ipea Instituto de Pesquisa Econômica Aplicada

| | |
|---|---|
| **Título do capítulo** | CAPÍTULO 4 – MODELAGEM PARA A IDENTIFICAÇÃO DE NÚCLEOS URBANOS INFORMAIS: UMA PROPOSTA METODOLÓGICA |
| **Autores(as)** | Flávia da Fonseca Feitosa<br>Luis Felipe Bortolatto da Cunha<br>Gilmara da Silva Gonçalves<br>Guilherme Frizzi Galdino da Silva<br>Pedro Reis Simões |
| **DOI** | DOI: http://dx.doi.org/10.38116/978-65-5635-044-8/capitulo4 |

| | |
|---|---|
| **Título do livro** | NÚCLEOS URBANOS INFORMAIS: ABORDAGENS TERRITORIAIS DA IRREGULARIDADE FUNDIÁRIA E DA PRECARIEDADE HABITACIONAL |
| **Organizadores(as)** | CLEANDRO KRAUSE<br>ROSANA DENALDI |
| **Volume** | - |
| **Série** | - |
| **Cidade** | Brasília |
| **Editora** | Instituto de Pesquisa Econômica Aplicada (Ipea) |
| **Ano** | 2022 |
| **Edição** | 1ª |
| **ISBN** | 978-65-5635-044-8 |
| **DOI** | DOI: http://dx.doi.org/10.38116/978-65-5635-044-8 |

CAPÍTULO 4

# MODELAGEM PARA A IDENTIFICAÇÃO DE NÚCLEOS URBANOS INFORMAIS: UMA PROPOSTA METODOLÓGICA

Flávia da Fonseca Feitosa[1]
Luis Felipe Bortolatto da Cunha[2]
Gilmara da Silva Gonçalves[3]
Guilherme Frizzi Galdino da Silva[4]
Pedro Reis Simões[5]

## 1 INTRODUÇÃO

As cidades brasileiras caracterizam-se pela exclusão de boa parte de sua população do mercado formal de provisão habitacional. A irregularidade fundiária, que não está necessariamente associada à pobreza (Amore e Moretti, 2018), apresenta seus efeitos mais perversos sobre os mais pobres e vulneráveis, submetidos a situações de precariedade urbana e habitacional. Para esses, a alternativa habitacional possível envolve ocupações irregulares do ponto de vista fundiário, edilício e urbanístico, realizadas em desacordo com a legislação, sujeitas a remoções, sem investimentos em infraestrutura e serviços e, frequentemente, em áreas ambientalmente sensíveis (Maricato, 2009).

A Lei Federal nº 13.465, sancionada em junho de 2017, representa o marco legal da regularização fundiária no Brasil e busca ampliar a ação sobre as informalidades urbanas em suas diversas formas, viabilizando ações de reconhecimento de posse e propriedade. A lei define o termo "núcleo urbano informal" como "aquele clandestino, irregular ou no qual não foi possível realizar, por qualquer modo, a titulação de seus ocupantes, ainda que atendida a legislação vigente à época de sua implantação ou regularização" (Brasil, 2017). Entre esses núcleos, prevalecem aqueles ocupados predominantemente por população de baixa renda, que apresentam uma enorme variedade de padrões e condições de ocupação, associados a distintas

1. Professora do Programa de Pós-Graduação em Planejamento e Gestão do Território da Universidade Federal do ABC (PPGPGT/UFABC). *E-mail*: <flavia.feitosa@ufabc.edu.br>.

2. Pesquisador associado ao Laboratório de Estudos e Projetos Urbanos e Regionais (Lepur) da UFABC. *E-mail*: <luis.cunha@ufabc.edu.br>.

3. Pesquisadora associada ao Lepur da UFABC. *E-mail*: <gilmara.goncalves94@gmail.com>.

4. Pesquisador associado ao Lepur da UFABC. *E-mail*: <guilhermefrizzi@gmail.com>.

5. Pesquisador do Subprograma de Pesquisa para o Desenvolvimento Nacional (PNPD) na Diretoria de Estudos e Políticas Regionais, Urbanas e Ambientais do Instituto de Pesquisa Econômica Aplicada (Dirur/Ipea). *E-mail*: <pedro.simoes@ipea.gov.br>.

dimensões e níveis de precariedade, e que podem ou não ser passíveis de consolidação. Identificar tais núcleos, caracterizá-los e reconhecer a multidimensionalidade e diversidade de condições de precariedade e irregularidade é, portanto, questão crucial para a elaboração de programas e estratégias de urbanização e regularização fundiária que promovam condições dignas de moradia.

Embora a informalidade e a precariedade atinjam uma notável parcela da população brasileira e estejam posicionadas como tema de reconhecida relevância nos âmbitos acadêmico e das políticas públicas, a disponibilidade e qualidade das informações sobre núcleos urbanos informais (NUIs) precários ainda é muito limitada. Com o intuito de contribuir nessa direção, este capítulo apresenta uma metodologia que busca auxiliar na identificação de NUIs precários, por meio da construção de modelos que integrem dados secundários de naturezas distintas, para a geração de superfícies de probabilidade relacionadas à presença desses núcleos. Essas superfícies de probabilidade podem ser utilizadas como um plano de informação para subsidiar trabalhos de campo e/ou para análise da qualidade de bases de dados existentes (tais como os aglomerados subnormais – AGSNs, do Instituto Brasileiro de Geografia e Estatística – IBGE ou informações fornecidas por prefeituras).

A metodologia proposta, denominada metodologia NUI, é fruto de uma das frentes de trabalho da Pesquisa de Núcleos Urbanos Informais no Brasil, realizada por meio de cooperação técnico-científica entre o Ipea e a Secretaria Nacional de Habitação (SNH) do Ministério do Desenvolvimento Regional (MDR), e que envolve a identificação e caracterização de núcleos urbanos informais localizados em seis regiões do país, denominadas polos: Brasília-DF, Belo Horizonte-MG, Recife-PE, Porto Alegre-RS, Marabá-PA e Juazeiro do Norte-CE. O desenvolvimento da metodologia NUI partiu da análise de metodologias relacionadas, voltadas à identificação de assentamentos precários no contexto brasileiro, que serão brevemente apresentadas na seção 2. A seção 3 apresenta uma visão geral da metodologia NUI e suas etapas, ao passo que a seção 4 traz os resultados de sua aplicação para os seis polos da pesquisa. A seção 5 apresenta as considerações finais do capítulo.

## 2 METODOLOGIAS RELACIONADAS

Esta seção apresenta uma breve descrição e comparação de duas metodologias que utilizam modelos estatísticos para a identificação de assentamentos precários, as quais serviram de referência para o desenvolvimento da metodologia NUI: a do Centro de Estudos da Metrópole – CEM (Marques, 2007; 2013) e a metodologia para identificação e caracterização de assentamentos precários em regiões metropolitanas paulistas – Mappa (CDHU e UFABC, 2019; Feitosa *et al.*, 2019).

Em 2007, o CEM desenvolveu e divulgou, a pedido da SNH/Ministério das Cidades, uma metodologia pioneira para auxiliar na identificação de assentamentos precários. A partir dos dados dos setores subnormais[6] divulgados pelo IBGE, a metodologia baseia-se na utilização da técnica de análise discriminante, para a construção de modelos que identifiquem setores censitários com características populacionais semelhantes às dos setores subnormais, mas que não haviam sido classificados pelo IBGE como tais (Marques, 2007). Parte-se do pressuposto de que "as características sociais da população não classificada como moradora de setores subnormais (e incluída em setores não-especiais), mas que habita setores precários, devem ser similares às dos indivíduos e famílias de setores classificados como subnormais" (Marques, 2007, p. 14).

A análise discriminante tem como objetivo distinguir categorias mediante o uso de diferentes variáveis que as descrevem. No contexto da metodologia proposta pelo CEM, significa utilizar variáveis que descrevem as características populacionais de setores subnormais para discriminar setores que são similares e que, por conseguinte, apresentam alta probabilidade de abranger assentamentos precários.

Considerando a diversidade de situações urbanas no país, foram desenvolvidos modelos específicos para distintas regiões. Para a estimativa dos modelos, foram utilizadas variáveis censitárias que representam condições de habitação e infraestrutura dos domicílios, renda e escolaridade do responsável pelo domicílio e aspectos demográficos. A metodologia foi aplicada para os dados censitários de 2000 de 561 municípios brasileiros, a maioria localizada em regiões metropolitanas e/ou com mais de 150 mil habitantes. Posteriormente, a metodologia foi aplicada para os municípios da Macrometrópole Paulista a partir dos dados de 2010 (Marques, 2013).

Outra metodologia relacionada à identificação de áreas precárias é a Mappa, desenvolvida em pesquisa contratada pela Companhia de Desenvolvimento Habitacional e Urbano do Estado de São Paulo (CDHU) e executada pela UFABC. A partir da utilização intensiva de técnicas de geoprocessamento, estatística e análise espacial, a Mappa propõe procedimentos para identificar, dimensionar e classificar assentamentos precários em distintas tipologias de tecido urbano – TECs (CDHU e UFABC, 2019; Feitosa *et al.*, 2019). A metodologia foi aplicada na Região Metropolitana (RM) da Baixada Santista, servindo de subsídio para aprimorar as informações sobre assentamentos precários que estavam sendo sistematizadas por intermédio de um processo de mapeamento colaborativo coordenado pela CDHU e realizado com o apoio da Agência Metropolitana da Baixada Santista – Agem-BS

6. Os setores subnormais compõem aglomerados subnormais que, segundo o IBGE, são "formas de ocupação irregular de terrenos de propriedade alheia (públicos ou privados) para fins de habitação em áreas urbanas e, em geral, caracterizados por um padrão urbanístico irregular, carência de serviços públicos essenciais e localização em áreas que apresentam restrição à ocupação" (IBGE, 2020, p. 5).

