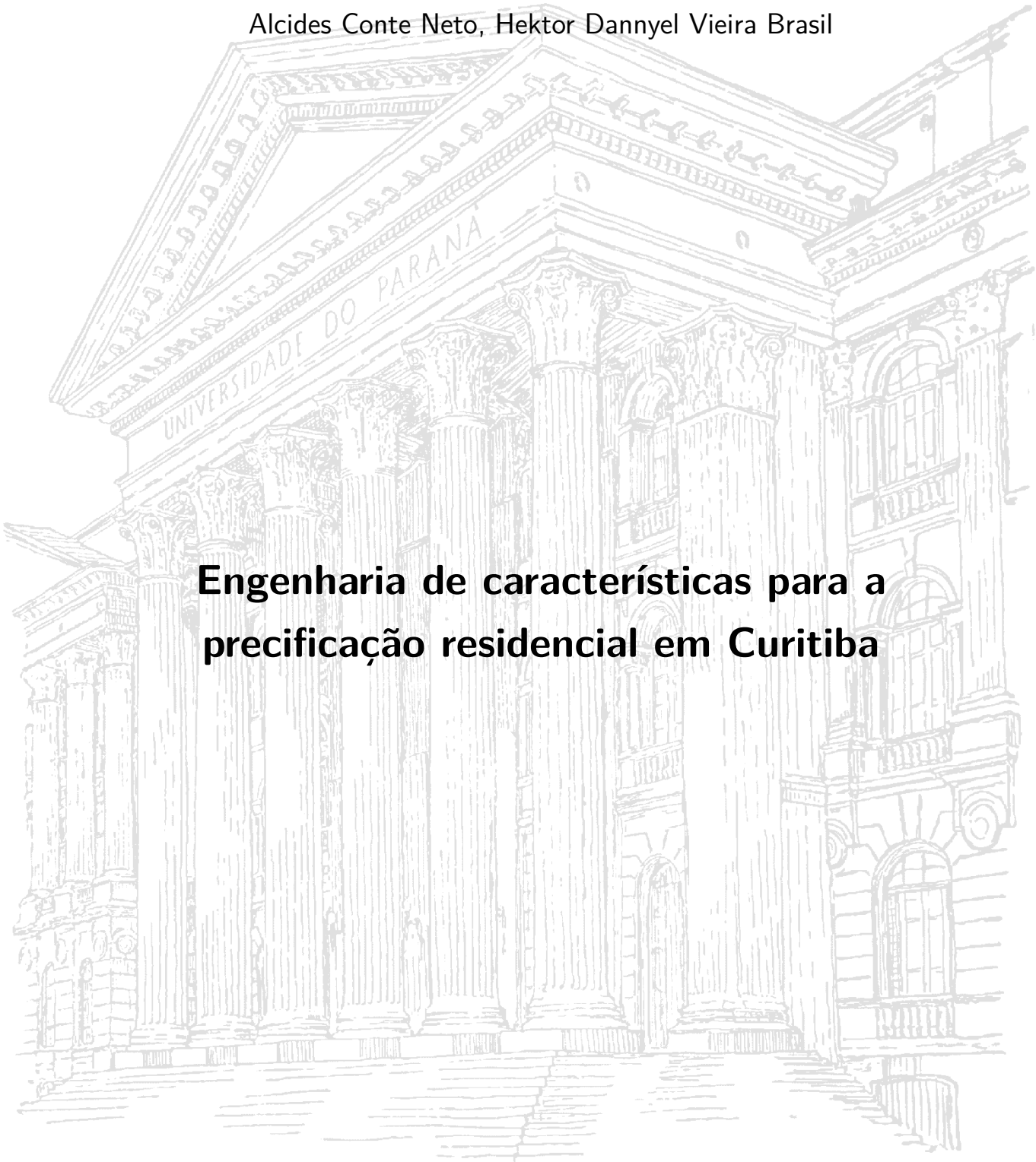
Universidade Federal do Paraná

Alcides Conte Neto, Hektor Dannyel Vieira Brasil



Engenharia de características para a precificação residencial em Curitiba

Curitiba

2018

Alcides Conte Neto, Hektor Dannyel Vieira Brasil

Engenharia de características para a precificação residencial em Curitiba

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da Universidade Federal do Paraná como requisito para a obtenção do grau de bacharel em estatística.

Universidade Federal do Paraná

Setor de Ciências Exatas

Departamento de Estatística

Curitiba

2018

Agradecimentos I

Ao fim dessa jornada de um ano de trabalho há muitas pessoas que contribuíram de alguma forma, seja pela paciência que dedicaram a mim em dias de estresse ou pela sua contribuição e dicas que enriqueceram o conteúdo aqui apresentado. Dessa forma sinto-me na obrigação de pedir desculpas as pessoas cujo os nomes que não estão aqui incluídos.

Primeiramente agradeço a minha família e a minha digníssima namorada, cuja a paciência fora infinita. Ao meu parceiro de jornada Hektor, que me aturou durante esse ano e suportou minhas chatices e peculiaridades. A todos os professores do Laboratório de Estatística e Geoinformação (LEG), em especial ao nosso orientador Walmes Marques Zeviani, cuja as idéias e discussões enriqueceram esse trabalho, aos professores Paulo Justiniano Ribeiro Junior e Cesar Augusto Taconeli que sempre estiveram a disposição para ajudar, não só durante esse período, mas durante todo o curso. Também não posso esquecer do professor Fernando Mayer, que gastou tempo em instalar os pacotes necessários para disponibilizar um ambiente online que possibilitasse o desenvolvimento do trabalho em qualquer lugar.

Por ultimo mas não menos importante, aos meus colegas de trabalho dos Institutos Lactec, em especial ao Lucas Jerszurki que me auxiliou quando tive dúvidas em relação aos tipos de coordenadas e projeções cartográficas, e ao pesquisador Bernardo Lipski que me apresentou o projeto OpenStreetMap.

Alcides Conte Neto

Agradecimentos II

Antes de tudo, eu preciso agradecer ao meu parceiro ao longo desse ano, por todo o empenho e a paciência durante o desenvolvimento deste trabalho. Alcides, you're the man.

Agradeço também à minha amada mãe, a quem eu devo praticamente tudo o que eu sou e que sempre esteve presente para me fazer olhar para a frente e continuar.

Aos meus amigos incríveis, que me proporcionaram os melhores anos da minha vida e que compreenderam as minhas ausências forçadas ao longo deste ano. Sem vocês, os finais de semana não seriam os mesmos.

A todos os professores que contribuíram de forma ímpar para a consolidação da minha carreira. Em especial:

Ao professor Walmes Zeviani, que acompanhou este trabalho de perto;

Ao professor Cesar Taconeli, que aceitou compor a banca avaliadora e disponibilizou do seu tempo sempre que pôde;

Ao professor Paulo Justiniano, por sua sabedoria e por todo o suporte prestado.

À equipe da DQSC-CT, da Renault, por me aturar das 8:07 às 14:40. Estas feras aí, meu.

Às Schweppes no armário da PC.

Hektor Dannyel Vieira Brasil

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem"

— John Tukey

Resumo

É comum o estudo das características que levam consumidores a adquirir bens para si. Dentre tais bens, a residência própria é bastante almejada, proporcionando aos estudiosos e profissionais da área imobiliária um campo muito interessante para exploração. Na economia, utiliza-se a modelagem hedônica, que parte do pressuposto que o preço de um bem para uma pessoa está diretamente relacionado às características que expressam algum valor emocional a ela. Pensando nisso, este estudo procurou analisar as características de maior relevância para o ajuste de modelos de precificação de imóveis em Curitiba, utilizando Engenharia de Características e Análise de Regressão. Os dados foram extraídos através de *web scraping*. A base é constituída de características dos imóveis, como quantidade de dormitórios, banheiros, e vagas na garagem, além de alguns atributos predeterminados pelo site Imovelweb, direcionados a informar melhor um possível comprador. Também foram adquiridas as descrições dos imóveis, escritas em texto livre, que foram processadas com técnicas de mineração de texto. Foram utilizados recursos para enriquecer os dados, como a inclusão da densidade de estabelecimentos comerciais próximos aos imóveis, a detecção de sinônimos e a classificação gramatical das palavras relacionadas às descrições dos anúncios, o zoneamento dos imóveis e a Regressão Lasso para auxiliar na construção dos modelos. Os modelos ajustados mostraram um desempenho preditivo abaixo do esperado, tendo um erro de aproximadamente oitocentos reais por metro quadrado. No entanto, as variáveis utilizadas na sua construção são promissoras e oferecem diversas possibilidades para aplicação de outros métodos buscando melhores resultados.

Palavras-chave: Web Scraping. Mineração de texto. Regressão Lasso. Engenharia de Covariáveis. Precificação de imóveis. Curitiba.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Há algumas décadas, a tecnologia podia ser vista como algo futurístico e muitas vezes poético. Nas antigas obras de ficção científica, muitas das coisas não passavam de meras interpretações de como os autores imaginavam ou esperavam o futuro. Isaac Asimov, professor de Bioquímica na Universidade de Boston, ficou muito famoso na cultura popular com seus livros, que retratavam os avanços tecnológicos e as esperanças da humanidade quanto à ciência. Em 1964, durante uma entrevista, lhe perguntaram como imaginava o mundo em 50 anos e uma de suas respostas estava relacionada à velocidade com que a informação chegaria para as pessoas.

Vinte anos atrás, ainda se descobria como a *internet* podia manter as pessoas mais conectadas entre si, com o mundo e principalmente com a informação; além de como os computadores serviriam para deixar as atividades do cotidiano mais rápidas. À época, um computador com *Gigabytes* de capacidade era considerado tecnologia de última geração. Os *downloads* de arquivos de tamanhos considerados irrisórios hoje, demoravam diversos minutos ou horas. A partir de então, o mundo se modernizou com uma velocidade inacreditável.

Hoje, há quem pense na informação como algo trivial. Se é preciso descobrir algo, prontamente pode-se pegar um dispositivo móvel e fazer uma pesquisa. Os resultados são mostrados em segundos, abertos como um livro aos olhos, sem que seja necessário um deslocamento até uma biblioteca. A demanda pela informação tomou proporções inimagináveis, e é necessário supri-la com velocidade igual ou superior.

O mercado imobiliário é um segmento de grande importância na economia de maneira geral. O investimento em imóveis não é algo novo (FERREIRA, 2005), seja para fins pessoais ou profissionais. Ao longo dos últimos três anos, o Índice Imobiliário (IMOB) têm demonstrado alta (BM&FBOVESPA, 2018), atingindo entre o fim de 2017 e o início de 2018 seu maior valor desde março de 2013. Com o mercado de imóveis tendo tal peso para a ciência econômica, é natural esperar a alta exigência na obtenção rápida e assertiva de dados.

Alguns estudos econométricos (Mingoti (2005), Hermann e Haddad (2005), Fávero, Belfiore e Lima (2008), por exemplo) definem que o preço de um imóvel envolve uma quantidade de características que acrescentem ao bem-estar do consumidor. Essas características são "somadas", subjetivamente, e juntas compõem o valor do imóvel.

Em resumo, este trabalho tem como objetivo investigar, em uma abordagem voltada à estatística, as características com maior capacidade de predizer o valor do imóvel, considerando na análise dados não estruturados, como a descrição da propriedade em

texto, feita pelo próprio anunciante, além de informações acerca da geolocalização.

Abordagens mais usuais de seleção de variáveis talvez não surtissem o efeito desejado. Uma vez que as descrições dos anúncios foram trabalhadas utilizando técnicas de Mineração de Texto, a quantidade de covariáveis se tornou muito grande e difícil de se trabalhar. Logo, a saída encontrada foi a utilização de Engenharia de Covariáveis, método que possui a robustez necessária para garantir que dentre milhares de covariáveis, as mais indicadas para a análise proposta fossem escolhidas no modelo final.

2 Revisão de Literatura

2.1 Modelos hedônicos

O trabalho de Lancaster (1966) levou a um avanço importante na maneira com que os economistas entendem o consumidor (WIERENGA, 1984). Nele, de acordo com o próprio autor, foi enfatizada a ideia de que a utilidade de um bem de consumo é derivada diretamente das características desse bem e não dele como um todo, o que discerne o modelo da metodologia clássica. Wierenga (1984) explica a ideia de Lancaster de forma mais simples, possibilitando o entendimento leigo, afirmando que nesse contexto "um produto é concebido como um pacote de características que tem propriedades satisfatórias para o consumidor".

Como bem cita Wierenga (1984), o modelo de Lancaster ignora o processo de percepção do consumidor, ou seja, o fato de que cada consumidor percebe as características do imóvel de forma diferente (características intangíveis). Além disso, a possibilidade de que o consumidor encontre utilidade em diferentes itens, aspecto que tem papel importante na tomada de decisão na hora da compra, também é desprezada. Isso não invalida o modelo, mas abre espaço para o surgimento de extensões do mesmo. Nesse sentido surgem outras obras, como a de Rosen (1974), que marcou os fundamentos da análise de preços hedônicos (ARRAES; FILHO, 2008), sendo o primeiro a inserir o problema dentro do contexto de mercado (FáVERO; BELFIORE; LIMA, 2008).

Os modelos hedônicos[1] são muito utilizados para avaliar quais as características que determinam os preços de produtos. Esse tipo de modelo busca medir atributos que causam aspectos de satisfação e prazer na compra de um bem, portanto diz-se que ele é relacionado com bens intangíveis e, por natureza, sem mercado, uma vez que sua venda explícita não é possível (JOHN; PORSSE, 2016). Para essa propriedade ficar mais clara, pode-se pensar em um comprador em potencial, que tem um sonho de morar de frente para um lago e, desse modo, o imóvel terá mais valor para ele se possuir essa qualidade. Esse tipo de modelo também possuem um forte apelo econômico, já que os imóveis muitas vezes representam boa parte da riqueza de uma família (JOHN; PORSSE, 2016) e, por esse motivo, muitos dos trabalhos realizados estão no campo da economia e da econometria.

De acordo com Hermann e Haddad (2005) a teoria dos modelos hedônicos não determina uma forma funcional nem as variáveis relevantes para a estimação,

[1] Consultando um dicionário, verifica-se que hedonismo é definido como "a busca incessante pelo prazer como propósito de vida".

decorrendo que uma das maneiras de escrever o modelo é na forma de regressão linear múltipla com alguma transformação na variável resposta, e estimar os coeficientes utilizando o procedimento de mínimos quadrados. Portanto, olhando desse modo, a análise é bem conhecida pelos estatísticos, que estudam exaustivamente modelos de regressão. A diferença desse para um modelo de regressão usual é a utilização das variáveis explicativas que mensuram, de alguma forma, as características intangíveis.

Uma análise de imóveis residenciais da região metropolitana de São Paulo foi realizado por Fávero, Belfiore e Lima (2008). Os autores afirmam que os modelos hedônicos tem sido utilizados para investigar a demanda e a oferta de produtos, uma vez que definem quais atributos, intrínsecos ou extrínsecos, são relevantes na composição do imóvel. Uma análise fatorial foi aplicada para definir perfis sociodemográficos similares entre 134 localidades (96 distritos do Município de São Paulo e 38 municípios da Região Metropolitana de São Paulo). Por meio do método de Rosen (1974), foram ajustados modelos para cada um dos três perfis definidos, em busca de determinar as equações de oferta e demanda. As covariáveis que compuseram os modelos foram selecionadas com base em questionários autoaplicados a clientes e profissionais da área imobiliária, sendo que elas apontam a presença de salão de festas, presença de hospitais próximos, mensuram a renda e o tamanho da família, o número de quartos, a quantidade de vagas na garagem, entre outras. Apesar de mostrar uma lista de formas funcionais utilizadas na modelagem, os autores também modelaram o preço do imóvel transformado pelo logaritmo neperiano.

O trabalho de John e Porsse (2016), dentre os pesquisados, é o que tem maior semelhança com o presente estudo. O estudo teve como objetivo avaliar o que determina os preços de apartamentos na cidade de Curitiba e, para isso, utiliza 8000 anúncios coletados do portal Imovelweb para modelar o logaritmo neperiano dos preços por três grupos de covariáveis: características intrínsecas do imóvel, amenidades e localidade. O primeiro conjunto refere-se a características como quantidade de quartos e de vagas na garagem. O segundo mensura o número de escolas, parques, hospitais, taxa de criminalidade e outras características das redondezas do imóvel. Já o último grupo busca de alguma forma incluir a localização espacial, utilizando o bairro para cumprir tal objetivo. A estatística I de Moran foi utilizada para verificar a dependência espacial dos coeficientes do modelo relacionados aos bairros. Um resultado interessante é que a estrutura espacial de Curitiba se encaixa no formato chamado por Brueckner (2011) de *Central Business District*[2] (CDB), implicando que quanto mais longe o imóvel se encontra do CDB menor é o seu custo.

[2] De forma simplificada, é a área em que estão concentrados os locais de trabalho da população.

3 Materiais e Métodos

A base de dados foi obtida do portal de imóveis Imovelweb por meio de um algoritmo de *web scraping* escrito no ambiente R (R Core Team, 2017). Todas as páginas HTML disponíveis no momento da coleta[1] foram guardadas e posteriormente analisadas com o auxílio do pacote *xml2* (WICKHAM; HESTER; OOMS, 2017), que extraiu somente as informações de interesse utilizando o conjunto de regras de sintaxe do XPath.

