REVISTA CAMINHOS DE GEOGRAFIA ISSN 1678-6343
http://www.seer.ufu.br/index.php/caminhosdegeografia/ DOI: https://doi.org/10.14393/RCG249668395

POTENCIALIDADES E LIMITAÇÕES DOS DADOS DE WEB SCRAPING PARA O MAPEAMENTO DOS PREÇOS DOS IMÓVEIS URBANOS

Thaís Góes de Souza
Universidade Federal da Bahia – UFBA
Programa de Pós-Graduação em Engenharia Civil
Salvador, BA, Brasil
thaisgdsouza@gmail.com

Vivian de Oliveira Fernandes
Universidade Federal da Bahia – UFBA
Programa de Pós-Graduação em Engenharia Civil
Salvador, BA, Brasil
vivian.fernandes@ufba.br

Julio César Pedrassoli
Universidade Federal da Bahia – UFBA
Programa de Pós-Graduação em Engenharia Civil
Salvador, BA, Brasil
jpedrassoli@ufba.br

Fernanda Doracy Rocha Fonseca
Universidade Federal da Bahia – UFBA
Escola Politécnica, Salvador, BA, Brasil
fernandafonseca2313@gmail.com

RESUMO

Embora a disponibilidade de dados online tenha aumentado, o tratamento e a avaliação destes dados incorrem em discussões acerca dos instrumentos científicos que permitem mensurar as principais características dos dados, minimizando inconsistências próprias relacionadas ao objeto. O propósito deste artigo foi analisar a representatividade dos dados obtidos por *web scraping* dos anúncios presentes em dois sites de extensão nacional ao aplicar uma forma de depuração sobre os dados de venda dos imóveis, bem como verificar sua proporção em relação ao cadastro imobiliário residencial municipal. O método propôs a utilização de dados disponíveis na *web*, recuperados através da técnica de raspagem de dados online (*web scraping*). Na abordagem, prestou-se atenção quanto aos preços das médias nos períodos de referência, como do levantamento das potencialidades e limitações dos dados de *big data*, assim como o mapeamento da concentração espacial dos preços do mercado imobiliário classificados por tipo e espacializados por bairro. A pesquisa concluiu que a base do site Olx apresentou menor completude e menor volume se comparado ao Imovelweb (Iw), porém maior variedade referente à cobertura espacial dos imóveis como o mapeamento da distribuição das médias dos preços do m² por bairro.

Palavras-chave: Mercado imobiliário. Preço da terra. Big data. Mapa dos preços.

POTENTIALS AND LIMITATIONS OF WEB SCRAPING DATA FOR MAPPING URBAN PROPERTY PRICES

ABSTRACT

While the availability of online data has increased, the processing and evaluation of this data give rise to discussions regarding the scientific instruments that enable the measurement of critical data characteristics while minimizing inherent inconsistencies related to the subject. This article analyzed the representativeness of data obtained through web scrapping from advertisements on two nationally recognized websites. It applied a data refinement technique to property sales data and examined its proportion concerning the municipal residential real estate registry. The method proposed the utilization of web-available data retrieved through the technique of online data scraping. The approach focused on average prices during reference periods, as well as on the assessment of the potentialities and limitations of big data, along with the spatial concentration mapping of real estate market prices categorized by type and spatialized by neighborhood. The research concluded that the Olx website dataset exhibited lower completeness and volume than Imovelweb (Iw) yet

demonstrated greater diversity regarding spatial coverage of properties, including mapping the distribution of average per-square-meter prices by neighborhood.

Keywords: Real estate market. Land price. Big data. Price map.

INTRODUÇÃO

A literatura tem sinalizado o cenário atual de maior profusão da informação espacial no mundo, mensurado em "alguns quintilhões de *bytes* de dados são criados todos os dias" (GRIFFIN; ROBINSON; ROTH, 2017, p. 5). Seguindo tal perspectiva, expande em conjunto, a complexidade cartográfica em projetar mapas que importam, bem como o de fornecer suporte à análise exploratória dos dados produzidos através de fontes não oficiais, de forma voluntária por usuários da internet. Em meio a esse contexto, emerge a necessária transferência do conhecimento interdisciplinar entre as ciências, devido à complexidade de *big data* em suas formas atuais e futuras (GRIFFIN; ROBINSON; ROTH, 2017; ROBINSON et al., 2017; DAVIS, 2018).

Sabe-se que em maior período da trajetória geográfica, as informações foram tradicionalmente produzidas de forma analógica através das representações de mapas por agências de mapeamento, difundidos através do papel para usuários pesquisadores e público em geral. O desenvolvimento tecnológico dos Sistemas de Informações Geográficas (SIG) transformou, primeiramente, a aquisição dos dados brutos e os processos de processamento e representação dos mesmos em posteriores informações. Diante deste cenário, considerou-se contributiva a disseminação da *Web* 2.0, quanto à disseminação de diversas áreas da ciência, sobretudo da cartografia em toda a sociedade. Isto pode ser pensado, sobretudo pelo modo acelerado e disseminado gerado pelo fenômeno de inclusão da população não-técnica na produção de informação cartográfica (GOODCHILD; GLENNON, 2010; MARTINS JUNIOR; SILVA, 2018).

A potencialidade de análise sobre *big data* vem sendo mostrada em aplicações de respostas imediatas. Nesta perspectiva, o estudo realizado diariamente durante a pandemia de Covid-19 mostrado por Bricongne et al (2021), a partir da mineração em tempo real dos dados de preço do mercado imobiliário forneceu uma alternativa ao atraso e nível agregado dos dados das estatísticas oficiais das áreas próximas à Londres. Neste estudo foi possível calcular indicadores que refletiram a perspectiva dos vendedores, tais como o número de novos anúncios publicados ou a forma como os preços flutuaram ao longo do tempo para os anúncios existentes.

Aplicado ao ambiente urbano, devido à própria dinâmica socioespacial, sabe-se que oferta de dados digitais georreferenciados sobre as cidades será cada vez maior. Por conta do desenvolvimento exponencial das tecnologias, espera-se que as técnicas de tratamento destes dados não sofram interrupções (RAMOS, 2002; GOODCHILD, 2007; GOODCHILD; GLENNON, 2010;). Há de se apontar, ainda, a bem-vinda evolução do conhecimento sobre os dados *crowdsourcing,* uma vez que a Cartografia se beneficia das práticas colaborativas e se relaciona com as necessidades de órgãos oficiais de mapeamento na utilização e interligação de bases de informações voluntárias para atualizações oficiais (BRAVO; SLUTER, 2015).

De acordo com a contextualização apresentada, propõe-se o questionamento quanto ao maior detalhamento da completude – ou seja, o quão completa está a base de dados após o processo de mineração dos mesmos, a partir da pesquisa em dois *websites* e volume dos dados obtidos a partir do *web scraping* dos anúncios dos imóveis urbanos. Nesta perspectiva buscou-se assegurar que os atributos coletados, a partir dos critérios definidos quanto ao objeto e transação, similaridade, área, localização e preço, estivessem integralmente preenchidos, retratando assim a completude dos dados obtidos. Em seguida foi possível analisar a distribuição espacial dos preços dos imóveis urbanos dos anúncios imobiliários, enquanto base observatória do mercado imobiliário. Por fim, o mapeamento permitiu verificar onde ocorrem os agrupamentos espaciais dos preços dos imóveis, conforme o estudo de caso, os bairros da cidade de Salvador, na Bahia.

CONVERGÊNCIA ENTRE A CARTOGRAFIA, OS DADOS DE INTERNET E O MERCADO IMOBILIÁRIO

A literatura cartográfica destaca os desafios em integrar dados oriundos de fontes distintas, específicos para aplicação e estudos urbanos (BATTY, 2013), tanto das vantagens da coleta de dados atualizados de *big data* (GOODCHILD, 2021) por rastreadores e APIs, especialmente acerca da estrutura, veracidade dos dados inseridos pelos usuários (WENCESLAU; DAVIS JUNIOR; SMARZARO, 2017), além do potencial das informações criadas pelos usuários através dos sistemas colaborativos (FERNANDES; ELIAS; ZIPF, 2020). Neste contexto, é visível o potencial de grandes bancos de dados online gerados por usuários para aplicações institucionais (LOBERTO; LUCIANI; PANGALLO, 2018), uma vez que a técnica de mineração é uma fonte ágil de ganho de dados aplicada em diversas áreas de pesquisa. Do ponto de vista científico, estes dados são oportunos também nas ciências sociais, aplicado às etnografias ou estudos de caso em amplas escalas (KITCHIN, 2014).