(Souza, Rossi e Rudge, 2018). O potencial de generalização da Mappa para outras regiões paulistas foi testado na Região do Grande ABC (Feitosa *et al.*, 2021).

O desenvolvimento da Mappa teve como ponto de partida a definição e caracterização de tipologias de tecido urbano (TECs) dos assentamentos precários da RM da Baixada Santista, identificadas em função das dimensões dos elementos urbanos (vias, lotes e habitações), do traçado regulador da ocupação e da compacidade do tecido. As TECs definidas foram:

- TEC 1 – Morros;
- TEC 2 – Palafitas;
- TEC 3 – Áreas úmidas (áreas de preservação permanente – APPs – de rios e córregos ou aterros de mangues e restingas);
- TEC 4 – Ocupação desordenada (sem traçado regulador prévio à ocupação);
- TEC 5 – Ocupação ordenada por traçado regulador; e
- TEC 6 – Ocupação esparsa ou pouco consolidada.

Para a classificação das TECs, a Mappa utiliza modelos de regressão logística para geração de superfícies de probabilidade da presença de cada tipologia de tecido de assentamentos precários. A partir das superfícies de probabilidade, constrói-se uma árvore de classificação das tipologias de assentamentos. A regressão logística é uma técnica estatística multivariada utilizada para a classificação de unidades de análise e que se assemelha à análise discriminante, adotada na metodologia proposta pelo CEM. Para a estimativa dos modelos, foram levantados e integrados dados de fontes diversas para a construção de variáveis representativas de aspectos físico-ambientais e presença de unidades de conservação, características da malha urbana e características populacionais, dos domicílios e entorno dos domicílios.

A partir da análise das metodologias desenvolvidas pelo CEM (aqui denominada como "metodologia do CEM") e CDHU e UFABC (Mappa), é possível realizar uma comparação de diferentes aspectos inerentes a esses estudos, um balanço sobre os avanços metodológicos alcançados, assim como algumas de suas limitações (quadro 1).

QUADRO 1
**Comparação entre as metodologias CEM e Mappa**

| Aspecto analisado | Metodologia do CEM (Marques, 2007) | Mappa (CDHU e UFABC, 2019) |
|---|---|---|
| Objetivo | Identificar assentamentos precários. | Identificar assentamentos precários e classificá-los em distintas TECs. |
| Abrangência geográfica | Território nacional. | Território do estado de São Paulo. |
| Disponibilidade dos dados | Dados abertos para o território nacional. | Dados abertos para o estado de São Paulo. |
| Tipo de dados | Uso exclusivo de dados censitários. | Dados provenientes de fontes e naturezas diversas. |
| Atualização dos resultados | Dependente da realização de levantamentos censitários (geralmente atualizados a cada dez anos). | Dados com diferentes resoluções temporais, incluindo alguns com atualização frequente (por exemplo, sensoriamento remoto e dados abertos de logradouros). |
| Unidade espacial de análise | Setor censitário: caráter operacional, que não necessariamente dialoga com a forma urbana; não é adequado para integração de dados. | Células e unidades homogêneas de uso e cobertura da terra (UHCTs): tem aderência à forma urbana (UHCT) ou flexibilidade para tal (célula); e facilita integração de dados (célula). |
| Apresentação dos resultados | Classificação categórica (precário ou não precário). | Resultados apresentados de forma contínua (superfícies de probabilidade) e de forma categórica (seis classes, uma para cada tipologia). |
| Técnica de classificação | Análise discriminante. | Análise de regressão logística. |

Elaboração dos autores.

Os primeiros aspectos analisados dizem respeito aos objetivos e à abrangência geográfica das metodologias. A metodologia do CEM busca contribuir para a identificação de assentamentos precários, sendo passível de aplicação em todo o território nacional. A Mappa foi desenvolvida para o estado de São Paulo e busca não apenas identificar assentamentos precários, como também classificá-los em tipologias de tecido urbano. Essas tipologias, embora possam servir de referência para distintas regiões, foram especificadas a partir da realidade da RM da Baixada Santista. Assim, a produção de estudos que ampliem o conhecimento sobre tipologias de assentamentos precários em outras regiões é necessária para que a Mappa possa ser aplicada a distintos contextos.

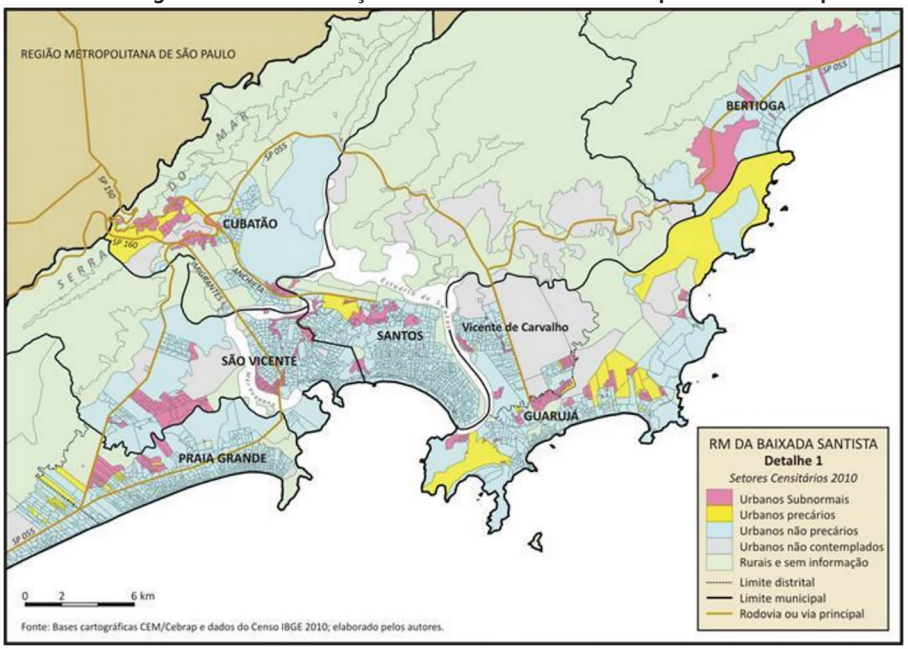
A abrangência geográfica das metodologias tem relação com a disponibilidade e o tipo de dados considerados no processo de classificação. A metodologia do CEM utiliza apenas informações do Censo Demográfico, que apresentam a vantagem de estarem disponíveis para todo o Brasil, mas não incluem aspectos urbanísticos e ambientais, conforme esclarecem Ferreira, Marques e Fusaro (2016). A Mappa integra dados de fontes diversas para construção de variáveis representativas de distintos aspectos territoriais. Quanto à disponibilidade dos dados, a Mappa utiliza

tanto dados disponíveis para todo o país como dados disponíveis apenas para o estado de São Paulo, visto que o estado possui uma infraestrutura de dados espaciais mais completa do que a maioria dos estados brasileiros. Este aspecto reforça as dificuldades de aplicação da Mappa para outras regiões do país.

A possibilidade de atualização dos resultados também está relacionada aos dados utilizados. As duas metodologias utilizam dados censitários, cujo levantamento deve ocorrer a cada dez anos (o próximo está previsto para 2022, atrasado por conta da pandemia e de restrições orçamentárias), o que dificulta a atualização em períodos intercensitários. A Mappa, entretanto, ao conciliar dados de fontes distintas, avança ao incluir alguns que podem ser atualizados com maior frequência, como os provenientes de produtos de sensoriamento remoto ou dados abertos sobre logradouros.

Outro aspecto metodológico relevante relacionado aos dados é a unidade espacial de análise. A metodologia do CEM adota o setor censitário como unidade espacial de análise, classificando-os como precários quando apresentam características semelhantes aos setores subnormais (figura 1). A escolha pelo setor censitário é coerente, visto que a metodologia utiliza apenas variáveis provenientes do questionário básico do Censo Demográfico, já agregadas por setor censitário. Entretanto, cabe salientar que a delimitação do setor censitário segue princípios operacionais que não dialogam necessariamente com a forma urbana e, por conseguinte, podem apresentar geometrias muito divergentes das poligonais dos assentamentos precários. Outra limitação do uso do setor censitário como unidade espacial de análise é que ele dificulta a integração de dados de fontes distintas, um aspecto que não era relevante no caso da metodologia do CEM.

