UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL

LAÍS FREITAS MOREIRA DOS SANTOS

ORIENTADORA: CONCEIÇÃO DE MARIA ALBUQUERQUE ALVES

DISSERTAÇÃO DE MESTRADO EM TECNOLOGIA AMBIENTAL E RECURSOS HÍDRICOS

Brasília – DF
2021

LAÍS FREITAS MOREIRA DOS SANTOS

PROCEDIMENTO PARA IDENTIFICAÇÃO DE ASSENTAMENTOS PRECÁRIOS COM USO DE INDICADORES DE ACESSO AOS SERVIÇOS DE ÁGUA E ESGOTO: uma aplicação no Distrito Federal - Brasil

Trabalho apresentado ao Programa de Pós-Graduação em Tecnologia Ambiental e Recursos Hídricos do Departamento de Engenharia Civil da Universidade de Brasília, como requisito para aprovação.

ORIENTADORA: CONCEIÇÃO DE MARIA ALBUQUERQUE ALVES

DISSERTAÇÃO DE MESTRADO EM TECNOLOGIA AMBIENTAL E RECURSOS HÍDRICOS

Brasília – DF

2021

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL

PROCEDIMENTO PARA IDENTIFICAÇÃO DE ASSENTAMENTOS PRECÁRIOS COM USO DE INDICADORES DE ACESSO AOS SERVIÇOS DE ÁGUA E ESGOTO: uma aplicação no Distrito Federal – Brasil

LAÍS FREITAS MOREIRA DOS SANTOS

DISSERTAÇÃO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA COMO PARTE DOS REQUISÍTOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM TECNOLOGIA AMBIENTAL E RECURSOS HÍDRICOS

APROVADA POR:

_____

Profª Conceição de Maria Alves de Albuquerque, PhD (UnB) (Orientadora)

_____

Prof. Carlos Henrique Ribeiro Lima, PhD (UnB) (Examinador Interno)

_____

Prof. Dr. Ricardo de Sousa Moretti (UFABC) (Examinador Externo)

BRASÍLIA/DF, 28 DE JUNHO DE 2021

FICHA CATALOGRÁFICA

DOS SANTOS, LAÍS FREITAS MOREIRA.
PROCEDIMENTO PARA IDENTIFICAÇÃO DE ASSENTAMENTOS PRECÁRIOS COM USO DE INDICADORES DE ACESSO AOS SERVIÇOS DE ÁGUA E ESGOTO: uma aplicação no Distrito Federal - Brasil [Distrito Federal] 2021.

131p., 210 x 297 mm (ENC/FT/UnB, Mestre, Tecnologia Ambiental e Recursos Hídricos, 2021). Dissertação de Mestrado – Universidade de Brasília. Faculdade de Tecnologia. Departamento de Engenharia Civil e Ambiental.

1. Assentamentos Precários
2. Abastecimento de Água
3. Esgotamento Sanitário
4. Modelagem Estatística
I. ENC/FT/UnB
II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

DOS SANTOS, L. F. M. (2021). Procedimento para Identificação de Assentamentos Precários com Uso de Indicadores de Acesso aos Serviços de Água e Esgoto: uma aplicação no Distrito Federal – Brasil. Dissertação de Mestrado em Tecnologia Ambiental e Recursos Hídricos, Publicação PTARH.DM 235/2021, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 131p.

CESSÃO DE DIREITOS

AUTOR: Laís Freitas Moreira dos Santos

TÍTULO: Procedimento para Identificação de Assentamentos Precários com Uso de Indicadores de Acesso aos Serviços de Água e Esgoto: uma aplicação no Distrito Federal – Brasil.

GRAU: Mestre ANO: 2021

É concedida à Universidade de Brasília permissão para reproduzir cópias desta dissertação de mestrado e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte dessa dissertação de mestrado pode ser reproduzida sem autorização por escrito do autor.

_____

Laís Freitas Moreira dos Santos
laismoreira.eng@gmail.com

DEDICATÓRIA

Ao meu pai, Delor,
grande esteio de conhecimento e ternura.

AGRADECIMENTOS

Ao Universo, por ser tão generoso comigo; às bênçãos e a todos os ensinamentos espirituais que atravessaram meu caminho, não por acaso, durante essa jornada.

À minha mãe, por todo o apoio incondicional de sempre que, mesmo de longe, fez toda a diferença em cada passo dado.

Ao meu pai, por toda a estrutura, acolhimento e paciência durante todo o processo de elaboração da dissertação. Pelo afeto genuíno, pelo estímulo e pela celebração a cada pequena vitória.

Às minhas irmãs, Vidya, Thais, Laura, Giselle e Amanda, base na qual me apoiei em todos os momentos de dificuldade, e aos meus sobrinhos (David, Raul, Daniel e Gael) e sobrinhas (Lys, Odara e Amora), a quem dedico meu eterno amor e de onde tiro forças para continuar.

À toda a família Freitas, responsável pela minha formação como pessoa e que guia meus valores no mundo.

À toda a família Moreira, de onde vem minha força emocional para lidar com os desafios da vida.

À Cilene, ao Cláudio e ao Gustavinho, por serem a família que Deus me presenteou de quem não abro mão da presença na celebração de minhas vitórias.

Ao Pedro, meu companheiro de jornada, com quem me permito sonhar e construo os caminhos para realizar.

À minha orientadora, professora Conceição, pelo direcionamento, pela confiança, por acreditar em mim e no tema que propus, e por suas ricas contribuições.

A todos os professores do PTARH, que me proporcionaram o acesso a conhecimentos que me acompanharão por toda a vida.

À toda equipe administrativa da Universidade de Brasília, onde fui bem acolhida e recebida.

Ao Centro de Estudos da Metrópole, nas figuras de Edgard Fusaro e Daniel Silva, por serem extremamente receptivos e me repassarem conhecimentos fundamentais à construção dessa pesquisa.

À FAP DF, pelo financiamento da partilha de minha pesquisa no exterior, cuja experiencia foi engrandecedora.

A todos os colegas de mestrado, pelo cotidiano compartilhado, em especial à Letícia Brito, pelo apoio em todos os momentos importantes.

Ao Yannick, grande amigo que fiz ao longo desse processo e que compartilha de minhas inquietações sociais e acadêmicas.

A todos os amigos de graduação e de vida. À Louise e Tamires, pela companhia cotidiana e afeto de sempre.

À Daniele Lameira, que acompanhou o processo de minha mudança para Brasilia desde a inscrição no mestrado até o momento presente, sendo essencial para tornar mais amena a vivência da pós-graduação, especialmente durante a pandemia.

Ao Márcio, que, pacientemente me ensinou a manipular o software R e me apoiou em toda a aventura autodidata que tive que encarar durante o processo da pesquisa.

Aos professores e mestres que passaram pela minha vida e a todos os pesquisadores que me inspiram e me fazem continuar a sonhar em impactar o mundo positivamente através da ciência.

A vocês, minha eterna gratidão.

Epígrafe

“Nenhuma solução será individual”.

Emicida

RESUMO

O acesso universal aos serviços de Abastecimento de Água e Esgotamento Sanitário (SAAES) é um direito humano reconhecido internacionalmente pela Organização das Nações Unidas (ONU), sendo o foco do sexto Objetivo de Desenvolvimento Sustentável (ODS). No entanto, a desigualdade de acesso a esses serviços nas áreas urbanas impacta diretamente as populações mais vulneráveis, residentes de Assentamentos Precários. Nesse sentido, o presente estudo teve como objetivo propor um procedimento metodológico para identificação desses assentamentos com base em dados do Censo Demográfico, partindo do método desenvolvido por Marques et al. (2007), no Centro de Estudos das Metrópoles (CEM). O procedimento foi aplicado para o Distrito Federal (DF), testando três técnicas estatísticas multivariadas: Análise Discriminante Linear (ADL), Análise Discriminante Quadrática (ADQ) e Regressão Logística (RL). O melhor desempenho preditivo foi encontrado no modelo de RL, que apresentou sensitividade de 88% e *Area Under the Curve* de de 0,96, de modo que seus resultados foram escolhidos como base para o cálculo dos índices de acesso aos SAAES no DF. Tais índices foram calculados para os Assentamentos Precários e para o DF como um todo, cujos resultados corroboraram com a hipótese levantada por Marques et al. (2007), Juliano et al. (2012) e Guimarães (2015), de que os índices de acesso, quando medidos para o conjunto da população, tendem a mascarar a realidade encontrada entre as populações mais vulneráveis.

Palavras-chave: Assentamentos Precários; Abastecimento de Água; Esgotamento Sanitário; Modelagem Estatística; Técnicas de Estatística Multivariada.

ABSTRACT

Universal access to water supply and sanitation services (WaSa) is a human right recognized internationally by the United Nations (UN), being the focus of the sixth Sustainable Development Goal (SDG). However, the inequality of access to these services in urban areas directly impacts the most vulnerable populations, residents of Precarious Urban Settlements. Thus, the present study aimed to propose a methodological procedure for identifying these settlements with data from the Demographic Census, based on the method developed by Marques et al. (2007), at the Center for Metropolis Studies (CEM). The procedure was applied to the Federal District of Brazil (FDB), testing three multivariate statistical techniques: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Logistic Regression (LR). The best predictive performance was found in the LR model, which presented a sensitivity of 88% and an Area Under the Curve of 0.96, so that its results were chosen as the basis for calculating the access rates to SAAES in the FDB. Such indexes were calculated for the Precarious Urban Settlements and also for the DF as a whole, whose results corroborated the hypothesis raised by Marques et al. (2007), Juliano et al. (2012) and Guimarães (2015), that WaSa access indexes, when measured for the population as a whole, tend to mask the reality found among the most vulnerable populations.

Keywords: Precarious Urban Settlements; Water supply; Sanitary Sewage; Statistical Modeling; Multivariate Statistical Techniques.

LISTA DE ABREVIATURAS, SIGLAS, SÍMBOLOS E UNIDADES

AD - Análise Discriminante

ADL - Análise Discriminante Linear

ADQ - Análise Discriminante Quadrática

AGSN - Aglomerados Subnormais

AIC - *Akaike Information Criteria*

AP - Assentamentos Precários

AUC - *Area Under the Curve*

CEBRAP - Centro Brasileiro de Pesquisa e Planejamento

CEM - Centro de Estudos das Metrópoles

DF – Distrito Federal

DHAES - Direito Humano à Água e ao Esgotamento Sanitário

DHARMA - *Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*

DPP – Domicílios Particulares Permanentes

ER - Regressão Logística em casos de Eventos Raros

FINEP - Financiadora de Estudos e Projetos

IBGE – Instituto Brasileiro de Geografia e Estatística

IPEA - Instituto de Pesquisa Econômica Aplicada

K-L - Divergência de Kullback-Leibler

LDA – *Linear Discriminant Analysis*

LIT - Levantamento de Informações Territoriais

LR – *Logit Regression*

MASS - *Modern Applied Statistics with S*

MVN - *Multivariate Normality Tests*

NA's - Not available

NE - Tipo de Setor - Não Especiais

ODM - Objetivos de Desenvolvimento do Milênio

ODS - Objetivos de Desenvolvimento Sustentável

Opas - Organização Pan-americana de Saúde

OR - *Odds ratio*

ONU – Organização das Nações Unidas

PNAD - Pesquisa Nacional por Amostra de Domicílios

PR – Percentual de Pessoas Responsáveis

QDA – *Quadratic Discriminant Analysis*

REGIC - Regiões de Influência das Cidades

RL – Regressão Logística

RM - Região Metropolitana

ROC - *Receiver Operating Characteristic Curve*

SAAES – Serviços de Água e Esgoto

SbN - Tipo de Setor – SubNormais

SERFHA - Serviço Especial de Recuperação das Favelas e Habitações Anti-Higiênicas

SNIS - Sistema Nacional de Informações sobre Saneamento Básico

SPSS - *Statistical Package for the Social Sciences*

UI - Universalização Inclusiva

USP - Universidade de São Paulo

VIF - *Variance Inflation Factor*

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

De acordo com dados do Censo Demográfico de 2010, mais de 160 milhões de brasileiros vivem nas cidades – cerca de 84% da população –, onde faltam moradias para 17 milhões de famílias, sendo de 5 milhões o déficit quantitativo (falta absoluta de moradia) e de 12 milhões o déficit qualitativo (habitações inadequadas) (Carrion, 2013; IBGE, 2011). Esse cenário representa o resultado de um crescimento desordenado, baseado no alto índice de migração campo-cidade, pela alta demanda de mão-de-obra urbana na segunda metade do século XX. Isso ocasionou um boom populacional nas cidades, agravando as questões de exclusão socioespacial, uma vez que a população que não tinha acesso ao mercado imobiliário formal acabou por ocupar áreas periféricas, de maneira irregular.

Assim, foram formados os denominados Assentamentos Precários, porções do território urbano com dimensões e tipologias variadas (cortiços, favelas, loteamentos), os quais apresentam irregularidade de posse da propriedade e/ou carência no acesso a equipamentos urbanos e serviços públicos básicos – como os serviços de água e esgoto (SAAES), que compõem o saneamento básico (Morais *et al*., 2016). Desse modo, a falta de políticas habitacionais e intervenções governamentais que garantissem o acesso à moradia e SAAES a essa parcela da população afetou diretamente as questões urbanas e de saúde pública, pela falta de acesso a água tratada e esgotamento sanitário.

De acordo com o Sistema Nacional de Informações sobre Saneamento Básico (SNIS), o índice de universalização urbana no Brasil até 2016 era de 93% para abastecimento de água e 59,7% para esgotamento sanitário (SNIS, 2016). Esses dados são obtidos por meio de entrevistas junto às prefeituras e, na maioria dos casos, solicitados à concessionária ou prestadora de serviços. Nesse contexto, Guimarães (2014) questiona os índices de atendimento existentes, afirmando que não contabilizam os Assentamentos Precários (AP), pelo fato de comumente apresentarem situação de irregularidade da propriedade, fazendo com que a companhia prestadora do serviço não tenha a obrigatoriedade de atendê-los, por não estarem em consonância com a lei – ainda que esteja consolidado o direito humano à água e ao esgotamento sanitário nacional e internacionalmente –, os quais tornam-se "invisíveis aos olhos do gestor".

Ou seja, de acordo com a autora, os indicadores não revelariam a realidade de acesso, na medida em que não incluiriam no seu universo total de domicílios a serem atendidos aqueles situados em Assentamentos Precários irregulares. Nessa linha, Juliano

et al. (2012) afirma que são diversos os dados disponíveis acerca da cobertura de serviços de abastecimento de água e de esgotamento sanitário em áreas legais, porém faltam indicadores que meçam esses serviços em Assentamentos Precários e ilustrem as disparidades nas questões da universalização do acesso aos SAAES. Isso porque as estatísticas de acesso à água e saneamento costumam ser estaduais, regionais, nacionais ou globais, isto é, os dados disponíveis se referem a grandes áreas, de modo que acabam por suavizar ou até mesmo mascarar a real situação real dos impactos sofridos pela população mais pobre e moradora de assentamentos.

Este cenário descrito por Guimarães (2015) e Juliano et al. (2014) leva à necessidade do mapeamento de Assentamentos Precários e o acesso de sua população aos SAAES, uma vez que o atendimento das populações mais vulneráveis – que, nos espaços urbanos, estão largamente concentradas em Assentamentos Precários – pode ser entendida como um requisito rumo à universalização. Nesse sentido, na literatura, observa-se uma necessidade de ampliar estudos que abordem a questão do acesso aos SAAES especificamente em Assentamentos Precários, os quais são provavelmente os mais afetados pela falta de acesso a esses serviços (UN-Habitat, 2003; Heller, 2009), especialmente em grandes cidades de países em desenvolvimento, onde as desigualdades em serviços sanitários são maiores entre ricos e pobres (HAWKINS et al., 2013; OMS/UNICEF, 2019).

Embora o acesso ao saneamento e sua relação com o status socioeconômico venha sendo amplamente explorado desde que a ONU reconheceu o direito humano à água e ao saneamento, em 2010, pouco se sabe sobre como isso é enfrentado nos assentamentos urbanos (Adams, 2018). Tal lacuna ocorre principalmente pela falta de padronização da identificação e mapeamento dos Assentamentos Precários, uma vez que, não havendo um consenso sobre o conceito e a caracterização dessas áreas, não há um *proxy* de detecção e tampouco um procedimento metodológico unificado e confiável que permita o estudo desse fenômeno de forma mais acurada, com a geração de dados que possam ser comparados entre si. Assim, a investigação do acesso aos SAAES em Assentamentos Precários perpassa a necessidade de, primeiramente, conceituá-los e mapeá-los.

No entanto, isso representa um desafio, uma vez que há uma grande variedade de definições para os Assentamentos Precários, pelo fato de que cada região ou país possui uma particularidade em relação às características estruturais e socioeconômicas da população que os ocupa, prejudicando o monitoramento da dimensão e evolução

desse fenômeno em todo o mundo (Patel et al., 2014). Assim, a falta de uma definição consolidada, indicadores apropriados e dados precisos, relevantes e representativos sobre os Assentamentos Precários no Brasil representam uma barreira informacional ao acesso de suas populações aos SAAES. Nesse sentido, um método para a identificação de Assentamentos Precários pode ser determinante no aprimoramento do conhecimento que os tomadores de decisão devem ter para agir no atendimento às necessidades e direitos dos residentes (Sinharoy et al. 2019; Queiroz Filho, 2015; Jones et al.; 2012). Santos & Gupta (2017) também incluíram a disponibilidade de dados como um desafio para o acesso aos SAAES em países de baixa renda, especialmente para a população mais vulnerável socioeconomicamente. Entende-se, portanto, que obter informações atualizadas sobre o acesso a esses serviços em Assentamentos Precários pode ser considerado um primeiro passo para o atendimento do direito humano a água e esgoto.

No caso do Brasil, em 2007, foi desenvolvido um método de detecção de Assentamentos Precários pelo Centro de Estudos das Metrópoles (CEM), realizado em parceria com o Ministério das Cidades, tendo como base o mapeamento de Aglomerados Subnomais realizado pelo IBGE no Censo Demográfico de 2000 (Marques et al., 2007). Dessa forma, pretende-se estudar o método desenvolvido pelo CEM e, com base nele, propor um procedimento metodológico para detecção de Assentamentos Precários a ser desenvolvido no software R Project e aplicado para o Distrito Federal, com base nos dados do Censo Demográfico de 2010. Os resultados da detecção permitirão, por fim, calcular o índice de Universalização Inclusiva, ou seja, o percentual de acesso aos SAAES dentro de Assentamentos Precários.

O desenvolvimento do método tem como objetivo, portanto, detectar setores censitários representativos de Assentamentos Precários com base em uma classificação binária pré-definida pelo IBGE, de setores Não Especiais (NE) e SubNormais (SbN), designados por 0 e 1, respectivamente; identificando os setores com perfil socioeconômico semelhante aos SbN mas que foram originalmente classificados como NE. Dessa forma, o método amplia a classificação realizada pelo IBGE, incluindo fatores socioeconômicos como base para a classificação, além dos estruturais utilizados no Censo Demográfico.

Para tal, serão testadas três técnicas estatísticas – sendo a primeira delas a mesma que foi utilizada pelo CEM em 2007: Análise Discriminante Linear (ADL), Análise Discriminante Quadrática (ADQ) e Regressão Logística (RL), para detectar os Setores Censitários pertencentes a Assentamentos Precários com base em um conjunto de

Indicadores sociais, demográficos e de infraestrutura calculados utilizando o Censo de 2010. Com isso, pretende-se testar qual técnica fornecerá resultados mais bem ajustados e, a partir isso, propor um procedimento metodológico baseado e um modelo estatístico com maior nível de acerto. Assim, será proposto um método ajustado e confiável que permita averiguar a situação de atendimento dos SAAES dentro dos Assentamentos Precários, de forma a buscar alternativas para solucionar a lacuna de mapeamento dos Assentamentos Precários no Brasil, cuja situação é alarmante, tanto no contexto da precariedade habitacional e quanto da garantia dos SAAES enquanto um direito humano.

2 OBJETIVOS

2.1 Objetivo Geral

- Avaliar o alcance da universalização do acesso aos serviços de Abastecimento de Água e Esgotamento Sanitário em Assentamentos Precários do Distrito Federal classificados conforme procedimento metodológico desenvolvido com base no método do Centro de Estudos da Metrópole (Marques et al., 2007).

2.2 Objetivos Específicos

- Propor procedimento para identificação de Assentamentos Precários com base no método de Marques (2007) do Centro de Estudos das Metrópoles utilizando a classificação do IBGE/Censo (2010) para Tipo de Setor – SubNormais (SbN) ou Não Especiais (NE) – por meio de modelagem estatística;
- Identificar Assentamentos Precários para os Setores Censitários Urbanos do Distrito Federal com base na nova metodologia proposta;
- Identificar e avaliar níveis de universalização do acesso aos serviços de Abastecimento de água e Esgotamento sanitário dos Assentamentos Precários do Distrito Federal obtidos pela metodologia proposta.

3 REVISÃO BIBLIOGRÁFICA

3.1 Urbanização Brasileira e o Surgimento dos Assentamentos Precários

Para conhecer a realidade dos assentamentos precários no Brasil, é fundamental que se entenda o contexto do seu surgimento e sua evolução ao longo das últimas décadas, juntamente aos desdobramentos das demandas habitacionais. Assim, é necessária a compreensão de que o déficit e a precariedade habitacional que permeiam a realidade urbana brasileira são resultados do processo de formação da sociedade.

O primeiro movimento migratório campo-cidade ocorreu no contexto da legalização da propriedade privada (1850) e abolição da escravidão (1888), onde os trabalhadores rurais partiram em busca de oportunidades de trabalho assalariado, formando assim a classe do proletariado urbano (Ministério das Cidades, 2010). No entanto, poucas eram as possibilidades de moradia para essa parcela da população, que recebia baixíssimos salários e não tinha acesso ao mercado imobiliário estabelecido nas cidades. Esse cenário acabou gerando um processo de segregação, marcado pelo descaso com a classe do proletariado, ocasionando a formação dos primeiros cortiços e representando o início da problemática questão habitacional urbana (Ferreira, 2010).

Os cortiços tradicionais eram moradias alugadas estruturadas como habitação coletiva, em edificações antigas e deterioradas, com a higiene sanitária de uso comum e, muitas vezes até alugados informalmente, comumente constituídos pela população mais pobre proveniente do campo. Eram entendidos como focos de transmissão de doenças, já que normalmente apresentavam alto adensamento e insalubridade, além de representarem potenciais obstáculos à expansão das áreas mais nobres da cidade. Por essas razões, muitas vezes eram demolidos e a "massa sobrante" (Villaça, 2001), realocada para regiões menos valorizadas pelo mercado, marcando as primeiras ações de remoção maciça de moradias populares das áreas centrais das cidades (Ministério das Cidades, 2010).

Dentro desse contexto de falta de acessibilidade ao mercado imobiliário formal por parte do proletariado, outras formas de ocupação começaram a surgir e, em 1920, foi popularizado o termo 'favela' no Rio de Janeiro, para designar as novas ocupações em situação de precariedade, que apresentavam conflitos ou irregularidades na posse da terra. As favelas ficaram conhecidas com os mais diversos nomes ao longo do país – como mocambos, ocupações, invasões, baixadas e palafitas – e foram produtos de

ocupação espontânea irregular em terras vazias centrais ou periféricas, públicas ou privadas, sem título de propriedade e normalmente com padrões urbanísticos deficitários, sem acesso aos serviços públicos básicos (Ministério das Cidades, 2010; Cardoso, 2016).

Nesse período, acreditava-se que, na economia de livre mercado, a questão da habitação não era responsabilidade do Poder Público e sim algo que se resolveria no setor privado. Essa lógica só começou a se modificar em 1930, quando diversos setores da sociedade passaram a cobrar uma postura do governo em relação à essas ocupações, levando a intervenções pontuais de remoção – erradicação –. No entanto, nas décadas posteriores a crise habitacional só se agravava, enquanto o modelo da "casa própria" ganhava força (Ministério das Cidades, 2010).

Esse modelo intensificou a precariedade habitacional, com novas ocupações em locais que não incorporavam nenhuma infraestrutura, dando início ao parcelamento do solo e venda de loteamentos clandestinos ou irregulares nas periferias, a preços muito baixos. Os loteamentos irregulares eram aqueles que não cumpriam integralmente as normas urbanísticas – cujo em muitos casos já havia sido feito o pedido de licença –, enquanto os clandestinos se referiam a parcelamentos efetuados sem qualquer tipo de licenciamento, não possuindo qualquer registro oficial pelo poder municipal.

Esse cenário não era previsto na legislação urbanística da época, sendo completamente negligenciado pelo Poder Público, o qual além de não se mobilizar para solucionar o problema da habitação, contribuía para seu agravamento barateando custos industriais pela retirada da moradia do custo da mão de obra, para favorecer o processo de industrialização (Bonduki, 1998).

A década de 1950 marcou o início de muitas transformações, que incluíram o fortalecimento da indústria, a construção de Brasília e o grande movimento de migração inter-regional. Até a década de 1980, foram gerados 25 milhões de ocupações no contexto da urbanização motivada pelas demandas de mão de obra, sendo que a maioria dos empregos gerados eram de caráter terciário, informal, e sem proteção social (Morais et al., 2016).

Nesse mesmo período, por conta dessas grandes mudanças, surgiram as primeiras propostas de melhoria dessas áreas ocupadas, com a implantação de infraestrutura e construção de moradias juntamente à população local, por iniciativa da Igreja Católica (Ministério das Cidades, 2010). O novo modelo proporia a urbanização de favelas, contrariando a postura anterior do governo, que agia no sentido ou de

erradicar essas áreas, ou de simplesmente tolerá-las, sem encarar o problema que representavam.

O primeiro programa que institucionalizou essa nova proposta foi o Serviço Especial de Recuperação das Favelas e Habitações Anti-Higiênicas (SERFHA), pelo governo do Distrito Federal, o qual não obteve resultados muito significativos. Os primeiros censos de favelas datam desse período e revelam números alarmantes em muitas cidades brasileiras: no Rio de Janeiro (1950), 58 favelas com 169,3 mil moradores; em Porto Alegre (1951), 56 favelas com 54,1 mil moradores; em Belo Horizonte (1955), 9,3 mil barracos com 36,4 mil moradores; em São Paulo (1957), 141 favelas e 8,4 mil barracos com 50 mil moradores (FINEP-GAP, 1985 apud Ministério das Cidades, 2010).

Na década de 1960, já na ditadura militar, surgiu a primeira iniciativa de criação de uma política nacional sobre a questão habitacional, mobilizando recursos para a habitação e construindo uma quantidade significativa de moradias – ainda que insuficientes à demanda estabelecida. No entanto, as construções foram feitas em áreas distantes, sem infraestrutura urbana e sem subsídios orçamentários, baseando-se em um sistema de financiamento bancário. Assim, surgiram entraves semelhantes aos do mercado imobiliário privado: as classes de menor renda ainda não conseguiam realizar os financiamentos e continuavam expandindo as ocupações nas favelas e periferias (Denaldi, 2010; Ministério das Cidades, 2010).

Nesse sentido, a produção de habitações sociais implementada na década de 1960 não foi tão eficaz na redução da expansão de loteamentos clandestinos e irregulares nas periferias, pelo contrário, o déficit ficou ainda maior, devido à grande explosão urbana marcante dessa década. Os grandes polos industriais tornaram as cidades extremamente atrativas pelas novas oportunidades de emprego, gerando um aumento populacional urbano sem precedentes, fazendo com que o Brasil migrasse de um país essencialmente rural para urbano em 1965 (Denaldi, 2010).

Durante esse período, o setor habitacional, junto a outros setores ligados às infraestruturas urbanas, foi negligenciado, propiciando a emergência do fenômeno que foi descrito por Piquet (1983 apud Morais et al., 2016) como “urbanização descapitalizada”.

Além disso, o crescimento econômico resultante da industrialização estava inserido no contexto do capitalismo internacional, pautado na desvalorização da mão de obra, com oferta de baixos salários, num modelo fundamentalmente concentrador de

renda. Enquanto isso, a política habitacional do regime militar não atingia a população de baixa renda com até três salários mínimos – justamente a que mais crescia –, por apresentar um recorte privatista, de favorecimento às grandes indústrias, com o objetivo de fortalecer o fenômeno chamado "milagre econômico" (Denaldi, 2010; Ministério das Cidades).

Esse processo intensificou a tendência de segregação sócioespacial, tornando a situação da precariedade e irregularidade habitacional cada vez mais preocupante, de modo que, em 1940, a população urbana no Brasil representava somente 26,34%, enquanto em 1980 ela passou para 68,86%, alcançando 81,20% em 2000. Em dez anos, de 1970 a 1980, as cidades com mais de um milhão de habitantes dobraram, passando de cinco para dez (Maricato, 1996 apud Ministério das Cidades, 2010).

Diante desses dados e da crescente precariedade habitacional, no final da década de 1970, foi criada a Lei Federal 6.766/79 de parcelamento do solo urbano, a qual criminalizou os loteamentos clandestinos e estabeleceu parâmetros urbanísticos para aprovação dos loteamentos futuros, tornando o processo de regularização desses empreendimentos mais rigoroso. Essa lei representou um grande um marco na atuação do Poder Público em relação à irregularidade habitacional, modificando a política de erradicação ou negligência para finalmente se propor a encarar a problemática e regulamentá-la.

Já que a maioria dos loteamentos existentes não estavam em conformidade com a nova lei, tornou-se necessário regularizá-los para atender aos novos requisitos estabelecidos, os quais abrangiam desde a manutenção das vegetações ciliares nos corpos hídricos urbanos até a implantação de projetos completos de infraestrutura de abastecimento de água, esgotamento sanitário e drenagem. A criminalização de loteamentos clandestinos foi positiva no sentido de impedir a formação de novas ocupações, no entanto, reduziu a oferta de novas áreas disponíveis à ocupação irregular, o que acabou intensificando a ocupação de loteamentos já existentes e, em alguns casos, agravando a precariedade habitacional (Ministério das Cidades, 2010; Morais et al., 2016)

A partir de então, na década de 1980, surgiram novas mobilizações populares de lutas por moradia e direito à cidade, culminando no Movimento Nacional de Reforma Urbana. Esse projeto, inserido no contexto do processo constituinte, apresentava uma proposta de Emenda Popular à Constituição, exigindo os direitos de todos os cidadãos à moradia e aos serviços e equipamentos urbanos, por meio da criação de instrumentos

jurídicos que os garantissem. A nova Constituição foi promulgada em 1988, mas a resposta a essas demandas só se efetivou no ano 2000, com a inclusão dos artigos 182 e 183 na Constituição e a aprovação do Estatuto da Cidade, em 2001 (Ministério das Cidades, 2010).

Durante essa lacuna entre as mobilizações populares e a efetivação do direito à cidade na Constituição, os problemas de moradia agravaram-se, pelo empobrecimento da população e a falta de recursos orçamentários federais, que marcaram as décadas de 1980 e 1990. No entanto, já embasados pela legislação, os programas de urbanização de assentamentos precários tornaram-se mais elaborados, apresentando um trabalho em conjunto com a população e aproveitando os investimentos já feitos pelos moradores na autoconstrução de suas casas.

Além disso, possuíam novas diretrizes, direcionadas para a regularização fundiária, focando mais na reurbanização das áreas ocupadas e no direito de permanência da população que ali habitava e menos na substituição absoluta das unidades habitacionais precárias, adotando instrumentos de monitoramento e avaliação baseados em critérios ambientais.

Nos anos 2000, foram grandes os avanços em termos de legislação. Além do Estatuto da Cidade, em 2001, foi aprovada a Medida Provisória nº 2.220/01, que definiu avanços significativos para a urbanização e regularização fundiária das ocupações irregulares. Em 2003, foram criados o Ministério e o Conselho das Cidades e, posteriormente, o Fundo Nacional de Habitação de Interesse Social, por meio da Lei nº 11.124/05. Com tudo isso, foi finalizado o arcabouço legal e institucional necessário à implementação da nova Política Nacional de Habitação, aprovada em 2004 e finalizada em 2008 (Brasil, 2001a; 2001b; 2005).

A Política foi institucionalizada e passou a incluir a necessidade de subsídios públicos para viabilizar a moradia urbana da população de baixa renda, bem como a prioridade para a integração urbana dos assentamentos precários, mediante o desenvolvimento de programas articulados entre os três níveis de governo e participação da sociedade civil. Desde então, após mais de um século de urbanização acelerada e marcada pelos Assentamentos Precários como solução predominante de moradia popular, projetos e programas vêm sendo desenvolvidos e implementados no sentido de identificar e mapear os Assentamentos Precários, finalmente compreendendo que "conhecer o universo dos assentamentos precários, para planejar e executar essa política, tornou-se uma necessidade inadiável" (Ministério das Cidades, 2010, p. 15).

3.2 Assentamentos Precários e Saneamento: Contexto Nacional

Uma vez compreendido o processo de formação e de espraiamento dos Assentamentos Precários, especialmente nas metrópoles brasileiras, há a necessidade de apontar suas variações e distintas particularidades. Há três principais tipos de assentamentos precários, de acordo com a literatura: os cortiços, os loteamentos irregulares ou clandestinos e as favelas, representadas pelo IBGE como Aglomerados Subnormais (AGSN).

Para maior clareza da diferença dessas três tipologias, o ministério das cidades (2010) definiu as principais variáveis a serem consideradas para suas classificações, mostradas na Figura 1:

Favelas, Mocambos, palafitas e assemelhados	Ocupação de terreno de propriedade alheia
	A maioria das unidades habitacionais não possui título de propriedade
	Vias de circulação estreitas e de alinhamento irregular
	Lotes de tamanho e forma desiguais
	Ocupação densa de unidades habitacionais
	Precariedade de serviços públicos essenciais
Cortiços, casas de cômodo ou cabeça de porco	Unidade de moradia de várias famílias
	Uso comum de instalações hidráulicas e sanitárias
	Nas unidades habitacionais o mesmo cômodo tem várias funções
	Construção em lotes urbanos
	Subdivisão de habitações em uma mesma edificação
	Unidades habitacionais geralmente alugadas, subalugadas ou cedidas sem contrato formal
Loteamentos irregulares	Sem aprovação prévia do poder público municipal
	Descumprimento de normas legais urbanísticas e/ou ambientais
	Falta de titulação correta da terra
	Falta de correspondência entre o projeto apresentado e o executado

Figura 1 - Critérios para a classificação de Assentamentos Precários. Fonte: Ministério das Cidades, 2010.

Não existem dados nacionais e comparáveis para assentamentos precários no Brasil, no entanto, há a detecção dos AGSN realizada pelo IBGE, pela qual pode se ter uma noção da situação brasileira a respeito dessas ocupações. No último Censo Demográfico, em 2010, foram identificados um total de cerca de 3,22 milhões de domicílios particulares ocupados em setores subnormais, o que corresponde a cerca de 5,6% dos domicílios do país, situados em 330 municípios brasileiros, correspondendo a 6% da sua população (Nadalin et al., 2014).

Assim, dos 160 milhões de brasileiros, 11,4 milhões viviam em AGSN em 2010, sendo que 6 milhões deles estavam situados em metrópoles. Nadalin et al. (2014) realizou um estudo da distribuição dos AGSN ao longo do país utilizando as tipologias de Regiões de Influência das Cidades (REGIC) (IBGE, 2008), que divide as cidades nas categorias: núcleo da metrópole, região de abrangência da metrópole, capital regional, região de abrangência da capital regional, centros sub-regionais, centros de zona e centros locais. E os dados mostraram que, somando-se a população total de AGSN em 2010, 74,83% estava concentrada nas metrópoles e suas regiões de abrangência, enquanto essas duas tipologias abarcavam apenas 38,5% da população total brasileira (Figura 2). Portanto, tem-se que os AGSN representam eventos essencialmente urbanos e, indo mais além, metropolitanos.

Tipologia REGIC	Municípios com AS			Todos os municípios	
	População em AS em 2010	Total %	Proporção da população em AS em 2010 %	População em 2010	**Total**
Metrópole	6.158.778	53,87	16,86	36.534.266	**22,68**
Metrópole – abrangência	2.396.649	20,96	11,32	25.482.212	**15,82**
Capitais regionais	1.627.380	14,24	8,49	26.162.382	**16,24**
Capital regional – abrangência	624.952	5,47	11,22	8.831.224	**5,48**
Centros sub-regionais	340.842	2,98	10,61	14.300.704	**8,88**
Centros de zona	106.508	0,93	5,58	16.638.878	**10,33**
Centros locais	177.054	1,55	11,10	33.163.102	**20,58**
Total	**11.432.163**	**100,00**	**12,82**	**161.112.768**	**100,00**

Figura 2 - Distribuição da população brasileira e da população em AGSN em 2010, segundo categorias adaptadas da REGIC. Fonte: Nadalin et al., 2014.

Quando analisada por região, a população residente em AGSN no Brasil se comporta como ilustrado na Figura 3.

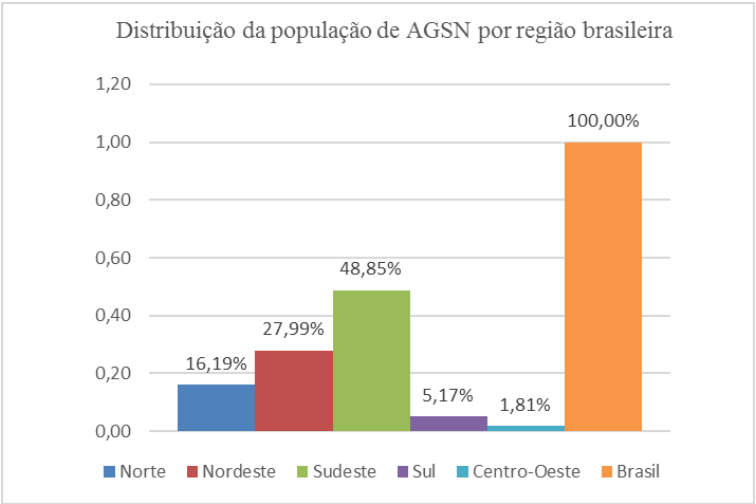

Figura 3 - Distribuição da população residente em AGSN no Brasil (IBGE, 2011).

