AVALIAÇÃO DOS IMPACTOS E VALORAÇÃO DOS DANOS SOCIOECONÔMICOS CAUSADOS PARA AS COMUNIDADES ATINGIDAS PELO ROMPIMENTO DA BARRAGEM DE FUNDÃO

Valor do Estoque Habitacional por Preços Hedônicos em Barra Longa

FGV

DEZEMBRO DE 2019

Dados Internacionais de Catalogação na Publicação (CIP)

Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Fundação Getulio Vargas

Valor do Estoque Habitacional por Preços Hedônicos em Barra Longa / Fundação Getulio Vargas. – Rio de Janeiro; São Paulo : FGV, 2019.

86 p.

Em colaboração com: André Portela Fernandes de Souza, Daniela Stucchi da Silva Maeji, Lucas Gerez Foratto, Paulo Picchetti, Reynaldo Fernandes, Ronan Cunha.

Acima do título: Projeto Rio Doce - Avaliação dos Impactos e Valoração dos Danos Socioeconômicos Causados para as Comunidades Atingidas pelo Rompimento da Barragem de Fundão.

Inclui bibliografia.

1. Projeto Rio Doce. 2. Fundão, Barragem de (MG). 3. Barragens e açudes – Aspectos sociais – Barra Longa (MG). 4. Barragens e açudes – Aspectos sociais – Mariana (MG). 5. Habitação – Aspectos econômicos. I. Título.

CDD – 627.8

EQUIPE TÉCNICA

André Portela Fernandes de Souza

Daniela Stucchi da Silva Maeji

Lucas Gerez Foratto

Paulo Picchetti

Reynaldo Fernandes

Ronan Cunha

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

SUMÁRIO EXECUTIVO

O presente trabalho apresenta um modelo de previsão por preços hedônicos do valor das propriedades residenciais, a partir dos dados de aluguel do Censo Demográfico de 2010, para uma amostra selecionada de imóveis de Barra Longa (MG), em momento no tempo anterior ao rompimento da Barragem de Fundão, outubro de 2015. A partir do modelo de preços hedônicos, estimam-se tanto o valor agregado, antes do rompimento, do estoque habitacional atingido pela lama de rejeito e/ou pelas obras de limpeza, como os valores esperados de algumas residências típicas ou representativas da região.

A mensuração do estoque de riqueza afetado é parte do diagnóstico socioeconômico e o presente estudo avança nessa direção ao mensurar a extensão do estoque habitacional potencialmente afetado pelo rompimento da Barragem de Fundão. A razão para essa escolha inicial é porque o imóvel residencial se constitui no principal ativo físico que em geral as pessoas possuem ao longo da vida. E dadas as restrições atuais de disponibilidade e/ou acesso a dados adicionais do setor imobiliário para a região atingida, o presente exercício está circunscrito às propriedades residenciais cadastradas pela Fundação Renova em Barra Longa.

A propriedade residencial pode ser usada adicionalmente como local para o desenvolvimento de pequenos negócios familiares. Portanto, o dano à propriedade residencial em decorrência de desastres pode representar uma perda dupla às populações mais pobres, pois envolve não somente a perda de abrigo, mas também a perda de renda. Contudo, considerando somente a perda associada a danos ou destruição do estoque de unidades habitacionais em decorrência do desastre, essa perda constitui-se da interrupção, temporária ou permanente, dos serviços de acomodação. A interrupção desses serviços corresponde ao valor dos aluguéis, pagos ou imputados, àquelas habitações danificadas ou destruídas.

Os valores dos imóveis são modelados a partir dos modelos de preços hedônicos comumente utilizados na literatura especializada. Os modelos de preços hedônicos são adequados nesse contexto pois eles consideram bens diferenciados que têm múltiplas características e que o valor de um bem diferenciado é dado pelos valores dessas características. Esse é precisamente o caso de imóveis habitacionais. Além disso, o imóvel é considerado um ativo econômico e seu valor é dado pelo valor presente do fluxo de seus rendimentos futuros, a saber, os aluguéis potencialmente gerados.

As estimativas empíricas são feitas a partir do modelo estatístico/econométrico de florestas aleatórias. Para tanto, são utilizadas as informações de aluguéis de imóveis

residenciais disponíveis para a proximidade de Barra Longa. As fontes de informações disponíveis são o Censo Demográfico de 2010 e o Cadastro da Fundação Renova. Todos os resultados são em reais de outubro de 2015[1]. São apresentados intervalos de valores para diferentes cenários de taxa desconto de 3%, 4%, 5% e 6% ao ano.

Existem no Cadastro da Fundação Renova 291 residências em Barra Longa declaradas atingidas com informações das variáveis utilizadas nos modelos de preços hedônicos. Para estas, utilizando os dados do Censo Demográfico, estimam-se valores agregados médios e intervalos de valores agregados dos níveis de confiança de 95%. A tabela seguinte apresenta os resultados. Para esse estoque habitacional estima-se um valor médio de cerca de R$ 54 milhões de reais com a taxa de desconto de 3% e de cerca de R$ 35 milhões com a taxa de desconto de 5%. Obviamente, são valores sujeitos às incertezas estatísticas e os valores dos intervalos de confiança inferior e superior são apresentados nos segundo e terceiro painéis da tabela seguinte, respectivamente.

Por fim, é importante salientar que os valores estimados correspondem a valores de mercado. Os valores subjetivos que os proprietários atribuem aos seus imóveis podem ser maiores que os valores de mercado (caso contrário, poderiam ter vendido seu imóvel) e, portanto, a perda de bem-estar que um proprietário ou uma proprietária tem com a perda do imóvel é maior que o valor de mercado do mesmo[2].

[1]Importa destacar que o ano utilizado como base (2015) não se confunde com a data de referência para o cálculo do valor da moradia para fins de reparação. Isso, pois diversas diretrizes internacionais de deslocamento compulsório, além da própria jurisprudência dos Tribunais Superiores, determinam como data de referência para o cálculo indenizatório o momento do efetivo do pagamento e não a data do deslocamento compulsório, de modo a impedir a subvaloração do quantum indenizatório.

[2]Muito embora este não seja o escopo do presente trabalho, vale destacar que, dentro da ótica da reparação integral, o cálculo do valor indenizatório com base deve considerar não só o "valor de mercado", mas também as ações e valores capazes de restaurar o modo de vida das populações atingidas. Nesse sentido, outros critérios e considerações devem ser consideradas para fins de reparação, como é o caso de dimensões relevantes ao modo de vida que não estão valoradas nos preços de mercado dos imóveis. Para uma análise da perspectiva de direitos humanos sobre o direito à moradia adequada, ver o produto Parâmetros para reparação do direito à moradia no contexto do rompimento da Barragem de Fundão da FGV (2019).

Tabela 1 — Valores totais estimados do estoque habitacional do Cadastro em Barra Longa em R$ milhões (preços de out./2015)

Tipos de residências consideradas em Barra Longa	Total de residências consideradas em Barra Longa	Valor total do estoque habitacional — taxa de desconto de 3% a.a.	Valor total do estoque habitacional — taxa de desconto de 4% a.a.	Valor total do estoque habitacional — taxa de desconto de 5% a.a.	Valor total do estoque habitacional — taxa de desconto de 6% a.a.
			Valor Médio		
Residências cadastradas	471	79,35	61,99	50,94	43,29
Residências atingidas	291	54,36	42,47	34,90	29,66
Residências afetação total	69	13,61	10,64	8,74	7,43
Residências afetação parcial	23	3,59	2,81	2,31	1,96
			Valor Limite Inferior		
Residências cadastradas	471	61,46	48,01	39,46	33,53
Residências atingidas	291	41,47	32,40	26,63	22,63
Residências afetação total	69	10,13	7,92	6,51	5,53
Residências afetação parcial	23	2,88	2,25	1,85	1,57
			Valor Limite Superior		
Residências cadastradas	471	97,24	75,97	62,43	53,06
Residências atingidas	291	67,25	52,54	43,17	36,69
Residências afetação total	69	17,10	13,36	10,98	9,33
Residências afetação parcial	23	4,31	3,37	2,77	2,35

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.

1 INTRODUÇÃO

Para populações de baixa ou média renda, a moradia constitui, em geral, o principal ativo da família. A propriedade residencial pode ser usada adicionalmente como local para o desenvolvimento de pequenos negócios familiares. Portanto, o dano à propriedade residencial em decorrência de desastres[3] pode representar uma perda dupla às populações mais pobres[4], pois envolve não somente a perda de abrigo, mas também a perda de renda (BANCO MUNDIAL, 2012a).

O setor habitacional inclui, além da propriedade residencial, outros elementos de infraestrutura e equipamentos públicos, tais como prestação de serviços de utilidade básica: água, saneamento, eletricidade, entre outros (CEPAL, 2014). No entanto, para que não ocorra dupla contagem na agregação de resultados setoriais, o valor agregado dos danos em habitação deve incluir somente aqueles relacionados com o acesso ao abastecimento de água, saneamento e eletricidade, entre outras utilidades, das unidades habitacionais. Isto é, os danos aos ativos nos sistemas de água e saneamento e na distribuição de energia, em geral, devem ser contabilizados sob os setores de água, saneamento e eletricidade, respectivamente (BANCO MUNDIAL, 2012b).

Portanto, considerando somente a perda associada a danos ou destruição do estoque de unidades habitacionais em decorrência do desastre, essa perda constitui-se da interrupção, temporária ou permanente, dos serviços de acomodação. A interrupção desses serviços corresponde ao valor dos aluguéis, pagos ou imputados, àquelas habitações danificadas ou destruídas (CEPAL, 2014).

Assim, para além das necessárias diretrizes e parâmetros para reparação de danos às moradias afetadas pelo desastre, interessa investigar a magnitude do impacto do rompimento da Barragem de Fundão no valor agregado do estoque habitacional como relevante componente setorial do impacto do desastre no estoque de capital/riqueza do

[3] Entendendo desastre como uma "(…) ruptura grave do funcionamento de uma comunidade ou de uma sociedade em qualquer escala devido a eventos perigosos interagindo com condições de exposição, vulnerabilidade e capacidade, levando a uma ou mais das seguintes consequências: impactos e perdas humanas, materiais, econômicas ou ambientais" (ASSEMBLY, 2016).

[4] Muito embora este não seja o escopo do presente trabalho, vale destacar que, dentro da ótica da reparação integral, o cálculo do valor indenizatório com base deve considerar não só o "valor de mercado", mas também as ações e valores capazes de restaurar o modo de vida das populações atingidas. Nesse sentido, outros critérios e considerações devem ser consideradas para fins de reparação, como é o caso de dimensões relevantes ao modo de vida que não estão valoradas nos preços de mercado dos imóveis. Para uma análise da perspectiva de direitos humanos sobre o direito à moradia adequada, ver o produto Parâmetros para reparação do direito à moradia no contexto do rompimento da Barragem de Fundão da FGV (2019).

território atingido. Para tal, o primeiro passo é estimar o valor do estoque habitacional no momento anterior ao rompimento da Barragem de Fundão, chamado aqui de valor na linha de base[5]. Isto é, qual o valor agregado das unidades habitacionais impactadas pela lama de rejeito imediatamente antes do rompimento da Barragem de Fundão?

Posteriormente, sujeito à disponibilidade de dados, pode-se mensurar o impacto do desastre sobre o setor habitacional, considerando a variação no valor médio das propriedades residenciais. Essa variação é composta pela perda ou destruição do valor das moradias atingidas[6] pela lama de rejeito, e/ou afetadas pelas obras de limpeza, e pela potencial valorização/desvalorização decorrente dos efeitos do desastre no mercado imobiliário.

De forma a recuperar o valor do estoque habitacional na linha de base, o objetivo do presente trabalho é desenvolver um modelo de previsão por preços hedônicos do valor dos imóveis, a partir dos dados de aluguel do Censo Demográfico 2010 para uma amostra selecionada, em momento no tempo anterior ao rompimento, outubro de 2015[7]. A partir do modelo de preços hedônicos para os imóveis da amostra, é possível estimar o valor agregado, antes do rompimento, do estoque habitacional atingido pela lama de rejeito e/ou pelas obras de limpeza. Contudo, dadas as restrições atuais de disponibilidade e/ou acesso a dados adicionais do setor imobiliário para a região atingida, o presente exercício será aplicado somente às propriedades residenciais cadastradas pela Fundação Renova em Barra Longa.

A mensuração dos valores de mercado de imóveis residenciais leva em consideração os preços intrínsecos das características dessas propriedades. O preço do imóvel será a soma dos preços de todas as características que proporcionam satisfação e são determinantes nas escolhas das pessoas. Imóveis, por sua vez, podem gerar fluxos de aluguéis ao longo do tempo. Assim, um imóvel é um ativo econômico e o valor de um ativo é determinado pelo valor presente do fluxo de rendimentos futuros[8]. Sob o ponto

[5] Estimar o valor da linha de base não implica posicionamento da FGV sobre valores de indenização destes imóveis no tempo presente. Para tanto, devem ser consideradas questões legais e da jurisprudência brasileira sobre o momento de cálculo destes valores.
[6] Para mais detalhes sobre o processo de cadastramento dos atingidos, ver Produto “Análise do Cadastro Socioeconômico” da FGV.
[7] Importa destacar que o ano utilizado como base (2015) não se confunde com a data de referência para o cálculo do valor da moradia para fins de reparação. Isso, pois diversas diretrizes internacionais de deslocamento compulsório, além da própria jurisprudência dos Tribunais Superiores, determinam como data de referência para o cálculo indenizatório o momento do efetivo do pagamento e não a data do deslocamento compulsório, de modo a impedir a subvaloração do quantum indenizatório.
[8] Nessa perspectiva, a aquisição de uma residência pode ser comparada a outros projetos de investimento, como a aplicação na poupança ou fundo de investimento. Nesse contexto, é preciso considerar suas rentabilidades, taxas de depreciação e possível

de vista de ativos, os imóveis residenciais pertencem à classe de bens de longo prazo, portanto, a taxa de desconto utilizada como critério de decisão e valoração deve refletir as perspectivas de longo prazo.

Esse argumento é importante de ser ressaltado, pois a escolha da taxa de desconto a ser utilizada é fundamental no cálculo final do valor do imóvel; quanto menor a taxa de desconto, maior será seu valor estimado. Posto isso, nas estimativas para os valores agregados do estoque habitacional desse estudo são apresentados cenários com taxas de descontos de 3%, 4%, 5% e 6% ao ano. Esses valores representam diferentes cenários para rentabilidades financeiras do ativo residencial, incorporando sua depreciação e possíveis perspectivas de valorização ou desvalorização. Além disso, também são apresentados os valores médios de algumas residências típicas ou representativas da região.

Este estudo está organizado em quatro seções, além desta introdução. Na seção 2, discutem-se os principais aspectos metodológicos relacionados com os componentes que determinam os valores dos domicílios por preços hedônicos, assim como as diferentes ferramentas para as estimativas empíricas dos valores agregados das propriedades residenciais. Na seção 3, apresentam-se as duas bases de dados utilizadas, os dados de domicílios do Censo Demográfico 2010 e os dados das propriedades cadastradas pela Fundação Renova. Na seção 4, mostra-se o embasamento empírico para o critério de escolha do melhor modelo de predição. Na seção 5, apresentam-se os resultados das estimativas dos valores agregados do estoque habitacional e dos valores médios de algumas residências típicas ou representativas da região. Por fim, a seção 6 conclui esse estudo.

1.1 Contexto

Propriedades residenciais em subdistritos do município de Mariana (MG) e de Barra Longa (MG) foram severamente afetadas pelo rompimento da Barragem de Fundão. Entre essas localidades, Bento Rodrigues e Paracatu de Baixo, em Mariana, e Gesteira, em Barra Longa, tiveram a maioria de suas unidades habitacionais completamente destruída pela lama de rejeito. Somam-se, a esses casos, outros de deterioração nas estruturas físicas das residências, em alegada decorrência não somente da passagem da lama, mas das obras de limpeza e reconstrução que se seguiram ao rompimento.

valorização/desvalorização causadas por diferentes motivos. Para a comparação e análise de viabilidade desses investimentos, é necessária uma taxa de desconto que leve essas dimensões em consideração.

No cadastramento realizado em 2016 pela Fundação Renova junto às famílias em Mariana[9], naquele momento consideradas elegíveis ao reassentamento, havia 269 moradias cadastradas em Bento Rodrigues, 143 em Paracatu de Baixo e outras 32 em localidades dispersas pela área rural do município de Mariana[10]. Contudo, no cadastro das propriedades residências atingidas em Mariana não constam as informações completas sobre as características dessas moradias, como é o caso para o cadastro completo da Fundação Renova para os demais municípios com propriedades cadastradas.

Em contraposição, muitas famílias, e suas respectivas propriedades, foram cadastradas nas campanhas de cadastramento da Fundação Renova, além da fase emergencial. A Figura 1 apresenta as localizações das propriedades cadastradas pela Fundação Renova no município de Barra Longa, totalizando 756 propriedades cadastradas[11], para as quais constam informações sobre as características do imóvel. Entre as moradias classificadas como totalmente destruídas pela Fundação Renova em Barra Longa, tem-se 37 famílias previstas para serem reassentadas no reassentamento coletivo de Gesteira.

À época do rompimento, Barra Longa era marcada por uma dinâmica rural, com 62,3% da população residente em zonas rurais. Apresentava também uma elevada taxa de pessoas em situação de pobreza[12], estando 25,4% dos munícipes nessa condição, taxa essa maior do que a média do estado de Minas Gerais, de 11,0%, e do Brasil, de 15,2%, no mesmo ano de comparação (IBGE, 2015). Não obstante o impacto nos distritos de Gesteira, Barretos, entre outras comunidades rurais, o município de Barra Longa foi o único atingido em seu centro urbano pela enxurrada de rejeito proveniente do rompimento da Barragem de Fundão.

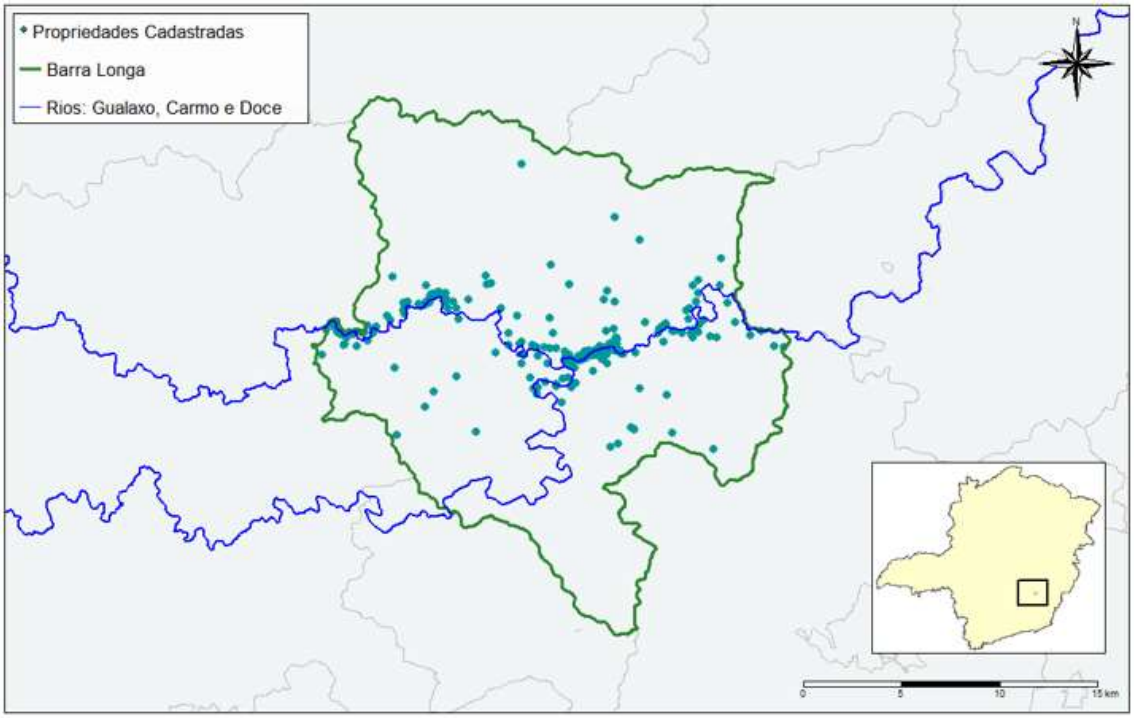
[9] A base do Cadastro foi obtida por meio do download do filtro 1322, que contém as informações levantadas no questionário socioeconômico aplicado durante o cadastramento, referentes às propriedades e aos indivíduos. A extração dos dados do sistema foi realizada no dia 28 de agosto de 2019.

[10] Para uma discussão detalhada sobre o cadastro realizado pela Fundação Renova, ver o produto "Analise do Cadastro Socioeconômico" da FGV.

[11] A base do Cadastro da Fundação Renova foi obtida por meio do download do filtro 1322, que contém as informações levantadas no questionário socioeconômico aplicado durante o cadastramento, referentes às propriedades e aos indivíduos. A extração dos dados do sistema foi realizada no dia 28 de agosto de 2019.

[12] Consideraram-se como em situação de pobreza os indivíduos com renda domiciliar *per capita* igual ou inferior a R$ 140,00 mensais, em valores de agosto de 2010, nos dados do Censo de 2010 (IBGE, 2015). O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.

Figura 1 — Propriedades cadastradas pela Fundação Renova em Barra Longa



Fonte: Elaboração própria (2019), a partir de dados do Cadastro da Fundação Renova acessados em 28/08/2019.

2 METODOLOGIA

Nesta seção, discutem-se os principais aspectos metodológicos relacionados com os componentes que determinam os valores dos domicílios por preços hedônicos, assim como as diferentes ferramentas para as estimações empíricas dos valores agregados das propriedades residenciais cadastradas pela Fundação Renova.

A subseção 2.1 apresenta a abordagem teórica dos preços hedônicos na literatura econômica, partindo da existência de um mercado de imóveis e como suas características podem afetar seus preços. A subseção 2.2 contém as informações dos modelos de Mínimos Quadrados Ordinários em sua forma mais geral de estimação. Na subseção 2.3 apresentam-se as implicações dos Modelos Hierárquicos e, por fim, na subseção 2.4 é introduzida a modelagem de Florestas Aleatórias que será a principal estratégia empírica para se obter os valores preditos dos aluguéis, base para o cálculo dos valores das propriedades residenciais.

2.1 Preços Hedônicos de imóveis

A literatura de Preços Hedônicos para imóveis ganhou força em meados da década de 1970, com contribuições importantes na concepção teórica e nos modelos econométricos (GOODMAN, 1978; HARRISON JR. e RUBINFELD, 1978; ROSEN, 1974). A mensuração dos valores de mercado de imóveis deve levar em consideração fatores intrínsecos dos imóveis, como: número de cômodos, banheiros, dormitórios, entre outros. Somam-se a isso fatores temporais e amenidades locais, como o conjunto de vizinhança presente nos arredores, infraestrutura local, pavimentação, entre outros (HARRISON JR. e RUBINFELD, 1978).

