UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ESTRUTURAL E CONSTRUÇÃO CIVIL
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL: ESTRUTURAS E CONSTRUÇÃO CIVIL

FELIPE FERNANDES MOREIRA

MODELO HEDÔNICO ESPACIAL PARA AVALIAÇÃO EM MASSA DE IMÓVEIS DE FORTALEZA.

FORTALEZA
2020

FELIPE FERNANDES MOREIRA

MODELO HEDÔNICO ESPACIAL PARA AVALIAÇÃO EM MASSA DE IMÓVEIS DE FORTALEZA

Dissertação de Mestrado apresentada à Coordenação do Programa de Pós-Graduação em Engenharia Civil: Estruturas e Construção Civil da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Mestre em Engenharia Civil. Área de Concentração: Construção Civil.

Orientadora: Profa. Dr. Vanessa Ribeiro Campos

Co-orientador: Prof. Dr. José de Paula Barros Neto

FORTALEZA
2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M837m Moreira, Felipe Fernandes.
MODELO HEDÔNICO ESPACIAL PARA AVALIAÇÃO EM MASSA DE IMÓVEIS DE FORTALEZA. / Felipe Fernandes Moreira. – 2020.
110 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Civil: Estruturas e Construção Civil, Fortaleza, 2020.
Orientação: Profa. Dra. Vanessa Ribeiro Campos.
Coorientação: Prof. Dr. José de Paula Barros Neto.

1. Avaliação de imóveis. 2. Mercado Imobiliário. 3. Modelos Hedônicos. 4. Big Data. 5. Aprendizado de Máquinas. I. Título.

CDD 624.1

FELIPE FERNANDES MOREIRA

MODELO HEDÔNICO ESPACIAL PARA AVALIAÇÃO EM MASSA DE IMÓVEIS DE FORTALEZA

Dissertação de Mestrado apresentada à Coordenação do Programa de Pós-Graduação em Engenharia Civil: Estruturas e Construção Civil da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Mestre em Engenharia Civil. Área de Concentração: Construção Civil.

Aprovada em: 01/10/2020.

BANCA EXAMINADORA

_____
Profa. Dr. Vanessa Ribeiro Campos (Orientadora)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. José de Paula Barros Neto (Coorientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Lucas Feitosa de Albuquerque Lima Babadopulos
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Luiz Fernando Mahlmann Heineck
Universidade Estadual do Ceará (UECE)

_____
Profa. Dra. Maria Carolina G. Oliveira Brandstetter
Universidade Federal de Goiás (UFG)

A minha esposa Richelle.

Aos meus pais, Rita e Marciano.

AGRADECIMENTOS

À Profa. Dra. Vanessa Ribeiro Campos, pelo zelo e esmero num trabalho que não se limitou à orientação das atividades do presente trabalho, mas que objetivava a formação de um pesquisador. Por esse empenho e pela confiança, sou bastante grato.

Ao Prof. Dr. José de Paula Barros Neto, por ter desde muito cedo me incentivado (ou desafiado) a aprimorar minha formação de engenharia por meio da pesquisa acadêmica.

Aos professores Lucas Feitosa de Albuquerque Lima Babadopulos, Luiz Fernando Mahlmann Heineck e Maria Carolina G. Oliveira Brandstetter pela disposição em participar na banca examinadora.

À minha esposa, Richelle, minha amada companheira de todos os momentos e principal fonte de inspiração.

Aos meus pais, Rita e Marciano, por terem construído aquilo que sou e por terem dado significado à palavra “Família”.

Ao meu sogro e sogra, Roberto e Fátima, por quem nutro profunda admiração.

Aos queridos colegas de mestrado de “Gerenciamento”, Clarissa, Lardner, Luiz, Renan e Yan, e “Materiais”, Ana, Bruna, Cristina e Gisela. Aos professores e profissionais do programa, meu muito obrigado.

RESUMO

A avaliação mais precisa de imóveis gera benefícios para todas as partes interessadas envolvidas na cadeia de valor do mercado imobiliário: clientes, poder público e empresas. A capacidade de elaborar modelos comparativos que permitam, com base em observações reais, prever o valor de um imóvel com maior precisão influencia positivamente, portanto, um setor com grande atuação na economia nacional. Na literatura, encontra-se a utilização de métodos capazes de modelar relações não-lineares entre variáveis, métodos que incorporam efeitos do fenômeno da autocorrelação espacial e a combinação de resultados de modelos distintos no intuito de se obter modelos com erros menores. O objetivo geral deste ensaio é construir um modelo hedônico espacial de avaliação em massa de imóveis para a cidade de Fortaleza, Ceará, com suporte em técnicas capazes de explorar potenciais relações não-lineares da autocorrelação espacial. É sabido, porém, que o processo de modelagem está sujeito a efeitos de aleatoriedade, sendo necessário garantir que as melhorias na precisão dos modelos são causadas pelas mudanças realizadas pelos pesquisadores. Dessa maneira, o modelo proposto tem o desempenho comparado com padrões de referência que utilizam métodos consolidados na literatura por meio de testes estatísticos que atestem a significância das diferenças nas métricas de erro encontradas. Os testes estatísticos revelam que a comparação pontual é falha e não permite atestar com a devida confiança o melhor desempenho dos algoritmos testados em termos relativos e absolutos. Denotou evidências de que o emprego de *ensemble*, em particular o algoritmo *Random Forest*, e que exploram o fenômeno da autocorrelação espacial, a saber modelo autoregressivo espacial, são mais precisos para problemas de avaliação em massa de imóveis. Não foram encontradas diferenças estatisticamente significantes entre os erros relativos dos modelos Random Forest e autoregressivo espacial. Os resultados advogam em favor dos modelos de regressão espacial, tendo em vista que além do desempenho obtido os mesmos são mais facilmente interpretados.

Palavras-chave: Avaliação de imóveis. Mercado Imobiliário. Modelos Hedônicos. *Big Data*. Aprendizado de Máquinas.

ABSTRACT

A more precise real estate appraisal generates benefits for all stakeholders involved in real estate value chain: customers, government and companies. Therefore, the capacity to construct comparative models capable of, with real market observations, predict more precisely the value of real estate assets influences positively a economic sector with large presence on national economy. Literature reveals that the use of methods capable of modelling non-linear relationships between variables, methods that incorporate the effects of special autocorrelation and emsemble of different models results with the purpose of reducing prediction errors. The main objective of the study is to build a mass real estate appraisal spatial hedonic model for the city of Fortaleza, Ceará, derived from techniques that allows exploring potential non-linear spatial autocorrelation effects. However, it is known that the modelling process is swayed by randomness, demanding that the betterment observed in error metrics results are produced by the changes applied by researchers. Thus, the performance of the proposed model will be compared with referential models built with well-established methods through statistical tests that can certify the significance of the difference between error metrics obtained. Statistical testing reveal that single error metrics comparison are not suited for evaluate model performance in relative and absolute terms. The study provide evidence that models constructed with ensembles (namely Random Forest) and models exploring spatial autocorrelation phenomenon (Spatial Autoregressive Models) are relativily more precise for real estate mass appraisal problems, although no found no evidence to state that there are significant difference between the performance of Random Forest and Autoregressive Spatial Models. Results point to advantages in the use of spatial autoregressive model, since in addition to its overall performance, such models are easily interpreted.

Keywords: real estate appraisal, real estate market, hedonic models, Big Data, machine learning.

LISTA DE FIGURAS

LISTA DE QUADROS E TABELAS

SUMÁRIO

1 INTRODUÇÃO

A tomada de decisão de investimentos no mercado imobiliário passa invariavelmente pela etapa de estimativa do valor que um projeto é capaz de gerar ao investidor seja pela receita oriunda da comercialização ou do aluguel das unidades (comerciais ou residenciais) do empreendimento. Independentemente da fonte de receitas, métodos comparativos, em que o empreendimento objeto de decisão tem seu valor estimado com suporte na comparação de suas características com as de outros empreendimentos do mercado, são amplamente utilizados e estudados na literatura especializada. Um caso particular de método comparativo é a avaliação em massa de imóveis, quando são estabelecidos modelos de regressão com base em dados de transações imobiliárias previamente registradas e são posteriormente utilizados para estimar o valor de venda de um imóvel ou o valor a ser cobrado pelo seu aluguel.

A avaliação em massa de imóveis beneficia-se de técnicas e ferramentas aplicáveis da tecnologia do "*Big Data*", que permite a estruturação e manipulação de bancos de dados de grande porte com finalidades específicas. Nessas circunstâncias, diversas organizações investem na formulação de sistemas capazes de empregar meios para coletar e processar grandes volumes de dados. Bancos de dados de transações imobiliárias são constituídos para auxiliar a tomada de decisão de empresas do mercado imobiliário, agentes financeiros que financiam empréstimos para aquisição de imóveis ou recebem imóveis como garantia em outras operações e o governo que cobra impostos incidentes sobre o valor justo de imóveis.

No caso específico de Fortaleza, o presente trabalho fez uso de diferentes bancos de dados que disponibilizam dados estruturados ou não e que, segundo a literatura, são capazes de explicar parte do valor de um ativo imobiliário. Tal fato baseia-se no que Rosen (1974) chamou de *hipótese hedônica*, cuja premissa fundamental é a de que o valor de um produto depende diretamente de suas características intrínsecas. Os modelos de avaliação em massa são comumente denominados modelos hedônicos, sendo capazes de representar matematicamente o valor de um ativo imobiliário.

Ademais, foi analisado um conjunto de registros imobiliários comparável às maiores bases encontradas na literatura consultada. Com esteio na aplicação de técnicas de "*Data Mining*" e de Sistemas de Informação Geográfica (SIG), a coleção de variáveis do registro inicial foi expandida com a integração entre distintas bases de dados georreferenciadas geridas por distintas agências governamentais, entidades privadas e *open source*. A base resultante foi então utilizada para a estruturação de um modelo hedônico com potencial de estimativa de preços com precisão comparável ao observado na literatura especializada e a

aplicação de um posterior procedimento diagnóstico dos erros de estimativa observados. O registro de transações imobiliárias usado no presente trabalho engloba 144.914 transações imobiliárias, tendo cada uma dessas transações 29 atributos registrados além do valor do imóvel. Em comparação às amostras encontradas na literatura, conforme observado no Quadro 1, o conjunto de dados utilizado no presente estudo é o segundo maior em número de observações.

Quadro 1 – Compilação de estudos relacionados e tamanho de amostra estudada.

Autores	**Tamanho da Amostra**	**Local**
(FERNANDEZ; MUKHERJEE; SCOTT, 2018)	148.000	Riverside e San Bernardino, Califórnia, EUA
(PETERSON; FLANAGAN, 2009)	46.467	Wake County, Carolina do Norte
(ZURADA; LEVITAN; GUAN, 2011)	16.366	Louisville, Kentucky, EUA
(LASOTA et al., 2013)	9.795	Polônia
(ČEH et al., 2018)	7.407	Liubliana, Eslovênia
(KEMPA et al., 2011)	5.303	Polônia
(NGUYEN; CRIPPS, 2001)	3.906	Rutherford County, Tennessee, EUA
(ANTIPOV; POKRYSHEVSKAYA, 2012)	2.848	São Petesburgo, Rússia
(MCCLUSKEY et al., 2013)	2.694	Não informa.
(YEH; HSU; WEIGHT, 2018)	1.963	Taipé e Nova Taipé, Taiwan
(SEYA; YAMAGATA; TSUTSUMI, 2013)	520	"Boston housing data"
(KOSTOV, 2010)	506	"Boston housing data"
(JANSSEN et al., 2001)	351	Estocolmo, Suécia
(WORZALA; LENK; SILVA, 1995)	288	Fort Collins, Colorado
(UBERTI et al., 2018)	113	Rio de Janeiro, Brasil
(KONTRIMAS; VERIKAS, 2011)	100	Lituânia
(HE et al., 2010)	73	Pequim, China

FONTE: Elaboração própria.

1.1 Justificativa

O mercado imobiliário recebe influência de uma série de variáveis macroeconômicas, como a taxa de juros, mercado de crédito, renda e confiança do consumidor (ALBUQUERQUE *et al*., 2018). No que diz respeito ao produto, ou seja, ao comportamento

do consumidor, não só os atributos do imóvel influenciam na tomada de decisão como também atributos referentes à sua localização. Locatelli *et al.* (2017) concluem que aspectos comportamentais do consumidor – busca pelo sentimento de autorrealização, ou demanda pela proteção patrimonial em decorrência de preocupações com turbulências econômicas causadas por um frágil rigor na condução das políticas econômicas no concerto federal – são tão importantes na tomada de decisão de compra quanto aspectos macroeconômicos. Brando e Barbedo (2016) encontraram efeitos expressivos no preço em modelos de curto prazo de imóveis motivados por variáveis comportamentais (como o sentimento de otimismo internacional com a economia brasileira) e inovações institucionais (lei do patrimônio de afetação, por exemplo) a partir de dados das cidades de São Paulo e Rio de Janeiro.

Campos e Almeida (2018) exprimem externalidades negativas (como congestionamento e criminalidade), serviços (como segurança e lazer) e infraestrutura (*e.g.*, saneamento e qualidade da estrutura viária) como fatores que influem no valor dos imóveis da cidade de São Paulo. Alertam, portanto, para a importância do estudo das amenidades, ou as qualidades inerentes ao espaço geográfico que irão exercer influxo negativo ou positivo no preço dos imóveis. Foram encontradas evidências de que mudanças nas amenidades de uma certa localidade são capazes de afetar o preço médio dos imóveis de lugares adjacentes (efeito que os autores denominam transbordamento), demonstrando as relações espaciais do fenômeno de formação de preço de imóveis.

Arraes e Filho (2008), com uma amostra contendo apartamentos residenciais, *flats* e salas comerciais distribuídos em 41 dos 112 bairros da cidade de Fortaleza, detectaram significância nas variáveis relacionadas com segurança para o segmento residencial, mas não tão influente no caso dos *flats* e comerciais. A disparidade seria explicada pela menor incidência de homicídios em localidades com predominância de *flats* e imóveis comerciais. Observou-se também a importância das amenidades (acessibilidade, educação, saúde e lazer) na formação do preço, com influência negativa das externalidades oriundas de equipamentos urbanos, como escolas e hospitais, e influxo positivo para o nível de lazer que o bairro oferece.

Seya, Yamagata e Tsutsumi (2013) apontam a importante influência que variáveis relacionadas à localização geográfica no valor dos imóveis, existindo, portanto, o fenômeno da autocorrelação espacial. Os autores fazem referência à expressão "Modelos Hedônicos Espaciais", creditando a Dubin (1988) e Can (1990) as primeiras referências ao termo, estes obtidos quando os modelos de avaliação em massa incorporam não apenas variáveis geográficas ao conjunto de características que descrevem os imóveis, como também permitem capturar a

influência exercida pela proximidade de outros imóveis.

Modelos mais precisos (ou seja, com menores erros) são de grande valia para consumidores de produtos imobiliários, empreendedores do setor, agentes financeiros do sistema nacional de habitação e esferas municipais e estaduais do governo (que cobram, respectivamente, impostos sobre transação e impostos sobre transmissão e doação de imóveis). A cadeia de valor do setor imobiliário possui grande influência na economia nacional. Em 2018, R$ 53 bilhões foram pagos como salários, retiradas e outras remunerações apenas por empresas ligadas a construção de edifícios no País (IBGE, 2020). A realidade é semelhante ao observado em outros países, especialmente os desenvolvidos, cujos ativos imobiliários são porção representativa do sistema financeiro (ISHIJIMA; MAEDA, 2015).

Para os agentes financeiros, entidades que alocam recursos no Sistema de Financiamento da Habitação (SFH), a avaliação de imóveis está diretamente associada ao risco do portfólio de investimentos da instituição (ZURADA; LEVITAN; GUAN, 2011). Para tal classe de *stakeholders* do mercado imobiliário, a precisão é fundamental, dadas as características do imóvel na qualidade de ativo financeiro. Nesse âmbito, a precisão nas avaliações do valor dos imóveis são fator-chave para analisar o risco inerente ao sistema financeiro brasileiro.

Projetos imobiliários têm tipicamente longos períodos de construção que, combinados com grande fluxo de capital, enseja lapsos extensos de retorno. Ainda, são projetos sujeitos a grandes incertezas, restrições e mudanças políticas e de legislação. Consequentemente, são projetos de investimentos de alto risco (YIJIAN; RUFU, 2008; MAO; WU, 2011; MINLI; WENPO, 2012). Para Mao e Wu (2011), a seleção do projeto é bem mais importante do que o seu posterior gerenciamento. Uma organização que desenvolve um processo de avaliação de imóveis robusto consegue aprimorar sua tomada de decisão, podendo o processo de avaliação fazer parte de um sistema de apoio à decisão.

A avaliação de imóveis é também importante fator para o garantir o equilíbrio entre a tributação justa e a arrecadação de impostos incidentes sobre propriedades. No caso particular da cidade de Fortaleza, no estado do Ceará, segundo o Relatório Contábil de Propósito Geral de 2019, emitido pela sua Secretaria Municipal das Finanças, 35% da arrecadação tributária do município procede de impostos sobre propriedade – Imposto sobre a Propriedade Predial e Territorial Urbana (IPTU) e Imposto sobre Transmissão de Bens e Imóveis (ITBI) (SEFIN, 2019).

Na literatura, muitos são os estudos comparativos entre modelos hedônicos de

avaliação em massa, principalmente comparações entre aqueles em redes neurais artificiais e modelos de regressão múltipla (NGUYEN; CRIPPS, 2001; PETERSON; FLANAGAN, 2009; KONTRIMAS; VERIKAS, 2011; ZURADA; LEVITAN; GUAN, 2011; MCCLUSKEY et al., 2013; YEH; HSU; WEIGHT, 2018), havendo porém utilização de métodos mais complexos como *Random Forest, Support Vector Machines*, krigagem e *ensembles* (combinação de modelos) (LAM; YU; LAM, 2009; BARBOZA *et al.*, 2011; KEMPA *et al.*, 2011; KONTRIMAS; VERIKAS, 2011; ANTIPOV; POKRYSHEVSKAYA, 2012; LASOTA et al., 2013; WIŚNIEWSKI, 2017; ČEH *et al.*, 2018; UBERTI *et al.*, 2018), enquanto outros estudos que focam nos modelos hedônicos espaciais (KOSTOV, 2010; SEYA; YAMAGATA; TSUTSUMI, 2013; WEN; BU; QIN, 2014; QU; LEE, 2015; ZHANG *et al.*, 2015). Percebe-se, porém, que a comparação dos modelos nos estudos consultados se dá com suporte de métricas como o erro quadrático médio, erro percentual absoluto médio, erro absoluto médio ou o total de estimativas consideradas como inaceitáveis (KONTRIMAS; VERIKAS, 2011). As diferenças de precisão entre os modelos são, portanto, avaliadas sem o uso de testes estatísticos capazes de detectar significância estatística entre as variadas métricas de erro obtidas para cada modelo em particular. Tais testes seriam capazes de fornecer indícios de que as melhorias nas avaliações feitas por um modelo não são resultadas do acaso.

1.2 Objetivos de Pesquisa

O objetivo geral do experimento sob relatório é propor uma metodologia de modelagem de avaliação em massa de imóveis para Fortaleza, Ceará. Os objetivos específicos configuram-se em:

- construir uma base georreferenciada de dados de transações imobiliárias integrada, agregando dados geográficos controlados e mantidos por distintos departamentos e organizações;
- avaliar a influência das variáveis intrínsecas e geográficas na formação de preço dos imóveis da cidade de Fortaleza, Ceará;
- investigar o desempenho de variados modelos de regressão no conjunto de dados obtido; e
- definir um protocolo de análise de erros com a finalidade de orientar a escolha de modelos e desenvolvimento de ações de melhoria dos resultados.

1.3 Estrutura da Dissertação

Esta dissertação é estruturada em cinco capítulos, contando com este, constituído de introdução do tema, justificativa, em que é mostrada a relevância do tema e são indicados os objetivos que a pesquisa busca lograr (tanto gerais como específicos).

O segundo contém uma revisão bibliográfica, onde dados sobre o mercado imobiliários são expressos, bem como se apontam os métodos utilizados na prática e discutidos na literatura, com ênfase na combinação de modelos e modelos hedônicos espaciais. Exprime-se, também, como os modelos são comumente comparados na literatura e analisa-se criticamente essas práticas.

O terceiro capítulo é constituído dos dados e métodos empregados nesta dissertação. Nesse módulo, é detalhado o procedimento seguido para a montagem do banco de dados e manipulação necessária com a fim de agregar variáveis geográficas. As variáveis estudadas são discutidas, apontando-se, também, a metodologia de treinamento e avaliação dos modelos, assim como o protocolo de testes.

No quarto capítulo, os resultados obtidos para cada modelo são expressos e discutidos, bem como comparados entre si. Em seguida, efetiva-se uma análise dos erros obtidos, ao passo que o desempenho dos modelos é comparado com o desempenho relatado na literatura especializada.

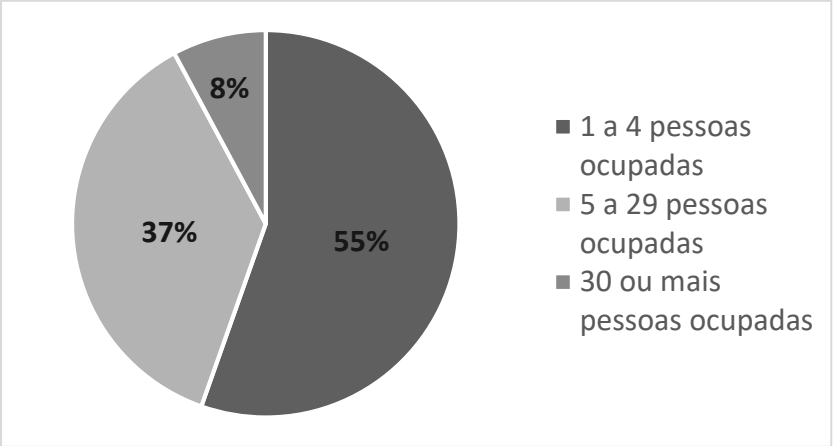
No quinto segmento estão as conclusões do experimento, assim como se demonstram os potenciais avanços recomendados para futuros trabalhos.

2 AVALIAÇÃO DE IMÓVEIS

2.1 Mercado Imobiliário

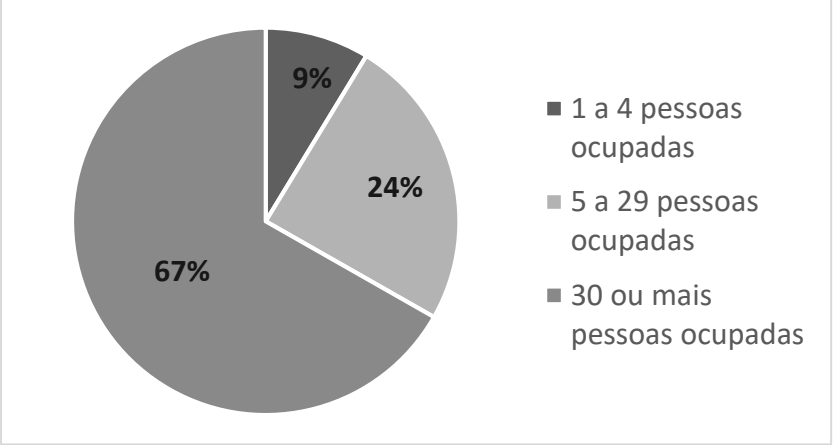
No Brasil, de acordo com a Pesquisa Anual da Indústria da Construção (PAIC) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2020), em 2018, o mercado da construção civil era composto por 124 mil empresas que perfaziam uma receita bruta anual de cerca de 290 bilhões de reais. Na Figura 1, nota-se que empresas com até 29 pessoas ocupadas representam 91% do total do mercado. Ainda, de acordo com a pesquisa, essas organizações concentram 33% do faturamento do mercado, e contabilizam faturamento médio de R$ 366 mil (de uma a quatro pessoas ocupadas) e R$ 1,55 milhões (de cinco a 29 pessoas ocupadas).

Figura 1 – Divisão das empresas de construção civil do País por tamanho.



FONTE: Pesquisa Anual da Indústria da Construção (IBGE, 2020).

Figura 2 – Divisão do faturamento do mercado por tamanho de empresa.

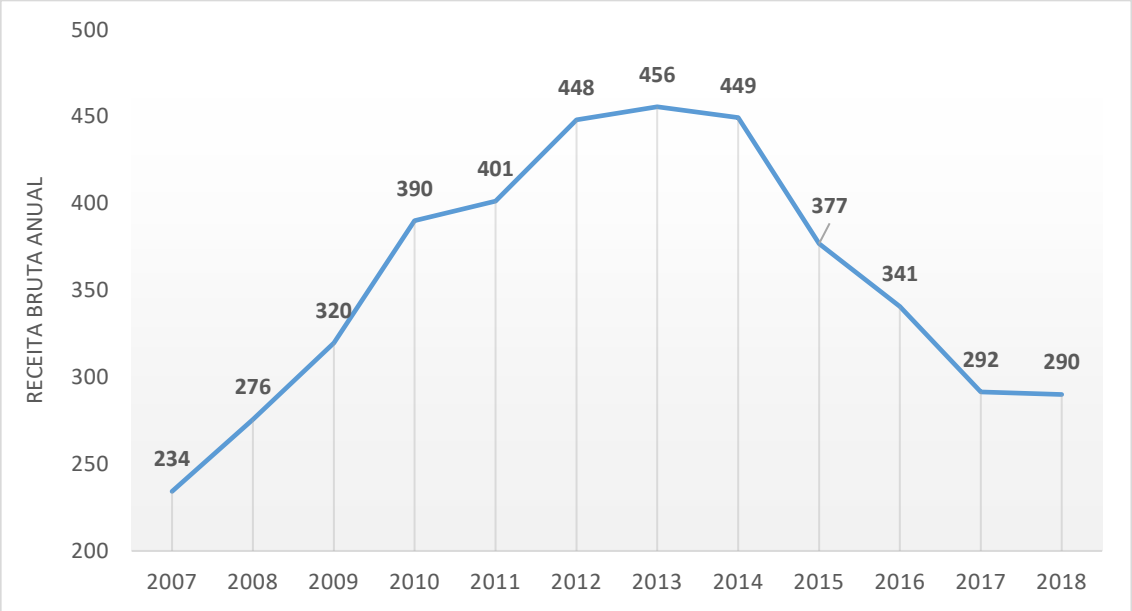


FONTE: Pesquisa Anual da Indústria da Construção (IBGE, 2018).

A construção civil (classificação de acordo com o Cadastro Nacional de Atividades

Econômicas, CNAE) compõe, além da construção de edifícios, as obras de infraestrutura e serviços especializados para construção (empresas que prestam serviços de demolição, instalações elétricas, hidráulicas, obras de acabamento, serviços de fundações, dentre outros). Na Figura 3 (onde os valores foram ajustados com base no IPCA a valores de 2018) observa-se que o setor experimentou dos anos de 2007 a 2013 um significativo crescimento (cerca de 95% no período, o que representa um crescimento anual de 11,76% ao ano). Nos cinco anos seguintes, porém, o setor enfrentou significativa queda (36,4% de queda no período, 5,7% ao ano).

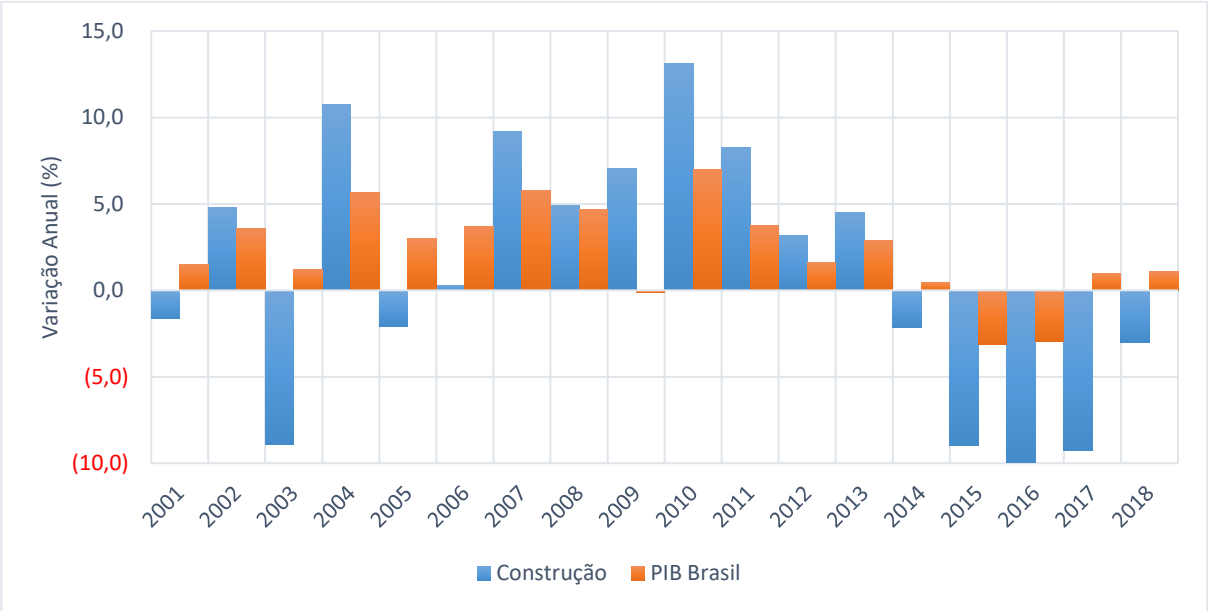
Figura 3 – Série histórica de receita bruta do setor a valores de 2018.



FONTE: Pesquisa Anual da Indústria da Construção (IBGE, 2009-2020)

O desempenho do setor da construção nacional está intensivamente correlacionado com o desempenho da economia nacional, conforme é divisado na Figura 4. A volatilidade do setor da construção, entretanto, é maior do que a volatilidade observada na economia brasileira. No período dos anos 2001 a 2018, a variabilidade do setor é superior a duas vezes e meia a variabilidade experimentada pela economia nacional. De 2014 a 2018, enquanto o País teve uma queda anual média no seu valor adicionado bruto de 0,7%, no setor de construção civil, a descensão foi de 6,2%. Os dados da Figura 4, por sua vez, alinham-se àqueles exibidos na Figura 3, notando-se consistentes crescimentos no valor adicionado bruto da construção, de 2007 a 2013, um crescimento médio anual de 7,2%.

Figura 4 – Variação anual do valor adicionado bruto pela construção e pela economia brasileira.



FONTE: Câmara Brasileira da Indústria da Construção (CBIC, 2002-2019).

2.2 Métodos de Avaliação em Massa

Rosen (1974) chamou de *hipótese hedônica* a percepção de que o valor de um certo produto é resultado da utilidade inerente as suas características ou atributos. Rosen (1974) desenvolve uma proposição de Lancaster (1966) de que produtos possuem ou são capazes de gerar características em proporções fixas e que consumidores exercem suas preferências por meio de tais características e não particularmente do produto em si.

Tomando particularmente como exemplo o modelo hedônico desenvolvido por Yeh, Hsu e Weight (2018) observa-se que os valores de imóveis são estimados com base em sua distância ao ponto de metrô mais próximo, do número de lojas de conveniência próximas, idade do imóvel e data da transação e suas coordenadas geográficas. Percebe-se que o modelo gerado busca relacionar o valor do imóvel com atributos e a utilidade gerada aos moradores. Tal utilidade é evidente para variáveis como distância ao ponto de metrô mais próximo (facilidade de transporte) e quantidade de lojas de conveniência próximas (acesso a produtos próximos de sua residência). Outras, trazidas, por exemplo, pelas coordenadas geográficas do imóvel, podem não ser tão evidentes assim.

A prática de avaliação de imóveis urbanos no Brasil é normatizada pela NBR 14653 – Avaliação de Bens. Enquanto sua primeira parte recomenda procedimentos gerais sobre avaliações, a segunda parte trata da avaliação de imóveis urbanos (ABNT, 2011). Na primeira parte, encontram-se a classificações dos métodos de avaliação (ABNT, 2001):

- Método comparativo direto de dados de mercado: que estima o valor do imóvel alvo com base na comparação dos seus atributos com os dos imóveis da amostra;
- Método involutivo: que estima o valor com base numa análise de viabilidade técnico-econômica de uma situação hipotética em que um imóvel compatível com as características do imóvel-alvo e com as condições postas pelo mercado, assumindo seu uso eficiente e formulando cenários viáveis para execução e comercialização do produto;
- Método evolutivo: quando a estimativa é feita a partir da soma de estimativas de cada componente do imóvel-alvo;
- Método de capitalização de renda: quando a estimativa se utiliza de estimativas da capitalização da renda líquida futura oriunda do uso do imóvel.

Alguns pontos relevantes sobre a NBR 14653:

- O avaliador deve, sempre que possível, dar preferência ao método comparativo;
- Estabelece a quantidade mínima de amostras para cada método utilizado e para cada grau de fundamentação pretendido, recomendando, porém, a obtenção do "máximo possível" de dados;
- Quando da utilização de modelos de regressão, o nível de significância mínimo para teste de hipótese de regressores é de 10% (teste bicaudal) para se obter maior grau de fundamentação;
- Ainda para modelos de regressão, o nível de significância para teste de hipótese nula do modelo é de 1% para se obter o maior grau de fundamentação;
- Não há na norma especificação de métricas máximas de erro do modelo;
- A norma apresenta um anexo informativo sobre o uso de redes neurais artificiais (ABNT, 2011).

Apesar da diferença na denominação usada, os métodos listados na NBR 14653 também podem ser categorizados pelas perspectivas financeiras e comparativas, estando a proposta de classificação de Yeh, Hsu e Weight (2018) alinhada com o normativo nacional.

Antipov e Pokryshevskaya (2012) informam que a escolha tradicional para modelos de avaliação em massa são modelos de regressão. Essa afirmativa é também feita por Kontrimas e Verikas (2011). Os modelos de regressão possuem forma geral conforme expressa na Equação (1).

$$y = X\beta + \varepsilon \quad (1)$$

Onde:

y é um vetor de dimensão “n x 1” com as variáveis dependentes;

X é uma matriz de dimensão “n x m” com as variáveis independentes;

β é um vetor de dimensão “m x 1” que reúne os coeficientes da regressão;

ε é um vetor de dimensão “n x 1” com os erros de estimativa;

n é o número de observações da amostra;

m é o número de variáveis independentes da amostra.

A equação representa o relacionamento da variável dependente (y) com um conjunto de variáveis independentes (X) ponderado por um conjunto de coeficientes (β). A construção do modelo de regressão se dá a partir da estimativa do vetor β tal que se obtenha a condição expressa na Equação (2).

$$\beta = \min_{b} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \,(b) = \sum_{i=1}^{n} (y_i - X_i b)^2 \quad (2)$$

Onde:

y é o vetor de variáveis dependentes;

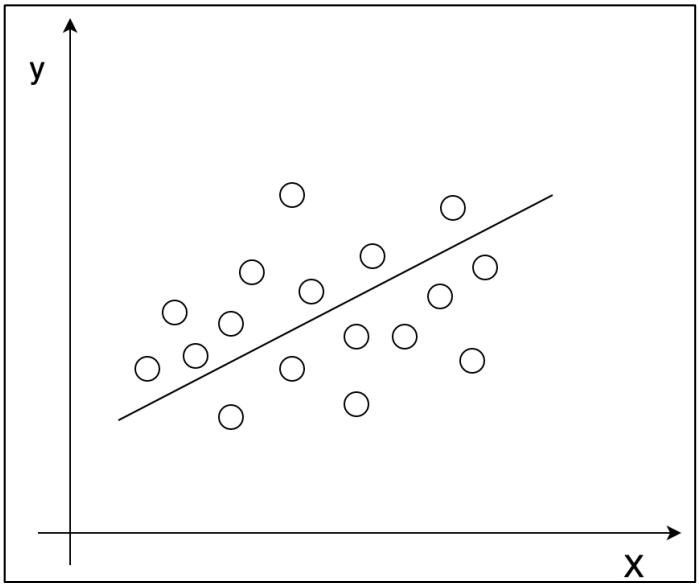
$\hat{y}$ é o vetor de valores estimados pelo modelo;

X é a matriz com as variáveis independentes;

b é o vetor com as estimativas dos coeficientes da regressão.

