



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Identificação de áreas precárias em Marabá – PA por Regressão Logística clássica e Regressão Logística Geograficamente Ponderada

Bruno Dias dos Santos

Monografia apresentada como requisito parcial da disciplina de Análise Espacial de Dados Geográficos, ministrado por Dr. Eduardo G. Camargo e Dr. Antônio Miguel Vieira Monteiro, do Curso de Pós-Graduação em Sensoriamento Remoto no Instituto Nacional de Pesquisas Espaciais

INPE
São José dos Campos
2021

RESUMO

A definição e caracterização de assentamentos precários não considera a realidade de grande parte dos municípios brasileiros, especialmente para os municípios da região amazônica. Além disso, as abordagens que utilizam técnicas de geoprocessamento e sensoriamento remoto para identificar assentamentos precários no Brasil estão concentradas em regiões metropolitanas. Portanto, este trabalho tem como objetivo geral o desenvolvimento de uma metodologia para identificação de áreas precárias em cidades da Amazônia, tendo como estudo de caso a cidade de Marabá, no sudeste do estado do Pará. Para isso, foram elaborados modelos de regressão logística, um clássico e outro geograficamente ponderado (GWLR), modelados a partir de variáveis agregadas em dimensões de análise (Morfológica, Ambiental e Domiciliar), baseados em dados censitários, de informação geográfica voluntária (VGI) e dados extraídos de sensoriamento remoto. As variáveis foram integradas em um espaço celular – composto de células hexagonais com 100m de diâmetro. A base de dados foi dividida em base desenvolvimento e em base de validação e dados não rotulados foram utilizados para previsão dos modelos. A partir das superfícies de probabilidades, foi aplicado um corte para classificação das áreas precárias. O corte escolhido maximizou o F1-Score de ambos os modelos e resultou em níveis de acurácias globais acima de 0,84 para as bases de validação, indicando resultados de boa qualidade. O uso de modelos de regressão logística possibilitou a identificação de níveis da precariedade e o modelo GWLR possibilitou espacializar as diferenças dos valores dos coeficientes β .

Palavras-chaves: Áreas precárias. Amazônia. GWLR. Regressão Logística. Urbanização.

SUMÁRIO

	<u>Pág.</u>
1 INTRODUÇÃO	1
2 ÁREA DE ESTUDO	2
3 MATERIAIS E MÉTODOS	3
3.1 Caracterização dos Assentamentos Precários	4
3.2 Preparação dos dados.....	5
3.3 Integração em espaço celular	8
3.4 Seleção de variáveis.....	9
3.5 Regressão Logística.....	10
3.6 Regressão Logística Geograficamente Ponderado	12
4 DISCUSSÃO DOS RESULTADOS	14
4.1 Avaliação do Modelo de Regressão Logística clássico	14
4.2 Avaliação do Modelo de Regressão Logística Geograficamente Ponderado (GWLR)	19
4.3 Comparação entre os modelos.....	22
5 CONSIDERAÇÕES FINAIS	26
REFERÊNCIAS BIBLIOGRÁFICAS	27

1 INTRODUÇÃO

A Política Nacional de Habitação (PNH) adotou sob peso de lei o termo “assentamentos precários” como uma categoria de assentamentos habitacionais urbanos que se encontram em situações de precariedade e inadequação (BRASIL, 2005). Conforme o Ministério das Cidades (BRASIL, 2010), os assentamentos precários são segmentos do território urbano de grandezas e tipologias variáveis, mas que mantêm em comum o caso de serem áreas predominantemente residenciais e ocupadas por moradores - em sua maioria - de baixa renda, caracterizadas por inúmeras carências e inadequações das condições de moradia.

Os moradores desses assentamentos utilizam múltiplas estratégias para viabilizar, de modo autônomo, soluções para as suas necessidades habitacionais, pois nem o Estado, nem o Mercado Formal atendem as suas demandas (BRASIL, 2010). Essa viabilização autônoma resulta num processo de autoconstrução da moradia que tem como resultado a apropriação de áreas consideradas irregulares nas cidades (MARICATO, 1982).

De modo geral, esses assentamentos englobam inúmeras tipologias, com destaque aos cortiços, favelas, loteamento irregulares de baixa renda e conjuntos habitacionais produzidos pelo poder público que se encontram degradados (DENALDI; ROSA, 2010). Dentre estas tipologias, os cortiços e as favelas são os dois tipos de assentamentos mais tratados pela literatura (ABREU, 1994; VAZ, 1994; PASTERNAK, 2006; QUEIROZ FILHO, 2015).

Alguns trabalhos que utilizam técnicas automatizadas de identificação desses assentamentos habitacionais incluem em sua metodologia o conceito de áreas precárias (DOS SANTOS; DE PINHO; DE JESUS, [s.d.]; CDHU; UFABC, 2018; GONÇALVES, 2018). Apesar de não existir uma definição oficial, áreas precárias podem ser descritas como regiões que concentram assentamentos precários. Dessa forma, todo assentamento precário está contido em uma área precária, mas nem todo território presente numa área precária pode ser considerado um assentamento precário.

No entanto, a definição e a caracterização dos assentamentos precários que foi instituída pela PNH não contempla a realidade de grande parte dos municípios brasileiros

(SAKATAUSKAS, 2020), especialmente para os municípios periurbanos¹ da região amazônica. Além disso, as abordagens metodológicas que utilizam técnicas de geoprocessamento e sensoriamento remoto empregadas para a identificação de áreas e assentamentos precários no Brasil estão concentradas em regiões metropolitanas ou em regiões de altas densidade construtiva (KUFFER; PFEFFER; SLIUZAS, 2016).

Segundo Santana e colaboradores (2012), existe a necessidade de identificar peculiaridades existentes nos assentamentos precários dos municípios da Amazônia, bem como incorporar diversidades regionais, municipais e urbanas quando se trata da elaboração de políticas públicas habitacionais para a região. O entendimento sobre como é caracterizada as áreas precárias em cidades amazônicas é importante para visibilizar as desigualdades sociais, econômicas e regionais, e auxilia na estruturação de instrumentos de luta pela democratização da política urbana e habitacional no Brasil (SAKATAUSKAS; SANTANA, 2015).

Motivado pelo contexto apresentado, questiona-se: Como as áreas precárias em municípios amazônicos são retratados na literatura? Quais dimensões de análise poderiam ser úteis para analisar essas áreas precárias? É possível identificar essas áreas utilizando uma regressão espacial?

Para responder essas questões, este trabalho tem como objetivo geral o desenvolvimento e implementação de uma metodologia para identificar áreas precárias em Marabá, no sudeste do estado do Pará. O trabalho propõe uma Regressão Logística clássica (RL) e uma Regressão Logística Geograficamente Ponderada – *Geographically Weighted Logistic Regression* (GWLR), modeladas a partir de informações extraídas de dados censitários, de informações geográficas voluntárias e de sensoriamento remoto.

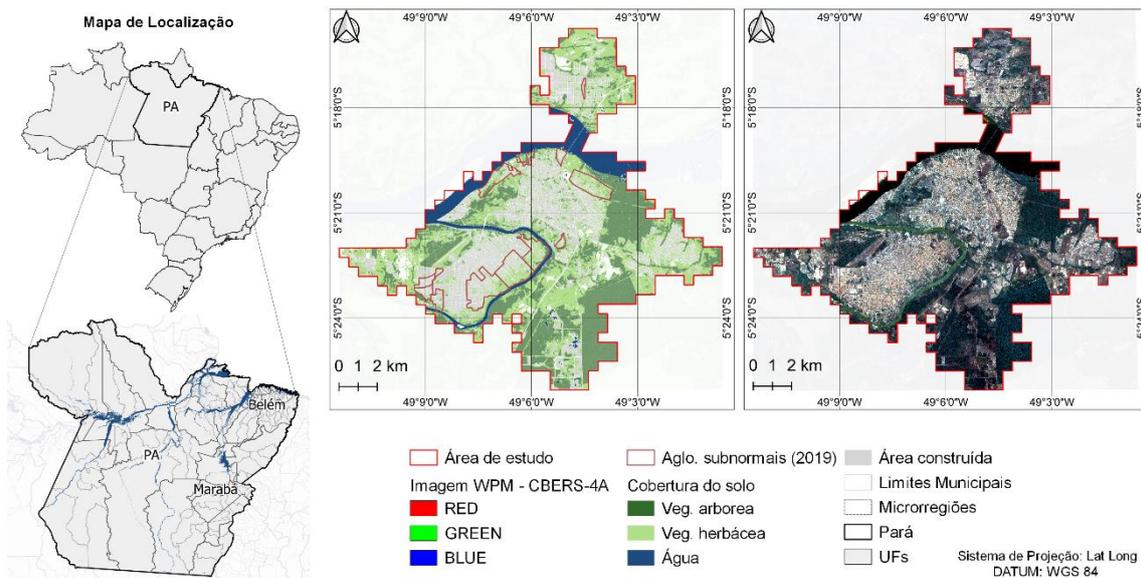
2 ÁREA DE ESTUDO

A Área de estudo refere-se a área urbana que cruza a sede de Marabá – definida a partir de dados de luzes noturnas. Marabá foi fundada em 1913 e localiza-se no Sudeste

¹ A concepção de “periurbano” se aproxima do conceito de “padrão periférico de urbanização desenvolvido por Ermínia Maricato, Raquel Rolnik, Nabil Bonduki e outros, durante as décadas de 1970 e 1980 (SAKATAUSKAS; SANTANA, 2015). Este termo também é utilizado para designar áreas de transições urbano-rurais e que podem ou não estar nos arredores das grandes cidades (NAKANO, 2012).

Paraense, entre dois grandes rios: Itacaiúnas e Tocantins. A cidade de Marabá divide-se em cinco núcleos urbanos, num município que possui mais de 15 mil km². A população estimada é de 205.753 habitantes (CAMARA MUNICIPAL DE MARABÁ, 2021). A cidade é cortada pelas rodovias Transamazônica (BR-230) e pela Rodovia Paulo Fonteles (BR-150). A porção delimitada como Área de estudo possui 1.473 km² e pode ser visualizada na Figura 2.1.

Figura 2.1: Mapa de localização da Área de estudo.



Fonte: Produção do autor.

3 MATERIAIS E MÉTODOS

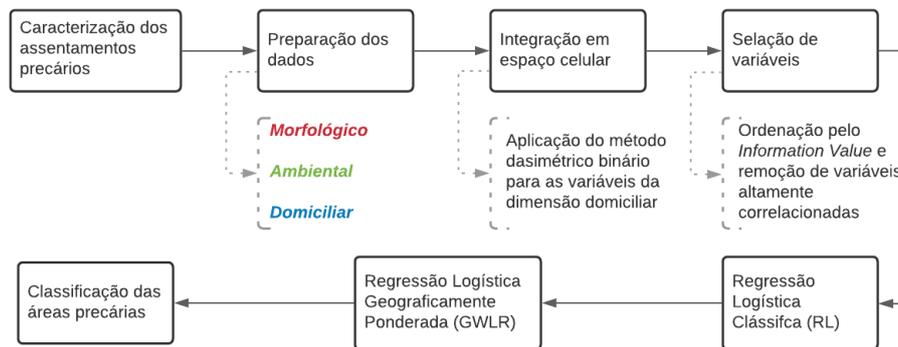
Neste Capítulo é apresentada a metodologia desenvolvida para a criação da Regressão Logística Tradicional (RL) e da Regressão Logística Geograficamente Ponderada (GWLR) e a posterior identificação de áreas precárias de Marabá – PA. Todas as etapas foram realizadas utilizando os softwares:

- QGIS 3.18 (TEAM, 2021): Para preparação da base de dados e construção dos mapas temáticos;
- GeoDMA 2.0.1 (KÖRTING; GARCIA FONSECA; CÂMARA, 2013): Para extração de atributos e preenchimento da grade celular;
- RStudio (ALLAIRE, 2012): Para construção e avaliação do modelo de regressão logística tradicional;
- GWR4 (NAKAYA et al., 2014): Para construção e avaliação do modelo de regressão logística geograficamente ponderado;

- GeoDA (ANSELIN; SYABRI; KHO, 2010): Para avaliação da correlação espacial entre os resíduos da regressão logística tradicional.