FIGURA 1
**Metodologia do CEM: classificação de setores censitários como precários ou não precários**




Fonte: Marques (2013).
Obs.: Figura cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

A Mappa, por utilizar dados diversos, envolve o desafio da integração de dados e demanda maior atenção na escolha da unidade espacial de análise. A metodologia explora duas unidades espaciais de análise distintas: as UHCTs (Estado de São Paulo, 2014) e células de 100 m x 100 m (figura 1). As UHCTs são unidades delimitadas por meio de interpretação visual de ortofotos, utilizando critérios morfológicos. Por conseguinte, apresenta a vantagem de dialogar diretamente com a forma urbana e frequentemente coincide com os limites de assentamentos e loteamentos, incluindo os precários. No entanto, assim como o setor censitário, a UHCT não representa uma unidade espacial adequada para a integração de dados diversos. Outra limitação da UHCT diz respeito ao fato de ela estar disponível apenas para o estado de São Paulo. Por seu turno, as células podem ser geradas facilmente para qualquer área de estudo e são unidades mais adequadas para a integração de dados provenientes de unidades geográficas distintas, como as político-administrativas, físicas ou operacionais. Caso não tenham dimensões muito grandes, as células constituem unidades espaciais flexíveis, *pixels* que podem indicar a presença de NUI e que, agregados, representam também sua abrangência em termos de área. Outra

vantagem da célula é sua estabilidade espaço-temporal, uma vez que suas unidades não estão sujeitas a alterações nos seus limites físicos.

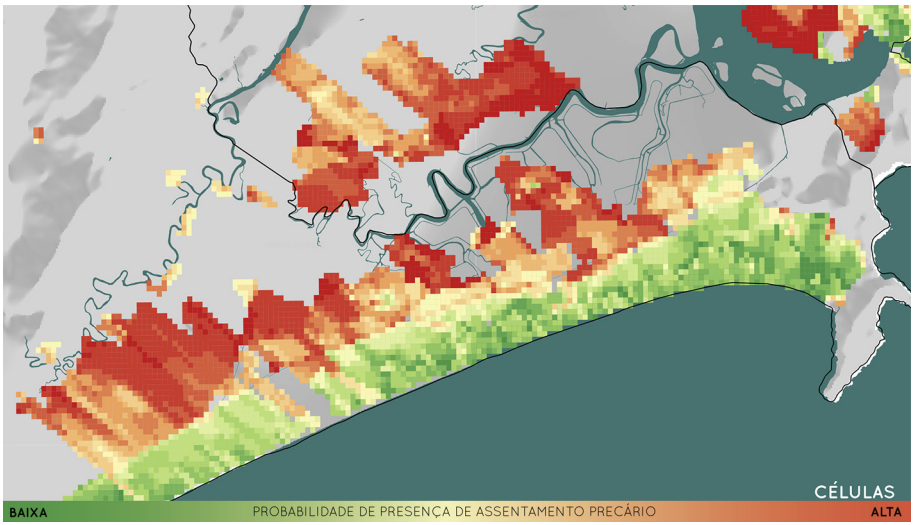
FIGURA 2
**Mappa: classificação de células ou UHCTs conforme tipologia de tecido de assentamento precário – município de Guarujá**

Fonte: CDHU e UFABC (2019).
Obs.: 1. Os polígonos com contorno preto representam as amostras de assentamentos precários.

Outro aspecto metodológico importante diz respeito à forma de apresentação dos resultados. Na metodologia do CEM, os resultados são apresentados de forma categórica, ou seja, os setores censitários urbanos são classificados como precários ou não precários. Na Mappa, os resultados são apresentados de duas maneiras distintas: i) contínua, como superfícies que indicam a probabilidade da presença de determinada tipologia de assentamento precário; ou ii) categórica, na qual as UHCTs ou células são classificadas como não precárias ou como uma das tipologias de assentamento precário estabelecidas.

A apresentação dos resultados na forma de superfícies de probabilidade (figura 2), de modo complementar à classificação categórica, é interessante por explicitar a heterogeneidade do território. Cabe salientar ainda que a escolha do limiar de probabilidade para a classificação pode gerar resultados muito distintos e é passível de controvérsias, principalmente em situações nas quais há desequilíbrio do tamanho das amostras pertencentes a cada classe e/ou incertezas sobre a ausência do evento investigado (por exemplo, existência de assentamentos precários em áreas que, na construção da variável dependente, estão classificadas como não precárias). Tais situações são típicas em aplicações voltadas à identificação de assentamentos precários, e os impactos na escolha dos limiares de classificação serão brevemente explorados na seção 4.

FIGURA 3
**Mappa: superfície de probabilidade de presença de assentamento precário**




Fonte: Feitosa *et al*. (2019).
Elaboração dos autores.
Obs.: Figura cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

A forma de apresentação dos resultados está relacionada à escolha da técnica de classificação adotada na metodologia. Na metodologia do CEM, optou-se pela análise discriminante, ao passo que a Mappa adotou a análise de regressão logística. Embora envolvam procedimentos estatísticos distintos, as duas técnicas são análogas quanto ao resultado produzido: ambas envolvem a estimativa de probabilidades e os resultados são de fácil interpretação. Quando as suposições básicas de ambas as técnicas são atendidas, seus resultados classificatórios e preditivos tendem a ser semelhantes. Entretanto, a regressão logística tem a vantagem de ser menos afe-

tada quando as suposições básicas não são satisfeitas – em específico, a suposição referente à distribuição normal das variáveis (Hair *et al.*, 2009).

## 3 A METODOLOGIA NUI

Os avanços promovidos pelas metodologias CEM e Mappa foram considerados no desenvolvimento da metodologia NUI, que tem como objetivo contribuir para a identificação de núcleos urbanos informais ocupados predominantemente por população de baixa renda. Utilizando dados de fontes e naturezas diversas disponíveis para todo o território nacional, a metodologia NUI baseia-se na construção de modelos para a geração de superfícies de probabilidade da presença de NUI, e sua aplicação é possível em distintas regiões do país. Além disso, a metodologia prioriza a utilização de dados passíveis de atualização mais frequente.

A metodologia consiste em três etapas principais: i) construção e integração de variáveis potencialmente relevantes para a identificação de NUI; ii) construção de modelos para a geração de superfícies de probabilidade; e iii) apresentação e análise dos resultados.

A etapa 1 envolve a construção de variáveis a partir de dados provenientes de fontes diversas, amplamente disponíveis e preferencialmente abertos, que representem aspectos potencialmente relacionados a irregularidade e precariedade:

- forma urbana, incluindo informações sobre irregularidade da forma de quadra ou bolsões de ocupação e cobertura de vias carroçáveis;
- características físico-territoriais, tais como declividade e curvatura do terreno, bem como áreas com restrições à ocupação (unidade de conservação, faixa de servidão de linhas de alta tensão e dutovias);
- características das edificações, entorno e infraestrutura; e
- características sociodemográficas, incluindo aspectos relacionados a densidade, renda, trabalho, educação e ciclo de vida da população.

Uma vez que a metodologia NUI utiliza dados diversos, que demandam integração, a célula apresenta-se como unidade espacial de análise mais adequada. Em particular, adotou-se uma grade celular com resolução de 100 m, cuja área é compatível com a de uma quadra média. Também se optou pela construção de uma grade que fosse compatível com a grade estatística[7] do IBGE, com o intuito de facilitar uma futura integração dos resultados (figura 3).

7. A grade estatística do IBGE, publicada em 2016, foi gerada com o intuito de disseminar dados estatísticos a partir dos microdados do universo do censo (IBGE, 2016). Disponibiliza dados populacionais em células de 200 por 200 m nas áreas urbanas e de 1 por 1 km nas áreas rurais.

FIGURA 4
**Grade celular**
4A – Imagem esquemática da integração de dados

4B – Sobreposição da grade celular NUI com a grade estatística do IBGE


Fonte: Frizzi e Pinho (2017).
Elaboração dos autores.
Obs.: Figura cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Para a definição da etapa 2, foi necessária a seleção da técnica que será utilizada para a construção dos modelos de identificação de NUI. Três diferentes técnicas de classificação foram analisadas a partir dos dados dos seis polos da pesquisa: i) regressão logística; ii) análise discriminante; e iii) árvore de decisão (algoritmo C5.0). As duas primeiras têm a vantagem de permitir fácil interpretação dos parâmetros gerados pelo modelo e são adequadas à construção de superfícies de probabilidade. Por sua vez, a terceira, a árvore de decisão, foi analisada em vir-

tude do destaque que as técnicas de *machine learning* vêm recebendo em estudos internacionais de identificação de assentamentos precários (Friesen *et al.*, 2018; Mahabir *et al.*, 2018; Ribeiro, 2015). Os resultados obtidos, considerando-se o contexto e os objetivos da pesquisa, indicaram a regressão logística como a mais vantajosa para essa aplicação, dado que concilia resultados de fácil interpretação, com maior potencial de concordância e adequados para a apresentação na forma de superfícies de probabilidade. Foi, assim, a técnica escolhida para a modelagem da probabilidade da presença de NUI.

A construção dos modelos demanda informações sobre a presença de NUIs, que podem ser provenientes de levantamentos locais. Na ausência desses, os dados sobre AGSNs do IBGE podem ser adotados como "amostras" de NUI, conforme será detalhado no capítulo 5 deste livro. Observou-se, para todos os polos da pesquisa, que distintas tipologias de NUI puderam ser identificadas na base de AGSN, embora a qualidade e a representatividade dos AGSNs como amostra de NUI varie. Por exemplo, em Recife-PE, os AGSNs representam uma amostra de 40% dos NUIs, ao passo que, em Juazeiro do Norte-CE, o percentual é de apenas 15%, o que se torna particularmente crítico nessa região, que, além de ser menos populosa e possuir um número menor de amostras de NUI, apresenta diferenças menos expressivas entre as áreas identificadas pela pesquisa de campo como NUI e as demais áreas.