Do total de habitantes de AGSN no Brasil, o Centro Oeste tem o menor percentual quando comparado às outras regiões brasileiras, no entanto, isso significa que, para atingir a universalização de seus serviços básicos, em especial o saneamento, deve focar nesse percentual da população que ainda não tem acesso. No caso do Distrito Federal, foram mapeados 175 AGSN, que representam 36504 mil domicílios ocupados e comportam 133.556 pessoas – o que corresponde a 5,2% da população total do DF (IBGE, 2011).

3.3 O Saneamento enquanto Direito Humano e a Universalização Inclusiva

Após várias décadas de debate internacional, em 2010, a ONU reconheceu o Direito Humano à Água e ao Esgotamento Sanitário (DHAES). Muitos países demonstraram resistência à consagração desse direito, cuja aprovação não só concluiu um debate que vinha se arrastando por um longo período, mas abriu novas frentes de discussão a respeito do tema. Isso porque a concretização desse direito pode representar, na prática, um aumento da responsabilidade do Estado no que se refere à sua implementação, em termos legais e de políticas públicas (IPEA, 2015).

Por essa razão é que a aprovação do direito é apenas uma etapa inicial do processo de universalização, uma vez que se sabe que a carência de acesso está nas áreas de maior vulnerabilidade e essa problemática apresenta aspectos variados envolvidos e difícil resolução. Em termos de América Latina, as injustiças sociais se manifestam significativamente nas condições de aceso a esses serviços. O relatório da Organização Pan-americana de Saúde (Opas) indicou que, em 2011, cerca de 40 milhões de pessoas – 7% da população latino-americana – não possuíam água segura para o consumo humano, enquanto mais de 117 milhões de pessoas – 20% da população – não usufruíam de instalações sanitárias que apresentassem condições mínimas necessárias (Opas, 2011 apud IPEA, 2015). Além disso, frequentemente, a desigualdade no acesso à água e saneamento, ou na proteção da população contra os perigos associados à água, são acentuadas mesmo em áreas onde há disponibilidade adequada – ou até mesmo abundante – do recurso, a exemplo do sul do México, da Região Amazônica e de regiões às margens dos grandes rios latino-americanos (IPEA, 2015).

No entanto, ocorreram avanços significativos no acesso à água e saneamento para população latino-americana nas últimas décadas. A meta estabelecida pelos Objetivos de Desenvolvimento do Milênio (ODM) de redução em 50% do número de

pessoas sem acesso à água de melhor qualidade foi atingida e a de redução de 50% do número de pessoas sem saneamento chegou muito próximo do sucesso. Porém, os ODM não abarcam todas as dimensões do acesso à água e ao saneamento, de modo que a acessibilidade financeira ou a falta de sustentabilidade devido às disparidades sociais são desafios à parte não evidenciados em estatísticas de cobertura dos serviços de água e esgotamento sanitário (IPEA, 2015).

Assim, em 2015, foi criada a Agenda 2030, com os ODS, os quais incluíram o termo 'sustentabilidade' nos objetivos, na tentativa de torná-los mais abrangentes e atuais na questão socioambiental. O ODS número seis consiste em "Assegurar a Disponibilidade e Gestão Sustentável da Água e Saneamento para todas e todos, independentemente de sua condição social, econômica ou cultural, de gênero ou etnia", por meio do indicador "Proporção da população que utiliza serviços de água potável gerenciados de forma segura" (IPEA, 2018, P. 163).

Dessa forma, o Estado tem a obrigação de proteger e implementar esse direito, não necessariamente por meio da provisão direta, mas monitorando e regulamentando os prestadores e garantindo que ele não seja violado. Porém, as populações vulneráveis são as mais afetadas, demandando políticas públicas específicas e foco direcionado à garantia de um nível aceitável de qualidade de vida e bem-estar (Silva-Neves & Heller, 2016).

Para demonstrar a desigualdade no acesso aos SAAES, pode-se verificar o Índice de Pobreza Hídrica (Water Poverty Index) proposto por Lawrence et al. (2002). O índice mede o bem-estar e a disponibilidade de água de forma integrada, avaliando 140 países, a partir de cinco indicadores, sendo eles: disponibilidade de recursos hídricos, acesso, capacidade instalada, uso e meio ambiente. O índice calculado para alguns dos países da América Latina sinalizou para o Brasil uma situação relativamente pior do que países como Bolívia, Peru, Venezuela, Colômbia, Costa Rica, Uruguai, Equador e Chile.

Além disso, em pesquisa realizada por Schneider et al. (2010), para propor indicadores para melhorar a gestão pública dos serviços de saneamento básico em áreas de concentração de populações vulneráveis em áreas urbanas e periurbanas, aplicada no município de São Carlos – SP utilizando dados do Censo de 2000, foi verificada a relação entre renda e acesso: quanto maior a renda total do responsável do domicílio, maior o acesso aos serviços de saneamento básico na área urbana; concluindo que a

parcela mais fragilizada da população em termos socioeconômicos foi exatamente a que possuía menor percentual de acesso aos serviços de saneamento.

Assim, embora haja uma tendência de melhoria das taxas de cobertura dos serviços de saneamento básico observada nos últimos anos (Castro & Heller, 2009), ainda não se encontraram formas consolidadas de medir ou avaliar como este setor se encontra no que se refere à prestação de SAAES em Assentamentos Precários urbanos.

Dentro desse contexto que Guimarães (2015) propõe o conceito de Universalização Inclusiva (UI), com o intuito de promover uma forma de avaliação da prestação desses serviços em áreas de vulnerabilidade social, que frequentemente são desconsideradas e/ou possuem suas realidades mascaradas em índices calculados para o território urbano como um todo.

Guimarães (2015, p.10) define o conceito de Universalização Inclusiva como:

> [...] a prestação de serviços públicos essenciais à vida de um subcidadão, sub-incluido nas Políticas Públicas, não contado nas metas setoriais, mediante um processo participativo, conduzido pela concessionária em parceria com Poder Concedente e demais atores da sociedade, para prover água e esgotamento sanitário, ainda que em áreas de exclusão social por meio de arranjos sociotécnicos em construções singulares, inclusivas e cuidadoras do Direito Humano.

Nesse sentido, propõe o Indicador de UI, que consiste basicamente na razão entre os domicílios atendidos em áreas de vulnerabilidade social e o número total de domicílios nessas áreas, a fim de garantir um olhar direcionado às demandas da população residente de Assentamentos Precários urbanos ou de outras áreas negligenciadas pelo Poder Público.

3.4 IDENTIFICAÇÃO E MAPEAMENTO DE ASSENTAMENTOS PRECÁRIOS

Conhecer o contexto dos assentamentos precários no Brasil é de extrema importância para a elaboração de políticas públicas, mas segue sendo um desafio para estudiosos, técnicos e gestores públicos, uma vez que a questão envolve questões sociais, ambientais, habitacionais e de saúde. Os órgãos públicos federais têm atuado sem ter o conhecimento do tamanho do problema de fato, assim como as diferentes formas que ele se apresenta ao longo do território brasileiro (Morais et al., 2016).

Satterthwaite (2003) aponta que a qualidade das estimativas existentes em relação a esses assentamentos ainda é muito insuficiente, especialmente observada a alta relevância da questão. Oliveira & Anjos (2004) afirmam que é surpreendente a falta de conhecimento do poder público brasileiro sobre os dados referentes ao número de residentes em favelas, destacando as dificuldades para estimar o crescimento dessa população, registrar as novas áreas ocupadas e atualizar o número dos seus domicílios.

Assim, imprescindível que haja uma coleta de dados nacionais confiáveis que possam ser comparáveis entre si, para que se tenha subsídios suficientes na tomada de decisões. Além disso, destaca-se também a importância de pesquisas locais a nível municipal, uma vez que a maioria dos municípios brasileiros não possuem estimativas e cartografias de assentamentos precários. Entende-se que é dever do Governo Federal a criação de incentivos e a padronização conceitual do fenômeno, de modo que os governos locais possam construir informações detalhadas de seus espaços e atualizá-las periodicamente (Ministério das Cidades, 2007).

Nesse sentido, os dados do Censo Demográfico representam fonte estratégica de obtenção e análise de dados para os fins da presente pesquisa, que se destina a estudar o acesso aos SAAES nos Assentamentos Precários e os desdobramentos que se seguem no mapeamento da carência desses serviços e na seleção de variáveis que permitam definir prioridades na tomada de decisão por parte do poder público.

Nas últimas décadas, muitos estudos têm sido feitos a respeito da detecção de assentamentos precários, utilizando as mais variadas ferramentas. A combinação de diversos métodos tem se mostrado um direcionamento possível para aprimorar a identificação e conhecimento dessas áreas, de modo que neste capítulo, serão mostrados alguns métodos de grande utilização.

3.4.1 Censo IBGE: Aglomerados Subnormais

No âmbito do Censo Demográfico, o IBGE desenvolveu uma metodologia própria para mapear os chamados Aglomerados Subnormais (AGSN), que consistem em conjunto de setores censitários contíguos. Para tal, o Censo Demográfico subdivide os setores censitários urbanos em dois tipos: os setores subnormais, que correspondem àqueles integrantes de um Aglomerado Subnormal (AGSN), e os setores Não Especiais (NE), os quais configuram os setores comuns.

Pelos critérios utilizados, é senso comum na literatura que os Aglomerados Subnormais mapeados pelo IBGE no Censo Demográfico correspondem principalmente às favelas (Nadalin & Mation, 2018), que consiste em apenas uma das tipologias dos Assentamentos Precários, os quais abrangem também os loteamentos clandestinos, loteamentos irregulares e cortiços (Morais et al., 2016).

A primeira contagem da população residente nessas áreas ocorreu em 1950, quando o termo utilizado ainda era 'favela' e sua classificação se baseava na identificação de características relacionadas à densidade populacional, tipo de habitação, serviços públicos, urbanização e a ocupação ilegal da terra. Em 1980, o termo 'favela' foi substituído por 'setor especial de aglomerado urbano', cujo conceito consistia em setores dotado de infraestrutura carente e geralmente localizados em terrenos não pertencentes aos moradores.

Por fim, em 1991, o termo foi atualizado para Aglomerados Subnormais, os quais foram definidos por um grupo de setores censitários "constituídos por um mínimo de 51 domicílios, ocupando ou tendo ocupado até período recente, terreno de propriedade alheia (pública ou particular), dispostos em geral de forma desordenada e densa e carente, em sua maioria, de serviços públicos essenciais" (IBGE, 2001).

A metodologia foi avançando ao longo do tempo, de forma que a delimitação dos setores nos primeiros censos era realizada por meio de dados secundários de cadastros prediais domiciliares e órgãos com atuação em assentamentos precários, passando a contar com idas a campo para atualização dos dados em 1980 e com delimitações efetuadas previamente em campo nos censos de 1991 e 2000, pela Base Operacional Geográfica, até chegar, finalmente, em 2010, quando houve grandes avanços com a inclusão do Levantamento de Informações Territoriais (LIT), o qual tornou possível a espacialização dos dados (Cardoso, 2016).

Assim, no que se refere aos mapeamentos de AGSN realizados nos censos do IBGE, o método mais recente é o desenvolvido em 2010, o qual utilizou recursos como imagens de satélite e visitas a campo, ampliando a metodologia de mapeamento. Por essa razão, o IBGE recomenda fortemente não utilizar os dados de AGSN do censo de 2010 para comparação com os dos anos anteriores. O conceito mais recente de AGSN e atualmente utilizado baseia-se no seguinte:

> [...] são setores censitários constituídos por um conjunto constituído de, no mínimo, 51 unidades habitacionais carentes, em sua maioria de serviços públicos essenciais, ocupando ou tendo ocupado, até período

> recente, terreno de propriedade alheia (pública ou particular) e estando dispostas, em geral, de forma desordenada e densa.
> A identificação dos aglomerados subnormais deve ser feita com base nos seguintes critérios:
> a) Ocupação ilegal da terra; e
> b) Possuírem pelo menos uma das seguintes características:
> • urbanização fora dos padrões vigentes (vias de circulação estreitas e de alinhamento irregular, lotes de tamanhos e formas desiguais e construções não regularizadas por órgãos públicos); ou
> • precariedade de serviços públicos essenciais (IBGE, 2013, p. 18).

As variáveis consideradas para a classificação de AGSN estão relacionadas à topografia (declive acentuado, moderado ou plano), localização (margens de rios, praias, unidades de conservação, etc), padrões urbanísticos (tipos de arruamento, vias de circulação, veículos de circulação, etc) e densidade de ocupação. As características são registradas como existentes em um setor censitário quando presentes em pelo menos 10% dos domicílios, sendo registradas até três características existentes em cada variável; e como predominante quando apresentou o maior número de domicílios em determinada situação. A caracterização espacial de AGSN do IBGE considera apenas os aspectos predominantes das variáveis consideradas.

No entanto, o mapeamento realizado pelo IBGE tem suas limitações. A primeira delas se trata da quantidade mínima de domicílios ocupados para a classificação como AGSN, impossibilitando a detecção de assentamentos precários menores, como cortiços e alguns loteamentos irregulares ou clandestinos. Desse modo, a classificação de AGSN é mais propícia para a caracterização de favelas, mostrando-se incompleta no mapeamento da precariedade habitacional como um todo, à medida que pode subestimar o número de assentamentos precários existentes.

Porém, ainda que haja um certo problema de subestimação, essa é a única informação e padronizada coletada nacionalmente de metodologia definida e confiável, de modo que configura uma excelente ferramenta de baixo custo para se estudar áreas de grande abrangência territorial. Por essa razão, é um dado amplamente utilizado pelas pesquisas e políticas habitacionais, como um norte para suprir a ausência de dados abrangentes sobre os assentamentos precários (Cardoso, 2016).

3.4.2 Método de Estimação dos Assentamentos Precários proposto pelo Centro de Estudos das Metrópoles (CEM)

Em 2007, o Centro de Estudos da Metrópole (CEM), instituição de pesquisa sediada na Universidade de São Paulo (USP) e no Centro Brasileiro de Pesquisa e Planejamento (CEBRAP), contratado pelo Ministério das Cidades, realizou estudo que propôs uma nova metodologia para a identificação de Assentamentos Precários Urbanos, publicada no documento oficial "Assentamentos Precários no Brasil Urbano", de Marques et al. (2007). O objetivo principal foi sistematizar o mapeamento e ampliar a caracterização dos Assentamentos Precários para além dos Aglomerados Subnormais mapeados pelo IBGE, no Censo Demográfico, informação que, ainda hoje, representa a única fonte de dados nacional e padronizada que evidencia a presença desse fenômeno.

No Censo Demográfico, a designação dos setores como SbN é prévia à pesquisa de campo, sendo baseada em critérios estruturais visualizados em imagens de satélite, como precariedade de serviços públicos, arruamento e densidade construtiva (IBGE, 2011). Assim, a metodologia proposta pelo CEM veio no sentido de complementar a informação fornecida pelo IBGE, partindo do princípio de que era possível traçar um perfil socioeconômico da população residente de setores SbN, admitindo-se que os setores censitários cuja população apresentasse tal perfil seguramente conteriam ou corresponderiam a Assentamentos Precários.

Dessa maneira, objetivou identificar indicadores que permitissem caracterizar esse perfil populacional, incluindo informações sociais, econômicas e demográficas, indo além de questões unicamente estruturais. Acreditou-se que tais informações poderiam ser úteis ao aprimoramento do mapeamento, especialmente se posteriores ao levantamento realizado pelo Censo, já em posse dos dados coletados sobre a população residente (Ferreira et al., 2007). Além disso, os setores censitários podem conter núcleos precários de pequeno porte dentro de sua extensão, os quais podem ser incluídos em áreas urbanas mais amplas e terem seus indicadores "diluídos" em médias heterogêneas, de modo que acabem não sendo classificados como SbN.

Assim, o objetivo teve foco em incluir no universo de Assentamentos Precários os setores censitários que correspondem ou abrangem tais áreas, mas que não entraram no rol de setores SbN classificados pelo IBGE, evitando assim a subestimação do fenômeno. Por essa razão é que o método do CEM se baseia no perfil socioeconômico e demográfico do setor censitário – e não somente em suas características físicas ou estruturais, sob o entendimento de que a população mais vulnerável socioeconomicamente se concentraria em áreas mais carentes de estrutura e serviços públicos (Ferreira et al., 2007; Silva et al., 2014).

A menor unidade do estudo, portanto, é o setor censitário, o qual consiste na desagregação territorial mínima utilizada na coleta e divulgação das informações levantadas no censo demográfico. Seu tamanho varia segundo as condições urbanas, as regiões do país e os recenseamentos, mas normalmente apresentam extensão reduzida e uma homogeneidade bastante razoável, uma vez que representam a unidade básica de análise do censo (Silva et al., 2014).

Em suma, a metodologia consiste em identificar, dentre os Setores Censitários NE, aqueles que mais se assemelham aos do tipo SbN em termos sociais, demográficas e de infraestutura da população residente. Ou seja, compara os conteúdos sociais médios dos setores SbN com os dos NE, discriminando aqueles que são similares aos SbN, mas que não foram originalmente classificados como tal pelo IBGE. Por fim, os setores detectados pelo modelo poderiam então demonstrar uma estimativa do total da população habitante de Assentamentos Precários (Ferreira et al., 2007; Marques et al., 2007).

Na metodologia proposta, a discriminação dos setores – e posterior classificação – é realizada por meio da Análise Discriminante Linear, técnica estatística capaz de determinar uma função discriminante entre os dois tipos de setores (NE e SbN) e, com isso, estabelecer critérios que devem ser atendidos para que um setor seja classificado como SbN, expressos na forma de indicadores considerados relevantes para traçar o perfil dos moradores de Assentamentos Precários. Os indicadores estão listados na Figura 4 e foram propostos considerando três dimensões: Habitação e Infraestrutura; Renda e Escolaridade e Aspectos Demográficos (Marques et al., 2007).

Dimensão	Variável
Habitação e infraestrutura	Porcentagem de domicílios sem coleta de lixo
	Porcentagem de domicílios sem ligação à rede de abastecimento de água
	Porcentagem de domicílios sem banheiros ou sanitários
	Porcentagem de domicílios sem ligação à rede de esgoto ou fossa séptica
	Porcentagem de domicílios do tipo cômodo
	Porcentagem de domicílios – outra forma de posse da moradia
	Porcentagem de domicílios – outra forma de posse do terreno
	Número de banheiros por habitante
Renda e escolaridade do responsável pelo domicílio	Porcentagem de responsáveis por domicílio não alfabetizados
	Porcentagem de responsáveis por domicílio com menos de 30 anos não alfabetizados
	Porcentagem de responsáveis por domicílio com renda de até 3 salários mínimos
	Porcentagem de responsáveis por domicílio com menos de 8 anos de estudo
	Anos médios de estudo do responsável pelo domicílio
	Renda média do responsável pelo domicílio
Aspectos demográficos	Número de domicílios particulares permanentes no setor censitário
	Número de domicílios improvisados no setor censitário
	Número de pessoas residentes no setor censitário
	Porcentagem de responsáveis por domicílios com menos de 30 anos
	Número médio de pessoas por domicílio

Figura 4 - Indicadores utilizados no estudo de Marques et al., 2007 para a identificação de Assentamentos Precários. Fonte: Silva et al. (2014).

As funções discriminantes correspondem a somas ponderadas das variáveis, de modo que: a(moradia) + b(instrução) + c(emprego) + d(renda) + k, em que a, b, c e d traduzem aos pesos atribuídos às variáveis para a classificação dos setores e k corresponde à constante (Perez et al, 1994 *apud* Marques et al., 2007; Silva et al., 2014). Com isso, são efetuadas as etapas de comparação e classificação dos setores pelo Modelo Discriminante Linear, de modo que:

a) Os setores originalmente classificados como NE pelo IBGE e classificados pelo modelo também como NE permanecem com essa classificação;
b) Os setores originalmente classificados como SbN pelo IBGE e assim também designados pelo modelo, permanecem como SbN; e configuram uma importante medida de ajuste do modelo; e, por fim,
c) Os setores censitários tidos como NE pelo IBGE e reclassificados para SbN pelo modelo recebem a denominação de Setores Precários – e consistem nos setores representativos de Assentamentos Precários (AP) por meio do perfil populacional traçado na função discriminante.

Por fim, a metodologia considera como Assentamentos Precários o conjunto formado pelos setores censitários que o modelo classificou como SbN, os quais são

chamados de Setores Precários; somado aos setores originalmente classificados pelo IBGE como SbN e que porventura não tenham sido detectados pelo modelo.

O segundo ponto crucial da metodologia proposta a ser destacada se refere à seleção dos indicadores: as variáveis independentes do modelo estatístico não necessariamente corresponderão a todos os indicadores listados na Figura 7. Como o Brasil apresenta grande diversidade histórica de ocupação, cada região apresenta contextos distintos no processo de urbanização e, consequentemente, dos núcleos precários de aglomeração. Isso faz com que as unidades federativas apresentem peculiaridades no perfil populacional dos residentes de Assentamentos Precários, de modo que cada espaço apresentará características próprias que marcam a forma como esse fenômeno se manifesta. Por essa razão, foi definida a utilização do procedimento *Stepwise*, de modo que, para cada região metropolitana, fossem selecionados os indicadores de maior relevância para discriminação de Assentamentos Precários, dentre os apresentados na Figura 4.

A fragmentação da análise foi proposta pelo agrupamento de municípios conforme os seguintes critérios:

a) Os agrupamentos de municípios deveriam apresentar, no mínimo, 20 setores censitários do tipo SbN;
b) As regiões metropolitanas foram consideradas agrupamentos de municípios, exceto quando o número de setores subnormais era inferior a 20;
c) Os municípios foram agrupados respeitando a Região e a Unidade da Federação onde se localizavam.

3.4.2.1 Resultados da Aplicação do Método de Estimação de Assentamentos Precários do Centro de Estudos das Metrópoles (CEM)

Seguindo o protocolo metodológico descrito acima, o método foi aplicado por Marques et al. (2007) para o território brasileiro, utilizando os dados do Censo Demográfico de 2000 como base para o cálculo dos indicadores. A amostra englobou todos os setores censitários NE e SbN do território nacional que estavam localizados em áreas definidas pelo IBGE como ‘Urbanas’ e ‘Rurais de Extensão Urbana’.

O método foi aplicado em 554 municípios brasileiros, englobando 98% dos setores classificados como SbN. À época, isso representava 47,3% dos setores censitários do país e 52% da população brasileira (Marques et al., 2007). Para que os

indicadores mais relevantes fossem selecionados, a área de estudo foi fragmentada em 21 agrupamentos de municípios, representados por Regiões Metropolitanas brasileiras e Demais Municípios da Região, com o objetivo de respeitar as peculiaridades que marcam os Assentamentos Precários de cada região.

Desse modo, os resultados foram expressos na forma de 21 modelos estatísticos, um para cada agrupamento de municípios. Como critério de ajuste do modelo, foi verificado o percentual de setores SbN detectados corretamente pelo modelo, o que foi chamado de "Taxa de acerto de setores SbN", embora a amostra de treino tenha sido igual à amostra de validação, uma vez que compreende a população total, por se tratar de dados do Censo.

Assim, esse percentual, expresso pela taxa de setores classificados pelo IBGE como SbN e classificados pelo modelo como tal, foi entendido como uma medida de acerto, que demonstraria que a função estava de fato representativa da realidade encontrada nos setores SbN e assim teria melhor ajuste ao identificar setores com perfil semelhante (Tabela 1).

Tabela 1 - Regiões de Aplicação do Método do CEM.

Agrupamentos de Municípios	% de Setores do Tipo SbN classificados como SbN pelo Modelo
Região Norte	
RM de Belém	76,8
Demais Municípios da Região Norte	74,6
Região Nordeste	
RM de Maceió	77,6
RM de Salvador	75,7
RM de Fortaleza	73,4
RM de São Luiz	64,6
RM de Recife	65,3
Demais Municípios do Nordeste - Litoral	59,5
Demais Municípios do Nordeste - Interior	71,4
Região Centro Oeste e Sudeste	
Distrito Federal e RM de Goiânia	52,0
RM de Belo Horizonte e Colar Metropolitano	81,3
RM do Rio de Janeiro	80,1
RM de São Paulo	77,8
RM de Campinas	80,1
RM da Baixada Santista	76,7
Demais Municípios de Minas Gerais e do Centro-Oeste	73,5
Demais Municípios do Rio de Janeiro e do Espírito Santo	57,7
Demais Municípios de São Paulo	71,1
Região Sul	

RM de Curitiba	77,7
RM de Porto Alegre	80,5
Demais Municípios da Região Sul	75,4

Fonte: Adaptado de Marques et al. (2007) a partir do Censo Demográfico - IBGE (2000)

É possível verificar que, para que fossem cumpridos os critérios estabelecidos, algumas regiões metropolitanas foram agregadas, como foi o caso do Distrito Federal e Região Metropolitana de Goiânia. Para cada um dos modelos, foi efetuado o Procedimento *Stepwise* de seleção das variáveis, para definir os indicadores mais relevantes na amostra populacional, de modo que cada agrupamento apresentou um grupo de indicadores distinto. Por exemplo, dos dezenove indicadores propostos originalmente, foram utilizados seis na RM de São Luís; oito para o Distrito Federal e RM de Goiânia; nove para a RM de Maceió; onze para a RM de Belém; doze na RM de Recife; treze na de Porto Alegre; quatorze em Campinas e Minas Gerais; quinze na Baixada Santista, Curitiba e Rio de Janeiro; dezesseis na RM de Belo Horizonte; dezessete em Fortaleza e dezoito para a RM de São Paulo (Ferreira et al., 2007).

Assim, fica claro que tal procedimento apresenta um papel fundamental no estudo, uma vez que, como já mencionado, uma das dificuldades metodológicas está em relação à própria definição de Assentamentos Precários, já que cada região apresentará especificidades e terá seu próprio perfil habitacional e estrutural em áreas de ocupação precária (Queiroz Filho, 2015).

Quanto ao percentual de acerto, observou-se que a maioria dos valores ficaram entre 60% e 80%, sendo melhores, de forma geral, nas RMs que concentravam os maiores municípios. Apenas quatro regiões apresentaram taxa de acerto menor que 60%, com destaque para a região formada pelo Distrito Federal e RM de Goiânia, onde houve menor aderência: apenas 52% dos setores SbN foram classificados como tal pelo modelo (Ferreira et al., 2007).

Outra importante análise realizada se deu na comparação de algumas características sociais entre os moradores de setores classificados como SbN pelo IBGE e os Assentamentos Precários, representados pela soma dos setores SbN mapeados pelo IBGE e detectados pelo modelo. Para os objetivos dessa pesquisa, foram selecionados os gráficos a seguir, que compara os acessos aos serviços de Abastecimento de Água e Esgotamento Sanitário (Tabela 2).

Tabela 2 - Proporção de domicílios sem acesso à rede de abastecimento de água, segundo tipo de setor censitário (em %).

Unidade de Análise	Percentual de Domicílios Sem Acesso à rede Abastecimento de Água	Percentual de Domicílios Sem Acesso à rede de Esgotamento Sanitário ou Fossa Séptica
Aglomerados Subnormais	12,5	38,7
Assentamentos Precários	17,1	40,6
Setores Não Especiais	8,1	17,1
Brasil	9,0	20,1

Fonte: (Ferreira et al., 2007).

É possível perceber valores similares entre os setores SbN e os assentamentos precários definidos no estudo. Nota-se também que ambos apresentam valores muito distantes da realidade encontrada nos setores comuns (NE) e no Brasil como um todo, razão pela qual merecem especial atenção. Além disso, destacam-se os altos índices encontrados de domicílios sem acesso à rede de esgoto e, ainda, um percentual significativo de falta de acesso à rede de abastecimento de água. Lembrando, ainda, que se esses percentuais estivessem separados por região, ou por estado, provavelmente os resultados mostrariam índices completamente distintos, o que denota a importância da análise em escalas menores.

Por fim, é importante pontuar que, como o método do CEM data de 2007, foi desenvolvido e aplicado para o último Censo da época, cuja realização ocorreu em 2000. Por isso, foi publicada uma atualização por Silva et al. (2014), a qual indica as adaptações necessárias a serem realizadas para aplicação ao Censo de 2010, especialmente no que se refere às variáveis que sofreram alterações de um censo para outro. Embora a metodologia tenha sido atualizada, não chegou a ser reaplicada ao último censo e, portanto, ainda não há resultados de Assentamentos Precários para o ano de 2010.

Além disso, é necessário ressaltar algumas limitações levantadas pelos autores. A primeira delas refere-se ao fato de que os dados são agregados e, por serem provenientes de setores censitários, não se pode discretizar as informações em escalas menores. Isso significa que muitos setores não seriam classificados como precários – como não foram, pelo IBGE –, mas apresentam núcleos de evidente precariedade e, portanto, foram contabilizados. Em segundo lugar, as informações se limitam aos dados do Censo, não incluindo possíveis dados municipais disponíveis. E, por fim, esse método indica a existência de precariedade e onde ela se encontra, porém não é capaz de

especificar que tipo de problema está envolvido, não substituindo trabalhos mais detalhados de campo e visitas *in loco* (Ferreira et al., 2007; Silva et al., 2014).

Entretanto, ainda se faz necessária a uniformização do mapeamento de Assentamentos Precários no Brasil, que possa ser utilizada como *proxy* para a detecção dessas áreas e apresente uma metodologia unificada que permita não só mapeá-las como realizar comparações entre cidades e regiões, para que, assim, políticas públicas nacionais possam ser direcionadas à precariedade habitacional sofrida pela população mais vulnerável socioeconomicamente.

4 REFERENCIAL TEÓRICO

4.1 TÉCNICAS ESTATÍSTICAS MULTIVARIADAS PARA CLASSIFICAÇÃO DE VARIÁVEIS CATEGÓRICAS BINÁRIAS

Neste tópico, serão apresentadas três técnicas estatísticas de análise multivariada aplicáveis a dados que possuem a variável dependente categórica – também chamada de não-métrica ou qualitativa – de natureza binária ou dicotômica, ou seja, que pode ser representada por 0 ou 1. No presente estudo, a variável dependente se enquadra nesse contexto, representada pelo tipo de setor censitário, NE (0) e SbN (1), enquanto as variáveis independentes são quantitativas ou métricas.

Assim, para analisar as observações e classificá-las em um dos grupos nesse caso, de acordo com a literatura, há duas principais técnicas estatísticas capazes de separar grupos e/ou alocar um novo elemento em um desses grupos: a Análise Discriminante – que se divide em Linear e Quadrática – e a Regressão Logística, as quais representam métodos que relacionam um conjunto de variáveis independentes a uma variável dependente categórica (Sharma, 1996; Hair et al., 1998; Morgan e Griego, 1998).

Neste tópico, portanto, serão detalhados os três métodos: Análise Discriminante Linear, Análise Discriminante Quadrática e Regressão Logística.

4.1.1 Análise Discriminante

A Análise Discriminante (AD) é uma técnica estatística inicialmente abordada por Fisher (1936) e utilizada para discriminar e classificar objetos em grupos previamente definidos, com base em variáveis que reflitam suas características gerais. Tem como objetivo identificar diferenças entre grupos, por meio de uma ou mais funções matemáticas compostas por combinações lineares que relacionam uma variável dependente, categórica, a variáveis independentes, não categóricas, ou métricas.

Assim, com base nessas variáveis independentes, é realizada a discriminação e posterior classificação. A discriminação corresponde à primeira etapa, atuando como uma análise exploratória dos dados, que consiste na busca das variáveis independentes que melhor discriminam os grupos. Já a classificação é definida como um conjunto de regras aplicadas à alocação de novas observações no grupo cujo com seu conjunto de

características mais se assemelha (Johnson & Wichern, 2007; Hair et. al, 2009; Mingoti, 2005).

Os três objetivos principais da técnica foram sintetizados por Favero et. al (2009) e Sharma (1996):

i) Identificar as variáveis que melhor discriminam os grupos;
ii) Usar as variáveis identificadas para desenvolver uma ou mais funções discriminantes que sejam capazes de representar as diferenças entre os grupos;
iii) Usar as funções construídas para o desenvolvimento de regras de classificação para alocar as observações nos grupos já existentes.

Assim, o propósito da Análise Discriminante é a relação entre um grupo ou classe, representada por uma variável dependente não-métrica, por meio da obtenção de funções matemáticas capazes de classificar um elemento ou observação em um dos grupos com base em suas características, representadas por um conjunto de *p* variáveis independentes métricas (Hair et. al, 2009), apresentando, portanto, a seguinte forma geral:

$$\underset{\text{(não-métrica)}}{Y} = \underset{\text{(métricas)}}{X_1 + X_2 + X_3 + \ldots + X_p} \qquad (1)$$

A população amostral pode conter dois ou mais grupos, de modo que o número de funções matemáticas corresponderá ao número de grupos menos um (*g – 1)*. Ou seja, para variáveis dependentes binárias, onde há apenas dois grupos, ter-se-á apenas uma função discriminante. Portanto, para compreender a técnica, considera-se um exemplo com dois grupos: o grupo 1, com n_1 elementos amostrais e probabilidade de ocorrência P_1 e um grupo 2, com n_2 elementos e probabilidade P_2, sendo que, para cada um dos seus $n_1 + n_2 = n$ elementos ou observações, tenham sido medidas *p* características, expressas na forma de variáveis independentes.

A análise estatística dessas variáveis permite identificar o perfil geral dos n_i elementos que compõem cada grupo e comparar cada nova observação com os perfis dos grupos 1 e 2, alocando-a, por fim, no grupo cujo perfil mais se assemelha com o seu, através da maximização da diferença entre-grupos e a minimização intra-grupos (Hair et. al, 2009; Mingoti, 2005). A alocação da observação em um dado grupo ocorre por meio da função discriminante, que gera um *score* discriminante Z para cada observação e, com base nos *scores* médios de cada grupo, é definida uma linha de corte

que separa os grupos, definindo as regiões de classificação (Figura 5). A linha pode reta ou curva, dependendo do tipo de função discriminante (linear ou quadrática).

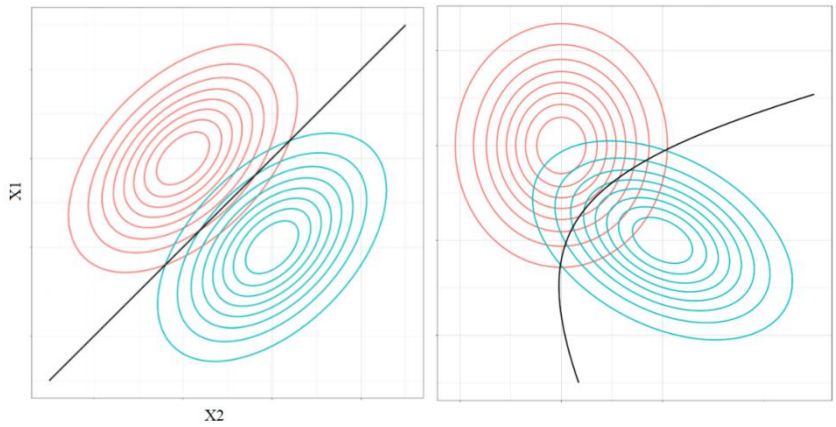

Figura 5 - Regiões de classificação dos grupos definidas por uma função linear – reta (à esquerda) e por uma função quadrática – curva (à direita). Fonte: Math for Machines, 2020[1].

Para o caso da Análise Discriminante Linear, as regiões de classificação são determinadas por um ponto de corte Z (*cut-off value*), o qual é situado no ponto de interseção entre as curvas de densidade dos *scores* discriminantes Z de cada grupo. A Figura 6 mostra as curvas de densidade plotadas em relação aos scores Z de cada grupo, expressando dois casos de discriminação: acima, observa-se boa separação entre os grupos, com uma área de interseção mínima (área sombreada); e, abaixo, uma discriminação de má qualidade, onde há maior área de sobreposição e, consequentemente, maior chance de erros (Hair et. al, 2009).

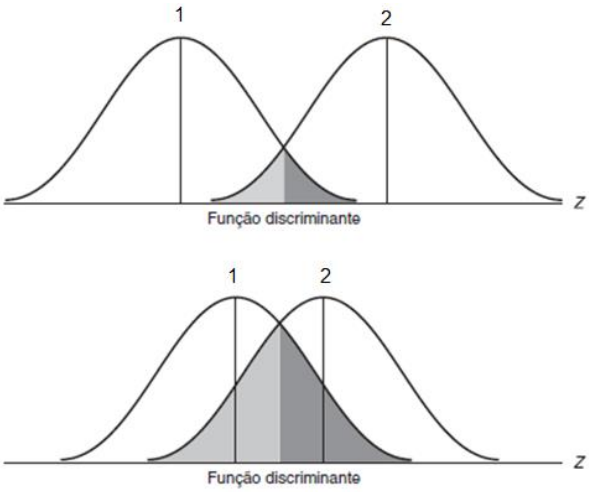

Figura 6 - Representação univariada de scores Z discriminantes. Fonte: Adaptado de Hair et. al, 2009.

[1]Website: https://mathformachines.com/posts/discriminant-analysis/.

Pressupostos da Análise Discriminante

A Análise Discriminante requer a satisfação de três principais premissas – ou pressupostos – que garantam a confiabilidade da sua aplicação. São elas: a) a existência de normalidade multivariada nas variáveis independentes; b) ausência de multicolinearidade e c) a homogeneidade de matrizes de covariância dos grupos. O último pressuposto é flexibilizado pela Análise Discriminante Quadrática, sendo este o principal critério que a diferencia da linear, sendo, portanto, recomendada sua utilização para os casos em que não se satisfaz essa premissa.