As características desses dados com intensa divulgação espaço-temporal podem contribuir juntamente aos dados das instituições públicas oficiais de forma atualizada numa série de estudos quanto aos preços do solo nas cidades, como na atualização das Plantas Genéricas de Valores (PGV). As informações geográficas produzidas por cidadãos, muitas vezes conhecidas como Informação Geográfica Voluntária ou *Volunteered Geographic Information* (VGI), forneceram uma alternativa interessante às informações provenientes de instituições responsáveis pelo mapeamento de referência e podem complementar a informação das estatísticas oficiais (GOODCHILD; GLENNON, 2010; GOODCHILD; LI, 2012).

Dentre as principais dificuldades no mapeamento e análise espacial do preço real da terra e dos imóveis, encontra-se a criação de um modelo que reflita sobre a lógica de mercantilização utilizadas pelo mercado imobiliário e que contemple os valores do cadastro imobiliário municipal. As informações municipais, muitas vezes desatualizadas e inexistentes sobre os preços imobiliários, são diferentes em termos dos preços e localizações, quando praticadas pelas incorporadoras e anunciantes. Nesse panorama, como apontado por Davis (2018), há uma lacuna em integrar dados oficiais aos dados produzidos por aqueles que publicam na web, devido aos desafios em confiabilidade, completude, sistematização e dinâmica entre os conjuntos de dados.

Os dados dos anúncios dos imóveis são capazes de refletir as nuances espaciais dos processos de acumulação na produção do espaço urbano, uma vez que, de acordo com Carlos (2015), o processo de acumulação vigente no espaço urbano o coloca enquanto espaço da localização e suporte das relações sociais de produção e da propriedade O espaço urbano capitalista, segundo Corrêa (2000), é fragmentado ao passo que é articulado e projetado por agentes que produzem e consomem o espaço, em repetido processo de reorganização espacial com agrupamentos complexos de usos da terra. O entendimento sobre os instrumentos legais e financeiros atuantes no espaço geográfico possibilita a oferta de subsídios para a gestão e planejamento territorial urbano de forma consciente.

Estudos sobre relação entre o mercado imobiliário e o espaço urbano já são amplamente conhecidos (ABRAMO, 1989, 2009; RONILK, 2015; MARICATO, 1985; SPOSITO; SPOSITO, 2020; SPOSITO, 2016; CARLOS, 2015, 2016). Atualmente, esforços iniciais são aplicados em estudos quanto à integração, completude e mapeamento dos dados de *big data* incluídos aos dados oficiais. Tal ponto, forma uma importante lacuna de pesquisa e apresenta potencial de complementação às análises geográficas, quanto à coleta dos dados das ofertas das vendas a partir da técnica de *web scraping* e mapeamento dos imóveis na cidade.

Dados de web scraping do mercado imobiliário

A técnica de *web scraping* ou *web crawling*, *data* scraping equivale a atividade de se recuperar o conteúdo de uma página da *Web* individualmente (FATMASARI; KUNANG; PURNAMASARI, 2018), dos dados por completo ou apenas do dado requisitado do *website* e recuperado em arquivo ou banco de dados para análise posterior (POONGODAI; SUHASINI, 2019). Trata-se, então, de um conceito exato, cuja aplicação, segundo Fatmasari; Kunang e Purnamasari (2018) possibilita a obtenção de dados de inúmeras áreas do conhecimento.

Dados e anúncios online têm sido cada vez mais utilizados em pesquisas, principalmente para estudar o mercado imobiliário (CHAPELLE; EYMÉOUD, 2022). A utilização da técnica havia sido aplicada no monitoramento sobre os preços dos imóveis dos aluguéis nos EUA, no estudo de Boeing e Waddell (2016). Este estudo foi realizado em momento inicial de disseminação da aplicabilidade da

mineração dos dados, bem como no exercício realizado para a espacialização dos dados dos preços dos imóveis.

Em sequência, investigações de caráter mais amplo podem ser verificadas nas investigações de Neder et al. (2017) e Grybauskas, Pilinkienė e Stundžienė (2021) (2021). Na proposta de construção de um índice da defasagem do Imposto Predial Territorial Urbano (IPTU) do município de Uberlândia (MG), Neder et al. (2017) racionalizou os valores médios do mercado imobiliário local via *web scraping* com os valores venais por bairro. Já Grybauskas, Pilinkienė e Stundžienė (2021) propuseram raspagens mensais em período inicial e final da quarentena do coronavírus em 2020, a partir da linguagem de programação em *Python* (pacotes *BeautifulSoup e Selenium*), aliado ao tratamento das variáveis desejadas por aprendizado de máquina das listagens obtidas dos apartamentos localizados nos bairros da capital Vilnius (Lituânia).

O uso da técnica de *web scraping*, assim como a espacialização dos resultados, revela padrões espaciais urbanos e possibilita aos gestores estimarem preços em escala local e temporal. Os resultados referentes ao enquadramento da base de dados destes anúncios podem ser analisados a partir de duas perspectivas. Primeiramente, os resultados incitam as discussões, desenvolvimento conceitual e procedimentos do trabalho pertinente ao tema de *big data*, uma vez que dados extraídos de sites e reformulados em um conjunto de dados estruturados são fontes não tradicionais de *big data*. Por outro lado, alguns trabalhos mostraram empiricamente a importância quanto a atenção da depuração das variáveis/campos que compõem a base explorada obtida através da mineração.

Boeing e Waddell (2016) ao monitorarem o mercado imobiliário dos Estados Unidos através das listagens dos preços dos aluguéis obtidos por *web scraping* apresentaram como motivação primária a propriedade desses conjuntos de dados serem menos investigados, à época, no meio científico. Os autores propõem sistematização das etapas de trabalho sobre os dados de *web scraping* em sequência lógica de ação, resumida em: (1) Coletar; (2) Organizar; (3) Analisar; (4) Mapear; (5) Visualizar, a serem cumpridas desde a aquisição até processo de visualização dos dados. Embora com abordagem semelhante, Zhao (2017) simplificou o processo de coleta de dados da internet em apenas duas etapas sequenciais: adquirir o recurso da *web* (1), para em seguida extrair os dados (2).

Quanto às repetições dos anúncios, podem estar associadas a uma necessidade de vender brevemente ou às dificuldades para encontrar um comprador. Ao se trabalhar com maiores amostras o problema se mostra menos grave, porém os anúncios duplicados são particularmente prejudiciais para medir as taxas de crescimento em pequenas amostras devido a representação em excesso da unidade habitacional específica a que foram associados (LOBERTO; LUCIANI; PANGALLO, 2018).

Tal colocação foi reforçada por Tomal (2020), ao apontar que um mesmo apartamento pode ser postado em uma plataforma de internet, tanto pelo proprietário, quanto por uma ou várias agências imobiliárias que cooperam entre si. Estes anúncios referentes à mesma residência podem ser simultaneamente ou em diferentes momentos postados pelos usuários, o que pode significar que o número de anúncios é muito maior do que o número real de habitações no mercado (LOBERTO; LUCIANI; PANGALLO, 2018).