Portanto, trata-se de um problema de otimização que busca minimizar a diferença entre o valor real (y) e o valor estimado pelo modelo ($\hat{y}$) a partir dos coeficientes que há em “b” (KONTRIMAS; VERIKAS, 2011). A intuição por trás do problema é explicada com o auxílio de uma simplificação na Figura 5. A situação ótima, ou seja, de erros mínimos, é obtida quando as somas das distâncias entre os dados e a reta é o menor valor possível. Este método é chamado de Método dos Quadrados Ordinários (MQO).

Figura 5 – Intuição do Método dos Quadrados Ordinários (MQO).



FONTE: Elaboração própria.

Nunes (2016) realizou o ajustamento de um modelo de regressão linear demonstrado na Equação (3) e Equação (4).

$$Valor\ de\ mercado\ da\ unidade = 10^{Potência} - 1 \quad (3)$$

$$Potência = 12{,}19 + 8{,}895 \log_{10}(VG + 1) + 9{,}66 \log_{10}(MQ + 1) - 4{,}751 \log_{10}(UV + 1) - 13{,}07seg1 - 12{,}409seg2 - 8{,}268seg3 \quad (4)$$

Onde:

VG é a quantidade de vagas de garagem do imóvel;

MQ é o valor do metro quadrado em dezembro de 2014;

UV é a quantidade de unidades vendidas;

seg1 é uma variável "*dummy*" que corresponde ao segmento econômico;

seg2 é uma variável "*dummy*" que corresponde ao segmento médio;

seg3 é uma varável "*dummy*" que corresponde ao segmento alto.

Conforme se observar, a variável dependente e três das variáveis independentes foram transformadas com base na transformação logarítmica de base 10. Essa transformação é realizada quando se percebe que não há indícios de relação linear entre as variáveis e deve seguir análise prévia do relacionamento entre variáveis dependentes e independentes. No caso,

a transformação foi feita com base no valor da variável adicionado de 1, já que, em tese, o valor zero é possível para as variáveis VG e UV, valor para o qual função logarítmica não está definida. Relações não lineares podem também ser modeladas com assento em relações polinomiais de ordem superior a um, conforme é visto no trabalho de Yeh, Hsu e Weight (2018). Nesse caso em particular, em que os autores realizam a avaliação de imóvel não pela abordagem hedônica, mas sim estritamente comparativa, utilizou-se um modelo de regressão polinomial de segunda ordem para encontrar coeficientes para ajuste de preço em virtude da posição geográfica do imóvel. Empregou-se também um modelo logarítmico para encontrar coeficiente de ajuste para preços decorrente da distância do imóvel a um ponto de metrô e da idade do imóvel.

Ressalta-se que as variáveis “seg1”, “seg2” e “seg3” não são independentes entre si. Um imóvel não pode pertencer a dois segmentos ao mesmo tempo – um imóvel não pode ser considerado como pertencente ao segmento econômico e ao segmento alto ao mesmo tempo. Assim, na hipótese de um imóvel pertencer ao segmento econômico, este teria registrado na variável “seg1” o valor 1, na mesma medida que as variáveis “seg2” e “seg3” teria o valor zero.

O modelo obtido por Nunes (2016) construído por meio de *Ridge Regression* (RR), foi escolhido para combater os efeitos da multicolinearidade (fenômeno que distorce os resultados do modelo e que é causado pela correlação entre variáveis independentes). Nunes (2016) acentua que a multicolinearidade é comum nos modelos de avaliação imobiliária. Por exemplo, duas variáveis removidas do modelo foram “quantidade de quartos” e “área útil” e que devem possuir uma considerável correlação – quanto maior um apartamento (maior sua área útil), maior a quantidade de quartos definidos no projeto de arquitetura. Na matriz de correlação mostrada no estudo, verifica-se que a correlação entre ambas é de 0,724, indicando uma correlação significativa. A partir do modelo, Nunes (2016) obteve 80% das previsões dentro de um intervalo de até 20% do valor real, sendo cerca de 27% das previsões dentro de um intervalo de até 5% do real. O erro percentual absoluto médio foi de 21,1%.

O RR, conforme Exterkate *et al.* (2016) argumentam, também é utilizado para evitar o *overfitting*. Em comparação ao que se observa nos critérios de regressão linear múltipla, o RR funciona com esteio no controle da magnitude da estimativa do vetor de coeficientes β mediante um termo de regularização, observado na Equação (5) (tal equação foi adaptada de Pereira, Basto e Silva (2016), já que no estudo os autores trabalhavam com problemas de classificação e não de regressão).

$$\min_{b} \sum_{i=1}^{n} (y_i - X_i b)^2 + \lambda (b)^2 \quad (5)$$

Onde:

y é o vetor de variáveis dependentes;

X é a matriz com as variáveis independentes;

b é o vetor com as estimativas dos coeficientes da regressão;

λ é chamado de parâmetro de penalização.

No RR, o parâmetro $\lambda > 0$ corresponde a uma penalização definida pelo usuário que é imposta à função de custos. Na prática, essa penalização força valores menores para os coeficientes de "b". Observa-se que, para um mesmo conjunto de variáveis dependentes e independentes, o valor mínimo das Equações (2) e (5) não se altera. Assim, sendo $\lambda > 0$, para se obter o mesmo valor mínimo para a função custo, é necessário reduzir os valores de b. Estes, por sua vez, em razão de a incidência de λ ser a mesma para cada componente do vetor b, todos são afetados igualmente. Pereira, Basto e Silva (2016) ressaltam que o RR aproxima os valores de b de zero, porém não induz que os coeficientes sejam zerados, motivo pelo qual tal método não funciona como selecionador de variáveis.

O Quadro 2 reúne informações básicas sobre a precisão dos modelos encontrados nos estudos revisados. Primeiramente, é notória uma predominância da utilização da regressão com MQO, que, segundo Nunes (2016), é vulnerável ao problema da multicolinearidade comum aos modelos de avaliação de imóveis. Dentre os métodos de regressão também comumente utilizados, observam-se outros dois tipos: *Least Absolute Shrinkage and Selection Operator* (LASSO) e *Reweighted Least Squares* (RLS), que serão discutidos mais à frente.

Quadro 2 – Precisão dos modelos regressão linear múltipla revisados.

Estudo	Tipo de Regressão	Métrica	Valor
(KONTRIMAS; VERIKAS, 2011)	MQO	EPAM* e valores inaceitáveis**	15,02% e 31
(PETERSON; FLANAGAN, 2009)	MQO	Média dos EPAM dos modelos, mínimo e máximo EPAM	23%, 19,3% e 28,5%
(MCCLUSKEY et al., 2013)	MQO	EPAM	12,27%
(WORZALA; LENK; SILVA, 1995)	MQO	EPAM (mais de um modelo)	Entre 11,1% e 15,2%
(JANSSEN; SÖDERBERG; ZHOU, 2001)	MQO RLS	Diferença percentual do valor estimado entre os dois modelos	Entre 1,0% e 10%
(YOU et al., 2017)	LASSO	EPAM (mais de um modelo)	16,92% e 24,83%
(NGUYEN; CRIPPS, 2001)	MQO	EPAM e percentual de erro acima de 15% (vários modelos)	Entre 10,9% e 18,3%, entre 21 e 31%.
(ANTIPOV; POKRYSHEVSKAYA, 2012)	MQO	EPAM (mais de um modelo)	18,3% e 20%

Fonte: Elaboração própria.

* Erro percentual absoluto médio.

** Valores cujo erro supera 20%.

Na compilação das métricas do Quadro 2, foi priorizada a observação da medida do erro percentual absoluto médio (EPAM) por ser uma medida relativa, não dependendo da ordem de grandeza da variável dependente (que, por ser uma variável de preço, é traduzida nas mais diversas moedas, além da utilização da medida de preço por unidade de área). Cabe salientar que a maioria dos trabalhos revisados utiliza não apenas uma métrica de erro para a avaliar a precisão do modelo, mas sim um conjunto de métricas. Para um mesmo conjunto de dados, no entanto, e para uma mesma configuração do modelo, usa-se apenas um conjunto de métricas. A comparação entre dois modelos, por exemplo, um hipotético modelo A e um outro hipotético modelo B, se dá mediante comparação do conjunto de métricas de A com o conjunto de métricas de B. Testes estatísticos que comprovem a significância das diferenças entre os conjuntos de métricas não foram feitos.

A regressão linear do tipo LASSO envolve, assim como no RR, um termo de regularização para estimativa dos coeficientes da equação de regressão b, porém, num formato diferente. O método LASSO, contrariamente ao RR, induz que coeficientes de variáveis menos representativas sejam zeradas, servindo também na função de seletor de variáveis independentes (PEREIRA; BASTO; SILVA, 2016). A Equação (6) exprime o formato da função de custo do método LASSO, que também foi adaptada de Pereira, Basto e Silva (2016) pelo mesmo motivo.

$$\beta = \min_{b} \sum_{i=1}^{n} (y_i - X_i b)^2 + \lambda |b| \quad (6)$$

Onde:

y é o vetor de variáveis dependentes;

X é a matriz com as variáveis independentes;

b é o vetor com as estimativas dos coeficientes da regressão;

λ é chamado de parâmetro de penalização.

A condição $\lambda > 0$ permanece para o LASSO, com a diferença de que em vez de usar os coeficientes elevados ao quadrado, utiliza-se o módulo. Pereira, Basto e Silva (2016) afirmam que no caso em que poucas observações (n) existem para a elaboração do modelo (quando o número de variáveis dependentes possui ordem de grandeza similar à ordem de grandeza de n), o LASSO tende a oferecer melhores resultados. Quando se observa que os coeficientes possuem a mesma ordem de grandeza, o RR tende a oferecer melhores resultados. Os autores recomendam a utilização da técnica de validação cruzada para verificar qual método se aplica melhor para um conjunto de dados qualquer.

O RLS é considerado por Janssen *et al.* (2001) como uma técnica de regressão robusta, sendo utilizado também no intuito de se minimizar os efeitos de *outliers*. Ao contrário do MQO, o RLS utiliza o estimador chamado de "Mínimos Quadrados Medianos" (MQM) calculado conforme expresso na Equação (7) (adaptada de Janssen et al., 2001).

$$\beta = \min_{b} mediana_i (y_i - X_i b)^2 \quad (7)$$

Onde:

y é o vetor de variáveis dependentes;

X é a matriz com as variáveis independentes;

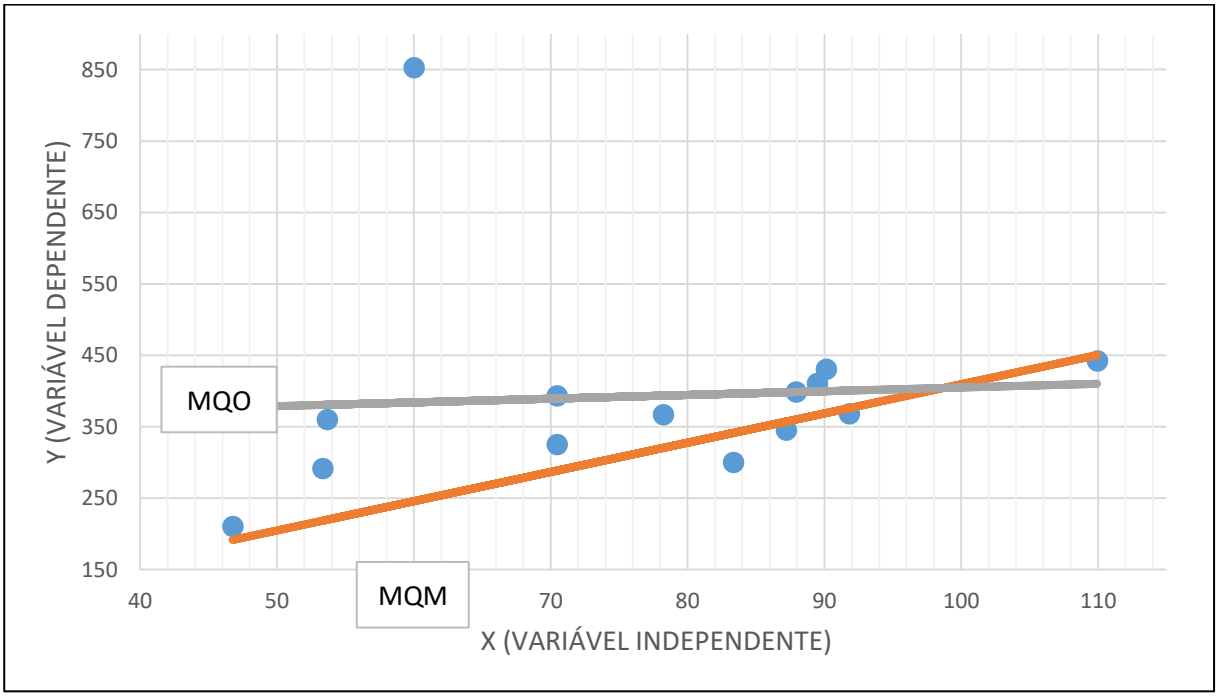
b é o vetor com as estimativas dos coeficientes da regressão.

Enquanto no MQO se minimiza a soma dos erros obtidos a partir da estimativa dos coeficientes b, no MQM objetiva-se a minimização da mediana de tais erros. A mediana é uma medida de tendência central robusta cujo ponto de ruptura é 50% e, no tocante ao estimador

MQM, oferece robustez contra *outliers* tanto no vetor y quanto na matriz X. O ponto de ruptura é a fração máxima que um estimador absorve de "contaminação" (valores discrepantes muito maiores ou muito menores do que os dados "não-contaminados") e ainda assim oferecer uma estimativa cuja ordem de grandeza semelhante ao que se obteria caso os *outliers* fossem removidos.

Na Figura 6, observa-se um exemplo do resultado de ambos os modelos com um valor bastante discrepante. Notam-se para um mesmo conjunto de dados equações bastante distintas para o modelo de regressão – o coeficiente angular do modelo estimado pelo MQM é cerca de 8 vezes o coeficiente do modelo estimado pelo MQO.

Figura 6 – Comparação entre MQO e MQM



FONTE: Elaboração própria.

Janssen *et al.* (2001) realizaram o que chamaram de "ponderação" da base de dados com suporte no resultado obtido com o método que utiliza a equação do MQM. A ponderação obedeceu às seguintes regras:

- cálculo dos erros oriundos do ajuste modelo MQM para cada observação do conjunto de dados (os erros são chamados de r);
- cálculo de um parâmetro de ponderação (chamado de s) para cada observação;
- caso a razão entre r_i e s_i seja inferior a 2,5, w_i, que é um vetor com os pesos para a ponderação, recebe o valor 1. Do contrário, w_i recebe o valor zero;

- estimar novamente os coeficientes desde a função de custo expressa na Equação (8).

$$\beta = \min_{b} \sum_{i=1}^{n} w_i (y_i - X_i b)^2 \quad (8)$$

O RLS constitui um MQO executado num subconjunto dos dados originais, sendo a seleção executada desde os coeficientes obtidos inicialmente com o MQM. No estudo de Janssen *et al.* (2001), o modelo MQO indicava a maior estimativa dentre ambos, e a diferença observada entre os valores previstos pelo MQO e pelo RLS variavam de 1 a 10%.

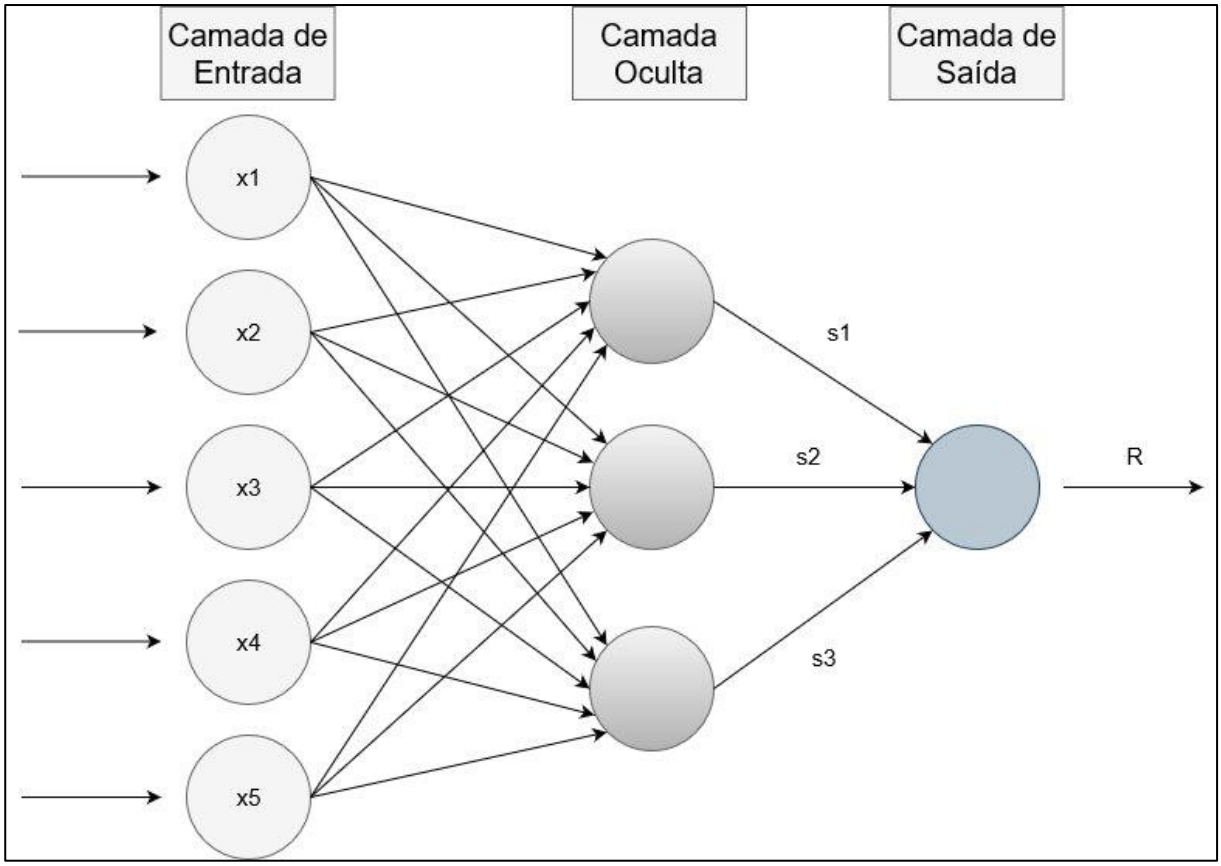
Em contraposição aos modelos de regressão linear, Peterson e Flanagan (2009) argumentaram que o desempenho de um modelo de preços hedônicos depende diretamente do quanto este é capaz de minimizar os erros de suas previsões. Estes argumentam que uma falha dos modelos de regressão linear para avaliação de imóveis está na própria premissa de relação linear entre o valor e seus atributos. Por exemplo, um modelo hipotético que relaciona o preço de venda de um imóvel com sua área construída assume que cada unidade adicional de área possui o mesmo valor das demais. Porém, afirma-se que há diferença entre uma unidade adicional de área alocada numa laje técnica e uma unidade adicional de área disposta numa sala de estar.

Peterson e Flanagan (2009) argumentam que modelos não lineares são preferíveis por não estarem sujeitos a essas premissas potencialmente falhas. Não obstante, os modelos não lineares adequam-se às relações lineares. Nos trabalhos de Nunes (2016) e Yeh, Hsu e Weight (2018) foi contornado o problema da não linearidade mediante a transformação de variáveis. Peterson e Flanagan (2009) argumentam, porém, que recorrer a esse artifício esconde riscos pois não existem passos práticos para se adereçar exatamente o relacionamento não-linear. Argumentam também que as Redes Neurais Artificiais (RNA) são alternativa prática para contornar tais dificuldades impostas pelas relações não-lineares e conseguem, através de uma implementação fácil, modelar de maneira eficiente tais relações.

Peterson e Flanagan (2009) obtiveram resultados mais precisos com a aplicação das RNA (medidos a partir do EPAM). Na simulação em que 10% das observações da amostra foram deixadas para testar as previsões do modelo, as medidas de EPAM para os anos de 1999 a 2005 do modelo RNA resultaram média igual a 21%, enquanto para o modelo MQO a média foi de 23%.

A Figura 7 mostra um exemplo de RNA com uma camada oculta com três nós, seis variáveis independentes e uma saída.

Figura 7 – Exemplo de Rede Neural Artificial.



FONTE: Adaptado de Peterson e Flanagan (2009).

Cada nó da camada oculta recebe todas as variáveis de entrada ponderadas por pesos e transmite o resultado de uma função de ativação definida pelo usuário. No caso, o primeiro nó da camada oculta transmite para a camada de saída o valor s1, calculado conforme expresso na Equação (9).

$$s1 = f_1(W_1^1 X + b_1) \tag{9}$$

Onde:

W_1^1 é o vetor de pesos do nó 1;

b_1^1 é o viés do nó 1 (equivalente ao intercepto dos modelos de regressão);

f_1 é a função de ativação da camada escondida.

O mesmo conjunto de operações ocorre para s2 e s3, que são combinados também por meio de pesos no nó da camada de saída, sendo também transformado por outra função de ativação. O resultado dessa transformação é mostrado como saída da rede R, que é calculado conforme a Equação (10).

$$R = f_2(W_2 S + b_2) = f_2(W_2 f_1(W_1 X + b_1) + b_2) \quad (10)$$

Onde:

S é o vetor que reúne os valores s1, s2 e s3;

W_1 é a matriz de pesos das entradas da camada escondida;

b_1 é o vetor de vieses da camada escondida;

b_2 é o vetor de vieses da camada de saída;

W_2 é a matriz de pesos de s1, s2 e s3 na camada de saída;

f_1 é a função de ativação da camada de saída;

f_2 é a função de ativação da camada de saída.

As RNAs são elaboradas conforme necessidades do estudo. A camada oculta da Figura 7 possui três nós, porém, a camada poderia ter sido construída com quatro ou cinco nós, ou toda a rede poderia ter sido feita contendo mais de uma camada oculta, entrando no âmbito específico da *Deep Learning* (MEYER et al., 2018). Além disso, várias são as escolhas de funções de ativação - Peterson e Flanagan (2009) utilizaram na camada oculta a função arco-tangente e na camada de saída uma função de soma. São outras funções de ativação: sigmóide, gaussiana, *LeakyRelu*, retificadora linear, dentre outras (ACHARYA et al., 2017; MCALLISTER et al., 2018).

O método, porém, é considerado por autores como uma "caixa-preta" (WORZALA; LENK; SILVA, 1995; ZURADA; LEVITAN; GUAN, 2011; MCCLUSKEY et al., 2013; MABU; OBAYASHI; KUREMOTO, 2015; WIŚNIEWSKI, 2017). A sucessão de ponderações e de aplicações de funções de ativação torna impossível, assim como se pode nos modelos de regressão linear discutidos previamente, reduzir a influência de uma variável a um só coeficiente. Contorna-se, entretanto, o problema utilizando simulações (PETERSON; FLANAGAN, 2009).

Kontrimas e Verikas (2011) oferecem, também, como método não-linear para problemas de regressão tais quais os modelos de avaliação em massa de imóveis o *Support*

Vector Machine (SVM), também chamado de *Support Vector Regression* (SVR). Com o SVM, os dados são utilizados para estimar uma função não-linear num espaço vetorial induzido por um "*kernel*", entendido como uma função que simula operações de produto escalar entre vetores e produz uma transformação espacial.

A função de regressão do SVM tem seus parâmetros α, α^* e b obtidos através da solução do problema dual de Lagrange, um problema de otimização. A função está na Equação (11) (KONTRIMAS; VERIKAS, 2011).

$$y = \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) \mathrm{K}(x_i, x) + b \tag{11}$$

Onde:

α e α^* multiplicadores de Lagrange não negativos;

$\mathrm{K}(x_i, x)$ é a função "*kernel*" escolhida;

b é o viés da equação (semelhante ao intercepto);

y é a variável dependente.

Kontrimas e Verikas (2011), utilizando um *kernel* polinomial cúbico, reportaram métricas de desempenho de um modelo de regressão com SVM para avaliação de imóveis usando dados da Lituânia superiores aos resultantes de um modelo com redes neurais e regressão múltipla. Kontrimas e Verikas (2011) creditam o melhor desempenho do SVM ao fato do algoritmo permitir lidar com não-linearidades, apesar do algoritmo RNA também lidar com não-linearidades. Lam, Yu e Lam (2009), que utilizaram *kernel* com função de base radial, oferecem para uma amostra de 21 imóveis resultados de erros absolutos menores comparativamente aos obtidos com RNA e regressão múltipla.

2.3 Análise Comparativa entre Modelos

No Quadro 3, há comparações entre resultados de modelos de avaliação de imóveis construídos com métodos de regressão e RNA. Em apenas dois dos seis trabalhos onde ambos modelos foram comparados, as redes neurais tiveram desempenho pior do que os modelos de regressão (KONTRIMAS; VERIKAS, 2011; ANTIPOV; POKRYSHEVSKAYA, 2012). As RNAs tiveram, portanto, melhor desempenho, ainda que sensível em alguns casos (WORZALA; LENK; SILVA, 1995; NGUYEN; CRIPPS, 2001; PETERSON; FLANAGAN,

2009; MCCLUSKEY et al., 2013). Ressalta-se que tipicamente os resultados dos estudos são mostrados desde uma medida de uma ou mais métricas de erro. Ou seja, Kontrimas e Verikas (2011) oferecem apenas uma medida de EPAM e erro médio absoluto (cuja unidade está na unidade monetária da Lituânia, "Lt") para cada um dos modelos testados (Quadro 4).

Quadro 3 – Comparação de métricas de erros dos modelos de regressão e de redes neurais.

Estudo	Métrica	Medida de Erro Regressão	Medida de Erro Redes Neurais
(KONTRIMAS; VERIKAS, 2011)	EPAM e UV	15,02% e 31	23,3% e 42
(PETERSON; FLANAGAN, 2009)	Média dos EPAM dos modelos, mínimo e máximo EPAM	23%, 19,3% e 28,5%	21,75%, 17,5% e 26%
(MCCLUSKEY et al., 2013)	EPAM	12,27%	11,97%
(WORZALA; LENK; SILVA, 1995)	EPAM	Entre 11,1% e 15,2%	Entre 10% e 14,4%
(NGUYEN; CRIPPS, 2001)	EPAM e percentual de erro acima de 15%	Entre 10,9% e 18,3%, entre 21% e 31%.	Entre 7,1% e 15,5%, entre 8,5% e 33,5%
(ANTIPOV; POKRYSHEVSKAYA, 2012)	EPAM	18,3% e 20,02%	20,53% e 19,79%

FONTE: Elaboração própria.

Quadro 4 – Típicos resultados de precisão de modelos de avaliação de imóveis.

Modelo	EPAM (%)	Erro médio absoluto (Lt)	Valores Inaceitáveis
Regressão MQO	15,02	65260.24	31
RNA	23,30	90255.57	42
SVM	13,62	59055.77	18

FONTE: Adaptado de Kontrimas e Verikas (2011).

Nguyen e Cripps (2001) realizaram 108 comparações ao todo entre modelos de regressão múltipla com RNA. Ao todo, seis modelos de regressão e RNA foram comparados sob 18 formatações diferentes de bases de dados de treino e validação. De posse de 3906 observações, os autores dividiram os dados com base em amostragens aleatórias em conjuntos T1 até T18 e V1 até V18. Os conjuntos de treino e validação variam em tamanho de 200 em 200 (T1 contém 306 observações, T2 contém 506, T3 contém 706 e assim por diante, enquanto V1 possui 3600, V2 possui 3400, V3 possui 3200 e assim sucessivamente).

Para cada par de modelos de regressão e RNA gerados, existem 18 parâmetros de EPAM, o que permite uma avaliação mais profunda da diferença de desempenho, ainda que os resultados sejam produzidos com configurações diferentes de conjunto de dados. Nguyen e Cripps (2001) mostraram que das 108 comparações realizadas a partir da métrica de erro EPAM,

apenas em cerca de 17% os modelos de regressão registraram melhor precisão. Nenhum teste estatístico foi aplicado nos erros calculados.

Demšar (2006) argumenta que, ao observar uma melhoria num resultado a partir de uma mudança num algoritmo, é necessário como passo seguinte testar a significância estatística da hipótese de que tal melhoria é fruto da alteração realizada e que não se deu por mera aleatoriedade, principalmente em casos em que tal melhoria é marginal (GOETZ et al., 2015). Demšar (2006), apesar de abordar problemas de classificação, propõe testes para averiguar as diferenças na precisão entre modelos adaptáveis a problemas de regressão. A comparação seria feita com apoio nos resultados obtidos para diferentes amostras. Dentre os testes possíveis, menciona-se a média dos resultados, teste t e de postos sinalizados de Wilcoxon para comparação em pares, ANOVA e testes de Friedman para comparação de múltiplos modelos.

O primeiro teste seria a comparação entre as médias do percentual de classificações corretas feitas pelos modelos. Neste teste, se o modelo A obteve classificações corretas em três conjuntos de dados distintos iguais a 85%, 87% e 90% (resultando numa média igual a 87,3%), enquanto o modelo B teve para os mesmos conjuntos 82%, 85% e 83% (cuja média é 83,3%), suspeita-se que o modelo A possui desempenho melhor que o modelo B. Conforme Demšar (2006) afirma, todavia, além de as médias serem sujeitas a *outliers*, há questionamentos sobre a possibilidade de se comparar resultados de classificação de distintos conjuntos de dados, a não ser quando os modelos são submetidos a problemas que guardam um certo relacionamento entre si – como leitura de imagens de diagnósticos médicos ou mineração de texto, por exemplo. Demšar (2006) considera o uso deste teste como pouco comum e não se trata de um teste de hipótese.

O teste t em pares checa se a média entre as diferenças obtidas entre os dois modelos é significativamente diferente de zero – ou seja, se há de fato de diferença entre os resultados entre os modelos. Sendo c_i^1 e c_i^2 a métrica de desempenho de dois modelos diferentes para um mesmo conjunto de dados i, e d_i a diferença entre os dois, a estatística t é computada pela razão $\bar{d}/\sigma_{\bar{d}}$ de acordo com a distribuição t de Student com N-1 graus de liberdade, sendo N o número de conjunto de dados usados no teste, $\bar{d}$ a média das diferenças entre as métricas e $\sigma_{\bar{d}}$ o desvio-padrão das diferenças. Assim como o caso da comparação das médias, o teste t também compara diretamente resultados de classificação de variados conjuntos de dados, ainda que na forma de diferença. Ainda, tal teste demanda que N seja superior a 30 caso não seja possível atestar que d_i segue distribuição normal (DEMŠAR, 2006).

O teste de postos sinalizados de Wilcoxon (denominado de teste de Wilcoxon) é um

teste não paramétrico. Classifica-se as diferenças das métricas de desempenho dos dois modelos em uma ordem (*ranks*) em termos absolutos, calculando a média em caso de empate (se em dois casos a diferença foi igual a zero, ambos os casos recebem 1,5, que é a média entre a primeira e segunda posição). Daí, calcula-se a soma das diferenças positivas e negativas conforme as Equações (12) e (13).

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{12}$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{13}$$

Observa-se que as posições atribuídas à diferença igual a zero são divididas entre a soma de diferenças positivas e negativas. Sendo $T = \min(R^+, R^-)$ e N o número de observações, obtêm-se valores críticos para o teste de Wilcoxon para a hipótese nula de que ambos os modelos possuem desempenho igual. Caso T seja menor ou igual ao valor crítico, rejeita-se a hipótese nula. Demšar (2006) afirma que o teste Wilcoxon é mais seguro pois não assume como sendo normal a distribuição das diferenças, sendo recomendado por ele caso não sejam atendidas as condições necessárias para o teste t. Quando, porém, as condições do teste t são atendidas, este deve ser preferido.

Para a comparação de três ou mais modelos entre si, Demšar (2006) recomenda a utilização da análise de variância (ANOVA) de medidas repetidas. O teste considera as métricas de desempenho de cada modelo para cada conjunto de dados, de preferência utilizando a mesma divisão entre grupos de treinamento e grupos de teste. O teste é feito para a hipótese nula de que não há diferença entre o desempenho de cada modelo e que as variações são fruto de mera aleatoriedade. Conforme Demšar (2006), a ANOVA segrega a variabilidade total em variabilidade entre modelos, entre os dados e entre os resíduos, portanto, caso a variabilidade entre modelos seja superior ao que se observa na variabilidade dos erros, rejeita-se a hipótese nula num dado nível de significância e se afirma que há diferença entre os modelos.

Sendo confirmada a rejeição da hipótese nula, sabe-se apenas que dentre os modelos, pelo menos um deles é diferente dos demais. Para identificar qual deles é o que possui desempenho destoante, realiza-se um teste post-hoc, sendo o teste de Tukey adequado por permitir a comparação pareada entre um modelo com cada um dos outros (DEMŠAR, 2006). Demšar (2006) ensina que as premissas necessárias para a utilização da ANOVA (as amostras partem de uma população que segue distribuição normal e esfericidade) são frequentemente

violadas nas aplicações de aprendizado de máquinas, não sendo seu uso recomendado.

O teste de Friedman é interpretado como um teste ao equivalente não paramétrico da ANOVA. Assim como no teste de Wilcoxon, o teste de Friedman também trabalha com ranqueamento das métricas, porém, cada modelo recebe uma classificação para cada conjunto de dados. Assim, caso os resultados expressados pelos modelos de classificação A, B e C sejam 83%, 91% e 87%, as posições destes modelos serão respectivamente 3, 1 e 2. Caso houvesse algum empate, os modelos empatados receberiam a média das posições. Sendo r_i^j a posição do j-ésimo modelo de um total de k modelos para o i-ésimo conjunto de dados de um total de N conjuntos, o teste é feito com suporte na comparação da média da posição dos modelos, $R_j = \frac{1}{N}\sum_i^N r_i^j$. A estatística de Friedman é calculada com a Equação (14) (DEMŠAR, 2006).

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j^k R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{14}$$

A estatística do teste de Friedman segue distribuição chi-quadrado com k-1 graus de liberdade, quando N e k são superiores a 10 e 5 respectivamente. Caso seja possível rejeitar a hipótese nula (todos os modelos possuem desempenho igual), parte-se para o teste post-hoc de Nemenyi, que, semelhante ao teste de Tukey para a ANOVA, permite a comparação de cada modelo com os outros. Segundo o teste de Nemenyi, há diferença significativa entre dois modelos caso a esta diferença seja superior à chamada diferença crítica (CD), calculada com a Equação (15), sendo q é o valor crítico da estatística do intervalo "studentizado" correspondente ao quantil de significância almejado dividido por $\sqrt{2}$ (DEMŠAR, 2006).

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{15}$$

Todas as aplicações sugeridas por Demšar (2006) se dão para comparações com variados conjuntos de dados, fator que é reconhecido pelo autor como limitante para aferir diferenças entre os modelos. Para contornar tal problema, Granatyr (2017) sugere um protocolo de teste para algoritmos utilização de comparações para um mesmo conjunto de dados. No protocolo, o autor faz a alteração da semente geradora (*seed*) de 1 a 30, obtendo assim 30 leituras do desempenho do modelo para a mesma base.

2.4 Abordagem Hedônica Espacial

O fenômeno da autocorrelação espacial é algo frequentemente observado em situações em que a localização é um fator determinante. No caso particular do estudo dos preços do mercado imobiliário, os efeitos da localização são frequentemente atribuídos a variáveis chamadas de fatores de vizinhança, condição que resulta em erros de estimativa (DUBIN, 1998).

