Instituto de Pesquisa Econômica Aplicada

Secretaria Nacional de Habitação do
Ministério do Desenvolvimento Regional

Termo de Execução Descentralizada n. 01/2019 SNH/MDR e Ipea

Pesquisa de Núcleos Urbanos Informais no Brasil

Produto 12 – APÊNDICES

Elaboração
Flávia da Fonseca Feitosa
Gilmara Gonçalves
Luis Felipe Bortolatto da Cunha
Pedro Reis Simões
Cleandro Krause
Juliana Gomes Petrarolli

Versão de 23 de agosto de 2021

Pesquisa de Núcleos Urbanos Informais no Brasil

Coordenação

Cleandro Henrique Krause (titular) e Marco Aurélio Costa (suplente) – Técnicos de Planejamento e Pesquisa da Diretoria de Estudos e Políticas Regionais, Urbanas e Ambientais (Dirur) do Ipea

Equipe da Pesquisa de Núcleos Urbanos Informais no Brasil – bolsistas e colaboradores

Alexandrina Saldanha Sobreira de Moura – FUNDAJ / Ipea (PNPD)
Ana Carolina Campos de Melo – Ipea (PNPD)
André Simionato Castro – Ipea (PNPD)
Bruno Gallina – UFRGS
Cátia Wanderley Lubambo – FUNDAJ / Ipea (PNPD)
David Melo Van Den Brule – Ipea (PNPD)
Elisa Escosteguy Utzig – Ipea (PNPD)
Fernanda Balestro – Ipea (PNPD)
Fernanda Carolina Vieira da Costa – Ipea (PNPD)
Flávia da Fonseca Feitosa – UFABC / Ipea (PNPD)
Gilmara Gonçalves – Ipea (PNPD)
Gabriel Moraes de Outeiro – UNIFESSPA / Ipea (PNPD)
Giuliana de Freitas – Ipea (PNPD)
Guilherme Frizzi Galdino da Silva – Ipea (PNPD)
Heleniza Ávila Campos – UFRGS / Ipea (PNPD)
Juliana Gomes Petrarolli – Ipea (PNPD)
Kaiena Thyelle Malaquias – FUNDAJ
Livia Salomão Piccinini – UFRGS
Luis Felipe Bortolatto da Cunha – Ipea (PNPD)
Manoela Guedes Ferreira Jordão de Vasconcelos – Ipea (PNPD)
Marcela Rodrigues Santos – Ipea (PNPD)
Mariana Roberti Bomtempo – Ipea (PNPD)
Miriam Francisca Rodrigues Couto – Ipea (PNPD)
Paulo Somlanyi Romeiro – Ipea (PNPD)
Pedro Reis Simões – Ipea (PNPD)
Rafael Gonçalves Gumiero – UNIFESSPA
Raquel de Mattos Viana – FJP / Ipea (PNPD)
Rosana Denaldi – UFABC / Ipea (PNPD)
Sergio Moreno Redón – UNIFESSPA / Ipea (PNPD)
Tatiana Mamede Salum Chaer – Ipea (PNPD)
Thaís Pires Rubioli – Ipea (PNPD)
Tiago Gonçalves da Silva – Ipea (PNPD)
Valéria Barroso da Silveira – Ipea (PNPD)

SUMÁRIO

Apêndice 1 - Quadro das variáveis construídas e integradas na grade celular NUI

Identificador	Descrição	Fonte	Eixo
ID	ID célula	IPEA (2021)	Identificação
Col	ID coluna célula	IPEA (2021)	Identificação
Row	ID linha célula	IPEA (2021)	Identificação
Polo	Polo de pesquisa	IPEA (2021)	Identificação
UF	UF do polo de pesquisa	IPEA (2021)	Identificação
AGSN	Presença de aglomerado subnormal	IBGE (2020)	Modelagem
NUI	Presença de núcleo urbano informal	IPEA (2021)	Modelagem
NUIDenDom	Estimativa do número de domicílios do NUI por ha	IPEA (2021)	NUI
NUIDom	Estimativa do número de domicílios do NUI	IPEA (2021)	NUI
NUIArea	Área do núcleo urbano informal (ha)	IPEA (2021)	NUI
Declividade	Declividade média do terreno por unidade de análise	SRTM (2020)	Físico-Territoriais
Curvatura	Curvatura média do terreno por unidade de análise	SRTM (2020)	Físico-Territoriais
APP30m	Porcentagem de área ocupada da unidade de análise a 30 metros do curso d'agua	FBDS (2018)	Físico-Territoriais
UCIntegral	Porcentagem de área ocupada da unidade de análise dentro de Unidade de Conservação de Proteção Integral	MMA (2020)	Físico-Territoriais
AltaTensao	Porcentagem de área ocupada da unidade de análise em faixas de servidão de Linhas de Alta Tensão	ANEEL (2020)	Físico-Territoriais
Vias50m	Porcentagem de área ocupada da unidade de análise dentro da faixa de 50 metros da via carroçavel	OSM (2020)	Físico-Territoriais
Dutovias	Porcentagem de área ocupada da unidade de análise em faixas de serviço de dutovias	ANP (2020)	Físico-Territoriais
IndiceForma	Média do índice de forma das quadras/bolsões de ocupação por unidade de análise - o índice de forma mede a regularidade das quadras e quanto mais próximo de 1, mais regulares são as quadras/bolsões de ocupação.	OSM (2020)	Físico-Territoriais
DomSIden	Porcentagem de domicílios sem identificação do logradouro	Censo (2010)	Infraestrutura e entorno
DomSIlu	Porcentagem de domicílios sem iluminação pública	Censo (2010)	Infraestrutura e entorno
DomSPav	Porcentagem de domicílios sem pavimentação	Censo (2010)	Infraestrutura e entorno
DomSCal	Porcentagem de domicílios sem calçada	Censo (2010)	Infraestrutura e entorno
DomSFio	Porcentagem de domicílios sem meio-fio	Censo (2010)	Infraestrutura e entorno
DomSBue	Porcentagem de domicílios sem bueiro	Censo (2010)	Infraestrutura e entorno
DomSArb	Porcentagem de domicílios sem arborização	Censo (2010)	Infraestrutura e entorno
DomSEne	Porcentagem de domicílios sem energia elétrica	Censo (2010)	Infraestrutura e entorno

DomSAgua	Porcentagem de domicílios sem abastecimento de água de rede geral	Censo (2010)	Infraestrutura e entorno
DomSMed	Porcentagem de domicílios sem medidor de uso exclusivo	Censo (2010)	Infraestrutura e entorno
DomSEsg	Porcentagem de domicílios com esgoto a céu aberto	Censo (2010)	Infraestrutura e entorno
DomSRedeEsg	Porcentagem de domicílios sem ligação à rede de esgoto ou fossa séptica	Censo (2010)	Infraestrutura e entorno
DomCLixAc	Porcentagem de domicílios com lixo acumulado nos logradouros	Censo (2010)	Infraestrutura e entorno
DomSColLix	Porcentagem de domicílios sem coleta de lixo	Censo (2010)	Infraestrutura e entorno
DomSColLixDir	Porcentagem de domicílios sem coleta de lixo direta	Censo (2010)	Infraestrutura e entorno
DomApto	Porcentagem de domicílios particulares permanentes do tipo apartamento	Censo (2010)	Infraestrutura e entorno
DomCasa	Porcentagem de domicílios particulares permanentes do tipo casa	Censo (2010)	Infraestrutura e entorno
DomVila	Porcentagem de domicílios particulares permanentes do tipo casa de vila ou em condomínio	Censo (2010)	Infraestrutura e entorno
DomImpr	Porcentagem de domicílios particulares improvisados	Censo (2010)	Infraestrutura e entorno
DomAdeq	Porcentagem de domicílios particulares permanentes com moradia adequada	Censo (2010)	Infraestrutura e entorno
DomAdeqSN	Porcentagem de domicílios particulares permanentes com moradia adequada - sem identificação do logradouro	Censo (2010)	Infraestrutura e entorno
DomAdeqCN	Porcentagem de domicílios particulares permanentes com moradia adequada - com identificação do logradouro	Censo (2010)	Infraestrutura e entorno
DomSemiAdeq	Porcentagem de domicílios particulares permanentes com moradia semi-adequada	Censo (2010)	Infraestrutura e entorno
DomInadeq	Porcentagem de domicílios particulares permanentes com moradia inadequada	Censo (2010)	Infraestrutura e entorno
DomPosseOutro	Porcentagem de domicílios – outra forma de posse da moradia	Censo (2010)	Infraestrutura e entorno
DomSBan	Porcentagem de domicílios particulares permanentes sem banheiro de uso exclusivo dos moradores	Censo (2010)	Infraestrutura e entorno
DomNBanDom	Média do número banheiros por domicílio	Censo (2010)	Infraestrutura e entorno
DomNBanHab	Média do número de banheiros por habitante	Censo (2010)	Infraestrutura e entorno
AguaRede	Porcentagem de domicílios particulares permanentes com abastecimento de água da rede geral	Censo (2010)	Infraestrutura e entorno
AguaNascente	Porcentagem de domicílios particulares permanentes com abastecimento de água de poço ou nascente na propriedade	Censo (2010)	Infraestrutura e entorno
AguaCisterna	Porcentagem de domicílios particulares permanentes com abastecimento de água da chuva armazenada em cisterna	Censo (2010)	Infraestrutura e entorno
AguaOutra	Porcentagem de domicílios particulares permanentes com outra forma de abastecimento de água	Censo (2010)	Infraestrutura e entorno
EsgotoRede	Porcentagem de domicílios particulares permanentes com esgotamento sanitário pela rede geral	Censo (2010)	Infraestrutura e entorno
EsgotoSeptica	Porcentagem de domicílios particulares permanentes com esgotamento sanitário por fossa séptica	Censo (2010)	Infraestrutura e entorno
EsgotoRudimentar	Porcentagem de domicílios particulares permanentes com esgotamento sanitário rudimentar	Censo (2010)	Infraestrutura e entorno
EsgotoVala	Porcentagem de domicílios particulares permanentes com esgotamento sanitário por vala	Censo (2010)	Infraestrutura e entorno