O primeiro pressuposto, de normalidade multivariada, pode ser avaliado por meio do teste de Mardia, que avalia assimetria e curtose e do teste de Henze-Zirkler, que utiliza a distância de Mahalanobis entre as observações e entre cada uma delas e o centroide. Ambos os testes têm como H_0 (hipótese nula) que a amostra provém de uma população normal, sendo frequentemente utilizado o nível de significância de 0,05.

É importante pontuar que, se a violação do pressuposto de normalidade multivariada for causada apenas por uma assimetria na distribuição, a eficácia não é comprometida, especialmente em grandes amostras (Favero et. al, 2009). Além disso, o impacto da violação reduz efetivamente quando a amostra possui 200 casos ou mais (Hair et. al, 2009), pelo Teorema do Limite Central (Magalhães, Lima, 2015; Moretin, 1999).

Já a multicolinearidade pode ser analisada pela matriz de correlação de Pearson, como proposto por Favero et al. (2009), que atesta que valores acima de 0,85 devem ser retirados do modelo, pois provavelmente indicam que duas variáveis independentes do modelo estão explicando o mesmo fenômeno, a não ser que represente uma correlação espúria.

Por fim, a homogeneidade de covariâncias é analisada pela estatística M de Box, o qual é particularmente sensível a desvios da normalidade. O teste tem como hipótese nula a igualdade entre as matrizes de covariâncias, a qual não é rejeitada se p-valor for maior que o nível de significância estabelecido – definido por padrão em 0,05.

Seleção de variáveis

Para obter um modelo bem ajustado, é necessário selecionar o melhor conjunto de variáveis independentes para compor a função discriminante, uma vez que estas fornecem a base para a discriminação entre os grupos. A etapa de seleção de variáveis

na AD é realizada por meio do procedimento *Stepwise*, o qual consiste basicamente em sucessivas iterações incluindo, excluindo – ou ambos – as variáveis independentes pré-selecionadas até alcançar um conjunto de variáveis que represente a melhor combinação entre elas, ou seja, até encontrar melhor ajuste do modelo, de acordo com o critério escolhido e a significância estatística estabelecida. O procedimento *Stepwise* é classificado de acordo com o método que utiliza a cada etapa do processo: a inclusão (*forward*), a exclusão (*backwards*) ou ambos (*both)* (Sharma, 1996).

Há alguns métodos comumente utilizados como critério para a seleção de variáveis, dentre eles o Lambda de Wilks, representado pela letra grega Λ. Seu cálculo é realizado por meio da divisão do determinante da matriz de Soma dos Quadrados e Produtos de Resíduos (W) e o determinante da matriz das Somas de Quadrados e Produtos Total (T) e tem o objetivo de avaliar as diferenças de médias entre os grupos para cada variável independente métrica. Seu valor varia entre 0 e 1, de modo que, quanto mais próximo de zero, melhor a capacidade de discriminação da variável em questão.

Dessa forma, o procedimento *Stepwise*, quando aplicado com o método do Wilks Λ, interrompe as iterações quando encontra o menor valor possível de Λ dentre todas as combinações de variáveis independentes dentro do modelo, além de excluir aquelas que apresentarem significância estatística fora do limite desejado, geralmente definido em 5%.

Estimação da Função Discriminante Linear

O problema da discriminação de grupos e posterior classificação foi abordado pela primeira vez por Fisher, em 1936, ao considerar a população de um grupo 1, com vetor de variáveis independentes X de distribuição normal, média μ_1 e matriz de covariância $\sum_1$; e a população de um grupo 2, também com média μ_2, matriz de covariância $\sum_2$ e um vetor X também com distribuição normal, representado por $X = [X_1, X_2, ..., X_p]$, propôs a obtenção da combinação linear das p características observadas que fornecesse a maior discriminação entre as populações, sendo capaz de classificar uma observação em uma delas, buscando minimizar a probabilidade de erro de classificação – ou seja, a chance de classificar no grupo da população 1 uma observação originalmente pertencente ao grupo da população 2 e vice-versa (Johnson & Wichern, 2007; Sharma, 1996).

Mingoti (2005) mostra que essa combinação linear é derivada da função densidade de probabilidade, de modo que, para cada valor de X, é possível calcular a razão entre as distribuições probabilidades dos dois grupos, a qual é denominada de razão de verossimilhança, definida por:

$$\lambda(x) = \frac{função\ densidade\ de\ x\ no\ grupo\ 1}{função\ densidade\ de\ x\ no\ grupo\ 2} \quad (2)$$

Ao substituir os valores das funções de densidade, o resultado é reescrito em função de $-2\ln(\lambda(x))$ e não mais de λ(x), para fins de simplificação de cálculo. Assim, após algumas transformações algébricas, chega-se em:

$$-2\ln(\lambda(x)) = [(x_1 - \mu_1)' \sum_1^{-1} (x_1 - \mu_1)] - [(x_2 - \mu_2)' \sum_2^{-1} (x - \mu_2)] + [\ln|\sum_1| - \ln|\sum_2|] \quad (3)$$

No entanto, ainda é possível simplificar ainda mais o resultado obtido, uma vez que, na ADL, pressupõe-se que a diferença entre os grupos está associada apenas ao vetor médio das variáveis independentes, assumindo que a matriz de covariância é igual entre os grupos. Assim, assumindo que as matrizes de covariância idênticas e vetores de média diferentes, iguala-se $\Sigma = \Sigma_1 = \Sigma_2$ e substituem-se os valores, simplificando para:

$$-2\ln(\lambda(x)) = (x - \mu_1)' \sum^{-1} (x - \mu_1) - (x - \mu_2)' \sum^{-1} (x - \mu_2)' \quad (4)$$

Que equivale à função conhecida como "função discriminante de Fisher", expressa como:

$$fd(x) = (\mu_1 - \mu_2)' \sum^{-1} x - \frac{1}{2} (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 + \mu_2) \quad (5)$$

De forma que a fronteira de decisão é definida por uma função linear, onde a classificação de determinado elemento dependerá da igualdade $fd(x) = 0$. Em outras palavras, tal igualdade representará o chamado ponto de corte ou *threshold*, de modo que determinada observação será classificada no grupo 1 se fd(x) for maior que zero ou, de forma equivalente, se for:

$$(\mu_1 - \mu_2)' \sum^{-1} x > \frac{1}{2} (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 + \mu_2) \quad (6)$$

E será classificada no grupo 2, se for:

$$(\mu_1 - \mu_2)' \sum^{-1} x < \frac{1}{2} (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 + \mu_2) \quad (7)$$

Ao estabelecer que $(\mu_1 - \mu_2)' \sum^{-1} = a'$, o lado esquerdo da função pode ser representado por $a'x$:

$$(\mu_1 - \mu_2)' \sum^{-1} x = a'x = a_1 x_1 + a_1 x_2 + \ldots + a_p x_p \quad (8)$$

Enquanto o lado direito, que corresponde ao ponto de corte, representado por C, corresponde à constante que delimita as regiões de classificação:

$$\frac{1}{2} (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 + \mu_2) = a' \frac{(\mu_1 + \mu_2)}{2} = C \quad (9)$$

Dessa forma, a razão de verossimilhança resulta em uma função matemática construída com base nas matrizes de covariância e vetores médios das variáveis independentes, formando, finalmente, a função discriminante linear, que assume a forma geral expressa pela equação:

$$Zn = a_1 x_1 + a_1 x_2 + \ldots + a_p x_p + C \quad (10)$$

Onde Z é o *score* discriminante de dada observação n, a_p são os coeficientes discriminantes de cada variável independente e x_p são os valores das variáveis independentes para a observação n. Por meio dessa forma geral, calcula-se o valor de Z para cada observação e, de acordo com os critérios definidos acima, será estabelecido um *score* de corte e um *score* médio por grupo (centroides). A classificação ocorre de tal forma que um novo elemento será alocado no grupo cujo *score* médio mais se aproxima do seu (Mingoti, 2005; Sharma, 1996; Klecka, 1980).

Estimação da Função Discriminante Quadrática

A função discriminante quadrática também é estimada com base na razão de verossimilhança, sendo originada da mesma base que a linear, com a exceção de que, neste caso, parte-se da premissa de que as matrizes de covariância dos grupos não são

iguais ($\Sigma_1 \neq \Sigma_2$). Assim, os termos da Equação 3 não são simplificados, resultando, portanto, em:

$$-2\ln(\lambda(x)) = [(x_1 - \mu_1)'\ \Sigma_1^{-1}\ (x_1 - \mu_1) + \ln|\Sigma_1|] - [(x_2 - \mu_2)'\ \Sigma_2^{-1}\ (x - \mu_2) + \ln|\Sigma_2|] \quad (9)$$

Onde x representa o vetor de variáveis independentes, μ é o vetor média da variável independente e Σ é a matriz de covariância dos grupos correspondentes. Para fins de aplicação da técnica, comumente o resultado é separado em dois, sendo expresso em uma função discriminante para cada grupo, representadas por δ_1 e δ_2:

$$\delta_1 = -\frac{1}{2}(x - \mu_1)'\ \Sigma_1^{-1}(x - \mu_1) - \frac{1}{2}\ln|\Sigma_1| \quad (10)$$

$$\delta_2 = -\frac{1}{2}(x - \mu_2)'\ \Sigma_2^{-1}(x - \mu_2) - \frac{1}{2}\ln|\Sigma_2| \quad (11)$$

A fronteira de decisão para classificação, que divide as regiões de classificação, será representada por uma curva composta pelo conjunto de pontos que satisfazem a igualdade $\delta_1 = \delta_2$. Assim, cada grupo terá seu *score* discriminante, de forma que se:

$$\delta_1 - \delta_2 > 0 \quad (12)$$

A observação será classificada no grupo 1 e, da mesma maneira, se

$$\delta_1 - \delta_2 < 0 \quad (13)$$

A observação será classificada no grupo 2.

Por fim, é necessário pontuar que os cálculos detalhados acima se referem à situação em que as probabilidades *a priori* de um elemento pertencer ao grupo 1 (p_1) e ao grupo 2 (p_2) são iguais entre si, ou seja, de 0,5, uma vez que a soma das probabilidades não deve exceder 1. Quando as probabilidades *a priori* diferem entre si, acrescenta-se ao final da função discriminante linear o termo $\ln(p_1/p_2)$. No caso das funções discriminantes quadráticas, acrescenta-se o termo $\ln p_1$ na função do grupo 1 e $\ln p_2$ na função do grupo 2.

4.1.2 Regressão Logística

A Regressão Logística (RL) é uma técnica estatística que, assim como a AD, é utilizada para analisar o comportamento de uma variável dependente categórica binária, em função de um conjunto de variáveis independentes. A variável dependente ou explicativa representa o sucesso ou a ocorrência do evento de interesse, geralmente denotada por 1; e o fracasso ou não-evento, geralmente representado por 0. As variáveis independentes podem ser métricas ou não métricas e o estudo consiste nos efeitos que suas variações causam na ocorrência – ou não – do evento de interesse (Favero; Belfiore, 2017; Favero et al., 2009; Hair et al., 2009).

O modelo de Regressão Logística pertence à família do Modelo Linear Generalizado, que consiste em uma generalização do modelo de Regressão Linear que permite que as variáveis independentes tenham outras distribuições além da normal. Nesse caso, como a variável dependente é binária, apresenta distribuição de Bernoulli. O modelo estima a probabilidade de ocorrência do evento de interesse e se relaciona às variáveis independentes por via de uma função de ligação chamada função *logit* (Favero et al., 2009; Belfiore, 2017; Hosmer & Lemeshow, 2000; Wasserman, 2004).

O objetivo do modelo de RL é prever as classes das observações, alocando-as em 0 ou 1 de acordo com a sua probabilidade de ocorrência, observada a definição de um limite – ou *threshold* – para a classificação na classe 0 ou 1. Usualmente, esse limite é padronizado em 0,5, mas pode ser modificado conforme o conhecimento da probabilidade *a priori* observada de ocorrência do evento. A curva logística é um sigmoide, apresentando formato de "S" de forma a restringir os valores entre 0 e 1, os quais correspondem às probabilidades de ocorrência do evento de interesse (Hosmer & Lemeshow, 2000) (Figura 7).

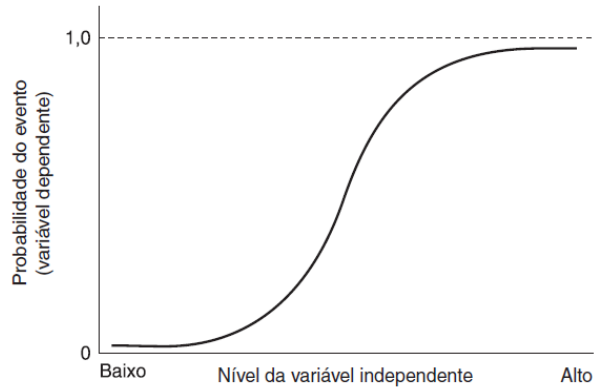

Figura 7 - Comportamento das probabilidades de ocorrência do evento em relação a uma variável independente. Fonte: Hair et al., 2009.

Para compreender melhor a técnica, é fundamental que se faça a diferenciação entre probabilidade, chance e razão de chances. Por exemplo, supondo que haja dois grupos de populações, A e B, e que, para cada uma de suas observações, tenhamos um vetor X com k características $X = [X_1\ X_2\ ...\ X_k]'$. A variável dependente Y é binária e codificada como 0 ou 1, sendo definida como 1 quando uma observação pertencer ao grupo A e, alternativamente, como 0 quando pertencer ao grupo B (Favero; Belfiore, 2017; Wasserman, 2004).

Dessa forma, se a probabilidade de uma observação pertencer à população A for dada por *P*, a probabilidade de pertencer à população B será de 1 – P. Já a chance – também conhecida como *odds* – de determinada observação pertencer à população A ou, equivalentemente, de Y = 1, será a razão entre a probabilidade de pertencer a esta população e a probabilidade de não pertencer, ou seja, $odds_{(Y=1)} = \mathrm{P}/(1-\mathrm{P})$. Da mesma maneira, a chance de Y=0 será dada por $odds_{(Y=0)} = (1-\mathrm{P})/\mathrm{P}$ (Favero; Belfiore, 2017; Favero et al., 2009).

Por fim, a razão de chances – também designada por *odds ratio* (OR) – será a divisão de uma chance pela outra e representará quantas vezes maior – ou menor – será a chance de uma determinada observação ser classificada no grupo A em relação ao grupo B (ou vice-versa) (Favero; Belfiore, 2017; Hosmer & Lemeshow, 2000). Tais conceitos são cruciais para a interpretação dos resultados da Regressão Logística, uma vez que essa técnica modela o logaritmo neperiano da razão de chances, como será detalhado nos tópicos posteriores deste capítulo.

Pressupostos da Regressão Logística

Os pressupostos da Regressão Logística são: a) a relação entre o termo logit e o vetor de variáveis independentes é linear; b) o valor esperado dos resíduos é igual a zero; c) ausência de valores extremos ou *outliers*; d) ausência de alta correlação entre as variáveis independentes (Favero et al., 2009; Hilbe, 2015; Kassambara, 2018). Alguns autores incluem também a premissa de ausência de heterocedasticidade, como Favero et al. (2009), também sustentada por Wooldridge (2016), que afirma que modelos logísticos heterocedásticos terão seus parâmetros enviesados, resultando em interpretações errôneas.

Como é possível observar, os pressupostos da Regressão Logística são mais flexíveis que os das técnicas anteriores, uma vez que não incluem a normalidade multivariada e a homogeneidade entre as matrizes de covariância, o que a torna um

método mais comumente utilizado na literatura quando se trata de análises de variáveis dependentes categóricas binárias (Favero et al., 2009; Hair et al., 2009).

Os dois primeiros pressupostos são hipóteses *ad hoc*, ou seja, no momento da utilização da técnica, parte-se do princípio de que são verdadeiros. Quanto aos demais, Kassambara (2018) sugere um conjunto de etapas a serem realizadas no software R Project para avaliar a satisfação das premissas. São elas:

i. Plotagem do termo logit *versus* as variáveis independentes contínuas
ii. Uso do método da distância de Cooks para avaliar a presença de *outliers*, considerando alarmantes aqueles que apresentam resultados maiores que 3, por meio do Pacote Broom.
iii. o cálculo do Variance Inflation Factor (VIF) para avaliar a presença de multicolinearidade entre as variáveis independentes, considerando alarmantes os valores de VIF acima de 5.

Seleção de Variáveis

O procedimento *Stepwise* de seleção de variáveis, como já mencionado, consiste na da adição e remoção iterativa de variáveis preditoras, resultando na seleção de um número reduzido de variáveis que formem o modelo de melhor performance. É uma técnica extremamente útil a modelos que contêm múltiplas variáveis, uma vez que reduz a complexidade e aprimora sua acurácia (Kassambara, 2018).

Na Regressão Logística, há alguns critérios que podem ser utilizados para determinar a inclusão ou exclusão das variáveis independentes. Kassambara (2018) e Hilbe (2015) dão destaque para o Akaike Information Criteria (AIC), por considerarem uma das informações estatísticas mais utilizadas para fins de pesquisa.

O critério de Akaike calcula a distância entre modelo avaliado para um modelo "real", que é desconhecido e representa o modelo que melhor descreve os dados. Tal distância é chamada divergência de Kullback-Leibler (K-L) e está associada à informação que se perde por usar o modelo aproximado e não o "real" (Hilbe, 2015; Wasserman, 2004). O cálculo do AIC se dá por:

$$AIC = -2L + 2K \tag{14}$$

Em que L é a função de máxima verossimilhança do modelo e K, o seu número de parâmetros. O AIC tenta selecionar o modelo que minimiza a divergência K-L, ou seja, melhor será o ajuste do modelo quanto menor for o valor resultante do AIC.

Estimação da Função Logística

A Regressão Logística é assim denominada pelo fato de se basear na função de distribuição logística (Wasserman, 2004), que é dada por:

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+\ e^{x}}\ , para\ x = \ -\infty, \dots, +\infty \tag{15}$$

Em que, no modelo de Regressão Logística, assume valores entre zero e um, que correspondem às probabilidades de ocorrência do evento de interesse (Y=1). O modelo é dado por:

$$P(Y=1|X) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{k}\beta_i X_i\right)}{1+\exp\left(\beta_0 + \sum_{i=1}^{k}\beta_i X_i\right)} \tag{16}$$

Onde P é a probabilidade de que Y=1 em dada observação, que possui seu vetor X_i com os valores das k variáveis independentes $(i = 1, 2, \dots, k)$; β_0 representa o intercepto e β_i, os k parâmetros do modelo.

Analogamente, $1 - \ P(Y = 1|X)$ será a probabilidade de não ocorrência do evento ou de que Y=0:

$$P(Y=0|X) = 1 - \ P(Y=1|X) = \ \frac{1}{1+\exp\left(\beta_0 + \sum_{i=1}^{k}\beta_i X_i\right)} \tag{17}$$

Assim, substituindo as expressões dadas nas equações 16 e 17, a chance (*odds*) de ocorrência do evento de interesse resultará em:

$$odds\ (Y=1) = \frac{P(Y=1)}{P(Y=0)} = e^{\beta 0 + \sum \beta i X i}$$

No entanto, ao utilizar o logaritmo natural da chance, tem-se um modelo linear para seus parâmetros, o qual é conhecido como termo *logit* e pode ser representado por Z:

$$Z = ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \quad (18)$$

De modo que a função do modelo de Regressão Logística toma a seguinte forma:

$$f(Z) = \frac{1}{1 + e^{-(Z)}} = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i X_i)}} \quad (19)$$

Onde o parâmetro β_0 é o primeiro logaritmo natural da chance quando todas as variáveis independentes são nulas; e β_i, os coeficientes das variáveis independentes para i = 1, 2, ... k. O coeficiente β_i corresponde à mudança no logaritmo natural da chance (*odds*) diante da variação de uma unidade na variável x_i (Favero et al., 2009). Destaca-se que a mudança que uma variável x_i ocasiona na chance poderá ser representada pelo coeficiente β_i exponenciado, ou seja, na forma e^{β_i} (Hair et al., 2009).

Tais coeficientes são estimados pelo método da máxima verossimilhança, onde uma função de verossimilhança é construída com a probabilidade de ocorrência do evento em função de parâmetros inicialmente desconhecidos, em que os coeficientes serão os estimadores de máxima verossimilhança dos parâmetros cujos valores maximizam a função (Wasserman, 2004). Dessa forma, tem-se os coeficientes do modelo, cuja interpretação possui um papel de alta relevância no estudo do modelo e nas inferências que podem ser feitas a partir dele (Hosmer & Lemeshow, 2000).

Os coeficientes das variáveis independentes do modelo trazem informações de como aquela variável se relaciona com a ocorrência do evento de interesse. Hair et al. (2009) divide a interpretação dos coeficientes em direção e magnitude. A primeira diz respeito aos sinais e revela se a variável tem uma relação positiva ou negativa com a probabilidade de ocorrência do evento de interesse, enquanto a magnitude se refere ao quanto essa variável impacta percentualmente na razão de chances.

Em termos de direção, tem-se que: se o sinal de βi for positivo, um aumento na variável explicativa *i* ocasiona um aumento na probabilidade de ocorrência do evento de interesse, ao passo que, se for negativo, causa uma redução. Já para fazer a análise da

magnitude, utiliza-se a razão de chances, também chamada de *odds ratio*, que consiste na chance de ocorrência do evento de interesse em relação à sua não ocorrência, ou seja:

$$OR_{1,0} = \frac{odds\ (Y = 1)}{odds\ (Y = 0)} = \frac{e^{\beta 0 + \sum \beta i Xi}}{e^{\beta 0}} = e^{(\beta 0 + e^{\beta i}) + \beta 0} = \ e^{\beta i} \quad (16)$$

A razão de chances expressa quantas vezes maior ou menor é a chance de ocorrer determinado evento em relação à sua não ocorrência diante da alteração em certa variável. O cálculo dessa variação é feito pela razão entre a chance de ocorrência do evento quando a variável $i = x + 1$ e quando $i = x$. Essa razão corresponde ao número de Euler elevado ao coeficiente da variável, $e^{\beta i}$, sendo expressa em fator. Apresenta uma relação diretamente proporcional à ocorrência do evento de interesse quando maior que um e, inversamente proporcional quando menor que um (Hosmer & Lemeshow, 2000).

Regressão Logística em casos de Eventos Raros (ER)

Amostras com Eventos Raros (ER) são aquelas em que o número de não eventos é dezenas ou centenas de vezes maior que o número de eventos, isto é, quando a população amostral é desbalanceada. No caso da Regressão Logística, a aplicação da técnica em amostras dessa natureza pode causar um enviesamento dos parâmetros, que pode resultar na subestimação do número de eventos de interesse na predição (King e Zeng, 2001). Nesse sentido, King & Zeng (2001) propuseram um método de estimação de modelos logísticos baseados na correção do viés quando os dados apresentam eventos raros.

Uma estratégia muito comum nesses casos é a reamostragem, também denominada *case-cohort*, que consiste em uma seleção aleatória e sem reposição de todas as observações para gerar uma subamostra contendo todos os eventos (uns) e uma quantidade pré-determinada de não-eventos (zeros). Em seu estudo, King e Zeng (2001) sugerem alcançar uma amostra que varie no máximo de duas a cinco vezes mais zeros do que uns, uma vez que, em teste realizado em uma amostra de 1000 observações, os autores notaram a ausência total de viés nos pontos onde foram descartados aproximadamente 55% e 78% zeros da amostra original.

Nesse caso, todos os parâmetros do modelo permanecem consistentes, com exceção do intercepto. Assim, os autores propõem um método de correção que permite

manter a consistência dos estimadores de máxima verossimilhança quando o modelo é construído com subamostras geradas por meio dessa técnica. O método consiste na correção do intercepto do modelo, sendo chamado de Correção a Priori, a qual envolve a correção do intercepto (β_0) do modelo de Regressão Logística com base na informação a priori da proporção de eventos encontrada na população e da proporção de eventos presentes na amostra. A técnica de Correção a Priori apresenta grande facilidade de uso, uma vez que se refere à correção apenas do intercepto, sendo os demais parâmetros estimados da forma usual.

4.2 Vantagens da Regressão Logística em relação à Análise Discriminante

As técnicas de discriminação buscam uma ou mais funções que discriminem os grupos definidos pela variável dependente categórica, visando minimizar erros de classificação. Em situações em que o conjunto de variáveis independentes apresentam normalidade multivariada, Sharma (1996) e Hair et al. (1998) apontam que a Análise Discriminante é adequada, por minimizar os erros de classificação. Portanto, a satisfação da premissa de normalidade multivariada é um dos fatores que asseguram que a Análise Discriminante apresente resultados satisfatórios

Já a Regressão Logística é mais utilizada por ser mais flexível, não impondo às variáveis independentes condições como homogeneidade de matrizes de covariância entre os grupos de classificação e normalidade multivariada. Outro fator que contribui para sua preferência é que a interpretação de seus resultados é realizada com base em valores de probabilidade do evento estudado ocorrer, diferente da Análise Discriminante, cuja interpretação é menos intuitiva, uma vez que apresenta seus resultados em *scores*. Tais *scores* são basicamente um dispositivo de discriminação de classificação ordinal, sem nenhum aspecto probabilístico embutido (Ohlson, 1980 apud Castro Jr, 2003). E, por fim, assim como a Regressão Múltipla, a Regressão Logística é capaz de incorporar efeitos não lineares (Hair, 1998).

A capacidade de interpretação dos coeficientes de forma direta, de forma que possa ser estendida para qualquer problema prático, é uma das maiores vantagens da Regressão Logística (Paula, 2004). Nesse sentido, a técnica se tornou um método padrão de análise de regressão de variáveis dependentes dicotômicas, especialmente na área de ciência da saúde. Ainda que o evento de interesse não seja binário, muitas

pesquisas tem adotado a dicotomização de variáveis para que a probabilidade de ocorrência possa ser calculada por meio da Regressão Logística (Paula, 2004).

O modelo logístico de discriminação acaba por ser um método mais abrangente, que funciona quase tão bem quanto a Análise Discriminante mesmo quando esta tem seus pressupostos satisfeitos, de forma que é consenso que, sendo as distribuições não normais, a Regressão Logística deve ser preferida (Press & Wilson, 1978; Krzanowski, 1988; McLachlan, 1992).

5 METODOLOGIA

A metodologia utilizada para identificar os Assentamentos Precários terá como base o método concebido pelo Centro de Estudos da Metrópole (CEM). O método consiste em traçar o perfil socioeconômico da população moradora de Aglomerados Subnormais utilizando variáveis do Censo Demográfico e utilizá-lo como base para, por meio de modelagem estatística, detectar setores censitários com perfis semelhantes, mapeados utilizando não só critérios estruturais, mas de perfil populacional. Desse modo, são identificados não só os setores censitários correspondentes a favelas, como é o caso da classificação do IBGE, que se baseia em critérios estruturais, uma vez que, valendo-se do perfil populacional dos habitantes, amplia-se o espectro de detecção e engloba-se as demais tipologias de Assentamentos Precários.

O método-base será adaptado e aplicado para a área urbana do Distrito Federal com os dados do Censo Demográfico de 2010, testando três técnicas distintas de estatística multivariada, a fim de avaliar qual apresentará maior acurácia na estimativa de detecção de Assentamentos Precários. Assim, a metodologia do presente estudo será dividida em quatro fases, as quais serão detalhadas neste tópico: Fase 1, que consistirá na realização de adaptações ao método-base; Fase 2, que se refere à aquisição e preparação dos dados para a aplicação do método definido na etapa anterior; Fase 3, de aplicação das técnicas estatísticas em ambiente de programação, utilizando o Software R; e, por fim, a Fase 4, de espacialização dos resultados e cálculo do Indicador de Universalização Inclusiva no DF (Figura 8).

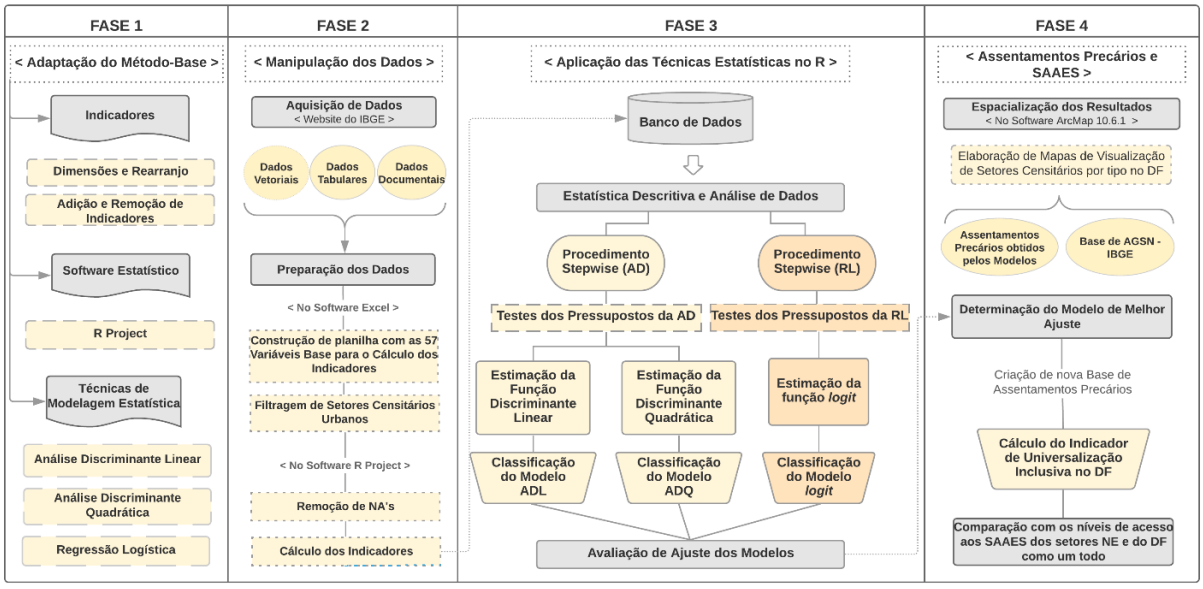

Figura 8 - Fluxorama Metodológico do Estudo

Finalmente, destaca-se que, embora o objetivo do estudo envolva, de fato, atualizar a estimativa dos Assentamentos Precários atualizada para 2010 (ano do último dado disponível), o intuito não é apenas adaptar o método desenvolvido pelo CEM para que seja aplicado às variáveis do Censo Demográfico de 2010, como orientam Silva et al. (2014); mas, na verdade, se valer de todo o extenso arcabouço teórico-prático trazido pelo estudo do CEM para obter, como resultado, um procedimento metodológico consolidado, que:

a) amplie o alcance do IBGE e padronize a detecção de Assentamentos Precários;

b) utilize dados acessíveis, considerando a grande dificuldade da construção de estimativas confiáveis, padronizadas e viáveis financeira e operacionalmente no mapeamento dessas áreas (Silva et al., 2014), tanto nacional quanto globalmente;

c) possa ser aplicado em qualquer área de estudo selecionada, desde que compreendida pelo território nacional inserido no levantamento de dados do Censo Demográfico.

Dessa forma, a metodologia do presente estudo foi conduzida no sentido de promover o aprimoramento da estimativa da identificação de Assentamentos Precários resultante do estudo do CEM (2007), para torná-la mais completa e acurada.

5.1 FASE 1: ADAPTAÇÃO AO PROCEDIMENTO BASE PROPOSTO PELO CEM (2007) PARA DETECÇÃO DE ASSENTAMENTOS PRECÁRIOS

A primeira fase da metodologia consistiu na realização de adaptações, acréscimos e reorganizações realizadas no método-base. Foram realizadas quatro principais modificações no método base, a saber: a) acréscimo, remoção e rearranjo de indicadores; b) adição de duas técnicas estatísticas de modelagem para discriminação e classificação dos setores censitários; c) alteração do software de aplicação do método; e, por fim, d) modificação da entidade federativa de aplicação do método. Assim, neste tópico, serão detalhadas as quatro principais alterações realizadas no método-base.

Quanto às alterações realizadas nos indicadores, tem-se que: dos dezenove indicadores propostos pelo método-base, quatro foram excluídos, por não serem aplicáveis ao Censo de 2010 e dois foram acrescidos, com o objetivo de tornar a análise mais completa. A não aplicabilidade dos indicadores excluídos se refere ao fato de que, de um Censo Demográfico para outro, é comum que haja alterações no questionário aplicado ao universo populacional, modificando assim a forma de divulgação das

informações (variáveis); e isso ocorreu com o Censo de 2010, que apresentou algumas mudanças em relação ao Censo de 2000, na forma de variáveis que foram eliminadas e/ou reformuladas. Por conta disso, alguns indicadores propostos inicialmente e aplicados ao Censo de 2000 pelo método do CEM (2007) não teriam como ser calculados com os dados do Censo de 2010. Esse foi o caso dos quatro indicadores listados abaixo excluídos no Censo 2010:

- "Anos médios de Estudo do Responsável pelo Domicílio"
- "Porcentagem de Responsáveis por Domicílio com menos de 8 Anos de Estudo"
- "Domicílios com Outro tipo de posse do Terreno".
- "Domicílios do Tipo Cômodo"

Silva et al. (2014), no documento de sistematização publicado pelo CEM contendo orientações para a replicação do método com os dados de 2010 chegaram a apontar a não aplicabilidade dos três primeiros indicadores acima, uma vez que, em 2010, não foram obtidas informações sobre anos de estudo ou sobre tipo de posse do terreno. No entanto, verificou-se que isso também ocorre com os domicílios tipo cômodo, que não foram contabilizados no último censo. Diante disso, os quatro indicadores acima foram removidos.

Após a remoção desses indicadores, dois novos foram inseridos, relativos à gênero e raça, fatores importantes na análise do perfil socioeconômico da população mais vulnerável e moradora de Assentamentos Precários. São eles:

- "Percentual de Pessoas Responsáveis do Sexo Feminino";
- "Percentual de Pessoas Residentes Brancas".

Acredita-se que o indicador referente aos domicílios chefiados por mulheres enriqueceria o estudo devido ao fenômeno de feminização da pobreza, que vem sendo identificado no Brasil e consiste no aumento da pobreza entre as mulheres relativamente aos homens. Ao pesquisar esse fenômeno utilizando dados do PNAD entre 2004 e 2015, Carmo (2019) identificou que o recorte de raça foi mais determinante que o gênero na

sua ocorrência; motivo pelo qual optou-se por também adicionar à análise o percentual de pessoas brancas residentes no setor.

Por fim, a nova lista totalizou dezessete indicadores, os quais foram reorganizados em três aspectos: (A) Demográficos, (B) Sociais e (C) Estruturais e numerados de acordo com sua categoria para facilitar a representação (Tabela 3).

Tabela 3 - Indicadores do perfil socioeconômico da população de Assentamentos Precários.

Indicadores (calculados por setor)	
A) Aspectos demográficos (ocupação e habitação)	A.1) Número de Domicílios Particulares Permanentes (DPP) A.2) Número de Moradores em DPP A.3) Número Médio de Moradores em DPP A.4) Percentual de Pessoas Brancas A.5) Percentual de Pessoas Responsáveis (PR) do Sexo Feminino A.6) Número de Domicílios Improvisados A.7) Percentual de DPP com outra forma de ocupação (diferente de comprado/alugado/cedido)
B) Aspectos sociais (renda e alfabetização)	B.1) Porcentagem de PR de até 30 anos B.2) Porcentagem de PR não alfabetizadas B.3) Porcentagem de PR de até 30 anos não alfabetizadas B.4) Porcentagem de PR por DPP com renda até 3 Salários Mínimos B.5) Renda média das PR por DPP
C) Aspectos estruturais (saneamento)	C.1) Porcentagem de DPP sem coleta de lixo C.2) Porcentagem de DPP sem ligação à rede de Abastecimento de Água C.3) Porcentagem de DPP sem Banheiros ou Sanitário C.4) Porcentagem de DPP sem ligação à Rede de Esgoto ou Fossa Séptica C.5) Número Médio de Banheiros por Habitante de DPP

No que se referem às técnicas estatísticas utilizadas, foram adicionadas duas técnicas estatísticas multivariadas, também úteis à discriminação e classificação de dados binários: a Análise Discriminante Quadrática (ADQ) e a Regressão Logística (RL), além da Análise Discriminante Linear (ADL), utilizada no método-base. Desse modo, foram testados três modelos estatísticos e obtiveram-se três resultados, a fim de avaliar as vantagens e desvantagens de cada um e avaliar qual teria melhor desempenho.

Com relação ao software de aplicação, no método-base, o modelo foi desenvolvido com uso do pacote estatístico SPSS, enquanto no presente estudo, foi utilizado o software R, livre e gratuito, onde foram elaborados scripts para o desenvolvimento do método e construção dos modelos, permitindo maior liberdade nas modificações e na partilha do conhecimento, facilitando reaplicações.

E, por fim, o método-base trabalhou com os limites geográficos das Regiões Metropolitanas brasileiras, gerando estimativas de setores precários para todo o país. No

caso do DF, uniu-o com a RM de Goiânia para fazer a análise; enquanto a presente pesquisa selecionou apenas o Distrito Federal para aplicar e avaliar o método desenvolvido.

5.2 FASE 2: MANIPULAÇÃO DOS DADOS E CONSTRUÇÃO DO BANCO DE DADOS

Tendo descrito as principais alterações realizadas no método-base, pode-se passar às etapas detalhadas da Fase 2, que consistiu na manipulação dos dados para garantir a aplicação das técnicas de modelagem estatística. A Fase 2, portanto, foi subdividida em três etapas, descritas na Figura 9: aquisição de dados, sua organização no software Excel e, por fim, a construção do banco de dados, já no software R, onde o método foi efetivamente desenvolvido e aplicado.