A metodologia pode ser sintetizada em oito etapas (Figura 3.1): elaboração de uma ontologia dos assentamentos precários, construção das variáveis, integração em espaço celular, seleção das variáveis, construção do modelo de regressão logística tradicional, construção do modelo de regressão logística geograficamente ponderada, geração das superfícies de probabilidades e classificação das áreas precárias.

Figura 3.1: Etapas para a elaboração do Índice de Precariedade.



Fonte: Produção do autor.

3.1 Caracterização dos Assentamentos Precários

Cardoso et al. (2016) delimitaram e descreveram os padrões morfológicos de assentamentos urbanos em seis cidades paraenses: Marabá (local da nossa área de estudo), Parauapebas, Canaã dos Carajás, São Felix do Xingu (as quatro contidas no Sudeste Paraense), Santarém (Baixo Amazonas, Sudeste Paraense) e Altamira (Sudoeste Paraense). Os autores apresentam as tipologias de assentamento tradicional e assentamento informal dentre outros padrões morfológicos discutidos.

Os assentamentos tradicionais possuem um traçado orgânico e moldado ao sítio natural, com predominância dos cheios sobre os vazios e uma forte conexão entre a área construída e os espaços livres (PONTES, 2015). As áreas livres presentes em espaços tradicionais funcionam como extensões das casas, entrelaçando a economia urbana à vida cotidiana. Por sua vez, os assentamentos informais podem ser entendidos como releituras reduzidas desses assentamentos tradicionais (CARDOSO; DE MELO; DO VALE GOMES, 2016), sendo divididos entre informais *espontâneos* e informais *organizados*.

Os assentamentos informais espontâneos ocupam áreas ambientalmente sensíveis nas periferias das cidades, como regiões de várzeas e beiras de rios e possuem uma maior irregularidade no traçado das quadras quando comparadas ao padrão tradicional. As ruas desses assentamentos chegam a possuir 400 metros de comprimento, dificultando a formação de centralidades e a mobilidade de pedestres. Nessas ocupações ocorre uma progressiva degradação ambiental, com aterramento de rios e remoção de mata ciliar e vegetação nativa (CARDOSO; DE MELO; DO VALE GOMES, 2016).

Sob a mesma lógica de racionalização do território que o setor privado utiliza, os assentamentos informais organizados são resultados da conversão prematura de áreas rurais com fragilidade fundiária para áreas urbanas (CARDOSO; DE MELO; DO VALE GOMES, 2016). Tantos os assentamentos informais organizados, como os assentamentos informais espontâneos são portadores de pobreza ao reduzirem o tamanho do lote, numa forma de ocupação que impede o aproveitamento do terreno e da rua para atendimento de necessidades de consumo e dificulta a atenuação da carência de renda vivida pelos moradores (PONTES, 2015; CARDOSO; DE MELO; DO VALE GOMES, 2016).

Assim como os assentamentos tradicionais, alguns assentamentos informais estão localizados em áreas de cotas mais baixas. No entanto, há uma postura diferente de um ribeirinho para um morador de um assentamento informal diante das inundações decorrentes das cheias sazonais. Ao contrário das áreas informais, no modo de vida ribeirinho as cheias não remetem a um risco ambiental e as edificações são pensadas para comportá-los inclusive nos períodos em que ocorrem inundações (PONTES, 2015; CARDOSO; DE MELO; DO VALE GOMES, 2016).

3.2 Preparação dos dados

Semelhante à metodologia apresentada pelo Projeto MAPPA (CDHU; UFABC, 2018), foram pensadas variáveis que pudessem representar as áreas precárias em seus diferentes aspectos. As variáveis foram agrupadas em três Dimensões de análises, de acordo com as suas características semânticas:

- **Morfológica:** Esta dimensão refere-se às características de ocupação, ordenamento territorial e condições de acesso pela infraestrutura viária. Essa dimensão foi inspirada no método de análise morfológica de Conzen (1960; 2004), fundador da escola inglesa de morfologia urbana, que combina padrões das

edificações, padrões de uso do solo e o plano da cidade – ou seja, a análise das ruas, lotes, quadras e edificações dentro das quadras. No entanto, as ferramentas de desenho urbano e sintaxe espacial foram formulados em sociedades com alto nível de urbanização e industrialização, e com forte tradição cadastral – diferentemente dos municípios da região amazônica (CARDOSO et al., 2020). A característica híbrida das cidades amazônicas (de diversidade de paisagens, complexa formação social e peculiaridades no processo de formação) requer adaptações do ferramental utilizado numa análise morfológica tradicional. A Tabela 3.1 descreve os indicadores selecionados para a representar esta dimensão e suas respectivas justificativas.

Tabela 3.1: Indicadores representantes da dimensão Morfológica.

Variável	Justificativa
Mapas axiais adaptados	Neste trabalho, foram utilizadas duas escalas de análise para produção de mapas axiais de <i>integração</i> e <i>escolha (choice)</i> : a escala da Microrregião de Marabá e a escala da área de estudo. Para a escala da Microrregião de Marabá, além das rede viária, é utilizado também a hidrografia no computo das métricas – já que os rios possuem um importante papel na locomoção da população. Além das análises que consideravam todas as vias, foram realizadas análises com os seguintes valores de raios (em metros): 400m, 800m, 2500m e 5000m. A base de vias utilizadas foi adquirida do Open Street Maps (OSM, 2021).
Área construída	Dentro da racionalidade urbano desenvolvimentista, milênios de convivência harmônica entre população e floresta são desconsiderados. A cidade se coloca como a antítese do mato, local da civilidade e do progresso e legítima a destruição da natureza e das populações originais (CARDOSO; DE MELO; DO VALE GOMES, 2016). Os assentamentos precários representam uma das tipologias morfológicas gestadas por esse tipo de racionalidade. O aumento de área construída sobre as áreas de mata representa: a diminuição da permeabilização do solo; dificulta a infiltração da água; aumenta do escoamento da água da chuva em direção aos rios, causando aumento da vazão de forma mais rápida do que natural; relaciona-se com processos de alagamento; e ocasiona aumento do temperatura por meio de ilhas de calor (BRAGA et al., 2005). Foi identificada como áreas construída os pixels que possuíam um valor de NDVI entre 0 e 0.3 e um valor acima de 0 para a diferença normalizada entre a banda azul e a banda verde, das imagens provenientes do Sensor WPM do satélite CBERS-4A.
Período da construção	Um palimpsesto urbano refere-se à ideia de períodos construtivos, onde mediante acumulações e substituições, as ações das diferentes gerações se superpõem. A forma urbana traduz o registro das ações civis e públicas e que delas pode-se aprender qual ideologia norteou a ocupação do solo ao longo do tempo (CONZEN, 1960; HARVEY, 1996; SANTOS, 2002; COSTA; NETTO, 2015). O período da área construída foi obtido a partir da base de Multi-temporal GHS-BUILT (PESARESI et al., 2016).
Métricas Texturais	Reub (2017) utilizou apenas métricas texturais, <i>Gray Level Co-occurrence Matrices</i> (GLCMs) (HARALICK; SHANMUGAM; DINSTEN, 1973), para identificar assentamentos precários nas regiões metropolitanas de São Paulo e Rio de Janeiro. O autor testou diferentes tamanhos de kernel para computo das métricas e obteve níveis de acurácia globais acima de 0,8 quando comparados aos aglomerados subnormais. Na metodologia proposta neste trabalho, utilizando uma imagem WPM do CBERS-4A e com Kernel de 3x3,

foram calculados: contraste, correlação, medidas de informação de correlação, uniformidade, variância e entropia.

Shape Index

Os assentamentos informais tendem a ter uma maior irregularidade das quadras quando comparados com assentamentos tradicionais (CARDOSO; DE MELO; DO VALE GOMES, 2016). Este dado permite a análise da regularidade geométrica das quadras em áreas mais regulares (quadras retangulares com tamanhos parecidos) e a irregularidade geométrica ou mesmo a ausência de subdivisões em quadras, comumente característicos de assentamentos precários (CDHU; UFABC, 2018). O SI foi calculado o pela Equação 1 (MCGARIGAL; MARKS, 1995), onde p é perímetro do polígono (ou da quadra, em metros) e a é a área do polígono (ou da quadra, em metros quadrados).

$$SI = \frac{p}{2\sqrt{\pi^{\circ} a}} \quad (1)$$

Fonte: Produção do autor.

- **Ambiental:** Variáveis do meio físico-ambiental relacionadas aos processos de inundação e alagamento, proximidade aos corpos hídricos e adensamento construtivo. Foram utilizadas como variáveis: A área verde, as Áreas de Proteção Permanente (APP) e a altura do terreno em relação a drenagem mais próxima. Tabela 3.2 descreve os variáveis escolhidas para a representar esta dimensão e suas respectivas justificativas.

Tabela 3.2: Indicadores representantes da dimensão Ambiental.

Variável	Justificativa
Áreas verdes	A presença de vegetação nas áreas urbanas oferece uma série de serviços ecossistêmicos, como a oferta de alimentos e matéria prima, regulação de erosão, de enchentes, de microclimas e da poluição, que direta ou indiretamente impactam a saúde humana (BRAGA et al., 2005). Neste trabalho, foi utilizado como indicativo da presença de áreas verdes o <i>Normalized Difference Vegetation Index</i> (NDVI).
Áreas de proteção permanente	Os assentamentos informais ocupam áreas ambientalmente sensíveis nas periferias das cidades, como regiões de várzeas e beiras de rios, com intensa degradação de rios e remoção de mata ciliar (CARDOSO; DE MELO; DO VALE GOMES, 2016). As APP, conforme definição da Lei n. 12.651/2012, é uma área protegida, coberta ou não por vegetação nativa, com a função ambiental de preservar os recursos hídricos, a paisagem, a estabilidade geológica e a biodiversidade, facilitar o fluxo gênico de fauna e flora, proteger o solo e assegurar o bem-estar das populações humanas.
Altura do terreno em relação a drenagem mais próxima (HAND)	Alguns assentamentos informais estão localizados em áreas de cotas mais baixas (PONTES, 2015; CARDOSO; DE MELO; DO VALE GOMES, 2016). No entanto, diferentemente das populações tradicionais, as habitações não são bem-preparadas para as inundações decorrentes das cheias sazonais.

Fonte: Produção do autor.

- **Domicílio e Entorno:** Inclui variáveis obtidas dos dados do universo, agregados por setor censitário, do Censo Demográfico de 2010 (INSTITUTO BRASILEIRO

DE GEOGRAFIA E ESTATÍSTICA, 2011). Compõem essa dimensão: a razão de dependência, o percentual de domicílios com destinação inadequada de lixo, o percentual de domicílio com rendimento per capita de até meio salário mínimo e a sobreposição à área urbana delimitada pelo Censo de 2010. A Tabela 3.3 descreve os indicadores selecionados para a representar esta dimensão e suas respectivas justificativas.

Tabela 3.3: Indicadores representantes da dimensão Domiciliar.

Variável	Justificativa
Proporção de domicílios com destinação inadequada de lixo	Segundo Sakatauskas (SAKATAUSKAS, 2020), o serviço de coleta de lixo é realizado de forma irregular em muitos assentamentos precários amazônicos, em especial nas áreas cujo acesso é dificultado, causando acúmulo de lixo nas áreas alagadas e nas ruas. Neste trabalho, foi adotado como destinação inadequada de lixo uma adaptação do conceito estabelecido pela Fundação João Pinheiro (FJP, 2013): o aterramento na propriedade; lixo jogado em rio, lago, mar ou outras destinações sem ser coleta.
Proporção de domicílios com renda per capita de até meio salário mínimo	Segundo a PNH (BRASIL, 2005), os assentamentos precários em todas as suas tipologias carregam em comum o fato de serem áreas predominantemente residenciais, habitadas por famílias de baixa renda.
Razão de Dependência	Segundo a Da Motta (2017), a razão de dependência em aglomerados subnormais é superior ao dos outros setores censitários. Isso significa que, além terem menos renda, os moradores de áreas precárias ainda precisam dar conta de mais dependentes na família, resultando em uma redução da renda per capita.
Sobreposição à área urbana	Segundo a PNH (BRASIL, 2005), os assentamentos precários são porções do território urbano.