O embasamento teórico é definido a partir de modelos econômicos de equilíbrio geral de bens diferenciados, de modo que os preços observados de diversos bens assim como suas características a eles associados definem um conjunto de preços implícitos, comumente denominados como hedônicos (ROSEN, 1974).

Sob a hipótese de que a diferenciação de produtos é baseada no nível de utilidade[13] ou satisfação proporcionado às pessoas envolvidas, os preços hedônicos podem ser denominados como os preços implícitos de cada característica do produto. No caso dos imóveis, é como uma determinada característica, por exemplo, um dormitório, é valorada pelos indivíduos, de acordo com o grau de utilidade ou satisfação que essa

[13] Esse termo é comumente utilizado em economia e sua definição se aproxima do nível de satisfação ou felicidade pela obtenção e uso de bens.

característica do imóvel pode proporcionar. O preço do imóvel será, portanto, a soma dos preços de todas as características envolvidas na escolha dos indivíduos.

Em um contexto de mercado competitivo, o equilíbrio de mercado corresponde a um plano com diversas dimensões onde ofertantes e demandantes alocam seus recursos. Considere um imóvel z, representado pelo conjunto de suas características em geral classificadas em três grupos: estrutura (número de cômodos, banheiros, dormitórios etc.); localização (local urbano, proximidade a serviços etc.); e ambiente (vista ao campo, iluminação etc.). Assim, $z = (z_1, z_2, \dots, z_n)$, sendo z_1 a primeira característica do imóvel em questão, z_2 a segunda característica, e assim por diante. O vetor z descreve todos os serviços providos pelo imóvel. Em um mercado competitivo, com variabilidade suficiente nas características dos imóveis, o preço do imóvel é $p(z) = p(z_1, z_2, \dots, z_n)$, definido no espaço n-dimensional das suas características.

De maneira simplificada, a escolha de $p(z)$ se dá a partir da maximização conjunta de todos as características presentes no imóvel, dada uma restrição orçamentária, ou seja, seleciona-se o nível de cada característica z_j de forma a dar máxima satisfação a cada pessoa. Pelas condições de equilíbrio de mercado, deve haver um balanço entre o saldo da demanda, o quanto os consumidores almejam obter imóveis com determinadas características, e o lado da oferta, isto é, o estoque de habitações disponíveis com as caraterísticas desejadas. Dessa forma, analisam-se as escolhas ótimas de ambos os lados.

Formalmente, do ponto de vista do consumidor, a modelagem de maximização da utilidade de um indivíduo representativo será: seja $U(x, z_1, z_2, \dots, z_n)$ estritamente côncava, e x é o conjunto de todos os demais bens consumidos. Para simplificar a álgebra, suponha que os preços dos demais bens x seja padronizado para um, de modo que a restrição orçamentária do indivíduos será: $y = x + p(z)$, onde x é uma cesta de outros bens, como alimentação, roupas etc., e o mercado é competitivo.

Pelas condições de primeira ordem, tem-se para $i = \{1, \dots, n\}$:

$$\frac{\partial p}{\partial z_i} = p_i'(z_i) = \frac{U_{z_i}}{U_x} \tag{1}$$

onde $\frac{\partial p}{\partial z_i}$ representa a primeira derivada do preço com relação à característica z_i, e pode ser interpretado como o quanto o preço varia se alterada uma unidade de z_i. U_{z_i} é a

derivada da função de utilidade com relação a z_i e U_x é a derivada com relação à cesta de bens x. Ambas as derivadas têm interpretação semelhante àquela feita para $\frac{\partial p}{\partial z_i}$.

Sob a ótica da produção ou lado da oferta, por sua vez, suponha um conjunto de custos para produzir e manter um imóvel denominado como $C(M, z; \beta)$ derivado da minimização dos custos dado um nível de produção M, o parâmetro β reflete os resultados de minimização dos custos. Adicionalmente, suponha que C seja uma função convexa. Então sendo a ótica da firma maximizar o lucro $\pi = Mp(z) - C(M, z_1, z_2, \dots, z_n)$ escolhendo M e z de maneira ótima, teremos z dado pela função implícita da função de preços hedônicos $p(z)$. A escolha ótima de M e z requer que:

$$\frac{\partial p}{\partial z_i} = p_i'(z) = \frac{C_{z_i}(M, z_1, z_2, \dots, z_n)}{M} \tag{2}$$

$$p_i(z) = C_M(M, z_1, z_2, \dots, z_n) \tag{3}$$

onde $C_M(.)$ e $C_{z_i}(.)$ são as derivadas de C com relação a M e z_i, respectivamente. Dessa forma, os preços são determinados resolvendo o sistema formado pelas equações 1, 2 e 3. Entende-se os valores que solucionam esse sistema como aqueles que equilibram o lado da demanda, cujo preço dos imóveis representa o equilíbrio entre o benefício que cada característica z_j traz para os consumidores e o lado da oferta por meio do custo para os ofertantes em oferecer a mesma característica.

Esse modelo básico é a base para trabalhos empíricos que buscam estimar o valor de imóveis e outros ativos cujo preço pode ser determinado pelo seu conjunto de características intrínsecas. Na literatura nacional, grande parte dos trabalhos tem como enfoque mercados imobiliários locais (DANTAS e CORDEIRO, 1988; GONZÁLEZ e FORMOSO, 1994; PAIXÃO, 2009).

É importante destacar que o modelo apresentado pode ser interpretado como a variável $p(z)$ representando o valor do fluxo de serviços providos pelo imóvel em um determinado intervalo de tempo. Ou seja, $p(z)$ é o valor do aluguel do imóvel. Por outro lado, precisamente por poder gerar um fluxo de aluguéis ao longo do tempo, um imóvel é um ativo econômico. E o valor de um ativo é determinado pelo valor presente do fluxo de rendimento futuros.

Dessa forma, considerando que a moradia é um ativo e pagaria fluxos anuais de aluguel caso fosse alugada (BELLMAN e KUSANAGI, 1973), o valor do imóvel K pode ser calculado por meio da seguinte equação:

$$(Kr = A - D + \Delta K,) \tag{4}$$

onde r é a taxa de juro real, A o valor anualizado dos aluguéis, D é o valor anualizado da depreciação e ΔK a variação esperada no valor do imóvel entre o período atual e o subsequente. Essa equação supõe que a moradia pode ser valorada como qualquer outro ativo que gere um fluxo de benefícios (A).

Admitindo-se que as hipóteses: (i) a depreciação ocorre a uma taxa constante do valor do imóvel, $D = gK$, onde g é a taxa de depreciação do imóvel; e (ii) a valorização ocorre a uma taxa constante do valor do imóvel, $\Delta K = cK$, onde c é a taxa de valorização do imóvel, são aproximações razoáveis para trajetórias de longo prazo, obtém-se o valor do imóvel dado pela equação (5):

$$K = \frac{A}{r + g - c} \tag{5}$$

Sob o ponto de vista de ativos, os imóveis pertencem à classe de bens de longo prazo, portanto, a taxa de juros utilizada como critério de decisão e valoração deve refletir as perspectivas de longo prazo. Esse argumento é importante de ser ressaltado, pois a escolha da taxa de desconto $(r + g - c)$ a ser utilizada é fundamental no cálculo final do valor do imóvel, pois quanto menor a taxa de desconto, maior será o valor estimado do imóvel. Nas estimativas para os valores dos imóveis na seção 5 deste estudo são apresentados cenários com taxas de descontos de 3%, 4%, 5% e 6% ao ano.

Por fim, cabe ressaltar a importância na determinação da forma funcional da equação do preço das características dos imóveis $p(z)$. Ampla é a gama de possibilidades a serem utilizadas. A fim de dar robustez a esta análise, analisam-se três categorias de modelos. A primeira e mais simples pressupõe que o valor dos imóveis/aluguéis tem uma relação linear com as suas características. A segunda e a terceira são mais flexíveis e permitem capturar relações não lineares e mais complexas. Para cada tipo de modelo, é utilizado um método de estimação diferente. Assim, na subseção 2.2 a seguir, apresenta-se o método de Mínimos Quadrados Ordinários para modelos lineares com parâmetros constantes. A subseção 2.3 introduz os Modelos Hierárquicos que são

modelos lineares mais flexíveis, e, por último, na subseção 2.4, apresenta-se o método de Florestas Aleatórias que tem a capacidade de capturar relações mais complexas entre as variáveis.

2.2 Mínimos Quadrados Ordinários

A estimação por Mínimos Quadrados Ordinários (MQO) é um método de regressão linear mais conhecido entre os economistas. Ele tem sido o principal método de estimação dos parâmetros de equações lineares (Equação 6).

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon_i, \tag{6}$$

onde β_0 é o intercepto e β_j para $j = \{1, \dots p\}$ são as inclinações para a variável x_j e p é a quantidade de variáveis explicativas. O MQO é um método de estimação dos parâmetros β_0 e β_j para $j = \{1, \dots p\}$. Sua abordagem se baseia em encontrar os valores dos β_0 e β_j para $j = \{1, \dots p\}$ que minimizam a soma dos resíduos ao quadrado (SRQ).

$$SRQ = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 = \sum_{i=1}^{N} (\epsilon_i)^2, \tag{7}$$

onde N é o tamanho da amostra. A solução para esse problema de minimização pode ser facilmente caracterizada em notação matricial. Seja $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$ o vetor de parâmetros $1 \times (p+1)$, $X = (\mathbf{1}, X_1, X_2, \dots, X_p)$, a matrix de variáveis $N \times (p+1)$ onde 1 é um vetor $N \times 1$ representando o intercepto e X_j para $j = \{1, \dots, p\}$ o vetor $N \times 1$ com os valores da variável explicativa j. Em notação matricial, a equação 6 se torna

$$Y = X\beta + \epsilon \tag{8}$$

onde Y é o vetor $N \times 1$ com os valores da variável dependente (valor dos aluguéis nesse estudo) e ϵ é o vetor $N \times 1$ com os termos de erro. Assim, resolvendo o problema de minimização para a equação 7, obtém-se o estimador $\hat{\beta}$.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{9}$$

Por fim, é possível estimar os valores de $\hat{y}_i$ condicional aos valores de x_j para $j = \{1, \dots, p\}$ por meio da equação 10.

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j \tag{10}$$

Esse método tem algumas limitações. Ele somente captura relações lineares ou quase lineares dependendo da forma como é tratada a variável, por exemplo, colocando y em logaritmo neperiano, é possível incluir não linearidade no método. Outra limitação é a hipótese de que todas as observações podem ser representadas pelos mesmos valores dos parâmetros. Por exemplo, esse modelo pressupõe que se se construir um cômodo a mais em uma casa localizada em Mariana (MG) e outra localizada em Barra Longa (MG), o aumento no valor do aluguel será o mesmo para as duas moradias.

De modo a contornar essas limitações, neste trabalho, aplicam-se dois outros métodos. O primeiro é o Modelo Hierárquico que flexibiliza a hipótese de parâmetros constantes em todos os municípios incluídos na amostra e o segundo é o método de Florestas Aleatórias, uma técnica de aprendizagem de máquina (*machine learning*) onde é possível considerar relações não lineares e mais complexas entre as variáveis.

2.3 Modelos Hierárquicos

Os Modelos Hierárquicos, ou, como também são conhecidos, Modelos Lineares Mistos, são modelos mais flexíveis com relação aos modelos lineares mais simples. Eles permitem que os parâmetros variem segundo grupos aos quais os ativos pertencem. Tendo por base o exemplo anterior, no caso dos Modelos Hierárquicos, é possível estimar parâmetros diferentes para domicílios em municípios diferentes. Por exemplo, a equação 11 mostra um modelo onde o intercepto varia de acordo com o grupo m e a equação 12 mostra um modelo ainda mais flexível em que todos os parâmetros variam de acordo com o grupo.

$$y_{im} = \beta_{0m} + \sum_{j=1}^{p} \beta_j x_{jim} + \epsilon_{im} \tag{11}$$

$$y_{im} = \beta_{0m} + \sum_{j=1}^{p} \beta_{jm} x_{jim} + \epsilon_{im} \quad (12)$$

Esses modelos podem ser difíceis de serem estimados, pois dependem de informações suficientes em todos os grupos, mas podem ser feitos utilizando método de máxima verossimilhança restrita ou técnicas de econometria Bayesiana (GELMAN; HILL, 2006).

Por simplicidade, explica-se o modelo com somente o intercepto variando (Equação 11). Supõe-se que $y_{im} \sim \mathcal{N}(\beta_{0m} + \sum_{j=1}^{p} \beta_j x_{jim}, \sigma_y^2)$ e $\beta_{0m} = \gamma_0 + U_m$, tal que $U_m \sim \mathcal{N}(0, \sigma_{\beta_0}^2)$ ou, de outra forma, $\beta_{0m} \sim \mathcal{N}(\gamma_0, \sigma_{\beta_0}^2)$. O parâmetro γ_0 é conhecido como o efeito fixo enquanto β_{0m} são os efeitos aleatórios nesse modelo. Gelman e Hill (2006) mostram que a estimativa de $\hat{\beta}_{0m}$ é uma combinação de γ_0, a média geral, e a estimativa para o coeficiente utilizando somente as informações da região m restringindo os outros parâmetros a serem constantes entre todas as regiões $(\bar{y}_{im} - \sum_{j=1}^{p} \beta_j \bar{x}_{jm})$. A equação 13 apresenta esse cálculo:

$$\hat{\beta}_{0m} = \frac{\frac{N_m}{\sigma_y^2}}{\frac{N_m}{\sigma_y^2} + \frac{1}{\sigma_{\beta_0}^2}} \left(\bar{y}_m - \sum_{j=1}^{p} \beta_j \bar{x}_{jm} \right) + \frac{\frac{1}{\sigma_{\beta_0}^2}}{\frac{N_m}{\sigma_y^2} + \frac{1}{\sigma_{\beta_0}^2}} \gamma_0, \quad (13)$$

onde N_j é o tamanho do município m. Intuitivamente, essa fórmula representa que o valor do intercepto do município m $(\hat{\beta}_{0m})$ depende da quantidade de observações nos municípios. Se os municípios possuem poucas observações de domicílios, então maior o peso relativo da média global (γ_0).

A previsão nesse método é feita de forma semelhante ao do MQO, por meio da esperança da equação 11 após estimados todos os parâmetros. Por fim, cabe ressaltar que, apesar da maior flexibilidade de modelos para cada município, os modelos hierárquicos também pressupõem uma relação linear entre a variável dependente e explicativas. Na próxima seção, apresenta-se o método de Florestas Aleatórias que consegue capturar relações não lineares e mais complexas entre as variáveis.

2.4 Florestas Aleatórias

O método de Florestas Aleatórias, proposto por Breiman (2001), é uma extensão da técnica de árvores de classificação e regressão (Classification and Regression Trees —

CART, em inglês) introduzido por Breiman et al. (1984) e da técnica de Agregação de Árvores Reamostradas (*Bootstrap Aggregation — bagging*, em inglês) desenvolvida por Breiman (1996). Assim, a fim de explicar o método de Florestas Aleatórias, apresentam-se inicialmente as técnicas de CART e *bagging*.

A vantagem desse método sobre os dois outros previamente apresentados é a capacidade de modelar relações não lineares mais complexas. Segundo James et al. (2013), se a relação entre as variáveis é próxima de uma função linear, regressões lineares podem ter um desempenho melhor do que técnicas de árvore de regressão. Entretanto, se essa relação for não linear e mais complexa, CART e extensões tendem a prever com mais exatidão. Exemplos dessas relações são as funções quadráticas, cúbicas ou de ordens superiores; funções mais complexas como funções trigonométricas (seno, cosseno etc.). Essas relações não são somente matemáticas, mas algumas tem fácil interpretação econômica, diga-se, as funções quadráticas côncavas. Interpretam-se essas funções do ponto de vista do consumidor, como o benefício de se ter um banheiro a mais, por exemplo, não é constante, mas decrescente, ou seja, para uma família de duas pessoas, o benefício de se ter um terceiro banheiro é menor do que o benefício de se ter o segundo.

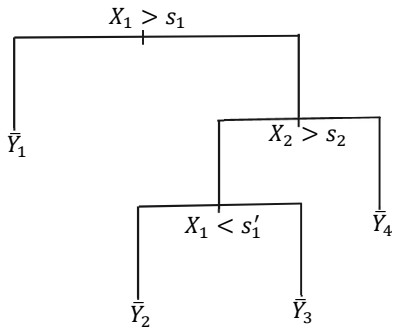
O CART é um modelo flexível de predição não paramétrico. Ele consiste em recursivamente particionar o espaço criado pelas variáveis explicativas em regiões menores. A partição é feita de forma binária, de modo que o modelo final pode ser expresso como uma árvore de decisão. A fim de entender como a árvore de decisão é construída, apresenta-se o seguinte exemplo adaptado de Hastie et al. (2009). Suponha que se deseja prever a variável Y, aluguel, com base nas variáveis X_1 e X_2, número de cômodos e de banheiros. Dessa forma, divide-se a variável X_1 em duas regiões. A região da esquerda tal que $X_1 \leq s_1$ e a região da direita é $X_1 > s_1$, por exemplo, número de banheiros menor ou igual a um e maior que um.

A Figura 2a apresenta essa primeira divisão no primeiro nó da árvore. Se $X_1 \leq s_1$, então o valor predito de Y é $\bar{Y}_1$, aluguel predito. Se $X_1 > s_1$, segue para a esquerda, no segundo nó, onde se divide X_2 em duas partes no ponto s_2. Se $X_2 > s_2$, o valor predito de Y é $\bar{Y}_4$. Caso contrário, segue para o terceiro nó, onde se divide novamente a variável X_1. Se $X_1 < s_1'$, o valor predito de Y é $\bar{Y}_2$, caso contrário, Y previsto é $\bar{Y}_3$. Outra forma de visualizar a árvore é por meio da divisão do espaço gerado por X_1 e X_2 na Figura 2b. Cada retângulo é uma região do espaço (R_m, $m = \{1,2,3,4\}$). A previsão de Y para cada região é a média aritmética para os valores de Y correspondentes aos respectivos valores de X_1 e X_2 pertencentes a cada região, $\bar{Y}_m = \sum_{i=1}^{N} I(x_{1i} \text{ e } x_{2i} \in R_m) y_i$, onde N é

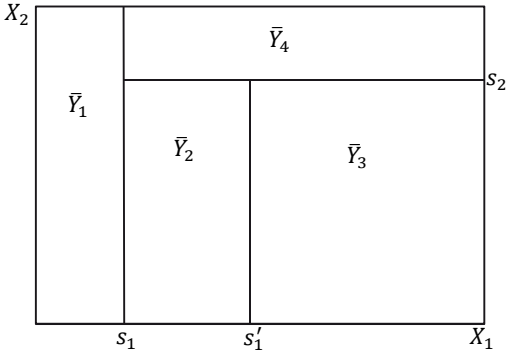
o tamanho da base de dados, x_{1i} e x_{2i} são realizações das variáveis X_1 e X_2 e $I(.)$ é uma função indicadora que assume valor igual a 1 se os valores x_{1i} e x_{2i} pertencem a região R_m e zero, caso contrário.

Figura 2 — Árvores de regressão

(a) Árvore de decisão



(b) Divisão do espaço gerado por X_1 e X_2



Fonte: Elaboração pelos próprios autores, adaptado de Hastie et al. (2009).

Teoricamente, cada região R_m não precisa ser necessariamente um retângulo, ela pode assumir qualquer formato (JAMES et al., 2013), mas se assume esse formato por simplicidade e facilidade na interpretação e explicação. Formalmente, o objetivo do CART é encontrar J regiões tal que minimize a soma dos resíduos ao quadrado (SRQ) dado por

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \bar{Y}_{R_j} \right)^2, \tag{14}$$

onde $\bar{Y}_{R_j}$ é a média de Y na região R_j. Entretanto, devido à complexidade computacional e a inviabilidade de se considerarem todas as possíveis partições do espaço, utiliza-se a partição binária recursiva, assim como apresentado na Figura 2a. Para tanto, é preciso escolher o ponto de corte para cada variável e a própria sequência com que as variáveis são particionadas.

Seguindo Hastie et al. (2009), suponha que há *p* variáveis explicativas e defina-se as regiões $R_1(j, s_j) = \{X | X_j < s_j\}$ e $R_2(j, s_j) = \{X | X_j \geq s_j\}$, onde $j = \{1, \ldots, p\}$. Escolhe-se j e s_j que resolve o problema:

$$\min_{j,s_j}\left[\min_{c_1}\sum_{i:x_i\in R_1(j,s_j)}(y_i-c_1)^2+\min_{c_2}\sum_{i:x_i\in R_2(j,s_j)}(y_i-c_2)^2\right] \quad (15)$$

Para qualquer escolha de *j* e s_j, as minimizações dentro dos colchetes são resolvidas com $\hat{c}_1=\frac{1}{N_{R_1}}\sum_{i:x_i\in R_1(j,s_j)}y_i$ e $\hat{c}_2=\frac{1}{N_{R_2}}\sum_{i:x_i\in R_2(j,s_j)}y_i$, onde N_{R_1} e N_{R_2} são os tamanhos das regiões R_1 e R_2, respectivamente. Intuitivamente, a cada nó, seleciona-se aquela variável *j* e seu respectivo ponto de corte s_j que minimiza o erro de previsão.

Um questionamento natural é sobre em qual ponto se deve parar a árvore (isto é, as partições binárias). Quanto maior a árvore, maior a probabilidade de o modelo ser sobreajustado aos dados. Isto é, não há erro para a previsão utilizando as observações dentro da base de dados. Entretanto, isso não significa que ele tenha o mesmo desempenho para observações fora da amostra. Árvores maiores são mais instáveis ou com alta variância, de forma que ao gerar uma outra árvore com uma base similar, mas não igual, pode resultar uma árvore muito diferente. Por outro lado, quanto menor a árvore, maiores as chances de ela deixar de capturar não linearidade e relações complexas importantes para a previsão. A variância da árvore deve ser menor, mas o desempenho tende a ser pior. Hastie et al. (2009) sugerem utilizar o método de poda custo-complexidade (*cost-complexity prunnig*). Entretanto, o método de Florestas Aleatórias utiliza outra estratégia para balancear o viés e a variância. Em essência, o autor utiliza a ideia de Agregação de Árvores Reamostradas (*bagging*), que reduz a variância fazendo a média da previsão de diferentes árvores. A fim de esclarecer esse ponto, apresenta-se o método de Agregação de Árvores Reamostradas.

O método de Agregação de Árvores Reamostradas é proposto por Breiman (1996) com o objetivo de reduzir a variância das árvores de decisão. Sua abordagem é relativamente simples. Criam-se *B* subamostras com reposição da base de dados original. Para cada subamostra se gera uma árvore de decisão, assim como apresentado anteriormente, sem utilizar técnicas de poda, de modo que elas tenham alta variância. A fim de reduzir a variância, a previsão da variável dependente é a média das previsões das *B* árvores.