Dois filtros foram adicionados na pesquisa direta do portal, um deles para mostrar somente imóveis a venda e outro para mostrar exclusivamente os que se localizam na cidade de Curitiba. O algoritmo só salva as páginas marcadas como classificados (apresentados com tarja roxa no site), o que significa que a estrutura da página é específica para exibir as informações de construções já finalizadas. A lista das variáveis que foram coletadas é a seguinte:

- **title**: título do anúncio;
- **description**: descrição textual do imóvel feita pelo anunciante;
- **address**: endereço do imóvel (rua, bairro, cidade);
- **lat**: latitude;
- **lon**: longitude;
- **pictures**: número de fotos colocadas no anúncio;
- **iptu**: valor de IPTU em reais;
- **condominium**: valor de condomínio em reais;
- **usefulArea**: área útil em m^2;
- **totalArea**: área total em m^2;
- **bedroom**: número total de quartos;
- **suite**: quantidade de suítes;
- **bathroom**: quantidade de banheiros;
- **garage**: quantidade de vagas na garagem;

[1] Os dados foram coletados como um recorte no tempo, ou seja, de forma transversal.

- **price**: preço do imóvel em reais;
- **years**: idade do imóvel em anos;
- **publishment**: dias desde a publicação;
- **advertiser**: código do anunciante;
- **CRECI**: número de registro no CRECI (Conselho Regional de Corretores de Imóveis) que identifica o corretor.

Além disso, foram extraídas as características padrão de cada imóvel (tags) que são dados semiestruturados, disponibilizados em uma lista predefinida pelo site para que os anunciantes estabeleçam as que mais convém com o anúncio. Por conta disso, espera-se que essas características possuam uma menor quantidade de erros se comparado às variáveis geradas pela descrição, cujo texto é livre.

Ao todo foram coletadas 71316 anúncios, sendo que 17,93% desses imóveis estavam sem informações de latitude e longitude e, além disso, 0,29% abrangiam coordenadas fora dos limites da cidade de Curitiba. Essas observações foram removidas. A figura 1 mostra cada um dos 58356 imóveis restantes dispostos nas suas respectivas localizações. Cada partição é um tipo de propriedade, decorrendo diretamente da classificação feita pelo site. A maioria dos dados corresponde a apartamentos e casas, e ainda há 3443 imóveis comerciais, 3575 terrenos e apenas 54 propriedades rurais [2].

Os objetos de interesse são os imóveis residenciais (casas e apartamentos), já que possuem características em comum, como o número de quartos, quantidade de banheiros e vagas na garagem. Portanto, os demais imóveis foram removidos da base, permanecendo 51284 dados a serem trabalhados.

3.1 Descritiva das variáveis numéricas

Ao longo de uma pré-análise, erros foram detectados nos dados coletados. Alguns foram relatados aos anunciantes e corrigidos, enquanto outros foram simplesmente eliminados da base de dados.

Devido ao grande volume de erros de digitação e à inviabilidade de espera da correção pelos anunciantes, foi decidido pela adoção de algumas regras para melhor representação ou eliminação de dados incorretos:

[2] De acordo com o INCRA (2010) o imóvel rural é uma área formada de uma ou mais matrículas de terras contínuas, do mesmo detentor, podendo ser localizada tanto na zona rural quanto urbana do município.

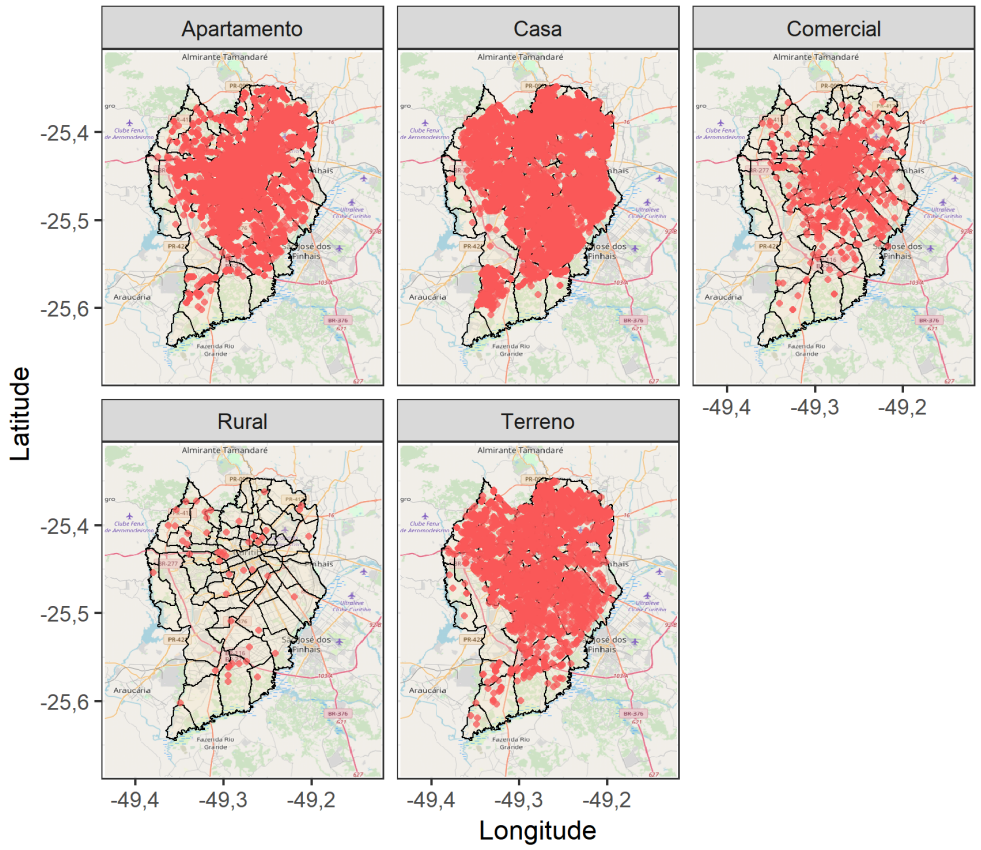


Figura 1 – Mapa político da cidade de Curitiba. Cada ponto representa a localização de um imóvel coletado do site Imovelweb.

- O valor da área total deve ser maior ou igual ao valor da área útil, ocasionando na remoção do valor de área total caso a regra não seja obedecida.

- Registros faltantes do valor de área útil foram substituídos pela área total correspondente.

- Os erros de digitação que ocorrem na variável *area* levam a ter valores elevados de metragem do imóvel, o que resulta em um preço por metro quadrado abaixo do comum. Já os problemas na variável *price* conduzem a preços elevados.

  Para diluir esse efeito, foi criada uma variável que expressa o preço por m^2 (*priceSM*) e foram removidos da base as propriedades com valores abaixo de R$ 1000,00 e acima de R$ 20000,00 por m^2.

- A idade do imóvel é uma característica que nem todos os corretores informam no anúncio (50,6%), portanto ela só foi utilizada para entender melhor as peculiaridades dos imóveis disponíveis para o estudo. Alguns anúncios foram verificados e dois dados chamaram a atenção, então foi decidido remover o valor de idade de propriedades declaradas com mais de 100 anos.

- Quando o dado do número de quartos, suítes e de vagas na garagem não estava presente, ele foi considerado zero, pois foi utilizado como pressuposto que eles não foram informados pelo simples fato de o imóvel não possuir essa característica. Já no caso da falta de informação de quantidade de banheiro o imóvel foi removido, uma vez que, por regra prática, toda residência deve ter pelo menos um banheiro.

Após a aplicação do filtro, o número de observações restantes foi de 47412 (10998 casas e 36414 apartamentos), significando que houve perda de apenas 7,55% dos dados.

Como o valor de IPTU e de condomínio não é especificado em grande parte dos imóveis (75,35% e 55,96%, respectivamente) essas variáveis não foram utilizadas nas análises subsequentes, pois agregam pouca ou nenhuma informação sobre as características gerais das propriedades.

A figura 2 mostra uma análise descritiva de algumas das variáveis presentes na base já filtrada. Dela, são extraídas algumas considerações interessantes:

- Apartamentos, no geral, possuem um menor número de quartos, banheiros e vagas na garagem se comparados a casas, mas em ambos os casos a moda do número de quartos é 3. Também foi constatado que há anunciantes que consideram um espaço no terreno grande o suficiente para um veículo como uma vaga na garagem. Isso inflaciona a variável.

- 54,51% dos imóveis tem somente uma suíte e apenas 0,04% possuem mais de 5. Observa-se ainda um comportamento estranho com relação aos apartamentos, onde imóveis com 3 suítes aparecem com maior frequência do que os com 2 suítes. Além disso, há imóveis registrados com 9 desses cômodos. Isso pode ser justificado por uma não-conformidade na hora do cadastro do número de suítes, de modo que não se sabe se o anunciante contabiliza quartos e suítes como coisas distintas.

- Com relação ao preço por m^2, observa-se que apartamentos dispõem de preços mais variáveis se comparados a casas. Além disso, o valor do m^2 é maior se utilizada a mesma comparação.

- Examinando o número de fotos, há algumas barras que aparecem em destaque (*e.g.* a de 50 fotos). Aparentemente, algum erro ocorre na hora do site criar os *links*. Sendo assim, apesar de existirem poucas imagens, o número coletado pelo algoritmo de *web scraping* é maior do que o número real de fotos.

- Apartamentos geralmente possuem uma área menor do que as casas, além de usualmente serem mais novos. Esse comportamento é comum e bem intuitivo,

considerando que o surgimento de edifícios é mais recente e que geralmente o espaço horizontal para construção de uma casa é maior.

- Pode-se observar em todos os gráficos de caixas mostrados na figura um número grande de pontos marcados como observações com valores extremos (*outliers*), o que é um comportamento esperado dado a assimetria da distribuição dessas variáveis.

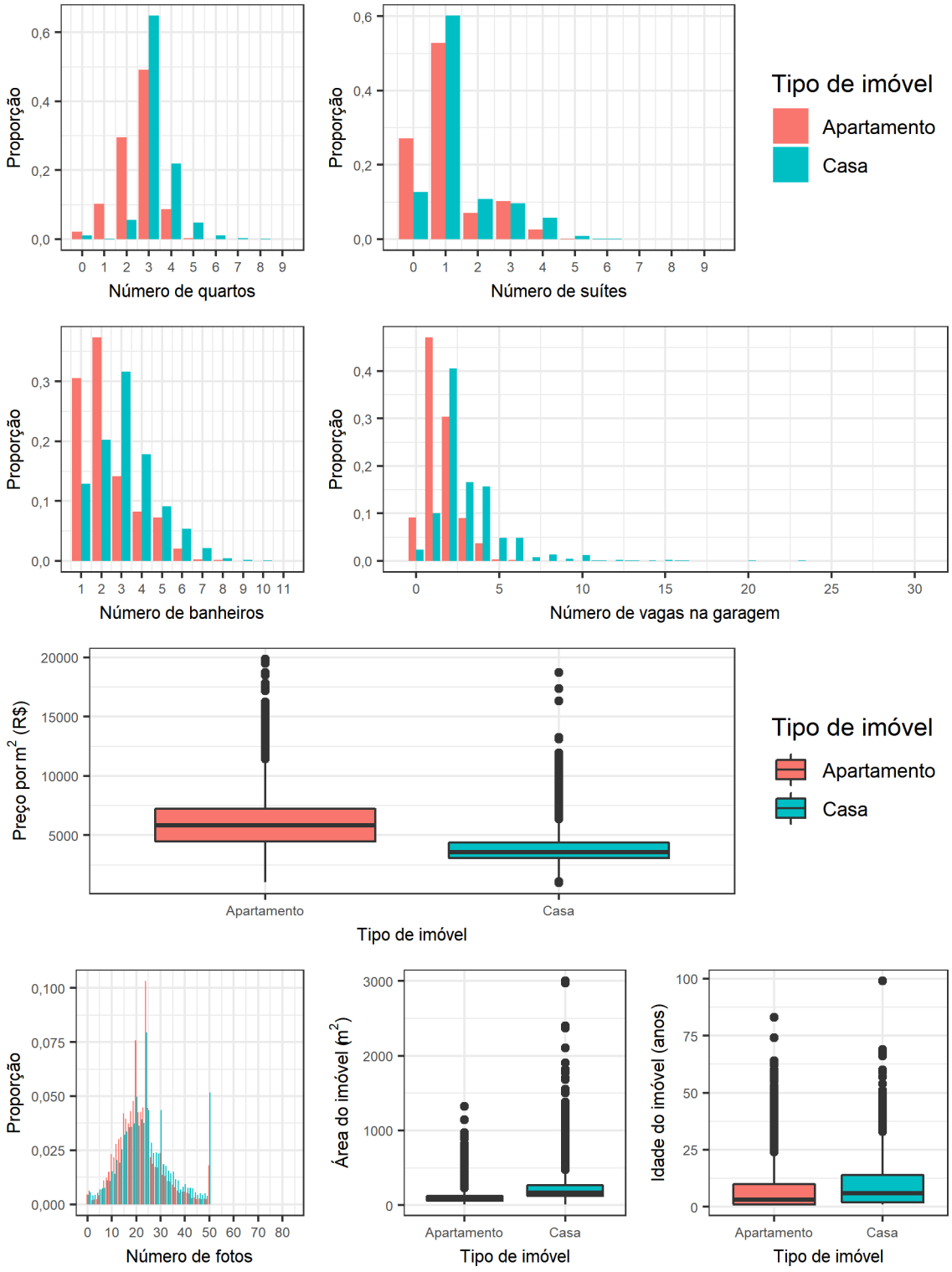


Figura 2 – Conjunto de gráficos descritivos das variáveis que caracterizam os imóveis anunciados.

Tabela 1 – Estatísticas descritivas das variáveis numéricas da base de dados.

Variável	Dados válidos	Média	Mínimo	Q1	Mediana	Q3	Máximo	Desvio padrão
pictures	47412	22,28	0,00	16,00	21,00	26,00	84,00	9,79
usefulArea	46478	130,44	10,00	63,00	96,00	160,00	3000,00	112,38
bedroom	47412	2,70	0,00	2,00	3,00	3,00	9,00	0,92
suite	47412	1,16	0,00	1,00	1,00	1,00	9,00	1,02
bathroom	47412	2,52	1,00	1,00	2,00	3,00	11,00	1,42
garage	47412	1,89	0,00	1,00	2,00	2,00	30,00	1,54
years	25226	8,35	1,00	1,00	3,00	11,00	99,00	10,28
publishment	43294	124,25	2,00	21,00	66,00	199,00	1416,00	137,59
area	47412	132,36	10,00	63,00	98,00	161,00	3000,00	117,64
priceSM	47412	5552,15	1000,00	3834,43	5186,67	6842,29	19844,44	2173,43

A tabela 1 mostra os resumos numéricos usuais para descrição das variáveis mostradas na figura 2, mas agora sem a estratificação pelo tipo de imóvel. Além dos indicativos de assimetria da distribuição, é notável a diferença na quantidade de dados válidos (coluna N na tabela).

A disposição espacial dos preços de imóveis é uma informação muito relevante para o estudo, pois indica se um componente ou uma covariável que captura esse comportamento no espaço deve ser levada em consideração. A figura 3 mostra que o preço por m^2 é maior em apartamentos[3], corroborando com o que foi visto na figura 2. Além disso, observa-se que os preços mais baixos ocorrem na periferia da cidade. Uma mancha mais avermelhada chama a atenção na região de divisa do bairro Umbará com o Campo de Santana, porém não foi encontrada uma explicação para esse fato.

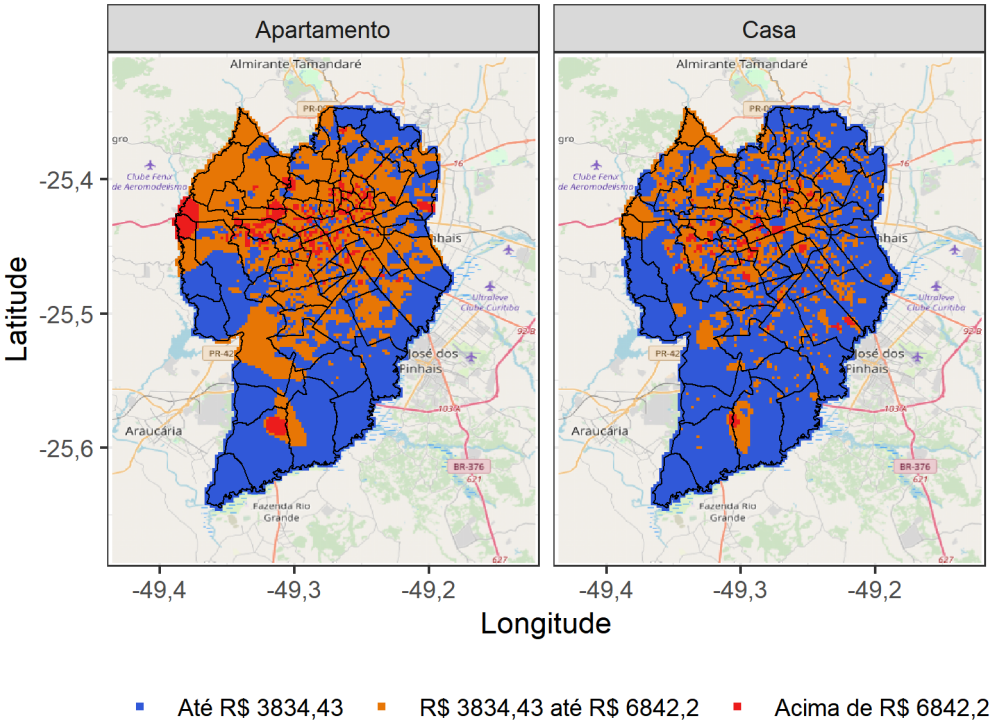


Figura 3 – Mapa mostrando o fatiamento do preço por m^2 (calculado com base nos quartis da distribuição marginal) estimado pelo método de ponderação pelo inverso da distância (IDW) no território da cidade de Curitiba.

