

UNIVERSIDADE FEDERAL DE VIÇOSA

TAYANE CRISTIELE RODRIGUES MESQUITA

**ANÁLISE DA RELAÇÃO SAÚDE, SANEAMENTO E AMBIENTE NAS EPIDEMIAS
DO ZIKA VÍRUS E MICROCEFALIA NO BRASIL**

**VIÇOSA - MINAS GERAIS
2023**

TAYANE CRISTIELE RODRIGUES MESQUITA

**ANÁLISE DA RELAÇÃO SAÚDE, SANEAMENTO E AMBIENTE NAS EPIDEMIAS
DO ZIKA VÍRUS E MICROCEFALIA NO BRASIL**

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

Orientador: André Pereira Rosa

Coorientadores: Alisson Carraro Borges
Elpídio Inácio Fernandes Filho

**VIÇOSA - MINAS GERAIS
2023**

Ficha catalográfica elaborada pela Biblioteca Central da Universidade
Federal de Viçosa - Campus Viçosa

T

M582a
2023 Mesquita, Tayane Cristiele Rodrigues, 1994-
Análise da relação saúde, saneamento e ambiente nas
epidemias do Zika Virus e microcefalia no Brasil / Tayane
Cristiele Rodrigues Mesquita. – Viçosa, MG, 2023.
1 tese eletrônica (157 f.): il. (algumas color.).

Inclui apêndice.

Orientador: André Pereira Rosa.

Tese (doutorado) - Universidade Federal de Viçosa,
Departamento de Engenharia Agrícola, 2023.

Inclui bibliografia.

DOI: <https://doi.org/10.47328/ufvbbt.2023.261>

Modo de acesso: World Wide Web.

1. Arboviroses. 2. Vírus da Zika. 3. Aprendizado do
computador. 4. Saneamento. 5. Algoritmos. I. Rosa, André
Pereira, 1983-. II. Universidade Federal de Viçosa.
Departamento de Engenharia Agrícola. Programa de
Pós-Graduação em Engenharia Agrícola. III. Título.

CDD 22. ed. 579.2562

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552


TAYANE CRISTIELE RODRIGUES MESQUITA

**ANÁLISE DA RELAÇÃO SAÚDE, SANEAMENTO E AMBIENTE NAS EPIDEMIAS
DO ZIKA VÍRUS E MICROCEFALIA NO BRASIL**


Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

APROVADA: 22 de março de 2023

Assentimento:

Documento assinado digitalmente
 TAYANE CRISTIELE RODRIGUES MESQUITA#
Data: 15/05/2023 10:58:33-0300
Verifique em <https://validar.itl.gov.br>

Tayane Cristiele Rodrigues Mesquita
Autora

Documento assinado digitalmente
 ANDRE PEREIRA ROSA
Data: 15/05/2023 19:17:12-0300
Verifique em <https://validar.itl.gov.br>

André Pereira Rosa
Orientador

*A todas as crianças acometidas pela síndrome
congênita do vírus Zika, com amor, dedico.*

AGRADECIMENTOS

Ao fim de mais uma etapa em minha vida, agradeço ao meu adorado Deus por Seu infinito amor e misericórdia.

Ao professor André Rosa, agradeço imensamente pela orientação, paciência, compreensão, pelo imenso conhecimento compartilhado e empenho na realização desse trabalho.

Aos professores Alisson Carraro e Elpídio Inácio Fernandes Filho, agradeço muitíssimo pelo imenso conhecimento compartilhado, por todas as sugestões, correções e pela disponibilidade para coorientação ao longo do desenvolvimento desse trabalho.

Aos meus pais Antonio e Filomena e à minha irmã Sandy, muito obrigada por todo amor, cuidado, companheirismo e por nunca medirem esforços para me ajudar em tudo.

Aos membros do Laboratório da Qualidade Ambiental (LQA), agradeço por toda a ajuda, pela convivência e pelas ricas discussões dos mais variados temas. Agradeço de maneira especial à Isabelle e à Maria Fernanda, minhas irmãs de orientação, que, de diversas formas, me auxiliaram durante toda essa jornada.

Ao Gustavo, agradeço por toda a ajuda e conhecimento compartilhado durante as etapas de processamento de dados.

À Universidade Federal de Viçosa (UFV) e ao Departamento de Engenharia Agrícola (DEA) pela infraestrutura disponibilizada e pela oportunidade de fazer parte do Programa de Pós-Graduação em Engenharia Agrícola.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela concessão da bolsa de doutorado*.

Aos membros da Sociedade São Vicente de Paulo, especialmente aos confrades e consócias das conferências Santo Antônio e Nossa Senhora do Rosário. Muito obrigada pela amizade, por todas as orações e por serem uma escola de caridade em minha vida.

Aos membros da banca, agradeço a disponibilidade e as contribuições para o aperfeiçoamento desse trabalho.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para que esse trabalho pudesse ser realizado.

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (processo nº 140419/2020-9).

*“Que a tua vida não seja uma vida estéril. - Sê útil. - Deixa
rastro. - Ilumina com o resplendor da tua fé e do teu amor.
(São Josemaria Escrivá)*

RESUMO

MESQUITA, Tayane Cristiele Rodrigues, D.Sc., Universidade Federal de Viçosa, março de 2023. **Análise da relação saúde, saneamento e ambiente nas epidemias do Zika vírus e microcefalia no Brasil.** Orientador: André Pereira Rosa. Coorientadores: Alisson Carraro Borges e Elpídio Inácio Fernandes Filho.

Os casos de Zika e microcefalia, nos anos de 2015 e 2016, relatados no Brasil apresentaram grande desigualdade na distribuição entre as regiões do país. Até os tempos recentes, as razões para tais diferenças não foram completamente elucidadas. Contudo, a existência de condições socioeconômicas, climáticas e de saneamento díspares entre as regiões brasileiras pode representar um ponto de partida para o entendimento dessas questões. Diante disso, nesse estudo teve-se como intuito realizar uma análise multi-abordagem acerca da relação entre as condições sanitárias, socioeconômicas, ambientais e climáticas e a ocorrência das epidemias do vírus Zika e microcefalia no Brasil. Para o procedimento metodológico, foram obtidos dados secundários relativos à incidência de Zika e microcefalia no Brasil, no ano de 2016, além de diversos indicadores socioeconômicos, climáticos e sanitários dos municípios brasileiros. Para avaliar a existência de correlação espacial entre as taxas de incidência de Zika e microcefalia nos municípios brasileiros foi utilizado o índice local de Moran (I_i) univariado e bivariado. A relação entre a incidência de Zika e microcefalia e as características sanitárias e socioeconômicas dos municípios foi investigada por meio da realização de uma análise de agrupamentos, utilizando métodos hierárquicos e não hierárquicos, e posterior aplicação do teste U de Mann-Whitney para identificação de diferenças entre as taxas de incidência de Zika e microcefalia nos grupos formados. Adicionalmente, uma investigação da existência de associações entre as taxas de incidência de Zika e microcefalia e a ocorrência de eventos climáticos extremos foi conduzida. Para isso, foi realizada a separação dos dados de Zika e microcefalia, em função dos diferentes eventos extremos notificados, e posterior aplicação do teste U para comparação entre os grupos formados. Por fim, para o processo de predição de casos de Zika e microcefalia nos municípios brasileiros, seis modelos de aprendizado de máquina foram selecionados, sendo estes: *Random Forest (RF)*, *Multivariate Adaptive Regression Spline (MARS)*, *K-nearest Neighbor (KNN)*, *Support Vector Machine (SVM)*, *Gradient Boosting Machine (GBM)* e regressão logística. Os resultados obtidos indicaram que as taxas de incidência de Zika e microcefalia nos municípios brasileiros estão associadas à localização geográfica de ocorrência dos eventos. As maiores taxas de incidência ocorreram, majoritariamente, nos agrupamentos formados por municípios que apresentam menores índices

socioeconômicos e de saneamento no país. Além disso, foi verificado que existem diferenças entre as taxas de incidência de Zika e microcefalia nos municípios que possuem histórico de ocorrência de eventos climáticos extremos, sendo os municípios que possuem eventos de seca aqueles mais atingidos pela epidemia do vírus Zika. O modelo de aprendizado de máquina *Random forest* apresentou os melhores resultados na predição de Zika e mostrou possuir capacidade de prever corretamente a incidência de Zika em 65% dos municípios contidos nos dados de teste a partir dos preditores avaliados. Por outro lado, o modelo para predição de microcefalia apresentou resultados de ajustes inferiores aos observados para o modelo de Zika, o que pode estar associado à maior complexidade de se modelar uma síndrome congênita. Os modelos avaliados indicaram forte influência das condições demográficas, socioeconômicas, climáticas e sanitárias dos municípios na predição da ocorrência de Zika. Diante do exposto, acredita-se que os resultados obtidos nesse estudo possam fornecer informações valiosas para a tomada de decisão em saúde pública e definição de estratégias para prevenção de epidemias associadas ao vírus Zika.

Palavras-chave: Arboviroses. Aprendizado de máquina. Saneamento básico. Predição. Modelagem.

ABSTRACT

MESQUITA, Tayane Cristiele Rodrigues, D.Sc., Universidade Federal de Viçosa, March, 2023. **Analysis of the relationship among health, sanitation and environment in the Zika virus and microcephaly epidemics in Brazil.** Advisor: André Pereira Rosa. Co-advisors: Alisson Carraro Borges and Elpídio Inácio Fernandes Filho.

Cases of Zika virus disease and microcephaly reported in Brazil (2015-2016) showed great difference in their distribution throughout the country. The reasons for such disparity have not been fully elucidated yet. However, the existence of different socioeconomic, climatic and sanitation conditions among Brazilian regions may represent a starting point for understanding these issues. This study carried out a multi-approach analysis on the relationship among health, socioeconomic, environmental and climatic conditions and the occurrence of Zika virus and microcephaly epidemics in Brazil. For the methodological procedure, secondary data were obtained, regarding the incidence of Zika and microcephaly in Brazil, in 2016, in addition to socioeconomic, climatic and health indicators of Brazilian municipalities. To evaluate the existence of spatial correlation between the incidence rates of Zika and microcephaly in Brazilian municipalities, both the univariate and bivariate Moran's index (I_i) were used. The relationship between the incidence of Zika and microcephaly and the sanitary and socioeconomic characteristics of the municipalities was investigated through cluster analysis, using hierarchical and non-hierarchical methods. The Mann-Whitney U test was used to identify differences between the incidence rates of Zika and microcephaly in the formed clusters. Additionally, an investigation of the potential association between the incidence rates of Zika and microcephaly and the occurrence of extreme weather events was conducted. The data on Zika and microcephaly cases were separated, according to the different extreme events reported, and the U test was applied to compare the formed groups. Finally, for the process of predicting cases of Zika and microcephaly in Brazilian municipalities, six machine learning models were selected: *Random Forest (RF)*, *Multivariate Adaptive Regression Spline (MARS)*, *K-nearest Neighbor (KNN)*, *Support Vector Machine (SVM)*, *Gradient Boosting Machine (GBM)* and logistic regression. The results indicated that the incidence rates of Zika and microcephaly in Brazilian municipalities are associated with the geographical location of the events. The highest incidence rates occurred, mostly, in the groups formed by municipalities with the lowest socioeconomic and sanitation indices in the country. Also, there are differences among the incidence rates of Zika and microcephaly in the municipalities with a history of extreme events, and those which present extreme droughts are the most affected by the Zika

virus epidemic. The machine learning model *Random forest* presented the best results in predicting Zika cases and the ability to correctly predict the incidence of Zika in 65% of the municipalities contained in the test data from the evaluated predictors. On the other hand, the model for predicting microcephaly showed worse fitting results than those observed for the Zika model, which may be associated with the complexity of modeling a congenital syndrome. The evaluated models indicated a strong influence of the demographic, socioeconomic, climatic and sanitary conditions of the municipalities in predicting the occurrence of Zika. Therefore, the results obtained in this study may provide valuable information for public health decision-making and the definition of strategies for the prevention of epidemics associated with the Zika virus.

Keywords: Arboviruses. Machine learning. Basic sanitation. Prediction. Modeling.

SUMÁRIO

1. INTRODUÇÃO GERAL	13
2. HIPÓTESES DE INVESTIGAÇÃO	16
3. OBJETIVO GERAL.....	18
3.1 Objetivos específicos	18
4. ESTRUTURAÇÃO DA TESE.....	19
CAPÍTULO 1	22
1.1 Arboviroses.....	22
1.1.1 Zika: origens	24
1.1.2 Evolução da epidemia do vírus Zika no Brasil.....	25
1.1.3 Arboviroses, saneamento e determinantes ambientais	28
1.2 Aprendizado de máquina	31
1.2.1 Aprendizado supervisionado	32
1.2.2 Aprendizado não supervisionado	46
1.3 Estatística espacial	49
1.4 Referências	51
CAPÍTULO 2	57
2.1 Introdução.....	57
2.2 Material e métodos	60
2.2.1 Estratégia de busca do material bibliográfico e critérios de seleção	60
2.2.2 Avaliação dos estudos selecionados	61
2.3 Resultados e discussão	61
2.3.1 Características gerais dos estudos avaliados	61
2.3.2 Distribuição de vetores e mudanças climáticas	66
2.3.3 Transmissão viral, incidência de arboviroses e mudanças climáticas	71
2.3.4 Limitações dos estudos.....	77
2.4 Conclusões.....	77
2.5 Referências	78
CAPÍTULO 3	81
3.1 Introdução.....	81
3.2 Material e métodos	83
3.2.1 Fonte de dados e seleção de indicadores	83
3.2.2 Análise preliminar dos dados	87

3.2.3 Distribuição da incidência de Zika e microcefalia no Brasil.....	88
3.2.4 Incidência de Zika e microcefalia em municípios com notificação de eventos climáticos extremos	89
3.2.5 Relação entre Zika, microcefalia e indicadores sanitários e socioeconômicos	91
3.3 Resultados e discussão	95
3.3.1 Distribuição da incidência de Zika e microcefalia no Brasil.....	95
3.3.2 Incidência de Zika e microcefalia em municípios com notificação de eventos climáticos extremos	99
3.3.3 Relação entre Zika, microcefalia e indicadores sanitários e socioeconômicos	102
3.4 Considerações finais	112
3.5 Referências	113
CAPITULO 4	118
4.1 Introdução.....	118
4.2 Material e métodos	120
4.2.1 Base de dados	122
4.2.2 Preparação dos dados e eliminação de covariáveis correlacionadas	128
4.2.3 Partição dos dados	129
4.2.4 Seleção das covariáveis por importância.....	130
4.2.5 Treinamento dos modelos.....	131
4.2.6 Avaliação do desempenho dos modelos	132
4.3 Resultados e discussão	132
4.3.1 Visão geral dos dados.....	132
4.3.2 Pré-processamento dos dados e seleção de covariáveis	135
4.3.3 Desempenhos dos modelos preditivos.....	136
4.3.4 Importância das covariáveis em cada modelo	142
4.4 Considerações finais	148
4.5 Referências	149
5 CONSIDERAÇÕES GERAIS.....	155
Apêndice A	157

1. INTRODUÇÃO GERAL

Nas últimas décadas, uma expansão inesperada da distribuição territorial de diversos arbovírus resultou em um expressivo aumento na atividade epidêmica global (HILGENFELD; VASUDEVAN, 2018). O aparecimento repentino e a disseminação a nível global fizeram com que as arboviroses passassem a ser consideradas uma grave e sucessiva ameaça de saúde pública (LIMA; SOARES; SILVA, 2022).

No cenário epidemiológico brasileiro, há a circulação - simultânea ou não - de quatro sorotipos do vírus dengue (DEV), além dos vírus Chikungunya (CHIKV) e Zika (ZIKV). Dentre os arbovírus citados, os problemas associados ao ZIKV foram mais amplamente discutidos, recentemente, no Brasil em função da ocorrência, nos anos de 2015 e 2016, de uma epidemia associada a esse vírus (BRASIL, 2019).

As arboviroses podem ser transmitidas por duas espécies de mosquitos do gênero *Aedes*: *Aedes aegypti* e de *Aedes albopictus*. No Brasil, as populações de *Aedes aegypti* possuem maior potencial de transmissão do Zika, sendo estas altamente antropofílicas. De acordo com Barrera *et al.* (2008), tais mosquitos desenvolvem-se, sobretudo, em regiões que apresentam condições de infraestrutura deficiente, favorecendo o seu estabelecimento e, conseqüentemente, disseminação dos sorotipos virais por eles veiculados.

A princípio, quando o vírus Zika foi identificado no Brasil, a principal preocupação dos especialistas estava associada à sua circulação concomitante com os vírus da dengue e Chikungunya, que poderia agravar os efeitos adversos desses patógenos (LOWY, 2019). No entanto, um aumento atípico do número de casos de nascidos vivos com microcefalia, durante o mesmo período de incidência de Zika, levantou a hipótese de existência de uma associação causal entre o ZIKV e a microcefalia. Posteriormente, investigações confirmaram tal afirmação (CALVET *et al.*, 2016; MELO *et al.*, 2016; XIMENES *et al.*, 2023), o que fez com que o Ministério da Saúde declarasse Emergência de Saúde Pública de Importância Nacional (ESPIN) (BRASIL, 2019).

Os casos de Zika e microcefalia relatados no Brasil, inicialmente, estiveram concentrados na região Nordeste e, posteriormente, se espalharam pelo país, de modo que, ao final do ano de 2015, haviam sido relatados casos em todas as cinco regiões brasileiras. No entanto, foi verificada uma grande desigualdade na distribuição de tais agravos entre as regiões do país (LOWY, 2019).

Até os tempos recentes, as razões para as diferenças espaciais na incidência de Zika e microcefalia no Brasil não foram completamente elucidadas (DIDERICHSEN; AUGUSTO, 2021). Contudo, a existência de condições socioeconômicas, climáticas e de saneamento díspares entre as regiões brasileiras pode representar um ponto de partida para o entendimento dessas questões. Chitolina *et al.* (2016), Han *et al.* (2016) e Silva e Machado (2018) já demonstraram que as condições de saneamento de uma região podem influir na incidência de arboviroses. Desse modo, é possível que a junção de fatores tradicionalmente avaliados em estudos de arboviroses, tais como climáticos e socioeconômicos, com fatores sanitários possa contribuir para uma identificação mais acurada de áreas de risco. Isso ocorre devido à distribuição espacial e densidade vetorial de mosquitos do gênero *Aedes* serem determinadas a partir interação entre os fatores citados (DIDERICHSEN; AUGUSTO, 2021). De acordo com Dhaka e Singh (2020), a junção de fatores diversos, que influenciam na propagação de arboviroses, pode auxiliar na criação de um sistema de alerta epidêmico capaz de identificar áreas de risco.

Diante das possibilidades descritas, verifica-se que as diferenças na prevalência de casos de Zika e microcefalia nas regiões brasileiras podem estar associadas às condições socioeconômicas e de saneamento distintas no território nacional. Contudo, ainda não foram desenvolvidos estudos que avaliaram tal relação causal com o uso de informações atualizadas dos municípios brasileiros. Adicionalmente, os estudos que avaliaram a relação entre indicadores socioeconômicos e de saneamento e outras arboviroses no Brasil concentraram-se em regiões específicas do país ou não fizeram uso de métodos estatísticos robustos (ALMEIDA; COTA; RODRIGUES, 2019; DIDERICHSEN; AUGUSTO; PEREZ, 2018; RODRIGUES *et al.*, 2018; SILVA; MACHADO, 2018; SOUZA *et al.*, 2018).

Nesse sentido, a utilização de métodos estatísticos multivariados, estatística espacial e, mais recentemente, de modelos de aprendizado de máquina, tem se mostrado uma ferramenta útil na predição de fenômenos complexos que demandam a avaliação simultânea de múltiplas variáveis (QUEIROZ; MEDRONHO, 2021; ZHAO *et al.*, 2020). A capacidade de tais métodos de lidar com dados com as características citadas, indica uma potencialidade para a predição da incidência de Zika e microcefalia, tendo em vista a utilização de diversas covariáveis socioeconômicas e de saneamento, cuja relação com a doença ainda não foi completamente elucidada.

Em trabalhos recentes, vários pesquisadores demonstraram a aplicabilidade de métodos estatísticos multivariados em estudos epidemiológicos e ambientais (CARDO *et al.*, 2018;

DESJARDINS *et al.*, 2018; MAITRA *et al.*, 2019; SILVA *et al.*, 2018). Modelos preditivos baseados em algoritmos de aprendizado de máquina também têm sido utilizados em diversos estudos na área da saúde, demonstrando potencial para atuar como uma ferramenta de apoio à tomada de decisão (DHAKA; SINGH, 2020; GALASSO *et al.*, 2022; ISMAIL *et al.*, 2022; LIMA *et al.*, 2021; RAHMAN *et al.*, 2021). No entanto, Rocha e Giesbrecht (2022) destacam que há uma carência de estudos voltados para predição de Zika e síndromes associadas a esse vírus, o que pode estar relacionado à escassez de dados, uma vez que, além das epidemias associadas ao vírus ainda poderem ser consideradas recentes, a Zika integra a lista de doenças tropicais negligenciadas. Adicionalmente, a qualidade dos dados contidos nos sistemas de informação e as falhas no processo de vigilância em saúde podem dificultar o desenvolvimento de estudos acerca da referida temática.

A utilização conjunta de métodos estatísticos multivariados e algoritmos de aprendizagem de máquina pode atuar como uma ferramenta eficaz na identificação de fatores de influência em problemas de saúde pública. Ainda, os resultados obtidos por meio dessas análises podem subsidiar a tomada de decisão e definição de estratégias para prevenção de epidemias. Diante disso, nesse estudo teve-se como intuito realizar uma análise multi-abordagem acerca da relação entre as condições sanitárias, socioeconômicas, ambientais e climáticas e a ocorrência das epidemias do vírus Zika e microcefalia no Brasil.

O conteúdo desse estudo foi estruturado em quatro capítulos independentes. O primeiro capítulo refere-se a uma revisão da literatura, que aborda os aspectos relativos à incidência de arboviroses e aos diferentes métodos que podem ser utilizados para a avaliação e predição dessas doenças. O segundo capítulo disponibiliza uma revisão sistemática da literatura que retrata os efeitos das mudanças climáticas na incidência e distribuição futura de arboviroses. No terceiro capítulo são avaliados três aspectos relativos à incidência de Zika e microcefalia no Brasil: a existência de correlação espacial entre as taxas de incidência de Zika e microcefalia; a existência de relações entre as taxas da incidência de Zika e microcefalia e variáveis sanitárias, socioeconômicas e ambientais; e a existência de diferenças entre as taxas da incidência de Zika e microcefalia nos municípios brasileiros que possuem histórico de ocorrência de eventos climáticos extremos. O quarto capítulo investiga a possibilidade de predizer quais municípios brasileiros apresentam tendência para ocorrência de eventos de Zika e microcefalia a partir do conhecimento das características sanitárias, socioeconômicas, climáticas e ambientais desses municípios e uso de algoritmos de aprendizado de máquina.

2. HIPÓTESES DE INVESTIGAÇÃO

A partir de uma análise da literatura e o estudo das lacunas do conhecimento, na presente tese são propostas as seguintes hipóteses:

Hipótese 1

As mudanças climáticas globais possuem o potencial de afetar a incidência de arboviroses, seja provocando o aumento destas ou a diminuição em locais distintos.

Hipótese 2

As taxas de incidência de Zika e microcefalia nos municípios brasileiros estão associadas à localização geográfica de ocorrência dos eventos.

Hipótese 3

A elevada variabilidade espacial das taxas de incidência de Zika e microcefalia no Brasil pode ser explicada a partir da avaliação das diferentes características sanitárias, socioeconômicas e ambientais dos municípios brasileiros.

Hipótese 4

Existem diferenças entre as taxas de incidência de Zika e microcefalia nos municípios brasileiros que possuem histórico de ocorrência de eventos climáticos extremos, tais como secas e cheias. Adicionalmente, os municípios que apresentaram eventos de secas ou estiagens estão localizados, majoritariamente, na região Nordeste do Brasil e apresentam as maiores taxas de incidência de microcefalia.

Hipótese 5

É possível prever quais municípios brasileiros apresentam tendência para ocorrência de eventos de Zika e microcefalia a partir do conhecimento das características sanitárias,

socioeconômicas, climáticas e ambientais desses municípios e uso de algoritmos de aprendizado de máquina.

3. OBJETIVO GERAL

Realizar uma análise multi-abordagem acerca da relação entre as condições sanitárias, socioeconômicas, ambientais e climáticas e a ocorrência das epidemias do vírus Zika e microcefalia no Brasil.

3.1 Objetivos específicos

1. Investigar se as mudanças climáticas globais possuem o potencial de afetar o padrão de incidência de arboviroses;
2. Avaliar a existência de correlação espacial entre as taxas de incidência de Zika e microcefalia nos municípios brasileiros;
3. Relacionar as taxas da incidência de Zika e microcefalia nos municípios brasileiros com variáveis sanitárias, socioeconômicas e ambientais;
4. Avaliar a existência de diferenças entre as taxas da incidência de Zika e microcefalia nos municípios brasileiros que possuem histórico de ocorrência de eventos climáticos extremos;
5. Predizer quais municípios brasileiros possuem maior propensão de apresentar eventos de Zika e microcefalia, a partir do conhecimento das características sanitárias, socioeconômicas e ambientais dos municípios e uso de algoritmos de aprendizado de máquina.

4. ESTRUTURAÇÃO DA TESE

De posse das hipóteses, apresenta-se a no Quadro 1 estruturação da tese.

Quadro 1 – Estruturação da tese

Hipótese 1 – Objetivo específico 1	
Aspectos avaliados	Resultados esperados
<ul style="list-style-type: none"> • Realização de uma revisão sistemática da literatura e descrição dos efeitos das mudanças climáticas na incidência e distribuição futura de arboviroses. • Avaliação dos possíveis efeitos das mudanças climáticas na distribuição de vetores e na transmissão viral, considerando-se estudos em escala global, nacional, regional e local. 	<ul style="list-style-type: none"> • As mudanças climáticas poderão afetar a faixa de distribuição geográfica do <i>A. aegypti</i> e <i>A. albopictus</i>, bem como a aptidão das regiões para transmissão viral. • Os efeitos das mudanças climáticas variam em diferentes regiões, podendo causar, no futuro, expansão ou contração das áreas atualmente consideradas aptas para o desenvolvimento de arboviroses.
Hipótese 2 - Objetivo específico 2	
Aspectos avaliados	Resultados esperados
<ul style="list-style-type: none"> • Realização de uma análise exploratória de dados espaciais, relativos à incidência de Zika e microcefalia em decorrência da Zika no Brasil, e construção de mapas temáticos, que serão utilizados como ferramenta de análise prévia da distribuição dos casos de Zika e microcefalia no Brasil no ano de 2016. 	<ul style="list-style-type: none"> • A ocorrência de altas ou baixas taxas de incidência de Zika e microcefalia nos municípios brasileiros está associada à localização geográfica de ocorrência dos eventos. • De maneira semelhante à verificada no período inicial da epidemia por Zika, os casos de microcefalia concentraram-se na região Nordeste do Brasil.

<ul style="list-style-type: none"> • Cálculo dos índices globais e locais de associação espacial para identificação de independência ou correlação espacial dos dados de Zika e microcefalia nos municípios brasileiros. • Construção de mapas que indicarão clusters espaciais dos municípios e <i>outliers</i> estatisticamente significativos na área de estudo. 	<ul style="list-style-type: none"> • A análise comparativa dos clusters formados para os eventos de Zika e microcefalia permite identificar se as maiores taxas de microcefalia ocorreram ou não nos mesmos agrupamentos espaciais que apresentaram maiores taxas de Zika.
Hipótese 3 - Objetivo específico 3	
Aspectos avaliados	Resultados esperados
<ul style="list-style-type: none"> • Agrupamento dos municípios brasileiros em que foram confirmados casos de Zika e microcefalia, levando em consideração os indicadores socioeconômicos, sanitários e ambientais. • Realização de uma análise exploratória de agrupamentos (utilizando procedimentos hierárquicos e não hierárquicos) e posterior aplicação de testes estatísticos para identificação de diferenças entre os dados de Zika e microcefalia nos grupos formados, a fim de verificar se os casos ocorreram, predominantemente, em regiões com menores índices socioeconômicos e de saneamento no país. 	<ul style="list-style-type: none"> • As maiores taxas de incidência de Zika e microcefalia ocorreram, majoritariamente, nos agrupamentos formados por municípios que apresentam menores índices socioeconômicos e de saneamento no país. • As diferentes características ambientais, sanitárias e socioeconômicas dos municípios brasileiros podem explicar a elevada variabilidade espacial das taxas de incidência de Zika e microcefalia no Brasil.
Hipótese 4 - Objetivo específico 4	
Aspectos avaliados	Resultados esperados
<ul style="list-style-type: none"> • Avaliação dos registros relacionados a eventos de cheias (inundações, enchentes, alagamentos e enxurradas) e secas (estiagens e secas) nos municípios que apresentaram casos confirmados de Zika e microcefalia. 	<ul style="list-style-type: none"> • Existem diferenças entre as taxas de incidência de Zika e microcefalia nos municípios brasileiros que possuem histórico de ocorrência de eventos climáticos extremos, tais como secas e cheias. Assim, a ocorrência de eventos

- Determinação da existência de diferenças entre as taxas da incidência de Zika e microcefalia nos municípios brasileiros que possuem histórico de ocorrência de eventos climáticos extremos.

climáticos extremos pode desempenhar um papel importante na propagação de arboviroses.

- Os municípios que apresentaram eventos de secas ou estiagens estão localizados, majoritariamente, na região Nordeste do Brasil e apresentam as maiores taxas de incidência de microcefalia no referido ano.

Hipótese 5 - Objetivo específico 5

Aspectos avaliados

- Avaliação de diferentes algoritmos de aprendizado de máquina e identificação do algoritmo com melhor capacidade de predição de eventos de Zika e microcefalia nos municípios brasileiros.
- Seleção das variáveis (sanitárias, socioeconômicas e climáticas) que melhor contribuem para a predição dos eventos de Zika e microcefalia nos municípios brasileiros.

Resultados esperados

- Desenvolvimento de um modelo para predição dos municípios brasileiros que apresentam tendência para ocorrência de eventos de Zika e microcefalia, a partir do conhecimento das características sanitárias, socioeconômicas e climáticas desses municípios e uso de algoritmos de aprendizado de máquina.
- Identificação do algoritmo de aprendizagem de máquina que melhor prediz a incidência de Zika e microcefalia nos municípios brasileiros.
- As variáveis relacionadas aos aspectos sanitários dos municípios estão entre as mais importantes para predição das taxas de incidência de Zika e microcefalia no Brasil.
- O modelo proposto pode ser utilizado como ferramenta eficaz na identificação de fatores de influência em problemas de saúde pública associados à Zika e/ou microcefalia.

CAPÍTULO 1

Revisão Bibliográfica

1.1 Arboviroses

Segundo Silva e Angerami (2008), o termo arbovírus foi concebido na década de 1930 e designa o conjunto de vírus carreados por artrópodes. Tais vírus possuem como característica comum o fato de parte do ciclo de replicação viral ocorrer no organismo do vetor. Adicionalmente, o ciclo dos arbovírus inclui um ou mais hospedeiros vertebrados e vetores artrópodes, sendo estes, majoritariamente, mosquitos hematófagos (LIMA; SOARES; SILVA, 2022).

Às doenças causadas por arbovírus dá-se o nome de arboviroses. Embora muitas arboviroses sejam causa de morbidade e mortalidade humana, Hilgenfeld e Vasudevan (2018) e Silva e Angerami (2008) destacam que, originalmente, elas não se tratam de uma doença de humanos. Os autores afirmam que a ocorrência de tais doenças em humanos se dá quando ocorre alguma alteração ecológica significativa nos habitats naturais, o que resulta em alterações de reservatórios e vetores.

Nos tempos recentes, arbovírus e arboviroses podem ser identificados em diversos tipos de ecossistemas, desde que seja possível a presença de artrópodes nestes. Nos últimos 30 anos, houve o surgimento (ou ressurgimento) de diversos arbovírus que representaram, globalmente, uma notável ameaça à saúde humana. A considerável expansão geográfica das áreas de incidência do vírus da dengue (DENV), aliada a ocorrência de surtos e epidemias do vírus do Nilo Ocidental (WNV), vírus Chikungunya (CHIKV) e, mais recentemente, do vírus Zika (ZIKV) despertaram a atenção dos governos e da população acerca da possibilidade de surgimento de arboviroses a qualquer momento (HILGENFELD; VASUDEVAN, 2018; SILVA; ANGERAMI, 2008). Além disso, o expressivo número de infecções identificadas nos últimos anos resultaram em uma mudança no entendimento epidemiológico de vírus, até então, pouco conhecido (HUANG; HIGGS; VANLANDINGHAM, 2019)

A maioria dos arbovírus que causam patologias em humanos pertence aos gêneros *Flavivirus*, *Alphavirus* e *Orthobunyavirus* (BRASIL, 2017). Os vírus dengue (DENV) e o Zika (ZIKV) pertencem ao gênero *Flavivirus* e destacam-se como vírus neuroinvasivos de interesse epidemiológico no Brasil (BRASIL, 2019). Em relação ao ZIKV, Brasil (2019) destacam que duas linhagens são conhecidas: africana e asiática.

