Construção de Banco de Dados do Mercado Imobiliário. Um estudo na Cidade de São Paulo

Celso G. A. Ribeiro[1], Flavio F. Helena[1], Flavio A. M. Cipparrone[1]

[1]Universidade de Sao Paulo, Departamento de Sistemas Eletrônicos. Av. Prof. Luciano Gualberto, trav. 3, n 158, CEP 05508-900, São Paulo/SP, Brasil

celso.ribeiro@usp.br, flavio.helena@usp.br, prof.cipparrone@gmail.com

Abstract. *Obtaining and organizing reliable data plays a key role in understanding and analyzing the real estate market. In this study, a specific methodology for obtaining and processing data applied in the city of São Paulo is proposed, with the aim of overcoming existing data limitations and providing comprehensive and scalable information for an in-depth analysis of this constantly evolving sector. The availability of this complete and reliable database provides valuable insights for sellers and buyers, facilitating informed decision-making and enriching understanding of the real estate market through benchmarking and other relevant analysis.*

Resumo. *A obtenção e organização de dados confiáveis desempenham um papel fundamental na compreensão e análise do mercado imobiliário. Neste estudo, propõe-se uma metodologia específica de obtenção e tratamento de dados aplicada na cidade de São Paulo, com o objetivo de superar as limitações de dados existentes e fornecer informações abrangentes e escaláveis para uma análise aprofundada desse setor em constante evolução. A disponibilidade dessa base de dados completa e confiável proporciona insights valiosos para vendedores e compradores, facilitando a tomada de decisões embasadas e enriquecendo a compreensão do mercado imobiliário por meio de análises comparativas de preços e outras análises relevantes.*

1. Introdução e Revisão Bibliográfica

A obtenção de dados confiáveis no mercado imobiliário é um desafio complexo e relevante. A escassez e precariedade das informações registradas nas prefeituras municipais, bem como as dificuldades de acesso aos Cartórios de Registros de Imóveis ou Receita Federal, contribuem para a falta de transparência e confiança nos dados desse setor (Abreu & Amorim, 2014).

Em alguns casos, os trabalhos de precificação enfrentam dificuldades na coleta de dados, recorrendo a pesquisas de campo e formulários manuais (Abreu & Amorim, 2014), o que limita o escopo da pesquisa. Outras vezes, usam bases virtuais, mas com restrições de tamanho e variáveis (Pinto & Fernandes, 2019; Paz & Nobre, 2020), diminuindo a assertividade. Segundo Mullainathan & Spiess (2017), overfit em análise de precificação é maior com amostras menores, reforçando a importância de bases de dados com número suficiente de amostras.

Diante dessas limitações, é crucial desenvolver uma abordagem sistemática para obter informações imobiliárias confiáveis. O objetivo deste estudo é estabelecer uma

metodologia que viabilize a coleta de dados atualizados e abrangentes por meio de listagens imobiliárias, de forma coesa. Essa abordagem não se limita a dados intrínsecos, como características dos imóveis, mas inclui também dados extrínsecos, como elementos do ambiente em que o imóvel se encontra.

Para construir uma base de dados sólida, é necessário integrar diversas fontes de informação, selecionadas por aspectos relevantes. Essa metodologia envolve mineração, filtragem e modelagem de dados, garantindo informações confiáveis, essenciais para análises do mercado imobiliário. Nesse sentido, São Paulo destaca-se como exemplo ideal para consolidar o estudo proposto, usando anúncios online, informações sobre amenidades públicas, registros de imóveis e outras fontes na construção de uma base de dados abrangente.

Uma base de dados completa e confiável do mercado imobiliário fornece insights valiosos, facilitando decisões embasadas e permitindo análises comparativas das características dos imóveis e dos padrões de preço. Um exemplo é o estudo de caso de D'acci (2019), que investigou o impacto das variáveis extrínsecas no preço utilizando uma base de dados abrangente em Torino, ilustrando a importância do uso de base de dados confiável na compreensão do mercado imobiliário.