3.2 Mineração de Texto

A mineração de texto consiste em utilizar um conjunto de ferramentas para analisar uma coleção de documentos, a fim de extrair alguma informação dos dados

[3] O primeiro item da legenda indica o valor ao qual 25% dos dados são menores. O segundo indica a faixa de valores em que se encontram os preços por m^2 de 50% (do segundo ao terceiro quartil) dos imóveis. Já o terceiro item da legenda indica o valor do m^2 ao qual 25% dos imóveis são mais caros.

textuais por meio da identificação e exploração de padrões (FELDMAN; SANGER, 2007).

3.2.1 Definições e Conceitos

Como em qualquer campo de estudo, há algumas terminologias específicas da mineração de texto:

- **Documento**: entende-se como qualquer registro de informação em texto.
- **Corpus**: é uma coleção de documentos.
- **Termo**: é uma única palavra, ou um conjunto de palavras, selecionada diretamente do *corpus* por uma metodologia de extração (FELDMAN; SANGER, 2007).
- **Vocabulário**: é um conjunto de termos.
- **Token**: é uma unidade de texto dotada de significado (*e.g.* uma palavra) que se tem interesse em utilizar para a análise (SILGE; ROBINSON, 2017).
- **Tokenização**: é o processo de dividir o texto em *tokens* (SILGE; ROBINSON, 2017).
- **Stop words**: são palavras que não acrescentam informação na análise, como palavras muito gerais (*e.g.* artigos, preposições, conjunções, ...) ou palavras utilizadas em grande parte dos documentos do *corpus*.
- **Stemming**: consiste em transformar as formas variantes de uma palavra em uma representação comum (HUYCK; ORENGO, 2001), útil para a eliminação de variações de escrita devido à inflexões de gênero, número (plural) e grau.

3.2.2 Vector Space Model (VSM)

Como os algoritmos de aprendizado não processam o texto na forma original, o documento é convertido em uma representação mais gerenciável, como vetor de características, onde as característica são simplesmente termos que descrevem o documento (FELDMAN; SANGER, 2007). Portanto, a representação algébrica dos documentos de texto para utilização em um método estatístico foi feita pelo modelo de espaço vetorial (*Vector Space Model*).

Se a matriz de dados é construída de forma que cada uma de suas linhas seja um documento e cada uma das colunas sejam termos, com o corpo da matriz representando a frequência (ou uma função da frequência) com que cada termo aparece no documento, diz-se que os documentos são retratados como um vetor no espaço indexado pelos termos (MANNING; SCHÜTZE, 1999). Formalmente, seja um espaço consistindo de

documentos D_i, de modo que cada documento seja identificado por um ou mais termos indexados T_j, então:

$$D_i = (w_{i1}, w_{i2}, \ldots, w_{it}), \qquad i = 1, 2, \ldots, n \tag{3.1}$$

onde $w_{ij} \geq 0$ representa o peso do j-ésimo termo no documento i e t corresponde à quantidade de termos indexados (SALTON; WONG; YANG, 1975).

Para entender melhor o conceito, a figura 4 ilustra dois documentos representados em um espaço bidimensional, onde cada um dos eixos simboliza um termo (T_1 e T_2) e os documentos são colocados nas respectivas coordenadas de acordo com a quantidade de vezes que o termo aparece no documento. Portanto, o peso (w_{ij}) foi definido como a frequência absoluta. Como mostra a figura 4, os documentos D_1 e D_2 são mais parecidos entre si, já que as palavras que os constituem apresentam frequências parecidas. O documento D_3 é pouco similar aos demais, uma vez que nele os dois termos aparecem com uma frequência menor.

A matriz que representa os documentos na figura 4 pode ser construída de modo que cada linha seja o vetor definido por 3.1, constituído de 2 termos indexados, ou seja, $t = 2$. Essa matriz é chamada de matriz documento-termo:

$$\text{documento-termo} = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \end{array} \begin{array}{c} \begin{array}{cc} T_1 & T_2 \end{array} \\ \begin{bmatrix} 10 & 4 \\ 9 & 7 \\ 1 & 3 \end{bmatrix} \end{array}$$

Inconvenientemente, a frequência com que cada termo ocorre em um documento tem uma relação diretamente proporcional com o tamanho do documento. Por isso são utilizadas funções da frequência para remover ou amenizar influências deste e de outros problemas nesta forma de representação.

3.2.3 Escolha do método de ponderação

A partir do momento em que a representação do texto de forma algébrica está bem definida, surge a questão de como ponderar os termos, ou seja, qual será a função da frequência utilizada como w_{ij}.

De acordo com MANNING; SCHÜTZE existem três quantidades que são comumente utilizadas como base para ponderar os termos: tf_{ij}, que é a frequência do termo j no documento i; df_j, que é o número de documentos em que o termo j ocorre; cf_j, representando o número de ocorrências do termo j dentro de um *corpus* de documentos[4].

[4] Observe que a quantidde cf_j, diferentemente do df_j, pode ultrapassar o valor de n, já que um termo

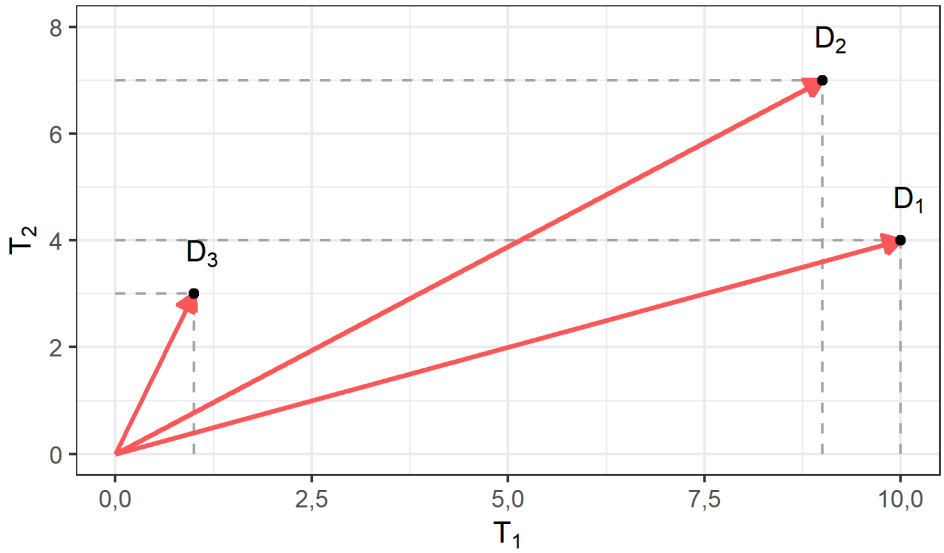


Figura 4 – Representação de três documentos em um espaço de termos bidimensional.

Cada uma dessas quantidades tem suas peculiaridades e as duas primeiras são largamente utilizadas, pois elas possuem interpretações práticas interessantes. Quanto maior for o valor de tf_{ij}, maior é a probabilidade do termo descrever o documento. Usualmente utiliza-se uma transformação desse valor, já que a importância do termo não cresce de forma linear à medida que sua frequência aumenta (MANNING; SCHÜTZE, 1999). Já a quantidade df_j pode ser considerada um indicador da quão comum ou não é o termo, visto que os termos que ocorrem em todos os documentos tendem a não trazer informação semântica relevante. Desse modo, geralmente utiliza-se uma função do inverso dessa medida chamada de *inverse document frequency* (idf) (SILGE; ROBINSON, 2017) que é definida pela equação 3.2.

$$\text{idf}_j = \ln\left(\frac{n}{\text{df}_j}\right) \tag{3.2}$$

Multiplicando a medida de frequência do termo (tf) pela função definida na equação 3.2 tem-se o tf-idf, que mensura a importância de um termo para um documento, levando em conta o *corpus* (SILGE; ROBINSON, 2017).

Sendo assim, para ponderar os termos que apareceram após a *tokenização* da descrição do anúncio, foi utilizado o tf-idf normalizado pelo tamanho do documento ($||D_i||$), escrito na equação 3.3, de forma similar ao que foi descrito por HIEMSTRA, que afirma que essa combinação gera os melhores resultados[5] para recuperação de

pode ocorrer várias vezes dentro do mesmo documento.

[5] Apesar da afirmação ter sido feita nos anos 2000, a métrica é muito utilizada ainda hoje, constando, senão em todas, na maiorias das bibliografias de mineração de texto consultadas para construção desse trabalho.

documentos[6], além de mostrar uma interpretação probabilística para a função.

$$w_{ij} = \text{tf-idf}_{ij} = \frac{\text{tf}_{ij}}{||D_i||} \ln\left(\frac{n}{\text{df}_j}\right) \tag{3.3}$$

Já para as tags, como elas são uma unidade e representam uma característica particular dos imóveis, foi utilizada a ponderação binária, a qual marca somente se o termo está ou não presente no documento, como mostra a equação 3.4.

$$w_{ij} = \begin{cases} 1 & \text{se } T_j \in D_i \\ 0 & \text{caso contrário} \end{cases} \tag{3.4}$$

3.2.4 Obtenção dos Termos

Antes de extrair e ponderar os termos faz-se necessário uma limpeza nos documentos do *corpus*, visto que a presença de termos irrelevantes cria dimensões desnecessárias na matriz documento-termo. Inicialmente os textos foram transformados para caixa baixa, já que não há necessidade de considerar a capitalização, além desse ser um procedimento muito empregado. Utilizando expressões regulares, foram também removidos os acentos, a pontuação e os símbolos possivelmente presentes.

Como muitos dos anúncios citam o bairro onde se localiza o imóvel, optou-se por remover todos os nomes de bairros de Curitiba que estavam presentes no texto, ou seja, esses nomes foram adicionados à lista de *stop words*. Uma lista de *stop words* foi obtida de Lopes (2013), e esta também foi incrementada com palavras específicas do contexto sob estudo, sendo ela utilizada para a limpeza do *corpus* (Apêndice A).

O algoritmo GloVe (PENNINGTON; SOCHER; MANNING, 2014) foi utilizado como uma forma de diminuir a esparsidade da matriz de documento-termo. Esse algoritmo é um método de aprendizado não supervisionado[7] que busca obter uma representação vetorial dos termos presentes no *corpus*. A partir disso, a correlação entre dois termos pode ser calculada para indicar se eles são sinônimos.

Buscando capturar melhor as relações entre os termos, além de preparar os objetos necessários para a aplicação do GloVe, alguns passos intermediários foram executados:

1. Após a limpeza do texto (como descrito acima), ainda foram excluídas as palavras com menos de 3 caracteres, já que elas geralmente são *stop words*.

[6] A recuperação de documentos consiste em representar a requisição textual de um usuário e, a partir disso, retornar os documentos relevantes à pesquisa.

[7] São aqueles métodos aplicados a situações nas quais observou-se o vetor de características mas nenhuma variável resposta associada a ele.

2. O algoritmo RSLP (HUYCK; ORENGO, 2001) foi utilizado[8] para obter o *stemming* da palavra, possibilitando que os sinônimos fossem encontrados de forma independente das variações da mesma palavra (*e.g.* singular e plural).

3. A matriz de co-ocorrência de termos (TCM) foi calculada, já que ela é a base para o GloVe.

O treinamento do modelo GloVe foi feito utilizando o pacote *text2vec* (SELIVANOV; WANG, 2018), disponível para o ambiente R (R Core Team, 2017). Desse modo, após a obtenção da representação vetorial das palavras, foi montado um algoritmo para fazer a substituição dos termos na base de dados:

1. Foi criada uma tabela com três colunas: a primeira contém as palavras, a segunda contém o *stemming* das palavras e a terceira é uma cópia da primeira. Na tabela, a primeira palavra é a mais frequente e a última é a menos frequente.

2. Para cada palavra na segunda coluna da tabela calcula-se a correlação desta com as outras palavras utilizadas no treinamento do modelo[9].

3. Seleciona-se as palavras que tem correlação maior do que 0,9 com a palavra que está sendo atualmente verificada no laço.

4. Se a nova palavra já apareceu anteriormente na lista, então substitui-se a palavra atual da terceira coluna pela palavra na terceira coluna da ocorrência anterior.

Todo processo descrito foi representado de forma visual e simplificada no fluxograma da figura 5. Essa sequencia de passos irá garantir que termos semelhantes sejam trocados por um único sinônimo. Além disso, como a tabela está organizada pela frequência, espera-se que as palavras que aparecem no início estejam corretas, já que muitos anúncios utilizaram das mesmas. Isso ajuda na não propagação de palavras erradas.

Uma consequência interessante da utilização desse algoritmo de substituição é que o termo será mantido da forma que mais aparece, por exemplo, se ele ocorre mais frequentemente no singular isso será propagado pelo restante da tabela.

3.3 Descritiva das variáveis textuais

Uma observação importante a ser feita é que as variáveis textuais mostradas nesta seção já foram tratadas com os passos descritos na seção 3.2.4. Além disso, optou-

[8] Pacote *rslp* (FALBEL, 2016).
[9] Note que o modelo GloVe foi ajustado com base no *stemming* da palavra.

se por remover as palavras que ocorrem em menos de 5 anúncios ($df_j < 5$), uma vez que elas não agregam muita informação sobre o *corpus*.

3.3.1 Título do anúncio

A figura 6 mostra as 30 palavras do título do anúncio com maior peso na representação ponderada pelo tf-idf. A palavra "vende" aparece seguida da palavra "residencial", sendo que elas são um reflexo dos imóveis sob estudo, já que, como citado no início do desse capítulo (capítulo 3), a base é constituída de imóveis residenciais à venda.

Grande parte dos termos que aparecem na figura 6 representam características do imóvel, sendo que alguns deles aparecem de forma abreviada (*e.g.* dorms e apto, que são abreviações para dormitório e apartamento). Entre as palavras mostradas, dois termos causam estranhamento a primeira vista: "axis" e "garden". O primeiro se refere a um nome de imobiliária (Axis 21) que "assina" os títulos de seus anúncios, já o último é o nome que se dá a apartamentos no térreo que possuem um quintal isolado da área de circulação.

Após a troca das palavras por seus sinônimos, o termo "dormitorio" foi substituído pelo termo "suite", mas isso não ocorre com sua abreviação ("dorms"). Isso acontece porque ela aparece poucas vezes na descrição do imóvel (utilizada para o ajuste do algoritmo GloVe), levando à não captura da relação.

Outro ponto interessante a ser observado é a relação entre as palavras "alto" e "padrao". Nos anúncios as duas frequentemente aparecem juntas na frase alto padrão, mas o primeiro termo também faz referência à apartamentos em andares altos.

3.3.2 Descrição do anúncio

A descrição do imóvel é parte crucial do anúncio. Nela o anunciante tem o espaço para convencer o cliente a pelo menos dar uma olhada no empreendimento. Por esse motivo, as covariáveis geradas a partir dela foram incluídas na modelagem do preço do metro quadrado do imóvel, sendo a descritiva do título deixado apenas como um passo para entender melhor o conjunto de dados.

Os 50 termos da descrição do imóvel com maior peso pela métrica do tf-idf estão sendo representados na figura 7. A palavra "previo" aparece em primeiro lugar nesse ranking, fazendo alusão a uma frese utilizada em vários anúncios:

"[...] dados e valores sujeitos a alteração sem aviso prévio."

Assim como no título do anúncio, muitos dos termos que aparecem indicam características do imóvel. Outros, bem como a palavra "previo", apontam características diretas do próprio anúncio. Logo abaixo alguns deles são explicados:

- "visita": é uma característica do anúncio, que pede pra agendar a visita.
- "dados": é uma característica do anúncio, já que o anunciante deixa um link escrito "ver dados" para acesso às informações de contato.
- "disponibilidade": é uma característica do anúncio, já que o anunciante pede ao cliente para consultar a disponibilidade do empreendimento.
- "unidade": é uma característica do anúncio que está associada a disponibilidade ("[...] consulte a disponibilidade da unidade.").
- "social": característica do imóvel, que possui "banheiro social".
- "contendo": geralmente utilizado antes de apontar as características do imóvel no anúncio.
- "face": associado a "face norte", que são apartamentos que, por estarem localizados no sul do país (Curitiba), pegam mais sol durante o dia.
- "total": faz referência a área total do imóvel.

É importante evidenciar que as palavras "cozinha" e "banheiro" foram classificas pelo algoritmo como sinônimos pois frequentemente aparecem dentro do mesmo contexto. Apesar dessa troca, de forma geral o algoritmo faz um bom trabalho, uma vez que captura bem a relação de grande parte dos termos, levando a uma redução de 28,3% no vocabulário (de 29114 palavras para 20876).

3.3.3 Descrição das Tags

Algumas das tags possuem informações mais detalhadas que outras. Um bom exemplo é a tag "Posição do Apto", que especifica também qual é a posição do apartamento em relação a entrada do edifício (*e.g.* tem a face voltada para: frente, fundos, lateral, ...).