EsgotoRio	Porcentagem de domicílios particulares permanentes com esgotamento sanitário pelo rio	Censo (2010)	Infraestrutura e entorno
EsgotoOutro	Porcentagem de domicílios particulares permanentes com outra forma de esgotamento sanitário	Censo (2010)	Infraestrutura e entorno
LixoLimpeza	Porcentagem de domicílios particulares permanentes com lixo coletado por serviço de limpeza	Censo (2010)	Infraestrutura e entorno
LixoQueimado	Porcentagem de domicílios particulares permanentes com lixo queimado na propriedade	Censo (2010)	Infraestrutura e entorno
LixoAterrado	Porcentagem de domicílios particulares permanentes com lixo enterrado na propriedade	Censo (2010)	Infraestrutura e entorno
LixoJogado	Porcentagem de domicílios particulares permanentes com lixo jogado em terreno baldio ou logradouro	Censo (2010)	Infraestrutura e entorno
LixoRio	Porcentagem de domicílios particulares permanentes com lixo jogado em rio, lago ou mar	Censo (2010)	Infraestrutura e entorno
LixoOutro	Porcentagem de domicílios particulares permanentes com outro destino do lixo	Censo (2010)	Infraestrutura e entorno
LixoCacamba	Porcentagem de domicílios particulares permanentes com lixo coletado em caçamba de serviço de limpeza	Censo (2010)	Infraestrutura e entorno
Ren0SM	Porcentagem de domicílios particulares sem rendimento nominal mensal domiciliar per capita	Censo (2010)	Emprego e renda
RenMeioSM	Porcentagem de domicílios particulares com rendimento nominal mensal domiciliar per capita de até 1/2 salários mínimos	Censo (2010)	Emprego e renda
Ren1a2SM	Porcentagem de domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1 a 2 salários mínimos	Censo (2010)	Emprego e renda
Ren3SM	Porcentagem de pessoas responsáveis com rendimento nominal mensal de até 3sm	Censo (2010)	Emprego e renda
RenPopDependente	Porcentagem da população dependente	Censo (2010)	Emprego e renda
RenPopAtiva	Porcentagem da população economicamente ativa	Censo (2010)	Emprego e renda
RenResp3SM	Porcentagem de responsáveis por domicílio com renda de até 3 salários mínimos	Censo (2010)	Emprego e renda
RenRespMedia	Renda média do responsável pelo domicílio	Censo (2010)	Emprego e renda
NDenDom	Densidade domiciliar por km²	Grade celular (Censo 2010)	Sociodemográficas
NDenPop	Densidade populacional por km²	Grade celular (Censo 2010)	Sociodemográficas
NMoradores	Média do número de moradores em domicílios particulares permanentes	Censo (2010)	Sociodemográficas
NPes10Alf	Porcentagem de pessoas alfabetizadas com 10 anos ou mais	Censo (2010)	Sociodemográficas
NRespAlf	Porcentagem de pessoas responsáveis alfabetizados	Censo (2010)	Sociodemográficas
NRespFem	Porcentagem de pessoas responsáveis do sexo feminino	Censo (2010)	Sociodemográficas
NRespIdade	Idade média dos responsáveis	Censo (2010)	Sociodemográficas
NResp30	Porcentagem de pessoas responsáveis com menos de 30 anos	Censo (2010)	Sociodemográficas
NResp30NAlf	Porcentagem de pessoas responsáveis com menos de 30 anos não-alfabetizadas	Censo (2010)	Sociodemográficas
mchefe_fmenor	Mulheres chefes de família e com filhos menores de 15 anos	Atlas	Atlas

vulner_dia	População ocupada vulnerável à pobreza que retorna diariamente do trabalho	Atlas	Atlas
dom_vulner_idoso	População em domicílios vulneráveis e com idoso	Atlas	Atlas
t_analf_18m	Taxa de analfabetismo - 18 anos ou mais	Atlas	Atlas
t_analf_25m	Taxa de analfabetismo - 25 anos ou mais	Atlas	Atlas
t_c0a5_fora	Porcentagem de crianças de 0 a 5 anos que não frequentam a escola	Atlas	Atlas
t_c6a14_fora	Porcentagem de pessoas de 6 a 14 anos que não frequentam a escola	Atlas	Atlas
t_m10a17_filho	Porcentagem de mulheres de 10 a 17 anos que tiveram filhos	Atlas	Atlas
t_analf_15m	Taxa de analfabetismo da população de 15 anos ou mais de idade	Atlas	Atlas
t_cdom_fundin	Porcentagem de crianças que vivem em domicílios em que nenhum dos moradores tem o ensino fundamental completo	Atlas	Atlas
t_desocup18m	Taxa de desocupação da população de 18 anos ou mais de idade	Atlas	Atlas
t_p18m_fundin_informal	Porcentagem de pessoas de 18 anos ou mais sem fundamental completo e em ocupação informal	Atlas	Atlas
t_pop18m_fundc	Porcentagem de 18 anos ou mais com fundamental completo	Atlas	Atlas
t_pop5a6_escola	Porcentagem de 5 a 6 anos na escola	Atlas	Atlas
t_pop11a13_ffun	Porcentagem de 11 a 13 anos nos anos finais do fundamental ou com fundamental completo	Atlas	Atlas
t_pop15a17_fundc	Porcentagem de 15 a 17 anos com fundamental completo	Atlas	Atlas
t_pop18a20_medioc	Porcentagem de 18 a 20 anos com médio completo	Atlas	Atlas
prosp_soc	Prosperidade Social	Atlas	Atlas
espvida	Esperança de vida ao nascer	Atlas	Atlas
t_mort2	Mortalidade até 1 ano de idade	Atlas	Atlas
t_fmor6	Mortalidade até 5 anos de idade	Atlas	Atlas
t_fectot	Taxa de fecundidade total	Atlas	Atlas
t_env	Taxa de envelhecimento	Atlas	Atlas
vulner15a25	População vulnerável de 15 a 24 anos	Atlas	Atlas
t_densidadem3	Porcentagem da população em domicílios com densidade > 2	Atlas	Atlas
i_gini	Índice de Gini	Atlas	Atlas
t_p15a24_nada	Porcentagem de pessoas de 15 a 24 anos que não estudam, não trabalham e possuem renda domiciliar per capita igual ou inferior a meio salário mínimo (de 2010)	Atlas	Atlas
t_vulner_depende_idosos	Porcentagem de pessoas em domicílios com renda per capita inferior a meio salário mínimo (de 2010) e dependentes de idosos	Atlas	Atlas
rdpc_def_vulner	Renda per capita dos vulneráveis à pobreza	Atlas	Atlas
t_nremunerado_18m	Porcentagem dos ocupados sem rendimento - 18 anos ou mais	Atlas	Atlas
t_vulner_mais1h	Porcentagem de pessoas que vivem em domicílios com renda per capita inferior a meio salário mínimo (de 2010) e que gastam mais de uma hora até o trabalho	Atlas	Atlas
t_renda_trab	Porcentagem da renda proveniente de rendimentos do trabalho	Atlas	Atlas
t_carteira_18m	Porcentagem de empregados com carteira - 18 anos ou mais	Atlas	Atlas
t_scarteira_18m	Porcentagem de empregados sem carteira - 18 anos ou mais	Atlas	Atlas
t_setorpublico_18m	Porcentagem de trabalhadores do setor público - 18 anos ou mais	Atlas	Atlas
t_contapropria_18m	Porcentagem de trabalhadores por conta própria - 18 anos ou mais	Atlas	Atlas

t_empregador_18m	Porcentagem de empregadores - 18 anos ou mais	Atlas	Atlas
t_formal_18m	Grau de formalização dos ocupados - 18 anos ou mais	Atlas	Atlas
t_atividade10a15	Taxa de atividade das pessoas de 10 a 14 anos de idade	Atlas	Atlas
P1	Renda média (per capita) da família	CadÚnico (2020)	CadÚnico
P2	Quantidade de pessoas nas famílias cadastradas	CadÚnico (2020)	CadÚnico
P3	Domicílios improvisados	CadÚnico (2020)	CadÚnico
P4	Quantidade de cômodos nos domicílios	CadÚnico (2020)	CadÚnico
P5	Domicílios com densidade excessiva (> 3 pessoas por dormitório)	CadÚnico (2020)	CadÚnico
P6	Material piso = Terra	CadÚnico (2020)	CadÚnico
P7	Material piso = Cimento	CadÚnico (2020)	CadÚnico
P8	Material piso = Madeira aproveitada	CadÚnico (2020)	CadÚnico
P9	Material piso = P6+P7+P8	CadÚnico (2020)	CadÚnico
P10	Material parede = alvenaria sem revestimento	CadÚnico (2020)	CadÚnico
P11	Material parede = taipa	CadÚnico (2020)	CadÚnico
P12	Material parede = taipa não revestida	CadÚnico (2020)	CadÚnico
P13	Material parede = madeira aproveitada	CadÚnico (2020)	CadÚnico
P14	Material parede = palha	CadÚnico (2020)	CadÚnico
P15	Material parede = P10 até C14	CadÚnico (2020)	CadÚnico
P16	Domicílio sem água encanada	CadÚnico (2020)	CadÚnico
P17	Abastecimento de água por poço ou nascente	CadÚnico (2020)	CadÚnico
P18	Abastecimento de água por cisterna	CadÚnico (2020)	CadÚnico
P19	Outro tipo de abastecimento de água	CadÚnico (2020)	CadÚnico
P20	Abastecimento de água = P16 A P19	CadÚnico (2020)	CadÚnico
P21	Domicílio sem banheiro (co_banheiro_domic_fam = 2)	CadÚnico (2020)	CadÚnico
P22	Esgotamento Sanitário = Fossa rudimentar (co_escoa_sanitario_domic_fam = 3)	CadÚnico (2020)	CadÚnico
P23	Esgotamento Sanitário = Vala a céu aberto (co_escoa_sanitario_domic_fam = 4)	CadÚnico (2020)	CadÚnico
P24	Esgotamento Sanitário = Direto para rio, lago ou mar (co_escoa_sanitario_domic_fam = 5)	CadÚnico (2020)	CadÚnico
P25	Esgotamento Sanitário = Outra forma (co_escoa_sanitario_domic_fam = 6)	CadÚnico (2020)	CadÚnico
P26	Esgotamento Sanitário = P22 até P25	CadÚnico (2020)	CadÚnico
P27	Coleta de Lixo = Coletado Indiretamente (co_destino_lixo_domic_fam = 2)	CadÚnico (2020)	CadÚnico
P28	Coleta de Lixo = Queimado ou enterrado na propriedade (co_destino_lixo_domic_fam = 3)	CadÚnico (2020)	CadÚnico