Os vírus descritos possuem como característica comum a transmissão por meio da picada de mosquitos do gênero *Aedes* infectados. Embora apresentem competência vetorial distinta, Chouin-Carneiro *et al.* (2020) explicam que os mosquitos *Aedes aegypti* e de *Aedes albopictus* podem ser infectados pelo vírus Zika e transmitir a doença. No Brasil, as populações de *Aedes aegypti* possuem maior potencial de transmissão do Zika, sendo estas altamente antropofílicas. Por esse motivo, Brasil (2019) menciona que as arboviroses transmitidas pelo *Aedes aegypti* representam um dos principais problemas de saúde pública no país.

Os arbovírus DENV, CHIK e ZIKV podem ser transmitidos ao homem pelas vias vetorial, vertical e transfusional. A primeira representa a principal forma de transmissão, ocorrendo por meio da picada de fêmeas de mosquitos do gênero *Aedes* infectadas, no denominado ciclo humano-vetor-humano. Em relação ao vírus Zika, a transmissão vertical pode dar-se em idades gestacionais distintas e resultar em malformações no feto. Embora sejam relatados casos de transmissão dos referidos vírus por via transfusional, a relevância de tal via ainda é objeto de estudos. Além das três vias de transmissão citadas, é relatado que, para o vírus Zika, a transmissão pode ocorrer, ainda, por via sexual (BRASIL, 2019).

O processo de transmissão de arbovírus compreende dois períodos: período de incubação intrínseco (PII), que acontece nos humanos, e período de incubação extrínseco (PIE), que ocorre nos vetores. Em relação ao vírus Zika, o PII pode variar de 2 a 7 dias. Após isso, inicia-se o período de viremia no homem, que pode se estender até o 5º dia após o início dos sintomas. Ao picar uma pessoa infectada e que se encontra no período virêmico, o vetor inicia o PIE. Tal período é referente ao tempo decorrido entre a ingestão do sangue virêmico e o surgimento do vírus na saliva do mosquito vetor. Após o PIE, o vetor permanece infectante até o final de sua vida, que compreende um período de 6 a 8 semanas. Vale ressaltar que o PIE pode ser influenciado por fatores ambientais, sobretudo, pela temperatura. (BRASIL, 2019).

No cenário epidemiológico brasileiro, há a circulação - simultânea ou não - de quatro sorotipos do vírus dengue (DENV-1, DENV-2, DENV-3 e DENV-4), além dos vírus Chikungunya e Zika. Tal fato resulta em um grande desafio para a vigilância em saúde, tanto em relação à identificação dos casos, quanto na elaboração de ações de controle e prevenção (BRASIL, 2019). De acordo com Lima, Soares e Silva (2022), as arboviroses podem provocar quadros clínicos distintos, que variam de doenças febris, artralgia, erupções na pele a síndromes neurológicas hemorrágicas. Todavia, Brasil (2017) destacam que grande parte das infecções humanas advindas de arbovírus são assintomáticas ou oligossintomáticas, o que torna complexo a elaboração de um diagnóstico preciso.

Conforme observado, o perfil epidemiológico dos arbovírus possui natureza complexa em função da especificidade do vetor artrópode envolvido na transmissão e da participação de humanos. Desse modo, diversas particularidades regionais podem ser associadas à incidência de arboviroses, tais como as características ecológicas e ambientais, o comportamento dos vetores e indivíduos em geral, além das características climáticas (LIMA; SOARES; SILVA, 2022).

Tendo em vista que as arboviroses são transmitidas por mosquitos vetores, Lima, Soares e Silva (2022) citam que os principais métodos para prevenção dessas doenças consistem no controle de vetores e no desenvolvimento de medidas farmacológicas. Por se tratar de um problema complexo e que pode resultar em pressão sobre os sistemas de saúde, a prevenção de arboviroses exige intervenções interdisciplinares e intersetoriais, sobretudo em regiões com maior vulnerabilidade (GOMES; JESUS; QUARESMA, 2023).

1.1.1 Zika: origens

O vírus Zika (ZIKV) foi identificado em 1947 em um macaco *rhesus* na floresta Zika, localizada em Uganda (LARRANDABURU *et al.*, 2017). No entanto, de acordo com Pinto Junior *et al.* (2015), análises filogenéticas realizadas anos mais tarde sugerem que o surgimento do vírus nessa localidade se deu por volta de 1920. Segundo Wikan e Smith (2016), a primeira descrição formal desse vírus foi publicada no ano de 1952. No entanto, ao longo dos 60 anos posteriores ao citado, os estudos a respeito do ZIKV estiveram restritos a poucos pesquisadores especializados nessa temática.

Por cerca de 30 anos após a identificação do ZIKV, a circulação do vírus esteve restrita ao continente africano. Após esse período, o vírus se espalhou pela Ásia, onde foram identificados episódios esporádicos da infecção em humanos (OPAS, 2022). Por se tratarem de episódios esporádicos, a importância do vírus Zika não era considerada significativa em termos de saúde pública (HUANG; HIGGS; VANLANDINGHAM, 2019).

Somente em 2007 foram relatados os primeiros casos de infecção pelo ZIKV fora da África e da Ásia. O primeiro relato ocorreu em abril de 2007, quando foi identificado o primeiro surto do vírus no estado de Yap, parte dos Estados Federados da Micronésia (LARRANDABURU *et al.*, 2017; WIKAN; SMITH, 2016). Nesse episódio, é relatado que cerca de 75% da população da ilha foi infectada pelo vírus Zika. Desde então, infecções pelo ZIKV foram relatadas em outras ilhas do Oceano Pacífico (OPAS, 2022).

Nas Américas, os primeiros casos de infecção pelo vírus Zika foram relatados no Brasil. No primeiro semestre de 2015, as autoridades de saúde pública brasileiras confirmaram a transmissão do ZIKV em alguns estados no nordeste do país. Nos meses posteriores, o vírus foi detectado em todas as regiões brasileiras e em diversos países do continente americano (BRASIL, 2019; CHOUIN-CARNEIRO *et al.*, 2016). Na Figura 1.1 o padrão de disseminação do vírus Zika no mundo pode ser observado.

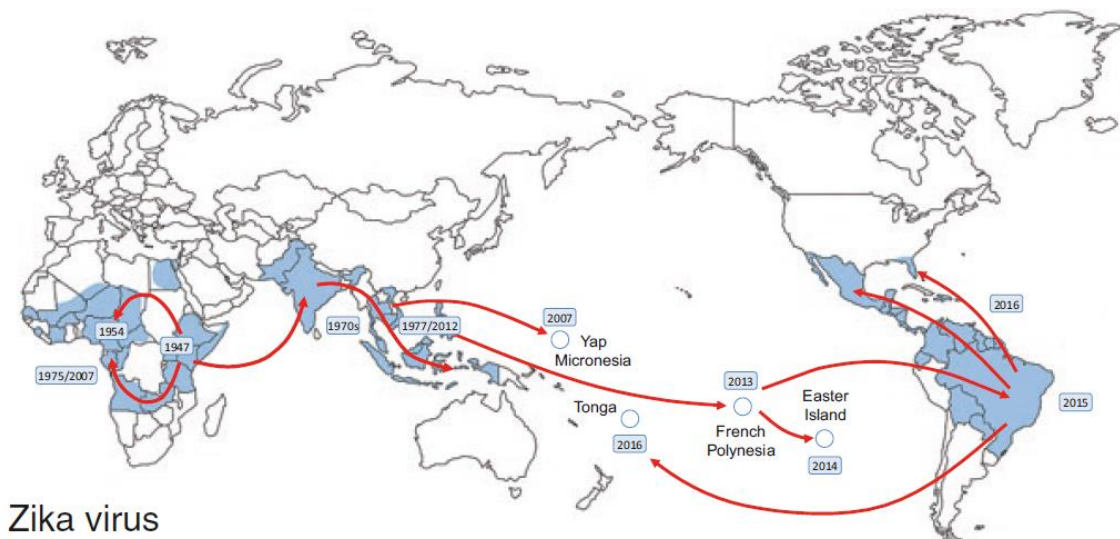


Figura 1.1. Padrão de disseminação do vírus Zika no mundo ao longo dos anos.

Fonte: Hilgenfeld e Vasudevan (2018).

No segundo semestre do ano de 2015 foi relatada a associação do vírus Zika com a síndrome de Guillain-Barré e com malformações do sistema nervoso central, incluindo microcefalia, em recém-nascidos cujas mães foram infectadas pelo ZIKV durante a gestação (OPAS, 2022). Esses fatos resultaram na declaração de Emergência de Saúde Pública de Importância Internacional pela Organização Mundial da Saúde (OMS) em fevereiro de 2016 (LARRANDABURU *et al.*, 2017).

1.1.2 Evolução da epidemia do vírus Zika no Brasil

Conforme anteriormente relatado, os primeiros casos de infecção pelo vírus Zika foram identificados em meados de março de 2015 na região Nordeste do Brasil. Nesse período, médicos identificaram uma rápida propagação de uma doença que apresentava sintomas semelhantes à dengue (BRASIL, 2019).

Os primeiros casos de Zika foram identificados no estado do Rio do Grande do Norte e, em seguida, nos estados da Paraíba e da Bahia. Nesse contexto, testes de biologia molecular foram realizados e estes excluíram o vírus da dengue e Chikungunya como possíveis causadores da infecção. Posteriormente, pesquisadores de Salvador (BA) identificaram que as infecções relatadas estavam associadas à circulação do vírus Zika no país (GARCIA, 2018; LOWY, 2019). Estudos filogenéticos identificam que a cepa em circulação no Brasil pertencia à linhagem asiática e apresentava características semelhantes às identificadas em amostras coletadas, anteriormente, na Polinésia Francesa (GARCIA, 2018).

Após a identificação da presença do ZIKV no Brasil, houve muita especulação a respeito de como o vírus teria sido introduzido no país. Virologistas levantaram diversas possibilidades, destacando-se o possível papel das competições esportivas. A discussão esteve focada em três eventos: Copa das Confederações (15 a 30 de junho de 2013), Copa do Mundo de Futebol (12 de junho a 13 de julho de 2014) e Campeonato Mundial de Canoagem (12 a 17 de agosto de 2014). Contudo, alguns pesquisadores acreditam que o ZIKV pode ter sido introduzido no país pela visita de uma ou mais pessoas infectadas, não estando, necessariamente, associado aos eventos citados (LOWY, 2019).

De acordo com Garcia (2018), o fato de a maioria das infecções pelo ZIKV ser assintomática, ou apresentar sintomas leves, fez com que o vírus circulasse na região Nordeste sem que fosse detectado. Aliado a isso, no mesmo período, houve a circulação concomitante de outros arbovírus na região, o que dificultou a realização de um diagnóstico preciso.

Em abril de 2015, a Zika foi incluída no monitoramento de doenças exantemáticas pela Secretaria de Vigilância em Saúde do Brasil. No mesmo período, a Organização Pan-Americana da Saúde (Opas) emitiu um alerta sobre a circulação do ZIKV no nordeste do Brasil (LOWY, 2019). Em menos de um ano após a sua identificação no país, o ZIKV se disseminou por todas as regiões brasileiras, embora com distribuição desigual entre estas, sendo predominante na região Nordeste (GARCIA, 2018).

A princípio, quando o vírus Zika foi identificado no Brasil, a principal preocupação dos especialistas estava associada à sua circulação concomitante com os vírus da dengue e Chikungunya, que poderia agravar os efeitos adversos desses patógenos (LOWY, 2019). Em outubro de 2015, foi identificado um aumento atípico do número de casos de nascidos vivos com microcefalia, inicialmente no estado do Pernambuco e, em seguida, nos demais estados da região Nordeste. Tendo em vista que tal aumento ocorreu após a identificação da febre pelo vírus Zika na região, foi levantada a hipótese de existência de uma associação causal entre o ZIKV e a microcefalia. Posteriormente, investigações confirmaram tal afirmação, o que fez

com que o Ministério da Saúde declarasse Emergência de Saúde Pública de Importância Nacional (ESPIN) (BRASIL, 2019).

A associação entre a infecção pelo vírus Zika em gestantes e o risco de malformações congênitas, até então, não havia sido descrita nos demais países que observaram a incidência de Zika. Nesse ponto, vale ressaltar a importância da atuação dos médicos e pesquisadores brasileiros, que foram pioneiros ao levantar a hipótese da existência da referida relação causal (BRASIL, 2019).

Entre 2015 e 2016, 2018 casos de microcefalia haviam sido confirmados em todo o Brasil. No entanto, a distribuição desses casos entre as regiões brasileiras ocorreu de maneira desigual. Cerca de 80% dos casos foram registrados na região Nordeste, 10% na região Sudeste, 4% no Centro-Oeste, 3% no Norte e 0,2% na região Sul. Segundo Garcia (2018), tais dados reforçam a existência da acentuada desigualdade regional existente no Brasil. O referido autor destaca que as razões para as diferenças espaciais e temporais observadas na incidência de microcefalia nas diferentes regiões brasileiras ainda não estão completamente elucidadas. Adicionalmente, os dados de vigilância indicam que o crescimento da prevalência de microcefalia no nordeste brasileiro foi muito superior ao identificado (tempos após os eventos) nos demais países acometidos pelo vírus Zika. Diante disso, Garcia (2018) cita a possibilidade de existência de condições ambientais e sociais propícias para a disseminação do ZIKV na região Nordeste e infecção de gestantes.

Segundo Brasil (2017), a disseminação do ZIKV no Brasil modificou o cenário epidemiológico de manifestações neurológicas. Além dos casos de microcefalia, após a detecção do vírus no país, foi observado um aumento no número de pacientes com manifestações neurológicas e histórico de infecção viral. Dentre tais manifestações, incluem-se os casos de encefalite, meningoencefalite, mielite e, sobretudo, síndrome de Guillain-Barré (SGB) (HUANG; HIGGS; VANLANDINGHAM, 2019).

Durante o período de ocorrência desses eventos, houve uma grande preocupação da população, sobretudo entre as gestantes e mulheres em idade fértil, acerca da possibilidade de contração da doença de Zika. Nesse cenário, a posição do governo brasileiro era de que a principal forma de controle dos casos de microcefalia estava associada ao combate ao vetor da doença (*Aedes aegypti*). Adicionalmente, foram desenvolvidos protocolos de vigilância da doença e de atenção às gestantes e crianças, além de outras ações, que incluíram o fomento à pesquisa e desenvolvimento de estratégias voltadas para o enfrentamento da Zika e síndromes associadas (GARCIA, 2018).

Segundo Huang, Higgs e Vanlandingham (2019), a epidemia do vírus Zika demonstrou como o surgimento de um patógeno pouco estudado pode afetar, negativamente, a saúde humana em escala global. No Brasil, a evolução da epidemia do ZIKV evidenciou a manifesta dificuldade do país em realizar o controle vetorial, além de expor falhas na atenção materno-infantil. Tais adversidades, associadas às expressivas desigualdades existentes no país, podem ser apontadas como fatores que contribuíram para que a Zika e a microcefalia atingissem, majoritariamente, famílias pobres e residentes das regiões menos desenvolvidas do país. (GARCIA, 2018).

Para Hilgenfeld e Vasudevan (2018), uma das formas de prevenção e de preparação para possíveis epidemias/pandemias futuras associadas ao ZIKV compreende o claro entendimento dos mecanismos que impulsionam as mudanças nos ciclos de transmissão de arboviroses. Associado a isso, o conhecimento do padrão de distribuição dos vetores pode auxiliar na identificação da população humana exposta ao vírus e, conseqüentemente, gerar subsídios para o combate a doenças emergentes.

1.1.3 Arboviroses, saneamento e determinantes ambientais

De acordo com a Comissão Nacional sobre os Determinantes Sociais da Saúde (CNDSS), grande parte da carga mundial de morbidade está associada às condições as quais as pessoas estão expostas ao longo da vida. Tais condições são abarcadas no conceito de Determinantes Sociais de Saúde (DSS). Os DSS englobam os fatores sociais, econômicos, culturais, étnicos/raciais, psicológicos e comportamentais que influenciam a ocorrência de problemas de saúde. Alguns exemplos de DSS incluem as condições de habitação, saneamento, educação e serviços de saúde dos locais em que as pessoas vivem e trabalham.

Diante disso, verifica-se que a compreensão das relações entre saneamento básico, saúde pública e meio ambiente revela-se um pressuposto fundamental para o planejamento de ações de promoção de saneamento e prevenção de doenças (SOARES; BERNARDES; NETTO, 2002). No Brasil, o déficit nos serviços de saneamento básico é acentuado, de modo que, parcela expressiva da população não dispõe de sistemas de coleta e tratamento de esgotos, abastecimento de água e disposição de resíduos sólidos apropriados (ANA, 2017).

A ausência de condições apropriadas de saneamento acarreta maior risco de incidência de doenças de veiculação hídrica e de transmissão vetorial. Dentre as últimas citadas, no contexto epidemiológico brasileiro, destacam-se as doenças transmitidas por mosquitos, principalmente por espécies do gênero *Aedes*. Tais mosquitos desenvolvem-se, sobretudo, em

regiões que apresentam condições de infraestrutura deficiente, favorecendo o seu estabelecimento e, conseqüentemente, a disseminação dos sorotipos virais por ele veiculados (BARRERA *et al.*, 2008)

Em áreas cujo acesso aos serviços públicos de saneamento básico é dificultado, a incidência do *Aedes aegypti* pode estar relacionada a práticas inapropriadas de armazenamento de água e de disposição de efluentes e resíduos sólidos (HAN *et al.*, 2016; SILVA; MACHADO, 2018). Além disso, um estudo desenvolvido por Mackay *et al.* (2009), identificou a presença de *Aedes aegypti* em fossas sépticas. Os autores verificaram uma maior taxa de reprodução e emergência de mosquitos de maior tamanho nesse ambiente do que em recipientes de superfície que continham água limpa.

Embora, comumente, seja citado que fêmeas de *Aedes aegypti* são atraídas para recipientes que contêm água com baixos índices de poluição, estudos têm demonstrado que quanto maior o grau de poluição da água, maior a atratividade da mesma ao mosquito (BESERRA *et al.*, 2010). De acordo com Chitolina *et al.* (2016), a presença de coliformes termotolerantes em esgotos brutos é capaz de aumentar a atratividade para a oviposição do *Aedes aegypti*.

Hai *et al.* (2021), ao avaliarem a capacidade de larvas do mosquito *Aedes aegypti* eclodirem e se reproduzirem em diferentes tipos de água, identificaram que o *Aedes aegypti* é capaz de sobreviver e se reproduzir de forma eficaz em águas poluídas. Corroborando com o referido estudo, Beserra *et al.* (2010) verificaram grande capacidade de desenvolvimento do *Aedes aegypti* em águas com elevados níveis de poluição, tais como esgoto bruto e efluentes de reator UASB, lagoa de polimento e filtro anaeróbio. Os autores explicam que a existência de altos níveis de turbidez e maior concentração de nutrientes em esgotos são fatores que podem contribuir na atração para oviposição do mosquito. As referidas informações possuem expressiva relevância para a orientação de ações de controle do *Aedes aegypti*, tendo em vista que estas apontam para a necessidade de inspeção de todos os tipos de coleções de água, sobretudo, em áreas que apresentam baixos índices de coleta e tratamento de esgotos.

A falta ou a intermitência do acesso à rede de distribuição de água também é apontada como um fator capaz de influenciar a incidência de doenças transmitidas por mosquitos. Isso ocorre devido à necessidade de armazenamento doméstico de água em situações de falta, o que resulta na criação de locais propícios para a reprodução dos mosquitos vetores (DADA *et al.*, 2013).

Adicionalmente, Gomes, Jesus e Quaresma (2023) avaliaram a relação entre a incidência de arbovirose e a infraestrutura urbana existente de saneamento básico. Os autores

identificaram uma relação inversa entre os fatores, onde a maioria dos municípios avaliados que possuíam alta incidência de arboviroses apresentava baixos níveis de saneamento. Além das características sanitárias, Gomes, Jesus e Quaresma (2023) destacam a existência de estudos que demonstram a correlação existente entre indicadores socioeconômicos, tais como produto interno bruto (PIB) per capita, índice de desenvolvimento humano (IDH) e taxa de pobreza, e elevadas taxas de incidência de arboviroses.

Fatores socioeconômicos também foram utilizados para a identificação do perfil das mães de nascidos com microcefalia, em decorrência do ZIKV, no estudo de Freitas *et al.* (2019). Os autores verificaram que tais mães, predominantemente, se autodeclaravam “não brancas”, residiam em áreas periféricas consideradas “precárias” e possuíam renda inferior a um salário mínimo. De acordo com Freitas *et al.* (2019), o perfil identificado das mães revelou que a epidemia do ZIKV não foi equânime, atingindo, majoritariamente, a população de estratos sociais desfavorecidos. Diante disso, os referidos autores mencionam que os agravos associados ao vírus Zika podem ser determinados a partir do conhecimento das desigualdades sociais presentes no país.

Para Hilgenfeld e Vasudevan (2018), a transmissão de arboviroses pode ser influenciada por diversos fatores, com destaque para os fatores climáticos, demográficos e sociais. Estudos indicam que fatores climáticos como temperatura, precipitação e umidade podem afetar a população de mosquitos (reprodução, desenvolvimento, distribuição), propagação de vírus (replicação) e a interação vetor-hospedeiro (comportamento de picada de fêmeas adultas de mosquitos do gênero *Aedes*) (BARBAZAN *et al.*, 2010; GAGE *et al.*, 2008; XAVIER *et al.*, 2021). Altas temperaturas estão associadas ao aumento na reprodução de vetores, comportamento de mordida mais frequente e redução do período de incubação do patógeno no interior dos vetores. Por outro lado, o tempo de sobrevivência do mosquito pode ser reduzido em condições muito quentes e secas (WHO, 2008).

Ainda, condições de umidade e precipitação locais são apontadas por Dada *et al.* (2013) como fatores capazes de afetar a disponibilidade de habitats necessários durante os estágios larvais dos vetores. O aumento nos índices de precipitação pluviométrica pode elevar a população de vetores quando ocorre o acúmulo de água de chuva em recipientes nas residências. No entanto, condições de seca e escassez hídrica também podem influenciar no aumento da população de vetores, tendo em vista a necessidade de armazenamento doméstico de água nesse período, que resulta na criação de locais propícios para a reprodução dos mosquitos.

Diante do exposto, é possível verificar que diversos fatores sanitários, socioeconômicos e climáticos podem estar associados à disseminação de doenças arbovirais. Desse modo, uma

melhor compreensão da interação de tais fatores com a incidência de arboviroses pode auxiliar no desenvolvimento de mecanismos de previsão de surtos e de estratégias para mitigação de epidemias (GOMES; DE JESUS; QUARESMA, 2023). Nesse sentido, a utilização de métodos estatísticos multivariados, estatística espacial e, mais recentemente, de modelos de aprendizado de máquina, tem se mostrado uma ferramenta útil na explicação de fenômenos complexos que demandam a avaliação simultânea de múltiplas variáveis.

1.2 Aprendizado de máquina

De acordo com Murphy (2012), aprendizado de máquina (AM) ou *machine learning* pode ser definido como um conjunto de métodos utilizados para detectar padrões em dados e, a partir disso, prever dados futuros ou realizar tomadas de decisão que envolvem incerteza. Desse modo, o objetivo dos métodos de AM é utilizar dados existentes para o desenvolvimento de modelos capazes de prever saídas para dados novos (GRUS, 2016).

O termo aprendizado de máquina foi concebido na década de 60, sendo configurado como o campo da inteligência artificial aplicado no aprendizado de padrões em bases de dados. Embora, inicialmente, as aplicações de AM estivessem restritas ao campo computacional, desde o final dos anos 90, a área se expandiu para diversas áreas do conhecimento (IZBICKI; SANTOS, 2020). Atualmente, as aplicações de AM são interdisciplinares, sendo encontrados estudos em áreas como saúde (ALBAHRA *et al.*, 2023), agricultura (MESHRAM *et al.*, 2021), marketing (NGAI; WU, 2022), meio ambiente (HAGGERTY *et al.*, 2023), ciência dos materiais (OKAFOR *et al.*, 2023), automação (DOĞAN; KORKMAZ; KIRMACI, 2023).

De acordo com Izbicki e Santos (2020), a ampla utilização de técnicas de AM nos dias atuais está associada aos avanços computacionais e à necessidade de processamento de banco de dados de grande dimensão. Diversas características que tais banco de dados podem apresentar, tais como existência de mais covariáveis que observações, distribuição não paramétrica, interações complexas e não lineares, dificultam a utilização de métodos tradicionais (IZBICKI; SANTOS, 2020).

Em função do tipo de aprendizado utilizado, Murphy (2012) cita que os problemas de AM podem ser divididos em dois grupos principais: aprendizado supervisionado e aprendizado não supervisionado. A seleção do melhor método a ser utilizado dependerá da natureza do banco de dados avaliado, bem como dos objetivos pretendidos com a realização da análise.

1.2.1 Aprendizado supervisionado

Os algoritmos utilizados no aprendizado supervisionado, também denominado aprendizado preditivo, possuem como principal característica a utilização de bancos de dados que contém rótulos, ou seja, possuem variável resposta. Desse modo, o objetivo do aprendizado supervisionado é aprender a fazer previsões a partir de um banco de dados que contém entradas x (covariáveis) e saídas y (variável resposta/rótulo), dado um conjunto rotulado de pares entrada-saída $D = \{(x_i, y_i)\}$ (IZBICKI; SANTOS, 2020). Um exemplo de banco de dados utilizado nesse tipo de aprendizado é indicado na Figura 1.2.

		Covariáveis			
		Cor	Comprimento	Espessura	Rótulo
Nº de casos	↓	Vermelho	11	0,01	1
		Rosa	10	0,03	0
		Rosa	12	0,05	0

Figura 1.2. Exemplo de banco de dados utilizado em métodos de aprendizado supervisionado.

Os métodos de aprendizado supervisionado podem ser utilizados para análise de problemas de regressão e de classificação. No primeiro, a variável resposta é categórica ou qualitativa, ou seja, contém um número finito de valores ou categorias distintas. Nos problemas de regressão, por outro lado, são utilizados bancos de dados cuja variável resposta é numérica ou quantitativa (MURPHY, 2012).

1.2.2.1 Modelos de aprendizado de máquina

Para a resolução de problemas de classificação ou de regressão utilizando aprendizado de máquina supervisionado faz-se necessária a construção de diferentes modelos e de diferentes algoritmos para ajustá-los. Isso ocorre devido ao fato de nenhum modelo ser considerado universalmente melhor. Em função da área de estudo, cada banco de dados apresenta características próprias, desse modo, o conjunto de suposições que funciona bem para alguns dados, pode não ser apropriado para dados diferentes. Como consequência, diferentes modelos são desenvolvidos e avaliados, a fim de abarcar a imensa variedade de dados que ocorrem no mundo real (MURPHY, 2012).

A construção de um modelo de predição é realizada a partir do ajuste de parâmetros realizado por um algoritmo. Após a sua construção, o modelo preditivo é utilizado para a predizer um rótulo, anteriormente desconhecido, a partir do mapeamento previamente descoberto. O processo de apresentar um novo exemplar para o modelo construído é denominado teste (SILVA; PERES; BOSCARIOLI, 2021).

Diversos algoritmos podem ser utilizados para a construção de modelos de aprendizado de máquinas. A principal diferença entre os diferentes tipos de algoritmos está associada à maneira que estes gerenciam os dados para obtenção de respostas (LABOISSIERE; FERNANDES; LAGE, 2015). Dentre os algoritmos utilizados em AM, podem ser mencionados os algoritmos *Random Forest*, *Gradient Boosting Machine*, *Multivariate Adaptive Regression Spline*, *K-nearest Neighbor*, *Support Vector Machine* e Regressão logística, cujas descrições são realizadas a seguir. A determinação de qual modelo melhor se ajusta ao banco de dados avaliado pode ser realizada utilizando diferentes métricas de avaliação de desempenho (MURPHY, 2012).

Random Forest (RF):

De acordo com Breiman (2001), *Random Forest* pode ser definido com uma algoritmo de *ensemble*. *Ensemble* é uma técnica de aprendizado de máquina que consiste na combinação do resultado de múltiplos modelos preditivos com o objetivo de melhorar a precisão e a estabilidade do modelo. Métodos de *ensemble* são conhecidos por aperfeiçoar os modelos baseados em árvore. Dentro da classe de modelos *ensemble* há duas técnicas comumente conhecidas como *Bagging* e *Boosting* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Bagging (*Bootstrap Aggregating*) é um método proposto por Breiman (1996), que consiste na geração de diversas versões de um preditor (árvores) e posterior uso das versões geradas para obtenção um preditor agregado. Em tal método, cada preditor é modelado em diferentes sub-amostras, geradas por amostragem *bootstrap* dos dados originais. Desse modo, K amostras, de mesmo tamanho, são retiradas do conjunto de dados original, com reposição. Isso significa que em cada amostra *bootstrap*, uma dada observação do conjunto original pode não aparecer ou aparecer mais de uma vez. Em geral, cada amostra *bootstrap* utiliza apenas 2/3 dos dados do conjunto original. Assim, os demais dados não utilizados em uma subamostra podem, posteriormente, serem utilizados como conjuntos de teste para avaliação dos modelos gerados (KUBAT, 2017).

Cada uma das subamostras obtidas é utilizada para o treinamento de um preditor, de forma independente. Como resultado, cada preditor se concentrará em um aspecto diferente do problema de aprendizagem (KUBAT, 2017). Em seguida, os K preditores são agregados através de um método de combinação apropriada, tal como a maioria de votos para problemas de classificação ou a média dos preditores em problemas de regressão (BREIMAN, LEO, 1996). De acordo com Breiman (1996), a agregação de preditores realizada em métodos de *Bagging* fornece um erro de predição ou classificação inferior aos modelos individuais utilizados para construí-los. Um esquema das etapas realizadas em métodos de *Bagging* é indicado na Figura 1.3.

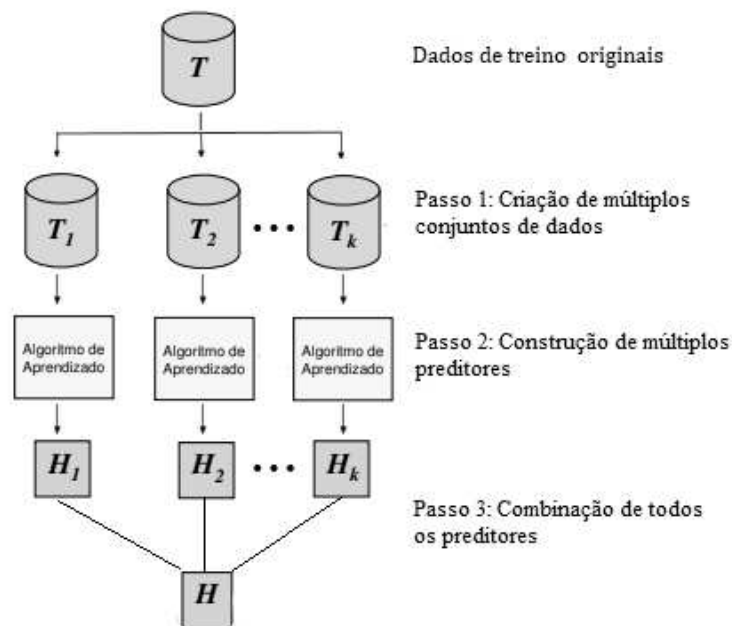


Figura 1.3. Esquema de funcionamento de métodos de bagging.

O algoritmo *Random Forest* utiliza a técnica de *bagging* para a construção de árvores de regressão ou classificação. Contudo, de acordo com Breiman (2001), tal algoritmo apresenta uma modificação substancial em relação à referida técnica, uma vez que, a utiliza em conjunto com a seleção aleatória de atributos preditivos. Com isso, o algoritmo *Random Forest* visa reduzir a correlação entre as árvores, sem aumentar a variação entre estas.

O RF faz uso de múltiplas árvores de decisão para construir modelos preditivos adaptados para resolver problemas de regressão ou classificação. Nesse algoritmo, cada árvore é ajustada usando um subconjunto, selecionado aleatoriamente, das covariáveis preditoras, o que confere uma correlação reduzida entre as árvores construídas (BREIMAN, 2001). Os modelos preditivos são obtidos particionando recursivamente o espaço de dados e ajustando um

modelo de previsão simples dentro de cada partição. Como resultado, o particionamento pode ser representado graficamente como uma árvore de decisão (LIAW; WIENER, 2002).

