratarem de transformações inversíveis, em se tendo conhecimento e acesso pleno ao conjunto de dados, caso deste trabalho.

4.2.8 Tratamento de Instâncias Coletadas

A remoção de *outliers* foi realizada por meio do Método de Tukey com limite inferior estabelecido como "$q1 - 1,5 * (intervalointerquatil)$" e limite superior como "$q3 + 1,5 * (intervalointerquatil)$". Tal método foi utilizado com base na variável *Valor Unitário Atualizado* na abordagem A_1.

Para todas as modelagens, foram identificados pontos influenciantes por meio do cálculo da distância de Cook, conforme Equação 4.3.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}j - \hat{Y}j(i))^2}{p * \text{MSE}} \tag{4.3}$$

para a qual D_i é a distância de Cook para a observação i, n são as instâncias do conjunto de dados, $\hat{Y}j$ é o valor predito por um modelo para a observação j, $\hat{Y}j(i)$ é o valor predito por um modelo para a observação j após a remoção da observação i, p é o número de parâmetros ou de variáveis explicativas no modelo e *MSE* é o erro quadrático médio.

R. Dennis Cook [37] define o ponto de corte $D_i > 1$ como um bom limite operacional para identificar pontos influenciantes.

Ao ser aplicado à amostra de dados urbanos do EB e da SPU, três instâncias foram identificadas como influenciantes, todas da SPU, sendo removidas do conjunto efetivamente utilizado nos experimentos.

A aplicação do algoritmo à amostra rural implicou a eliminação de somente uma ocorrência, administrada pelo EB. A despeito de sua distância de Cook ser inferior ao limite operacional, mas por ser próxima a ele, realizamos alguns testes e os modelos rurais tiveram comportamento mais consistente sem a referida instância.

As Figuras 4.12 e 4.13 ilustram as distâncias de Cook calculadas e que levaram aos conjuntos finais de 4595 imóveis urbanos e 23 imóveis rurais.

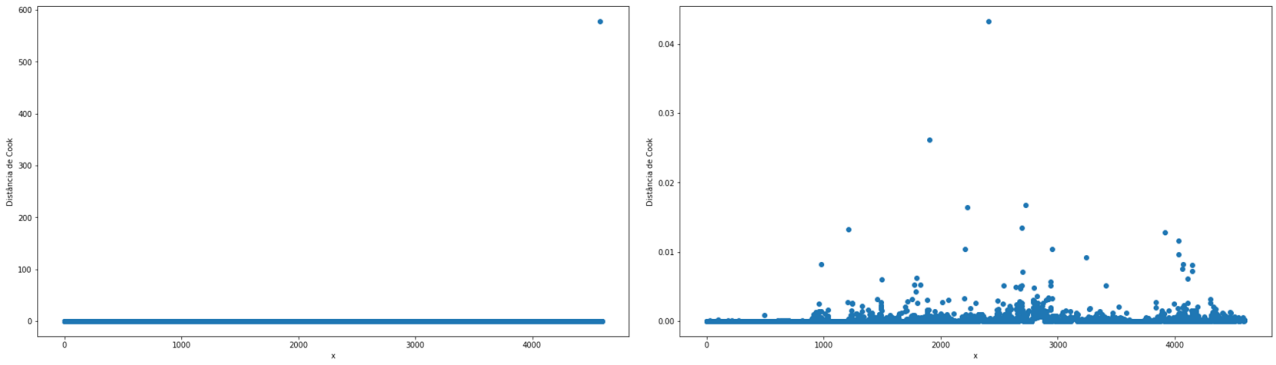

Figura 4.12: Distâncias de Cook calculadas antes (à esquerda) e depois (à direita) da remoção de instâncias influenciantes urbanas.

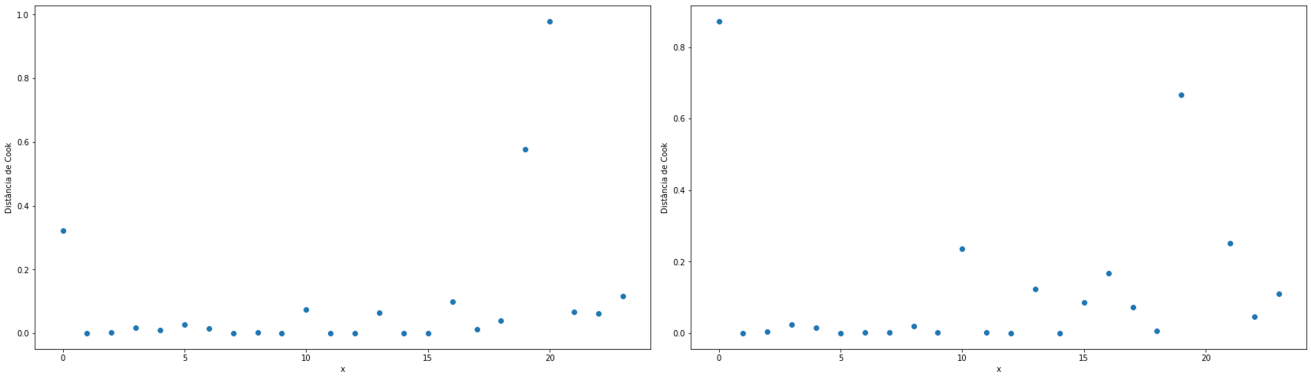

Figura 4.13: Distâncias de Cook calculadas antes (à esquerda) e depois (à direita) da remoção de instâncias influenciantes rurais.

4.2.9 Métricas de Avaliação

As principais métricas de avaliação dos modelos implementados foram o coeficiente de determinação (R^2), conforme Equação 4.4, e a raiz do erro quadrático médio ($RMSE$), conforme Equação 4.5, respeitando-se a necessidade de interpretabilidade e limites de tempo de processamento razoáveis.

$$R^2(Y) = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{4.4}$$

em que n é o número de instâncias do conjunto de dados considerado, Y_i é o valor observado da variável dependente característica do elemento i da amostra, $\hat{Y}_i$ é o valor projetado pelo modelo para a variável explicada de i e $\bar{Y}$ é a média dos valores observados.

$$RMSE(Y) = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{4.5}$$

em Y_i é o valor observado e $\hat{Y}_i$ é o valor calculado pelo modelo avaliado.

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.

4.2.10 Intervalo de Confiança

Os intervalos de confiança (IC) foram estimados pressupondo normalidade dos resíduos e fundamentado na distribuição t de Student bicaudal, de acordo com a Equação 4.6.

$$res_{max,min} = res_{medio} \pm \frac{t * \sigma}{\sqrt{n}} \tag{4.6}$$

na qual, res_{max} é o limite superior dos resíduos, res_{min} é o limite inferior dos resíduos, res_{medio} é a média dos resíduos calculados pelo modelo e σ é o desvio padrão dos resíduos ou erros do modelo. O valor crítico t foi determinado em função do nível de confiança de 90% e de $(n-1)$ graus de liberdade, em que n representa do número de instâncias analisadas.

Considerando a transformação matemática ln aplicada à variável *Valor do Imóvel Atualizado*, denotada como y em sua forma original e como Y em sua forma transformada, na abordagem selecionada A_2, tem-se as Equações 4.7 e 4.8.

$$\frac{\hat{y}_{max,90\%}}{\hat{y}} = e^{(res_{medio} + \frac{t*\sigma}{\sqrt{n}})} \tag{4.7}$$

$$\frac{\hat{y}_{min,90\%}}{\hat{y}} = e^{(res_{medio} - \frac{t*\sigma}{\sqrt{n}})} \tag{4.8}$$

Capítulo 5

Experimentos e Resultados

Neste capítulo, são apresentados os dados efetivamente utilizados, a configuração dos experimentos, os cenários de aplicação, os resultados dos modelos e uma análise das saídas à luz das principais métricas de avaliação adotadas.

5.1 Bases de Dados e Cenários de Aplicação

Conforme etapas de pré-processamento discutidas no Capítulo 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), interseção espacial com camada de potencialidade agrícola natural do solo do IBGE (no caso de imóveis rurais) e detecção e remoção de elementos influenciantes da amostra levaram a reduções quantitativas das bases originais.

A amostra urbana efetivamente utilizada nos experimentos descritos nas seções a seguir contemplou 4595 instâncias; já a rural, 23 ocorrências. As Figuras 5.1 e 5.2 ilustram a distribuição geográfica dos elementos da amostra urbana, com destaque para o Distrito Federal, que concentra aproximadamente 29% dos dados. A Figura 5.2 ilustra, também, a densidade regional de dados rurais utilizados nos experimentos.