Fonte: Produção do autor.

3.3 Integração em espaço celular

As variáveis foram integradas em uma grade celular hexagonal de 100 metros de diagonal principal e área de 8.661 m². Os hexágonos são as unidades espaciais de projeto utilizadas como referência para agregar os dados de diferentes tipos e fontes. A grade hexagonal foi escolhida pois nos permite entender melhor os fenômenos espaciais com um componente de vizinhança e conectividade. Definimos o tamanho do hexágono com base em estudos que identificaram assentamentos precários em municípios brasileiros usando grades celulares (DOS SANTOS; DE PINHO; DE JESUS, [s.d.]; CDHU; UFABC, 2018; GONÇALVES, 2018; FEITOSA et al., 2021) e por 100 m ser o comprimento de uma quadra formal. Além disso, esse tamanho de célula permite a diferenciação entre o vizinho sem comprometer a visualização dos padrões internos.

Para o preenchimento das células, foi adotado o operador de *média ponderada pela área* para as variáveis do tipo vetorial² e para as variáveis em formato matricial, foram extraídos os valores de *média, máximo, mínimo, soma e variância*. Foi aplicado o método dasimétrico binário para as variáveis da dimensão domiciliar, onde os dados foram redistribuídos somente entre as células com presença de área construída. Após o processo de preenchimento de atributos, cada célula possuía 152 variáveis.

Já nas células, os atributos foram normalizados pelo método *MinMax*, conforme Equação 2. Desse modo, todos os atributos passaram a variar entre 0 (mínimo) e 1 (máximo).

$$\text{Valor Normalizado} = \frac{(\text{valor do pixel} - \text{valor mínimo})}{(\text{valor máximo} - \text{valor mínimo})} \quad (2)$$

A grade celular hexagonal é formada por 17.632 hexágonos. Foi realizada uma amostragem aleatória estratificada, sendo selecionadas 1.007 amostras de área precárias – definidas a partir da sobreposição com os aglomerados subnormais IBGE, 2020) – e 3.013 amostras de áreas não precárias. Os outros 13.612 hexágonos são células não rotuladas utilizadas para a previsão dos modelos de regressão.

Os dados rotulados foram divididos em dois: (a) a base de desenvolvimento, representando 70% das amostras utilizados na definição dos modelos e (b) a base de validação, onde os modelos desenvolvidos foram validados para avaliar se a metodologia pode ser replicada se as mesmas condições forem mantidas.

3.4 Seleção de variáveis

Para reduzir a quantidade de atributos de cada célula, foram adotadas duas abordagens de seleção de variáveis: o ranqueamento pelo *Information Value (IV)* e a remoção de variáveis correlacionadas. De acordo com Lin (2013), o *Information Value* visa eliminar variáveis independentes que possuem baixo poder explicativo sobre a variável resposta, quanto esta última possui uma distribuição binária. Nesse método, as variáveis são categorizadas em faixas que amplifiquem a medida de IV de acordo com a Equação 3:

$$\sum_1^{n^{\circ} \text{ de categorias}} (\% \text{ evento} - \% \text{ não evento}) \times \ln \frac{\% \text{ evento}}{\% \text{ não evento}} \quad (3)$$

² Apenas as métricas GLCMs, NDVI e HAND não são variáveis do tipo vetorial

Como resultado, para cada atributo foi obtido um valor de IV igual ou acima de 0, com as variáveis independentes com maior poder explicativo sobre a variável resposta possuindo os valores mais altos. Apesar da categorização das variáveis para cálculo do IV, foram utilizadas as variáveis em seu formato contínuo.

Neste trabalho, as variáveis com IV inferior a 0,25 foram excluídas. Após essa primeira retirada das variáveis, foi realizada uma segunda eliminação, retirando as variáveis com correlação (*de Pearson*) superior a 0,65, analisadas duas a duas.

Todas essas eliminações de variáveis visam reduzir a redundância de informações, uma vez que variáveis correlacionadas tendem a medir as mesmas informações e variáveis com baixo IR têm baixo poder preditivo, não sendo úteis para modelagem. Ao final do processo, restavam apenas 18 variáveis.

3.5 Regressão Logística

Um modelo de regressão logística resulta na probabilidade de ocorrência de um evento que pode ser estimada a partir de um conjunto de variáveis independentes. A variável dependente possui uma distribuição binomial, com dois estados possíveis (1 ou 0) e há um conjunto de n variáveis independentes categóricas. O modelo de regressão logística pode ser descrito como Equação 4 (STRAUSS, 1992):

$$P(Y = 1 | X_1 = x_1, \dots, X_n = x_n) = \frac{1}{1 + e^{-g(x_1, \dots, x_n)}}$$
$$e \ g(x_1, \dots, x_n) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \varepsilon \quad (4)$$

Os coeficientes $\beta_0, \beta_1, \dots, \beta_n$ são estimados a partir do conjunto de dados, buscando maximizar a probabilidade de a amostra ter sido observada. Os coeficientes representam a importância de cada variável explicativa para a ocorrência do evento (HAIR et al., 2009) e suas estimativas geralmente são calculadas através do método da máxima verossimilhança (HOSMER; LEMESHOW; STURDIVANT, 2000). Probabilidades próximas a 1 indicam que o estado de $Y = 1$ é muito provável e probabilidades próximas a 0 indicam o contrário.

O modelo de regressão logística foi desenvolvido no RStudio com base nas variáveis integradas na grade celular. O modelo foi gerado apenas com dados referentes à base de desenvolvimento. A variável dependente (Y) foi definida como 1 se a célula for precária

e 0 se for não precária. O resultado da equação é a probabilidade da ocorrência de Y, variando entre 0 (não precária) e 1 (precária). As variáveis que caracterizam a precariedade são as variáveis independentes ($x_1..x_n$).

Foi utilizado a técnica *Stepwise* (modo *both*) para selecionar variáveis. Nesse método, é avaliado a influência no modelo da entrada e da retirada de cada variável independente. Foi utilizado o Critério de Informação de Akaike (AIC) para decidir qual o melhor modelo a ser utilizado. O AIC pode ser calculado pela Equação 5:

$$AIC = 2p + n \log\left(\frac{SQE}{n}\right) \quad (5)$$

O AIC é uma combinação entre uma medida de ajuste (*SQE*) e uma medida de simplicidade do modelo (dado pelo número de parâmetro p). Quanto menor for o valor de AIC, melhor o modelo.

A avaliação dos resíduos do modelo de regressão logística foi realizada pela técnica gráfica de “envelope simulado”. De acordo com Atkinson (1981), é possível gerar o envelope com os seguintes passos³:

1. Obtenha os resíduos padronizados para o modelo de regressão logística;
2. Gere valores simulados para estes resíduos;
3. Construa bandas de confiança para comparar os resíduos observados com percentis da distribuição normal padrão; e,
4. Se o modelo estiver adequado, os resíduos padronizados do modelo devem estar dentro das bandas de confiança.

Além da avaliação dos resíduos da regressão logística pela técnica gráfica do envelope, foi avaliado a existência de correlação espacial dos resíduos a partir do Índice Global de Moran e do Índice Local de Associação Espacial - *Local Indicators of Spatial Association* (LISA), usando uma matriz de vizinhança considerando os seis vizinhos mais próximos.

O Índice Global de Moran, fornece uma medida geral da associação espacial existente no conjunto dos dados, variando de -1 a +1. Dados com baixa associação espacial, resultam

³ Neste trabalho foi utilizado o script em R fornecido por (PAULA, 2004) para a geração do envelope.

em um índice baixo, próximos a 0. Já valores positivos e negativos significam autocorrelação espacial positiva e negativa, respectivamente (CÂMARA et al., 2004).

Enquanto o Índice de Moran nos fornece um valor único como medida da associação espacial do conjunto de dados, o LISA nos fornece indicadores locais ao produzir um valor específico para cada objeto, permitindo a identificação de agrupamentos de objetos com valores de atributos semelhantes (*clusters*), *outliers* e de mais de um regime espacial (CÂMARA et al., 2004).

3.6 Regressão Logística Geograficamente Ponderado

O modelo de regressão logística clássica é um modelo global que não leva em consideração a dimensão e a variação espacial, produzindo um único valor de coeficiente β para cada variável independente. Em contraste, nos modelos de GWLR a relação entre a variável dependente e as variáveis independentes variam localmente. Cada observação é modelada separadamente e inclui apenas outras observações dentro de uma vizinhança (*kernel*), determinada por uma distância fixa ou uma distância adaptativa (com base na densidade do ponto de amostra) da observação modelada, onde os pontos mais próximos do ponto de regressão são mais pesados do que aqueles mais longe (MEDINA, 2016; MAYFIELD et al., 2018; NUR; AL HARIS, 2021).

O modelo GWLR produz múltiplos coeficientes β , um para cada observação na amostra. Brundson e colaboradores (2002, 1996) reescreveram a Equação 4, incorporando as coordenadas geográficas (Equação 5):

$$g(x_1, \dots, x_n)_i = \beta_0(\mu_i, v_i) + \beta_1(\mu_i, v_i)x_1 + \dots + \beta_n(\mu_i, v_i)x_n + \varepsilon_i \quad (5)$$

Onde (μ_i, v_i) é o elemento espacial que representa a localização geográfica do ponto i , e $\beta_0, \beta_1, \dots, \beta_n$ são os valores dos coeficientes para a observação de localização i . No GWLR estimação dos coeficientes β pode é descrita pela Equação (6):

$$\hat{\beta}(\mu_i, v_i) = (x^t w(\mu_i, v_i) x)^{-1} x^t w(\mu_i, v_i) g \quad (6)$$

Onde:

$$w(\mu_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{bmatrix}$$

Note que $w(\mu_i, v_i)$ é uma matriz diagonal e distinta para cada ponto i de coordenadas (μ_i, v_i) , contendo uma diagonal principal de pesos w_{ij} obtidas por meio das funções de ponderação (*kernel*). As duas principais funções de ponderação encontradas na literatura são as funções Gaussiana e a função Biquadrática. As fórmulas de ambas as funções estão contidas na Tabela 3.4:

Tabela 3.4: Funções de Ponderação para GWLR.

Funções de Ponderação	Fórmula das Funções de Ponderação
Gaussiana Fixa	$w_{ij} = \exp\left\{-\frac{1}{2}(d_{ij}/b)^2\right\}$
Biquadrática Fixa	$w_{ij} = [1 - (d_{ij}/b)^2]^2$ se $d_{ij} < b$, e $w_{ij} = 0$ caso contrário
Gaussiana Variável	$w_{ij} = \exp\left\{-\frac{1}{2}(d_{ij}/b_{i(k)})^2\right\}$
Biquadrática Variável	$w_{ij} = [1 - (d_{ij}/b_{i(k)})^2]^2$ se $d_{ij} < b_{i(k)}$, e $w_{ij} = 0$ caso contrário

Fonte: (BRUNSDON; FOTHERINGHAM; CHARLTON, 2002), adaptado de (MEDINA, 2016).

Tanto para função Gaussiana, quanto para a função Biquadrática, há uma diferenciação por meio da escolha do parâmetro b (*bandwidth*) a ser utilizado (se fixo ou variável). O parâmetro d_{ij} contido nas funções de ponderação representa a distância do ponto i ao ponto j , o parâmetro b é o *bandwidth* (parâmetro de suavização) fixo e o parâmetro $b_{i(k)}$ representa o *bandwidth* variável, sendo que a letra k representa o número de vizinhos mais próximos do ponto i .