A etapa 3 consiste em apresentar os resultados dos modelos logísticos estimados na forma de superfícies de probabilidade da presença de NUI e analisá-los. Caso seja relevante apresentar os resultados de forma binária (presença ou ausência de NUI), pode-se selecionar um limiar de probabilidade para a classificação. Cabe salientar, entretanto, que uma análise prévia para a seleção desse limiar deve ser adotada, conforme será demonstrado na próxima seção.

## 4 APLICAÇÃO DA METODOLOGIA NUI

A metodologia NUI foi aplicada aos seis polos da Pesquisa de Núcleos Urbanos Informais no Brasil, que constituem um total de 157 municípios, com 19.783.220 habitantes (IBGE, 2018). A diversidade de características dos polos – um em cada região do país, além do Distrito Federal – auxilia na análise sobre a possibilidade de aplicação da metodologia para todo o país.

### 4.1 Construção e integração de variáveis

Os dados considerados para a construção dos modelos estão disponíveis para todo o território nacional[8] e são os seguintes:

8. Todos os dados são abertos, com exceção do Cadastro Único, que demanda cuidados específicos para a manutenção do sigilo.

- *Aglomerados subnormais 2019* (IBGE, 2020);
- Cadastro Único para Programas Sociais do Governo Federal – Cadastro Único (Brasil, 2020);
- Censo Demográfico 2010 (IBGE);[9]
- Atlas do Desenvolvimento Humano (PNUD, Ipea e FJP, 2016);
- modelos digitais de terreno provenientes da Shuttle Radar Topography Mission – SRTM (National Aeronautics and Space Administration – Nasa);[10]
- logradouros (OpenStreetMap);[11]
- dados de hidrografia (Fundação Brasileira para o Desenvolvimento Sustentável – FBDS);[12]
- unidades de conservação de proteção integral (Ministério do Meio Ambiente – MMA);[13]
- faixas de servidão de linhas de alta tensão (Agência Nacional de Energia Elétrica – Aneel);[14] e
- faixas de servidão de dutos (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP).[15]

A partir desses dados, foram construídas e integradas variáveis que representam distintos aspectos relevantes para a caracterização dos núcleos urbanos informais. O quadro 2 apresenta uma seleção dessas variáveis.[16] A construção das variáveis demandou uma série de processamentos, tais como o cálculo de declividade a partir de modelos digitais de terreno, a geração de faixas de distância a partir de vias carroçáveis ou cursos d'água, a aplicação de métricas para representar a irregularidade das quadras e bolsões de ocupação, a geocodificação de endereços do Cadastro Único, entre outros.

A metodologia NUI inova ao enfatizar o uso de dados que são constantemente atualizados, em particular os dados georreferenciados do Cadastro Único. O Cadastro Único tem como objetivo o cadastramento e a manutenção de informações atualizadas das famílias de baixa renda em todos os municípios brasileiros. O Cadastro

9. Disponível em: <https://bit.ly/38lqZ64>.
10. Disponível em: <https://go.nasa.gov/3K5BlVB>. Acesso em: 9 jun. 2020.
11. Disponível em: <https://bit.ly/3IWY199>. Acesso em: 17 ago. 2020.
12. Disponível em: <https://bit.ly/36G27FL>. Acesso em: 17 set. 2020.
13. Disponível em: <http://mapas.mma.gov.br/i3geo/datadownload.htm>.
14. Disponível em: <https://bit.ly/3NEIgXQ>. Acesso em: 15 jun. 2020.
15. Disponível em: <https://bit.ly/3qTueId>. Acesso em: 25 jun. 2020.
16. Essa seleção inclui as variáveis que foram utilizadas nos modelos de regressão logística para a geração das superfícies de probabilidade.

Único tem mais de 28,5 milhões de famílias cadastradas e constitui uma importante base de dados para caracterização da parcela mais pobre da população brasileira, agregando informações de renda, características do domicílio, escolaridade, entre outras (Brasil, 2019). Seus dados incluem os endereços das famílias cadastradas, o que potencializa o uso dessas informações para a identificação de NUIs. Para que esses dados possam ser espacializados, é necessário geocodificá-los, o que pode ser problemático para algumas regiões, conforme revelaram os experimentos realizados. Do total de 2,272 milhões de registros do Cadastro Único nos seis polos, 284.169 (12,5%) não puderam ser geocodificados. A qualidade da geocodificação, entretanto, variou muito entre os polos: em Porto Alegre-RS, apenas 1,3% dos endereços não foram geocodificados, ao passo que, em Brasília-DF e Juazeiro do Norte-CE, esse percentual foi de 28% e 26%, respectivamente (tabela 1).

QUADRO 2
**Variáveis selecionadas construídas para a aplicação da metodologia NUI nos polos de Belo Horizonte-MG, Brasília-DF, Juazeiro do Norte-CE, Marabá-PA, Porto Alegre-RS e Recife-PE**

| Tipo | Variável | Descrição | Fonte |
|---|---|---|---|
| Dependente | AGSN | Aglomerados subnormais 2019 do IBGE. | IBGE (2020) |
| | NUI | Núcleos urbanos informais levantados em campo. | - |
| Independente – forma urbana | IndiceForma | Média do índice de forma das quadras/bolsões de ocupação por unidade de análise – o índice de forma mede a regularidade das quadras: quanto mais próximo de 1, mais regulares são as quadras/bolsões de ocupação. | OpenStreetMap. Disponível em: <https://bit.ly/3lWY199>. Acesso em: 17 ago. 2020. |
| | Vias50m | Porcentagem de área ocupada da unidade de análise dentro da faixa de 50 m da via carroçável. | OpenStreetMap. Disponível em: <https://bit.ly/3lWY199>. Acesso em: 17 ago. 2020. |
| Independente – características físico-territoriais | Declividade | Declividade média do terreno por unidade de análise. | Nasa. Disponível em: <https://go.nasa.gov/3K5BlVB>. Acesso em: 9 jun. 2020. |
| | APP30m | Porcentagem de área ocupada da unidade de análise a 30 m do curso d'água. | FBDS. Disponível em: <https://bit.ly/36G27FL>. Acesso em: 17 set. 2020. |
| | AltaTensao | Porcentagem de área ocupada da unidade de análise em faixas de servidão de linhas de alta tensão. | Aneel. Disponível em: <https://bit.ly/3NEIgXQ>. Acesso em: 15 jun. 2020. |

(Continua)

(Continuação)

| Tipo | Variável | Descrição | Fonte |
|---|---|---|---|
| Independente – características das edificações, entorno e infraestrutura | DomSIden | Porcentagem de domicílios sem identificação do logradouro. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomSIlu | Porcentagem de domicílios sem iluminação pública. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomSPav | Porcentagem de domicílios sem pavimentação. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomSArb | Porcentagem de domicílios sem arborização. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomSMed | Porcentagem de domicílios sem medidor de uso exclusivo. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomSEsg | Porcentagem de domicílios com esgoto a céu aberto. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomSRedeEsg | Porcentagem de domicílios sem ligação à rede de esgoto ou fossa séptica. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomApto | Porcentagem de domicílios particulares permanentes do tipo apartamento. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomImpr | Porcentagem de domicílios particulares improvisados. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomAdeq | Porcentagem de domicílios particulares permanentes com moradia adequada. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DomNBanDom | Média do número de banheiros por domicílio. | Disponível em: <https://bit.ly/38lqZ64>. |
| | DomNBanHab | Média do número de banheiros por habitante. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | AguaRede | Porcentagem de domicílios particulares permanentes com abastecimento de água da rede geral. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | LixoQueimado | Porcentagem de domicílios particulares permanentes com lixo queimado na propriedade. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | LixoAterrado | Porcentagem de domicílios particulares permanentes com lixo enterrado na propriedade. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | LixoCacamba | Porcentagem de domicílios particulares permanentes com lixo coletado em caçamba de serviço de limpeza. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | CadAlvenaria | Quantidade de famílias cadastradas no Cadastro Único com material da parede do tipo alvenaria sem revestimento. | Brasil (2020) |
| | CadSMed | Quantidade de famílias cadastradas no Cadastro Único com iluminação elétrica sem medidor. | Brasil (2020) |

(Continua)

(Continuação)

| Tipo | Variável | Descrição | Fonte |
|---|---|---|---|
| Independente – características sociodemográficas | DenDom | Densidade domiciliar. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | DenPop | Densidade populacional. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | RenMeioSM | Porcentagem de domicílios particulares com rendimento nominal mensal domiciliar *per capita* de até 1/2 salário mínimo. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | Ren3SM | Porcentagem de pessoas responsáveis com rendimento nominal mensal de até 3 salários mínimos. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | RenRespMedia | Renda média do responsável pelo domicílio. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | NMoradores | Média do número de moradores em domicílios particulares permanentes. | IBGE. Disponível em: <https://bit.ly/38lqZ64>. |
| | Mort1 | Mortalidade até 1 ano de idade. | PNUD, Ipea e FJP (2016) |
| | FecTot | Taxa de fecundidade total. | PNUD, Ipea e FJP (2016) |
| | CadNPessoas | Quantidade de pessoas cadastradas no Cadastro Único por família. | Brasil (2020) |
| | CadNFamilias | Quantidade de famílias cadastradas no Cadastro Único por domicílio. | Brasil (2020) |

Elaboração dos autores.