P29	Coleta de Lixo = Jogado em terreno baldio ou logradouro (co_destino_lixo_domic_fam = 4)	CadÚnico (2020)	CadÚnico
P30	Coleta de Lixo = Jogado em rio ou mar (co_destino_lixo_domic_fam = 5)	CadÚnico (2020)	CadÚnico
P31	Coleta de Lixo = Outro Destino (co_destino_lixo_domic_fam = 6)	CadÚnico (2020)	CadÚnico
P32	Coleta de Lixo = P27 até P31	CadÚnico (2020)	CadÚnico
P33	Coleta de Lixo = P28 até P31	CadÚnico (2020)	CadÚnico
P34	Iluminação = Elétrica com medidor comunitário	CadÚnico (2020)	CadÚnico
P35	Iluminação = Elétrica sem medidor	CadÚnico (2020)	CadÚnico
P36	Iluminação = Óleo, querosene, gás, vela ou outra forma	CadÚnico (2020)	CadÚnico
P37	Iluminação = P43 a P36	CadÚnico (2020)	CadÚnico
P38	Calçamento = Não existe	CadÚnico (2020)	CadÚnico
P39	Pessoas por domicílio	CadÚnico (2020)	CadÚnico
P40	Famílias por domicílio	CadÚnico (2020)	CadÚnico
P41	Preço do aluguel	CadÚnico (2020)	CadÚnico
P42	Ônus Excessivo com Aluguel (Preço Aluguel > 30Porcentagem da renda domiciliar)	CadÚnico (2020)	CadÚnico
P43	Responsável não alfabetizados	CadÚnico (2020)	CadÚnico
P44	Escolaridade Responsável = ensino fundamental ou menos	CadÚnico (2020)	CadÚnico
P45	Escolaridade Responsável = frequentou ensino superior	CadÚnico (2020)	CadÚnico

Fonte: Elaboração própria, 2021.

Apêndice 2 - Geocodificação do CadÚnico

Os dados do CadÚnico foram georreferenciados pelo processo de geocodificação de endereços. A geocodificação é uma operação em Sistemas de Informação Geográfica (SIG) de conversão de endereços em coordenadas geográficas. A depender da quantidade de dados, a geocodificação pode ser um processo custoso e lento e esse foi um dos desafios deste estudo: realizar a geocodificação de um grande volume de endereços (cerca de 2,2 milhões de endereços) em pouco tempo e sem custos adicionais.

Cabe também destacar que as operações de geocodificação podem apresentar níveis de precisão distintos para cada um dos elementos inseridos. Esta variação pode estar associada a uma série de fatores: formação do endereço, abreviações e apelidos para endereços, correta grafia do endereço, existência do registro do logradouro na base de dados, entre outros. A metodologia adotada nesta etapa foi a combinação de ferramentas de geocodificação e os detalhes serão discutidos a seguir.

Dados Cadúnico

O universo de endereços a serem georreferenciados totalizou o número de 2,272 milhões para os 6 polos da pesquisa, distribuídos regionalmente da seguinte forma:

TABELA A2.1 – Número de endereços do CadÚnico por polo.

Polos	Porto Alegre	Juazeiro do Norte	Brasília	Recife	Belo Horizonte	Marabá
Endereços CadÚnico	319.938	200.590	336.014	804.921	464.038	146.684

Fonte: Elaboração própria, 2021.

A2.1 Métodos

O processo de geocodificação dos endereços do Cadúnico foi realizado em três etapas principais por três ferramentas distintas de geocodificação:

1. **Etapa 1: Galileo**

O Galileo é um software proprietário de solução para espacialização ou geocodificação de endereços no Brasil. O sistema Galileo apresenta, ao final da execução, o nível de precisão individual de cada endereço, os níveis de precisão e seus significados são classificados conforme abaixo. (Quadro A2.1)

QUADRO A2.1 – Níveis de precisão do Galileo.

Classificação	Definição
4 Estrelas	Endereço completo, composto por logradouro e número, ou cep com oito dígitos e número.
3 Estrelas	O endereço foi encontrado em nível de rua ou cep com oito dígitos, sem inclusão do número. O ponto encontrado representa o centro da rua.
2 Estrelas	A geocodificação foi realizada e o endereço foi encontrado utilizando uma das seguintes estratégias: cep com 7, 6 ou 5 dígitos, ou nome da rodovia.
1 Estrela	A geocodificação foi realizada e indica como resultados o município ou estado encontrado.
Erro	Ocorreu algum erro no processamento do endereço.

Fonte: Galileo, 2021.

Para esta primeira etapa foi necessário realizar a adequação da base de dados dos endereços do Cadúnico para utilização do software Galileo, procedimentos como a inclusão do nome do município a partir do código municipal do IBGE e a separação por coluna de cada uma das informações dos endereços (rua, número, bairro, cep) foram realizados. Após estes procedimentos iniciais, utilizou-se o software proprietário Galileo para atribuir as coordenadas geográficas de cada família cadastrada no Cadúnico a partir do seu endereço completo.

2. Etapa 2: HERE API

A HERE é uma empresa de tecnologias de soluções de mapeamento digital. Um dos serviços disponíveis é a ferramenta de geocodificação "HERE Geocoding and Search" para desenvolvedores. A empresa permite acesso a duas chaves API[1] que permitem a geolocalização de até 250.000 endereços por mês gratuitamente.

O processamento da ferramenta foi realizado no ambiente R, no qual é possível acessar a aplicação do HERE Geocoding and Search através da chave API, o script está disponível no GitHub do projeto. Os resultados, assim como os do Galileo, apresentam métrica de avaliação dos endereços geocodificados, conforme listados abaixo (Quadro A2.2).

QUADRO A2.2 - Níveis de precisão do HERE Geocoding and Search

Classificação	Definição
Country	Qualidade do resultado no nível de informação do país.
State	Qualidade do resultado no nível de informação do estado.
County	Qualidade do resultado no nível de informação do condado.
City	Qualidade do resultado no nível de informação do município/cidade.
District	Qualidade do resultado no nível de informação do distrito.
Subdistrict	Qualidade do resultado no nível de informação do subdistrito.
PostalCode	Qualidade do resultado no nível de informação do CEP.
Street	Qualidade do resultado no nível de informação do logradouro.
HouseNumber	Qualidade do resultado no nível de informação do número da casa.
Building	Qualidade do resultado no nível de informação dos edifícios.

Fonte: HERE, 2021.

Devido ao grande volume de dados, nesta segunda etapa optou-se por realizar a geocodificação apenas para os endereços que o processamento do Galileo retornou com qualidade abaixo das "4 estrelas", cerca de 942 mil famílias cadastradas no Cadúnico para os 6 polos da pesquisa.

[1] Para desenvolver uma programação mais prática e organizada, retornando somente os dados que o programa necessita é preciso utilizar um sistema externo para obtenção de dados, umas das formas é através de API. Em seu site (https://developer.here.com/ documentation/ geocoding-search -api/) é possível encontrar toda a sua documentação em relação à ferramenta.

3. **Etapa 3: *Geocoding with Google Sheets***

Geocoding with Google Sheets é um google script que permite a geocodificação de endereços para coordenadas de latitude e longitude ou atribuir coordenadas para endereços. Esta terceira etapa foi realizada concomitante a segunda etapa (HERE), e também devido ao volume de dados, optou-se pelo processamento apenas dos endereços resultantes do Galileo com menos de "4 estrelas", cerca de 942 mil famílias cadastradas no Cadúnico para os 6 polos da pesquisa.

A2.2 Resultados

1. **Galileo**

O tempo de execução da geocodificação para cada um dos polos ficou no intervalo de 40 minutos à 1h 30 minutos. Ressalta-se as duas principais vantagens desta ferramenta: métrica de avaliação dos resultados e tempo de execução. Entretanto, muitos endereços (41,5% do total) tinham baixa qualidade de georreferenciamento - igual ou abaixo da avaliação de 3 estrelas - por isso buscou-se outras ferramentas de geocodificação.

2. **HERE API**

Para atender o número de transações necessárias e permitidas por mês gratuitamente, foi necessário a utilização de mais de uma conta cadastrada na plataforma HERE. Em relação ao tempo de execução do processamento, esse método é consideravelmente mais lento que o Galileo. Destaca-se também que apesar de permitir cerca de 250 mil operações por mês, não foi possível executá-las em um único processamento, sendo assim, realizou-se a divisão em "pacotes" de 30.000 endereços para o processamento no ambiente R, com cerca de 30 minutos de execução para cada "pacote".