Quando se utiliza um *bandwidth* fixo, o mesmo deve ser especificado por seu valor em unidade de distância. No entanto, na utilização do *bandwidth* variável, deve-se definir um número k (fixo) de vizinhos mais próximos a serem utilizados no modelos e, com base nessa quantidade k , o valor do *bandwidth* varia entre as regiões do estudo.

O modelo GWLR foi elaborado a partir da base de desenvolvimento, com um *bandwidth* ótimo de 2805 metros definido pelo critério de otimização de AICc⁴ e com um kernel do

⁴ Uma adaptação proposta por Brundson e colaboradores (2002) do AIC para modelos geograficamente ponderados.

tipo Biquadrática Variável. As variáveis independentes utilizadas foram as mesmas que resultaram do modelo de regressão logística clássica após o método *Stepwise*.

4 DISCUSSÃO DOS RESULTADOS

4.1 Avaliação do Modelo de Regressão Logística clássico

O modelo de regressão logística clássico (RL) pode ser visualizado na Tabela 4.1. As variáveis que utilizadas no modelo RL são àquelas resultados do processo de seleção de variáveis e do método *Stepwise*. Todas as variáveis possuem um p-valor abaixo de 0,05 (nível de significância $\alpha = 5\%$).

Tabela 4.1: Modelo de regressão logística clássico (RL) para identificação de áreas precárias.

Variáveis	Descrição	β	Std. Error	z value	Pr(> z)
<i>Intercept</i>	-	-7,3139	0,5545	-13,191	< 2e-16
<i>Urbano</i>	Célula localizada em área urbana de acordo com o Censo 2010	3,4389	0,1983	17,339	< 2e-16
<i>SI_médio</i>	Média da Shape Index	-3,0038	1,3707	-2,191	0,02842
<i>Lixo Inadequado</i>	Proporção de domicílios com coleta de lixo inadequada	2,7312	0,4736	5,767	8,07e-09
<i>MC_Int_R400</i>	Integração regional das vias (e rios) considerando um raio de 400m	10,7631	2,9391	3,662	0,00025
<i>Contraste (INTERVALO)</i>	Intervalo entre o valor máximo e mínimo da métrica GLCM de Contraste	-3,2660	0,7885	-4,142	3,44e-05
<i>MC_Ch_R5000</i>	Potencial de escolha regional das vias (e rios) considerando um raio de 5 km	-14,8859	2,9783	-4,998	5,79e-07
<i>Correlação (SOMA)</i>	Soma da métrica GLCM de Correlação	6,4911	0,8069	8,045	8,64e-16
<i>Period_2000_20014</i>	Área construída durante os anos de 2000 e 2014	2,0352	0,2362	8,615	< 2e-16
<i>TABO (Área Construída)</i>	Área do maior objeto da classe área construída que intercepta a célula	-10,5469	1,5164	-6,955	3,52e-12
<i>HAND (MAX)</i>	Valor máximo do HAND	-2,4050	0,3378	-7,119	1,09e-12

Fonte: Produção do autor.

A partir da análise do modelo RL, podemos afirmar que as variáveis resultantes e os seus coeficientes traduzem os assentamentos precários de Marabá (CARDOSO; DE MELO; DO VALE GOMES, 2016). As células localizadas em áreas consideradas urbanas possuem maior probabilidade de serem precárias, quando comparadas com áreas consideradas rurais. Isso vai ao encontro com a própria definição de assentamentos precários instituída pela PNH (BRASIL, 2005). Além disso, áreas com maior atendimento dos serviços de coleta de lixo tendem a ser menos precárias – resultado que é coerente com outros estudos sobre a precariedade habitacional de municípios amazônicos

(SAKATAUSKAS, 2020). Essas duas variáveis, *Urbano* e *Lixo Inadequado*, fazem parte da dimensão Domiciliar e são as únicas representantes desta dimensão.

A dimensão Morfológica é aquela que mais possui variáveis no modelo estatístico – 7 das 10 variáveis pertencem a essa dimensão. De acordo com o modelo RL, quanto maior a conformidade da quadra, avaliado pelo *Shape Index*, maior a possibilidade de ser precário – o que pode significar um maior nível de organização dos assentamentos precários de Marabá. Esse alto nível de organização associado às áreas precárias pode ser resultado dos planos federais modernistas aplicados em Marabá durante as décadas de 60 e 70. Esses projetos urbanos modernistas são propostas bidimensionais que priorizaram a técnica em detrimento das experiências culturais e das características do bioma local, rompendo com o urbano pré-estabelecido pela região (PONTES, 2015).

Duas métricas texturais apareceram no modelo RL, sendo elas o *contraste* e a *correlação*. A correlação mede a dependência linear dos níveis de cinza no entorno de um pixel. Já o contraste representa a medida do quão diferente é o nível de cinza de um pixel em relação aos pixels adjacentes (HARALICK, 1979). De acordo com o modelo RL, a correlação está relacionada positivamente com a variável dependente (Y). O contraste, por sua vez, está relacionado negativamente com a variável dependente (Y). Os coeficientes dessas duas variáveis texturais demonstram que as áreas precárias possuem uma maior homogeneidade do que as áreas (ditas) formais. A alta correlação e baixo contraste significam que os pixels das áreas precárias possuem níveis de cinza similares, podendo ser uma medida da uniformidade do padrão construtivo das edificações e até mesmo um indicativo de uma baixa diversidade de uso desses assentamentos.

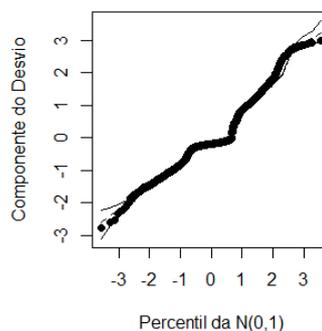
Ainda segundo o modelo RL, as áreas precárias possuem vias com alta integração na escala local, ao considerarmos um raio de 400m. No ponto de vista do pedestre, quando adotamos um raio grande, de 5 km, as vias (e rios) das áreas precárias possuem baixa potencialidade de escolha. Essas duas medidas, calculadas considerando toda a microrregião de Marabá, demonstram um isolamento dos assentamentos precários e uma menor integração com o restante da cidade, constatações já obtidas por estudos em outras cidades brasileiras (MEDEIROS; HOLANDA; BARROS, 2011; BARROS; MEDEIROS; MORAIS, 2016).

As áreas construídas entre os anos 2000 e 2014 possuem maior probabilidade de serem precárias, indicando um período de precarização da cidade de Marabá. Além disso, a relação negativa entre a variável *TABO - Total Area of the Biggest Object* e a variável dependente (Y), demonstra que os assentamentos precários tendem a ter objetos da classe “área construída” de menor tamanho (em m²), podendo ser um indicativo das pequenas edificações desses locais – características já relatadas anteriormente (CARDOSO; DE MELO; DO VALE GOMES, 2016; SAKATAUSKAS, 2020).

A única variável da dimensão ambiental é o *HAND* (o seu valor máximo). A relação negativa entre o máximo do *HAND* e a variável dependente (Y), significa que as áreas precárias tendem a estar localizadas em áreas mais baixas, com uma altura do terreno similar ao nível da drenagem mais próxima.

Segundo (FERNANDES, 2019), existem duas formas de avaliar o ajuste do modelo de regressão a partir do método gráfico do envelope simulado: (i) pela proporção de pontos fora do envelope e (ii) pela distância entre a banda do envelope e o ponto mais afastado desse limite – esta última considerada apenas quando há a ocorrência de resíduos fora do envelope. Conforme Figura 4.1, percebemos que todos os resíduos estão dentro das bandas limites evidenciando a suposição de distribuição da variável resposta adequada para o modelo de regressão logística clássico. A inexistência de pontos fora dos limites também indica que os erros não são correlacionados⁵, são dispersos aleatoriamente em torno de zero, possuem variância constante e há poucos ou nenhum resíduo com valor extremo, não havendo indícios de inadequação do modelo obtido.

Figura 4.1: Gráfico do envelope simulado dos resíduos do modelo RL.



Fonte: Produção do autor.

⁵ O método não avalia uma possível correlação espacial dos resíduos.

Também foi realizada uma avaliação do modelo de regressão RL comparando as bases de desenvolvimento e validação. Foram utilizadas a estatística de Kolmogorov-Smirnov (KS)⁶ e a área sob a Curva de Lorenz. O KS mede a separação máxima entre a frequência relativa acumulada de 0 (não precário) e 1 (precário). Obtivemos um mesmo valor de KS para as bases de desenvolvimento e validação, sendo 65%, não existindo perda de desempenho do modelo para a base de validação.

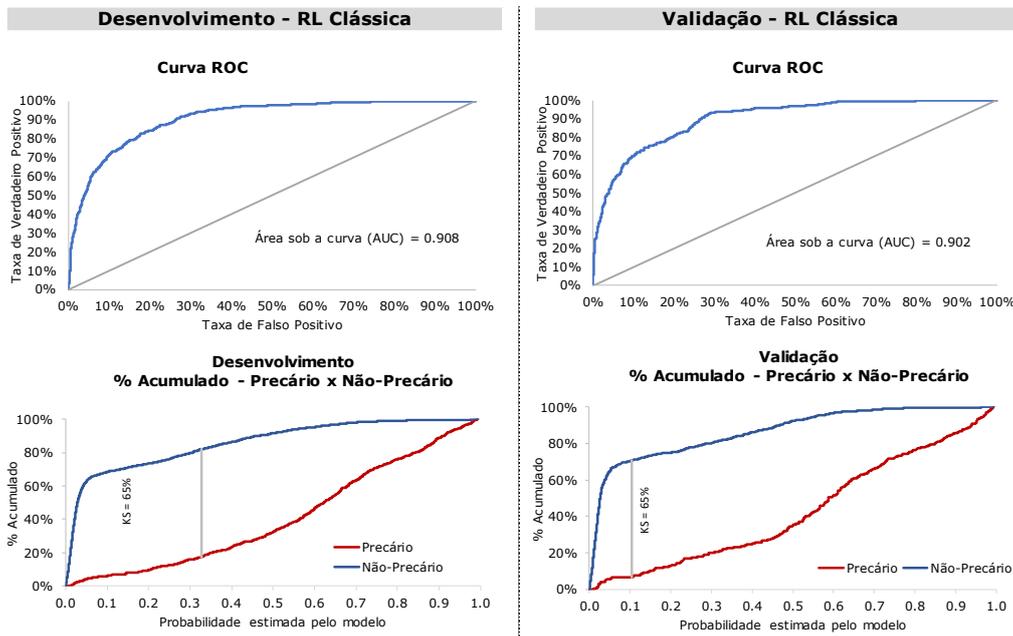
Embora o KS seja amplamente utilizado na medição de modelos estatísticos, seu uso por si só não é indicado, pois altos valores de KS não significam que teremos modelos bem ajustados, uma vez que altos valores podem ser obtidos para uma determinada faixa de frequência. Por esse motivo, é recomendável usar essa métrica em conjunto com pelo menos algum outro indicador de desempenho, como a área sob a curva de Lorenz (SIDDIQI, 2012).

A curva ROC é uma representação gráfica que ilustra o desempenho de um sistema classificador binário conforme seu limite de discriminação varia (HANLEY; MCNEIL, 1982). Ele compara a frequência de verdadeiros positivos e falsos positivos para cada ponto da curva.

A partir da Curva ROC, é possível calcular a Área sob a Curva (AUC), obtida pela área total de uma distribuição acumulada independente de um limiar. O valor de AUC é uma avaliação do desempenho geral de um teste, variando de 0 (onde todas as previsões no modelo estão erradas) a 1 (onde todas as previsões no modelo estão corretas) (HANLEY; MCNEIL, 1982). Para as bases de desenvolvimento e validação, os valores obtidos da AUC ficaram próximos a 1, sendo 0,908 para desenvolvimento e 0,902 para validação, concernindo mais uma evidência de que não houve perda de desempenho do modelo para a base de validação. A Figura 4.2 apresentam os KS e as Curvas ROCs do modelo RL.