Um ponto fraco dessa metodologia é a alta correlação entre as árvores e, consequentemente, das previsões. A intuição para isso é que as variáveis mais importantes serão escolhidas para os primeiros nós na maioria, se não em todas as árvores (JAMES et al., 2013). O método de Florestas Aleatórias é uma maneira de reduzir a correlação entre as árvores e melhorar o desempenho nas previsões fora da amostra.

Para tanto, Breiman (2001) adiciona outra etapa na construção das árvores. A cada nó, um subconjunto $mtry$ das p variáveis explicativas é escolhido aleatoriamente. Desse subconjunto, a melhor variável preditora é escolhida por meio da equação 15. Dessa forma, evita-se que as mesmas variáveis sejam escolhidas nos primeiros nós de todas ou quase todas as árvores.

Para exemplificar todo o procedimento da técnica de Florestas Aleatórias, suponha uma base de dados com N=500 observações onde a variável dependente é o valor do aluguel e há cinco variáveis explicativas $X = \{X_1, X_2, X_3, X_4, X_5\}$, por exemplo, número de cômodos, número de dormitórios, número de banheiros, uma variável binária que assume valor igual a 1 se a moradia é um apartamento e zero se for uma casa, uma variável binária que assume valor igual a 1 se a moradia tem rede de esgoto e zero, caso contrário. Define-se que $mtry = 3$, ou seja, a cada nó, três das cinco variáveis explicativas serão selecionadas aleatoriamente, para posteriormente escolher aquela que melhor prevê o valor do aluguel. Define-se uma quantidade de subamostras geradas por reamostragem com repetição e, consequentemente, o número de árvores (*ntree*) igual a 1.000. Por último, define-se que ao final de cada ramo da árvore tem que se ter no mínimo cinco observações[14]. A técnica de Florestas Aleatórias segue o seguinte procedimento:

I Inicia-se na primeira subamostra $(b = 1)$;

II No primeiro nó, selecionam-se aleatoriamente três variáveis no conjunto X, supõe-se que sejam as variáveis X_1, X_3 e X_4.

III Resolve o problema 15 com as três variáveis selecionadas no item anterior, de forma que o resultado é a variável mais relevante e seu respectivo ponto de corte (s_j) para definir as regiões $R_1(j, s_j) = \{X | X_j < s_j\}$ e $R_2(j, s_j) = \{X | X_j \geq s_j\}$;

IV Segue-se para o segundo nó, onde novamente é selecionado aleatoriamente um segundo subconjunto de variáveis, por exemplo, X_2, X_3 e X_5;

V Resolve o problema 15 para as variáveis selecionadas no item anterior;

[14] O número de árvores (*ntree*), o número de variáveis consideradas potencialmente (*mtry*) e o tamanho mínimo de observações ao final de cada ramo da árvore (*nodesize*) são conhecidos como hiperparâmetros do modelo e não há um consenso sobre quais devem ser seus valores. Os números ótimos dependem da natureza do problema e dos dados. Nesse estudo, testam-se diferentes combinações de valores para esses três hiperparâmetros e seleciona-se aquele com melhor desempenho na previsão. Apresenta-se esse exercício na subseção 4.2.3.

VI Repete o procedimento IV e V para todos os nós da árvore e não é aplicado nenhum método de poda, mas até que cada ramo tenha no mínimo cinco observações;

VII Repetem-se os procedimentos de I a VI para $b = \{2, \dots 1000\}$, de forma que se tenham 1.000 árvores ao final;

VIII Prevê-se o valor do aluguel para observações selecionadas em cada uma das 1.000 árvores;

IX A previsão final é a média das previsões das 1.000 árvores.

Cabe ressaltar que o desempenho do método de florestas aleatórias se baseia na não linearidade, relações complexas e seleção de variáveis. Para esse último, um indicador de importância relativa de cada variável explicativa pode ser derivado. Esse indicador mostra o quanto o erro quadrático médio aumentaria se cada variável fosse substituída por uma outra cujos valores foram gerados aleatoriamente. Desse modo, quanto maior seu valor, mais importante é a variável para prever o valor do aluguel.

Por fim, apresenta-se a estratégia de construção do intervalo de confiança para as previsões. Interpreta-se esse intervalo da seguinte forma: se novos dados de aluguéis forem coletados, os valores para propriedades residenciais com características semelhantes estarão contidos nesse intervalo com probabilidade predeterminada. Por exemplo, se se determinar que a probabilidade é 95%, diz-se que o intervalo de confiança tem 95% de nível de confiança. Constrói-se um intervalo para cada valor estimado $\bar{y}$ utilizando a fórmula $\hat{y} \pm t_{c,1-\alpha}\sqrt{\hat{V}^B}$, onde $t_{c,1-\alpha}$ é o valor crítico para u nível de confiança $1-\alpha$ da distribuição normal[15] e $\hat{V}^B$ é o erro-padrão da previsão $\hat{y}$. Utiliza-se o estimador de Wager, Hastie e Efron (2014) para calcular a variância da predição (V^B):

$$\hat{V}^B = \sum_{i=1}^{n} \widehat{Cov}_i^2, \tag{16}$$

$$\widehat{Cov}_i = \frac{\sum_{b=1}^{B}(N_{bi} - 1)\left(\hat{y}_b - \bar{\hat{y}}_b\right)}{B} \tag{17}$$

[15] Por exemplo, para um nível de confiança de 95%, $t_{c,0,95} = 1{,}96$.

Onde B é a quantidade de reamostragens, n é a quantidade de observações, N_{bi} é a quantidade de vezes que a observação i aparece na reamostragem b, $\hat{y}_b$ é a previsão da reamostragem b e $\bar{\hat{y}}_b = \frac{\sum_{b=1}^{B} \hat{y}_b}{B}$.

3 DADOS

Nesta seção apresentam-se as duas bases de dados utilizadas no presente estudo: (i) os dados de domicílios do Censo Demográfico de 2010 (IBGE, 2010); e (ii) os dados das propriedades cadastradas pela Fundação Renova. Os dados de domicílios do Censo Demográfico são a base para o desenvolvimento do modelo de previsão do valor das habitações por preços hedônicos de uma amostra selecionada. Os dados do cadastro da Fundação Renova, por sua vez, compõem a única base de dados estruturada disponível com as informações das propriedades residenciais danificadas, total ou parcialmente, no município de Barra Longa[16].

3.1 Dados de domicílios do Censo Demográfico

Os modelos de previsão do valor das habitações por preços hedônicos tiveram por base os dados do Censo Demográfico de 2010 (IBGE, 2010) que são coletados por meio de dois questionários: um aplicado em todos os setores censitários[17] do Brasil, chamado Questionário Básico[18], e outro que contém mais perguntas e é aplicado em um subconjunto de domicílios a partir de uma amostragem representativa em nível de área de ponderação, chamado Questionário da Amostra[19].

A pergunta sobre valor do aluguel, para o caso de domicílios alugados, consta apenas no Questionário da Amostra, assim como as características completas dos domicílios, como tipo de habitação (casa, apartamento etc.) e números de cômodos e dormitórios, além da forma de abastecimento de água, destinação de lixo e fonte de energia elétrica, entre outras informações de interesse para caracterização dos domicílios. Assim, utilizam-se no presente estudo somente dados dos domicílios da amostra do Censo

[16] Conforme mencionado anteriormente, no cadastramento realizado pela Fundação Renova das propriedades residenciais atingidas, não constam informações completas sobre as características das moradias localizadas no município de Mariana (MG). Dessa forma, neste estudo não estimamos o valor agregado do estoque habitacional, em momento anterior ao rompimento da Barragem de Fundão, para as localidades atingidas de Mariana.

[17] "O setor censitário é a unidade territorial estabelecida para fins de controle cadastral, formado por área contínua, situada em um único quadro urbano ou rural, com dimensão e número de domicílios que permitam o levantamento por um recenseador (IBGE, 2015).

[18] O Questionário Básico do Censo Demográfico de 2010 que foi aplicado em todos os setores censitários do Brasil pode ser encontrado no seguinte link do IBGE: <https://censo2010.ibge.gov.br/images/pdf/censo2010/questionarios/questionario_basico_cd2010.pdf>.

[19] O Questionário da Amostra do Censo Demográfico de 2010 pode ser encontrado no seguinte link do IBGE: <https://censo2010.ibge.gov.br/images/pdf/censo2010/questionarios/questionario_amostra_cd2010.pdf>.

Demográfico, que contém as informações necessárias para a previsão, por preços hedônicos, do valor das propriedades residenciais.

A Tabela 1 apresenta o total de domicílios da amostra do Censo Demográfico de 2010 com informação sobre o valor de aluguel no município de Barra Longa (MG) e para todo o estado de Minas Gerais. Em Barra Longa, tem-se o valor do aluguel declarado em 2010 para 43 domicílios da amostra, sendo nove localizados em zonas rurais e 34 em áreas urbanas. Para o estado de Minas Gerais, tem-se o valor do aluguel declarado em 2010 para 121.691 domicílios da amostra, sendo 3.623 em zonas rurais e 118.068 em áreas urbanas.

Ainda na Tabela 1, a coluna "Percentual de domicílios alugados" corresponde à porcentagem dos domicílios com valores de aluguel entre todos os domicílios da amostra do Censo Demográfico, ponderado pelo fator de expansão, utilizado para calcular o resultado para os dados do universo. Assim, em Barra Longa, somente 6,3% dos domicílios são alugados, enquanto para o estado de Minas Gerais, 17,8% dos domicílios da amostra do Censo têm informações sobre valor de aluguel. Também se apresenta o valor médio, em reais, do aluguel declarado para agosto de 2010, ponderado pelos seus respectivos pesos amostrais. Em Barra Longa, o valor médio do aluguel, a preços de agosto de 2010, era de R$ 158,72, sendo R$ 90,16 para zonas rurais e R$ 175,18 para as áreas urbanas. Esse valor é inferior ao valor médio do aluguel para o estado de Minas Gerais, de R$ 292,89 a preços de 2010, sendo R$ 236,29 para a média dos domicílios da amostra com valores de aluguel nas zonas rurais e R$336,83 nas áreas urbanas.

Tabela 1 — Domicílios da amostra do Censo Demográfico com informações de valor de aluguel para 2010 em Barra Longa (MG) e Minas Gerais

Local dos domicílios	Total de domicílios com valor de aluguel	Percentual de domicílios alugados	Valor médio do aluguel em 2010 (R$)
Barra Longa (MG)	**43**	**6,3%**	**158,72**
Rural	*9*	*2,0%*	*90,16*
Urbana	*34*	*13,0%*	*175,18*
Minas Gerais	**121.691**	**17,8%**	**292,89**
Rural	*3.623*	*2,2%*	*236,29*
Urbana	*118.068*	*20,3%*	*336,83*

Fonte: Elaboração própria, a partir de dados do Censo Demográfico (2010).

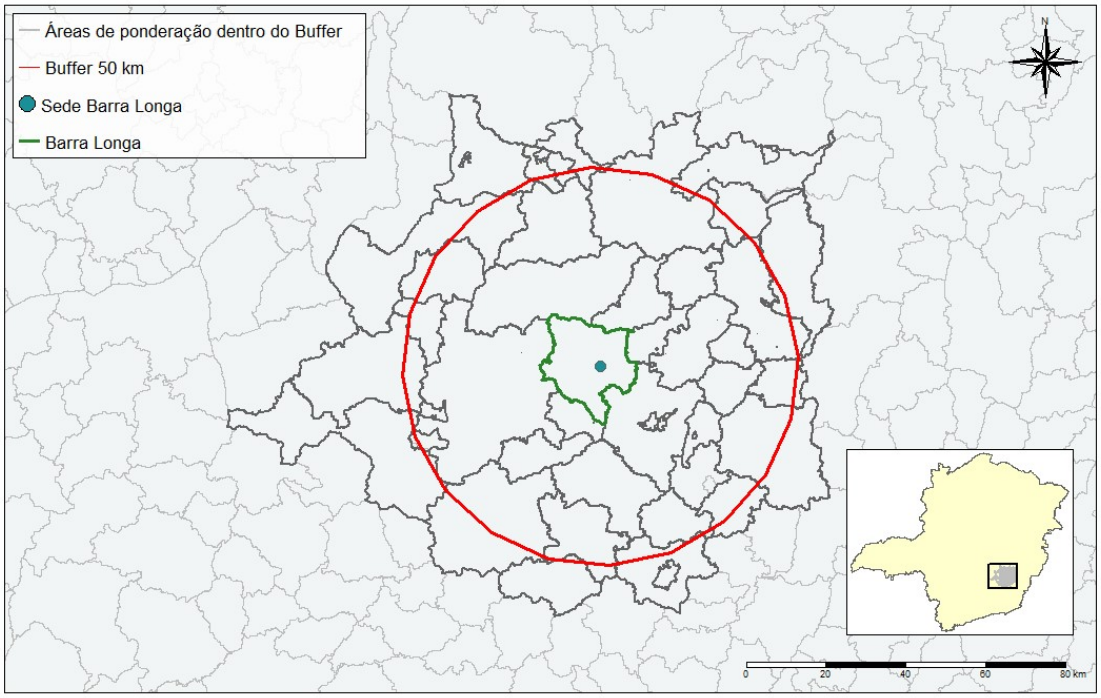
Assim, uma restrição da base de dados do Censo Demográfico de 2010 para estimar o valor agregado das propriedades residenciais atingidas em Barra Longa é o pequeno número de domicílios com informação de valor de aluguel para esse município. Por outro lado, na tentativa de se manter uma amostra com certas características locais não observadas, adota-se uma estratégia de delimitação da amostra total do Censo Demográfico para o estado de Minas Gerais aos domicílios em áreas de ponderação no entorno do município de Barra Longa.

3.1.1 Delimitação da amostra de domicílios do Censo Demográfico

Para este estudo, delimitou-se a amostra de domicílios do Censo Demográfico com informações de valor de aluguel para o estado de Minas Gerais pelas áreas de ponderação de até 50 quilômetros da sede municipal de Barra Longa. A área de ponderação é a menor unidade geográfica, formada por um agrupamento de setores censitários, no qual se têm informações do Questionário da Amostra do Censo Demográfico. As áreas de ponderação são menores ou iguais às fronteiras municipais. Por exemplo, para o município de Barra Longa há apenas uma área de ponderação que compreende todo o município.

Utilizando a base de coordenadas geográficas das sedes municipais e das bases cartográficas municipais do IBGE de 2015 (IBGE, 2015), delimitou-se uma área em um raio de 50 quilômetros a partir da sede do município de Barra Longa, conforme apresentado na Figura 3. Desta forma, todas as áreas de ponderação com ao menos uma parte de seu território dentro desse raio de 50 quilômetros foram incluídas na amostra de domicílios selecionados. Soma-se, assim, na base de dados restrita, um total de 40 áreas de ponderação, localizadas em 35 municípios, com informações de valor de aluguel e respectivas características do domicílio alugado.

Figura 3 — Delimitação da amostra de domicílios do Censo Demográfico com informações de valor de aluguel para o estado de Minas Gerais



Fonte: Elaboração própria, a partir de dados do Censo Demográfico (2010).

3.1.2 Seleção das variáveis do Censo Demográfico

Para desenvolver o modelo de previsão dos valores dos domicílios por preços hedônicos, selecionaram-se as variáveis[20] coletadas pelo Questionário da Amostra do Censo Demográfico. Para a variável dependente, tem-se o valor de aluguel declarado para os domicílios alugados da amostra, em reais, a preço de agosto de 2010. As variáveis explicativas são um conjunto de características atribuídas aos domicílios amostrados, a saber: números de cômodos, dormitórios e banheiros de uso exclusivo, tipos de domicílio (casa, apartamento etc.), tipo do setor censitário (urbana ou rural), de localização (coordenadas geográficas), formas de esgotamento sanitário, abastecimento de água e destinação de lixo e fonte de energia elétrica.

Para algumas características dos domicílios, a coleta de dados pelo Censo Demográfico no Questionário da Amostra é feita por meio de perguntas de múltipla escolha. Nesses casos, a variável foi construída, conforme critérios apresentados no APÊNDICE A, de forma a captar um padrão de resposta binário. Por exemplo, para a variável “lixo

[20] A escolha das variáveis se respalda na literatura de preços hedônicos internacional (BOURASSA et al., 1999; HILL, 2011; SHEPPARD, 1999) e estudos nacionais (MORAIS e CRUZ, 2003; VINHAIS, 2006).

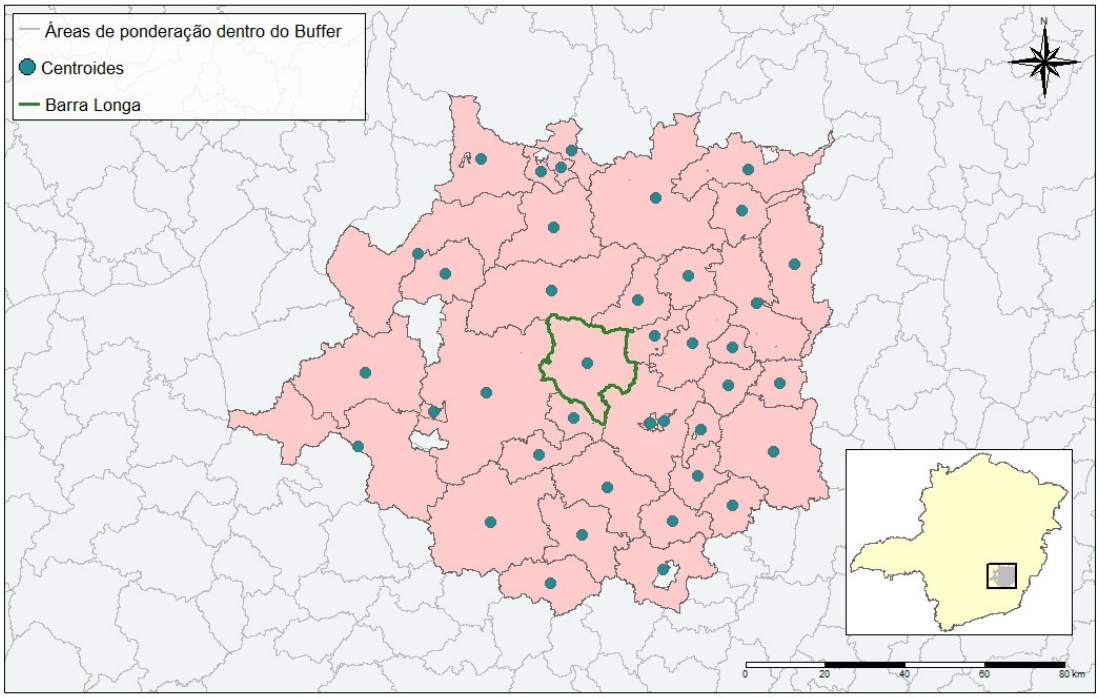
coletado por serviço de limpeza", considerou-se como "sim" (ou valor igual a um), quando escolhida a opção de resposta "coletado diretamente por serviço de limpeza"; do contrário, como "não" (ou valor igual a zero), todas as demais alternativas disponíveis, quais sejam: colocado em caçamba, queimado, enterrado, jogado em terreno baldio ou logradouro, rio, lago ou mar, e outro destino.

Uma limitação, contudo, entre as informações disponíveis para o desenvolvimento do modelo de previsão é a ausência de coordenadas geográficas para os domicílios da amostra do Censo Demográfico. Como sugerem Fletcher et al. (2004) e Shepparde (1999), as questões relacionadas com as amenidades locais são importantes fatores para os modelos de previsão de valores de imóveis por preços hedônicos. Assim, quando se conhece a localização geográfica exata de cada domicílio, podem-se incluir no modelo não só suas respectivas latitude e longitude, mas também variáveis de distância a importantes equipamentos sociais, entre outras amenidades.

No caso da base de domicílios do Censo Demográfico, a melhor aproximação que se tem da localização dos domicílios são as coordenadas geográficas (latitude e longitude) de sua área de ponderação correspondente. Desta forma, para todos os domicílios de uma dada área de ponderação, atribuíram-se as coordenadas geográficas dos centroides[21] da respectiva área, conforme ilustrado no mapa da Figura 4

[21] O centroide geográfico equivale ao centro de massa de um polígono no espaço. Seu cálculo pode ser obtido a partir da sua divisão em triângulos das delimitações geográficas, usando suas áreas como peso.

Figura 4 — Centroides das áreas de ponderação da amostra do Censo Demográfico



Fonte: Elaboração própria, a partir de dados do Censo Demográfico (2010).

3.1.3 Estatísticas descritivas das variáveis do Censo Demográfico

Dadas as variáveis selecionadas, apresentadas na subseção anterior 3.1.2, para o modelo de previsão dos valores dos domicílios por preços hedônicos, a Tabela 2 apresenta as estatísticas descritivas de cada uma delas na amostra de domicílios com valores de aluguel do Censo Demográfico restrita às áreas de ponderação definidas na subseção 3.1.1.

Tabela 2 — Estatísticas descritivas para as variáveis selecionadas nas áreas de ponderação da amostra delimitada do Censo Demográfico de 2010

Estatísticas	N	Média	Desvio-Padrão	Min.	Percentil (25)	Percentil (75)	Max.
Valor do Aluguel (em reais)	3.353	264,51	189,14	1,00	150,00	350,00	3200,00
Número de Cômodos	3.353	5,81	1,75	1,00	5,00	7,00	20,00
Número de Dormitórios	3.353	1,95	0,78	1,00	1,00	2,00	5,00
Número de Banheiros de Uso Exclusivo	3.353	1,16	0,45	0,00	1,00	1,00	5,00
Apartamento	3.353	0,10	0,29	0,00	0,00	0,00	1,00
Urbano	3.353	0,96	0,20	0,00	1,00	1,00	1,00
Rede Geral de Esgoto	3.353	0,87	0,34	0,00	1,00	1,00	1,00
Rede Geral de Água	3.353	0,95	0,22	0,00	1,00	1,00	1,00
Poço ou Nascente na Propriedade	3.353	0,02	0,14	0,00	0,00	0,00	1,00
Lixo Coletado por Serviço de Limpeza	3.353	0,87	0,33	0,00	1,00	1,00	1,00
Energia Elétrica	3.353	1,00	0,05	0,00	1,00	1,00	1,00
Latitude	3.353	-20,25	0,25	-20,77	-20,41	-20,05	-19,80
Longitude	3.353	-43,04	0,31	-43,62	-43,31	-42,80	-42,57

Fonte: Elaboração própria, a partir de dados do Censo Demográfico (2010).

Na Tabela 2 a coluna “N” corresponde ao número total de observações (isto é, número total de domicílios na amostra), a coluna “Média” corresponde à média aritmética para cada variável, a coluna “Desvio-Padrão” apresenta o desvio-padrão das variáveis como medida de dispersão, a coluna “Min.” corresponde ao menor valor da variável em questão, a coluna “Percentil (25)” indica qual é o percentil 25, ou primeiro quartil da distribuição em ordem crescente, enquanto a coluna “Percentil (75)” corresponde ao percentil 75, ou terceiro quartil da distribuição, e, por fim, a coluna “Max.” é o valor máximo da distribuição.