A tag "Andares" é a única na qual o anunciante é livre para compor a informação, que neste caso é o número do andar. Por esse motivo foram detectados alguns valores errados (*e.g.* valores negativos). Sem a informação extra ela simplesmente indicaria que o imóvel é um apartamento, o que já é apontado pelo tipo do imóvel. Sendo assim, optou-se por remover ela da análise.

Na figura 8 as tags são apresentadas por meio de uma nuvem de palavras. Nessa visualização, tags que aparecem com fonte maior são as mais utilizadas. Observa-se que a tag "Churrasqueira" é a mais frequente, seguida por "Área de Serviço" e "Salão de Festas". Também fica claro que há várias delas com frequência muito baixa, como é o caso da tag "Lavoura", que aparece apenas uma vez (Apêndice B).

3.4 Inclusão do zoneamento urbano

De acordo com o Ministério do Meio Ambiente o zoneamento urbano é um instrumento que divide a cidade em zonas, definindo regras para o uso e ocupação do solo em cada uma delas. A cidade de Curitiba é dividida em 14 classes de zoneamento, sendo que a prefeitura disponibiliza os *shapefiles*[10] com informação de zoneamento, divisa de bairros, e outras. Desse modo, os dados foram obtidos pelo site do IPPUC (2018), e a informação de zona e bairro foi anexada ao conjunto de características dos imóveis (*zone* e *neighborhood*, respectivamente).

Na tabela 2 estão as porcentagens de imóveis da base de dados localizados em cada uma das zonas, estando estas estratificadas pelo tipo de empreendimento. Observa-se que grande parte dos imóveis estão em zonas residenciais, seguida por setores especiais, que são aqueles nos quais são estabelecidos critérios especiais para o uso e ocupação do solo, condicionais as características do local [11]. O mapa do zoneamento urbano de Curitiba, disponibilizado pelo IPPUC (2018), pode ser visto no Anexo A.

Tabela 2 – Porcentagem de apartamentos e casas em cada zona urbana.

Zona	Apartamento	Casa
ZONAS RESIDENCIAIS	55,50%	76,97%
SETORES ESPECIAIS	31,32%	9,10%
OPERAÇÃO URBANA CONSORCIADA LINHA VERDE	4,00%	6,13%
ZONA CENTRAL	3,76%	0,13%
ZONAS DE TRANSIÇÃO	3,41%	1,59%
SETORES ESPECIAIS DOS EIXOS DE ADENSAMENTO	1,04%	1,75%
ZONAS ESPECIAIS	0,38%	0,14%
ZONAS DE SERVIÇOS	0,17%	0,98%
UNIDADE DE CONSERVAÇÃO	0,12%	0,68%
ZONAS INDUSTRIAIS	0,10%	0,05%
ÁREA DE PROTEÇÃO AMBIENTAL DO PASSAÚNA	0,09%	2,27%
ZONAS DE USO MISTO	0,06%	0,06%
ZONA DE CONTENÇÃO	0,04%	0,01%
ÁREA DE PROTEÇÃO AMBIENTAL DO IGUAÇU	0,01%	0,14%

Há poucos imóveis nos grupos de zonas que aparecem nas últimas linhas da tabela 2. Por esse motivo optou-se por criar uma nova categoria que engloba: zonas especiais, zonas de serviços, unidade de conservação, zonas industriais, área de proteção ambiental do Passaúna, zona de uso misto, zona de contenção e área de proteção ambiental do Iguaçu. Essa nova categoria foi chamada de "OUTRAS".

[10] Arquivo contendo os dados para armazenar a posição, forma e atributos de feições geográficas (ARCGIS,).

[11] A lei municipal N° 9800/2000 dispõe sobre o zoneamento e define o uso e ocupação do solo para cada uma das zonas (CURITIBA, 2015)

3.5 Inclusão dos Alvarás

Não é difícil supor que, além das características do próprio imóvel, detalhes acerca das redondezas de onde se vai morar podem ter um peso fundamental na escolha de uma residência. Para mensurar essas características, foi utilizada uma base do acervo de dados abertos do Centro de Computação Científica e Software Livre (C3SL, 2018), onde se encontram dados de todos os alvarás comerciais da cidade de Curitiba, estejam eles ainda em atividade ou não. Foram obtidos dados de 1204758 alvarás, com registros emitidos de dezembro de 1899 até junho de 2018. As seguintes informações constituem a base:

- Nome empresarial;
- Data de início das atividades;
- Número de registro do alvará;
- Nome da empresa [12];
- Data de emissão da licença;
- Data de expiração da licença (em branco caso ainda esteja em vigor);
- Atividade principal do estabelecimento;
- Primeira atividade secundária (opcional);
- Segunda atividade secundária (opcional);
- Endereço;
- Número predial;
- Unidade;
- Andar;
- Complemento;
- Bairro;
- CEP.

Alguns critérios foram adotados para filtrar os alvarás:

[12] Pelo dicionário de dados, disponibilizado pelo próprio C3SL, não consta uma diferença entre esta variável e a referente ao nome empresarial. O responsável pelos dados foi contactado a respeito, sem resposta.

Categoria	Palavras-chave
Alimentação	quitanda mercearia panificadora lanchonete restaurante acougue pizzaria petiscaria
Estacionamento	estacionamento
Exercício	natacao musculacao ginastica ballet marciais[13] danca
Saúde	clinica[14]
Educação	escola creche educacao colegio

Tabela 3 – Relação de palavras utilizadas para criar as categorias de alvarás.

- Ele deve estar ativo;
- Deve ter um valor plausível para a data de início de suas atividades e data de emissão do alvará;
- O número do alvará deve ser diferente de zero;
- A descrição principal do alvará não pode estar em branco;
- Deve constar endereço ou CEP;
- Não deverá haver réplicas de registros. Como é muito difícil de saber qual o alvará original no caso de um valor repetido, foi optado por remover todos os números de licença duplicados.

Após a filtragem inicial, foram definidas categorias para padronizar os tipos de alvarás disponíveis, com base em palavras-chave. A tabela 3 demonstra essa relação.

[13] Refere-se a Artes marciais.
[14] Aponta para estabelecimentos de saúde, como hospitais, que também estão registrados como "clínicas".

Registros cujas descrições de atividades, seja primária ou secundária, contenham quaisquer das palavras-chave descritas na tabela 3, se mantém na base. Por fim, os endereços foram padronizados e utilizados para a coleta de latitude e longitude.

3.5.1 Construção da covariável por kernel

Espera-se que imóveis localizados nas proximidades de estabelecimentos que fornecem alimentação, serviços de saúde, e outros serviços (como aqueles utilizados para categorizas os alvarás) tenham um valor de mercado mais elevado. Para que os modelos propostos levassem em conta, de alguma forma, a quantidade de estabelecimentos próximos a cada imóvel, foram criadas novas variáveis utilizando um método de estimação de densidade por kernel.

O método consistiu em centrar uma distribuição normal bivariada (equação 3.5) na localização do imóvel e, a partir disso, calcular o valor da densidade das categorias estabelecidas na tabela 3. O valor retornado pela função de densidade por kernel é a média desses valores (3.6). Repetindo esses passos foram obtidas as densidades estimadas para cada um dos imóveis em cada uma das categorias.

$$K(h|\sigma) = \left(\frac{1}{2\pi\sigma_{lat}\sigma_{long}}\right) exp\left\{-\frac{1}{2}\left[\left(\frac{h_{lat}}{\sigma_{lat}}\right)^2 + \left(\frac{h_{long}}{\sigma_{long}}\right)^2\right]\right\} \tag{3.5}$$

em que h_{lat} e h_{long} são as diferenças entre as coordenadas de um imóvel e um estabelecimento específico, e σ_{hat} e σ_{long} representam desvios padrão da normal bivariada e definem até que ponto considera-se um ponto comercial próximo suficiente do imóvel.

$$\hat{f}_j(x) = \frac{1}{n_j}\sum_{i=1}^{n_j} K(h_{ij}|\sigma) \tag{3.6}$$

em que n_j é a quantidade de estabelecimentos encontrados dentro da categoria j. Já x é um vetor bidimensional de coordenadas do imóvel e h_{ij} representa a diferença entre x e as coordenadas do estabelecimento i da categoria j dos alvarás.

O valor da constante normalizadora (quantidade que multiplica a parcela exponencial na equação 3.5) faz com que a área abaixo da curva integre 1, mas para os propósitos pretendidos ela não faz diferença, já que suas quantidades são fixas para todas as observações. Retirando essa constante da equação obtêm-se um índice que atinge o valor 1 quando todos os estabelecimentos estão no mesmo local do imóvel e decai pra 0 a medida que os estabelecimentos se distanciam e se espalham pelo mapa.

Os valores de σ_{lat} e σ_{long} foram escolhidos de modo que eles representassem um

raio de mais ou menos 0,5 km de distância[15]. Portanto o índice atribuído as observações decai de forma que os empreendimentos com uma distância, em relação ao imóvel, maior do que 1,5 Km tem peso praticamente nulo.

Pela figura 9 fica claro que os alvarás estão concentrados na região central da cidade, além da notável similaridade entre o comportamento entre as categorias. Na figura 10 observa-se que as variáveis geradas são altamente correlacionadas, visto que a que medem a quantidade de estabelecimentos de saúde e a que mede a quantidade de estabelecimentos da categoria exercício são as que possuem a menor correlação, e ainda assim o valor estimado é de 0,87.

3.6 Modelo

3.6.1 Ajustes nos dados

Optou-se por remover as tags com frequência menor do que 5, por conta da estimativa do coeficiente relacionado a elas ser muito errônea. Esse critério retirou as tags: cachoeira, celeiro, divisórias, energia elétrica (aparentemente energia elétrica é algo tão trivial que não é informado nas tags), estacionamento, lavoura, murado, pasto, refeitório e a tag rio.

Com relação às variáveis textuais, foi escolhido deixar somente as 1000 palavras com maior soma do tf-idf por termo, pois colocando todas elas do menor para o maior valor do somatório em um gráfico, observa-se que a partir desse número os valores do tf-idf tem baixa taxa de decaimento (figura 11).

3.6.2 Notação das variáveis

Para conseguir distinguir qual a origem da variável, uma *string* foi adicionada ao final do nome da mesma. As seguintes regras foram adotadas:

- Para as variáveis que vieram diretamente do site Imovelweb, bairro e zona, não foi adicionado nada aos nomes, portanto manteve-se os mesmos nomes citados anteriormente.

- Para as covariáveis geradas com a análise textual da descritiva foi adicionado a string "_txt" ao final do nome.

[15] O centroide da cidade de Curitiba (Latitude: -25,441105 Longitude: -49,276855) foi utilizado para calcular quanto vale 1 unidade da coordenada geográfica no local. Foi encontrado que 1 unidade de latitude equivale a aproximadamente 111,21 Km e uma de longitude 72,76 Km. A partir disso as devidas conversões foram realizadas.

- Para as covariáveis geradas com base nas tags informadas no anúncio do imóvel, foi adicionado a string "_tag" ao final do nome.

- Para as covariáveis geradas a partir da base de dados dos alvarás, foi adicionado a string "_alv" ao final do nome.

3.6.3 Remoção de covariáveis correlacionadas

As covariáveis incluídas no modelo devem ser altamente correlacionadas com a variável resposta e possuírem baixa correlação entre elas para não causarem problemas na estimação. Como visto na seção 3.5.1 (figura 10), todas as variáveis construídas a partir da base de dados dos alvarás tem forte correlação, o que aumenta a variância dos coeficientes estimados. Ademais, é de interesse prático remover essas variáveis do modelo, já que elas não agregam mais informação e diminuem a parcimônia, dificultando a interpretação de forma geral.

Para esse fim, foi montado um algoritmo que calcula as correlações e remove variáveis a partir de um ponto de corte. Ele consiste nos seguintes passos:

- Calcula-se o coeficiente de correlação de Pearson (3.7) entre cada uma das variáveis independentes e a dependente. No caso em que o cálculo é realizado entre a variável dependente, que é numérica, e uma variável proveniente das tags (nominal dicotômica), esse coeficiente é chamado de correlação ponto-bisserial.

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.7}$$

  onde $\bar{x}$ e $\bar{y}$ são as médias da variável x e y, respectivamente.

- Ordena-se a lista, de forma que a primeira covariável que aparece é a mais correlacionada com a resposta.

- Percorre-se a lista, sendo que em cada passo calcula-se uma medida de similaridade entre a variável atual e todas as outras que estão abaixo dela, removendo-se as variáveis muito similares a ela ($|\text{coeficiente}| > 0,6$[16]).

A similaridade entre duas variáveis numéricas ou entre uma numérica e uma dicotômica foi calculada utilizando a correlação de Pearson, que é dada pela equação 3.7.

[16] Outros valores foram testados e verificou-se que a escolha entre esse ponto de corte ou um que considera correlações mais fortes não mudava muito a escolha das variáveis.

Já entre duas variáveis dicotômicas, foi calculado o índice de similaridade de Jaccard (3.8).

$$J(x,y) = \frac{\sum_{i=1}^{n} I(x_i = 1)I(y_i = 1)}{\sum_{i=1}^{n} [I(x_i = 1)I(y_i = 0) + I(x_i = 0)I(y_i = 1) + I(x_i = 1)I(y_i = 1)]} \tag{3.8}$$

onde $I(\cdot)$ é a função indicadora, que assume valor 1 quando a condição passada como seu argumento é verdadeira e zero caso contrário.

3.6.4 Regressão Lasso

A regressão lasso faz parte de um conjunto de métodos chamados de Métodos de Encolhimento (*Shrinkage Methods*), nos quais se ajusta o modelo completo (com as p preditoras) e aplica-se restrições nos coeficientes do mesmo, de forma que os "encolhe" para zero (JAMES et al., 2013).

A lasso é uma técnica aplicada com objetivos similares aos métodos de seleção *forward stepwise*, e *backward stepwise*, mas diferente desses, o modelo é ajustado somente uma vez, minimizando-se a equação 3.9, e as variáveis selecionadas são aquelas em que os coeficientes tem valor estimado diferente de zero.

$$S_\lambda(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = SQR + \lambda R \tag{3.9}$$

em que:

- SQR é a soma do quadrado dos resíduos;
- R é o o termo de regularização;
- λR é chamado de penalidade de encolhimento (*shrinkage penalty*);
- λ é chamado de parâmetro de *tuning* e controla o impacto relativo da regularização nos parâmetros estimados.

Quando $\lambda = 0$ a equação restringe-se ao caso da regressão linear múltipla, ao passo que quanto maior o valor de λ maior é a penalização sob os valores de β, forçando alguns deles a chegarem a zero. Usualmente, o valor desse parâmetro é escolhido por meio de validação cruzada (*cross-validation*), técnica na qual o modelo é ajustado com base em diferentes partições da base de dados, escolhidas sistematicamente, e o erro

quadrático médio (EQM) é calculado. Isso é feito para um conjunto de valores para λ e escolhe-se aquele que retorna o menor EQM.

A não ser quando $\lambda = 0$, as estimativas dos coeficientes variam substancialmente com a mudança de escala das preditoras e, por esse motivo, recomenda-se que as preditoras sejam padronizadas antes da aplicação do método, de forma que todas elas tenham desvio padrão igual a 1.

O modelo foi ajustado por intermédio do pacote *glmnet* (FRIEDMAN; HASTIE; TIBSHIRANI, 2010), utilizando a função *cv.glmnet*, que realiza a validação cruzada 10-fold por padrão e já faz a padronização das variáveis preditoras. A matriz do modelo foi passada como a matriz de preditoras. Dessa forma também leva-se em consideração as variáveis categóricas, representadas como *dummies*.

3.6.5 Dados de treino e de validação

Essa abordagem consiste em dividir a base de dados aleatoriamente em duas partes. Uma delas é utilizada para o ajuste do modelo (base de treino) e outra para fazer predição do modelo já ajustado (base de validação). O erro quadrático médio calculado a partir da base de validação fornece uma estimativa do erro de predição em dados novos (dados de teste) (JAMES et al., 2013).

O conjunto de dados foi dividido de modo que 90% das observações ficassem para treino (42670) e 10% para a validação (4742).

3.6.6 Especificação do modelo

Os modelos propostos tem como fundamentação a teoria dos modelos de regressão linear múltipla. Nessa família de modelos há a suposição de que a variável dependente segue uma distribuição normal com a média μ_i e variância σ^2 (equação 3.10). Portanto, a média é uma função linear de um conjunto de variáveis (equação 3.11), mas a variância do modelo é constante.

$$y_i \sim Normal(\mu_i, \sigma^2) \tag{3.10}$$

$$\mu_i = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p \tag{3.11}$$

Para estimação dos coeficientes é utilizado o método de mínimos quadrados ordinários, que no caso do modelo normal é equivalente ao método de verossimilhança. A função *lm* foi empregada para o ajuste e os gráficos de diagnóstico de resíduos usuais foram analisados.

Os modelos propostos estão definidos de forma genérica pelas equações 3.12 e 3.13. A variável resposta está transformada pelo logaritmo natural, já que além do procedimento ser usual em modelos hedônicos, testes revelaram que essa transformação é necessária do ponto de vista estatístico, já que sem ela o pressuposto de variância constante dos resíduos não é atendido.

$$\ln(priceSM)_i \sim Normal(\mu_i, \sigma^2) \tag{3.12}$$

$$\mu_i = \beta_0 + \sum_{j \in I} \beta_j \cdot x_{ij} + \sum_{j \in J} \beta_j \cdot x\_tag_{ij} + \sum_{j \in K} \beta_j \cdot x\_alv_{ij} + \sum_{j \in L} \beta_j \cdot x\_txt_{ij} \tag{3.13}$$

em que I, J, K e L são, respectivamente, o conjunto de índices das variáveis originais (coletadas diretamente do site) em conjunto com o bairro e a zona em que o imóvel se encontra, das tags, das variáveis construídas com base nos alvarás dos comércios e das baseadas na descrição.