No que se refere ao estudo do potencial de *big data* da habitação aplicado ao mercado imobiliário italiano, a partir da extração dos anúncios de venda de imóveis de um portal *online* popular de serviços imobiliários, Loberto; Luciani e Pangallo (2018), argumentaram que o conjunto, apesar dos problemas de completude e repetições dos anúncios para mesma unidade, possibilitam análises em nível geográfico, bem como preencheu uma grande lacuna nas estatísticas do mercado imobiliário italiano, quanto à ausência dos dados das características físicas das casas vendidas. Os resultados dessa pesquisa mostraram que, apesar das distorções da base, a qual foi corrigida a partir do aprendizado por máquina, os dados minerados forneceram evidências sobre sua consistência e relevância quantitativa. Na investigação de Tomal (2020), cujo objetivo foi identificar os determinantes que afetam os preços de aluguel na Cracóvia (Polônia) foram eliminadas da base de dados, as observações discrepantes, as quais não foram informadas as localizações exatas da propriedade.

Nesta linha, tanto o estudo de Bricongne; Meunier e Sylvain (2021) (2021), quanto o de Chapelle e Eyméoud (2022), chamaram atenção sobre recuperar o campo de descrição geral ou da parte não estruturada do anúncio, presente em maior parte da arquitetura dos anúncios. De acordo com os autores, obter este campo possibilita enriquecer a base, uma vez que pode permitir a busca em seu conteúdo, das palavras-chave referentes às instalações adicionais, por exemplo, se há varanda no imóvel, qual o número de vagas de garagem, número do andar, dentre outros. Ainda assim, pontuaram uma atenção quanto a disponibilização do dado de metragem como das unidades de área, em que devem estar expressos na mesma medida.

Outrossim, posto que se trata de um tema atual e relevante em contribuição prática, social e acadêmica, de acordo com o objetivo acadêmico do presente trabalho, pontua-se a relevância e objeto da Lei N° 13.709/2018 – Lei Geral de Proteção dos Dados[1] (LGPD), a fim de atingir o aspecto jurídico sobre a extração dos dados dispostos em rede. Com a proteção aos dados pessoais da pessoa física, esta lei compreende tanto a área jurídica, quanto a área de recursos humanos e de sistema da informação. É uma lei complexa, mas que trata dos dados virtuais da pessoa física à medida em que força uma melhor curadoria sobre as relações físicas com a tecnologia.

Os dados divulgados nestes *websites* de comercialização imobiliária podem ser considerados dados públicos, uma vez que suas publicações são intencionadas para tal finalidade. Tais dados não se conceituam enquanto dados sensíveis[2] descritos na LGPD. A Lei enfoca a proteção aos dados sensíveis da pessoa física, em que para todo e qualquer dado ou informação de pessoa física coletado por parte de empresas, torna-se necessário a comunicado aos seus titulares, bem como sua guarda, de acordo com a referida legislação.

Ainda assim, sob o ponto de vista do VGI, esses dados divulgados nos *websites* de comercialização imobiliária, também podem ser considerados enquanto dados passivos, ao passo que são publicados por usuários-anunciantes sem a intencionalidade de serem colaboradores da informação. De acordo com Lima e Silva (2021), a regulamentação dos usos dos dados ainda requer novas dimensões de análise de coleta e tratamento, visto a necessidade do uso e análises por dados atualizados e divulgados na internet.

METODOLOGIA

A metodologia empregada nesta pesquisa compreendeu os procedimentos de coleta, tratamento, análise descritiva dos dados minerados e a espacialização das médias dos preços do m² por tipo de imóveis nos bairros da cidade de Salvador - BA, de acordo com a Figura 1. Para tal, são utilizadas duas variáveis fundamentais para análise: **i)** *web scraping* dos preços dos anúncios de venda (apartamentos, casas e terrenos); **ii)** quantitativo dos imóveis residenciais presentes no cadastro imobiliário municipal.

Conforme os objetivos desta pesquisa, desenvolveu-se uma metodologia de três etapas para aquisição de dados dos anúncios a partir do *web scraping (ws)* em plataformas de anúncios de imóveis, utilizando o raspador online *Data Miner.* A etapa de seleção consistiu, a partir dos sites do Imovelweb (Iw) e do Olx, na filtragem dos anúncios de venda dos imóveis residenciais, os quais foram recuperados dados com atributos localizacionais, em formato de texto à nível da descrição de “bairro” para a cidade. A etapa de configuração da ferramenta de extração equivaleu a identificação dos dados disponíveis na página, dado de área e preço por produto imobiliário e do campo “descrição geral” dos anúncios, para posterior cálculo do m² e espacialização da distribuição das médias dos preços do produto por bairro da cidade. Obteve-se assim, os dados para ofertas dos imóveis residenciais dos anúncios em arquivo *.xls* dos dados das páginas buscadas.

Por conseguinte, utilizou-se a etapa de depuração ou pós-scraping para assegurar a qualidade e completude da base obtida, a qual consiste em editar dos campos vazios, ao recuperar dados como nome do bairro, área, a partir do campo descrição do anúncio; identificar e excluir anúncios repetidos, com dados inconsistentes quanto à localização, anúncios sem o dado preço e metragem do imóvel.

O outro conjunto de dados processados em ambiente SIG foram da base vetorial e tabular com campo nome_de_bairro e geocodigo dos limites oficiais de bairros do município, obtidos através da Secretária de Desenvolvimento Urbano de Salvador (SEDUR, 2020). Assim, para atender ao critério de localização foram correlacionados os dados de bairros obtidos na extração com a nomenclatura oficial de bairros definidos por Lei municipal. Dessa forma, as etapas realizadas permitiram a organização dos dados não estruturados da mineração online dos anúncios para espacialização dos padrões das médias dos preços dos imóveis.

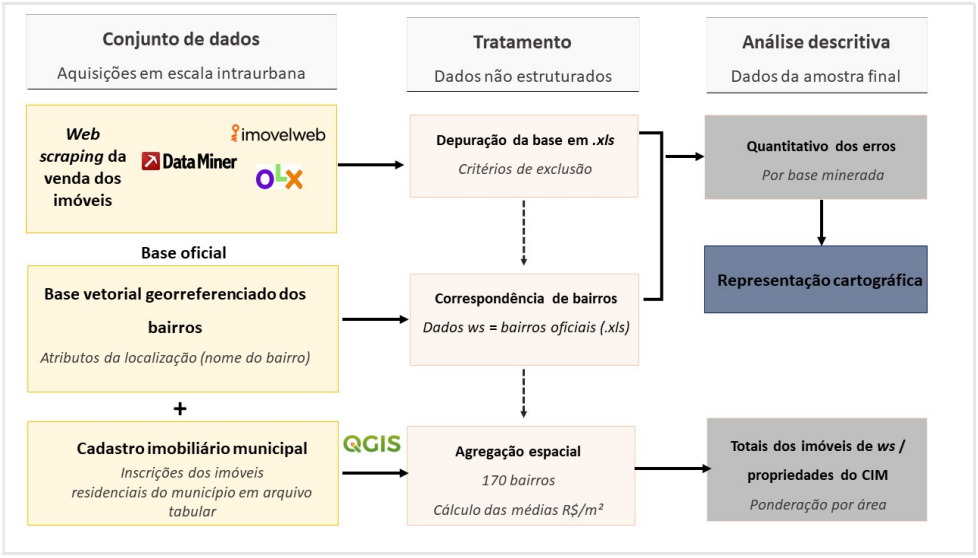
Em sequência, para verificar a representatividade por bairro da base por *web scraping* das ofertas dos imóveis, utilizou-se a base cadastral imobiliária residencial do município (SALVADOR, 2020), para calcular as razões entre volume dos anúncios extraídos dos imóveis (por tipo) e as propriedades presentes no sistema cadastral. Contudo, para compatibilizar os quantitativos das inscrições

[1] Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em 13/02/2022.

[2] Em seu art. 5º, § II, os dados sensíveis são relacionados à personalidade do indivíduo, escolhas pessoais, origem étnica e racial, convicção religiosa, opinião filosófica e política, dado referente à saúde (LIMA; SILVA, 2021; BRASIL, 2018).

residenciais do Cadastro Imobiliário Municipal (CIM) agregadas por setores fiscais aos bairros foram inicialmente calculadas as áreas dos setores fiscais, bem como das áreas dos bairros, em ambiente SIG. Com este fim, os quantitativos das inscrições residenciais imobiliárias distribuídos nos setores fiscais do município foram ponderados, de acordo aos limites definidos por Lei de delimitação dos bairros. Por fim, mapeou-se as distribuições das médias dos preços do m², agregada por unidade espacial de bairro.