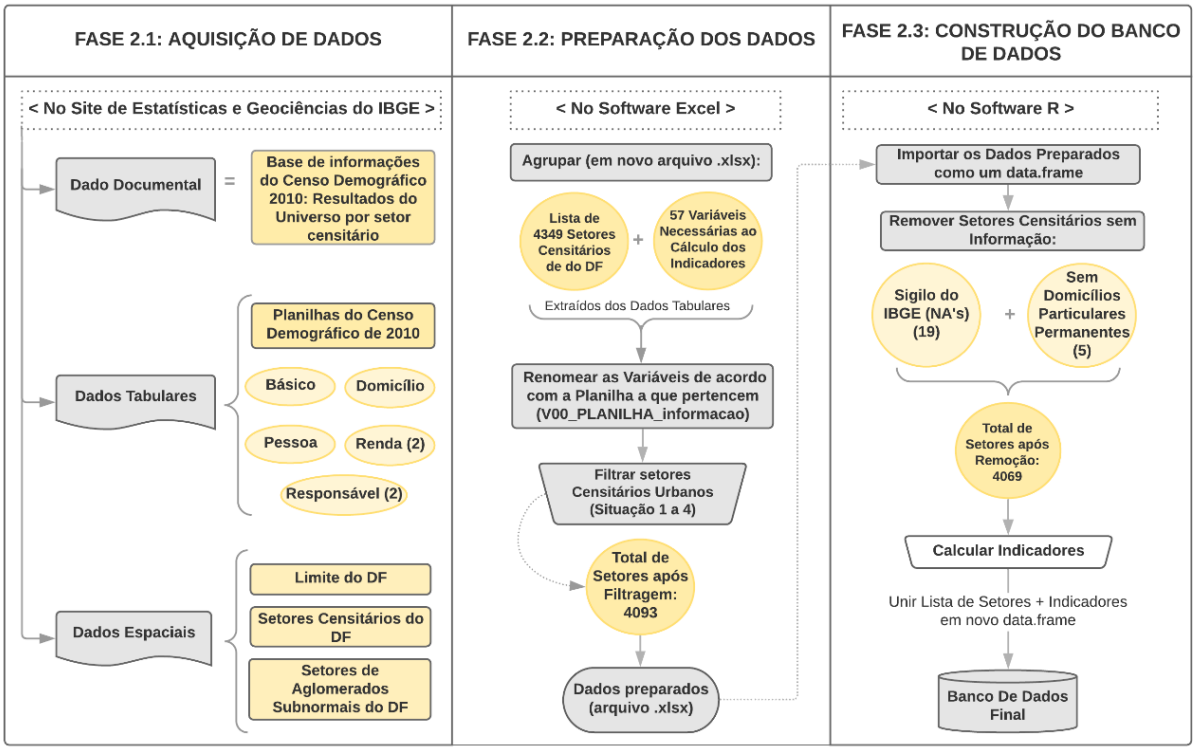

Figura 9 - Fluxograma de Detalhamento da Fase 2 referente à montagem do banco de dados para detecção de Assentamentos Precários no DF.

5.2.1 Aquisição dos dados

Os dados foram baixados do site de Estatística e Geociências do IBGE, onde estão disponíveis em três formas: textual, representado pela “Base de informações do Censo

Demográfico 2010: Resultados do Universo por setor censitário", documento explicativo do IBGE que contém a descrição detalhada da base de dados; tabular, na forma de planilhas que contêm as variáveis coletadas por setor censitário, divididas por categoria de informação e por unidade federativa, que contêm todas as variáveis coletadas por setor censitário; e, por fim, espacial, que corresponde aos arquivos que contém os limites geográficos das unidades federativas e setores censitários.

Quanto aos dados tabulares, são disponibilizadas vinte e seis planilhas por unidade federativa, distribuídas em seis categorias, de acordo com o tipo de informação que contêm: Básico (1), Domicílio (2), Responsável (2), Renda (3), Pessoa (13) e Entorno (5). No entanto, com o auxílio do documento explicativo do IBGE (dado textual), foi possível identificar que as variáveis necessárias ao cálculo dos indicadores estavam disponíveis em sete das vinte e seis planilhas disponibilizadas para o Distrito Federal: Básico (1), Domicílio (2), Responsável (1), Renda (2), Pessoa (1). Assim, foi realizado o *download* do documento textual, de sete planilhas e dos limites geográficos do DF, de seus setores censitários e de seus aglomerados subnormais.

Os dados vetoriais foram organizados em um mapa de localização da área de estudo com os setores censitários urbanos do DF, por tipo, de acordo com a classificação realizada pelo IBGE (Figura 10).

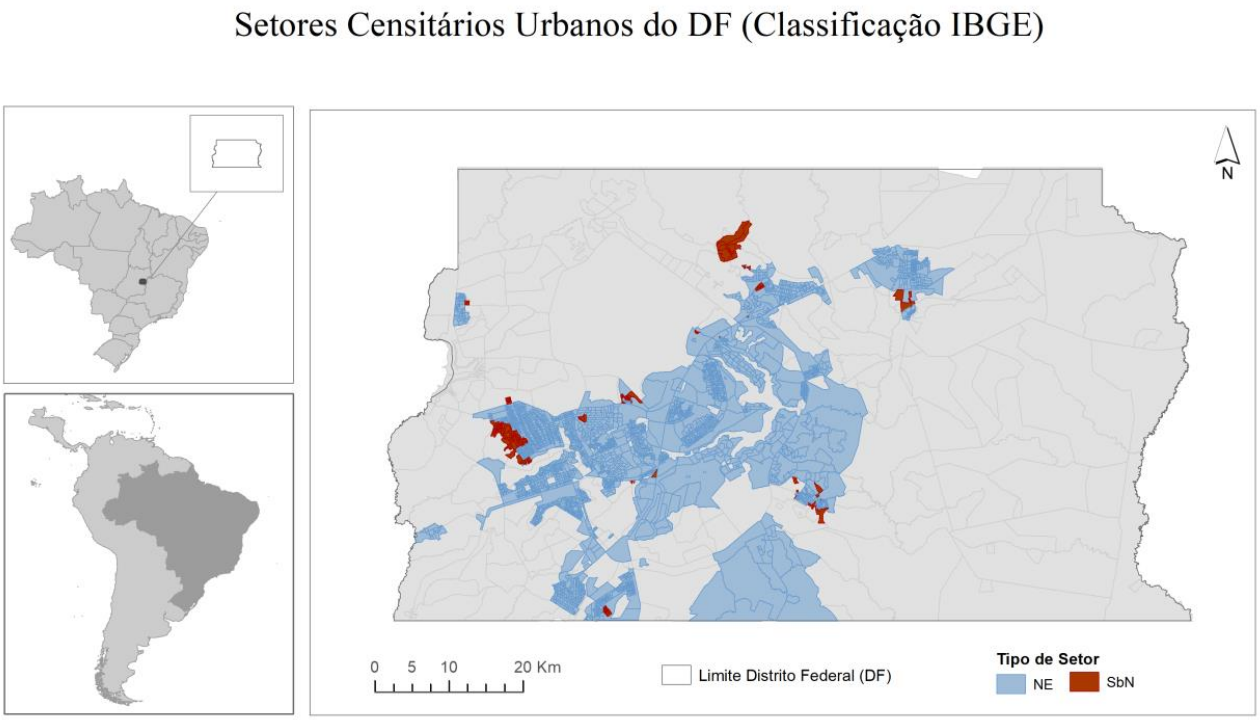

Figura 10 - Mapa de localização do DF e Setores censitários urbanos por tipo. Fonte: IBGE, 2011.

Esta etapa consistiu na organização das informações retiradas dos dados tabulares, os quais consistem nas sete planilhas listadas no Quadro 1. Foram extraídas, no total, 57 variáveis do total de 4.349 setores censitários do DF, as quais foram agrupadas em uma única planilha no software Excel. Em seguida, as variáveis foram renomeadas para que fosse possível identificar de que planilha foram retiradas e a informação que contêm, com o intuito de facilitar a montagem das fórmulas para o cálculo dos indicadores.

Quadro 1 - Planilhas do Censo Demográfico de 2010 utilizadas para a Coleta das Variáveis.

Categoria da Informação	**Título da Planilha**	**Nome do Arquivo (IBGE)**	**Código da Planilha (utilizado no estudo)**
Básico	Básico	"Básico_DF"	Bas
Domicílio	Domicílio, características gerais	"Domicílio01_DF"	D01
Responsável	Responsável pelo Domicílio, mulheres	"Responsável01_DF"	R01
	Responsável pelo domicílio, total e homens	"Responsável02_DF"	R02
Renda	Renda da Pessoa Responsável	"ResponsávelRenda_DF"	RR
	Renda dos Domicílios	"DomicílioRenda_DF"	DR
Pessoa	Cor ou Raça, idade e gênero	"Pessoa03_DF"	P03

Assim, o novo nome dado a cada variável consistiu na sua numeração original na planilha de onde foi retirada (V00); no código que designa a planilha da qual foi extraída a variável, apresentado na última coluna do Quadro 1 e criado com base no nome dado ao arquivo pelo IBGE; e, por fim, de abreviação que denotasse seu significado (Quadro 2).

Quadro 2 - Variáveis Necessárias ao Cálculo dos Indicadores e suas Denominações.

Nome da Variável	**Numeração da Variável na Planilha de Origem**	**Código da Planilha de Origem**	**Abreviação da Variável**
Código do Setor Censitário	Codigo Setor	-	Cod_Setor
Situação do Setor Censitário (1 a 4)	Situação Setor	-	Situacao_setor
Tipo do Setor Censitário (0 ou 1)	Tipo Setor	-	Tipo_setor
Número de DPPs	V001	Bas	V001_Bas_Num_DPP

Número de Domicílios Improvisados	V001	DR	V001_DR_Num_Dom_Imp
Número de Moradores em DPP	V002	Bas	V002_Bas_Num_Mor_em_DPP
Número de Pessoas Residentes	V001	P03	V001_P03_Pessoas_Residentes
Número de Pessoas Residentes Brancas	V002	P03	V002_P03_Pes_Res_brancas
Número de Moradores em DPP com outra condição de ocupação	V011	D01	V011_D01_Mor_DPP_outra_cond_ocup
Número de Pessoas Responsáveis	V001	R02	V001_R02_Pes_Resp
Número de Pessoas Responsáveis do Sexo feminino	V001	R01	V001_R01_PR_sexo_fem
Número de Pessoas Responsáveis com 10 anos	V002	R02	V002_R02_PR_com_10
Número de Pessoas Responsáveis com 11 anos	V003	R02	V003_R02_PR_com_11
Número de Pessoas Responsáveis com 12 anos	V004	R02	V004_R02_PR_com_12
Número de Pessoas Responsáveis com 13 anos	V005	R02	V005_R02_PR_com_13
Número de Pessoas Responsáveis com 14 anos	V006	R02	V006_R02_PR_com_14
Número de Pessoas Responsáveis com 15 anos	V007	R02	V007_R02_PR_com_15
Número de Pessoas Responsáveis com 16 anos	V008	R02	V008_R02_PR_com_16
Número de Pessoas Responsáveis com 17 anos	V009	R02	V009_R02_PR_com_17
Número de Pessoas Responsáveis com 18 anos	V010	R02	V010_R02_PR_com_18
Número de Pessoas Responsáveis com 19 anos	V011	R02	V011_R02_PR_com_19
Número de Pessoas Responsáveis com 20 anos	V012	R02	V012_R02_PR_com_20
Número de Pessoas Responsáveis com 21 anos	V013	R02	V013_R02_PR_com_21
Número de Pessoas Responsáveis com 22 anos	V014	R02	V014_R02_PR_com_22
Número de Pessoas Responsáveis com 23 anos	V015	R02	V015_R02_PR_com_23
Número de Pessoas Responsáveis com 24 anos	V016	R02	V016_R02_PR_com_24
Número de Pessoas Responsáveis com 25 anos	V017	R02	V017_R02_PR_com_25
Número de Pessoas Responsáveis	V018	R02	V018_R02_PR_com_26

com 26 anos			
Número de Pessoas Responsáveis com 27 anos	V019	R02	V019_R02_PR_com_27
Número de Pessoas Responsáveis com 28 anos	V020	R02	V020_R02_PR_com_28
Número de Pessoas Responsáveis com 29 anos	V021	R02	V021_R02_PR_com_29
Número de Pessoas Responsáveis alfabetizadas	V093	R02	V093_R02_PR_alf
Número de Pessoas Responsáveis alfabetizadas de 10 a 14 anos	V094	R02	V094_R02_PR_alf_10_a_14
Número de Pessoas Responsáveis alfabetizadas de 15 a 19 anos	V095	R02	V095_R02_PR_alf_15_a_19
Número de Pessoas Responsáveis alfabetizadas de 20 a 24 anos	V096	R02	V096_R02_PR_alf_20_a_24
Número de Pessoas Responsáveis alfabetizadas de 25 a 29 anos	V097	R02	V097_R02_PR_alf_25_a_29
Número de Pessoas Responsáveis com Renda até meio Salário Mínimo	V067	RR	V067_RR_PR_DPP_Rend_ate_1-2_SM
Número de Pessoas Responsáveis com Renda entre meio e um Salário Mínimo	V068	RR	V068_RR_PR_DPP_Rend_1-2_a_1_SM
Número de Pessoas Responsáveis com Renda entre um e dois Salários Mínimos	V069	RR	V069_RR_PR_DPP_Rend_1_a_2_SM
Número de Pessoas Responsáveis com Renda entre dois e três Salários Mínimos	V070	RR	V070_RR_PR_DPP_Rend_2_a_3_SM
Número de Pessoas Responsáveis sem Renda	V076	RR	V076_RR_PR_DPP_Sem_Rend
Renda média das Pessoas Responsáveis por DPP	V005	Bas	V005_Bas_Renda_Med_PR_DPP
Número de DPPs com Coleta de Lixo	V035	D01	V035_D01_DPP_com_col_lixo
Número de DPPs com Abastecimento de Água	V012	D01	V012_D01_DPP_com_ab_agua
Número de DPPs sem Banheiro ou Sanitário	V023	D01	V023_D01_DPP_sem_ban_san
Número de DPPs com Banheiro ou Sanitário	V016	D01	V016_D01_DPP_com_ban_san
Número de DPPs com Banheiro ligado à Rede de Esgoto	V017	D01	V017_D01_DPP_com_ban_esg_red
Número de DPPs com Banheiro ligado à Fossa Séptica	V018	D01	V018_D01_DPP_com_ban_esg_fos
Número de DPPs com Banheiro	V024	D01	V024_D01_DPP_com_ban

Número de DPPs com 1 Banheiro	V025	D01	V025_D01_DPPcom_1_ban
Número de DPPs com 2 Banheiros	V026	D01	V026_D01_DPP_com_2_ban
Número de DPPs com 3 Banheiros	V027	D01	V027_D01_DPP_com_3_ban
Número de DPPs com 4 Banheiros	V028	D01	V028_D01_DPP_com_4_ban
Número de DPPs com 5 Banheiros	V029	D01	V029_D01_DPP_com_5_ban
Número de DPPs com 6 Banheiros	V030	D01	V030_D01_DPP_com_6_ban
Número de DPPs com 7 Banheiros	V031	D01	V031_D01_DPP_com_7_ban
Número de DPPs com 8 Banheiros	V032	D01	V032_D01_DPP_com_8_ban
Número de DPPs com 9 Banheiros	V033	D01	V033_D01_DPP_com_9_ban

As três primeiras variáveis presentes no Quadro 2 são qualitativas e não apresentam planilha de origem, pois estão presentes em todas. São elas: o código do setor, que corresponde ao identificador numérico do setor censitário; a situação do setor, que se refere ao seu caráter urbano ou rural, expressa por números de 1 a 8 (Quadro 3); e, por fim, o tipo de setor, que indica se o setor é Não Especial ou SubNormal (0 ou 1) e corresponde à principal variável para o presente estudo.

Quadro 3 - Situação dos Setores Censitários.

Situação do Setor	
Situação urbana – códigos: 1, 2 ou 3	1 - Área urbanizada de cidade ou vila 2 - Área não-urbanizada de cidade ou vila 3 - Área urbana isolada
Situação rural – códigos: 4, 5, 6, 7 ou 8	4 - Aglomerado rural de extensão urbana 5 - Aglomerado rural isolado - povoado 6 - Aglomerado rural isolado - núcleo 7 - Aglomerado rural isolado - outros aglomerados 8 - Zona rural, exclusive aglomerado rural

Fonte: IBGE, 2011.

A situação do setor foi importante para a última etapa da preparação dos dados, uma vez que, no presente estudo, não há interesse nos setores rurais já que os Assentamentos Precários são definidos como um fenômeno de natureza urbana. Dessa forma, foram filtrados os setores que apresentaram números de 1 a 4 – sendo incluídos também os setores com aglomerados rurais de extensão urbana, para garantir que fossem abrangidas todas as áreas de influência urbana presentes nos entornos da cidade (Quadro 3). Após a filtragem, os dados preparados resultaram em uma planilha final constituída por 4.094 setores censitários.

5.2.2 Cálculo dos Indicadores

A planilha dos dados preparados foi importada ao software R, onde o primeiro comando aplicado foi para a remoção de setores censitários que apresentassem algum campo sem informação. No R, essas células vazias são denominadas *NA's* (*not availabe*). Foram então removidos 24 setores censitários no total, dos quais: 19 representavam sigilo do IBGE, procedimento comum que consiste em ocultar dados de alguns setores para garantir a anonimidade; e 5 representavam setores sem domicílios, os quais não permitiriam o cálculo dos indicadores que possuíam o total de domicílios no denominador. Assim, após a remoção desses 24 setores, restaram 4.069 setores censitários, para os quais foram calculados os 17 indicadores por meio das fórmulas apresentadas na Tabela 4.

O Banco de Dados utilizado como base para a realização de todas as análises seguintes consiste em uma planilha contendo todos os elementos da amostra – chamados de observações – e representados pelos setores censitários urbanos do DF. Assim, na primeira coluna, têm-se os códigos do setor censitário; na segunda, o grupo ao qual pertencem (NE ou SbN), indicado por 0 ou 1, respectivamente, que corresponde à variável dependente ou categórica; e, nas colunas seguintes, os dezessete indicadores, que correspondem às variáveis independentes ou explicativas da análise.

Para a construção desse banco, foi necessário, inicialmente, importar a planilha compilada com os dados do IBGE, contendo todas as variáveis do censo que seriam necessárias aos cálculos dos indicadores. Tal planilha foi importada à plataforma R Studio contendo 4069 setores censitários, como mencionado anteriormente.

Tabela 4 - Fórmulas de Cálculos dos Indicadores.

	Indicador (por Setor Censitário)	Abreviação do Indicador no Banco de Dados	Cálculo
A.1	Número de DPP	i_num_dpp	$V001_{Bas}$
A.2	Número de Moradores em DPP	i_num_mor_dpp	$V002_{Bas}$
A.3	Número Médio de Moradores por DPP	i_num_med_mor_por_dpp	$\frac{V002_{Bas}}{V001_{Bas}}$
A.4	Percentual de Pessoas Brancas	i_perc_pes_res_brancas	$\frac{V002_{P03}}{V001_{P03}} * 100$
A.5	Percentual de PR do Sexo Feminino	i_perc_PR_sexo_fem	$\frac{V001_{R01}}{V001_{R02}} * 100$
A.6	Número de Domicílios Improvisados	i_num_dom_imp	$V001_{DR}$
A.7	Percentual de DPP com outra forma de ocupação (não comprado/alugado/cedido)	i_perc_dpp_outra_ocup	$\frac{V011_{D01}}{V002_{Bas}} * 100$
B.1	Porcentagem de PR de até 30 anos	i_perc_pr_ate_30	$\frac{V002_{R02} + V003_{R02} + \cdots + V021_{R02}}{V001_{R02}} * 100$
B.2	Porcentagem de PR não alfabetizadas	i_perc_PR_nalf	$1 - \frac{V093_{R02}}{V001_{R02}} * 100$

B.3	Porcentagem de PR de até 30 anos não alfabetizadas	i_perc_PR_ate_30_nalf	$\frac{(V002_{R02} + V003_{R02} + \cdots + V021_{R02}) - (V094_{R02} + \cdots + V097_{R02})}{V001_{R02}} * 100$
B.4	Porcentagem de PR por DPP com renda até 3 Salários Mínimos	i_perc_pr_ren_ate_3_SM	$\frac{V067_{RR} + V068_{RR} + \cdots + V070_{RR} + V076_{RR}}{V001_{Bas}} * 100$
B.5	Renda média das PR por DPP	i_Renda_Med_PR_DPP	$V005_{Bas}$
C.1	Porcentagem de DPP sem coleta de lixo	i_perc_dpp_sem_col_lixo	$1 - \frac{V035_{D01}}{V001_{Bas}} * 100$
C.2	Porcentagem de DPP sem ligação à rede de Abastecimento de Água	i_perc_DPP_sem_ab_agua	$1 - \frac{V012_{D01}}{V001_{Bas}} * 100$
C.3	Porcentagem de DPP sem Banheiros ou Sanitário	i_perc_DPP_sem_ban_san	$\frac{V023_{D01}}{V001_{Bas}} * 100$
C.4	Porcentagem de DPP sem ligação à Rede de Esgoto ou Fossa Séptica	i_perc_DPP_sem_esg_fos	$1 - \frac{(V017_{D01} + V018_{D01})}{V016_{D01}} * 100$
C.5	Número Médio de Banheiros por Habitante de DPP	i_num_med_ban_hab	$\frac{(V025_{D01} * 1 + \cdots + V026_{D01} * 2 + V33_{D01} * 9)}{V024_{D01}}$

Os resultados do cálculo dos indicadores foram unidos em um *data frame*, formando finalmente o Banco de Dados do presente estudo, contendo as 4.069 observações (setores censitários), os grupos a que pertencem (tipo de setor - NE ou SbN) e as 17 respectivas variáveis explicativas (indicadores listados na Figura 7). Essa foi a base de dados utilizada para todas as análises posteriores.

5.3 FASE 3: APLICAÇÃO DOS MODELOS ESTATÍSTICOS PARA IDENTIFICAÇÃO DE ASSENTAMENTOS PRECÁRIOS NO DF

A aplicação das técnicas de modelagem estatística para detecção dos Assentamentos Precários foi realizada conforme descrito na Fase 3 do fluxograma ilustrado na Figura 8. As etapas serão detalhadas a seguir, na ordem em que aparecem no fluxograma, juntamente com os pacotes e respectivos comandos utilizados para a sua execução, os quais estarão sumarizados em quadros explicativos (Quadro 4 a 8).

5.3.1 Estatística Descritiva e Análise Prévia dos dados

A estatística descritiva consistiu no cálculo da Média, Mediana, Desvio Padrão, Valor Mínimo e Valor Máximo, por meio do comando *sumarize* do pacote *stats*, próprio do software R. Em seguida, os valores foram organizados em quadros e analisados separadamente por tipo de setor, com o intuito de analisar o comportamento geral dos indicadores por grupo e identificar diferenças entre os perfis populacionais.

5.3.2 Procedimentos Stepwise (AD e RL)

Em seguida, foi realizada a seleção de variáveis independentes pelo procedimento *Stepwise*, com o objetivo de escolher as mais significantes para serem incluídas nos modelos, por meio dos critérios Lambda de Wilks para a AD e Critério de Informação de Akaike (AIC - *Akaike Information Criterion*) para a RL (Quadro 4).

O processo consiste em adicionar e/ou remover as variáveis independentes, uma a uma, à medida que calcula o valor do critério, interrompendo o processo quando o melhor ajuste é encontrado. No caso dos critérios utilizados, quanto menor o seu valor, melhor o ajuste do modelo. Os tipos de procedimento realizados foram o *forward* para AD e o *both* para a RL, os quais se distinguem pelo primeiro efetuar, a cada etapa, apenas a inclusão de variáveis, enquanto o segundo realiza tanto a inclusão quanto a exclusão, simultaneamente, de modo que o processo de iterações é encerrado, nesse caso, quando os modelos atingem o menor valor de critério dentre todas as combinações possíveis.

Quadro 4 - Método de Aplicação do Procedimento Stepwise para as técnicas da AD e RL.

Técnica	Procedimento	Método	Comando	Pacote
AD	Stepwise – método *forward*	Lambda de Wilks	*greedy.wilks()*, tipo *forward*	klaR
RL	Stepwise – método *both*	AIC (*Akaike Information Criterion*)	*stepAIC()*, tipo *both*	MASS

5.3.3 Teste de Pressupostos

Os pressupostos da AD de ausência de multicolinearidade, normalidade multivariada e, no caso específico da ADL, homogeneidade das matrizes de covariâncias foram avaliados por meio da correlação de Pearson e dos testes de Mardia, Henze-Zirkler e M de Box, respectivamente (Quadro 5). Quanto aos pressupostos da RL, a linearidade dos indicadores em relação ao termo *logit* e a ausência de *outliers* foram avaliadas por análise gráfica, utilizando a distância de *Cooks* para a definição de valores extremos.

Já a ausência de multicolinearidade, embora já tenha sido avaliada pela AD, requere método distinto para a RL. Foi utilizado, portanto, o cálculo do VIF (*Variance Inflation Factor)*, o qual deve apresentar valores abaixo de 10 para garantir que a premissa seja atendida. Por fim, optou-se por incluir testes de superdispersão e heterocedasticidade, casos nos quais os modelos de RL apresentam grandes riscos de viés, o que poderia reduzir a confiabilidade e/ou inviabilizar a predição.

Quadro 5 - Método de Teste dos Pressupostos das técnicas das AD e RL.

Técnica	Pressupostos	Método	Comando	Pacote
AD	Ausência de Multicolinearidade	Matriz de Correlação de *Pearson*	*rcorr(), type* = "*Pearson*"	Hmisc
	Normalidade Multivariada	Testes de Mardia e de Henze-Zirkler	*mvn()*	MVN
	Homogeneidade das Matrizes de Covariância*	Teste M de Box	*boxM()*	biotools
RL	Relação linear entre o termo logit e as variáveis independentes	Plotagem de gráficos termo logit vs. indicador	*ggplot()*	ggplot2
	Ausência de Outliers	Distância de Cooks	*augment()* *top_n()* *filter()*	broom dplyr
	Ausência de Multicolinearidade	Cálculo do VIF	*vif()*	car
	Ausência de Heterocedasticidade	Teste de Heterocedasticidade	*simulateResiduals()*	DHARMa
	Ausência de Superdispersão	Estatística Qui Quadrado/ Grau de Liberdade do Resíduo	*residuals()*	stats

5.3.4 Construção das Funções

Testados os pressupostos e selecionados os indicadores, pode-se passar à construção das funções propriamente ditas. As funções discriminantes foram criadas por meio do pacote MASS, utilizando os indicadores selecionados pelo procedimento Stepwise. Já a função logística foi formada pela construção de um modelo linear generalizado, definindo o tipo do modelo como *logit* e a família como binomial.

Foi gerado, também, um modelo logístico ajustado com uma subamostra, apenas para fins de análise, não tendo sido utilizado para a predição das classes. O intuito foi comparar seus coeficientes com o modelo gerado tradicionalmente e, assim, analisar a presença discrepâncias muito significativas que justificassem um possível enviesamento por desbalanceamento da amostra. A subamostragem seguiu o estudo de King e Zelig (2001), que obtiveram ausência total de viés nos pontos onde foram descartados aproximadamente 30%, 55% e 78% zeros de uma amostra original com 1000 observações. Como os autores sugerem alcançar uma subamostra que varie no máximo de duas a cinco vezes mais zeros do que uns, optou-se por realizar uma subamostragem que resultasse na última opção, ou seja, com uma proporção de 875 setores NE para os 175 setores SbN, com o objetivo de descartar o menor número possível de observações.

Dessa forma, 77,5% dos zeros da amostra do estudo foram descartados totalizando uma amostra de 1050 observações, valores próximos daqueles que resultaram em ausência de viés nos achados de King e Zeng (2001). O pacote utilizado para a geração da subamostra e para a geração do modelo ajustado estão descritos no Quadro 6, com as demais informações dessa etapa no software R Project.

Quadro 6 - Métodos Utilizados para a Construção das Funções dos Modelos.

Técnica	**Construção da Função**	**Método**	**Comando**	**Pacote**
ADL	Função Discriminante Linear	Método dos Momentos (*Sem Cross Validation*)	*lda()*	MASS
	Carga Discriminante	Matriz de Estrutura	*statsBy()*	psych
ADQ	Função Discriminante Quadrática	Método dos Momentos (*Sem Cross Validation*)	*qda()*	MASS
RL	Função Logística	Família Binomial, link = *logit*	*glm()*	stats
RL Ajustada	Rebalanceamento da Amostra	Subamostragem Aleatória	*ubUnder()*	unbalanced
	Função Logística Ajustada	Correção a Priori	*relogit()*	Zelig

5.3.5 Predição dos Modelos AD e RL

As predições dos tipos de setor pelos modelos desenvolvidos foram todas realizadas pelo comando *predict()*, próprio do software R. Tal comando se baseia funções previamente construídas para realizar as predições, sendo que, no caso da AD, parte das probabilidades a priori, no caso da RL, delimita-se uma probabilidade mínima como ponto de corte (*threshold)* a partir do qual um setor deve ser alocado no grupo 1, ou seja, classificado como SbN. Tanto a probabilidade a priori de um setor ser SbN quanto o ponto de corte da RL consistiram na proporção de observações do grupo 1 encontrada na amostra, ou seja, de 0,043 (175 setores SbN em uma população de 4069 setores).

Quadro 7 - Métodos de Predição para os Modelos.

Técnica	Método	Comando	Pacote
ADL	*Prior* = 0.04300811	*predict()*	stats
ADQ			
RL	Se probabilidade > 0.04300811, alocar em grupo 1, caso contrário, 0.	*predict()* *ifelse()*	

5.3.6 Avaliação do Ajuste dos Modelos

Para avaliar o ajuste dos modelos da AD, foi utilizada a sensitividade, que representa a taxa de acerto na classificação de setores SbN e o valor de AUC (*Area Under the Curve)*, onde se avalia o desempenho preditivo geral dos modelos. Como para a RL há uma maior variedade de testes de ajuste, foram aplicados, complementarmente aos anteriores, o Teste de Hosmer-Lemeshow e o Teste de Pearson Qui Quadrado.

Quadro 8 - Critérios de Avaliação do Ajuste dos Modelos.

Técnica	Ajuste do Modelo	Método	Comando	Pacote
AD	Sensitividade (Taxa de Acerto de Setores SbN)	Matriz de Confusão	*confusionMatrix()*	caret
	AUC (*Area Under the Curve)*	Curva ROC	*roc (), plot()*	Graphics
RL	Sensitividade (Taxa de Acerto de Setores SbN)	Matriz de Confusão	*table()*	Base R
	AUC (*Area Under the Curve)*	Curva ROC	*plot() print.auc=TRUE, auc.polygon=TRUE*	Graphics
	Qualidade do Modelo na Explicação do fenômeno	Teste de *Hosmer-Lemeshow*	*hoslem.test*	ResourceSelection
	Qualidade de Ajuste do Modelo	Teste de *Pearson* Qui Quadrado	*pchisq*	stats

5.4 FASE 4: ESPACIALIZAÇÃO DOS RESULTADOS E ANÁLISE DO ACESSO AOS SAAES NOS ASSENTAMENOS PRECÁRIOS DO DF

A quarta e última fase da metodologia consiste na escolha do modelo de melhor desempenho preditivo e na elaboração de mapas de visualização dos resultados obtidos. Os mapas trarão os setores censitários da área urbana do DF por tipo, sendo aqueles classificados como SbN os representativos de Assentamentos Precários. Em seguida, utilizando a base de Assentamentos Precários do modelo escolhido, será calculado o Indicador de Universalização Inclusiva do DF, bem como os índices de acesso nos setores NE e no DF como um todo, para fins de comparação. Este tópico trará, portanto, o detalhamento dessa fase metodológica, seguindo os passos descritos acima e ilustrados na Figura 8.

5.4.1 Representação Espacializada da Identificação De Assentamentos Precários No Distrito Federal

O modelo de melhor desempenho preditivo foi escolhido para representar visualmente a configuração geográfica dos Assentamentos Precários no Distrito Federal, com o objetivo de comparar o mapeamento realizado pelo IBGE com o resultado obtido pelo modelo.

Além disso, a elaboração de mapas com as áreas que apresentaram perfil similar aos setores SbN e por isso são consideradas Assentamentos Precários possibilita a continuidade da investigação por meio de visitas de campo *in loco*. Tal etapa não foi inclusa no presente estudo, que se limitou à proposição do método e cálculo dos indicadores inclusivos de acesso aos serviços de abastecimento de água e esgotamento sanitário. A elaboração dos mapas foi realizada no Sofware ArcMap 10.6.1, utilizando dados espaciais coletados no site do IBGE.

5.4.2 Indicadores de Universalização Inclusiva do Acesso aos SAAES no Distrito Federal

Os Assentamentos Precários do DF foram representados pelos chamados Setores Precários, os quais corresponderam aos setores censitários SbN previamente classificados pelo IBGE, somados àqueles identificados como SbN pelo modelo de melhor desempenho preditivo (setores que foram realocados da classe 0 para a classe 1).

Assim, foi possível realizar os somatórios do número total de DPP, moradores em DPP, número de DPP com acesso ao serviço de abastecimento de água e número de DPP com acesso à rede coletora de esgoto ou fossa séptica por tipo de setor, permitindo o cálculo dos percentuais de acesso dos setores Precários, NE e também de todos os setores conjuntamente, tornando possível a comparação do nível de acesso dos Assentamentos Precários com as demais áreas da entidade federativa.

Os indicadores de Universalização Inclusiva (UI) foram propostos por Guimarães (2015) e consistem na razão entre o número de domicílios em Assentamentos Precários atendidos pelo serviço em questão e o número total de domicílios em Assentamentos Precários, permitindo uma análise mais acurada da carência dos serviços especificamente nas localidades urbanas de maior vulnerabilidade socioeconômica. Portanto, utilizando esse conceito e os dados gerados pelo modelo, foram calculados os Indicadores de UI para os serviços de abastecimento de água e esgotamento sanitário no DF para o ano de 2010.

6 DISCUSSÃO E ANÁLISE DE RESULTADOS

Os resultados serão apresentados e analisados em seis seções, sendo a primeira a Estatística Descritiva dos dados, seguida de uma seção para cada método de análise multivariada aplicado à classificação dos setores censitários: a Análise Discriminante Linear (ADL), a Análise Discriminante Quadrática (ADQ) e a Regressão Logística (RL). A quinta seção se referirá à espacialização dos dados e comparação dos resultados obtidos nos modelos estatísticos; e, por fim, a última seção corresponderá ao cálculo dos Índices de Acesso aos SAAES no DF com base nos resultados do modelo de melhor desempenho preditivo, incluindo o cálculo do Indicador de Universalização Inclusiva.

6.1 Análise de Estatísticas Descritivas

A estatística descritiva consiste na primeira etapa analítica dos dados, onde é possível visualizar o comportamento de cada indicador.

6.1.1 Média, Mediana e Desvio Padrão

Para que um indicador tenha alta capacidade de discriminação, os valores de suas médias devem apresentar a maior diferença possível entre os grupos, de modo que os centroides estejam distantes entre si, facilitando a diferenciação de setores NE e SbN. Além disso, baixos valores de variância ou desvio padrão de um indicador dentro dos grupos também são favoráveis quando associados a médias distantes entre os grupos, pois sugerem que os grupos são diferentes entre si, mas similares internamente. Ou seja, que há uma certa homogeneidade dentro dos setores caracterizados como NE ou como SbN e que, de fato, faz sentido reunir tais observações em uma determinada classificação. Isso indica qualidade do indicador, que possivelmente contribuirá para um bom ajuste do modelo. A Tabela 5 mostra os valores de Média, Mediana e Desvio padrão de cada um dos indicadores em cada grupo de setores censitários.

Tabela 5 - Estatística Descritiva dos Indicadores da Amostra: Média, Mediana e Desvio Padrão.

Indicadores	**Abreviação**	**Média**		**Mediana**		**Desvio Padrão**	
		NE	SbN	NE	SbN	NE	SbN
A.1	i_num_dpp	183,56	208,41	179,00	195,00	62,51	87,77
A.2	i_num_mor_dpp	601,66	762,71	590,00	720,00	224,55	319,96
A.3	i_med_mor_por_dpp	**3,28**	**3,67**	3,39	3,64	0,52	0,24
A.4	i_perc_pes_res_brancas	45,01	28,10	40,64	28,57	16,03	**5,81**
A.5	i_perc_PR_sexo_fem	**43,70**	**42,43**	43,71	40,32	11,42	13,07

A.6	i_num_dom_imp	2,36	7,56	1,00	5,00	4,11	8,90
A.7	i_perc_dpp_outra_ocup	0,58	1,87	0,00	0,00	3,00	9,87
B.1	i_perc_PR_ate_30	16,62	24,50	15,87	23,33	8,88	6,87
B.2	i_perc_PR_nalf	3,80	7,43	2,71	6,82	4,00	4,21
B.3	i_perc_PR_ate_30_nalf	0,13	0,45	0,00	0,34	0,35	0,66
B.4	i_perc_PR_ren_ate_3SM	56,47	90,64	63,95	92,50	29,73	**7,23**
B.5	i_Renda_med_PR_dpp	3.023,80	809,41	1.839,17	758,31	2804,63	283,81
C.1	i_perc_dpp_sem_col_lixo	**0,70**	**13,07**	0,00	0,46	5,68	27,51
C.2	i_perc_DPP_sem_ab_agua	3,31	6,29	0,00	0,56	15,00	16,66
C.3	i_perc_DPP_sem_ban_san	0,06	0,23	0,00	0,00	0,42	0,73
C.4	i_perc_DPP_sem_esg_fos	**6,79**	**65,59**	0,00	91,58	21,39	40,30
C.5	i_num_med_ban_hab	1,80	1,25	1,49	1,23	0,78	0,13

Com base nos valores apresentados, verifica-se que as médias dos indicadores referentes ao Número de Moradores por DPP (A.1) e a Pessoas Responsáveis do Sexo Feminino (A.5) são muito próximas, indicando baixo poder de discriminação; enquanto as médias dos indicadores referentes a Domicílios sem Coleta de Lixo (C.1) e Acesso à rede de Esgoto ou Fossa séptica (C.4) apresentaram as maiores diferenças entre si, sugerindo que eles provavelmente serão determinantes na diferenciação entre os dois grupos estudados. Em relação ao desvio padrão, os valores devem ser baixos especialmente no grupo SbN, onde se espera maior homogeneidade, por se referir à classe de estudo. Já quanto ao grupo de setores NE, tolera-se maior discrepância entre suas características, já que abrangem um grupo mais geral de classificação, que representaria os setores não-SbN.