Árvores de classificação são projetadas para a avaliação de variáveis resposta categóricas. O resultado da classificação pode ser obtido avaliando-se o voto de classe de cada árvore e, em seguida, identificando o voto da maioria, ou seja, a moda das classes indicadas pelas diferentes árvores.

Na Figura 1.4 é apresentado um exemplo de construção de uma árvore de classificação, considerando-se dois atributos discriminativos para o problema. O objetivo da construção da árvore é determinar a qual espécie (*Íris setosa*, *Íris versicolor* ou *Íris virginica*) uma nova flor de *Íris* pertence, dada suas características (comprimento e largura da pétala). Nesse exemplo, inicialmente, o espaço bidimensional (Figura 1.4a) é dividido em duas regiões, considerando-se a covariável (característica) e o ponto de divisão que permite obter maior pureza (homogeneidade) nas sub-regiões resultantes. Cada divisão do espaço é representada por um nó na árvore de decisão (Figura 1.4b). O espaço é particionado recursivamente, por meio de cortes ortogonais que maximizem a pureza das sub-regiões resultantes, até que todas as folhas sejam puras ou alguma regra de parada seja aplicada (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Quando isso acontece, os objetos são rotulados com a classe associada a cada folha (KUBAT, 2017).

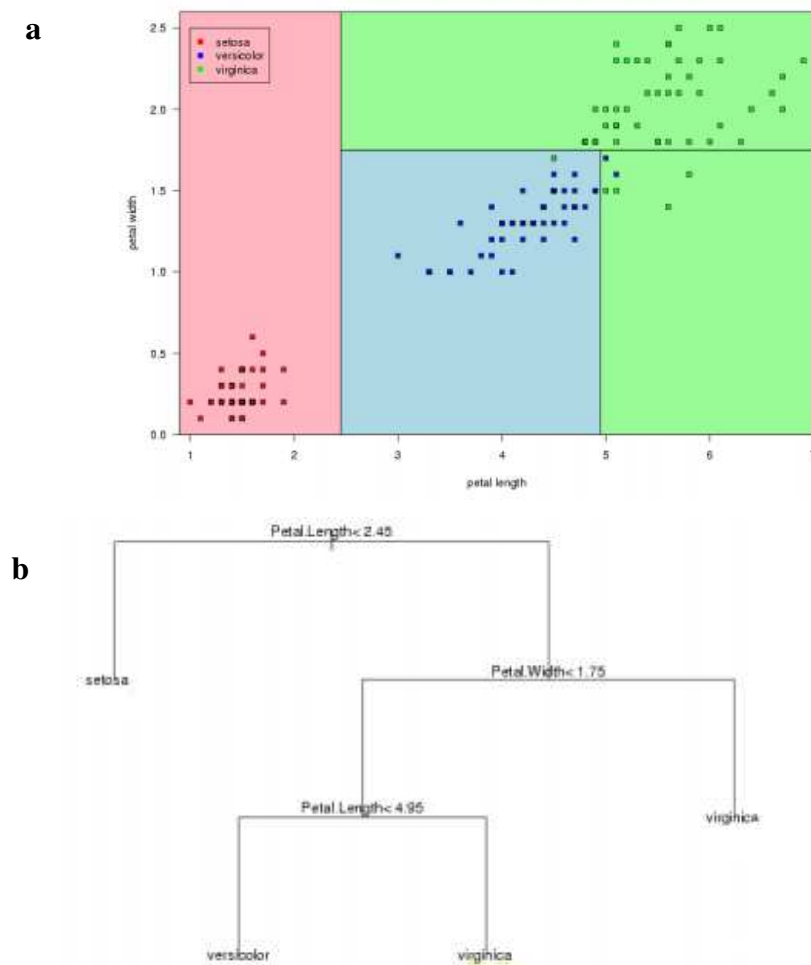


Figura 1.4. Representação do processo de construção de uma árvore de classificação. Em a) é indicada a partição do espaço de recurso bidimensional por divisão binária recursiva e em b) é indicada a árvore correspondente à partição realizada.

Fonte: Breiman *et al.* (2017)

Na literatura, diversas vantagens atribuídas ao uso do algoritmo *Random Forest* são citadas, destacando-se:

- Aplicabilidade para problemas de regressão e classificação (BREIMAN, 2001);
- Alta precisão de classificação (CUTLER *et al.*, 2007);
- Capacidade de modelar interações complexas e não lineares entre variáveis preditoras (CHAGAS *et al.*, 2016; CUTLER *et al.*, 2007; GRIMM *et al.*, 2008);
- Possibilita a análise de grande quantidade de dados (RODRIGUEZ-GALIANO *et al.*, 2012; STAVROPOULOS *et al.*, 2020);
- O algoritmo é menos propenso a sobreajuste, devido à natureza aleatória de construção de cada árvore (CHAGAS *et al.*, 2016; GRIMM *et al.*, 2008);
- Possui relativa robustez à presença de *outliers* e ruídos (BREIMAN, L., 2001);

- Fornece estimativas internas úteis de erro (erro Out-of-bag), importância de variáveis e proximidade de observações (BREIMAN, L., 2001). A capacidade de determinar a relevância das variáveis utilizadas é útil em estudos onde a dimensionalidade dos dados é alta e é relevante identificar como cada variável preditiva influencia o modelo de classificação/regressão, a fim de selecionar as melhores variáveis (BREIMAN, L., 2001; RODRIGUEZ-GALIANO *et al.*, 2012);
- A estimativa de erro usando os dados out-of-bag (OOB) é tão precisa quanto usar um conjunto de teste de tamanho similar ao conjunto de treinamento. Portanto, o uso da estimativa OOB elimina a necessidade de separação de parte dos dados para teste (BREIMAN, L., 2001).
- As estimativas de proximidade de observações fornecidas pelo algoritmo podem ser usadas para identificar *outliers* e imputar valores ausentes (RODRIGUEZ-GALIANO *et al.*, 2012);
- Exige o ajuste de poucos hiperparâmetros, quando comparado a outros algoritmos de aprendizado de máquina. Adicionalmente, os hiperparâmetros possuem fácil compreensão e é possível obter bons resultados de predição utilizando-se os valores *default* (CHAGAS *et al.*, 2016; GRIMM *et al.*, 2008; LIAW; WIENER, 2002);
- Possui natureza não paramétrica, desse modo, não faz suposições acerca da distribuição das variáveis preditoras ou de resposta (CUTLER *et al.*, 2007; RODRIGUEZ-GALIANO *et al.*, 2012);
- Possui método para imputação de valores ausentes (CUTLER *et al.*, 2007; RODRIGUEZ-GALIANO *et al.*, 2012);
- As árvores são construídas utilizando-se diferentes subconjuntos dos dados de treinamento e diferentes covariáveis das observações escolhidas. Desse modo, todas as árvores são construídas de modo independente (STAVROPOULOS *et al.*, 2020);
- O algoritmo pode lidar com situações em que o número de variáveis preditoras excede o número de observações (CUTLER *et al.*, 2007).

Apesar das vantagens citadas, o algoritmo *Random Forest* apresenta algumas desvantagens, tais como:

- O algoritmo *Random Forest* e métodos de *ensemble*, em geral, possuem interpretabilidade limitada. A relação entre os preditores e as respostas não pode ser

examinada individualmente para todas as árvores da floresta, por esse motivo, essa técnica é frequentemente denominada abordagem de “caixa preta” (GRIMM *et al.*, 2008; RODRIGUEZ-GALIANO *et al.*, 2012);

- Algoritmos como o *Random Forest* não apresentam relações simples, como uma equação matemática ou representação gráfica, entre covariáveis preditoras e variáveis resposta, o que pode tornar sua interpretação difícil (CUTLER *et al.*, 2007);
- A utilização de uma grande quantidade de árvores (visando obter uma predição com mais acurácia) pode tornar o processo de previsão demorado (STAVROPOULOS *et al.*, 2020);
- Métodos baseados em árvores, como *Random Forest* e Árvores de Decisão, não realizam extrapolações, desse modo, não apresentam bom desempenho com dados muito fora do alcance do conjunto de treinamento (HARTSHORN, 2016).

Gradient Boosting Machine (gbm):

Assim como o RF, o algoritmo *Gradient Boosting Machine* também é um algoritmo baseado em árvores, que utiliza técnicas de *ensemble*. No entanto, neste não são utilizadas técnicas de *Bagging*, mas técnicas de *Boosting*.

Ambas as técnicas consistem na utilização de K subamostras, produzidas por amostragem a partir do conjunto original de dados de treinamento. No entanto, diferente da técnica de *Bagging*, onde qualquer elemento tem a mesma probabilidade de aparecer em uma subamostra, em *Boosting*, as observações são ponderadas. Sendo assim, algumas observações estarão presentes, com maior frequência, nas subamostras geradas. A distinção entre as observações em uma subamostra é feita a partir da atribuição de maior peso para aquelas que forem incorretamente classificadas em preditores anteriores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

De acordo com Freund e Schapire (1999), inicialmente, todos os pesos são definidos igualmente mas, a cada rodada, os pesos das observações classificadas incorretamente são aumentados, de modo que o preditor fraco é forçado a se concentrar nas observações difíceis do conjunto de treinamento. Desse modo, é possível verificar que, na técnica de *Boosting*, a construção dos preditores é feita de modo sequencial e não independente, como ocorre em *Bagging*.

O princípio da técnica de *Boosting* é construir hipóteses sucessivas, de tal modo que elementos classificados incorretamente por hipóteses anteriores sejam melhor classificados por

hipóteses seguintes. Para que isso ocorra, após cada etapa do treinamento, os pesos são redistribuídos. Ao final, os resultados de todos os preditores são combinadas por meio de um voto majoritário ponderado para produzir a previsão final (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Multivariate Adaptive Regression Spline (MARS):

Multivariate Adaptive Regression Splines (*splines* de regressão adaptativa multivariada) é uma técnica de regressão multivariada não paramétrica proposta por Friedman (1991). O termo “MARS” é uma marca registrada da empresa *Salford Systems*. Por esse motivo, tal termo não é utilizado por *softwares* de análise, sendo substituído por outros termos, como é o caso do *software* R, que utiliza “*earth*” para se referir a tal técnica de aprendizado.

De acordo com Hastie, Tibshirani e Friedman (2009), MARS pode ser caracterizado como um procedimento adaptativo para regressão, sendo visto como uma generalização da regressão linear. Tal método é apropriado para problemas de alta dimensão, ou seja, aqueles que apresentam grande número de entradas.

O princípio de funcionamento do MARS consiste na partição dos dados de treinamento e posterior modelagem utilizando funções lineares. O modelo criado é constituído por uma série de retas, onde cada uma destas possui uma inclinação. Desse modo, o algoritmo é capaz de contornar a não linearidade dos dados e maximizar a precisão preditiva. A variável resposta é então estimada a partir da soma das funções de base ajustadas a cada uma das retas, ou segmentos, obtidas a partir do particionamento recursivo dos dados (FRIEDMAN, 1991).

Na Figura 1.4 são indicados exemplos da aplicação e funcionamento do método MARS. Os pontos de corte, que separam dois segmentos (*splines*) são denominados “nós”. A posição de cada “nó” é definida pelo algoritmo, que avalia todos os “nós” potenciais nos dados avaliados e, em seguida, reduz o número destes, baseado em uma medida de erro.

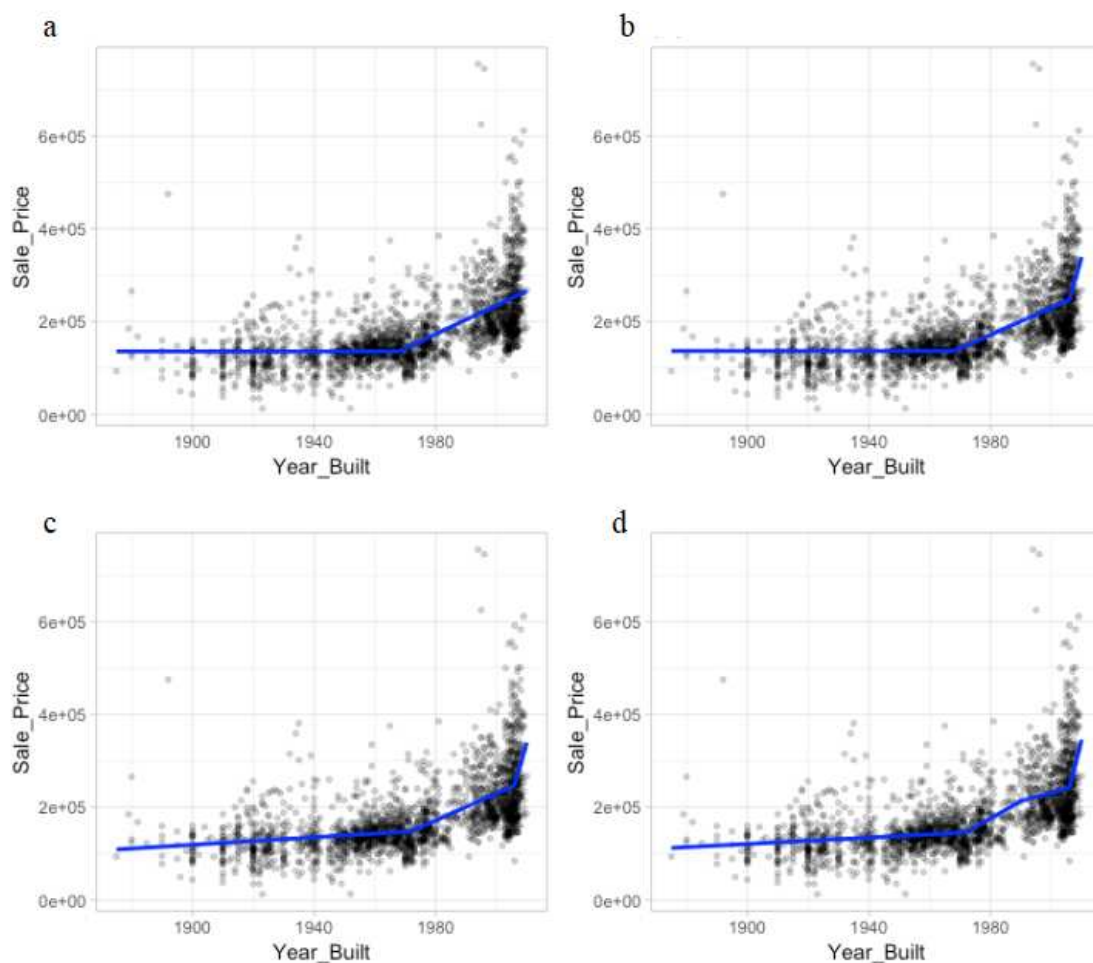


Figura 1.4. Exemplos utilização do algoritmo MARS em funções de um (a), dois (b), três (c) e quatro (d) nós.

Fonte: University of Cincinnati (2023).

À medida que mais “nós” são adicionados, melhor se torna o ajuste do modelo aos dados. No entanto, o ajuste excessivo pode resultar em uma dificuldade do modelo em generalizar respostas ao ser exposto a dados novos e não vistos na etapa de treinamento. Visando reduzir o erro de generalização, o algoritmo MARS realiza um processo denominado “poda”, que visa reduzir o número de “nós” em um determinado modelo. Uma vez criado o conjunto completo de “nós”, o algoritmo realiza uma remoção sequencial dos “nós” que não contribuem significativamente para a precisão preditiva. Como consequência da redução dos nós, as funções básicas que apresentam menor contribuição para o ajuste do modelo são removidas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MURPHY, 2012). As técnicas MARS possuem aplicações em diversas áreas e podem ser utilizadas em problemas de classificação e de regressão, sendo capazes de gerar modelos com alto desempenho para variáveis-resposta binárias (FRIEDMAN, 1991).

K-nearest Neighbor (KNN)

O método *K-nearest Neighbor* (k-vizinhos mais próximos) é um dos métodos de aprendizado de máquina mais simples e populares (MURPHY, 2012). KNN não possui premissas matemáticas e apresenta baixo custo computacional, o que contribuiu para a sua popularidade na comunidade de aprendizado de máquina (GRUS, 2016).

O princípio de funcionamento do método KNN se baseia em estimar o rótulo, ou variável resposta, de um determinado elemento a partir de uma análise de seus k vizinhos mais próximos (KUBAT, 2017). Tais vizinhos são os elementos que apresentam maior semelhança a um dado elemento, sendo esta semelhança relativa às informações contidas nas covariáveis disponíveis no banco de dados. A identificação dos k vizinhos mais próximos pode ser realizada a partir do cálculo de métricas de distância ou similaridade entre pontos, tal como a distância euclidiana. A definição do número de vizinhos a ser considerada nos cálculos é, em geral, realizada pelo usuário (KUBAT, 2017; MURPHY, 2012).

KNN pode ser utilizado em problemas de classificação e de regressão. Para os primeiros, após a identificação dos k vizinhos mais próximos ao elemento avaliado, a classificação é realizada a partir da contabilização da classe majoritária indicada pelos k vizinhos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Por outro lado, em problemas de regressão, o valor da variável resposta de um elemento é estimado a partir do cálculo da média local das respostas dos k vizinhos mais próximos ao elemento no espaço das covariáveis (MURPHY, 2012).

Support Vector Machine (SVM)

Support Vector Machine (máquina de vetor de suporte, em tradução livre) é um algoritmo de AM supervisionado desenvolvido por Gunn (1998). SVM pode ser utilizado em problemas de classificação e regressão.

O princípio da técnica consiste em identificar um hiperplano que “melhor” separe as classes desejadas em um conjunto de dados de treinamento. Para isso, inicialmente, cada item do banco de dados é plotado como um ponto no espaço n-dimensional, sendo n o número de recursos ou covariáveis. Em seguida, o classificador é executado para identificação de um hiperplano que maximize a distância entre pontos de classes distintas (GRUS, 2016). A técnica SVM é ilustrada na Figura 1.5. A distância entre pontos de classes distintas é denominada “margem” e os vetores de suporte são os pontos, ou coordenadas, de cada observação no espaço. A linha sólida indicada na Figura 1.5b indica o hiperplano que maximiza a separação entre

classes. A identificação do hiperplano ideal é realizada por meio de um processo de otimização, que visa identificar a maior margem de separação das classes (KUBAT, 2017).

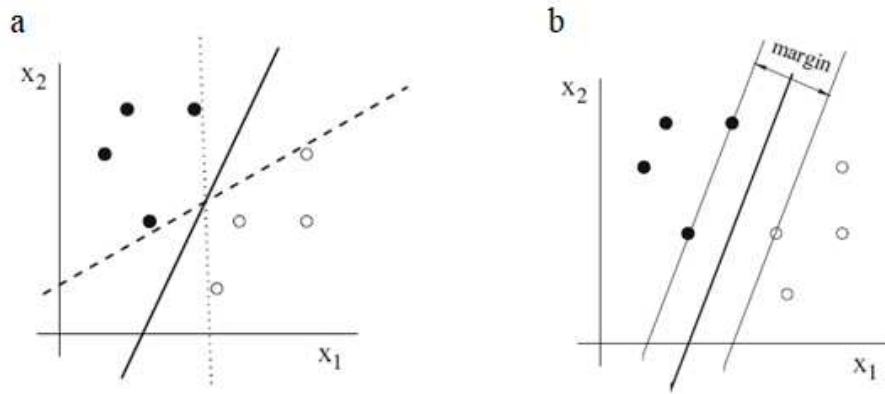


Figura 1.5. Princípio básico de funcionamento da técnica *Support Vector Machine*. Em a) são indicadas duas classes de observações e três hiperplanos possíveis para separação e em b) é indicado o hiperplano que maximiza a separação entre classes.

Fonte: Kubat (2017).

Existem diversas situações em que, devido à natureza dos dados, não é possível estabelecer um limite linear capaz de separar duas classes. A solução adotada para problemas como esse consiste em transformar o conjunto de dados de formato original (não linear) para um espaço linear de maior dimensão. Tal procedimento é denominado truque do kernel. Após a utilização de uma função não linear do kernel, as observações passam a ser linearmente separáveis, o que permite a utilização dos métodos anteriormente citados (SVM linear) (KUBAT, 2017; SHAW-TAYLOR; CRISTIANINI, 2004).

Regressão logística (RL)

Apesar de possuir a palavra “regressão” em seu nome, a regressão logística é uma técnica de aprendizado supervisionado utilizada em problemas de classificação. A referida nomenclatura está associada ao fato dos algoritmos de regressão logística construírem modelos a partir de uma transformação, denominada função logística, sobre a regressão linear. Desse modo, as funções possuem formato em S, denominado sigmoide. O objetivo da utilização de algoritmos de regressão logística é prever rótulos categóricos de elementos, a partir do conhecimento do comportamento de um conjunto de covariáveis (MURPHY, 2012).

Regressão logística é considerada um dos algoritmos de AM supervisionado mais simples para classificação em problemas com valores resposta binários, tais como 0/1 e sim/não. De modo geral, a regressão logística prevê a probabilidade de ocorrência de um evento e, para isso, realiza ajustes dos dados a uma função logit. Em função disso, tal modelo é, por vezes, denominado regressão logit. A partir do uso de algoritmos de regressão logística é possível identificar quais covariáveis influenciam nos resultados da variável resposta e como se dá essa influência. Diferente de outros algoritmos de classificação, que apresentam como resposta um valor binário para a classe (0 ou 1), a regressão logística retorna valores entre 0 e 1, que indicam a probabilidade de um evento ocorrer (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MURPHY, 2012).

Modelos aplicados na predição de arboviroses

A partir das informações apresentadas, é possível verificar que os modelos de AM apresentam diferenças conceituais que fazem com que um banco de dados possa ser gerenciado de diferentes maneiras. No campo de predição de arboviroses, autores têm avaliado a capacidade de modelos distintos de realizar predições assertivas (HOYOS; AGUILAR; TORO, 2021).

De modo geral, grande parte dos estudos na referida temática estão voltados para a modelagem e a previsão da população de mosquitos vetores de arboviroses (JOSHI; MILLER, 2021). No entanto, Colón-González *et al.* (2017) salientam que apenas a presença de vetores em uma área não é suficiente para inferir o aumento na carga de doenças. A expansão do risco de transmissão depende, ainda, da introdução de patógenos, padrões de uso da terra, urbanização e outros fatores que mediam a distribuição dos vetores e a ocorrência de contato vetor-humano (RYAN *et al.*, 2019).

Visando obter predições mais assertivas, outros estudos avaliaram a distribuição espacial de arboviroses baseada na correlação entre as localidades de ocorrência e as condições ambientais das localidades. Tal abordagem utiliza informações de ocorrência geográfica de doenças disponíveis em bancos de dados, que são então utilizadas como variáveis resposta (TJADEN, NILS BENJAMIN *et al.*, 2018). Exemplos de estudos baseados na metodologia citada são indicados na Tabela 1.1.

É possível verificar que grande parte dos estudos possui abrangência nacional ou municipal para a predição de arboviroses. As covariáveis utilizadas em tais estudos incluem, sobretudo, condições climáticas, índices de vegetação e mobilidade humana. Além disso,

verifica-se que a arbovirose de maior destaque em tais estudos é a dengue. De acordo com Rocha e Giesbrecht (2022), há uma carência de estudos voltados para predição de Zika, o que pode estar relacionado à escassez de dados, uma vez que, as epidemias associadas ao vírus ainda podem ser consideradas recentes.

Tabela 1.1. Listagem de estudos que utilizaram aprendizado de máquina para predição de casos de arboviroses.

Arbovirose	Local	Modelo ¹	Referência
Dengue, Chikungunya Zika	Recife (Brasil)	SVM, RF, MLP, ELM, LR e ESM	(SILVA <i>et al.</i> , 2022)
Dengue	São Luís (Brasil)	LR, NB, DT, RF	(ROCHA; GIESBRECHT, 2022)
Dengue	Colômbia	RF e ANN	(ZHAO <i>et al.</i> , 2020)
Dengue	Cingapura	RF	(ONG <i>et al.</i> , 2018b)
Dengue	Guangdong (China)	RF, SVR, GBM, NMB e LASSO	(GUO <i>et al.</i> , 2017)
Dengue	Costa Rica	RF	(BARBOZA <i>et al.</i> , 2023)
Dengue	Malásia	BN, SVM, RBF e Decision table	(NEJAD;VARATHAN, 2021)

Nota: ¹ *Support Vector Machine* (SVM); *Random Forest* (RF); Perceptron Multicamada (MLP); Máquina de Aprendizagem Extrema (ELM); Regressão Linear (LR); Máquina de Estado de Eco (ESM); *Naive Bayes* (NB); Árvore de decisão (DT); redes neurais artificiais (ANN); regressão de vetor de suporte (SVR), modelo de regressão binomial negativa (NBM), Least absolute shrinkage and selection operator (LASSO); *Gradient Boosting Machine* (gbm); Bayes network (BN); modelo de função de base radial (RBF).

1.2.2.2 Medidas de avaliação de modelos de aprendizado de máquina

De acordo com Izbicki e Santos (2020), diversas métricas podem ser utilizadas para a avaliação do desempenho de algoritmos de aprendizado de máquina em problemas de classificação. Uma forma comum consiste na construção de matrizes de confusão e posterior cálculo de métricas como: acurácia, acurácia balanceada, sensibilidade, especificidade, f1-score e índice Kappa. Todos os índices citados são adimensionais e apresentam valores no intervalo de 0 a 1, de modo que, valores próximos a 1 indicam uma maior concordância entre classes originais e as classes preditas pelos modelos de AM.

Uma matriz de confusão tem como objetivo indicar os valores reais de uma variável resposta e os valores preditos para essas variáveis pelos modelos de AM avaliados. Na Tabela 1.2 é apresentado um tipo de matriz de confusão. É possível observar que a matriz, basicamente,

relaciona as predições das classes “sim” (positiva) e “não” (negativa). As linhas dessa matriz representam as classes verdadeiras (originais) e as colunas as predições realizadas pelos modelos de AM. A partir dessas informações, torna-se possível calcular diversas métricas que possibilitam avaliar o desempenho dos modelos e identificar aqueles que melhor se ajustam aos dados (IZBICKI; SANTOS, 2020).

Tabela 1.2. Matriz de confusão para um problema de classificação binária.

Matriz de confusão		Classe predita (f(x))	
		Positiva	Negativa
Classe original (y)	Positiva	Verdadeiro positivo (VP)	Falso negativo (FN)
	Negativa	Falso positivo (FP)	Verdadeiro negativo (VN)

A acurácia calcula o desempenho do modelo, ou seja, indica quantas observações o modelo classificou corretamente. O cálculo de tal métrica é realizado conforme indicado na Equação 1.1.

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FN} + \text{FP}} \quad \text{Equação 1.1}$$

No entanto, em situações onde o banco de dados avaliado apresenta uma distribuição desbalanceada da variável resposta, a utilização da acurácia pode não ser apropriada. Isso ocorre devido ao cálculo de tal métrica basear-se na identificação de todos os acertos realizados pelo modelo (VP + VN), independente da classe da variável resposta (IZBICKI; SANTOS, 2020). Em situações como a descrita, a utilização da acurácia balanceada pode ser mais apropriada. Na acurácia balanceada, os cálculos são realizados considerando a frequência de acertos em cada classe, o que faz com este índice seja menos influenciado pelo desbalanceamento dos dados e, conseqüentemente, forneça resultados mais confiáveis (BRODERSEN *et al.*, 2010). A obtenção da referida métrica é feita a partir da Equação 1.2.

$$\text{Acurácia balanceada} = \frac{1}{2} * \left(\frac{\text{VP}}{\text{VP} + \text{FN}} + \frac{\text{VN}}{\text{VN} + \text{FP}} \right) \quad \text{Equação 1.2}$$

Outras métricas, tais como sensibilidade e especificidade, podem ser utilizadas para avaliar a capacidade do modelo de reconhecer uma determinada classe. A primeira mensura a taxa de classes positivas corretamente identificadas pelo modelo. A especificidade, por sua vez, calcula a taxa de classes negativas que foram identificadas corretamente. O cálculo para

obtenção da sensibilidade e da especificidade é indicado na Equação 1.3 e Equação 1.4, respectivamente.

$$\text{Sensibilidade} = \frac{VP}{VP+FN} \quad \text{Equação 1.3}$$

$$\text{Especificidade} = \frac{VN}{VN+FP} \quad \text{Equação 1.4}$$

A métrica F1-score combina a precisão e a sensibilidade de um classificador em uma única métrica. Tal métrica é, comumente, utilizada para comparação do desempenho de dois ou mais modelos de classificação. O F1-score pode ser obtido por meio da utilização da Equação 1.5.

$$\text{F1-score} = \frac{VP}{VP + 1/2 *(FP + FN)} \quad \text{Equação 1.5}$$

Além destas métricas citadas, a avaliação do desempenho de um modelo de classificação também pode ser realizada utilizando-se o índice Kappa. Tal índice avalia o grau de concordância entre as classes originais e previstas pelo modelo. O índice Kappa pode ser calculado por meio da Equação 1.6.

$$\text{Kappa} = \frac{Po - Pe}{1 - Pe} \quad \text{Equação 1.6}$$

Onde,

$$Po = VP + VN;$$

$$Pe = (VP + FN) / (VP + FP) + (FN + VN) / (FP + VN).$$

1.2.2 Aprendizado não supervisionado

O segundo principal tipo de aprendizado de máquina é definido como aprendizado não supervisionado. Nesse tipo de aprendizado não há uma definição prévia de classes que identifiquem o padrão dos dados, ou seja, os dados avaliados não possuem rótulos. Desse modo, o objetivo desses métodos é encontrar padrões naturais dos dados, o que é, por vezes, definido como descoberta de conhecimento (IZBICKI; SANTOS, 2020).

Devido à ausência de rótulos, torna-se necessário “aprender mais” sobre a estrutura dos dados. Comparados aos problemas de aprendizado supervisionado, os problemas de aprendizado não supervisionado são menos bem definidos, tendo em vista o desconhecimento do tipo de padrão procurado e a pouca disponibilidade de métricas de erro para avaliação de desempenho (MURPHY, 2012).

Kubat (2017) cita que a tarefa mais frequentemente utilizada para análise de dados não rotulados é a análise de agrupamentos (*clusters*). Tais análises consistem na divisão de amostras em grupos a partir da identificação de semelhanças entre estas. A divisão é realizada de modo que os grupos formados sejam diferentes entre si, porém, que os elementos pertencentes ao mesmo grupo sejam semelhantes (IZBICKI; SANTOS, 2020). Segundo Hair, Black e Sant’anna (2009) em função do método de particionamento de dados utilizado, os algoritmos de agrupamento podem ser subdivididos em hierárquicos e não hierárquicos.

1.2.1.1 Métodos de agrupamento hierárquicos e não hierárquicos

De acordo com Härdle e Simar (2015), existem, essencialmente, dois tipos de métodos de análise de agrupamentos: hierárquicos e não hierárquicos, também denominados de procedimentos de particionamento.

Métodos hierárquicos criam uma hierarquia de relacionamentos entre os elementos e geralmente são apresentados graficamente usando um dendrograma. Em função das metodologias adotadas para a construção de agrupamentos, tais métodos podem ser divididos em procedimentos aglomerativos e divisivos (HAIR; BLACK; SANT’ANNA, 2009). O primeiro tipo de agrupamento hierárquico assume que cada observação em uma matriz de similaridade ou distância é um *cluster* individual, desse modo, inicialmente, o número de agrupamentos é igual ao de observações. Nas etapas sucessivas do procedimento, os *clusters* são combinados por meio de algum algoritmo e esse processo continua até que todas as observações tenham sido combinadas para formar um *cluster* (TINSLEY; BROWN, 2000). Os algoritmos aglomerativos mais, frequentemente, utilizados são: *single linkage*, *complete linkage*, *average linkage*, *centroid method* e *Ward's method* (HAIR; BLACK; SANT’ANNA, 2009). Por outro lado, em procedimentos divisivos, todas as observações iniciam com um único *cluster* e são sucessivamente divididas, até que cada observação seja um agrupamento unitário. Segundo Manly e Alberto (2019), métodos hierárquicos divisivos são utilizados com menor frequência do que os aglomerativos.

Hair, Black e Sant'anna (2009) citam que os métodos hierárquicos foram os primeiros métodos de agrupamento desenvolvidos, desse modo, são amplamente utilizados. Os autores destacam que tais métodos oferecem diversas vantagens, tais como: simplicidade para análise e interpretação, possibilidade de serem aplicados a quase todo tipo de questão de pesquisa e rapidez para gerar o conjunto de soluções de agrupamento, o que permite a investigação de uma vasta gama de soluções alternativas, variando medidas de similaridade e métodos de ligação de uma maneira eficiente. Contudo, os mesmos autores destacam que algumas desvantagens associadas ao uso de métodos hierárquicos podem afetar as soluções de agrupamento encontradas. Dentre as desvantagens citadas pelos autores, destacam-se: formação de combinações iniciais indesejáveis que podem persistir na análise e conduzir a resultados artificiais, impacto substancial de observações atípicas e limitação para análise de amostras muito grandes. De modo geral, embora o método hierárquico de agrupamento facilite uma avaliação abrangente de uma vasta gama de soluções, tais soluções podem ser afetadas por uma característica do método: uma vez que os agrupamentos são unidos, eles não são separados no processo.

