Vazios de Informação & Informalidade: Web-Scraping para Identificação de Assentamentos Precários

Ana Carolina C. Carneiro[1], Flávia da Fonseca Feitosa[1], Luis Felipe B. da Cunha [1]

[1]Laboratório de Estudos e Projetos Urbanos e Regionais (LEPUR) - Programa de Pós-Graduação em Planejamento e Gestão Territorial da Universidade Federal do ABC (UFABC) – São Bernardo do Campo, SP - Brasil

carolina.carneiro@ufabc.edu.br, flavia.feitosa@ufabc.edu.br, luis.cunha@ufabc.edu.br

Abstract. *This article presents an exploratory study to test the use of data from the construction of alternative web models for the identification of precarious settlements. The informal character of real estate transactions in AP, in which AP properties predominate, to the low density and absence of specialized property advertisements on the web, associated with the presence of AP. Thus, such "information gaps" represent information relevant to the identification of AP. The methodology used consists of mining 21,067 real estate advertisements in the city of São Bernardo do Campo (SP) on the ImóvelWeb portal. From the data, logistic models were built to estimate the probability of PA presence. The tests carried out confirmed that, as dynamic, the formal real estate markets were formed in a localized way in the territory and mostly in the perimeters of the AP, confirming an initial hypothesis*

Resumo. *Este artigo apresenta um estudo exploratório para testar a utilização de dados alternativos, provenientes de web-scraping na construção de modelos voltados para a identificação de assentamentos precários (AP). Dado o caráter informal das transações imobiliárias em AP, nos quais predominam situações de insegurança da posse, partiu-se da hipótese de que áreas caracterizadas por baixa densidade e até mesmo ausência de anúncios de imóveis em portais web especializados, estão associadas à presença de AP. Assim, tais "vazios de informação" representam, na verdade, informação relevante para a identificação de AP. A metodologia utilizada consistiu na mineração de 21.067 anúncios imobiliários da cidade de São Bernardo do Campo (SP) no portal ImóvelWeb. A partir desses dados, foram construídos modelos logísticos para estimar a probabilidade de presença de AP. Os testes realizados confirmaram que as dinâmicas do mercado imobiliários formal se dão de forma concentrada no território e majoritariamente fora dos perímetros dos AP, confirmando a hipótese inicial do trabalho.*

1.Introdução

Os assentamentos precários (AP) diferem ao longo de todo o território nacional no que se refere à tipologia, métodos construtivos, localização, cultura e nível de precariedade. Este fato dificulta sua identificação e caracterização com base em um único método para todo o país, o que impacta o desenvolvimento de políticas públicas voltadas para a regularização fundiária e programas habitacionais (DENALDI et al., 2020). Em que pese a relevância dos problemas habitacionais brasileiros, o país apresenta escassez de

informação sobre seus AP, o que vem mobilizando o desenvolvimento de distintas metodologias para sua identificação e caracterização – (MARQUES et al., 2007); (DENALDI; FEITOSA, 2016); (FEITOSA et al., 2019); (FEITOSA et al., no prelo). Em âmbito internacional, os AP também mobilizam esforços de identificação, com destaque para a pesquisa de (MAHABIR et al., 2020) de Detecção e Mapeamento de favelas utilizando dados abertos no Quênia, que buscou identificar fontes de dados abertos alternativos, incluindo a mineração de dados imobiliários, que poderiam ser associados a dados de sensoriamento remoto para identificar favelas com carência de dados.

O presente trabalho busca verificar se dados imobiliários extraídos de portais web podem ser úteis no contexto brasileiro, contribuindo para o desenvolvimento das metodologias existentes. Com o alcance cada vez maior da internet, aumenta-se o volume, variedade e velocidade de dados gerados a todo instante, os quais podem subsidiar inúmeras aplicações. Cabe salientar, entretanto, que estes dados – muitos deles denominados "big data" – apresentam limitações, entre as quais destaca-se o fato de estarem restritos a usuários e mercados (SHEARMUR, 2015). Ao tratar do uso destes dados em estudos urbanos, Feitosa (2020) ressalta que tal característica pode restringir nossa visão de "urbano", assim como de sociedade. A autora afirma que estes dados são tendenciosos por natureza e ignoram pessoas que estão alheias aos mercados e atividades específicas que estão sendo rastreadas; excluem boa parte da sociedade e contribuem para deixar invisível aqueles que mais precisam de visibilidade (Feitosa, 2020, p. 4). Ao mesmo tempo, a vasta gama de dados geoespaciais disponíveis, criam um potencial para geovisualização capaz de apoiar o planejamento urbano voltado para esta parcela invisível da sociedade (MACEACHREN, 2004)