São duas as principais causas de erros ocasionados pelos fatores de vizinhança. Primeiramente, a vizinhança é algo não observável, sendo necessário o uso de variáveis *proxy* para incorporar os efeitos da localização. Por exemplo, os efeitos da localização podem ser incorporados ao modelo através de fatores como taxas de criminalidade ou características socioeconômicas dos residentes. Com relação ao segundo ponto, que age complementarmente ao primeiro, ocorre que, para se empregar os *proxies*, deve-se impor alguma fronteira geográfica. Por exemplo, informações sobre criminalidade e características socioeconômicas podem ser compiladas por bairros – assim, um imóvel teria assinalada uma variável correspondente às informações de criminalidade do bairro em que está localizado, por exemplo. Incorporar tais *proxies* num nível de bairro assume que toda as vizinhanças de um tal bairro possuem comportamento uniforme, algo que não é acurado, principalmente para bairros maiores e heterogêneos (DUBIN, 1998).

Wen, Bu e Qin (2014) afirmam que uma forma de lidar com o fenômeno da heterogeneidade espacial no mercado imobiliário era o ajuste de vários modelos correspondentes a subsetores geográficos do mercado. A modelagem realizada por Codes (2018) segue em conformidade com essa diretriz. O autor treina diferentes redes neurais artificiais para anos diferentes e bairros diferentes da cidade de Fortaleza.

Outros estudos incorporam variáveis do tipo *dummy* que modificam a equação de regressão tanto no intercepto quanto nos coeficientes (WEN; BU; QIN, 2014). Tal solução é observada no trabalho de Nunes (2016), quando ao todo 18 variáveis *dummy* foram incorporadas inicialmente ao modelo para simular cartoze zonas de valor e quatro segmentos de imóveis para o mercado de Fortaleza. Pelo que se observa nas Equações (3) e (4), apenas três das variáveis *dummy* exibiram significância estatística num nível adequado.

Wen, Bu e Qin (2014) comentam que a alternativa para se tratar da ocorrência da dependência espacial na análise dos imóveis que tem gerado bons resultados é regressão espacial, realizada através de modelos de defasagem espacial (*Spatial Lag Model* – SLM), ou modelos de erros espaciais (*Spatial Error Model* – SEM). O SEM é recebe de Anselin (2003) a

denominação de modelo espacial autoregressivo (*Spatial Autoregressive Model* – SAM), e é o termo que será utilizado no presente trabalho. O formato geral de um SLM é apresentado na Equação (16).

$$P = \rho WP + X\beta + \mu \tag{16}$$

Onde:
P é a variável dependente;
ρ é o parâmetro de correlação espacial;
W é a matriz de ponderação espacial;
X é a matriz de variáveis independentes;
β é o vetor de coeficientes;
μ é o vetor de erros do modelo espacial.

Comparando a Equação (16) com a Equação (1), observa-se a Matriz de Ponderação Espacial (*Spatial Weight Matrix* – SWM), sendo multiplicada pelo preço dos imóveis (variável dependente). Assim, WP representa um conjunto de variáveis dependentes que representam a defasagem espacial do preço de cada imóvel – em outras palavras, WP reúne a contribuição que o preço de um dado imóvel i recebe de seus vizinhos seguindo as regras definidas durante a modelagem. As variáveis WP são ponderadas pelo parâmetro ρ, chamado de correlação espacial ou dependência espacial, com função semelhante ao β, coeficiente das variáveis dependentes que caracterizam o imóvel. Caso a correlação ou dependência espacial seja próxima de zero, ou não denote significância estatística durante a regressão, assegura-se que o fenômeno da dependência espacial é não representativo na formação do preço do imóvel (WEN; BU; QIN, 2014).

No SAM, o efeito da dependência espacial é aplicado no termo de erros do modelo de regressão – termo ε encontrado na Equação (1) – através da matriz espacial. Assim, em vez da contribuição das observações mais próximas atuarem diretamente no preço, elas influenciam os desvios produzidos pelo modelo de regressão múltipla comum. A equação geral do SAM é expressa na Equação (17).

$$P = X\beta + \lambda W\varepsilon + \mu \tag{17}$$

Onde:

P é a variável dependente;

λ é o parâmetro de correlação espacial;

W é a matriz de ponderação espacial;

X é a matriz de variáveis independentes;

β é o vetor de coeficientes;

ε é o vetor de erros do modelo de regressão;

μ é o vetor de erros do modelo espacial.

A SWM é essencial para a modelagem da dependência especial, como pode-se observar, porém, surge nos estudos com notória arbitrariedade, conforme percebe Rincke (2010). Sendo uma imposição a priori, a SWM atribui uma medida do grau de "vizinhança" entre as observações. Portanto, caso exista um conjunto de dados com N observações, W terá dimensão N x N e seu termo w_{ij} representa uma medida da influência que a observação i terá na observação j. Uma consequência dessa definição conduz para a necessidade de se ter na diagonal principal da matriz preenchida com valores nulos, para que uma observação não gere influência nela mesma no modelo ajustado. Na Figura 8 há um exemplo de SWM para um conjunto de N observações.

Figura 8 – Exemplo de Matriz Espacial de Ponderação.

$$W = \begin{bmatrix} 0 & w_{12} & w_{13} & \cdots & w_{1N} \\ w_{21} & 0 & w_{23} & \cdots & w_{2N} \\ w_{31} & w_{32} & 0 & \cdots & w_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & w_{N3} & \cdots & 0 \end{bmatrix}$$

FONTE: Elaboração própria.

Rincke (2010) ainda cita que duas regras para a construção da SWM surgiram entre os estudiosos do tema. Uma delas é que há uma forte correlação entre a influência entre as observações e a distância entre ambas, sendo os pesos da matriz inversamente proporcionais à distância, obtendo assim pesos maiores entre observações mais próximas. Esta regra conduz à utilização de decaimentos obtidos desde a distância euclidiana entre as observações ou com procedência em conjuntos dos *k-Nearest Neighbors* (os k vizinhos mais próximos, cuja sigla em inglês é kNN). A segunda regra afirma que os pesos da SWM devem ser exógenos ao modelo, o que significa que há independência dos pesos em relação ao termo de erros do modelo

linear. Segundo Kostov (2010), o uso massivo da distância geográfica na estruturação da SWM se dá em grande parte por conta da garantia automática de exogeneidade.

Anselin (2002), em seu trabalho, também afirma que faltam diretrizes formais para a obtenção dos elementos da SWM que modelam fielmente a relação de dependência espacial. Em certos casos, a obtenção se dá de forma *ad hoc*, demandando análises de sensibilidade dos resultados. Kostov (2010) sugere um algoritmo de seleção da melhor estrutura para a SWM a partir do teste de várias combinações. O autor utiliza a função inversa da distância elevada a uma certa potência (p), utilizando um conjunto de vizinhos mais próximos (k), conforme expresso na Equação (18). Fazendo k variar de 1 a 50 vizinhos mais próximos e a potência p variar entre 0,4 e 4 em intervalos de 0,1, o algoritmo testou 1850 possíveis matrizes (50 possíveis combinações de vizinhos mais próximos multiplicados por 35 possíveis combinações de potências para o inverso da distância).

$$w_{ij} = \frac{1}{(distância)_{ij}^{p}} \tag{18}$$

Assim, com a SWM cujo k é igual a 10 e p é igual a 2, o algoritmo irá selecionar as dez distâncias menores dentre todas as possíveis distâncias entre uma dada observação i e o restante do conjunto de dados, calcular o inverso do quadrado das medições e agrupá-las na linha i da matriz na correspondente coluna. Todos os outros elementos da matriz receberão o valor zero.

Seya, Yamagata e Tsutsumi (2013) também conduzem experimentos para a seleção de uma SWM dentre 34 candidatas com amparo em modelos complexos de simulação. Dentre as 34 possíveis, 4 utilizavam o formato expresso pela Equação (18), com a variação da potência p entre 1 e 4. Nesses casos, não se limitou o número de elementos por linha da matriz assim como feito por Kostov (2010). As outras 30 candidatas usaram a abordagem kNN, com k variando entre 1 e 30. O cálculo dos pesos da SWM com o kNN se dá de acordo com a Equação (19).

$$w_{ij} = \begin{cases} 1, & j \in k\ vizinhos\ mais\ próximos \\ 0, & restante\ dos\ valores \end{cases} \tag{19}$$

O estudo de seleção de SWM conduzido por Seya, Yamagata e Tsutsumi (2013) utilizou um modelo ajustado para um conjunto de dados imobiliários do Japão, mais

precisamente o preço de condomínios de 23 bairros de Tóquio. As variáveis independentes utilizadas no modelo caracterizam acessibilidade ao sistema de transporte, aspectos do imóvel, zoneamento e bairro onde se encontra o imóvel, sujeição a riscos ambientais, dentre outras características.

A variável dependente utilizada foi o preço dividido por unidade de área transformado com uma função logarítmica. Os autores utilizaram uma variação do modelo SLM em que a SWM é aplicada também a certas variáveis dependentes. Trata-se do modelo espacial de Durbin ("*spatial Durbin model*" – SDM), representado na Equação (20). Nela, $\tilde{X}$ representa a matriz de variáveis independentes sem o intercepto selecionadas para a ponderação espacial e γ são os parâmetros de dependência espacial.

$$P = \rho WP + X\beta + W\tilde{X}\gamma + \mu \tag{20}$$

Onde:

P é a variável dependente;

ρ é o parâmetro de correlação espacial;

W é a matriz de ponderação espacial;

X é a matriz de variáveis independentes;

$\tilde{X}$ é a matriz de variáveis independentes;

γ é o vetor de parâmetros de dependência espacial;

β é o vetor de coeficientes;

μ é o vetor de erros do modelo espacial.

O termo $W\tilde{X}$ no estudo em particular abrigou apenas as variáveis numéricas, nenhuma das variáveis do tipo "*dummy*" foi incluída. Tal modelo, com 1200 observações ao todo, com as 34 possibilidades de SWM e com diversas variáveis dependentes, levou mais de 3 dias ininterruptos para calcular todos os resultados (SEYA; YAMAGATA; TSUTSUMI, 2013).

Utilizando-se de um SAM, Fernandez, Mukherjee e Scott (2018) estudaram os efeitos de políticas de conservação ambiental, precisamente o impacto do "*Riverside County Integrated Plan*" (RCIP), que define diretrizes para a proteção ambiental, no preços de imóveis no Condado de Riverside, na Califórnia. Utilizando um conjunto de dados de registro de vendas de imóveis em Riverside e um conjunto análogo oriundo do Condado de San Bernardino como

grupo-controle, os autores puderam ajustar um modelo capaz de mostrar a evolução temporal dos preços de imóveis tendo como variável dependente distâncias a parques e áreas de proteção mais próximas. O modelo desenvolvido pelo autor é um modelo logarítmico com variável dependente o preço de venda registrado para o imóvel e engloba 52 variáveis independentes ao todo com caraterísticas como área, tamanho do lote, distância para aparelhos de lazer e áreas de preservação, percentual de densidade urbana, dentre outros.

O modelo de Fernandez, Mukherjee e Scott (2018) demonstrou indícios de influência positiva no valor dos imóveis no período posterior ao RCIP. Segundo os autores, um aumento de 1 milha na distância entre um certo imóvel e uma certa área de proteção ambiental em Riverside corresponde a uma penalidade de cerca de U$ 36 mil. A mesma condição em San Bernardino corresponde a uma perda cerca de 40% menor, por volta de U$ 20 mil.

2.5 Combinação de Modelos

Kontrimas e Verikas (2011) exibem uma modalidade de conjugação de três modelos hedônicos – regressão linear, rede neural artificial e *Support Vector Machine* (SVM) – que chamaram de "comitê de previsores". Segundo os autores, é de pleno conhecimento o fato de que tais comitês são capazes de aprimorar a acurácia das previsões, principalmente na combinação de modelos de classificação.

Gader, Mohamed e Keller (1996) afirmam que a "fusão" de múltiplos modelos classificadores é uma prática bastante recompensadora, sendo capaz de melhorar o desempenho através da diminuição dos erros. Afirmam ainda que pesquisas mostram que a combinação de diferentes modelos simples confere melhores resultados que a busca por um modelo único mais sofisticado. A mesma constatação acerca dos ganhos na combinação de classificadores é reforçada por outros estudos (KITTLER et al., 1998; ALEXANDRE; CAMPILHO; KAMEL, 2001; LIU, 2005; NEMMOUR; CHIBANI, 2005; MABU; OBAYASHI; KUREMOTO, 2015).

Anish, Majhi e Majhi (2016) afirmam, num contexto de estudos de previsão de valores de ativos financeiros, que a principal maneira de se agregar os resultados de previsores é a partir de uma equação de combinação linear. A Equação (21) apresenta o formato geral de uma combinação de m modelos de previsão, em que os resultados fornecidos por parte de cada modelo (My) são agregado com base em pesos (w).

$$My_n = w_1 My_n^1 + w_2 My_n^2 + \cdots + w_m My_n^m = \sum_{i=1}^{m} w_i My_n^i \qquad (21)$$

Segundo os autores, são três modalidades bem conhecidas de realizar a ponderação são:

- Média simples dos resultados dos modelos;
- Média "podada" dos resultados, quando uma certa porção de valores com desvios extremos é removida;
- Mediana dos resultados.

No trabalho em foco, os autores se utilizam de algoritmos genéticos e otimização por enxame de partículas (*Particle Swarm Optimization* – PSO). As métricas de erros apresentadas revelam que apesar do algoritmo optimizado pelo PSO ter logrado melhor desempenho, a variação entre este e o desempenho obtido pela combinação linear através da média, por exemplo, é pequena. Os erros obtidos pela otimização com algoritmos genéticos foram consideravelmente superiores. A Tabela 1 apresenta a comparação entre o EPAM dos três métodos de combinação obtidos para a previsão do valor de um fundo de investimento.

Tabela 1 – Comparação do EPAM de três métodos de combinação (em %).

Método de Combinação	**Horizonte de Previsão (dias)**				
	1	3	5	7	15
Média	1.3670	1.7842	2.2350	2.4120	3.3723
Algoritmo Genético	2.0211	3.1000	3.7595	4.5184	5.6716
PSO	1.0015	1.5500	2.0142	2.3628	3.3723

FONTE: Anish, Majhi e Majhi (2016).

Kempa *et al.* (2011) também se utilizam da combinação de modelos para obter melhores resultados em termos de desvio. A técnica empregada pelos autores foi o *bagging*, termo oriundo de *boostrap aggregation*, que gera aleatoriamente diversos subconjuntos de treino a partir do conjunto de dados original chamados de "bolsas" (tradução de *bags*). Tais "bolsas" são por sua vez submetidas a modelos de regressão e os resultados dos modelos são agregados com apoio em funções algébricas. Segundo Kempa *et al.* (2011), análises teóricas e resultados experimentais demonstraram que o *bagging* é capaz de reduzir a variância e garantir melhoria estável tanto para problemas de classificação como para regressão.

O *bagging* é, portanto, uma forma de combinação de modelos que exige uma manipulação do conjunto de dados. É uma técnica que, inclusive, é usada como alternativa para a mudança da semente geradora proposta por Granatyr (2017) para avaliação de modelos. Gerando 30 "bolsas" e calculando as métricas de erro, realiza-se a comparação entre os modelos e verificar a significância das diferenças de desempenho.

O efeito da redução da variância é observado nos resultados expressos por Kontrimas e Verikas (2011). Após a organização dos modelos em comitês, o resultado de EPAM permaneceu parecido ao encontrado pelo modelo em SVM, sendo o EPAM do comitê 1 igual a 13,61% e do comitê 2 é igual a 13,36%. O efeito maior incidiu sobre os valores inaceitáveis, com cada comitê exibindo apenas um valor inaceitável. Haja vista que apenas os valores com desvio superior a 20% são considerados como inaceitáveis, fica evidente o efeito de redução da variância dos resíduos.

Dentre os métodos considerados como *ensemble*, o *Random Forest* (RF) é definido por Breiman (2001) como um modelo classificador que consiste na coleção de outros classificadores do tipo árvore de decisão elaborados com base numa seleção aleatória das variáveis independentes e que cada árvore de decisão "vota" em apenas uma das possíveis classes. Breiman (2001) afirma que, em decorrência da Lei dos Grandes Números, o RF não está sujeito ao *ovefitting* com o aumento do número de árvores. Idealizado inicialmente como um método para modelos de classificação, também é utilizado como modelo de regressão, quando o resultado final do modelo é a média do resultado obtido em cada uma das "n" árvores de decisão geradas aleatoriamente.

Segundo Rudžianskaite-Kvaraciejiene, Apanavičiene e Gelžinis (2015), um RF é construído seguindo os passos listado a seguir:

1. Escolha do n, número de árvores que serão geradas, e q, o número de atributos que serão usadas em cada árvore, que deve ser menor ou igual ao número de variáveis do conjunto de dados;
2. Uma amostra do conjunto de dados total é selecionada com repetição, assim como um conjunto de q variáveis são aleatoriamente escolhidas;
3. A amostra é utilizada para desenvolver uma árvore de decisão;
4. Os passos 2 e 3 são repetidos até que se tenham n árvores.

Assim, uma observação k pertencente ao conjunto de dados é utilizada para o treinamento de A de árvores, não participando do treinamento de n – A árvores. k é, portanto, considerada como parte da amostra *out-of-bag* (OOB) das n – A árvores, amostra que será usada para avaliar o desempenho destas. Todas as n árvores possuem uma amostra OOB composta por todos as observações não utilizadas no processo de treinamento (BREIMAN, 2001).

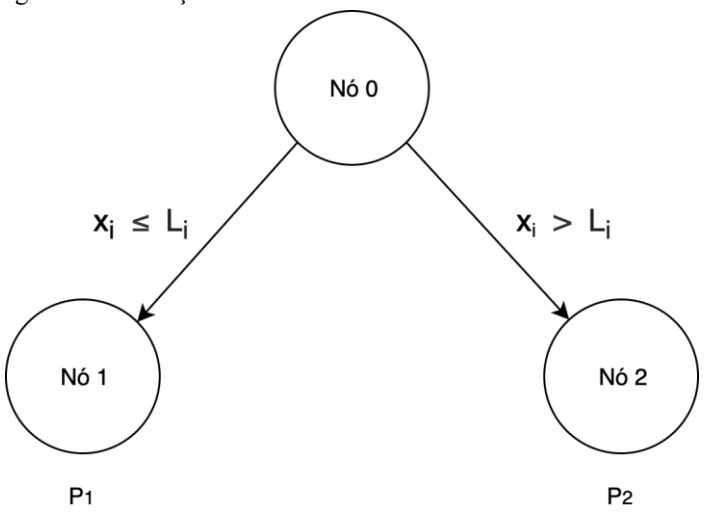
Os nós de uma árvore são escolhidos automaticamente pelo algoritmo após a avaliação de quais são os limiares que fornecem menor impureza de nó para as variáveis q. No caso de modelos de regressão, a medida de impureza do nó utilizada é variância dos erros de

estimativa, sendo o alvo do algoritmo reduzir a variância final em cada nó da árvore. As divisões na árvore vão progredindo até que se chegue a um tamanho mínimo de nó (um parâmetro de entrada fornecido ao algoritmo) ou até que a criação de um novo nó não resulte em resultados mais precisos (os nós do nível inferior possuem impureza maior que aqueles do nível superior).

A partir de um nó raiz (denominado "nó zero" na Figura 9), o algoritmo busca dentre as variáveis pertencentes ao conjunto q qual o limiar L_i que minimiza a variância ou soma dos erros quadrados (SEQ) nos nós seguintes. A escolha de L_i segrega o conjunto de dados, alocando parte no "Nó 1" e o restante no "Nó 2". Esse processo de otimização determina valores alvo P_1 e P_2 para os respectivos nós. Assim, caso uma observação i possua x_i inferior ao limiar L_i, o algoritmo estima seu preço com o valor P_1. Caso contrário, estima-se seu preço com o valor P_2. O algoritmo seleciona para a divisão no nó raiz a variável do conjunto q que promove a menor redução de variância ou SEQ. Após a escolha da variável para o então nó raiz, o processo repete-se nos nós criados até que se atinja algum critério de parada do algoritmo.

Antipov e Pokryshevskaya (2012) afirmam que apesar de ser um método consolidado na literatura para problemas de classificação, pouco se sabe sobre o uso do RF para problemas de regressão. Relatam também que o RF é indicado para problemas com alta presença de variáveis categóricas, característica típica dos problemas de avaliação em massa de imóveis, ou com base de dados incompletas, bem como lida bem com relações não-lineares e *outliers*. Nos resultados obtidos por Antipov e Pokryshevskaya (2012), o modelo RF e outros também estruturados com árvores de decisão demonstraram melhor desempenho quando comparados com modelos de regressão e *multilayer perceptron* (uma utilização específica de rede neural artificial).

Figura 9 – Ilustração dos nós de uma árvore de decisão.



FONTE: Elaboração própria.

Čeh *et al.* (2018) reportaram desempenho significativamente superior para o RF na estimativa do preço de venda de imóveis na capital da Eslovênia, Liubliana, em comparação ao modelo construído com regressão múltipla. O RF obteve métricas de erro inferiores à metade daquelas logradas com o MQO. A comparação foi feita pontualmente entre os resultados, não havendo aplicação de testes estatísticos que comprovem significância na diferença de desempenho.

2.6 Síntese da Revisão

O Quadro 5 reproduz compilação dos estudos revisados para a formulação desta pesquisa, com uma breve descrição do estudo e características do modelo. Nela estão citados apenas os estudos que envolveram modelagem.

Quadro 5 – Compilação da Revisão Bibliográfica.

Autores	**Descrição do Estudo**	**Características do Modelo**
Worzala, Lenk e Silva (1995)	Estudo comparativo de modelos de avaliação de imóveis.	Regressão múltipla e redes neurais artificiais.
Gader, Mohamed e Keller (1996)	Modelo de classificação construído para reconhecimento de escrita.	Métodos "fuzzy integrals" para combinação de resultados de modelos de classificação.
Janssen, Söderberg e	Definição de um modelo robusto para lidar com outliers a partir da combinação de	"*Reweighted Least Median*

(continuação)

Autores	Descrição do Estudo	Características do Modelo
Zhou (2001)	regressão robusta e matriz de ponderação espacial.	*Squares*"
Nguyen e Cripps (2001)	Comparação entre resultados de modelos de regressão.	Regressão múltipla e redes neurais artificiais.
Granitto, Verdes e Ceccatto (2005)	Estudo focado no processo de combinações de modelos para melhoria dos resultados, testando diferentes arquiteturas de redes para 10 conjuntos de dados distintos.	Combinações de redes neurais artificiais.
Arraes e Filho (2008)	Construção de modelos hedônicos segmentados por tipo de imóvel utilizando variáveis com características físicas, locacionais, econômicas e financeiras.	Modelo de regressão múltipla com proxies representando a localização.
García, Gámez e Alfaro (2008)	Criação de um sistema automatizado para avaliação de imóveis da cidade de Albacete, Espanha.	Redes neurais artificiais e Sistemas de Informação Geográfica.
Peterson e Flanagan (2009)	Comparação entre modelos de regressão para avaliação de imóveis.	Regressão múltipla e redes neurais artificiais.
He *et al.* (2010)	Análise de variáveis estruturais (preço do terreno e distância para a área central por exemplo) na formação de preços em Pequim.	Regressão múltipla com variáveis geográficas.
Kostov (2010)	Desenvolvimento de um método para definição de matriz de ponderação espacial.	Regressão múltipla e dependência espacial.
Kempa *et al.* (2011)	Utilização de combinação de métodos e comparação com métodos utilizados profissionalmente.	Redes neurais artificiais genéticas, sistemas fuzzy e "*bagging*".
Kontrimas e Verikas (2011)	Comparação entre modelos de regressão e combinação em comitês com práticas oficiais a partir de dados de imóveis da Lituânia.	Regressão múltipla, redes neurais artificiais, "SVM", comitês ponderados por "*SOM*".
Pontes, Paixão e Abramo (2011)	Estudo dos efeitos de criminalidade nos preços de imóveis.	Regressão múltipla.
Zurada, Levitan e	Comparação entre modelos de regressão para avaliação de imóveis.	Regressão múltipla e redes

(continuação)

Autores	Descrição do Estudo	Características do Modelo
Guan (2011)		neurais artificiais.
Antipov e Pokryshevskaya (2012)	Comparação de dez tipos de algoritmos para avaliação em massa de imóveis com dados de São Petersburgo, Rússia.	Árvores de decisão CHAID, *random forest*, redes neurais artificiais, "kNN", regressão múltipla.
Lasota *et al.* (2013)	Utilização de combinação de modelos de árvore de decisão com adição de ruídos ao modelo.	"*Random forest*", "*Random subspace*", "*Rotation forest*" e "*bagging*".
McCluskey *et al.* (2013)	Comparação entre modelos de regressão e modelos de regressão espacial.	Regressão múltipla, modelos de defasagem espacial, modelos autoregressivos, redes neurais artificiais.
Seya, Yamagata e Tsutsumi (2013)	Estudo sobre a especificação de matrizes de ponderação espacial.	Modelo espacial de Dubin e modelo de defasagem espacial.
Wen, Bu e Qin (2014)	Avaliação do efeito do ambiente lacustre no preço de imóveis.	Regressão múltipla.
Coral *et al.* (2015)	Construção de um comitê de modelos para previsão de desempenho de compressores.	Comitê de redes neurais artificiais.
Kettani e Oral (2015)	Desenvolvem uma metodologia para seleção de um subconjunto de casos para uma avaliação comparativa de imóveis.	Método foi denominado como "Regressão análoga".
Mabu, Obayashi e Kuremoto (2015)	Combinação de modelos para previsão de preços de ações.	Redes neurais artificiais e algoritmos genéticos.
Qu e Lee (2015)	Construção de modelos autoregressivos e de matrizes de ponderação espacial.	Modelos autoregressivos.
Zhang *et al.* (2015)	Especificação da matriz de ponderação espacial a partir de operações fuzzy.	Modelo espacial autoregressivo e regressão múltipla.
Anish, Majhi e Majhi (2016)	Combinação de modelos para previsão de valores de ativos.	Combinação de modelos com três tipos de estruturas

(continuação)

Autores	Descrição do Estudo	Características do Modelo
		adaptativas.
Nunes (2016)	Estudo de preço de imóveis para a cidade de Fortaleza, Ceará.	"*Stepwise ridge regression*".
Acharya *et al.* (2017)	Construção de um classificador de cinco tipos de batimentos cardíacos a partir de dados de exames de eletrocardiograma.	Rede neural artificial convolucional de 9 camadas
Campos e Almeida (2018)	Modelo construído com dados do município de São Paulo simulando efeitos de dependência espacial entre imóveis e entre distritos municipais.	Método hierárquico linear espacial.
Codes (2018)	Estudo dos imóveis da cidade de Fortaleza, Ceará, com análise segregada por bairros.	Redes neurais artificiais.
Fernandez, Mukherjee e Scott (2018)	Estudo dos efeitos de políticas de conservação ambiental no preço de imóveis.	Regressão múltipla.
Yeh, Hsu e Weight (2018)	Proposta de um método comparativo para avaliação de imóveis.	Abordagem comparativa quantitativa, regressão múltipla e redes neurais artificiais.

FONTE: Elaboração própria.

Dentre os estudos, nota-se predominância do uso de métodos de regressão linear múltipla e redes neurais artificiais na construção dos modelos. Na listagem, observa-se que estudos sobre os efeitos da localização, seja através dos modelos espaciais ou por meio de variáveis geográficas, não empregam modelos não-lineares, como as redes neurais artificiais ou *Random Forest*.

É perceptível que não há convergência entre o tema da combinação de modelos e do tema dos modelos de regressão espacial. Em estudos que modelos são combinados no intuito de se obter melhor precisão nas avaliações, modelos feitos sob métodos de regressão espacial não estão entre os modelos eleitos. Da mesma forma, em estudos cujo foco é a aplicação de métodos de regressão espacial, os três métodos de regressão espacial (defasagem espacial, autoregressivo e Durbin) não foram combinados num comitê, por exemplo.

Por fim, em nenhum dos estudos revisados foi empregado teste estatístico para atestar se as diferenças nas métricas de erros obtidas possuem significância estatística, limitando a comparação de métricas singulares para cada modelo ajustado.

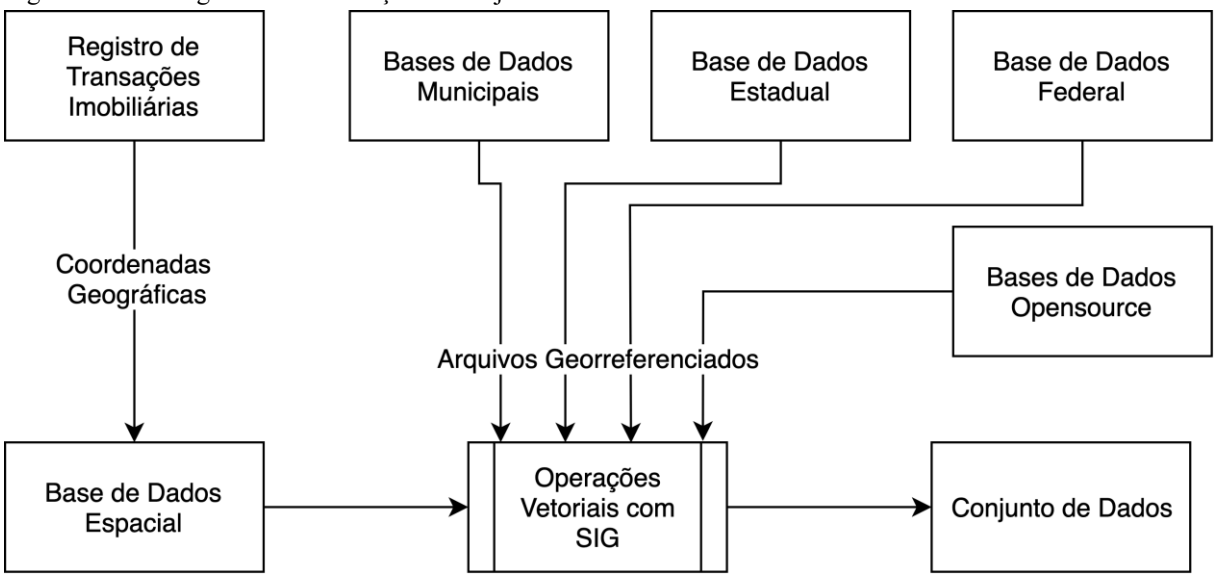
3 DADOS E MÉTODOS

A pesquisa iniciou por revisão bibliográfica a respeito de avaliação de imóveis, bem como sobre os fundamentos das abordagens de avaliação, os métodos quantitativos aplicados e as métricas utilizadas para caracterização do desempenho dos valores preditos pelos modelos construídos na literatura correspondente. A revisão também reforçou a estruturação do método de pesquisa, que consiste na construção do banco de dados, seleção e transformação de variáveis, análise exploratória, ajuste dos modelos, avaliação de desempenho e análise dos erros.

O desenvolvimento do conjunto de dados passou inicialmente pela coleta do registro de transações imobiliárias dos anos de 2009 a 2016 disponibilizado pela Secretaria Municipal das Finanças de Fortaleza (SEFIN) em formato de planilha eletrônica com extensão XLSX. Dentre as cerca de 145 mil transações, encontram-se imóveis de variados usos (residenciais, comerciais, destinados à hotelaria, industriais) e distintas classificações arquitetônicas (apartamentos, casas, lojas, galpões). No estudo sob relatório, foram filtradas apenas as transações decorrentes num período de cinco anos (de 2012 a 2016) e classificados como apartamentos. O horizonte de cinco anos foi escolhido pois esse é o período utilizado no processo de avaliação na SEFIN. A escolha pelos apartamentos se deu por ser essa a classificação arquitetônica com maior número de registros. Após essa filtragem, reduziu-se a amostra para 43.617 observações. Removendo registros incompletos restaram 39.181 transações. O conjunto resultante é também comparável aos maiores estudos encontrados na literatura, sendo a terceiro maior de acordo com o encontrado na revisão de literatura (Quadro 1, presente na “Introdução”).

Dentre os 29 atributos cadastrados para cada transação, dois são as coordenadas geográficas do imóvel na projeção Universal Transversa de Mercator (UTM), Sul, Zona 24. A disponibilidade das coordenadas permite a exportação desses dados para um pacote computacional SIG e a criação de uma base de dados espacial. Neste trabalho foram utilizados pacotes disponíveis para a linguagem de programação R: *sf* e *sp*. A construção do conjunto de dados seguiu a lógica do fluxograma da Figura 10.

Figura 10 – Fluxograma de construção do conjunto de dados.



FONTE: Elaboração própria.

Desde então, se sobrepõe a base espacial estabelecida com outras bases de dados georreferenciadas, o que agrega, por meio de operações vetoriais, outros atributos aos registros de transações de imóveis da base original. As bases de dados consultadas são mantidas por órgãos municipais, estaduais, federais e plataformas colaborativas *opensource*.

As bases municipais consultadas foram o Sistema de Informações Geográficas "Fortaleza em Mapas", mantido pelo Instituto de Planejamento de Fortaleza (Iplanfor), o Sistema de Informações de Acidentes de Trabalho de Fortaleza (SIAT-FOR), mantido pela Autarquia Municipal de Trânsito e Cidadania (AMC), e registro de pontos de ônibus da cidade de Fortaleza mantidos pela Empresa de Transporte Urbano de Fortaleza (ETUFOR). O SIG "Fortaleza em Mapas" disponibiliza informações como o posicionamento e traçado das linhas de metrô, base de logradouros e edificações da cidade, IDH dos bairros, localização de lagoas e lagos, praças e áreas verdes, zonas de assentamento precário, dentre outros, informações essas registradas em variados arquivos georreferenciados de extensão "SHP" (chamados de *Shapefiles*) utilizando o *Datum* WGS84.

No SIAT-FOR extraíram-se dados de acidentes de trânsito ocorridos anualmente na cidade de Fortaleza, dados que poderiam dar origem a uma variável *proxy* para identificar regiões com maior movimentação de pessoas. Foi utilizado o registro de acidentes do ano de 2017, também disponibilizado num arquivo georreferenciado de extensão "SHP", utilizando o *Datum* WGS84. Da base da ETUFOR foi obtida a localização de todos os pontos de ônibus na cidade de Fortaleza (referente ao ano de 2019), gerando uma variável que poderia medir a

conexão do imóvel com a malha viária da cidade. A localização dos pontos de ônibus é disponibilizada em arquivos georreferenciados de extensão "KML" (*Keyhole Markup Language*), na forma de latitudes e longitudes num formato decimal.

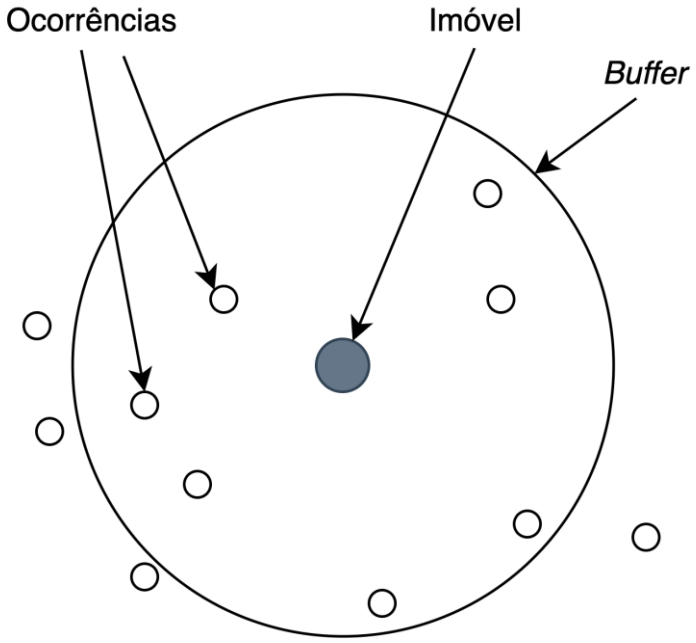
A base estadual consultada é mantida pela Secretaria da Segurança Pública e Defesa Social (SSPDS) e consiste num repositório de relatórios diários de ocorrências policiais registradas de acordo com o zoneamento feito pela Instituição. Foi adaptado um script de mineração de dados em Python[1] para extrair, tabular e agregar as informações dos relatórios (disponibilizados em formato de texto em PDF). Posteriormente, com base no endereço extraído, realiza-se um georreferenciamento (em inglês a técnica é chamada de *geocoding*) que atribui latitudes e longitudes decimais a cada uma das ocorrências.