TABELA 1
**Resultados da geocodificação dos endereços do Cadastro Único para os seis polos**

| | Porto Alegre-RS | Juazeiro do Norte-CE | Brasília-DF | Recife-PE | Belo Horizonte-MG | Marabá-PA | **Todos** |
|---|---|---|---|---|---|---|---|
| Total de famílias | 319.938 | 200.590 | 336.014 | 804.921 | 464.038 | 146.684 | **2.272.185** |
| Endereços não encontrados | 4.106 (1,3%) | 53.497 (26%) | 97.358 (28%) | 87.707 (10,9%) | 10.929 (2,36%) | 30.572 (21%) | **284.169 (12,5%)** |

Fonte: Brasil (2020).
Elaboração dos autores.

### 4.2 Construção de modelos e análise dos resultados

Foram estimados, para os seis polos da pesquisa, modelos de regressão logística que utilizam como variável dependente tanto os aglomerados subnormais do IBGE ($Y$ = AGSN) quanto os NUIs levantados pela pesquisa por meio de trabalho de campo ($Y$ = NUI). As duas variáveis foram utilizadas com o objetivo de comparar os resultados dos modelos e avaliar o uso da informação dos AGSNs em regiões onde não há disponibilidade de dados mais completos sobre NUIs. Na prática, deve-se adotar como variável dependente a informação mais completa possível e

utilizar os resultados da modelagem como um plano de informação adicional, que sirva de referência para a revisão da base existente. As tabelas 2 e 3 apresentam os parâmetros estimados para os modelos dos seis polos da pesquisa.

TABELA 2
**Parâmetros estimados dos modelos logísticos: polos de Belo Horizonte-MG, Brasília-DF e Juazeiro do Norte-CE**

| Variável | Parâmetros estimados e erro-padrão – *b*(*EP*) | | | | | |
|---|---|---|---|---|---|---|
| | Belo Horizonte | | Brasília | | Juazeiro do Norte | |
| | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) |
| Constante | -11,703<br>(0,057) | -5,962<br>(0,057) | -6,516<br>(0,063) | -5,514<br>(0,063) | -6,212<br>(0,127) | -3,445<br>(0,127) |
| Vias50m | 1,513<br>(0,016) | 0,568<br>(0,016) | 1,183<br>(0,024) | 1,281<br>(0,024) | 0,978<br>(0,03) | 0,672<br>(0,03) |
| Declividade | 0,083<br>(0,001) | 0,047<br>(0,001) | - | - | 0,128<br>(0,004) | 0,025<br>(0,004) |
| APP30m | 0,972<br>(0,023) | 0,303<br>(0,023) | - | - | - | - |
| AltaTensao | - | - | - | - | 1,491<br>(0,245) | 1,494<br>(0,245) |
| DomSArb | 0,01<br>(2e-04) | 0,009<br>(2e-04) | - | - | - | - |
| DomSMed | 0,016<br>(5e-04) | 0,008<br>(5e-04) | - | - | - | - |
| DomSEsg | 0,008<br>(4e-04) | 0,005<br>(4e-04) | 0,029<br>(5e-04) | 0,007<br>(5e-04) | 0,009<br>(4e-04) | 0,004<br>(4e-04) |
| DomCLixAc | - | - | - | - | 0,014<br>(1e-03) | 0,005<br>(1e-03) |
| DomApto | - | - | -0,09<br>(0,001) | -0,028<br>(0,001) | - | - |
| DomImpr | - | - | 0,085<br>(0,002) | 0,048<br>(0,002) | - | - |
| DomAdeq | -0,012<br>(2e-04) | -0,005<br>(2e-04) | -0,007<br>(3e-04) | -0,001<br>(3e-04) | - | - |
| DomNBanDom | - | - | -0,957<br>(0,023) | -0,487<br>(0,023) | - | - |
| LixoAterrado | - | - | - | - | 0,274<br>(0,008) | 0,136<br>(0,008) |
| LixoCacamba | - | - | 0,019<br>(3e-04) | 0,005<br>(3e-04) | - | - |
| CadAlvenaria | 0,243<br>(0,014) | 0,197<br>(0,014) | - | - | - | - |
| DenPop | 2e-04<br>(2e-06) | 6e-05<br>(2e-06) | - | - | - | - |
| RenMeioSM | - | - | - | - | 0,015<br>(0,002) | 0,034<br>(0,002) |

(Continua)

(Continuação)

| Variável | Parâmetros estimados e erro-padrão – *b*(*EP*) | | | | | |
|---|---|---|---|---|---|---|
| | Belo Horizonte | | Brasília | | Juazeiro do Norte | |
| | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) |
| RenRespMedia | - | - | - | - | -3e-04<br>(8e-05) | -0,002<br>(8e-05) |
| NMoradores | 0,209<br>(0,013) | 0,312<br>(0,013) | - | - | - | - |
| Mort1 | - | - | 0,204<br>(0,002) | 0,201<br>(0,002) | - | - |
| FecTot | 2,372<br>(0,018) | 1,462<br>(0,018) | - | - | - | - |
| CadNPessoas | 0,092<br>(0,005) | 0,131<br>(0,005) | - | - | - | - |
| CadNFamilias | - | - | - | - | 0,304<br>(0,038) | 0,357<br>(0,038) |
| AIC | 38.147 | 183.817 | 40.870 | 98.555 | 9.673 | 37.715 |
| Cox & Snell R2 | 0,087 | 0,13 | 0,095 | 0,126 | 0,027 | 0,101 |
| Nagelkerke R2 | 0,332 | 0,189 | 0,339 | 0,252 | 0,111 | 0,153 |

Elaboração dos autores.
Obs.: Todos os coeficientes são significativos ao nível de 1%.

TABELA 3
**Parâmetros estimados dos modelos logísticos: polos de Marabá-PA, Porto Alegre-RS e Recife-PE**

| Variável | Parâmetros estimados e erro-padrão – *b*(*EP*) | | | | | |
|---|---|---|---|---|---|---|
| | Marabá | | Porto Alegre | | Recife | |
| | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) |
| Constante | -5,503<br>(0,078) | -3,156<br>(0,078) | -9,758<br>(0,136) | -8,201<br>(0,136) | -4,401<br>(0,03) | -2,23<br>(0,03) |
| IndiceForma | - | - | 0,468<br>(0,013) | 0,455<br>(0,013) | 0,634<br>(0,011) | 0,382<br>(0,011) |
| Vias50m | 1,453<br>(0,035) | 0,908<br>(0,035) | 0,422<br>(0,029) | 0,928<br>(0,029) | 0,961<br>(0,024) | 0,726<br>(0,024) |
| Declividade | 0,105<br>(0,005) | 0,044<br>(0,005) | - | - | 0,088<br>(0,002) | 0,067<br>(0,002) |
| APP30m | 1,139<br>(0,05) | 0,498<br>(0,05) | - | - | - | - |
| DomSIden | 0,009<br>(4e-04) | 0,002<br>(4e-04) | - | - | - | - |
| DomSIlu | 0,026<br>(6e-04) | 0,018<br>(6e-04) | - | - | - | - |
| DomSPav | - | - | 0,004<br>(3e-04) | 0,004<br>(3e-04) | - | - |
| DomSMed | - | - | 0,022<br>(6e-04) | 0,01<br>(6e-04) | 0,042<br>(6e-04) | 0,009<br>(6e-04) |

(Continua)

(Continuação)

| Variável | Parâmetros estimados e erro-padrão – *b*(*EP*) | | | | | |
|---|---|---|---|---|---|---|
| | Marabá | | Porto Alegre | | Recife | |
| | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) | (*Y* = AGSN) | (*Y* = NUI) |
| DomSEsg | - | - | 0,011<br>(6e-04) | 0,007<br>(6e-04) | - | - |
| DomSRedeEsg | - | - | 0,252<br>(0,036) | 0,248<br>(0,036) | 1,119<br>(0,02) | 0,152<br>(0,02) |
| DomCLixAc | - | - | 0,013<br>(6e-04) | 0,003<br>(6e-04) | - | - |
| DomNBanHab | - | - | -3,291<br>(0,095) | -1,888<br>(0,095) | - | - |
| AguaRede | -0,021<br>(4e-04) | -0,009<br>(4e-04) | - | - | - | - |
| LixoQueimado | - | - | - | - | -0,049<br>(4e-04) | -0,013<br>(4e-04) |
| CadAlvenaria | - | - | 0,163<br>(0,008) | 0,163<br>(0,008) | - | - |
| CadSMed | - | - | - | - | 0,184<br>(0,008) | 0,091<br>(0,008) |
| DenDom | - | - | - | - | 6e-06<br>(4e-06) | 1e-04<br>(4e-06) |
| DenPop | 4e-05<br>(4e-06) | 8e-05<br>(4e-06) | 5e-06<br>(2e-06) | 4e-05<br>(2e-06) | - | - |
| Ren3SM | 0,008<br>(9e-04) | 0,016<br>(9e-04) | 0,022<br>(8e-04) | 0,011<br>(8e-04) | - | - |
| RenRespMedia | - | - | - | - | -0,001<br>(1e-05) | -4e-04<br>(1e-05) |
| NMoradores | - | - | 1,296<br>(0,033) | 1,155<br>(0,033) | - | - |
| CadNPessoas | - | - | 0,283<br>(0,007) | 0,327<br>(0,007) | - | - |
| CadNFamilias | - | - | - | - | 0,478<br>(0,019) | 0,683<br>(0,019) |
| AIC | 13.051 | 38.482 | 41.626 | 84.718 | 58.217 | 110.213 |
| Cox & Snell R2 | 0,102 | 0,096 | 0,072 | 0,106 | 0,148 | 0,161 |
| Nagelkerke R2 | 0,269 | 0,139 | 0,291 | 0,255 | 0,288 | 0,227 |

Elaboração dos autores.
Obs.: Todos os coeficientes são significativos ao nível de 1%.