3.7 Análise de Componentes Principais

É um método que visa explicar a estrutura de variância e covariância de um vetor aleatório, formando novas variáveis, que são combinações lineares das variáveis originais, chamadas de componentes principais (MINGOTI, 2005). Os componentes são não correlacionados, sendo que o primeiro componente captura a maior parte da variação nos dados, o segundo componente captura a segunda maior e assim por diante. Desse modo, pode-se obter uma redução no número de variáveis originais, já que os primeiros componentes podem explicar grande parte da variação, não sendo necessário a utilização dos demais.

A extração dos componentes principais é usualmente feita com base na matriz de correlação, o que remove a influência da escala da variável no resultado obtido. Como produto da aplicação do método, tem-se as cargas (*loadings*), que são os coeficientes da combinação linear para cada um dos componentes, além da porcentagem da variação total que é explicada pelo mesmo.

Esse método foi aplicado no conjunto de variáveis derivadas dos alvarás comerciais, com o objetivo de verificar se hávia a necessidade de utilizar a categorização proposta na tabela 3.

3.8 POS tagging

O ambiente R (R Core Team, 2017) possui diversas ferramentas para auxiliar na análise de texto e, entre elas, está o pacote *udpipe* (WIJFFELS, 2018), que faz interface com o software livre UDpipe (STRAKA; STRAKOVÁ, 2017) que possui uma implementação do *POS tagging* (*Part of speech tagging*), sendo esse um software que lê um conjunto de texto e classifica cada um dos *tokens* na sua respectiva classe gramatical. O pacote já fornece um conjunto de modelos treinados em diferentes linguagens, de forma que o correspondente ao português do Brasil foi utilizado.

O algoritmo foi aplicado aos dados da descritiva e, a partir disso, foi criada um lista em que cada *token* é pareado com a classe gramatical na qual aparece com mais frequência. A aplicação do método visa verificar, dentre as palavras que aparecem no modelo selecionado pela regressão lasso, qual a classe gramatical mais frequente, e quanto a inclusão de palavras dessa classe melhora o ajuste do modelo.

Como adjetivos, substantivos e verbos qualificam o objeto descrito, espera-se que contribuam mais para a predição se comparado a outras classes gramaticais.

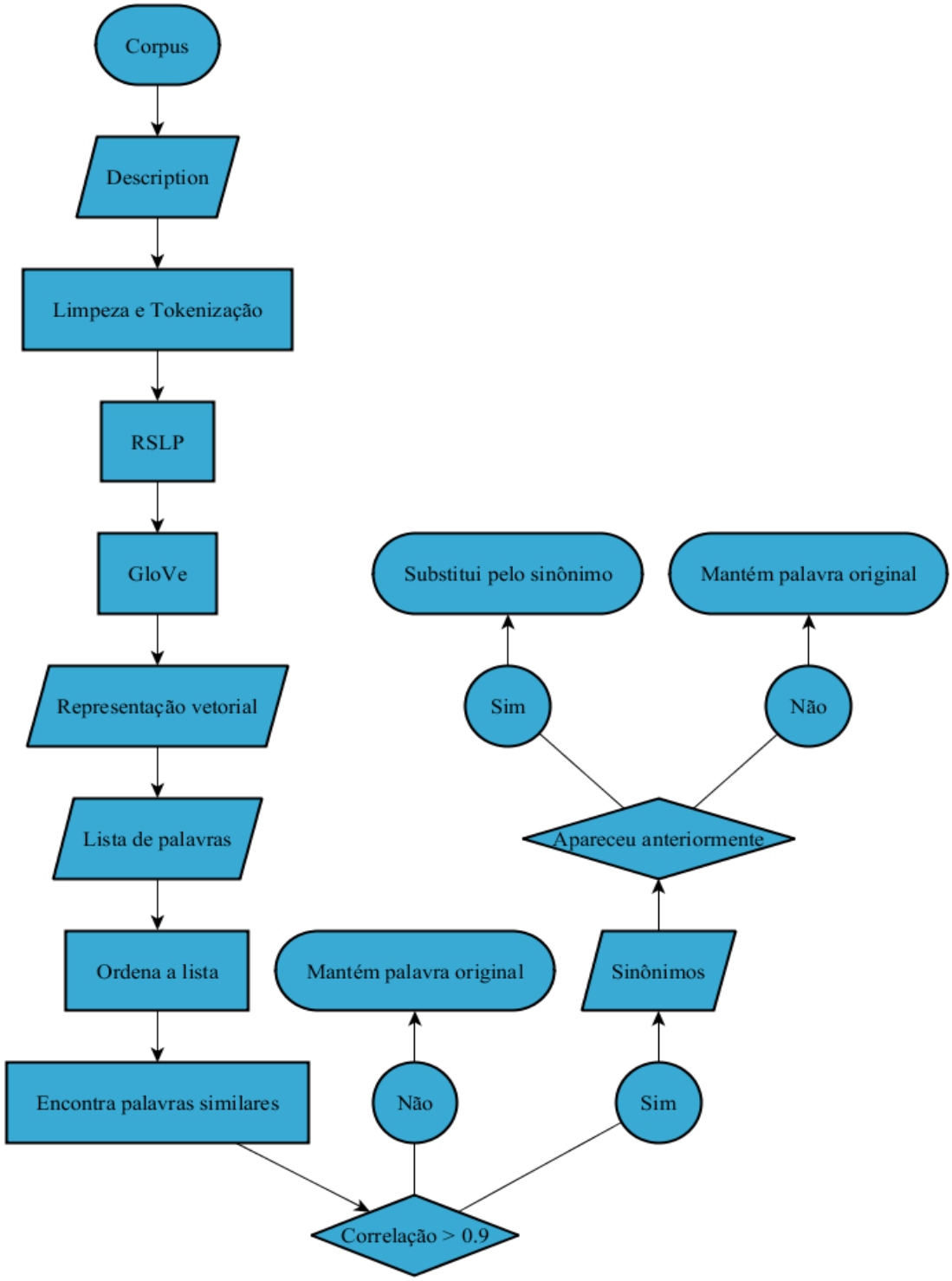


Figura 5 – Fluxograma representando a metodologia de processamento do texto, obtenção de sinônimos e regra de substituição de palavras.

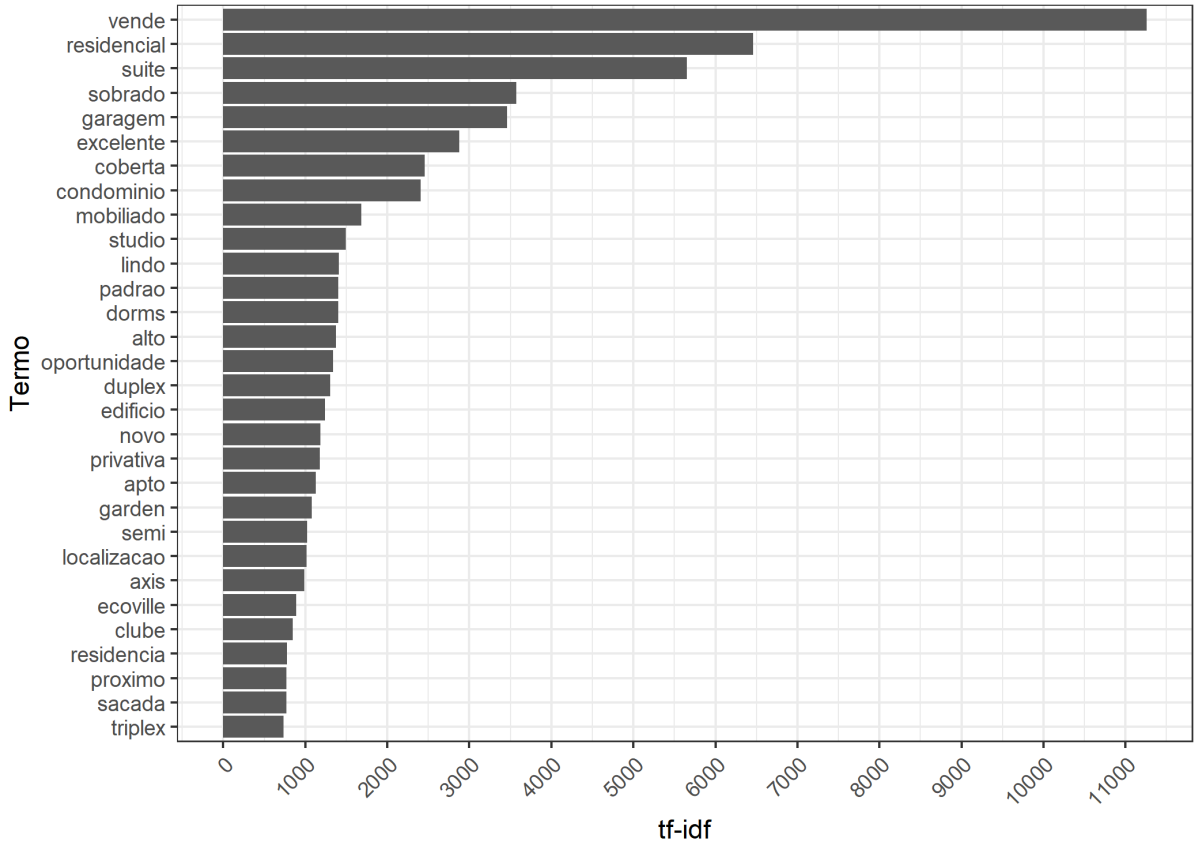


Figura 6 – Representação das 30 palavras do título do imóvel com maior peso segundo a métrica do tf-idf.

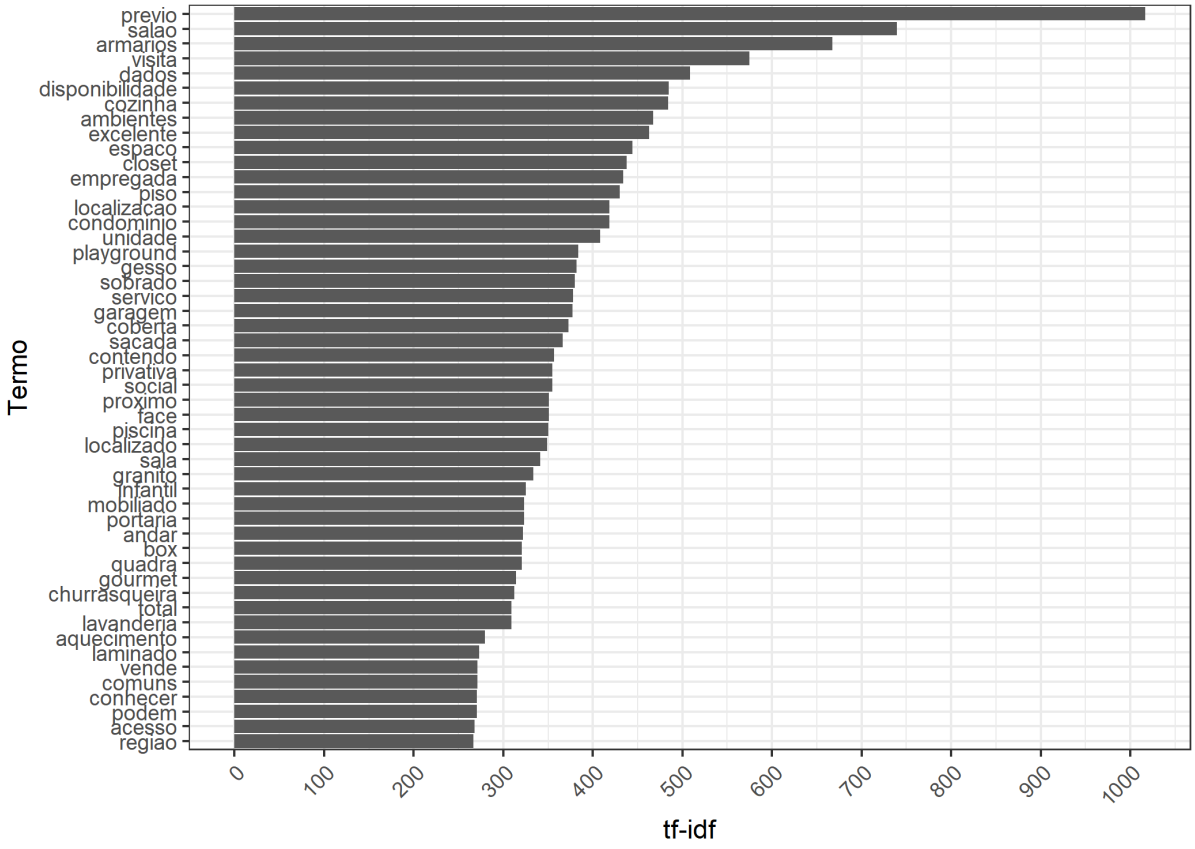


Figura 7 – Representação das 50 palavras da descrição do imóvel com maior peso pela métrica do tf-idf.

Figura 8 – Nuvem de palavras mostrando as tags utilizadas para descrever o imóvel. O tamanho do termo corresponde à quantidade de imóveis marcados com essa característica.

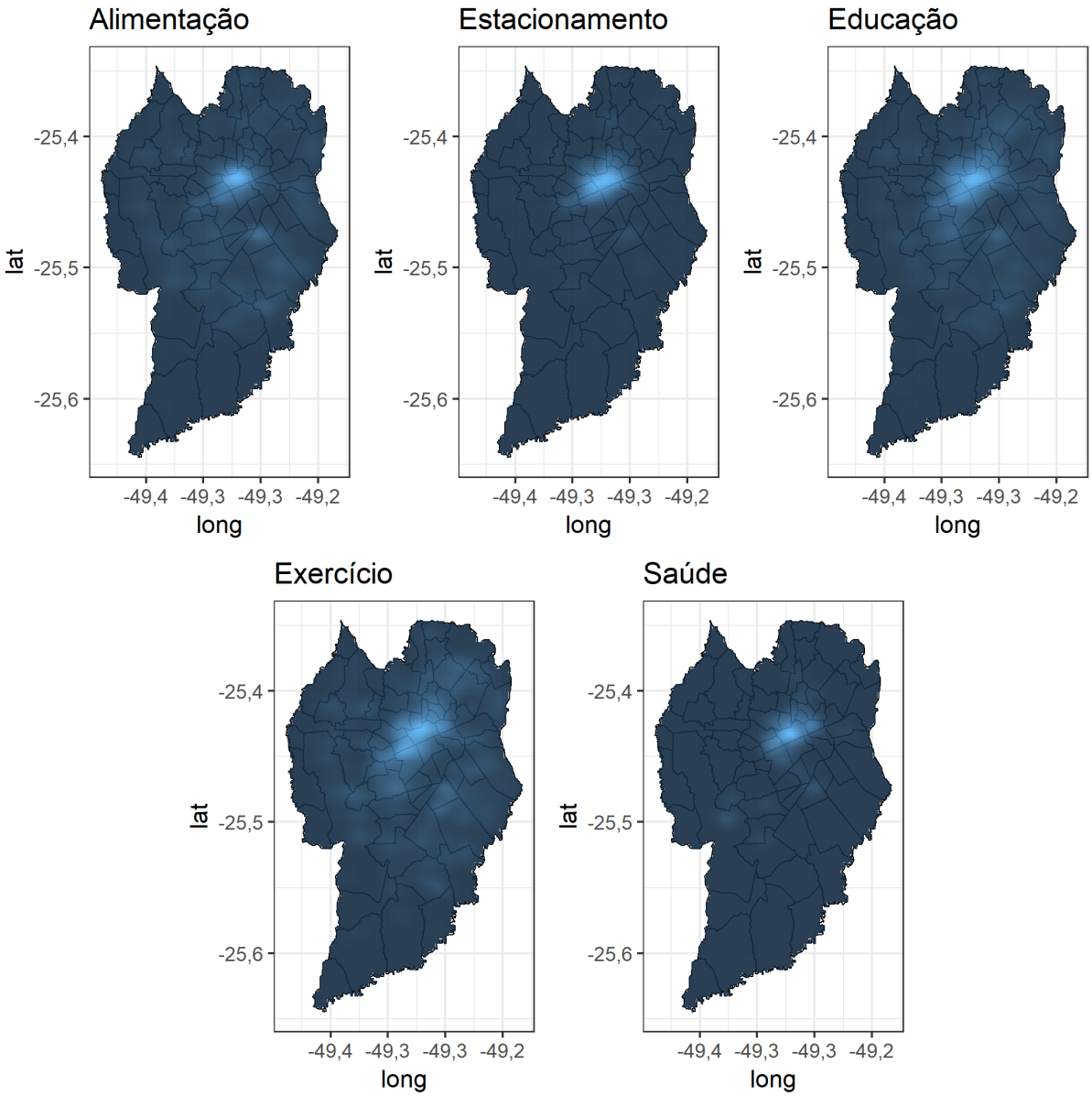


Figura 9 – Índices calculados para uma malha de pontos. Quanto mais clara a cor, maior é a concentração de alvarás encontrados na região.