Foram minerados dados disponíveis para os anos de 2014 a 2016, resultando em 16.987 ocorrências extraídas ao todo. Utilizando apenas os dados de 2016, foi criada uma variável *proxy* que mede a localização do imóvel quanto a segurança pública. A variável consiste numa operação vetorial envolvendo *buffers* e interseções após a superposição de ambas bases (a Figura 11 ilustra a operação). Para cada imóvel da base espacial, cria-se um buffer com um certo raio, delimitando uma região geográfica (para essa variável, especificamente, foi criada uma área com 500 m de raio). Em seguida, conta-se (por operação de interseção) quantos pontos representando ocorrências policiais estão dentro da zona de influência criada pelo *buffer* e armazena-se o resultado dessa contagem como uma variável na base espacial.

O mesmo procedimento foi utilizado, por exemplo, para a contagem de pontos de ônibus estão próximos aos imóveis (em particular, dentro de um raio de 300 m cujo centro é a coordenada registrada) ou o número de acidentes de trânsito (na mesma área de influência de raio de 300 m). Procedimento semelhante foi utilizado para determinar se o imóvel está localizado a uma certa distância de estações de metrô (foram utilizadas as distâncias de 300 e 500 m) ou de praças e áreas verdes (distância usada foi de 200 m), por exemplo. Nesses últimos casos, porém, cria-se um *buffer* nos pontos de interesse com as distâncias especificadas e avalia-se se cada um dos imóveis está localizado dentro das zonas de influência definidas para cada um dos pontos de interesse. As distâncias utilizadas foram distâncias euclidianas e as zonas de influências foram definidas com base na revisão bibliográfica (WANG et al., 2015; ČEH et al., 2018; UBERTI et al., 2018) e em análise crítica.

[1] Código adaptado de http://github.com/netodelino

Figura 11 – Operação com *buffers* e interseções.



FONTE: Elaboração própria.

A base federal consultada foi a Base Territorial Estatística de Áreas de Risco, mantida pelo Instituto Brasileiro de Geografia e Estatísticas (IBGE), que fornece as áreas do município classificadas com área de risco, definida como área passível de ser atingida por fenômeno ou processos naturais e/ou induzidos que causem efeito adverso (IBGE, 2018). Tal base, armazenada em arquivo georeferrenciado utilizando *Datum* SIRGAS2000, enseja a criação de variável *dummy* que caracteriza se um imóvel está localizado proximamente a uma área de risco.

Por fim, a fonte *opensource* consultada foi o *Open Street Map* (OSM), de onde foram coletadas localização de amenidades registradas na plataforma colaborativa: lojas, bares, restaurantes, lanchonetes, clínicas, igrejas, cafés, farmácias, postos de gasolina, supermercados, escolas, hospitais e bancos. Cada um desses tipos de amenidades possui uma *tag* específica demandada por meio de um API especialmente desenvolvido para consultas[2]. Do API extraem-se latitudes e longitudes em formato decimal das amenidades registradas na base do OSM. Ao todo, 3.925 amenidades cadastradas foram encontradas nas buscas realizadas e usadas para criar uma variável que indica a densidade de amenidades próximas ao imóvel (usando a operação vetorial descrita anteriormente, usando distância euclidiana de 500 metros).

Para realizar a integração e gerar as variáveis descritas, foi necessária uniformizar

[2] Link do API: http://overpass-turbo.eu

o sistema de referência de coordenadas. Assim, logo após o carregamento das bases com auxílio de pacotes *sf* e *sp* da linguagem de programação R, realizou-se a transformação do sistema de referências para o *Datum* WGS84, coordenadas UTM, Zona 24, Sul. Finalizada a integração das bases de variadas origens, criou-se uma variável que mede a distância (distância euclidiana) de cada um dos imóveis para a Avenida Beira Mar, localizada numa região de mais elevado valor imobiliário. No remate, obteve-se um conjunto de dados com 53 variáveis ao todo, descritas Quadro 6.

Quadro 6 – Definição das Variáveis.

Id	Variável	Descrição	Fonte
1	numero_pavimentos	Número de pavimentos do edifício onde localiza-se o imóvel	SEFIN
2	num_unidades_lote	Número de unidades no terreno onde localiza-se o imóvel	SEFIN
3	testada_principal	Medida da frente principal do terreno onde localiza-se o imóvel	SEFIN
4	fator_edificacao	Parâmetro relacionado com o aproveitamento do terreno pela edificação	SEFIN
5	fator_lote	Parâmetro relacionado com o aproveitamento do terreno pela edificação	SEFIN
6	area_terreno_gi	Área do terreno onde localiza-se o imóvel	SEFIN
7	fracao_ideal	Parcela do empreendimento (em termos de área) que corresponde ao imóvel – divisão entre área construída do imóvel e área construída total da edificação	SEFIN
8	area_edificada_gi	Área do imóvel	SEFIN
9	valor_base_calculo_itbi	Valor da transação	SEFIN
10	x	Coordenada geográfica na projeção UTM	SEFIN
11	y	Coordenada geográfica na projeção UTM	SEFIN
12	preco_area	Valor do imóvel por metro quadrado	Razão entre “9” e “8”
13	amenidades	Quantidade de estabelecimentos num raio de 500 m do imóvel	OSM
14	pontos_de_onibus	Quantidade de pontos de ônibus que localizam-se num raio de 300 m do imóvel	ETUFOR
15	OP	Número de ocorrências policiais registradas numa raio de 500 m do imóvel no período descrito	SSPDS
16	AT2017	Número de acidentes de trânsito que ocorreram no ano de 2017 num raio de 300 m	SIAT-FOR (AMC)

(continuação)

Id	Variável	Descrição	Fonte
17	prox_metro_300	Variável que indica se o imóvel está a menos de 300 m de uma estação de metro da cidade	Iplanfor
18	prox_metro_500	Variável que indica se o imóvel está a menos de 500 m de uma estação de metro da cidade	Iplanfor
19	prox_BATER	Variável que indica se o imóvel está a menos de 300 m de uma área de risco da cidade	IBGE
20	Assent_Precario	Variável que indica se o imóvel está a menos de 150 m de um assentamento precário	Iplanfor
21	parques_urbanos	Variável que indica se o imóvel está a menos de 200 m de um parque urbano da cidade	Iplanfor
22	Areas_Verdes	Variável que indica se o imóvel está a menos de 200 m de praças ou áreas verdes da cidade	Iplanfor
23	lagoas	Variável que indica se o imóvel está a menos de 150 m de uma lagoa da cidade	Iplanfor
24	dist_BM	Distância do imóvel para a Avenida Beira Mar	Calculado
25	IDH	Medida do Índice de Desenvolvimento Humano do bairro onde localiza-se o imóvel	Iplanfor
26	AREAP_RES	Percentual da área edificada do bairro onde localiza-se o imóvel que corresponde a área destinada a imóveis residenciais	Iplanfor
27	AREAP_COM	Percentual da área edificada do bairro onde localiza-se o imóvel que corresponde a área destinada a imóveis com fins comerciais	Iplanfor
28	AREAP_SER	Percentual da área edificada do bairro onde localiza-se o imóvel que corresponde a área destinada a imóveis destinados a prestação de serviços	Iplanfor
29	AREAP_IND	Percentual da área edificada do bairro onde localiza-se o imóvel que corresponde a área destinada a imóveis com fins industriais	Iplanfor
30	idade	Idade do imóvel (usado o ano de 2016 como base)	SEFIN
31	A2012	Variável que indica que a transação se deu no ano de 2012	SEFIN
32	A2013	Variável que indica que a transação se deu no ano de 2013	SEFIN
33	A2014	Variável que indica que a transação se deu no ano de 2014	SEFIN
34	A2015	Variável que indica que a transação se deu no ano de 2015	SEFIN
35	ALAMEDA	Variável que indica que o imóvel localiza-se numa alameda	SEFIN
36	AVENIDA	Variável que indica que o imóvel localiza-se numa avenida	SEFIN
37	ESTRADA	Variável que indica que o imóvel localiza-se numa estrada	SEFIN
38	RODOVIA	Variável que indica que o imóvel localiza-se numa rodovia	SEFIN

(continuação)

Id	Variável	Descrição	Fonte
39	TRAVESSA	Variável que indica que o imóvel localiza-se numa travessa	SEFIN
40	A1	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão elevado (primeiro nível)	SEFIN
41	A2	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão elevado (segundo nível)	SEFIN
42	A3	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão elevado (terceiro nível)	SEFIN
43	B1	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo popular (primeiro nível)	SEFIN
44	B2	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo popular (segundo nível)	SEFIN
45	B3	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo popular (terceiro nível)	SEFIN
46	L1	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo luxuoso (primeiro nível)	SEFIN
47	L2	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo luxuoso (segundo nível)	SEFIN
48	L3	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo luxuoso (terceiro nível)	SEFIN
49	N2	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo normal (segundo nível)	SEFIN
50	N3	Variável que indica padrão construtivo do imóvel, indicando uma classe de padrão construtivo normal (terceiro nível)	SEFIN
51	Sit_Esquina	Variável que indica que o imóvel está localizado numa esquina	SEFIN
52	Sit_Gleba	Variável que indica que o imóvel está localizado numa gleba	SEFIN
53	Sit_Quadra	Variável que indica que o imóvel está localizado numa quadra inteira	SEFIN

FONTE: Elaboração própria.

Alerta-se para o fato de que conjuntos de variáveis do tipo *dummy* não englobam todos os possíveis resultados, por exemplo, a inexistência de uma variável indicativa de que a transação foi realizada no ano de 2016. Essas variáveis foram excluídas para evitar que haja a

já mencionada *dummy trap*, quando uma variável é representada como uma combinação linear de outras. Assim, transações que foram realizadas em 2016 receberão o valor zero para as variáveis A2012, A2013, A2014 e A2015. O mesmo ocorre para um imóvel em situação normal, quando as variáveis de 51 a 53 recebem valor zero, ou um imóvel num primeiro nível de padrão construtivo normal, quando as variáveis 40 a 50 recebem valor zero, ou um imóvel localizado numa rua, quando as variáveis 35 a 39 recebem valor zero.

De posse do conjunto de dados, partiu-se para uma análise exploratória do conjunto de dados para melhor compreensão do fenômeno do extenso conjunto de dados resultante. Na análise, foram executadas as seguintes etapas não sequenciais: visualização e análise espacial dos dados, análise da influência das variáveis no preço dos imóveis e análise de agrupamento (*clustering*). A visualização e análise espacial dos dados permite uma maior compreensão do conjunto de dados: entender a variabilidade espacial do valor dos imóveis, detectar concentrações, *outliers* e zonas de valor. Foram feitas comparações entre os bairros e de bairros com o conjunto de dados totais também no intuito de avaliar variabilidade.

A análise da influência das variáveis no preço dos imóveis foi conduzida por meio da correlação e dos coeficientes obtidos por uma regressão linear. O objetivo dessa etapa é estudar quais variáveis mais influenciam no valor dos imóveis, permitindo uma maior compreensão do fenômeno. Esse entendimento guia etapas posteriores de análise da performance dos modelos e dos erros obtidos. Por fim, o *clustering* permite o agrupamento de imóveis semelhantes em grupos. A divisão em grupos mais uniformes pode aprimorar a precisão dos modelos e auxiliar na detecção de fatores que prejudicam o seu desempenho.

Em seguida, foi a vez da preparação do conjunto de dados para aplicações dos algoritmos de regressão. Seguindo prática corrente na literatura relacionada, os algoritmos serão “treinados” e “testados” com grupos diferentes de observações, sendo o primeiro chamado de conjunto de treino e o segundo de conjunto de teste. Além disso, para se verificar se há significância estatística nas diferenças entre métricas de erro obtidas para cada algoritmo, o conjunto de dados será particionado em 30 subconjuntos, conforme protocolo proposto por Granatyr (2017). Assim sendo, para cada algoritmo, será obtida amostra com 30 leituras para cada uma das métricas de erro, o que permitirá a aplicação do teste não paramétrico de Friedman com o teste posterior de Nemenyi, descrito na seção 2.3. As métricas de erros calculadas serão EPAM, REQM (raiz do erro quadrático médio, tradução do termo *Root Mean Squared Error*) e valores inaceitáveis, seguindo além do limiar de 20% recomendado por Kontrimas e Verikas (2011), também os limiares de 50% e 100%. O EPAM e o REQM são calculado de acordo com

as Equações (22) e (23).

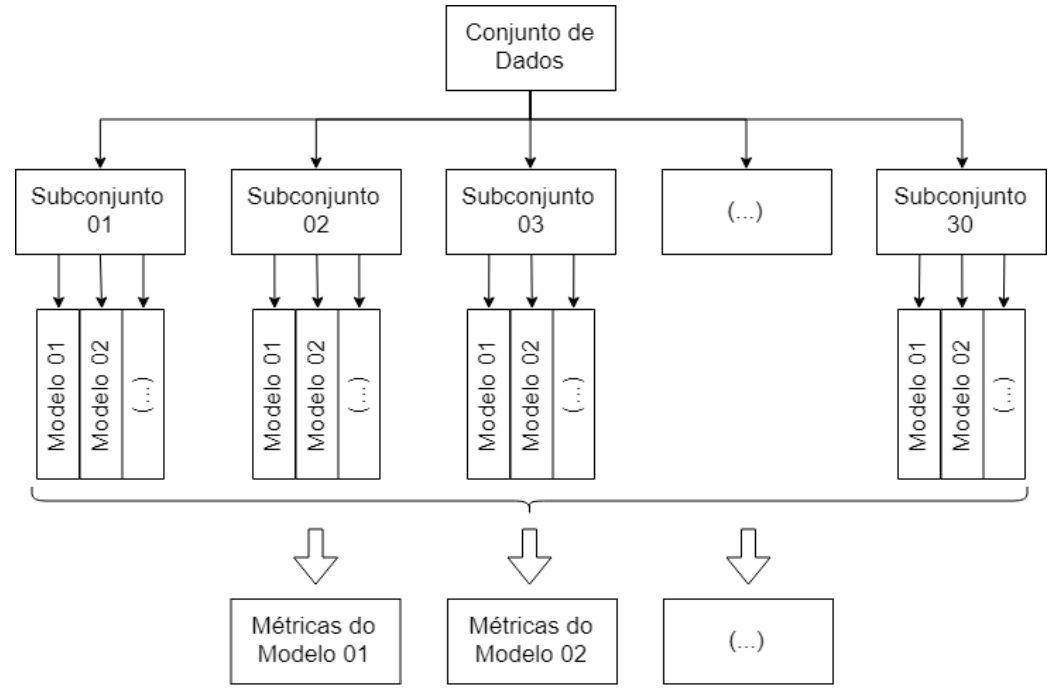
$$EPAM(\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i^{real} - y_i^{modelo}\right|}{y_i^{real}} \tag{22}$$

$$REQM = \sqrt{\frac{\sum_{i=1}^{n}\left|y_i^{real} - y_i^{modelo}\right|}{n}} \tag{23}$$

EPAM e REQM são métricas muito utilizados na literatura e representam medidas dos erros relativos e absolutos, respectivamente. A filtragem dos valores inaceitáveis em distintos limiares permite uma análise do perfil de erros do modelo, servindo como diagnóstico das causas de um EPAM elevado, por exemplo. Um EPAM resulta elevado em razão de erros absolutos em média elevados ou valores extremos de alta magnitude ou em alta frequência.

A Figura 12 ilustra o processo de preparações dos subconjuntos e geração das métricas de erro. Cada um dos 30 subconjuntos é composto por particionamentos do conjunto de dados total na fração de 80%/20% - 80% das observações são alocadas no conjunto de treino e 20% no conjunto de testes. Os mesmos subconjuntos são aplicados em cada um dos algorítimos de regressão, que, nesse estudo, serão *Random Forest* (RF), descrito na seção 2.5, *Spatial Autoregressive Model* (MAE), descrito na sessão 2.4, Redes Neurais Artificiais (RNA) e *Support Vector Machine* (SVM), ambos descritos na sessão 2.2. A regressão linear multivariada será utilizada como base (Modelo Base) de comparação para os algoritmos supracitados.

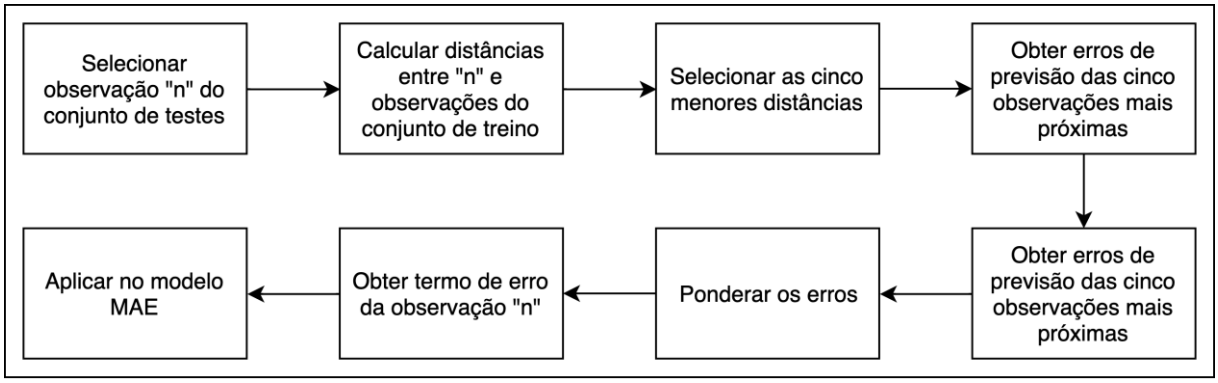
Figura 12 – Processo de Modelagem dos Dados



FONTE: Elaboração própria.

O modelo MAE foi construído com uma SWM obtida com o método kNN, usando os cinco vizinhos mais próximos de cada observação (k = 5). Com a matriz, os erros oriundos do modelo Base são ponderados e incorporados como variável independente no modelo MAE, que é gerado com os conjuntos de treino. Para as previsões, buscam-se as cinco observações do conjunto de treino mais próximas de cada um dos componentes do conjunto de testes, obtendo assim o termo de erro ponderado para ser usado no modelo MAE.

Figura 13 – Obtenção dos termos de erros ponderados do conjunto de testes.



FONTE: Elaboração própria.

Na obtenção dos cinco vizinhos mais próximos, por ser um conjunto de dados tão extenso, há a possibilidade de que haja pelo menos um imóvel com, por exemplo, mais de cinco

observações com distâncias iguais. No código desenvolvido para o cálculo dos valores de erros ponderados (descrito na Figura 13) foi incorporado um “critério de desempate” que pinça aleatoriamente dentre as observações com distâncias iguais.

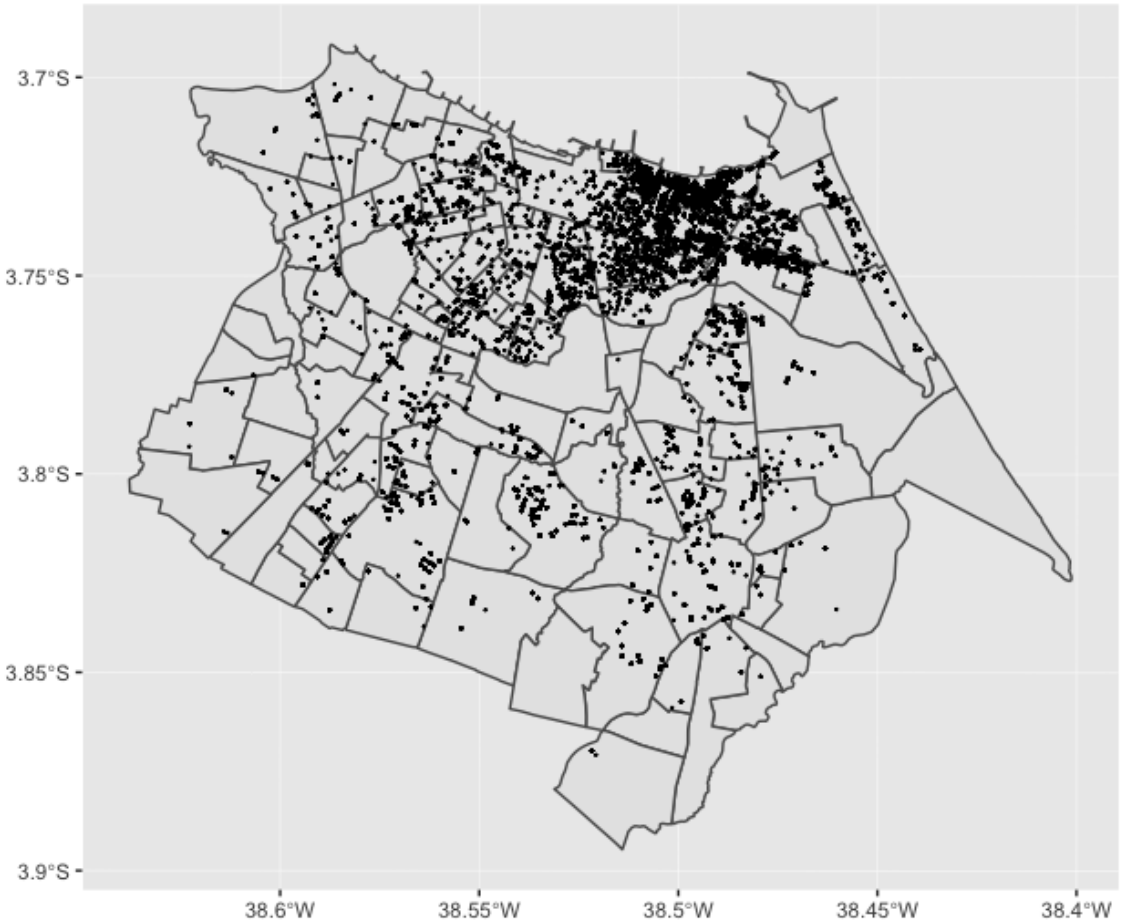
4 RESULTADOS

Neste capítulo, a análise exploratória dos dados é expressa, bem como os particionamentos da base de dados e dos resultados do modelo Base. Em seguida, são mostrados, respectivamente, os resultados dos modelos RNA, MAE, SVM e RF. É indicada, também, a arquitetura e configuração dos hiperparâmetros de cada algoritmo.

4.1 Análise Exploratória dos Dados

Na Figura 14, é notória a grande concentração de transações nos bairros Meireles, Aldeota, Varjota e Mucuripe, que englobam 8.777 transações (22,40% de todo o conjunto de dados), com os dois primeiros englobando 7.648 transações (19,52%). Outros bairros que também concentram grande número de transações são Cocó (5,88%), Messejana (4,89%), Bairro de Fátima (3,92%), Passaré (3,74%), Mondubim (3,67%), Parque Iracema (2,97%) e Cambeba (2,61%).

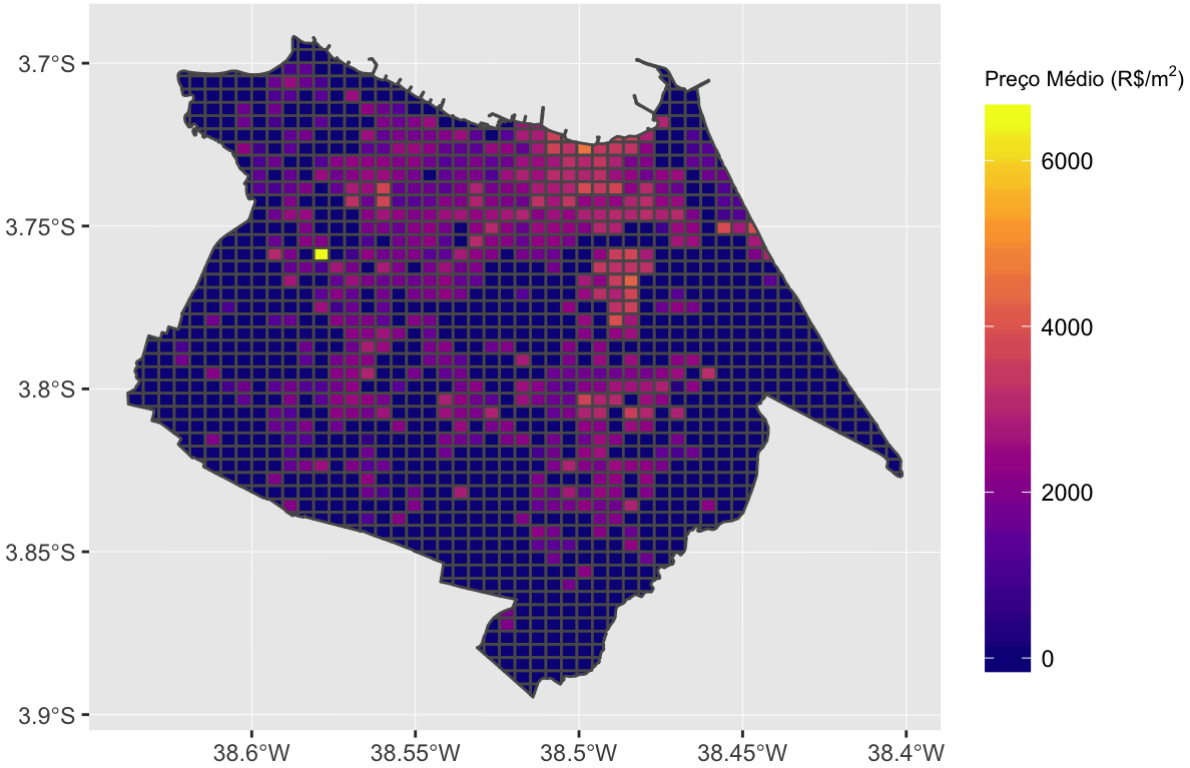
Figura 14 – Distribuição espacial das transações do conjunto de dados.



FONTE: Elaboração própria.

A Figura 15 mostra a variação geográfica do preço por área dos imóveis na cidade. Observa-se que os imóveis de maior valor se concentram na região norte da cidade, na região da orla, onde situa-se a Avenida Beira Mar. Os valores vão, assim, reduzindo no vetor norte-sul, havendo na região adjacente à orla e na região leste-sudeste concentrações de imóveis de alto valor. O mapa evidencia a existência de valores médios elevados nos bairros Meireles, Aldeota, Mucuripe, Varjota, Guararapes, Luciano Cavalcante e Bairro de Lourdes. Na região oeste da cidade existe também uma concentração de imóveis de maior valor (ainda que menor que nas outras regiões), onde se nota um ponto de elevado valor no bairro Jóquei Clube que discrepa do observado na região – o que indica um *outlier*.

Figura 15 – Mapa com distribuição de imóveis, por valor por metro quadrado.



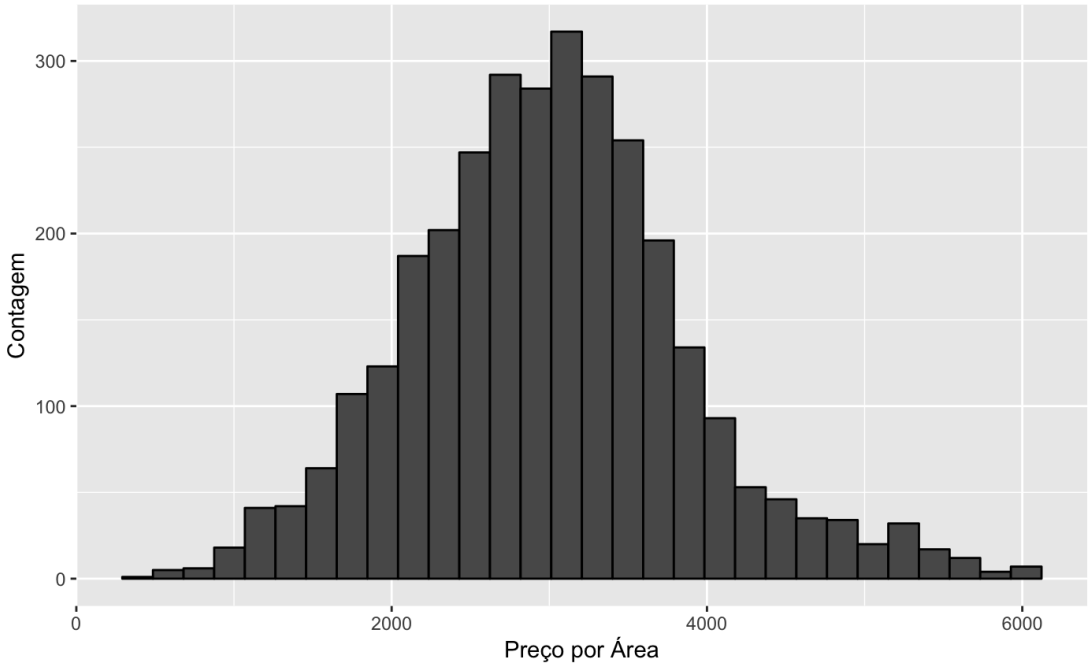
FONTE: Elaboração própria.

Ao analisar os dados agregados em bairros, percebe-se, entretanto, que há grande variabilidade intrarregional. Por exemplo, os bairros Aldeota e Meireles indicaram valor médio de R$ 3.000 e R$ 3.535 e desvio-padrão R$ 881 e R$ 1.348 respectivamente – coeficientes de variação (razão entre desvio-padrão e média) de cerca de 29% e 38% respectivamente. No bairro Aldeota, foram registradas transações entre R$ 464 e R$ 6.100 por metro quadrado. No bairro Meireles, o intervalo possui amplitude ainda maior, entre R$ 69 e R$ 30.526. Na Figura 16 é apresentado um histograma com os valores de transações do bairro Aldeota, onde ilustra-

se a grande amplitude de valores. Já na Figura 17 vê-se histograma semelhante para o bairro de Lourdes (Dunas), onde há maior uniformidade de valores. Na primeira, detecta-se um padrão semelhante com a curva gaussiana, enquanto na segunda se observa um comportamento praticamente aleatório dentro do intervalo.

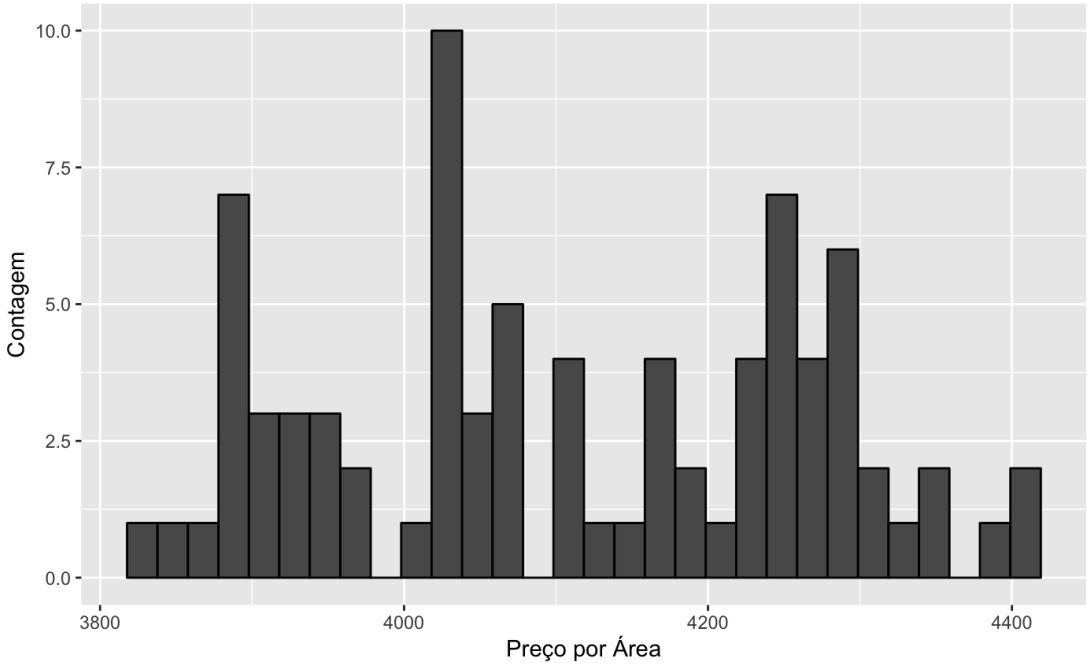
As maiores variabilidades intra-regionais foram detectadas nos bairros Amadeu Furtado (63%), Vicente Pinzon (45%), Pan Americano (41%) e Henrique Jorge (38%). 73% dos bairros apresentaram coeficiente de variação superior a 20%. Ainda que não apresente delimitações dos bairros, pela Figura 15, notadamente entre os paralelos 3,70ºS – 3,75ºS e 38,6ºW – 38,5ºW, nota-se a proximidade entre células de elevador valor com células de baixo valor o que denota variação espacial abrupta. O Quadro 7 retrata os dez bairros que registraram transações com maiores valores médios por área, mostrando, juntamente a esses valores, o desvio-padrão medido.

Figura 16 – Histograma com valores transacionais do bairro Aldeota.



FONTE: Elaboração própria.

Figura 17 – Histograma com valores transacionais do bairro de Lourdes.



FONTE: Elaboração própria.

Quadro 7 – Bairros com maiores valores médios de transações registradas.

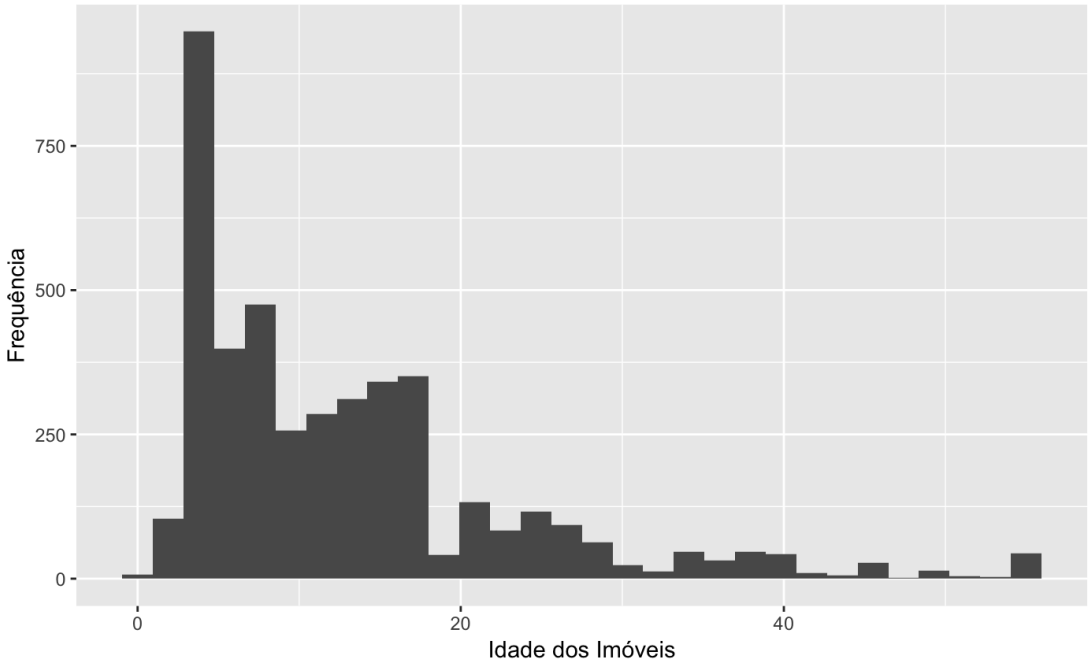
Bairro	Registros	Valor Médio (R$)	Desvio-Padrão (R$)
Bairro de Lourdes (Dunas)	82	4.113	158
Meireles	4.484	3.535	1.348
Guararapes	711	3.445	793
Mucuripe	867	3.322	1.005
Eng. Luciano Cavalcante	628	3.106	817
Cais do Porto (Serviluz)	93	3.067	828
Autran Nunes	1	3.063	-
Aldeota	3.164	3.000	881
Cocó	2.303	2.996	811
Parquelândia	177	2.981	908

FONTE: Elaboração própria.