Nas 40 áreas de ponderação da amostra do Censo Demográfico restrita para a região de Barra Longa, há 3.353 domicílios com informações de valor de aluguel. O valor do aluguel, em reais e a preço de agosto de 2010, será a variável explicada nos modelos de preços hedônicos do capítulo 4. Em média, os valores de aluguel da amostra de Barra Longa e seu entorno correspondem a R$ 264,50 mensais.

Entre as variáveis explicativas para os modelos de preços hedônicos da seção 4, os domicílios na região de Barra Longa, em média, possuem 5,8 cômodos, sendo dois deles, em média, utilizados como dormitórios e 1,2 utilizado como banheiro de uso exclusivo. As demais variáveis explicativas apresentadas correspondem a um conjunto

de informações binárias, atribuindo-se valor igual a um quando certa característica do domicílio está presente e valor zero, caso contrário.

A variável “Apartamento” é uma variável binária que corresponde ao valor um quando o domicílio é apartamento e valor zero quando o domicílio é casa. Grande parte dos domicílios na região são casas, com somente 10% dos domicílios amostrados sendo apartamentos. A variável “Urbano” é também uma variável binária que tem valor igual a um quando para o domicílio está localizado em um setor censitário urbano e valor zero caso esteja em zona rural. Nota-se que 96% dos domicílios com informações de aluguel localizam-se em setores urbanos na amostra analisada, uma porcentagem muito acima do número de domicílios que estão em áreas urbanas na região, que correspondem a 72% dos 24.429 da amostra[22] de domicílios no raio de 50 quilômetros da sede de Barra Longa.

Para a variável “Rede Geral de Esgoto” atribui-se o valor um para os domicílios que possuem rede geral de esgoto ou pluvial e zero para os casos de fossa séptica, fossa rudimentar, vala, rio, lago ou mar, outro ou em branco; 87% dos domicílios da região amostrada com valores de aluguel fazem lançamento de esgoto na rede geral de esgoto ou pluvial. Analogamente, 95% desses domicílios têm acesso à rede geral de distribuição para abastecimento de água em pelo menos um cômodo e zero tem poço ou nascente na propriedade.

Para a variável “Lixo Coletado por Serviço de Limpeza” atribuiu-se o valor um para os casos em que o destino do lixo residencial é coletado diretamente por serviço de limpeza e zero para os casos em que o lixo é coletado em caçamba de serviço de limpeza, queimado (na propriedade), enterrado (na propriedade), jogado no terreno baldio ou logradouro, jogado em rio, lago ou mar, tem outro destino, ou ficou em branco. A variável “Energia Elétrica” corresponde ao valor um para os domicílios que possuem energia elétrica fornecida por companhia distribuidora e zero caso tenham energia elétrica por outras fontes ou caso não tenham.

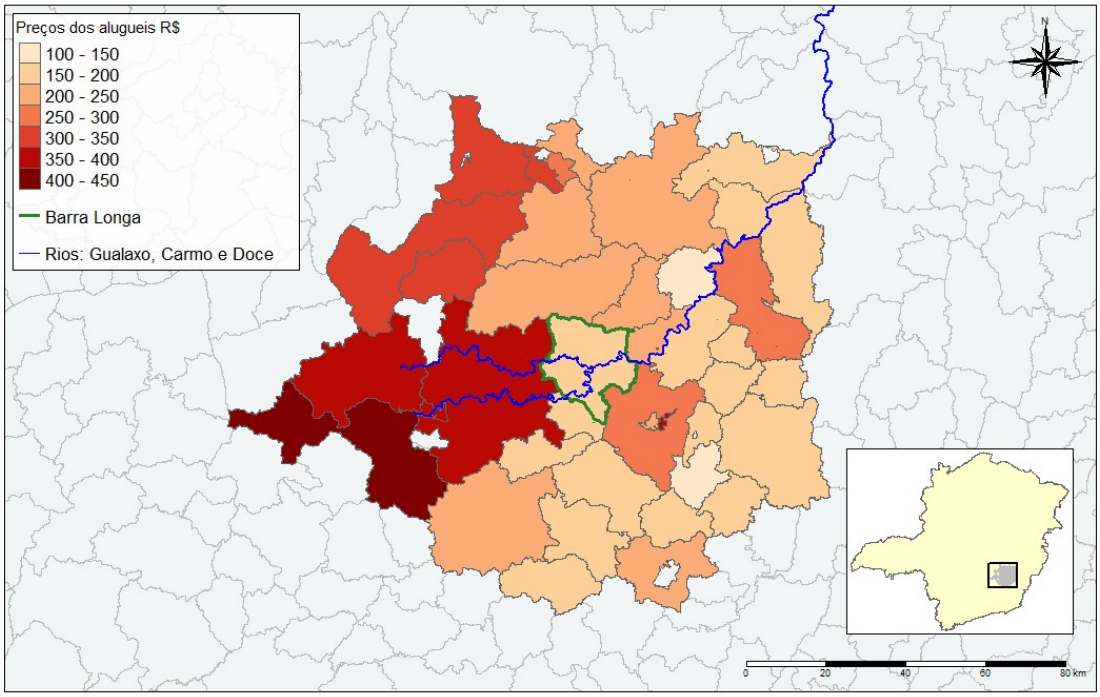
Para os domicílios com valor de aluguel da amostra na região de interesse, 87% têm coleta de lixo residencial realizada diretamente pelo serviço de limpeza e praticamente 100% têm energia fornecida por companhia distribuidora.

Adicionalmente, a fim de ilustrar os preços médios dos aluguéis nas áreas de ponderação, a Figura 5 apresenta a disposição espacial dos preços de aluguel na amostra tratada. Nota-se que as áreas de ponderação do município de Mariana, na

[22] Essa característica de aluguéis se concentrarem em áreas urbanas já foi ressaltada (KILSZTAJNA et al., 2009; PASTERNAK e BÓGUS, 2014).

região sudoeste da amostra, possuem os valores médios mais elevados a preços de agosto de 2010.

Figura 5 — Preço médio dos aluguéis da amostra do Censo Demográfico a preços de agosto de 2010 por área de ponderação



Fonte: Elaboração própria, a partir de dados do Censo Demográfico (2010).

3.2 Dados de domicílios do Cadastro da Fundação Renova

Após o rompimento da Barragem de Fundão, foi iniciado o procedimento de cadastramento das pessoas atingidas por meio da realização do chamado "Cadastro Emergencial", ocorrido entre os meses de novembro de 2015 e maio de 2016, e realizado em quatro campanhas. Devido à urgência na delimitação imediata dos atingidos, esse procedimento foi realizado por diversas empresas, utilizando-se de formulários diversos[23].

Em 2 de março de 2016, a União, estados e suas respectivas autarquias firmaram um acordo, intitulado de Termo de Transação de Conduta (TTAC), com as empresas Samarco Mineração S.A., Vale S.A. e BHP Billiton Brasil Ltda., em que as partes dispuseram sobre a constituição de uma fundação, à qual incumbiria, de forma

[23] Para mais detalhes sobre o processo de cadastramento dos atingidos, ver Produto "Análise do Cadastro Socioeconômico".

centralizada, a gestão das ações com o objetivo de "recuperar o meio ambiente e as condições socioeconômicas da ÁREA DE ABRANGÊNCIA impactada pelo EVENTO, de forma a restaurar a SITUAÇÃO ANTERIOR" (Cláusula 5). Tais ações seriam organizadas sob o formato de programas socioambientais e socioeconômicos.

É nesse contexto que o TTAC dispõe sobre o levantamento e cadastro de "impactados", a servir de referência de dimensionamento e quantificação de todos os programas socioeconômicos. Segundo a Cláusula 21 do TTAC:

> O cadastro se refere às pessoas físicas e jurídicas (neste último caso, apenas micro e pequenas empresas), famílias e comunidades, devendo conter o levantamento das perdas materiais e das atividades econômicas impactadas.

Constituída a Fundação Renova, em 28 de junho de 2016, coube-lhe a apresentação do planejamento e descrição de seus programas, inclusive aquele voltado ao levantamento e cadastro de atingidos. Houve, então, a divisão do cadastro em dois momentos: o "Cadastro Emergencial", iniciado antes da celebração do TTAC, e o "Cadastro Integrado".

Esse segundo momento de cadastramento realizado já após a constituição da Fundação Renova, em agosto de 2016, deveria seguir as diretrizes e os procedimentos constantes da primeira versão do referido programa, que passaria a ser denominado de "Cadastro Integrado". Foi consolidado o formulário a ser aplicado na entrevista e no levantamento dos dados — com início em agosto de 2016 e previstas para encerramento em 31 de março de 2017 (ver Nota Técnica 6/2016 CTOS e Deliberação CIF n. 32/2016) — ampliando o prazo original do TTAC que previa oito meses para a conclusão do cadastro em sua cláusula 19. As famílias já registradas no "Cadastro Emergencial" seriam novamente cadastradas no "Cadastro Integrado".

Dito isso, até o momento, 93.848 indivíduos e 32.588 propriedades[24] foram incluídos no "Cadastro Integrado" da Fundação Renova. Especificamente, as propriedades residenciais cadastradas pela Fundação Renova em Barra Longa, por meio de seu "Cadastro Integrado", serão a base das estimativas do valor agregado do estoque habitacional atingido pela lama de rejeito e/ou pelas obras de limpeza em Barra Longa em momento no tempo anterior ao rompimento, ao estimar valores de referência para outubro de 2015.

[24] A base do Cadastro da Fundação Renova foi obtida por meio do download do filtro 1322, que contém as informações levantadas no questionário socioeconômico aplicado durante o cadastramento, referentes às propriedades e aos indivíduos. A extração dos dados do sistema foi realizada no dia 28 de agosto de 2019.

3.2.1 Delimitação dos dados do Cadastro para Barra Longa

Entre as propriedades cadastradas pela Fundação Renova, existem diversos tipos de imóveis, como estabelecimentos públicos (escolas), estabelecimentos comerciais, terrenos, residências, entre outros. A identificação das residências de Barra Longa se deu a partir de informações contidas na base de dados do Cadastro da Fundação Renova e em documento correlato contendo a titularidade das propriedades cadastradas[25].

Para este estudo, consideraram-se apenas as residências cadastradas que:

- Não são estabelecimentos com CNPJ;
- Há ao menos um núcleo familiar; e
- Há ao menos um cômodo (para excluir propriedades que são apenas terrenos).

Adicionalmente, no conjunto de informações disponíveis no Cadastro da Fundação Renova existem perguntas relacionadas com afetação da propriedade devido ao rompimento. Dessa forma, o critério de inclusão das propriedades consideradas em Barra Longa teve por base um conjunto de variáveis obtidas no próprio cadastramento, a saber:

- **Propriedades de Barra Longa:** a identificação das propriedades que se encontram no município, obtidas a partir das coordenadas geográficas disponibilizadas na base de dados de propriedades. Assim, restringiu-se o conjunto de propriedades que se localizam dentro dos limites geográficos do município de Barra Longa. Constatam-se 756 propriedades cadastradas no município;
- **Edificação principal:** identificação por meio da variável “10.10.2.8_a” que contém as informações do tipo da edificação principal da propriedade. Para o estudo foram consideradas as propriedades com edificação principal como “casa”, “casa de cômodo”, “apartamento” e “apartamento de cômodo”[26];

[25] O documento GOV3493, disponibilizado pela Fundação Renova em 09/10/2019, por meio de seu repositório (Intralinks), contém as informações relacionadas com as titularidades das propriedades (para identificação dos equipamentos públicos e ou dos proprietários) e a identificação se o imóvel possui ou não CNPJ.
[26] Não foram incluídas edificações como: Barracão, Galpão, Armazém" e "Outro".

- **Propriedades atingidas:** a variável "1.1.43.2", por sua vez, contém as informações de perdas e/ou danos declarados a edificações e construções civis (benfeitorias não reprodutivas);
- **Afetação total ou parcial:** duas perguntas mais específicas marcam as propriedades com declaração de impossibilidade de uso, sendo elas[27]:

I 10.10.32.1: afetação total com impossibilidade de uso da edificação;

II 10.10.32.2: afetação parcial estrutural com impossibilidade de uso da edificação.

A Tabela 3 apresenta o total de propriedades cadastradas em Barra Longa observando cada uma das categorias e o total de propriedades residenciais efetivamente incluídas nas estimativas do valor agregado do estoque habitacional de Barra Longa pelos modelos de preços hedônicos.

Tabela 3 — Propriedades cadastradas no Cadastro da Fundação Renova em Barra Longa

Categorias de propriedades cadastradas em Barra Longa	Total	Incluídas na base para estimação de valores por preços hedônicos
Propriedades cadastradas	759	471
Residências cadastradas	505	471
Residências cadastradas com perda e/ou danos a edificações e construções civis	310	291
Residências com afetação total com impossibilidade de uso da edificação	72	69
Residências com afetação parcial estrutural com impossibilidade de uso da edificação	26	23

Fonte: Elaboração própria a partir do Cadastro da Fundação Renova, versão 28/08/2019.

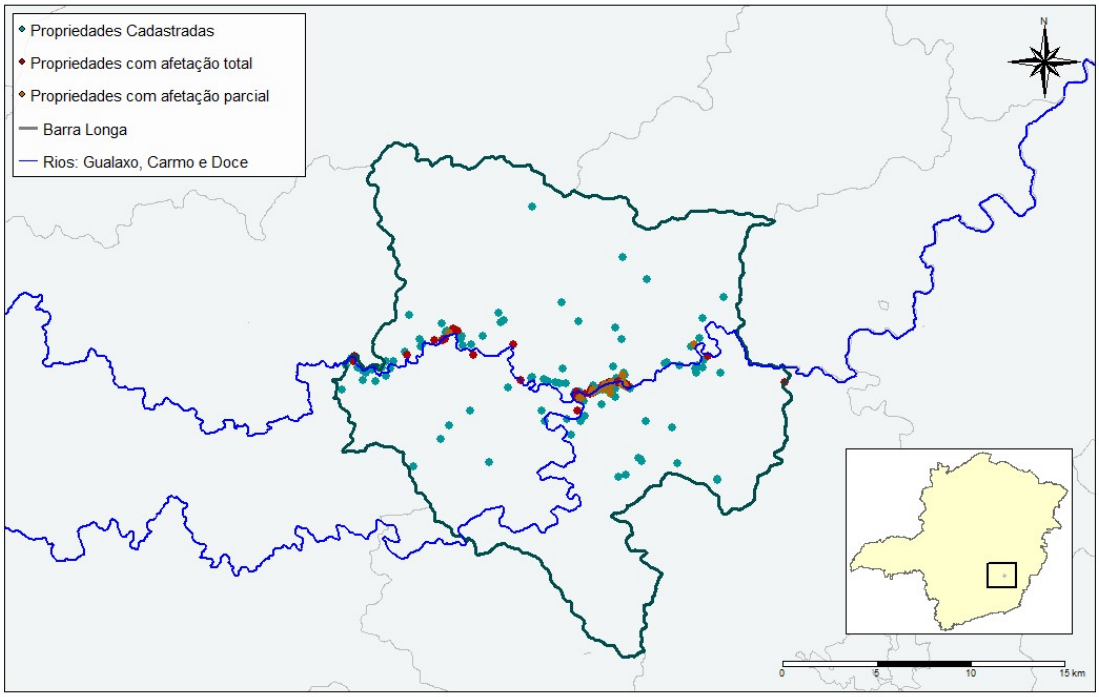
Portanto, do total de 759 propriedades cadastradas em Barra Longa, 505 são consideradas residências e, dessas, somente 471 têm informações completas para todas as variáveis utilizadas no modelo de preços hedônicos. Pelo mesmo motivo, no caso das propriedades atingidas com registro de perda e/ou danos a edificações e

[27] Uma ressalva importante de ser apontada foram algumas incongruências nos fluxos de respostas, por exemplo, existem seis propriedades que constaram "sim" tanto na pergunta *10.10.32.1* como na pergunta *10.10.32.2*, quando na verdade deveria ser um "ou" excludente, de modo que a declaração como "sim" em uma deveria ser "não" na outra. Assim, nos casos em que ambas as classificações foram atribuídas para a mesma propriedade, considerou-se como afetação total com impossibilidade de uso da edificação.

construções civis, das 310 residências presentes no cadastro foram incluídas apenas 291.

Entre as 72 residências com afetação total com impossibilidade de uso da edificação, incluem-se 69 e, no caso das 26 residências com afetação parcial, 23 foram incluídas. Na Figura 5 pode-se observar a disposição espacial do total de propriedades cadastradas no Cadastro da Fundação Renova em Barra Longa.

Figura 6 — Residências com edificação principal em Barra Longa



Fonte: Elaboração própria a partir do Cadastro da Fundação Renova, versão 28/08/2019.

3.2.2 Seleção das variáveis do Cadastro

Os dados do Censo Demográfico de 2010 correspondem à base de dados primária na qual foram estimados os modelos de preços hedônicos. No caso dos dados do Cadastro da Fundação Renova, as informações referentes às moradias foram utilizadas para prever os valores dos imóveis. Como será introduzido na seção 4, a predição para um determinado imóvel, com base nos modelos utilizados, exige que todas as informações utilizadas nos modelos estejam presentes nas residências em questão.

Ilustrativamente, uma variável que será utilizada é o número de cômodos das residências; assim, é possível prever o valor dos imóveis que possuam a variável de número de cômodos do Cadastro da Fundação Renova preenchido. Portanto,

residências sem as informações das variáveis utilizadas não são consideradas, por isso, das 505 residências presentes no Cadastro, estimaram-se os valores para 471.

Outra questão importante é a compatibilização das variáveis, de modo que as informações presentes no Censo e no Cadastro precisam corresponder às mesmas variáveis. Para isso, utilizou-se o conjunto de variáveis do Cadastro que mais se aproximam com as informações disponibilizadas no Censo.

Como muitas variáveis são simples, com pouca variação na pergunta, como número de cômodos, de dormitórios, banheiros exclusivos, é possível realizar essa análise conjunta, sem grandes perdas de interpretação (ver Apêndice B).

3.2.3 Estatísticas descritivas das variáveis do Cadastro

A Tabela 4 apresenta as informações das propriedades cadastradas no Cadastro da Fundação Renova em Barra Longa utilizadas no cálculo do valor dos imóveis por preços hedônicos.

Tabela 4 — Estatísticas descritivas para as variáveis selecionadas para as propriedades residenciais do município de Barra Longa no Cadastro da Fundação Renova

Estatísticas	N	Média	Desvio-Padrão	Min.	Percentil (25)	Percentil (75)	Max.
Número de Cômodos	471	9.76	8.89	2,0	7,0	10,0	101,0
Número de Dormitórios	471	3.47	3.28	1,0	2,0	4,0	40,0
Número de Banheiros de Uso Exclusivo	471	1.45	1.19	1,0	1,0	2,0	10,0
Apartamento	471	0.03	0.18	0,0	0,0	0,0	1,0
Urbano	471	0.70	0.46	0,0	0,0	1,0	1,0
Rede Geral de Esgoto	471	0.90	0.30	0,0	1,0	1,0	1,0
Rede Geral de Água	471	1.00	0.00	1,0	1,0	1,0	1,0
Poço ou Nascente na Prop.	471	0.12	0.33	0,0	0,0	0,0	1,0
Lixo Coletado por Serviço de Limpeza	471	0.92	0.27	0,0	1,0	1,0	1,0
Energia Elétrica	471	1.00	0.05	0,0	1,0	1,0	1,0
Latitude	471	-20.28	0.01	-20.33	-20.29	-20.28	-20.20
Longitude	471	-43.06	0.04	-43.18	-43.06	-43.04	-42.95

Fonte: Elaboração própria a partir do Cadastro da Fundação Renova, versão 28/08/2019.

No município de Barra Longa, 471 residências possuíam todas as informações necessárias para se preverem os valores das habitações com base nos modelos de

preços hedônicos a partir dos dados do Censo Demográfico. Entre as variáveis explicativas, as residências da região de Barra Longa, em média, possuem 9,76 cômodos, sendo 3,47 deles, em média, utilizados como dormitórios e 1,45 utilizado como banheiro de uso exclusivo. As demais variáveis explicativas apresentadas correspondem a um conjunto de informações binárias, atribuindo-se valor igual a um quando certa característica do domicílio está presente, e valor zero do contrário.

A variável “Apartamento” é uma variável binária que corresponde ao valor um quando o domicílio é apartamento e valor zero quando o domicílio é casa. Grande parte dos domicílios na região são casas, com somente 3% dos domicílios amostrados sendo apartamentos. A variável “Urbano” é também uma variável binária que tem valor igual a um quando o domicílio está localizado em um setor censitário urbano e valor zero caso esteja em zona rural. Nota-se que 70% das residências se situam em setores urbanos, uma porcentagem abaixo da verificada na amostra de domicílios com aluguel do Censo (96%).

Para a variável “Rede Geral de Esgoto” atribui-se o valor um para os domicílios que possuem rede geral de esgoto ou pluvial e zero para os casos de fossa séptica, fossa rudimentar, vala, rio, lago ou mar, outro ou em branco; 90% dos domicílios da região amostrada com valores de aluguel fazem lançamento de esgoto na rede geral de esgoto ou pluvial. Analogamente, praticamente 100% desses domicílios têm acesso à rede geral de distribuição para abastecimento de água em pelo menos um cômodo e zero tem poço ou nascente na propriedade.

Para a variável “Lixo Coletado por Serviço de Limpeza” atribuiu-se o valor um para os casos em que o destino do lixo residencial é coletado diretamente por serviço de limpeza e zero para os casos em que o lixo é coletado em caçamba de serviço de limpeza, queimado (na propriedade), enterrado (na propriedade), jogado no terreno baldio ou logradouro, jogado em rio, lago ou mar, tem outro destino, ou ficou em branco. A variável “Energia Elétrica” corresponde ao valor um para os domicílios que possuem energia elétrica fornecida por companhia distribuidora e zero caso tenham energia elétrica por outras fontes ou caso não tenham.

Para os domicílios em questão, 92% têm coleta de lixo residencial realizada diretamente pelo serviço de limpeza e praticamente 100% têm energia fornecida por companhia distribuidora.

4 ESCOLHA DO MODELO DE PREDIÇÃO

Nesta seção, mostra-se o embasamento empírico para o critério de escolha do melhor modelo de predição. Na seção 2 de metodologia apresentaram-se os aspectos gerais dos modelos de estimação de preços hedônicos, assim como suas implicações teóricas. Tendo em vista o conjunto de dados levantados (ver seção 3), nesta seção apresentam-se explicitamente as formas funcionais para os modelos estimados, a estratégia para comparação de poder preditivo dos mesmos e os resultados dos exercícios de comparação utilizando os dados do Censo Demográfico.