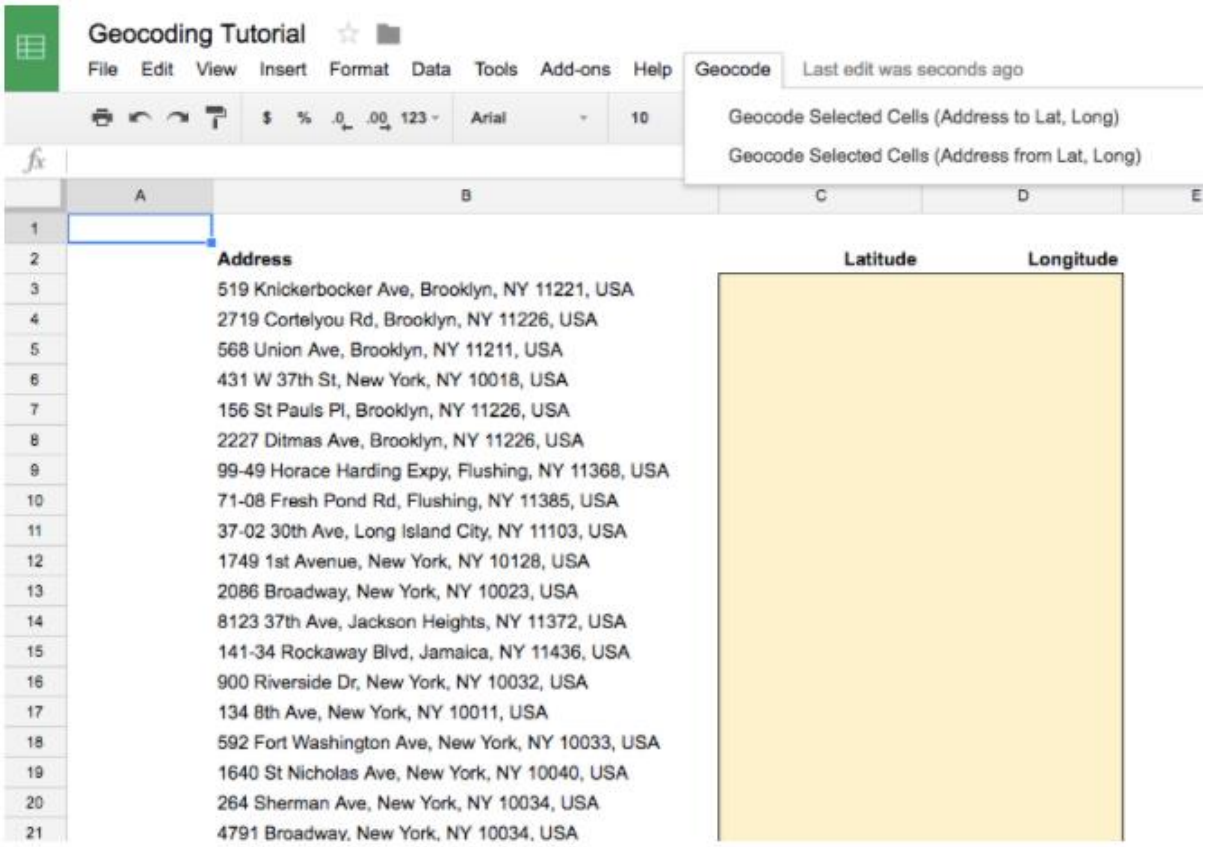
3. **Google Geocode**

Comparativamente aos outros dois métodos apresentados, é o mais simples para execução e possui maior tempo de processamento. Assim como para o método HERE, foi necessária a utilização de mais de uma conta institucional, nesse caso para reduzir o tempo de processamento devido ao grande número de dados e a limitação de geocodificação diária. Entre as vantagens e desvantagens deste método destacam-se:

Vantagens

- Acesso ao amplo banco de dados da Google, geocodificando endereços não encontrados pelos métodos anteriores, como endereços mais recentes (novas ruas e avenidas, novos bairros);
- Método gratuito para 10.000 endereços por dias, a partir do uso de contas institucionais;
- Aplicação simples, após inserir o script no Google Sheets, uma nova aba é criada, o 'Geocode', e é possível escolher qual o tipo de geocodificação desejada a partir da seleção da coluna de endereços.

FIGURA A2.1 - Aba do Google Sheets para o Geocode.



Fonte: http://willgeary.github.io/data/2016/11/04/Geocoding-with-Google-Sheets.html.

Desvantagens

- Não possui método de avaliação da geocodificação;
- Tempo de processamento lento, por vezes é difícil atingir os 10.000 endereços por dia em cada uma das contas utilizadas.

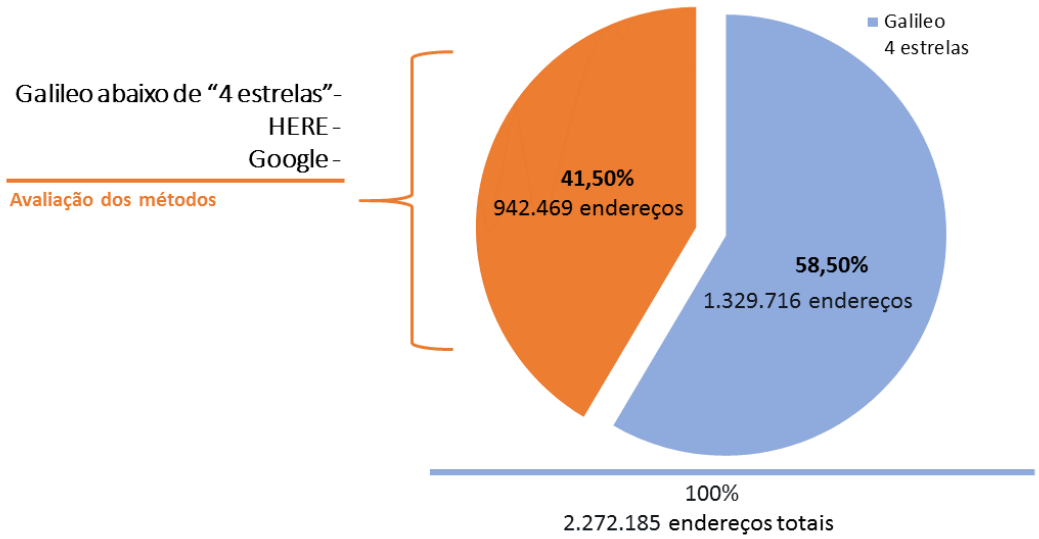
A2.3 Seleção e avaliação dos métodos - alocando os endereços geocodificados

Ao final das três etapas de geocodificação, obteve-se os endereços para quase toda a base de famílias do CadÚnico para os 6 polos da pesquisa. Na maioria dos casos, existem três opções (Galileo, HERE e Google) de coordenadas lat long para o mesmo endereço cadastrado. Por conseguinte, realizou-se uma avaliação dos métodos para escolha da coordenada mais precisa para cada um dos endereços. Em virtude do grande volume de dados a serem avaliados, foram estabelecidos critérios para alocação das coordenadas/métodos para cada endereço cuja descrição é apresentada a seguir.

Em relação aos três métodos, a geocodificação utilizando o Galileo foi a única a não atribuir coordenadas fora dos limites municipais dos polos, enquanto o HERE e o Google atribuíram coordenadas até fora dos limites do Brasil. Os erros grosseiros, como coordenadas fora do município cadastrado no endereço, foram eliminados em um primeiro filtro de resultados.

O segundo filtro de resultados foi conduzido para a realização do processamento da geocodificação pelo HERE e pelo Google, no qual foram aceitos todos os resultados "4 estrelas" do Galileo e foram geocodificados novamente os outros resultados pelos outros métodos. A avaliação resulta, portanto, da comparação da precisão das coordenadas resultantes dos métodos para um número menor de endereços (41,5%) (Figura A2.2).

FIGURA A2.2 - Etapas de Geocodificação

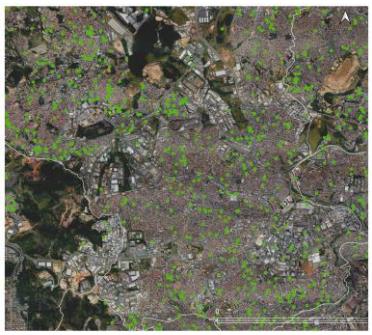


Fonte: Elaboração própria, 2021.

A2.3.1 Avaliação dos métodos

A Figura A2.3 apresenta os resultados das geocodificações para o polo Belo Horizonte pelos três métodos. Observa-se como o método Galileo possui menor cobertura do território porque está sendo analisado apenas os resultados de baixa qualidade, no qual, endereços diferentes são georreferenciados em um mesmo ponto (Ex: na mesma rua - 3 estrelas; no mesmo cep ou rodovia - 2estrelas; no município - 1 estrela ou erro). Ainda na Figura A2.3 destaca-se o resultado pelo Google sheets, que apesar de não ter métrica de avaliação própria, possui a maior cobertura do território ao ser comparado com os outros dois métodos.

FIGURA A2.3 - Endereços do Cadúnico do polo Recife Geocodificados

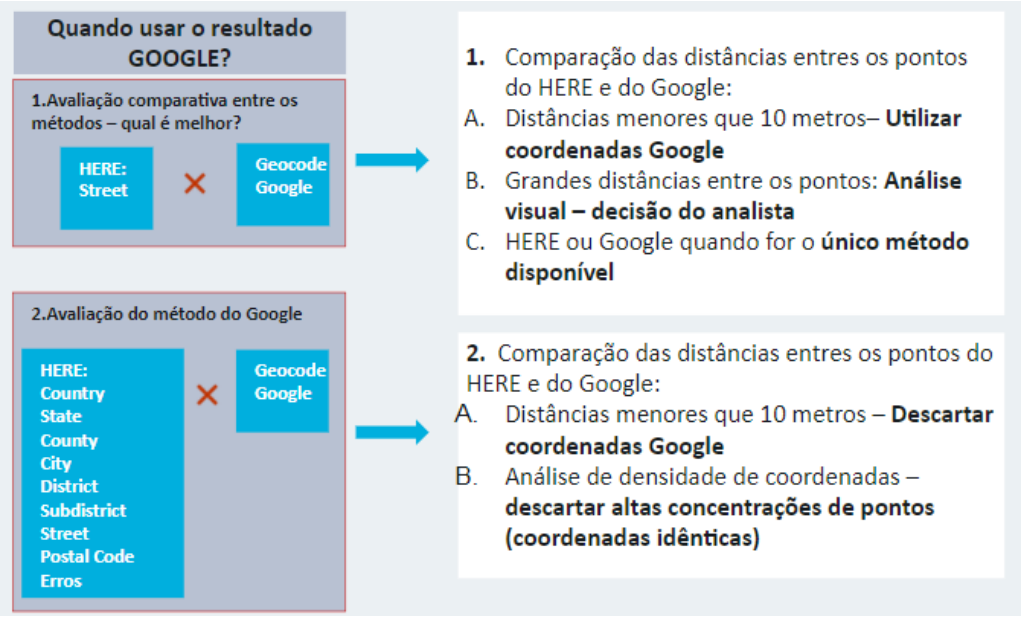
		
Galileo	HERE	Google

Fonte: Elaboração própria, 2021.

A primeira decisão nesta primeira análise visual foi a de dispensar os resultados do Galileo e avaliar comparativamente os resultados HERE e Google. É importante lembrar que os resultados do HERE também possuem métrica própria de avaliação, semelhante ao Galileo. Portanto, optou-se por aceitar todos os endereços HERE avaliados como “Building” ou “House Number”, que são 338.799 endereços, ou seja, 14,9% de todos os endereços.