Partindo para a análise, a Tabela 5 mostra que os indicadores que apresentaram menor desvio em relação à média dentro do grupo SbN foram: o Percentual de Pessoas Residentes Brancas (A.4) e o percentual de PR com Renda até três Salários Mínimos (B.4). Ou seja, no que se refere a esses temas, os setores SbN são similares entre si, apresentando valores que se reúnem em torno da média. Quanto à mediana, frisa-se aqui os indicadores com mediana igual a zero, por conta da inflação de zeros nos dados de setores NE, que será explicado em tópico seguinte (6.1.3). No mais, além das características mencionadas, é válido analisar também os valores máximo e mínimo de cada indicador, informações que auxiliam na compreensão de seus comportamentos e das suas disparidades entre os grupos.

6.1.2 Valor Mínimo e Valor Máximo

Como mencionado, dados complementares da estatística descritiva que também podem fornecer informações importantes a respeito do comportamento das variáveis independentes são os valores mínimos e máximos, dispostos na Tabela 6.

Tabela 6 - Valor Máximo e Valor Mínimo dos Indicadores por Tipo de Setor.

Indicadores	**Abreviação**	**Valor mínimo**		**Valor máximo**	
		NE	SbN	NE	SbN
A.1	i_num_dpp	6,00	**45,00**	781,00	543,00
A.2	i_num_mor_dpp	19,00	160,00	3.067,00	1.951,00
A.3	i_med_mor_por_dpp	1,04	3,13	5,44	4,73
A.4	i_perc_pes_res_brancas	5,06	10,54	89,33	**43,29**
A.5	i_perc_PR_sexo_fem	0,98	14,75	86,46	73,47
A.6	i_num_dom_imp	0,00	0,00	96,00	68,00
A.7	i_perc_dpp_outra_ocup	0,00	0,00	86,12	81,20
B.1	i_perc_PR_ate_30	0,00	5,41	92,00	49,69
B.2	i_perc_PR_nalf	0,00	0,51	46,15	21,84
B.3	i_perc_PR_ate_30_nalf	0,00	0,00	4,92	5,34
B.4	i_perc_PR_ren_ate_3SM	0,00	**49,03**	100,00	100,00
B.5	i_Renda_Med_PR_dpp	307,94	285,07	21.343,82	**2.411,17**
C.1	i_perc_dpp_sem_col_lixo	0,00	0,00	100,00	97,58
C.2	i_perc_DPP_sem_ab_agua	0,00	0,00	100,00	95,18
C.3	i_perc_DPP_sem_ban_san	0,00	0,00	16,67	4,93
C.4	i_perc_DPP_sem_esg_fos	0,00	0,00	100,00	100,00
C.5	i_num_med_ban_hab	1,00	1,01	6,16	**1,92**

Dentre os indicadores, destaca-se o valor mínimo do Número de DPP (A.1) em setores SbN, o qual corresponde a 45 domicílios, em contraste aos 6 encontrados em setores NE. Isso mostra que a densidade é uma característica de um setor SbN, que provavelmente contribuirá para o traçado de um perfil com base nas características dos setores SbN, como já explicitado anteriormente.

Além desse indicador, destaca-se que, no mínimo, 49% das Pessoas Responsáveis de setores SbN possuem Renda Mensal de até três Salários Mínimos, frisando a característica de vulnerabilidade econômica e de concentração de pessoas com menor renda nesse tipo de setor. Nessa mesma linha de raciocínio, destaca-se também o valor máximo do indicador de Renda Média (B.5), que é de 2.411,17 reais dentro de setores SbN. Ademais, tem-se que o percentual máximo de pessoas brancas residindo em setores SbN (A.4) é de 43%, denotando a divisão racial que ocorre em conglomerados urbanos, bem como o número de banheiros por habitante (C.5), que não chega a 2, em contraste aos 6 encontrados em setores NE.

6.1.3 Inflação de Zeros

Na análise do comportamento dos indicadores dentro de cada grupo de setores (NE e SbN), observou-se que alguns indicadores apresentaram elevado número de zeros, pelo fato de a amostra conter muito mais setores do tipo NE do que do tipo SbN. Tal fato ocasionou uma inflação de zeros na base de dados, especialmente nos indicadores que se referem ao percentual de domicílios sem determinado serviço, uma vez que muitos setores NE não apresentam nenhum domicílio descoberto, zerando o valor do indicador.

Desse modo, os indicadores correspondentes aos percentuais de domicílios particulares permanentes sem acesso à abastecimento de água (C.2), sem banheiro ou sanitário (C.3) e sem ligação à rede de esgoto ou fossa séptica (C.4) apresentaram, respectivamente, 3621, 3762 e 2778 zeros, de um total de 4069 observações na amostra. O mesmo ocorreu com os indicadores que revelavam o percentual de pessoas responsáveis até 30 anos não alfabetizadas (B.3), com 3315 zeros, e o percentual de domicílios com outra forma de ocupação (A.7), com 2966 zeros. Tal fato afetou, evidentemente, a distribuição dos dados que afetaria a normalidade multivariada, pelo elevado número de zeros.

Em uma distribuição normal, a média é igual ou muito próxima da mediana; no entanto, será possível observar, nos resultados da estatística descritiva, que os dados não seguem esse padrão, apresentando, na verdade, mediana igual a zero, distanciando-se muito de suas respectivas médias. Assim, é esperado que os dados não apresentem distribuição normal – o que será avaliado na aplicação do teste de normalidade. No entanto, a questão da normalidade nos indicadores mencionados não impedirá o prosseguimento do estudo, uma vez que, para o método de Análise Discriminante, é possível assumir normalidade assintótica pelo grande tamanho amostral e, para o método de Regressão Logística, não há a premissa de normalidade da distribuição.

6.2 Análise Discriminante Linear

Esta seção apresentará os resultados da aplicação da técnica da AD, cujo objetivo é discriminar os setores com base nos indicadores fornecidos, por meio de um modelo que gera um *score* para cada setor e, assim, permite reclassificá-lo no grupo com o qual o perfil socioeconômico de sua população tem mais similaridade. Foram aplicados dois tipos de AD: a ADL e a ADQ, as quais se diferenciam basicamente pela

ordem da equação do modelo, sendo a primeira linear e a segunda, quadrática. Os dois primeiros tópicos apresentados consistem em etapas iniciais comuns aos dois métodos: a realização do procedimento Stepwise, cujo objetivo é selecionar as variáveis independentes que entrarão nos modelos e o teste dos pressupostos para a aplicação das técnicas.

6.2.1 Procedimento Stepwise: Seleção das Variáveis Independentes para a Análise Discriminante

Como foi possível verificar pela análise descritiva dos indicadores, nem todos indicaram ter impacto significativo na discriminação de setores SbN. Assim, foi aplicado o procedimento Stepwise de seleção de variáveis, cujo resultado pode ser visualizado na Tabela 7, onde os indicadores estão dispostos na ordem em que foram selecionados e incluídos ao modelo.

Tabela 7 - Etapas do Procedimento Stepwise.

Indicador	Abreviação	Lambda de Wilks	P-valor
C.4	i_perc_DPP_sem_esg_fos	0,781	9,62E-221
C.1	i_perc_dpp_sem_col_lixo	0,753	3,58E-251
C.2	i_perc_DPP_sem_ab_agua	0,729	9,91E-278
A.6	i_num_dom_imp	0,714	1,01E-295
B.1	i_perc_pr_ate_30	0,709	1,24E-300
A.2	i_num_mor_dpp	0,705	2,93E-303
A.1	i_num_dpp	0,705	3,99E-303
B.2	i_perc_PR_nalf	0,704	3,88E-303
B.4	i_perc_pr_ren_ate_3_SM	0,702	2,07E-303
B.5	i_Renda_Med_PR_DPP	0,699	1,22E-306
C.5	i_num_med_ban_hab	0,695	0
B.3	i_perc_PR_ate_30_nalf	0,693	0

O P-valor revela a significância estatística de cada etapa e deve ser menor que 0,05 para que seja significante a 5%. Nesse caso, como observado, todas as variáveis apresentaram P-valor igual a zero, atingindo, portanto, a significância desejada.

A visualização dos resultados mostra a seleção de doze indicadores, atingindo o Lamba de Wilks de 0,693. Como as variáveis são adicionadas ao modelo de acordo com sua ordem de significância, é possível verificar que os indicadores referentes ao acesso à rede coletora de esgoto, coleta de lixo e abastecimento de água se destacam como os três primeiros mais relevantes para o modelo.

6.2.2 Testes dos Pressupostos da Análise Discriminante

Uma vez definidas as variáveis independentes pelo procedimento Stepwise, a próxima etapa consiste no teste dos pressupostos da AD, os quais são: ausência de multicolinearidade, normalidade multivariada e, no caso específico da ADL, homogeneidade das matrizes de covariância – a ADQ flexibiliza esse último pressuposto, permitindo matrizes de covariâncias distintas entre os grupos. O primeiro pressuposto se refere à ausência de multicolinearidade. Para checar a satisfação dessa premissa, foi utilizada a matriz de correlação de Pearson (Tabela 8).

Tabela 8 - Matriz de Correlação de Pearson das variáveis selecionadas pelo Procedimento Stepwise.

	A.1	A.2	A.6	B.1	B.2	B.3	B.4	B.5	C.1	C.2	C.4	C.5
A.1	1	**0,906**	0,331	0,052	0,042	0,046	0,092	-0,070	-0,078	-0,141	-0,001	-0,104
A.2	0,906	1	0,437	-0,002	0,214	0,115	0,292	-0,221	-0,040	-0,101	0,092	-0,072
A.6	0,331	0,437	1	0,306	0,448	0,267	0,462	-0,349	0,059	-0,032	0,201	-0,295
B.1	0,052	-0,002	0,306	1	0,298	0,229	0,537	-0,523	0,078	-0,038	0,117	-0,600
B.2	0,042	0,214	0,448	0,298	1	0,420	0,729	-0,577	0,216	0,131	0,230	-0,447
B.3	0,046	0,115	0,267	0,229	0,420	1	0,291	-0,216	0,160	0,105	0,211	-0,163
B.4	0,092	0,292	0,462	0,537	0,729	0,291	1	**-0,875**	0,128	0,006	0,209	-0,663
B.5	-0,070	-0,221	-0,349	-0,523	-0,577	-0,216	-0,875	1	-0,084	0,000	-0,162	0,789
C.1	-0,078	-0,040	0,059	0,078	0,216	0,160	0,128	-0,084	1	0,340	0,303	-0,067
C.2	-0,141	-0,101	-0,032	-0,038	0,131	0,105	0,006	0,000	0,340	1	0,293	0,088
C.4	-0,001	0,092	0,201	0,117	0,230	0,211	0,209	-0,162	0,303	0,293	1	-0,066
C.5	-0,104	-0,072	-0,295	-0,600	-0,447	-0,163	-0,663	0,789	-0,067	0,088	-0,066	1

De acordo com Favero et al. (2009), duas variáveis podem ser consideradas multicolineares a partir de uma correlação de 0,85. Embora não haja uma unanimidade na literatura a respeito de um valor específico, considera-se que valores acima de 0,8 devem ser analisados com cautela, uma vez que ainda há a possibilidade de correlação espúria. Com base nisso, foram observadas altas correlações em dois pares de indicadores (grifo na Tabela 8): A.1 e A.2, com correlação de 0,9 e B.4 e B.5, com 0,87.

O primeiro par se refere ao Número de DPP (A.1) e ao Número de Moradores em DPP (A.2) em um dado setor censitário, os quais provavelmente apresentam uma relação linear entre si, revelando-se colineares, na medida em que explicam eventos semelhantes: quanto mais domicílios, mais moradores em domicílio. Há sentido também na alta correlação entre os dois indicadores que se referem à renda (B.4 e B.5), os quais apresentam correlação negativa, uma vez que, quanto maior o número de Pessoas Responsáveis com Renda de até três Salários Mínimos, menor a renda média de determinado setor.

Diante dessa situação, um indicador de cada par deverá ser excluído do modelo, com o intuito de garantir a premissa da não multicolinearidade entre as variáveis independentes na AD. Para tomar a decisão de exclusão das variáveis, foi utilizado o teste F (tabela 9), que consiste na razão entre a variação entre os grupos e a variação dentro dos grupos, de forma que, quanto maior o seu valor, mais a variável contribuirá para a distinção entre os grupos no modelo.

Tabela 9 - Resultados do Teste F

Indicadores com Alta Correlação	Abreviação	Valor de F
A.1	i_num_dpp	25,40
A.2	i_num_mor_dpp	82,51
B.4	i_perc_pr_ren_ate_3_SM	230,54
B.5	i_Renda_Med_PR_DPP	109,02

Ao comparar os indicadores, pode-se verificar que A.2 e B.4 apresentaram maior valor de F, sendo então mais efetivos para a discriminação entre os grupos, o que resultou na exclusão de A.1 e B.5 do modelo – mantendo, evidentemente, as demais variáveis independentes selecionadas. Dessa forma, o modelo final totalizou dez variáveis independentes.

O segundo pressuposto – existência de normalidade multivariada nas variáveis independentes – foi analisado por meio do teste de Mardia e Henze-Zirkler, utilizando pacote *MVN*, cuja hipótese nula atesta que a amostra vem de uma distribuição normal multivariada em ambos os testes. Os resultados obtidos estão expressos na Tabela 10.

Tabela 10 - Resultados do Teste de Normalidade Multivariada de Mardia e Henze-Zirkler.

Tipo de Setor	Estatística de Mardia		Estatística de Henze-Zirkler	P-valor
	Curtose	Assimetria		
Setor Tipo 0 (NE)	1.176,63	217.143,76	57,06	0
Setor Tipo 1 (SbN)	28,60	1.951,41	1,77	0

No entanto, como pode ser verificado, os p-valores obtidos em ambos os testes foram iguais a zero, menores que o nível de significância de 0,05. Desse modo, rejeita-se a hipótese nula, o que significa que o teste não nos permite afirmar que a amostra vem, de fato, de uma população normal.

No entanto, há alguns fatores que explicam as possíveis razões pelas quais os resultados tenham rejeitado a hipótese de normalidade. O primeiro foi mencionado em tópico anterior e se trata da inflação de zeros presente em algumas variáveis independentes, o que poderia causar um desvio da normalidade multivariada dos dados.

Além disso, é importante pontuar que, ao contrário do que ocorre com as amostras pequenas – em que os testes de normalidade têm pouco poder de rejeição da hipótese nula – em grandes amostras, basta um pequeno desvio da normal para que os resultados rejeitem a hipótese de normalidade, ainda que este não afete os resultados de testes paramétricos (Field, 2009; Otzuna et al., 2006).

No entanto, em casos de grande tamanho amostral – como é o caso no presente estudo (n = 4069) –, a violação do pressuposto de normalidade não causa grandes problemas para a aplicação da ADL (Favero, 2009; Pallant, 2007). Para grandes amostras com n > 40, é possível aplicar procedimentos paramétricos ainda que os dados não sejam normalmente distribuídos e, quando n > 100, é possível inclusive ignorar a distribuição dos dados (Elliot, 2007; Altman, 1995). Isso ocorre porque, de acordo com o Teorema do Limite Central, em amostras grandes, a distribuição amostral tende para uma normal, independentemente da forma dos dados, pelo fato de que a variância convergirá para zero (Field, 2009; Elliot, 2007). Tal fato nos permite assumir normalidade assintótica dos dados, a qual ocorre quando o alto número de observações na amostra atua na redução da dispersão dos dados e consequente convergência para uma distribuição normal.

O terceiro e último pressuposto é exclusivo da ADL e se trata da homogeneidade das matrizes de variância e covariância dos grupos. Para avaliar sua satisfação, foi realizado o teste de M de Box, cuja hipótese nula atesta que as matrizes de covariâncias dos grupos são iguais. Para não rejeitar tal hipótese – garantindo uma significância estatística de 5% –, o p-valor deve ser maior que 0,05. O resultado do teste está disposto na Tabela 11.

Tabela 11 - Resultados do Teste M de Box.

Teste Box's M para Homogeneidade de Matrizes de Covariância	
Qui-quadrado (aprox.)	5741.8
Graus de liberdade (df)	78
P-valor	< 2.2e-16

Como pode-se verificar, o p-valor foi muito inferior ao desejado e a hipótese nula foi rejeitada, apontando para a hipótese alternativa, de não homogeneidade das matrizes. No entanto, tal resultado é compreensível – e até mesmo esperado –, pois o teste M de Box é sensível a desvios da normalidade e a tamanhos muito grandes de amostra (Favero, 2009) e, como foi mostrado anteriormente, a hipótese de normalidade

foi rejeitada nos testes de Mardia e Henze-Zirkler e a amostra utilizada apresenta um alto número de observações.

Além disso, há mais um ponto relevante: o teste M de Box avalia a covariância – medida relacionada à dispersão –, de modo que quanto menor a variação de cada indicador em relação à média de seu grupo, menor será a dispersão dentro dos grupos e, por conseguinte, maior a propensão de apresentarem covariâncias homogêneas entre si (hipótese nula). Porém, os dados da presente pesquisa apresentam alta variabilidade e, assim, tendem a trilhar o caminho contrário – de maior dispersão. Isso ocorre principalmente dentro do grupo de setores NE, que abrange setores de todo tipo, com uma enorme variedade de características, inclusive de Assentamentos Precários, cuja identificação é objetivo que orienta a presente pesquisa.

De todo modo, diante da rejeição da hipótese nula do teste M de Box, aplica-se o mesmo argumento do teorema do limite central, uma vez que grandes tamanhos de amostra tendem a atuar na redução da dispersão dos dados, por meio da redução da variância – já que esta tende a zero quando a população amostral tende ao infinito (Magalhaes, 1999; Woolridge, 2016). E, ainda, a técnica da AD possui grande robustez diante da violação das premissas, desde que o tamanho amostral dos grupos seja superior ao número de variáveis independentes (Favero, 2009).

6.2.3 Estimação da Função Discriminante Linear

Após testadas as premissas e discutidas suas violações, foi construída a função discriminante linear, responsável por diferenciar os grupos com base nas dez variáveis explicativas previamente selecionadas. O modelo da ADL gerou os seguintes resultados: as probabilidades a priori de cada grupo (Tabela 12), calculadas basicamente por meio da proporção da amostra de cada grupo pela amostra total; os centroides da função (Tabela 13) e os coeficientes não padronizados (Tabela 14).

Tabela 12 - Probabilidades a Priori dos Setores (AD).

Tipo de Setor	Grupo	Probabilidade
NE	0	0.957
SbN	1	0.0437

Os dados da Tabela 12 mostram que, no modelo da ADL, a probabilidade a priori de um determinado setor ser classificado no grupo 0, representado pelos setores do tipo NE, é de 95,7% enquanto a de ser classificado como SbN – grupo 1 – é de apenas 4,3%. Isso é resultado do fato da amostra total conter um valor muito maior de

setores do tipo NE (3894) do que SbN (175). É possível alterar essas probabilidades na função, para que sejam iguais entre os grupos ou até mesmo definir outros valores arbitrariamente. No entanto, optou-se por manter a proporção realista da quantidade de setores do tipo SbN em relação a todos os setores do Distrito Federal, uma vez que alterar as probabilidades originais poderia enviesar o modelo.

O segundo resultado da construção da função discriminante corresponde aos centroides dos grupos, os quais representam as médias das variáveis independentes em cada grupo (Tabela 13).

Tabela 13 - Centroides das Variáveis Independentes por Grupo (Tipo de Setor).

Indicadores	Abreviação	Grupo 0 (NE)	Grupo 1 (AS)
A.2	i_num_mor_dpp	601,664	762,714
A.6	i_num_dom_imp	2,362	7,560
B.1	i_perc_pr_ate_30	16,619	24,498
B.2	i_perc_PR_nalf	3,798	7,430
B.3	i_perc_PR_ate_30_nalf	0,128	0,452
B.4	i_perc_pr_ren_ate_3_SM	56,466	90,640
C.1	i_perc_dpp_sem_col_lixo	0,703	13,072
C.2	i_perc_DPP_sem_ab_agua	3,307	6,290
C.4	i_perc_DPP_sem_esg_fos	6,793	65,593
C.5	i_num_med_ban_hab	1,804	1,254

Os centroides são utilizados para o cálculo da constante da função discriminante linear, representada por C. A constante é inserida no modelo como um ajuste, um reescalonamento, para garantir na função a média 0 e desvio padrão 1, reposicionando a origem dos coeficientes não padronizados na grande média – ou seja, no centroide dos centroides (Sharma, 1996; Klecka, 1975). Portanto, a constante C corresponde ao valor que, quando somado à média ponderada entre os dois *scores* médios ("média das médias"), o iguala à zero e, como a função discriminante linear assume a forma Z = a'x + C, o cálculo da constante se dá por:

$$\frac{n1a'\overline{x1} + n2a'\overline{x2}}{n1 + n2} + C = 0$$

Onde:

n1: número da população amostral de setores NE;

n2: número da população amostral de setores SbN;

$\overline{x1}$: centroides das variáveis independentes nos setores NE;

$\overline{x2}$: centroides das variáveis independentes nos setores SbN;

Assim, para os grupos aqui estudados, tem-se que:

$$C = -\frac{3894 * 0{,}972 + 175 * 4{,}19}{3894 + 175} = -1{,}11$$

Ao passo que a função ajustada de Fisher resulta em:

$$Z = a'x - 1{,}11$$

O valor de C encontrado consta na Tabela 14, onde estão dispostos os coeficientes não padronizados, que formam a estrutura da função discriminante linear.

Tabela 14 - Coeficientes não Padronizados das Variáveis Independentes.

Indicadores	Abreviação	Coeficientes Não Padronizados
A.2	i_num_mor_dpp	0,0005
A.6	i_num_dom_imp	0,0387
B.1	i_perc_pr_ate_30	0,0121
B.2	i_perc_PR_nalf	-0,0477
B.3	i_perc_PR_ate_30_nalf	0,2374
B.4	i_perc_pr_ren_ate_3_SM	0,0065
C.1	i_perc_dpp_sem_col_lixo	0,0571
C.2	i_perc_DPP_sem_ab_agua	-0,0198
C.4	i_perc_DPP_sem_esg_fos	0,0350
C.5	i_num_med_ban_hab	-0,0212
C	Constante	-1,11

Assim, seguindo a forma a'x + C, a função pode ser escrita da seguinte maneira:

$$Z = 0{,}0005\ I_{A.2} + 0{,}0387\ I_{A.6} + 0{,}0121\ I_{B.1} - 0{,}0477\ I_{B.2} + 0{,}2374\ I_{B.3} + 0{,}0065\ I_{B.4} + 0{,}0571\ I_{C.1} - 0{,}0198\ I_{C.2} + 0{,}035\ I_{C.4} - 0{,}0212\ I_{C.5} - 1{,}11$$

A função formada pelos coeficientes não padronizados é utilizada para o cálculo dos *scores* discriminantes Z de cada observação, os quais serão utilizados na etapa de classificação para determinar em que grupo dada observação será alocada. Os coeficientes padronizados, por outro lado, são úteis apenas para a análise do peso das variáveis independentes na função discriminante, funcionando como um parâmetro de medida do poder discriminante de cada variável (Hair et al., 2005).

Há, entretanto, outra forma de avaliar a magnitude da influência de cada variável na função, denominada de carga discriminante. Essas cargas são obtidas por meio da chamada matriz de estrutura, formada pela correlação de cada variável independente com os *scores* discriminantes da função. Variando entre -1 e 1, fornecem uma

hierarquia do poder discriminante das variáveis e auxiliam na interpretação dos pesos encontrados nos coeficientes padronizados, os quais, se analisados isoladamente, podem apresentar certa confusão na interpretação caso haja algum grau de multicolinearidade entre as variáveis (Favero et al., 2009; Maroco, 2007).

Klecka (1975) mostra que, em casos de grande tamanho amostral – por exemplo em casos em que há 20 vezes mais observações do que variáveis –, as cargas discriminantes são mais estáveis do que os coeficientes padronizados da função, permitindo uma análise mais acurada. Assim, juntamente aos coeficientes padronizados, foram geradas também as cargas discriminantes, para fins de comparação (Tabela 15).

Tabela 15 - Coeficientes Padronizados e Cargas Discriminantes da Função Discriminante.

Indicadores	Abreviação	Coeficientes Padronizados	Cargas Discriminantes
A.2	i_num_mor_dpp	0,111	0,218
A.6	i_num_dom_imp	0,171	0,365
B.1	i_perc_pr_ate_30	0,106	0,279
B.2	i_perc_PR_nalf	-0,191	0,282
B.3	i_perc_PR_ate_30_nalf	0,087	0,273
B.4	i_perc_pr_ren_ate_3_SM	0,188	0,365
C.1	i_perc_dpp_sem_col_lixo	0,454	0,484
C.2	i_perc_DPP_sem_ab_agua	-0,298	0,062
C.4	i_perc_DPP_sem_esg_fos	0,788	0,812
C.5	i_num_med_ban_hab	-0,016	-0,222

Na análise da Tabela 15, a primeira questão que se destaca se trata dos sinais. Os indicadores com sinais positivos indicariam que, quanto maior o valor daquele indicador, maior será o *score* discriminante da função e maior chance aquela observação terá de ser classificada como um setor SbN. Na Tabela 15, verifica-se que, nas cargas discriminantes, o único indicador que apresenta sinal negativo é o C.5. Isso significa que essa variável contribui para uma redução no *score* da função e, por conseguinte, reduz as chances de um dado setor ser classificado como SbN. Essa relação é coerente com o que se espera empiricamente, uma vez que, quanto maior o número médio de banheiros por habitante, menor a chance de um setor ser SbN. Já nos coeficientes padronizados, o sinal negativo também ocorre nos indicadores B.2 e C.2, onde não se observa essa mesma coerência.

Em termos de valores absolutos, foi possível verificar que, tanto nos coeficientes padronizados quanto na matriz de estrutura, o indicador que apresentou maior poder discriminante foi o C.4, o percentual de DPP sem rede de esgoto ou fossa séptica, seguido do C.1, referente à coleta de lixo, com 0,48. Isso significa que, dentre as

variáveis presentes na função, a que se refere ao acesso à rede de esgoto é aquela que apresenta maior contraste entre um setor do tipo NE e um setor do tipo SbN. Isso sugere que, de acordo com o modelo construído, o acesso a esse serviço de saneamento é uma questão de destaque nos assentamentos precários do DF.

Com relação ao abastecimento de água, o coeficiente padronizado se apresenta negativo, provavelmente porque o Distrito Federal, diferente das demais unidades federativas, se destaca em termos de atendimento a esse serviço, água, cobrindo um total de 97,32% da população em 2010 (IBGE, 2011), ano de referência dos dados de estudo. Isso pode ter levado a muitos índices altos dentro do grupo SbN e confundido o modelo. Por essa mesma razão, a carga discriminante desse indicador não foi muito significativa, apresentando um valor de 0,06.

Os indicadores A.6 e B.4, referentes ao número de domicílios improvisados e à renda da população também se destacaram com as maiores cargas discriminantes subsequentes, o que se alinha com características esperadas de setores SbN, ocupado por pessoas mais pobres e com improviso de habitação. O B.2, referente ao percentual de pessoas responsáveis não alfabetizadas seguiu com uma relevância de 0,28, denotando a diferença de escolaridade da população habitante de setores NE e SbN. Por fim, o percentual de pessoas responsáveis até 30 anos (B.1) e o percentual de pessoas responsáveis até 30 anos não alfabetizadas (B.3) tiveram carga discriminante similar, em torno de 0,27.

6.2.4 Predição do Modelo Discriminante Linear: Classificação

Com a função da ADL construída, pôde-se realizar a predição do modelo, que consistiu na reclassificação da amostra total de setores censitários originalmente classificados pelo IBGE. Para melhor visualização dos dados, foram calculados os *scores* discriminantes médios por tipo de setor, pois são dados que baseiam a predição realizada pela ADL. Seu cálculo é realizado pela substituição dos centroides na função discriminante linear, resultando, portanto, nos seguintes valores:

$$Z = \overline{Z0} = 0{,}0005 * 601{,}664 + 0{,}0387 * 2{,}362 + 0{,}0121 * 16{,}619 - 0{,}0477 * 3{,}798 + 0{,}2374 * 0{,}128 + 0{,}0065 * 56{,}466 + 0{,}0571 * 0{,}703 - 0{,}0198 * 3{,}307 + 0{,}035 * 6{,}793 - 0{,}0212 * 1{,}804 - 1{,}11 = -0{,}1381$$

$$Z = \overline{Z1} = 0{,}0005 * 762{,}714 + 0{,}0387 * 7{,}560 + 0{,}0121 * 24{,}498 - 0{,}0477 * 7{,}430 + 0{,}2374 * 0{,}452 + 0{,}0065 * 90{,}640 + 0{,}0571 * 13{,}072 - 0{,}0198 * 6{,}290 + 0{,}035 * 65{,}593 - 0{,}0212 * 1{,}254 - 1{,}11 = 3{,}077$$

Diante dos valores obtidos, tem-se que o *score* discriminante médio dos setores do tipo NE é de $\overline{Z0} = -0{,}14$ enquanto o dos setores do tipo SbN corresponde a $\overline{Z1} = 3{,}08$. Nos gráficos de densidade que mostram os *scores* discriminantes dos tipos de setor antes e depois da predição (Figura 11), é possível observar que, ao serem projetados no eixo X, os picos das curvas dos setores do tipo NE estão em torno de zero, enquanto os dos SbN apontam para 3, aproximadamente.

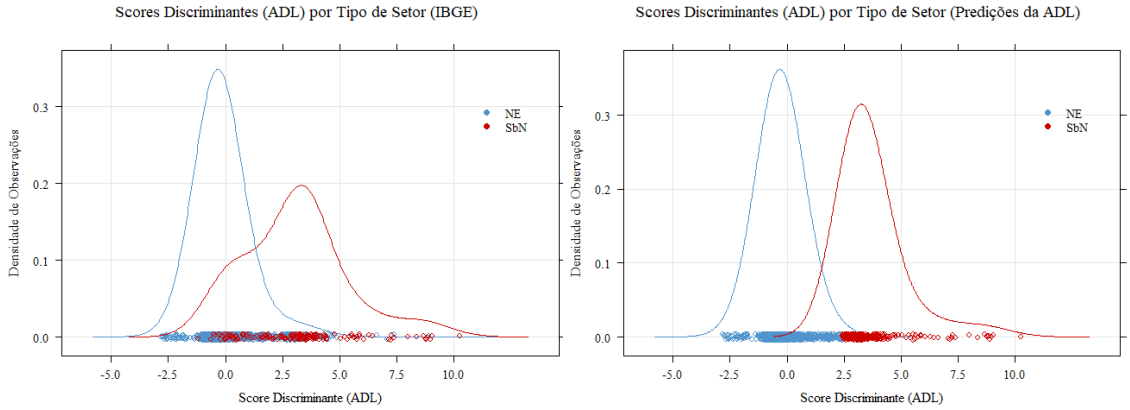

Figura 11 - Gráficos do Scores Discriminantes da ADL x Densidade de Observações por Tipo de Setor (IBGE e ADL).

Nota-se que, nos *scores* dos tipos de setor mapeados pelo IBGE, isto é, antes da classificação, à esquerda, há uma grande área de interseção entre as curvas de densidade, mostrando uma região de incerteza na classificação, ao passo que, nos *scores* dos tipos de setor preditos pelo modelo, após a classificação, à direita da Figura 11, a região de interseção foi significativamente reduzida, uma vez que o cruzamento das linhas foi deslocado para o ponto que corresponde a 1,47 no eixo X. Esse ponto de cruzamento é o limiar do modelo, representado por $\hat{m}$ e consiste na média entre os dois *scores* médios:

$$\hat{m} = \frac{-0{,}1381 + 3{,}077}{2} = 1{,}47.$$

O limiar seria o ponto de corte (*cut-off*) do modelo se as probabilidades a priori tivessem sido definidas em 0,5, de modo que os setores que apresentassem *score* acima de 1,47 teriam sido classificados como SbN. No entanto, como a probabilidade a priori

de um setor censitário ser classificado como NE foi definida como 0,96, seguindo a proporção da amostra, o ponto de corte foi deslocado para a direita e a região de interseção ficou dominada por setores NE, passando a apresentar setores SbN a partir de um valor próximo de 2,5.

Outra forma de representação dessa interseção é o histograma, onde são visualizados os *scores* discriminantes no eixo X e quantas vezes eles ocorreram dentre todas as observações, no eixo Y, antes e depois da predição (Figura 12).

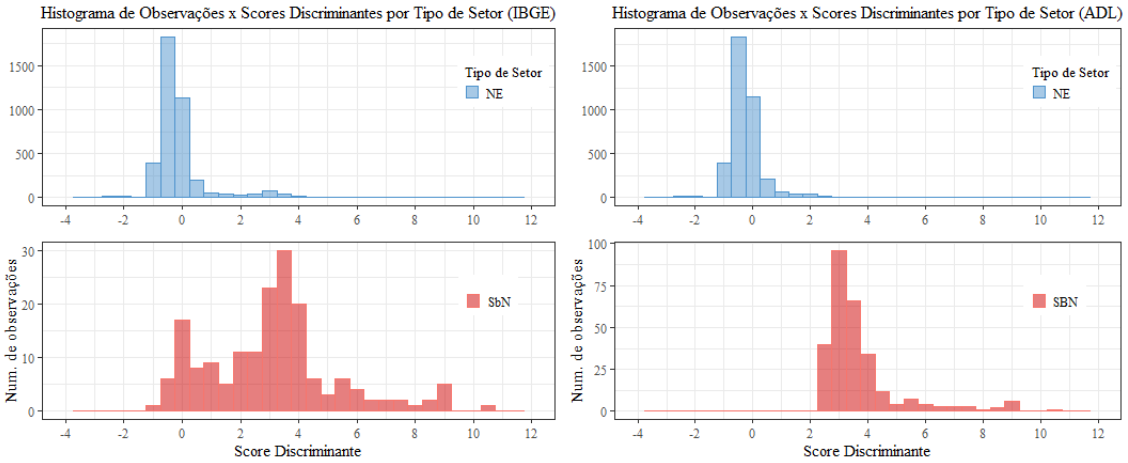

Figura 12 - Gráficos de Score Discriminante x Número de Observações por Tipo de Setor (IBGE e ADL).

A análise dos histogramas mostra que os *scores* discriminantes dos setores do tipo NE mapeados pelo IBGE estavam concentrados até 2, aproximadamente, a partir de onde começam a ocorrer a maior parte dos setores do tipo SbN. Nota-se, também, que os setores NE com *scores* maiores que 2 foram reclassificados como SbN pelo modelo da ADL, representando setores que, embora previamente classificados como NE, apresentaram perfil mais próximo dos setores SbN, os quais são o foco do presente estudo.Evidentemente, ao mencionar "perfil mais próximo", considera-se aqui as médias das dez variáveis estudadas e inseridas na construção da função. Não há, portanto, pretensão de afirmar que um assentamento precário se limita unicamente às características extraídas dessas variáveis; mas, que, dentre as dezessete escolhidas, essas foram as que apresentaram significância para a região do DF e, dessa forma, representam as referências de análise.

As duas variáveis de maior poder preditivo, portanto, foram plotadas em relação aos *scores* discriminantes antes e depois da predição (Figura 13), com o objetivo de ilustrar as regiões de classificação. Tais varáveis correspondem aos indicadores de

Renda (B.4) e de Acesso a serviços de Esgotamento Sanitário (C.4), que demonstraram ter grande influência na classificação dos setores.

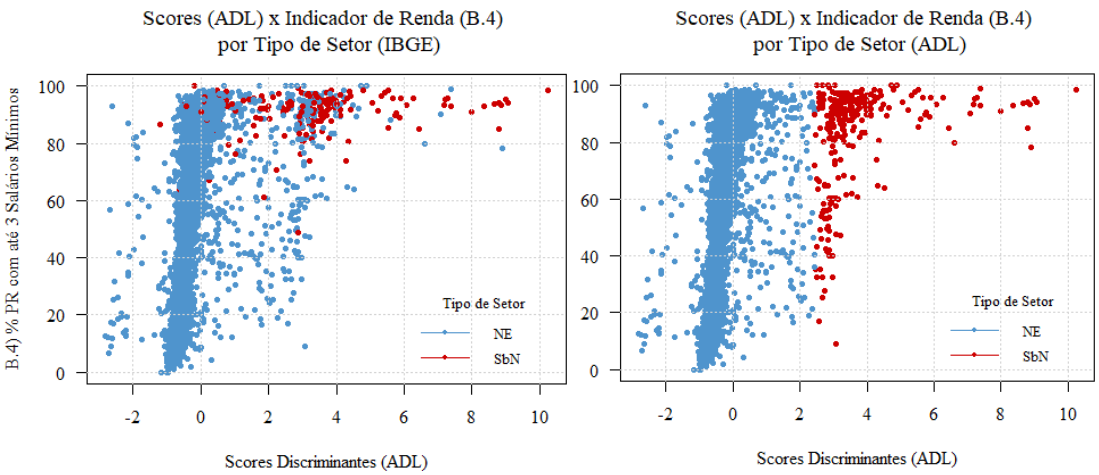

Figura 13 - Gráficos de Score Discriminante x Indicador de Renda (B.4) por Tipo de Setor (IBGE e ADL).