Diferente dos métodos hierárquicos, métodos não hierárquicos não envolvem o processo de construção de um dendrograma, em vez disso, designam objetos a um número específico de *clusters* (TINSLEY; BROWN, 2000). De acordo com Hair, Black e Sant'anna (2009) e Tinsley e Brown (2000), o processo de agrupamento em métodos não hierárquicos, essencialmente, possui os seguintes passos:

1. Especificar sementes de agrupamento: Identificar os centros de cada agrupamento, denominados de sementes. Uma semente de agrupamento pode ser pré-especificada ou selecionada de modo aleatório;
2. Designação: Com as sementes de agrupamento definidas, cada observação é designada a uma das sementes de agrupamento, com base em similaridade. O objetivo desse passo é designar cada observação à semente mais próxima;
3. Recalcular o centro de cada agrupamento e repetir o processo até que nenhuma atribuição ou reatribuição de *cluster* ocorra entre as observações ou até que um número predeterminado de iterações seja alcançado.

Segundo Tinsley e Brown (2000), os métodos não hierárquicos, também frequentemente referidos como particionamento *k-means*, foram desenvolvidos, em parte, em resposta a uma das principais deficiências dos métodos hierárquicos: uma vez que grupos são encontrados e observações são atribuídas e esses grupos, essa atribuição não pode ser alterada em algum estágio subsequente do processo. De acordo com Härdle e Simar (2015), a capacidade dos

métodos não hierárquicos de redesignar observações a outros grupos, durante a aplicação do algoritmo, é a principal característica que diferencia os dois métodos de agrupamento. Tal procedimento tem como objetivo minimizar a variabilidade dentro do agrupamento e, ao mesmo tempo, maximizar a variabilidade entre estes.

Hair, Black e Sant'anna (2009) citam que os métodos não hierárquicos possuem vantagens em relação aos métodos hierárquicos, tais como: os resultados são menos suscetíveis às observações atípicas nos dados, à medida de distância usada e à inclusão de variáveis irrelevantes; métodos não-hierárquicos podem analisar conjuntos com muitos dados, uma vez que, não demandam o cálculo de matrizes de similaridade entre todas as observações, mas apenas a similaridade de cada observação com os centroides de agrupamento.

Apesar das vantagens citadas, Hair, Black e Sant'anna (2009) e Tinsley e Brown (2000) ressaltam que os benefícios dos métodos não-hierárquico são evidenciados somente quando pontos sementes não-aleatórios (ou seja, especificados) são empregados. Os resultados obtidos a partir da utilização de métodos não hierárquicos com pontos sementes aleatórios são, geralmente, considerados inferiores aos obtidos em métodos hierárquicos. Desse modo, a aplicação de métodos não hierárquicos depende da seleção de pontos sementes apropriados, seguindo alguma base prática, objetiva ou teórica.

Hair, Black e Sant'anna (2009) afirmam que a combinação de ambos os métodos (hierárquicos e não hierárquicos) permite extrair os benefícios e minimizar as desvantagens associadas a cada abordagem. A utilização conjunta dos métodos consiste em duas etapas: 1) um método hierárquico é utilizado para selecionar o número de agrupamentos e caracterizar os centros de agrupamentos que servirão como sementes iniciais no procedimento não-hierárquico; 2) um método não-hierárquico agrega as observações, utilizando os centros de grupos dos resultados hierárquicos como os pontos sementes iniciais. Tal procedimento, que visa a obtenção de alocações mais precisas, permite que as vantagens dos métodos hierárquicos sejam complementadas pela habilidade dos métodos não hierárquicos de refinar os resultados.

1. 3 Estatística espacial

As três principais áreas reconhecidas na estatística espacial utilizam representações de padrão de pontos, superfície contínua (geoestatística) e dados de área, cada qual dotado de características específicas que demandam a utilização de abordagens metodológicas distintas (BIVAND, 2008). Segundo Bivand, Pebesma e Gómez-Rubio e Diggle (2013), um processo pontual pode ser definido como um processo estocástico que gera um conjunto contável de

eventos dentro de uma região delimitada. A localização dos eventos gerados por um processo pontual em uma área de estudo é denominada padrão de pontos.

Druck (2004) cita que o termo evento se refere a qualquer tipo de fenômeno de interesse localizável no espaço, que possa estar associado a uma representação pontual. Tais eventos possuem localização espacial conhecida por meio de coordenadas. Os dados na forma de um conjunto de pontos podem surgir em contextos distintos. Alguns exemplos incluem a localização de ocorrências criminais, espécies vegetais e residências onde casos de doenças foram observados.

Contudo, em diversas situações, em função da necessidade de confidencialidade e privacidade dos registros de dados, não é possível obter a localização exata dos eventos. Diante disso, os conjuntos de dados espaciais têm sido frequentemente reportados em níveis espaciais agregados, especificamente sob a forma de dados de área (OYANA; MARGAI, 2016).

Comumente, as informações espaciais em nível individual são agregadas por municípios, bairros, setores censitários, áreas de serviços de saúde, entre outras escalas. Desse modo, a localização exata dos eventos não é disponibilizada, mas somente um valor por área. Os exemplos mais comuns de dados de área referem-se a contagens, tais como contagens de crimes por região, doenças por municípios, número de pessoas com uma determinada faixa de renda por setor censitário; proporções, como percentual de adultos desempregados; e médias, como renda média do chefe da família por setor censitário (DRUCK, 2004).

Na análise espacial de dados de área são avaliados eventos agregados em áreas delimitadas por polígonos. Druck *et al.* (2004) citam que, usualmente, os dados agregados por áreas são apresentados por meio de mapas temáticos, que contém o padrão espacial do fenômeno analisado. Em tais mapas, a variação de um atributo no espaço não é representada por uma superfície contínua e pode apresentar variações abruptas de valor. Segundo os mesmos autores, para a análise espacial de dados agregados por áreas, o primeiro passo consiste na escolha do modelo inferencial a ser utilizado. A hipótese mais comum é supor que as áreas são diferenciadas, o que implica que cada área avaliada apresenta uma distribuição de probabilidade diferente das demais, o intitulado modelo espacial discreto.

Ao se analisar dados espaciais de áreas busca-se realizar uma descrição de um fenômeno ocorrido no espaço, investigar a existência de um padrão espacial (agrupado, disperso ou aleatório) e, caso o padrão não seja aleatório, explorar os processos subjacentes ao padrão observado (OYANA; MARGAI, 2016).

De acordo com Brasil (2007), a mensuração da dependência espacial entre observações é uma etapa essencial na análise espacial de dados de área. Tal etapa visa estimar o quanto um

determinado elemento está relacionado com seus vizinhos, considerando-se uma variável de interesse. Para a avaliação da dependência espacial em dados de área é necessário construir uma matriz de vizinhança, também denominada matriz de proximidade. Tal matriz indica a relação espacial de cada área com as demais e pode ser composta pela lista de vizinhos de cada área (baseada na contiguidade), pela distância entre áreas ou, ainda, pela conectividade ponderada pelo comprimento da fronteira comum das áreas.

A estrutura da dependência espacial entre os valores observados nas áreas em estudo é analisada pela função de autocorrelação espacial. A autocorrelação espacial estima o quanto o valor observado de uma variável de interesse em uma região é dependente dos valores desta mesma variável nas localizações vizinhas (ALMEIDA, E., 2012; BRASIL, 2007). Para estimar a magnitude da autocorrelação espacial entre áreas podem ser utilizados indicadores globais, que apresentam uma única medida de tendência espacial para toda a região em estudo. Tais indicadores, que possuem caráter exploratório, permitem efetuar testes de hipótese, cuja hipótese nula é que não há dependência espacial entre as áreas avaliadas, considerando-se a variável em estudo.

Em situações em que um grande número de polígonos é analisado, é presumível que diferentes regimes de autocorrelação espacial ocorram na região em estudo. Nessas circunstâncias, avaliar somente os índices globais pode não ser suficiente para o entendimento da estrutura espacial dos dados avaliados. Para evidenciar estes regimes espaciais, torna-se relevante apresentar também uma medida de autocorrelação espacial para cada polígono. Em geral, tais estatísticas locais derivam de índices de dependência global (BRASIL, 2007).

1. 4 Referências

ALBAHRA, S. *et al.* Artificial Intelligence and Machine Learning Overview in Pathology & Laboratory Medicine: A General Review of Data Preprocessing and Basic Supervised Concepts. **Seminars in Diagnostic Pathology**, 2023.

ALMEIDA, E. **Econometria espacial aplicada**. Campinas: Alínea, 2012.

BARBAZAN, P. *et al.* Modelling the effect of temperature on transmission of dengue. **Medical and Veterinary Entomology**, v. 24, p. 66–73, 2010.

BARBOZA, L. A. *et al.* Assessing dengue fever risk in Costa Rica by using climate variables and machine learning techniques. **PLOS Neglected Tropical Diseases**, v. 17, n. 1, p. e0011047, 2023.

BARRERA, R. *et al.* Unusual productivity of *Aedes aegypti* in septic tanks and its implications for dengue control. **Medical and Veterinary Entomology**, v. 22, n. 1, p. 62–69,

2008.

BESERRA, E. B. *et al.* Efeito da qualidade da água no ciclo de vida e na atração para oviposição de *Aedes aegypti* (L.) (Diptera: Culicidae). **Neotropical Entomology**, v. 39, n. 6, p. 1016–1023, 2010.

BIVAND, R. Implementing representations of space in economic geography*. **Journal of Regional Science**, v. 48, n. 1, p. 1–27, 2008.

BIVAND, R. S.; PEBESMA, E. J.; GÓMEZ-RUBIO, V. **Applied Spatial Data Analysis with R**. New York: Springer, 2008.

BRASIL. **Introdução à Estatística Espacial para a Saúde Pública**. Brasília: Ministério da Saúde, 2007.

BRASIL.. **Manual de vigilância sentinela de doenças neuroinvasivas por arbovírus**. Brasília: Ministério da Saúde, 2017.

BRASIL.. **Guia de Vigilância em Saúde : volume único**. 3. ed. Brasília: Ministério da Saúde, 2019.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. ST4_Method_Random_Forest. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BRODERSEN, K. H. *et al.* The balanced accuracy and its posterior distribution. **Proceedings - International Conference on Pattern Recognition**, p. 3121–3124, 2010.

CHAGAS, C. DA S. *et al.* Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. **Catena**, v. 139, p. 232–240, 2016.

CHITOLINA, R. F. *et al.* Raw sewage as breeding site to *Aedes* (*Stegomyia*) *aegypti* (Diptera, culicidae). **Acta Tropica**, v. 164, p. 290–296, 2016.

CHOUIN-CARNEIRO, T. *et al.* Differential Susceptibilities of *Aedes aegypti* and *Aedes albopictus* from the Americas to Zika Virus. **PLOS Neglected Tropical Diseases**, v. 10, n. 3, p. e0004543, 2016.

CHOUIN-CARNEIRO, T. *et al.* Zika virus transmission by Brazilian *Aedes aegypti* and *Aedes albopictus* is virus dose and temperature-dependent. **PLOS Neglected Tropical Diseases**, v. 14, n. 9, p. e0008527, 2020.

COLÓN-GONZÁLEZ, F. J. *et al.* After the epidemic: Zika virus projections for Latin America and the Caribbean. **PLoS Neglected Tropical Diseases**, v. 11, n. 11, p. 1–19, 2017.

CUTLER, D. R. *et al.* Random forests for classification in ecology. **Ecology**, v. 88, n. 11, p. 2783–2792, 2007.

DADA, N. *et al.* Relationship between *Aedes aegypti* production and occurrence of *Escherichia coli* in domestic water storage containers in rural and sub-urban villages in Thailand and Laos. **Acta Tropica**, v. 126, n. 3, p. 177–185, 2013.

- DIGGLE, P. J. **Statistical analysis of spatial and spatio-temporal point patterns**. Boca Raton: CRC press, 2013.
- DOĞAN, A.; KORKMAZ, M.; KIRMACI, V. Estimation of ranque-hilsch vortex tube performance by machine learning techniques. **International Journal of Refrigeration**, 2023.
- DRUCK, S. **Análise Espacial de Dados Geográficos**. Brasília: Embrapa, 2004.
- FREITAS, P. S. S. *et al.* Síndrome congênita do vírus Zika: perfil sociodemográfico das mães. **Revista Panamericana de Salud Pública**, v. 43, p. 1, 2019.
- FREUND, Y.; SCHAPIRE, R. E. A Short Introduction to Boosting. **Journal of Japanese Society for Artificial Intelligence**, v. 14, n. 5, p. 771–780, 1999.
- FRIEDMAN, J. H. Multivariate Adaptive Regression Splines. **The Annals of Statistics**, v. 19, n. 1, p. 1–67, 1991.
- GAGE, K. L. *et al.* Climate and Vectorborne Diseases. **American Journal of Preventive Medicine**, v. 35, n. 5, p. 436–450, 2008.
- GARCIA, L. P. **Epidemia do vírus Zika e microcefalia no Brasil: emergência, evolução e enfrentamento**. Rio de Janeiro: Ipea, 2018.
- GOMES, H.; JESUS, A. G. DE; QUARESMA, J. A. S. Identification of risk areas for arboviruses transmitted by *Aedes aegypti* in northern Brazil: A One Health analysis. **One Health**, p. 100499, 2023.
- GRIMM, R. *et al.* Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. **Geoderma**, v. 146, n. 1–2, p. 102–113, 2008.
- GRUS, J. **Data Science do Zero**. Rio de Janeiro: Alta Books, 2016.
- GUNN, S. R. Support Vector Machines for Classification and Regression. **ISIS technical report**, v. 14, n. 1, p. 5–16, 1998.
- GUO, P. *et al.* Developing a dengue forecast model using machine learning: A case study in China. **PLOS Neglected Tropical Diseases**, v. 11, n. 10, p. e0005973, 2017.
- HAGGERTY, R. *et al.* Application of machine learning in groundwater quality modeling - A comprehensive review. **Water Research**, v. 233, p. 119745, 2023.
- HAI, N. A. *et al.* A study on Adaptation of *Aedes Aegypti* Mosquito Larvae in Sewage, Boring and Sea Water. **Proceedings of 18th International Bhurban Conference on Applied Sciences and Technologies, IBCAST 2021**, p. 481–485, 2021.
- HAIR, J. J. F.; BLACK, W. C.; SANT'ANNA, A. S. **Análise multivariada de dados** (6a. ed.) 2009.
- HAN, M. *et al.* Novel integrated systems for controlling and prevention of mosquito-borne diseases caused by poor sanitation and improper water management. **Journal of Environmental Chemical Engineering**, v. 4, n. 4, p. 3718–3723, 2016.

HÄRDLE, W. K.; SIMAR, L. **Applied multivariate statistical analysis**. 4 ed, Springer Nature, 2019.

HARTSHORN, S. **Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners**. p. 74, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **Linear Methods for Classification**. Vol. 2. New York: Springer, 2009.

HILGENFELD, R.; VASUDEVAN, S. G. **Dengue and Zika : Control and Antiviral Treatment Strategies**. Singapura: Springer Nature, 2018.

HOYOS, W.; AGUILAR, J.; TORO, M. Dengue models based on machine learning techniques: A systematic literature review. **Artificial Intelligence in Medicine**, v. 119, p. 102157, 2021.

HUANG, Y. J. S.; HIGGS, S.; VANLANDINGHAM, D. L. Emergence and re-emergence of mosquito-borne arboviruses. **Current Opinion in Virology**, v. 34, p. 104–109, 2019.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina : uma abordagem estatística**. 1. ed. São Carlos: Rafael Izbicki, 2020.

JOSHI, A.; MILLER, C. Review of machine learning techniques for mosquito control in urban environments. **Ecological Informatics**, v. 61, p. 101241, 2021.

KUBAT, M. **An Introduction to Machine Learning**. Cham, Switzerland: Springer International Publishing, 2017.

LABOISSIERE, L. A.; FERNANDES, R. A. S.; LAGE, G. G. Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. **Applied Soft Computing**, v. 35, p. 66–74, 2015.

LARRANDABURU, M. *et al.* Zika virus infection and congenital anomalies in the Americas: opportunities for regional action. **Rev Panam Salud Publica**, v. 41, p. 1–8, 2017.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002.

LIMA, L. R.; SOARES, L. O. L.; SILVA, P. S. P. F. **Arboviroses**. Teresina: SCISAUDE, 2022.

LOWY, I. **Zika no Brasil: história recente de uma epidemia**. Rio de Janeiro: Editora Fiocruz, 2019.

MACKAY, A. J. *et al.* Dynamics of *Aedes aegypti* and *Culex quinquefasciatus* in Septic Tanks. **Journal of the American Mosquito Control Association**, v. 25, n. 4, p. 409–416, 2009.

MESHARAM, VISHAL *et al.* Machine learning in agriculture domain: A state-of-art survey. **Artificial Intelligence in the Life Sciences**, v. 1, p. 100010, 2021.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. London: The MIT Press,

2012.

NGAI, E. W. T.; WU, Y. Machine learning in marketing: A literature review, conceptual framework, and research agenda. **Journal of Business Research**, v. 145, p. 35–48, 2022.

OKAFOR, C. E. *et al.* Advances in machine learning-aided design of reinforced polymer composite and hybrid material systems. **Hybrid Advances**, v. 2, p. 100026, 2023.

ONG, J. *et al.* Mapping dengue risk in Singapore using Random Forest. **PLOS Neglected Tropical Diseases**, v. 12, n. 6, p. e0006587, 2018.

OPAS, O. P.-A. DA S. **Zika - OPAS/OMS | Organização Pan-Americana da Saúde**. Disponível em: <<https://www.paho.org/pt/topicos/zika>>. Acesso em: 27 fev. 2023.

OYANA, T. J.; MARGAI, F. M. **Spatial Analysis: Statistics, Visualization, and Computational Methods**. Boca Raton: CRC press, 2016.

PINTO JUNIOR, V. L. *et al.* Vírus Zika: revisão para clínicos. **Acta Médica Portuguesa**, v. 28, n. 6, p. 760–765, 2015.

ROCHA, F. P.; GIESBRECHT, M. Machine learning algorithms for dengue risk assessment: a case study for São Luís do Maranhão. **Computational and Applied Mathematics**, v. 41, n. 8, p. 1–27, 2022.

RODRIGUEZ-GALIANO, V. F. *et al.* An assessment of the effectiveness of a random forest classifier for land-cover classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 67, n. 1, p. 93–104, 2012.

RYAN, S. J. *et al.* Global expansion and redistribution of Aedes-borne virus transmission risk with climate change. **PLoS Neglected Tropical Diseases**, v. 13, n. 3, p. 1–20, 2019.

SHAWE-TAYLOR, J.; CRISTIANINI, N. **Kernel Methods for Pattern Analysis**. Cambridge University Press, 2004.

SILVA, C. C. *et al.* Spatiotemporal forecasting for dengue, chikungunya fever and Zika using machine learning and artificial expert committees based on meta-heuristics. **Research on Biomedical Engineering**, v. 38, n. 2, p. 499–537, 2022.

SILVA, J. C. B.; MACHADO, C. J. S. Associations between dengue and socio-environmental variables in capitals of the Brazilian northeast by cluster analysis. **Ambiente & Sociedade**, v. 21, p. 1332, 2018.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados - Com Aplicações em R**. 1. ed. Rio de Janeiro: LTC, 2021.

SILVA, L. J.; ANGERAMI, R. N. **Arboviroses no Brasil contemporâneo**. Rio de Janeiro: Fiocruz, 2008.

SOARES, S. R. A.; BERNARDES, R. S.; NETTO, O. DE M. C. Relações entre saneamento, saúde pública e meio ambiente: elementos para formulação de um modelo de planejamento em saneamento. **Cadernos de Saúde Pública**, v. 18, n. 6, p. 1713–1724, 2002.

STAVROPOULOS, G. *et al.* **Random Forest and Ensemble Methods**. 2. ed. Elsevier Inc., 2020.

TINSLEY, H. E. A.; BROWN, S. D. **Multivariate Statistics and Mathematical Modeling**. Academic Press, 2000.

TJADEN, N. B. *et al.* Mosquito-Borne Diseases: Advances in Modelling Climate-Change Impacts. **Trends in Parasitology**, v. 34, n. 3, p. 227–245, 2018.

WIKAN, N.; SMITH, D. R. Zika virus: history of a newly emerging arbovirus. **The Lancet. Infectious diseases**, v. 16, n. 7, p. e119–e126, 2016.

XAVIER, L. L. *et al.* Analysis of climate factors and dengue incidence in the metropolitan region of Rio de Janeiro, Brazil. **PLOS ONE**, v. 16, n. 5, p. e0251403, 2021.

YAVARI NEJAD, F.; VARATHAN, K. D. Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction. **BMC Medical Informatics and Decision Making**, v. 21, n. 1, p. 1–12, 2021.

ZHAO, N. *et al.* Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. **PLOS Neglected Tropical Diseases**, v. 14, n. 9, p. e0008056, 2020.

CAPÍTULO 2

Mudanças climáticas e seu impacto na distribuição de arboviroses: uma revisão sistemática de estudos recentes¹

¹ Capítulo publicado na Revista Brasileira de Geografia Física, v. 14, n. 6, p. 3361-3377, 2021.

2.1 Introdução

Ao longo das duas últimas décadas, a notável expansão da faixa territorial de diversos arbovírus resultou em um aumento expressivo na atividade epidêmica global (HILGENFELD; VASUDEVAN, 2018). De acordo com Maniero *et al.* (2021), os principais arbovírus que apresentam ampla distribuição geográfica estão relacionados a doenças emergentes e reemergentes, sendo eles: Dengue (DEN), Zika (ZIK), Febre Amarela (YF), e Chikungunya (CHIK). Tais doenças podem ser transmitidas por duas espécies de mosquitos do gênero *Aedes*: *Aedes aegypti* e *Aedes albopictus* (HILGENFELD; VASUDEVAN, 2018). Por esse motivo, estudos têm avaliado a distribuição global e os requisitos ecológicos dessas espécies (LETA *et al.*, 2018).

Segundo Lippi *et al.* (2019), a distribuição atual de mosquitos do gênero *Aedes* é orientada em função das características climáticas da região. Contudo, as mudanças climáticas globais possuem o potencial de impulsionar a expansão geográfica desses vetores e, conseqüentemente, alterar a distribuição geográfica das diversas doenças que estes transmitem (LUBINDA *et al.*, 2019; LWANDE *et al.*, 2020). À vista disso, Iwamura *et al.* (2020) alegam que a compreensão do comportamento dos vetores em resposta aos fatores climáticos é central para a caracterização da distribuição espaço-temporal das arboviroses e previsão das mudanças potenciais no risco devido às mudanças climáticas.

El-Sayed e Kamel (2020) afirmam que o aquecimento global influencia diretamente a abundância, a sobrevivência e a distribuição de patógenos e seus vetores. Além dos fatores climáticos, Ordoñez-Sierra *et al.* (2021) citam que a transmissão de arboviroses também pode ser influenciada por fatores demográficos, sociais e entomológicos.

Estudos revelam que fatores climáticos como temperatura, precipitação e umidade podem afetar a população de mosquitos (reprodução, desenvolvimento, distribuição), propagação de vírus (replicação) e a interação vetor-hospedeiro (comportamento de picada de fêmeas adultas de mosquitos do gênero *Aedes*) (GAGE *et al.*, 2008; BARBAZAN *et al.*, 2010).

Temperaturas altas estão associadas ao aumento na reprodução de vetores, comportamento de mordida mais frequente e redução do período de incubação do patógeno no interior dos vetores. Por outro lado, o tempo de sobrevivência do mosquito pode ser reduzido em condições muito quentes e secas (CIOTA e KEYEL, 2019; WHO, 2008).

Adicionalmente, condições de umidade e precipitação locais são apontadas como fatores capazes de afetar a disponibilidade de habitats necessários durante os estágios larvais dos vetores (DICKENS *et al.*, 2018.; LEMOS *et al.*, 2021). O aumento nos índices de precipitação pluviométrica pode elevar a população de vetores quando ocorre o acúmulo de água de chuva em recipientes nas residências (ROBERT *et al.*, 2020). No entanto, condições de seca e escassez hídrica também podem influenciar no aumento da população de vetores, tendo em vista a necessidade de armazenamento doméstico de água nesse período, o que resulta na criação de locais propícios para a reprodução dos mosquitos (AKANDA *et al.*, 2020; DADA *et al.*, 2013).

Um estudo desenvolvido por Ryan *et al.* (2019) indicou que as mudanças climáticas poderão alterar, no próximo século, o padrão de expansão de mosquitos do gênero *Aedes* e de transmissão dos vírus por eles veiculados. Tal projeção é crítica do ponto de vista da saúde humana, uma vez que, a expansão dos arbovírus em regiões sem exposição prévia pode resultar em surtos explosivos de doenças (EQUIHUA *et al.*, 2017).

No entanto, é importante ressaltar que os efeitos do clima na transmissão de arboviroses não são lineares. De acordo com Caldwell *et al.* (2021), embora o clima seja considerado um dos fatores mais relevantes no estudo de processos ecológicos, os efeitos direcionais e dinâmicos do clima nos sistemas são, geralmente, mal compreendidos ou difíceis de prever. Diante disso, Ciota e Keyel, (2019) destacam que as mudanças climáticas poderão aumentar ou até mesmo diminuir o potencial de transmissão de doenças em alguns contextos.

Adicionalmente, Colón-González *et al.* (2018) salientam que apenas a presença de vetores em uma área não é suficiente para inferir o aumento na carga de doenças. A expansão do risco de transmissão depende, também, da introdução de patógenos, padrões de uso da terra, urbanização e outros fatores que mediam a distribuição dos vetores e a ocorrência de contato vetor-humano (RYAN *et al.*, 2019).

Segundo Morgan *et al.* (2021) os fatores socioeconômicos podem contribuir para a perpetuação do risco de doenças causadas por arbovírus. Isso decorre do fato que, em geral, áreas de elevada pobreza possuem abundância de habitats apropriados para mosquitos vetores, uma vez que, em tais locais, comumente, são observadas condições precárias de habitação e saneamento. Desse modo, as previsões de risco de transmissão de doenças baseadas em adequação climática dependem não apenas da incerteza acerca das mudanças climáticas mas,

ainda, da incerteza advinda dos demais fatores ambientais e socioeconômicos que podem influenciar a incidência de doenças (CAUSA *et al.*, 2020; LWANDE *et al.*, 2020; RYAN *et al.*, 2019).

Ainda, de acordo com Souza *et al.* (2021), o sistema de saúde e a infraestrutura existente em um local também irão influenciar a vulnerabilidade ao risco de doenças infecciosas induzido pelas mudanças climáticas. Por conseguinte, países em desenvolvimento, que não possuem recursos para uma efetiva resposta dos sistemas de saúde aos surtos explosivos de doenças, tendem a apresentar risco elevado à saúde causado pelas mudanças climáticas.

Diante dos potenciais impactos das mudanças climáticas no comportamento global da distribuição de arboviroses, diversos estudos, tais como Williams *et al.* (2016), Campbell *et al.* (2015) e Ryan *et al.* (2019), têm realizado projeções considerando diferentes cenários de emissão e seus efeitos na incidência de arboviroses, na distribuição futura de vetores e na transmissão viral. Em função da utilização de diferentes abordagens metodológicas, bem como de cenários de emissão distintos, tais estudos apresentam uma heterogeneidade considerável nas projeções realizadas.

Com o intuito de analisar as diferenças entre as projeções da distribuição da dengue existentes, Xu *et al.* (2020) reuniram estudos com a referida temática. Contudo, a revisão realizada pelos autores não abrangeu estudos que avaliaram a distribuição dos vetores da dengue em função das mudanças climáticas. Para El-Sayed e Kamel (2020), a avaliação da expansão geográfica dos vetores é um passo essencial para o entendimento da distribuição de doenças infecciosas e, por esse motivo, não deve ser subestimada. Uma revisão do futuro global da dengue, em cenários de mudanças climáticas, também foi realizada por Messina *et al.* (2015). No entanto, na referida revisão, estudos em escala regional ou local não foram incluídos.

Diante disso, a fim de descrever, de forma ampla, os efeitos das mudanças climáticas na incidência e distribuição futura de arboviroses, buscou-se nessa revisão avaliar os possíveis efeitos das mudanças climáticas na distribuição de vetores e na transmissão viral, considerando-se estudos em escala global, nacional, regional e local. A hipótese levantada no estudo é que as mudanças climáticas globais possuem o potencial de afetar a incidência de arboviroses, seja provocando o aumento destas ou a diminuição em locais distintos.

2.2 Material e métodos

2.2.1 Estratégia de busca do material bibliográfico e critérios de seleção

A pesquisa teve como foco avaliar evidências científicas a respeito de possíveis impactos das mudanças climáticas na i) distribuição geográfica de vetores de arboviroses, ii) capacidade de transmissão viral e iii) incidência de arboviroses como dengue, Zika e Chikungunya. Para isso, uma revisão sistemática da literatura foi realizada, em julho de 2020, utilizando-se as bases de pesquisa *PubMed*, *Science Direct*, *Web of Science*, *Scielo* e *LILACS*. Tais bases foram selecionadas em função da relação do conteúdo indexado com o tema avaliado nesse estudo. A fim de analisar as pesquisas mais recentes relacionadas ao tema, foram avaliados artigos científicos publicados no período de 2015 a 2020.

A busca dos artigos nas bases citadas foi realizada utilizando-se uma combinação de palavras-chave, incluindo “*arbovirus*”, “*arboviruses*”, “*climate*”, “*climate change*”, “*prediction*”, “*forescast*”, “*projection*”, “*scenario*”, “*zika*”, “*dengue*” e “*chikungunya*”. Para um melhor desempenho da busca realizada, foram utilizados os operadores booleanos “AND” “OR” e “NOT”.

Como primeira etapa para seleção, foram avaliados os títulos dos documentos identificados nas buscas. Os estudos que apresentaram títulos condizentes com o tema desse estudo foram selecionados para leitura dos resumos. Em seguida, foram selecionados documentos para leitura completa e análise de conteúdo. Adicionalmente, nessa etapa, foram excluídos artigos duplicados.

Por fim, alguns critérios de elegibilidade foram utilizados na seleção dos artigos a serem avaliados, sendo eles:

- Restrição da pesquisa a artigos de revistas revisadas por pares;
- Utilização de artigos classificados apenas como “artigo de pesquisa” ou “estudo de caso”;
- Seleção de artigos que incluía variáveis climáticas e apresentavam pelo menos uma projeção climática futura para avaliação da influência do clima sob arboviroses.

2.2.2 Avaliação dos estudos selecionados

Os artigos selecionados, utilizando-se os critérios citados anteriormente, foram analisados, resumidos e organizados em um banco de dados contendo os seguintes itens: autores, ano de publicação, título da publicação, área de estudo, objetivo, variáveis avaliadas, abordagens analíticas, cenários climáticos adotados, principais resultados e demais observações pertinentes.

Para uma melhor discussão do conteúdo, a avaliação dos artigos foi dividida em duas categorias, selecionadas em função de seus objetivos principais, sendo elas: artigos que realizaram a projeção da distribuição geográfica dos vetores *A. aegypti* e *A. albopictus* e artigos com foco na projeção da distribuição ou incidência de arboviroses.

2.3 Resultados e discussão

2.3.1 Características gerais dos estudos avaliados

As buscas iniciais realizadas nas bases de pesquisa identificaram 2.136 artigos que possuíam quaisquer relações com as palavras-chave utilizadas. Seguindo os critérios de elegibilidade, citados no tópico anterior, foram selecionados 28 artigos para revisão. Um resumo do número de artigos selecionados em cada etapa da seleção é apresentado na Figura 2.1.

Dentre os 28 estudos selecionados, 8 realizaram projeções climáticas a nível global. Os demais estudos realizaram previsões em escala nacional, regional ou local. Os estudos apresentaram diferenças quanto às variáveis de resultado avaliadas, sendo verificadas previsões baseadas na adequação espacial para propagação dos vetores (*A. aegypti* e *A. albopictus*), transmissão viral e no padrão espacial da incidência de arboviroses. Um resumo das características dos estudos avaliados é realizado na Tabela 2.1.