Figura 1 - Procedimentos metodológicos e etapas da pesquisa.



Fonte - Autores (2023).

Por fim, para as operações espaciais de agregação dos dados tabulares, resultantes da extração dos imóveis e do cadastro imobiliário para unidade de bairro, utilizou-se o Sistema de Informação Geográfica de licença gratuita, Quantum GIS (QGIS), versão 3.10. O método utilizado para a classificação dos dados mapeados foi o método de quebras naturais, para maximizar a variância entre as 5 classes agrupadas. As variáveis são descritas no Quadro 1.

Quadro 1 - Relação das variáveis utilizadas na pesquisa.

Símbolo	**Variável**	**Fonte**	**Período/a**	**Descrição (unidade)**	**Análise por agregado de bairro**
X1	Média dos preços dos apartamentos	www.imovelweb.com www.olx.com.br	Dez. (2020) /Jan. (2021). Jan. (2022)	Preço unitário de venda por amostra (R$) /área	Distribuição das médias de preço de venda por m²
X2	Média dos preços das casas				
X3	Média dos preços dos terrenos				
X4	Cadastro imobiliário municipal	Secretaria Municipal da Fazenda	2020	Totais dos imóveis (nº) extraídos por base	Correlação entre a variável X_1/ X_2/X_3

Fonte - Autores (2023).

RESULTADOS E DISCUSSÃO

Na plataforma de comercialização imobiliária: https:\\www.imovelweb.com.br, extraiu-se 32.639 observações durante o período de 28/12/2020 a 09/01/2021 dos anúncios de venda dos apartamentos e casas. Já os anúncios dos terrenos datam de 25 de fevereiro de 2021. Na plataforma online de comercialização de produtos gerais: https:\\www.olx.com.br, extraiu-se 25.386 anúncios durante o mês de janeiro de 2022 para todos os tipos de imóveis. A Tabela 1 mostra uma sumarização deste universo por tipologia de imóvel para cada base extraída e o resultado da depuração do conjunto dos anúncios.

Tabela 1 - Quantitativo dos anúncios extraídos e depurados dos bairros por tipologia.

Website	**Imóvel**	**Universo extraído**		**Amostra depurada**		**Bairro[1]**	
		(n)	(%)	(n)	(%)	(n)	(%)
Imovelweb	Apartamento	28.090	86,06	26.055	88,31	120	70,59
	Casa	3.801	11,65	2.737	9,28	111	65,29
	Terreno	748	2,29	713	2,42	81	47,65
Total das observações		32.639	100%	29.505	100%	--	--
Olx	Apartamento	18.634	73,40	14.648	78,06	135	79,41
	Casa	5.000	19,70	2.961	15,78	140	82,35
	Terreno	1.752	6,90	1.155	6,16	129	75,88
Total das observações		25.386	100%	18.764	100%	--	--

Fonte: - Autores (2023) com base nos dados extraídos e depurados das plataformas *web*.

[1] Cobertura percentual calculada a partir do total definido por Lei oficial dos 170 bairros em Salvador (BA).

Cumpre destacar que para minimizar os erros próprios deste conjunto de dados foram excluídas as inconsistências associadas à base obtida para cada plataforma online de anúncio coletada. As exclusões ocorreram devido à não caracterização do tipo do imóvel anunciado, às similaridades, ausência do dado de preço e área, ausência do nome do bairro e atribuição de localização em outro município que não seja o de Salvador - BA. O intuito desta depuração foi retirar do conjunto de dados os anúncios em que os dados não foram recuperados após a consulta ao campo de descrição geral. A recuperação destes dados de área, preço e localização (por toponímia), seguem as propostas de Bricongne; Meunier e Sylvain, (2021) e Chapelle e Eyméoud (2022). A etapa de depuração das inconsistências associadas à base de *web scraping* seguiu os cinco critérios de exclusão especificados abaixo:

i. Quanto ao **objeto** (residencial) e **transação** (venda): excluídos as propriedades com caracterização divergente do tipo do imóvel anunciado;

ii. Quanto à **similaridade** (repetições): excluídos os anúncios em que o conjunto de todos os caracteres dos campos: título, área, localização, preço e descrição foram similares, indicando duplicidades. Porém, as exclusões não significaram a retirada efetiva (real) dos anúncios repetidos, haja vista que uma mesma unidade imobiliária pode ser publicada por diferentes anunciantes, como em períodos e com dados divergentes de preços (WENCESLAU; DAVIS JUNIOR; SMARZARO, 2017; LOBERTO; LUCIANI; PANGALLO, 2018);

iii. Quanto à **área** (m²): retirados os anúncios que não foram especificados os dados de metragem do imóvel, com dados discrepantes, abaixo de 10 m² (imóveis casa e apartamento), valores em casa do bilhar, ou ainda aqueles em que constam apenas a área do terreno em sua descrição e não o dado da área privativa;

iv. Quanto à **localização** (nomenclatura de bairro): neste critério foram excluídos erros referentes à toponímia e ausência de dados textual do nome do bairro. De modo geral, este critério deve obedecer a um dado que indique a sua localização (TOMAL, 2020), tais quais, dado textual de endereço, nome do bairro ou dado em formato GMS;

v. Quanto ao **preço** (R$): excluídos os anúncios, os quais não foram especificados a precificação do imóvel, ou que apresentam somente o "valor de entrada do parcelamento" (na descrição do anúncio), ou preenchidos com valor igual a zero (0), e ainda aqueles abaixo da casa do milhar, ou com valores discrepantes na casa do bilhar. A completude deste dado foi imprescindível para o cálculo do preço médio por área (m²).

Nesta pesquisa, buscou-se assegurar que os atributos coletados, quanto ao objeto e transação, similaridade, área, localização e preço, estivessem integralmente preenchidos. O campo descrição geral do anúncio foi extraído a fim de verificar e recuperar dados não preenchidos, quando possível. Assim, entendeu-se que seria garantido a completude do dado por base analisada.

Observou-se que o maior quantitativo de anúncios decorre dos imóveis do tipo apartamento, seguido dos anúncios de casas e dos terrenos, este com menor expressão de publicação, obtidos nas duas plataformas, conforme visto na Tabela 1. As bases totalizam 58.025 anúncios de venda como resultado prévio da mineração, ao final do módulo de depuração obteve-se 48.269 registros das ofertas. Dessa totalidade foram excluídos 9.756 (Iw = 3.134, Olx = 6.622) anúncios de todo o conjunto de dados coletados. As exclusões da amostra depurada representam aproximadamente 16,81% dos anúncios somados dos sites.

A maior taxa do total dos anúncios ao final da depuração foi da base do Olx com variação de depuração de 35,3% em relação ao universo. Já a base do Imovelweb varia 10,6% do universo apresentado na amostra depurada resultante. Seguindo os valores pontuados, a coleta dos dados do Iw apresenta melhor resposta quanto a completude dos dados, confirmada através da depuração.

No que se refere ao volume total dos anúncios extraídos por *web scraping*, este representou cerca de 6% do quantitativo representado pelas 785.241 inscrições residenciais presentes na base oficial confiável (WENCESLAU; DAVIS JUNIOR; SMARZARO, 2017), do cadastro imobiliário municipal. Destes, os anúncios do Imovelweb apresentaram maior volume total (3,76%), em comparação à base do Olx (2,39%), conforme a Tabela 2. Os valores percentuais mostraram uma distribuição heterogênea por imóvel no espaço intraurbano relacionado à base cadastral. Isto, notadamente pode indicar uma sujeição prevista na valoração das áreas da cidade, com efeitos de uma ocupação formal e informal territorial.