Como a base final rural conta com um número muito reduzido de dados e, considerando que cada imóvel possui características próprias que o conferem unicidade, pretende-se apresentar como resultado tão somente uma proposta de metodologia. As métricas de avaliação apresentadas para o caso rural (R^2 e $RMSE$) são ilustrativas.

As correlações de Pearson entre as variáveis explicativas utilizadas nos experimentos deste capítulo e listadas nas Tabelas 4.7 e 4.9 foram calculadas e encontram-se ilustradas nas Figuras 5.3 e 5.4. Não se observam correlações superiores a 80%, o que representa um atendimento aos requisitos normativos da NBR 14653.

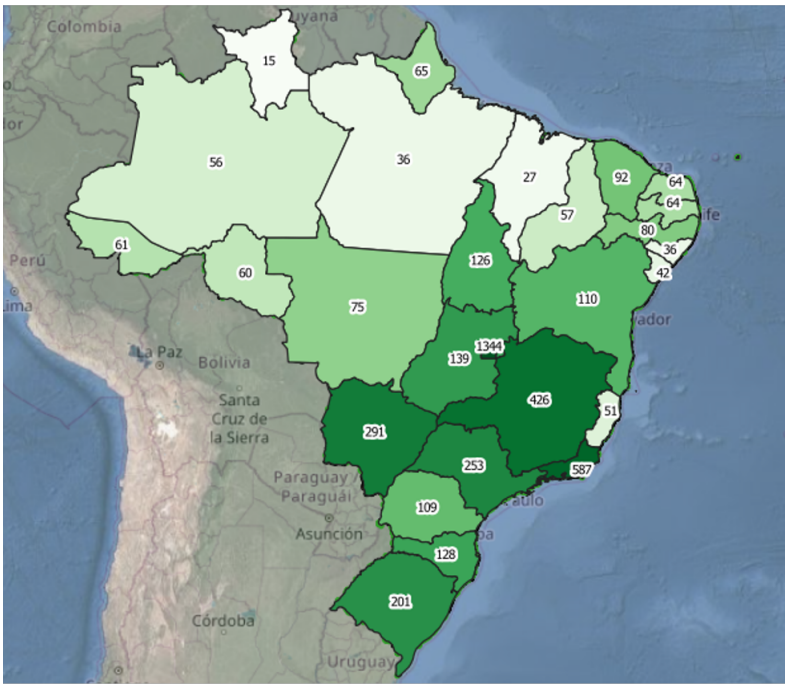

Figura 5.1: Distribuição das 4595 instâncias urbanas por UF com suas respectivas quantidades absolutas.

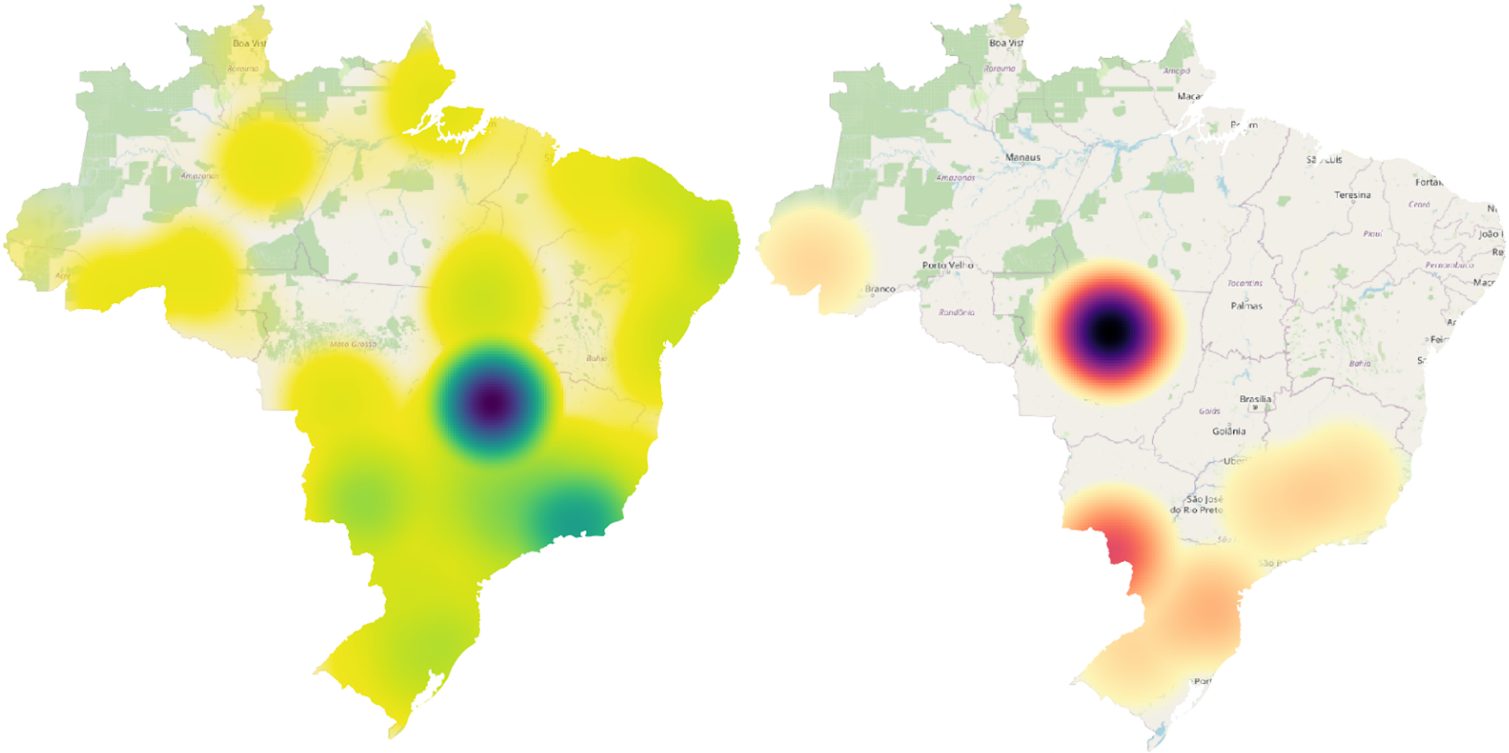

Figura 5.2: Mapa de densidade dos imóveis que compõem a amostra urbana (à esquerda) e rural (à direita).

A aplicação dos algoritmos de aprendizagem de máquina ao conjunto de dados urbanos foi realizada em 2 cenários: **Cenário A** (treinamento, validação e teste em base de dados híbrida do EB e da SPU) e **Cenário B** (treinamento e validação em base híbrida e teste em base exclusiva do EB composta por 130 instâncias, aproximadamente 50% do total de ocorrências com valores observados de imóveis administrados pelo EB).

5.2 Configuração dos Experimentos

A Tabela 5.1 contém a configuração dos principais parâmetros para os três níveis de modelo urbano 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 5.1: Parâmetros utilizados nos modelos de aprendizagem de máquina.

Parâmetro	***SGDRegressor* linear**	***ANN MLPRegressor* não linear**	***XGBRegressor* 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

A Figura 5.5 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.

Já para os imóveis rurais, as regressões OLS e espacial foram aplicadas em cenário único, com os 23 imóveis da base mista EB/SPU.

5.3 Resultados Obtidos

Os resultados obtidos neste trabalho são apresentados para o conjunto de dados urbanos e rurais, separadamente, com foco nas métricas de avaliação indicadas na Subseção 4.2.9.

5.3.1 Imóveis Urbanos

Foram obtidos os coeficientes de determinação (R^2) e as raízes dos erros quadráticos médios (*RMSE*) especificados nas Tabelas 5.2, 5.3 e 5.4 para os modelos U_1, U_2 e U_3, respectivamente.

Figura 5.3: Gráfico de calor de correlações entre as variáveis explicativas quantitativas urbanas já tratadas e selecionadas.