No contexto da pesquisa sobre obtenção de dados confiáveis no mercado imobiliário, a metodologia se destaca em relação aos estudos previamente explorados. Enquanto muitos lidam com restrições na coleta de dados, esta abordagem abrange várias fontes, integrando dados intrínsecos e extrínsecos por meio de listagens imobiliárias e informações públicas. Especificamente no cenário brasileiro, onde esta metodologia ainda não é amplamente usada, este estudo busca preencher essa lacuna, introduzindo uma perspectiva original e eficaz para servir de subsídio a análises do setor imobiliário.

2. Metodologia

Como discutido anteriormente, a existência de um banco de dados contendo as informações de características intrínsecas e extrínsecas de imóveis é um passo primordial para a construção de modelos que possam ajudar a analisar e explicar o comportamento do mercado imobiliário, com diversas aplicações para tais modelagens. O Brasil utiliza um sistema de registro de transações por meio Cartorial, onde as informações são pertencentes ao poder público. (Tierno et al., 2007). Dessa forma, não há uma base de dados oficial disponibilizada de forma sistemática e com abrangência nacional que contenha os preços reais negociados pelos imóveis.

Dessa forma propõe-se a metodologia exemplificada a seguir, utilizando uma forma engenhosa de obter essas informações, por meio de anúncios imobiliários online. Nesse sentido, a cidade de São Paulo possui uma vasta base de dados com informações e dados sobre o "ambiente construído" (qualidade do pavimento, calçadas, arborização), sendo, portanto, o local escolhido para o desenvolvimento deste trabalho. Nesse sentido, seis distritos da cidade foram escolhidos, garantindo tanto a variabilidade da realidade socioeconômica quanto a representatividade estatística da amostra. Os distritos escolhidos foram: Moema, Pinheiros, Vila Mariana, Vila Carrão, Vila Andrade e Jaguaré. Cabe destacar que, neste estudo a Cidade de São Paulo foi escolhida, no entanto tal metodologia é replicável para outras localidades, na medida da disponibilidade de bases de dados com informações similares nestes locais.

Como próximo passo, a escolha das variáveis e fontes de dados também foi realizada. Aqui, objetivando uma representatividade do contexto em que o imóvel está inserido, foram selecionadas variáveis comumente associadas ao preço e à qualidade de um imóvel, tanto intrínsecas quanto extrínsecas. As variáveis foram provenientes de bases de dados distintas, onde diversas premissas para suas uniões foram adotadas. A Figura 1 ilustra essas bases e como foram consolidadas no banco de dados final.

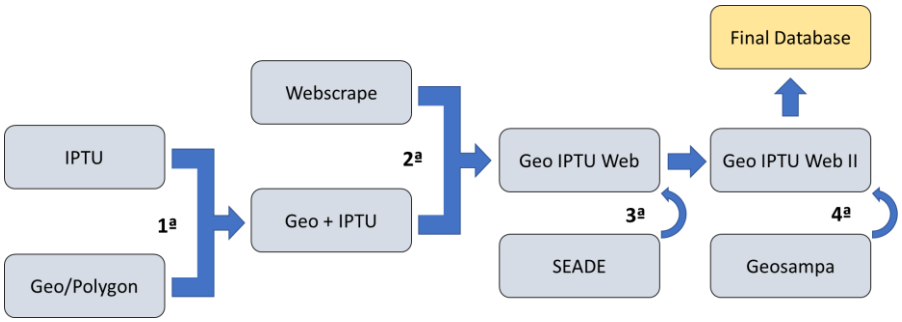

Figura 1. Processo de Junção das bases de dados

A Primeira Etapa da junção inicia-se com a "Base de dados IPTU", que reúne diversas informações das propriedades imobiliárias da cidade (ano de construção, número de pavimentos, área construída, etc.), disponibilizadas pela prefeitura na plataforma *GeoSampa*[1]. À esta base são adicionadas as geolocalizações dos polígonos representativos dos imóveis (figuras geolocalizadas dos terrenos), através da junção com a "Base de dados Geo/Polígono". O critério de junção aqui adotado foi o uso do "número de contribuinte", campo que está presente nas duas bases de dados e pode ser usado, com restrições, como chave indexadora. Há uma particularidade a destacar: como já referido, existe um número de contribuinte relativo ao terreno e outro relativo ao imóvel. Quando um terreno contém apenas uma propriedade, ambos os números são iguais, mas se o terreno contém mais de um imóvel (edifícios, por exemplo), a situação muda. Consequentemente, para unir as duas bases de dados, o número do contribuinte é modificado na "Base de dados IPTU", removendo o código da unidade toda vez que for identificado que o registro é um apartamento, e substituindo pelo código representativo do terreno. Por fim, por meio de um *left join* utilizando esta nova chave, as informações dos polígonos são trazidas para o banco de dados do IPTU, garantindo que cada registro do IPTU tenha um polígono associado a ele.