Os próximos critérios para escolha entre o método HERE (com menor qualidade que “Building” e HouseNumber”) e o Google dependeu da decisão do analista para cada polo. Os passos adotados são descritos na Figura A2.4.

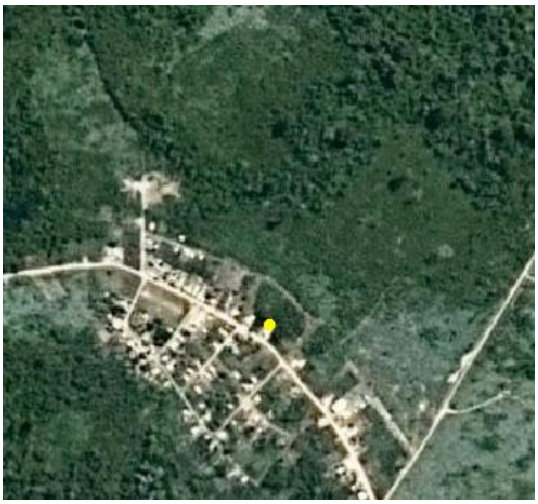
FIGURA A2.4 - Critérios de seleção das coordenadas



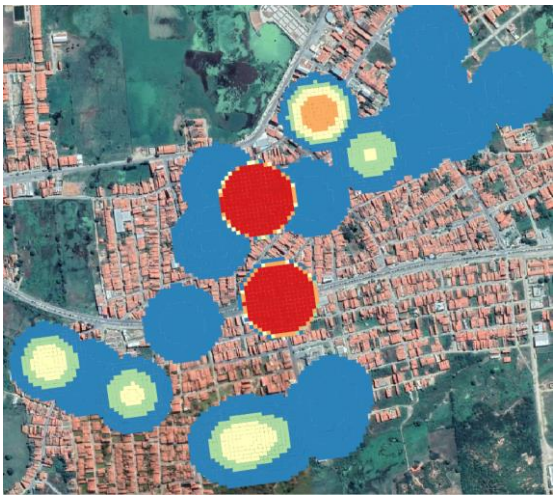
Fonte: Elaboração própria, 2021.

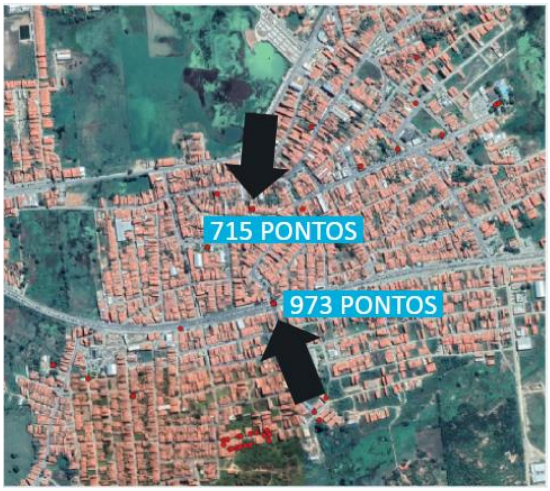
Algumas situações encontradas são exemplificadas na Figura A2.5. A Figura A2.6 apresenta uma visão geral do processo de escolha entre as três estratégias de geocodificação.

FIGURA A2.5 - Avaliação dos métodos Google e HERE.

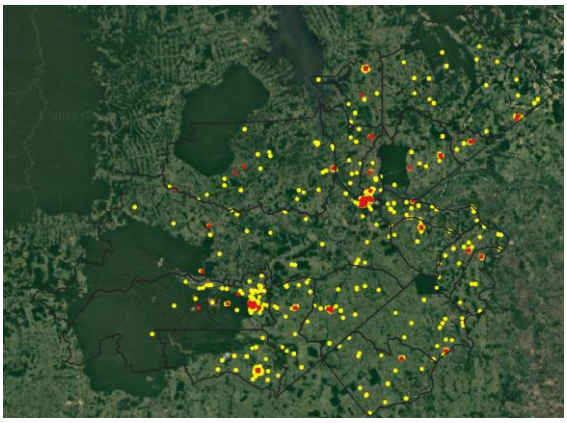


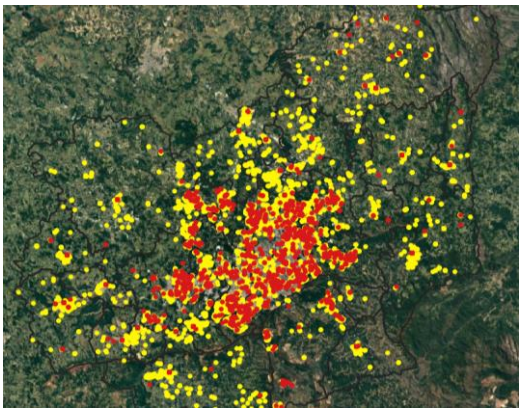
(a) Pontos geocodificados pelo Google. A decisão de favorecer a geocodificação do Google se deu pelo amplo alcance territorial da ferramenta, principalmente nas áreas de expansão urbana.





(b) Descarte de altas concentrações de pontos (geocodificações de baixa qualidade)

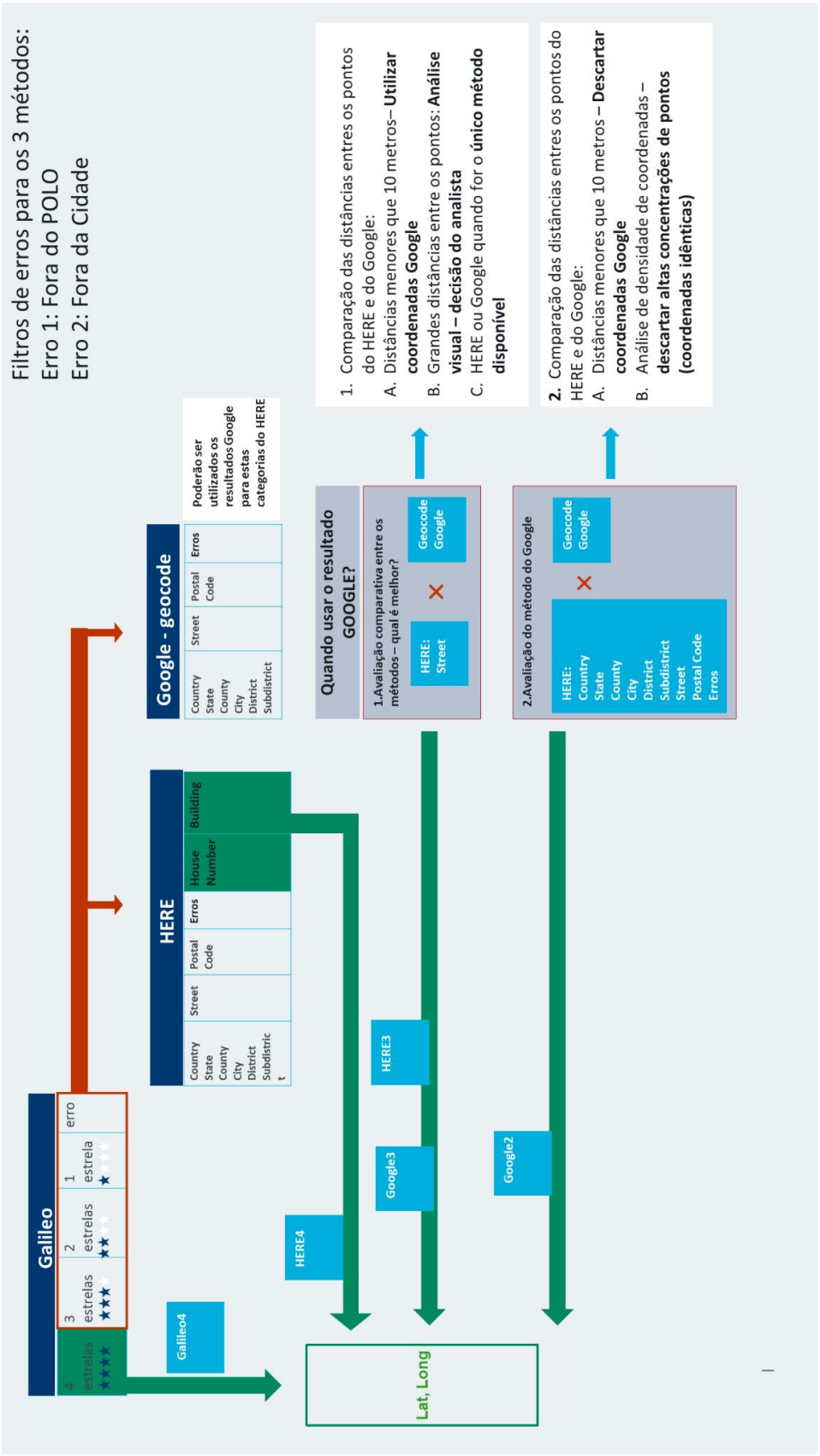




(c) Comparação visual dos resultados

Fonte: Elaboração própria, 2021.

FIGURA A2.6 - Resumo dos critérios de escolha



Fonte: Elaboração própria, 2021.

A2.4 Resultados

As Tabelas A2.2 e A2.3 apresentam os resultados do processo de geocodificação dos endereços no Cadúnico, distinguindo o total de pontos geocodificados pelos distintos métodos, bem como o total de pontos não geocodificados (erros por polo e total).

TABELA A2.2 Resultados da Geocodificação dos endereços do Cadúnico para os 6 polos.

Polos	Porto Alegre	Juazeiro do Norte	Brasília	Recife	Belo Horizonte	Marabá
TOTAL	319.938 (100%)	200.590 (100%)	336.014 (100%)	804.921 (100%)	464.038 (100%)	146.684 (100%)
GALILEO4	229.736 (72%)	96.672 (48%)	57.363 (17%)	514.371 (64%)	367.373 (79%)	64.201 (44%)
HERE4	58.166 (18%)	10.666 (12%)	106.254 (31%)	94.462 (12%)	68.338 (15%)	913 (1%)
HERE3	418 (0,1%)	14.706 (0,5%)	0	0	312 (0,1%)	25.610 (17%)
GOOGLE2	15.184 (4,7%)	24.029 (5%)	49.104 (14%)	89.183 (11%)	10.280 (2%)	9.308 (6%)
GOOGLE3	12.328 (3,9%)	1.020 (7%)	25.935 (7%)	19.198 (2%)	6.806 (1%)	16.080 (11%)
Erros	4.106 (1,3%)	53.497 (26%)	97.358 (28%)	87.707	10.929 (2,36%)	30.572 (21%)

Fonte: Elaboração própria, 2021.