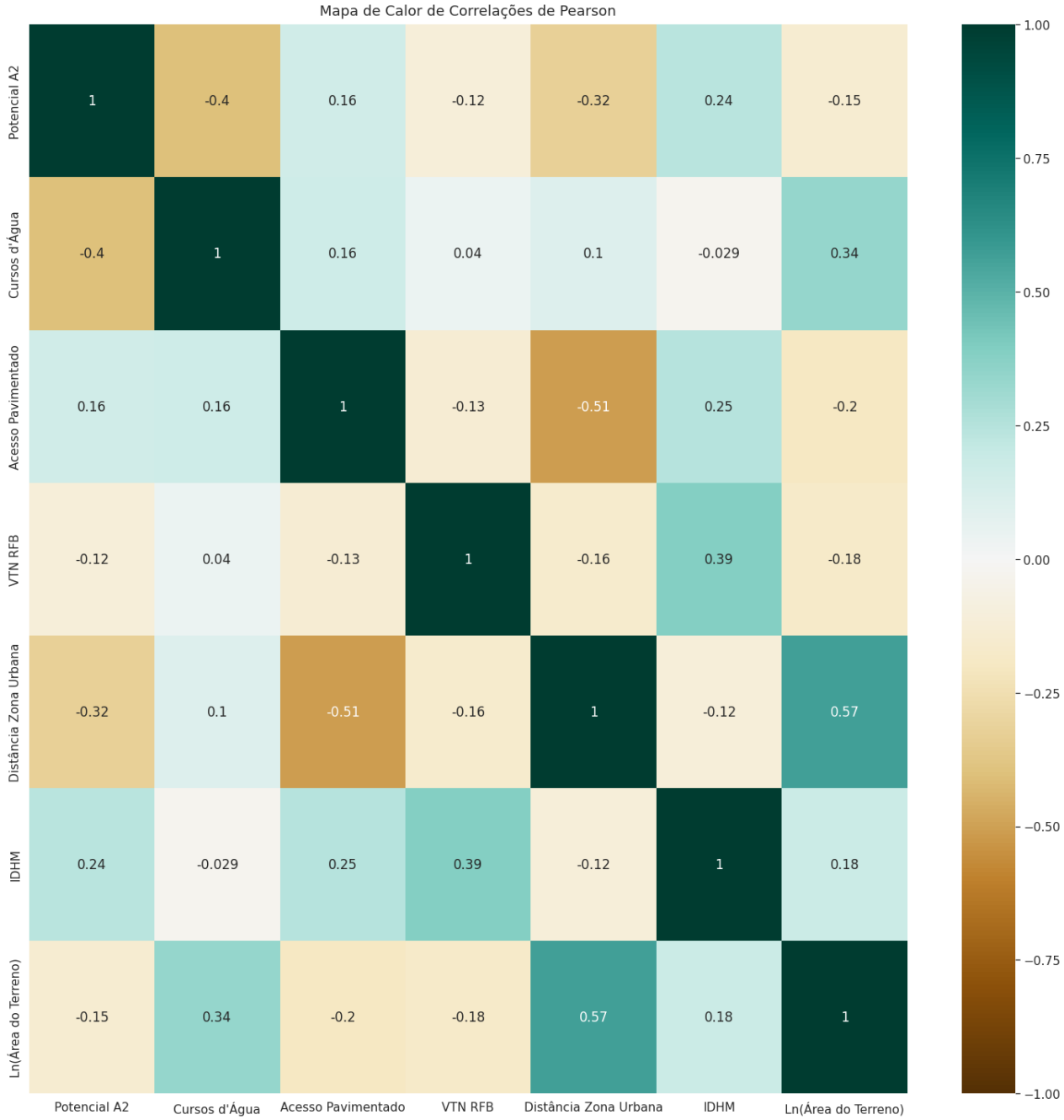

Figura 5.4: Gráfico de calor correlações entre as variáveis explicativas rurais já tratadas e selecionadas.

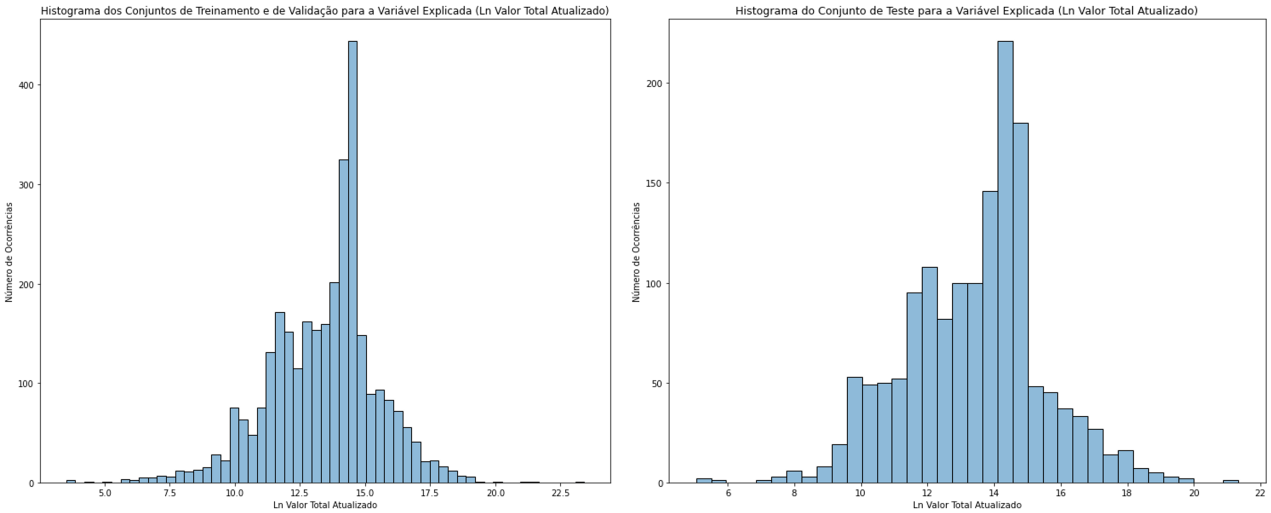

Figura 5.5: Histograma da variável dependente *Ln(Valor Total Atualizado)* nos conjuntos de treinamento e de validação (à esquerda) e de teste (à direita).

Tabela 5.2: Resultados das métricas de avaliação dos modelos urbanos básicos para os dois cenários considerados.

Modelos U_1 (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	57,1%	49,8%	1,36	1,82
Regressão Espacial (A)	NA	NA	58,7%	51,1%	1,33	1,45
SGDRegressor linear (A)	56,4%	5,9%	57,0%	56,8%	1,35	1,38
SGDRegressor linear (B)	55,2%	6,2%	55,5%	48,1%	1,39	1,47
ANN não linear (A)	66,1%	4,8%	67,8%	66,3%	1,17	1,22
ANN não linear (B)	69,5%	2,8%	70,1%	69,6%	1,14	1,05
XGBRegressor não linear (A)	83,3%	2,0%	91,5%	80,3%	0,58	0,94
XGBRegressor não linear (B)	83,8%	3,6%	91,9%	81,8%	0,59	0,81

NA: não se aplica

A maior concentração de pontos próximo às bissetrizes, destacadas em vermelho nos gráficos constantes na Figura 5.6, aliada às medidas de dispersão expostas nas tabelas desta subseção, são indícios de razoável assertividade das predições dos modelos não lineares mais específicos. 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.

Para os modelos básico, intermediário e específico 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 4.2.5, são significantes.

Tabela 5.3: Resultados das métricas de avaliação dos modelos urbanos intermediários para os dois cenários considerados.

Modelos U_2 (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	59,8%	50,1%	1,32	1,78
Regressão Espacial (A)	NA	NA	61,9%	54,1%	1,28	1,41
SGDRegressor linear (A)	59,3%	5,2%	59,5%	59,1%	1,31	1,34
SGDRegressor linear (B)	58,7%	5,5%	58,3%	48,9%	1,33	1,40
ANN não linear (A)	67,9%	3,8%	69,6%	66,7%	1,14	1,21
ANN não linear (B)	69,0%	3,2%	69,2%	69,1%	1,15	1,16
XGBRegressor não linear (A)	83,6%	2,2%	92,5%	80,5%	0,57	0,93
XGBRegressor não linear (B)	84,4%	3,3%	92,0%	81,7%	0,59	0,82

NA: não se aplica

Tabela 5.4: Resultados das métricas de avaliação dos modelos urbanos específicos para os dois cenários considerados.

Modelos U_3 (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
SGDRegressor linear (A)	61,2%	4,7%	61,9%	61,1%	1,27	1,31
SGDRegressor 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
XGBRegressor não linear (A)	86,2%	1,9%	95,7%	83,6%	0,43	0,85
XGBRegressor não linear (B)	86,7%	3,5%	94,8%	80,6%	0,47	0,84

NA: não se aplica

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 para cada um dos três níveis. As semiamplitudes inferiores e superiores dos intervalos calculados encontram-se na Tabela 5.5.