O Quadro 7 mostra que dentre os dez bairros de maiores preços médios de transação apenas o Bairro de Lourdes exibiu baixo desvio-padrão, com um coeficiente de variação (CV) de 3,84%. O bairro possui o sétimo menor desvio-padrão encontrado dentre todos os bairros da cidade e o terceiro menor CV. Parte do desvio-padrão pode ser atribuído a uma elevada variabilidade inter-regional de atributos – por exemplo, a correlação entre o desvio-padrão dos preços de transação e o desvio-padrão das áreas edificadas e da idade dos imóveis em cada bairro é de 0,590 e 0,404 respectivamente. Tais valores de correlação indicam que a variabilidade das variáveis idade e área edificada (representadas pelo desvio-padrão) em cada

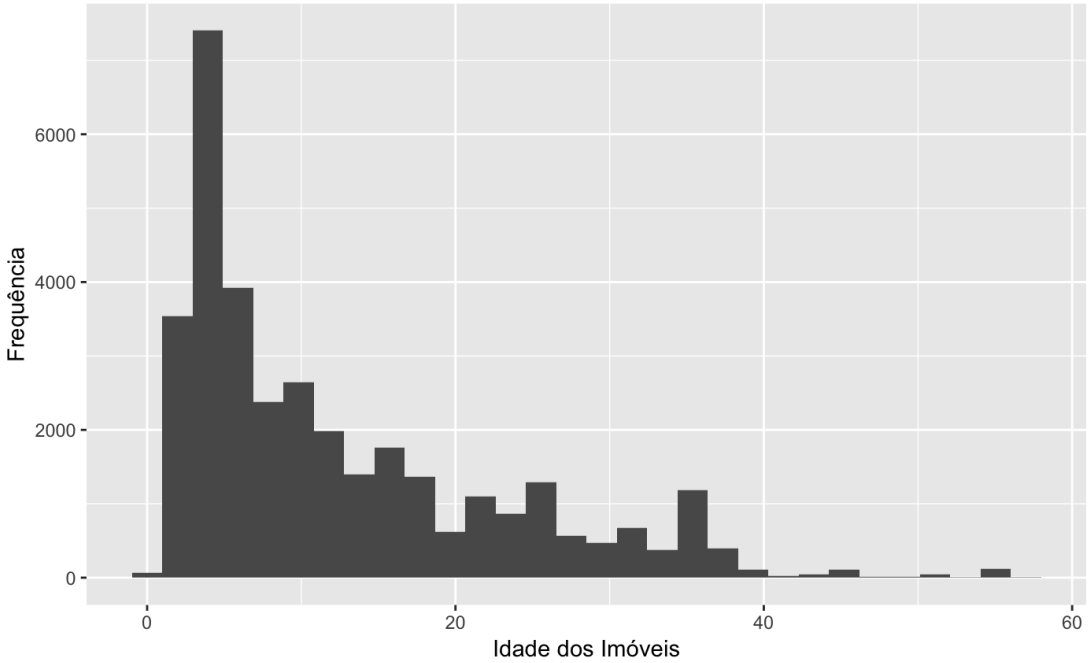
bairro possuem influência positiva na variabilidade dos preços também em cada bairro. No bairro Meireles, onde se observa elevado desvio-padrão nos valores de transação, observa-se elevado desvio-padrão no atributo idade – o valor encontrado é de 10,55 anos, bastante elevado quando comparado com a idade média do bairro igual a 12,56 anos. Na Figura 18 é possível visualizar a variabilidade para o atributo idade dos imóveis transacionados nesse bairro. Apesar de uma maior concentração de imóveis com idade entre 0 (imóveis novos, recém construídos) e 20 anos, 7% possuem idade superior a 30 anos, sendo 44 com idade registrada igual a 55 anos. Pouco mais da metade das transações registra imóveis com idade inferior a 10 anos. A Figura 19 revela que o perfil de idade dos imóveis no bairro Meireles não destoa tanto do observado na base completa.

Figura 18 – Histograma com idade dos imóveis no bairro Meireles.



FONTE: Elaboração própria.

Figura 19 – Histograma com idade de todos os registros da base.



FONTE: Elaboração própria.

A correlação das variáveis dependentes com a transformação logarítmica natural do valor do imóvel por metro quadrado (denominada desde então como variável dependente) é notada no Quadro 8. Observa-se que as variáveis “idade” e “area_edificada_gi” possuem importante correlação com a variável dependente. A matriz de correlação completa (onde é possível observar não só a correlação entre variáveis independentes e variável dependente mas também a correlação entre as variáveis independentes) é apresentada no Anexo A.

Quadro 8 – Correlação com variável dependente.

Variável	Correlação	Variável	Correlação
numero_pavimentos	0,559	AT2017	0,239
num_unidades_lote	0,058	prox_metro_300	-0,049
testada_principal	-0,007	prox_metro_500	-0,079
fator_edificacao	0,227	prox_BATER	-0,124
fator_lote	0,009	Assent_Precario	-0,175
area_terreno_gi	-0,046	parques_urbanos	-0,051
fracao_ideal	-0,158	Areas_Verdes	0,095
area_edificada_gi	0,331	lagoas	-0,037
X	0,299	dist_BM	-0,404
Y	0,290	IDH	0,431
amenidades	0,097	AREAP_RES	0,392
pontos_de_onibus	-0,138	AREAP_COM	0,257
OP	-0,142	AREAP_SER	0,244
AREAP_IND	-0,132	A2	-0,292
Idade	-0,447	A3	0,079
A2012	-0,300	B1	-0,001
A2013	-0,085	L1	0,186
A2014	0,085	L2	0,345
A2015	0,205	L3	0,116
ALAMEDA	-0,031	N2	-0,011
AVENIDA	0,051	N3	-0,036
ESTRADA	-0,069	Sit_Esquina	0,050
RODOVIA	-0,015	Sit_Gleba	0,036
TRAVESSA	-0,005	Sit_Quadra	0,029
A1	-0,325		

FONTE: Elaboração própria.

O Quadro 9 contém o resultado de uma regressão multivariada entre as variáveis do banco de dados com a variável dependente referida. Nele estão as transformações aplicadas às variáveis independentes “dist_BM”, “AT2017”, “OP”, “x”, “y”, “area_edificada_gi”, “area_terreno_gi”. As transformações foram aplicadas com o intento de normalizar as diferenças na escala das variáveis com a da variável dependente. O valor do “valor p” da regressão, usado para indicar se há significância estatística no valor encontrado para o coeficiente (utilizado o limiar de significância de 5%), também é reportado. Em particular, as

variáveis “x” e “y”, por terem magnitudes bastante superiores (são da ordem de dezenas de milhares e de milhões, respectivamente), foram padronizadas (transformadas para que a média e desvio-padrão possuam valores zero e 1 respectivamente). As outras foram transformadas com a função logarítmica. A regressão apresentou R^2-ajustado de 0,6152. A equação da regressão linear está na Equação (24).

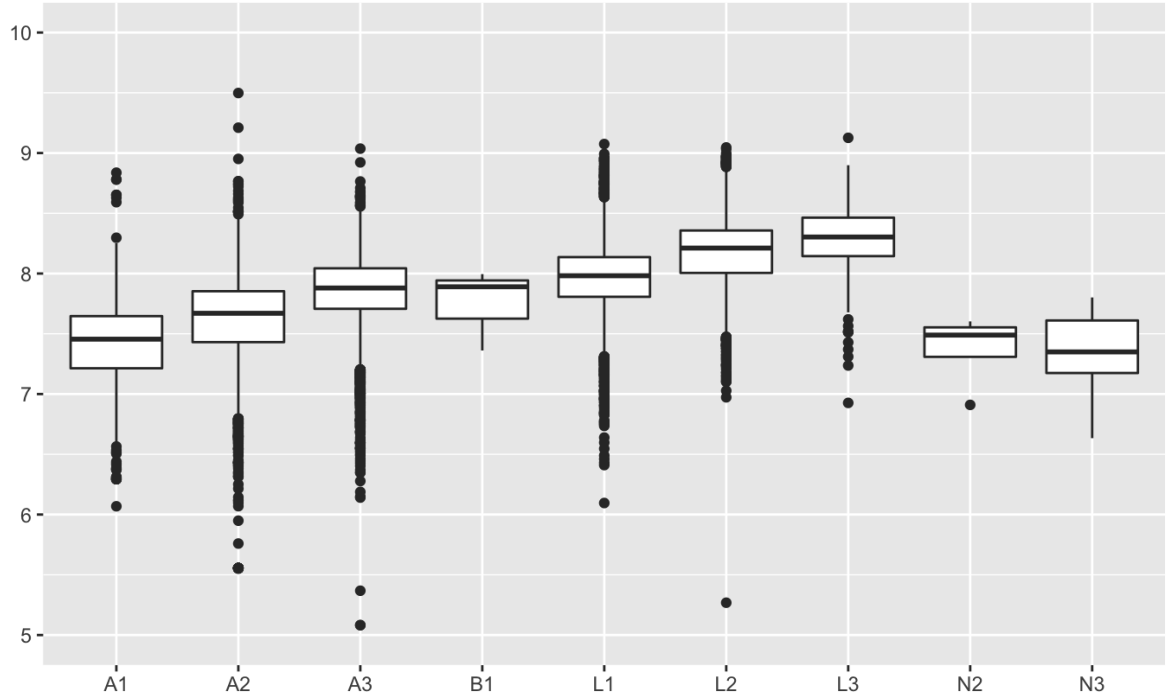
$$
\begin{aligned}
\log(\text{preco\_area}) = {} & 7.88 - 0.07 \times \log(1 + \text{OP}) - 0.05 \times (\text{prox\_metro\_300}) + 0.02 \times log(1 + AT2017) \\
& + 0.01 \times (\text{prox\_BATER}) - 0.04 \times log(1 + \text{pontos\_de\_onibus}) + 16.28 \times (\text{dist\_BM}^{-1}) + 0.04 \times (\text{x\_sc}) \\
& + 0.07 \times (\text{y\_sc}) - 0.34 \times (\text{A2012}) - 0.2 \times (\text{A2013}) - 0.1 \times (\text{A2014}) \\
& - 0.03 \times (\text{A2015}) - 0.01 \times (\text{idade}) + 0.04 \times (\text{Areas\_Verdes}) - 0.09 \times log(\text{area\_edificada\_gi}) \\
& + 0.003 \times (\text{AVENIDA}) + 0.09 \times (\text{RODOVIA}) + 0.12 \times (\text{TRAVESSA}) - 1.07 \times (\text{A1}) \\
& - 1.07 \times (\text{A2}) - 1.09 \times (\text{A3}) - 1.14 \times (\text{L1}) - 1.15 \times (\text{L2}) - 1.03 \times (\text{L3}) - 1.04 \times (\text{N3}) + 0.09 \times (\text{B1}) \\
& - 1.04 \times (\text{N2}) + 0.34 \times \log(1 + \text{IDH}) + 0.07 \times \log(1 + \text{numero\_pavimentos}) \\
& + 0.02 \times (\text{Sit\_Esquina}) - 0.03 \times (\text{Sit\_Gleba}) - 0.04 \times (\text{lagoas}) + 0.03 \times \log(1 + \text{amenidades})) \\
& + 0.07 \times (\text{parques\_urbanos}) + 0.04 \times (\text{Sit\_Quadra}) - 0.11 \times (\text{ESTRADA}) - 0.88 \times (\text{ALAMEDA}) \\
& + 0 \times (\text{testada\_principal}) + 0.09 \times (\text{fator\_lote}) + 0.22 \times (\text{fracao\_ideal}) \\
& - 0.04 \times \log(1 + \text{num\_unidades\_lote}) - 0.06 \times (\text{Assent\_Precario}) + 0.08 \times \log(\text{area\_terreno\_gi}) \\
& + 0.03 \times (\text{AREAP\_RES}) - 0.33 \times (\text{AREAP\_COM}) + 1.5 \times (\text{AREAP\_SER}) - 0.56 \times (\text{AREAP\_IND}) \\
& + 0.84 \times (\text{fator\_edificacao}) + \epsilon
\end{aligned}
\tag{24}
$$

Com arrimo nesses resultados, verifica-se que não há significância estatística que suporte a influência da proximidade a uma área de risco no valor de um imóvel, tendo em vista que o coeficiente da variável “prox_BATER” obteve valor p de 0,091. Do mesmo jeito, não há indícios de que a localização numa avenida (variável “AVENIDA”, com p valor de 0,424) ou numa rodovia (variável “RODOVIA”, com p valor de 0,390) influencie no valor do imóvel. Nesse aspecto, sendo o padrão de comparação utilizado a localização numa rua, um imóvel numa rua teria o mesmo valor de um com as mesmas características, porém localizado numa avenida ou numa rodovia.

A última variável independente cujo coeficiente não demonstrou significância estatística foi a “B1”, que faz parte de uma escala qualitativa definida pela fonte dos dados que visa a qualificar o padrão construtivo dos imóveis. Foram definidas quatro classes: padrão popular (B), padrão normal (N), padrão elevado (A) e padrão luxuoso (L), sendo cada classe dividida em três níveis. “B1” pertence ao primeiro nível do padrão popular, enquanto as variáveis “B2” e “B3” pertencem ao segundo e terceiro níveis, respectivamente (e que não estão na regressão pois não existem observações registrados com esse padrão construtivo). A Figura 20 mostra um *boxplot* com a variável dependente em função do padrão construtivo, onde é notório o fato de que, apesar de haver uma escala crescente em termos de mediana, o elevado desvio-padrão dentro de cada classe faz com que várias observações fiquem de fora dos limites

inferiores e superiores da caixa.

Figura 20 – *Boxplot* com variável dependente para cada classe de padrão construtivo.



FONTE: Elaboração própria.

Tais variáveis visam a quantificar o diferencial de valor em comparação ao primeiro nível da classe normal, a variável “N1”, que, como mencionado anteriormente, foi removida para evitar problemas do tipo *dummy trap*. Assim, a variável “B1” representa o quanto, em média, imóveis dessa classe são mais ou menos valiosos do que os imóveis da classe “N1”. O valor p encontrado de 0,189 indica que não há significância estatística encontrada para o coeficiente encontrado de 0,186. Ou seja, segundo os dados, imóveis da classe “B1” possuem valores maiores que aqueles da classe “N1” (já que o coeficiente é positivo), porém não evidência suficiente para julgar essa diferença como algo que não seria causado por aleatoriedade. Notavelmente, apenas 3 observações foram classificadas como “B1”, quantidade não representativa em comparação ao conjunto de dados total.

Um fator que colabora para os coeficientes não revelarem essa ordem entre as classes é possibilidade de que a inserção de outras variáveis na análise faça com que o maior valor de imóveis de padrão construtivo de luxo esteja sendo “capturado” por outra variável independente ou por uma combinação de variáveis independentes. Outras possibilidades são o elevado desvio-padrão observado dentro de cada classe ou a baixa importância de cada uma das variáveis em explicar a variável dependente.

Quadro 9 – Regressão multivariada com todo o conjunto de dados.

Variável	Coef.	Valor p	Variável	Coef.	Valor p
Intercepto	7.880	0.000	**L3**	-1.029	0.000
log(1 + OP)	-0.071	0.000	**N3**	-1.044	0.000
prox_metro_300	-0.045	0.000	**B1**	0.088	0.533
log(1 + AT2017)	0.016	0.000	**N2**	-1.042	0.000
prox_BATER	0.006	0.268	**Sit_Esquina**	0.018	0.000
log(1 + pontos_de_onibus)	-0.042	0.000	**Sit_Gleba**	-0.026	0.000
I(1/dist_BM)	16.285	0.000	**lagoas**	-0.039	0.000
x_sc	0.044	0.000	**log(1 + amenidades)**	0.029	0.000
y_sc	0.065	0.000	**parques_urbanos**	0.072	0.000
A2012	-0.341	0.000	**Sit_Quadra**	0.037	0.000
A2013	-0.203	0.000	**ESTRADA**	-0.114	0.000
A2014	-0.098	0.000	**ALAMEDA**	-0.880	0.000
A2015	-0.034	0.000	**testada_principal**	0.000	0.000
Idade	-0.012	0.000	**fator_lote**	0.089	0.002
Areas_Verdes	0.041	0.000	**fracao_ideal**	0.223	0.000
log(area_edificada_gi)	-0.091	0.000	**log(1 + numero_pavimentos)**	0.066	0.000
AVENIDA	0.003	0.352	**log(1 + num_unidades_lote)**	-0.036	0.000
RODOVIA	0.094	0.263	**Assent_Precario**	-0.058	0.000
TRAVESSA	0.119	0.000	**log(area_terreno_gi)**	0.080	0.000
A1	-1.068	0.000	**log(1 + IDH)**	0.340	0.000
A2	-1.068	0.000	**AREAP_RES**	0.034	0.000
A3	-1.091	0.000	**AREAP_COM**	-0.332	0.000
L1	-1.138	0.000	**AREAP_SER**	1.505	0.000
L2	-1.151	0.000	**AREAP_IND**	-0.562	0.005
fator_edificacao	0.836	0.000			

FONTE: Elaboração própria.

Os resultados apontam para uma relação inversa entre a a variável que mede o número de ocorrências policiais próximas ao imóvel (“OP”) e o preço do imóvel, apontando para um efeito depreciativo relacionado à insegurança na região. Relação inversa também foi detectada para a proximidade a estações de metrô (“prox_metro_300”), para a quantidade de pontos de ônibus próximos ao imóvel (“pontos_de_onibus”), para a proximidade de assentamentos precários (“Assent_Precario”), para a proximidade a lagos e lagoas (“lagoas”) e para a localização numa gleba (“Sit_Gleba”). Já proximidade a áreas verdes e praças (“Areas_Verdes”), parques urbanos (“parques_urbanos”) e a contagem de amenidades próximas ao imóvel (“amenidades”) influem positivamente no preço do imóvel.

Destaca-se que por ser utilizada uma transformação logarítmica na variável dependente, a percepção da magnitude dos coeficientes não é automática. A contribuição do coeficiente de uma variável depende do total agregado pelas outras variáveis. Tomando como exemplo a proximidade a parques urbanos (cujo coeficiente é da ordem de 0,07), seu efeito num imóvel de R$ 3.000,00/m^2 seria de cerca de R$ 200,00/m^2, enquanto num imóvel de R$ 5.000,00/m^2 seria de cerca de R$ 360,00/m^2.

É perceptível uma valorização dos imóveis no curso dos anos no período em foco. Sendo o ano de 2016 o padrão, a variável A2015 receber valor negativo indica que os valores transacionados em 2015 possuem valores em média menores que os valores transacionados em 2016. Os coeficientes para os anos de 2014 a 2012 são sucessivamente menores, demonstrando que há essa progressiva valorização à extensão desses anos. Assim, a transação de um imóvel avaliado em R$ 3.000,00/m^2 no ano de 2016, caso fosse transacionado em 2015 teria o valor de cerca de R$ 2.900,00/m^2, teria valor de R$ 2.700,00/m^2 caso fosse transacionado em 2014, teria valor de R$ 2.430,00/m^2 caso fosse transacionado em 2013 e valor de R$ 2.139,00/m^2 caso fosse transacionado em 2012.

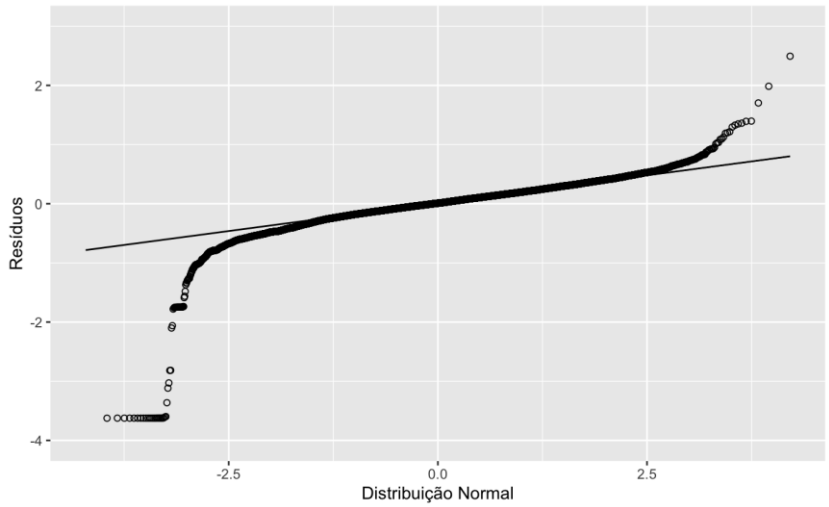
Os coeficientes das variáveis das coordenadas X e Y são positivos, o que indica maiores valores nos imóveis localizados no norte e no leste da cidade (mesma direção do sistema de coordenadas). Da mesma maneira, o inverso da distância do imóvel para a Avenida Beira Mar (medido em metros) possui coeficiente positivo, denotando que uma maior distância reduz o valor da variável dependente e diminui o efeito aplicado pelo coeficiente. A variável de acidentes de trânsito (“AT2017”), que foi empregada como “*proxy*” para caracterizar regiões de elevado fluxo de pessoas, obteve coeficiente positivo. Aqui cabe ressaltar que o relacionamento expresso por esse coeficiente não é que maiores taxas de acidentes de trânsito ocorrendo próximos aos imóveis exerce influência positiva no preço, mas que essas regiões onde existe um maior número de acidentes de trânsito são áreas (segundo premissa assumida) de maior fluxo de pessoas, fator que de fato teria influência positiva nos preços.

Ainda com objetivos de caracterizar as regiões onde se localizam os imóveis, os coeficientes de “AREAP_RES”, “AREAP_COM”, “AREAP_SER” e “AREAP_IND”, variáveis que caracterizam cada bairro da cidade quanto a densidade de imóveis residenciais, comerciais, de prestação de serviços e industriais, respectivamente, revelam que regiões com maior predominância residencial e de serviços influem positivamente no valor do imóvel. Com efeito oposto, atividade comercial e industrial influem negativamente no valor. Observa-se que o IDH do bairro onde se localiza o imóvel também influencia positivamente no valor do imóvel.

A variável de área edificada (“area_edificada_gi”), incorporada à regressão como *proxy* para outros importantes atributos de um imóvel, como vagas de garagem – afinal, conforme observado por Nunes (2016), ambas são fortemente correlacionadas – resultou num coeficiente negativo. Isso foi observado apesar da correlação entre área e o valor do imóvel ser positiva (0,3314). De maneira oposta, a área do terreno exibiu um coeficiente positivo. Na literatura revisada, encontra-se incidência de coeficientes negativos em Kontrimas e Verikas (2011), Campos e Almeida (2018) e Uberti et al. (2018). Nos estudos citados utilizou-se como variável dependente o preço unitário (preço transacional dividido pela área do imóvel). Nguyen e Cripps (2001) detectam coeficiente negativo para componente não linear envolvendo área (área do imóvel elevada ao quadrado). Dentre os estudos revisados, porém, são mais numerosos aqueles que apontam coeficiente positivo para a área do imóvel (ARRAES; FILHO, 2008; ZURADA; LEVITAN; GUAN, 2011; MCCLUSKEY et al., 2013; GLAESENER; CARUSO, 2015; WANG et al., 2015; NUNES, 2016).

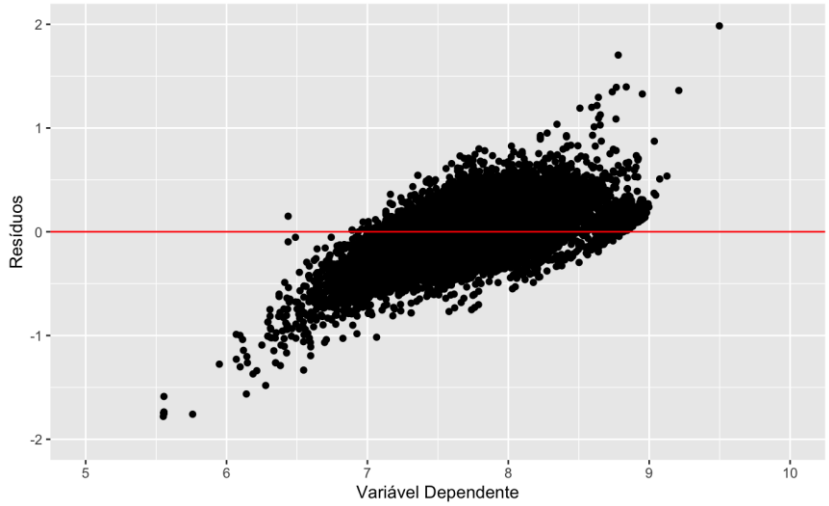
Um ponto relevante é que os resíduos obtidos pela regressão demonstram um padrão incompatível com a premissa de normalidade dos desvios para a regressão linear. Conforme se nota na Figura 21, nos extremos da escala há um forte desvio do comportamento esperado para uma distribuição normal, com discrepância ainda maior no extremo inferior. Igualmente, observa-se que os resíduos não parecem constantes ao longo de todo o intervalo de valores previstos e que a regressão fornece erros de previsão substanciais também para os extremos da escala, ainda permitindo a suspeita de que há violação da premissa de heterocedasticidade (Figura 22). Confirmada a hipótese de heterocedasticidade, os resultados dos testes de significância das variáveis independentes não são confiáveis em consequência de erro na estimativa dos erros-padrão de cada coeficiente.

Figura 21 – Checagem da normalidade dos resíduos.



FONTE: Elaboração própria.

Figura 22 – Resíduos dos respectivos valores da variável dependente.

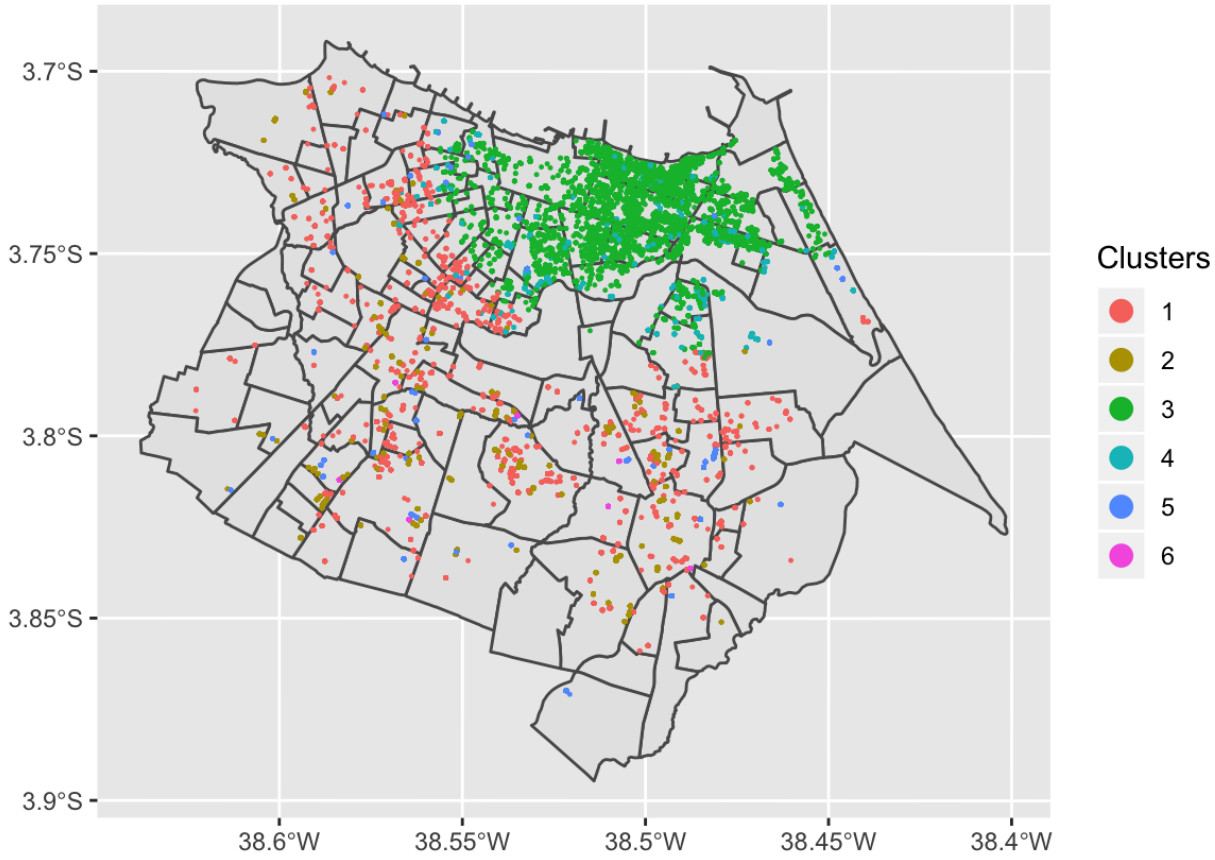


FONTE: Elaboração própria.

Utilizando a técnica *k-Means*, os dados foram separados em k *clusters* de acordo com a proximidade entre eles – nesta análise, apenas as coordenadas geográficas foram utilizadas para cálculo das distâncias. Assim, caso seja fornecido como parâmetro de entrada k = 3, três grupos são gerados aleatoriamente com os dados da amostra. Em seguida, uma observação é escolhida e calcula-se a distância euclidiana desta para os três centróides. Caso ela no início do processo ela tenha sido alocada no *cluster* de número 3 e observa-se que a menor distância é realmente para o centroide do *cluster* 2, tal observação deixa de pertencer ao *cluster* 3 e passa a integrar o *cluster* 2. Como houve nova reconfiguração dos *clusters*, os novos centroides são calculados. Esse processo é então repetido até que não seja possível a realocação de nenhuma observação. Neste estudo, foi empregado k = 6, número de clusters ótimo para o

conjunto de dados que foi encontrado desde o algoritmo proposto por Tibshirani, Walther e Hastie (2001). A divisão das observações em seis *clusters* está na Figura 23.

Figura 23 – Resultado da análise de agrupamento (*K Means*).



FONTE: Elaboração própria.

Quadro 10 – Descrição dos *clusters* obtidos (valores descritivos médios e desvio-padrão).

Cluster	Total	Bairros (10 principais)	Valor	Área	IDH	Idade
1	6.044 (15,4%)	Passaré, Mondubim, Maraponga, Messejana, Parque Iracema, Itaperi, Damas, Montese, Eng. Luciano Cavalcante, Presidente Kennedy.	R$ 2.127 (R$ 655)	77 m^2 (34 m^2)	0,3612 (0,1215)	9 anos (9,2)
2	6.954 (17,7%)	Messejana, Passaré, Mondubim, Parque Iracema, Itaperi, Jangurussu, Jóquei Clube, Parangaba, Cidade dos Funcionários, Maraponga.	R$ 2.197 (R$ 633)	77 m^2 (33 m^2)	0,3684 (0,1128)	6 anos (8,8)
3	15.947 (40,7%)	Meireles, Aldeota, Cocó, Centro, Dionísio Torres, Mucuripe, Fátima, Papicu, Praia de Iracema, Joaquim Távora.	R$ 2.796 (R$ 969)	150 m^2 (92 m^2)	0,7931 (0,1744)	11 anos (11,5)
4	5.410 (13,8%)	Meireles, Cocó, Jacarecanga, Praia do Futuro II, Fátima, Joaquim Távora, São Gerardo, Guararapes, Eng. Luciano Cavalcante, Edson Queiroz.	R$ 3.126 (R$ 1.200)	128 m^2 (93 m^2)	0,6625 (0,2362)	4 anos (9,9)
5	4.348 (11,1%)	Cambeba, Fátima, Messejana, Papicu, São Gerardo, Maraponga, Parque Iracema, Cristo Redentor, Jangurussu, Mondubim	R$ 2.621 (R$ 690)	83 m^2 (32 m^2)	0,5050 (0,1567)	4 anos (8,8)
6	479 (1,2%)	Messejana, Mondubim, Cajazeiras, Itaperi, Manoel Sátiro, Parangaba, Barroso, Novo Mondubim, Meireles	R$ 1.652 (R$ 411)	79 m^2 (30 m^2)	0,3757 (0,0688)	21 anos (12,6)
Base	**39.182 (100%)**	Fortaleza	R$ 2.542 (R$ 964)	108 m^2 (83 m^2)	0,5224 (0,2506)	7 anos (10,6)

FONTE: Elaboração própria.

O Quadro 10 contém uma descrição dos dez principais bairros que compõem cada um dos seis *clusters*. Os bairros são listados na ordem decrescente quanto ao total de imóveis em cada bairro (por exemplo, no cluster de número 1, o bairro que possui o maior número de observações é o bairro Passaré, seguido de Mondubim, Maraponga e assim por diante). Como pode ser observado, as fronteiras dos clusters não são delimitadas por fronteiras dos bairros, tendo em vista que imóveis do bairro Messejana estão inseridos em ambos *clusters* 1 e 2.

Ainda de acordo com o Quadro 10, os *clusters* 1 e 2 são bastante semelhantes, em valor, área e IDH medianos, com diferença sensível na idade dos imóveis. Numa busca mais detalhada, percebe-se que o *cluster* 1 também engloba bairros com maior atividade industrial (cerca de 55% maior) e imóveis localizados em terrenos de maior área em comparação ao *cluster* 2 (área de terreno mediana de cerca de 10 mil m^2 e 2,8 mil m^2 respectivamente). Os *clusters* 3 e 4 englobam imóveis de maior valor e área, com importante diferença na idade dos

imóveis - *cluster* 4 possui imóveis mais novos. De igual modo, a mediana das áreas de terreno no *cluster* 4 é bem superior às áreas do *cluster* 3 (respectivamente, 7,2 mil m^2 e 1,8 mil m^2). O *cluster* 5 representa o terceiro em termos de valor, com bairros de médio IDH, porém com apartamentos menores e mais novos. Já o *cluster* 6, por sua vez, representa os imóveis de menor valor, em bairros de menor IDH e mais antigos.

O Quadro 11 revela que a razão entre a mediana e o desvio-padrão (uma métrica semelhante ao coeficiente de variação que revela a magnitude da variabilidade) para as variáveis de preço, área, IDH e idade nos *clusters* é, em geral, menor que a razão agregada da base de dados. O *cluster* 4 experimentou a menor redução relativa na razão em foco para as variáveis do Quadro 11, porém, na variável de área do terreno houve uma importante redução de variabilidade (de cerca de 300% existente na base integral para 24,65%) e reduções sensíveis no número de pavimentos. O *clusters* 3 e 4 são os mais concentrados em termos geográficos, assim, haveria também uma menor variabilidade quando considerados os efeitos de autocorrelação espacial.

Quadro 11 – Razão entre mediana e desvio-padrão nos *clusters*.

Cluster	Valor	Área	IDH	Idade
1	30,79%	44,16%	33,64%	102,22%
2	28,81%	42,86%	30,62%	146,67%
3	34,66%	61,33%	21,99%	104,55%
4	**38,39%**	72,66%	35,65%	**247,50%**
5	26,33%	38,55%	31,03%	**220,00%**
6	24,88%	37,97%	18,31%	60,00%
Base	**37,92%**	**76,85%**	**47,97%**	**151,43%**

FONTE: Elaboração própria.