⁶ O KS é amplamente utilizado no mercado financeiro para medir bons e maus pagadores em modelos logísticos de crédito (SIDDIQI, 2012). Não foram encontradas referências que utilizaram o KS para identificação de áreas precárias.

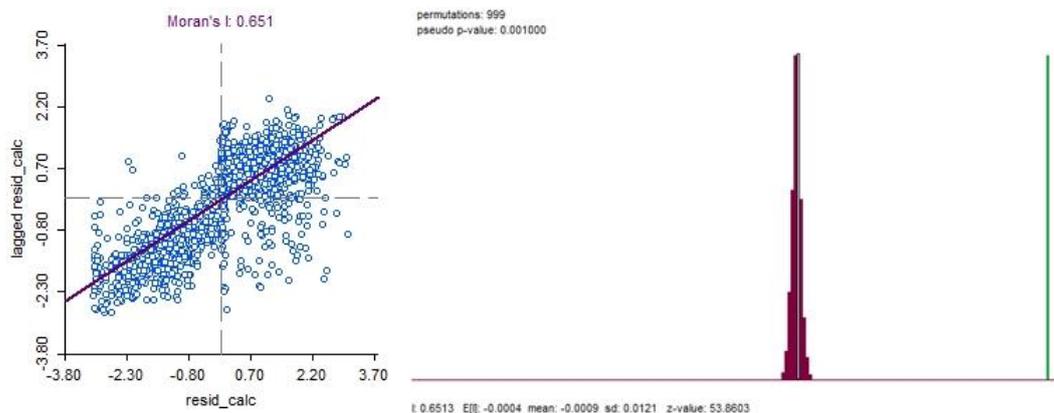
Figura 4.2: Curva ROC e estatística KS das bases de desenvolvimento e validação do modelo RL.



Fonte: Produção do autor.

Apesar do modelo RL apresentar um bom ajuste pelas técnicas clássicas de diagnóstico, os resíduos da regressão logística apresentaram uma dependência espacial (Figura 4.3). O Índice de Moran Local aplicado sobre os resíduos padronizados da regressão RL foi de 0,651, dado um teste de pseudo significância com 999 permutações e um p-valor igual a 0,001. Assim, pode-se dizer que os resíduos apresentam uma autocorrelação espacial positiva e significativa.

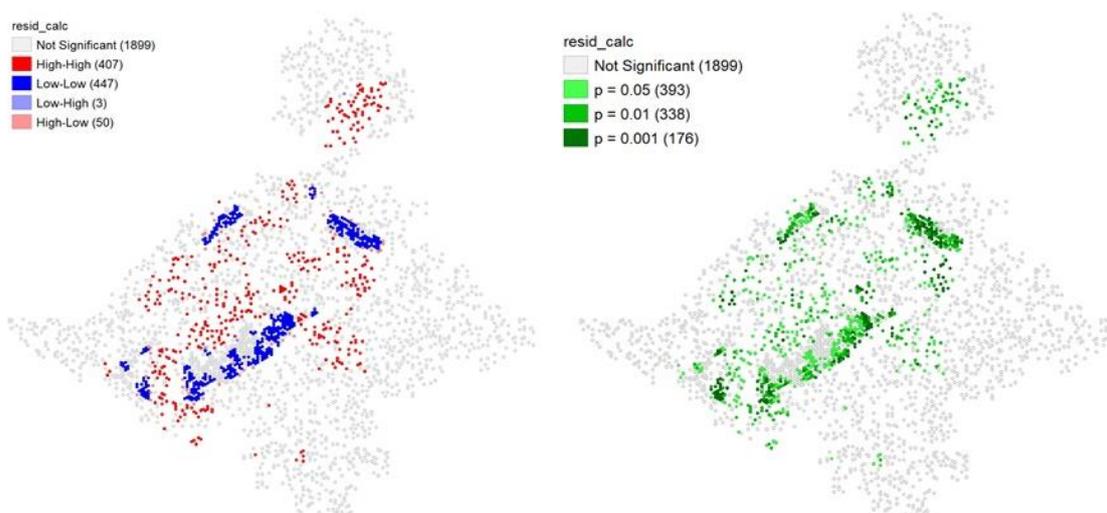
Figura 4.3: Índice de Moran Local dos resíduos padronizados da regressão RL.



Fonte: Produção do autor.

Constatada a dependência espacial dos resíduos, é possível verificar a existência de *clusters* por meio de um LISA Map. Os cluster em vermelho vivo (*High-High*) representam resíduos cujos vizinhos têm valores altos. Já aqueles em azul vivo (*Low-Low*) representam resíduos cujos vizinhos têm valores baixos. Por fim, aqueles de coloração em tom mais desbotado (em vermelho e azul), são resíduos cujos vizinhos apresentam valores discrepantes (ou seja, correlação negativa). Além da verificação dos *clusters*, é possível também visualizar a significância dos mesmos a partir da Figura 4.3.

Figura 4.4: LISA Map dos resíduos padronizados do modelo RL.



Fonte: Produção do autor.

4.2 Avaliação do Modelo de Regressão Logística Geograficamente Ponderado (GWLR)

Conforme descrito na metodologia, foi desenvolvido um modelo de Regressão Logística Geograficamente Ponderado (GWLR), com um *bandwidth* ótimo de 2805 metros definido pelo critério de otimização de AICc e com um *kernel* do tipo Biquadrática Variável. As variáveis utilizadas para compor o modelo GWLR foram as mesmas selecionadas no modelo RL e expostas na Tabela 4.1.

A Tabela 4.2 apresenta as estatísticas descritivas dos coeficientes estimados do modelo GWLR. É possível perceber que a média dos coeficientes ficaram próximos do modelo RL. Além disso, os valores médios dos coeficientes são congruentes (fazem sentido teórico) e possuem o mesmo sinal (positivo ou negativo) dos coeficientes da regressão RL.

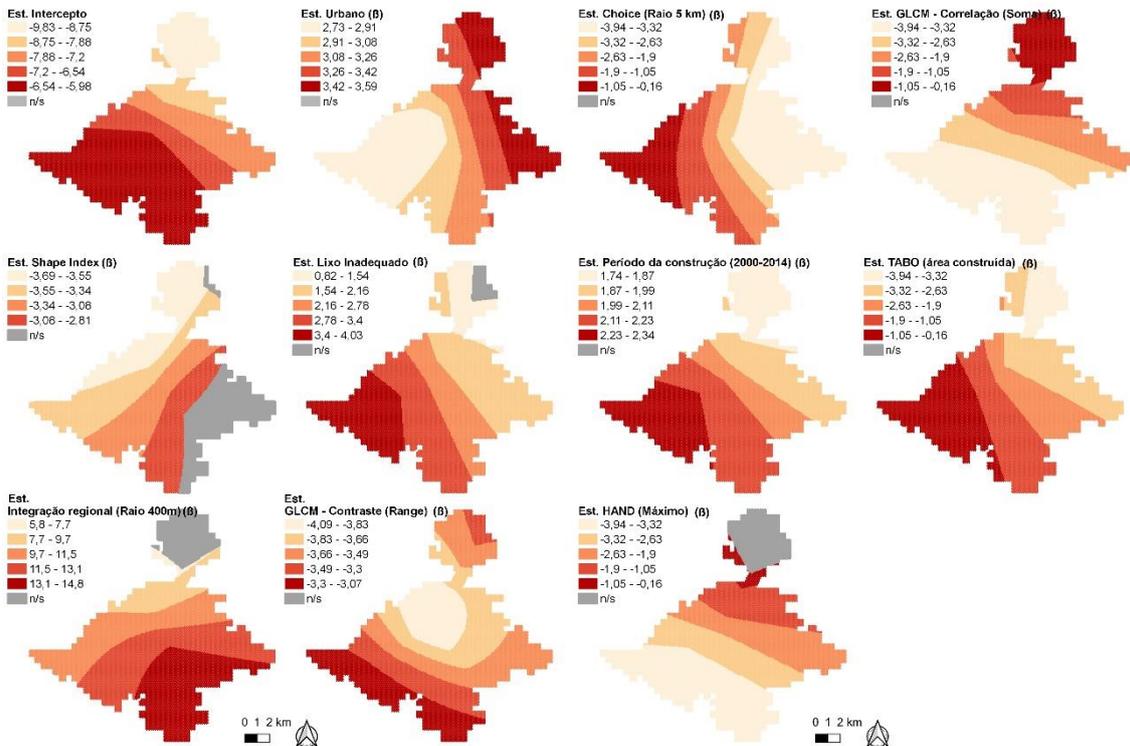
Tabela 4.2: Modelo de regressão logística geograficamente ponderado (GWLWR) para identificação de áreas precárias.

Variáveis	Média de β	Std. Error	Min	Max	Range	Q1	Mediana	Q3
<i>Intercept</i>	-6,95	1,00	-9,83	-5,99	3,84	-7,55	-6,45	-6,19
<i>Urbano</i>	3,09	0,26	2,73	3,59	0,86	2,86	3,04	3,34
<i>SI_médio</i>	-3,22	0,32	-3,69	-2,49	1,20	-3,49	-3,25	-2,98
<i>Lixo Inadequado</i>	2,65	0,81	0,82	4,02	3,20	1,89	2,79	3,31
<i>MC_Int_R400</i>	11,47	2,19	5,86	14,75	8,88	10,41	12,08	13,00
<i>Contraste (INTERVALO)</i>	-3,56	0,24	-4,09	-3,09	1,00	-3,76	-3,57	-3,36
<i>MC_Ch_R5000</i>	-15,74	0,85	-17,02	-13,89	3,13	-16,59	-15,73	-15,06
<i>Correlação (SOMA)</i>	6,69	1,13	5,55	9,95	4,40	5,81	6,22	7,28
<i>Period_2000_20014</i>	2,10	0,16	1,74	2,34	0,60	1,95	2,15	2,23
<i>TABO (Área Construída)</i>	-13,96	4,70	-27,21	-7,69	19,52	-17,72	-12,58	-10,00
<i>HAND (MAX)</i>	-2,64	1,05	-3,93	-0,16	3,77	-3,55	-2,97	-1,89

Fonte: Produção do autor.

A variação dos valores dos coeficientes e suas significâncias estão expostas na Figura 4.5. Nota-se a variação espacial dos coeficientes de todas as variáveis na grade celular. Apenas as variáveis Shape Index, Lixo Inadequado, Integração (400m) e Hand Máximo apresentaram regiões sem significância do coeficiente β – estas áreas apresentam valor-t fora do intervalo $-1,96 < \text{valor-t} < 1,96$.

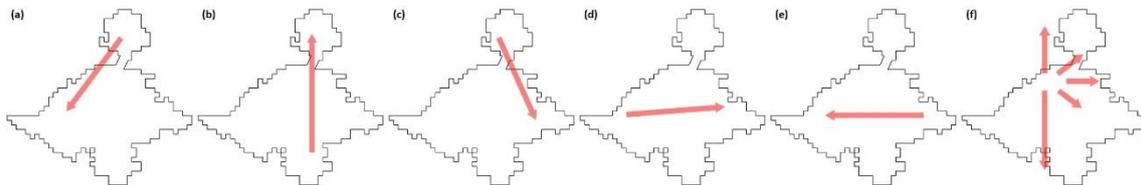
Figura 4.5: Distribuição espacial do intercepto, dos coeficientes β e suas significâncias.



Fonte: Produção do autor.