Buscando dar outro sentido para estas considerações, o presente trabalho, parte da hipótese de que a baixa concentração ou até mesmo a ausência de dados pode representar informação relevante para a identificação de áreas precárias. Para tanto, apresenta um estudo exploratório similar ao realizado por (MAHABIR et al., 2020) para a cidade de São Bernardo do Campo. O estudo testa a utilização de dados imobiliários extraídos de portais web de imóveis (*web scraping*[1]) na construção de modelos para identificar AP. Acredita-se que, ao construir uma representação da 'formalidade urbana', constituída pela concentração de imóveis anunciados em portais web por imobiliárias, seja possível identificar vazios de informação que caracterizam a informalidade. Busca-se, assim, verificar se tais vazios representam um aspecto territorial relevante para estudos sobre a precariedade habitacional.

2.Método

O estudo consistiu na construção de modelos de regressão logística para a identificação de AP a partir da densidade de endereços de imóveis anunciados na internet. A regressão logística é um modelo estatístico formulado para prever e explicar uma variável categórica binária (dependente) a partir da sua relação com variáveis independentes.

[1] Web scraping é a coleta automatizada de dados da internet, também conhecida como raspagem de dados ou mineração de dados em português (MITCHELL, 2018).

As etapas do estudo foram: 1) coleta de dados; 2) tratamento e integração dos dados; 3) criação de superfícies de densidade Kernel; 4) construção de modelos de regressão logística; 5) Mapeamento da superfície de probabilidade[2].

A coleta de dados primários foi realizada através da mineração de dados de anúncios imobiliários do portal ImóvelWeb ("imovelweb", 2022), entre os meses de julho e agosto de 2022, com o auxílio do software Power Automate. Os anúncios coletados foram tratados, filtrados e geocodificados com o auxílio da Interface de Programação de Aplicação (API) do Google Planilhas. Em seguida, foram estimadas duas superfícies de densidade kernel a partir destes dados, ambas com largura de banda de 150, sendo uma delas simples (Dens_Simples) e a outra ponderada pelo valor do imóvel anunciado (Dens_Pond). Estas superfícies de densidade de anúncios de imóveis serviram como variáveis independentes dos modelos logísticos.

Para representar a presença de AP (variável dependente dos modelos), considerou-se o mapeamento realizado pelo Diagnóstico Habitacional da Região do Grande ABC – DHABC (DENALDI; FEITOSA, 2016). Este levantamento identifica quatro tipologias de AP:

- T1: assentamentos urbanizados, consolidados e irregulares, que demandam ações de regularização fundiária;
- T2: AP, irregulares e consolidáveis, que demandam obras de infraestrutura, podendo necessitar de alguma remoção;
- T3: AP, irregulares e consolidáveis, que demandam obras complexas de urbanização e/ou percentual elevado de remoção;
- T4: AP, irregulares e não consolidáveis, cuja solução é a remoção total dos domicílios.

A partir dos kernels, foram construídos quatro modelos de regressão logística com o auxílio do Software R para estimar a probabilidade da presença de AP a partir da densidade de anúncios imobiliários, considerando-se a densidade de ocupação das células (Quadro 1).

Para construção dos modelos, considerou-se duas alternativas de variável dependente: a primeira inclui apenas as tipologias T2, T3, T4 (AP_234), ao passo que a outra inclui todas as tipologias, inclusive a T1, que abarca assentamentos já urbanizados, que aguardam apenas regularização (AP_1234). As variáveis geradas foram integradas em uma base celular vetorial com resolução de 100m e utilizadas para a construção de 4 modelos de regressão logística, conforma apresentado no Quadro 1.

Quadro 1 – Variáveis consideradas nos modelos de regressão logística

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Variável Dependente (Y)	AP_1234	AP_234	AP_1234	AP_234
Variável Independente (X)	Dens_Simples	Dens_Simples	Dens_Pond	Dens_Pond

[2] Superfícies de probabilidade são obtidas pela regressão logística e apresentam uma visão mais informativa do que que apenas o mapa de classificação (sim ou não), obtido pela análise discriminante (FEITOSA et al., 2019).

Os modelos de regressão logística foram construídos para estimar a probabilidade de uma área (célula) ser (1) ou não (0) um AP a partir de sua relação a densidade de imóveis anunciados em portais web. Para cada modelo de regressão logística foram geradas e mapeadas superfícies de probabilidade.

3.Resultados

A mineração de dados obteve um retorno de 58.007 anúncios, dos quais 21.067 puderam ser geocodificados, por possuírem endereços completos. Os anúncios válidos e geocodificados são apresentados na Figura 1. A partir destes dados gerou-se a estimativa de densidade Kernel apresentada na Figura 2, na qual se encontram sobrepostos os assentamentos precários de São Bernardo do Campo (DENALDI; FEITOSA, 2016). Os resultados apresentados na Figura 3 apontam uma concentração dos anúncios imobiliários em áreas sem assentamentos precários.