A Segunda Etapa da junção envolve agregar à esta base de dados gerada anteriormente as informações de anúncios de imóveis online do site Viva Real[2]. Foi desenvolvido um código do tipo *webscrape* em linguagem *Python* para obter todos os anúncios das 100 primeiras páginas do site, repetindo o processo para cada um dos seis bairros escolhidos, reunindo todas as informações relevantes de cada anúncio (área, preço, número de quartos, etc.). Neste código foi adotado um filtro que seleciona apenas imóveis com endereços completos (Nome do Logradouro e Número do imóvel) para registro no banco de dados. Este filtro foi aplicado pois, utilizando este endereço completo e também a biblioteca *googlemaps* no *Python*, é possível obter o CEP do

[1] O GeoSampa é o portal cartográfico oficial da Cidade de São Paulo e reflete a infraestrutura municipal em dados geográficos. A plataforma traz mais de 240 tipos de informações, como fotos aéreas, dados de equipamentos públicos, rede de transporte, etc. É a maior coleção de dados geoespaciais da cidade de SP.
[2] Viva Real é o maior portal online de anúncios imobiliários do Brasil. <https://www.vivareal.com.br/>.

imóvel, o código do logradouro (CodLog), bem como sua Latitude e Longitude. Assim, com o CEP, CodLog e o Número do imóvel, usados como chave indexadora, pode-se fazer a composição dos dois bancos de dados obtidos até o momento.

Dessa forma, o banco de dados contendo todas as características intrínsecas do imóvel está pronto. Esse novo banco de dados será chamado de "Base de dados de Geo IPTU Web", passando para a Terceira Etapa do fluxo: agregar as informações sobre características extrínsecas disponíveis na "Base de dados SEADE". Esta base de dados proveniente da SEADE[3] reúne dados e informações referentes aos equipamentos de uso comum disponibilizados à população (pontos de ônibus, estações de metrô, escolas, faculdades, etc.), mais especificamente sua localização na cidade, com latitude e longitude. O processo de agregação dessas informações consiste em contar quantos pontos de cada variável existem nas proximidades de cada imóvel, considerando uma distância de até 500 metros do centróide do polígono do imóvel. Por exemplo, contar quantas Universidades Públicas estão a 500 metros de cada imóvel. Neste processo foi adotada a distância de *haversine*[4].

Essa mesma lógica também foi adotada para incluir informações sobre características extrínsecas da "Base de dados GeoSampa" (Quarta Etapa), listando quantos itens de suas variáveis estão nas proximidades dos imóveis. Existe apenas uma particularidade nesta base de dados que consiste no fato de que determinadas variáveis, como as relacionadas à Infraestrutura Cicloviária ou à Infraestrutura de Ônibus, são linhas geolocalizadas, contendo a latitude e longitude de seu contorno. Isso leva ao cálculo de quantas linhas existem nas proximidades de um determinado imóvel, em vez do cálculo de pontos. Além disso, uma vez que a geometria da linha está disponível na "Base de dados GeoSampa", seu comprimento total foi calculado, e essas informações foram agregadas ao cruzar os dois bancos de dados, finalizando a junção.