TABELA A2.3 Resumo dos Resultados por método.

TOTAL	2.272.185	100%
GALILEO4	1.329.716	58,5%
HERE4	338.799	14,9%
HERE3	41.046	1,8%
GOOGLE2	197.088	8,7%
GOOGLE3	81.367	3,6%
Erros	284.169	12,5%

Fonte: Elaboração própria, 2021.

Apêndice 3 - Análise Comparativa de Técnicas de Classificação

Uma vez finalizado o processo de construção e integração de variáveis em uma base celular comum, três diferentes técnicas de classificação foram analisadas e comparadas sob as mesmas condições, ou seja, considerando o mesmo conjunto de variáveis e as mesmas áreas. O objetivo desta etapa é apontar a técnica de classificação mais adequada para a identificação de Núcleos Urbanos Informais.

Duas dessas técnicas, regressão logística e análise discriminante, respondem diretamente às diretrizes estabelecidas para o desenvolvimento da Metodologia NUI por serem adequadas para a construção de **superfícies de probabilidade** da presença de NUI e permitirem a fácil interpretação dos parâmetros gerados pelo modelo. Já a terceira técnica, árvore de decisão (algoritmo C5.0), foi analisada em virtude do destaque que as técnicas de *machine learning* vem recebendo em estudos internacionais de identificação de assentamentos precários (MAHABIR et al., 2018; FRIESEN et al., 2018; RIBEIRO, 2015).

Dessa forma, a comparação foi conduzida em três etapas: **(1) seleção de variáveis** relevantes em cada polo de pesquisa; **(2) treinamento dos modelos**; e **(3) calibração e avaliação dos modelos**. Considerando o fato de que a Metodologia NUI busca fornecer subsídios para o aprimoramento dos dados sobre NUI a partir de informações incompletas, os aglomerados subnormais (AGSN) foram utilizados como variável dependente no treino. Já os dados do levantamento de campo (NUI), mais completos, são utilizados para a avaliação dos modelos.

A descrição de cada uma das etapas da análise comparativa dos modelos é apresentada nas Seções A3.2, A3.3 e A3.4. Antes disso, uma breve apresentação das três técnicas comparadas é realizada (Seção A3.1). Por fim, a Seção A3.5 apresenta as conclusões da análise.

A3.1. Regressão Logística, Análise Discriminante e Árvore de Decisão

A **regressão logística** é uma técnica estatística multivariada utilizada para a modelagem da probabilidade de ocorrência de determinado evento (neste caso, a Presença de Núcleos Urbanos Informais) ou classificação (Presença de NUI ou Ausência de NUI). Sua fórmula apresenta diversas semelhanças com a regressão linear múltipla, mas ao invés de prever o valor da variável (Y), ela estima a sua probabilidade de ocorrência (P(Y)). Para gerar a classificação binária, deve ser escolhido um limiar de probabilidade, que, quando não especificado, assume-se como sendo de 0,5 ou 50%. Considerando, portanto, o limiar padrão de 0,5, P(Y) >= 0,5 corresponderia à Presença de NUI e a probabilidade de Y menor que 0,5 corresponderia à Ausência de NUI.

A **análise discriminante** é uma técnica utilizada para encontrar as combinações lineares de atributos que melhor separam duas ou mais classes. Ela é baseada numa função discriminante que busca minimizar o erro de classificação, de tal maneira que as classes sejam o mais distintas possível. A saída do modelo também descreve a probabilidade de Y (P(Y)) e quando usada para a classificação de uma variável dependente binária, o limiar de corte padrão é o mesmo usado na regressão logística (0,5 ou 50%).

A regressão logística e a análise discriminante são consideradas duas formas diferentes de chegar em um mesmo resultado (FIELD, 2012, p. 738; HILBE, 2009, p. 540). Ambas são técnicas que permitem

classificar (e, portanto, identificar) espacialidades como AGSN ou NUI. Entretanto, embora sejam análogas, cada uma das técnicas possui características próprias, vantagens e desvantagens.

A análise discriminante assume, para estimar os coeficientes da função de classificação resultante, que os dados de cada uma das classes que se pretende classificar são provenientes de uma população de distribuição aproximadamente Gaussiana. Já a regressão logística não necessita dessa premissa, usando o estimador de máxima verossimilhança para as estimativas (GARETH et al., 2014). De maneira geral, a análise discriminante é preferida à regressão logística quando: (1) há previsão perfeita entre a variável resposta e os preditores; (2) quando há quase igualdade de matrizes de covariância no grupo; e (3) quando todas as variáveis preditoras possuem distribuição normal (HILBE, 2009, p. 532). Em todos os outros casos, a regressão logística é considerada uma técnica mais robusta e generalizável.

Algoritmos de **árvore de decisão** são os modelos de *machine learning* mais utilizados para classificação (WU et al., 2008), sendo considerados classificadores rápidos, versáteis e confiáveis. Eles partem de uma base de dados de treino para criar uma estrutura, chamada de "árvore", que pode ser utilizada para classificar novos casos. Os "nós" da árvore possuem um teste lógico, sendo o resultado usado para decidir qual "galho" a observação deve seguir a partir daquele nó. Os nós finais, chamados de "folhas", possuem uma classe ao invés de um teste lógico, que é atribuída às observações que cumprem todos os testes lógicos anteriores (QUINLAN, 1986; QUINLAN, 1993).

O algoritmo escolhido, C5.0 (ou see5), assim como os seus precursores C4.5 e ID3, utiliza fórmulas baseadas na teoria da informação para escolher os testes que apresentam o maior ganho de informação. O maior dilema relacionado à árvore de decisão é o *overfitting*: uma árvore de decisão que classifica corretamente cada observação da base de dados de treinamento geralmente possui um desempenho abaixo do esperado na classificação de dados reais e, consequentemente, na avaliação do modelo. É por isso que os algoritmos mais modernos de árvore de decisão incluem o crescimento de uma árvore maior, seguido pela "poda" da árvore, removendo galhos quando há baixo ganho de informação (WU et al., 2008). Mesmo após a poda, as árvores geradas pelo algoritmo são tão grandes que é difícil a sua representação gráfica para interpretação, sendo comum a análise apenas dos primeiros níveis da árvore, onde existe o maior ganho de informação.

Ao contrário dos modelos regressão logística e análise discriminante, o resultado da árvore de decisão é essencialmente categórico e não contínuo (probabilidade). É possível gerar uma superfície de probabilidade baseada na distribuição observada, embora estudos apontem que essa probabilidade dificilmente será estável, ou seja, uma pequena perturbação na base de dados como a remoção de uma observação ou a validação cruzada pode alterar significativamente a árvore de decisão e, consequentemente, as probabilidades resultantes (WU et al., 2008). Apesar de suas limitações, a superfície de probabilidade da árvore de decisão será comparada sob as mesmas condições das superfícies geradas pelos outros modelos.

Essa consideração é importante porque quando aplicamos um modelo de classificação que consegue identificar a probabilidade de ocorrência de um evento, desejamos que as probabilidades estimadas reflitam a verdadeira probabilidade subjacente da amostra. Em geral, utiliza-se um limiar de classificação equivalente a 0,5 para classificar a ocorrência de um evento (no caso, a Presença de NUI). Entretanto, em situações em que se tem conhecimento da presença do evento, mas há incertezas

sobre a ausência, denominadas na análise de classificação como "pseudo-ausência" (HIJMANS; ELITH, 2017), os limiares relativos à presença/ausência do evento tornam-se muito incertos se definidos apenas pela base de dados (PHILIPS et al., 2009). Nesses casos, Philips e Elith (2011) recomendam que os **limiares de classificação** das superfícies de probabilidade sejam complementados por informações externas à base de dados original.

Dessa forma, buscamos definir novos limiares de classificação a partir das superfícies de probabilidade, de forma a explorar o desempenho potencial dos modelos na etapa de classificação. A métrica escolhida para a calibrar os modelos é a mesma utilizada na etapa de avaliação: o coeficiente de concordância Kappa.

O **coeficiente Kappa** é uma métrica desenhada originalmente para avaliar a concordância entre dois avaliadores, levando em consideração a probabilidade de a concordância ocorrer ao acaso. Ela também é muito utilizada para avaliar modelos quando a distribuição de classes é desigual – sendo este o caso dos Núcleos Informais Urbanos.

Sua fórmula é $\kappa = \frac{O-E}{1-E}$, onde O é a acurácia observada e E é a acurácia esperada com base nos totais marginais da matriz de confusão. 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, 2013).

O coeficiente Kappa é considerado uma medida robusta e conservadora, já que dificilmente a concordância observada é alta. Há controvérsias em relação ao seu uso devido à dificuldade de interpretação, visto que situações de alta concordância podem ser identificadas mesmo quando o Kappa não está próximo de 1. Além disso, como ele leva em consideração a probabilidade de a concordância ocorrer ao acaso, não é recomendado comparar os coeficientes gerados para diferentes conjuntos de dados, que possuem distribuições diferentes (STEHMAN, 1997; FOODY, 2020). Apesar das controvérsias, o coeficiente Kappa é uma métrica adequada para comparar modelos estimados para o mesmo conjunto de dados.

A3.2 Seleção de variáveis

A seleção de variáveis foi conduzida antes da aplicação dos modelos, considerando critérios que envolveram análises da correlação e multicolinearidade, modelagem em etapas (*stepwise*), diversidade de fontes de dados e experiência e trocas com a equipe da Pesquisa. As variáveis selecionadas para cada polo de pesquisa são apresentadas no Quadro A3.1.