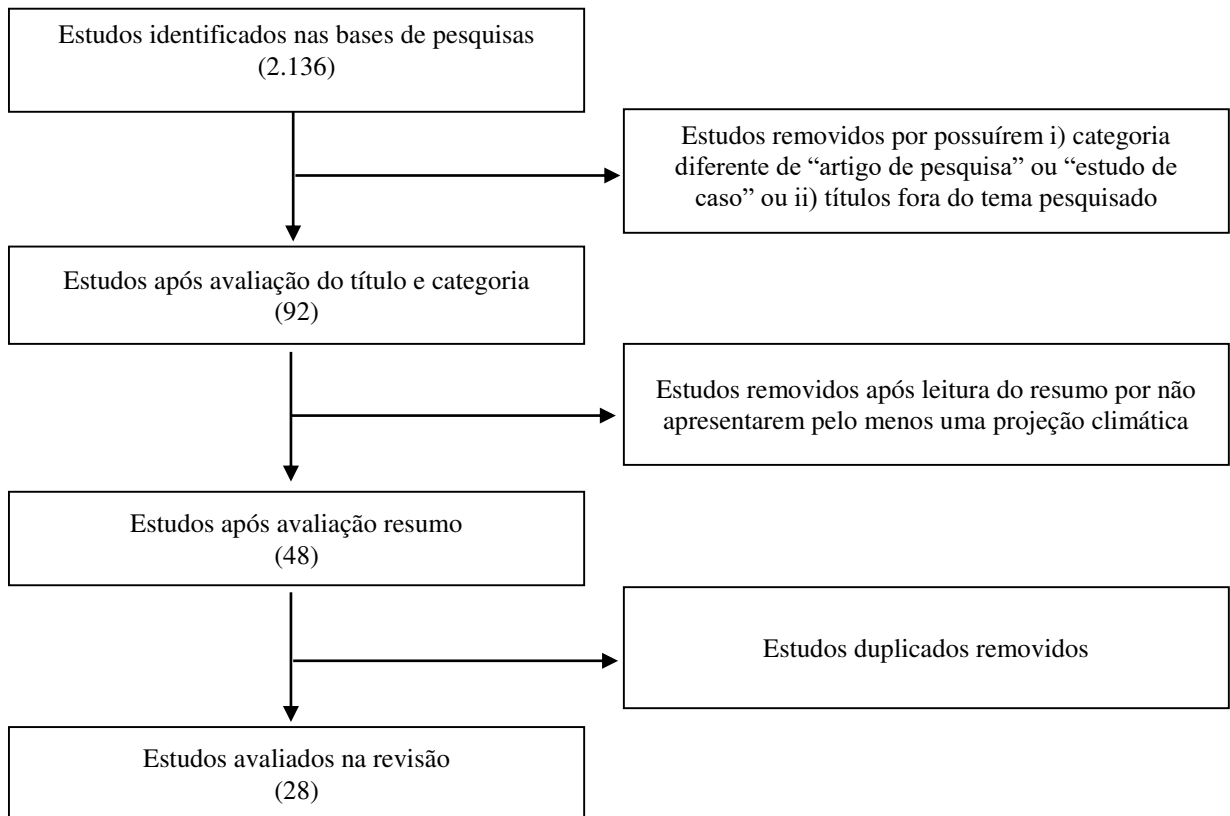


Figura 2.1. Fluxograma do procedimento de seleção de artigos para revisão.

Tabela 2.1. Resumo das informações dos estudos avaliados na revisão sistemática.

Estudo	Área de estudo	Período projetado	Cenários climáticos	Tipo de modelo	Variável resposta
Campbell <i>et al.</i> (2015)	Global	2050	A2, B1 e A1B	Correlativo	Distribuição geográfica do <i>A. aegypti</i> e <i>A. albopictus</i>
Kamal <i>et al.</i> (2018)	Global	2050 e 2070	RCP (2,6; 4,5; 6,0; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i> e <i>A. albopictus</i>
Kraemer <i>et al.</i> (2019)	Global	2020, 2050 e 2080	RCP (4,5; 6,0; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i> e <i>A. albopictus</i>
Messina <i>et al.</i> (2019)	Global	2020, 2050 e 2080	RCP (4,5; 6,0; 8,5)	Correlativo	Áreas com adequação ao vírus e população humana em risco
Monaghan <i>et al.</i> (2018)	Global	2061-2080	RCP (4,5; 8,5)	Correlativo	Exposição humana ao <i>A. aegypti</i>
Proestos <i>et al.</i> (2015)	Global	2045-2054	RCP8,5	Mecanicista	Áreas com adequação de habitat para o <i>A. albopictus</i>
Ryan <i>et al.</i> (2019)	Global	2050 e 2080	RCP (4,5; 6,0; 8,5)	Mecanicista	Áreas com adequação para transmissão viral por <i>A. aegypti</i> e <i>A. albopictus</i> e população em risco de exposição ao vírus
Tjaden <i>et al.</i> (2017)	Global	2021-2040 2041-2060 2061-2080	RCP (4,5; 8,5)	Correlativo	Áreas adequadas à transmissão de chikungunya
Escobar <i>et al.</i> (2016)	Equador	2030, 2050 e 2100	A2	Correlativo	Áreas adequadas para vetores e risco para a exposição humana aos vetores
Lippi <i>et al.</i> (2019)	Equador	2050	RCP (2,6; 4,5; 6,0; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i>
Portilla Cabrera e Selvaraj (2020)	Colômbia	2050 e 2070	RCP (4,5; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i>
Khan <i>et al.</i> (2020)	Canadá e EUA	2011-2100	RCP (4,5; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i> e <i>A. albopictus</i> e risco para a exposição humana aos vetores
Fan e Liu (2019)	China	2020, 2030, 2050 e 2100	RCP (2,6; 4,5; 6,0; 8,5)	Mecanicista	População e áreas com risco de transmissão de dengue
Carbajo <i>et al.</i> (2019)	Argentina	2025-2030	RCP (4,5; 8,5)	Correlativo	Ocorrência de <i>A. aegypti</i> por localidade

Williams <i>et al.</i> (2016)	Austrália	2046-2064	A2 e B1	Mecanicista	Incidência de dengue
Mweya <i>et al.</i> (2016)	Tanzânia	2020 e 2050	não informado	Correlativo	Distribuição dos vetores da dengue em relação às áreas de risco epidêmico
Lee <i>et al.</i> (2018)	Coréia	2020-2099	RCP (2,6; 4,5; 6,0; 8,5)	Mecanicista	Risco de surto de dengue
Acharya <i>et al.</i> (2018)	Nepal	2050 e 2070	RCP (2,6; 6,0; 8,5)	Correlativo	População exposta em áreas climaticamente adequadas à dengue
Butterworth <i>et al.</i> (2017)	Sudeste dos EUA	2045-2065	A1B	Mecanicista	Número de casos de dengue
Liu <i>et al.</i> (2019a)	China continental	2030 e 2050	RCP (2,6; 4,5; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i> e <i>A. albopictus</i>
Liu-Helmersson <i>et al.</i> (2019a)	Europa continental	2006–2099	RCP (2,6; 8,5)	Mecanicista	Distribuição geográfica do <i>A. aegypti</i>
Liu-Helmersson <i>et al.</i> (2016)	Cidades da Europa	2070-2099	RCP (2,6; 4,5; 6,0; 8,5)	Mecanicista	Potencial de surtos de dengue
Colón-González <i>et al.</i> (2017)	América latina	2040–2069 2086–2115	Aumento da temperatura média global em 1,5 °C, 2,0 °C ou 3,7 °C	Correlativo	Número de casos de dengue
Jia <i>et al.</i> (2017)	Guangzhou - China	Aquecimento ao longo de quatro estações climáticas	Aumento da temperatura variando de 0,5 a 5 °C	Mecanicista	Reação térmica teórica de <i>A. albopictus</i> adultos ao longo de quatro estações do ano
Li <i>et al.</i> (2017)	Guangzhou - China	2020-2070	RCP (2,6; 4,5; 6,0; 8,5)	Correlativo	Número de casos de dengue
Equihua <i>et al.</i> (2017)	Veracruz - México	2015-2039 2075-2099	RCP (4,5; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i> e número de pessoas potencialmente expostas
Bal e Sodoudi (2020)	Calcutá - Índia	2020-2100	RCP (4,5; 8,5)	Correlativo	Número de casos de dengue
Liu <i>et al.</i> (2019b)	Taiwan - China	2030, 2050 e 2070	RCP (2,6; 6,0; 8,5)	Correlativo	Distribuição geográfica do <i>A. aegypti</i>

Os estudos também apresentaram diferenças quanto às abordagens metodológicas utilizadas, sendo adotadas abordagens mecanicistas ou correlativas. Modelos mecanicistas requerem o conhecimento de parâmetros relacionados à morfologia, fisiologia e comportamento de vetores e patógenos e sua associação com variáveis ambientais. Tais informações, comumente, são derivadas de experimentos laboratoriais. Os modelos correlativos, por sua vez, são empíricos e utilizam informações de ocorrência geográfica, disponíveis em bancos de dados, para prever a distribuição espacial de espécies ou doenças (TJADEN *et al.*, 2018).

Além das diferenças quanto à abordagem de modelagem utilizada, os estudos apresentaram diferença quanto aos cenários de mudança climática adotados. Predominantemente, os modelos climáticos adotados nos estudos avaliaram quatro Caminhos Representativos de Concentração (RCP) – 2,6; 4,5; 6,0 e 8,5 – desenvolvidos Painel Intergovernamental sobre Mudanças Climáticas (IPCC), que simulam diferentes cenários de emissão de gases de efeito estufa no clima futuro. Os RCPs são nomeados de acordo com os níveis de forçamento radiativo total aproximado no ano de 2100, em relação ao ano de 1750, sendo estes de: 2,6 W m⁻² para RCP 2,6; 4,5 W m⁻² para RCP 4,5; 6,0 W m⁻² para RCP 6,0 e 8,5 W m⁻² para RCP 8,5 (IPCC, 2013). Uma descrição dos cenários citados é realizada na Tabela 2.2.

Tabela 2.2. Descrição dos cenários representativos das vias de concentração de gases de efeito estufa avaliados nos artigos revisados.

Cenário	Descrição
RCP 2,6	Cenário de mitigação que leva a um nível de forçamento muito baixo - Patamar onde o forçamento radiativo atinge, aproximadamente, 3 W m ⁻² antes de 2100 e, em seguida, declina
RCP 4,5	Cenário de estabilização - Patamar de estabilização intermediário em que o forçamento radiativo está estabilizado a aproximadamente 4,5 W m ⁻² após 2100
RCP 6,0	Cenário de estabilização - Patamar de estabilização intermediário em que o forçamento radiativo está estabilizado a aproximadamente 6,0 W m ⁻² após 2100
RCP 8,5	Cenário com emissões de gases de efeito estufa muito elevadas - Patamar elevado em que o forçamento radiativo é superior a 8,5 W m ⁻² em 2100 e continua a aumentar durante algum tempo

Fonte: IPCC (2013).

Além dos RCPs, foram verificados estudos que utilizaram os cenários de mudança climática A1B, A2 e B1 (que correspondem, aproximadamente, aos cenários RCP 6,0, RCP 8,5 e RCP 4,5, respectivamente) (CAMPBELL *et al.*, 2015; ESCOBAR *et al.*, 2016; WILLIAMS *et al.*, 2016) e que utilizaram cenários que assumem, baseado nas estratégias de mitigação, que a temperatura média global poderá aumentar em 1,5 °C, 2,0 °C ou 3,7 °C em 2100, em relação aos níveis pré-industriais (COLÓN-GONZÁLEZ *et al.*, 2018).

2.3.2 Distribuição de vetores e mudanças climáticas

Os artigos avaliados nesse estudo, predominantemente, referem-se à projeção da distribuição dos vetores *A. aegypti* e *A. albopictus* em diferentes cenários de mudanças climáticas. Embora, isoladamente, os fatores climáticos não possam explicar o risco de exposição humana às arboviroses, estes podem limitar a distribuição de vetores ou, ainda, afetar a sua faixa de distribuição geográfica. Morin *et al.* (2013) desenvolveram um esquema que contempla as relações existentes entre os mosquitos do gênero *Aedes*, transmissão de dengue, clima e tempo (Figura 2.2).

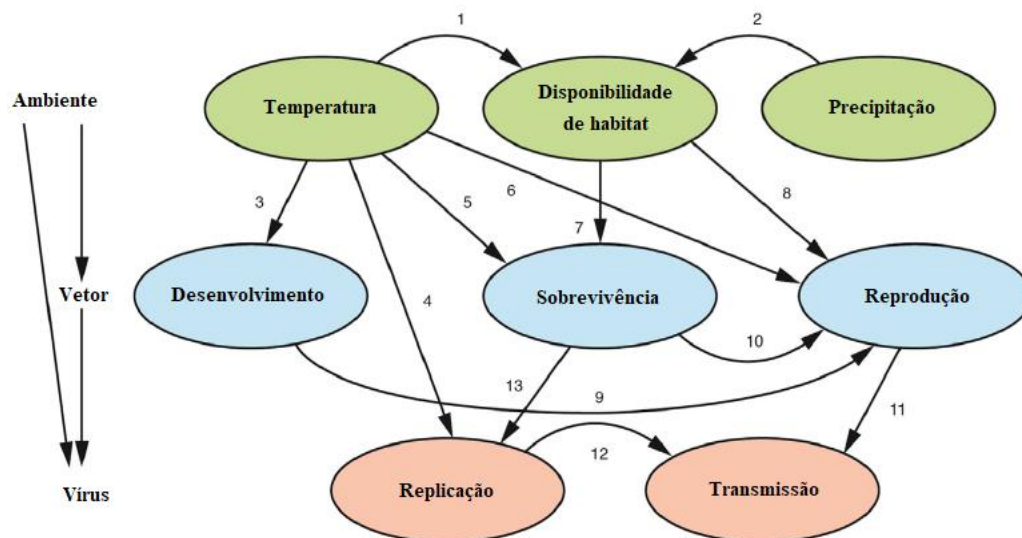


Figura 2.2. Interações entre variáveis climáticas, vetores e vírus da dengue.

Os números indicados na figura identificam as relações existentes entre as variáveis. A disponibilidade de habitat para larvas de mosquitos é influenciada pela (1) temperatura, por meio da evaporação e transpiração, e pela (2) precipitação de entrada. A temperatura é um importante regulador do (3) desenvolvimento do mosquito, (4) da replicação viral em mosquitos infectados, (5) da sobrevivência e (6) do comportamento reprodutivo dos mosquitos. A disponibilidade de habitat é necessária para a (7) sobrevivência das formas imaturas de mosquitos e (8) reprodução dos mosquitos adultos. A reprodução dos mosquitos é acelerada pelo (9) desenvolvimento mais rápido e (10) aumento da sobrevivência dos mosquitos. O aumento reprodutivo do mosquito aumenta a (11) probabilidade de transmissão, por meio do aumento do número de picadas, e a (12) replicação viral mais rápida aumenta a transmissão, uma vez que, encurta o período de incubação extrínseco. Por fim, o aumento da sobrevivência dos mosquitos adultos aumenta a (13) quantidade de replicação viral.

Fonte: Adaptado de Morin *et al.* (2013).

Campbell *et al.* (2015) e Kamal *et al.* (2018) investigaram a distribuição global de *A. aegypti* e *A. albopictus* utilizando variáveis bioclimáticas aplicadas a um modelo de nicho ecológico. Os autores utilizaram cenários de mudança climática distintos para projetar a futura distribuição dos vetores. Campbell *et al.* (2015) realizaram projeções para o ano de 2050, considerando os cenários A2, B1 e A1B. Kamal *et al.* (2018), por outro lado, adotaram quatro RCPs, com projeções para os anos de 2050 e 2070.

Apesar das diferenças metodológicas entre os trabalhos citados, as conclusões obtidas em termos globais foram, em geral, coincidentes. Os resultados indicaram rearranjos globais complexos nas áreas de distribuição para *A. aegypti* e *A. albopictus*. De modo geral, considerando-se mudanças em nível global, com exceção de mudanças em escalas regionais, houve semelhança entre os padrões distributivos gerais dos mosquitos nas condições atuais e futuras. Para Campbell *et al.* (2015), tal resultado pode ser explicado em função da ampla tolerância climática e das diferenças entre as duas espécies, o que faz com que as mudanças climáticas nas próximas décadas possam não se traduzir em grandes mudanças distributivas. Kamal *et al.* (2018) citam que o *A. aegypti* apresentou potencial de distribuição marcadamente mais amplo nas regiões tropicais e subtropicais do que o *A. albopictus*. Por outro lado, Reinhold *et al.* (2018) destacam que o *A. albopictus* é capaz de sobreviver em climas muito mais frios do que o *A. aegypti*. Corroborando com essa afirmação, Leta *et al.* (2018) verificaram que as regiões temperadas apresentam maior adequação para distribuição do *A. albopictus*.

Vale ressaltar que estes são resultados em termos globais dos estudos de Campbell *et al.* (2015) e Kamal *et al.* (2018). Ao se avaliar regiões individuais são verificados potenciais de expansão ou contração de ambas as espécies de vetores. No estudo de Campbell *et al.* (2015), por exemplo, verifica-se que em cenários futuros, a área de distribuição para o *A. aegypti* pode aumentar na América do Sul e, ao mesmo tempo, diminuir na Austrália. Ainda, no estudo de Kamal *et al.* (2018), foram verificadas diferenças entre os cenários de emissão de gases de efeito estufa utilizados, sendo previstos aumentos na distribuição de ambas espécies do RCP 2,6 para o RCP 8,5. Adicionalmente, considerando-se os padrões distributivos individualmente, os resultados previstos por Campbell *et al.* (2015) diferiram daqueles previstos por Kamal *et al.* (2018) em muitas regiões. O estudo de Campbell *et al.* (2015), por exemplo, subestimou a ocorrência dos vetores em áreas em que estes foram considerados abundantes por Kamal *et al.* (2018). Desse modo, os resultados obtidos em estudos globais devem ser avaliados com cautela.

Ainda em termos globais, Proestos *et al.* (2015) avaliaram a adequação global de habitats para o *A. albopictus*, considerando os impactos das mudanças climáticas. Os autores verificaram que, para o período de 2045 a 2054 (RCP 8,5), são projetadas reduções

significativas de adequação em áreas onde as temperaturas no verão podem ser tornar muito altas. Desse modo, estima-se que as condições ambientais dos trópicos serão menos favoráveis para o *A. albopictus*. Ao mesmo tempo, outras regiões, que apresentarão aumento das temperaturas no inverno, poderão se tornar mais predispostas para o desenvolvimento desse vetor, o que permitirá que a espécie compense as perdas de território. Diante disso, os autores estimam que cerca de 2,4 bilhões de pessoas serão potencialmente expostas ao *A. albopictus* no período projetado, caso medidas de mitigação de emissões de gases de efeito estufa não sejam implementadas.

Em conjunto com os fatores climáticos, Kraemer *et al.* (2019) e Monaghan *et al.* (2018) realizaram a previsão global de vetores considerando fatores demográficos. Assim como Campbell *et al.* (2015) e Kamal *et al.* (2018), Kraemer *et al.* (2019) observaram uma prevalência da distribuição de *A. aegypti* em regiões tropicais. Mesmo ao se considerar os cenários de projeção mais extremos (RCP 8,5 em 2080), o estudo de Kraemer *et al.* (2019) previu que o *A. aegypti* deve se estabelecer na Europa apenas em regiões isoladas do sul da Itália e Turquia.

Diferente do comportamento observado para o *A. aegypti*, os resultados do estudo de Kraemer *et al.* (2019) indicaram que o *A. albopictus* se espalhará amplamente por toda a Europa. Além disso, o estabelecimento dessa espécie é previsto em regiões ao norte dos Estados Unidos e regiões montanhosas da América do Sul e África Oriental nos próximos 30 anos. Ao mesmo tempo, prevê-se que áreas onde os modelos climáticos indicam aumento de aridez se tornarão menos apropriadas para o estabelecimento do *A. albopictus*. A projeção realizada prevê que, até o ano de 2080, cerca de 197 países relatarão a presença desse vetor, dos quais 20 a reportarão pela primeira vez.

Kraemer *et al.* (2019) destacam que as previsões revelam que os esforços para conter as mudanças climáticas poderão ser insuficientes para impedir a expansão de mosquitos do gênero *Aedes*. Uma expansão expressivamente maior dos vetores é prevista para os anos de 2050 a 2080. O impacto dessa previsão é considerável, tendo em vista que a avaliação do crescimento populacional realizada pelos autores indicou uma tendência de concentração da população em áreas onde o *A. aegypti* e o *A. albopictus* já estarão estabelecidos, o que resultará em um aumento da população global exposta a arboviroses.

Além das projeções climáticas, utilizando-se os cenários RCP (4,5 e 8,5), Monaghan *et al.* (2018) avaliaram a exposição humana futura ao *A. aegypti* considerando-se projeções populacionais globais. Avaliando a área de exposição humana ao *A. aegypti*, Monaghan *et al.* (2018) projetaram um aumento global da área de cerca de 8% (RCP 4,5) e 13% (RCP 8,5) entre

2061-2080, em relação ao período de referência avaliado. Por outro lado, os autores verificaram que apenas em algumas regiões as mudanças climáticas desempenham um papel importante no aumento da exposição humana ao *A. aegypti*.

Ao se avaliar globalmente a exposição humana ao vetor, foi verificado que as mudanças climáticas são um fator pouco influente no aumento da exposição quando comparado ao aumento decorrente do crescimento populacional. Esse comportamento foi observado, sobretudo, em países em desenvolvimento e de renda média, tais como os países da África, Ásia e América do Sul, onde o elevado crescimento populacional supera o impacto das mudanças climáticas na projeção da população exposta. Como resultado, em tais países, as diferenças na exposição humana ao *A. aegypti* entre os cenários de mudança climática RCP 4,5 e 8,5 não são estatisticamente significativas. Tais resultados revelam não haver uma vantagem aparente na adoção do cenário com menores emissões de gases de efeito estufa nessas regiões, em termos de redução da população humana exposta ao vetor. Em contrapartida, para a América do Norte e Europa, foram verificadas diferenças significativas entre os cenários de emissão RCP 4,5 e RCP 8,5, o que sugere que, nessas regiões, a adoção de estratégias de mitigação (RCP 4,5) pode diminuir a população exposta ao *A. aegypti*.

A projeção da distribuição de vetores nos demais artigos foi realizada em escala nacional ou local. A avaliação de estudos em tais escalas é muito importante, uma vez que, embora os estudos globais sejam informativos, são as análises locais que possuem o potencial de enfatizar as evidências de que o clima e a transmissão de arboviroses estão intimamente relacionados (ROBERT *et al.*, 2020).

De modo geral, os estudos desenvolvidos em regiões de clima tropical indicaram que o aumento da temperatura poderá reduzir as áreas, atualmente, consideradas aptas para o desenvolvimento de mosquitos do gênero *Aedes*. Na Colômbia, as projeções climáticas previram uma redução de cerca de 30% das áreas de distribuição do *A. aegypti* (PORTILLA-CABRERA; SELVARAJ, 2020). Tal redução está associada à elevação da temperatura acima dos níveis ideais para o ciclo de vida e capacidade reprodutiva do mosquito. Por esse motivo, para a caracterização das distribuições espaço-temporais das arboviroses, é fundamental compreender como os vetores respondem aos fatores climáticos. Tal compreensão permite, ainda, antecipar e responder às mudanças potenciais no risco de incidência de arboviroses, em decorrência das mudanças climáticas (IWAMURA *et al.*, 2020).

No Equador, os resultados obtidos por Escobar *et al.* (2016) sugerem que as mudanças climáticas podem reduzir a carga de vetores em regiões do país em que estes encontram-se, atualmente, presentes. Em contrapartida, no mesmo país, Lippi *et al.* (2019) verificaram que os

efeitos das mudanças climáticas poderão estender a distribuição do *A. aegypti* a regiões com altitudes mais elevadas onde, nas condições atuais, estes não são previstos. A tendência de deslocamento dos vetores para regiões de maior altitude também foi verificada na cidade de Veracruz (México) (EQUIHUA *et al.*, 2017). De acordo com uma revisão de literatura realizada por Ciota e Keyel (2019), de modo geral, espera-se que os arbovírus mudem na direção dos polos e em altitudes mais elevadas sob as mudanças climáticas. Segundo Castillo-Quino *et al.* (2018), a alteração do nicho ecológico dos vetores, influenciada pelas mudanças climáticas, contribui para que o *A. aegypti* supere a adaptação a altitudes acima daquelas atualmente descritas na literatura.

Por outro lado, estudos realizados no Canadá e Estados Unidos (KHAN *et al.*, 2020) e China (LIU *et al.*, 2019a; LIU *et al.*, 2019b) indicaram que, devido às mudanças climáticas, os mosquitos do gênero *Aedes* poderão expandir a faixa de distribuição atual. Ainda, na China, Jia *et al.* (2017) destacam que projeções realizadas revelam que os efeitos do aquecimento na distribuição de vetores são diferentes ao longo das estações. No inverno e na primavera, tais efeitos facilitam o desenvolvimento das espécies, encurtando o período de diapausa. Em contrapartida, no verão, o efeito do aquecimento inibe o desenvolvimento do mosquito.

Por fim, na Europa continental, Liu-Helmersson *et al.* (2019) verificaram a influência dos cenários de emissão na alteração da distribuição do *A. aegypti* ao longo do século XXI. No cenário de baixa emissão (RCP 2,6), foram identificadas diferenças sutis em comparação à distribuição atual. No entanto, ao avaliar o cenário de alta emissão (RCP 8,5), foi verificado que grande parte da região sul da Europa apresenta risco potencial de sofrer invasão do *A. aegypti*.

Baseado nos resultados dos estudos avaliados é possível verificar que as mudanças climáticas poderão afetar a faixa de distribuição geográfica do *A. aegypti* e *A. albopictus*. No entanto, os efeitos das mudanças climáticas variam em diferentes regiões, podendo causar, no futuro, expansão ou contração das áreas atualmente consideradas aptas para o desenvolvimento dos vetores. De modo geral, em decorrência das mudanças climáticas, verifica-se um potencial crescimento das áreas aptas para o desenvolvimento de vetores em regiões temperadas, sobretudo para o *A. albopictus*. Regiões de clima tropical, por outro lado, poderão apresentar uma redução das áreas aptas ou, ainda, migração dos vetores para regiões com maior altitude.

Vale ressaltar que, embora a presença de vetores seja essencial para a transmissão de arboviroses, o aumento da exposição da população não se traduz, necessariamente, em aumento do risco de transmissão dessas doenças. Como exemplo disso, o estudo desenvolvido por Leta *et al.* (2018) mostrou que, dos 250 territórios considerados, 215 (86%) eram potencialmente

apropriados para a existência e desenvolvimento de *A. aegypti* ou *A. albopictus*. Contudo, a ocorrência de pelo menos uma das arboviroses avaliadas (Zika, dengue, Chikungunya, febre amarela e febre do Vale do Rift) foi relatada em, somente, 146 (58%) territórios, segundo os dados de fontes disponíveis publicamente. Apesar disso, entender a distribuição futura de vetores é um passo essencial para entendimento da futura distribuição das doenças transmitidas por estes (XU *et al.*, 2020).

Visando contornar as limitações das projeções baseadas apenas na distribuição de vetores, Mweya *et al.* (2016) utilizaram um modelo de nicho ecológico que incorporou dados de *A. aegypti* infectados com o vírus da dengue para estimar a distribuição de áreas de risco epidêmico. Messina *et al.* (2019), por sua vez, para investigar a distribuição futura de dengue, avaliaram dados históricos de incidência da doença em conjunto com informações sobre o vírus e disseminação de vetores. A incorporação desses fatores na modelagem de distribuição de espécies pode contribuir para o aperfeiçoamento da avaliação do real risco epidêmico em regiões identificadas com alta aptidão para o desenvolvimento de vetores.

2.3.3 Transmissão viral, incidência de arboviroses e mudanças climáticas

Os demais artigos identificados nessa revisão avaliaram a relação entre as mudanças climáticas e os componentes do ciclo de transmissão viral ou distribuição geográfica observada de arboviroses.

O estudo desenvolvido por Butterworth *et al.* (2017), no sudeste dos Estados Unidos, destacou a influência do clima tanto para os vetores quanto para os vírus. Os autores enfatizaram que o aumento do risco de doença em um local depende tanto da adequação deste para a sobrevivência do vetor, quanto da existência de condições ótimas para a transmissão viral. Desse modo, existem locais que poderão observar a presença de vetores e, todavia, permanecer à margem da transmissão viral. Contudo, as condições favoráveis ao vírus surgem apenas em períodos mais quentes.

Segundo Ryan *et al.* (2019), a faixa de temperatura ideal para transmissão viral do *A. aegypti* é de 21,3 a 34,0 °C, para o *A. albopictus*, por outro lado, tal faixa é de 19,9 a 29,4 °C. As faixas descritas apresentam valores muito superiores aos limites de sobrevivência, relatados na literatura, dos vetores, o que reforça as informações descritas por Butterworth *et al.* (2017)

De acordo com Reinhold *et al.* (2018), os vetores atuam dentro de uma faixa de temperaturas, de modo que, se extrapolados os pontos críticos mínimos e máximos, sua atividade é cessada e o risco de morte aumenta. Na revisão de literatura realizada pelos referidos

autores, é destacado que o limite inferior de temperatura para *A. aegypti* é de cerca de 10 °C. Em temperaturas inferiores à citada, é relatado que os vetores ficam entorpecidos e incapazes de se mover. Tal limite de temperatura pode apresentar valores distintos em função da espécie de vetor e do estágio de desenvolvimento deste. No estudo realizado por Tsai *et al.* (2018) foi verificado que a temperatura crítica para limitar a ocorrência de larvas de *A. aegypti* é de 13,8 °C. Tais resultados demonstram que larvas e adultos de *A. aegypti* possuem a capacidade de se aclimatar a baixas temperaturas. Essa capacidade de aclimação ao frio pode facilitar a disseminação dessa espécie para latitudes mais altas, sobretudo, considerando-se um cenário de mudanças climáticas (JASS *et al.*, 2019).

Em relação ao vetor *A. albopictus*, Tippelt *et al.* (2020) verificaram que os ovos de todas as cepas utilizadas foram capazes de sobreviver a temperaturas constantes tão baixas quanto -5 °C por um período de exposição de 30 dias. Adicionalmente, foi verificada a eclosão dos ovos do vetor mesmo com uma temperatura mínima de -10 °C. Segundo os autores, o conhecimento de tais limites é relevante para determinação dos limites ecológicos dessa espécie.

Avaliando a capacidade de transmissão viral, por meio de um modelo mecanicista, Ryan *et al.* (2019) realizaram um estudo em escala global, considerando os vetores *A. aegypti* e *A. albopictus*. Os autores identificaram impactos das mudanças climáticas na transmissão viral, sendo estes causados, sobretudo, pelas diferenças entre o *A. aegypti* e *A. albopictus*, que apresentam maior e menor tolerância térmica, respectivamente. Desse modo, em um cenário de mudança climática, tais diferenças poderão modificar o padrão de transmissão de vírus para ambos os vetores.

O período de incubação extrínseco (PIE) – tempo necessário para que o vetor infectado se torne infectivo - é influenciado pela temperatura, sendo encurtado com pequenos aumentos desta. Um PIE mais curto aumenta o potencial de transmissão viral para os seres humanos (BUTTERWORTH *et al.*, 2017). Assim, ligeiros aumentos na temperatura poderão levar ao aumento da transmissão de arboviroses. Por outro lado, os vetores possuem também um limite superior de temperatura para transmissibilidade que, quando ultrapassado, pode suprimi-la.

De modo geral, Ryan *et al.* (2019) verificaram que as mudanças climáticas poderão alterar o risco atual de transmissão viral para espécies do gênero *Aedes*. Contudo, os padrões identificados diferem entre as espécies de mosquitos, localidades e cenários. No cenário atual, para ambos os vetores, a maior parte da região nos trópicos é ideal para a transmissão viral durante todo o ano e apresenta declínio da adequação ao longo dos gradientes latitudinais.

Por meio da Figura 3.3 é possível verificar que, no ano de 2050, é previsto que a elevação da temperatura resulte em mudanças no risco de transmissão do *Aedes*. Para o *A.*

aegypti, espera-se uma expansão do risco de transmissão, por pelo menos dois meses do ano, em regiões temperadas. Nos trópicos, a expansão do risco é verificada durante todo o ano, até mesmo em regiões de elevada altitude que, anteriormente, não eram consideradas aptas para o vetor. Ainda no ano de 2050, prevê-se que o risco de transmissão viral do *A. albopictus* se expanda em regiões temperadas durante alguns meses. No entanto, nos trópicos, onde o aquecimento excederá os limites térmicos superiores desse vetor, é prevista uma redução do risco sazonal ou durante todo o ano.