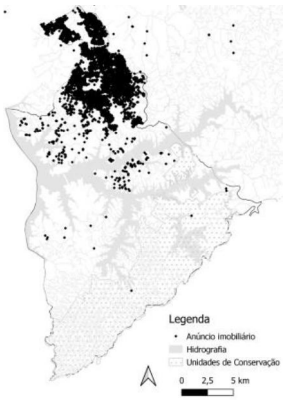

Figura 1- Mapeamento dos anúncios imobiliários coletados para São Bernardo do Campo

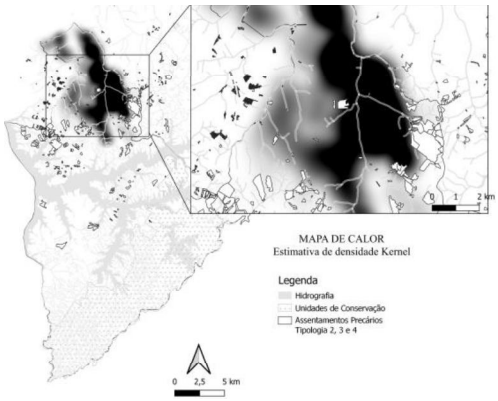

Figura 2 – Estimativa de densidade Kernel

Os resultados dos quatro modelos logísticos estimados são apresentados na Tabela .

Tabela 1 - Coeficientes Estimados

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Constante	-1,566 ***	-2,058 ***	-1,563 ***	-2,057 ***
Dens_Simples	-0,044 ***	-0,056 ***	-	-
Dens_Pond	-	-	$-6{,}546.10^{-8}$ ***	$-7{,}852.10^{-8}$ ***
Erro Padrão (EP)	0,016	0,019	$1{,}605.10^{-2}$	$1{,}922.10^{-2}$
AIC	26552	20176	26515	20158

*** *indica nível de significância estatística de 0,001*

Em todos os modelos, os coeficientes estimados para a variável independente (Dens_Simples ou Den_Ponderada) foram negativos e significativos, o que corrobora a

hipótese inicial de que os dados de anúncios de imóveis estão negativamente correlacionados com a presença de assentamentos precários.

O Akaike Information Criterion (AIC) fornece um método para avaliar a qualidade do modelo por meio de comparação com outros relacionados, sendo o mais indicado aquele que apresentar o menor coeficiente. No caso deste estudo, os modelos que apresentaram os melhores resultados foram os que desconsideraram a tipologia 1 de AP na análise (Modelos 2 e 4), restringindo-se aos assentamentos mais precários. Entre os dois modelos, o Modelo 4, que considera a superfície de densidade kernel ponderada pelos preços dos imóveis (Dens_Pond) apresentou resultados melhores.

A Figura 4 apresenta a superfície de probabilidade de presença de AP gerada a partir dos resultados do Modelo 4. No mapa da Figura 3, as células mais escuras estão mais próximas de 1 e apresentam maiores probabilidades de serem AP, ao passo que as células mais claras apresentam menores probabilidades. Uma análise qualitativa foi conduzida a partir da sobreposição dos polígonos dos AP (em vermelho), indicando a concentração de AP em áreas escuras, ou seja, apontadas pelo modelo como de baixa probabilidade.