Tabela 2 - Totais dos imóveis extraídos em relação ao cadastro imobiliário municipal (CIM)

***Imovelweb*/Cadastro Imobiliário Municipal**				
Imóvel	Apartamentos (n=26.055)	Casas (n=2.737)	Terrenos (n=713)	Percentual dos anúncios
Totais	3,32%	0,35%	0,09%	3,76%
***Olx*/Cadastro Imobiliário Municipal**				
Imóvel	Apartamentos (n=14.648)	Casas (n=2.961)	Terrenos (n=1.155)	Percentual dos anúncios
Totais	1,87%	0,38%	0,15%	2,39%

Fonte - Autores (2023) com base na depuração da base e dados do CIM (SEFAZ SALVADOR, 2021).

Web scraping da plataforma de anúncios imobiliários Imovelweb

A plataforma do Imovelweb, em específico, oferece somente anúncios imobiliários com a prevalência dos anunciantes corretores, uma vez que os anúncios exigem o pagamento para publicização. Entende-se, devido a essas características, que o site é voltado para a comunicação digital para a comercialização do mercado imobiliário formal nas cidades.

A partir do total da amostra inicial dos anúncios, verificou-se o percentual de aproximadamente 9,60% referente às exclusões deste *website*. Ainda assim, a menor parte dos anúncios depurados decorreram dos terrenos, com exclusão de 4,68% em relação ao seu total extraído, o que indica menor ocorrência de erros associados à divulgação deste tipo de produto. Em ordem decrescente,

segundo os critérios de exclusão dos anúncios, as taxas dos erros foram nos imóveis do tipo apartamento com 7,25% e, os de casa com 28,00% em maior taxa de inconsistências associadas. De acordo com a Tabela 3, indicam-se as porcentagens das amostras obtidas após a depuração dos anúncios totais, a partir dos critérios de exclusão.

Tabela 3 - Erros depurados por imóvel no Imovelweb

Produto	Anúncios extraídos	Objeto	Similar	Área	Localização	Preço	Amostra final	Resultado (%)
Apartamentos	28.090	46	1.561	285	119	24	26.055	92.75%
Casas	3.801	10	838	143	66	4	2.737	72,00%
Terreno	748	11	19	0	3	1	713	95.32%
Total	32.639	67	2.418	428	188	29	29.505	90,40%

Fonte - Autores (2023) com base nos dados depurados após extrações.

Quanto às proporções das inconsistências associadas aos anúncios de todos os imóveis extraídos do site, verificou-se os 7,4% referentes ao critério de similaridade (anúncios duplicados), seguindo de 1,3% das inconsistências de acordo à área, 0,6% para localização, 0,2% para objeto (anúncios diferentes ao imóvel anunciado) e 0,09% para anúncios com inconsistência quanto ao dado de preço. Ao analisar isoladamente por tipo de imóvel, as maiores inconsistências ocorreram nos imóveis dos apartamentos, devido ao seu maior volume coletado. Todos os imóveis desse conjunto seguiram o mesmo padrão de verificação dos cincos critérios observados em todo o conjunto.

No que se refere à cobertura dos anúncios nos bairros na amostra final, a distribuição dos apartamentos nos 120 bairros representou maior percentual, seguido da distribuição dos anúncios de casas e terrenos. De maneira geral, os percentuais de cobertura espacial variaram de acordo com a classificação por produto (ou tipo de imóvel), em que os anúncios coletados dos apartamentos e casas (71% e 65%, respectivamente) recobrem a maior parte dos bairros da cidade. Todavia, ao verificar a distribuição total dos anúncios dos terrenos em Salvador, as publicações dos 713 anúncios deste imóvel dispuseram da menor cobertura espacial, com 48% em relação aos 170 bairros oficiais.

Web scraping da plataforma de comercialização Olx

A partir dos critérios estabelecidos na depuração, a base Olx apresentou maior quantificação de inconsistências, quando comparado a base do Imovelweb. Do total da amostra inicial dos anúncios, verificou-se a exclusão de 26,08% do seu conjunto. Ainda assim, a menor parte dos anúncios depurados foram dos apartamentos devido a exclusão de 21,40% em relação ao seu total extraído, o que indicou menor ocorrência de erros associados. Ainda em relação ao universo extraído, em ordem decrescente, segundo os critérios de exclusão, os anúncios dos terrenos obtiveram 34,08% e os de casa 40,78% - imóvel com maior proporção de inconsistências associadas. De acordo com a Tabela 4, indicam-se as proporções da amostra obtida após a depuração dos anúncios extraídos por imóvel.

Tabela 4 - Erros depurados por imóvel do Olx

Produto	Anúncios extraídos	Objeto	Similar	Área	Localização	Preço	Amostra final	Resultado (%)
Apartamentos	18.634	60	825	1.299	267	1.535	14.648	78,60%
Casas	5.000	23	116	986	19	895	2.961	59,22%
Terreno	1.752	27	24	164	210	172	1.155	65,92%
Total	25.386	110	964	2.449	496	2.602	18.765	73,92%

Fonte - Autores (2023) com base nos dados depurados após extrações.

Quanto às proporções das inconsistências associadas aos anúncios de todos os imóveis extraídos do site, verificou-se cerca de 10,25% referente ao critério de preço, seguido de aproximadamente 9,65% das inconsistências de área, 3,8% de similaridade (anúncios duplicados), 1,95% para a localização, 0,43% para objeto (anúncios diferentes ao imóvel anunciado). Neste aspecto, a base do Olx apresentou diferença no resultado da depuração, quando comparado aos resultados do Iw. O critério com maior proporção foi referente ao dado preço (no Imovelweb, o critério foi o de similaridade).

Outras diferenças observadas quanto às duas bases, pôde ser destacada ao analisar isoladamente o tipo do imóvel. As maiores inconsistências foram dos dados referentes aos imóveis do tipo casa, apesar do segundo menor volume coletado da sua base (Olx). Somente os apartamentos seguiram o mesmo padrão, quanto a verificação dos cinco critérios observados em todo o conjunto, enquanto, os imóveis de casa e terreno apresentaram divergência entre si. As casas apresentaram, em ordem decrescente, inconsistências associadas à área, ao preço, similaridade, objeto e localização. Já os terrenos apresentam inconsistências associadas à localização, ao preço, área, objeto e similaridade. Em termos quantitativos, tais padrões e maior depuração da base do Olx mostrou menor confiabilidade, quanto à completude dos dados desta base, embora apresente maior variedade, quanto à cobertura de bairro.

Nesse sentido, quanto à distribuição dos anúncios nos 170 bairros prevalece o volume dos anúncios dos apartamentos, seguidos das casas e terrenos. Pontua-se que todos os anúncios estiveram acima dos 75% de abrangência dos bairros da cidade. Assim, comparando as duas bases mineradas, a base do Olx apresentou maior cobertura espacial de anúncios para todos os tipos de imóveis nos bairros da cidade. Esta representatividade pode ser entendida devido a alguns fatores: variedade de nicho de mercado, a gratuidade do anúncio e, dessa forma, uma maior popularização.

A plataforma *web* Olx oferece anúncios de produtos variados, incluindo os imobiliários por diversos anunciantes, tendo ou não referência a agência ou atividade imobiliária na plataforma, uma vez que os anúncios não exigem o pagamento para publicização. Sob esta perspectiva, entende-se que esta caracterização possibilita a publicação dos anúncios do mercado informal de imóveis.

Limitações da base de dados

Problemas quanto à construção da base de *web scraping* foram identificados como limitações, tais quais: a completude dos dados publicados, que requerem depuração da base, assim como a inexistência de dados, como o ano de construção do bem imobiliário ofertado. Esse processo denota uma realidade de um mercado formal, sendo um nicho de mercado, o qual separa as áreas de alta rotação imobiliária das áreas populares e periféricas da cidade - estas com observações escassas.