Nos modelos estimados para o Polo de Belo Horizonte-MG, as variáveis físico-territoriais relacionadas à alta declividade (*Declividade*), presença de área de proteção permanente (*APP30*) e proximidade a vias carroçáveis (*Vias50m*) revelaram-se significantes para a identificação de NUI. No que se refere às variáveis do Cadastro Único, os modelos indicaram que a presença de NUI está relacionada ao total de pessoas em famílias cadastradas (*CadNPessoas*) e à presença de edifi-

cações com alvenaria sem revestimento (*CadAlvenaria*). Também apresentaram efeito positivo as variáveis:

- maior proporção de domicílios sem arborização (*SArb*);
- sem medidor de uso exclusivo (*SMed*);
- com esgoto a céu aberto (*SEsg*);
- densidade populacional (*DenPop*);
- média do número de moradores (*NMoradores*); e
- taxa de fecundidade (*FecTot*).

Apenas uma variável apresentou efeito com sinal negativo, ou seja, que diminui a probabilidade de presença de NUI: a porcentagem de domicílios com moradia adequada (*DomAdeq*).

No Polo de Brasília-DF, indicam maior probabilidade de presença de NUI: i) áreas próximas a vias carroçáveis (*Vias50m*); ii) com esgoto a céu aberto (*SEsg*); iii) com alta proporção de domicílios improvisados (*DomImpr*); iv) com lixo coletado em caçamba de serviço de limpeza (*LixoCacamba*); e v) com maiores taxas de mortalidade infantil (*Mort1*). Os modelos indicam ainda que uma maior proporção de domicílios do tipo apartamento (*DomApto*), com moradia adequada (*DomAdeq*) e com maior média do número de banheiros por domicílio (*DomNBanDom*), diminuem a chance de presença de NUI.

Nos modelos construídos para o Polo de Juazeiro do Norte-CE, destacaram-se, com maior efeito, as variáveis que indicam a precariedade no destino do lixo (*LixoAterrado*), no caso do modelo *Y* = AGSN, e a baixa renda dos moradores (*RenMeioSM*), no modelo *Y* = NUI. Os modelos indicaram ainda que áreas com alta declividade (*Declividade*), localizadas em faixas de servidão de linhas de alta tensão (*AltaTensao*) e próximas a vias carroçáveis (*Vias50m*) têm maior probabilidade de serem NUI. A presença de famílias cadastradas no Cadastro Único também está relacionada à presença de NUI (*CadNFamilias*). No que diz respeito à infraestrutura, os NUIs estão associados à presença de esgoto a céu aberto (*DomSEsg*) e lixo acumulado nas ruas (*DomCLixAc*). Uma maior concentração de responsáveis de baixa renda (*RenRespMedia*) também aumenta a probabilidade da presença de NUI.

No caso do Polo de Marabá-PA, a variável domicílios sem iluminação (*DomSIlu*) apresentou o maior efeito em ambos os modelos (*Y* = AGSN e *Y* = NUI), estando associada à presença de NUI. Os modelos também indicaram que apresentam maiores chances de serem NUIs as áreas com ausência de rede de água (*AguaRede*) e identificação de logradouros (*DomSIden*), elevada densidade populacional (*DenPop*), próximas a vias carroçáveis (*Vias50m*) e com presença de APP

(*APP30m*) e altas declividades (*Declividade*). A predominância de responsáveis pelo domicílio de baixa renda (*Ren3SM*) também aumenta a probabilidade de presença de NUI em Marabá-PA.

Nos modelos estimados para o Polo de Porto Alegre-RS, a variável que representa a proporção de domicílios sem medidor de eletricidade (*DomSMed*) destacou-se para ambas as variáveis dependentes (*Y* = AGSN e *Y* = NUI). Os modelos indicaram ainda que a presença de famílias cadastradas no Cadastro Único (*CadNFamilias*) – em particular, aquelas que residem em edificações de alvenaria sem revestimento *(CadAlvenaria*) – está relacionada à presença de NUI. Áreas que também relacionam-se à maior probabilidade da presença de NUI foram:

- alta densidade populacional (*DenPop*);
- forma irregular de quadras ou bolsões de ocupação (*IndiceForma*);
- com esgoto a céu aberto (*DomSEsg*);
- sem ligação à rede de esgoto ou fossa séptica (*DomSRedeEsg*);
- com lixo acumulado nos logradouros (*DomCLixAc*);
- sem pavimentação (*DomSPav*); e
- próximas de vias carroçáveis (*Vias50m*).

Situações relacionadas à alta densidade no domicílio, representadas nos modelos pelas variáveis média de moradores por domicílio (*NMoradores*) e número de banheiros por habitante (*DomNBanHab*), também aumentam a probabilidade da presença de NUI. Por fim, os modelos indicam ainda que um maior percentual de responsáveis com renda de até 3 salários mínimos (*Ren3SM*) aumenta a probabilidade de a célula ser NUI.

No caso do Polo de Recife-PE, as variáveis que apresentaram maior poder explicativo foram as indicadoras da irregularidade da forma das quadras ou bolsões de ocupação (*IndiceForma*) e da presença de famílias cadastradas no Cadastro Único (*CadNFamilias*), especialmente aquelas famílias com precariedade no acesso à eletricidade (*CadSMed*, energia elétrica sem medidor). As áreas com: i) moradores mais pobres (*RenRespMedia*); ii) de alta densidade de domicílios (*DenDom*); iii) alta declividade (*Declividade*); e/ou iv) próximas a vias carroçáveis (*Vias50m*) apresentam maior probabilidade de serem NUIs. Em relação às variáveis de infraestrutura, destacou-se a ausência de medidor de eletricidade (*DomSMed*) e a ausência de ligação à rede de esgoto ou fossa séptica (*DomSRedeEsg*). Chamou ainda atenção o coeficiente negativo associado à variável porcentagem de domicílios com lixo queimado na propriedade (*LixoQueimado*), associado a áreas mais rurais, que

comumente apresentam piores condições de renda e infraestrutura, mas que não estão necessariamente identificadas como NUI.

A avaliação dos modelos foi realizada com base em duas métricas: o coeficiente de concordância kappa e a área sob a curva (*area under the curve* – AUC). A curva característica de operação do receptor (*receiver operating characteristic curve* – ROC curve) é uma figura que ilustra a *performance* de um classificador binário de acordo com diferentes limiares de classificação. Ela é influenciada por duas métricas: a razão de verdadeiros positivos e a razão de falsos positivos. Um resultado de classificação perfeito possui uma taxa de verdadeiro positivo igual a 1 e uma taxa de falso positivo igual a 0, produzindo um ponto no canto superior esquerdo do gráfico, enquanto um resultado de classificação *aleatório* é coincidente com a linha diagonal e indica a incapacidade de discriminação do modelo. Portanto, quanto maior a AUC, que varia de 0 a 1, maior a capacidade preditiva do modelo. Para este estudo, o uso da AUC torna-se particularmente interessante, pelo fato de tratar-se de uma métrica invariante em escala (não trabalha com valores absolutos, e, sim, com a precisão das classificações) e por medir a qualidade das previsões do modelo independente do limiar de classificação.

Por sua vez, o coeficiente kappa é uma métrica desenhada para avaliar a concordância entre dois avaliadores, levando em consideração a probabilidade de a concordância ocorrer ao acaso. Ela é uma métrica muito utilizada para avaliar modelos quando a distribuição de classes é desigual – sendo este o caso da presença de NUIs. Sua fórmula é a que se segue:

$$Kappa = (O-E)/(1-E),$$

em que $O$ é a acurácia observada e $E$ é a acurácia esperada com base nos totais marginais da matriz de confusão.[17] O coeficiente kappa pode apresentar valores entre -1 e 1, sendo que 0 representa a ausência de concordância entre as classes observadas e previstas, enquanto o valor 1 indica a concordância perfeita entre a previsão do modelo e as classes observadas. Valores negativos indicam que a predição está na direção oposta da verdade, mas raramente ocorrem em modelos preditivos (Kuhn e Johnson, 2013).

Ao contrário da ROC e da AUC, o coeficiente de concordância kappa avalia os resultados de classificação categórica (por exemplo, presença/ausência de NUI) específica, sendo o limiar de classificação geralmente utilizado igual a 0,5. Entretanto, em situações em que se tem conhecimento da presença do evento, mas há incertezas sobre a ausência, denominadas na análise de classificação como

17. Matriz de confusão é um recurso utilizado em análise preditiva que consiste em uma tabela que relata o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. No caso desta pesquisa, os totais marginais da matriz de confusão indicam o número de células mapeadas como NUI e ausência-NUI (totais marginais de linha) e o número de células cujos modelos classificaram como NUI e ausência-NUI (totais marginais de coluna).

"pseudoausência" (Hijmans e Elith, 2017), os limiares de classificação relativos à presença/ausência do evento tornam-se muito incertos se definidos apenas pela base de dados (Phillips *et al.*, 2009). Nesses casos, Phillips e Elith (2011) recomendam que a seleção dos limiares de probabilidade adotados na classificação seja realizada com auxílio de informações externas à base de dados original.