A partir da análise dos valores do intercepto e coeficientes β , é possível identificar alguns padrões espaciais (Figura 4.6). O intercepto e os β das variáveis *Lixo Inadequado*, *Período da Construção* e *TABO* possuem um padrão espacial de aumento de seu valores de norte para oeste da área de estudo (padrão A); os valores dos β das variáveis *Correlação (GLCM)* e *HAND (Máximo)* aumentam de sul para norte (padrão B); os valores dos β das variáveis *Shape Index* e *Integração regional* aumentam do norte para o leste (padrão C); os valores de β da variável *Urbano* aumentam do oeste para o leste (padrão D); os valores de β da variável *Choice* aumentam do leste para o oeste (padrão E); e os valores de β da variável *Constraste (GLCM)* aumentam radialmente do porção centro oeste em direção aos limites da área de estudo (padrão F).

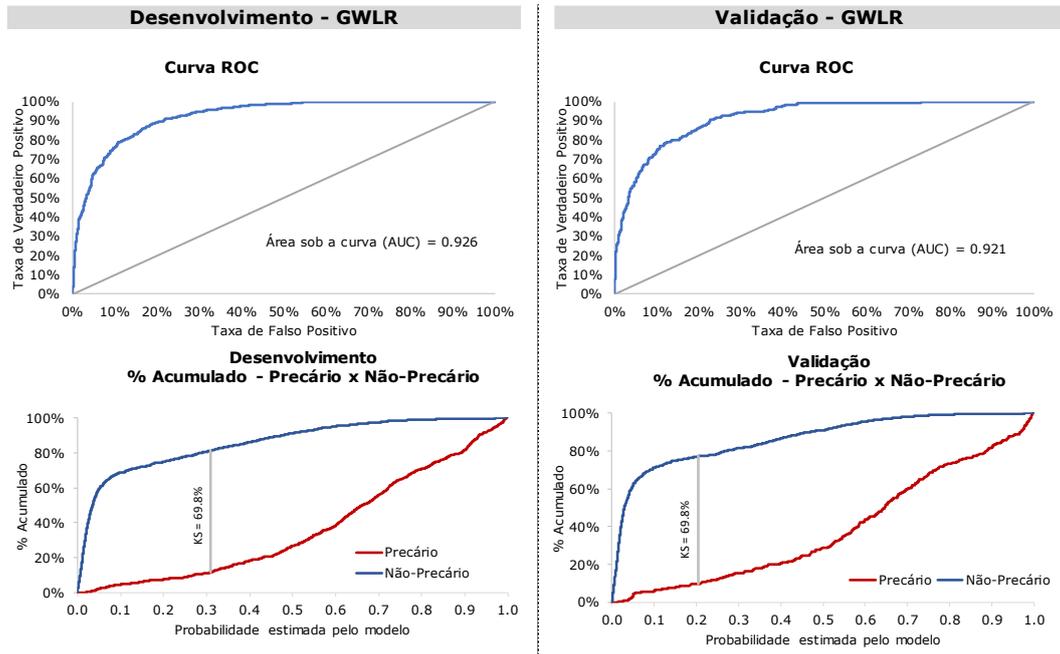
Figura 4.6: Padrões espaciais de crescimento dos coeficientes β .



Fonte: Produção do autor.

Assim como no modelo RL, foi realizada uma avaliação do modelo de regressão GWLR comparando as bases de desenvolvimento e validação a partir do KS e da AUC. Obtivemos um mesmo valor de KS para as bases de desenvolvimento e validação (69,8%). Os valores obtidos da AUC também ficaram próximos a 1, sendo 0,926 para desenvolvimento e 0,921 para validação. A análise do KS e do AUC evidenciam que não houve perda de desempenho do modelo GWLR para a base de validação. A Figura 4.6 apresentam os KS e as Curvas ROCs do modelo GWLR.

Figura 4.7: Curva ROC e estatística KS das bases de desenvolvimento e validação do modelo RL.

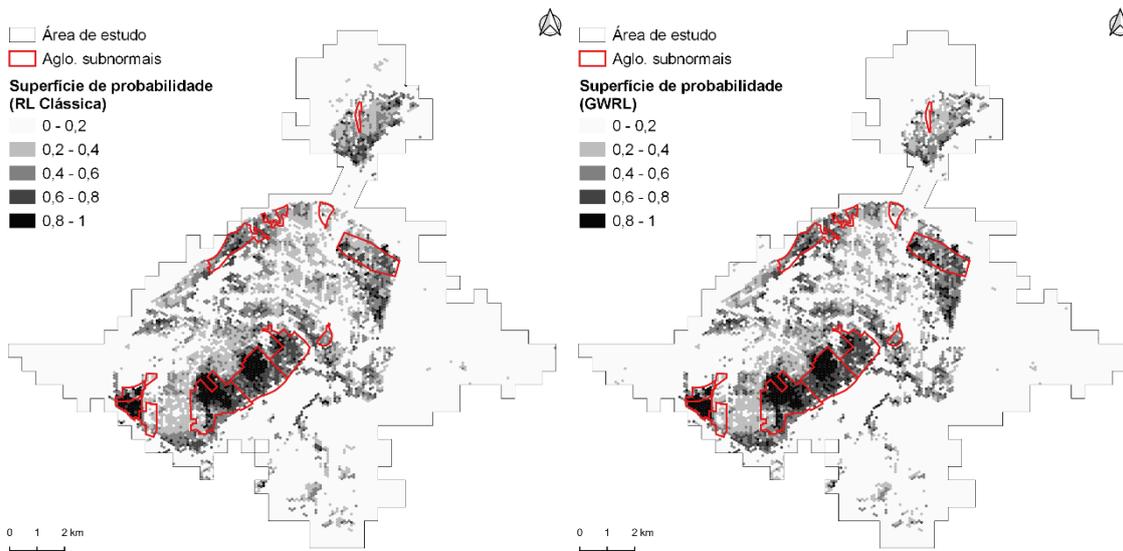


Fonte: Produção do autor.

4.3 Comparação entre os modelos

Para os dois modelos, RL e GWLR, calculamos a probabilidade de cada célula ser uma célula precária, retornando duas superfícies de probabilidades compostas por células com valores entre 0 (não precária) e 1 (precária). Os resultados são mostrados na Figura 4.7.

Figura 4.8: Superfícies de probabilidades de ser uma área precária.

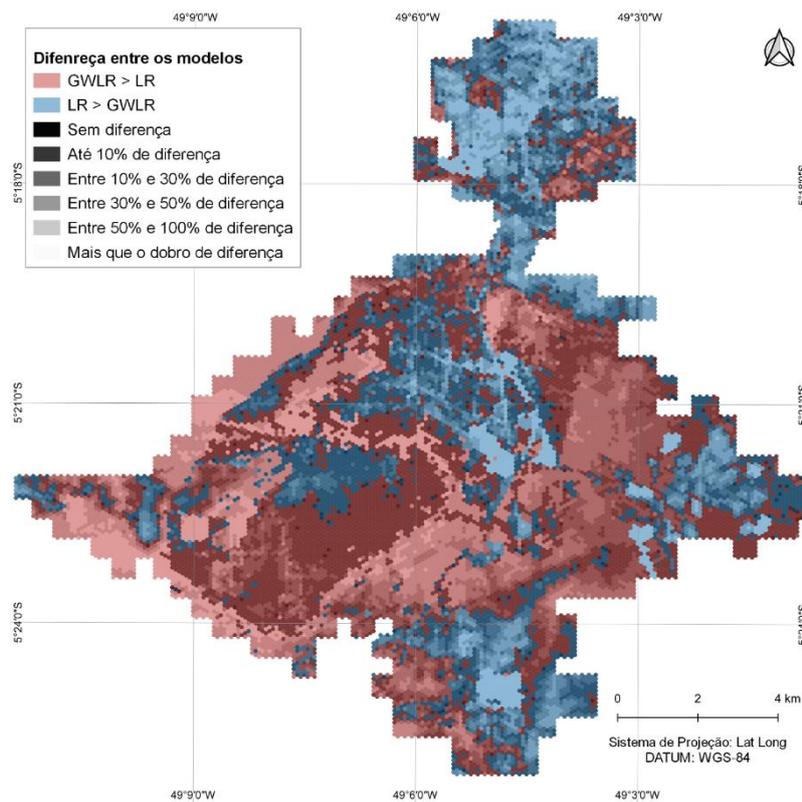


Fonte: Produção do autor.

Estatisticamente, foi calculada a correlação de Pearson entre as duas superfícies de probabilidades, resultando numa correlação de 0,9917. Isso significa que os dois modelos estão muito próximos de uma relação linear perfeita. Visualmente, as duas superfícies apresentaram probabilidades muito parecidas.

No entanto, ainda é possível identificar diferenças entre os dois modelos, em especial na porção superior da área de estudo. Em algumas regiões, os modelos apresentaram mais de 100% de diferença entre as probabilidades. A Figura 4.9 apresenta em azul as áreas onde o modelo GWLR resultou numa probabilidade (de ser precário) maior do que a calculada pelo modelo RL e em vermelho, as áreas onde o modelo RL resultou numa probabilidade (de ser precário) maior do que a calculada pelo modelo GWLR. A intensidade da cor indica o quão diferente são os dois modelo.

Figura 4.9: Diferença entre os modelos RL e GWLR.



Fonte: Produção do autor.

A comparação entre os dois modelos se deu a partir do Critério Informacional de AICc. Enquanto o modelo RL resultou num AICc de 1836.92, o modelo GWRL resultou num AICc de 1697.28. Dessa forma, o modelo GWLR foi o que apresentou o melhor desempenho, por possui um menor AICc. Além dos AICc, a Tabela 4.3 apresenta os

valores de Pseudo-R² de cada um dos modelos e o Índice de Moran dos Resíduos de ambos os modelos. Apesar de possuir um Índice de Moran menor, o modelo GWRL ainda apresenta correlação espacial entre os resíduos padronizados – com teste de pseudo-significância de 999 permutações e p-valor de 0,001.

Tabela 4.3: Comparação entre os modelos RL e GWRL.

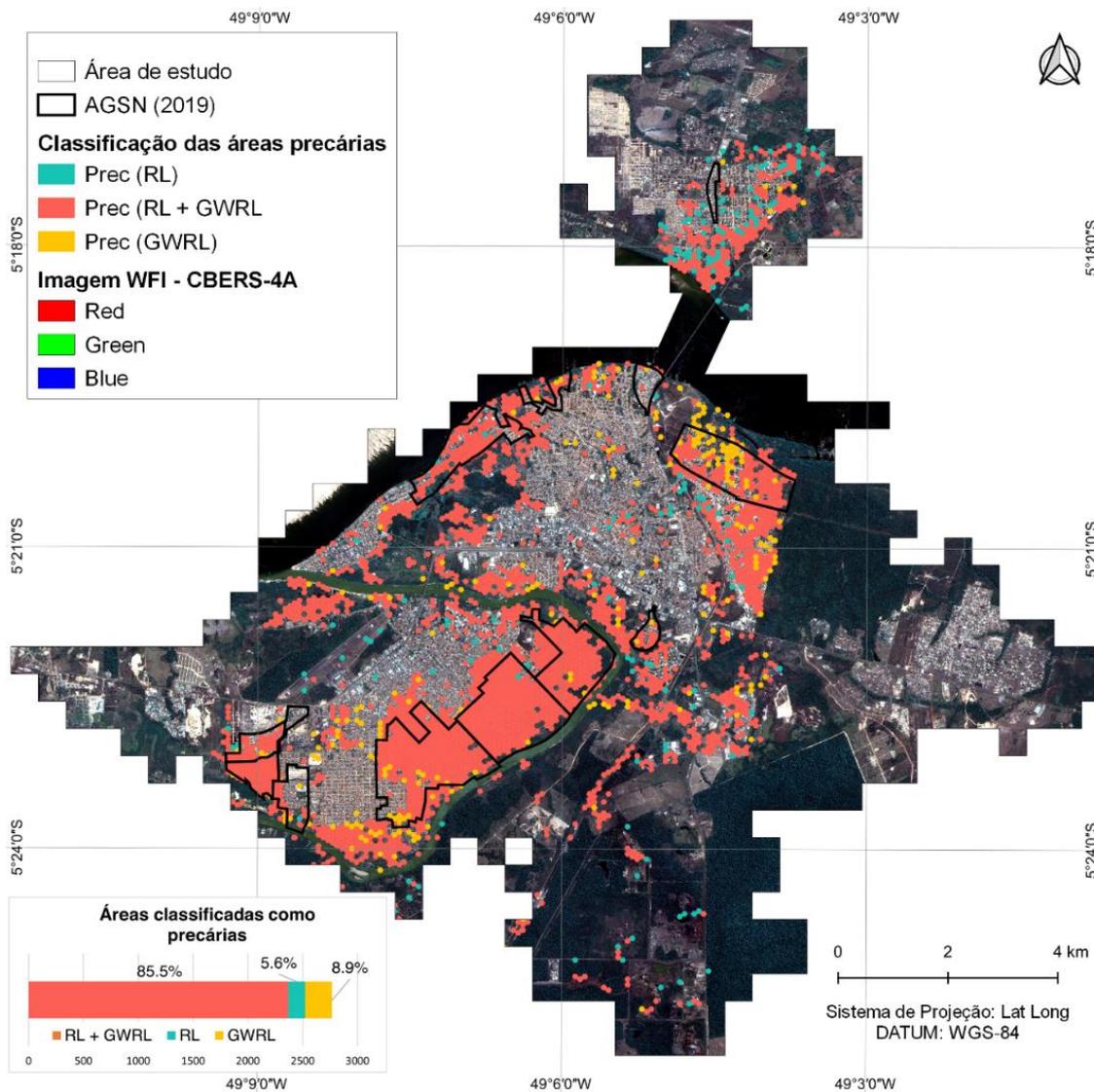
Modelo	Índice de Moran dos Resíduos	Pseudo-R²	AICc
RL	0,65	0,43	1.836,92
GWLR	0,58	0,47	1.697,28