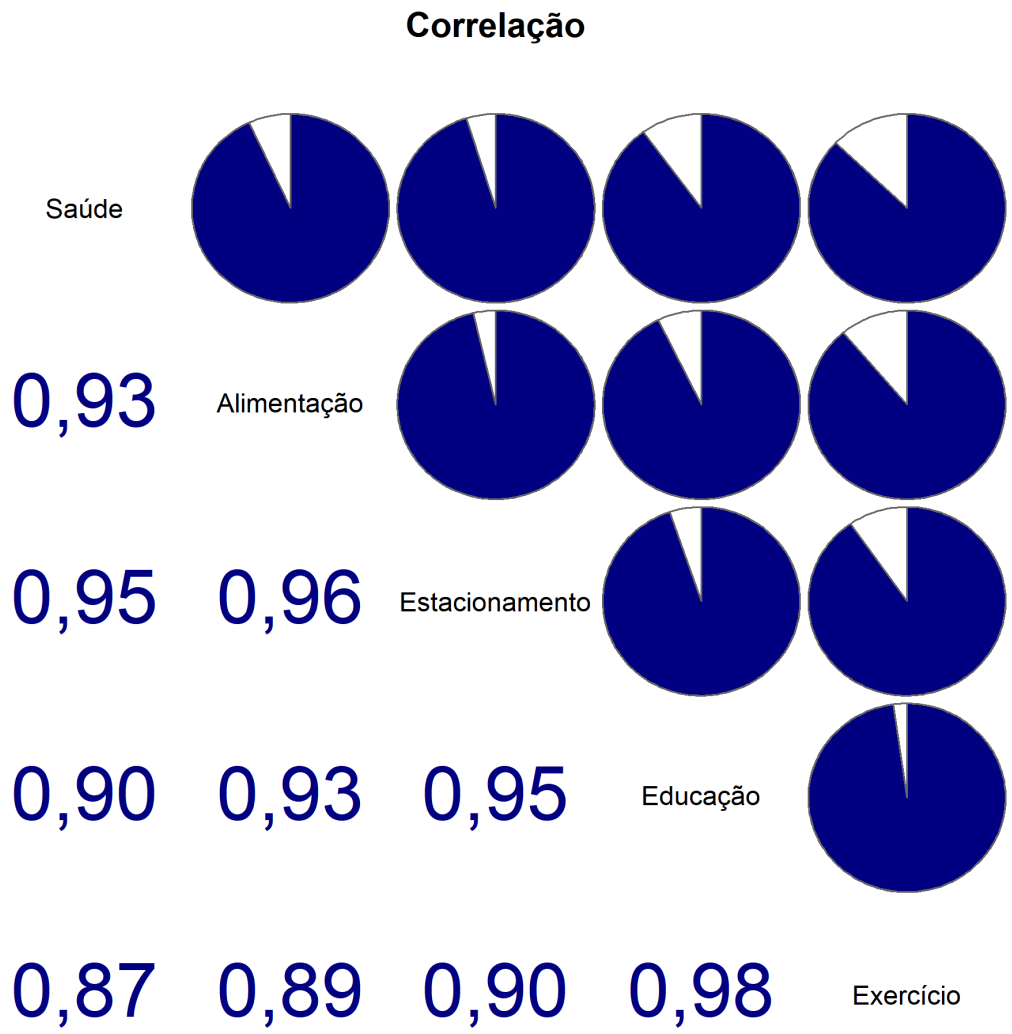


Figura 10 – Correlação entre as variáveis que mensuram a quantidade de alvarás de determinada categoria que estão próximos ao imóvel.

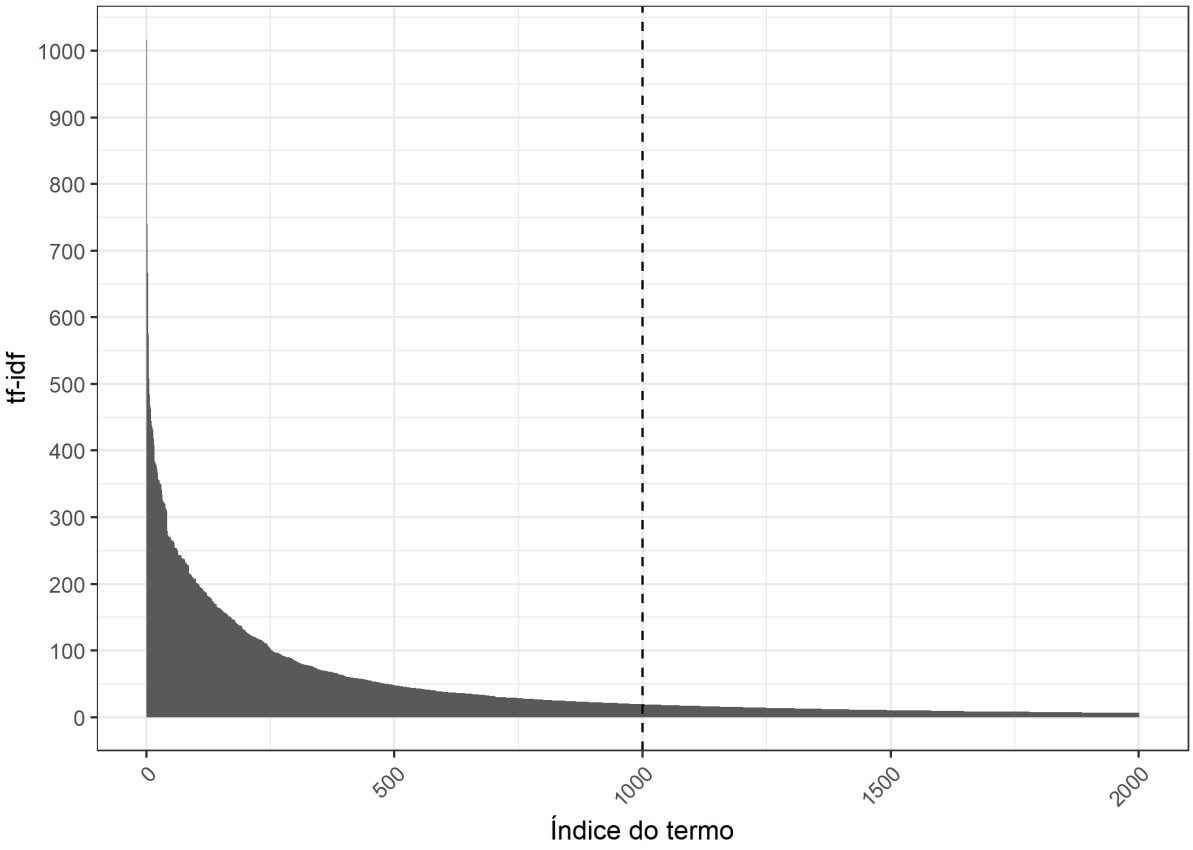


Figura 11 – Termos ordenados pelo valor do somatório do tf-idf. A linha pontilhada representa o corte nos 1000 primeiros termos.

4 Resultados e Discussão

Uma análise de componentes principais foi feita nas variáveis de alvará. Observou-se que o primeiro componente explica 94,11% da variação total e as variáveis que o compõe possuem praticamente o mesmo peso, como mostra a tabela 4. Portanto, o método de kernel foi reaplicado na base de alvarás, mas agora sem a utilização da estratificação por categoria, criando assim a variável *Alvaras_alv*. O resultado aparece na figura 12, em que percebe-se uma mancha mais clara no centro da cidade, similar à vista na figura 9, representando a maior quantidade de estabelecimentos no local.

Tabela 4 – Cargas (*loadings*) para o primeiro componente da análise de componentes principais nas variáveis de alvará.

Variável	Componente 1
Alimentacao_alv	0,45
Estacionamento_alv	0,45
Educacao_alv	0,45
Exercicio_alv	0,44
Saude_alv	0,44

Aplicando a remoção de variáveis correlacionadas pelo método descrito na sessão 3.6.3, 60 preditoras foram removidas, sendo que grande parte delas provém da descrição do imóvel englobando até mesmo nomes próprios. Dentre as remoções há somente uma covariável relacionada às tags (*Posição_do_Sol_tag*). Além disso, as variáveis *area* e *bathroom* também foram removidas pelo algorítimo. A variável *bathroom* tem correlação de 0,802 com a variável *suite*. Esse comportamento está de acordo com o esperado, uma vez que suíte é definida como a integração de quarto e banheiro. Já a variável *bathroom* é correlacionada com o número de quartos (correlação de 0,602), suítes (correlação de 0,802) e vagas na garagem (correlação de 0,61).

A regressão lasso (descrita na sessão 3.6.4) foi então aplicada aos dados. O parâmetro λ que implicou no menor EQM da validação cruzada foi $\lambda = 0,0004$ e o modelo ajustado nessa configuração tem 170 coeficientes nulos, mas 4 dos coeficientes são relacionados com a covariável que indica o bairro e, como a lasso aponta a remoção somente desses níveis (bairro Bacacheri, Campo Comprido, São Miguel e São João), essa variável tem que ser mantida. Muitas das variáveis relacionadas a coeficientes nulos são da descrição textual (84,34%), enquanto o restante delas são das tags[1].

As bases de treino e de validação foram formadas por 42670 e 4742 observações, respectivamente, totalizando nos 47412 imóveis que restaram após todo o processo

[1] A contagem não está incluindo a variável relativa aos bairros.

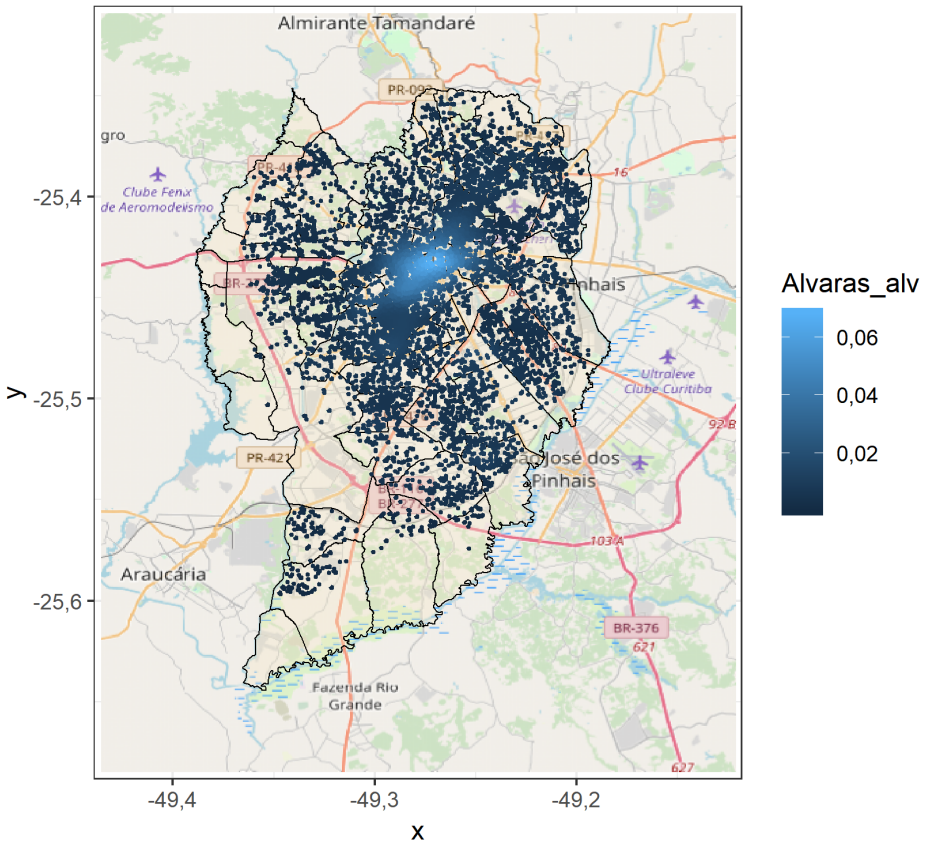


Figura 12 – Gráfico demosntrando o valor variável *Alvaras_alv* para os imóveis da base de dados.

de limpeza (10998 casas e 36414 apartamentos). Desse modo o modelo de regressão foi ajustado considerando as 1096 preditoras restantes, já que estas, por terem sido mantidas no modelo lasso, trazem o menor erro médio de predição.

Além das tags, as variáveis *type*, *neighborhood* e *zone* também entraram no modelo como um conjunto de varáveis indicadoras, de modo que a categoria base da primeira é representada pelos apartamento, da segunda pelo bairro Centro e da última por zonas residenciais. A figura 13 mostra os gráficos de diagnóstico para o modelo com todas as covariáveis. O gráfico (a) revela 3 observações mais afastadas da nuvem de pontos, mas mostra resíduos simétricos com variância aproximadamente constante. Os três pontos foram verificados e removidos, já que o anúncio possuía inconsistências, como a falta de vagas na garagem e o preço bem abaixo ou acima do valor de mercado para as características que possuem. Já o gráfico (b) mostra que a distribuição dos resíduos tem caudas mais pesadas do que a normal. O gráfico (c) apresenta outra maneira de ver que os resíduos tem variância constante. No (d) a distância de Cook é mostrada, indicando 3 pontos discrepantes dos demais, apesar de eles estarem bem abaixo de 1, que é um ponto de corte sugerido, eles foram verificados e não foram encontrados problemas no anúncio.

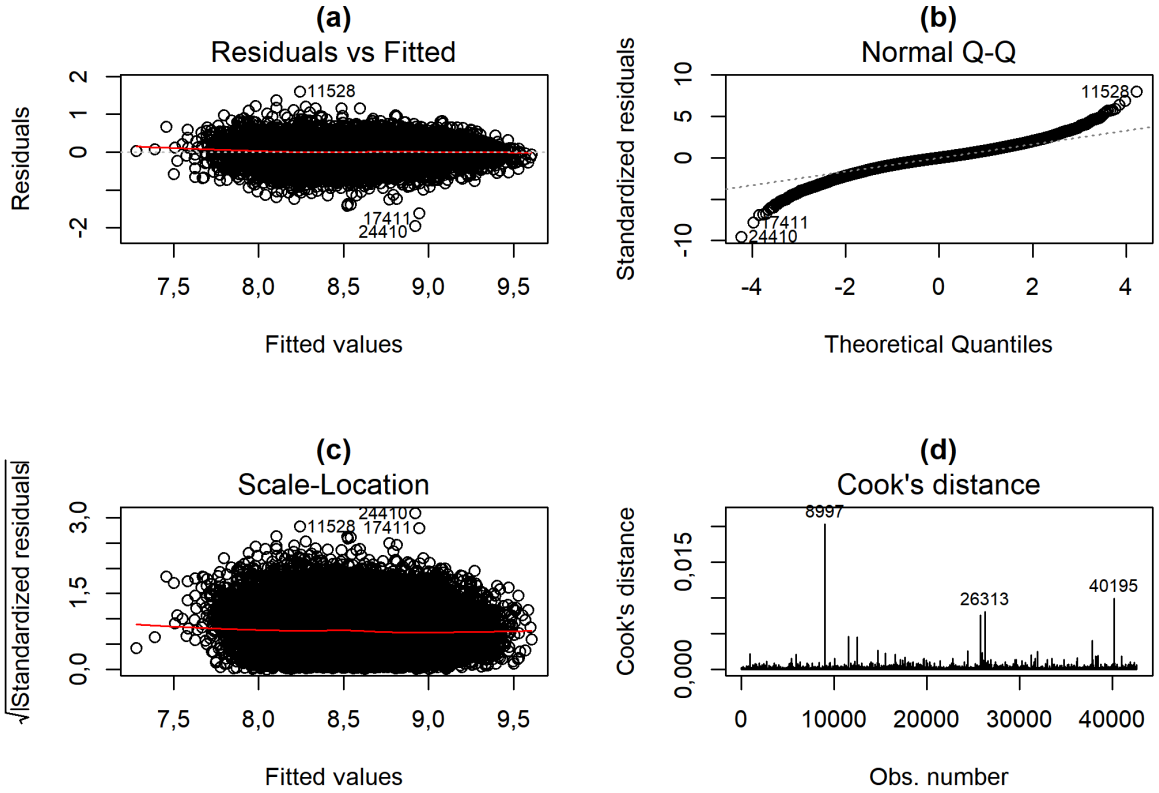


Figura 13 – Gráfico de diagnóstico do modelo completo.

O algoritmo de *POS tagging* foi aplicado ao *corpus*. A partir da tabela de resultados, foi atribuída a classe gramatical para cada uma das palavras que compõem a parte textual do modelo. A tabela 5 mostra a proporção de palavras pertencentes a cada uma das classes gramaticais. Observa-se que verbos (VERB) aparecem em maior proporção, seguido por substantivos (NOUN) e por adjetivos (ADJ), somando 49,84% das variáveis textuais.