5.3.2 Imóveis Rurais

As regressões lineares múltipla e espacial aplicadas ao contexto de imóveis rurais tiveram a probabilidade de nulidade simultânea de todos os regressores calculada por meio de teste F de Snedecor. Todas os testes indicaram chance inferior a 0,1%.

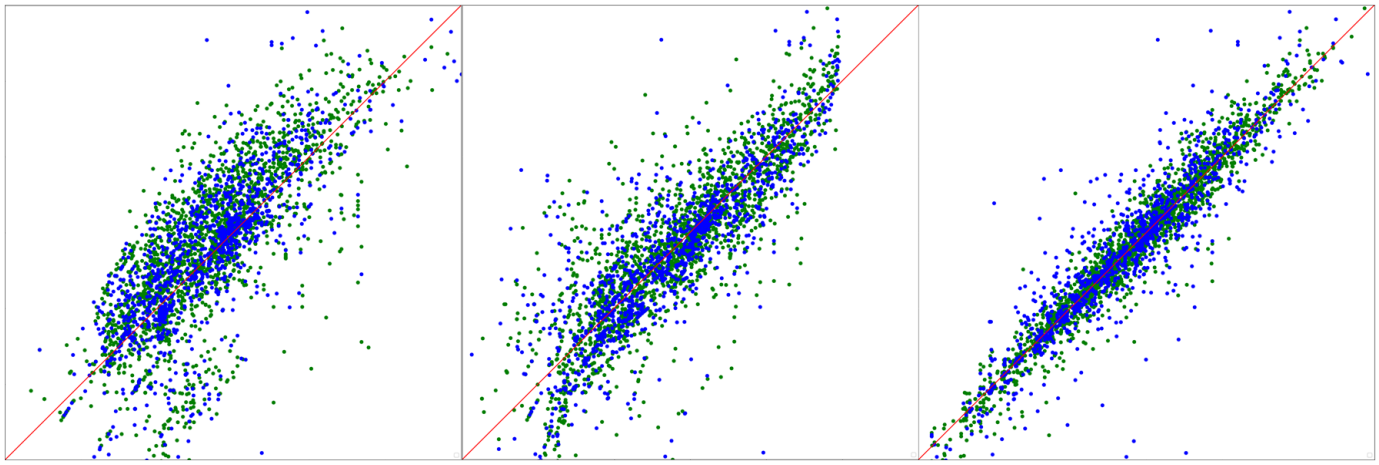

Figura 5.6: Gráficos de dispersão de *ln* de valores observados (ordenadas) versus *ln* de valores projetados pelos modelos U_3 (abscissas). Implementado com os algoritmos *SGDRegressor* (à esquerda), *ANN MLPRegressor* (ao centro) e *XGBRegressor* (à direita).

Tabela 5.5: Semiamplitudes dos intervalos de confiança 90% calculados para os modelos urbanos multiníveis.

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

A distribuição dos resíduos se assemelha à distribuição normal, pelo teste de Jarque-Bera. Os modelos se mostraram homocedásticos, pelo teste de Breusch-Pagan, e inconclusivos quanto à autocorrelação dos erros, pelo teste de Durbin-Watson.

A regressão OLS foi aplicada à amostra rural e calculou-se a diferença relativa entre os valores observados dos imóveis y_i e os valores $\hat{y}_i$ estimados pelos modelos R_1 e R_2, calculados como frações de $\hat{y}_i$. Ou seja, os valores constantes na Tabela 5.6 representam fatores que, multiplicados por $\hat{y}_i$, devem ser a ele acrescidos para se chegar ao valor de avaliação y_i. Os resultados da aplicação do modelo OLS R_2 caso a caso estão registrados na Tabela 5.7; destacando-se em vermelho as instâncias para as quais o modelo sobreprecificou os imóveis, ou seja, cujos valores de IC mínimo foram superiores a y_i. Tais casos correspondem a aproximadamente 13% da amostra.

A autocorrelação global espacial foi calculada 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 21 km,

Tabela 5.6: Diferenças relativas entre entre os valores observados dos imóveis y_i e valores $\hat{y}_i$ projetados pelos modelos, calculados como alíquotas de $\hat{y}_i$. Valores positivos indicam valores observados maiores que predições.

Instância	**OLS R_1**	**OLS R_2**
1	-38,5%	+35,7%
2	+136,1%	-2,4%
3	-14,8%	-19,1%
4	-32,2%	-37,1%
5	-27,5%	-29,9%
6	+52,0%	+4,2%
7	+29,2%	-11,7%
8	-7,4%	+16,6%
9	+23,8%	+34,6%
10	+3,2%	+11,8%
11	-70,2%	-76,3%
12	-8,2%	-18,6%
13	-23,0%	-12,8%
14	+209,9%	+246,7%
15	+9,3%	+1,9%
16	+8,0%	+100,9%
17	-73,6%	-56,11%
18	+99,1%	+169,8%
19	-65,4%	+17,5%
20	-90,6%	-68,1%
21	+466,7%	-26,7%
22	+390,6%	+88,6%
23	+95,7%	+79,5%
Média	+46,5%	+19,5%
Desvio Padrão	137,2%	74,5%

Tabela 5.7: Valores observados dos imóveis y_i e valores $\hat{y}_i$ estimados pelo modelo OLS R_2. Encontram-se tabulados as medidas de tendência central dos valores projetados, os mínimos e os máximos calculados.

Instância	Valor Observado	Valor Estimado Médio OLS R_2	Valor Estimado Mínimo OLS R_2	Valor Estimado Máximo OLS R_2
1	R$ 834 714,47	R$ 615 342,72	R$ 368 590,29	R$ 1 055 312,77
2	R$ 1 199 179,41	R$ 1 228 625,00	R$ 735 946,38	R$ 2 107 091,88
3	R$ 874 828,62	R$ 1 080 895,60	R$ 647 456,46	R$ 1 853 735,95
4	R$ 766 829,19	R$ 1 219 336,29	R$ 730 382,44	R$ 2 091 161,74
5	R$ 261 636,01	R$ 373 404,37	R$ 223 669,22	R$ 640 388,49
6	R$ 3 966 069,63	R$ 3 806 030,75	R$ 2 279 812,42	R$ 6 6 527 342,74
7	R$ 2 879 788,85	R$ 3 261 182,81	R$ 1 953 448,51	R$ 5 592 928,53
8	R$ 836 382,03	R$ 717 428,00	R$ 429 739	R$ 1 230 389,02
9	R$ 9 105 772,10	R$ 6 768 423,13	R$ 4 054 285,46	R$ 11 607 845,67
10	R$ 6 745 016,37	R$ 6 030 508,03	R$ 3 612 274,31	R$ 10 342 321,27
11	R$ 180 091,94	R$ 760 578,37	R$ 455 586,44	R$ 1 304 391,91
12	R$ 721 716,75	R$ 886 105,22	R$ 530 777,03	R$ 1 519 670,45
13	R$ 687 613,95	R$ 788 555,10	R$ 472 344,50	R$ 1 352 371,99
14	R$ 2 078 416,09	R$ 599 436,73	R$ 359 062,60	R$ 1 028 033,99
15	R$ 1 146 652,78	R$ 1 124 987,89	R$ 673 867,75	R$ 1 929 354,23
16	R$ 523 055,72	R$ 260 382,45	R$ 155 969,09	R$ 446 555,91
17	R$ 27 173,83	R$ 61 925,91	R$ 37 093,62	R$ 106 202,94
18	R$ 1 059 580,60	R$ 392 678,53	R$ 235 214,44	R$ 673 443,67
19	R$ 95 630,02	R$ 81 418,91	R$ 48 769,93	R$ 139 633,44
20	R$ 191 260,04	R$ 600 150,71	R$ 359 490,27	R$ 1 029 258,47
21	R$ 3 414 821,54	R$ 4 661 654,82	R$ 2 792 331,24	R$ 7 994 738,01
22	R$ 2 194 104,79	R$ 1 163 459,90	R$ 696 912,48	R$ 1 995 333,73
23	R$ 494 415,05	R$ 275 463,47	R$ 165 002,62	R$ 472 419,85

Tabela 5.8: Coeficientes de determinação (R^2) e raízes dos erros quadráticos médios ($RMSE$) calculados para os modelos lineares rurais.