3. Discussão e Resultados

Após realizar todas as etapas descritas no Capítulo anterior, para junção de todas as informações selecionadas, o banco de dados resultante está completo. Contém 43 variáveis que representam as características de cada imóvel. São elas: *Preço/m², Área Construída, Quartos por Banheiros, Número de Garagens, Número de Esquinas, Fração Ideal, Área do Terreno, Área útil, Ano de Construção, Fator de obsolescência, Número de andares, Testada do terreno, Número de UBS, Número de CREAS, Número de CRAS, Número de Escolas Privadas, Número de Escolas Públicas Estaduais, Número de Escolas Públicas Municipais, Número de Escolas Públicas Federais, Número de Escolas (outras), Número de FATECs, Número de Universidades Particulares, Número de Universidades Públicas, Número de Museus, Número de unidades do Poupatempo, Número de Centros Populares, Número de Hospitais Públicos, Número de Hospitais Privados, Número Escalado de Consultórios Médicos, Número Escalado de Clínicas Médicas, Número Escalado de Reparos no Pavimento, Número Escalado de Árvores, Número de imóveis com acessibilidade nas proximidades, Número de Bicicletários Públicos, Número de Estações de metrô,*

[3] SEADE (Fundação Sistema Estadual de Análise de Dados Estatísticos) vinculada ao Governo do Estado de SP, é referência nacional na produção e divulgação de análises e estatísticas socioeconômicas.

[4] A fórmula de Haversine determina a distância do superficial circular entre dois pontos em uma esfera, dadas suas longitudes e latitudes.

Número de pontos de ônibus, Número ponderado de Ciclovias, Número ponderado de Ciclofaixas, Número ponderado de Ciclorrotas, Número ponderado de Faixas de ônibus Locais, Número ponderado de Faixas de ônibus Coletoras, Número ponderado de Faixas de ônibus Arteriais, Número ponderado de BRT.

A Base de dados resultante contém 4.019 registros, cada um correspondendo a um único Imóvel e reunindo todas estas informações descritas. Vale ressaltar que o processo de *webscraping* do site VivaReal reuniu 5.479 registros na base de dados inicial (Base de dados Webscrape) e, após todas as regras, validações e limpeza dos dados para junção dessas bases totalmente distintas, restaram 4.019 registros, resultando em um índice de assertividade para o processo de junção de cerca de 73%.

3.1. Breve Análise Exploratória de Dados.

A partir do banco de dados resultante, abre-se um leque para realização de diversas análises ligadas ao mercado imobiliário que não são o escopo deste trabalho. No entanto, apenas para ilustrar parte dessas aplicações, realizou-se aqui uma breve análise exploratória, ilustrada com a Figura 2, que é uma representação visual desses 4.019 imóveis, geolocalizados no mapa da cidade de São Paulo. As cores do mapa representam a variação do preço por metro quadrado em todos os imóveis, em uma escala verde-vermelha, onde o verde representa os preços mais baixos e o vermelho os mais altos