Considerando-se as dificuldades inerentes à identificação de um limiar de probabilidade "ideal", bem como a relevância de se explicitar a heterogeneidade do território analisado, recomenda-se que os resultados da metodologia NUI sejam apresentados na forma de superfícies de probabilidade. Para a comparação dos resultados da classificação por meio do coeficiente kappa, foram considerados distintos limiares de classificação (gráfico 1), sendo reportado aquele que apresentou a maior concordância com os dados sobre NUI levantados em campo. Esse kappa foi denominado "kappa potencial". Sabe-se que, em uma situação real, dados que atuem como "verdade de campo" não estarão disponíveis para a identificação do limiar mais adequado, que resulte no maior kappa possível. Entretanto, caso seja relevante a realização de uma classificação categórica, o analista pode utilizar a própria variável dependente (por exemplo, os AGSNs) como referência para a seleção do limiar.

O kappa foi considerado nessa análise por conta de sua popularidade, embora apresente uma série de limitações. Feinstein e Cicchetti (1990) destacam, por exemplo, a alta sensibilidade do coeficiente kappa aos totais marginais da matriz de confusão. Mesmo com altos valores de acurácia observada ($O$), baixos valores de kappa são observados quando os totais marginais são drasticamente desbalanceados, o que é o caso dos experimentos apresentados, dado que a maior parte das células não são classificadas como NUI. Por conseguinte, o principal objetivo do uso do coeficiente kappa no estudo é o de identificar limiares de classificação mais adequados, dado que, nesses casos, são comparados resultados cujas condições dos experimentos são as mesmas.

No gráfico 1, identifica-se que os limiares ótimos para a classificação de NUI são sempre inferiores a 0,5, especialmente nos modelos "$Y$ = AGSN" (em vermelho). Esse resultado coincide com observações realizadas no âmbito do desenvolvimento da Mappa (Feitosa *et al.*, 2021) e evidencia a limitação de utilização deste valor-padrão (0,5) para a classificação de NUIs ou assentamentos precários. Reforça-se, assim, a recomendação de análise das superfícies de probabilidade somada ao conhecimento local e a outras fontes de dados disponíveis.

Na maioria dos polos, é pequena a diferença do maior coeficiente kappa obtido a partir dos modelos "$Y$ = AGSN" e "$Y$ = NUI" (valor máximo das curvas vermelhas e azuis gráfico 1), reforçando a conclusão de que os aglomerados subnormais podem ser uma boa amostra para a identificação de NUI. Essa diferença varia, sendo muito

menor nos polos de Belo Horizonte-MG e Marabá-PA e maior no Polo de Juazeiro do Norte-CE. A diferença constatada neste polo pode ser justificada pelo fato de os dados dos AGSNs de Juazeiro do Norte-CE apresentarem limitações maiores para seu uso como amostra de NUI, destacando-se o ruído da amostra (32% dos AGSNs não são NUIs), combinado com a baixa representatividade (os AGSNs representam 15% dos NUIs) e o baixo número de AGSNs no polo (capítulo 5).

GRÁFICO 1
**Coeficiente kappa dos modelos, de acordo com o limiar de classificação**

Elaboração dos autores.

O gráfico 2 apresenta as curvas ROC e a tabela 3 sumariza as métricas AUC e kappa obtidas para cada um dos modelos estimados. As métricas AUC destacam, novamente, que é pequena a diferença entre os modelos "$Y$ = AGSN" e "$Y$ = NUI" nos polos, com exceção de Juazeiro do Norte-CE. De maneira geral, a AUC dos modelos variou entre 0,63 e 0,83. Os modelos estimados para Porto Alegre-RS apresentaram os melhores resultados, o que pode ser justificado pelo fato de tratar-se de um centro urbano denso, consolidado, com muitas amostras de AGSNs/NUIs. Também é o polo com melhores condições de infraestrutura e onde as diferenças entre áreas precárias e não precárias tendem a ser bem definidas.

O modelo construído para Brasília-DF, mesmo o baseado nos AGSNs, cuja amostra apresenta problemas (33% dos AGSNs não coincidem com NUIs – capítulo 5), apresentou bons resultados em comparação com os demais polos.

É possível que isso resulte do fato de os NUIs da região apresentarem características bastante distintas das áreas não delimitadas como NUIs.

Os modelos estimados para Juazeiro do Norte-CE e Marabá-PA apresentaram os piores resultados. Ambos os polos representam regiões menores, menos densas e com muitos NUIs isolados. Apresentam também piores indicadores socioeconômicos e de infraestrutura, que, como caracterizam os polos como um todo, incluindo áreas não delimitadas como NUIs, dificultam a distinção dos NUIs. O modelo "*Y* = AGSN" de Juazeiro do Norte-CE destaca-se como o modelo com pior ajuste. Trata-se de um resultado esperado, pois a amostra de AGSNs do polo, além de pequena, também apresenta muito ruído (como já dito, 32% dos AGSNs não coincidem com NUIs).

GRÁFICO 2
**Curva ROC dos modelos**

Elaboração dos autores.

TABELA 4
**Métricas obtidas pelos modelos**

| Métrica | Belo Horizonte | | Brasília | | Juazeiro do Norte | | Marabá | | Recife | | Porto Alegre | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Y* = AGSN | *Y* = NUI | *Y* = AGSN | *Y* = NUI | *Y* = AGSN | *Y* = NUI | *Y* = AGSN | *Y* = NUI | *Y* = AGSN | *Y* = NUI | *Y* = AGSN | *Y* = NUI |
| AUC | 0,717 | 0,727 | 0,78 | 0,803 | 0,632 | 0,72 | 0,676 | 0,692 | 0,726 | 0,755 | 0,808 | 0,828 |
| Kappa | 0,288 | 0,296 | 0,29 | 0,339 | 0,171 | 0,267 | 0,266 | 0,274 | 0,33 | 0,379 | 0,31 | 0,353 |

Elaboração dos autores.

### 4.3 Superfícies de probabilidade

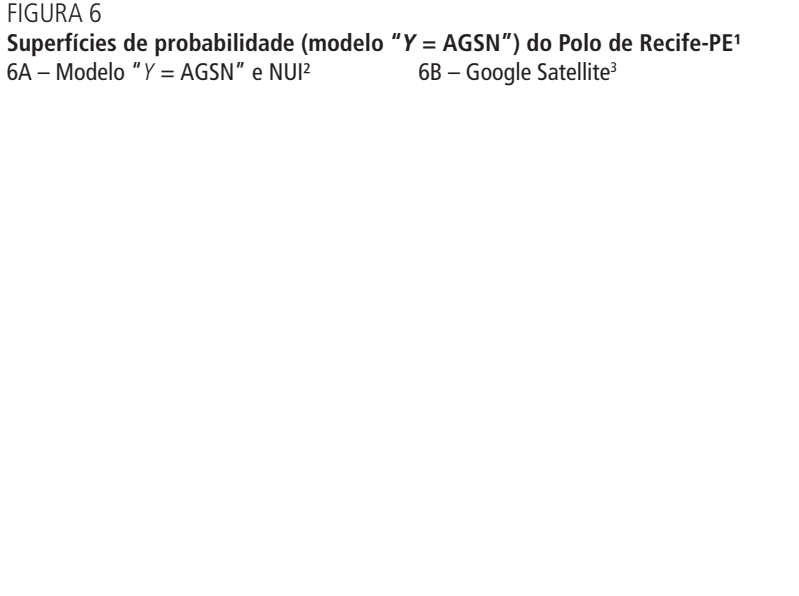
Os resultados dos modelos estimados podem ser mapeados e apresentados como superfícies de probabilidade, conforme apresentado nas figuras 5 e 6. As figuras revelam áreas indicadas com alta probabilidade de presença de NUI e que não foram originalmente mapeadas pelo IBGE como AGSN, mas que são, de acordo com a pesquisa de campo, núcleos urbanos informais. A figura 5 mostra os resultados do modelo "*Y* = AGSN" de Belo Horizonte-MG apresentado na forma de superfície de probabilidade, ilustrando como a superfície pode auxiliar na identificação de inúmeros NUIs não presentes na base de AGSN.

FIGURA 5
**Superfícies de probabilidade (modelo "*Y* = AGSN") do Polo de Belo Horizonte-MG[1]**
5A – Modelo "*Y* = AGSN" e AGSN[2] 5B – Modelo "*Y* = AGSN" e NUI[3]

Elaboração dos autores.
Notas: [1] Municípios de Ribeirão das Neves, Contagem e Esmeraldas.
[2] Sobreposição dos polígonos dos AGSNs (em preto).
[3] Sobreposição dos polígonos dos NUIs levantados em campo (em preto).

A figura 6 apresenta o bairro Brasília Teimosa, situado na zona sul de Recife-PE, que surgiu com a ocupação de uma área antes denominada Areal Novo, iniciada em 1947. Esse NUI, embora ausente na base de AGSN de 2019, foi destacado pela superfície de probabilidade estimada a partir desse dado.

FIGURA 6
**Superfícies de probabilidade (modelo "*Y* = AGSN") do Polo de Recife-PE[1]**
6A – Modelo "*Y* = AGSN" e NUI[2] 6B – Google Satellite[3]



Elaboração dos autores.
Notas: [1] Bairro Brasília Teimosa.
[2] NUI sobre a superfície de probabilidade (modelo *Y* = AGSN).
[3] Imagem orbital da área.