Nesse sentido, a subseção 4.1 apresenta a estratégia de comparação dos modelos, as estatísticas utilizadas e a construção das bases de treino e de teste. A seção 4.2 apresenta os modelos testados nas estimações por meio de Mínimos Quadrados Ordinários, por Modelos Hierárquicos e por Florestas Aleatórias. Em sequência, na seção 4.3, discutem-se os resultados de predição de cada especificação e a decisão do modelo preferido em termos de poder de predição a partir dos dados do Censo Demográfico.

4.1 Estratégia para a comparação do poder preditivo dos modelos

Em linha com a literatura de predição, o exercício de comparação do poder preditivo aqui apresentado tem como base as estatísticas de predição calculadas a partir de dados que não foram utilizados para as estimações dos modelos, ou seja, constroem-se duas subamostras: i) base de treino ou de estimação; e ii) base de teste. Intuitivamente, suponha que gostaríamos de estabelecer um modelo de previsão do tempo. O teste para saber qual modelo prevê melhor deve ser aplicado nos dias em que não temos informação, representado pela base de teste. Assim, utilizamos todas as informações das variáveis determinantes para a temperatura até hoje (base de treino) para prever a temperatura de amanhã.

Posto isso, para cada especificação apresentada, elaboraram-se 10 simulações nas quais, em cada uma delas, a base de dados da amostra foi dividida em dois blocos: 90% das observações são utilizadas na estimação dos modelos (base de treino), e os 10% restantes (base de teste) são postos à prova, de modo que para cada modelo aplica-se o valor predito do aluguel nesses domicílios e compara-se com os valores de aluguel declarados no Questionário da Amostra do Censo Demográfico. A escolha dos domicílios que entram na base de treino e de teste em cada uma das 10 simulações é

feita de maneira aleatória, mantendo apenas a semente[28] da simulação fixa para fins de replicabilidade.

Assim, constroem-se as estatísticas de erro de predição para os modelos. A fim de comparação do ajuste amostral fora da amostra, computam-se essas estatísticas para ambas as bases de treino e teste. Entretanto, a escolha do melhor modelo de predição é baseada somente nas estatísticas para a base de treino.

Na literatura de previsão, as estatísticas mais comumente utilizadas para comparar o poder preditivo dos modelos são (ZAMBRANO-BIGIARINI, 2017): i) Erro Médio (ME); ii) Erro Médio Absoluto (MAE); iii) Erro Quadrático Médio (MSE); iv) Raiz Quadrada do Erro Médio (RMSE); v) Percentual de Viés (PBIAS %); e vi) R-quadrado (R^2). Em todos os casos, as siglas são advindas dos termos em inglês. Na sequência, cada uma das estatísticas será explicada em mais detalhes:

I Erro Médio (ME): a estatística ME corresponde à média entre os valores previstos (S_i) e os observados (O_i) para cada modelo, como se pode observar na equação 18 a seguir:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i), \tag{18}$$

onde N corresponde ao número de observações na amostra. Uma propriedade derivada das condições de momentos dos modelos de Mínimos Quadrados Ordinários é que o ME calculado com todas as informações da amostra é zero (WOOLDRIDGE, 2010). Vale reforçar que essa propriedade é válida apenas para os valores previstos dentro da amostra; para os casos de previsão para fora da amostra, como se realizou nos 10%, em cada uma das simulações isso não pode ser afirmado.

II Erro Médio Absoluto (MAE): a estatística MAE corresponde à média absoluta entre os valores previstos (S_i) e os observados (O_i) no modelo, como se pode observar na equação 19 a seguir:

[28] Como se pode notar nos códigos de programação apresentados no apêndice, mantém-se o comando “set.seed()” como fixo, pois assim a sequência de números sorteados é feita de maneira aleatória, porém sempre da mesma maneira, para que se obtenham os mesmos resultados. Esse procedimento é muito comum nos estudos de simulações em diversas áreas da ciência.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|S_i - O_i|, \quad (19)$$

onde N o número de observações na amostra.

III Erro Quadrático Médio (MSE): a estatística MSE, por sua vez, corresponde ao quadrado da média dos valores previstos (S_i) e os observados (O_i) no modelo, como se pode observar na equação 20 a seguir:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2, \quad (20)$$

onde N o número de observações na amostra.

IV Raiz Quadrada do Erro Médio (RMSE): a estatística RMSE é um dos critérios mais utilizados como critério de decisão nos modelos de aprendizado de máquina ("*Machine Learning*" do termo em inglês). Essa estatística corresponde à raiz quadrada do MSE, como se pode observar na equação 21 a seguir:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2} \quad (21)$$

onde S_i é o valor previsto do modelo e O_i o resultado observado, N o número de observações na amostra. Nas etapas de seleção dos hiperparâmetros dos modelos de florestas aleatórias, o conjunto preferido de número de árvores e do número de variáveis consideradas potencialmente (mtry) é obtido do modelo que possui o menor RMSE.

V Percentual de Viés (PBIAS %): a estatística PBIAS %, corresponde ao desvio percentual entre valores previstos (S_i) e os observados (O_i) no modelo, como pode ser observado na equação 22 a seguir:

$$PBIAS = 100 \times \frac{\sum_{i=1}^{N}(S_i - O_i)}{\sum_{i=1}^{N} O_i} \quad (22)$$

Em todas as estatísticas apresentadas (ME, MAE, MSE, RMSE, PBIAS %) até o momento, quanto menor seu valor, maior o poder assertivo do modelo.

VI R-quadrado (R^2): a estatística R^2, também conhecida como coeficiente de determinação, indica o percentual de explicação das variáveis explicativas *vis-à-vis* o total explicado, incluindo os termos de erro. Dados os valores previstos (S_i) e os observados (O_i) no modelo, define-se:

$$\overline{O} = \frac{1}{N}\sum_{i=1}^{N} O_i$$

Então, a Soma dos Quadrados Totais (SQT) será:

$$SQT = \sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2$$

E a Soma dos Quadrados Explicados (SQE):

$$SQE = \sum_{i=1}^{N}\left(S_i - \overline{O}\right)^2$$

Assim, o R-quadrado pode ser calculado da seguinte maneira:

$$R^2 = 1 - \frac{SQE}{SQT} \qquad (23)$$

4.2 Apresentação dos modelos

Na perspectiva de selecionar o modelo com maior poder preditivo para os dados fora da amostra, esta subseção contém as especificações de 13 modelos separados em três grandes grupos: os modelos de Mínimos Quadrados Ordinários, os modelos Hierárquicos e o modelo de Florestas Aleatórias.

No caso dos dois primeiros grupos, há uma subdivisão da forma funcional, sendo uma parte estimada no formato nível-nível, com a variável de aluguel na sua forma original como variável dependente, e outra no formato log-nível, sendo o logaritmo neperiano do aluguel ($\ln(aluguel)$) como a forma funcional. Essa especificação segue a literatura de preços hedônicos e sugere que a distribuição dos fatores idiossincráticos, os termos de erro ϵ, seja log-normal (AGUIRRE e DE FARIA, 1997). Utilizar essa segunda forma funcional também é uma estratégia para trazer não linearidade ao modelo. Para o método de Florestas Aleatórias, não é necessária essa transformação logarítmica, pois o método consegue capturar formas mais abrangentes de não linearidade.

4.2.1 Modelos de Mínimos Quadrados Ordinários

O grupo de modelos estimados pelo método de Mínimos Quadrados Ordinários é identificado pela sigla OLS, que varia de OLS 1 a OLS 6. Seguindo as especificações dos modelos OLS 1 a OLS 3, a variável $aluguel$ em nível é considerada dependente. Os três modelos possuem as variáveis comuns, sendo elas:

I O número de Cômodos: $Cômodos$;

II O número de Dormitórios: $Dormitórios$;

III O número de banheiros de uso exclusivo: $Banheiros\ Exclus$;

IV A variável indicadora[29] se o domicílio é apartamento: $Apartamento$;

V A variável indicadora se o domicílio está em setores censitários urbanos: $Urbano$;

VI A variável indicadora se o domicílio possui Rede Geral de Esgoto: $Esgoto$;

VII A variável indicadora se o domicílio possui Rede Geral de Água: $Água\ e\ Dist$;

VIII A variável indicadora se o domicílio possui Poço ou nascente na propriedade: $Poço$;

IX A variável indicadora se Lixo Coletado é por Serviço de Limpeza: $Lixo$;

X A variável indicadora se o domicílio possui energia elétrica: $Energia$;

XI Latitude da área da ponderação: $Latitude$;

XII Longitude da área da ponderação: $Longitude$.

[29] Uma variável indicadora é uma variável binária que assume o valor igual a 1 se determinada condição é satisfeita e 0, caso contrário.

Como se pode verificar na equação 24, o modelo OLS 1 contém apenas essas variáveis explicativas supracitadas:

OLS 1:

$$\begin{aligned} aluguel_d = {} & \beta_0 + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d \\ & + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d \\ & + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d \\ & + \beta_{12} Longitude_d + \epsilon_d \end{aligned} \quad (24)$$

O modelo OLS 2 da equação 25, por sua vez, contém as variáveis citadas e acrescenta as variáveis de interação entre o componente da latitude com a variável indicadora de urbano ($Urbano \times Latitude$), assim como a interação entre a longitude e urbano ($Urbano \times Longitude$). Incluem-se essas variáveis para testar de formas alternativas o efeito da localização das moradias.

OLS 2:

$$\begin{aligned} aluguel_d = {} & \beta_0 + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d \\ & + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d \\ & + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d \\ & + \beta_{12} Longitude_d + \beta_{13} Urbano \times Latitude_d + \beta_{14} Urbano \\ & \times Longitude_d + \epsilon_d \end{aligned} \quad (25)$$

O modelo OLS 3 da equação 26, por sua vez, contém as variáveis do OLS 2 e os termos quadráticos das variáveis cômodos, dormitórios e banheiros de uso exclusivo. A inclusão desses termos quadráticos busca trazer não linearidade ao modelo. A intuição para esses termos é que o aumento no valor do aluguel com o acréscimo de um dormitório pode ser cada vez menor, pois, a partir de um determinado número de dormitórios, há menor necessidade de uma unidade adicional. Isto é, as pessoas estariam dispostas a pagar um adicional menor por um dormitório adicional a partir desse ponto.

OLS 3:

$$aluguel_d = \beta_0 + \beta_1 Cômodos_d + \beta_2 Cômodos_d^2 + \beta_3 Dormitórios_d + \beta_4 Dormitórios_d^2 + \beta_5 Banheiros\ Exclus_d + \beta_6 Banheiros\ Exclus_d^2 + \beta_7 Apartamento_d + \beta_8 Urbano_d + \beta_9 Esgoto_d + \beta_{10} Água\ e\ Dist_d + \beta_{11} Poço_d + \beta_{12} Lixo_d + \beta_{11} Energia_d + \beta_{12} Latitude_d + \beta_{13} Longitude_d + \beta_{14} Urbano \times Latitude_d + \beta_{15} Urbano \times Longitude_d + \epsilon_d \quad (26)$$

O modelo OLS 4 da equação 27 contém as mesmas variáveis explicativas do modelo OLS 1. A única diferença é o fato de a variável explicada ser $\ln(aluguel)$, ou seja, ser log linearizada. Cabe destacar aqui que a previsão do aluguel utilizando modelos log-linearizadas tem suas peculiaridades. Essa explicação está descrita no Apêndice B.

OLS 4:

$$\ln(aluguel)_d = \beta_0 + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d + \beta_{12} Longitude_d + \epsilon_d \quad (27)$$

O modelo OLS 5 da equação 28 contém as mesmas variáveis explicativas do modelo OLS 2. Novamente, com a diferença de a variável explicada ser $\ln(aluguel)$.

OLS 5:

$$\ln(aluguel)_d = \beta_0 + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d + \beta_{12} Longitude_d + \beta_{13} Urbano \times Latitude_d + \beta_{14} Urbano \times Longitude_d + \epsilon_d \quad (28)$$

Por fim, o modelo OLS 6 da equação 29 contém as mesmas variáveis explicativas do modelo OLS 3. Novamente, com a diferença de a variável explicada ser $\ln(aluguel)$.

OLS 6:

$$\begin{aligned}\ln(aluguel)_d = {} & \beta_0 + \beta_1 Cômodos_d + \beta_2 Cômodos_d^2 + \beta_3 Dormitórios_d \\ & + \beta_4 Dormitórios_d^2 + \beta_5 Banheiros\ Exclus_d \\ & + \beta_6 Banheiros\ Exclus_d^2 + \beta_7 Apartamento_d + \beta_8 Urbano_d \\ & + \beta_9 Esgoto_d + \beta_{10} Água\ e\ Dist_d + \beta_{11} Poço_d + \beta_{12} Lixo_d \\ & + \beta_{11} Energia_d + \beta_{12} Latitude_d + \beta_{13} Longitude_d + \beta_{14} Urbano \\ & \times Latitude_d + \beta_{15} Urbano \times Longitude_d + \epsilon_d\end{aligned} \quad (29)$$

4.2.2 Modelos Hierárquicos

Os modelos hierárquicos, conforme introduzidos na seção 2 de metodologia, se diferem dos mínimos quadrados em alguns sentidos, sobretudo, no fato de cada município ter um intercepto diferente. Isso traz mais flexibilidade para o modelo se ajustar aos dados. No que tange às variáveis explicativas, os números dos modelos subsequentes seguem a mesma ordem dos mínimos quadrados, ou seja, as variáveis explicativas do modelo "multi-level" ML 1 equivale à do OLS 1, a ML 2 tem as variáveis semelhantes a OLS 2 e, por último, ML 3 é equivalente a OLS 3.

ML 1:

$$\begin{aligned}aluguel_d = {} & \beta_{0m[d]} + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d \\ & + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d \\ & + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d \\ & + \beta_{12} Longitude_d + \epsilon_d\end{aligned} \quad (30)$$

ML 2:

$$aluguel_d = \beta_{dm[d]} + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d + \beta_{12} Longitude_d + \beta_{13} Urbano \times Latitude_d + \beta_{14} Urbano \times Longitude_d + \epsilon_d \quad (31)$$

ML 3:

$$aluguel_d = \beta_{0m[d]} + \beta_1 Cômodos_d + \beta_2 Cômodos_d^2 + \beta_3 Dormitórios_d + \beta_4 Dormitórios_d^2 + \beta_5 Banheiros\ Exclus_d + \beta_6 Banheiros\ Exclus_d^2 + \beta_7 Apartamento_d + \beta_8 Urbano_d + \beta_9 Esgoto_d + \beta_{10} Água\ e\ Dist_d + \beta_{11} Poço_d + \beta_{12} Lixo_d + \beta_{11} Energia_d + \beta_{12} Latitude_d + \beta_{13} Longitude_d + \beta_{14} Urbano \times Latitude_d + \beta_{15} Urbano \times Longitude_d + \epsilon_d \quad (32)$$

onde $\beta_{0m[d]}$ representa o intercepto do município m onde a moradia d está localizada.

Novamente, a diferença dos modelos ML 4 a ML 6 a seguir é a log-linearização da variável explicada, que nos seguintes casos é $\ln(aluguel)$.

ML 4:

$$\ln(aluguel)_d = \beta_{0m[d]} + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d + \beta_{12} Longitude_d + \epsilon_d \quad (33)$$

ML 5:

$$\ln(aluguel)_d = \beta_{0m[d]} + \beta_1 Cômodos_d + \beta_2 Dormitórios_d + \beta_3 Banheiros\ Exclus_d + \beta_4 Apartamento_d + \beta_5 Urbano_d + \beta_6 Esgoto_d + \beta_7 Água\ e\ Dist_d + \beta_8 Poço_d + \beta_9 Lixo_d + \beta_{10} Energia_d + \beta_{11} Latitude_d + \beta_{12} Longitude_d + \beta_{13} Urbano \times Latitude_d + \beta_{14} Urbano \times Longitude_d + \epsilon_d \quad (34)$$

ML 6:

$$\ln(aluguel)_d = \beta_{0m[d]} + \beta_1 Cômodos_d + \beta_2 Cômodos_d^2 + \beta_3 Dormitórios_d + \beta_4 Dormitórios_d^2 + \beta_5 Banheiros\ Exclus_d + \beta_6 Banheiros\ Exclus_d^2 + \beta_7 Apartamento_d + \beta_8 Urbano_d + \beta_9 Esgoto_d + \beta_{10} Água\ e\ Dist_d + \beta_{11} Poço_d + \beta_{12} Lixo_d + \beta_{11} Energia_d + \beta_{12} Latitude_d + \beta_{13} Longitude_d + \beta_{14} Urbano \times Latitude_d + \beta_{15} Urbano \times Longitude_d + \epsilon_d \quad (35)$$

4.2.3 Modelo de Florestas Aleatórias

Introduzidos na seção 2, os modelos de Florestas Aleatórias não partem de uma hipótese sobre o processo gerador de dados, tais como nos modelos de Mínimos Quadrados Ordinário e nos Hierárquicos. Para a estimação nesse caso, é necessário preestabelecer um conjunto de hiperparâmetros (KUHN; 2008), sendo eles:

I Número de árvores;

II Número de variáveis consideradas potencialmente em cada nó; e

III Número mínimo de observações ao final de cada ramo.

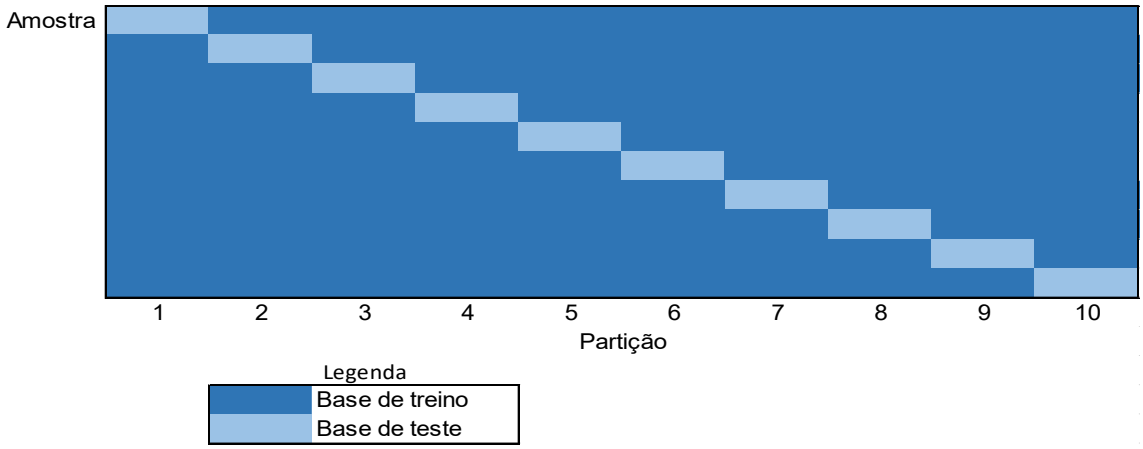
No seguinte exercício[30], foram simulados modelos para determinar esses três hiperparâmetros. Assim, foram estimadas todas as combinações possíveis do número de árvores variando de 1.000 a 5.000, com intervalo de 50; o número de variáveis consideradas potencialmente em cada nó variando de 1 a 12; e três tamanhos para o conjunto de dados do final de cada ramo (5, 10 e 15). Dessa forma, ao todo, foram

[30] Para esse exercício foi utilizado o pacote Caret de Kuhn (2008) no software R Project.

testados 324 tipos de combinações de hiperparâmetros. Para a escolha dos hiperparâmetros foi utilizada a estatística RMSE (equação 21).

Para esse exercício, utilizou-se a técnica de validação cruzada de K partições repetidas (em inglês, *Reapeated K-fold cross-validation*). Essa técnica é baseada em dividir a amostra em K pedaços de tamanhos semelhantes, como mostrado na Figura 7. Define-se K=10. Para cada partição, gera-se uma floresta aleatória utilizando a base de treino e computa-se o RMSE utilizando a base de teste. O RMSE final é a média do RMSE das 10 partições. Para dar mais robustez ao exercício, repete-se a validação cruzada três vezes, de modo que o RMSE final é a média de 30 indicadores.

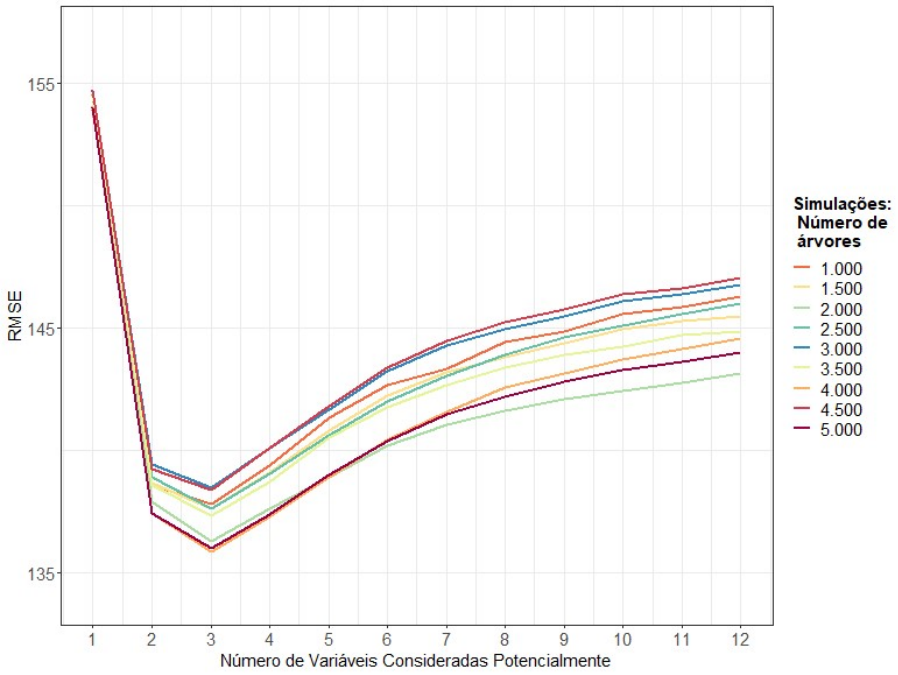
Figura 7 — Validação cruzada de K partições repetidas



Fonte: Elaboração própria (2019).