Para visualizar o aumento da performance preditiva do modelo causada pela inclusão das covariáveis, um modelo base foi ajustado (modelo 0), contendo somente as variáveis: *type* (tipo do imóvel), *bedroom* (quantidade de quartos), *suite* (quantidade de suítes), *garage* (quantidade de vagas na garagem), que são facilmente mensuradas e obtidas diretamente do site (equação 4.1), e as outras variáveis foram adicionadas em uma sequência de modelos, indo do mais simples ao mais complexo, formando o modelo 1 (equação 4.2), 2 (equação 4.3), 3 (equação 4.4), 4 (equação 4.5) e os modelo 5.1 (equação 4.6), modelo 5.2 (equação 4.7), modelo 5.3 (equação 4.8) e modelo 5.4 (equação 4.9). Nos modelos 5.j, sendo $j = 1, 2, 3, 4$, as variáveis textuais são adicionadas sequencialmente, de acordo com a seguinte ordem: primeiro adiciona-se os adjetivos, que devem definir qualidades do imóvel; depois adiciona-se os substantivos, que podem definir locais próximos; dai adiciona-se os verbos, que definem ações; e por fim

Tabela 5 – Proporção das classes gramaticais das palavras incluídas como covariáveis na parte textual do modelo.

Classe	Proporção
Verbo	0,173
Substantivo	0,171
Adjetivo	0,155
Advérbio	0,091
Adposição	0,064
Auxiliar	0,060
Pronome	0,058
Determinante	0,056
Outras	0,046
Nome próprio	0,045
Numeral	0,043
Partícula	0,024
Conjunção coordenativa	0,014

as palavras pertencentes às outras classes gramaticais foram incluídas.

$$\mu_0 = \beta_0 + \beta_1 \cdot type + \beta_2 \cdot bedroom + \beta_3 \cdot suite + \beta_4 \cdot garage \tag{4.1}$$

$$\mu_1 = \mu_0 + \beta_5 \cdot zone \tag{4.2}$$

$$\mu_2 = \mu_1 + \beta_6 \cdot neighborhood \tag{4.3}$$

$$\mu_3 = \mu_2 + \sum_{j \in J} \beta_j \cdot x\_tag_j \tag{4.4}$$

$$\mu_4 = \mu_3 + \sum_{j \in K} \beta_j \cdot x\_alv_j \tag{4.5}$$

$$\mu_{5.1} = \mu_4 + \sum_{j \in \{\mathrm{A} | \mathrm{A} \subset L\}} \beta_j \cdot x\_txt_j \tag{4.6}$$

$$\mu_{5.2} = \mu_{5.1} + \sum_{j \in \{\mathrm{N} | \mathrm{N} \subset L\}} \beta_j \cdot x\_txt_j \tag{4.7}$$

$$\mu_{5.3} = \mu_{5.2} + \sum_{j \in \{\mathrm{V} | \mathrm{V} \subset L\}} \beta_j \cdot x\_txt_j \tag{4.8}$$

$$\mu_{5.4} = \mu_{5.2} + \sum_{j \in \{O|O \subset L\}} \beta_j \cdot x\_txt_j \tag{4.9}$$

em que A, N, V e O denotam o conjunto de índices para os adjetivos, substantivos, verbos e outras classes, respectivamente.

Tabela 6 – Performance dos modelos na base de validação, considerando a proporção de variação explicada (R^2), a Raiz do Erro Quadrático Médio (REQM) e o Erro Absoluto Médio (EAM) do modelo com a resposta em escala log e em reais. A coluna de melhora refere-se a porcentagem de diminuição do EAM em relação ao modelo da linha anterior. Já a última coluna refere-se a melhora com relação ao modelo μ_0.

Modelo	R^2	REQM	EAM	EAM (R$)	Melhora (%)	Acumulada (%)
μ_0	0,434	0,287	0,228	1245,322		0,00%
μ_1	0,446	0,284	0,225	1228,404	1,29%	1,29%
μ_2	0,561	0,252	0,195	1076,611	15,54%	17,03%
μ_3	0,617	0,236	0,181	1000,375	7,62%	25,95%
μ_4	0,617	0,236	0,181	1000,019	0,06%	26,02%
$\mu_{5.1}$	0,656	0,224	0,170	936,829	6,24%	33,88%
$\mu_{5.2}$	0,670	0,219	0,166	913,338	2,42%	37,12%
$\mu_{5.3}$	0,679	0,216	0,163	894,979	1,87%	39,69%
$\mu_{5.4}$	0,705	0,207	0,156	852,312	4,74%	46,31%

A tabela 6 mostra a proporção da variação total explicada pelo modelo (R^2), a raiz do erro quadrático médio e o erro absoluto médio, tanto na escala da resposta log-transformada quanto em reais, como uma forma de avaliar a performance preditiva do modelo aplicado na base de dados de validação. As ultimas colunas contém a porcentagem de melhora em relação ao EAM do modelo na linha anterior e com relação ao modelo μ_0, respectivamente. Observa-se que a inclusão do bairro causa bastante impacto no erro de predição (15,54%), mas a variável de alvará só reduz o EAM em 0,06%, sendo que esse efeito só é notado se o valor do EAM em reais for verificado (o valor é reduzido em poucos centavos). Nota-se também que o erro de predição parece pequeno se observado na escala logarítmica, mas quando observado na escala da variável *priceSM* o modelo mais complexo leva a um erro absoluto médio de 852,312 reais. Outro ponto a se destacar é a melhora causada pela inclusão das variáveis textuais. Quando os adjetivos foram adicionados ao modelo houve uma redução de 6,24% no EAM, sendo esta a maior redução dentre as covariáveis textuais. A figura 14 mostra de forma gráfica os resultados supracitados.

A figura 15 exibe um semivariograma empírico com envelope[2], calculado com base nos resíduos do modelo $\mu_{5.4}$. Ela mostra que o padrão espacial não foi capturado

[2] Foi utilizadas as funções do pacote *geoR* (Ribeiro Jr; DIGGLE, 2018) para o cálculo.

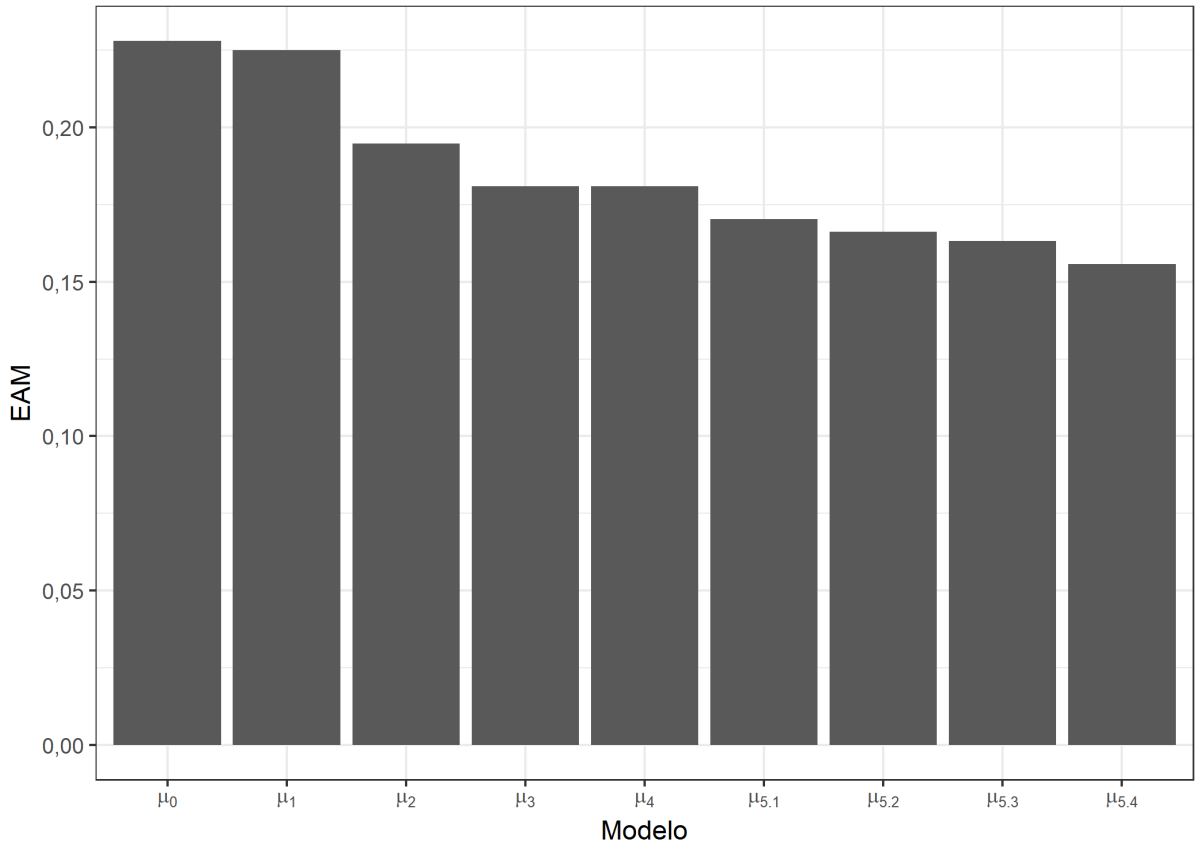


Figura 14 – Erro Absoluto Médio (EAM) para avaliar a performance preditiva dos modelos propostos.

pelo conjunto de covariáveis. Pode ser que a inclusão de outras covariáveis consiga capturar essa dependência, ou alternativamente um modelo espacial poderia ser utilizado.

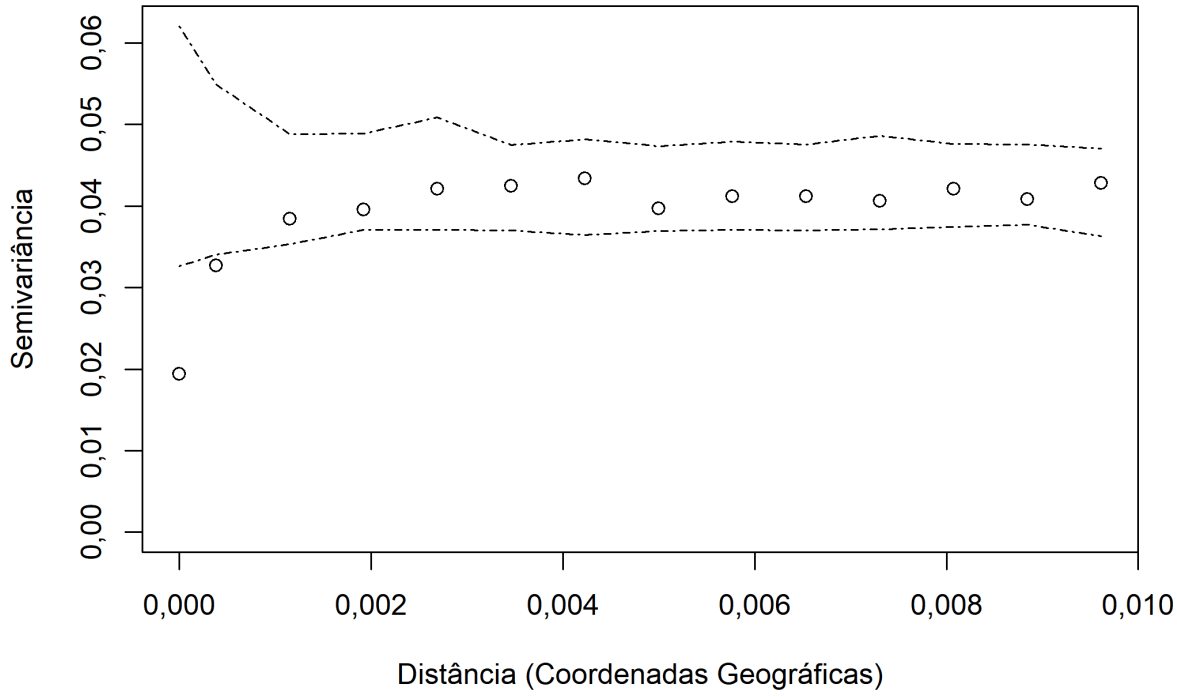


Figura 15 – Variograma dos resíduos do modelo $\mu_{5.4}$.

5 Considerações Finais

Enquanto a coleta dos dados é uma tarefa relativamente simples, uma vez que as ferramentas adequadas são utilizadas, como os pacotes que possibilitaram a implementação do algorítimo de *web scraping*, trabalhar com esses dados já não é trivial, já que a falta de cuidado na digitação do anúncio dos imóveis levam a sérias inconsistências, das quais as mais gritantes foram eliminadas por meio de algumas regras, mas outras não tão evidentes necessitam de algorítimos avançados para correção ou eliminação do dado.

Apesar da dificuldade do acesso automatizado à informação contida na descrição do imóvel em texto, tanto pelos erros de escrita, quanto pela quantidade de palavras que podem ser utilizadas para expressar o mesmo conceito, sem contar o esforço que deve ser empregado na escolha adequada da medida para os dados tornarem-se gerenciáveis do ponto de vista de modelagem, as covariáveis formadas a partir dela mostraram-se promissoras. A utilização dessas variáveis trouxe uma melhorara na performance preditiva do modelo, de modo que o último modelo apresentado, se comparado com o modelo base, reduz em aproximadamente 46,3% o erro de predição[1].

Embora o modelo de regressão normal com transformação logarítmica na variável resposta seja amplamente utilizado na literatura, geralmente não para modelar o preço do metro quadrado mas sim o valor do imóvel como um todo, o mesmo resultou em um ajuste pobre para a presente aplicação, apresentando resíduos com caldas pesadas. Isso impossibilita a inferência uma vez que a distribuição da variável resposta não esta bem especificada. Outro ponto é o erro de predição, que torna o modelo inviável para aplicações práticas. Parte do problema pode residir na qualidade dos dados, que, como já citado, apresentam várias inconsistências.

Três possíveis soluções para melhoria das predições são: a utilização de um modelo geoestatístico para levar em consideração a dependência espacial entre as observações; a utilização de modelos mais flexíveis como os da metodologia de Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS); ou se ainda houver insistência no modelo de regressão linear múltipla, outras formas funcionais podem ser especificadas para o modelo, o que pode melhorar a porcentagem da variação explicada e as predições. Alguns testes foram realizados com essa última opção. Neles foram construídos modelos incluindo a interação entre o tipo de imóvel e as outras covariáveis, o que aumenta a flexibilidade do modelo, não forçando um mesmo efeito das covariáveis para os dois tipos de empreendimento. Naturalmente a porcentagem

[1] Considerando-se o EAM como medida de erro.

de explicação aumenta, já que a quantidade de coeficientes praticamente dobra, mas as conclusões gerais acerca das covariáveis importantes se mantem, além de que o erro de predição do metro quadrado continua maior do que 800 reais.

Como resultado geral, a engenharia de covariáveis feitas no presente estudo pode servir como base para futuros trabalhos. Os resultados aqui apresentados mostram que, se o objetivo for a predição do preço do metro quadrado dos imóveis, modelos diferentes devem ser investigados, como redes neurais artificiais e outros métodos de *machine learning*.

REFERÊNCIAS

ARCGIS. *Shapefiles*. ArcGIS Online. Disponível em: <https://doc.arcgis.com/pt-br/arcgis-online/reference/shapefiles.htm>. Citado na página 36.

ARRAES, R. A.; FILHO, E. d. S. Externalidades e formação de preços no mercado imobiliário urbano brasileiro: um estudo de caso. *Economia Aplicada*, scielo, v. 12, p. 289–319, 2008. ISSN 1413-8050. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-80502008000200006>. Citado na página 19.

BM&FBOVESPA. *Evolução mensal do Índice Imobiliário (IMOB)*. 2018. Disponível em: <http://www.bmfbovespa.com.br/pt_br/produtos/indices/indices-setoriais/indice-bm-fbovespa-imobiliario-imob-estatisticas-historicas.htm>. Citado na página 17.

BRUECKNER, J. K. *Lectures on Urban Economics*. MIT Press, 2011. (The MIT Press). ISBN 9780262300315. Disponível em: <https://books.google.com.br/books?id=-NXxCwAAQBAJ>. Citado na página 20.

C3SL. *Relação de alvarás para liberação de atividades comercias e edificações dentro do município de Curitiba*. 2018. Disponível em: <http://dadosabertos.c3sl.ufpr.br/curitiba/BaseAlvaras/2018-08-01_Alvaras-Base_de_Dados.CSV>. Citado na página 37.

CURITIBA. *Uso do Solo, Lei 9.800 e Leis Complementares da Legislação de Uso do Solo*. 2015. IPPUC. Disponível em: <http://www.ippuc.org.br/planodiretor2014/arquivos/lei{_}9800{_}e{_}complemen>. Citado na página 36.

FALBEL, D. *rslp: A Stemming Algorithm for the Portuguese Language*. [S.l.], 2016. Disponível em: <https://cran.r-project.org/package=rslp>. Citado na página 33.

FELDMAN, R.; SANGER, J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007. ISBN 9780521836579. Disponível em: <https://books.google.com.br/books?id=U3EA{_}zX3Z>. Citado na página 29.

FERREIRA, J. S. W. A cidade para poucos: breve história da propriedade urbana no brasil. In: *Anais do Simpósio "Interfaces das representações urbanas em tempos de globalização"*. UNESP Bauru e SESC Bauru, 2005. Disponível em: <http://www.academia.edu/download/33263826/Cidadeparapoucos_Propriedade_Urbana_Joaosetter.pdf>. Citado na página 17.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, v. 33, n. 1, p. 1–22, 2010. Disponível em: <http://www.jstatsoft.org/v33/i01/>. Citado na página 43.

FáVERO, L. P. L.; BELFIORE, P. P.; LIMA, G. A. S. F. d. Modelos de precificação hedônica de imóveis residenciais na região metropolitana de São Paulo: uma abordagem sob as perspectivas da demanda e da oferta. *Estudos Econômicos São Paulo*, scielo, v. 38, p. 73–96, 2008. ISSN 0101-4161. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-41612008000100004>. Citado 3 vezes nas páginas 17, 19 e 20.