Modelo	R^2	$RMSE$
Regressão OLS R_1	52,7%	0,92
Regressão Espacial R_1	74,7%	0,67
Regressão OLS R_2	78,9%	0,62
Regressão Espacial R_2	83,6%	0,54

conforme detalhado na Subseção 4.2.5, são significantes.

Os coeficientes de determinação (R^2) e as raízes dos erros quadráticos médios ($RMSE$) calculados para os modelos destinados a imóveis rurais encontram-se especificados nas Tabelas 5.8.

As semiamplitudes inferiores e superiores dos intervalos de confiança calculados para os modelos rurais encontram-se na Tabela 5.9.

5.4 Interpretabilidade dos Resultados

Para as modelagens urbanas e rurais lineares específicas, 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 das Tabelas 5.10 e 5.11, é possível se

Tabela 5.9: Semiamplitudes dos intervalos de confiança 90% calculados para os modelos rurais multiníveis.

Modelo	**Limite IC Inferior**	**Limite IC Superior**	**Amplitude IC**
Regressão OLS R_1	-53,6%	+124,1%	177,7%
Regressão Espacial R_1	-43,1%	+91,6%	134,7%
Regressão OLS R_2	-40,1%	+71,5%	111,6%
Regressão Espacial R_2	-36,4%	+72,6%	109,0%

identificar as variáveis mais relevantes aos modelos específicos mais convencionais, regressão linear múltipla (OLS) e regressão espacial; destacam-se *Ln(Área do Terreno)* no caso de imóveis urbanos e *VTN (RFB)* e *Distância Zona Urbana* no caso de imóveis rurais.

Tabela 5.10: Coeficientes e *p-valores* associados calculados para os modelos específicos convencionais U_3. *RE* como abreviação de Regressão Espacial.

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

NA: não se aplica

Cada ponto nos gráficos das Figuras 5.7, 5.8 e 5.9 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. Os atributos utilizados em cada nível de modelagem podem ser observados na Subseção 4.2.6.

Quanto mais intenso o tom cor de rosa, maior o valor do atributo associado a determinada instância; quanto mais azul, menor. Analisando-se os gráficos ponto a ponto, a assimetria em relação ao valor de Shapley nulo, zero do eixo das abscissas, indica quanto aquela variável foi capaz de deslocar o *Ln(Valor Total Atualizado)* do valor esperado, como medida de tendência central probabilística, em comparação ao caso hipotético de ela ser desconsiderada na modelagem.

Tabela 5.11: Coeficientes e *p-valores* associados calculados para os modelos específicos R_2. *RE* como abreviação de Regressão Espacial.

Atributo	**Coef. OLS**	***p-valor* OLS**	**Coef. RE**	***p-valor* RE**
Constante / Intercepto	11,30	0,00	11,65	0,00
Potencial A2	1,17	0,03	0,37	0,45
Cursos d'Água	1,47	0,00	1,67	0,00
Acesso Pavimentado	1,05	0,03	1,27	0,00
Distância Zona Urbana	-2,51	0,01	-3,56	0,00
VTN (RFB)	2,70	0,03	2,39	0,00
IDHM	1,21	0,15	0,43	0,51
Ln(Área do Terreno)	2,13	0,03	2,86	0,00
Matriz de Vizinhança W	NA	NA	-0,09	0,02

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 5.9. 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.

Ressalta-se que o algoritmo *ANN MLPRegressor* também apresentou boa aderência à realidade nos níveis básico e intermediário, ilustrados nas Figuras 5.10 e 5.11, respectivamente.

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 nos três níveis de modelagem, lineares e não lineares.

Já para os imóveis rurais, o uso do imóvel (agricultura, pecuária, silvicultura ou preservação) representado pelo atributo *VTN (RFB)* ganhou posição de destaque.

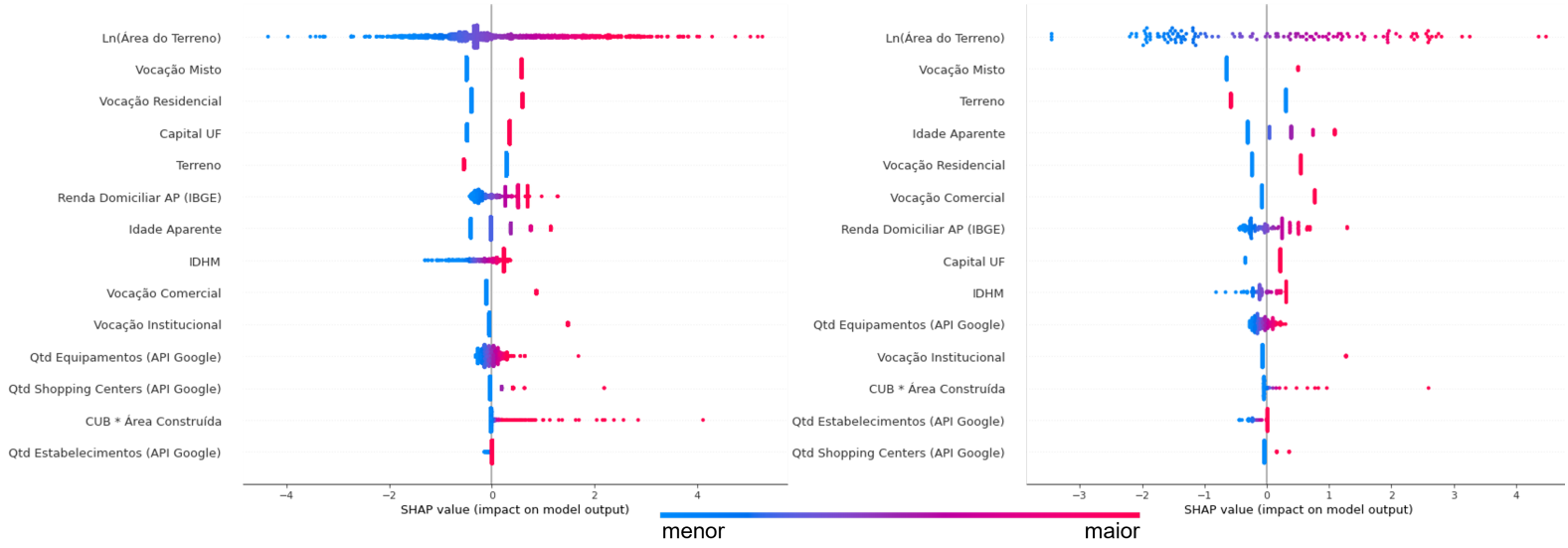

Figura 5.7: Valores de Shapley calculados para o modelo U_3 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *SGDRegressor*.