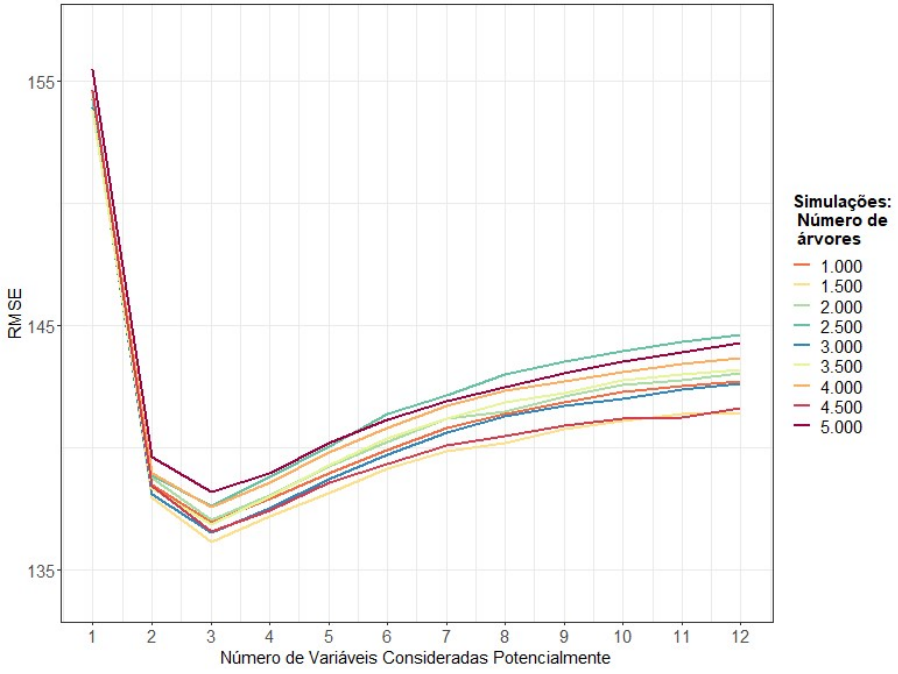
A Figura 8, a Figura 9 e a Figura 10 apresentam o RMSE para os diferentes números de variáveis selecionadas potencialmente (mtry) e número de árvores (ntree), respectivamente para o número mínimo de observações 5, 10 e 15. Observa-se um padrão semelhante nas três figuras. O número de variáveis selecionadas potencialmente em cada nó que foi selecionado para todas as combinações dos outros dois hiperparâmetros é mtry=3.

Figura 8 — RMSE para as simulações das Florestas Aleatórias com ao menos cinco observações em cada nó



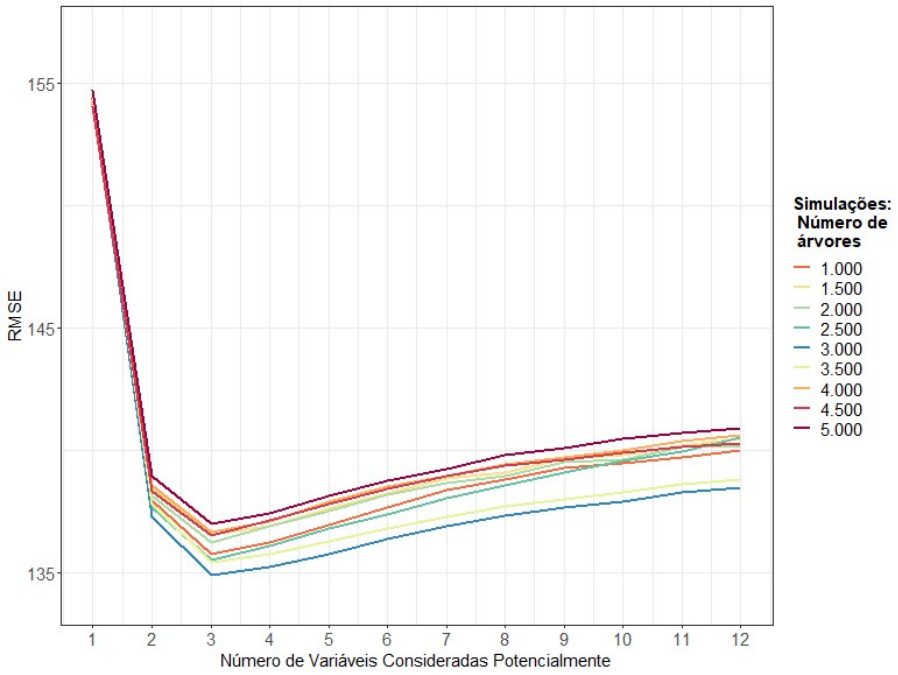
Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

Figura 9 — RMSE para as Simulações das Florestas Aleatórias com ao menos 10 observações em cada nó



Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

Figura 10 — RMSE para as Simulações das Florestas Aleatórias com ao menos 15 observações em cada nó



Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

O procedimento de escolha do número de variáveis potencialmente escolhidas é realizado fixando o número fixo de árvores e o número mínimo de observações em cada nó final. Assim, a partir de um conjunto de simulações, variando o número de árvores e variando o número de observações no nó, obteve-se o modelo preferido, sob a ótica de resultar em um menor RMSE como sugerido pela literatura (KUHN, 2012, p.201; KUHN, 2008). Os hiperparâmetros escolhidos estão resumidos a seguir:

Valores dos hiperparâmetros escolhidos para o modelo de Florestas Aleatórias com os dados do Censo Demográfico de 2010:

- Número de árvores: 3.000;
- Número de variáveis consideradas potencialmente em cada árvore: 3;
- Número mínimo de observações em cada nó: 15.

Esses valores serão utilizados na próxima seção para comparar o poder de predição dos diferentes modelos apresentados na subseção 4.2.

4.3 Escolha do melhor modelo

Nesta seção, compara-se o poder de predição dos diferentes modelos apresentados utilizando as estatísticas e estratégias mostradas na subseção anterior. Os valores das estatísticas de comparação para a base de treino e, mais importante, a base de teste estão na Tabela 5. A escolha dos modelos deve ser ponderada pelas suas condições de assertividade, como os indicadores ME, MAE, MSE, RMSE, PBIAS % — que, quanto menor forem os resultados, melhor para a predição — e nos indicadores de qualidade de ajustamento[31] tais como o R² que, quanto maior, melhor. Em caso de grande semelhança entre os resultados, seguem-se também os critérios de informação (BOZDOGAN, 1987).

Inicia-se analisando o desempenho dentro da amostra, na base de treino. Observa-se que não há um modelo que é selecionado por todas as estatísticas calculadas. Entretanto, o método de Florestas Aleatórias é selecionado por quatro indicadores: MAE, MSE, RMSE e R^2. Para o PBIAS %, esse método apresentou segundo menor valor, atrás dos modelos nível-nível. Selecionando o modelo pelo seu desempenho dentro da amostra, há fortes indícios de superioridade do método de Florestas Aleatórias. Entretanto, neste estudo, seu desempenho fora da amostra é mais importante e não necessariamente aquele modelo que apresenta melhor ajuste interno é o que melhor prevê fora da amostra.

Analisando as estatísticas de erro de previsão fora da amostra (base de treino), observa-se um padrão parecido com o anterior, porém com mais indícios a favor do método de Florestas Aleatórias. Esse método foi selecionado por todas as estatísticas utilizadas. Em termos absolutos, o erro médio é bem abaixo em relação aos outros modelos. Por exemplo, o RMSE para os modelos nível é em torno de 145, para os modelos log-linear é, aproximadamente, 153, e para o Método de Florestas Aleatórias é 120. O mesmo se observa para o R^2. Para todos os modelos (nível-nível e log-nível) o R^2 é próximo de 0,45 e para o Método de Florestas Aleatórias é 0,64. Lembrando que para esse indicador, quanto maior o valor, melhor é sua previsão. Esse resultado mostra que há importantes relações entre as variáveis explicativas e o valor do aluguel que não são capturadas pelos métodos lineares.

Dessa forma, entre os modelos testados, há fortes indícios de superioridade do método de Florestas Aleatórias para a previsão do valor do aluguel dos dados do Censo Demográfico 2010. Portanto, esse é o método selecionado para prosseguir com a

[31] "*Goodness-of-fit*", do termo em inglês.

previsão dos aluguéis e, posteriormente, os valores dos imóveis cadastrados pela Fundação Renova.

Tabela 5 — Valor previsto dentro e fora da amostra

	90% (Estimado)						10% (Fora da Amostra)					
Modelo	**ME**	**MAE**	**MSE**	**RMSE**	**PBIAS %**	**R²**	**ME**	**MAE**	**MSE**	**RMSE**	**PBIAS %**	**R²**
Modelos Nível-Nível												
OLS 1	0,0	91,7	19.508	139,7	0,0	0,45	-2,0	93,0	22.135	147,8	-0,6	0,44
OLS 2	0,0	91,7	19.461	139,5	0,0	0,45	-1,9	92,9	22.050	147,5	-0,6	0,44
OLS 3	0,0	91,4	18.962	137,7	0,0	0,46	-1,8	93,1	22.697	149,3	-0,6	0,43
ML 1	0,0	84,3	17.505	132,3	0,0	0,51	-2,3	87,3	20.467	142,1	-0,8	0,48
ML 2	0,0	84,3	17.452	132,1	0,0	0,51	-2,2	87,2	20.364	141,7	-0,8	0,48
ML 3	0,0	84,1	17.011	130,4	0,0	0,52	-2,1	87,7	21.175	144,1	-0,7	0,47
Modelos Log-Nível												
OLS 4	3,8	91,1	20.059	141,6	1,4	0,44	2,7	92,7	25.711	156,4	1,1	0,41
OLS 5	3,7	91,0	19.978	141,3	1,4	0,44	2,7	92,7	25.557	156,0	1,1	0,41
OLS 6	2,5	89,8	19.272	138,8	0,9	0,46	0,8	92,0	23.899	153,2	0,4	0,40
ML 4	3,8	83,7	18.545	136,1	1,4	0,49	2,9	87,1	26.576	156,4	1,2	0,45
ML 5	3,7	83,5	18.427	135,7	1,4	0,49	2,8	87,0	26.319	155,7	1,2	0,45
ML 6	2,4	82,1	17.221	131,2	0,9	0,51	0,6	86,1	22.461	148,3	0,3	0,44
Modelos RF												
RF	0,4	73,7	12.776	113,0	0,1	0,66	-1,6	75,7	14.750	120,8	-0,5	0,64

Fonte: Elaboração própria (2019), a partir de dados do IBGE (2010).
Nota: A seleção dos 90% utilizados nos modelos e dos 10% que foram utilizados para a previsão fora da amostra se deu de maneira aleatória. A construção de cada um dos indicadores apresentados (ME, MAE, MSE, RMSE, PBIAS% e R²) está na seção 4.1.

5 RESULTADOS

Nesta seção apresentam-se os valores agregados estimados das propriedades residenciais cadastradas no Cadastro da Fundação Renova em Barra Longa com base no modelo de previsão por preços hedônicos escolhido na seção anterior. A primeira etapa do cálculo do valor dos imóveis consiste em prever o valor de aluguel, caso fosse alugado, a preços de agosto de 2010, pois os modelos foram estimados a partir dos dados do Censo Demográfico de 2010 (IBGE, 2010).

O procedimento subsequente trata de ajustar os valores previstos dos aluguéis para preços de outubro de 2015, mês anterior ao rompimento da Barragem de Fundão. Por fim, utilizam-se taxas de desconto para realizar o cálculo do valor das residências, a preços de outubro de 2015.

5.1 Hiperparâmetros do modelo escolhido

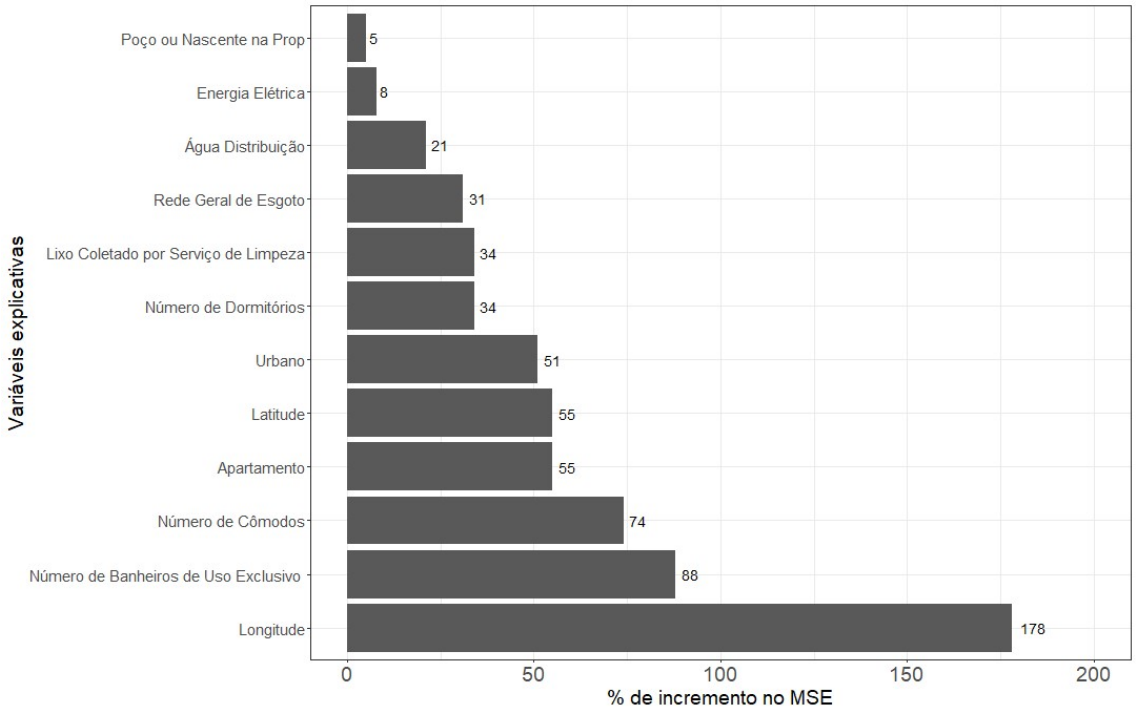
A partir do procedimento de escolha do modelo para previsão apresentado na subseção 4.3, o modelo escolhido corresponde ao de Floresta Aleatória com os seguintes hiperparâmetros:

- Número de árvores: 3.000;
- Número de variáveis consideradas potencialmente em cada árvore: 3;
- Número mínimo de observações em cada nó: 15.

A Figura 11 a seguir apresenta o impacto marginal das variáveis na redução do erro de previsão (RMSE) nos dados do Censo Demográfico. Intuitivamente, cada variável possui um ganho explicativo na predição do valor do aluguel, por exemplo, utilizar a variável número de cômodos pode ajudar a explicar as variações de aluguéis na amostra.

Assim, quanto maior o valor do impacto marginal das variáveis, maior é seu poder de explicação na previsão, a saber, menor o RMSE. Como se pode notar, a variável que possui maior explicação entre as utilizadas é a longitude, de modo que uma má especificação, como a substituição dessa variável por uma sequência aleatória de números, aumenta em 173% o erro de previsão (RMSE). De outra forma, se, por acaso, os domicílios de um local, digamos, Barra Longa, assumissem os valores da longitude dos domicílios de Mariana, o aumento no MSE seria de 173%, o que indica um componente importante para a determinação do aluguel previsto. Essa interpretação é análoga para as demais variáveis apresentadas.

Figura 11 — Impacto marginal das variáveis na redução do erro de previsão

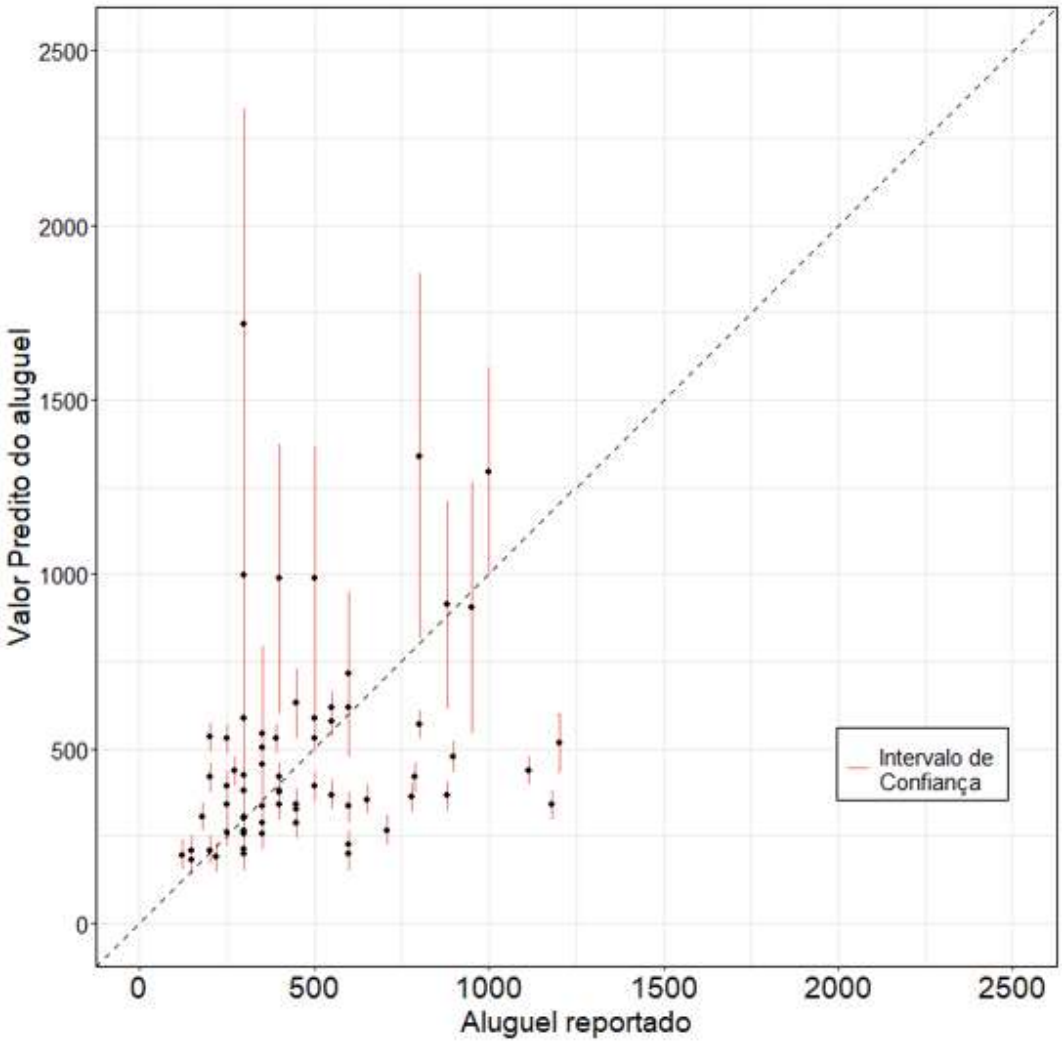


Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova.

5.2 Comparação entre valores de aluguel previstos pelo modelo e declarados no Cadastro

Para fins de comparação, o Cadastro da Fundação Renova possui informações declaradas do valor que os indivíduos pagavam em aluguel, em média, nos últimos 12 meses antes do rompimento da barragem. Dessa forma, estimaram-se os valores preditos para os domicílios cadastrados que estão na lista de propriedades atingidas em Barra Longa com informações de aluguel utilizando a Floresta Aleatória estimada com os dados do Censo Demográfico de 2010. Esse exercício mostra o desempenho do método em outra base de dados. O resultado gráfico pode ser visto na Figura 12, a seguir:

Figura 12 — Valores previstos para os domicílios de Barra Longa do Cadastro da Fundação Renova com os parâmetros do Modelo RF



Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

No gráfico anterior, apresentam-se as comparações entre os valores previstos e os declarados nos casos das propriedades que declararam pagar aluguel nos 12 meses anteriores ao rompimento. No eixo das abcissas estão os valores declarados de aluguéis nas residências de Barra Longa, no eixo das ordenadas se encontram os valores previstos para o mesmo conjunto de domicílios.

A linha tracejada possui uma inclinação de 45º e corresponde como uma linha de referência, pois esses seriam os resultados cujos valores preditos pelo modelo correspondem exatamente aos valores declarados. Portanto, quanto mais próximos os pontos estiverem dessa linha, menor seria o erro de previsão para essas observações. Uma ressalva importante é que o método de predição foi aplicado nas 471 propriedades enquanto no gráfico apresentado observamos apenas os 68 casos com informações de aluguel presentes no Cadastro.

Assim, observa-se que as previsões estão próximas, tanto para cima quanto para baixo da linha de 45º, de modo que a média dos erros de previsão está próxima de zero. Os intervalos de confiança mostram que o erro-padrão das previsões é heterogêneo, com alguns intervalos muito maiores do que outros. Provavelmente, essa diferença pode ser explicada pelos valores das variáveis explicativas que apresentam valores próximos aos limites dos seus respectivos suportes.

5.3 Valor agregado do estoque habitacional do Cadastro em Barra Longa

Recapitulando os passos da análise até esse ponto, estimaram-se os parâmetros do modelo-base de preços hedônicos com os dados do Censo Demográfico de 2010 para o conjunto de domicílios em áreas de ponderação de até 50 km da sede de Barra Longa. Com esses parâmetros, calcularam-se os valores preditos dos aluguéis dos domicílios presentes no Cadastro da Fundação Renova de Barra Longa a preços de 2010.

Tendo isso em vista, o próximo passo consiste em ajustar os preços dos aluguéis previstos para outubro de 2015, o mês anterior ao rompimento da Barragem de Fundão em Mariana (MG). No curto prazo, a dinâmica dos aluguéis segue reajustes periódicos com base em índices de preços; no caso dos contratos formais de aluguel, com reconhecimento de firma, o reajuste se dá, majoritariamente, com base no Índice de Preços do Mercado (IGPM) ou no Índice Nacional do Custo da Construção (INCC), ambos a cada 12 meses.

É importante ressaltar que, no caso específico de Barra Longa, os contratos formais de aluguel são raros, o que não significa não ter residências alugadas. Em grande medida, os contratos informais são feitos diretamente entre o locador e o locatário, sem intermédio das instituições legais. Entretanto, os reajustes podem não seguir uma periodicidade de exatamente 12 meses, mas por constituírem um mercado de aluguéis, com ofertantes e demandantes, supõe-se que existe um reajuste, e em um período mais longo, como é o caso dos cinco anos entre os dados do Censo e a data do rompimento, admite-se que esse reajuste se aproxime dos indicadores utilizados nos contratos formais.

Para o presente estudo, considerou-se o INCC acumulado de agosto de 2010 a outubro de 2015. A Tabela 6 apresenta a evolução de três índices de preços de referência, sendo o Índice de Preços ao Consumidor Amplo (IPCA), utilizado muitas vezes como o parâmetro da inflação, e os dois já citados, IGPM e INCC.

Tabela 6 — Índices de preços acumulados (IPCA, IGPM e INCC)

Índices	Valores acumulados dos índices de preços (Ago.-10 a Out.-15)
IPCA	32,0%
IGPM	40,2%
INCC	43,8%

Fonte: Elaboração própria a partir de dados IBGE e FGV (2015).

Como se pode observar na tabela, o índice acumulado do INCC de agosto de 2015 a outubro de 2015 foi de 43,8%. De maneira ilustrativa, espera-se que os imóveis alugados por R$ 100,00 em agosto de 2010 devam pagar aluguéis de R$ 143,80 em outubro de 2015.