4.2 Particionamento e Estimativa com Regressão Linear (Modelo Base)

O particionamento do conjunto de dados em 30 subconjuntos foi realizado com auxílio do pacote "*caret*" para R. O pacote permite o arranjo de 80% do total das observações (com reposição) em 30 grupos diferentes. O Quadro 12 agrupa as métricas de erro obtidas com regressão linear (mínimo, máximo, média, desvio-padrão e coeficiente de variação). Além das métricas EPAM e REQM, são apresentados também o percentual de erros absolutos superiores a 20%, 50% e 100% e o R^2 ajustado.

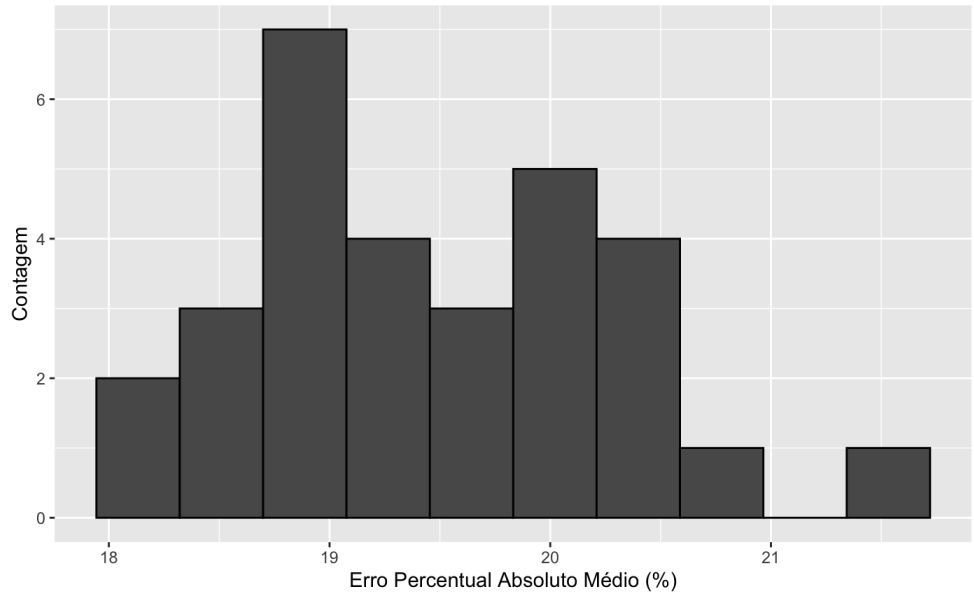
Quadro 12 – Métricas do modelo de regressão linear.

Métrica	Mínimo	Máximo	Média	DP	CV (%)
R^2 ajustado	0,6172	0,6319	0,6227	0,0032	0,5
EPAM (%)	18,11	21,51	19,45	0,15	0,4
REQM (R$)	522	627	548	29	5,0
Erro > 20% (%)	27,9	30,1	28,9	0,47	1,8
Erro > 50% (%)	3,9	4,7	4,3	0,19	4,4
Erro > 100% (%)	0,4	0,8	0,6	0,08	13,9

FONTE: Elaboração própria.

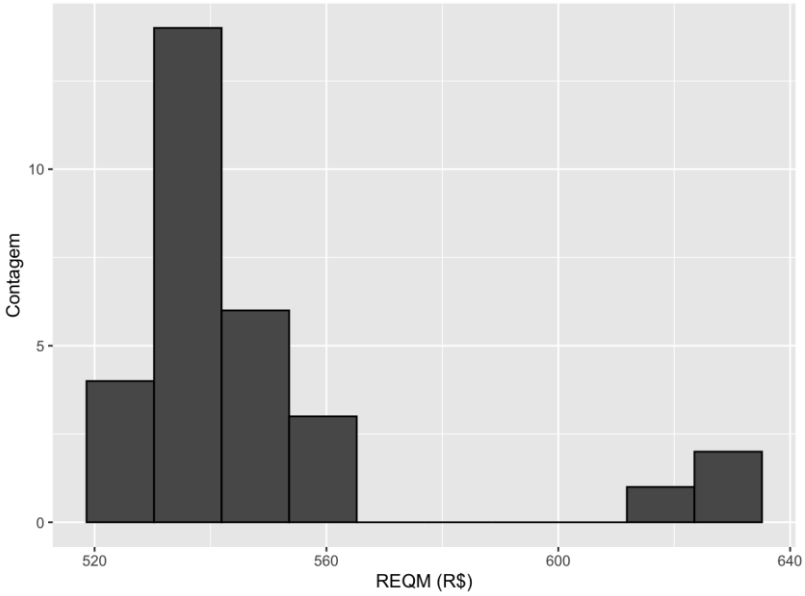
A Figura 24 e Figura 25 mostram as métricas de erro “EPAM” e “REQM” do modelo Base de regressão linear. Nelas vê-se que há uma descontinuidade nos perfis de erro, o que sugere importante sensibilidade do modelo à composição do conjunto de testes (um valor extremo em um dos subconjuntos traz impacto à média dos erros). Na Figura 26, em que os resultados do conjunto de testes da primeira partição gerada com os dados e os valores previstos pelo modelo para esse conjunto, é possível perceber uma concentração de erros superiores a 20% em imóveis de valores próximos a R$ 1.250,00 /m^2 (quando o modelo subestimou os valores reais), bem como ocorrências importantes de erros em imóveis de valores superiores a R$ 5.000,00/m^2 (quando o modelo superestimou os valores dos imóveis).

Figura 24 – Erros absolutos percentuais médios da regressão linear.



FONTE: Elaboração própria.

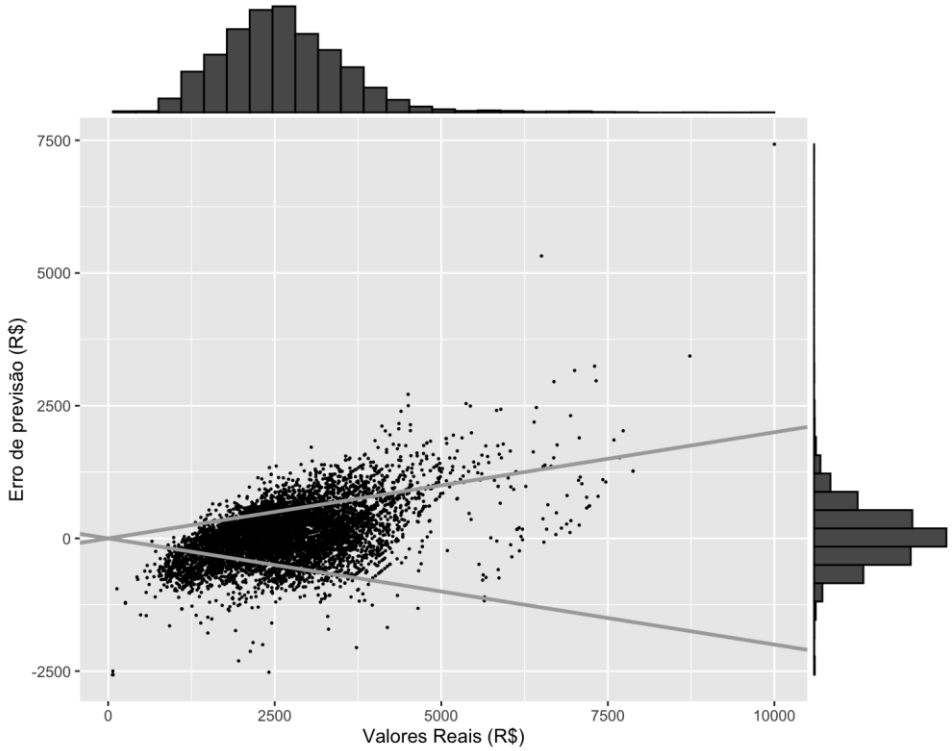
Figura 25 – Raiz do erro quadrado médio da regressão linear.



FONTE: Elaboração própria.

Caso sejam filtrados apenas os imóveis com valores inferiores a R$ 4.000/m^2, o EPAM obtido é de 21,8%, enquanto o EPAM para imóveis com valores superiores a R$ 4.000/m^2 é de 17,6%. Tal disparidade denota que não há desempenho uniforme do modelo ao longo de todo o intervalo de valores da base. Ressalta-se que a Figura 26 apresenta as diferenças não em erros percentuais, mas com valores positivos no caso em que resultado previsto é inferior ao valor real e negativo caso contrário. Nesta, as retas representam os limiares de erro de 20%, seja positivo ou negativo. Percebe-se, portanto, que apesar de serem maiores em magnitude os erros obtidos nos imóveis de valores superiores a R$ 4.000/m^2 não são em termos relativos superiores aos que se observam para os imóveis de valores inferiores. Dessa forma, observa-se que uma parte significativa do percentual de erros superiores a 20, 50 e 100% incide em imóveis de menor valor.

Figura 26 – Erros do modelo Base na primeira partição dos dados.



FONTE: Elaboração própria.

Com EPAM variando entre 18 e 21,5%, o desempenho do modelo de regressão foi semelhante ao encontrado nos estudos de Peterson e Flanagan (2009) e Antipov e Pokryshevskaya (2012). Com a regressão, o máximo percentual de erros relativos absolutos superiores a 20% foi de cerca de 30%, comparável às 31 observações dentre 100 consideradas como inaceitáveis por Kontrimas e Verikas (2011). Nguyen e Cripps (2001) reportam entre 21 e 31% de erros absolutos percentuais superiores a 15% para os modelos de regressão testados, comparáveis também ao desempenho obtido pelo modelo base.

4.3 Redes Neurais Artificiais (RNA)

O modelo em RNA foi desenvolvido com uma arquitetura rede composta por uma camada de entrada com 48 nós (composta pelas 48 variáveis independentes de entrada utilizadas no modelo: “numero_pavimentos”, “num_unidades_lote”, “testada_principal”, “fator_edificacao”, “fator_lote”, “area_terreno_gi”, “fracao_ideal”, “area_edificada_gi”, “preco_area”, “amenidades”, “pontos_de_onibus”, “OP”, “AT2017”, “prox_metro_300”, “prox_BATER”, “Assent_Precario”, “parques_urbanos”, “Areas_Verdes”, “lagoas”, “dist_BM”, “IDH”, “AREAP_RES”, “AREAP_COM”, “AREAP_SER”, “AREAP_IND”, “idade”, “A2012", “A2013”, “A2014”, “A2015”, “ALAMEDA”, “AVENIDA”, “ESTRADA”, “RODOVIA”, “TRAVESSA”, “A1”, “A2”, “A3”, “B1”, “L1”, “L2”, “L3”, “N2”, “N3”,

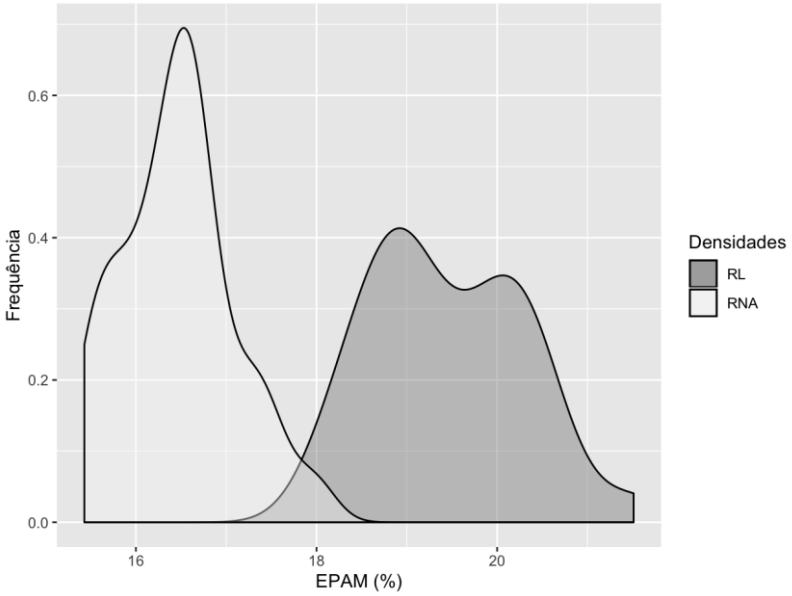
“Sit_Esquina”, “Sit_Gleba”, “Sit_Quadra”, “x” e “y”), três camadas intermediárias cada uma composta por 25 nós e função de ativação sigmóide e uma camada de saída com um nó com função de ativação linear. Essa arquitetura foi escolhida após uma série de testes com variadas configurações. Utilizando o processo de validação cruzada, em que, no caso, 20% do conjunto de treino é dividido para evitar risco de *overfitting*, e 100 iterações para cada partição de dados, os resultados obtidos estão no Quadro 13.

Quadro 13 – Resultados do modelo RNA.

Métrica	Mínimo	Máximo	Média	DP	CV (%)
EPAM (%)	15,43	17,98	16,46	0,63	3,8
REQM (R$)	487	609	528	31	6,0
Erro > 20% (%)	22,6	26,3	24,3	0,93	0,4
Erro > 50% (%)	2,0	4,1	2,8	0,39	13,9
Erro > 100% (%)	0,3	0,8	0,5	0,12	22,2

FONTE: Elaboração própria.

Figura 27 – Comparativo do EPAM entre Base e RNA.

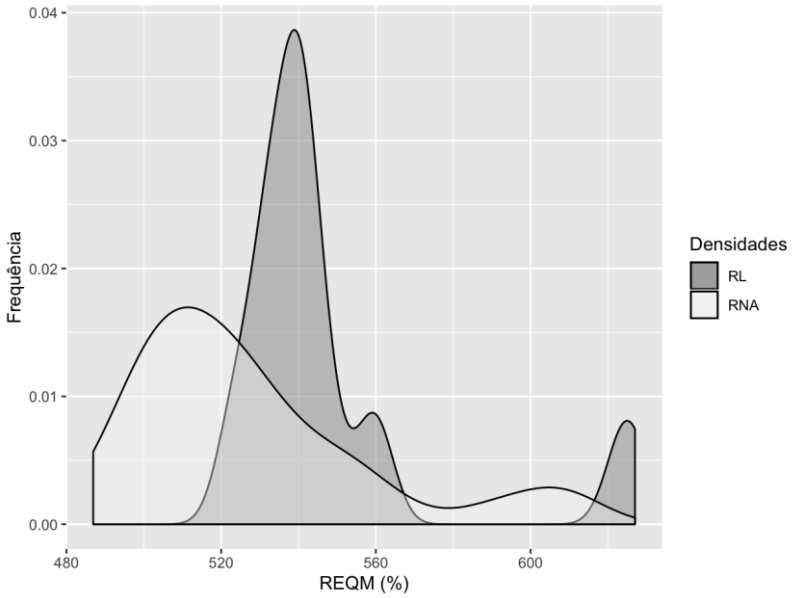


FONTE: Elaboração própria.

No modelo RNA, o EPAM varia entre 15,43% e 17,98%, enquanto no modelo Base varia entre 18,11% e 21,51%, havendo uma ligeira superposição entre os intervalos (A Figura 27 apresenta os gráficos de densidade desta métrica e revela tal superposição). Apesar dos resultados relativos sensivelmente menores para o modelo RNA, para a métrica REQM, que é absoluta, há uma maior superposição entre os intervalos – entre R$ 487 e R$ 609 para a RNA e entre R$ 522 e R$ 627 para a Base (tal superposição dos REQM pode ser observada na Figura

28). É possível perceber também que a RNA tende a apresentar menores observações com erros superiores aos limiares de 20%, 50% e 100% que o modelo Base. Percebe-se que ambos possuem perfis semelhantes de erros absolutos, porém os picos ocorridos no modelo RNA são menores que os obtidos no modelo Base, fator que explica parte dos menores erros relativos obtidos com a RNA.

Figura 28 – Comparativo do REQM entre Base e RNA.

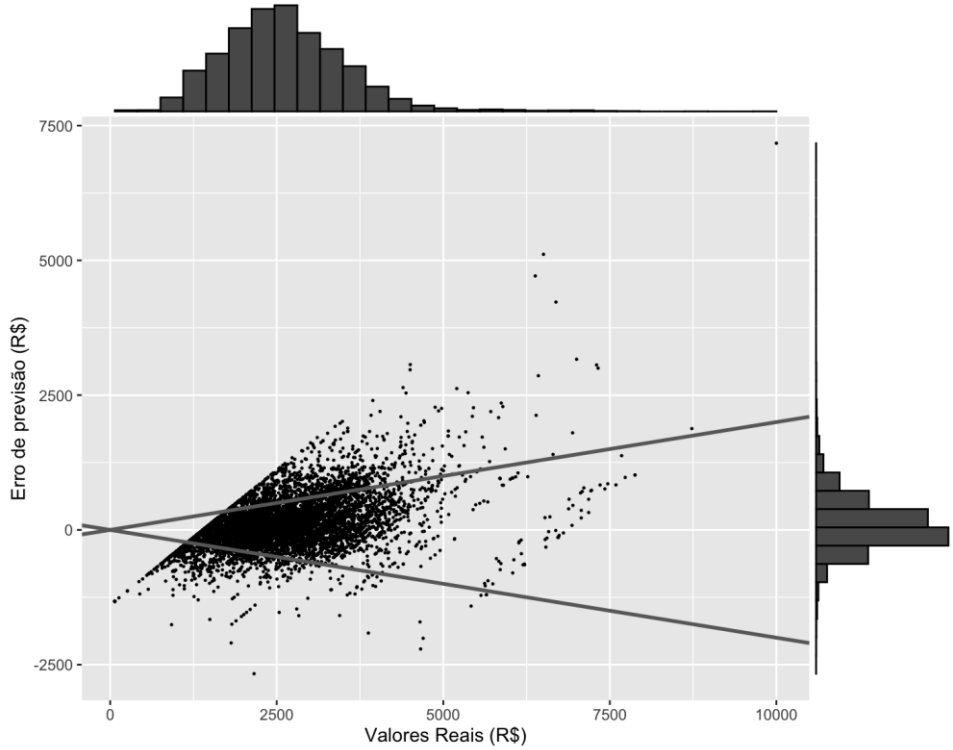


FONTE: Elaboração própria.

Comparando os erros da Figura 29 com aqueles da Figura 26 é possível perceber que no modelo RNA há uma menor concentração de erros que ultrapassam o limiar de 20% dentre os imóveis de valor inferior a R$ 2.500,00/m^2, porém ainda se percebe uma concentração de erros positivos (modelo superestimando valores) em imóveis de valor superior a R$ 5.000,00/m^2. No presente modelo persiste um menor EPAM para imóveis de valor superior a R$ 4.000,00/m^2 (no caso, 16%) quando comparado com imóveis de valor inferior (17%), porém com diferença significativamente menor, proximidade tal que permite suspeitar que no modelo não há diferença de desempenho ao longo de todo o intervalo de valores.

Ainda comparando a as duas figuras, comprova-se que o histograma dos erros (localizado na extremidade inferior direita) do modelo RNA denota uma assimetria para valores positivos quando comparado com os erros do modelo Base.

Figura 29 – Erros da primeira partição dos dados com modelo RNA.



FONTE: Elaboração própria.

4.4 Modelo Autorregressivo Espacial (MAE)

Os resultados obtidos com o modelo MAE estão no Quadro 14 e demonstram que o MAE é significativamente mais preciso que o RNA e o Base. A Figura 30 e a Figura 31, que apresentam gráficos de densidade comparando os EPAM e REQM do MAE com o do modelo Base, evidenciam que há grande ganho de precisão com o uso do MAE. Não há qualquer superposição dos intervalos tanto para erros relativos como para erros absolutos. O MAE retornou erros relativos com uma concentração em torno de um pico próximo da média, num padrão totalmente diferente daquele observado com a RNA e o modelo Base. Já no erro absoluto, observa-se um perfil significativamente menor, porém com um traçado semelhante ao observado no modelo Base.

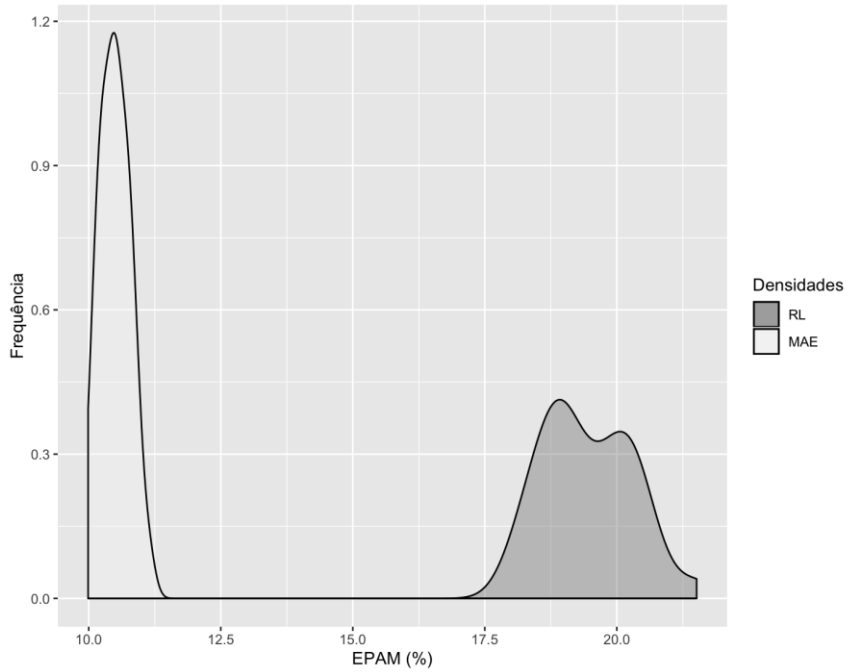
Quadro 14 – Resultados do modelo MAE.

Métrica	Mínimo	Máximo	Média	DP	CV (%)
EPAM (%)	9,98	11,12	10,49	0,28	2,7
REQM (R$)	343	484	371	31	10,4
Erro > 20% (%)	10,8	12,2	11,4	0,38	0,03
Erro > 50% (%)	1,1	1,6	1,4	0,12	8,6
Erro > 100% (%)	0,2	0,4	0,3	0,06	19,3

FONTE: Elaboração própria.

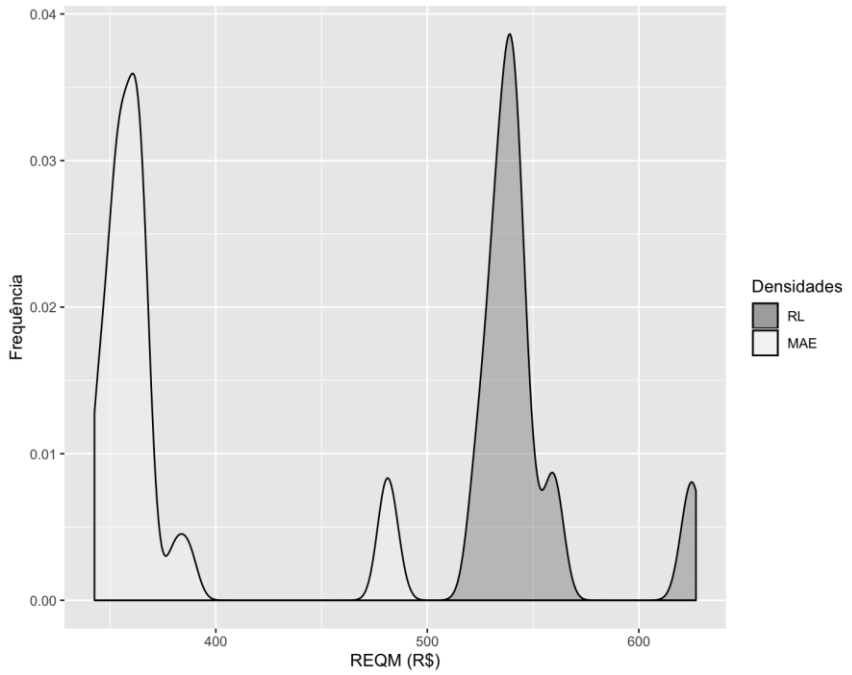
Na Figura 32 é possível ver os erros obtidos com a primeira partição de dados ao longo de todo o intervalo de valores. Comparando com a Figura 26, percebe-se que há uma concentração significativamente menor de observações além dos limiares positivos e negativos de 20% de erro, havendo, porém, ainda um acúmulo dentre imóveis de valores inferiores a R$ 2.500,00/m^2. Possível visualizar no gráfico a redução absoluta dos erros do modelo reduzindo a escala da ordenada. Com o MAE, há indícios fortes de estabilidade do modelo, com EPAM muito próximos para valores inferiores e superiores a R$ 4.000,00/m^2 (respectivamente, 10,3% e 10,0%).

Figura 30 – Comparativo dos EPAM entre Base e MAE.



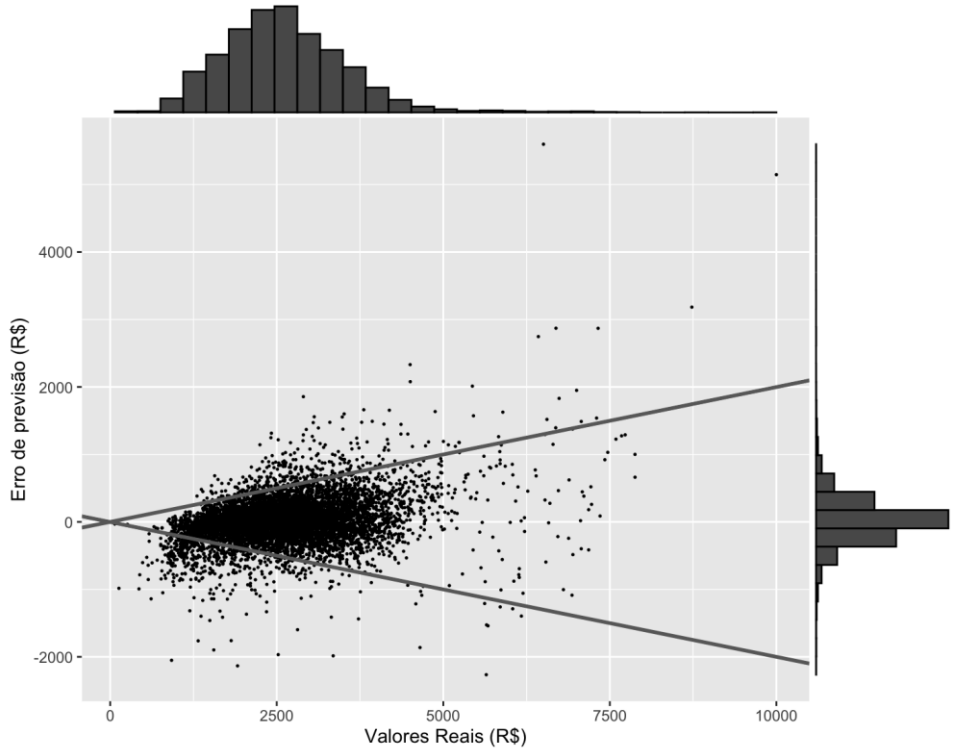
FONTE: Elaboração própria.

Figura 31 – Comparativo dos REQM entre Base e MAE.



FONTE: Elaboração própria.

Figura 32 – Erros da primeira partição dos dados com modelo MAE.



FONTE: Elaboração própria.

A utilização do MAE acarretou num grande ganho de desempenho - em média erros absolutos 32% menores, erros relativos 46% menores e incidência de erros superiores a 20%

em termos absolutos 60% menor. Os resultados apontam para a relevância do fenômeno da autocorrelação espacial. A significância do melhor desempenho percebido só poderá ser atestada, porém, após a realização do protocolo de testes.

4.5 *Support Vector Machine* (SVM)

Os resultados da regressão com o *Support Vector Machine*, ajustado com as mesmas variáveis do modelo RNA, estão apresentados no Quadro 15. Nele, é possível perceber que além consistentemente mais preciso que o modelo Base, o SVM apresentou erros relativos menos precisos e erros absolutos ligeiramente menores que o MAE. Ademais, o percentual de observações com erros superiores que 20% é ligeiramente menor no SVM que no MAE, ainda que o MAE possua menos observações com erros superiores a 50% e 100%. A Figura 33 e Figura 34 apresentam gráficos de densidade comparando o desempenho em termos relativos e absolutos do modelo Base, MAE e SVM.

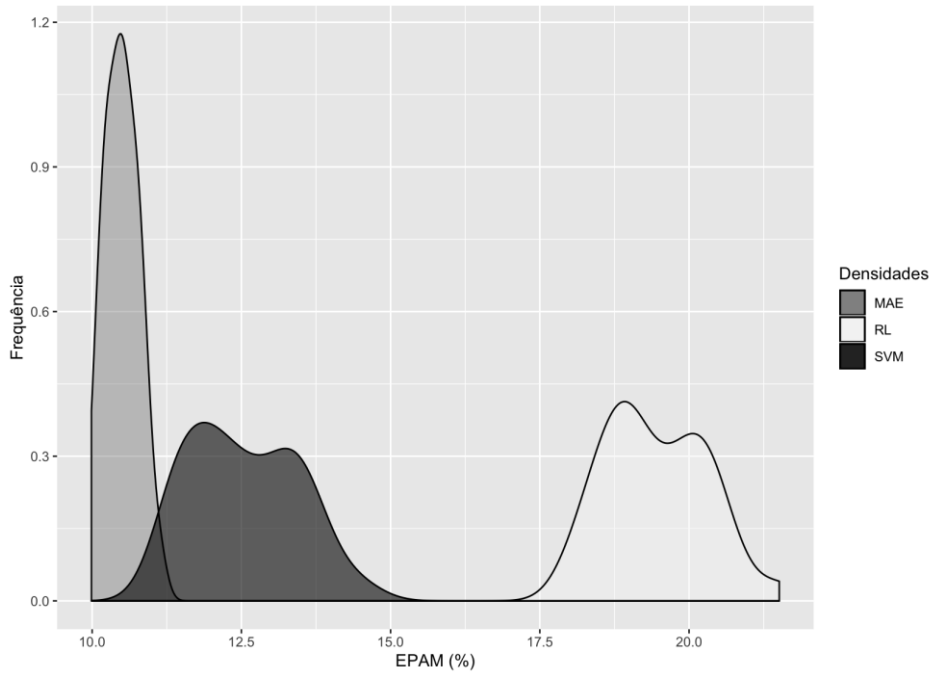
Quadro 15 – Resultados do modelo SVM.

Métrica	Mínimo	Máximo	Média	DP	CV (%)
EPAM (%)	11,30	14,47	12,54	0,89	7,1
REQM (R$)	338	466	367	35	9,4
Erro > 20% (%)	10,5	11,8	11,3	0,38	0,03
Erro > 50% (%)	1,4	2,1	1,7	0,14	8,2
Erro > 100% (%)	0,3	0,6	0,5	0,08	17,2

FONTE: Elaboração própria.

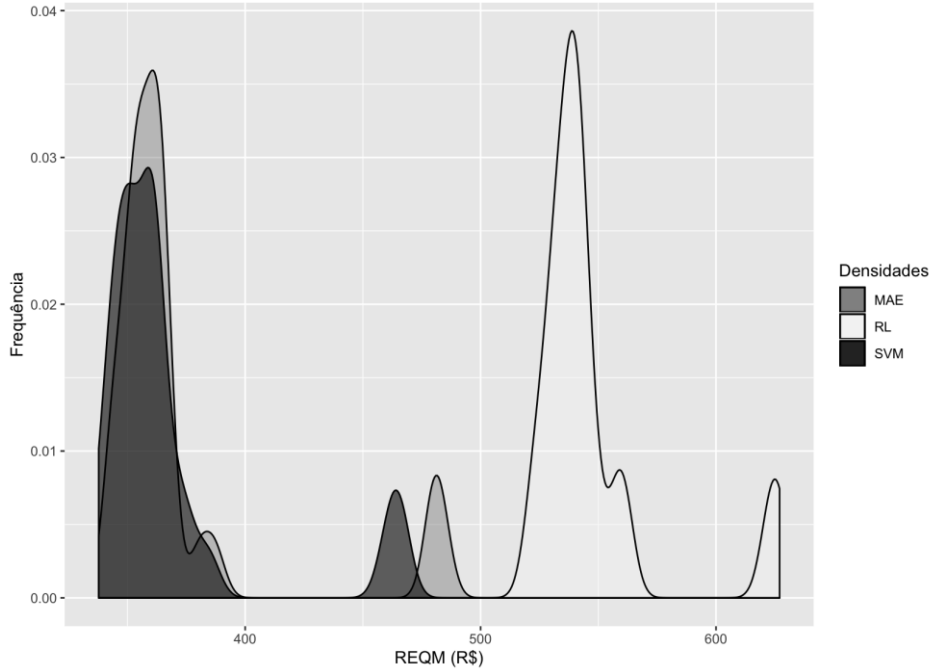
Analisando a Figura 34, percebe-se à primeira vista um padrão semelhante ao observado no modelo MAE. Porém, o SVM não desempenha uniformemente ao longo de todo o intervalo de valores, diferentemente do MAE. Em valores superiores a R$ 4.000,00/m^2, o EPAM é de 9,5%, enquanto em valores inferiores o EPAM é de 14,8%. A diferença no desempenho do modelo nas duas faixas de valores explica a pior precisão em comparação ao MAE.

Figura 33 – Comparativo dos EPAM entre Base, MAE e SVM.



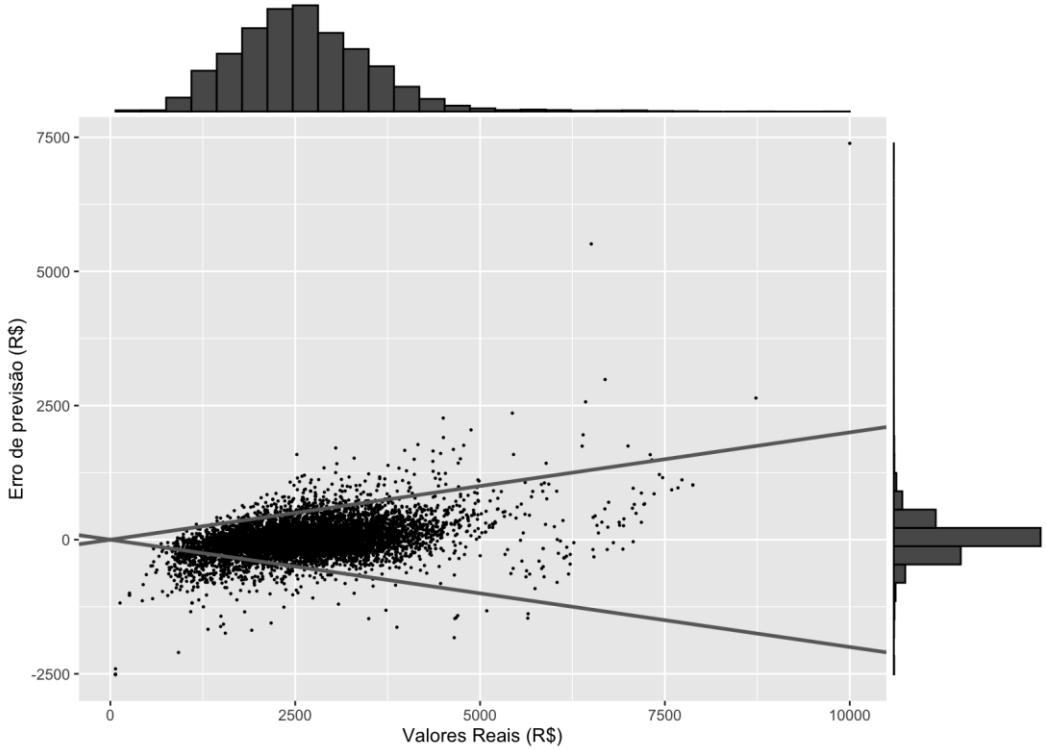
FONTE: Elaboração própria.

Figura 34 – Comparativo dos REQM entre Base, MAE e SVM.



FONTE: Elaboração própria.

Figura 35 – Erros da primeira partição dos dados com modelo SVM.