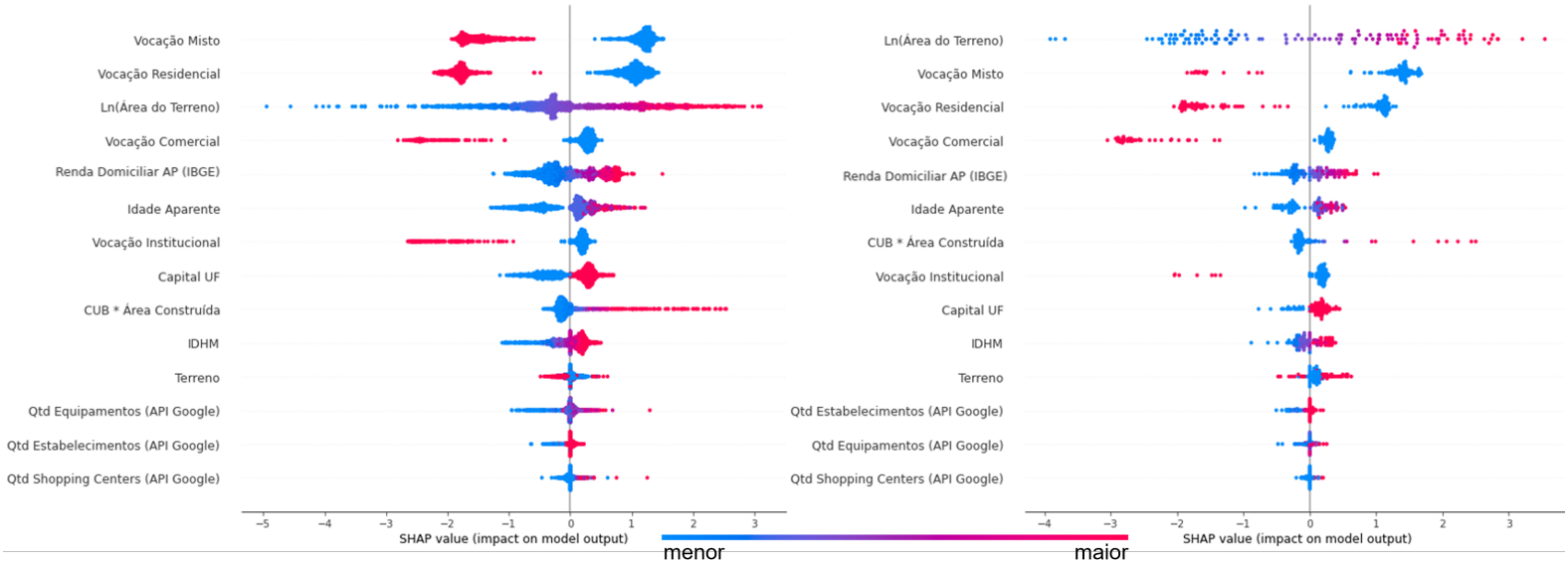

Figura 5.8: Valores de Shapley calculados para o modelo U_3 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *ANN MLPRegressor*.

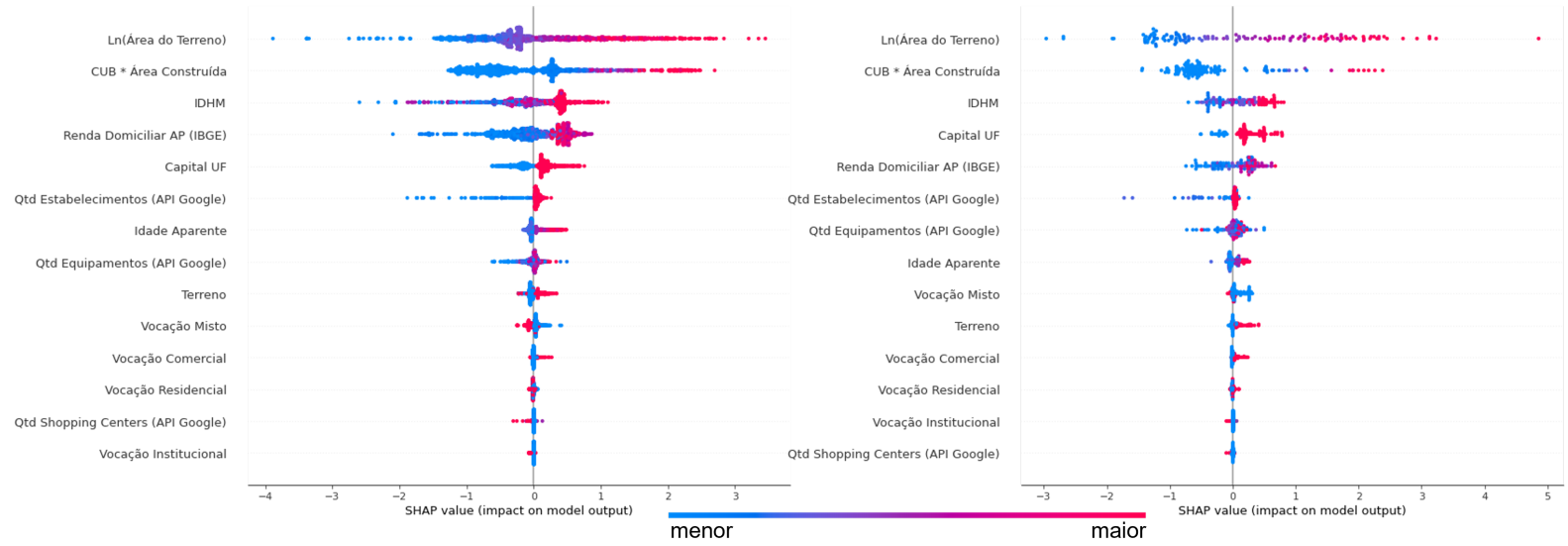

Figura 5.9: Valores de Shapley calculados para o modelo U_3 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *XGBRegressor*.

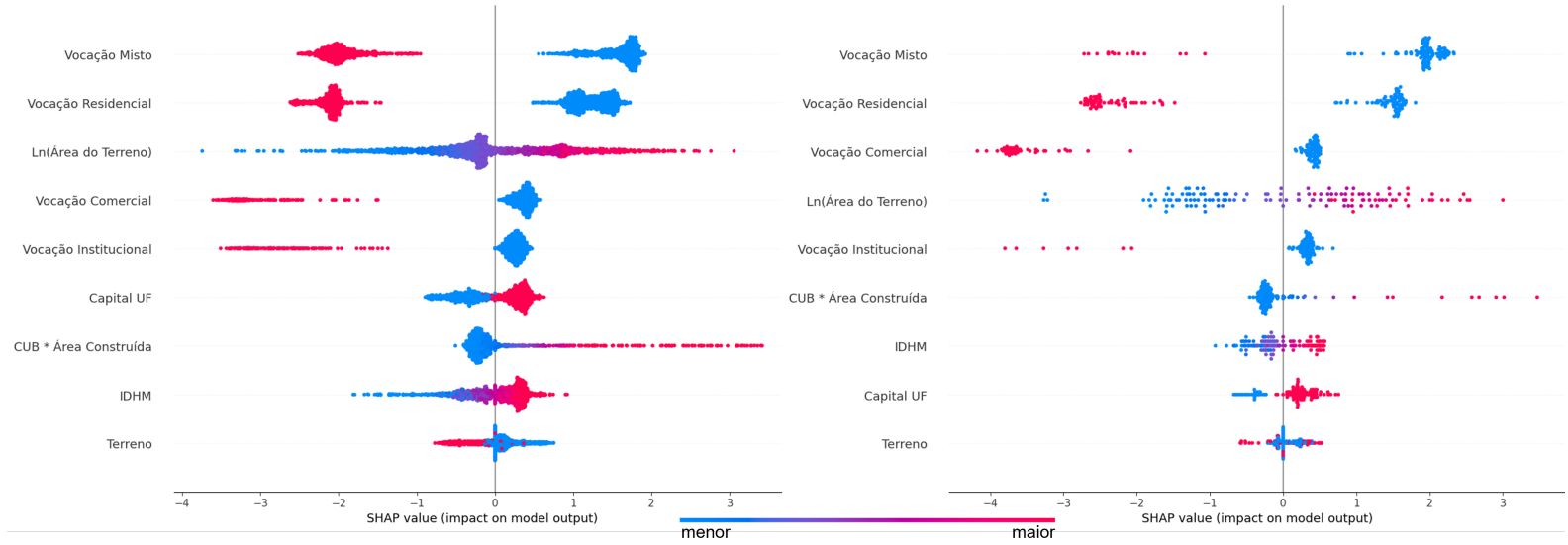

Figura 5.10: Valores de Shapley calculados para o modelo U_1 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *ANN MLPRegressor*.