## 5 CONSIDERAÇÕES FINAIS

A ausência ou baixa qualidade dos dados sobre informalidade e precariedade habitacional no país representa um grande entrave para a elaboração de políticas que promovam condições dignas de moradia. A metodologia NUI busca contribuir para a elaboração ou revisão de levantamentos voltados à identificação de núcleos urbanos informais ocupados predominantemente por populações de baixa renda. Para alcançar esse objetivo, baseia-se em procedimentos de modelagem que envolvem a integração de dados secundários de várias fontes e naturezas distintas.

Para além de classificações binárias, que apontam se determinada área é ou não um NUI, optou-se pelo desenvolvimento de uma metodologia cujos resultados possam ser apresentados de forma contínua, como superfícies de probabilidade da presença de NUI, o que explicita tanto os diferentes níveis de precariedade quanto as incertezas inerentes aos resultados da classificação. Caso o analista prefira uma categorização binária, poderá fazê-lo a partir do estabelecimento de um limiar de probabilidade que julgue adequado. Considerou-se também fundamental a obtenção de modelos de fácil interpretação, que destacam características que predominam nos NUIs de cada região analisada.

Outro aspecto considerado importante para a construção da metodologia NUI é a capacidade de ser generalizável para todo o território nacional, o que demanda a utilização de dados disponíveis para todo o Brasil e de procedimentos aplicáveis a diferentes realidades, tais como a estimativa de modelos que possam combinar diferentes variáveis dependendo da área de estudo. Cabe salientar, entretanto, que a qualidade dos resultados obtidos pode variar, segundo as características da região analisada. Os estudos realizados para os seis polos da pesquisa indicam, por exemplo, que os modelos apresentam piores resultados em regiões onde a precariedade é generalizada e é mais difícil identificar as diferenças entre os NUIs e as demais áreas.

Os produtos resultantes da aplicação da metodologia NUI deverão servir como um plano de informação complementar aos levantamentos existentes sobre NUIs (por exemplo, levantamentos municipais e perímetros dos aglomerados subnormais do IBGE). A associação das superfícies de probabilidade a esses dados deverá dar origem a novos planos de informação que indiquem áreas onde é possível ter maior/menor incerteza sobre a presença/ausência de NUI e sirvam de referência para o planejamento de verificações em campo ou remota (por meio de análise de imagens orbitais). Busca-se, assim, contribuir para a produção de informações mais precisas e atualizadas, que evidenciem a diversidade de condições de irregularidade e precariedade, e possam subsidiar a elaboração e o aprimoramento de políticas e programas habitacionais.

## REFERÊNCIAS

AMORE, C. S.; MORETTI, R. "Gelo não é pedra!": informalidade urbana e alguns aspectos da regularização fundiária de interesse social na Lei 13.465/2017. **Cadernos da Defensoria Pública do Estado de São Paulo**, v. 3, n. 17, p. 73-83, ago. 2018.

BRASIL. Lei nº 13.465, de 11 de julho de 2017. Dispõe sobre a regularização fundiária rural e urbana, sobre a liquidação de créditos concedidos aos assentados da reforma agrária e sobre a regularização fundiária no âmbito da Amazônia Legal; e dá outras providências. **Diário Oficial da União**, Brasília, 8 set. 2017.

_____. Cadastro Único completa 18 anos. **Informe Bolsa e Cadastro**, n. 669, ago. 2019. Disponível em: <https://bit.ly/3mrYXtc>.

_____. Ministério do Desenvolvimento Social. **Famílias cadastradas no CadÚnico em 2019**. Brasília: MDS, 2020. Disponível em: <https://bit.ly/3Hbvcqf>.

CDHU – COMPANHIA DE DESENVOLVIMENTO HABITACIONAL E URBANO; UFABC – UNIVERSIDADE FEDERAL DO ABC. **Desenvolvimento e aplicação de metodologia para identificação, caracterização e dimensionamento de assentamentos precários**. São Bernardo do Campo: EdUFABC, 2019.

ESTADO DE SÃO PAULO. **Unidades homogêneas de uso e ocupação do solo urbano (UHCT) do Estado de São Paulo**. Secretaria do Meio Ambiente, 2014. Disponível em: <https://bit.ly/3pi3Cj9>. Acesso em: 12 maio 2022.

FEINSTEIN, A. R.; CICCHETTI, D. V. High agreement but low Kappa: I. the problems of two paradoxes. **Journal of Clinical Epidemiology**, v. 43, n. 6, p. 543-549, Jan. 1990.

FEITOSA, F. F. *et al*. (Ed.). **Metodologia para identificação e caracterização de assentamentos precários em regiões metropolitanas paulistas (Mappa)**. São Bernardo do Campo: Ed. UFABC, 2019.

FEITOSA, F. F. *et al*. IMMerSe: An integrated methodology for mapping and classifying precarious settlements. **Applied Geography**, v. 133, p. 102-494, Aug. 2021.

FERREIRA, M. P.; MARQUES, E. C. L.; FUSARO, E. R. Assentamentos precários no Brasil: uma metodologia para estimação e análise. *In*: MORAIS, M. P.; KRAUSE, C.; LIMA NETO, V. C. (Ed.). **Caracterização e tipologia de assentamentos precários**: estudos de caso brasileiros. Brasília: Ipea, 2016. p. 53-74.

FRIESEN, J. *et al*. Determining factors for slum growth with predictive data mining methods. **Urban Science**, v. 2, n. 3, 29 Aug. 2018.

FRIZZI, G.; PINHO, C. M. D. Índice de vulnerabilidade socioecológica para avaliação das remoções na cidade de São Paulo. *In*: ENCONTRO NACIONAL DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM PLANEJAMENTO URBANO E REGIONAL, 17., 2017, São Paulo. **Anais**... São Paulo, 2017.

HAIR JUNIOR, J. F. *et al*. **Análise multivariada de dados**. 6. ed. Tradução: Adonai Schlup Sant'Anna. Porto Alegre: Bookman, 2009.

HIJMANS, R. J.; ELITH, J. **Species distribution modeling with R. CRAN**. [*s.l.*]: 2017. Disponível em: <https://bit.ly/39hkQsn>.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Grade estatística**. Rio de Janeiro: IBGE, 2016. Disponível em: <https://bit.ly/3aCqAx5>.

_____. **Estimativas da população 2018**. Rio de Janeiro: IBGE, 2018. Disponível em: <https://bit.ly/3qUcZqj>.

_____. **Aglomerados subnormais 2019**: classificação preliminar e informações de saúde para o enfrentamento à covid-19. Rio de Janeiro: IBGE, 2020. Disponível em: <https://bit.ly/3wSmeex>. Acesso em: 23 ago. 2021.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. New York: Springer, 2013.

MAHABIR, R. *et al*. Detecting and mapping slums using open data: a case study in Kenya. **International Journal of Digital Earth**, p. 1-25, 4 Dec. 2018.

MARICATO, E. Informalidade urbana no Brasil: a lógica da cidade fraturada. *In*: WANDERLEY, L. E. W.; RAICHELIS, R. (Ed.). **A cidade de São Paulo**: relações internacionais e gestão pública. São Paulo: Educ, 2009. p. 269-294.

MARQUES, E. (Coord.). **Assentamentos precários no Brasil urbano**. São Paulo: CEM/Cebrap; Brasília: SNH/MCidades, 2007.

_____. **Diagnóstico dos assentamentos precários nos municípios da macrometrópole paulista** – segundo relatório. São Paulo: CEM/Cebrap; Fundap, ago. 2013. Disponível em: <https://bit.ly/3uJW8HN>.

PHILLIPS, S. J. *et al*. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. **Ecological Applications**, v. 19, n. 1, p. 181-197, Jan. 2009.

PHILLIPS, S. J.; ELITH, J. Logistic methods for resource selection functions and presence-only species distribution models. *In*: ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE, 25., 2011, San Francisco, California. **Anais**... 2011.

PNUD – PROGRAMA DAS NAÇÕES UNIDAS PARA O DESENVOLVIMENTO; IPEA – INSTITUTO DE PESQUISA ECONÔMICA APLICADA; FJP – FUNDAÇÃO JOÃO PINHEIRO. **Desenvolvimento humano nas macrorregiões brasileiras**: 2016. Brasília: PNUD, 2016. Disponível em: <https://bit.ly/3DyjHHO>. Acesso em: 8 abr. 2020.

RIBEIRO, B. M. G. Mapping informal settlements using WorldView-2 imagery and C4.5 decision tree classifier. *In*: JOINT URBAN REMOTE SENSING EVENT, 2015, Lausanne. **Anais**… June 2015. Disponível em: <https://bit.ly/3uMeJTD>. Acesso em: 28 jul. 2020.

SOUZA, M. C. P.; ROSSI, M. T. B.; RUDGE, M. de S. Mapeamento colaborativo de assentamentos precários em regiões metropolitanas paulistas. *In*: SEMINÁRIO NACIONAL SOBRE URBANIZAÇÃO DE FAVELAS, 3., 2018, Salvador, Bahia. **Anais**… Salvador: Ed. UCSal, 2018. Disponível em: <https://bit.ly/3LBtY8y>. Acesso em: 24 jun. 2020.