FONTE: Elaboração própria.

Na comparação com os resultados obtidos por Kontrimas e Verikas (2011), os autores obtiveram com o uso do SVM desempenho aparentemente superior em comparação tanto ao modelo de regressão como à RNA, obtendo EPAM de 15% e 18% de erros percentuais superiores a 20%. Assim como nos resultados obtidos por Lam, Yu e Lam (2009), o desempenho do SVM foi aparentemente superior ao do modelo base de regressão e aos da RNA. Lam, Yu e Lam (2009) não encontraram erros absolutos superiores a 20%, todavia, o que pode decorrer da uniformidade do conjunto usado pelos autores (o coeficiente de variação encontrado para a variável dependente na amostra de testes é pouco superior 5%).

4.6 *Random Forest* (RF)

O modelo *Random Forest* (RF) foi ajustado utilizando as mesmas variáveis utilizadas no modelo RNA, com 300 árvores criadas aleatoriamente com um tamanho mínimo de nós terminais igual a 3 e 19 variáveis candidatas aleatoriamente selecionadas para guiar as ramificações em cada árvore (sendo o critério de escolha da variável para a ramificação a redução da impureza do nó). O algoritmo seleciona 70% das observações para a criação das árvores e realiza o teste em cada uma com os 30% restante. Ao aplicar os conjuntos de testes no modelo RF, as previsões obtidas são as médias das 300 árvores da "floresta aleatória" criada.

Os resultados do RF estão apresentados no Quadro 16. O desempenho do RF em

termos de erros relativos e absolutos é comparado com o modelo Base e o MAE na

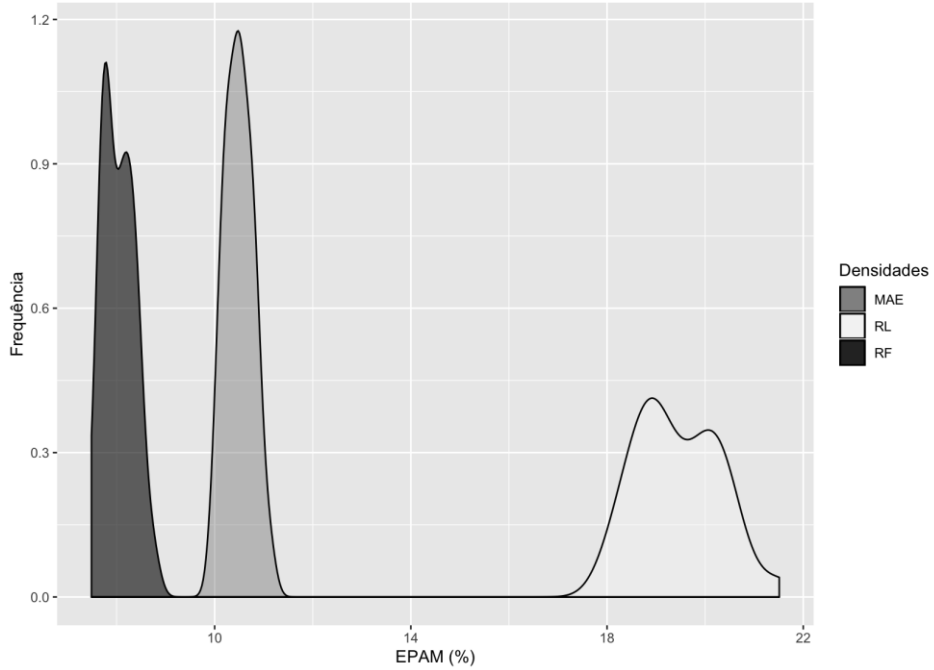
Figura 36 e Figura 37, respectivamente. Delas, apreende-se que o desempenho do RF é superior dentre os modelos testados em todos os aspectos, apesar da superposição entre os gráficos de densidade do erro absoluto do RF e MAE. Notável a semelhança entre os perfis de erro absoluto resultante dos três tipos de modelo, mostrando que apesar do ganho de precisão absoluta os modelos respondem de maneira semelhante aos dados de cada partição.

Quadro 16 – Resultados do modelo RF.

Métrica	Mínimo	Máximo	Média	DP	CV (%)
EPAM (%)	7,50	8,74	8,04	0,31	3,9
REQM (R$)	278	435	312	43	13,7
Erro > 20% (%)	6,9	7,9	7,5	0,25	3,4
Erro > 50% (%)	0,6	1,1	0,9	0,10	11,3
Erro > 100% (%)	0,1	0,4	0,2	0,06	23,5

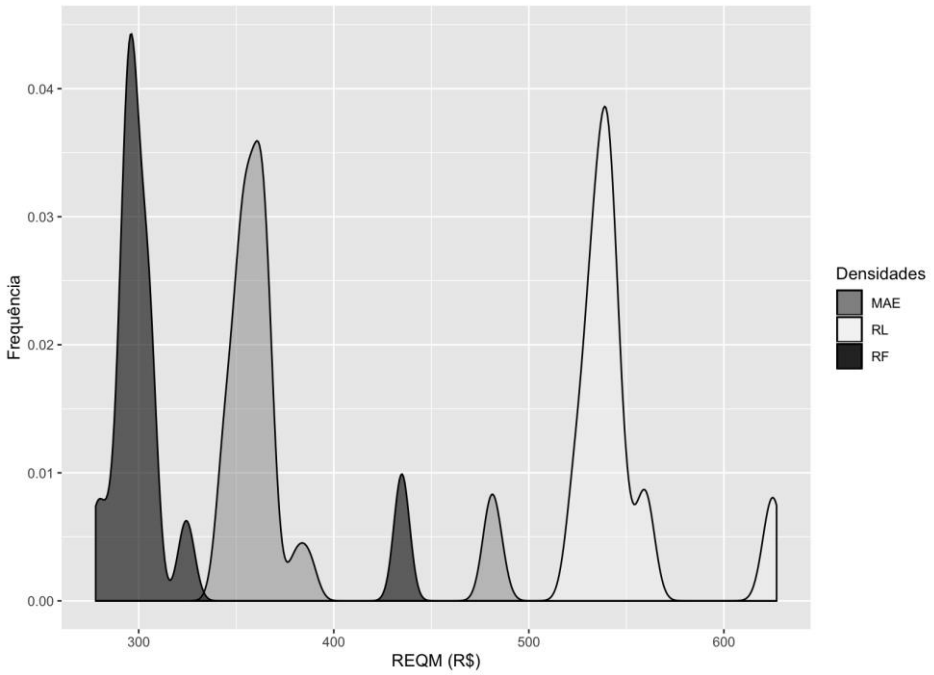
FONTE: Elaboração própria.

Figura 36 – Comparativo dos EPAM entre Base, MAE e RF.



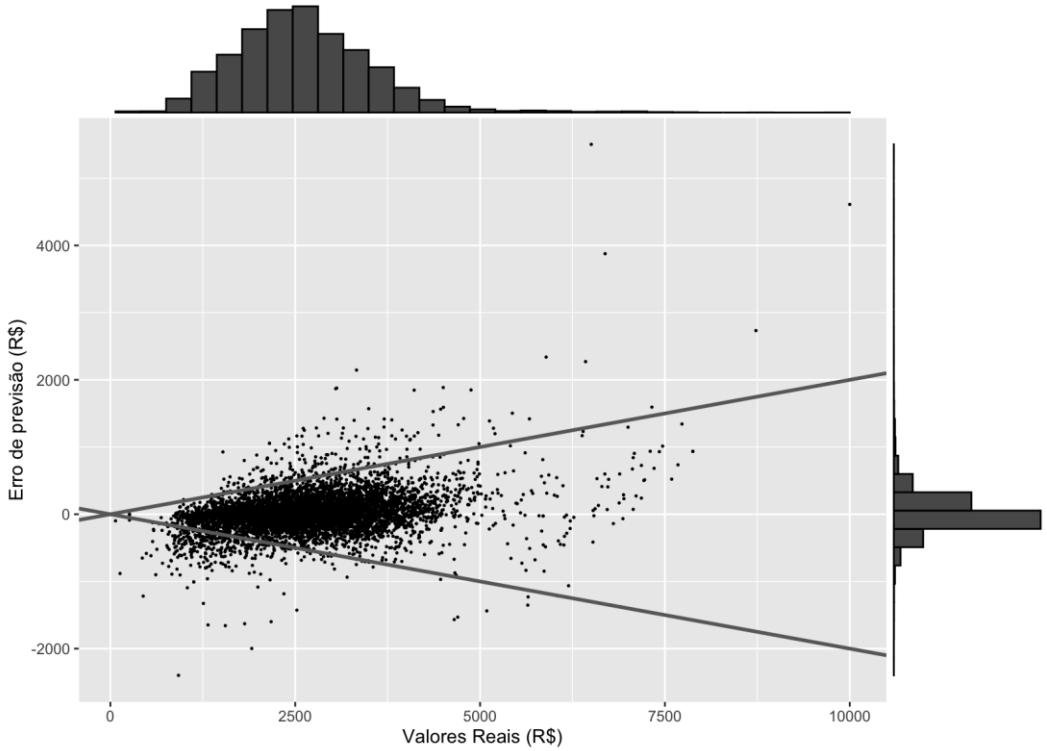
FONTE: Elaboração própria.

Figura 37 – Comparativo dos REQM entre Base, MAE e SVM.



FONTE: Elaboração própria.

Figura 38 – Erros da primeira partição dos dados com modelo RF.



FONTE: Elaboração própria.

Na Figura 38, com os erros do RF em todo o intervalo dos valores, é nítido que há uma quantidade menor de pontos além dos limiares de 20%, quando em comparação com os

erros do modelo Base (Figura 26) e até mesmo do MAE (Figura 32). Ainda se observa um acúmulo de erros negativos superiores a 20% em observações de valores inferiores a R$ 1.250,00/m^2 e uma redução significativa de erros negativos superiores a 20% em valores próximos a R$ 2.500,00/m^2 quando em comparação com o MAE e Base. O RF também demonstra estabilidade ao longo de todo o intervalo de valores, com EPAM para valores superiores a R$ 4.000,00/m^2 igual a 7,9% e para inferiores igual a 7,7%.

Com o RF, foi obtido um patamar de desempenho superior ao reportado por Antipov e Pokryshevskaya (2012), quando o ganho de desempenho em comparação ao modelo de regressão é da ordem de 20% (de um EPAM de 18% para um de cerca de 15%). É um desempenho semelhante ao obtido por Čeh *et al.* (2018), quando obteve-se um EPAM no patamar de 7% com o RF, enquanto com a regressão múltipla obteve a métrica de cerca de 17% (aprimoramento de cerca de 59%).

4.7 Análise Comparativa entre os Modelos

Dando sequência ao protocolo de testes de Granatyr (2017), foi aplicado o teste não-paramétrico de Friedman às métricas de erro relativo e absoluto. Para o EPAM, o teste revelou que existe evidência suficiente (estatística chi-quadrado igual a 120, correspondente a um p valor inferior a 5%) para rejeitar a hipótese nula que afirma que os perfis de erro são idênticos. O mesmo foi observado para a REQM (estatística chi-quadrado igual a 114, correspondente a um p valor inferior a 5%).

Com as significâncias atestadas pelo teste de Friedman, procede-se para o teste "*post-hoc*" de Nemenyi. Os Quadro 17 e Quadro 18 apresentam o valor p dos testes pareados que comparam modelo a modelo para os erros relativos e absolutos (neles, os valores em negrito mostram os testes pareados que não obtiveram significância para rejeitar hipótese de igualdade).

No tocante ao erro relativo, o modelo RNA não obteve performance diferente do modelo Base e do modelo SVM, havendo, portanto, evidências que o MAE e o modelo RF são de fato mais precisos que o modelo RNA. Não há evidências de que o MAE possua desempenho diferente do modelo SVM e do modelo RF, todavia, é superior ao modelo Base e modelo RNA. O modelo SVM não demonstrou desempenho significativamente diferente do MAE e do RNA, porém distinto do modelo Base e modelo RF. O modelo RF apenas não tem desempenho significativamente distinto do MAE.

A Figura 39 representa graficamente a comparação entre os modelos. Nela, as retas unem os modelos que possuem desempenho estatisticamente indistintos. Dessa forma, em ordem de menor erro relativo, o modelo RF é numericamente mais preciso que o MAE, ainda

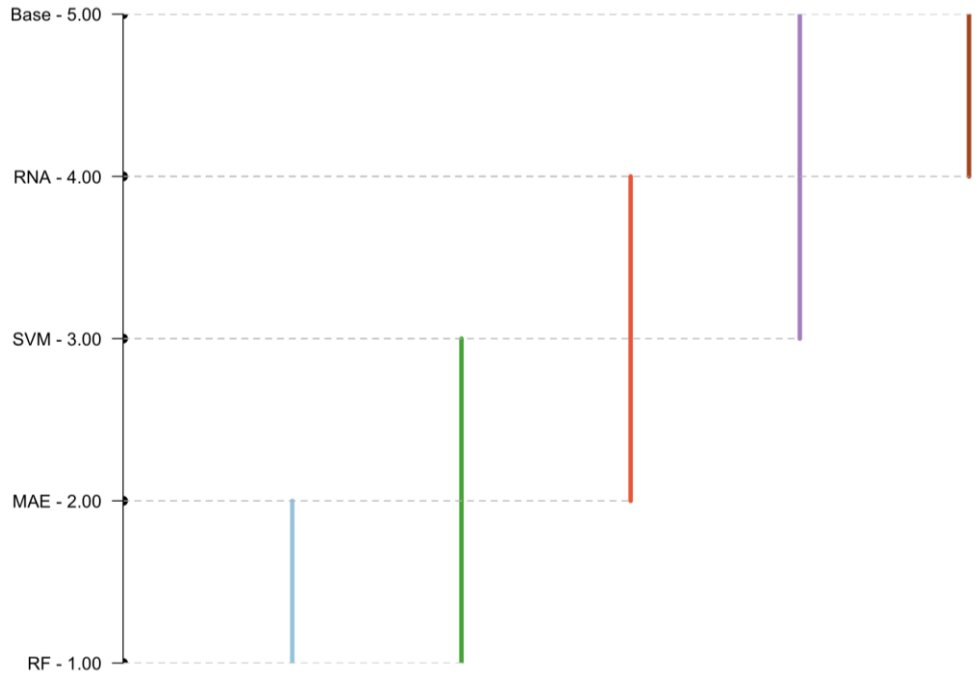
que não haja diferença estatisticamente significante entre eles. MAE, por sua vez, assume uma posição intermediária entre o modelo RF e o modelo SVM, não sendo estatisticamente distinto de ambos. Em seguida, o modelo SVM assume uma posição intermediária entre MAE e o modelo RNA, não havendo distinção significante com cada um deles. Por fim, o modelo RNA e o modelo Base, estatisticamente idênticos, sendo o último numericamente menos preciso. Destaca-se que, no caso do MAE e modelo RF, os gráficos de densidade dos erros relativos não se sobrepõem, mas o p-valor obtido de 0,10 não permite rejeitar a hipótese nula de identidade entre os resultados com 95% de grau de confiança.

Quadro 17 – Comparação pareada entre EPAM dos modelos.

	Base	**RNA**	**MAE**	**SVM**	**RF**
Base	–	**0,10**	$2{,}0 \times 10^{-12}$	$9{,}5 \times 10^{-6}$	$5{,}2 \times 10^{-14}$
RNA	**0,10**	–	$9{,}5 \times 10^{-6}$	**0,10**	$2{,}0 \times 10^{-12}$
MAE	$2{,}0 \times 10^{-12}$	$9{,}5 \times 10^{-6}$	–	**0,10**	**0,10**
SVM	$9{,}5 \times 10^{-6}$	**0,10**	**0,10**	–	$9{,}5 \times 10^{-6}$
RF	$5{,}2 \times 10^{-14}$	$2{,}0 \times 10^{-12}$	**0,10**	$9{,}5 \times 10^{-6}$	–

FONTE: Elaboração própria.

Figura 39 – Ordenação e comparação entre os modelos para o EPAM.



FONTE: Elaboração própria.

No tocante aos erros absolutos, percebe-se que em apenas duas comparações pareadas não houve como rejeitar a hipótese nula de resultados idênticos. As diferenças encontradas entre o modelo Base e o modelo RNA, assim como a diferença entre o modelo

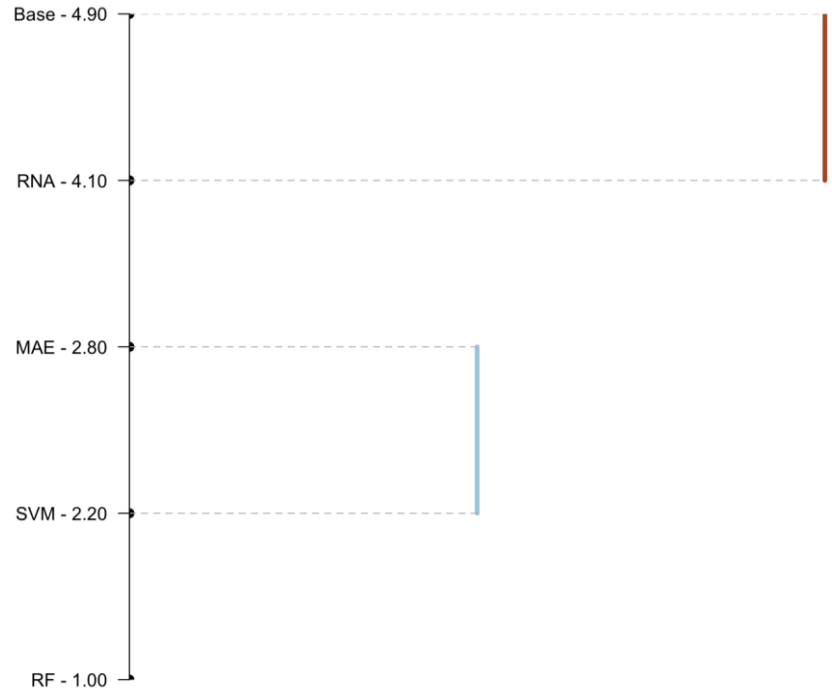
SVM e MAE, não são suficientes para rejeitar a hipótese de identidade entre os pares. Os resultados dos testes são bastante distintos do observado no teste dos erros relativos. De acordo com o apresentado na Figura 40, apesar de não haver sido rejeitada a hipótese nula na comparação entre o modelo SVM e MAE, o primeiro apresentou erros absolutos numericamente menores. Das comparações, resulta que o modelo RF é isoladamente o mais preciso em termos absolutos dentre todos os testados.

Quadro 18 – Comparação pareada entre REQM dos modelos

	Base	**RNA**	**MAE**	**SVM**	**RF**
Base	–	**0,29**	$2,7 \times 10^{-6}$	$3,7 \times 10^{-10}$	$4,0 \times 10^{-14}$
RNA	**0,29**	–	$1,2 \times 10^{-2}$	$3,2 \times 10^{-5}$	$3,4 \times 10^{-13}$
MAE	$2,7 \times 10^{-6}$	$1,2 \times 10^{-2}$	–	**0,58**	$1,0 \times 10^{-4}$
SVM	$3,7 \times 10^{-10}$	$3,2 \times 10^{-5}$	**0,58**	–	$2,7 \times 10^{-2}$
RF	$4,0 \times 10^{-14}$	$3,4 \times 10^{-13}$	$1,0 \times 10^{-4}$	$2,7 \times 10^{-2}$	–

FONTE: Elaboração própria.

Figura 40 – Ordenação e comparação entre os modelos para o REQM.



FONTE: Elaboração própria.

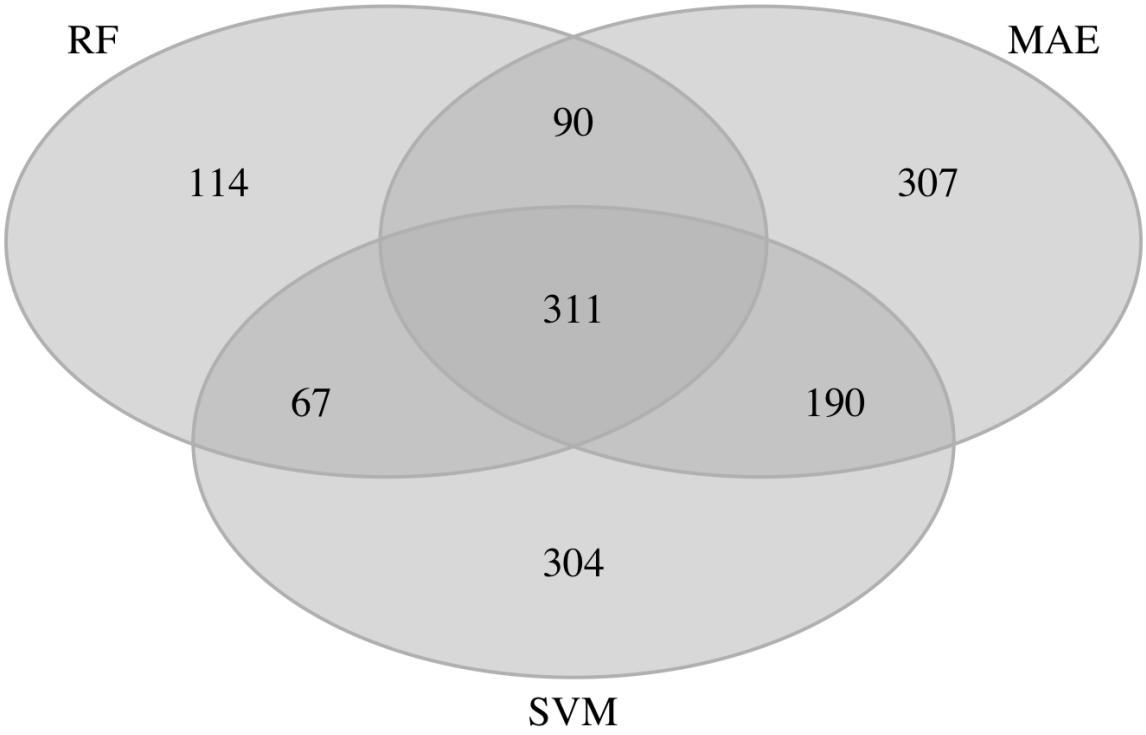
4.8 Análise dos Erros

Tomando os três modelos testados com melhor desempenho e apenas o primeiro subconjunto particionado, foram extraídas e comparadas as observações que resultam em cada modelo erros absolutos superiores a 20% (limiar utilizado para filtro), obtendo o perfil retratado

na Figura 41. Ao todo, 1.383 observações do primeiro subconjunto geram erro superior ao limiar de 20%, o que corresponde a um percentual superior ao que foi obtido em cada um dos modelos (17,65%). Isso se dá pois, conforme pode ser apreendido da Figura 41, existem observações únicas a cada um dos modelos (cerca de metade das 1.383 observações extremas). RF, tendo o melhor desempenho aferido, possui 114 observações extremas únicas (8,2% das observações extremas), enquanto MAE e SVM possuem respectivamente 307 e 304 observações extremas (22,2% e 22%).

Os três modelos possuem 311 observações extremas em comum (22,5%). Isolando apenas tais observações extremas comuns, o RF, MAE e SVM atingem, respectivamente, EPAM igual a 45%, 49% e 61%. Já no âmbito de erros absolutos, o REQM resultante é de R$ 942, R$ 989 e R$ 1.030. As métricas são, portanto, muito superiores ao limiar de 20% e superiores aos valores reportados para cada modelo.

Figura 41 – Indicativo de observações extremas dos modelos e suas interseções.



FONTE: Elaboração própria.

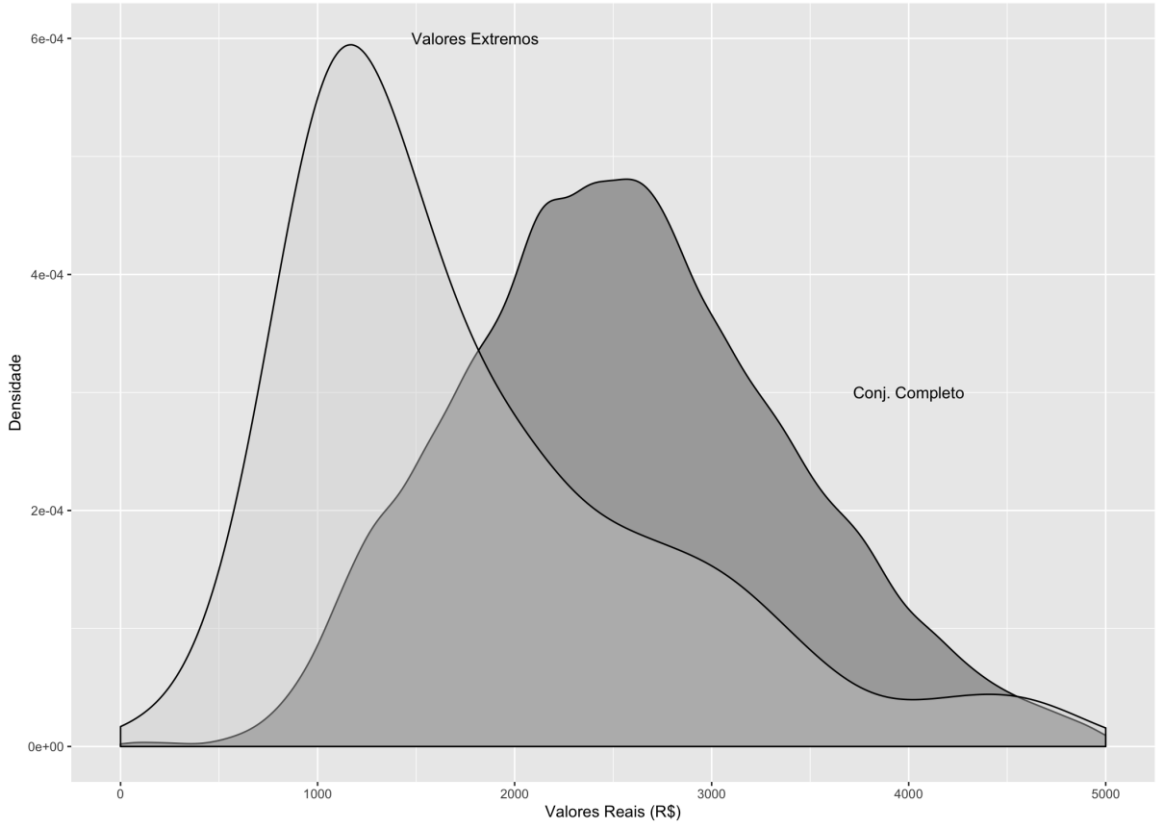
Comparando essas 311 observações (que receberá a partir de então a denominação “grupo de extremos”) com todo o conjunto de dados, percebe-se que no grupo de extremos têm-se uma maior predominância de baixos valores unitários de imóveis e mais antigos. No Quadro 19 são apresentados os 15 bairros com maior incidência percentual de imóveis extremos e ao

lado a incidência observada no conjunto completo. Em média, o grupo dos extremos possui preço unitário de R$ 1.942, enquanto na base completa é de R$ 2.628, e idade de 22,3 anos, enquanto na base completa é de 11,3 anos.

Gráficos de densidade podem ser observados na Figura 42 e Figura 43, de onde apreende-se que são poucas as observações extremas que podem ser consideradas como novas. Apenas 15 dos 311 imóveis (4,8%) possuem idade inferior a 5 anos, quando ainda vigoram garantias. O percentual de imóveis nessa faixa de idades na base completa é de cerca de 37%. Da mesma forma, verifica-se uma forte presença de imóveis com preço unitário inferior a R$ 1.000 – 55 de 311, correspondente a 18%, quando na base completa o percentual correspondente é de 1,6%.

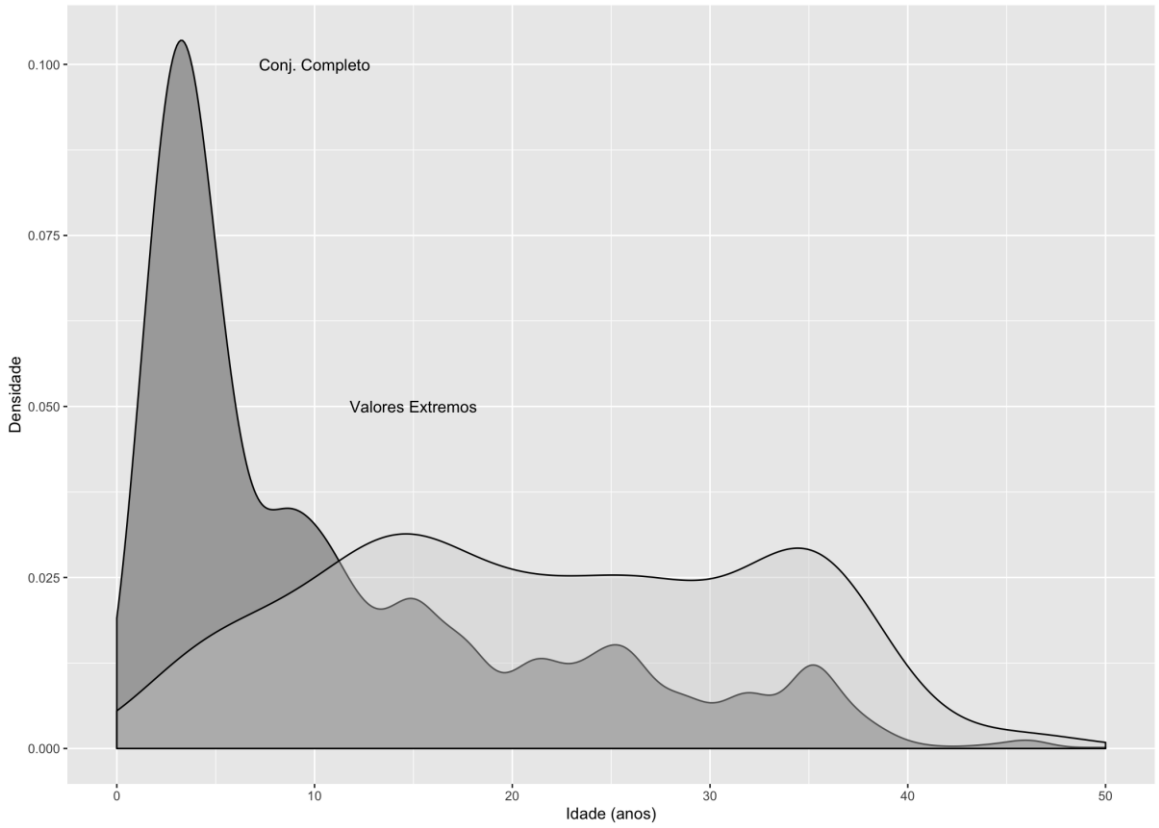
O grupo dos extremos é também em média mais próximo da Avenida Beira Mar (variável “dist_BM”), valor de 4,6 km para o grupo de extremos e 5,6 km da base completa, e composto por imóveis localizados em imóveis mais baixos (número de pavimentos em média de 8,6 no grupo dos extremos e 13,1 na base completa) e com menor quantidade de unidades no lote em média (“num_unidades_lote”, com médias 94,6 e 204,3 unidades respectivamente).

Figura 42 – Preços unitários do grupo extremo e base completa.



FONTE: Elaboração própria.

Figura 43 – Comparativo de idade do grupo extremo e base completa.



FONTE: Elaboração própria.

Quadro 19 – Comparativo do grupo de extremos com o conjunto total.

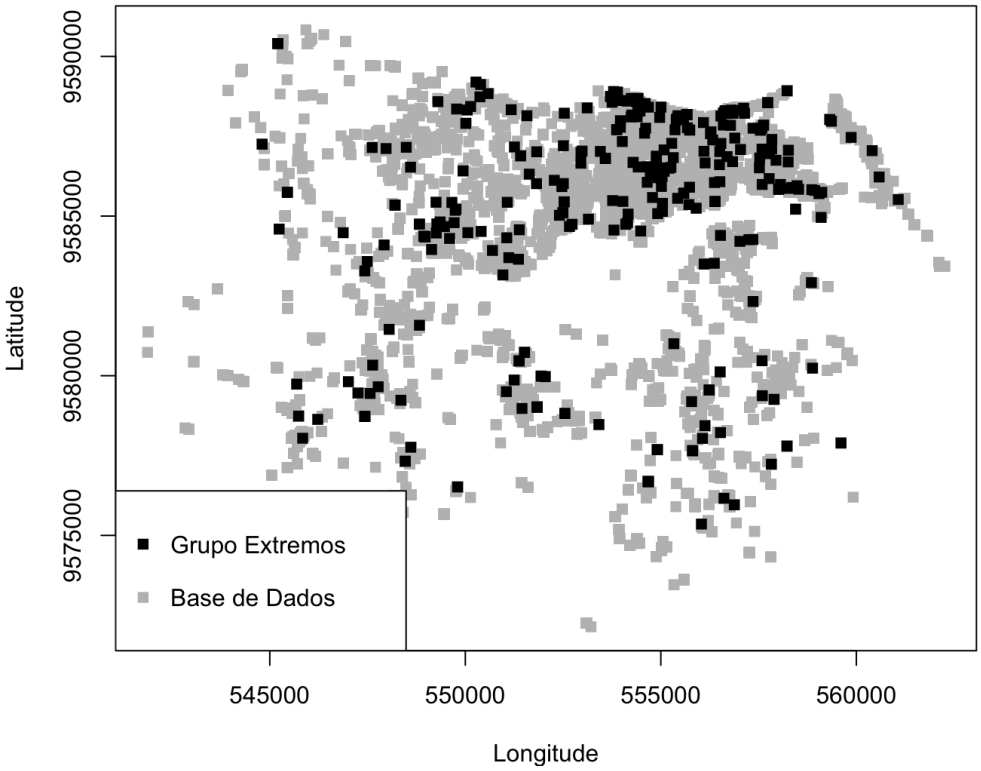
Bairro	Grupo Extremo			Conjunto de Dados		
	%	Idade (anos)	Preço Unitário (R$)	%	Idade (anos)	Preço Unitário (R$)
Meireles	12,86	20,88	3360	11,44	12,56	3534
Aldeota	8,36	27,04	1715	8,08	12,83	3000
Centro	4,50	30,29	1465	2,29	16,20	2332
Cocó	4,50	21,86	2009	5,88	9,70	2995
Mucuripe	4,50	23,50	2385	2,21	13,13	3322
Passaré	4,18	13,31	1726	3,74	6,93	2231
Dionísio Torres	3,86	33,67	2264	2,11	14,65	2867
Fátima	3,86	23,17	1766	3,92	11,33	2738
S. João do Tauape	3,54	35,09	1267	0,97	18,99	2210
Papicu	3,22	24,60	1333	2,05	12,89	2613
Messejana	2,57	13,88	1615	4,89	6,14	2155
Mondubim	2,57	17,00	1465	3,67	8,39	2068
Barroso	2,25	10,71	2084	0,39	10,71	1704
Praia de Iracema	2,25	22,29	2208	1,21	13,52	2722
Itaperi	1,93	14,50	1232	1,89	10,41	2008

FONTE: Elaboração própria.

Os dados do Quadro 19 mostram ligeira discrepância na distribuição geográfica entre o grupo de extremos e a base completa. As coordenadas do centroide do grupo de extremos é (553.773,7; 9.584.961) enquanto o da base completa é (553.649,3; 9.584.040). Ou seja, em

média os imóveis do grupo de extremos concentram-se mais para a região leste e norte da cidade, porém, como distância entre os dois centroides é igual a 796,6 m, a diferença não é tão significativa. Como pode ser visto na Figura 44, algumas regiões da cidade expressam incidência percentual menor, porém não se percebe distribuição muito distinta da base original.

Figura 44 – Distribuição geográfica de extremos e base de dados.



FONTE: Elaboração própria.