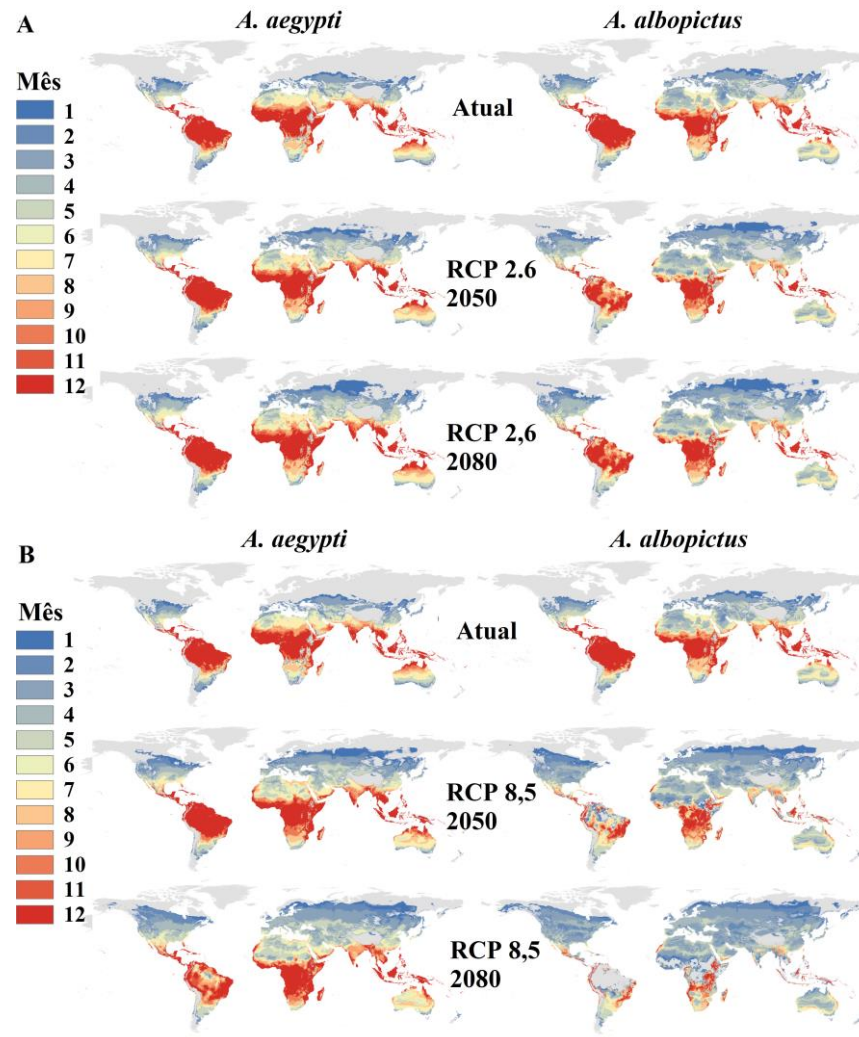


Figura 3.3. Mapeamento da adequação da temperatura para transmissão viral do *A. aegypti* e *A. albopictus* nos cenários de emissão RCP 2,6 (a) e RCP 8,5 (b).

Fonte: Adaptado de Ryan *et al.* (2019).

Desse modo, verifica-se que, para o *A. aegypti*, o gradiente atual de elevada transmissão nos trópicos e baixa transmissão em regiões temperadas é mantido em climas futuros. Por outro lado, o aumento da temperatura nos trópicos diminuirá as áreas com risco de transmissão do *A. albopictus*, sobretudo no RCP 8,5. Nesse cenário, considerando-se a projeção para o ano de 2080, as áreas com adequação (durante todo o ano) de temperatura para transmissão viral por *A. albopictus* será limitada às regiões de alta altitude, sul da África e costa atlântica do Brasil. O *A. aegypti*, por sua vez, perderá uma área central de adequação durante todo o ano, sobretudo na bacia Amazônica.

Diante do exposto, verifica-se que o estudo desenvolvido por Ryan *et al.* (2019) prevê que a mitigação parcial das mudanças climáticas, representada no cenário RCP 2,6, pode ocasionar maior risco de exposição para transmissão viral do que nenhuma mitigação. Embora não seja intuitivo, esse resultado se fundamenta no conhecimento dos limites térmicos para transmissão dos mosquitos do gênero *Aedes*, que poderão ser ultrapassados em cenários de elevada emissão.

Os resultados obtidos por Ryan *et al.* (2019) são consistentes com as observações de Williams *et al.* (2016). Tais autores utilizaram um modelo mecanicista de transmissão de vírus para determinar se as mudanças climáticas irão alterar a transmissão da dengue na Austrália. Os autores verificaram que cenários que consideram emissões de gases de efeito estufa elevadas podem resultar em menor potencial para transmissão de dengue. Por outro lado, tal potencial pode ser elevado caso ocorra um aquecimento climático ameno, como previsto em cenários de baixa emissão. Segundo Williams *et al.* (2016) esse resultado demonstra que a relação entre clima e incidência de dengue não é trivial, sendo incorreto afirmar que o aquecimento climático resultará, necessariamente, no aumento da doença.

Ainda em relação à transmissão de dengue, os estudos desenvolvidos na China (FAN; LIU, 2019) e na Europa (LIU-HELMERSSON *et al.*, 2016) identificaram que, como efeito das mudanças climáticas, a distribuição da área de transmissão de dengue se expandirá de áreas de baixa para alta latitude. A expansão da transmissão viral para regiões sem exposição prévia é considerada crítica, dado o potencial de ocorrência de surtos explosivos quando uma população é exposta pela primeira vez.

Os demais artigos identificados nessa revisão baseiam-se no treinamento prévio de modelos, com dados de ocorrência de arboviroses, para prever a distribuição futura dessas doenças. Acharya *et al.* (2018), Bal e Sodoudi (2020), Colón-González *et al.* (2018) e Li *et al.* (2017) avaliaram a previsão de casos de dengue e Tjaden *et al.* (2017), a adequação global da transmissão de Chikungunya. Assim como verificado nos estudos citados anteriormente, os

resultados da modelagem realizada por Tjaden *et al.* (2017) indicaram que, no cenário de alta emissão, a adequação climática poderá se expandir em algumas áreas e contrair em outras. Os autores ressaltam que a adequação de uma área à transmissão de Chikungunya não significa, necessariamente, que esta estará sujeita a surtos no futuro. A ocorrência da doença em uma área está também associada à adoção de medidas de controle de saúde pública e níveis gerais de desenvolvimento socioeconômico.

Ainda de modo similar aos estudos citados, Acharya *et al.* (2018) verificaram que no Nepal, como efeito das mudanças climáticas, a área com maior adequação para incidência de dengue se expandirá para regiões com elevada altitude. Como consequência disso, prevê-se que a população em risco aumentará cerca de 20% em relação ao cenário atual. Por outro lado, os estudos desenvolvidos por Bal e Sodoudi (2020), Colón-González *et al.* (2018) e Li *et al.* (2017) apresentaram resultados distintos daqueles anteriormente indicados. Os autores avaliaram o efeito das mudanças climáticas na incidência de dengue em países de clima tropical ou subtropical e concluíram que cenários de maior aquecimento refletirão na expansão da incidência para áreas, atualmente, não endêmicas. De modo geral, os estudos avaliados anteriormente indicaram uma tendência de diminuição de áreas aptas para o desenvolvimento de arboviroses, em cenários de alta emissão, nas regiões tropicais.

Os estudos apresentados nessa revisão indicaram os potenciais impactos das mudanças climáticas na transmissão viral e incidência de arboviroses em todo o mundo. É possível verificar que a relação entre esses fatores não é linear e pode variar em diferentes cenários e locais. Vale ressaltar que a maior parte dos estudos realizou projeções considerando apenas a influência de variáveis bioclimáticas.

Para Messina *et al.* (2019), as mudanças demográficas e sociais da população humana nas últimas décadas foram responsáveis pela incidência e reincidência de arboviroses em diversas partes do mundo. Segundo os autores, fatores como rápida urbanização, desmatamento e o conseqüente contato mais próximo de populações de vetores e comunidades humanas, contribuíram para a propagação de arboviroses. Diante disso, verifica-se que a avaliação conjunta de fatores climáticos, sanitários, demográficos e sociais pode auxiliar o entendimento da propagação ou recessão futura de doenças e, desse modo, fornecer subsídios para a implementação de estratégias de controle que visam reduzir o risco epidêmico de arboviroses.

Diante das informações levantadas nesse estudo, pode-se verificar que o desenvolvimento de estudos estatísticos e de modelagem são fundamentais para uma melhor compreensão das ligações entre o clima e a transmissão de arbovírus. Segundo Robert *et al.*

(2020), tal compreensão é particularmente relevante em regiões à margem da emergência de arbovírus.

2.3.4 Limitações dos estudos

Embora os estudos avaliados forneçam importantes subsídios para a avaliação do impacto das mudanças climáticas na incidência de arboviroses, algumas limitações podem ser apontadas, tais como:

- As previsões realizadas, em geral, não consideram o efeito de determinantes não climáticos tais como: urbanização, saneamento, uso da terra e desenvolvimento socioeconômico, que podem produzir mudanças expressivas nos níveis de riscos de exposição identificados;
- As projeções dos estudos avaliados, por vezes, apresentam resultados contraditórios, o que pode estar relacionado ao uso de modelos distintos, à qualidade dos dados utilizados e à escolha das variáveis explicativas para a distribuição de vetores ou doenças;
- Alguns estudos apresentaram falhas na identificação da distribuição atual de vetores/doenças que podem refletir nos resultados projetados. Tais falhas podem ser consequência da dificuldade de obtenção de dados para calibração dos modelos em algumas regiões;
- Embora alguns estudos citem a existência de interação competitiva entre o *A. Aegypti* e o *A. Albopictus*, a competição entre espécies, que pode afetar a taxa de expansão de alcance de cada um, não foi considerada nos estudos avaliados. Desse modo, conforme destacam Lwande *et al.* (2020), seria relevante prever se a distribuição global do *A. aegypti* e *A. albopictus* resultará em uma coexistência simpátrica ou exclusão competitiva entre as espécies.

2.4 Conclusões

Baseado nos estudos avaliados nessa revisão, em diferentes escalas geográficas, verifica-se que as mudanças climáticas poderão afetar a faixa de distribuição geográfica do *A. aegypti* e *A. albopictus*, bem como a aptidão das regiões para transmissão viral. No entanto, os efeitos das mudanças climáticas variam em diferentes regiões, podendo causar, no futuro,

expansão ou contração das áreas atualmente consideradas aptas para o desenvolvimento de arboviroses.

De modo geral, em decorrência das mudanças climáticas, verifica-se um potencial crescimento das áreas aptas para o desenvolvimento de vetores em regiões temperadas, sobretudo para o *A. albopictus*. Em regiões de clima tropical, por outro lado, o aumento da temperatura acima do limite de sobrevivência de vetores ou patógenos poderá limitar as áreas aptas para incidência de arboviroses. Prevê-se, ainda, uma migração dos vetores para regiões com altitude elevada, consideradas, atualmente, inapropriadas para a transmissão de arbovírus.

Embora não seja intuitivo, alguns estudos previram que a mitigação parcial das mudanças climáticas pode ocasionar maior risco de exposição para transmissão viral por mosquitos do gênero *Aedes* do que nenhuma mitigação. Desse modo, é possível verificar que a relação entre as mudanças climáticas e a transmissão viral não é linear e pode variar em diferentes cenários e locais.

Vale ressaltar que a maior parte dos estudos realizou projeções considerando apenas a influência de variáveis bioclimáticas. Contudo, outros determinantes como desenvolvimento socioeconômico, saneamento, urbanização e implantação de intervenções podem alterar os níveis de risco enfrentado pela população.

2.5 Referências

ACHARYA, B. K. *et al.* Present and future of dengue fever in nepal: Mapping climatic suitability by ecological niche model. **International Journal of Environmental Research and Public Health**, v. 15, n. 2, 2018.

BAL, S.; SODOUDI, S. Modeling and prediction of dengue occurrences in Kolkata, India, based on climate factors. **International Journal of Biometeorology**, n. 382, 2020.

BARBAZAN, P. *et al.* Modelling the effect of temperature on transmission of dengue. **Medical and Veterinary Entomology**, v. 24, p. 66–73, 2010.

BUTTERWORTH, M. K.; MORIN, C. W.; COMRIE, A. C. An analysis of the potential impact of climate change on dengue transmission in the southeastern United States. **Environmental Health Perspectives**, v. 125, n. 4, p. 579–585, 2017.

CAMPBELL, L. P. *et al.* Climate change influences on global distributions of dengue and chikungunya virus vectors. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 370, n. 1665, p. 1–9, 2015.

COLÓN-GONZÁLEZ, F. J. *et al.* Limiting global-mean temperature increase to 1.5-2 °C could reduce the incidence and spatial spread of dengue fever in Latin America. **Proceedings of the National Academy of Sciences of the United States of America**, v. 115, n. 24, p.

6243–6248, 2018.

EL-SAYED, A.; KAMEL, M. Climatic changes and their role in emergence and re-emergence of diseases. **Environmental Science and Pollution Research**, v. 27, n. 18, p. 22336–22352, 2020.

EQUIHUA, M. *et al.* Establishment of *Aedes aegypti* (L.) in mountainous regions in Mexico: Increasing number of population at risk of mosquito-borne disease and future climate conditions. **Acta Tropica**, v. 166, p. 316–327, 2017.

ESCOBAR, L. E. *et al.* Declining Prevalence of Disease Vectors Under Climate Change. **Scientific Reports**, v. 6, n. 1, p. 1–8, 2016.

FAN, J. C.; LIU, Q. Y. Potential impacts of climate change on dengue fever distribution using RCP scenarios in China. **Advances in Climate Change Research**, v. 10, n. 1, p. 1–8, 2019.

GAGE, K. L. *et al.* Climate and Vectorborne Diseases. **American Journal of Preventive Medicine**, v. 35, n. 5, p. 436–450, 2008.

HILGENFELD, R.; VASUDEVAN, S. G. **Dengue and Zika : Control and Antiviral Treatment Strategies**. Singapura: Springer Nature, 2018.

JIA, P. *et al.* How does the dengue vector mosquito *Aedes albopictus* respond to global warming? **Parasites and Vectors**, v. 10, n. 1, p. 1–12, 2017.

KAMAL, M. *et al.* Mapping the global potential distributions of two arboviral vectors *Aedes aegypti* and *Ae. Albopictus* under changing climate. **PLoS ONE**, v. 13, n. 12, p. 1–21, 2018.

KHAN, S. U. *et al.* Current and projected distributions of *Aedes aegypti* and *Ae. Albopictus* in Canada and the U.S. **Environmental Health Perspectives**, v. 128, n. 5, p. 1–13, 2020.

KRAEMER, M. U. G. *et al.* Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. **Nature Microbiology**, v. 4, n. 5, p. 854–863, 2019.

LETA, S. *et al.* Global risk mapping for major diseases transmitted by *Aedes aegypti* and *Aedes albopictus*. **International Journal of Infectious Diseases**, v. 67, p. 25–35, 2018.

LI, C. *et al.* Modeling and projection of dengue fever cases in Guangzhou based on variation of weather factors. **Science of the Total Environment**, v. 605–606, n. 19, p. 867–873, 2017.

LIPPI, C. A. *et al.* Geographic shifts in *Aedes aegypti* habitat suitability in Ecuador using larval surveillance data and ecological niche modeling: Implications of climate change for public health vector control. **PLoS Neglected Tropical Diseases**, v. 13, n. 4, p. 1–19, 2019.

LIU-HELMERSSON, J. *et al.* Climate Change and *Aedes* Vectors: 21st Century Projections for Dengue Transmission in Europe. **EBioMedicine**, v. 7, p. 267–277, 2016.

LIU-HELMERSSON, J. *et al.* Climate change may enable *Aedes aegypti* infestation in major European cities by 2100. **Environmental Research**, v. 172, n. December 2018, p. 693–699, 2019.

LIU, B.; JIAO, Z.; *et al.* Modelling the potential distribution of arbovirus vector *Aedes*

aegypti under current and future climate scenarios in Taiwan, China. **Pest Management Science**, v. 75, n. 11, p. 3076–3083, 2019.

LIU, B.; GAO, X.; *et al.* Modeling the present and future distribution of arbovirus vectors *Aedes aegypti* and *Aedes albopictus* under climate change scenarios in Mainland China. **Science of the Total Environment**, v. 664, p. 203–214, 2019.

MESSINA, J. P. *et al.* The many projected futures of dengue. **Nature Reviews Microbiology**, v. 13, n. 4, p. 230–239, 2015.

MESSINA, J. P. *et al.* The current and future global distribution and population at risk of dengue. **Nature Microbiology**, v. 4, n. 9, p. 1508–1515, 2019.

MONAGHAN, A. J. *et al.* The potential impacts of 21st century climatic and population changes on human exposure to the virus vector mosquito *Aedes aegypti*. **Climatic Change**, v. 146, n. 3–4, p. 487–500, 2018.

MWEYA, C. N. *et al.* Climate change influences potential distribution of infected *Aedes aegypti* co-occurrence with dengue epidemics risk areas in Tanzania. **PLoS ONE**, v. 11, n. 9, p. 1–13, 2016.

PORTILLA CABRERA, C. V.; SELVARAJ, J. J. Geographic shifts in the bioclimatic suitability for *Aedes aegypti* under climate change scenarios in Colombia. **Heliyon**, v. 6, n. 1, p. e03101, 2020.

PROESTOS, Y. *et al.* Present and future projections of habitat suitability of the Asian tiger mosquito, a vector of viral pathogens, from global climate simulation. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 370, n. 1665, p. 1–16, 2015.

RYAN, S. J. *et al.* Global expansion and redistribution of *Aedes*-borne virus transmission risk with climate change. **PLoS Neglected Tropical Diseases**, v. 13, n. 3, p. 1–20, 2019.

TJADEN, N. B. *et al.* Modelling the effects of global climate change on Chikungunya transmission in the 21 st century. **Scientific Reports**, v. 7, n. 1, p. 1–11, 2017.

TJADEN, N. B. *et al.* Mosquito-Borne Diseases: Advances in Modelling Climate-Change Impacts. **Trends in Parasitology**, v. 34, n. 3, p. 227–245, 2018.

WILLIAMS, C. R. *et al.* Projections of increased and decreased dengue incidence under climate change. **Epidemiology and Infection**, v. 144, n. 14, p. 3091–3100, 2016.

XU, Z. *et al.* Projecting the future of dengue under climate change scenarios: Progress, uncertainties and research needs. **PLoS Neglected Tropical Diseases**, v. 14, n. 3, p. 1–11, 2020.

CAPITULO 3

Relação entre a incidência de Zika e microcefalia e as condições socioeconômicas, climáticas e sanitárias dos municípios brasileiros

3.1 Introdução

No contexto epidemiológico do Brasil, diversos arbovírus apresentam potencial de disseminação, sendo dengue (DENV), Chikungunya (CHIKV) e Zika (ZIKV) os de maior circulação no país (DONALISIO; FREITAS; ZUBEN, 2017). Dentre os arbovírus citados, os problemas associados ao ZIKV foram mais amplamente discutidos recentemente no Brasil em função da ocorrência, nos anos de 2015 e 2016, de uma epidemia associada a esse vírus.

Os casos de Zika relatados no Brasil, inicialmente, estiveram concentrados na região Nordeste e posteriormente se espalharam pelo país, de modo que, ao final do ano de 2015, haviam sido relatados casos em todas as cinco regiões brasileiras, embora com distribuição desigual entre essas. A ocorrência da epidemia no território nacional teve como agravante a rápida proliferação do vetor em decorrência das condições climáticas propícias para o seu desenvolvimento no país (GARCIA, 2018).

Inicialmente, as manifestações clínicas do Zika foram consideradas brandas. Contudo, um aumento nos casos de microcefalia em recém-nascidos e o surto coincidente de Zika levantou a hipótese da existência de uma associação causal da referida malformação congênita com o vírus Zika. Segundo Wen, Song e Ming (2017), a microcefalia pode ser definida como distúrbio do neurodesenvolvimento, que resulta em uma redução acentuada no tamanho do cérebro e deficiência intelectual em recém-nascidos. A hipótese da relação causal Zika x microcefalia foi, posteriormente, confirmada por meio de avaliações espaço-temporais dos casos relatados, estudos laboratoriais e critérios diagnósticos (CALVET *et al.*, 2016; OLIVEIRA *et al.*, 2016; XIMENES *et al.*, 2023).

De maneira semelhante à verificada no período inicial da epidemia por Zika, os casos de microcefalia concentraram-se na região Nordeste do Brasil. Contudo, mesmo após o aumento da incidência de casos de Zika nas demais regiões brasileiras, não foi observado um aumento proporcional nos casos de microcefalia nessas regiões. Adicionalmente, a incidência de microcefalia na região Nordeste foi muito superior à observada nos demais países onde ocorreram epidemias associadas ao vírus Zika. Tal fato evidenciou a necessidade de

investigação da influência de fatores ambientais, sanitários e socioeconômicos na propagação do vírus Zika e ocorrência de microcefalia (GARCIA, 2018).

Até os tempos recentes, as razões para as diferenças espaciais na incidência de Zika e microcefalia no Brasil não foram completamente elucidadas (DIDERICHSEN; AUGUSTO, 2021). Contudo, a existência de condições socioeconômicas, climáticas e de saneamento díspares entre as regiões brasileiras pode representar um ponto de partida para o entendimento dessas questões.

Segundo Lippi *et al.* (2019), a distribuição atual de mosquitos do gênero *Aedes* é orientada em função das características climáticas de uma região. O aumento nos índices de precipitação pluviométrica pode elevar a população de vetores quando ocorre o acúmulo de água de chuva em recipientes nas residências (ROBERT; STEWART-IBARRA; ESTALLO, 2020). No entanto, condições de seca e escassez hídrica também podem influenciar no aumento da população de vetores, tendo em vista a necessidade de armazenamento doméstico de água nesse período, o que resulta na criação de locais propícios para a reprodução dos mosquitos (AKANDA *et al.*, 2020; DADA *et al.*, 2013). Desse modo, verifica-se que a incidência de eventos climáticos extremos pode ter atuado como um cofator na incidência de Zika em municípios brasileiros. (MAN *et al.*, 2022)

Estudos indicaram que condições apropriadas de saneamento básico podem atuar na prevenção de arboviroses (HAN *et al.*, 2016; SILVA; MACHADO, 2018). Isso ocorre devido à distribuição espacial e densidade vetorial de mosquitos do gênero *Aedes* serem determinadas a partir interação entre condições ambientais, socioeconômicas e meteorológicas (DIDERICHSEN; AUGUSTO, 2021).

Diante das possibilidades descritas, verifica-se que as diferenças na incidência de casos de Zika e microcefalia nas regiões brasileiras podem estar associadas às condições climáticas, socioeconômicas e sanitárias distintas no território nacional. Contudo, ainda não foram desenvolvidos estudos que avaliaram tal relação causal com o uso de informações atualizadas dos municípios brasileiros. Adicionalmente, os estudos que avaliaram a relação entre indicadores socioeconômicos e de saneamento e outras arboviroses no Brasil concentraram-se em regiões específicas do país ou não fizeram uso de métodos estatísticos robustos (ALMEIDA *et al.*, 2020; DIDERICHSEN *et al.*, 2018; RODRIGUES *et al.*, 2018; SOUZA *et al.*, 2018).

Nesse sentido, a utilização de métodos estatísticos multivariados tem se mostrado uma ferramenta útil na explicação de fenômenos complexos, que demandam a avaliação simultânea de múltiplas variáveis. Vários pesquisadores demonstraram a aplicabilidade de tais técnicas em estudos epidemiológicos e ambientais (CARDO *et al.*, 2018; SILVA *et al.*, 2018;

DESJARDINS *et al.*, 2018; MAITRA *et al.*, 2019). Desse modo, os resultados obtidos por meio dessas análises podem subsidiar a tomada de decisão e definição de estratégias para prevenção de epidemias.

Diante do exposto, acredita-se que a utilização de estatística multivariada, aplicada a um banco de dados representativo da população vulnerável brasileira, possa auxiliar no entendimento das relações espaciais associadas à incidência de Zika e microcefalia no Brasil. Considerando-se a problemática associada ao entendimento dessas relações, nesse estudo teve-se como objetivo avaliar a relação causal entre condições socioeconômicas, climáticas e sanitárias e a ocorrência da epidemia do vírus Zika e microcefalia no Brasil.

3.2 Material e métodos

3.2.1 Fonte de dados e seleção de indicadores

O universo da pesquisa foi composto pelos municípios brasileiros aos quais foram notificados casos de Zika e microcefalia, em decorrência da infecção congênita pelo vírus Zika (ZIKV), no ano de 2016. Foram excluídos os casos que, embora tenham cumprido os critérios para notificação, após investigação, não se enquadraram nas definições de confirmado, provável, inconclusivo ou excluído/inativo (BRASIL, 2017).

Os referidos dados, anonimizados, foram obtidos junto à Secretaria de Vigilância em Saúde do Ministério da Saúde por meio do Serviço de Informação ao Cidadão (SIC). Para Zika, os dados brutos continham a listagem de todas as notificações de Zika - presentes no Sistema de Informações de agravo de notificações (SINAN) - ao longo do ano de 2016. Os casos de microcefalia, por sua vez, apresentavam informações extraídas do Painel de Monitoramento de Malformações Congênitas, Deformidades e Anomalias Cromossômicas (D180 e Q00-Q99).

As notificações referentes à doença aguda pelo vírus Zika (agravo A928) foram agrupadas em função do município provável da fonte de infecção (COMUNINF) e da classificação final dos casos notificados (CLASSI_FIN). A partir do número de casos notificados e dos dados de população residente em cada município, obtidos junto ao Instituto Brasileiro de Geografia e Estatística (IBGE), foi calculada a taxa de incidência bruta de Zika, por 100.000 habitantes (Equação 3.1). A taxa de incidência de microcefalia foi calculada a partir do número de casos confirmados de microcefalia a cada 1.000 nascidos vivos, no ano de 2016, registrados no Sistema de Informações sobre Nascidos Vivos (Sinasc) (Equação 3.2).

$$\text{Taxa de incidência de Zika} = \frac{\text{número de casos de Zika em residentes}}{\text{população total residente}} \times 100.000 \quad \text{Equação 3.1}$$

$$\text{Taxa de incidência de microcefalia} = \frac{\text{número de nascidos com microcefalia}}{\text{total de nascidos vivos}} \times 1.000 \quad \text{Equação 3.2}$$

Tendo em vista que o Brasil possui um elevado número de municípios de pequeno porte, foi realizada uma suavização das taxas de incidência. A obtenção de taxas suavizadas foi necessária, uma vez que, taxas brutas podem apresentar elevada instabilidade para medir o risco de um evento quando este é raro ou quando a população da região de ocorrência é pequena (BRASIL, 2007). Desse modo, a incidência de poucos casos de um evento em uma população pequena tem um impacto grande no cálculo das taxas brutas. Para controlar tal instabilidade, foi realizada a suavização matemática das taxas de incidência por meio do método bayesiano empírico local (BRASIL, 2007), assim como realizado em outros estudos epidemiológicos, tais como Bermudi *et al.* (2020), Souza *et al.* (2018), Teixeira *et al.* (2021).

As taxas suavizadas apresentam maior estabilidade em virtude de levarem em conta no seu cálculo a informação sobre a vizinhança associada à área. O estimador bayesiano empírico local trabalha com a média de incidência encontrada na vizinhança do município. Por esse motivo, foi criada uma matriz de vizinhança utilizando o critério de contiguidade, que considera que um objeto está próximo apenas daqueles com os quais compartilha uma fronteira. As taxas bayesianas locais foram calculadas utilizando-se o *software* livre TerraView versão 5.4.3.

A investigação da relação entre a ocorrência dos casos de Zika e microcefalia e as condições socioeconômicas, climáticas e sanitárias nos municípios brasileiros foi realizada por meio da avaliação de indicadores obtidos por meio de dados secundários para o período avaliado. Um resumo dos indicadores utilizados, bem como suas respectivas fontes de consulta, é indicado na Tabela 3.1.

Tabela 3.1. Indicadores sanitários, climáticos, econômicos e sociais utilizados na análise de agrupamentos

Classe indicador	Indicador	Fonte
<i>Indicador sanitário</i>	Percentual de domicílios atendidos por rede geral de distribuição de água	CadÚnico – 2016
	Percentual de domicílios atendidos por rede coletora de esgotos ou pluvial	
	Percentual de domicílios atendidos com coleta direta ou indireta de resíduos sólidos	
<i>Indicador de extremo climático</i>	Incidência cumulativa de eventos extremos de cheias nos municípios brasileiros	ANA/SEDEC – 2003 a 2016
	Incidência cumulativa de eventos extremos de secas nos municípios brasileiros	
<i>Indicador econômico e social</i>	Percentual da população inscrita no CadÚnico	CadÚnico/IBGE – 2016
	Percentual de famílias inscritas no CadÚnico em situação de extrema pobreza ¹	
	Percentual de responsáveis familiares inscritos no CadÚnico com nível de educação classificado como “sem instrução”	CadÚnico – 2016
	Percentual de pessoas inscritas no CadÚnico com cor/raça preta ou parda	
	Localização do domicílio – Relação domicílio urbano/domicílio rural	
Produto interno bruto <i>per capita</i> do município	IBGE – 2016	

Nota: ANA (Agência Nacional de águas); CadÚnico (Cadastro Único para Programas Sociais); IBGE (Instituto Brasileiro de Geografia e Estatística); SEDEC (Secretaria Nacional de Proteção e Defesa Civil).

¹São consideradas famílias em situação de extrema pobreza aquelas que possuem renda mensal *per capita* de até R\$ 89,00 (BRASIL, 2018).

Em relação à fonte de dados socioeconômicos e sanitários, comumente, os estudos que envolvem arboviroses no Brasil utilizam dados do censo demográfico, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Embora tais informações sejam confiáveis e retratem a população brasileira, os dados do último censo disponibilizados pelo IBGE referem-se ao ano de 2010, desse modo, pode haver uma defasagem quanto às informações disponibilizadas. Além disso, os dados do IBGE caracterizam os municípios ou setores censitários do Brasil como um todo, não havendo a possibilidade de realizar recortes para avaliação de uma população específica.

Partindo do princípio que existem fortes indícios de que houve uma desigualdade social na ocorrência da infecção pelo ZIKV (DIDERICHSEN; AUGUSTO, 2021), considerar essa exposição diferencial ao vírus pode auxiliar no entendimento da interação existente entre o Zika e as condições ambientais e socioeconômicas da população. Para isso, é necessário avaliar

informações que retratam a população vulnerável e, potencialmente, mais exposta ao vírus. Nesse sentido, o banco de dados do Cadastro Único para Programas Sociais (CadÚnico) pode ser apontado como um meio eficaz de obtenção de dados que retratam a população mencionada.

O CadÚnico, ligado ao Ministério dos Direitos Humanos e da Cidadania, é um instrumento de coleta de dados para identificação e caracterização das famílias brasileiras que possuem renda per capita inferior a meio salário mínimo ou renda familiar total de até 3 salários mínimos. Famílias com renda superior à citada também podem ser incluídas no CadÚnico, contanto que sua inclusão esteja relacionada à participação em programas sociais ou serviços implementados em nível federal, estadual ou municipal. Tal possibilidade visa o atendimento, em programas sociais, de famílias cuja situação de vulnerabilidade não está necessariamente vinculada à renda. De acordo com Brasil (2017), o CadÚnico conta com informações de cerca de 40% da população brasileira. Por esse motivo, tal banco de dados é considerado um mapa representativo das famílias vulneráveis do país.

Para esse estudo, os dados do CadÚnico referentes ao ano de 2016 foram obtidos junto à Secretaria de Avaliação e Gestão da Informação (SAGI), por meio de consulta ao VIS DATA 3 Beta (BRASIL, 2016). Para tanto, foram selecionados os dados agregados por municípios, a fim de avaliar suas relações com as taxas de incidência de Zika e microcefalia nos municípios brasileiros. Dentre as variáveis disponíveis no banco de dados do CadÚnico foram utilizadas aquelas relativas à infraestrutura domiciliar (forma de abastecimento de água, escoamento sanitário e coleta de resíduos sólidos) e às características socioeconômicas e demográficas de seus moradores (população inscrita no CadÚnico, famílias em situação de extrema pobreza, pessoas sem instrução formal, pessoas inscritas no CadÚnico com cor/raça preta ou parda e localização do domicílio –urbano ou rural).

Os indicadores obtidos a partir do banco de dados do CadÚnico, citados na Tabela 3.1, foram selecionados baseado em sua relação, direta ou indireta, com a transmissão de arboviroses. Para o indicador de acesso à água, por exemplo, o percentual de domicílios que possuem rede geral de distribuição de água foi selecionado baseando-se na definição de criadouros prováveis de mosquitos, contida no Manual de Instruções de combate à Dengue da Fundação Nacional da Saúde (FUNASA) (BRASIL, 2001). No referido manual, o depósito de água em caixas d'água, tanques, poços e cisternas é considerado como possível criadouro ou foco de mosquitos.