A depuração da base é imprescindível devido a inconsistência associadas à base, como anúncios com registros incompletos de endereço, bairro e metragem de área, ajuste dos campos e exclusão de anúncios repetidos. Assim, preserva-se a possibilidade de recuperar os dados em campos vazios da tabela com a consulta do campo “descrição” do anúncio. Neste campo é comum os anunciantes qualificarem os atributos físicos e localizacionais do imóvel. Em termos de qualidade lógica, a verdade ou não do dado é peculiar ao modo operacional, não sendo objeto do estudo essa verificação por outra técnica de coleta, seja a partir de uma amostra ou universo dos dados. Tais dados provenientes da rede virtual colocam os seus usuários como produtores de informação.

Quanto às limitações analisadas por tipo de produto (imóvel) anunciado nesta base de *web scraping*, consideram-se algumas características gerais. Na base não estão discriminados os terrenos de grande área e maiores preços, situados às margens da cidade, tais quais os terrenos voltados à construção de habitações populares de programas instituídos pelo Estado. Sabe-se do menor estoque de imóveis do tipo terreno numa grande cidade, seguido do número de casas e maior volume de dados de apartamentos.

Em outro aspecto, pontua-se quanto aos preços extraídos dos anúncios. Os números das vendas das propriedades urbanas referem-se aos preços de oferta e não contabilizam a margem de negociação que possa ocorrer durante o trâmite comercial. Isto não é uma limitação em si, mas uma característica inerente ao dado. Todavia, utilizar a técnica de *web scraping* dos dados imobiliários é uma forma de minerar dados de acesso aberto e tem o mérito de indicar a partir dos preços, quais as localizações apresentaram maior número ofertas, maiores e menores médias dos preços do m², dos imóveis urbanos. Ainda assim, torna-se possível obter dados sobre características internas.

Análise da distribuição espacial dos dados de web scraping

Quanto a espacialização dos dados minerados foram identificados nos mapas abaixo (Figuras 2, 3 e 4), os bairros com as maiores médias do preço do m² calculados através das variáveis preço e variável estrutural área extraídas dos anúncios foram mais significativos (em maior intensidade da cor vermelha). As distribuições representadas em cor amarelo mostraram os bairros onde os valores foram os mais baixos da cidade. O bairro popular foi aquele com menor expressão de publicação e preço, já o nobre com maior volume e preços médios mais altos do m².

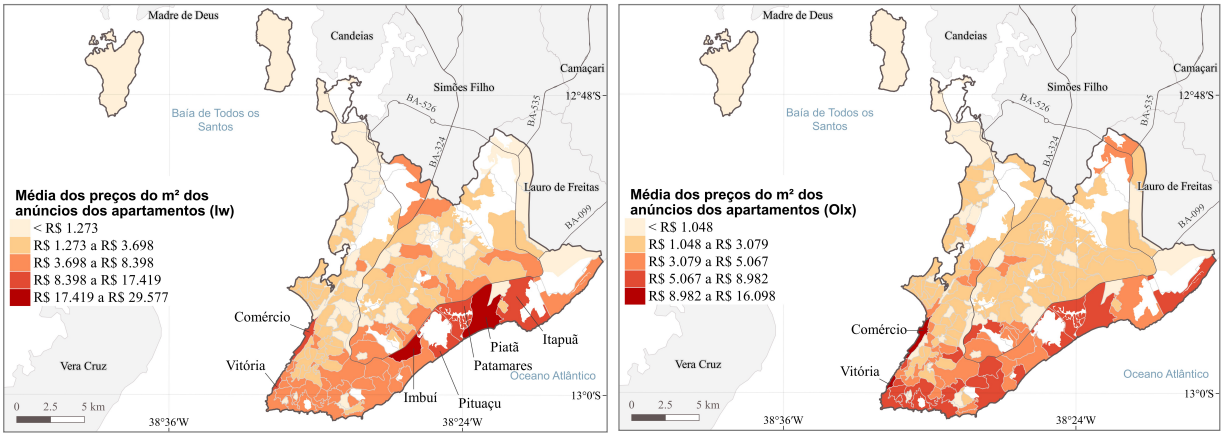
Acrescentadas às diferenças em volume dos anúncios já analisadas, verificou-se divergências dos preços entre os imóveis das bases. Observou-se uma forte concentração das médias baixas do preço dos imóveis nos bairros populares localizados próximos a orla da Baía de Todos os Santos, enquanto a concentração dos preços médios do m² mais altos nos bairros localizados na fachada sudeste em direção a Orla Atlântica de Salvador, que pode ser entendido certamente devido a permanente produção do espaço urbano de áreas nobres já com residências privilegiadas, a um interesse do mercado imobiliário nestes bairros da cidade.

Ao analisar as diferenças entre as bases dos anúncios dos imóveis de apartamento, notou-se menores intervalos dos valores das médias de preços do m² na base do Olx em comparativo ao Iw, com maiores preços. Entre as bases dos anúncios dos imóveis de apartamento e casa, notou-se menores intervalos dos valores das médias dos preços do m² na base do Olx, em comparativo ao Iw, com maiores preços. Porém, ao analisar as diferenças entre as bases dos anúncios dos imóveis dos terrenos, verificaram-se menores intervalos dos valores das médias de preços do m² na base do Olx em comparativo ao Iw, este com maiores preços.

O resultado da distribuição das médias dos preços dos m² dos imóveis na cidade, corroborou com a leitura que se dá sobre uma possível produção formal e informal do espaço urbano entre áreas de maiores e menores valores do solo urbano. Sobre isto, revelou-se um ponto crítico, ao verificar uma altos preços por imóveis nos bairros com proximidade do mar, devido às maiores médias concentradas nos bairros da orla atlântica.

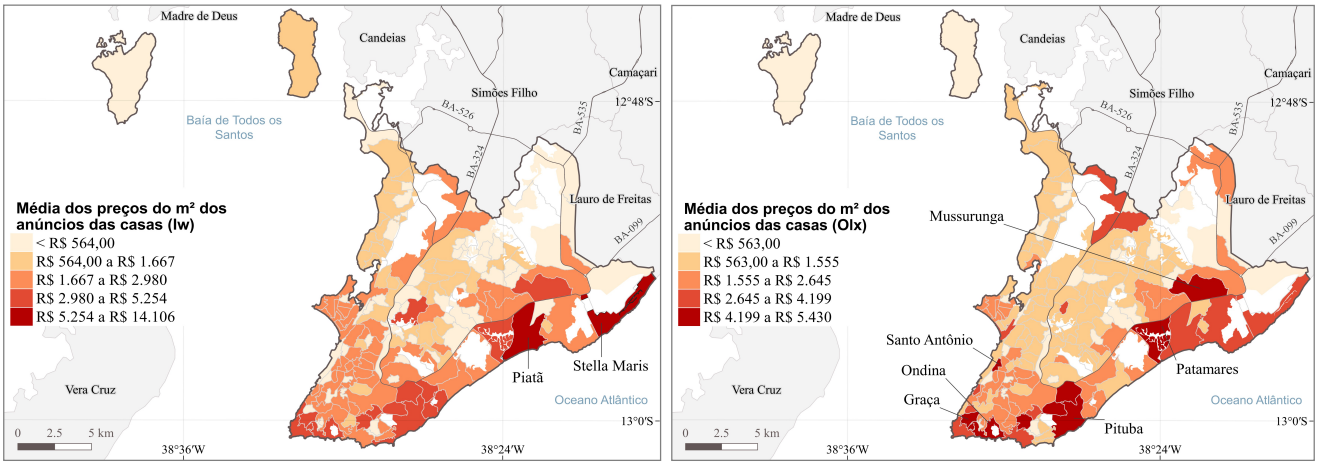
Este desenho apresentou exceção das áreas do subúrbio da cidade, onde, segundo Carvalho e Pereira (2008), há um histórico de menor valorização da terra, e onde foram vistos os menores volumes dos anúncios e preços médios do m². Segundo os autores, trata-se, então, de uma fragmentação socioespacial dos tipos de ocupação formal e informal entrelaçados na superfície no município, onde as ocupações informais de habitações populares na zona sul de Salvador, e na zona da orla do oceano Atlântico são consolidadas, segmentadas e menores, já as ocupações formais, estão localizadas majoritariamente na extensão da orla oceânica.