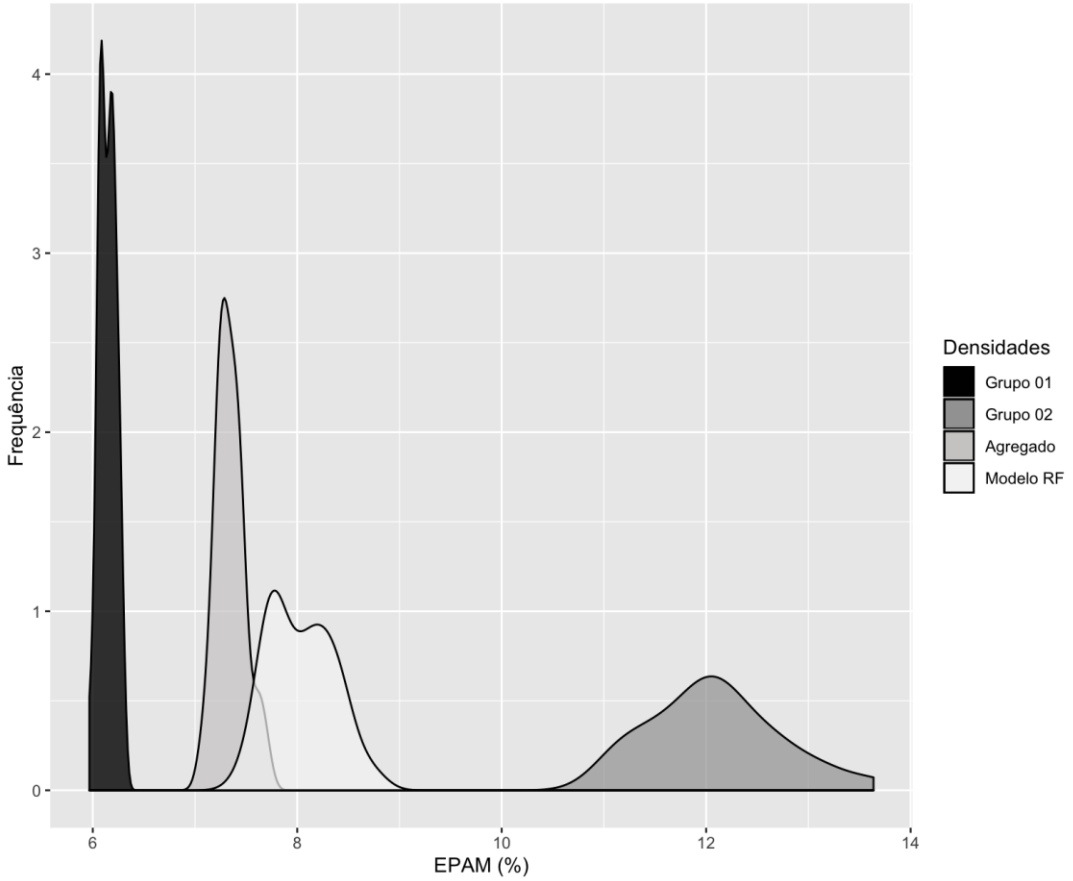
Com as informações advindas da análise do perfil dos erros, foi segregada a base de dados completa em duas: o primeiro grupo, denominado “Grupo 01”, com imóveis com valores superiores a R$ 1.500 e idade inferior a 25 anos (correspondente a 31.258 observações, ou cerca de 80%) e outro, denominado “Grupo 02”, com valores inferiores ou iguais a R$ 1.500 e idade superior ou igual a 25 anos (correspondente a 7.923 observações, ou cerca de 20%). Com base nesses dois grupos, foram ajustados um modelo *Random Forest* para o Grupo 01 e outro para o Grupo 02. Os erros relativos e absolutos de cada grupo estão no Quadro 20. Na Figura 45 e Figura 46 são expostos os gráficos de densidade das métricas de relativa e absoluta, respectivamente.

Quadro 20 – Comparativo entre grupos, agregado e modelo RF.

Grupo	Métrica	Mínimo	Máximo	Média	DP	CV (%)
Grupo 01	EPAM (%)	5,97	6,28	6,14	0,08	1,3
	REQM (R$)	258	449	302	68	22
Grupo 02	EPAM (%)	11,0	13,6	12,1	0,64	5,3
	REQM (R$)	300	444	354	42	12,0
Agregado	EPAM (%)	7,08	7,69	7,34	0,14	1,9
	REQM (R$)	271	429	315	55	17
Modelo RF	EPAM (%)	7,50	8,74	8,04	0,31	3,9
	REQM (R$)	278	435	312	43	13,7

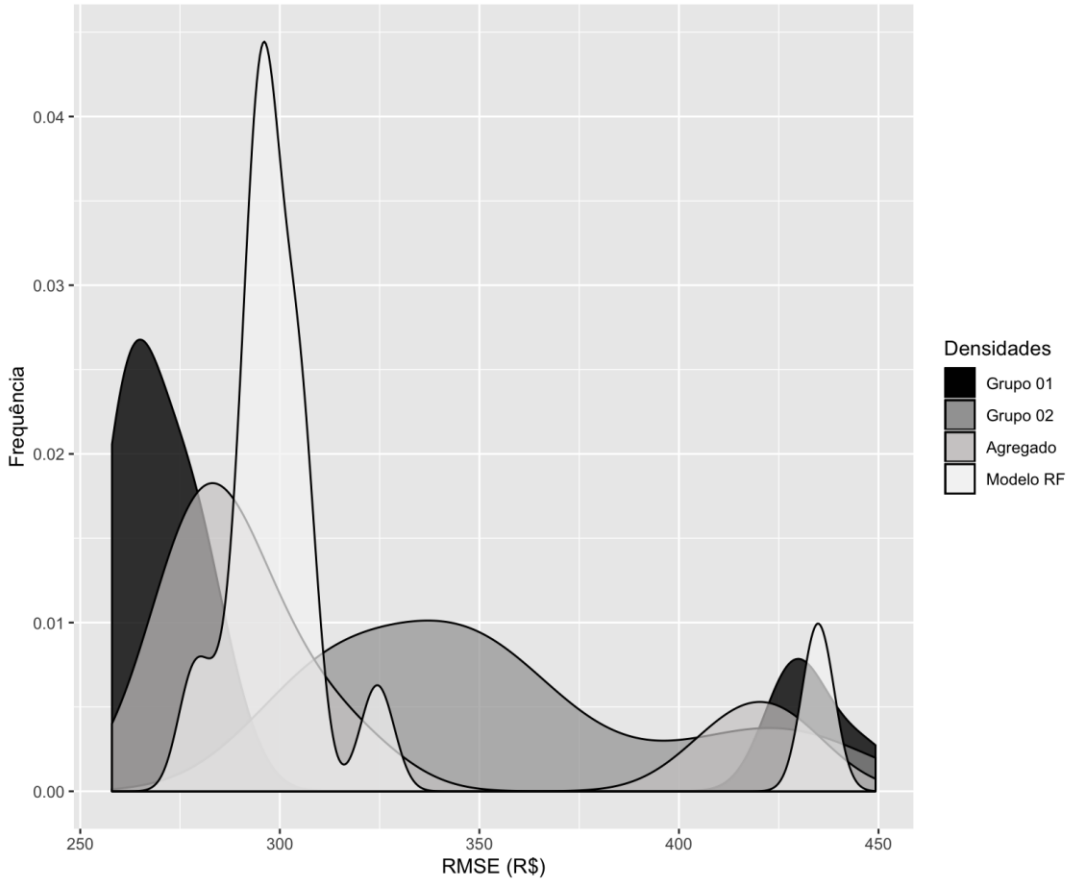
FONTE: Elaboração própria.

Figura 45 – Comparativo dos resultados relativos dos grupos, agregado e RF.



FONTE: Elaboração própria.

Figura 46 – Comparativo dos resultados absolutos dos grupos, agregado e RF.



FONTE: Elaboração própria.

Nota-se grande diferença de desempenho do algoritmo para os dois grupos, porém com um efeito agregado de desempenho (quando os resultados dos dois grupos foram reunidos para o cálculo das métricas) numericamente superior ao que foi observado no modelo RF apresentado anteriormente no âmbito do erro relativo (EPAM). Já o desempenho agregado no âmbito dos erros absolutos não parece ter sido alterado significativamente, ainda que haja sensível mudança no perfil do gráfico de densidade.

Aplicando o protocolo de testes de Friedman e "*post-hoc*" de Nemenyi para atestar significância estatística nas diferenças dos resultados agregados dos dois grupos e no modelo RF, encontra-se que para os erros relativos há evidência para rejeitar a hipótese nula de identidade entre os dois perfis de resultados (p-valor de $4{,}32 \times 10^{-8}$, inferior ao limiar de 5%). O "*post-hoc*" de Nemenyi, por sua vez, não se mostra necessário, já que o teste de Friedman é feito entre um par.

4.9 Comparação com Resultados da Literatura

O modelo RF obteve erro relativo (EPAM) entre 7,50 e 8,74%, com média 8,04%, um desempenho superior ao reportado por Antipov e Pokryshevskaya (2012) – 17,25 e 14,86% nos grupos de testes - e comparável ao reportado por Čeh *et al.* (2018) – 7,27% no grupo de testes. O MAE, por sua vez, obteve também resultados numericamente superiores ao encontrado na literatura – McCluskey *et al.* (2013) relatam EPAM de 13,69% enquanto o MAE no presente estudo obteve resultados entre 9,98 e 11,12%. Apenas a comparação pontual é possível, tendo em vista que apenas resultados pontuais são relatados pelos estudos, não sendo possível realizar testes de significância estatísticos.

Os modelos Base e RNA apresentaram erros relativos entre 18,11 e 21,51% e 15,43 e 17,98%, respectivamente, que são mais próximos ao limite superior identificado na literatura – conforme o Quadro 3, ambos modelos lineares e em RNA obtiveram erros relativos entre 10 e 23%. Os erros relativos obtidos por Codes (2018) inferiores a 10% com o uso das RNA para imóveis também de Fortaleza é explicado pelo fato de que o autor segregou a análise por bairro e por ano de registro da transação, o que resulta numa amostra de treino e testes mais homogênea. Como explicitado na análise exploratória, o banco de dados do presente trabalho possui elevada variabilidade intra-regional. Os erros obtidos com o modelo RF para toda a base possuem magnitude comparável àqueles obtidos com RNAs relatados por Codes (2018) (erros entre 4 e 7%).

Nunes (2016), com modelo de regressão linear e com dados de imóveis de Fortaleza, expressa medição do erro médio relativo de predição (métrica idêntica ao EPAM) de 11 imóveis que compõem o conjunto de testes, sendo os valores variando entre 2,4 e 92,0%, com média de 21,1% e desvio-padrão de 23,80%. Nunes (2016) também reporta 20% previsões cujo erro é superior a 20% e 1,77% com erros superiores a 40%. O modelo linear do presente trabalho obteve erro relativo comparável com o obtido por Nunes (2016) – média pouco abaixo de 20%, pouco menos de 30% estimativas com erro superior a 20% e 4,3% de estimativas com erro superior a 50%.

No âmbito da análise da influência das variáveis na formação de preço, detectou-se um efeito negativo de variáveis relacionadas com conectividade da malha de transporte urbano – a saber, “pontos_de_onibus”, “prox_metro_300” e “prox_metro_500” – no preço dos imóveis. Contrariamente ao observado, Wang *et al.* (2015) detectou influência positiva entre o preço dos móveis e a conexão com a rede de ônibus na cidade de Cardiff, no Reino Unido. Čeh *et al.* (2018) para imóveis na Liubliana, na Eslovênia, detectou que a proximidade de linhas de

trem impactava negativamente ao valor dos imóveis. Shin, Washington e Choi (2007) detectaram que maiores distâncias para estações de metrô implicam em desvalorização nos imóveis na região metropolitana de Seoul, na Coreia do Sul. Resultados semelhantes ao de Shin, Washington e Choi (2007) foram encontrados por Hess e Almeida (2007) para imóveis de Buffalo, Nova Iorque, onde a proximidade da rede de veículo leve sobre trilhos (VLT) influencia positivamente no valor dos imóveis. Hamidi, Kittrell e Ewing (2016), em meta-análise com dados de casas dos EUA e Canadá, detectou em média um gradiente de valor de 0,168% para cada 30,48 m (100 pés) de proximidade a pontos de metrô ou transporte regional.

Efthymiou e Antoniou (2015) detectaram influência negativa da proximidade a linhas de trem e linhas de metrô para o preço de compra de imóveis de Tessalônica, no norte da Grécia. Observaram que não houve influência de proximidade ao metrô para preços de aluguel e influência negativa da proximidade de linhas de trem e proximidade ao aeroporto. Efthymiou e Antoniou (2015) listam como possível motivo para a influência negativa o fato das linhas de metrô na região estarem submetidas a obras de ampliação. Revisão bibliográfica reportada pelos autores demonstra que não há uniformidade na influência entre fatores ligados a acessibilidade ao sistema de transporte – proximidade a linhas de trem, metrô, avenidas e rodovias, aeroportos e LVT. Compilação bibliográfica de estudos conduzidos com imóveis em cidades americanas feita por Hess e Almeida (2007) apontam para a possibilidade de influência da proximidade a estações de trem distinta em função do poder aquisitivo.

He *et al*. (2010), em análise para a cidade de Pequim, na China, não encontraram significância estatística na influência de variáveis de acessibilidade ao valor dos imóveis, apesar os coeficientes de correlação entre estas e o preço terem se mostrado relevantes. Também não detectaram significância na proximidade a parques e amenidades. No tocante a influência da proximidade a parques e áreas verdes, os resultados apresentados por Zoppi, Argiolas e Lai (2015) são semelhantes ao deste estudo: influência positiva no valor.

Foi encontrado um coeficiente positivo para o inverso da variável “dist_BM”, o que implica que uma maior distância da avenida Beira Mar – que resulta num valor menor para a razão “1/dist_BM” – impacta negativamente no valor do imóvel. Resultados semelhantes foram reportados por Zoppi, Argiolas e Lai (2015) para a cidade Cagliari, uma cidade litorânea da Itália. Semelhante influência também foi detectada por Efthymiou e Antoniou (2015) para a também cidade litorânea de Tessalônica, na Grécia.

5 CONCLUSÃO

Para construção do banco de dados que incorporaram os modelos, foi desenvolvido um algoritmo que relaciona o posicionamento do imóvel do registro municipal com outras bases de dados independentes, integrando com apoio em operações de sistema de informações geográficas 53 variáveis para cada registro imobiliário. A base de dados de segurança pública não estruturada precisou ser processada e georreferenciada mediante um procedimento automatizado com auxílio de um *script* desenvolvido em Python. Há, portanto, na construção do banco de dados interfaces geradas que oferecem base para uma futura integração desde a origem. A estruturação do banco de dados, ainda que automatizada por via de *scripts*, depende de manipulação, uma potencial fonte de ruídos e erros.

Analisando os dados do mercado da cidade de Fortaleza, percebeu-se elevada variabilidade nos valores dos imóveis. Isolando os dados por bairro, nota-se elevada incidência de coeficientes de variabilidade superiores a 20% - mais de 70% dos bairros apresentaram resultados superiores a esse limiar, sendo possível visualizar amplitudes de R$ 5.600 a R$ 30.400. A variabilidade nos valores médios por área é explicada em parte pela variabilidade (ainda maior) de variáveis dependentes cuja correlação com o preço é importante, como é o caso da variável “idade”.

Por via da regressão múltipla, foram detectados quais relacionamentos lineares entre as variáveis independentes e o preço do imóvel não apresentam significância estatística, permitindo também detectar, desde a comparação entre coeficientes, que é necessário repensar a classificação dos imóveis com respeito ao padrão (no caso, popular, normal, elevado e de luxo). Os coeficientes encontrados pela regressão múltipla não conferem a estas variáveis a escala pretendida em face do conjunto completo das variáveis. O aprofundamento do estudo desse relacionamento é uma possível novo tema de pesquisa.

Observou-se que divisão do banco de dados em *clusters* segregou os imóveis em categorias distintas que não levam em conta apenas posicionamento geográficos, mas todos os aspectos sobre os quais os imóveis são caracterizados. Outro aprofundamento do presente estudo é a avaliação da precisão dos modelos de regressão propostos nas observações integrantes de cada “*cluster*”, o que dá ensejo a uma melhor avaliação dos pontos fracos de cada aplicação, bem como do teste de qual a melhor quantidade de “*clusters*” para a categorização, ou se devem todas as variáveis integrar o processo de “clusterização” ou não.

A segmentação da base em diferentes grupos permitirá também uma análise mais aprofundada do relacionamento entre valor e as variáveis. Conforme explorado na comparação

com resultados da literatura, efeitos de fatores de acessibilidade no valor dos imóveis depende de particularidades regionais, sendo possível que para uma mesma cidade haja influência divergente em regiões de diferentes rendas médias. Um estudo mais aprofundado do tema também é uma outra recomendação para futuros trabalhos.

Dos tipos de modelos de regressão testados, o "*Random Forest*" (RF) e o "*Modelo Autoregressivo Espacial*" (MAE) retornaram as menores métricas de erro relativos e absolutos dentre todos, sendo o RF numericamente mais preciso que todos os cinco modelos testados. Da mesma maneira, demonstraram desempenho indistintos entre si do ponto de vista estatístico para a métrica relativa, a saber, o erro percentual absoluto médio (EPAM). Para o erro absoluto, a raiz do erro quadrático médio (REQM), os resultados melhores do RF em comparação ao MAE possuem significância estatística. O MAE, por sua vez, não demonstrou desempenho estatisticamente distinto do modelo "*Support Vector Machine*" (SVM) tanto para os erros relativos e do absoluto, ainda que tenha demonstrado desempenho numericamente melhor.

O MAE, ao contrário do RF, é de mais fácil interpretação, haja vista que o modelo é traduzido numa equação linear, fator importante no âmbito da tributação municipal e estadual. Considerando que não foi encontrada evidência para rejeitar a hipótese de que o MAE e o RF possuem o mesmo desempenho nas métricas relativas, a utilização do MAE para tal fim não é, a rigor, desaconselhada.

A comparação do desempenho dos modelos com testes estatísticos só foi possível porque foi seguida uma adaptação do protocolo proposto por Granatyr (2017) para modelos de regressão. Tal prática permite a comparação rigorosa do desempenho dos modelos, evitando a comparação pontual de métricas.

Os erros dos modelos RF, MAE e SVM foram isolados e analisados no intuito de identificar potenciais melhorias ao processo de modelagem. Foi identificado que erros maiores concentravam-se entre imóveis de menor valor e maior idade, tendo sido arbitrado o limiar de preço por área de R$ 1.500/m^2 e idade de 25 anos, o que dividiu o banco de dados em dois grupos. Foi ajustado um modelo RF em cada um dos grupos e os resultados agregados de cada um destes modelos foi comparado ao obtido anteriormente para todo o banco de dados. A segregação permitiu uma melhoria estatisticamente significante de desempenho.

Tendo em vista que a comparação entre modelos não encontrou diferenças significantes entre o modelo RF e MAE, recomenda-se para trabalhos futuros aprofundar o estudo de ambos modelos em distintos padrões de imóveis e idades, buscando identificar se há ganho de desempenho ou não, bem como a diferença de performance de ambos. Recomenda-

se também o estudo para definir um método para seleção de imóveis tendo em vista a estimativa de valor venal de novos imóveis.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ACHARYA, U. R.; OH, L. S.; HAGIWARA, Y.; TAN, J. H.; ADAM, M.; GERTYCH, A.; TAN, R. S. A deep convolutional neural network model to classify heartbeats. **Computers in Biology and Medicine**, v. 89, p. 389–396, 2017.

ALBUQUERQUE, P. H. M.; NADALIN, V. G.; NETO, V. C. L.; MONTENEGRO, M. R. Construção de índices de preços de imóveis para o Distrito Federal por meio de vendas repetidas e GWR. **Nova Economia**, v. 28, n. 1, p. 181–212, 2018.

ALEXANDRE, L. A.; CAMPILHO, A. C.; KAMEL, M. On combining classifiers using sum and product rules. **Pattern Recognition Letters**, v. 22, n. 12, p. 1283–1289, 2001.

ANISH, C. M.; MAJHI, B.; MAJHI, R. Development and evaluation of novel forecasting adaptive ensemble model. **The Journal of Finance and Data Science**, v. 2, n. 3, p. 188–201, 2016.

ANSELIN, L. Under the hood Issues in the specification and interpretation of spatial regression models. **Agricultural Economics**, v. 27, p. 247–267, 2002.

ANSELIN, L. Spatial Externalities, Spatial Multipliers, And Spatial Econometrics. **International Regional Science Review**, v. 26, n. 2, p. 153–166, 2003.

ANTIPOV, E. A.; POKRYSHEVSKAYA, E. B. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. **Expert Systems with Applications**, v. 39, n. 2, p. 1772–1778, 2012.

ARRAES, R. A.; FILHO, E. de S. Externalidades e formação de preços no mercado imobiliário urbano brasileiro: um estudo de caso. **Economia Aplicada**, v. 12, n. 2, p. 289–319, 2008.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR 14.653-01:** Avaliação de bens. Parte 1: Procedimentos gerais. Rio de Janeiro, 2001.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR 14.653-2:** Avaliação de bens. Parte 2: Imóveis urbanos. Rio de Janeiro, 2011.

BARBOZA, A. da S. R.; SILVA, M. M. C. P.; SILVA, L. L. da; ARAÚJO JÚNIOR, J. C. de. A técnica da coordenação modular como ferramenta diretiva de projeto. **Ambiente Construído**, v. 11, n. 2, p. 97–109, 2011.

BRANDO, L.; BARBEDO, C. H. Há Fatores Não Econômicos na Formação do Preço de Imóveis? **Revista de Administração Contemporânea**, v. 20, n. 1, p. 106–130, 2016.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CAMPOS, R. B. A.; ALMEIDA, E. S. de. Decomposição Espacial nos Preços de Imóveis

Residenciais no Município de São Paulo. **Estudos Econômicos**, v. 48, n. 1, p. 5–38, 2018.

CAN, A. The Measurement of Neighborhood Dynamics in Urban House Prices. **Economic Geography**, v. 66, n. 3, p. 254–272, 1990.

ČEH, M.; KILIBARDA, M.; LISEC, A.; BAJAT, B. Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. **ISPRS International Journal of Geo-Information**, v. 7, n. 5, 2018.

CODES, B. N. **Avaliação dos Preços de Imóveis na Cidade de Fortaleza, com a Utilização de Redes Neurais Artificiais, para a Composição do ITBI**. Dissertação (Mestrado em Construção Civil) – Programa de Pós-Graduação em Engenharia Civil: Estruturas e Construção Civil da Universidade Federal do Ceará. Ceará, 2018.

CORAL, R.; FLESCH, C. A.; PENZ, C. A.; BORGES, M. R. Development of a Committee of Artificial Neural Networks for the Performance Testing of Compressors for Thermal Machines. **Metrology and Measurment Systems**, v. 22, n. 1, p. 79–88, 2015.

DEMŠAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. **Journal of Machine Learning Research**, v. 7, p. 1–30, 2006.

DUBIN, R. A. Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. **The Review of Economics and Statistics**, v. 70, n. 3, p. 466–474, 1988.

DUBIN, R. A. Spatial Autocorrelation: A Primer. **Journal of Housing Economics**, v. 7, n. 4, p. 304–327, 1998.

EFTHYMIOU, D.; ANTONIOU, C. Measuring the effects of transportation infrastructure location on real estate prices and rents: investigating the current impact of a planned metro line. **EURO Journal on Transportation and Logistics**, v. 3, n. 3, p. 179–204, 2015.

EXTERKATE, P.; GROENEN, P. J. F.; HEIJ, C.; VAN DIJK, D. Nonlinear forecasting with many predictors using kernel ridge regression. **International Journal of Forecasting**, v. 32, n. 3, p. 736–753, 2016.

FERNANDEZ, L.; MUKHERJEE, M.; SCOTT, T. The effect of conservation policy and varied open space on residential property values: A dynamic hedonic analysis. **Land Use Policy**, v. 73, n. December 2016, p. 480–487, 2018.

GADER, P. D.; MOHAMED, M. A.; KELLER, J. M. Fusion of handwritten word classifiers. **Pattern Recognition Letters**, v. 17, n. 6, p. 577–584, 1996.

GARCÍA, N.; GÁMEZ, M.; ALFARO, E. ANN+GIS: An automated system for property valuation. **Neurocomputing**, v. 71, n. 4–6, p. 733–742, 2008.

GLAESENER, M. L.; CARUSO, G. Neighborhood green and services diversity effects on land prices: Evidence from a multilevel hedonic analysis in Luxembourg. **Landscape and Urban Planning**, v. 143, p. 100–111, 2015.

GOETZ, J. N.; BRENNING, A.; PETSCHKO, H.; LEOPOLD, P. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. **Computers and Geosciences**, v. 81, p. 1–11, 2015.

GRANATYR, J. **Modelo afetivo de confiança e reputação utilizando personalidade e emoção**. Tese (Doutorado em Informática) - Programa de Pós-Graduação em Informática, Pontifícia Universidade Católica do Paraná., 2017.

GRANITTO, P. M.; VERDES, P. F.; CECCATTO, H. A. Neural network ensembles: Evaluation of aggregation algorithms. **Artificial Intelligence**, v. 163, n. 2, p. 139–162, 2005.

HAMIDI, S.; KITTRELL, K.; EWING, R. Value of transit as reflected in U.S. single-family home premiums: A meta-analysis. **Transportation Research Record**, v. 2543, n. 1, p. 108–115, 2016.

HE, C.; WANG, Z.; GUO, H.; SHENG, H.; ZHOU, R.; YANG, Y. Driving forces analysis for residential housing price in Beijing. **Procedia Environmental Sciences**, v. 2, n. 5, p. 925–936, 2010.

HESS, D. B.; ALMEIDA, T. M. Impact of proximity to light rail rapid transit on station-area property values in Buffalo, New York. **Urban Studies**, v. 44, n. 5–6, p. 1041–1068, 2007.

IBGE. **Pesquina Anual da Indústria da Construção, 2018**. Rio de Janeiro: IBGE, 2020. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2017**. Rio de Janeiro: IBGE, 2019. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2016**. Rio de Janeiro: IBGE, 2018. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2015**. Rio de Janeiro: IBGE, 2017. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2014**. Rio de Janeiro: IBGE, 2016. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2013**. Rio de Janeiro: IBGE, 2015. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2012**. Rio de Janeiro: IBGE, 2014. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2011**. Rio de Janeiro: IBGE, 2013. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2010**. Rio de Janeiro: IBGE, 2012. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2009**. Rio de Janeiro: IBGE, 2011. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2008**. Rio de Janeiro: IBGE, 2010. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **Pesquina Anual da Indústria da Construção, 2007**. Rio de Janeiro: IBGE, 2009. Disponível em: <https://biblioteca.ibge.gov.br/ >. Acesso em: 28 de novembro de 2020.

IBGE. **População em áreas de risco no Brasil**. [s.l: s.n.].

ISHIJIMA, H.; MAEDA, A. Real Estate Pricing Models: Theory, Evidence, and Implementation. **Asia-Pacific Financial Markets**, v. 22, n. 4, p. 369–396, 2015.

JANSSEN, C.; SÖDERBERG, B.; ZHOU, J. Robust estimation of hedonic models of price and income for investment property. **Journal of Property Investment & Finance**, v. 19, n. 4, p. 342–360, 2001.

KEMPA, O.; LASOTA, T.; TELEC, Z.; TRAWIŃSKI, B. Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 6592 LNAI, n. PART 2, p. 323–332, 2011.

KETTANI, O.; ORAL, M. Designing and implementing a real estate appraisal system: The case of Québec Province, Canada. **Socio-Economic Planning Sciences**, v. 49, p. 1–9, 2015.

KITTLER, J.; HATEF, M.; DUIN, R. P. W.; MATAS, J. On Combining Classifiers. **IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,** v. 20, n. 3, p. 226–239, 1998.

KONTRIMAS, V.; VERIKAS, A. The mass appraisal of the real estate by computational intelligence. **Applied Soft Computing Journal**, v. 11, p. 443–448, 2011.

KOSTOV, P. Model boosting for spatial weighting matrix selection in spatial lag models. **Environment and Planning B: Urban Analytics and City Science**, v. 37, n. 3, p. 533–549, 2010.

LAM, K. C.; YU, C. Y.; LAM, C. K. Support vector machine and entropy based decision support system for property valuation. **Journal of Property Research**, v. 26, n. 3, p. 213–233, 2009.

LANCASTER, K. J. . A New Approach to Consumer Theory. **Journal of Political Economy**, v. 74, n. 2, p. 132–157, 1966.

LASOTA, T.; ŁUCZAK, T.; NIEMCZYK, M.; OLSZEWSKI, M.; TRAWIŃSKI, B. Investigation of property valuation models based on decision tree ensembles built over noised data. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 8083 LNAI, p. 417–426, 2013.

LIU, C.-L. Classifier combination based on confidence transformation. **Pattern Recognition**, v. 38, n. 1, p. 11–28, 2005.

LOCATELLI, R. L.; INÊS, H. M.; LARA, J. E.; NOGUEIRA, F. T. P. Real estate market of a brazilian metropolis: sustained growth or speculative bubble? **Revista de Administração Mackenzie**, v. 18, n. 2, p. 211–236, 2017.

MABU, S.; OBAYASHI, M.; KUREMOTO, T. Ensemble learning of rule-based evolutionary algorithm using multi-layer perceptron for supporting decisions in stock trading problems. **Applied Soft Computing Journal**, v. 36, p. 357–367, 2015.

MAO, Y.; WU, W. Fuzzy Real Option Evaluation of Real Estate Project Based on Risk Analysis. **Systems Engineering Procedia**, v. 1, n. 1, p. 228–235, 2011.

MCALLISTER, P.; ZHENG, H.; BOND, R.; MOORHEAD, A. Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets. **Computers in Biology and Medicine**, v. 95, p. 217–233, 2018.

MCCLUSKEY, W. J.; MCCORD, M.; DAVIS, P. T.; HARAN, M.; MCILHATTON, D. Prediction accuracy in mass appraisal : a comparison of modern approaches. **Journal of Property Research**, v. 30, n. 4, p. 239–265, 2013.

MEYER, P.; NOBLET, V.; MAZZARA, C.; LALLEMENT, A. Survey on deep learning for radiotherapy. **Computers in Biology and Medicine**, v. 98, p. 126–146, 2018.

MINLI, Z.; WENPO, Y. Fuzzy Comprehensive Evaluation Method Applied in the Real Estate Investment Risks Research. **Physics Procedia**, v. 24, p. 1815–1821, 2012.

NEMMOUR, H.; CHIBANI, Y. Neural Network Combination by Fuzzy Integral for Robust Change Detection in Remotely Sensed Imagery. **EURASIP Journal on Applied Signal Processing**, v. 14, p. 2187–2195, 2005.

NGUYEN, N.; CRIPPS, A. Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. **Journal of Real Estate Research**, v. 22, n. 3, p. 313–336, 2001.

NUNES, D. B. **Proposição de um Modelo de Regressão Linear para Avaliação do Valor de**

Mercado de Apartamentos Residenciais. Dissertação (Mestrado em Construção Civil) – Programa de Pós-Graduação em Engenharia Civil: Estruturas e Construção Civil da Universidade Federal do Ceará. Ceará, 2016.

PEREIRA, J. M.; BASTO, M.; SILVA, A. F. da. The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. **Procedia Economics and Finance**, v. 39, n. November 2015, p. 634–641, 2016.

PETERSON, S.; FLANAGAN, A. B. Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. **Journal of Real Estate Research**, v. 31, n. 2, p. 147–164, 2009.

PONTES, E.; PAIXÃO, L. A.; ABRAMO, P. O mercado imobiliário como revelador das preferências pelos atributos espaciais: uma análise do impacto da criminalidade urbana no preço de apartamentos em Belo Horizonte. **Revista de Economia Contemporânea**, v. 15, n. 1, p. 171–197, 2011.

QU, X.; LEE, L. F. Estimating a spatial autoregressive model with an endogenous spatial weight matrix. **Journal of Econometrics**, v. 184, n. 2, p. 209–232, 2015.

RINCKE, J. A commuting-based refinement of the contiguity matrix for spatial models, and an application to local police expenditures. **Regional Science and Urban Economics**, v. 40, n. 5, p. 324–330, 2010.

ROSEN, S. Hedonic Prices and Implicit Markets : Product Differentiation in Pure Competition. **The Journal of Political Economy**, v. 82, n. 1, p. 34–55, 1974.

RUDŽIANSKAITE-KVARACIEJIENE, R.; APANAVIČIENE, R.; GELŽINIS, A. Modelling the effectiveness of PPP road infrastructure projects by applying random forests. **Journal of Civil Engineering and Management**, v. 21, n. 3, p. 290–299, 2015.

SEFIN, S. M. de F. **Relatório Contabil de Propósito Geral - Prestação de Contas do Governo**. Disponível em: <https://www.sefin.fortaleza.ce.gov.br/contas-publicas/balanco-geral>. Acesso em: 27 ago. 2018.

SEYA, H.; YAMAGATA, Y.; TSUTSUMI, M. Automatic selection of a spatial weight matrix in spatial econometrics: Application to a spatial hedonic approach. **Regional Science and Urban Economics**, v. 43, n. 3, p. 429–444, 2013.

SHIN, K.; WASHINGTON, S.; CHOI, K. Effects of transportation accessibility on residential property values application of spatial hedonic price model in Seoul, South Korea, Metropolitan Area. **Transportation Research Record**, v. 1994, n. 1, p. 66–73, 2007.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B**, v. 63, n. 2, p. 411–

423, 2001.

UBERTI, M. S.; ANTUNES, M. A. H.; DEBIASI, P.; TASSINARI, W. Mass appraisal of farmland using classical econometrics and spatial modeling. **Land Use Policy**, v. 72, p. 161–170, 2018.

WANG, Y.; POTOGLOU, D.; ORFORD, S.; GONG, Y. Bus stop, property price and land value tax: A multilevel hedonic analysis with quantile calibration. **Land Use Policy**, v. 42, p. 381–391, 2015.

WEN, H.; BU, X.; QIN, Z. Spatial effect of lake landscape on housing price: A case study of the West Lake in Hangzhou, China. **Habitat International**, v. 44, p. 31–40, 2014.

WIŚNIEWSKI, R. Modeling of residential property prices index using committees of artificial neural networks for PIGS, the European-G8, and Poland. **Argumenta Oeconomica**, v. 38, n. 1, p. 145–194, 2017.

WORZALA, E.; LENK, M.; SILVA, A. An Exploration of Neural Networks and Its Application to Real Estate Valuation. **Journal of Real Estate Research**, n. September 1994, p. 185–201, 1995.

YEH, I.; HSU, T.; WEIGHT, C. Building real estate valuation models with comparative approach through case-based reasoning. **Applied Soft Computing Journal**, v. 65, p. 260–271, 2018.

YIJIAN, S. U. N.; RUFU, H. Fuzzy Set-Based Risk Evaluation Model for Real Estate Projects. **Tsinghua Science and Technology**, v. 13, n. October, p. 158–164, 2008.

YOU, Q.; PANG, R.; CAO, L.; LUO, J. Image Based Appraisal of Real Estate Properties. **IEEE Transactions on Multimedia**, v. 19, n. 12, p. 2751–2759, 2017.

ZHANG, R.; DU, Q.; GENG, J.; LIU, B.; HUANG, Y. An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study. **Habitat International**, v. 46, p. 196–205, 2015.

ZOPPI, C.; ARGIOLAS, M.; LAI, S. Factors influencing the value of houses: Estimates for the city of Cagliari, Italy. **Land Use Policy**, v. 42, p. 367–380, 2015.

ZURADA, J.; LEVITAN, A. S.; GUAN, J. A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. **Journal of Real Estate Research**, v. 33, n. 3, p. 349–387, 2011.

7 ANEXO A – MATRIZ DE CORRELAÇÃO DE VARIÁVEIS