Verifica-se que a renda, antes da predição, representou um fator determinante na caracterização de setores SbN, com todos os pontos em vermelho posicionados acima de 50 no eixo Y, o que significa que todos os setores SbN mapeados pelo IBGE apresentaram mais de 50% de suas Pessoas Responsáveis ganhando até 3 Salários Mínimos. No entanto, após a predição, observou-se a detecção de setores SbN posicionados abaixo de 50 no eixo Y, reclassificados à direita da linha vertical que passa pelo ponto 2,5 no eixo X (ponto de corte), reduzindo assim a especificidade do modelo da ADL.

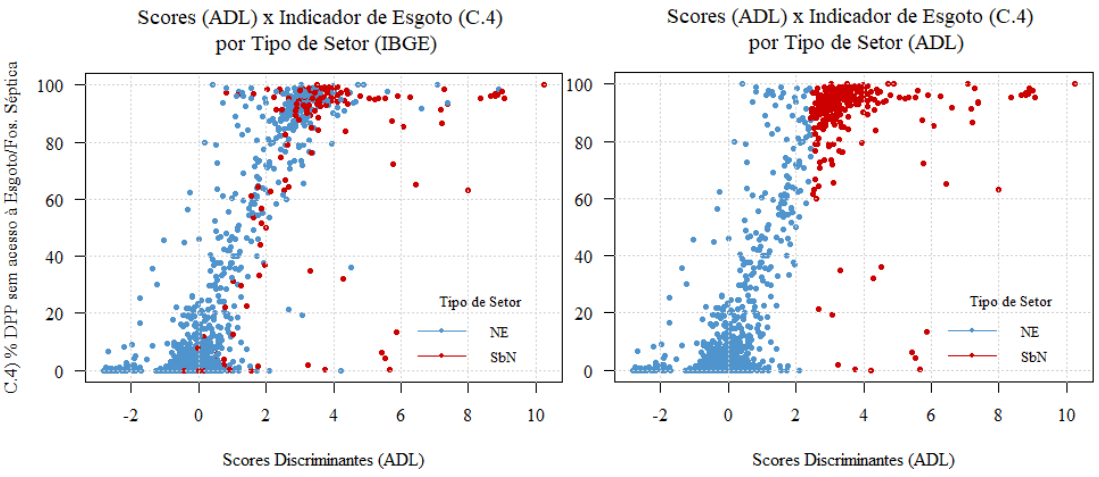

Figura 14 - Gráficos de Score Discriminante x Indicador de Esgoto (C.4) por Tipo de Setor (IBGE e ADL).

Já em relação ao indicador C.4, ocorreu o inverso: após a predição, ao classificar como SbN os setores com *score* acima de 2,5, alguns setores originalmente SbN não foram detectados pelo modelo da ADL, por apresentarem *score* abaixo do ponto de corte, ocasionando, portanto, uma redução na sensitividade do modelo.

Por fim, os indicadores B.4 e C.4 foram plotados entre si, para visualizar o comportamento da falta de acesso à esgotamento sanitário em relação às menores rendas por tipo de setor, antes e depois da predição das classes pelo modelo da ADL (Figura 15).

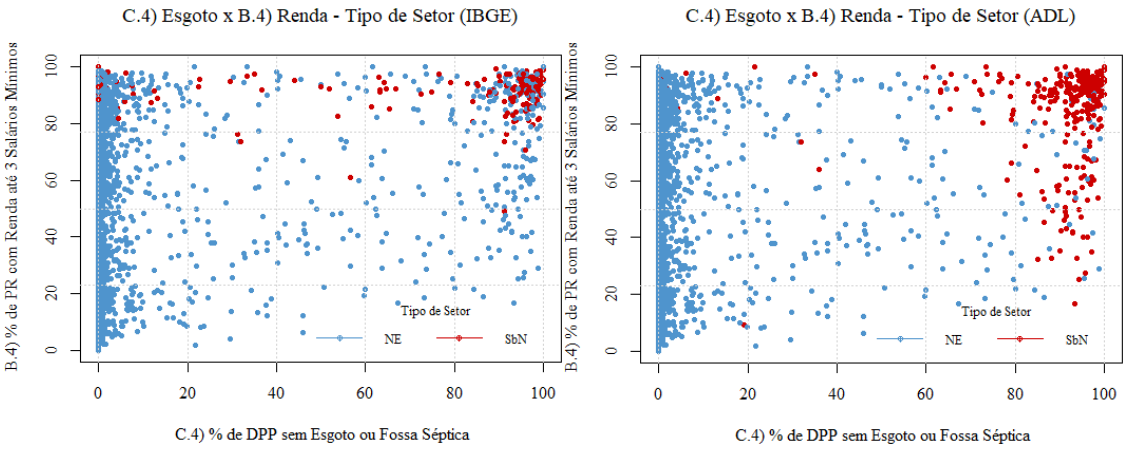

Figura 15 - Indicador B.4 x Indicador C.4 por Tipo de Setor (IBGE e ADL).

Foi possível observar que, em ambos os casos, antes e depois da predição, os setores SbN estão concentrados no canto superior direito dos gráficos, apontando para uma clara relação entre a renda e o número de domicílios sem ligação à rede de esgoto ou sem fossa séptica em Assentamentos Precários.

6.2.5 Avaliação da Predição do Modelo Discriminante Linear: Matriz de Confusão

Os resultados da predição estão expressos na matriz de confusão (Tabela 16).

Tabela 16 - Matriz de confusão das classes originais x classes preditas pelo modelo ADL.

		Classificação de origem (IBGE, 2011)		
		0 (NE)	1 (SbN)	Total
Classificação pelo modelo ADL	0 (NE)	3726	61	3787
	1 (SbN)	168	114	282
	Total	3894	175	**4069**

Verifica-se que o modelo classificou 282 setores como SbN, contra os 175 detectados previamente pelo IBGE. Dos 3894 setores censitários do tipo NE, 168 foram realocados para a classe SbN, ao passo que dos 175 setores originalmente SbN, 114 foram classificados como tal.

Alternativamente, isso pode ter acontecido com setores que não apresentavam valores muito drásticos nas variáveis consideradas e acabaram apresentando resultados ligeiramente mais próximos da classe NE do que da classe SbN. Já o número de setores NE classificados como NE pelo modelo não representa, para fins do presente estudo, um nível de acerto, vez que o objetivo final era justamente avaliar o número de setores que se deslocariam dessa classe por apresentar mais similaridade com um setor SbN, pelas características descritas nas variáveis independentes.

6.3 Análise Discriminante Quadrática

O segundo método de análise multivariada aplicado ao conjunto de dados foi a Análise Discriminante Quadrática (ADQ), que consiste em um modelo não linear cujo limiar de classificação é representado não mais por uma reta, mas por uma curva. Essa técnica difere da ADL especialmente pelo fato de não requerer a premissa de homogeneidade de matrizes de covariância. Nesse caso, portanto, não seria necessária a realização do teste M de Box, mantendo-se apenas o pressuposto de normalidade assintótica.

6.3.1 Estimação da Função Discriminante Quadrática

A construção da função discriminante quadrática gerou os seguintes resultados: probabilidades a priori, os centroides dos indicadores por grupo – idênticos aos gerados pela ADL – e os coeficientes da função. Destaca-se, no entanto, que os sinais e valores dos coeficientes das funções não podem ser interpretados como contribuição da variável à função, como é o caso da ADL (Erket et al., 2014; Reberšek et al., 2011). Na ADQ, os coeficientes são calculados de forma que, quando multiplicados pela matriz de variáveis independentes dentro de cada grupo, resultem em uma matriz identidade, com diagonal igual a um e os demais valores iguais a zero.

Além disso, pelo fato de considerar matrizes de covariância distintas, são geradas duas funções discriminantes, Q_0 e Q_1, uma para cada grupo. Os coeficientes dessas funções estão organizados em forma de matriz, dispostos nas Tabelas 17 e 18,

onde os valores da diagonal principal multiplicam os indicadores elevados ao quadrado e os demais multiplicam um indicador pelo outro, seguindo as linhas e colunas de cada matriz.

Tabela 17 - Matriz de coeficientes da QDA para o grupo 0 (NE).

Indicadores	A.2	A.6	B.1	B.2	B.3	B.4	C.1	C.2	C.4	C.5
A.2	0,004	-0,002	0,001	0,000	0,000	-0,001	0,001	0,000	0,000	-0,001
A.6	-	0,266	-0,091	-0,106	0,012	-0,019	0,018	0,010	-0,016	-0,002
B.1	-	-	0,119	-0,022	0,011	-0,071	0,000	0,007	0,001	0,055
B.2	-	-	-	0,285	0,097	-0,266	-0,108	-0,010	-0,004	0,013
B.3	-	-	-	-	-3,135	0,277	-0,004	-0,202	-0,285	-0,213
B.4	-	-	-	-	-	0,059	0,005	0,001	-0,001	0,025
C.1	-	-	-	-	-	-	0,187	-0,103	-0,030	0,005
C.2	-	-	-	-	-	-	-	0,078	-0,018	-0,008
C.4	-	-	-	-	-	-	-	-	0,051	-0,004
C.5	-	-	-	-	-	-	-	-	-	1,869

Tabela 18 - Matriz de coeficientes da QDA para o grupo 1 (SbN).

Indicadores	A.2	A.6	B.1	B.2	B.3	B.4	C.1	C.2	C.4	C.5
A.2	0,003	-0,002	-0,001	0,000	-0,001	0,000	0,000	-0,001	-0,001	0,001
A.6	-	0,136	0,053	0,032	0,004	-0,004	0,012	0,019	0,026	-0,042
B.1	-	-	-0,156	0,008	-0,045	-0,075	0,006	-0,004	-0,020	-0,017
B.2	-	-	-	-0,246	-0,139	-0,100	0,027	0,063	-0,032	-0,026
B.3	-	-	-	-	1,905	0,246	-0,386	0,234	0,157	-0,202
B.4	-	-	-	-	-	0,166	-0,019	-0,043	0,018	-0,162
C.1	-	-	-	-	-	-	0,037	-0,001	0,000	-0,004
C.2	-	-	-	-	-	-	-	-0,066	0,007	0,003
C.4	-	-	-	-	-	-	-	-	-0,026	0,003
C.5	-	-	-	-	-	-	-	-	-	-13,511

Portanto, as funções do grupo 0 (NE) e grupo 1 (SbN) apresentam, respectivamente, o seguinte formato:

$$Q_0 = 0{,}004\ I_{A.2}^2 + 0{,}266\ I_{A.6}^2 + 0{,}119\ I_{B.1}^2 + 0{,}285\ I_{B.2}^2 - 3{,}135\ I_{B.3}^2 + 0{,}059\ I_{B.4}^2 + 0{,}187\ I_{C.1}^2 + 0{,}078\ I_{C.2}^2 + 0{,}051\ I_{C.4}^2 + 1{,}869\ I_{C.5}^2 - 0{,}002\ I_{A.2}\ I_{A.6} + 0{,}001\ I_{A.2}\ I_{B.1} + \ldots - 0{,}004\ I_{C.4}\ I_{C.5}$$

$$Q_1 = 0{,}003\ I_{A.2}^2 + 0{,}136\ I_{A.6}^2 - 0{,}156\ I_{B.1}^2 - 0{,}246\ I_{B.2}^2 + 1{,}905\ I_{B.3}^2 + 0{,}166\ I_{B.4}^2 + 0{,}037\ I_{C.1}^2 - 0{,}066\ I_{C.2}^2 - 0{,}026\ I_{C.4}^2 - 13{,}511\ I_{C.5}^2 - 0{,}002\ I_{A.2}\ I_{A.6} - 0{,}001\ I_{A.2}\ I_{B.1} + \ldots + 0{,}003\ I_{C.4}\ I_{C.5}$$

Cada função terá seus respectivos *scores*, sendo a curva que delimita a decisão de alocação no grupo 0 ou 1 representada pelo vetor que satisfaz o critério $Q_0 = Q_1$. Os *scores* médios das funções Q_0 e Q_1 para cada grupo estão expressos na Tabela 19, os quais representam o intervalo no qual esse critério é satisfeito para cada função, por onde a curva de separação dos grupos passará.

Tabela 19 - Médias dos Scores Q0 e Q1 por Tipo de Setor com as Classes Originais (IBGE).

Média	Grupo 0 (NE)	Grupo 1 (SbN)
Scores Q0	- 203,73	-513,05
Scores Q1	183,5	-40,72

6.3.2 Predição do Modelo Discriminante Quadrático: Classificação

Dando sequência ao estudo, após a construção da função e cálculo dos *scores* discriminantes dos setores censitários da amostra, foi realizada a predição de classes utilizando o modelo discriminante quadrático construído. As densidades de observação em cada grupo antes e depois da predição estão dispostas na Figura 16, onde é possível observar as médias dos *scores* de Q_1 (Tabela 19) projetadas sob os picos das curvas.

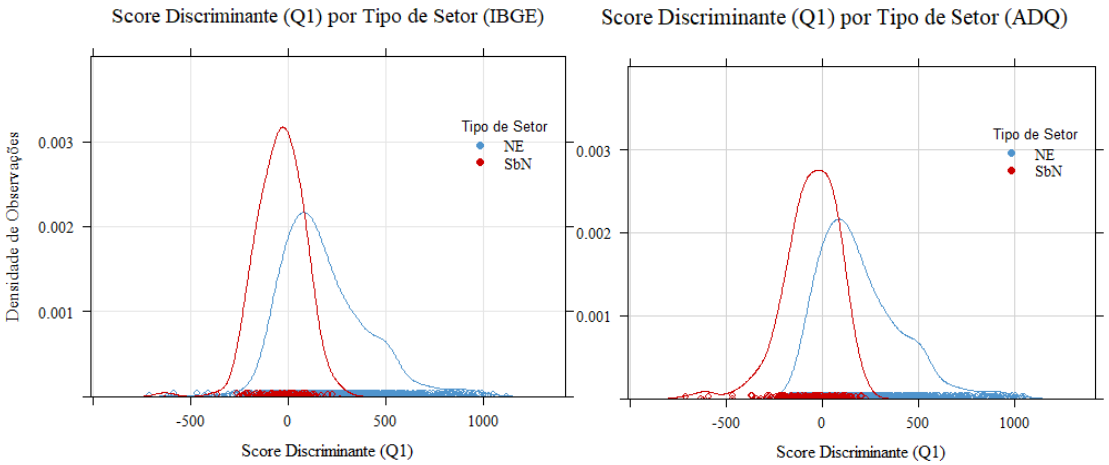

Figura 16 - Gráficos do Scores Discriminantes da ADQ (Q1) x Densidade de Observações por Tipo de Setor (IBGE e ADQ).

Embora seja possível observar uma área de interseção significativa entre os grupos, é importante pontuar que, na ADQ, esse não é um parâmetro válido para avaliar a qualidade da predição. A melhor forma de ilustrar a região de interseção, nesse caso, é pela plotagem dos *scores* das duas funções discriminantes quadráticas antes e depois da predição pelo modelo da ADQ (Figura 17).

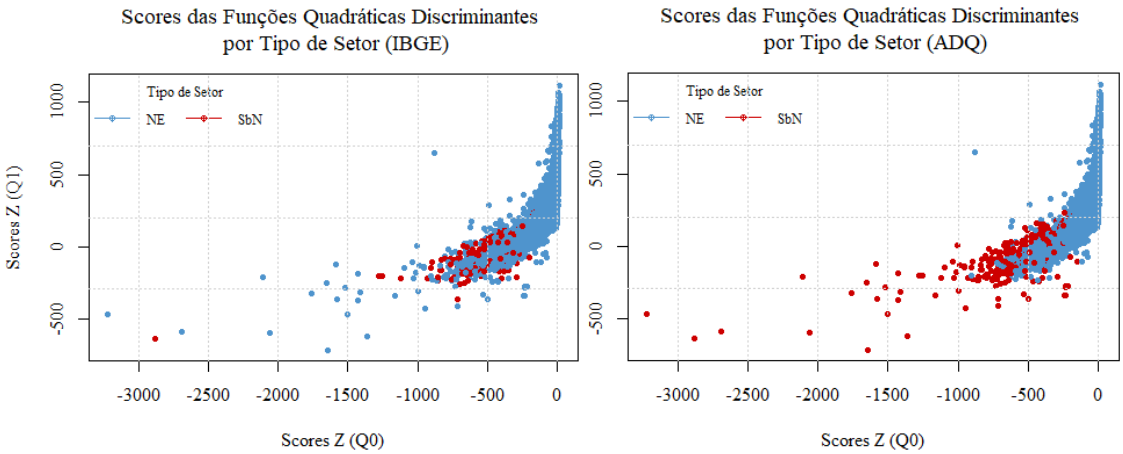

Figura 17 - Gráficos de Scores Discriminante Q1 x Q2 por Tipo de Setor (IBGE e ADQ).

Os gráficos dos *scores* das funções quadráticas Q_0 e Q_1 mostram que a região de interseção está em torno do ponto [0, -500] no gráfico, que contém valores próximos das médias para o grupo 1 nas duas funções, como descrito na Tabela 19. No gráfico referente à classificação original, verifica-se uma área de mistura entre setores NE e SbN em torno do ponto mencionado, ao passo que, após a predição, é possível observar uma separação nessa região, com maior presença de setores SbN. Os gráficos também permitem visualizar o padrão curvo da ADQ.

Para melhor visualizar a discriminação dos grupos pelo do modelo da ADQ, os *scores* foram novamente plotados em relação ao Percentual de DPP sem acesso à Esgoto (C.4) e ao Percentual de PR com Renda até 3 SM, (B.4), antes e depois da predição (Figura 18).

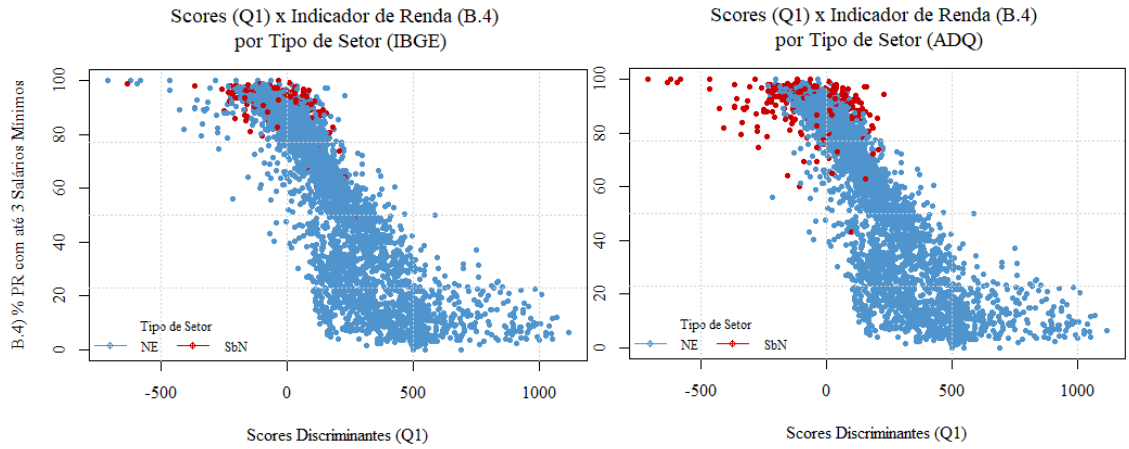

Figura 18 - Gráficos de Score Discriminante (Q1) x Indicador de Renda (B.4) por Tipo de Setor (IBGE e ADQ).

O gráfico do comportamento da renda em relação aos *scores* mostra que a técnica da ADQ permite uma maior flexibilização nos parâmetros de classificação. Os

setores reclassificados pelo modelo como SbN foram aqueles que apresentaram *scores* em torno da sua média na função Q_1, de -40. Todos os setores SbN detectados pelo modelo apresentaram valor de B.4 maior que 50, seguindo o padrão do gráfico das classes originais, apresentando, nesse sentido, melhor desempenho que a ADL.

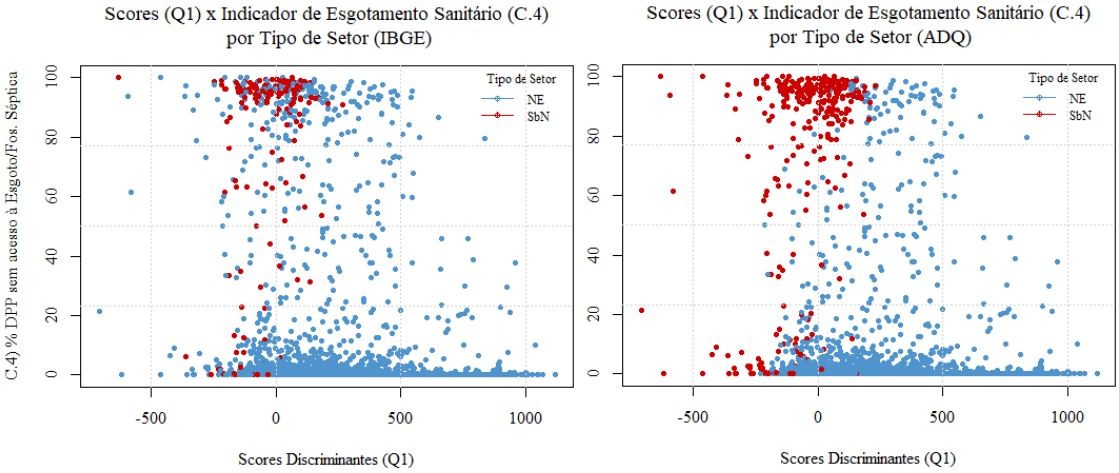

Figura 19 - Gráficos de Score Discriminante x Indicador de Esgoto (C.4) por Tipo de Setor (IBGE e ADQ).

No caso do indicador C.4, observou-se uma tendência similar, onde setores com valores menores que 0 foram reclassificados como SbN pelo modelo e aqueles presentes na região de interseção foram, em sua maioria, mantidos em suas classificações originais.

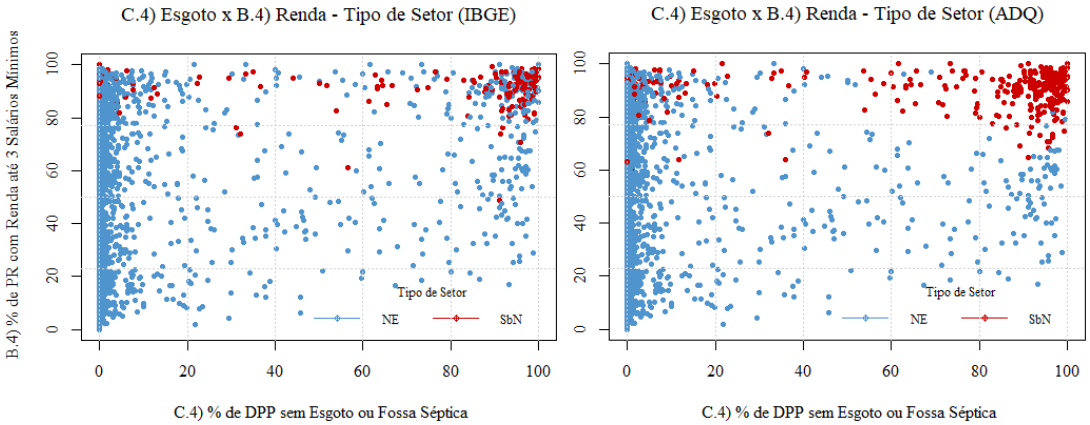

Figura 20 - Indicador B.4 x Indicador C.4 por Tipo de Setor (IBGE e ADQ).

No caso da plotagem dos indicadores B.4 e C.4 entre si, foi possível observar um padrão de concentração de setores SbN onde o Percentual de PR com renda até 3 SM estava acima de 80%, em ambos os gráficos. No caso do Percentual de DPP sem acesso à esgotamento sanitário ou fossa séptica, todos os setores classificados como SbN pelo

modelo apresentam valores acima de 60% de C.4, mostrando como esse indicador é determinante para a identificação de Assentamentos Precários, juntamente com a renda. Assim, similarmente à ADL, o comportamento padrão de setores SbN no canto superior direito do gráfico reforça, novamente, a relação da falta de acesso ao serviço de esgotamento sanitário com rendas populacionais mais baixas.

6.3.3 Avaliação da Predição do Modelo Discriminante Quadrático: Matriz de Confusão

Dos 4069 setores censitários da amostra, 352 foram classificados pelo modelo como SbN, dos quais 135 já pertenciam a esse grupo originalmente e 217 foram realocados (ver Tabela 20).

Tabela 20 - Matriz de confusão das classes originais x classes preditas pelo modelo da ADQ.

		Classificação de origem (IBGE, 2011)		
		0 (NE)	1 (SbN)	Total
Classificação pelo modelo ADQ	0 (NE)	3677	40	3717
	1 (SbN)	217	135	352
	Total	3894	175	**4069**

Dentre os setores classificados pela ADQ como SbN, 61% representam setores originalmente classificados como NE, mostrando uma quantidade significativa desses setores que apresentaram perfil semelhante ao dos SbN.

A sensitividade do modelo da ADQ foi de 0,77, ou seja, 77% dos setores SbN foram classificados corretamente, o que revela uma taxa de acerto superior à verificada no método da ADL. Uma maior taxa de acerto em relação aos setores SbN garante mais confiabilidade em relação aos setores do tipo NE que foram reclassificados como SbN – que representam um total de 217 na amostra.

Além disso, o número de setores censitários SbN reclassificados como NE também foi significativamente menor (40) se comparado ao resultado da ADL (61), mostrando assim um melhor ajuste.

6.4 Regressão Logística

A terceira e última técnica aplicada foi a Regressão Logística (RL), a qual foi incluída por ser amplamente utilizada na literatura para casos como este, em que a variável dependente é binária. O modelo logístico utilizado foi o modelo *logit*, cuja função de ligação calcula, ao invés de *scores*, probabilidades. Ou seja, os resultados

gerados pela função consistem na probabilidade de cada setor censitário ser do tipo SbN, com base nas suas respectivas características socioeconômicas e habitacionais, representadas pelas variáveis explicativas.

As etapas de aplicação da RL foram semelhantes às da AD, com exceção de uma particularidade: no uso dessa técnica, um alto desbalanceamento na base de dados pode gerar viés no modelo – situação que a literatura descreve como Eventos Raros (ER) – e, os Setores Censitários do DF para o ano de 2010, banco de dados utilizado na pesquisa, se enquadra nesse contexto, uma vez que apresenta apenas 4,3% da ocorrência do evento de interesse (setores SbN), sendo 22 vezes menor que sua não ocorrência (setores NE). Assim, foi incluído um tópico extra com a geração de um modelo RL ajustado, para fornecer uma base de comparação com o modelo construído e possibilitar a investigação de viés. Já na etapa classificação, lidar com a questão do desbalanceamento, foi utilizado o método de alteração do ponto de corte (*threshold*) para a predição das classes.

Isto posto, a aplicação da RL seguiu as seguintes etapas: seleção das variáveis independentes por procedimento Stepwise; teste dos pressupostos; construção do modelo *logit*; geração do modelo *logit* ajustado; e, por fim, a predição das classes pelo modelo.

6.4.1 Procedimento Stepwise: Seleção das Variáveis Independentes para o modelo logit

A primeira etapa da RL, similarmente à AD, consistiu na realização do procedimento Stepwise de seleção de variáveis, porém com método distinto, denominado AIC (Akaike Information Criterion), compatível com a técnica. O procedimento Stepwise foi aplicado em um modelo *logit* preliminar contendo os dezessete indicadores propostos inicialmente e gerou os resultados dispostos na Tabela 21.

Tabela 21 - Resultados do Procedimento Stepwise no modelo logit.

Indicadores	Abreviação	P valor
A.2	i_num_mor_dpp	3,03E-05
A.3	i_num_med_mor_por_dpp	0,001
A.5	i_perc_PR_sexo_fem	4,02E-05
B.1	i_perc_pr_ate_30	0,001
B.2	i_perc_PR_nalf	3,99E-05
B.5	i_Renda_Med_PR_DPP	0,001

C.1	i_perc_dpp_sem_col_lixo	1,66E-05
C.2	i_perc_DPP_sem_ab_agua	0,143
C.4	i_perc_DPP_sem_esg_fos	6,79E-39
C.5	i_num_med_ban_hab	0,004

Foram selecionadas dez variáveis explicativas, as quais foram consideradas mais relevantes para identificar o evento de interesse, isto é, os setores SbN (Tabela 21). De início, um ponto válido a ser observado é que as variáveis selecionadas utilizando o AIC diferem um pouco daquelas escolhidas no procedimento Stepwise anterior, realizado com base no Lambda de Wilks. De fato, não era esperado que os dois selecionassem exatamente as mesmas variáveis; é importante relembrar que as técnicas são teórica e matematicamente distintas. No entanto, cabe avaliar as semelhanças e diferenças entre as variáveis que compõem cada modelo.

Seis variáveis são comuns a todos os modelos, sendo que, no *logit*, a variável selecionada relativa à renda foi a Renda Média (B.5), enquanto na AD, foi o Percentual de Pessoas Responsáveis com Renda até 3 Salários Mínimos (B.4). No entanto, tendo em vista que as duas variáveis de renda explicam o mesmo fenômeno, pode-se considerar que os modelos apresentam sete variáveis equivalentes.

As duas variáveis que aparecem somente no modelo *logit* correspondem ao Número médio de moradores por domicílio (A.3) e o Percentual de Pessoas Responsáveis do Sexo Feminino (A.5), as quais não foram consideradas relevantes em termos de poder discriminante pela AD. Por outro lado, o Número de Domicílios Improvisados (A.6) e Percentual de Pessoas Responsáveis até 30 anos Não Alfabetizadas (B.3) estão apenas nas funções discriminantes e não compõem o modelo *logit*, embora se tenha o entendimento de que o indicador A.6, referente a domicílios improvisados, pode ser bastante útil à identificação de setores precários.

No entanto, de acordo com a análise da Tabela 21, nem todas as variáveis selecionadas pelo procedimento Stepwise comporão o modelo *logit*. Essa seleção ocorre pela análise do p-valor. Para fins de confiabilidade da pesquisa aqui desenvolvida, admitiu-se a significância estatística de 5%, que especifica que o p-valor deve apresentar valor máximo de 0,05. Das dez variáveis selecionadas, apenas o Percentual de Domicílios sem acesso à Abastecimento de Água (C.2) não atendeu ao critério, uma vez que apresentou p-valor de 0,143, não sendo então estatisticamente significante. No mais, todas as demais tiveram p-valor inferior a 5%.

De posse disso, o indicador C.2 foi excluído e o modelo atualizado, passando a ser composto pelas nove variáveis explicativas significantes selecionadas pelo método Stepwise. Uma vez pronto o modelo, pode-se partir para a etapa de verificação dos pressupostos da aplicação da Regressão Logística, detalhada a seguir.

6.4.2 Pressupostos da Regressão Logística

Kassambara (2018) sugere um conjunto de etapas para a avaliação da satisfação dos pressupostos. São eles:

iv. a plotagem do termo logit *versus* o valor da variável explicativa para avaliar a premissa linearidade na relação entre eles;

v. a utilização da distância de Cooks para avaliar a presença de *outliers;*

vi. o cálculo do VIF para avaliar a presença de multicolinearidade entre as variáveis explicativas.

A primeira etapa consiste na avaliação do cumprimento da premissa de linearidade. Assim, foram gerados gráficos de visualização do comportamento das variáveis independentes (indicadores) de acordo com o aumento do termo *logit*, a fim de investigar a presença de algum comportamento inusitado, que fuja muito do padrão de linearidade e impeça a aplicação do modelo (Figura 21).

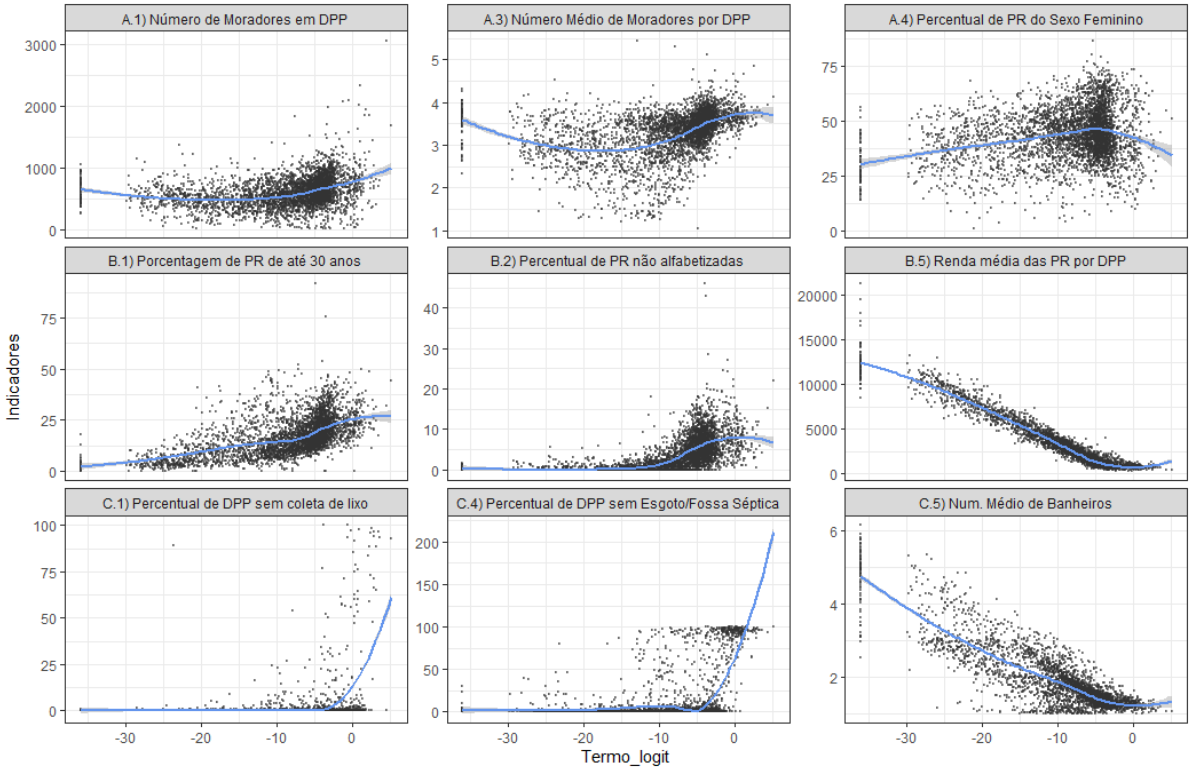

Figura 21 - Comportamento do Termo logit em relação aos Indicadores do Estudo.

Ao visualizar o comportamento das variáveis independentes, representadas pelos nove indicadores que compõem o modeo *logit*, foi possível perceber que o Percentual de PR do Sexo Feminino (A.5) se destacou pela presença de uma aglomeração de pontos significativa próximo do valor de 50%, onde o valor do termo *logit* varia entre -10 e 0. Isso indica que a variável não contribui para a predição de forma eficaz, uma vez que não apresenta uma relação clara com o aumento do termo *logit* e, assim, com a probabilidade de determinado setor ser precário.

Uma inferência possível a respeito é a de que, se o percentual de Pessoas Responsáveis do sexo feminino não apresentou linearidade com o aumento do termo *logit*, provavelmente casas chefiadas por mulheres não são necessariamente associadas a setores com perfis socioeconômicos mais vulneráveis, indicando que, nesse caso, o fator de gênero não é determinante na identificação de setores precários. De todo modo, como houve um comportamento muito discrepante da linearidade, a presença dessa variável no modelo pode agir de forma a "confundir" os critérios utilizados para a predição, motivo pelo qual optou-se por removê-la do modelo logístico. Assim, o modelo foi novamente atualizado, passando agora a conter oito variáveis independentes para, assim, seguir com os testes dos pressupostos

A segunda etapa consiste na verificação da presença de outliers. Inicialmente, foi plotado o gráfico dos resíduos *versus* as probabilidades preditas pelo modelo, onde foi identificado um único valor de resíduo muito alto (>2000), o qual foi removido para melhor visualização do gráfico de comportamento dos resíduos, obtendo assim o resultado ilustrado na Figura 22.

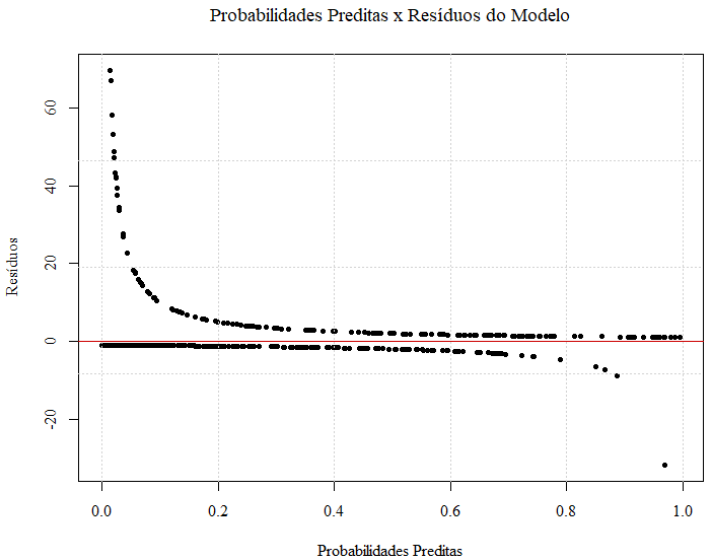

Figura 22 - Gráfico Residuos x Probabilidades Preditas.