QUADRO A3.1 – Variáveis consideradas em cada polo da pesquisa

Variável	Descrição	Belo Horizonte	Brasília	Juazeiro do Norte	Marabá	Porto Alegre	Recife
Decliv	Declividade média do terreno	X		X	X		X
APP	Faixa de 30 metros de curso d'água (% área ocupada)	X			X		
AltaTensao	Faixa de servidão de linha de alta tensão (% área ocupada)			X			
Dist50	Faixa de 50 metros de vias carroçáveis (% área ocupada)	X	X	X	X	X	X
Shape	Índice de forma (McGarigal & Marks, 1994) de quadras ou bolsões de ocupação (média): SHAPE>=1 (máxima regularidade = 1)					X	X
P2	Quantidade de pessoas nas famílias cadastradas no CadÚnico	X				X	
P10	Famílias cadastradas no CadÚnico em domicílio com alvenaria sem revestimento (material de parede)	X			X	X	
P35	Famílias cadastradas no CadÚnico sem medidor (iluminação elétrica)						X
P40	Famílias por domicílio (CadÚnico)			X			X
DenPop	Densidade populacional (IBGE, 2010)	X			X	X	
DenDom	Densidade domiciliar (IBGE, 2010)						X
Mmor	Média do número de moradores nos domicílios	X				X	X
Adequa	Domicílios com infraestrutura adequada (%)	X	X				
AguaRede	Domicílios com abastecimento de água da rede geral (%)				X		
Smed	Domicílios sem medidor (%)					X	X
Sesg	Domicílios com esgoto a céu aberto (%)	X		X		X	
Clix	Domicílios com lixo acumulado na rua (%)			X			
Silu	Domicílios sem iluminação pública (%)				X		
Sarb	Domicílios sem arborização (%)	X					
Siden	Domicílios sem identificação do logradouro (%)				X		
LixoCacamba	Domicílios com lixo coletado em caçambas de limpeza (%)	X	X				
LixoAterrado	Domicílios com lixo aterrado na propriedade (%)			X			
LixoQueimado	Domicílios com lixo queimado na propriedade (%)						X
MeioSM	Domicílios com renda per capita de até ½ salário mínimo (%)		X	X			
Res3SM	Responsáveis com renda de até 3 salários mínimos (%)				X	X	
IdaMedRes	Idade média do responsável			X	X		
MedBan	Média do número de banheiros no domicílio		X				
DomApar	Domicílios do tipo apartamento (%)		X				

DomImp	Domicílios improvisados (%)		X				
CEM08	Quantidade de banheiros por habitante					X	
CEM10	Responsáveis por domicílios não alfabetizados e com menos de 30 anos (%)		X				
CEM14	Renda média do responsável			X			X
t_fectot	Taxa de fecundidade total	X					
t_mort	Taxa de mortalidade total		X				

Fonte: Elaboração própria, 2021.

A3.3 Construção dos modelos

A construção dos três modelos – regressão logística, análise discriminante e árvore de decisão (C5.0), nos seis polos de pesquisa – Belo Horizonte, Brasília, Juazeiro do Norte, Marabá, Porto Alegre e Recife, foi realizado utilizando os aglomerados subnormais (AGSN) como variável dependente.

Os modelos regressão logística e análise discriminante apresentam coeficientes significativos para todas as variáveis selecionadas e podem ser facilmente interpretados, conforme os resultados apresentados para o polo Belo Horizonte na Tabela 3.1.

TABELA A3.1 – Coeficientes da regressão logística e análise discriminante (Belo Horizonte)

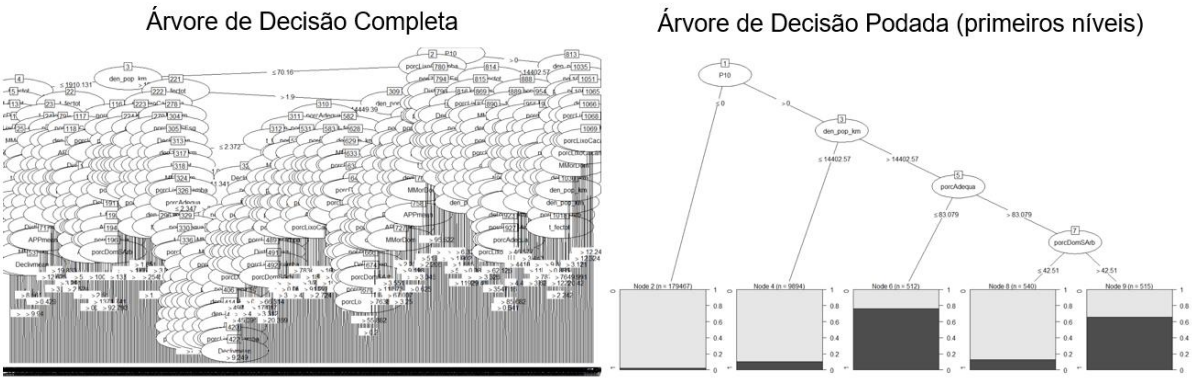
Variável	Coeficiente estimado e erro padrão (regressão logística)	Coeficientes de discriminantes lineares (análise discriminante)
Constante	-12,809 (0,148)	-
Decliv	0,074 (0,003)	0,043
APP	1,023 (0,056)	0,447
Dist50	1,609 (0,064)	0,333
P2	0,099 (0,010)	0,010
P10	0,236 (0,018)	0,754
DenPop	$1{,}75.10^{-4}$ (3.10^{-6})	2.10^{-4}
Mmor	0,309 (0,020)	0,141
Adequa	-0,01 (0,001)	-0,009
DomSEsg	0,013 (0,001)	0,013
DomSArb	0,011 (0,0005)	0,009
LixoCacamba	0,026 (0,001)	0,016
t_fectot	2,594 (0,053)	0,907

* Todos os coeficientes estimados são significativos ao nível de 1%.

Fonte: Elaboração própria, 2021.

Já a árvore de decisão completa, resultado do algoritmo C5.0 após a poda inicial que remove as decisões que apresentam baixo nível de ganho de informação, é muito grande em todos os polos, tornando a sua interpretação muito difícil. É possível realizar uma nova poda, que permite visualizar com clareza as decisões nos primeiros níveis da árvore, onde ocorrem as decisões mais significativas. Entretanto, essa nova poda representa uma grande redução de concordância (mensurada pelo coeficiente Kappa). A Figura A3.1 ilustra a árvore de decisão completa e a árvore de decisão podada (primeiros níveis da árvore de decisão completa) em Belo Horizonte.

FIGURA A3.1 – Árvores de Decisão Completa e Podada (Belo Horizonte)



Fonte: Elaboração própria, 2021.

Ao contrário dos modelos regressão logística e análise discriminante, que sempre incluem todas as variáveis explicativas no modelo, a árvore de decisão podada exibe apenas as variáveis com maior ganho de informação. Por exemplo, em Belo Horizonte, existe maior probabilidade de presença de núcleos urbanos informais quando ali residem famílias em domicílio com alvenaria sem revestimento (P10), existe alta densidade populacional (den_pop_km) e baixa porcentagem de domicílios com infraestrutura adequada (porcAdequa) ou alta porcentagem de domicílios sem arborização (porcDomSArb).

Dessa forma, destacamos que os modelos regressão logística e análise discriminante são mais inteligíveis, permitindo a identificação das variáveis significativamente associadas à presença dos núcleos urbanos informais, incluindo seus respectivos coeficientes. A fácil compreensão dos resultados desses modelos viabiliza uma melhor caracterização dos NUI de cada polo, bem como das diferenças observadas entre os polos.

A3.4 Calibração e avaliação dos modelos

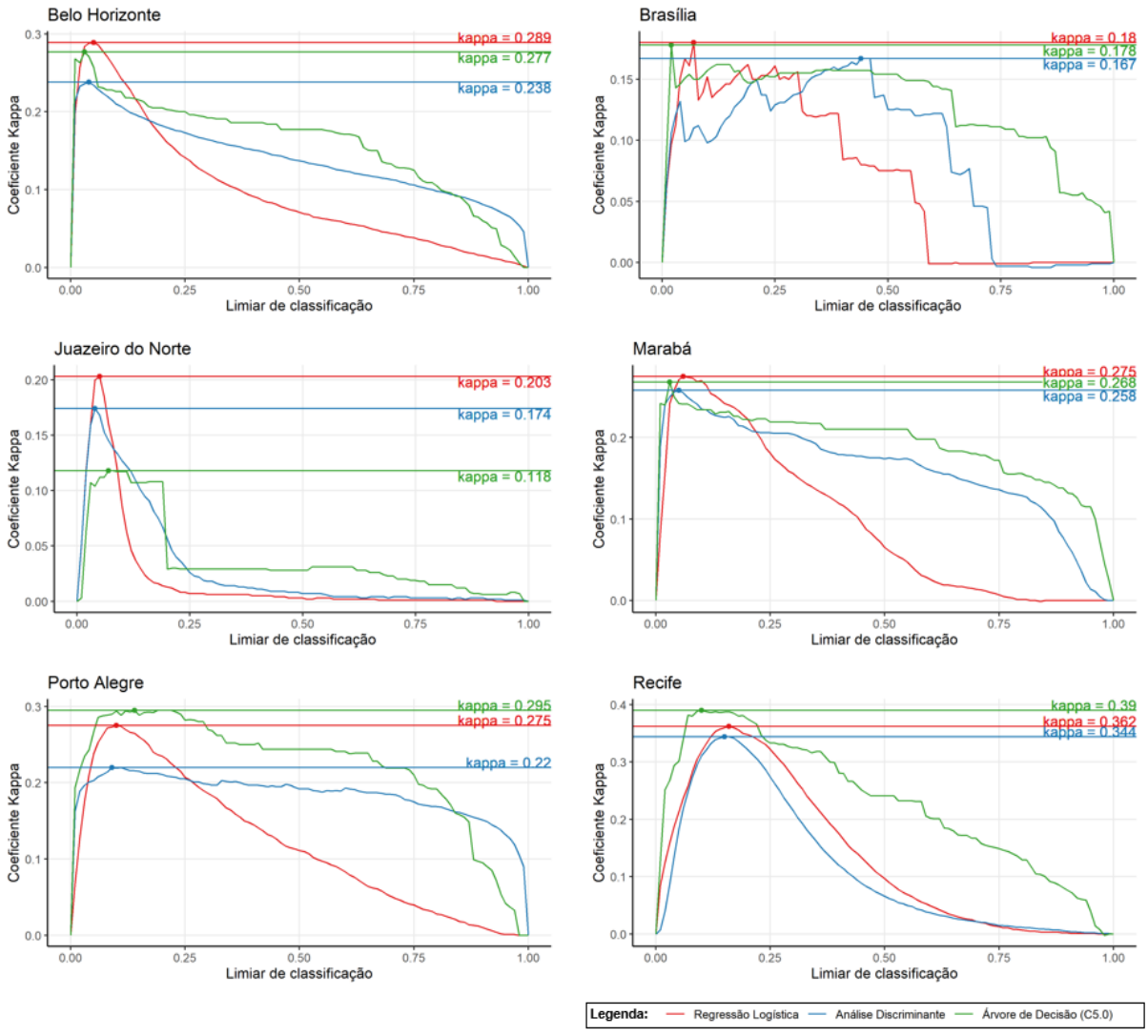
Após a construção dos modelos, iniciou-se um processo de calibração cujo objetivo era identificar, para cada modelo, o limiar de probabilidade que resultasse em uma classificação com a maior concordância possível com os dados de NUI levantados em campo pela Pesquisa. Esse processo consistiu no cálculo do coeficiente Kappa para cada limiar de probabilidade (de 0 a 1, a cada 0,01 unidade (probabilidade de 1%) – tradicionalmente, utiliza-se o limiar de 0,50.