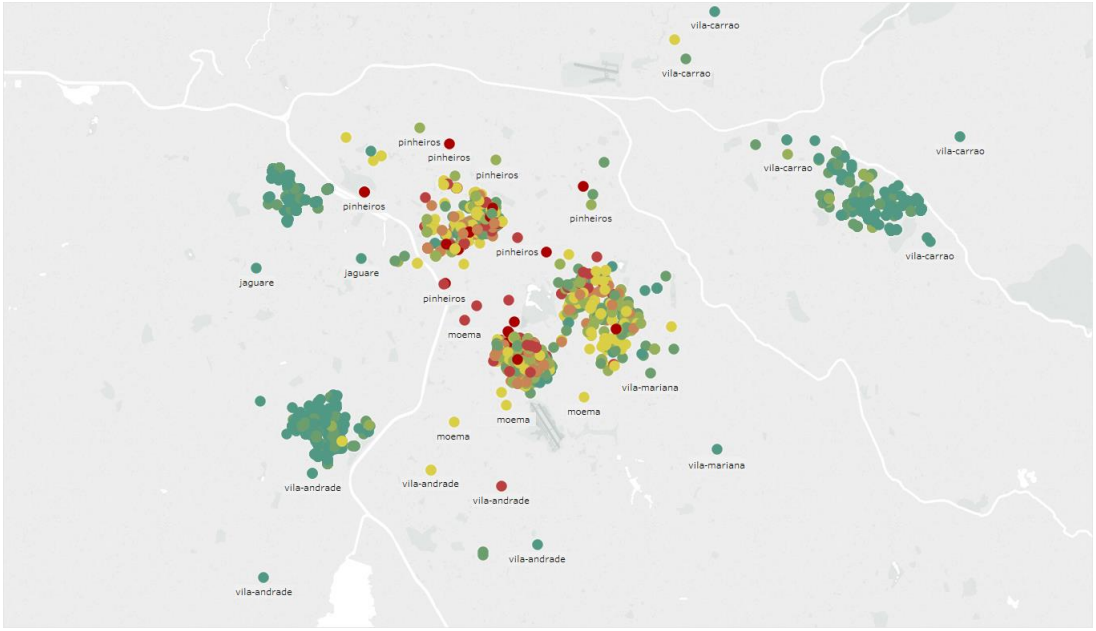

Figura 2. Propriedades Imobiliárias do banco de dados, geolocalizadas no mapa da cidade de São Paulo. A cor representa a variação do preço por metro quadrado (vermelho é alto, verde é baixo).

Nota-se, ao observar a Figura 2, que os bairros mais centralizados da cidade possuem preços/metro quadrado mais elevados, enquanto os bairros mais distantes possuem preços mais acessíveis. Essa tendência de variação de preços em função da distância do centro da cidade já foi abordada por D'acci (2019), citado anteriormente neste trabalho. Na literatura, é possível encontrar uma grande quantidade de trabalhos mostrando os efeitos de uma gama de características extrínsecas na avaliação de

imóveis. Cervero & Kang (2011), por exemplo, mostraram que o transporte público, especificamente Bus Rapid Transit (BRT), oferece prêmios de até 10% para residências dentro de 300 metros de paradas de BRT e mais de 25% para varejo e outros usos não residenciais, dentro de um raio de 150 metros de distância. Troy et al. (2008) mostraram que quando a taxa de criminalidade é relativamente baixa, os parques têm um impacto positivo nos valores das propriedades. Estes trabalhos demonstram a utilidade de um banco de dados reunindo características relacionadas a imóveis como um ponto de partida para a identificação dos comportamentos do mercado imobiliário, assim como o banco de dados consolidado neste presente estudo.

4. Conclusões

A partir do que foi discutido neste artigo, conclui-se que é possível criar uma base de dados unindo diversas características intrínsecas e extrínsecas de propriedades imobiliárias, a partir de anúncios disponibilizados em sites de compra e venda imobiliária e de informações geolocalizadas de características do ambiente urbano. Como limitação deste trabalho, cabe ressaltar que este processo foi feito apenas para a cidade de São Paulo, com dados de 2021, sendo dependente dos valores inseridos pelos usuários dos sites de anúncio, sujeitos a erros. Trabalhos futuros podem tanto expandir a metodologia usada para a construção do banco de dados aqui proposta, assim como utilizar a base de dados na modelagem de preço regional de propriedades imobiliárias.

Referencias

Abreu, M. A., & Amorim, W. V. (2014). O estudo do mercado imobiliário em cidades médias: procedimentos para coleta e sistematização dos dados. Geo UERJ, 2(25), 297-323.

Boulic, R. and Renault, O. (1991) “3D Hierarchies for Animation”, In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England.

Cervero, R., & Kang, C. D. (2011). Bus rapid transit impacts on land uses and land values in Seoul, Korea. Transport policy, 18(1), 102-116.

D’Acci, L. (2019). Quality of urban area, distance from city centre, and housing value. Case study on real estate values in Turin. Cities, 91, 71–92.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.Pinto, V. H. L., & Fernandes, R. A. S. (2019). Análise de preços hedônicos no mercado imobiliário residencial de Conselheiro Lafaiete, MG. Interações (Campo Grande), 20, 627-643.

Paz, R.R., Nobre, L.H., & Nobre, F.C. (2020). Determinantes De Preços No Mercado Imobiliário À Luz Do Modelo Hedônico. Revista Gestão em Análise, 9, 60.

Tierno, R., Carvalho, P. A., & MINISTÉRIO DAS CIDADES. (2007). O registro imobiliário: Conceitos e Bases Legais. PINHEIRO, OM et al. Acesso à terra urbanizada: implementação de planos diretores e regularização fundiária plena. Florianópolis: UFSC, 239-278.

Troy, A., & Grove, J. M. (2008a). Property values, parks, and crime: A hedonic analysis in Baltimore, MD. Landscape and Urban Planning, 87(3), 233–245.