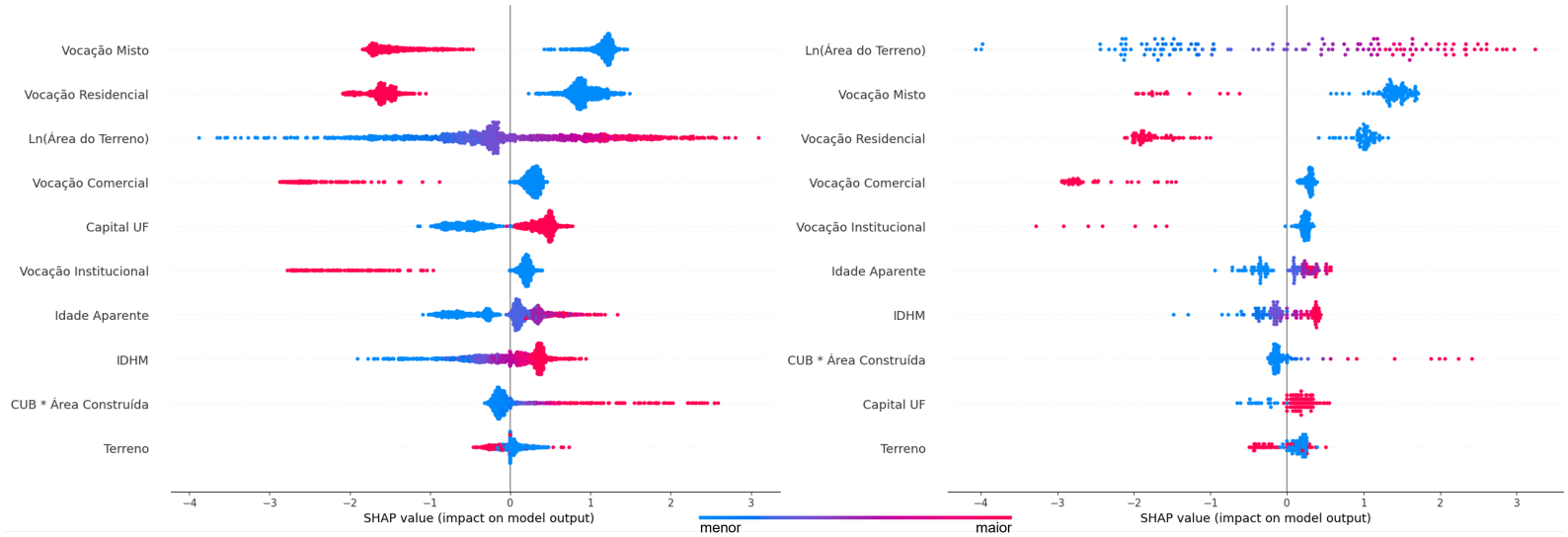

Figura 5.11: Valores de Shapley calculados para o modelo U_2 no Cenário A (à esquerda) e no Cenário B (à direita) implementado com o algoritmo *ANN MLPRegressor*.

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 e os modelos específicos U_3 alcançaram resultados melhores que os básicos e intermediários. Entendem-se por resultados melhores: 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 U_3 no Cenário A representado na Figura 5.12, as *ANN MLPRegressor* desenvolvidas parecem mais adequadas à valoração dos imóveis administrados pelo EB. Os pontos levantados na Seção 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 dos *p-valores* associados às matrizes de vizinhança W e com a autocorrelação espacial dos valores dos imóveis urbanos e rurais atestada pelo I de Moran, fica ainda mais evidente a importância da localização na formação das precificações imobiliárias.

Em decorrência da importância de fatores locacionais, os modelos destinados a imóveis urbanos e rurais tendem a performar de maneira superior nas regiões de maior densidade de bens anotados, ilustradas na Figura 5.2.

Para os modelos urbanos, as estimativas de valor são mais robustas, com menor grau

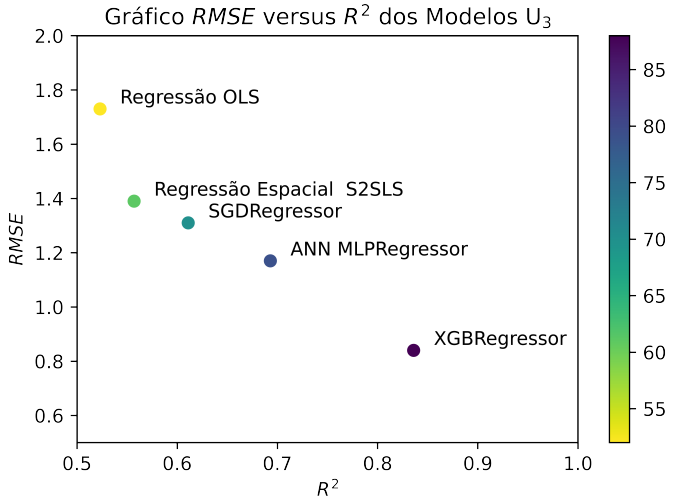

Figura 5.12: Gráfico comparativo entre os modelos urbanos U_3 no Cenário A considerando as métricas de avaliação R^2 e $RMSE$.

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, são mais imprecisas.

Furtado [38] sugere, a partir de estudo de caso no município de Belo Horizonte, que as modelagens econométricas espaciais de mercado imobiliário devem considerar a importância de bairros como entidades cognitivamente percebidas e que o uso deles como agregadores de informações de amenidades urbanas disponíveis nas cidades melhora a performance dos modelos.

Aplicando-se a conclusão de Furtado [38] a este trabalho, percebe-se uma limitação. Há bairros e AP sem ocorrências registradas na base de treinamento, o que acaba por produzir unidades espaciais isoladas, ilhas de Anselin [33]. Consequentemente, as estimativas produzidas pelos modelos implementados quando aplicados a imóveis situados em regiões isoladas ficam enfraquecidas, com maior grau de incerteza associado.

Com a largura de banda de 21 km testada e efetivamente utilizada para construir a matriz de pesos espaciais da amostra de imóveis rurais, produziram-se zonas de amortecimento poligonais (*buffers* poligonais) no *QGIS 3.16.3 Hannover*, ilustradas na Figura 5.13, que indicam regiões para as quais as modelagens básica e específica do campo devem atingir resultados mais satisfatórios. Ressalta-se que os dados rurais coletados não têm cobertura geográfica nacional e que os resultados obtidos sugerem muito mais um conjunto de variáveis passíveis de utilização em uma abordagem futura mais completa que um modelo preditivo pronto para uso. O desempenho dos modelos rurais se mostrou superior para ocorrências distantes mais de 15 km de centros urbanos.

Imóveis com valor histórico-cultural associado, tais como fortalezas históricas, podem

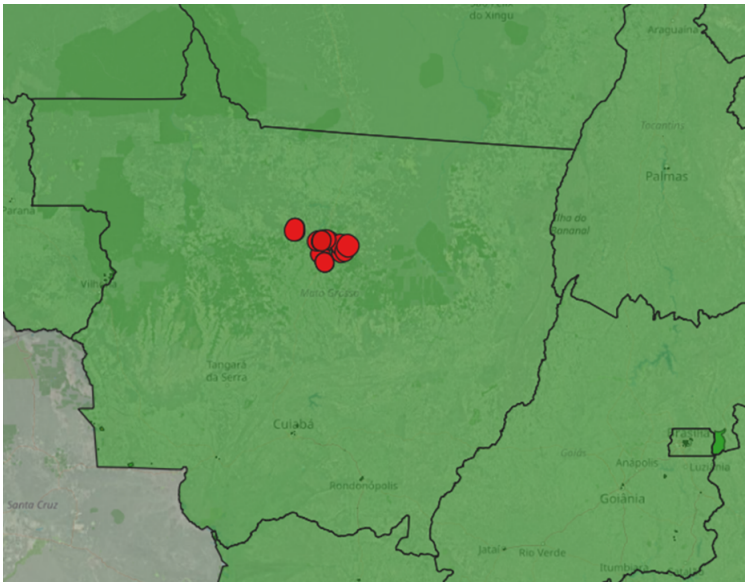

Figura 5.13: Zonas de amortecimento criadas para os modelos específicos rurais na região de Sinop-MT.

ter componente intangível e valor de acessórios específicos, como pinturas e monumentos, acrescidos às estimativas deste trabalho.

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

Capítulo 6

Conclusões e Trabalhos Futuros

Esta dissertação explorou o problema de se estimar, em níveis iniciais, valores de imóveis 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. Como consequência, tanto recursos laborais quanto recursos financeiros poderão ser poupados.

Os modelos destinados a valorar imóveis urbanos, em especial o específico com *ANN MLPRegressor*, são passíveis de utilização institucional, respeitando suas limitações de performance regional. 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.

Os modelos implementados com os algoritmos *ANN MLPRegressor* e *XGBRegressor* tiveram coeficiente de determinação consistentemente superior a 57%, em todos os cenários desenhados. Os básicos e intermediários são mais simples em termos de estruturação de atributos, mas apresentaram performance inferior aos específicos nos experimentos realizados.

Já os modelos com foco em imóveis rurais requerem treinamento em um conjunto de dados mais volumoso e com abrangência geográfica nacional. Esta pesquisa pode, entretanto, auxiliar na seleção de variáveis a utilizar.

6.1 Contribuições

Com o desenvolvimento desta pesquisa, contribui-se em vertente de inovação ao EB, mais especificamente à área de valoração imobiliária e de computação aplicada à resolução de problemas reais. A documentação de metodologia clara, respeitando os preceitos das áreas de conhecimento de mineração de dados e de estatística quanto à dinâmica de valoração

patrimonial de imóveis, corrobora para o desenvolvimento de uma abordagem institucional inovadora.