Sabe-se que, em uma situação real, os dados do levantamento em campo não estariam disponíveis, tendo em vista que a Metodologia NUI almeja fornecer subsídios para a identificação de núcleos urbanos informais a partir de dados incompletos (AGSN). Dessa forma, esse resultado refere-se ao

que denominamos aqui como "kappa potencial", ou seja, a melhor concordância entre os resultados de modelos construídos a partir de uma base de dados de treino (AGSN), reconhecidamente incompleta, e a base de dados de avaliação (NUI), mais próxima da realidade.

A Figura A3.2 apresenta uma curva do nível de concordância (coeficiente Kappa) considerando cada limiar de classificação. Dessa forma, é possível visualizar qual é a maior concordância possível em cada modelo (reta horizontal).

FIGURA A3.2 – Coeficiente de concordância Kappa para cada limiar de classificação.



Fonte: Elaboração própria, 2021.

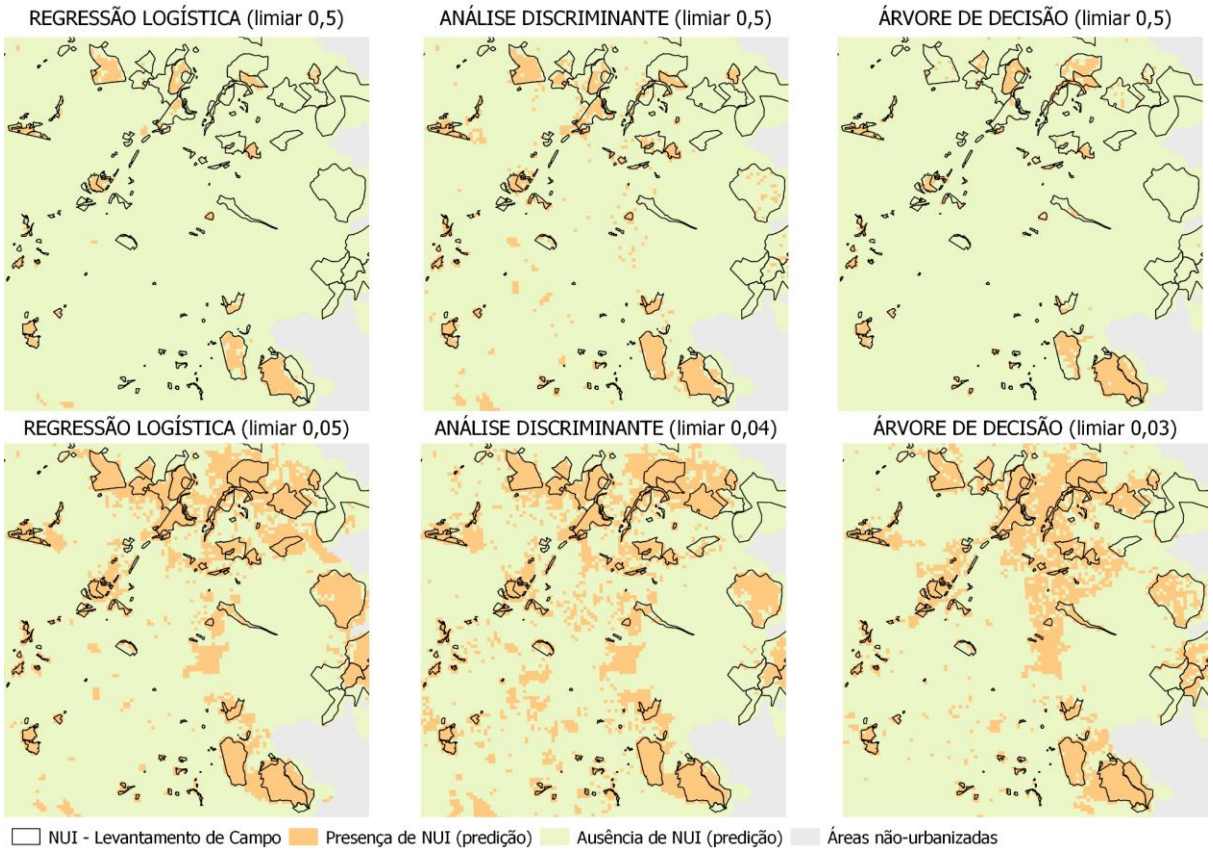
Nota-se pela Figura A3.2 que o melhor limiar de todos os modelos está deslocado à esquerda. Isso acontece devido à uma característica intrínseca à essa análise – a “pseudo-ausência”, que resulta na subestimativa da presença de núcleos urbanos informais quando o limiar padrão é adotado (0,50). Portanto, **para obter uma classificação com maior concordância utilizando os modelos analisados, é necessário realizar uma análise das superfícies de probabilidade de presença de NUI de forma a ajustar os limiares de classificação**. Os limiares da Figura A3.2 foram obtidos com a base de dados de validação (levantamento em campo), mas eles também podem ser encontrados a partir da base de

dados de treino (Aglomerados Subnormais) ou com o uso de outros recursos, como a interpretação visual das superfícies de probabilidade em conjunto com imagens orbitais. Nesses casos, atinge-se níveis de concordância melhores do que as do limiar padrão de 0,5, mas não tão altas quanto as apresentadas.

A Figura A3.2 também mostra que **a regressão logística sempre possui um nível de concordância potencial maior que a análise discriminante e maior ou similar quando comparado à árvore de decisão**. Deve-se considerar ainda que as superfícies de probabilidade da árvore de decisão não são consideradas estáveis, embora nenhum experimento tenha sido conduzido neste sentido.

Pelo mapa da classificação realizada pelos modelos (Figura A3.3), com o limiar padrão (0,5) e o limiar ajustado, vemos que enquanto os modelos com limiar de 0,5 tendem a subestimar a presença de NUI, os modelos com limiar ajustado tendem a superestimar a presença de NUI. Acreditamos que o segundo caso esteja mais alinhado com os objetivos da Metodologia NUI, ou seja, permitir que um especialista delimite os núcleos urbanos informais usando uma camada que aponta onde os NUI podem estar (superestimativa), ou ainda probabilidade de presença de NUI.

FIGURA A3.3 – Mapa da classificação dos modelos (Belo Horizonte)



Fonte: Elaboração própria, 2021.

Portanto, **o resultado da classificação (Presença de NUI ou Ausência de NUI) não explora todo o potencial desses modelos. Recomenda-se apresentar os resultados como superfície de probabilidade** e, existindo a necessidade de gerar uma resposta binária, o especialista pode chegar a um limiar de classificação por meio de duas estratégias: (a) utilizando os dados dos AGSN como referência para identificar o limiar de probabilidade associado a uma classificação com maior kappa;

ou (b) utilizando o conhecimento local associado à interpretação visual de imagens de alta resolução e levantamentos de campo.

A3.5 Conclusões sobre a Análise Comparativa

A partir da análise comparativa de técnicas apresentada, podem ser levantadas as seguintes considerações sobre o processo de identificação de núcleos urbanos informais:

1. **Interpretabilidade:** os resultados das análises de regressão logística e discriminante apresentam são mais inteligíveis do que os da árvore de decisão, sendo, portanto, mais úteis no processo de caracterização dos núcleos urbanos informais;

2. **Limiar de classificação:** independentemente da técnica considerada, o aprimoramento dos resultados da classificação de NUI depende da realização de procedimentos adicionais para a escolha de um limiar de classificação adequado. Esses procedimentos podem incluir análises quantitativas baseadas em medidas de concordância (como, por exemplo, o coeficiente kappa) ou qualitativas baseadas no conhecimento local e interpretação visual de imagens orbitais;

3. **Maior potencial de concordância:** considerando os dados dos polos da pesquisa, a regressão logística apresentou um nível de concordância potencial maior que do que o análise discriminante; e maior ou similar do que o da árvore de decisão;

4. **Superfície de probabilidade:** classificações binárias (Presença de NUI ou Ausência de NUI) não exploram todo o potencial dos modelos apresentados, sendo recomendada a análise de superfícies de probabilidade da presença de NUI, preferencialmente somadas a outras camadas de informação.

Os resultados apontam que a regressão logística é o modelo com maior potencial para identificar núcleos urbanos informais sob as condições analisadas. Para atingir os resultados esperados, faz-se necessária a disponibilização da superfície de probabilidade da presença de NUI que, mediante análise de um especialista complementada por outros planos de informação, pode ser utilizada na identificação de NUI.

Apesar das evidências a favor do uso da regressão logística para a identificação de núcleos urbanos informais neste estudo, a análise comparativa de técnicas pode ser aprofundada em diferentes direções. Se a facilidade de interpretação dos coeficientes e decisões não fosse uma das premissas da Metodologia NUI, seria possível explorar outras técnicas de *machine learning*, como a árvore de decisão com *boosting* ou *cross-validation* (validação cruzada), ou ainda o modelo *random forest* (floresta aleatória). Esse último, ao contrário da árvore de decisão, tende a gerar superfícies de probabilidade estáveis. Obter uma predição que seja territorialmente contígua também permanece um desafio, o que pode ser explorado ao incluir mais variáveis da forma urbana (e.g. métricas derivadas de sensoriamento remoto) ou através de modelos preditivos espaciais (e.g. *geographically weighted logistic model*; *spatial generalized mixed model*).