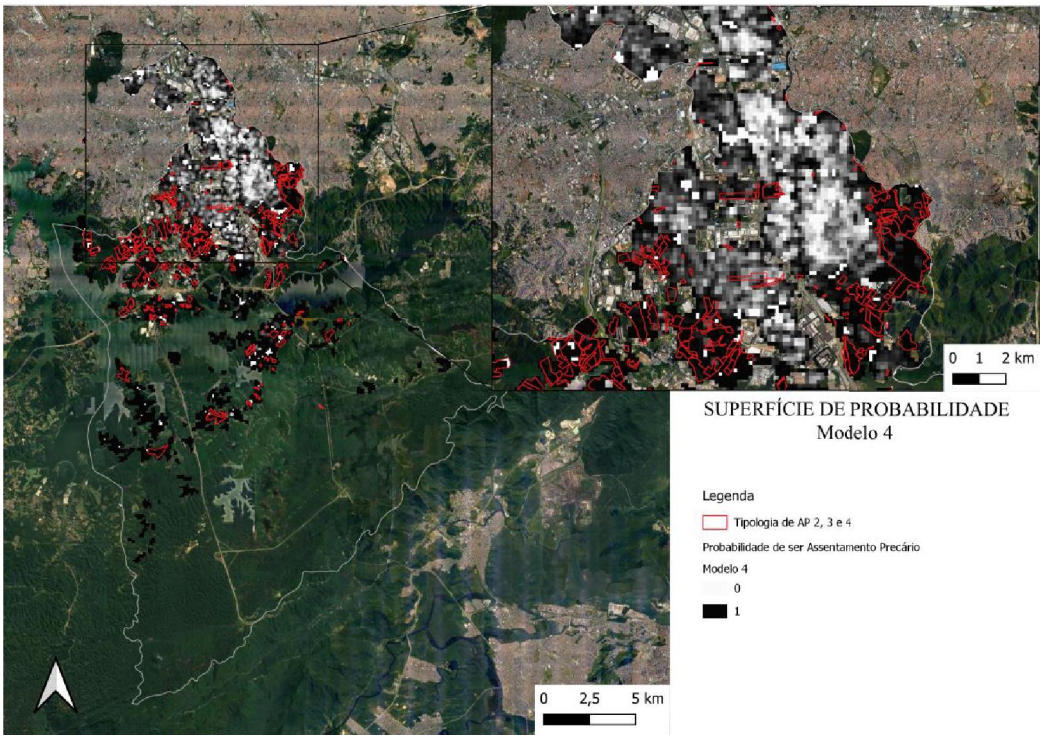

Figura 3: Mapa de superfície de probabilidade para o modelo 4

4. Considerações Finais

A identificação de assentamentos precários é uma tarefa essencial na elaboração de programas habitacionais, definição de prioridades e adequação de financiamento à distintas realidades. Apesar de seu carácter essencial, às metodologias de identificação de AP são fortemente impactadas pelas inúmeras limitações existentes na produção de informações e até mesmo da escassez de informações (FEITOSA et al., 2019). Dentro

deste cenário, este estudo se mostrou eficaz ao trazer a possibilidade de uma nova variável que possa auxiliar na identificação de assentamentos precários brasileiros. Devido ao aumento exponencial da produção de dados e o seu carácter desigual dentro do território, verifica-se que é possível extrair informações importantes através da espacialização destes dados, incluindo, os vazios informacionais gerados.

Ao analisar a relação entre a densidade de imóveis anunciados em portais web e a presença de assentamentos precários, verifica-se que a movimentação do mercado formal em ambiente online é concentrada no tecido urbano e não engloba grandes porções da periferia da cidade, onde, geralmente, encontram-se os assentamentos precários. Os resultados encontrados, corroboram a hipótese inicial.

Pelo caráter exploratório do estudo, entende-se que há uma vasta possibilidade de aperfeiçoamento da abordagem proposta, como o aumento do tempo de coleta de dados de modo a capturar as dinâmicas espaço-temporais do mercado imobiliário formal online e a elaboração de modelos utilizando outras variáveis, como renda e uso e ocupação do solo.

Referências

DENALDI, R. et al. Produto 1 – Estudo conceitual e metodológico: Pesquisa de Núcleos Urbanos Informais no Brasil. [s.l.] SNH/MDR e IPEA, 6 abr. 2020.

DENALDI, R.; FEITOSA, F. Diagnóstico Habitacional Regional do Grande ABC.pdf. Santo André: Consórcio Intermunicipal Grande ABC; Universidade Federal do ABC, set. 2016.

FEITOSA, F. et al. MAPPA: Metodologia para Identificação e Caracterização de Assentamentos Precários em Regiões Metropolitanas Paulistas. ABC: UFABC, 2019.

FEITOSA, F. et al. Modelagem para a Identificação de Núcleos Urbanos Informais: Uma Proposta Metodológica. Em: Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional. Brasília: [s.n.].

imovelweb. Portal Imobiliário. Disponível em: <https://www.imovelweb.com.br/>.

MACEACHREN, A. et al. Geovisualization for Knowledge Construction and Decision Support. IEEE Computer Graphics and Applications, 2004

MAHABIR, R. et al. Detecting and mapping slums using open data: a case study in Kenya. International Journal of Digital Earth, v. 13, n. 6, p. 683–707, 2 jun. 2020.

MARQUES, E. et al. (EDS.). Assentamentos precários no Brasil urbano. São Paulo, Brazil] : [Brasília, Brazil: Centro de Estudos da Metrópole / CEBRAP ; Secretaria Nacional de Habitação / Ministério das Cidades, 2007.

MITCHELL, R. E. Web scraping with Python: collecting more data from the modern web. Second edition ed. Sebastopol, CA: O'Reilly Media, 2018.

SHEARMUR, R. Dazzled by data: Big Data, the census and urban geography. Urban Geography, v. 36, n. 7, p. 965–968, 3 out. 2015.