O gráfico do comportamento dos resíduos em relação às probabilidades preditas serve como base para a análise prévia da presença de *outliers*, da homocedasticidade dos resíduos e da superdispersão dos dados. Embora os dois últimos não sejam premissas da RL, a inflação de zeros pode gerar superdispersão e muitos autores reforçam que a presença de heterocedasticidade nos modelos logísticos pode provocar efeitos negativos, como a inconsistência e enviesamento de seus parâmetros (Greene, 2003; Wooldridge, 2016), afetando a confiabilidade das inferências estatísticas. Assim, antes de partir à investigação dos *outliers* propriamente dita, esses critérios foram avaliados.

Pela observação do gráfico (Figura 22), já é possível verificar, de início, que os resíduos estão distribuídos majoritariamente no entorno da região que marca o zero no eixo Y, representada pela linha vermelha que corta o gráfico na horizontal. Esse comportamento é identificado especialmente na faixa compreendida pelos valores de probabilidade de 0,15 a 0,8 e se assemelha ao padrão homocedástico,

No entanto, as faixas de valores de probabilidade apresentados entre 0 e 0,15 e maiores que 0,8, que marcam os extremos do eixo X, abrigam pontos resíduos de valor muito alto, discrepantes do restante, fugindo do padrão da homocedasticidade. Nessas faixas, provavelmente, serão identificados *outliers*. No entanto, antes de partir à investigação dos *outliers* propriamente dita, para refinar a análise e garantir a ausência de heterocedasticidade, foi realizado um teste do pacote DHARMA (Florian Hartig, 2020) (Figura 23).

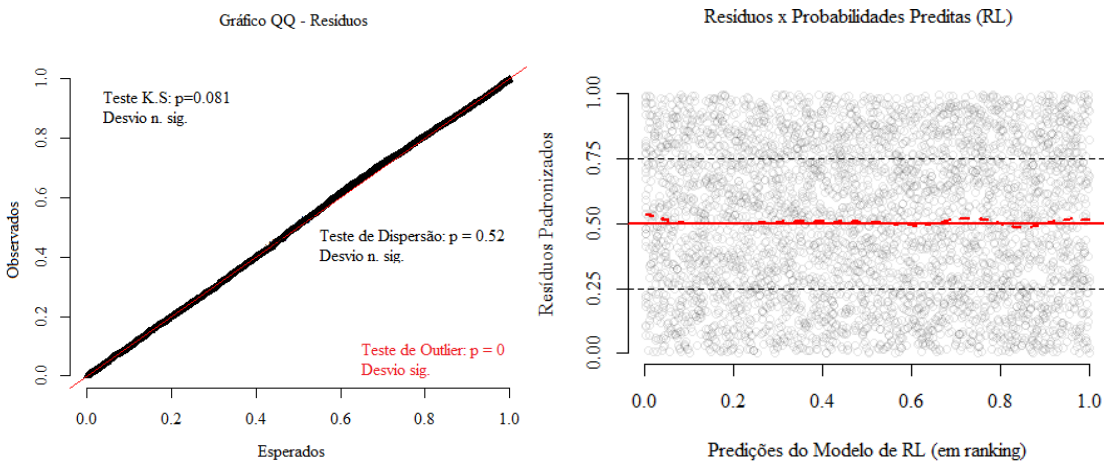

Figura 23 - Gráficos de Análise dos resíduos produzidos pelo pacote DHARMa.

O resultado do teste foi significativo apenas para a presença de *outliers*, não tendo sido detectado comportamento heterocedástico nos resíduos. A presença de

outliers já era esperada pela análise gráfica dos resíduos, onde foram identificados alguns pontos que destoaram da média encontrada. Porém, isso não pareceu interferir no comportamento geral dos resíduos, uma vez que o teste provou a homocedasticidade dos resíduos, no gráfico à direita da Figura 22.

Quanto à dispersão, apresentou p-valor de 0,5 no gráfico de desvio padrão dos resíduos ajustados vs. simulados, apontando para uma dispersão não significante. Além disso, foi calculado o valor da razão entre a estatística de Pearson Qui Quadrado e os Graus de Liberdade que, de acordo com Hilbe (2015), tende a indicar superdispersão quando maior que 1. O resultado para o modelo foi de 0,94 e, portanto, não indicativo de superdispersão. Diante disso, pode-se assegurar de que o modelo não está sujeito a sofrer fortes alterações de viés por dispersão e heterocedasticidade.

Já quanto à presença dos *outliers*, para uma melhor visualização e posterior avaliação, foi utilizada a distância de Cooks, que considera como *outliers* valores acima de 3 (Figura 24).

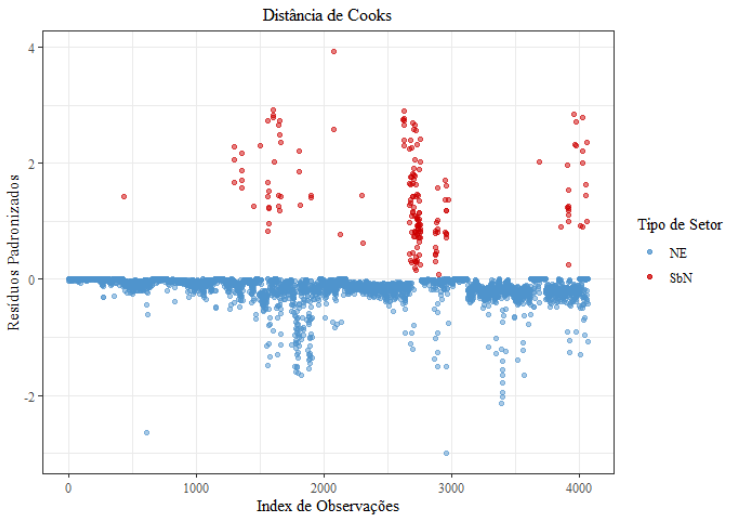

Figura 24 - Gráfico Residuos x Probabilidades Preditas.

Ao filtrar as observações que apresentaram valor maior que 3 ou menor que -3, o software apontou as observações 2076 e a 2963, correspondentes às mais afastadas da linha central do gráfico, sendo uma do setor tipo 0 e a outra do setor tipo 1. Porém, dependendo da técnica utilizada, podem ser identificados mais ou menos *outliers*. De acordo com o sugerido por Hilbe (2015), que consiste no quadrado do desvio residual padronizado *versus* as probabilidades preditas, onde valores maiores que 4 estariam acima do esperado, o banco de dados apresentaria 43 outliers (Figura 25).

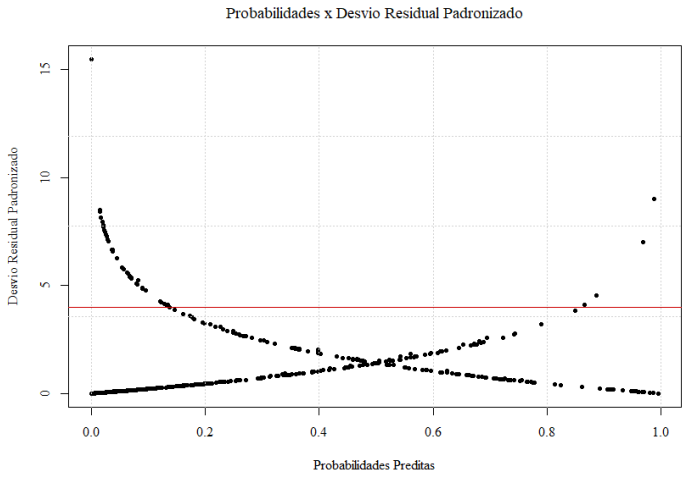

Figura 25 - Gráfico Residuos x Probabilidades Preditas.

Embora a ausência de *outliers* seja um pressuposto da Regressão Logística, era esperado que, no caso do objeto de estudo, fosse possível a ocorrência valores extremos, uma vez que os tipos de setores a analisados apresentam, muitas vezes, características opostas. Por exemplo, há alguns setores em que 100% de seus domicílios não têm acesso à coleta de lixo, enquanto outros não possuem nenhum em tal situação. Essa é a razão pela qual o banco de dados apresenta número inflacionado de zeros, fato que também tornava provável a possível presença de *outliers* no modelo. Por isso, optou-se por manter o banco de dados completo, sem a remoção dos *outliers* identificados, uma vez que que a quantidade elevada de zeros faz parte das características dos elementos estudados e retirá-los poderia acarretar uma perda de informação para o modelo.

Passando ao último pressuposto, a verificação da ausência de multicolinearidade é realizada por meio do cálculo do *Variance Inflation Factor* (VIF), cujos valores devem ser menores que 10 para garantir que não há variáveis multicolineares. Assim, foi calculado o VIF de todas as nove variáveis utilizadas no modelo, apresentadas na Tabela 22.

Tabela 22 - VIF das Variaveis Independentes do modelo logit.

Indicador	Abreviação	VIF
A.2	i_num_mor_dpp	1,07
A.3	i_num_med_mor_por_dpp	1,37
B.1	i_perc_pr_ate_30	1,37
B.2	i_perc_PR_nalf	1,38
B.5	i_Renda_Med_PR_DPP	2,07
C.1	i_perc_dpp_sem_col_lixo	1,12
C.4	i_perc_DPP_sem_esg_fos	1,09
C.5	i_num_med_ban_hab	1,59

Nota-se que, quanto à multicolinearidade, o modelo apresentou valores satisfatórios e, portanto, não precisa de ajustes. Todos os valores estão todos muito abaixo de 10, sendo o maior encontrado no B.5, referente à Renda Média, de 2,45. Conclui-se, assim, que o modelo atende ao pressuposto de ausência de multicolinearidade.

6.4.3 Estimação da Função Logística (modelo *logit*)

O modelo *logit* calcula probabilidades – e não *scores*, como é o caso da função discriminante –, formando uma curva em forma de S, dada pela função logística, que possui o seguinte formato:

$$f(Z) = \frac{1}{1 + e^{-(Z)}}$$

Sendo que:

$$Z = ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Onde p é a probabilidade de ocorrência do evento de interesse, α é o intercepto e β_i os coeficientes das variáveis independentes. O evento de interesse, na presente pesquisa, é a precariedade do setor censitário, ou seja, sua identificação como SbN, representado pela classe 1; enquanto a classe 0 representa o setor NE, cuja probabilidade de ocorrência é 1 - p.

De posse dos resultados dos testes dos pressupostos, foi possível realizar os ajustes necessários, de modo que o modelo *logit* final (Tabela 23) foi composto por oito variáveis explicativas, das dez selecionadas pelo procedimento Stepwise; uma vez que o indicador A.6 sobre o sexo feminino foi retirada por não satisfazer o primeiro pressuposto e o C.2 não foi inserido por não atingir a significância estatística de 5%.

Tabela 23 - Modelo logit final.

Indicador	Abreviação	Coeficiente	Razão de Chances	Desvio Padrão	Z-valor	P-valor
C	Intercepto	-2,733	-	1,9177	-1,43	1,54E-01
A.2	i_num_mor_dpp	0,002	1,00	0,0004	4,37	1,23E-05
A.3	i_num_med_mor_por_dpp	1,297	3,66	0,3962	3,27	1,06E-03
B.1	i_perc_PR_ate_30	0,028	1,03	0,0158	1,74	8,10E-02
B.2	i_perc_PR_nalf	-0,121	0,89	0,0279	-4,32	1,57E-05
B.5	i_Renda_Med_PR_DPP	-0,001	0,99	0,0004	-2,74	6,06E-03
C.1	i_perc_dpp_sem_col_lixo	0,033	1,03	0,0062	5,33	9,65E-08
C.4	i_perc_DPP_sem_esg_fos	0,034	1,03	0,0024	14,05	7,78E-45
C.5	i_num_med_ban_hab	-4,083	0,02	0,9414	-4,34	1,45E-05

Desvio: 1443.59 (4068 g.l.); Desvio do Resíduo: 702.71 (4060 g.l.); AIC: 720.71.

Os coeficientes do modelo logístico somados ao intercepto formam a parte representada pelo Z na função logística, de termo *logit* ou *logito*, essencial para o cálculo das probabilidades. A função logística, responsável por calcular tais probabilidades, está expressa na equação X, cujo termo *logit* apresenta os coeficientes das oito variáveis que compõe o modelo.

$$P = \frac{1}{1 + e^{-(-2{,}7 + 0{,}002\,A.2 + 1{,}3 + 0{,}03\,A.3 + 0{,}028\,B.1 - 0{,}12\,B.2 - 0{,}001\,B.5 + 0{,}03\,C.1 + 0{,}034\,C.4 - 4{,}08\,C.5)}}$$

A interpretação dos coeficientes β da função logística nos permite fazer inferências estatísticas, importantes para a análise das relações entre as variáveis independentes e a ocorrência do evento de interesse. Assim, o valor de cada coeficiente revela relações entre o seu indicador e o fenômeno estudado, as quais devem ser analisadas para verificar se apresentam coerência com o conhecimento acadêmico e empírico sobre o tema. Isso porque, evidentemente, as informações levam em consideração apenas as informações que compõe o modelo, de modo que é provável que existam fatores externos ao modelo que interfiram nessas relações, motivo pelo qual essa análise deve ser realizada com ressalva, especialmente ao analisar um fenômeno tão complexo quanto o dos Assentamentos Precários.

É importante destacar que, ao analisar os coeficientes de um modelo, é possível que se encontre relações inesperadas, tanto nos sinais dos coeficientes quanto nos efeitos expressos pelas razões de chances, demonstrando resultados que não condizem com o que é observado empiricamente e na literatura. Situações como essa podem ser resultado de heterocedasticidade, viés por amostras desbalanceadas, omissão de variáveis relevantes, alta variância, presença de valores extremos, dentre outros (Kennedy, 2002). Como o modelo aqui estudado se trata de um modelo com amostra consideravelmente desbalanceada e presença de *outliers*, tais ocorrências podem ser atribuídas principalmente a esses fatores.

No caso aqui estudado, a situação mencionada acima foi encontrada em apenas uma das variáveis explicativas do modelo: o Percentual de Pessoas Responsáveis não alfabetizadas (B.2), cujo coeficiente apresentou, inesperadamente, sinal negativo, indicando relação inversamente proporcional da variável com a probabilidade de um

setor ser SbN. Tal relação não condiz com o que foi proposto pelo estudo do CEM (Marques et al., 2007), que utilizou esse parâmetro como indicador de áreas precárias, partindo da premissa que a população de baixa renda apresenta maior suscetibilidade à falta ou precariedade do ensino formal. Assim, a relação esperada era de que setores censitários com maior Percentual de Pessoas Responsáveis não alfabetizadas tivessem maior probabilidade de serem precários, contrariando o que foi sugerido pelo modelo. No entanto, o sinal não foi só negativo como apresentou a terceira maior magnitude no efeito causado na chance de o setor ser SbN, de 11%.

A esse resultado, pode-se atribuir a omissão de alguma variável explicativa correlacionada importante para o modelo. No estudo realizado pelo CEM (2007), foram utilizadas também informações referentes à escolaridade e anos de estudo, mas tais variáveis, presentes no Censo Demográfico de 2000, já não estavam mais disponíveis em 2010. No entanto, existem muitas causas possíveis para esse resultado inesperado além da omissão de escolaridade e ano de estudo, como idade, raça e até fatores até então desconhecidos.

É provável que esse fator externo ao modelo influencie nessa variável de modo que ela tenha que subtrair da probabilidade calculada pelo modelo, uma vez que, avaliando seu comportamento graficamente, os valores do indicador B.2 na verdade aumentam com a probabilidade do setor ser SbN até um determinado valor e depois se mantém constantes (Figura 21).

Quanto aos coeficientes das demais variáveis, todos expressaram relações coerentes, em consonância com o objetivo da pesquisa. Mais duas apresentaram coeficientes com sinais negativos – Número Médio de Banheiros (C.5) e Renda Média (B.5) –, porém, nesses casos, a relação inversa era esperada. Quanto menos banheiros por habitante e quanto menor a renda média de determinado setor, maior a probabilidade de ser classificado como precário.

Em termos de magnitude, observa-se a razão de chances, interpretada em fator. Por exemplo, de acordo com a Tabela 23, quando o Percentual de Domicílios sem acesso à Coleta de Esgoto (C.4) aumenta em uma unidade, a chance de um setor ser SbN passaria a ser 1,03 vezes maior, ou aumentaria em 3% (subtraindo 1 da razão de chances e multiplicando por 100).

Verifica-se que o número médio de banheiros apresentou o segundo maior impacto na chance de um setor ser SbN em termos de magnitude, alterando-a em 98%, ao passo que a redução em uma unidade na renda média aumenta essa chance em 0,1%.

Isso se dá pelo fato de a unidade da renda estar em reais e, de fato, uma mudança de apenas um real na renda média de um setor não gera grande impacto. No entanto, se colocado em escala maior, tem-se que uma redução de mil reais na renda média causaria um aumento de 100% na chance de um setor ser precário.

O indicador que apresentou maior destaque foi o Número Médio de Moradores por Domicílio (A.3), cujo aumento de uma unidade faria a chance de ocorrência do evento de interesse crescer em 308%. As duas maiores razões de chances ocorreram nos indicadores que estavam em unidades de pessoas e reais, respectivamente, justamente porque o aumento está em termos absolutos. As demais variáveis apresentaram menores razões de chances por estarem em unidades percentuais, como foi o caso do Percentual de Domicílios sem acesso à Coleta de Esgoto (C.4), à Coleta de Lixo e do Percentual de Pessoas Responsáveis até 30 anos (C.1), cujos aumentos marginais acresceriam a chance em 3% cada. Isso não significa que eles não tenham efeitos significativos sobre o evento de interesse, pelo contrário, revela que se apenas um porcento a mais de todo o setor censitário tiver carência no acesso a esses serviços e/ou tiver pessoas responsáveis até 30 anos, isso já afeta a chance de precariedade do setor com um aumento percentual da ordem de unidades.

Assim, observou-se coerência nas relações expressas nas razões de chances de todas as variáveis do modelo *logit*, excetuando-se apenas o Percentual de Pessoas Responsáveis não Alfabetizadas. Dessa forma, considera-se que o modelo logístico pode apresentar fragilidade na utilização das inferências estatísticas para afirmar o comportamento de determinado indicador e sua relação com o evento de interesse. Como já mencionado, isso pode ter sido impacto da amostra desbalanceada, da presença de *outliers* ou mesmo da omissão de alguma variável no modelo.

Para fazer uma breve análise da possível causa da incoerência encontrada, foi gerado um modelo ajustado com uma subamostra balanceada, pelo método proposto por King e Zeng (2001).

6.4.4 Geração do Modelo Logístico pelo Método de King e Zeng para Amostras com Eventos Raros

Com o intuito de realizar uma investigação mais acurada do possível viés causado pelo desbalanceamento da amostra, foi gerado um modelo *logit* ajustado utilizando o método de ajuste de parâmetros com correção a priori proposto por King e Zeng (2001), onde foi realizado o rebalanceamento da amostra por meio do

undersampling ou subamostragem, que consiste no descarte aleatório de zeros de modo a alcançar proporções menos discrepantes entre zeros (setores NE) e uns (setores SbN).

Foram descartados descartar 77,5% dos zeros da amostra, resultando em uma subamostra composta por cinco vezes mais zeros que uns, com 1050 observações no total, formada pelos 175 setores SbN e por 875 setores NE escolhidos randomicamente. Assim, foi estimado o modelo *logit* com a correção de viés, utilizando o pacote *Zelig*, criado pelos próprios autores que propuseram o método (Tabela 24).

Tabela 24 - Modelo Logit pelo Método de King e Zeng.

Indicador	Abreviações	Coeficientes	Razão de Chances	Desvio Padrão	z-valor	p-valor
C	Intercepto	-1,093	-	2,479	-0,44	0,659
A.2	i_num_mor_dpp	0,001	1,001	0,000	2,41	0,016
A.3	i_num_med_mor_por_dpp	0,620	1,860	0,514	1,21	0,227
B.1	i_perc_PR_ate_30	0,023	1,023	0,020	1,11	0,266
B.2	i_perc_PR_nalf	-0,070	0,932	0,036	-1,92	0,054
B.5	i_Renda_Med_PR_DPP	-0,001	0,999	0,001	-2,33	0,020
C.1	i_perc_dpp_sem_col_lixo	0,027	1,027	0,010	2,81	0,005
C.4	i_perc_DPP_sem_esg_fos	0,031	1,032	0,003	10,30	<2e-16
C.5	i_num_med_ban_hab	-3,444	0,032	1,154	-2,99	0,003

Como pode-se observar na Tabela 24, os indicadores apresentaram comportamento similar ao modelo *logit* anterior, com valores muito próximos de razão de chances, com exceção do número médio de moradores por domicílio que, nesse modelo, foi de 1,8 – metade do valor do modelo *logit* com amostra desbalanceada –, provavelmente por conta da redução da amostra.

Por essa razão, optou-se por não realizar predições com o modelo gerado pela sub amostragem. No entanto, sua geração foi útil para auxiliar na investigação do possível viés que poderia estar causando erros nas inferências estatísticas, o que fragilizaria sua utilização. Ao descartar o viés por amostras desbalanceadas, depreende-se que o modelo gerado pelo método anterior mostra ser o mais adequado, já que utiliza todas as informações disponíveis.

6.4.5 Predição do Modelo Logístico: Classificação pelo Método de alteração do *threshold*

As predições foram realizadas pela função *predict* do pacote stats, utilizando uma alteração no *treshold* do modelo, ou seja, no valor de probabilidade a partir do qual o modelo deve considerar um setor como SbN. Tradicionalmente, na literatura, o limite (*treshold*) do valor de probabilidade utilizado para classificar determinada observação

na classe de interesse é de 0,5. No entanto, diante de uma amostra desbalanceada, Favero et al. (2009) e Maalouf e Trafalis (2011) sugerem o deslocamento desse valor para a fração de ocorrência do evento de interesse na amostra como uma forma de ajuste da predição ao desbalanceamento da amostra. Ou seja, para o caso estudado, foi utilizada a razão entre 175 e 4069, que corresponde a 0,043.

A Regressão Logística não calcula *scores* e sim probabilidades, portanto, sua visualização é realizada plotando o comportamento das probabilidades preditas, cujo comportamento apresenta a forma de sigmoide, juntamente aos tipos de setor de acordo com os valores dos indicadores. Foram gerados gráficos para os indicadores de Renda (B.5), Esgoto (C.4), Coleta de Resíduos Sólidos (C.1) e Número médio de banheiros por habitante (C.5).

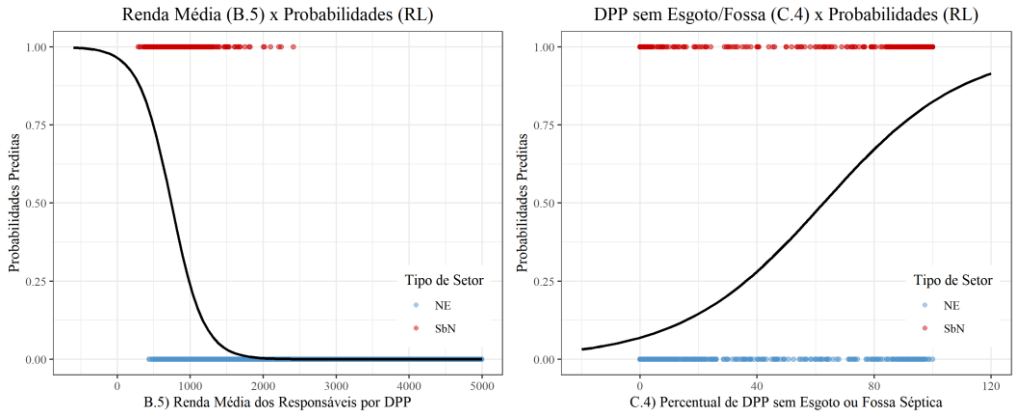

Figura 26 - Comportamento das Probabilidades Preditas (RL) x Indicadores B.5 e C.4 por tipo de setor.

Conforme observado na Figura 26, a curva sigmoide é crescente ou decrescente a depender da relação entre o indicador e da ocorrência do evento de interesse. No caso de B.5, o coeficiente do indicador é negativo, como mostra a Tabela 24, o que significa que quanto maior a renda média mensal dos responsáveis por DPP, menor a probabilidade de um setor ser SbN. Os pontos são plotados de acordo com o valor do seu indicador, mostrando que as receitas médias dos setores do SbN estão concentradas em valores de até R $ 2.500 reais. Todos os pontos acima desse valor foram classificados como NE.

Em relação a C.4, é possível observar que é diretamente proporcional à probabilidade de um setor ser SbN. No entanto, existem setores em ambos os grupos ao longo de todo o eixo X, o que significa que, embora a falta de acesso à coleta de esgoto ou fossa séptica seja um indicador relevante para classificar um setor como SbN pelo modelo logístico, não é determinante quando analisado sem as demais características.

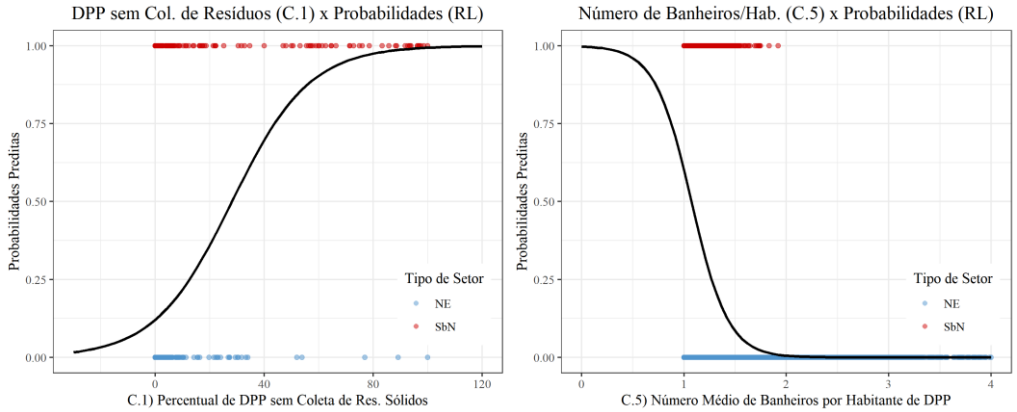

Figura 27 - Comportamento das Probabilidades Preditas (RL) x Indicadores C.1 e C.5 por tipo de setor.

A Figura 27 mostra que a grande maioria dos setores NE apresenta baixo percentual (40% no máximo) de domicílios sem coleta de resíduos sólidos, enquanto a maior parte dos setores SbN apresenta alto percentual neste indicador. Quanto ao número médio de banheiros por habitante, não há setores SbN com mais de 2 banheiros, enquanto nos setores NE, o indicador chega à marca de 6. Portanto, percebe-se que menos de dois banheiros por habitante e falta de acesso à coleta de resíduos sólidos acima de 40% são encontrados principalmente em setores SbN, o que indica que, para o modelo, essas características aumentam muito a probabilidade de ocorrência do evento de interesse.

O mesmo gráfico de receita e esgoto gerado para LDA e QDA é apresentado para LR, embora o indicador de renda não tenha sido o mesmo para as técnicas - as funções da LDA e QDA continham o indicador B.4. No entanto, ainda é possível visualizar sua relação antes e depois da classificação (Figura 28).

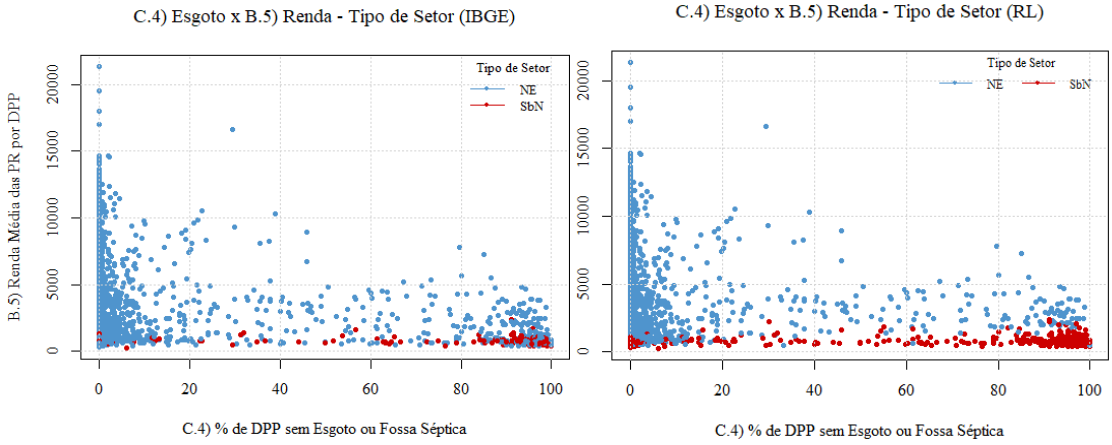

Figura 28 - Gráficos do Indicador de Acesso à Coleta de Esgoto (C.4) x Indicador de Renda (B.5) por Tipo de Setor (IBGE e RL).

A configuração dos grupos nos gráficos dos indicadores C.4 e B.5 mostram, com clareza, a concentração dos setores censitários SbN no canto inferior direito dos gráficos, que indicam menores rendas – abaixo de 2.500 – e maiores taxas de falta de acesso, ao passo os setores NE estão aglomerados de maior cobertura desse serviço, entre os valores de 0 e 20 do eixo X. Isso reforça a combinação dos fatores renda e acesso a esgotamento sanitário na definição do perfil de um setor SbN, como também verificado nas predições dos demais modelos.

6.4.6 Avaliação da Predição do Modelo Logístico: Matriz de Confusão

Como já mencionado, a predição de classes pelo modelo *logit* foi efetuada de modo que todas as observações que apresentaram probabilidade maior que 0,043 foram classificadas como SbN (classe 1) e, caso contrário, como NE (classe 0). Os resultados obtidos estão dispostos na Tabela 25.

Tabela 25 - Matriz de confusão do modelo logit.

		Classificação de origem (IBGE, 2011)		
		0 (NE)	1 (SbN)	Total
Classificação pelo modelo *logit*	0 (NE)	3502	20	3522
	1 (SbN)	392	155	547
	Total	3894	175	**4069**

Os resultados das predições do modelo logístico mostraram a classificação de 547 setores precários no DF, sendo 155 deles já previamente classificados como tal pelo IBGE, de modo que foram adicionados 392 setores cujo perfil populacional foi considerado similar ao dos moradores de setores SbN. Assim, 155 setores foram classificados corretamente como SbN, resultando em um percentual de 88,5% de acerto, o maior dos três modelos testados. Além disso, apenas 20 setores originalmente SbN foram classificados erroneamente como NE, também o menor valor entre os modelos analisados.

6.4.7 Ajuste do modelo

Para avaliar o ajuste do modelo, foram aplicados dois testes, a fim de garantir que as predições realizadas podem ser consideradas relevantes para o estudo: o teste de Pearson Qui-quadrado, que avalia se as probabilidades preditas desviam das probabilidades observadas de forma que a distribuição binomial não prediz; e o teste de

Hosmer-Lemeshow, que identificar se existe diferença entre os valores observados e preditos da variável dependente.

O resultado do teste de Hosmer Lemeshow (X-squared = 11.301, df = 8, p-value = 0.1852) apresentou p-valor superior a 0,5, indicando que a hipótese nula não foi rejeitada, indicando que o modelo está bem ajustado. O resultado do teste de Pearson Qui-quadrado (Chi2 = 3816.6222, df = 4060, p-value = 0.997) corroborou com essa hipótese, uma vez que quanto mais próximo de 1 estiver o valor de p, melhor a qualidade de ajuste do modelo, rejeitando a hipótese nula de que há desvios inesperados entre as probabilidades observadas e preditas do modelo.

É importante destacar, todavia, que o fato de os testes não terem rejeitado as hipóteses nulas indica apenas que não encontraram nenhuma evidência de que essas hipóteses estivessem incorretas. Ou seja, embora não tenha sido rejeitada a hipótese de que o modelo não apresenta função de ligação incorreta, termo de ordem mais alta omitido para variáveis no modelo, preditora omitida e super dispersão; não se pode afirmar que o modelo certamente não apresentará nenhum dos problemas descritos. No entanto, há, de fato, um indicativo de que o modelo estimado se ajusta aos dados a um nível satisfatório.

6.5 Comparação dos Resultados da ADL, ADQ e RL

Após a aplicação das três técnicas de estatística multivariada, buscou-se identificar as vantagens e desvantagens do uso de cada uma com base nos seus desempenhos preditivos e nos obstáculos encontrados ao longo da pesquisa. O primeiro critério considerado se trata da curva ROC (*Receiver Operating Characteristic*), que consiste na representação gráfica da sensitividade versus (1- especificidade) (Figura 29)

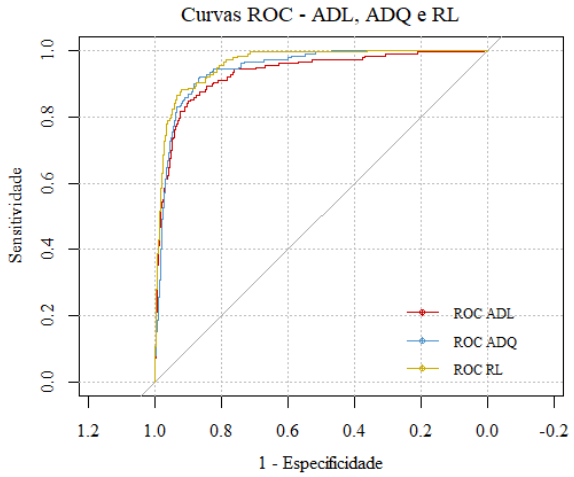

Figura 29 - Curva ROC dos modelos da ADL, ADQ e RL.

A sensitividade do modelo corresponde à taxa de acertos de setores do grupo 1 (SbN) e a especificidade, a taxa de acerto de setores do grupo 0 (NE). O primeiro fator de análise da curva ROC é a chamada AUC (*área under the curve*), a qual representa a área compreendida pela curva e varia entre 0 e 1, indicando que, quanto mais próxima de 1, melhor desempenho preditivo do modelo, uma vez que são maximizadas as duas taxas de acerto.

O segundo critério foi o valor da sensitividade somente, uma vez que a taxa de acerto dos setores do tipo SbN representam o principal método de ajuste do modelo utilizado pelo CEM (2007), pois quanto mais acurado o modelo tiver sido na detecção de setores SbN, mais confiáveis serão as predições que realocaram setores NE na classe dos SbN, indicando que aquele determinado setor apresenta perfil similar aos pertencentes a áreas de Assentamento Precário. Esse perfil é traçado por meio dos indicadores do modelo, os quais cumprem papel de extrema importância, uma vez que representam toda a base de características cujo modelo se baseia para guiar as decisões de classificação realizadas pela predição. Dessa forma, também foi considerado o número de indicadores de cada modelo. As informações descritas acima estão detalhadas na Tabela 26, separadas por tipo de técnica aplicada.

Tabela 26 - Métricas utilizadas para Avaliação e Comparação dos Modelos Estatísticos.

Medidas	ADL	ADQ	RL
AUC (Área sob a Curva ROC)	0,931	0,947	0,961
Taxa de Acerto de Setores SbN (%)	66%	77%	88,5%
Número de Variáveis Independentes	10	10	8

Como é possível verificar, as predições realizadas pelos três modelos apresentaram alto desempenho no que se refere à AUC, atingindo valores superiores a 0,9. O percentual de setores SbN classificados corretamente pelo modelo da ADL foi superior na RL, com 88%, seguido da ADQ, com 77% e por fim, a ADL, com 66%. De acordo com essa taxa, o modelo de maior capacidade preditiva foi da RL, ainda que todos tenham apresentados taxas de acerto satisfatórias, levando em consideração que o CEM (2007), quando realizou estudo similar para o ano de 2000, atingiu uma taxa de 52% de acerto para a região do DF e RM de Goiânia.

Já em relação ao número de indicadores, os dois modelos da AD tiveram a mesma estrutura, enquanto o da RL diferiu por incluir o número médio de moradores por DPP (A.3) e excluir os indicadores relativos ao Percentual de PR de até 30 anos não Alfabetizadas (B.3), o Número de Domicílios Improvisados (A.6) e o Percentual de DPPs sem Abastecimento de Água (C.2). Quanto ao último, a análise da estatística descritiva mostrou que era esperado que essa variável não apresentasse impacto significativo na identificação de setores SbN, uma vez que o DF possui alto nível de cobertura desse serviço. Já o B.3 não é essencial à análise, uma vez que há outro indicador que fornece informações sobre PR não alfabetizadas, ao passo que o A.6, referente aos domicílios improvisados, é único indicador que pode ter acarretado certa perda de informação para a RL, uma vez que contém um dado útil à caracterização de setores SbN. Além disso, apresentou, no modelo da ADL, o mesmo peso (carga discriminante) de B.4, indicador de renda, os quais configuraram, juntos, o segundo maior peso, perdendo apenas para o indicador relativo ao acesso à esgotamento sanitário (C.4). No entanto, ainda assim, o modelo da RL efetuou a melhor predição.

Após a análise dos critérios numéricos, os pontos positivos e negativos da aplicação das três técnicas foram sumarizados na Tabela 27, contendo, nos pontos positivos, os pontos elucidados nesse tópico e, nos negativos, as dificuldades encontradas no cumprimento de todas as premissas de cada técnica, já detalhadas nos tópicos que descreveram suas aplicações.

Tabela 27 - Pontos Positivos e Pontos Negativos de cada Método Estatístico.