HERMANN, B. M.; HADDAD, E. A. Mercado imobiliário e amenidades urbanas: a view through the window. *Estudos Econômicos São Paulo*, scielo, v. 35, p. 237–269, 2005. ISSN 0101-4161. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-41612005000200001>. Citado 2 vezes nas páginas 17 e 19.

HIEMSTRA, D. A probabilistic justification for using tf-idf term weighting in information retrieval. *International Journal on Digital Libraries*, v. 3, n. 2, p. 131–139, 2000. ISSN 1432-5012. Disponível em: <https://doi.org/10.1007/s007999900025>. Citado na página 31.

HUYCK, C.; ORENGO, V. A Stemming Algorithmm for the Portuguese Language. In: *String Processing and Information Retrieval, International Symposium on(SPIRE)*. [s.n.], 2001. v. 00, p. 186. Disponível em: <doi.ieeecomputersociety.org/10.1109/SPIRE.2001.10024>. Citado 2 vezes nas páginas 29 e 33.

INCRA. *O que é Imóvel Rural nos termos da legislação agrária?* 2010. Portal do Governo do Brasil. Disponível em: <http://www.incra.gov.br/o-que-e-imovel-rural-nos-termos-da-legislacao-agraria>. Citado na página 22.

IPPUC. *Downloads*. 2018. IPPUC. Disponível em: <http://ippuc.org.br/geodownloads/geo.htm>. Citado 3 vezes nas páginas 11, 36 e 79.

JAMES, G.; WITTEN, D.; TIBSHIRANI, R.; HASTIE, T. *An Introduction to Statistical Learning with Applications in R*. [S.l.: s.n.], 2013. 431 p. ISSN 01621459. ISBN 1461471389. Citado 2 vezes nas páginas 42 e 43.

JOHN, E. M. C.; PORSSE, A. A. Análise de preços hedônicos no mercado imobiliário de apartamentos em curitiba. *Revista Paranaense de Desenvolvimento - RPD*, v. 37, n. 130, 2016. ISSN 2236-5567. Disponível em: <http://www.ipardes.pr.gov.br/ojs/index.php/revistaparanaense/article/view/765>. Citado 2 vezes nas páginas 19 e 20.

LANCASTER, K. J. A New Approach to Consumer Theory. *Journal of Political Economy*, v. 74, 1966. Disponível em: <https://econpapers.repec.org/RePEc:ucp:jpolec:v:74:y:1966:p:132>. Citado na página 19.

LOPES, A. *Portuguese stop words*. [S.l.]: GitHubGist, 2013. \url{https://gist.github.com/alopes/5358189}. Citado na página 32.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. (Mit Press). ISBN 9780262133609. Disponível em: <https://books.google.com.br/books?id=YiFDxbEX3SUC>. Citado 3 vezes nas páginas 29, 30 e 31.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Editora UFMG, 2005. ISBN 9788570414519. Disponível em: <https://books.google.com.br/books?id=W7sZllHmmGIC>. Citado 2 vezes nas páginas 17 e 44.

Ministério do Meio Ambiente. *Outros tipos de zoneamento*. 2018. Brasil. Disponível em: <http://www.mma.gov.br/informma/item/8188-outros-tipos-de-zoneamento>. Citado na página 36.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. GloVe: Global Vectors for Word Representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*. [s.n.], 2014. p. 1532–1543. Disponível em: <http://www.aclweb.org/anthology/D14-1162>. Citado na página 32.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: [s.n.], 2017. Disponível em: <https://www.r-project.org/>. Citado 3 vezes nas páginas 21, 33 e 45.

Ribeiro Jr, P. J.; DIGGLE, P. J. *geoR: Analysis of Geostatistical Data*. [S.l.], 2018. R package version 1.7-5.2.1. Disponível em: <https://CRAN.R-project.org/package=geoR>. Citado na página 57.

ROSEN, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, v. 82, n. 1, p. 34–55, 1974. Disponível em: <https://econpapers.repec.org/RePEc:ucp:jpolec:v:82:y:1974:i:1:p:34-55>. Citado 2 vezes nas páginas 19 e 20.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. *Communications of the ACM*, v. 18, n. 11, p. 613–620, 1975. ISSN 00010782. Disponível em: <http://portal.acm.org/citation.cfm?doid=361219.361220>. Citado na página 30.

SELIVANOV, D.; WANG, Q. *text2vec: Modern Text Mining Framework for R*. [S.l.], 2018. Disponível em: <https://cran.r-project.org/package=text2vec>. Citado na página 33.

SILGE, J.; ROBINSON, D. *Text Mining with R: A Tidy Approach*. O'Reilly Media, 2017. ISBN 9781491981627. Disponível em: <https://books.google.com.br/books?id=qNcnDwAAQBAJ>. Citado 2 vezes nas páginas 29 e 31.

STRAKA, M.; STRAKOVÁ, J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 88–99. Disponível em: <http://www.aclweb.org/anthology/K/K17/K17-3009.pdf>. Citado na página 45.

WICKHAM, H.; HESTER, J.; OOMS, J. *xml2: Parse XML*. [S.l.], 2017. Disponível em: <https://cran.r-project.org/package=xml2>. Citado na página 21.

WIERENGA, B. Empirical test of the Lancaster characteristics model. *International Journal of Research in Marketing*, v. 1, n. 4, p. 263–293, 1984. ISSN 0167-8116. Disponível em: <http://www.sciencedirect.com/science/article/pii/0167811684900168>. Citado na página 19.

WIJFFELS, J. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. [S.l.], 2018. Disponível em: <https://cran.r-project.org/package=udpipe>. Citado na página 45.

Apêndices

APÊNDICE A – *Stop words*

Tabela 7 – Lista de *stop words* obtida da internet.

1	2	3	4	5
de	pelo	este	estejam	fossemos
a	pela	fosse	estivesse	fossem
o	ate	dele	estivessemos	for
que	isso	tu	estivessem	formos
e	ela	te	estiver	forem
do	entre	voces	estivermos	serei
da	era	vos	estiverem	seremos
em	depois	lhes	hei	serao
um	sem	meus	havemos	seria
para	mesmo	minhas	hao	seriamos
com	aos	teu	houve	seriam
nao	ter	tua	houvemos	temos
uma	seus	teus	houveram	tinhamos
os	quem	tuas	houvera	tinham
no	nas	nosso	houveramos	tive
se	me	nossa	haja	teve
na	esse	nossos	hajamos	tivemos
por	eles	nossas	hajam	tiveram
mais	estao	dela	houvesse	tivera
as	voce	delas	houvessemos	tiveramos
dos	tinha	estes	houvessem	tenha
como	foram	estas	houver	tenhamos
mas	essa	aquele	houvermos	tenham
foi	num	aquela	houverem	tivesse
ao	nem	aqueles	houverei	tivessemos
ele	suas	aquelas	houveremos	tivessem
das	meu	isto	houverao	tiver
tem	minha	aquilo	houveria	tivermos
seu	numa	estou	houveriamos	tiverem
sua	pelos	estamos	houveriam	terei
ou	elas	estive	sou	tera
ser	havia	esteve	somos	teremos

quando	seja	estivemos	sao	terao
muito	qual	estiveram	eramos	teria
ha	sera	estava	eram	teriamos
nos	tenho	estavamos	fui	teriam
ja	lhe	estavam	fomos	
esta	deles	estivera	fora	
eu	essas	estiveramos	foramos	de
tambem	esses	esteja	sejamos	a
so	pelas	estejamos	sejam	o

Tabela 8 – Palavras específicas adicionadas à lista de Lista de *stop words*.

1	2	3	4	5
rua	whatsapp	bem	apartamento	ligue
bairro	todo	area	dois	pra
curitiba	ver	imovel	estar	valor
mts	bwc	imoveis	possui	
av	bom	casa	sendo	

APÊNDICE B – *Proporção de Tags*

Tabela 9 – Frequência relativa de aparecimento de cada uma das tags coletadas do site Imovelweb.

Termo	n	Frequência relativa
Divisórias	1	0,00002
Estacionamento	1	0,00002
Lavoura	1	0,00002
Energia elétrica	2	0,00005
Pasto	2	0,00005
Celeiro	3	0,00007
Refeitório	3	0,00007
Rio	3	0,00007
Cachoeira	5	0,00012
Horta	9	0,00022
Marina	9	0,00022
Casa sede	15	0,00036
Heliponto	16	0,00039
Muro de escalada	25	0,00060
Salão de convenções	29	0,00070
Central de limpeza e governança	30	0,00072
Central telefônica	33	0,00080
Frente para o mar	35	0,00084
Lago	35	0,00084
Poço artesiano	36	0,00087
Casa de caseiro	42	0,00101
Car Wash	43	0,00104
Campo de golfe	46	0,00111
Lazer no Pilotis	63	0,00152
Posição do Apto (Meio)	64	0,00154
Cobertura Coletiva	66	0,00159
No pool de locação	96	0,00232
Depósito	99	0,00239
Pista de Skate	103	0,00249
Geminada	122	0,00294
Quadra de squash	126	0,00304

Carpete	130	0,00314
Bar na piscina	154	0,00372
Canil	155	0,00374
Restaurante	182	0,00439
Pista de cooper	195	0,00470
Cerca	196	0,00473
Redario	196	0,00473
Posição do Apto	199	0,00480
Quadra de tênis	200	0,00483
Deck	207	0,00499
Posição do Sol	207	0,00499
Pomar	212	0,00511
Vestiário para diaristas	223	0,00538
Freezer	229	0,00553
Piso elevado	235	0,00567
Árvores frutíferas	249	0,00601
Aquecimento Solar	286	0,00690
Casa de Boneca	289	0,00697
Guarita	293	0,00707
Centro de estética	309	0,00746
Vestiário	313	0,00755
Posição do Apto (Fundos)	317	0,00765
Próximo ao Metro	327	0,00789
Office Space	362	0,00873
Antena parabólica	367	0,00885
Edícula	373	0,00900
Bar	395	0,00953
Microondas	420	0,01013
Quadra de futebol de salão	431	0,01040
Sala de Massagem	435	0,01050
Mezanino	447	0,01078
Geladeira	489	0,01180
Pet Care	537	0,01296
Carpete de madeira	544	0,01313
Posição do Apto (Lateral)	552	0,01332
Pet Play	582	0,01404
Biblioteca	616	0,01486
Fitness ao ar livre	618	0,01491
Gerador	678	0,01636

Campo de futebol	679	0,01638
Fogão	680	0,01641
Depósito no subsolo	683	0,01648
Posição do Sol (Poente)	689	0,01662
Mini quadra	696	0,01679
Entrada lateral	704	0,01699
Porte Cochère	786	0,01896
Deck molhado na piscina	825	0,01990
Adega	842	0,02032
Varanda Fechada com vidro	940	0,02268
Pé direito elevado	965	0,02328
Piso de madeira	1056	0,02548
Posição do Sol (Perpendicular)	1104	0,02664
Arvorismo	1114	0,02688
Forno de pizza	1158	0,02794
Estuda permuta	1230	0,02968
Ronda/Vigilancia	1294	0,03122
Posição do Sol (Nascente)	1435	0,03462
Ocupado	1496	0,03609
Cinema	1558	0,03759
SPA	1565	0,03776
Automação predial	1600	0,03860
Entrada de serviço	1632	0,03938
Luminárias	1708	0,04121
De esquina	1713	0,04133
Posição do Apto (Frente)	1720	0,04150
Estacionamento para visitantes	1914	0,04618
Espaço zen	2007	0,04842
Praça	2068	0,04990
Cozinha Gourmet	2177	0,05252
Solarium	2704	0,06524
Circuito de TV	2730	0,06587
Escritório	3078	0,07426
Esgoto	3114	0,07513
Permite animais	3156	0,07615
Sauna	3272	0,07894
Piscina infantil	3277	0,07906
Aquecedor	3286	0,07928
Piscina coberta	3335	0,08046

Area verde	3541	0,08543
WC para empregados	3598	0,08681
Quintal	3657	0,08823
Gas encanado	3669	0,08852
Lareira	3671	0,08857
Mobiliado	3727	0,08992
Jardim	3740	0,09024
Piscina aquecida	3778	0,09115
Aquecimento central	3779	0,09118
Despensa	3951	0,09533
Closet	4148	0,10008
Dependência de empregados	4166	0,10051
Bicicletário	4211	0,10160
Hidromassagem	4272	0,10307
Armário embutido	4905	0,11834
Cozinha americana	4907	0,11839
Sistema de alarme	5475	0,13210
Varanda Gourmet	5789	0,13967
Aceita FGTS	5989	0,14450
Ar condicionado	6047	0,14590
Piso laminado	6116	0,14756
Câmeras de segurança	6470	0,15610
Copa	6713	0,16197
Quadra poliesportiva	7184	0,17333
Brinquedoteca	7479	0,18045
Interfone	7648	0,18452
Piso frio	7837	0,18908
Área de Lazer	9063	0,21866
Lavanderia	9395	0,22668
Acesso para deficientes	9498	0,22916
Fitness/Sala de Ginástica	9993	0,24110
Espaço Gourmet	10048	0,24243
Varanda	10049	0,24245
Aceita Financiamento	10291	0,24829
Armário de cozinha	10562	0,25483
Salão de Jogos	11461	0,27652
Piscina	11548	0,27862
Playground	12036	0,29039
Portaria 24 horas	12818	0,30926

Portão Eletrônico	12883	0,31083
Acesso asfaltado	17359	0,41882
Elevador	19058	0,45982
Suites	19207	0,46341
Salão de festas	19313	0,46597
Área de serviço	22876	0,55193
Churrasqueira	28363	0,68432

Anexos

ANEXO A – *Zoneamento urbano de Curitiba*

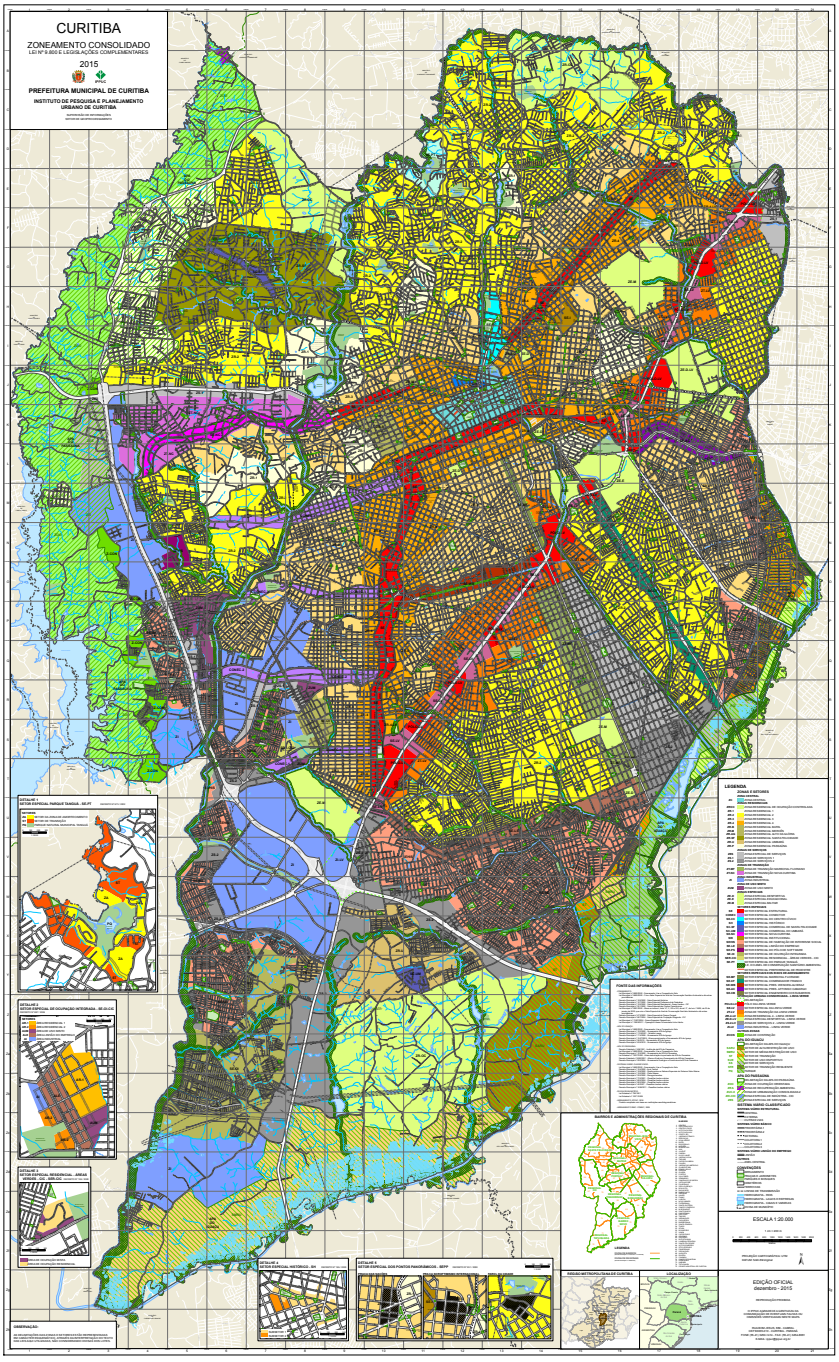


Figura 16 – Mapa do zoneamento urbano de Curitiba disponibilizado pelo IPPUC (2018).