Fonte: Produção do autor.

Assim como a Curva ROC, foi gerada uma curva de F1-Score para encontrarmos o ponto de corte que maximizasse a taxa de verdadeiro positivo, trazendo a maior qualidade de classificação para a área de desenvolvimento e validação de ambos os modelos. Após a análise das curvas ROC e F1-Score, da análise visual das superfícies de probabilidades de ambos os modelos e dos perímetros dos assentamentos precários, foi adotado um limiar de 0,45 para separar as áreas não precárias das potencialmente precárias (ANEXO A). O ANEXO A apresenta as curvas ROC e F1-Score com limiar de 0,45 e a Figura 4.10 mostra o resultado da classificação.

Do ponto de vista qualitativo, é possível verificar que a classificação foi capaz de identificar a maioria das células precárias. No entanto, ocorreram erros de comissão (áreas não precárias foram incluídas como potencialmente precárias) e erros de omissão (áreas precárias não identificadas). Cerca de 16% da área de estudo foi identificada como precária em ao menos um dos modelos. A maior parte das áreas precárias identificadas (85%) foram classificadas como precárias em ambos os modelos, 8,9% foram classificadas apenas pelo modelo GWLR e 5,6% foi classificado como precário apenas no modelo RL.

Figura 4.10: Classificação das áreas precárias.



Fonte: Produção do autor.

Para ambos os modelos foram calculadas matrizes de confusão para avaliar a classificação final. Essas informações são mostradas na Tabela 4.4. O corte de 0,45 aplicado maximizou os F1-Scores, retornando um F1-Score de 72,5% na base de validação GWRL e um F1-Score de 69,6% para a base de validação RL. Dessa forma, o modelo GWRL resultou numa classificação melhor do que o modelo RL.

Tabela 4.4: Matriz de confusão da classificação em áreas precárias (amostras de validação).

		Classificação		Acurácia do Usuário (%)	Modelo
		Não - Prec	Precário		
Referência	Não - Prec	803	83	90.63%	RL (F1 69.58%; GA 84.41%)
	Precário	105	215	67.19%	
	Acurácia do Prod. (%)	88.4%	72.1%	-	
	Não - Prec	810	73	91.73%	GWLR (F1 72.46%; GA 85.82%)
	Precário	98	225	69.66%	
Acurácia do Prod. (%)	89.21%	75.50%	-		

Fonte: Produção do autor.

5 CONSIDERAÇÕES FINAIS

Os assentamentos precários de Marabá – PA podem ser classificados entre assentamentos organizados e assentamentos espontâneos. As áreas precárias da cidade localizam-se em áreas baixas, sobrepondo áreas de potencial alagável, mas num padrão construtivo não adaptado às condições ambientais.

Do ponto de vista metodológico, este trabalho contribui para um debate interdisciplinar sobre a precariedade em cidades amazônicas, propondo a identificação de áreas precárias a partir de um modelo de regressão logística clássico (RL) e outro geograficamente ponderado (GWLR). Foram utilizadas variáveis agrupadas nas dimensões de análise Morfológica, Ambiental e Domiciliar. A partir das superfícies de probabilidades dos dois modelos, foi aplicado um corte para identificação das áreas precárias que resultou em acurácias globais acima de 84% - demonstrando um potencialidade do uso de modelos de regressão para identificação de áreas precárias.

Contudo, este trabalho é um estudo preliminar sobre a identificação de áreas precárias amazônicas. O conceito e a caracterização do que é precário em cidades amazônicas, bem como as variáveis utilizadas para descrever este fenômeno, devem ser reavaliadas em função das particularidades da região.

Por fim, as propostas de intervenção dos assentamentos precários de Marabá – PA devem levar em consideração as características culturais da população residente. A resolução da precariedade habitacional não pode ser solucionada apenas com a oferta de novas moradias, necessitando uma ação articulada entre as políticas urbanas e ambientais – que só podem ser implementadas pelo poder público.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, M. A. Reconstruindo uma história esquecida: origem e expansão inicial das favelas do Rio de Janeiro. **Espaço & Debates**, v. 14, n. 37, p. 34–46, 1994.

ALLAIRE, J. RStudio: integrated development environment for R. **Boston, MA**, v. 770, n. 394, p. 165–171, 2012.

ANSELIN, L.; SYABRI, I.; KHO, Y. GeoDa: an introduction to spatial data analysis. In: **Handbook of applied spatial analysis**. [s.l.] Springer, 2010. p. 73–89.

ATKINSON, A. C. Two graphical displays for outlying and influential observations in regression. **Biometrika**, v. 68, n. 1, p. 13–20, 1981.

BARROS, A. P. B. G.; MEDEIROS, V. A. S. DE; MORAIS, M. DA P. A configuração espacial para o diagnóstico dos assentamentos precários no Brasil. 2016.

BRAGA, B.; HESPANHOL, I.; CONEJO, J. G. L.; MIERZWA, J. C.; DE BARROS, M. T. L.; SPENCER, M.; PORTO, M.; NUCCI, N.; JULIANO, N.; EIGER, S. **Introdução à engenharia ambiental: o desafio do desenvolvimento sustentável**. [s.l.] Pearson Prentice Hall, 2005.

BRASIL. **Lei nº 11.124**. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/111124.htm>. Acesso em: 31 out. 2021.

BRASIL. **Guia para o Mapeamento e Caracterização de Assentamentos Precários**. Brasília: Ministério das Cidades, maio 2010.

BRUNSDON, C.; FOTHERINGHAM, A. S.; CHARLTON, M. Geographically weighted summary statistics — a framework for localised exploratory data analysis. **Computers, Environment and Urban Systems**, v. 26, n. 6, p. 501–524, 2002.

BRUNSDON, C.; FOTHERINGHAM, A. S.; CHARLTON, M. E. Geographically weighted regression: a method for exploring spatial nonstationarity. **Geographical analysis**, v. 28, n. 4, p. 281–298, 1996.

CÂMARA, G.; CARVALHO, M. S.; CRUZ, O. G.; CORREA, V. Análise espacial de áreas. **Análise espacial de dados geográficos**, v. 2, 2004.

CAMARA MUNICIPAL DE MARABÁ. **História do Município — Câmara Municipal de Marabá**. Disponível em: <<http://maraba.pa.leg.br/institucional/maraba/historia>>. Acesso em: 7 dez. 2021.

CARDOSO, A. C. D.; DE MELO, A. C.; DO VALE GOMES, T. O urbano contemporâneo na fronteira de expansão do capital: padrões de transformações espaciais

em seis cidades do Pará, Brasil. **Revista de Morfologia Urbana**, v. 4, n. 1, p. 5–28, 2016.

CARDOSO, A. C. D.; LIMA, J. J. F.; PONTE, J. P. X.; VENTURA, R. DA S.; RODRIGUES, R. M. Morfologia urbana das cidades amazônicas: a experiência do Grupo de Pesquisa Cidades na Amazônia da Universidade Federal do Pará. **urbe. Revista Brasileira de Gestão Urbana**, v. 12, 2020.

CDHU; UFABC. **MAPPA - METODOLOGIA PARA IDENTIFICAÇÃO E CARACTERIZAÇÃO DE ASSENTAMENTOS PRECÁRIOS EM REGIÕES METROPOLITANAS PAULISTAS**. São Bernardo do Campo. 2018.

CONZEN, M. R. G. Alnwick, Northumberland: a study in town-plan analysis. **Transactions and Papers (Institute of British Geographers)**, n. 27, p. iii–122, 1960.

COSTA, S. DE A. P.; NETTO, M. M. G. **Fundamentos de morfologia urbana**. [s.l.] C/Arte, 2015. ISBN(8576543028).

DA MOTTA, E. M. P. L. ST 5 Medindo a desigualdade socioespacial: os aglomerados subnormais em Belo Horizonte. **Anais ENANPUR**, v. 17, n. 1, 2017.

DENALDI, R.; ROSA, J. S. **Ações integradas de urbanização de assentamentos precários**. Brasília: Ministério das Cidades/Cities Alliance, 2010. Disponível em: <<https://bibliotecadigital.seplan.planejamento.gov.br/handle/iditem/272>>. Acesso em: 31 out. 2021

DOS SANTOS, B. D.; DE PINHO, C. M. D.; DE JESUS, T. B. **NÍVEIS DE CONSOLIDAÇÃO DE ASSENTAMENTOS PRECÁRIOS A PARTIR DE DADOS DE SENSORIAMENTO REMOTO**. [s.d.].

FEITOSA, F. DA F.; VASCONCELOS, V. V.; DE PINHO, C. M. D. DE; DA SILVA, G. F. G.; DA SILVA GONÇALVES, G.; DANNA, L. C. C.; LISBOA, F. S. **IMMerSe: An integrated methodology for mapping and classifying precarious settlements**. **Applied geography**, v. 133, p. 102494, 2021.

FERNANDES, V. V. Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão. 2019.

GONÇALVES, G. **Identificação de Assentamentos Precários na Região do Grande ABC: Uma abordagem estatística**. 2018. Santo André: [s.n.].

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. [s.l.] Bookman editora, 2009. ISBN(8577805344).

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.

HARALICK, R. M. Statistical and structural approaches to texture. **Proceedings of the IEEE**, v. 67, n. 5, p. 786–804, 1979.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. H. Textural features for image classification. **IEEE Transactions on systems, man, and cybernetics**, n. 6, p. 610–621, 1973.

HARVEY, D. Espaços urbanos na aldeia global: reflexões sobre a condição urbana no capitalismo no final do século XX. **Cadernos de Arquitetura e Urbanismo, Minas Gerais, PUC**, v. 4, p. 171–189, 1996.

HOSMER, D. W. .; LEMESHOW, S.; STURDIVANT, R. X. . **Applied logistic regression**. [s.l.] Wiley New York, 2000. ISBN(0471356328).

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico 2010**. Rio de Janeiro. 2011. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/periodicos/93/cd_2010_caracteristicas_populacao_domicilios.pdf>. Acesso em: 1 nov. 2021.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Aglomerados Subnormais 2019: Classificação preliminar e informações de saúde para o enfrentamento à COVID-19**. Rio de Janeiro. 2020. Disponível em: <<https://covid19.ibge.gov.br/>>. Acesso em: 1 nov. 2021.