Modelos	Pontos Positivos	Pontos Negativos
ADL	- Técnica utilizada pelo CEM - Contém o indicador A.6	- Assumiu-se a normalidade - Violação de Premissa de Homog. de Covariâncias
ADQ	- Maior taxa de acerto que ADL - Flexibiliza a Premissa de Homogeneidade das Matrizes de Covariância - Contém o indicador A.6	- Assumiu-se a normalidade
RL	- Maior percentual de acerto de setores SbN	- Violação de Premissa da Ausência de Outliers - Possibilidade de Viés por Eventos Raros

Em resumo, verificou-se que a ADL é a técnica menos recomendada dentre as três, não só por apresentar menor sensitividade, mas principalmente pela violação da premissa de Homogeneidade das Matrizes de Covariância que, embora possa ser relativizada pelo tamanho da amostra, assumir que ambos os tipos de setor terão a

mesma covariância quando não têm pode resultar em generalizações que têm potencial de comprometer as predições.

Já a ADQ é mais recomendada, uma vez que, além de flexibilizar esse pressuposto, mostrou melhor desempenho na separação dos grupos. No entanto, ambas as técnicas exigem a normalidade multivariada, a qual nem sempre é encontrada em bancos de dados reais, especialmente em situações que apresentem indicadores com valores extremos, como é o caso das características dos Assentamentos Precários.

A RL, por sua vez, também apresentou uma violação de premissa, uma vez que foram encontrados três *outliers* no banco de dados. Porém, a presença de valores extremos não foi determinante no desempenho da predição do modelo. Assim, os achados da presente pesquisa apontam para uma recomendação da RL como técnica mais efetiva na identificação de setores censitários com perfil semelhantes aos SbN, correspondendo, portanto, aos Assentamentos Precários.

Corrobora-se, portanto, com Muaualo (2013), que concluiu que a Regressão Logística é mais eficiente que a Análise Discriminante de dois ou mais grupos, exceto quando as variáveis explicativas satisfazem suas premissas, em que a Análise Discriminante se torna assintoticamente mais eficiente que a Regressão Logística. Finalmente, portanto, o modelo da RL selecionado para o seguimento da análise e medida do acesso aos serviços de abastecimento de água e esgotamento sanitário em Assentamentos Precários no Distrito Federal.

Espacialmente, é possível visualizar os resultados de classificação das três técnicas na Figura 30, comparadas à classificação realizada pelo IBGE. Em seguida, tem-se, com mais detalhes, a visualização dos resultados da Regressão Logística (Figura 31), os quais foram considerados como mais acurados para a identificação de Assentamentos Precários.

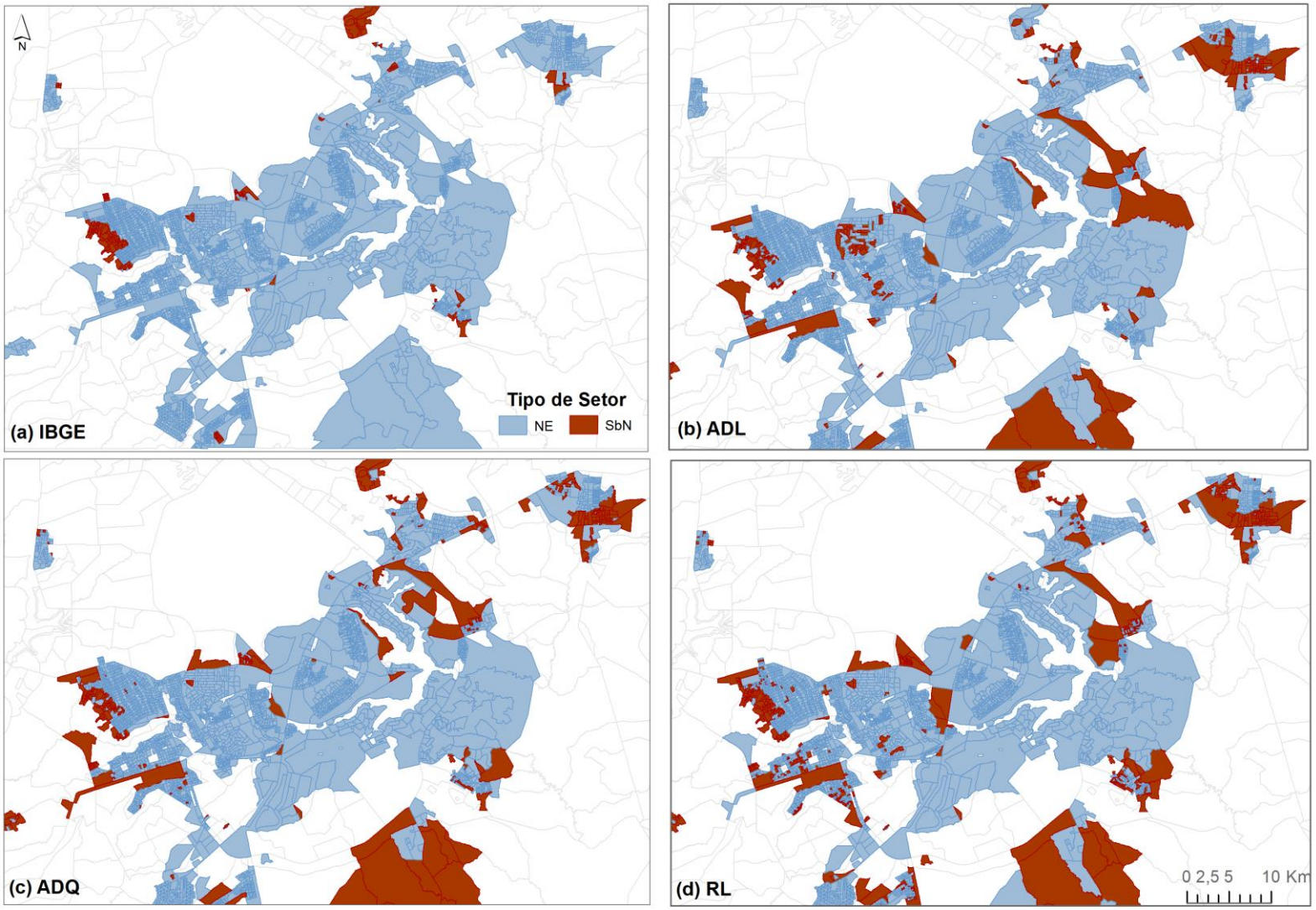

Figura 30 - Setores SbN no DF de acordo com o IBGE (a) e com as predições da ADL (b), ADQ (c) e RL (d).

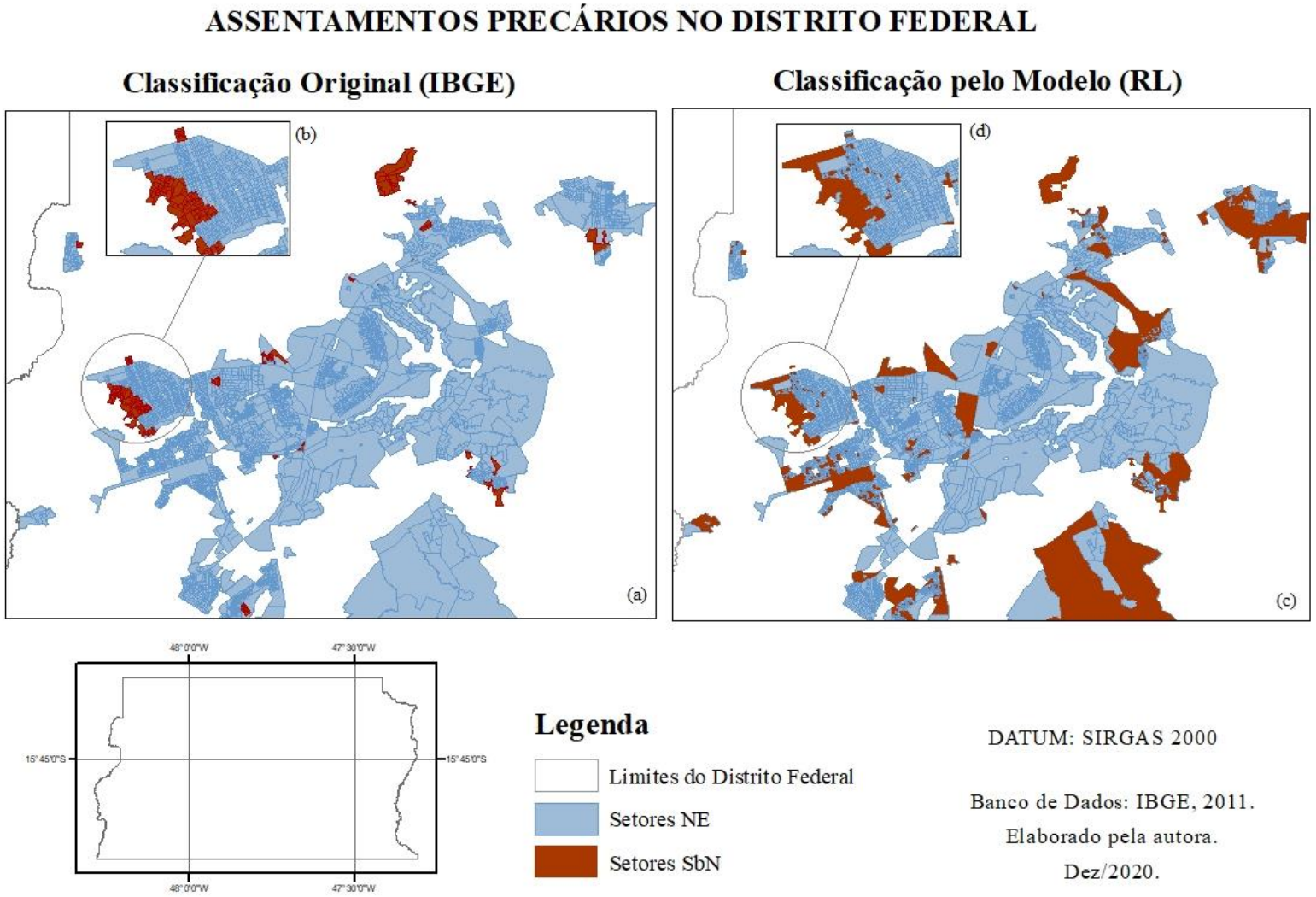

Figura 31 - Assentamentos Precários no DF: classificação realizada pelo IBGE (a e b) e classificação realizada pelo modelo de RL (c e d).

6.6 Cálculo do Indicador de Universalização Inclusiva: Acesso aos Serviços de Água e Esgotamento Sanitário nos Assentamentos Precários

Após o estudo das técnicas de estatística multivariada, foi selecionado o modelo de Regressão Logística para apresentação dos resultados encontrados, por ter apresentado melhor desempenho preditivo. Assim, foram considerados como Assentamentos Precários os setores identificados pelo modelo como tal, somados àqueles previamente classificados como SbN pelo IBGE, mas que porventura não tenham sido detectados pelo modelo. Dessa forma, a base final foi chamada de Setores Precários, composta por 567 setores censitários: sendo 175 classificados previamente pelo IBGE e mais 392 realocados da classe 0 (NE) para a classe 1 (SbN) pelo modelo de Regressão Logística. Assim, foi possível calcular as taxas percentuais de acesso aos serviços de abastecimento de água e esgotamento sanitário, objetivo que motivou a proposição do método desenvolvido.

Conforme os resultados obtidos, o Distrito Federal apresentou 14% de seus setores censitários urbanos como precários e não mais 4,3%, como mapeado pelo IBGE, confirmando a hipótese de que os Assentamentos Precários nas áreas urbanas brasileiras são subestimados se analisados somente com base no mapeamento de AGSN. Os 567 setores mapeados configuram, portanto, todas as áreas representativas de Assentamentos Precários no DF e compreenderam 422.904 pessoas em 115.944 domicílios, representando cerca de 15,4% dos DPP e 17% da população moradora em DPP no DF em 2010.

Assim, partindo da base de Setores Precários do DF para o ano de 2010 resultante do estudo, foram calculados os indicadores de Universalização Inclusiva (UI), formulados por Guimarães (2015), os quais consistem nos percentuais de domicílios com acesso aos serviços de abastecimento de água e esgotamento sanitário dentro dos Assentamentos Precários – ou de áreas de vulnerabilidade socioeconômica em geral. Em suma, os indicadores de UI buscam fornecer uma perspectiva mais realista da dimensão da carência dos serviços de abastecimento de água e esgotamento sanitário para a parcela mais vulnerável da população, evitando uma possível camuflagem desses dados, característica de quando as taxas de acesso são observadas para a população em geral. Foram gerados também, a título de comparação, os percentuais de acesso somente para os setores NE e também para todos os setores.

Os resultados obtidos foram expressos primeiramente em relação à população (Quadro 9), seguidos do cálculo dos índices por domicílio (Quadro 10), que inclui a obtenção do Indicador de UI proposto por Guimarães (2015).

Quadro 9 - População com Acesso a Abastecimento de Água e Esgotamento Sanitário ou Fossa séptica no DF em 2010, por Tipo de Setor classificado pela Regressão Logística.

Unidade de Análise	**População Total**	**População com Acesso à Abastecimento de Água pela Rede Geral**		**População com Acesso à Rede de Esgotamento Sanitário ou Fossa Séptica**	
		Quantidade	Percentual	Quantidade	Percentual
Setores Precários	422.904	404.125	95,6 %	246.786	58,4 %
Setores NE	2.053.450	2.001.308	97,5 %	1.974.028	96,1 %
Todos os Setores	2.476.354	2.405.433	97,1 %	2.220.814	89,7 %

Com base no Quadro 9, tem-se que, em 2010, o DF possuía 70.921 pessoas sem acesso à rede geral de abastecimento de água e 255.540 pessoas sem acesso à rede de esgotamento sanitário ou fossa séptica, das quais 18.779 e 176.118 eram habitantes de Assentamentos Precários, respectivamente. Os percentuais mostram que os índices de acesso foram menores em setores SbN mesmo no serviço de abastecimento de água, que cobre mais de 97% da população do DF.

No que se refere ao Indicador de UI, este se baseia no percentual de acesso dos domicílios e está expresso no Quadro 10, juntamente aos índices de acesso dos domicílios dos setores NE e no DF como um todo.

Quadro 10 - Índices de Acesso dos Domicílios aos Serviços de Abastecimento de Água e Esgotamento Sanitário e Cálculo do Indicador de Universalização Inclusiva.

Indicadores de Acesso	**Abastecimento de Água**	**Esgotamento Sanitário**
Percentual de Acesso Geral (em todos os Setores)	97,3%	90,6%
Percentual de Acesso em Setores NE	97,6%	96,5%
Percentual de Acesso em Setores Precários (Indicador de Universalização Inclusiva)	95,5%	58,5 %

A partir dos resultados obtidos, observa-se que os indicadores de acesso encontrados corroboram com o entendimento de Guimarães (2015), quando afirma que a situação de acesso aos serviços de saneamento básico em áreas urbanas de maior vulnerabilidade socioeconômica é mascarada quando os dados são disponibilizados para toda a população, uma vez que mostraram que, embora o DF tenha apresentado bons

índices de acesso em 2010, a universalização ainda segue sendo um desafio em Assentamentos Precários, especialmente no caso do esgotamento sanitário.

Verifica-se que 96,5% dos domicílios de setores NE e 90,6% dos domicílios de todos os setores urbanos do DF apresentaram ligação à rede de coleta de esgoto ou fossa séptica, enquanto esse valor foi de apenas 58,5% nos Setores Precários. No caso do serviço de abastecimento de água, não houve uma grande discrepância, uma vez que o DF apresenta alto índice de acesso nesse serviço como já mencionado. Ainda assim, o indicador de UI foi inferior aos indicadores gerais em dois pontos percentuais.

Dessa forma, frisa-se a necessidade de estudos que priorizem a análise da prestação dos serviços de água e esgoto em áreas de Assentamentos Precários urbanos, uma vez que foi provado que, de fato, a população que mais sofre com a carência desses serviços é aquela mais vulnerável social e economicamente, as quais, nos centros urbanos, concentra-se em áreas precárias de ocupação desordenada.

7 CONCLUSÕES E RECOMENDAÇÕES

O presente estudo teve como objetivo geral obter um método consolidado de identificação dos Assentamentos Precários no Brasil que pudesse ser aplicado em qualquer unidade federativa do Brasil com base nos dados disponibilizados pelo Censo Demográfico, partindo do método desenvolvido pelo CEM em 2007 e publicado na obra "Assentamentos Precários no Brasil Urbano", do Ministério das Cidades. O intuito do estudo é promover uma forma de mapeamento desse fenômeno no território brasileiro, com vistas a possibilitar o monitoramento de sua ocorrência e a avaliação do acesso aos serviços de abastecimento de água e esgotamento sanitário nessas áreas.

Para atingir esse objetivo, foram realizadas modificações no método-base, o qual foi aplicado no território urbano do Distrito Federal, como forma de estudo de caso. Assim, a informação foi atualizada para o dado disponível mais recente, que é o Censo Demográfico de 2010, além de terem sido implementados aprimoramentos em sua aplicação. Nesse sentido, foram testadas três técnicas estatísticas distintas para a identificação de Assentamentos Precários: ADL, ADQ e RL, tendo sido escolhida a técnica de melhor desempenho preditivo. A classificação foi realizada partindo da categorização prévia realizada pelo IBGE, sendo os setores detectados pelo modelo como SbN definidos como os correspondentes a Assentamentos Precários.

Na análise da Estatística Descritiva dos dados do IBGE, verificou-se que o indicador que apresentou maior discrepância de médias entre setores SbN e NE foi o referente aos domicílios sem acesso a esgotamento sanitário ou fossa séptica (C.4), onde a média para setores SbN foi de 65% de domicílios sem acesso, enquanto, para setores NE, foi de apenas 6,8%. O indicador de domicílios sem acesso à coleta de resíduos sólidos (C.1) foi o segundo com maior discrepância, com média de 13% para setores SbN e de apenas 0,7% para setores NE. O indicador de carência de abastecimento de água (C.2) apresentou valores pequenos para ambos (6% e 3%, respectivamente), o que se deve ao alto alcance de cobertura desse serviço no Distrito Federal. Os indicadores de renda (B.4 e B.5) também apresentaram médias bastante discrepantes entre os grupos, mostrando que os setores SbN possuem, em média, 90% de sua população com renda menor que 3SM, contra 56% em setores NE. A renda média dos setores SbN foi de R$809, e de R$3024 nos setores NE.

Dessa forma, pela análise da estatística descritiva, foi possível observar que as características expressas pelos indicadores B.4/B.5, C.4 e C.1 (renda, falta de acesso a

esgotamento sanitário/fossa séptica e falta de acesso à coleta de resíduos sólidos) seriam as de maior impacto nos modelos, uma vez que foram as mais marcantes no perfil da população residente em setores SbN e, consequentemente, na identificação de Assentamentos Precários no DF. Por outro lado, a carência no acesso ao abastecimento de água (C.2) não apresentou impacto significativo, juntamente com o indicador referente ao percentual de pessoas responsáveis do sexo feminino (A.5) e o de número médio de moradores por DPP (A.3), que mostraram ter a mesma média entre os setores SbN e NE.

Os resultados obtidos nos três modelos revelaram uma alta relação entre os *scores* discriminantes e probabilidade de ocorrência do evento de interesse – no caso da Regressão Logística – com o indicador de renda (B.4 na AD e B.5 na RL) que, quando plotado junto ao indicador C.4, mostrou ser decisivo na identificação de Assentamentos Precários, principalmente quando menor renda estava associada a altos índices de carência de acesso ao esgotamento sanitário. A propósito, o indicador de renda se mostrou ainda mais determinante que o de acesso a esgotamento sanitário, nos três modelos, especialmente quando observadas as representações gráficas, onde mesmo os setores censitários com mais de 80% de seus domicílios sem esgotamento sanitário ou fossa séptica não foram classificados como SbN quando apresentaram rendas maiores (superiores a R$2500 de renda média ou com menos de 20% de sua população ganhando até 3 SM). Nesse sentido, a carência de acesso, quando associada a maiores rendas, não faz parte do perfil de Assentamentos Precários no DF.

De toda forma, os indicadores de saneamento se destacaram como os mais significantes na identificação de Assentamentos Precários em todos os modelos, mostrando que o acesso aos SAAES configura uma questão marcante nessas áreas.

Com relação ao desempenho dos modelos das três técnicas testadas, a sensitividade, ou seja, taxa de acerto de setores SbN originalmente classificados como tal pelo IBGE, foi de 66% na ADL, de 77% na ADQ, e de 88% na RL. Tal medição foi a principal medida de ajuste sugerida no método-base, uma vez que, para identificar setores censitários de perfil semelhante aos setores SbN, o modelo deve apresentar uma boa taxa de acerto desses setores, garantindo um afinamento do modelo para o objetivo. Tais valores foram bem maiores ao obtido no estudo aplicado em 2007, cuja taxa de acerto de setores censitários SbN para o DF e RM de Goiânia foi de 52%, utilizando a técnica da ADL.

Dentre as três técnicas aplicadas, a que apresentou melhor desempenho preditivo foi a de Regressão Logística, que além de obter maior percentual de acerto de setores SbN, resultou em melhores medidas gerais, como a área da curva ROC. Por essa razão, foi a técnica escolhida para gerar a base de Assentamentos Precários do estudo, para a qual foram calculados os Índices de Acesso aos SAAES. Além disso, a técnica requer a satisfação de pressupostos mais flexíveis, ao não exigir nem a homogeneidade de matrizes de covariância, nem a normalidade multivariada. Isso porque, na aplicação da técnica da ADL, foi violado o pressuposto da homogeneidade de matrizes de covariância entre os grupos e, embora tenha se admitido o Teorema do Limite Central pelo tamanho significativo da amostra, é possível que os resultados sejam afetados pela falta de cumprimento da premissa. Dessa forma, após a RL, a ADQ seria mais recomendada para estudos desse escopo, uma vez que não exige o requisito da igualdade de matrizes.

Assim, partindo dos resultados obtidos pela RL, que identificou 547 setores representativos de Assentamentos Precários no DF, foi calculado o Índice de Universalização Inclusiva, bem como os índices de acesso em setores NE e no DF como um todo, incluindo todos os setores. Os resultados obtidos corroboraram com a hipótese do acesso aos SAAES em Assentamentos Precários serem mascarados quando diluídos em dados gerais levantada por Marques et al. (2007) e Guimarães (2015). O indicador de UI foi de 95,5% para abastecimento de água e 57,4% para esgotamento sanitário, em face aos índices gerais de acesso de 97,3% e 90,6%, respectivamente; apresentando, portanto, uma discrepância de mais de 30% para o esgotamento sanitário.

Foi identificado que, em 2010, 255 mil pessoas não tinham acesso à rede de esgotamento sanitário ou fossa séptica, das quais 175 mil eram residentes de Assentamentos Precários. É importante destacar que, como datam de 2010, ano de realização do último Censo Demográfico, os dados se encontrem relativamente desatualizados, de modo que não retratarão a realidade atual de maneira precisa, especialmente em se tratando do dinâmico processo de transformações urbanas. No entanto, ainda representam a única informação espacializada disponível nacionalmente com metodologia padronizada e confiável para o mapeamento de Assentamentos Precários (Silva et al., 2014), uma vez que reúnem informações referentes à ocupação territorial associadas às características socioeconômicas dos habitantes.

No entanto, recomenda-se que, como futuras complementações do procedimento metodológico realizado, incluam-se métodos de validação do modelo, como visitas de

campo e/ou uso de imagens de satélite. Além disso, é necessário que os resultados do método proposto sejam atualizados para todas as regiões metropolitanas brasileiras com os dados do Censo Demográfico de 2010, a fim de comparar com os resultados obtidos por Marques et al. (2007). Nesse sentido, recomenda-se, ainda, reaplicar o método com os dados do próximo Censo Demográfico, a ocorrer em 2022, de forma a utilizar o método proposto para monitorar a ocorrência desse fenômeno no território brasileiro.

Entende-se que a informação sobre o estado da arte dos Assentamentos Precários no Brasil pode ser decisiva na orientação de discussões sobre o tema, tanto a respeito do cumprimento do direito humano ao acesso à água e esgoto, que passa por questões relacionadas à infraestrutura e gestão, especialmente no que se referem aos contratos das concessionárias prestadoras dos serviços e na promoção de tarifa social; quanto no âmbito de outras questões não menos relevantes que atravessam essa temática, como a consolidação do conceito de Assentamentos Precários no Brasil, a discussão do arcabouço legal que abrange o tema da ocupação informal; a orientação para elaboração e implementação de políticas públicas direcionadas ao atendimento dessa população mais vulnerável, sejam elas direcionadas à realocação ou à regularização fundiária; e ao planejamento urbano, de forma geral.

REFERÊNCIAS

Adams, E. A. (2018). Intra-urban inequalities in water access among households in Malawi's informal settlements: Toward pro-poor urban water policies in Africa. In: *Environmental Development, 26*, 34-42.

Altman DG, Bland JM (1995). Statistics notes: the normal distribution. In: *Bmj*, 310(6975), 298.

Batista, M. E. M.; Silva, T. C. (2006). O modelo ISA/JP-indicador de performance para diagnóstico do saneamento ambiental urbano. In: *Revista de Engenharia Sanitária e Ambiental*, 11(1), 55-64.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York - USA.

Bonduki, N. (1998). *Origens da Habitação Social no Brasil*. Editora Estação Liberdade e FAPESP, São Paulo.

Brasil (1998). *Constituição da República Federativa do Brasil de 1988.* Diário Oficial da União, 1988.

Brasil. (2001). *Lei nº 10.257, de 10 de Julho de 2001*. Regulamenta os Arts. 182 e 183 da Constituição Federal, estabelece diretrizes gerais da política urbana e dá outras providências. Brasília: Diário Oficial da União, 2001a.

Brasil (2001). *Medida Provisória nº 2.220, de 04 de Setembro de 2001*. Dispõe sobre a concessão de uso especial de que trata o § 1o do art. 183 da Constituição, cria o Conselho Nacional de Desenvolvimento Urbano - CNDU e dá outras providências. Brasília: Diário Oficial da União, 2001b.

Brasil. (2005). *Lei nº 11.124, de 16 de Junho de 2005*. Dispõe sobre o Sistema Nacional de Habitação de Interesse Social – SNHIS, cria o Fundo Nacional de Habitação de Interesse Social – FNHIS e institui o Conselho Gestor do FNHIS. Brasília: Diário Oficial da União, 2005.

Brasil (2007). L*ei nº 11.445, de 5 de Janeiro de 2007.* Estabelece diretrizes nacionais para o saneamento básico. Brasília: Diário Oficial da União, 2007.

Cardoso, A. L. (2016). Assentamentos Precários No Brasil: Discutindo Conceitos. In: Morais, M. P., Krause, C. e Lima Neto, V. C. (2016). *Caracterização e tipologia de assentamentos precários: estudos de caso brasileiros.* Brasília: IPEA, 2016. 540 p. ISBN: 978-85-7811-276-9.

Carrion, R. (2013). *A inadiável Reforma Urbana*. Ministério das Cidades, V Conferência Nacional de Cidades, Textos Complementares. Out. 2013. Brasília-DF. Disponível em: <https://goo.gl/AEQmrc>. Acesso em: 05 out. 2018.

Castro, J.E.; Heller, L. (2009). *Water and Sanitation Services: public policy and management.* Earthscan, United Kingdom.

Castro Junior, F. H. F. (2003). *Previsão de insolvência de empresas brasileiras usando análise discriminante, regressão logística e redes neurais*. Dissertação de Mestrado em Administração, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo – SP.

Denaldi, R. (Org.) (2010). *Ações Integradas de Urbanização de Assentamentos Precários.* Brasília/São Paulo: Ministério das Cidades/Aliança de Cidades. ISBN : 978-85-7958-006-2.

dos Santos, R., & Gupta, J. (2017). Pro-poor water and sanitation: operationalising inclusive discourses to benefit the poor. Current opinion in environmental sustainability, 24, 30-35.

Elliott AC, Woodward WA. (2007). *Statistical analysis quick reference guidebook with SPSS examples*. 1st ed. Sage Publications, London - UK.

Fávero, L. P., Belfiore, P., Silva, F. D., & Chan, B. L. (2009). *Análise de dados: modelagem multivariada para tomada de decisões*. Elsevier, Rio de Janeiro.

Fávero, L. P., & Belfiore, P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. Elsevier Brasil.

Faya, O. E. N. (2014). *O efeito das ações de saneamento em aglomerados subnormais no litoral*. Dissertação de Mestrado. Universidade de São Paulo, 2014.

Ferreira, J. S. W. (2010). O processo de urbanização brasileira e a função social da propriedade urbana. In: Denaldi, R. (Org.) (2010). *Ações Integradas de Urbanização de Assentamentos Precários.* Brasília/São Paulo: Ministério das Cidades/Aliança de Cidades. ISBN : 978-85-7958-006-2.

Ferreira, M. P., Marques, E. C. L., Fusaro, E. R (2007). Assentamentos Precários no Brasil: Uma Metodologia Para Estimação e Análise. In: Morais, M. P., Krause, C. e Lima Neto, V. C. (2016). *Caracterização e tipologia de assentamentos precários: estudos de caso brasileiros.* Brasília: IPEA, 2016. 540 p. ISBN: 978-85-7811-276-9.

Field A. *Discovering statistics using SPSS*. 3 ed. SAGE publications Ltd, London - UK. p. 822.

Guimarães, E. F. (2015). *Modelo inclusivo para a universalização do saneamento básico em áreas de vulnerabilidade social*. Tese de Doutorado. Escola de Engenharia de São Carlos, Universidade de São Paulo, 486 p.

Guimarães, E. F., Coutinho, S. M. V., Malheiros, T. F., & Philippi Jr., A. (2014). Os indicadores do saneamento medem a universalização em áreas de vulnerabilidade social? In: *Engenharia Sanitaria e Ambiental*, *19*(1), 53–60. https://doi.org/10.1590/S1413-41522014000100006.

Hair, J.F.J.; Anderson, R.E.; Tatham, R.L.; Black, W.C. (1998). *Multivariate Data Analysis*, 5th ed. Prentice Hall, Upper Saddle River, New Jersey.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman editora, Porto Alegre – RS.

Hawkins, P., Blackett, I., & Heymans, C. (2013). Poor-inclusive urban sanitation: An overview. Water and Sanitation Programme.

Hilbe, M. (2015). *Practical Guide to Logistic Regression*. Taylor & Francis Group.

Heller, L. (2009). Water and sanitation policies in Brazil: historical inequalities and institutional change. Water and sanitation services: public policy and management. Londres: Earthscan, 2009, p. 321-337.

IBGE. *Censo Demográfico 2000*: resultados do universo. Rio de Janeiro: IBGE, 2001. Disponível em: <https://goo.gl/fwj5JH>. Acesso em: 10 set. 2017.

IBGE. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Censo Demográfico 2010:* resultado do universo. Rio de Janeiro: IBGE, 2011. Disponível em: <http://goo.gl/b83uhZ>. Acesso em: 10 set. 2018.

IPEA (2015). *O direito à água como política pública na América Latina : uma exploração teórica e empírica. Brasília*. Ipea, Brasília – DF, 322 p. ISBN: 978-85-7811-238-7

IPEA (2018). ODS – *Metas Nacionais dos Objetivos de Desenvolvimento Sustentável.* Proposta de adequação. IPEA, Brasília - DF.

Johnson, R. A.; Wichern, D. W. (2007) *Applied Multivariate Stattistical Analysis*. 6 ed. Pearson Prentice Hall, Upper Saddle River – NJ.

Jones, H., Fisher, J., & Reed, R. (2012, September). Water and sanitation for all in low-income countries. In Proceedings of the Institution of Civil Engineers-Municipal Engineer (Vol. 165, No. 3, pp. 167-174). Thomas Telford Ltd.

Juliano, F. G.; Feuerweker, L., Coutinho, S.; Malheiros, T. F. (2012). Racionalidade e Saberes na Produção de Modelos Organizativos para a Universalização do

Saneamento em Áreas Urbanas do Brasil. In: *Ciência & Saúde Coletiva*, 17(11):3037-3046, 2012.

Kassambara, A. (2018). *Machine learning essentials: Practical guide in R*. Sthda.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2), 137-163.

Klecka, W. R., Iversen, G. R., & Klecka, W. R. (1980). *Discriminant analysis* (Vol. 19). Sage, London – UK.

Krzanowsky, W. J. (1988) *Principles of multivariate analysis*. Clarendon Press, Oxford.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modelling*. Springer, New York – USA.

Magalhães, M. N.; Lima, A. C. P. (2015). *Noções de Probabilidade e Estatística*. 7 ed. Editora da Universidade de São Paulo, São Paulo – SP. ISBN 13: 9788531406775

Marques, E., Gomes, S., Gonçalves, R., Toledo, D., Moya, E., Cazolato, D., & Ferreira, P. (2007). *Assentamentos precários no Brasil urbano.* Brasília: Ministério das Cidades/CEM, 390 p.

Mclachlan, G. (1992). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, New York - USA.

Mingoti, S. A. (2005). Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In: *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*, 295.

Ministério das Cidades (2007). *Assentamentos precários no Brasil urbano.* Secretaria Nacional de Habitação, Ministério das Cidades e do Centro de Estudos da Metrópole.

Ministério das Cidades (2010). *Guia para o Mapeamento e Caracterização de Assentamentos Precários.* Brasília: Ministério das Cidades. 84 p. ISBN: 978-85-7958-015-4.

Morais, M. P., Krause, C. e Lima Neto, V. C. (2016). *Caracterização e tipologia de assentamentos precários: estudos de caso brasileiros.* IPEA, Brasília – DF, 540 p. ISBN: 978-85-7811-276-9.

Morettin, L. G. (1999). Est*atistica Basica: Probabilidade*. Vol I. 7 ed. Makron Books.

Morgan, G. A.; Griego, O. V. (1998*). Easy use and interpretation of SPSS for Windows: Answering research questions with statistics*. Psychology Press.

Nadalin, V.G.; Krause, C.; Lima Neto, V. C. *Distribuição de Aglomerados Subnormais na Rede Urbana e nas Grandes Regiões Brasileiras.* Brasília: Instituto de

Pesquisa Econômica Aplicada, 2014. Disponível em: <http://goo.gl/9LFTSy>. Acesso em 18 ago. 2018.

Nadalin, V. G.; Mation, L. F. (2018). Localização intraurbana das favelas brasileiras: O papel dos fatores geográficos. *Texto para Discussão, No. 2390*, Instituto de Pesquisa Econômica Aplicada (IPEA), Brasília – DF.

Office of the High Commissioner for Human Rights (OHCHR) (2010). *General Comment No. 15: The Right to Water* (Arts. 11 and 12 of the Covenant). OHCHR, Geneva.

Oliveira, S. D. R., & Anjos, R. S. (2004). A organização de dados de favelas para o planejamento territorial: uma proposta metodológica. In: *L'Espace Geographique*, 7(1), 99-131.

World Health Organization/Unicef. (2019). Progress on household drinking water, sanitation and hygiene 2000-2017: special focus on inequalities. World Health Organization.

Oztuna D, Elhan AH, Tuccar E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. In: *Turkish Journal of Medical Sciences*, 36(3), 171- 176.

Pallant J. (2007). *SPSS survival manual, a step by step guide to data analysis using SPSS for windows*. 3 ed. McGraw Hill, Sydney.

Patel, A., Koizumi, N., & Crooks, A. (2014). Measuring slum severity in Mumbai and Kolkata: a household-based approach. In: *Habitat International*, 41, 300-306.

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional* (pp. 28-55). IME-USP, São Paulo - SP.

PNUMA - Programa das Nações Unidas para o Meio Ambiente (2004). *Metodologia para elaboração de relatório GEO Cidades.* PNUMA, México, 181p.

Press, J.; Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. In: *Journal of the American Statistical Association,* 73 (364), 699- 705.

Queiroz Filho, A. P. D. (2015). As definições de assentamentos precários e favelas e suas implicações nos dados populacionais: abordagem da análise de conteúdo. urbe. In: *Revista Brasileira de Gestão Urbana*, 7(3), 340-353.

Satterthwaite, D. (2003). The Millennium Development Goals and urban poverty reduction: great expectations and nonsense statistics. In: *Environment and Urbanization*, 15(2), 170-190. http://dx.doi.org/10.1177/095624780301500208

Schneider, D. D., Dos Santos, R., Martínez, Coutinho, S. M. R. C., Malheiros, T. F.,

Temóteo, T. G. (2010). Indicadores para serviços de abastecimento de água e esgotamento sanitário voltados às populações vulneráveis. In: *Revista Brasileira de Ciências Ambientais*, 17, 65-76.

Sharma, S. (1996). *Applied multivariate techniques*. John Wiley & Sons Inc., New York.

Silva, D. W. T., Fusaro, D., Marques, E. C, L., Cazolato, J. D. (2014). *Assentamentos Precários no Brasil Urbano: Metodologia de identificação de assentamentos precários urbanos.* Centro de Estudos da Metrópole, São Paulo - SP.

Silva-Neves, P. e Heller, L (2016). O direito humano à água e ao esgotamento sanitário como instrumento para promoção da saúde de populações vulneráveis. In: *Ciência & Saúde Coletiva*, 21(6), 1861-1869.

Sinharoy, S. S., Pittluck, R., & Clasen, T. (2019). Review of drivers and barriers of water and sanitation policies for urban informal settlements in low-income and middle-income countries. Utilities policy, 60, 100957.

SNIS - Sistema Nacional de Informações sobre Saneamento (2016). *Diagnóstico anual água e esgotos.* Brasília - DF.

Un-Habitat. 2013. *Water and sanitation in the world's cities: Local action for global goals*. London: Earthscan.

Villaça (2001). *Espaço intraurbano no Brasil*. Studio Nobel/Fapesp/LILP, São Paulo - SP.

Wasserman, L. (2004). *All of Statistics: a concise course on statistical inference*. Springer, New York – USA. ISBN-13 : 978-0387402727

Wooldridge, J. M (2016). *Introdução à econometria: uma abordagem moderna*. Cengage Learning, São Paulo – SP.

World Health Organization (WHO), UNICEF (2015). *Joint monitoring program for water supply and sanitation. Progress on drinking water and sanitation. Update 2015*. WHO, UNICEF, Geneva.