Figura 2 - Distribuição espacial das médias dos preços do m² dos anúncios dos apartamentos por bairro.



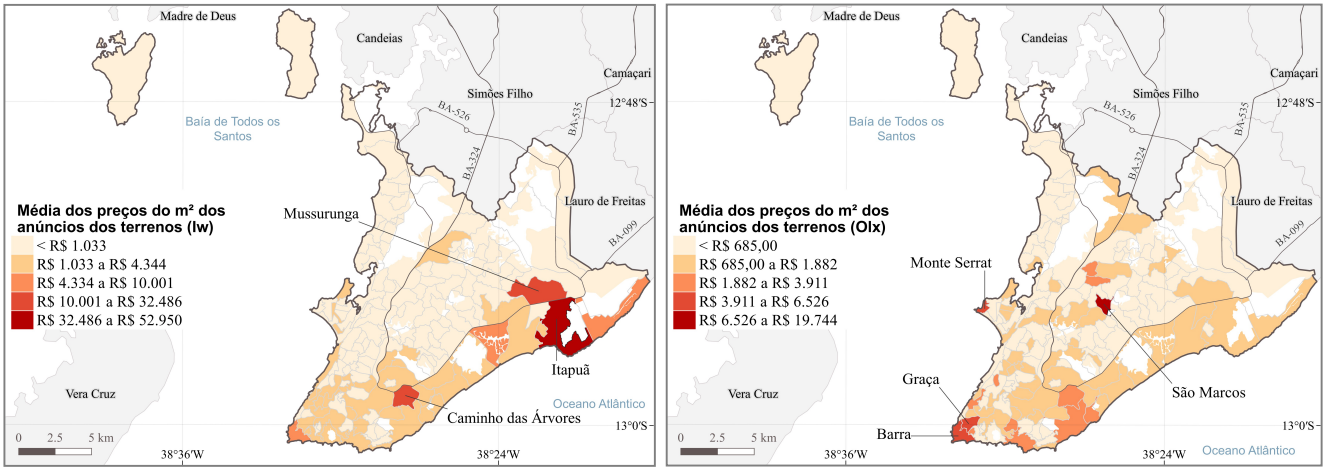
Fonte - Autores (2022) a partir dos dados depurados.

Figura 3 - Distribuição espacial das médias dos preços do m² dos anúncios das casas por bairro.



Fonte - Autores (2023) a partir dos dados depurados.

Figura 4 - Distribuição espacial das médias dos preços do m² dos anúncios dos terrenos por bairro.



Fonte - Autores (2023) a partir dos dados depurados.

CONSIDERAÇÕES FINAIS

O presente estudo pretendeu contribuir com o desenvolvimento de uma abordagem metodológica empregando critérios para depuração e formação de um conjunto de dados com maior nível de completude sobre os preços das vendas dos imóveis residenciais urbanos obtidos por meio da técnica de *web scraping* dos anúncios. Em conjunto, realizou-se discussões sobre a importância da utilização destes dados presentes na internet para os estudos urbanos.

Quanto ao mapeamento da distribuição desses preços médios dos imóveis de Salvador, notou-se um padrão heterogêneo por unidade espacial no espaço urbano. Ainda que a unidade de análise espacial por bairro tenha inconsistências, haja vista por sua homogeneização, concluiu-se que a segmentação dos preços do mercado imobiliário por bairro foi significativa no estudo dos preços da habitação. As análises dos achados da pesquisa apontaram para uma distribuição heterogênea dos preços dos imóveis verificados na cidade, com diferenças em quantidades e preços entre os territórios (bairros) populares e os nobres.

A metodologia utilizada no trabalho pode ser aplicada para demais grandes centros urbanos, uma vez que seja possível a recuperação dos dados presentes nos anúncios online do mercado imobiliário. A interlocução entre aplicação do conhecimento científico-acadêmico sobre o território, a utilização de

técnicas contemporâneas de aquisição de dados e todo o conhecimento adquirido neste trabalho, quando divulgado a sociedade chama atenção as desigualdades urbanas e ao direito à cidadania consciente, que em muitas vezes é relegado em detrimento ao perfil capitalista do direito ao consumidor.

Outros aspectos relevantes a serem apontados nesta pesquisa são sobre as complexidades empíricas devido a aquisição de extenso conjunto de dados dos anúncios online, pensada enquanto base observatória de preços e da obtenção de dados públicos com agregação espacial específica. Embora a limitação dos dados obtidos através da técnica de coleta por *web scraping* em sites específicos possam enviesar a análise sobre o território pelas distinções aqui tratadas, ainda assim foram empregados critérios de exclusão de inconsistências inerentes a este conjunto de dados para minimizar essas limitações. Por fim, observou-se que a base do Olx apresentou menor completude, menor volume se comparado ao Imovelweb, porém maior variedade referente a cobertura espacial dos imóveis por bairro.

As inconsistências mensuradas foram realizadas sobre os dados tabulares resultantes das extrações mostrando ainda, limitações quanto à comparação do processamento da requisição resposta do site. Pontua-se que não foram monitoradas em sua totalidade se os campos em branco ou com deslocamentos de dados foram devido ao preenchimento incorreto por parte do anunciante ou se ocorreram devido a perda de dados no processo de extração, sobre o tempo de resposta do site, capacidade de processamento computacional.

Cabe salientar ainda, a necessidade emergencial das políticas públicas municipais utilizarem ferramentas modernas não somente de obtenção, como de tratamento e divulgação dos dados sobre o mercado imobiliário na cidade. Em tempo presente, de alta competitividade e tecnologias, torna-se fundamental a comunicação à sociedade, a fim de garantir o desenvolvimento gradativo do conhecimento local quanto a importância da justiça espacial de precificação.

A aquisição e atualização da base (sobretudo, quanto à variável preço) exigem esforços empíricos para que se busquem novos resultados e evidências científicas sobre o espaço urbano. O conjunto de dados mostrou um pouco das dinâmicas mercadológicas que, entre continuidades e descontinuidades, desenha o atual mapa das distribuições dos preços em escala intraurbana da cidade. Assim, o método aplicado forneceu uma visão coexistente sobre apropriação espacial dos bairros em discordantes precificação e criações de solo. Isto leva à multiplicação das desigualdades socioespaciais e à quebra do direito igual à moradia, as infraestruturas mínimas das populações da grande cidade.

Sugere-se que pesquisas futuras possam replicar os procedimentos deste trabalho, com o objetivo de mapear a distribuição dos preços dos imóveis e dos padrões presentes, relacionar e discutir as segregações urbanas encontradas e fornecer um apanhado temporal dos dados sobre o espaço urbano, como uma forma de responder ao questionamento, quanto à falta de inserção histórica nas análises atuais sobre segregação urbana.

AGRADECIMENTOS

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro na modalidade de bolsa de mestrado, ao incentivo à pesquisa em políticas fundiárias e desenvolvimento urbano do Programa para América Latina e Caribe (*Latin America Program Graduate Student Fellowship Support*) do *Lincoln Institute Land of Policy*. Da mesma forma, agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) através do apoio financeiro na modalidade de bolsa de iniciação científica.

REFERÊNCIAS

ABRAMO, P. A dinâmica imobiliária. Elementos para o entendimento da espacialidade urbana. **Cadernos IPPUR**, Ano III, Nº especial, 1989.

ABRAMO, P. **Favela e mercado informal:** a nova porta de entrada dos pobres nas cidades brasileiras. Org. Pedro Abramo, Porto Alegre: ANTAC, Coleção Habitare, v. 10, 2009.

BATTY, M. Big data, smart cities and city planning. **Dialogues in human geography**, v. 3, n. 3, p. 274-279, 2013. https://doi.org/10.1177/2043820613513390

BOEING, G; WADDELL, P. New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. **Journal of Planning Education and Research**, v. 37, n. 4. 2016.

BRASIL. **Lei Federal nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018. Disponível em https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 23 jun. 2022.

BRAVO, J. V. M; SLUTER, C. R. O problema da qualidade de dados espaciais na era das informações geográficas voluntárias. **Boletim de Ciências Geodésicas**, v. 21, p. 56-73, 2015.