Em termos de contribuição bibliográfica, a presente pesquisa gerou a publicação do artigo "*Value Estimation of Properties Administered by the Brazilian Army Using Machine Learning and Spatial Components*" (Neto, J. N. A. S., Ladeira, M.) que apresenta resultados parciais deste trabalho com uso dos algoritmos *SGDRegressor* e *XGBRegressor*. Foi publicado e apresentado no KDMiLe 2022 (*Symposium on Knowledge Discovery, Mining and Learning*), realizado em Campinas-SP.

6.2 Trabalhos Futuros

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 à cobrança de Imposto Predial e Territorial Urbano (IPTU) extraído das plantas de valores genéricos dos municípios talvez constitua um atributo significante e, consequentemente, 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, incluindo estados de conservação, por meio de redes convolucionais, com potencial de aprimoramento dos modelos preditivos construídos neste trabalho.

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 significante. O desemprego local pode exercer influência sobre este aspecto.

Abordar o problema à luz da teoria da causalidade, conforme Pearl [39], incluindo a construção de *directed acyclic graphs* (DAG), pode melhorar a interpretabilidade dos modelos. Possivelmente, ajudando a preencher lacunas informacionais do mercado de imóveis no Brasil, que carece de dados disponíveis, diferentemente daquele dos EUA, onde muitas informações imobiliárias são disponibilizadas bairro a bairro e onde há rastreabilidade de um mesmo bem em séries temporais por meio de índices, como o residencial Case-Shiller[1], já que as transações de venda e compra são mais recorrentes.

A construção de amostras urbanas e rurais mais robustas, preferencialmente homologadas à luz da NBR 14653 ou contendo propriedades efetivamente transacionadas, repre-

[1] `https://www.spglobal.com/spdji/en/index-family/indicators/sp-corelogic-case-shiller/sp-corelogic-case-shiller-composite`

senta um desafio a transpor a fim de desenvolver modelos de valoração imobiliária mais acurados e que atendam às necessidades do EB e da SPU.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoa de Nível Superior (CAPES).

Referências

[1] Williams, Joseph: *Housing markets with endogenous search: Theory and implications.* Journal of Urban Economics, 105:107–120, dezembro 2017. 3

[2] Rosen, Sherwin: *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.* Journal of Political Economy, 82(1):34 – 55, 1974. 4, 7, 25

[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):891–916, 2010. 4, 7, 25, 28

[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(6):2928–2934, 2015. 4, 25, 28

[5] Tchuente, 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. 4, 27, 28

[6] Furtado, Bernardo Alves: *Policyspace2: modeling markets and endogenous public policies.* Journal of Artificial Societies and Social Simulation, 25(1), 2022. 4

[7] Barros Antunes Campos, Rodger e Eduardo Almeida: *Decomposição espacial nos preços residenciais no município de São Paulo.* Estudos Econômicos (São Paulo), 48(1):5–38, 2018. 7

[8] Camargo Lima, Marcelo Rossi de: *Engenharia de Avaliações Aplicada em Propriedades Rurais.* Leud, 2021, ISBN 978-85-745-6390-9. 15

[9] Getis, Arthur e Jared Aldstadt: *Constructing the spatial weights matrix using a local statistic.* Geographical Analysis, 36:147–163, agosto 2010. 16

[10] Anselin, Luc: *A Local Indicator of Multivariate Spatial Association: Extending Geary's c.* Geographical Analysis, 51:133–150, 2019. 16

[11] Shalev-Shwartz, Shai e Shai Ben-David: *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, USA, 2014, ISBN 978-1-107-05713-5. 17

[12] J. E. Dennis, Jr. e Robert B. Schnabel: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Society for Industrial and Applied Mathematic, 1996, ISBN 978-0-898-71364-0. 17

[13] Bottou, Léon e Noboru Murata: *Stochastic approximations and efficient learning*. Em Arbib, M. A. (editor): *The Handbook of Brain Theory and Neural Networks, Second edition,*. The MIT Press, Cambridge, MA, 2002. `http://leon.bottou.org/papers/bottou-murata-2002`. 17

[14] Zhang, Tong: *Solving large scale linear prediction problems using stochastic gradient descent algorithms*. Em *ICML*, volume 69, janeiro 2004. 17

[15] Tibshirani, Robert: *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288, 1996. 17

[16] Hoerl, Arthur E. e Robert W. Kennard: *Ridge regression: Biased estimation for nonorthogonal problems*. Technometrics, 12(1):55–67, 1970. 17

[17] Zou, Hui e Trevor Hastie: *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2):301–320, 2005. 17

[18] Hinton, Geoffrey E.: *Connectionist learning procedures*. Artificial Intelligence, 40(1-3):185–234, 1989. 18

[19] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay: *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12:2825–2830, 2011. 19

[20] Chen, Tianqi e Carlos Guestrin: *Xgboost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, páginas 785–794, 2016. 20

[21] Breiman, Leo, Jerome Friedman, Charles J. Stone e R.A. Olshen: *Classification and Regression Trees*. Taylor & Francis, 1984, ISBN 978-0-412-04841-8. 21

[22] Hair, Joseph, G. Tomas M. Hult, Christian Ringle e Marko Sarstedt: *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Sage Publishing, janeiro 2022, ISBN 978-1-5443-9640-8. 22, 23

[23] Basco, Rodrigo, Joseph Hair, Christian Ringle e Marko Sarstedt: *Advancing family business research through modeling nonlinear relationships: Comparing pls-sem and multiple regression*. Journal of Family Business Strategy, 1(3), setembro 2022. 23

[24] Becker, Jan Michael, Christian Ringle e Marko Sarstedt: *Estimating moderating effects in pls-sem and plsc-sem: Interaction term generation*data treatment*. Journal of Applied Structural Equation Modeling, 2:1–21, junho 2018. 23

[25] 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, 2:56–67, 2020. 24, 26

[26] Chakraborty, Debaditya e Hazem Elzarka: *A novel construction cost prediction model using hybrid natural and light gradient boosting.* Advanced Engineering Informatics, 2020. 24, 26

[27] Lundberg, Scott M. e Su In Lee: *A unified approach to interpreting model predictions.* Em *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, página 4768–4777, Red Hook, NY, USA, dezembro 2017. Curran Associates Inc., ISBN 9781510860964. 24

[28] Kiely, Timothy e Nathaniel Bastian: *The spatially conscious machine learning model.* Statistical Analysis and Data Mining: The ASA Data Science Journal, 13(1):31–49, 2020. 25, 28

[29] 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 (PerCom Workshops)*, páginas 783–789, Kyoto, Japan, 2019. 26

[30] Hagenauer, Julian e Marco Helbich: *A geographically weighted artificial neural network.* International Journal of Geographical Information Science, 36(2):215–235, 2022. 26, 28

[31] El-Geneidy, Ahmed, Michael Grimsrud, Wasfi Rania, Paul Tétreault e Julien Surprenant-Legault: *New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas.* Transportation, 41:193–210, janeiro 2014. 30

[32] Furtado, Bernardo Alves: *NT DISET 78 - Gerando Famílias Artificiais Intraurbanas: censo 2010.* Ipea, 2020. 32

[33] Anselin, Luc: *Spatial Econometrics: Methods and Models.* Springer Dordrecht, 1988, ISBN 978-94-015-7799-1. 43, 44, 68

[34] Anselin, Luc e Sergio Rey: *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL.* Geoda Press LLC, novembro 2014, ISBN 978-0-986-34210-3. 44

[35] Cohen, Jacob: *Statistical Power Analysis for the Behavioral Sciences.* Routledge, junho 1988, ISBN 978-0-203-77158-7. 46

[36] Hair, Joseph, William Black, Barry Babin e Rolph Anderson: *Multivariate Data Analysis: A Global Perspective.* Pearson Education, março 2010, ISBN 978-0-135-15309-3. 46

[37] Cook, R. Dennis e Sanford Weisberg: *Residuals and Influence in Regression.* Chapman and Hall, 1982. 50

[38] Furtado, Bernardo Alves: *TD 1570 - Análise Quantílica-Espacial de Determinantes de Preços de Imóveis Urbanos com Matriz de Bairros: Evidências do Mercado de Belo Horizonte.* Ipea, 2011. 68

[39] Pearl, Judea: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009, ISBN 978-0-521-89560-6. 71