O indicador de esgotamento sanitário, por sua vez, foi selecionado baseado nas informações disponíveis em estudos, tais como Beserra *et al.* (2010), Chitolina *et al.* (2016) e MacKay *et al.* (2009), que avaliaram a presença de mosquitos vetores de arboviroses em

esgotos. Nesses estudos, os autores citam a presença de mosquitos do gênero *Aedes* em esgotos brutos dispostos a céu aberto e em fossas sépticas e rudimentares. Desse modo, foi avaliado o percentual de domicílios cujo escoamento sanitário é realizado por meio de redes coletoras, considerando que tal indicador é representativo do cenário que minimiza a proliferação de mosquitos vetores. O indicador de extremo climático, por sua vez, foi definido a partir da avaliação de estudos que apontavam relações entre a elevação da população de vetores de arboviroses e a ocorrência de eventos de seca e escassez hídrica ou de aumento nos índices de precipitação pluviométrica (AKANDA *et al.*, 2020; DADA *et al.*, 2013; ROBERT; STEWART-IBARRA; ESTALLO, 2020)

3.2.2 Análise preliminar dos dados

Após a coleta dos dados descritos na sessão anterior, foi realizada uma análise exploratória a fim de identificar possíveis falhas no banco de dados, tais como dados faltantes, duplicados e ocorrência de valores fora da escala de medida adotada para o indicador. A partir da avaliação do percentual de falhas, foi realizada a exclusão dos dados citados.

Adicionalmente, foi avaliada a normalidade multivariada dos dados por meio dos testes de HenzeZirkler e de Mardia, considerando-se um nível de significância de 5%. Os referidos testes foram realizados no *software* R, utilizando-se o pacote “MNV” (KORKMAZ; GOKSULUK; ZARARSIZ, 2014). Para a detecção de *outliers* nos dados multivariados foi utilizada a distância de Mahalanobis robusta, proposta por Leys *et al.* (2018).

Tendo em vista que os indicadores avaliados no estudo possuem diferentes unidades de medida (dimensões), foi realizada a padronização dos dados com transformação para escala Z (média zero e desvio padrão 1), conforme indicado na Equação 3.3. A adoção desse procedimento é necessária para a realização de análises multivariadas, a fim de evitar a ocorrência de classificações equivocadas em função das diferenças naturais na dimensão dos dados. Além disso, a padronização tende a aumentar a influência de variáveis cuja variância é pequena e reduzir a influência de variáveis cuja variância é grande.

$$Z = \frac{(X-\mu)}{\sigma} \quad \text{Equação 3.3}$$

3.2.3 Distribuição da incidência de Zika e microcefalia no Brasil

Após a obtenção e avaliação preliminar dos dados, foi realizada uma análise exploratória da incidência de Zika e microcefalia nos municípios brasileiros. Para isso, foram elaborados mapas temáticos utilizando-se o *software* ArcGIS 10.4.1. Tais mapas foram construídos a partir das informações relativas às taxas de incidência de Zika e microcefalia nos municípios brasileiros e informações de localização geográfica dos referidos municípios. Os mapas gerados foram utilizados como ferramenta de análise prévia da distribuição dos casos de Zika e microcefalia no Brasil no ano de 2016.

A fim de verificar a existência de associação entre as taxas de incidência de Zika e microcefalia e a localização geográfica de ocorrência dos eventos, foi realizada uma análise exploratória dos dados espaciais. A avaliação da existência de correlação espacial entre os dados foi realizada, inicialmente, por meio da obtenção do índice global de associação espacial, sendo utilizado o Índice de Moran (I), que apresenta uma medida única para toda a área analisada. De forma geral, o Índice de Moran presta-se a um teste cuja hipótese nula é de independência espacial. Para o cálculo do índice I, uma matriz de vizinhança foi construída, adotando-se o critério de contiguidade do tipo “queen” e considerando os vizinhos de primeira ordem. A significância do índice foi avaliada por meio de um teste de pseudo significância, onde foram geradas 999 permutações aleatórias dos valores dos atributos associados a cada região. Vale ressaltar que o pressuposto básico para a aplicação do Índice de Moran é a existência de estacionariedade dos dados de primeira e segunda ordem. Desse modo, caso aplicado em dados não estacionários, o índice perde sua validade.

Conforme anteriormente citado, o índice global de Moran fornece um valor único para toda a área avaliada, o que é fundamental para a caracterização da região como um todo. No entanto, considerando-se o grande número de municípios existentes no Brasil, é provável que existam diferentes regimes de associação espacial e locais onde a dependência espacial é mais pronunciada. Desse modo, após constatada a existência de associação espacial, os padrões dos dados foram examinados, em detalhe, por meio de um índice local de associação espacial, sendo este o índice local de Moran (I_i). O referido índice gera valores específicos para cada área, permitindo assim, a identificação de agrupamentos.

Nessa análise, as taxas de incidência de Zika e microcefalia foram avaliadas individualmente e, em seguida, a auto correlação espacial foi investigada entre as duas variáveis, utilizando-se, para isso, o índice de Moran bivariado. Tal análise bivariada foi realizada com o objetivo de verificar se as taxas de incidência de Zika, observadas em um

determinado município, guardam uma associação com as taxas de incidência de microcefalia verificadas em municípios vizinhos.

O último passo da análise exploratória dos dados espaciais de Zika e microcefalia consistiu na obtenção de mapas do tipo *cluster/outlier* (*Moran map*), sendo estes construídos a partir dos valores determinados pelo Índice local de Moran. Tais mapas ilustram as observações em cada um dos municípios e indicam *clusters e outliers* estatisticamente significativos, considerando-se o nível de significância de 5%. As análises foram realizadas utilizando-se o *software* GeoDa versão 1.18.0.10, sendo obtidos mapas cuja representação espacial apresenta a seguinte tipologia:

- HH: *High-High* (*cluster* significativo de valores altos e próximos - municípios com altas taxas de incidência, cercados de outros municípios também com altas taxas);
- LL: *Low-Low* (*cluster* significativo de valores baixos e próximos - municípios com baixas taxas de incidência, cercados de outros municípios também com baixas taxas);
- HL: *High-Low* (*outlier* de valores altos que não se agrupam, pois se encontram em meio a valores baixos - municípios com altas taxas de incidência, cercados por municípios com baixas taxas);
- LH: *Low-High* (*outlier* de valores baixos que não se agrupam, pois se encontram em meio a valores altos - municípios com baixas taxas de incidência, cercados por municípios com altas taxas);
- Não significativos: municípios que não se enquadram nos agrupamentos, pois apresentam níveis variados, assim como os valores dos vizinhos.

3.2.4 Incidência de Zika e microcefalia em municípios com notificação de eventos climáticos extremos

Uma investigação da existência de associações entre as taxas de incidência de Zika e microcefalia e a ocorrência de eventos climáticos extremos ocorridos nos municípios brasileiros foi conduzida. Os dados sobre a incidência de eventos extremos nos municípios brasileiros, no período de 2003 a 2016, foram obtidos a partir dos registros da Secretaria Nacional de Proteção e Defesa Civil (SEDEC), vinculada ao Ministério do Desenvolvimento Regional e disponíveis através do Sistema Integrado de Informações sobre Desastres - S2ID. Tais registros são baseados nos decretos de declaração de Situação de Emergência (SE) e de Estado de Calamidade Pública (ECP) expedidos pelos municípios, que são posteriormente homologados

pelos estados e reconhecidos pela União. Foram avaliados registros relacionados a eventos de cheias (inundações, alagamentos e enxurradas) e secas (estiagens e secas). A definição dos referidos desastres, de acordo com a Codificação Brasileira de Desastres (COBRADE), é indicada na Tabela 3.2.

Os municípios que apresentaram taxas de incidência de Zika e microcefalia superiores a 0 foram divididos em grupos em função dos eventos extremos relatados à SEDEC, conforme indicado na Tabela 3.3.

Após a separação dos dados de Zika e microcefalia, em função dos diferentes eventos extremos notificados, foram realizados testes de comparação entre os grupos formados (A e B). Nessa etapa, foi investigado se os valores de Zika ou microcefalia de um grupo A são superiores aos valores de um grupo B.

Tabela 3.2. Definição dos desastres naturais relacionados a eventos de cheias e secas

Desastre	Definição	Referência¹
Inundação	Submersão de áreas fora dos limites normais de um curso de água em zonas que normalmente não se encontram submersas. O transbordamento ocorre de modo gradual, geralmente ocasionado por chuvas prolongadas em áreas de planície.	1.2.1.0.0
Alagamento	Extrapolação da capacidade de escoamento de sistemas de drenagem urbana e consequente acúmulo de água em ruas, calçadas ou outras infraestruturas urbanas, em decorrência de precipitações intensas.	1.2.3.0.0
Enxurrada	Escoamento superficial de alta velocidade e energia, provocado por chuvas intensas e concentradas, normalmente em pequenas bacias de relevo acidentado. Caracterizada pela elevação súbita das vazões de determinada drenagem e transbordamento brusco da calha fluvial. Apresenta grande poder destrutivo.	1.2.2.0.0
Estiagem	Período prolongado de baixa ou nenhuma pluviosidade, em que a perda de umidade do solo é superior à sua reposição.	1.4.1.1.0
Seca	A seca é uma estiagem prolongada, durante o período de tempo suficiente para que a falta de precipitação provoque grave desequilíbrio hidrológico.	1.4.1.2.0

Fonte: Classificação e Codificação Brasileira de Desastres (COBRADE)

¹Número de referência do desastre tipificado no COBRADE.

Tabela 3.3. Organização dos dados para aplicação do teste U de Mann-Whitney.

Agravo	Grupos de municípios
Zika	A - Municípios com notificação de eventos extremos de cheias
	B - Municípios sem notificação de eventos extremos de cheias
Microcefalia	A - Municípios com notificação de eventos extremos de cheias
	B - Municípios sem notificação de eventos extremos de cheias
Zika	A - Municípios com notificação de eventos extremos de secas
	B - Municípios sem notificação de eventos extremos de secas
Microcefalia	A - Municípios com notificação de eventos extremos de secas
	B - Municípios sem notificação de eventos extremos de secas

A verificação da existência de diferenças estatísticas entre os dados avaliados foi realizada por meio da utilização do teste U de Mann-Whitney, no *software* R, sendo adotado o nível de significância de 5% na aplicação do teste. O teste U de Mann-Whitney não possui pressupostos sobre a distribuição dos dados e é utilizado para testar se dois grupos independentes foram ou não extraídos da mesma população. Esse teste é, portanto, uma alternativa não paramétrica para o teste t de Student para amostras independentes. De maneira distinta ao teste t, que testa a igualdade das médias, o teste U de Mann-Whitney avalia a igualdade das medianas de dois grupos. Vale ressaltar que, para a utilização do teste de Mann-Whitney, é necessário que as amostras avaliadas sejam aleatórias e independentes entre si.

3.2.5 Relação entre Zika, microcefalia e indicadores sanitários e socioeconômicos

A relação entre a incidência de Zika e microcefalia e as características sanitárias e socioeconômicas dos municípios brasileiros foi investigada por meio da realização de uma análise de agrupamentos ou *clusters* (AC) e posterior aplicação de testes estatísticos para identificação de diferenças entre as taxas de incidência de Zika e microcefalia nos grupos formados. A análise de clusters foi utilizada para agrupar os municípios brasileiros em que foram notificados casos de Zika e microcefalia, levando em consideração os indicadores sanitários, econômicos e sociais, descritos na Tabela 3.1, associados a cada um deles. A análise teve como fundamento a divisão dos municípios brasileiros em grupos e posterior interpretação dos grupos formados. Os dados referentes a cada um dos indicadores descritos foram utilizados após a realização dos procedimentos de análise preliminar descritos no tópico 3.2.2 (análise preliminar dos dados).

Para a realização da análise de agrupamentos foi utilizada uma combinação da abordagem hierárquica seguida de um método não-hierárquico, assim como descrito por Chen *et al.* (2005). Hair *et al.* (2009) afirmam que a combinação de ambos os métodos permite extrair os benefícios e minimizar as desvantagens associadas a cada abordagem. Segundo Hair *et al.* (2009), embora técnicas hierárquicas sejam comumente utilizadas, algumas desvantagens associadas a elas podem afetar as soluções de agrupamento encontradas. Dentre as desvantagens citadas pelos autores, destacam-se: formação de combinações iniciais indesejáveis que podem persistir na análise e conduzir a resultados artificiais, impacto substancial de observações atípicas e limitação para análise de amostras muito grandes.

De modo geral, embora o método hierárquico de agrupamento facilite uma avaliação abrangente de uma vasta gama de soluções, tais soluções podem ser afetadas por uma característica do método: uma vez que os agrupamentos são unidos, eles não são separados no processo. Nesse sentido, os métodos não-hierárquicos possuem a vantagem de serem capazes de otimizar as soluções pela redesignação de observações até que seja obtida uma heterogeneidade mínima dentro dos conglomerados. Contudo, os benefícios de qualquer método não-hierárquico são percebidos apenas com o emprego de pontos sementes não-aleatórios (ou seja, especificados) (CHEN *et al.*, 2005; HAIR *et al.*, 2009)

Diante do exposto, o procedimento de análise consistiu em duas etapas: i) um método hierárquico foi utilizado para selecionar o número de agrupamentos e para caracterizar os centros de agrupamento que serviram como sementes iniciais no procedimento não-hierárquico; ii) um método não-hierárquico agregou as observações utilizando os centros de grupos dos resultados hierárquicos como os pontos sementes iniciais para fornecer alocações mais precisas.

Para a realização da análise de agrupamento hierárquico foi adotado o método aglomerativo, utilizando-se a distância euclidiana e a distância euclidiana quadrática como medidas de dissimilaridade. A fim de comparar o desempenho de diferentes algoritmos aglomerativos na classificação dos municípios avaliados, foram testados os resultados obtidos por meio dos algoritmos *single linkage*, *complete linkage*, *average linkage*, *Ward's method* e *centroid method* (BOEHMKE; GREENWELL, 2020).

A seleção do algoritmo mais apropriado para o agrupamento dos municípios se deu por meio da comparação dos dendrogramas gerados pelos diferentes algoritmos. Inicialmente, foi realizada uma inspeção visual dos dendrogramas, visando identificar aqueles com melhor adequação ao fenômeno estudado. A fim de auxiliar na escolha do algoritmo a ser utilizado, foram calculados os coeficientes de correlação cofenética para os resultados obtidos a partir de cada um dos algoritmos avaliados. Tal correlação é uma medida de validação utilizada,

frequentemente, em métodos de agrupamento hierárquicos (SARAÇLI; DOĞAN; DOĞAN, 2013).

O coeficiente de correlação cofenética mensura o grau de semelhança entre a matriz de distâncias do dendrograma elaborado (matriz cofenética) e a matriz de distâncias originalmente observadas entre os objetos (matriz de dissimilaridade). Os algoritmos que proporcionam a obtenção de coeficientes com valores mais próximos de 1 indicam menor distorção provocada pelo agrupamento dos indivíduos (SARAÇLI; DOĞAN; DOĞAN, 2013). Desse modo, o algoritmo selecionado foi aquele que apresentou o maior coeficiente de correlação cofenética. A significância do coeficiente de correlação cofenética foi calculada por meio do Teste de Mantel, considerando-se 999 permutações e nível de significância de 5% na aplicação do teste.

Após a seleção do algoritmo aglomerativo, foi definido o número de grupos (*clusters*) a serem utilizados. Para auxiliar na determinação do número de grupos, que definiram a partição do conjunto de dados analisados, foram avaliados os índices disponíveis no pacote *NbClust* (CHARRAD *et al.*, 2014). O pacote *NbClust* contém 30 índices capazes de estimar o número ótimo de *clusters* para um determinado conjunto de dados. A utilização do referido pacote permite o cálculo simultâneo de todos os índices aplicáveis ao conjunto de dados, considerando o algoritmo de agrupamento selecionado. Cada um dos índices realiza agrupamentos com o conjunto de dados, variando o número de grupos formados, e, posteriormente, analisa a qualidade dos agrupamentos e define o número ideal de *clusters* baseado em algum indicador de dissimilaridade. Após o cálculo de todos os índices, o número ideal de *clusters* é escolhido de acordo com a regra da maioria, ou seja, o número apontado pela maioria dos índices avaliados. Mais informações sobre os índices utilizados estão disponíveis em Charrad *et al.* (2014).

Após a definição do número de agrupamentos, por meio do método hierárquico, foram calculados os centroides de cada agrupamento, sendo estes, posteriormente, utilizados como pontos sementes para a execução do algoritmo não hierárquico de agrupamento *K-means*.

Visando identificar se os indicadores sanitários, econômicos e sociais, utilizados na AC, apresentavam diferenças significativas entre os grupos formados, foi utilizado o teste estatístico não paramétrico de Mann-Whitney, com nível de significância (α) igual a 5%. Por fim, o mesmo teste foi aplicado para avaliar a existência de diferenças significativas entre as incidências de Zika e microcefalia nos agrupamentos de municípios obtidos na AC. Por meio da aplicação dos referidos testes foi verificado se os casos notificados de Zika e microcefalia ocorreram, predominantemente, em regiões com menores índices socioeconômicos e de saneamento no país.

Um fluxograma das etapas realizadas para a avaliação da existência de relação entre a incidência de Zika e microcefalia e indicadores sanitários e socioeconômicos é apresentado na Figura 3.1. Todas as análises foram realizadas utilizando o *software* livre R.

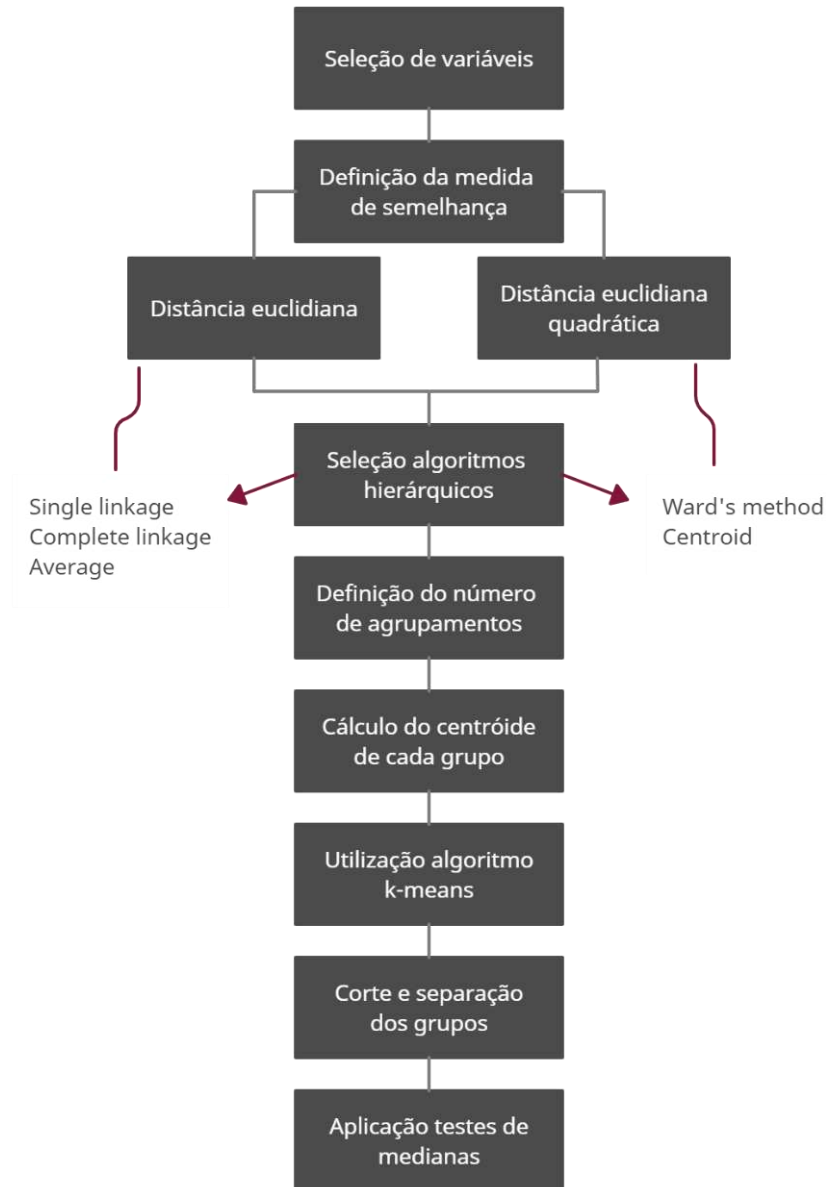


Figura 3.1. Fluxograma das etapas realizadas ao longo das análises de agrupamentos.

3.3 Resultados e discussão

3.3.1 Distribuição da incidência de Zika e microcefalia no Brasil

Os dados levantados nesse estudo indicaram que, no ano de 2016, cerca de 40% dos municípios brasileiros registraram casos de infecção pelo vírus Zika. A síndrome congênita associada à infecção pelo vírus Zika (microcefalia) em recém-nascidos, por sua vez, foi notificada em 15% dos municípios do país. Foi verificado que, no ano de 2016, houve a notificação de 281.464 casos de Zika no Brasil. Destes, 219.266 se enquadraram nas definições de classificação de caso confirmado, provável ou inconclusivo. Em relação à incidência de microcefalia, o painel de monitoramento de malformações congênitas, deformidades e anomalias cromossômicas identificou 2.276 casos no ano de 2016.

O estudo da distribuição espacial da incidência de Zika e microcefalia no Brasil buscou verificar a existência de associação entre tais taxas de incidência e a localização geográfica de ocorrência dos eventos. Por meio das análises realizadas, foi possível identificar a existência de agrupamentos espaciais para Zika e microcefalia, sendo obtidos valores do Índice de Moran (I) de 0,313 e 0,442, respectivamente. Ambos valores apresentaram significância estatística, considerando-se 999 permutações aleatórias (pseudo p-valor = 0,001).

A representação espacial da análise dos padrões de *cluster* e *outliers* das taxas de incidência de Zika e microcefalia é apresentada na Figura 3.2. Para Zika, é possível observar que aglomerados de altas taxas estão localizados, sobretudo, nas regiões Centro-Oeste e Nordeste do Brasil. Padrão semelhante foi identificado por Vissoci *et al.* (2018), que identificaram que a incidência dos casos de Zika, no início da epidemia, esteve concentrada na região nordeste e, em seguida, teve o seu padrão espacial alterado, deslocando-se para a região Centro-Oeste.

Embora um padrão semelhante tenha sido observado para as taxas de incidência de microcefalia, é possível verificar que uma maior concentração de aglomerados de altas taxas é verificada nos municípios pertencentes à região nordeste do país. De acordo com Barbeito-Andrés, Schuler-Faccini e Garcez (2018), nessa região, a contagem de microcefalia a cada 10.000 gestantes infectadas pelo ZIKV foi de 1.300 casos. A região Sudeste, por sua vez, para o mesmo número de gestantes infectadas, registou 145 casos de microcefalia. Tais resultados reforçam a assimetria verificada na incidência de microcefalia no Brasil e ressaltam a necessidade de investigação de outros fatores que podem ter atuado, em conjunto com o vírus ZIKV, na maior incidência dessa malformação congênita na região Nordeste.

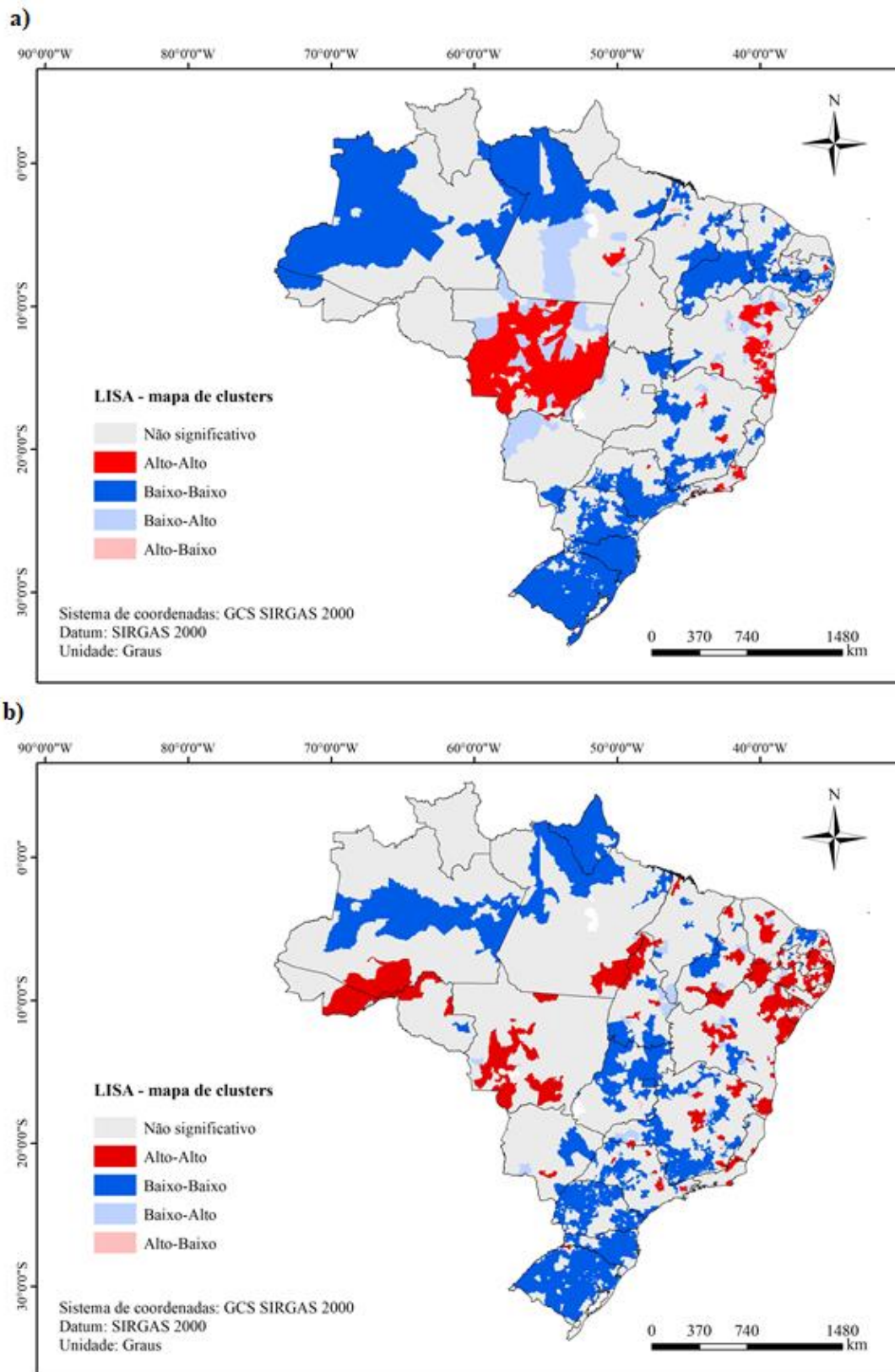


Figura 3.2. Análise *cluster* e *outlier* das taxas de incidência de a) Zika e b) microcefalia nos municípios brasileiros.

Mapas coropléticos foram criados utilizando uma categorização de três quantis, que visou simular uma divisão dos municípios em grupos de baixas, médias e altas taxas de incidência de Zika e microcefalia. Na Figura 3.3, o mapa bivariado da distribuição espacial conjunta de Zika e microcefalia pode ser visualizado. Por meio do mapa, é possível verificar que os municípios que apresentaram distribuição conjunta de altas taxas de microcefalia e baixas taxas de Zika estão localizados, majoritariamente, na região Nordeste. Em termos percentuais, os municípios nordestinos representam cerca de 75% do total de municípios na categoria citada. Corroborando com tais resultados, a análise bivariada de Moran Local indicou a existência de um maior número de aglomerados de municípios com baixas taxas de Zika cercados por municípios com altas taxas de microcefalia na região Nordeste do país (Figura 3.4).

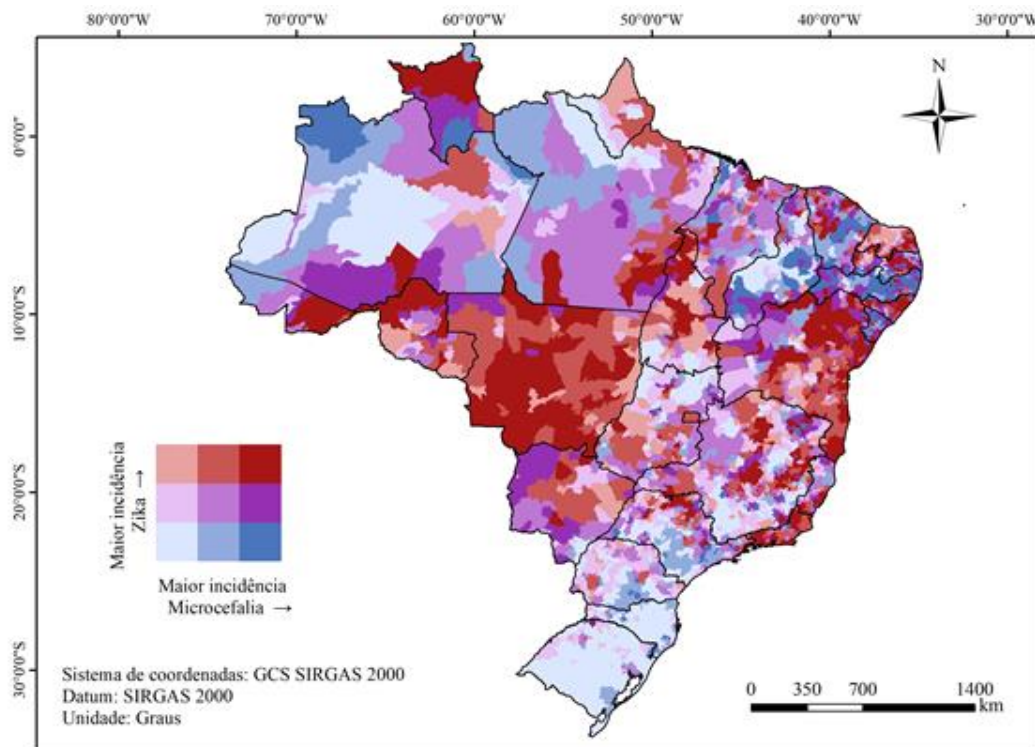


Figura 3.3. Mapas coropléticos bivariados para as taxas de incidência de Zika e microcefalia no Brasil.

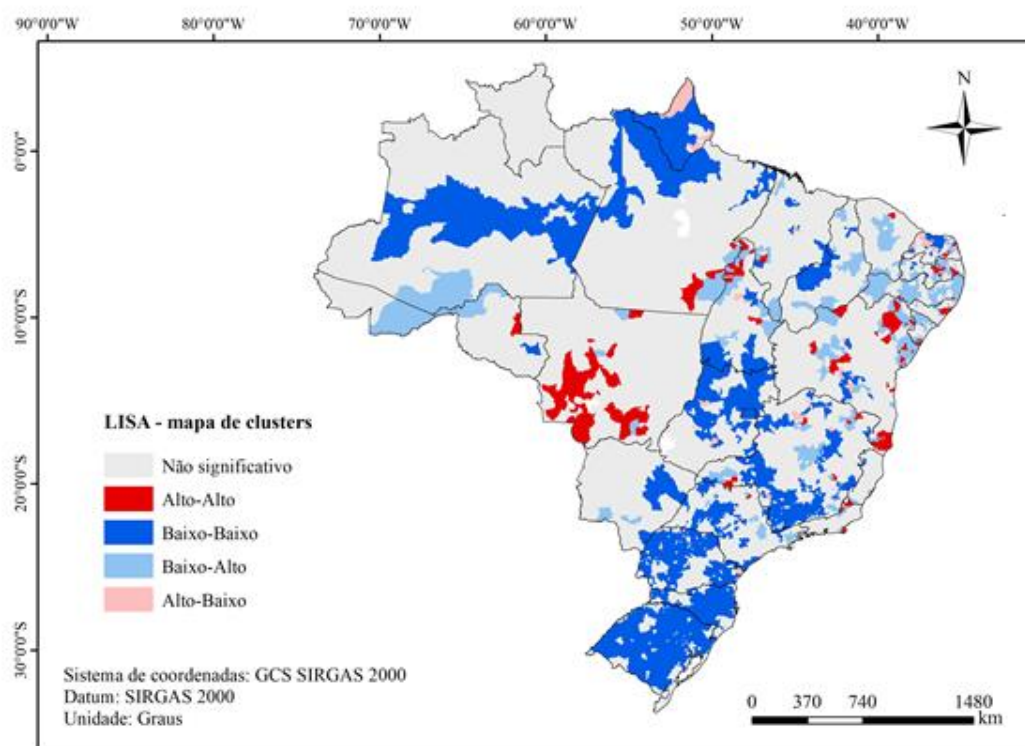


Figura 3.4. Análise Bivariada Local de Moran I. indicando a relação espacial entre os padrões de disseminação de Zika e microcefalia nos municípios brasileiros.

A partir dos resultados obtidos na análise espacial exploratória, é possível observar que, de maneira semelhante à verificada no período inicial da epidemia por Zika, os casos de microcefalia concentraram-se na região Nordeste do Brasil. Contudo, mesmo após o aumento da incidência de casos de Zika nas demais regiões brasileiras, não foi observado um aumento proporcional nos casos de microcefalia nessas regiões.