5.3.1 A escolha da taxa de desconto

As taxas de desconto refletem os preços de bens de consumo ao longo do tempo, em diferentes cenários de taxas de juros, depreciação e valorização do ativo. Por exemplo, nos casos de baixas taxas de juros, o ganho de restringir o consumo presente é pequeno quando comparado com consumir no período atual. Ao contrário, quanto maior as taxas de juros, maior o retorno financeiro do investimento e maior a possibilidade de consumo futuro.

Dessa forma, a taxa de juros tem canais de atuação nas decisões de consumo, investimento e poupança dos agentes da economia. O mercado imobiliário está inserido nesse contexto; assim, do ponto de vista dos imóveis como ativos, os quais poderiam render fluxos mensais de aluguéis, o valor do estoque do capital imobilizado na compra do imóvel leva em consideração a taxa de juros de longo prazo da economia.

No cálculo do valor do imóvel a preços de mercado, vale retomar a fórmula utilizada:

$$K = \frac{A}{r + g - c}$$

De modo que A representa o valor do aluguel pago no mês e o denominador representa a taxa de desconto $(r + g - c)$, onde g é a taxa mensal da depreciação, r é a taxa mensal do juro real e c é a taxa de valorização do imóvel.

Não há consenso para os valores das taxas supracitadas $(r + g - c)$. Elas são dependentes do estado da economia, condições das moradias e preferências de

consumo intertemporais[32]. Tendo em vista que escolhas para r, g e c são determinantes para o valor final das habitações, elaboram-se diferentes cenários de taxas de desconto a fim de minimizar as incertezas dos resultados. Neste estudo são apresentados cenários com taxas de descontos de 3%, 4%, 5% e 6% ao ano.

Taxa de depreciação

A partir da definição do Glossário de Terminologia Básica Aplicável à Engenharia de Avaliações e Perícias do IBAPE/SP, depreciação pode ser interpretada como uma perda de valor de um bem, devido às modificações em seu estado ou qualidade. No caso dos imóveis, por exemplo, a não manutenção do domicílio pode comprometer o valor do imóvel uma vez que o estado de preservação se alterou. A taxa de depreciação de um produto depende de suas características intrínsecas, da finalidade e circunstâncias de uso.

Quanto maior a taxa de depreciação, em um dado período, maior será a perda de valor do produto. Bens de capital como um caminhão que transporta carga excessiva, por exemplo, se depreciam rapidamente conforme o uso, por isso as frotas são renovadas com maior frequência (CASAROTTO FILHO e KOPITTKE, 2010). Por outro lado, para o caso de imóveis residenciais, a taxa de depreciação é razoavelmente pequena, dada a durabilidade da construção.

O valor do coeficiente de depreciação sugerido para “residenciais proletário rústico a médio comercial” (LOPES, 1995) é de 1,5% ao ano. Entretanto, como discutido anteriormente, ela pode ser maior ou menor dependendo das características e uso dos imóveis.

5.3.2 Cenários para os valores agregados

Esta subseção apresenta os resultados das estimativas para os valores agregados das residências de Barra Longa que constam na base de dados de propriedades do Cadastro da Fundação Renova para o município de referência, tendo por base diferentes cenários de taxas de desconto.

Conforme a subseção 3.2.1, a partir das informações obtidas no Cadastro da Fundação Renova, classificou-se o conjunto de propriedades em quatro grupos nos quais foram

[32] É importante também salientar a questão do horizonte temporal do ativo imóvel, pois sua comparação deve ser realizada com outros ativos no mesmo horizonte, como títulos de longo prazo do governo. Uma discussão relacionada é a taxa de desconto social a qual também leva em consideração outros custos impostos à sociedade e que poderiam ser considerados no cômputo do valor presente (DASGUPTA, 2007; NORDHAUS, 2007; STERN e STERN, 2007).

estimados os valores das propriedades residenciais com informações válidas em todas as covariadas do modelo de Floresta Aleatória.

I **Residências cadastradas:** total de propriedades residenciais do Cadastro;

II **Residências atingidas:** propriedades residenciais com declaração de perda/dano nas características referentes a moradia;

III **Residências com afetação total:** propriedades residenciais com afetação total e impossibilidade de uso;

IV **Residências com afetação parcial:** propriedades residenciais com afetação parcial e impossibilidade de uso.

Desta forma, as tabelas a seguir apresentam os resultados agregados das estimativas para todas as propriedades residenciais consideradas em cada grupo em diferentes cenários de taxas de desconto. Cabe ressaltar aqui que estimações pontuais estão sujeitas a diferentes fontes de incerteza, por exemplo, o método, o modelo e a própria base de dados a serem utilizados. Ademais, esse problema é potencializado pela falta de dados disponíveis, haja visto que os valores dos imóveis a preços de mercados são determinados por diferentes dimensões, como discutido anteriormente, e a inclusão de todas é inviável devido à indisponibilidade de informação.

Assim, para contornar esse problema e mitigar o impacto das diferentes fontes de incerteza nas estimações pontuais, utilizou-se a metodologia proposta por Wager, Hastie e Efron (2014) para construir os intervalos de confiança com um nível de confiança de 95% para as estimativas dos valores dos imóveis de Barra Longa. Esse intervalo de confiança significa que a cada 100 casas com as mesmas características, em 95% delas, o valor do aluguel estimado estará dentro dos limites desse intervalo. Desta forma, além dos valores das estimativas pontuais médias presentes, por exemplo, na Tabela 7, as Tabelas 8 e 9 contêm as estimativas dos limites inferiores e superiores dos intervalos de confiança, respectivamente.

Os valores dos limites inferiores correspondem aos menores valores contidos nos intervalos de confiança. Analogamente, os valores dos limites superiores são os maiores valores presentes no mesmo intervalo. Assim, no resultado propriedades residenciais cadastradas, o intervalo de confiança com 95% indica que o valor total do estoque habitacional considerado varia de R$ 61,46 milhões (Tabela 8) a R$ 97,24 milhões (Tabela 9).

Nos demais resultados das tabelas, pode-se observar que, por exemplo, na Tabela 7, consideraram-se 471 propriedades residenciais cadastradas em Barra Longa,

totalizando um montante de R$ 79,35 milhões, a preços de outubro de 2015, no cenário de taxa de desconto de 3% a.a. Com o aumento da taxa de desconto para 6% a.a., o valor agregado do estoque habitacional se reduz para R$ 43,29 milhões.

Mantendo o grupo de análise de propriedades residenciais cadastradas, na Tabela 7, pode-se observar que o valor estimado está diretamente relacionado com a escolha da taxa de desconto. Assim, o aumento dessa taxa, de 3% a.a. para 4% a.a. diminui o valor esperado do estoque habitacional de R$ 79,35 milhões para R$ 61,99 milhões.

Tabela 7 — Valores totais estimados do estoque habitacional do Cadastro em Barra Longa – Valor Médio (preços de out./2015)

Tipos de residências consideradas em Barra Longa	Total de residências consideradas em Barra Longa	Valor total do estoque habitacional — taxa de desconto de 3% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 4% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 5% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 6% a.a. (R$ milhões)
Residências cadastradas	471	79,35	61,99	50,94	43,29
Residências atingidas	291	54,36	42,47	34,90	29,66
Residências afetação total	69	13,61	10,64	8,74	7,43
Residências afetação parcial	23	3,59	2,81	2,31	1,96

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.

Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

Tabela 8 — Valores totais estimados do estoque habitacional do Cadastro em Barra Longa – Limite Inferior (preços de out./2015)

Tipos de residências consideradas em Barra Longa	Total de residências consideradas em Barra Longa	Valor total do estoque habitacional — taxa de desconto de 3% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 4% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 5% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 6% a.a. (R$ milhões)
Residências cadastradas	471	61,46	48,01	39,46	33,53
Residências atingidas	291	41,47	32,40	26,63	22,63
Residências afetação total	69	10,13	7,92	6,51	5,53
Residências afetação parcial	23	2,88	2,25	1,85	1,57

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.

Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

Tabela 9 — Valores totais estimados do estoque habitacional do Cadastro em Barra Longa – Limite Superior (preços de out./2015)

Tipos de residências consideradas em Barra Longa	Total de residências consideradas em Barra Longa	Valor total do estoque habitacional — taxa de desconto de 3% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 4% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 5% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 6% a.a. (R$ milhões)
Residências cadastradas	471	97,24	75,97	62,43	53,06
Residências atingidas	291	67,25	52,54	43,17	36,69
Residências afetação total	69	17,10	13,36	10,98	9,33
Residências afetação parcial	23	4,31	3,37	2,77	2,35

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.

Nota: $\boldsymbol{taxa\ de\ desconto}\ (\boldsymbol{d}) = \boldsymbol{taxa\ de\ juros}\ (\boldsymbol{r}) + \boldsymbol{taxa\ de\ depreciação}\ (\boldsymbol{g}) - \boldsymbol{taxa\ de\ valorização}\ (\boldsymbol{c})$

5.3.3 Construção de nova Floresta Aleatória com dados do Cadastro

A título de comparação, construiu-se a Floresta Aleatória utilizando os dados de aluguéis e as características das propriedades residenciais do Cadastro da Fundação Renova que se situam em todo o estado de Minas Gerais.

A partir do procedimento de escolha para os dados do Cadastro, o modelo escolhido corresponde ao de Floresta Aleatória com os seguintes hiperparâmetros:

- Número de árvores: 2.000;
- Número de variáveis consideradas potencialmente em cada árvore: 2;
- Número mínimo de observações em cada nó: 15.

Analogamente à seção anterior, os resultados estão nas Tabelas 10, 11 e 12 para os valores Médio, Limite Inferior e Superior, respectivamente. Nota-se que os valores encontrados são relativamente próximos àqueles estimados utilizando para a construção da Floresta Aleatória os dados do Censo Demográfico 2010.

O valor agregado do estoque habitacional considerado utilizando essa nova Floresta Aleatória, construída a partir dos dados do Cadastro da Fundação Renova, é de R$ 88,71 milhões com limite inferior e superior de R$ 60,57 milhões e R$ 116,85 milhões, respectivamente, utilizando a taxa de desconto de 3% a.a. No cenário com taxa de desconto de 6% a.a. o valor agregado estimado é de R$ 48,40 milhões para os valores médios com limites inferiores e superiores de R$ 33,05 milhões e R$ 63,76 milhões, respectivamente.

Tabela 10 — Valores totais estimados com Floresta Aleatória derivada dos dados do Cadastro – Valor Médio (preços de out./2015)

Tipos de residências consideradas em Barra Longa	Total de residências consideradas em Barra Longa	Valor total do estoque habitacional — taxa de desconto de 3% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 4% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 5% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 6% a.a. (R$ milhões)
Residências cadastradas	471	88,71	69,30	56,95	48,40
Residências atingidas	291	57,57	44,97	36,96	31,41
Residências afetação total	69	13,55	10,58	8,70	7,39
Residências afetação parcial	23	4,38	3,42	2,81	2,39

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.

Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

Tabela 11 — Valores totais estimados com Floresta Aleatória derivada dos dados do Cadastro – Limite Inferior (preços de out./2015)

Tipos de residências consideradas em Barra Longa	Total de residências consideradas em Barra Longa	Valor total do estoque habitacional — taxa de desconto de 3% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 4% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 5% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 6% a.a. (R$ milhões)
Residências cadastradas	471	60,57	47,32	38,89	33,05
Residências atingidas	291	39,82	31,11	25,57	21,73
Residências afetação total	69	9,62	7,52	6,18	5,25
Residências afetação parcial	23	3,00	2,34	1,93	1,64

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.

Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

Tabela 12 — Valores totais estimados com Floresta Aleatória derivada dos dados do Cadastro – Limite Superior (preços de out./2015)

Tipos de residências consideradas em Barra Longa	Total de residências consideradas em Barra Longa	Valor total do estoque habitacional — taxa de desconto de 3% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 4% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 5% a.a. (R$ milhões)	Valor total do estoque habitacional — taxa de desconto de 6% a.a. (R$ milhões)
Residências cadastradas	471	116,85	91,29	75,02	63,76
Residências atingidas	291	75,31	58,83	48,35	41,09
Residências afetação total	69	17,48	13,65	11,22	9,54
Residências afetação parcial	23	5,77	4,51	3,70	3,15

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.

Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

5.4 Simulações de valores para residências típicas em Barra Longa

A partir do modelo de Floresta Aleatória estimado com os dados do Censo Demográfico, a predição para residências que estão fora da amostra necessita do mesmo conjunto de variáveis utilizadas. Para ilustrar o procedimento realizado no caso da predição dos imóveis residenciais do Cadastro da Fundação Renova, apresentam-se nesta subseção alguns resultados para residências típicas nessa mesma base de dados.

O exercício aqui proposto se propõe a fornecer valores representativos de imóveis residenciais com diferentes características. Assim, a tabela 13 apresenta características de quatro residências hipotéticas. É importante salientar que essas informações se basearam nas distribuições de cada variável no Cadastro, porém não correspondem a nenhum caso declarado.

Tabela 13 — Variáveis para a simulação de residências típicas em Barra Longa

Variáveis utilizadas	Residência Típica 1	Residência Típica 2	Residência Típica 3	Residência Típica 4
Número de Cômodos	8	10	8	7
Número de Dormitórios	2	3	2	1
Número de Banheiros de Uso Exclusivo	1	2	1	1
Apartamento	0	0	0	0
Urbano	1	1	0	0
Rede Geral de Esgoto	1	1	1	1
Rede Geral de Água	1	1	1	1
Poço ou Nascente na Prop.	0	0	0	1
Lixo Coletado por Serviço de Limpeza	1	1	1	0
Energia Elétrica	1	1	1	1
Latitude	-20,285	-20,285	-20,257	-20,303
Longitude	-43,043	-43,031	-43,123	-43,001

Fonte: Elaboração própria (2019). Valores meramente ilustrativos, não correspondentes às residências do Cadastro da Fundação Renova.

A Residência Típica 1, por exemplo, indica um imóvel residencial com oito cômodos, entre eles, dois dormitórios e um banheiro exclusivo. A variável indicadora Apartamento, que corresponde ao valor um se o imóvel é um apartamento, indica que é uma casa. No que tange ao setor censitário, o imóvel está em uma região urbana, por isso a variável indicadora Urbano equivale a 1, com Rede Geral de Esgoto e de Água, ambas as variáveis indicadoras são 1, sem presença de poço ou nascente na propriedade, com

Coleta de Lixo, Energia Elétrica, e localização (Latitude e Longitude) próxima à sede municipal do município.

Uma ressalva importante é que as variáveis de coordenadas geográficas, Latitude e Longitude, possuem seis dígitos, mas foram apresentadas na tabela apenas com um dígito, para evitar a identificação exata do ponto de referência, por isso, na aproximação, os valores aparentam ser os mesmos.

Essa interpretação pode ser feita de maneira análoga para as outras três residências típicas restantes. Em linhas gerais, a Residência Típica 1 indica uma residência média, em área urbana e próxima à sede municipal, a Residência Típica 2 simula um imóvel maior, com mais cômodos, dormitórios e banheiros, também na área urbana, porém mais afastado da sede.

A Residência Típica 3, por sua vez, corresponde a um imóvel com características semelhantes à Residência Típica 1, com exceção de estar em uma zona rural mais afastada da sede municipal. Por fim, a Residência Típica 4 corresponde a um imóvel menor, em número de cômodos, dormitórios e banheiros que se situa em zona rural.

Os resultados estimados para os preços a valores de mercado nesses casos podem ser analisados nas Tabelas 14, 15 e 16 a seguir. Como se pode observar na Tabela 14, sob uma taxa de desconto de 3% a.a., no Cenário do Valor Médio, a Residência Típica 1, espera-se um valor de R$ 140.132, o que seria acima das Residências Típicas 3 e 4 que estão em zonas rurais, com R$ 110.068 e R$ 91.308, respectivamente, e abaixo do valor da Residência Típica 2, com valor esperado de R$ 237.873.

Nota-se, novamente, que na adoção de taxas de desconto mais altas, como nos cenários de 4% a.a. a 6% a.a., os valores esperados dos imóveis residenciais são menores. Adicionalmente, a Tabela 15 apresenta os resultados dos cenários calculados a partir do Limite Inferior dos intervalos de confiança, enquanto a Tabela 16 corresponde aos valores estimados no Limite Superior de cada caso.

Assim, pode-se dizer que a Residência Típica 1 teria um valor de mercado, a preços de outubro de 2015, sob uma taxa de desconto de 3% a.a. entre R$ 123.859 (Tabela 15) e R$ 156.405 (Tabela 16).

Tabela 14 — Valores totais estimados de residências típicas do Cadastro em Barra Longa – Cenário do Valor Médio (preços de out./2015)

Residências típicas consideradas em Barra Longa	Valor esperado da residência — taxa de desconto de 3% a.a.	Valor esperado da residência — taxa de desconto de 4% a.a.	Valor esperado da residência — taxa de desconto de 5% a.a.	Valor esperado da residência — taxa de desconto de 6% a.a.
Residência típica 1	R$ 140.132	R$ 105.568	R$ 84.829	R$ 71.002
Residência típica 2	R$ 237.873	R$ 179.201	R$ 143.996	R$ 120.524
Residência típica 3	R$ 110.068	R$ 82.919	R$ 66.629	R$ 55.769
Residência típica 4	R$ 91.308	R$ 68.787	R$ 55.273	R$ 46.263

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.
Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

Tabela 15 — Valores totais estimados de residências típicas do Cadastro em Barra Longa – Cenário do Limite Inferior (preços de out./2015)

Residências típicas consideradas em Barra Longa	Valor esperado da residência — taxa de desconto de 3% a.a.	Valor esperado da residência — taxa de desconto de 4% a.a.	Valor esperado da residência — taxa de desconto de 5% a.a.	Valor esperado da residência — taxa de desconto de 6% a.a.
Residência típica 1	R$ 123.859	R$ 93.309	R$ 74.978	R$ 62.757
Residência típica 2	R$ 221.433	R$ 166.816	R$ 134.044	R$ 112.195
Residência típica 3	R$ 93.946	R$ 70.774	R$ 56.870	R$ 47.600
Residência típica 4	R$ 75.067	R$ 56.552	R$ 45.442	R$ 38.035

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.
Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

Tabela 16 — Valores totais estimados de residências típicas do Cadastro em Barra Longa – Cenário do Limite Superior (preços de out./2015)

Residências típicas consideradas em Barra Longa	Valor esperado da residência — taxa de desconto de 3% a.a.	Valor esperado da residência — taxa de desconto de 4% a.a.	Valor esperado da residência — taxa de desconto de 5% a.a.	Valor esperado da residência — taxa de desconto de 6% a.a.
Residência típica 1	R$ 156.405	R$ 117.828	R$ 94.680	R$ 79.247
Residência típica 2	R$ 254.313	R$ 191.586	R$ 153.948	R$ 128.854
Residência típica 3	R$ 126.189	R$ 95.065	R$ 76.389	R$ 63.937
Residência típica 4	R$ 107.549	R$ 81.022	R$ 65.104	R$ 54.492

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010) e do Cadastro da Fundação Renova, versão 28/08/2019.
Nota: $\boldsymbol{taxa\ de\ desconto\ (d) = taxa\ de\ juros\ (r) + taxa\ de\ depreciação\ (g) - taxa\ de\ valorização\ (c)}$

6 CONCLUSÕES

Entre os impactos econômicos significativos de desastres naturais ou tecnológicos está a destruição de parte do estoque de riqueza existente nas regiões atingidas. Esse estoque de riqueza é composto, entre outros, de recursos naturais, infraestrutura física, capital físico, bens duráveis etc. A mensuração do estoque de riqueza afetado é parte do diagnóstico socioeconômico e o presente estudo avança nessa direção ao mensurar a extensão do estoque habitacional potencialmente afetado pelo rompimento da Barragem de Fundão. A razão para essa escolha inicial é porque o imóvel residencial se constitui no principal ativo físico que em geral as pessoas possuem ao longo da vida. Para tal, estima-se o valor do estoque habitacional no momento anterior ao rompimento da Barragem de Fundão, chamado aqui de valor na linha de base para um dos municípios atingidos, Barra Longa. Além disso, também são apresentados os valores médios de algumas casas típicas ou representativas da região.

Os valores dos imóveis são modelados a partir dos modelos de preços hedônicos comumente utilizados na literatura especializada. Os modelos de preços hedônicos são adequados nesse contexto pois eles consideram bens diferenciados que têm múltiplas características e que o valor de um bem diferenciado é dado pelos valores dessas características. Esse é precisamente o caso de imóveis habitacionais. Além disso, o imóvel é considerado um ativo econômico e seu valor é dado pelo valor presente do fluxo de seus rendimentos futuros, a saber, os aluguéis potencialmente gerados.

As estimativas empíricas são feitas a partir do modelo estatístico/econométrico de Florestas Aleatórias. Para tanto, são utilizadas as informações de aluguéis de imóveis residenciais disponíveis para a proximidade de Barra Longa. As fontes de informações disponíveis são o Censo Demográfico de 2010 e o Cadastro. Todos os valores são em reais de outubro de 2015. São apresentados intervalos de valores para diferentes cenários de taxa desconto de 3%, 4%, 5% e 6% ao ano.

Existem no Cadastro 291 residências em Barra Longa declaradas atingidas com informações das variáveis utilizadas nos modelos de preços hedônicos. Para estas, utilizando os dados do Censo Demográfico, estimam-se valores agregados médios de cerca de R$ 54 milhões de reais com a taxa de desconto de 3% e de cerca de R$ 35 milhões com a taxa de desconto de 5%. Por sua vez, utilizando os dados do Cadastro, estimam-se valores agregados médios de cerca de R$ 58 milhões de reais, com taxa de desconto de 3%, a preços de outubro de 2015, e de cerca de R$ 37 milhões, com taxa de desconto de 5%, a preços de outubro de 2015. Obviamente, são valores sujeitos

às incertezas estatísticas e os valores com intervalos de confiança são apresentados ao longo do texto.

Ademais, utilizando a taxa de desconto de 3% a.a., estimam-se os valores médios R$ 156.405, R$ 254.313, R$ 126.189 e R$ 107.507 para as residências típicas 1, 2, 3 e 4, respectivamente, em outubro de 2015. Para a taxa de desconto de 5% a.a., estimam-se os valores médios de R$ 94.680, R$ 153.948, R$ 76.389 e R$ 65.104, respectivamente, para as residências típicas 1, 2, 3 e 4.

Posteriormente, sujeito à disponibilidade de dados, pretende-se mensurar o impacto do rompimento sobre o setor habitacional, considerando a variação no valor médio das propriedades residenciais. Essa variação é composta pela perda ou destruição do valor das moradias atingidas pela lama de rejeito, e/ou afetadas pelas obras de limpeza, e pela potencial valorização/desvalorização decorrente dos efeitos do desastre no mercado imobiliário.

Por fim, é importante salientar que os valores estimados correspondem a valores de mercado. Os valores subjetivos que os proprietários atribuem aos seus imóveis podem ser maiores que os valores de mercado (caso contrário, poderiam ter vendido seu imóvel) e, portanto, a perda de bem-estar que um proprietário ou uma proprietária tem com a perda do imóvel é maior (ou igual) que o valor de mercado do mesmo[33].