BRICONGNE, J. C; MEUNIER, B; SYLVAIN, P. Web scraping housing prices in real-time: the covid-19. **Crisis in the UK**. Banque de France Working Paper, n. 827, 2021. https://doi.org/10.2139/ssrn.3916196

CARLOS, A. F. A. A virada espacial. **Mercator**, v. 14, p. 7-16, 2015.

CARLOS, A. F. A. Em nome da cidade (e da propriedade). COLOQUIO INTERNACIONAL DE GEOCRÍTICA: Las utopías y la construcción de la sociedad del futuro, 14., Barcelona, 2016. **Actas...** Barcelona: Universidad de Barcelona, v. 27, 2016.

CARVALHO, I. M. D.; PEREIRA, G. C. As "cidades" de Salvador. In: CARVALHO, I. M. M. D.; PEREIRA, G. C. (Org.). **Como anda Salvador e sua região metropolitana**. 1ª edição, Salvador: EDUFBA, 2008. https://doi.org/10.7476/9788523209094

CHAPELLE, G, EYMÉOUD, J. B. Can big data increase our knowledge of local rental markets? A dataset on the rental sector in France. **PloS one**, v. 17, n. 1, p. 1-21, 2022. https://doi.org/10.1371/journal.pone.0260405

CORRÊA, R. L. **O espaço urbano.** São Paulo: Editora Ática, 4ª edição, 2000.

DATA MINER**. Data Miner software**. Disponível em: https://data-miner.io/. Acesso em: 7 nov. 2022.

DAVIS, C. A. Challenges in crowdsourcing geospatial data to replace or enhance Official Sources. **Disegnarecon**, v. 11, n. 20, p. 1-1-1.15, 2018.

FATMASARI; KUNANG, Y. N.; PURNAMASARI, S. D. Web scraping techniques to collect weather data in South Sumatera. INTERNATIONAL CONFERENCE ON ELECTRICAL ENGINEERING AND COMPUTER SCIENCE (ICECOS), 2018, Pangkal, Indonesia. **Proceedings...** IEEE, 2018. p. 385-390. https://doi.org/10.1109/ICECOS.2018.8605202

FERNANDES, V. de O.; ELIAS, N. E.; ZIPF, A. Integration of authoritative and volunteered geographic information for updating urban mapping: challenges and potentials. **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, Volume XLIII-B4-2020, 2020 XXIV ISPRS Congress. 2020. https://doi.org/10.5194/isprs-archives-XLIII-B4-2020-261-2020

GOODCHILD, M, F. Citizens as Sensors: the world of volunteer geography. **GeoJournal**, v. 69, n. 4, p. 211-221, 2007. https://doi.org/10.1080/17538941003759255

GOODCHILD, M; GLENNON, A. Crowdsourcing geographic information for disaster response: a research frontier. **International Journal of Digital Earth**, v.3. p. 231-241, 2010. https://doi.org/10.1007/s10708-007-9111-y

GOODCHILD, M; Li, L. Assuring the quality of volunteered geographic information. **Spatial Statistics**, v. 1, p.110-120, 2012.

GOODCHILD, M. F. Introduction to Urban Big Data Infrastructure. **Urban Informatics**. Springer, Singapore, p. 543-545, 2021.

GRIFFIN, A. L; ROBINSON, A. C.; ROTH, R. E. Envisioning the future of cartographic research. **International Journal of Cartography**, v. 3, n. sup1, p. 1-8, 2017.

GRYBAUSKAS, A; PILINKIENĖ, V; STUNDŽIENĖ, A. Predictive analytics using big data for the real estate market during the covid-19 pandemic. **Journal of big data**, v. 8, n. 1, p. 1-20, 2021. https://doi.org/10.1186/s40537-021-00476-0

KITCHIN, R. Big data, new epistemologies, and paradigm shifts. **Big Data e Society**, v. 1, n. 1, p. 1-12, 2014.

LIMA, J. C. D. S; SILVA, A. R. da. O Decreto n. 10.046 de 2019 frente à legislação brasileira de proteção de dados. **Revista Cientifica Multidisciplinar Núcleo do Conhecimento**. Ano 06, ed. 07, v. 03, p. 21-39, jul. 2021.

LOBERTO, M.; LUCIANI, A.; PANGALLO, M. The potential of big housing data: An application to the Italian real-estate market. **Banca d'Italia Working Papers**, n. 1171, 2018.

MARICATO, E. Direito à terra ou direito à cidade. **Revista de Cultura Vozes**, v. 89, n. 6, 1985.

MARTINS JÚNIOR, O. G; SILVA, L, F. C. F. D. Proposta de Hierarquia para Conceitos de Cartografia Colaborativa. **Anuário do Instituto de Geociência** – UFRJ. Vol. 41, 3. ed., p. 560-567, 2018.

NEDER, H. D.; SANTOS, J. F. C.; DA SILVA, G. J. C.; PIORSKI, C. R. L. Índice de defasagem do Imposto Predial e Territorial Urbano (IPTU) dos municípios de Minas Gerais: um estudo de caso para Uberlândia (MG). Brasil. **Revista Espacios**, v.38, n. 46, p. 25-39, 2017.

POONGODAI, A.; SUHASINI, R. A command line tool for tracking error details of program using web scraper. **International Journal of Recent Technology and Engineering (IJRTE)**, v. 8, 2019. https://doi.org/10.35940/ijrte.B1276.0982S1119

QGIS. **QGIS Geographic Information System**. Versão 3.10. Open Source Geospatial Foundation Project. Disponível em: http://qgis.osgeo.org. Acesso em: maio 2022.

RAMOS, F. R. **Análise espacial de estruturas intra-urbanas:** o caso de São Paulo. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos: INPE, 2002.

ROBINSON, A. C; DEMŠAR, U; MOORE, A. B; BUCKLEY, A; JIANG, B; FIELD, K.; KRAAK, M.; CAMBOIM, S; P; SLUTER, C. R. Geospatial big data and cartography: research challenges and opportunities for making maps that matter. **International Journal of Cartography**, v. 3, n. sup1, p. 32-60, 2017.

ROLNIK, R. **Guerra dos lugares:** a colonização da terra e da moradia na era das finanças. 1. ed. São Paulo: Boitempo, 2015.

SALVADOR. Secretaria Municipal da Fazenda (SEFAZ). **Mapeamento cartográfico de Salvador**. 2020. Disponível em: http://cartografia.salvador.ba.gov.br/. Acesso em: maio 2022.

SALVADOR. Secretaria Municipal da Fazenda (SEFAZ). Coordenadoria de Cadastros (CCD). **Cadastro Imobiliário Municipal** (CIM). 2020.

SPOSITO, E. S; SPOSITO, M. E. B. Fragmentação Socioespacial. **Mercator**, Fortaleza, v.19, 2020. DOI: https://doi.org/10.4215/rm2020.e19015. https://doi.org/10.4215/rm2020.e19015

SPOSITO, M. E; GÓES, E. **Espaços fechados e cidades:** Insegurança urbana e fragmentação socioespacial. São Paulo: Editora Unesp, 2016.

TOMAL, M. Modelling housing rents using spatial autoregressive geographically weighted regression: a case study in Cracow, Poland. **ISPRS International Journal of Geo-Information**, v. 9, n. 6, p. 346, 2020. https://doi.org/10.3390/ijgi9060346

WENCESLAU, R; DAVIS JUNIOR, C. A; SMARZARO, R. Challenges for matching spatial data on economic activities from official and alternative sources. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA (GEOINFO), 18., 2017, Salvador. **Anais...** Salvador, 2017.

ZHAO, B. Web scraping. Encyclopedia of Big Data. **Springer International Publishing**, 2017. College of Earth, Ocean, and Atmospheric Sciences, University Corvallis, Oregon State, USA, 2017. https://doi.org/10.1007/978-3-319-32001-4_483-1

Recebido em: 17/02/2023
Aceito para publicação em: 11/09/2023