KÖRTING, T. S.; GARCIA FONSECA, L. M.; CÂMARA, G. GeoDMA—Geographic Data Mining Analyst. **Computers & Geosciences**, v. 57, p. 133–145, 2013.

KUFFER, M.; PFEFFER, K.; SLIUZAS, R. Slums from space—15 years of slum mapping using remote sensing. **Remote Sensing**, v. 8, n. 6, p. 455, 2016.

LIN, A. Z. Variable reduction in SAS by using weight of evidence and information value. In: SAS Global Forum, 2013, **Anais...** 2013

MARICATO, E. Autoconstrução, a arquitetura possível. **A produção capitalista da casa (e da cidade) no Brasil industrial**, v. 2, p. 71–93, 1982.

MAYFIELD, H. J.; LOWRY, J. H.; WATSON, C. H.; KAMA, M.; NILLES, E. J.; LAU, C. L. Use of geographically weighted logistic regression to quantify spatial variation in the environmental and sociodemographic drivers of leptospirosis in Fiji: a modelling study. **The Lancet Planetary Health**, v. 2, n. 5, p. e223–e232, 2018.

MCGARIGAL, K.; MARKS, B. J. Spatial pattern analysis program for quantifying landscape structure. **Gen. Tech. Rep. PNW-GTR-351. US Department of Agriculture, Forest Service, Pacific Northwest Research Station**, p. 1–122, 1995.

MEDEIROS, V. A. S. DE; HOLANDA, F. R. B. DE; BARROS, A. P. B. G. O labirinto das cidades brasileiras: heranças urbanísticas e configuração espacial. 2011.

MEDINA, F. A. S. Regressão Logística Geograficamente Ponderada Aplicada a Modelos de Credit Scoring. 2016.

NAKANO, A. K. Desenvolvimento Urbano e Territorial em municípios periurbanos: Diferenciações no Universo dos municípios brasileiros. **SANTANA, J. V, HOLANDA, AC G, MOURA, AS F (organizadoras). A questão da habitação em municípios periurbanos na Amazônia. Belém: editora UFPA, 2012.**

NAKAYA, T.; CHARLTON, M.; LEWIS, P.; BRUNSDON, C.; YAO, J.; FOTHERINGHAM, S. GWR4 user manual. **Windows Application for Geographically Weighted Regression Modelling**, 2014.

NUR, I. M.; AL HARIS, M. Geographically Weighted Logistic Regression (GWLR) with Adaptive Gaussian Weighting Function in Human Development Index (HDI) in The Province of Central Java. **Journal of Physics: Conference Series**, v. 1776, n. 1, p. 012048, 2021.

OSM. **OpenStreetMap**. Disponível em: <<https://www.openstreetmap.org/#map=4/-15.13/-53.19>>. Acesso em: 1 nov. 2021.

PASTERNAK, S. São Paulo e suas favelas. **Pós. Revista do Programa de Pós-Graduação em Arquitetura e Urbanismo da FAUUSP**, n. 19, p. 176–197, 2006.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [s.l.] IME-USP São Paulo, 2004.

PESARESI, M.; EHRLICH, D.; FERRI, S.; FLORCZYK, A.; FREIRE, S.; HALKIA, M.; JULEA, A.; KEMPER, T.; SOILLE, P.; SYRRIS, V. Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. **Publications Office of the European Union**, p. 1–62, 2016.

PINHEIRO, J. Fundação. Déficit habitacional municipal no Brasil 2010. **Belo Horizonte: Fundação João Pinheiro-Centro de Estatística e Informações**, 2013.

PONTES, L. B. **Cidade, desenho e natureza: uma reflexão sobre os espaços livres de Marabá**. 2015. 1–129 p. Belém do Pará: [s.n.].

QUEIROZ FILHO, A. P. DE. As definições de assentamentos precários e favelas e suas implicações nos dados populacionais: abordagem da análise de conteúdo. **urbe. Revista Brasileira de Gestão Urbana**, v. 7, n. 3, p. 340–353, 2015.

REUB, F. **Detection of favelas in Brazil using texture parameters and machine learning**. 2017. 94 p. Graz: [s.n.]. Disponível em: <https://elib.dlr.de/115220/1/Reuss_MA.pdf>. Acesso em: 7 dez. 2021. Acesso em:

SAKATAUSKAS, G. D. L. B. **ESPECIFICIDADES DA PRECARIIDADE HABITACIONAL NA AMAZÔNIA RIBEIRINHA: um olhar sobre a região do Baixo Tocantins**. 2020. 1–285 p. São Bernardo do Campo: [s.n.].

SAKATAUSKAS, G. DE L. B.; SANTANA, J. V. ST 2 PARTICULARIDADES DAS HABITAÇÕES NOS PEQUENOS MUNICÍPIOS PARAENSES. **Anais ENANPUR**, v. 16, n. 1, 2015.

SANTANA, J. V.; HOLANDA, A. C. G.; DE MOURA, A. DO S. F. **A questão da habitação em municípios periurbanos na Amazônia**. [s.l.] Ed. UFPA, 2012. ISBN(8524705124).

SANTOS, M. **A natureza do espaço: técnica e tempo, razão e emoção**. [s.l.] Edusp, 2002. v. 1 ISBN(8531407133).

SIDDIQI, N. **Credit risk scorecards: developing and implementing intelligent credit scoring**. [s.l.] John Wiley & Sons, 2012. v. 3 ISBN(1118429168).

STRAUSS, D. The many faces of logistic regression. **The American Statistician**, v. 46, n. 4, p. 321–327, 1992.

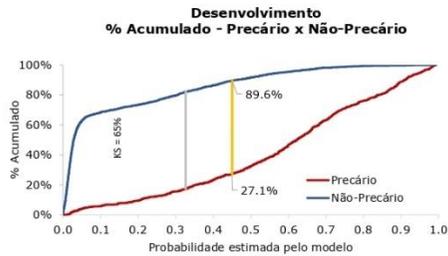
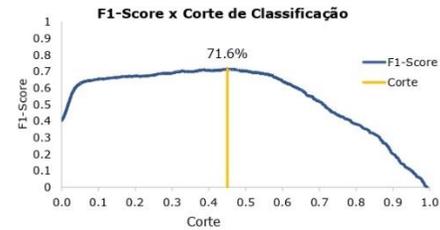
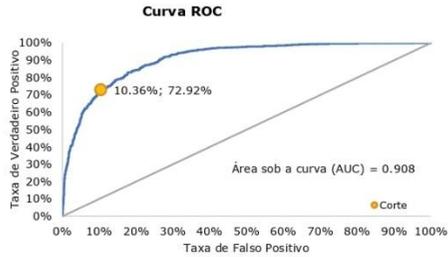
TEAM, Q. D. **QGIS Geographic Information System**. QGIS Association, 2021. Disponível em: <<https://www.qgis.org>>

VAZ, L. F. Dos cortiços às favelas e aos edifícios de apartamentos—a modernização da moradia no Rio de Janeiro. **Análise social**, p. 581–597, 1994.

ANEXO A – Curvas ROC, KS e Curva F1-Score das bases de desenvolvimento e validação dos modelos RL e GWLR.

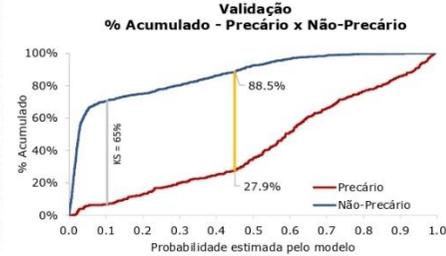
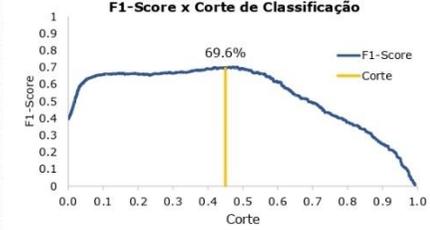
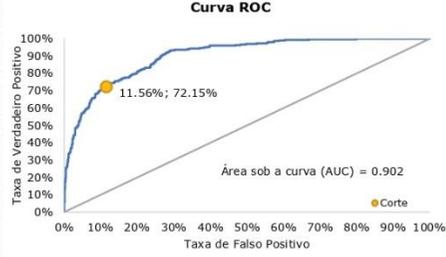
Desenvolvimento - RL Clássica

% por tipo	Não-Precário	Precário	Total	
< Corte de 0.45	90%	27%	74%	Verdadeiro Positivo 73%
>= Corte de 0.45	10%	73%	26%	Falso Positivo 10%
Total	100%	100%	100%	Falso Negativo 27%



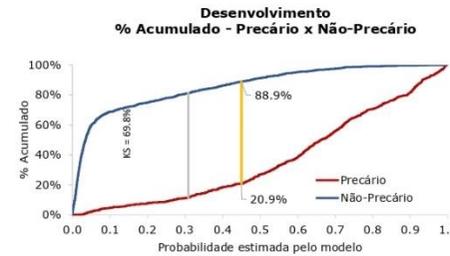
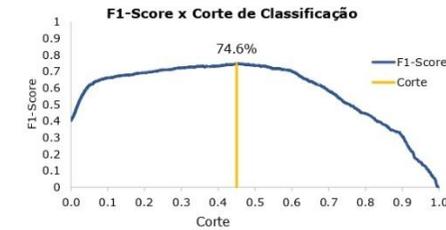
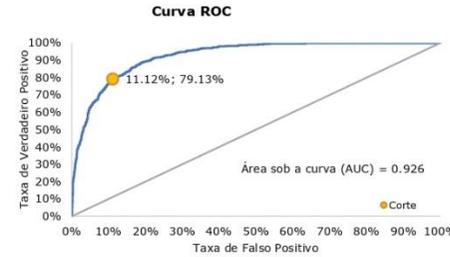
Validação - RL Clássica

% por tipo	Não-Precário	Precário	Total	
< Corte de	88%	28%	73%	Verdadeiro Positivo 72%
>= Corte de	12%	72%	27%	Falso Positivo 12%
Total	100%	100%	100%	Falso Negativo 28%



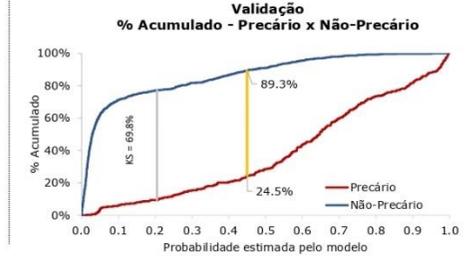
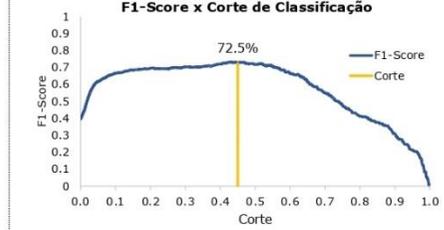
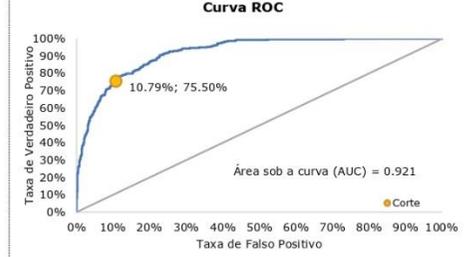
Desenvolvimento - GWLR

% por tipo	Não-Precário	Precário	Total	
< Corte de 0.45	89%	21%	72%	Verdadeiro Positivo 79%
>= Corte de 0.45	11%	79%	28%	Falso Positivo 11%
Total	100%	100%	100%	Falso Negativo 21%



Validação - GWLR

% por tipo	Não-Precário	Precário	Total	
< Corte de	89%	24%	73%	Verdadeiro Positivo 76%
>= Corte de	11%	76%	27%	Falso Positivo 11%
Total	100%	100%	100%	Falso Negativo 24%



Fonte: Produção do autor.