[33] Os valores estimados neste estudo não implicam posicionamento da FGV sobre valores de indenização desses imóveis no tempo presente. Para tanto devem ser consideradas questões legais e da jurisprudência brasileira sobre o momento de cálculo desses valores. O produto "Parâmetros para reparação do direito à moradia no contexto do rompimento da barragem de Fundão" da FGV mostra diversas dimensões ligadas ao direito à moradia adequada.

REFERÊNCIAS

AGUIRRE, A.; DE FARIA, D. M. A Utilização de “preços hedônicos” na avaliação social de projetos. **Revista Brasileira de Economia**, v. 51, n. 3, p. 391-411, 1997.

ASSEMBLY, U. G. Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction. **United Nations General Assembly**, Nova York, v. 41, 2016.

BANCO MUNDIAL. **PDNA guidelines volume B** — housing. 2012a.

____. **PDNA guidelines volume B** — macroeconomics impact of disasters. 2012b.

BELLMAN, R.; KUSANAGI, M. **Methods of nonlinear analysis, vol. I (mathematics in science and engineering volume 61-I)**. s.l.: American Society of Mechanical Engineers, 1973.

BOZDOGAN, H. Model selection and Akaike’s information criterion (AIC): the general theory and its analytical extensions. **Psychometrika**, v. 52, n. 3, p. 345-370, 1987.

BREIMAN, L. et al. Classification and regression Trees (Belmont, CA: Wadsworth International Group). **Biometrics**, v. 40, n. 3, p. 17-23, 1984.

____. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123-140, 1996.

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

CASAROTTO FILHO, N.; KOPITTKE, B. H. **Análise de investimentos**. s.l.: Vertice, 2010.

CEPAL. **Handbook for disaster assessment**. 2014.

DANTAS, R. A.; CORDEIRO, G. Uma nova metodologia para avaliação de imóveis utilizando modelos lineares generalizados. **Revista Brasileira de Estatística**, v. 49, n. 191, p. 27-46, 1988.

DASGUPTA, P. The Stern review’s economics of climate change. **National Institute Economic Review**, v. 199, n. 1, p. 4-7, 2007.

FLETCHER, M.; MANGAN, J.; RAEBURN, E. Comparing hedonic models for estimating and forecasting house prices. **Property Management**, v. 22, n. 3, p. 189-200, 2004.

GELMAN, A.; HILL, J. **Data analysis using regression and multilevel/hierarchical models**. s.l.: Cambridge University Press, 2006.

GONZÁLEZ, M. A. S.; FORMOSO, C. T. Especificação de modelos de preços hedônicos para locação residencial em Porto Alegre. **Cadernos Ippur/UFRJ**, v. 8, n. 1, p. 59-72, 1994.

GOODMAN, A. C. Hedonic prices, price indices and housing markets. **Journal of Urban Economics**, v. 5, n. 4, p. 471-484, 1978.

HARRISON JR., D.; RUBINFELD, D. L. Hedonic housing prices and the demand for clean air. **Journal of Environmental Economics and Management**, v. 5, n. 1, p. 81-102, 1978.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning:** data mining, inference, and prediction. s.l.: Springer Science & Business Media, 2009.

IBGE. **Censo demográfico, 2010**. 2010, v. 13. Acesso em: 16 out. 2019.

IBGE. **Censo demográfico, 2010**. 2015, v. 13. Acesso em: 16 out. 2019.

JAMES, G. et al. **An introduction to statistical learning**. s.l.: Springer, 2013. v. 112.

KUHN, M. **Variable selection using the caret package**. 2012. Disponível em:<http://cran. cermin. lipi. go. id/web/packages/caret/vignettes/caretSelection.pdf>.

____. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, n. 5, p. 1-26, 2008.

LOPES, J. T. D. Depreciação de edificações. In: CONGRESSO BRASILEIRO DE ENGENHARIA DE AVALIAÇÕES E PERÍCIAS, VIII. **Anais...** ICAPÈ, 1995.

NORDHAUS, W. D. A review of the Stern review on the economics of climate change. **Journal of Economic Literature**, v. 45, n. 3, p. 686-702, 2007.

PAIXÃO, L. A. R. O impacto da violência no preço dos imóveis comerciais de Belo Horizonte: uma abordagem hedônica. **Economia Aplicada**, v. 13, n. 1, p. 125-152, 2009.

ROSEN, S. Hedonic prices and implicit markets: product differentiation in pure competition. **Journal of Political Economy**, v. 82, n. 1, p. 34-55, 1974.

SHEPPARD, S. Hedonic analysis of housing markets. In: CHESHIRE, P.; MILLS, E. S. (Ed.). **Handbook of regional and urban economics**. 1999. v. 3, p. 1595-1635.

STERN, N.; STERN, N. H. **The economics of climate change:** the Stern review. s.l.: Cambridge University Press, 2007.

WAGER, S.; HASTIE, T.; EFRON, B. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. **The Journal of Machine Learning Research**, v. 15, n. 1, p. 1625-1651, 2014.

WOOLDRIDGE, J. M. **Econometric analysis of cross section and panel data**. s.l.: MIT Press, 2010.

ZAMBRANO-BIGIARINI, Mauricio. **HydroGOF:** goodness-of-fit functions for comparison of simulated and observed hydrological time series. s.l: s.n., 2017.

APÊNDICE A — Variáveis construídas a partir dos dados do Censo 2010

Neste presente apêndice constam a sistematização e o tratamento dos dados do Censo Demográfico de 2010 para a construção das variáveis explicativas dos modelos de preços hedônicos apresentados ao longo do documento.

As variáveis explicativas são um conjunto de características atribuídas aos domicílios amostrados, a saber: números de cômodos, dormitórios e banheiros de uso exclusivo, tipos de domicílio (casa, apartamento etc.), tipo do setor censitário (urbana ou rural), localização (coordenadas geográficas), formas de esgotamento sanitário, abastecimento de água e destinação de lixo e fonte de energia elétrica.

A variável indicadora Lixo Coletado por Serviço de Limpeza foi criada a partir da V0210 do Censo Demográfico e assume o valor 1 para os casos de lixo “Coletado diretamente por serviço de limpeza”, zero para os demais casos, exceto nas respostas em “Branco”, que não recebem informação.

Tabela 1 — Construção da variável Lixo Coletado por Serviço de Limpeza

Categorias de resposta na variável original V0210	Categorias de resposta na variável criada: Lixo Coletado por Serviço de Limpeza
1. Coletado diretamente por serviço de limpeza	1
2. Colocado em caçamba de serviço de limpeza	0
3. Queimado (na propriedade)	0
4. Enterrado (na propriedade)	0
5. Jogado em terreno baldio ou logradouro	0
6. Jogado em rio, lago ou mar	0
7. Tem outro destino	0
Branco	.

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável indicadora Rede Geral de Esgoto foi criada a partir da V0207 do Censo Demográfico e assume o valor 1 para os casos de resposta “Sim”, e zero para os casos de resposta “Não”.

Tabela 2 — Construção da variável Rede Geral de Esgoto

Categorias de resposta na variável original V0207	Categorias de resposta na variável criada: Rede Geral de Esgoto
1. Sim	1
2. Não	0

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável indicadora Rede Geral de Água foi criada a partir da V0208 do Censo Demográfico e assume o valor 1 para os casos de resposta “Rede geral de distribuição”, e zero para os demais casos.

Tabela 3 — Construção da variável Rede Geral de Água

Categorias de resposta na variável original V0208	Categorias de resposta na variável criada: Rede Geral de Água
01. Rede geral de distribuição	1
02. Poço ou nascente na propriedade	0
03. Poço ou nascente fora da propriedade	0
04. Carro-pipa	0
05. Água da chuva armazenada em cisterna	0
06. Água da chuva armazenada de outra forma	0
07. Rios, açudes, lagos e igarapés	0
08. Outra	0
09. Poço ou nascente na aldeia	0
10. Poço ou nascente fora da aldeia	0

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável indicadora Poço ou Nascente na Prop. foi criada a partir da V0208 do Censo Demográfico e assume o valor 1 para os casos de resposta “Poço ou nascente na propriedade”, e zero para os demais casos.

Tabela 4 — Construção da variável Poço ou Nascente na Propriedade

Categorias de resposta na variável original V0208	Categorias de resposta na variável criada: Poço ou Nascente na Prop.
01. Rede geral de distribuição	0
02. Poço ou nascente na propriedade	1
03. Poço ou nascente fora da propriedade	0
04. Carro-pipa	0
05. Água da chuva armazenada em cisterna	0
06. Água da chuva armazenada de outra forma	0
07. Rios, açudes, lagos e igarapés	0
08. Outra	0
09. Poço ou nascente na aldeia	0
10. Poço ou nascente fora da aldeia	0

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável indicadora Energia Elétrica foi criada a partir da V0211 do Censo Demográfico e assume o valor 1 para os casos de resposta “Sim, de companhia distribuidora” e “Sim, de outras fontes”, e zero para os demais casos.

Tabela 5 — Construção da variável Energia Elétrica

Categorias de resposta na variável original V0211	Categorias de resposta na variável criada: Energia Elétrica
01. Sim, de companhia distribuidora	1
02. Sim, de outras fontes	1
03. Não existe energia elétrica	0
04. Branco	0

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável indicadora Urbano foi criada a partir da V1006 do Censo Demográfico e assume o valor 1 para os casos de classificação “Urbana”, e zero para os casos “Rural”.

Tabela 6 — Construção da variável Urbano

Categorias de resposta na variável original V1006	Categorias de resposta na variável criada: Urbano
1. Urbana	1
2. Rural	0

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável indicadora Apartamento foi criada a partir da V4002 do Censo Demográfico e assume o valor 1 para os casos de classificação “Apartamento”, zero para os casos “Casa”, “Casa de vila ou em condomínio”, “Habitação em: casa de cômodos, cortiço ou cabeça de porco” e sem informação para os demais casos.

Tabela 7 — Construção da variável Apartamento

Categorias de resposta na variável original V4002	Categorias de resposta na variável criada: Apartamento
11. Casa	0
12. Casa de vila ou em condomínio	0
13. Apartamento	1

Categorias de resposta na variável original V4002	Categorias de resposta na variável criada: Apartamento
14. Habitação em: casa de cômodos, cortiço ou cabeça de porco	0
15. Oca ou maloca	.
51. Tenda ou barraca	.
52. Dentro de estabelecimento	.
53. Outro (vagão, trailer, gruta etc.)	.
61. Asilo, orfanato e similares com morador	.
62. Hotel, pensão e similares com morador	.
63. Alojamento de trabalhadores com morador	.
64. Penitenciária, presídio ou casa de detenção com morador	.
65- Outro com morador	.

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável numérica Valor do Aluguel (em reais) foi criada a partir da V2011 do Censo Demográfico. A variável Número de Cômodos, por sua vez, foi criada a partir da V0203 do Censo Demográfico.

Tabela 8 — Construção da variável Número de Cômodos

Categorias de resposta na variável original V0203	Categorias de resposta na variável criada: Número de Cômodos
1 a 30	Número
Branco	.

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável numérica Número de Dormitórios foi criada a partir da V0204 do Censo Demográfico.

Tabela 9 — Construção da variável Número de Dormitórios

Categorias de resposta na variável original V0204	Categorias de resposta na variável criada: Número de Dormitórios
1 a 15	Número
Branco	.

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

A variável Número de Banheiros de Uso Exclusivo foi criada a partir da V0205 do Censo Demográfico.

Tabela 10 — Construção da variável Número de Banheiros de Uso Exclusivo

Categorias de resposta na variável original V0205	Categorias de resposta na variável criada: Número de Banheiros de Uso Exclusivo
0. Zero banheiros	0
1. Um banheiro	1
2. Dois banheiros	2
3. Três banheiros	3
4. Quatro banheiros	4
5. Cinco banheiros	5
6. Seis banheiros	6
7. Sete banheiros	7
8. Oito banheiros	8
9. Nove ou mais banheiros	9
Branco	.

Fonte: Elaboração própria (2019), a partir de dados IBGE (2010).

APÊNDICE B — Variáveis construídas a partir dos dados do Cadastro da Fundação Renova

Os dados do Cadastro da Fundação Renova utilizados pela FGV foram obtidos no Sistema de Gerenciamento de Stakeholder (SGS) da Fundação Renova, que armazena e gerencia as informações da população cadastrada e atendida pelos programas de reparação. A base do Cadastro foi obtida através do download do filtro 1322, que contém as informações levantadas no questionário socioeconômico aplicado durante o cadastramento, referentes às propriedades e aos indivíduos. A extração dos dados do sistema foi realizada no dia 28 de agosto de 2019.

A partir dos *downloads* de cada dimensão do Cadastro, construiu-se a Tabela 1 a seguir que resume os dados disponibilizados no banco. A primeira coluna indica o nome da dimensão originalmente atribuído pela Fundação Renova, a segunda coluna, por sua vez, contém o número de variáveis disponíveis para a respectiva dimensão e na terceira coluna encontra-se o número de observações. Por exemplo: a "dimensão_1" contém 115 variáveis e 32.589 observações.

A unidade referente ao número de observações depende da dimensão à qual se refere. No caso da "dimensão_1", a unidade de análise é propriedade, o que indica informações de 32.589 propriedades apenas nessa dimensão. A "dimensão_people", por sua vez, contém 233 variáveis a respeito de 97.658 indivíduos cadastrados.

Tabela 1 — Número de variáveis e de observações em cada dimensão do Cadastro — Parte 1

Dimensões	Número de Variáveis	Número de Observações
dimensao_01	115	32.589
dimensao_03	138	28.934
dimensao_04	17	6
dimensao_06	69	103
dimensao_07	157	4.801
dimensao_08	31	2.492
dimensao_09	80	124
dimensao_10	102	32.586
dimensao_11	71	4.127
dimensao_12	100	3.483
dimensao_13	394	1.343
dimensao_14	576	23.649

Dimensões	Número de Variáveis	Número de Observações
dimensao_15	356	610
dimensao_16	227	2.344
dimensao_17	108	20
dimensao_18	13	5
dimensao_19	65	2
dimensao_20	15	279
dimensao_21	15	3
dimensao_22	7	7.360
dimensao_23	88	511
dimensao_24	74	306
dimensao_25	23	39
dimensao_26	93	586
dimensao_27	137	2.677
dimensao_28	30	137
dimensao_29	143	2.305
dimensao_30	40	32.586
dimensao_31	93	32.586
dimensao_34	7	32.588
dimensao_cultivars	15	20.322
dimensao_cultivations	26	15.426
dimensao_improvements	18	86.532
dimensao_living_spaces	18	2.842
dimensao_machineries	11	1.465
dimensao_people	233	97.658
dimensao_realties	22	48.010
dimensao_vehicles	15	780
propriedades	15	33.726
dimensao_buildings	46	3.499

Fonte: Elaboração dos autores com base nos dados do Cadastro da Fundação Renova (2019).
Nota: versão do cadastro: 28/08/2019.

Limpeza e tratamento de dados

O procedimento de limpeza de dados consiste em, a partir de uma base de dados “bruta” (ou seja, como vem da instituição que a cria e/ou alimenta), selecionar o conjunto de informações a serem utilizadas para a análise e realizar os tratamentos necessários nestas. Os procedimentos consistem, por exemplo, em transformação de variáveis categóricas, criação de novas variáveis e identificação das unidades de observação que serão excluídas da análise a ser realizada.

A título de ilustração, um caso comum de classificação é a criação de variáveis do tipo indicadoras (*dummies*), quando se geram variáveis binárias, que assumem o valor 1 quando a observação satisfizer determinada condição e 0 quando não satisfizer. Por exemplo, a variável "dummy_mulher", apresentada à frente, foi criada a partir das informações contidas na variável "2.2.4" — "Qual o sexo?" —, de modo que todos os indivíduos classificados como "Feminino" em "2.2.4" foram classificados como 1 na variável "dummy_mulher", enquanto os classificados como "Masculino" receberam o valor 0 na variável "dummy_mulher".

Georreferenciamento das propriedades

O georreferenciamento das propriedades teve como base as informações das colunas "c6" e "c7" da dimensão propriedades do cadastro. A coluna "c6" corresponde à latitude da propriedade cadastrada enquanto a "c7" corresponde à longitude. A partir do conjunto de informações disponíveis, atribuíram-se as variáveis dos setores censitários sob os quais os pontos estão contidos; nas técnicas de georreferenciamento, isso é chamado de interseção entre pontos e polígonos.

Uma ressalva importante é a questão das projeções das malhas digitais: para melhorar a precisão da interseção, utilizaram-se as projeções de UTM (*Universal Transversa Mercator*) que utiliza coordenadas cartesianas bidimensionais para a superfície da terra. Para cada fuso-horário, tem-se uma projeção de UTM que melhor se adequa, motivo pelo qual no presente estudo foram utilizadas as projeções:

- *+proj=utm +zone=21 +south +ellps=WGS84 +datum=WGS84 +units=m +no_defs*
- *+proj=utm +zone=22 +south +ellps=WGS84 +datum=WGS84 +units=m +no_defs*
- *+proj=utm +zone=23 +south +ellps=WGS84 +datum=WGS84 +units=m +no_defs*
- *+proj=utm +zone=24 +south +ellps=WGS84 +datum=WGS84 +units=m +no_defs*

O georreferenciamento foi realizado com as propriedades que foram cadastradas pela Fundação Renova, portanto, propriedades do cadastramento realizado pela Cáritas em Mariana foram excluídas. Do total de 32.589 propriedades listadas na base do cadastro, apenas uma não possui informações de coordenadas geográficas, sendo 32.588 o total

de propriedades com coordenadas. Além da propriedade que não possui informação de latitude e longitude, 14 propriedades estão com as coordenadas fora do limite costeiro do Brasil, possivelmente por equívoco no momento do preenchimento. Estas foram desconsideradas, de forma que o total de propriedades georreferenciadas utilizadas no documento foi de 32.574.

Para essas 32.574 propriedades com coordenadas e que foram consideradas, atribuíram-se, a partir de seu georreferenciamento, as informações de Unidade da Federação, município (segundo a malha municipal de 2015), setor censitário do IBGE e a variável **"**TIPO**"**, que corresponde ao tipo do setor (rural ou urbano). É necessário destacar que a localização das propriedades cadastradas corresponde ao momento do cadastramento.

Criação de variáveis com sufixos "pre" e "pos"

Entre o conjunto de informações disponíveis no Cadastro, existem variáveis relacionadas com características "pré-rompimento", as quais se referem ao período anterior ao rompimento da Barragem de Fundão e cujas perguntas relacionadas mencionam "nos últimos 12 meses antes do evento (nov./14 a nov./15)". No tratamento dos dados, essas variáveis "pré-rompimento" foram, em todos os casos, indexadas com o sufixo "pre".

Já as informações "pós-rompimento" têm como referência o momento em que os indivíduos foram cadastrados ou, em alguns casos, o mês anterior ao do cadastramento ("mês anterior da data da entrevista"). Nesses casos, no tratamento dos dados, essas variáveis foram indexadas com o sufixo "pos". É importante salientar que a aplicação das entrevistas para o Cadastro se iniciou em 2015 (fase emergencial) e continua em processo até hoje. Assim, as variáveis com o sufixo "pos" correspondem ao momento em que os indivíduos foram entrevistados que não é único para todos os cadastrados.

Variáveis utilizadas no presente estudo

Variáveis sobre os impactos sofridos: as variáveis entre "1.1.43.1" e "1.1.43.20" estão relacionadas com os impactos sofridos devido ao rompimento ("Liste quais foram os impactos sofridos com o evento"). Especificamente, são elas: "Perda total ou parcial do terreno/lote", "Perda e/ou dano a edificações e construções civis (benfeitorias não reprodutivas)", "Perda e/ou dano a cultivos agrícolas ou frutíferos (benfeitorias reprodutivas)", "Perda e/ou dano a pastagens/áreas de criação de animais (benfeitorias reprodutivas)", "Perda e/ou dano a áreas de vegetação nativa (APP e/ou RL)", "Perda

e/ou dano a animais (semoventes)", "Perda e/ou interrupção temporária ou permanente de atividades econômicas (pesca / criação de peixes no rio / lavadeira)", "Perda e/ou interrupção temporária ou permanente de atividades econômicas (comércio / serviço)", "Perda e/ou interrupção temporária ou permanente de atividades econômicas (indústria / agroindústria)", "Perda e/ou interrupção temporária ou permanente de atividades extrativistas (extrativismo mineral / garimpo/ jazidas de areia)", "Perda e/ou interrupção temporária ou permanente de atividades extrativistas (extrativismo vegetal)", "Perda e/ou dano de eletrodomésticos / mobiliários e/ou vestuários", "Perda e/ou dano de máquinas, implementos e equipamentos (bens móveis)", "Perda e/ou dano de veículos (bens móveis)", "Perda e/ou dano de documentos pessoais", "Afetação na ocupação, trabalho e/ou renda da pessoa física", "Afetação de animais domésticos (não produtivos)", "Perda de acesso a equipamentos públicos", "Perda/ desaparecimento de familiares ou pessoas sem parentesco com relação de dependência financeira" e "Outras perdas e/ou danos'. Para cada uma delas, foi gerada uma variável binária, seguindo o padrão:

Tabela 2 — Atribuição dos valores de acordo com as respostas

Categorias de resposta na variável original	Categorias de resposta na variável criada
Sim	1
Não	0

Fonte: Elaboração própria (2019).

APÊNDICE C — Predição por meio dos modelos log-linearizados

Os modelos especificados no formato log-nível (OLS 4-6 e ML 4-6) passam por uma transformação particular na variável dependente, no caso o $\ln(aluguel)$. Deste modo, o termo de erro em cada caso continua a seguir uma distribuição normal, com média zero e variância σ_ϵ^2, mas a distribuição do aluguel passa a ser uma log-normal com média $X\beta + \frac{\sigma_\epsilon^2}{2}$ e variância σ^2.

Estimado os modelos log-níveis, o passo subsequente é realizar a predição para observações fora da amostra. No caso para os dados do Censo Demográfico, seria o cálculo do valor de aluguel para os imóveis que foram retirados da base de dados utilizada para a estimação.

Sob a especificação do modelo log-nível, sabemos que o valor previsto do $\ln(aluguel)$ para um dado domicílio p fora da amostra pode ser escrito como:

$$\ln(\widehat{aluguel})_p = x'_p\hat{\beta}$$

Sendo x'_p o conjunto de covariadas para o domicílio p e $\hat{\beta}$ os parâmetros estimados em cada modelo. Sob a hipótese de a variável de aluguel seguir uma distribuição log-normal, o valor previsto pode ser calculado da seguinte maneira:

$$\widehat{aluguel}_p = \exp\left\{x'_p\hat{\beta} + \frac{\hat{\sigma}^2}{2}\right\}$$