Segundo Vissoci *et al.* (2018) a utilização do padrão espacial de incidência do vírus Zika auxilia no melhor entendimento de áreas de risco de microcefalia e fornece subsídios para a elaboração de políticas de vigilância. No entanto, os resultados obtidos no presente estudo sugerem que, apesar de haver indícios de uma dependência espacial entre Zika e microcefalia, a utilização única do padrão espacial de incidência de Zika pode não ser suficiente para identificar as áreas de risco para microcefalia. Esses resultados corroboram com o estudo de Amaral *et al.* (2019). Os autores verificaram que a incidência de microcefalia varia, consideravelmente, entre as regiões brasileiras quando correlacionada apenas com o ZIKV, ou seja, somente a presença do vírus não é capaz de explicar as diferenças regionais no padrão de distribuição da microcefalia. Assim como citado por Garcia (2018), os resultados mencionados

evidenciam a necessidade de investigação da influência de fatores ambientais, sanitários e socioeconômicos na propagação do vírus Zika e ocorrência de microcefalia no Brasil.

3.3.2 Incidência de Zika e microcefalia em municípios com notificação de eventos climáticos extremos

Os testes estatísticos realizados nesse estudo indicaram diferenças significativas ($\alpha = 5\%$) nas taxas de incidência de Zika e microcefalia entre os grupos de municípios que possuíam e não possuíam notificações de secas e estiagens registradas. Conforme indicado na Tabela 3.4, os municípios que apresentaram eventos de secas ou estiagens possuíam as maiores taxas de incidência de Zika e microcefalia. Tais municípios estão localizados, majoritariamente, na região Nordeste do Brasil. Nessa região, entre os anos de 2003 e 2016, a SEDEC registrou 16.076 notificações de Situação de Emergência (SE) e de Estado de Calamidade Pública (ECP) expedidos pelos municípios, em virtude da ocorrência de secas e estiagens. Os resultados obtidos podem estar associados ao fato de que a ocorrência de secas e estiagens resulta na adoção de práticas de armazenamento improvisado de água em barris e demais recipientes domésticos, que também atuam como locais de proliferação de vetores (ALMEIDA; COTA; RODRIGUES, 2020).

Tabela 3.4. Resultados do teste de Mann-Whitney (U).

Agravo	Grupos de municípios	Mediana	W	p-valor
Zika	Com notificação de cheias	8,75	44742	0.0398
	Sem notificação de cheias	23,96		
Microcefalia	Com notificação de cheias	1,05	5032	0.1191
	Sem notificação de cheias	1,43		
Zika	Com notificação de secas	31,06	509373	$9,87 \times 10^{-6}$
	Sem notificação de secas	21,13		
Microcefalia	Com notificação de secas	1,73	84646	$1,59 \times 10^{-6}$
	Sem notificação de secas	1,31		

De acordo com Eakin *et al.* (2014) e Marengo Orsini *et al.* (2018), no nordeste do Brasil, secas e estiagens afetam, principalmente, a população vulnerável da região semiárida, o que resulta em eventos de déficit hídrico e seus consequentes impactos na segurança hídrica, energética e alimentar. Por esse motivo, Souza *et al.* (2018) apontam que é comum na região a construção de estruturas, ou a utilização de recipientes, para armazenamento de água. Embora

tais medidas tenham como objetivo a ampliação da capacidade da região de se adaptar aos eventos climáticos extremos, Lowe *et al.* (2018) e Meason e Paterson (2014) citam que estas podem ter como consequência o aumento do risco de doenças arbovirais. Isso ocorre em virtude das referidas estruturas e recipientes, quando em mau estado de conservação ou indevidamente fechadas, poderem se tornar habitat larval para mosquitos do gênero *Aedes*.

Além disso, Lowe *et al.* (2018) citam que o aumento do uso de recipientes para armazenamento de água nos períodos de seca pode aumentar o risco de transmissão de arboviroses nos meses posteriores, quando é iniciada a estação de chuvas e, por esse motivo, diminui-se a necessidade de armazenamento doméstico de água. Em tais situações, os recipientes de armazenamento de água não são usados e mantidos regularmente mas servem, apenas, como uma fonte reserva de água, o que aumenta o risco de transmissão de arboviroses. Diante disso, alguns autores têm destacado que, dependendo das práticas locais de armazenamento de água e características de habitação, a seca pode aumentar, potencialmente, a disponibilidade de habitat larval (ALMEIDA; ARAÚJO; LÍDIA, 2019; LOWE *et al.*, 2018, 2021; IBARRA *et al.*, 2013).

Lowe *et al.* (2021), ao avaliarem os efeitos de eventos hidrometeorológicos extremos na transmissão de dengue, verificaram que o risco associado às secas extremas foi acentuado em áreas demasiadamente urbanizadas e que apresentaram maior frequência de escassez hídrica ou abastecimento de água intermitente durante a estação seca. Ainda segundo os autores, no período de 3 a 5 meses posteriores à ocorrência de secas extremas no Brasil, houve uma elevação do risco de dengue no país. Diante disso, verifica-se que áreas urbanas de maior densidade populacional e que possuem abastecimento de água intermitente durante períodos de escassez hídrica são mais susceptíveis a surtos de dengue após os períodos de seca. Acredita-se que os resultados apresentados podem ser extrapolados para os eventos de incidência de Zika no Brasil, tendo em vista que ambas arboviroses são transmitidas pelo mesmo vetor, cujas condições para desenvolvimento são intensificadas em períodos de seca.

Conforme anteriormente mencionado, a região Nordeste do país apresentou o maior número de notificações de eventos climáticos de secas e estiagens no ano de 2016. Ao mesmo tempo, tal região apresentou a maior incidência de microcefalia no Brasil, ainda que outras regiões do país tenham notificado maior taxa de incidência de Zika. Dentre os municípios que apresentaram casos de microcefalia e registraram tais eventos, cerca de 91% estão localizados na região Nordeste. Diante disso, pressupõem-se que a incidência de eventos climáticos extremos também pode ter atuado como um cofator na incidência de microcefalia em municípios brasileiros.

Além de reduzir os reservatórios a volumes críticos e resultar em deficiência no suprimento de água, os eventos de seca e estiagem favorecem a proliferação de cianobactérias capazes de produzir neurotoxinas com implicações para a saúde (BRASIL *et al.*, 2016). Tal fato é relevante no contexto da incidência de Zika e microcefalia, tendo em vista que um estudo conduzido por Pedrosa *et al.* (2020) indicou que o consumo de água contaminada com cianobactérias produtoras de saxitoxina pode ter atuado como um cofator da incidência de microcefalia no Brasil.

Segundo Pedrosa *et al.* (2020), o aumento da concentração de cianotoxinas nos reservatórios de abastecimento está, também, associado à ocorrência de eventos de seca e estiagem no Brasil, que acarretam a redução dos reservatórios a volumes críticos e favorecem a proliferação de cianobactérias. Adicionalmente, Oliver e Ribeiro (2020) apontam que o aumento da temperatura atmosférica, a diminuição de nebulosidade em períodos de secas e estiagens e a concentração de nutrientes (advindos de efluentes não tratados), em virtude do menor volume de água, resultam em maior floração das cianobactérias em reservatórios de água e, conseqüentemente, elevação da concentração das toxinas por elas produzidas.

De acordo com Pedrosa *et al.* (2020) os dados de vigilância da qualidade de água indicaram um aumento da população de cianobactérias e uma maior concentração de saxitoxina nos reservatórios de água para abastecimento humano na região nordeste, em comparação com outras regiões do Brasil. Desse modo, a grave seca enfrentada na região Nordeste, que resultou na redução dos níveis dos reservatórios, aliada ao baixo tratamento de esgoto (apenas 33,7% do esgoto gerado na região é tratado, segundo dados do SNIS) em diversas áreas da região Nordeste podem ter auxiliado na proliferação das cianobactérias. Adicionalmente, Oliver e Ribeiro (2020) destacam que há grande similaridade nas condições que favorecem a proliferação de mosquitos do gênero *Aedes* e de florações de cianobactérias produtoras das toxinas que podem estar associadas à síndrome congênita do Zika. Assim, além dos fatores associados à proliferação do *Aedes aegypti*, a qualidade da água de abastecimento pode ter influenciado na predominância de microcefalia na região Nordeste do Brasil.

Em relação à ocorrência de eventos climáticos extremos de cheias, não foram verificadas diferenças significativas ($\alpha = 5\%$) nas taxas de incidência de microcefalia entre os municípios que notificaram ou não a ocorrência dos referidos eventos (Tabela 3.4). Por outro lado, avaliando-se as taxas de incidência de Zika nos mesmos municípios, foram identificadas diferenças significativas, tendo os municípios que apresentaram eventos de cheia a menor mediana. De acordo com Souza *et al.* (2018), ainda que as arboviroses estejam estreitamente

associadas à ocorrência de precipitação, tais doenças são complexas e demandam uma avaliação holística de fatores climáticos e não climáticos para a compreensão de sua dinâmica.

A ocorrência de eventos climáticos extremos, tais como secas, estiagens e chuvas intensas, pode desempenhar um papel importante na propagação de arboviroses. De acordo com Meason e Paterson (2014), o aumento da intensidade de chuvas em um local acarreta o acúmulo de água em recipientes e, conseqüentemente, eleva a oferta de criadouros para oviposição de mosquitos. Adicionalmente, os alagamentos, decorrentes de precipitações intensas e extrapolção da capacidade de escoamento de sistemas de drenagem urbana, favorecem a proliferação de mosquitos vetores de arboviroses em virtude do acúmulo de água em ruas, calçadas ou outras infraestruturas urbanas.

Contudo, segundo Ibarra *et al.* (2013), a avaliação do efeito das chuvas na incidência de arboviroses é mais complexa do que para eventos de secas. A ocorrência de chuvas pode aumentar a disponibilidade de habitats larvais, em recipientes que permanecem cheios de água de chuva, e, conseqüentemente, aumentar a densidade de mosquitos (LOWE *et al.*, 2021). No entanto, quando as chuvas ocorrem em excesso, pode haver a diminuição da densidade de mosquitos, pois o excesso de chuva pode resultar na eliminação das larvas nos recipientes, lavando-as para áreas distantes dos locais de reprodução e, ainda, reduzindo a sobrevivência dos adultos (SEIDAHMED; ELTAHIR, 2016).

Seidahmed e Eltahir (2016) e Chien e Yu (2014) verificaram que chuvas fortes podem, no curto prazo, diminuir temporariamente o risco de arboviroses, o que ocorre em função da descarga de recipientes de água que estão a céu aberto. Tais informações corroboram com os resultados observados nesse estudo, tendo em vista que a avaliação foi realizada considerando a ocorrência de eventos climáticos extremos de chuva, que resultaram na necessidade de declaração de estado de calamidade ou emergência pelos municípios. A avaliação da relação entre a incidência de Zika e microcefalia e a ocorrência de eventos climáticos extremos aqui realizada pode contribuir no entendimento das associações entre as previsões climáticas e desfechos de risco de arboviroses.

3.3.3 Relação entre Zika, microcefalia e indicadores sanitários e socioeconômicos

A existência de diferenças expressivas nas incidências de Zika e microcefalia nas regiões do Brasil evidenciaram a necessidade de investigação da influência de fatores ambientais e socioeconômicos na propagação do vírus Zika no país. Nesse estudo, a relação

entre os fatores citados foi investigada por meio da realização de uma análise de agrupamentos (AC) e posterior aplicação de testes estatísticos.

Tendo em vista que os resultados de alguns métodos hierárquicos de agrupamento podem ser afetados pela presença de *outliers* no banco de dados, uma análise preliminar, em busca de tais valores atípicos, foi conduzida. O método de detecção de *outlier* foi selecionado após a averiguação da distribuição multivariada dos dados. Por meio da aplicação dos testes de HenzeZirkler e de Mardia foi verificado que os bancos de dados de Zika e microcefalia não possuíam distribuição normal multivariada, tendo ambos apresentado valor de $p < 0,001$. Diante disso, a identificação dos *outliers* nos bancos de dados foi realizada utilizando-se a distância de Mahalanobis robusta, uma vez que, a distância de Mahalanobis clássica assume que os conjuntos de dados possuem distribuição normal. Desse modo, caso esta seja aplicada em dados que seguem uma distribuição atípica, com assimetria moderada ou forte, poderá gerar resultados equivocados (COUSINEAU; CHARTIER, 2010). Uma representação gráfica da análise de *outliers* realizada é apresentada na Figura 3.5.

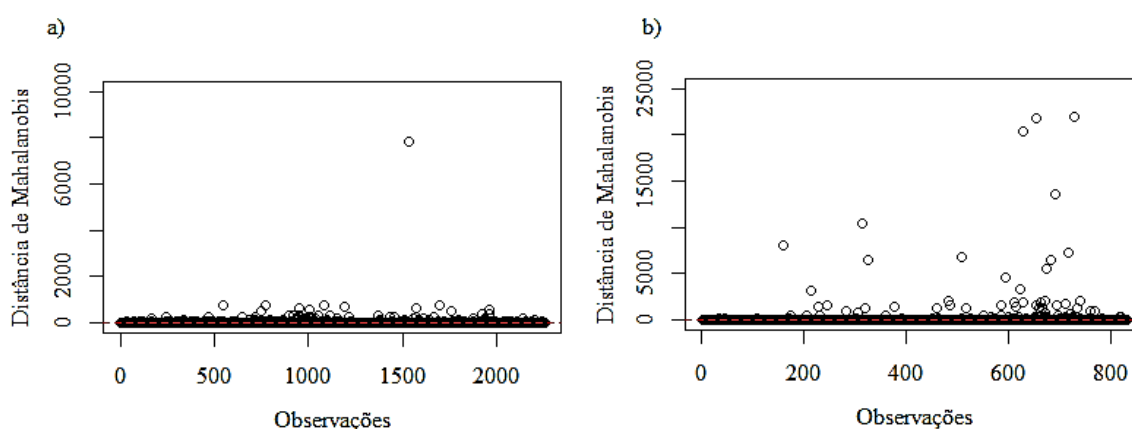


Figura 3.5. Identificação da presença de *outliers* no banco de dados de (a) Zika e (b) microcefalia por meio da distância robusta de Mahalanobis.

Nota: A linha tracejada (em vermelho) indica a distância limite, acima da qual, as observações são consideradas *outliers*.

Para o banco de dados referente aos municípios que apresentaram casos de Zika, foram identificados 401 *outliers*, o que corresponde a cerca de 7,2% dos municípios brasileiros. Os dados dos municípios que apresentaram casos de microcefalia, por sua vez, apresentaram 187 *outliers*, valor este referente a 3,3% dos municípios do Brasil. Dentre as observações (municípios) identificadas como *outliers*, foi verificada a presença de indicadores com valores fora da escala de medida adotada ou com valores espúrios, por exemplo, municípios que apresentavam população rural superior à população total. A vista disso, e considerando-se que

os valores atípicos identificados decorriam de erros de procedimento, tais como erros na entrada de dados, falhas na codificação ou erros na coleta dos dados, os *outliers* identificados no estágio de limpeza dos dados foram eliminados. O procedimento citado segue as recomendações de Hair *et al.* (2009) para *outliers* classe 1.

Com base na matriz de dissimilaridade, obtida a partir dos dados padronizados, foram gerados os dendrogramas para os cinco algoritmos aglomerativos avaliados. De modo geral, os dendrogramas obtidos a partir dos algoritmos *Ward* e *complete linkage* apresentaram uma melhor separação entre os municípios, indicando a existência de dois grupos distintos. Embora os dendrogramas citados tenham apresentado diferenças nos níveis em que os municípios foram agrupados, a estrutura geral dos agrupamentos mostrou-se bastante similar. Desse modo, para a seleção do algoritmo a ser utilizado para a obtenção dos *clusters*, os coeficientes de correlação cofenética foram avaliados (Tabela 3.5).

Tabela 3.5. Coeficientes de correlação cofenética obtidos para os métodos *Ward* e *complete linkage*.

Algoritmo de agrupamento		Coeficiente de correlação cofenética	Teste de Mantel (p-valor) ¹
Zika	<i>Complete linkage</i>	0,62	<0,001
	<i>Ward's</i>	0,66	<0,001
Microcefalia	<i>Complete linkage</i>	0,61	<0,001
	<i>Ward's</i>	0,64	<0,001

Nota: Valor de p obtido pelo teste de Mantel com 999 permutações aleatórias.

O teste de Mantel indicou que as correlações existentes entre a matriz de distância original e a matriz cofenética, oriunda dos dois algoritmos avaliados, foram significativas (considerando-se $\alpha = 5\%$). Tendo em vista que o algoritmo *Ward* apresentou os maiores coeficientes de correlação cofenética, tanto para os dados de Zika quanto para os de microcefalia, este foi selecionado como método para separação dos *clusters*. Tal algoritmo também foi identificado por Ferreira e Hitchcock (2009) como o que apresentou o melhor desempenho, dentre os outros métodos hierárquicos.

Os dendrogramas obtidos por meio do método de *Ward* são indicados na Figura 3.6. Por meio dos índices disponíveis no pacote NbClust (CHARRAD *et al.*, 2014) foi estimado como 2 o número ótimo de *clusters* para os conjuntos de dados, de acordo com a regra da maioria (Figura 3.7). A partição dos conjuntos de dados analisados em dois grupos corrobora com os objetivos da análise realizada, que visa separar os municípios em função dos seus índices

sanitários e socioeconômicos. Desse modo, os dois grupos formados podem ser categorizados como municípios com altos e baixos índices, respectivamente.

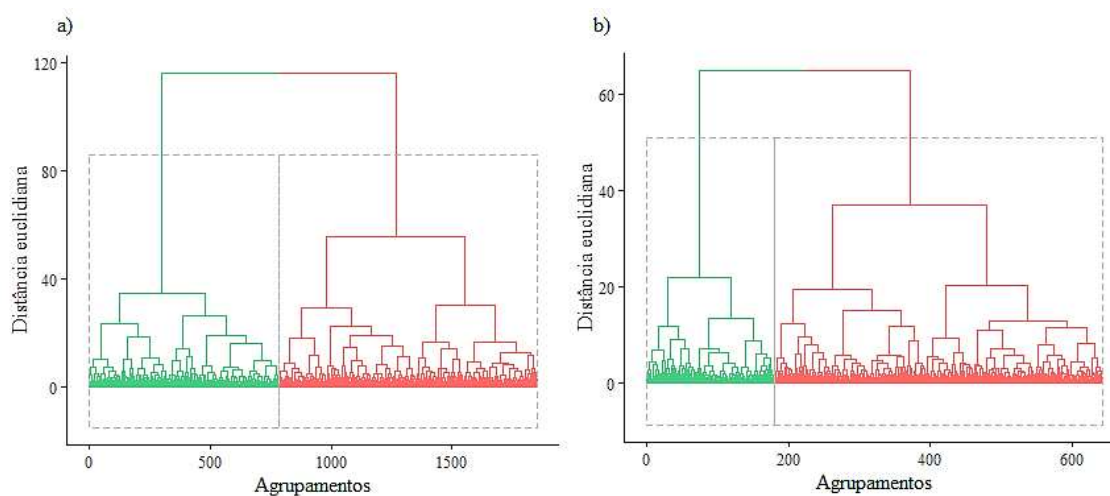


Figura 3.6. Dendrogramas obtidos a partir do método de Ward para o banco de dados de a) Zika e b) microcefalia.

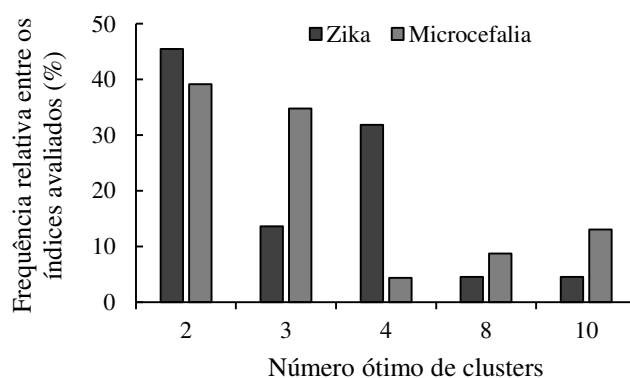


Figura 3.7. Número ótimo de clusters para os bancos de dados avaliados segundo índices disponíveis no pacote NbClust.

Após a definição dos grupos por meio do método hierárquico, foram calculados os centroides de cada agrupamento e estes foram utilizados como pontos sementes de partida na análise não hierárquica realizada pelo algoritmo *k-means*. Os resultados obtidos com a aplicação do referido algoritmo podem ser observados na Figura 3.8.

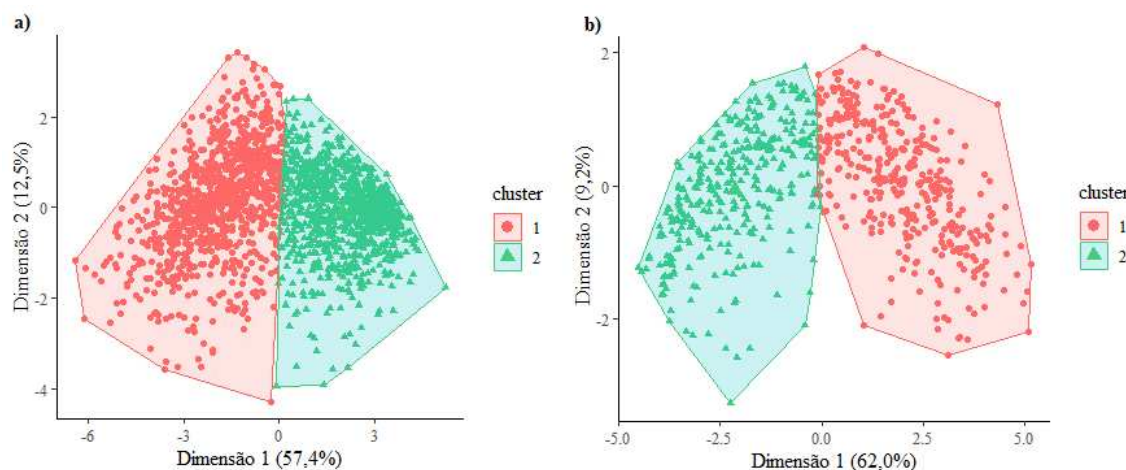


Figura 3.8. Resultados de saída aplicando o algoritmo *k-means* aos dados de a) Zika e b) microcefalia.

A partir da avaliação dos indicadores associados a cada um dos municípios pertencentes aos grupos formados, por meio do algoritmo *k-means*, foi verificado que todos os indicadores apresentaram diferenças significativas entre os grupos 1 e 2 (valor de $p < 2,2 \times 10^{-16}$). Conforme indicado na Tabela 3.6, os municípios do grupo 2 caracterizam-se por possuir os melhores índices em termos sanitários (maiores percentuais de atendimento com coleta de lixo, rede coletora de esgotos e rede geral de distribuição de água) e econômicos (menor percentual da população inscrita no CadÚnico, em situação de extrema pobreza, com responsáveis familiares sem instrução e maiores PIBs per capita).

Tabela 3.6. Caracterização dos agrupamentos de municípios que apresentaram casos de Zika e microcefalia quanto aos indicadores sanitários, econômicos e sociais.

Indicador ¹	Zika ²		Microcefalia ³	
	Mediana G1	Mediana G2	Mediana G1	Mediana G2
Percentual de domicílios atendidos por rede geral de distribuição de água	65,4	89,8	65,3	90,7
Percentual de domicílios atendidos por rede coletora de esgotos	9,0	72,4	11,9	78,1
Percentual de domicílios atendidos com coleta direta ou indireta de resíduos sólidos	66,1	94,8	67,4	96,1
Percentual da população inscrita no CadÚnico	67,2	34,4	67,3	31,6
Percentual de famílias em situação de extrema pobreza	64,9	25,2	66,9	29,4
Percentual de responsáveis familiares sem instrução	22,6	17,9	23,3	18,4
Percentual de pessoas com cor/raça preta ou parda	83,3	53,5	85,3	62,7
Localização do domicílio – Relação domicílio urbano/domicílio rural	0,6	0,1	0,6	0,1
Produto interno bruto <i>per capita</i> do município	9.392,40	24.047,10	9.428,80	25.871,70

Nota: ¹Todos os indicadores apresentaram diferenças significativas entre os grupos 1 e 2 pelo teste de Mann-Whitney ($p < 2,2 \times 10^{-16}$). ²Número de elementos no grupo 1 (G1): 941 e no grupo 2 (G2): 910. ³Número de elementos no grupo 1 (G1): 360 e no grupo 2 (G2): 330.

Segundo Aduña (2023), a falta de acesso aos serviços de saneamento básico (abastecimento de água por rede geral, esgotamento sanitário por rede coletora e coleta domiciliar direta ou indireta de lixo) está associado à pobreza monetária. Ao se analisar a situação sanitária dos domicílios no Brasil, é possível identificar que, quanto menor a renda agregada dos moradores, maior é o déficit no atendimento aos serviços de saneamento, o que corrobora com os resultados obtidos na análise de agrupamentos realizada nesse estudo. Ainda, de acordo com o IBGE (2020), a ausência de serviços públicos de saneamento, associada à pobreza monetária, resulta em uma vulnerabilidade maior, uma vez que, a adoção de soluções individuais de saneamento, tais como poços artesianos e tanques sépticos, implicam em custos financeiros que, por vezes, não são suportados pela população.

A análise de agrupamentos realizada também revelou que os municípios pertencentes ao grupo 2 apresentam menor percentual de pessoas autodeclaradas como pretas ou pardas. Tal população apresenta maiores níveis de vulnerabilidade econômica, o que reflete no menor

acesso aos serviços de saneamento básico. Essa relação foi evidenciada em estudos anteriores, tais como Jesus (2020), Gasteyer *et al.* (2016) e Leker e Gibson (2018).

Em relação à localização dos municípios pertencentes aos agrupamentos 1 e 2, por meio dos mapas indicados nas Figuras 3.9 e 3.10, é possível verificar que os municípios do grupo 2 localizam-se, predominantemente, nas regiões Sul e Sudeste do Brasil. Por outro lado, os municípios com os menores índices sanitários e econômicos estão situados no Norte e Nordeste do país. Os melhores índices de saneamento verificado nas regiões Sul e Sudeste do país podem estar associados ao maior nível de desenvolvimento econômico das referidas regiões, o que é evidenciado pela presença dos maiores PIB *per capita* do país e possibilita um maior investimento no setor de saneamento (MESQUITA; ROSA; BORGES, 2022).

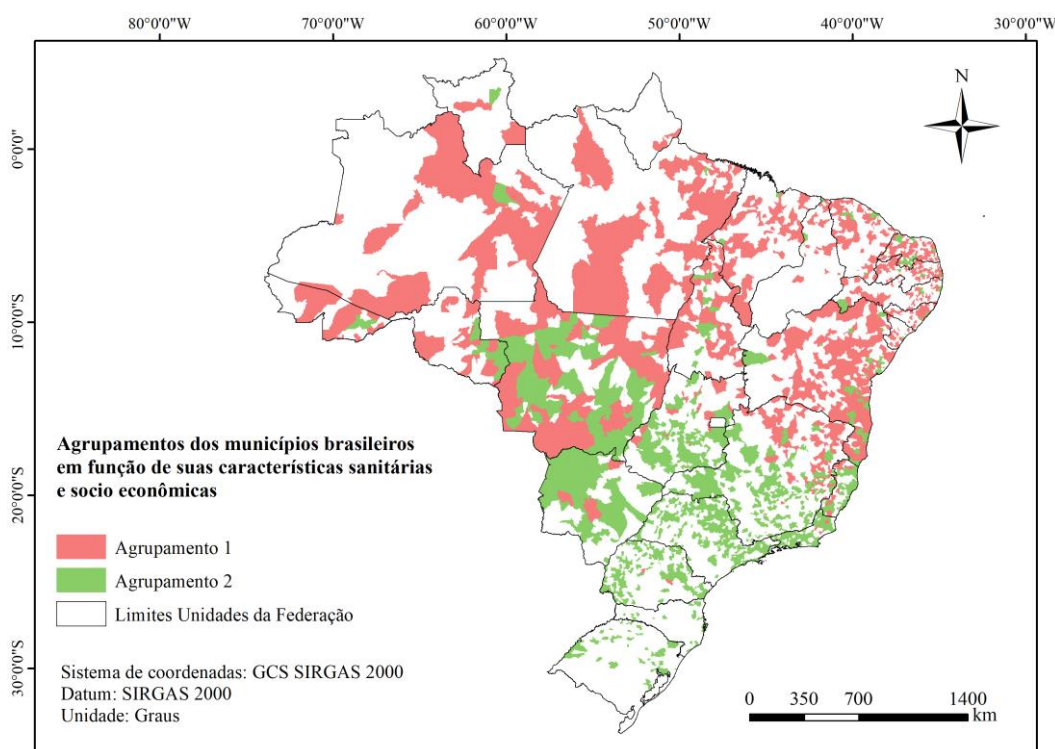


Figura 3.9. Distribuição dos agrupamentos dos municípios brasileiros que apresentaram casos de Zika em função dos de suas características sanitárias e socioeconômicas.

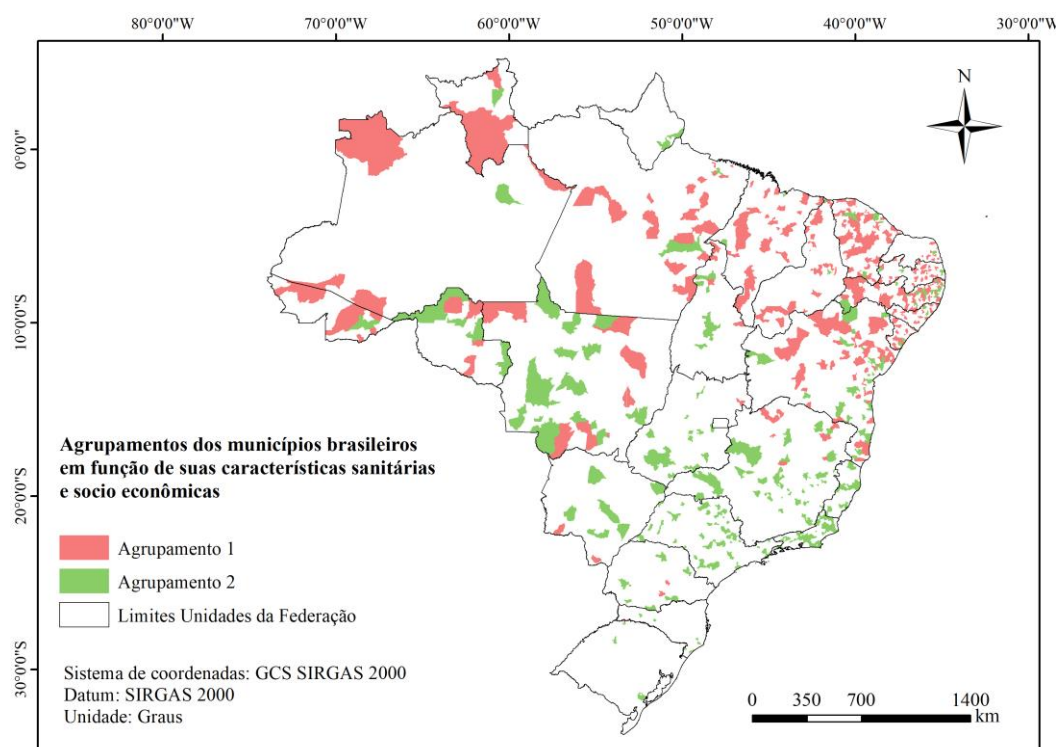


Figura 3.10. Distribuição dos agrupamentos dos municípios brasileiros que apresentaram casos de microcefalia em função dos de suas características sanitárias e socioeconômicas.

Após a avaliação das características sanitárias, econômicas e sociais dos agrupamentos de municípios formados, foi investigado se as taxas de incidência de Zika e microcefalia nos municípios pertencentes a cada um dos grupos formados também apresentavam diferenças significativas entre si. Por meio da aplicação do teste de Mann Whitney foi verificado que os casos notificados de Zika e microcefalia ocorreram, predominantemente, nos municípios pertencentes ao grupo 1, ou seja, em regiões com menores índices socioeconômicos e de saneamento no país (Tabela 3.7).

Tabela 3.7. Caracterização dos agrupamentos de municípios que apresentaram casos de Zika e microcefalia quanto às taxas de incidência dos referidos agravos.

	Grupo	n	Média	Mediana	W ¹	p-valor
Zika	1	941	215,9	43,9	545538	< 2,2x10 ⁻¹⁶
	2	910	124,2	13,4		
Microcefalia	1	360	2,4	1,4	75145	< 2,2x10 ⁻¹⁶
	2	330	1,8	1,0		

Nota: ¹Estatística W do teste de Mann-Whitney.

Os resultados obtidos corroboram com a hipótese desse estudo que avalia se as condições sanitárias, sociais e econômicas dos municípios brasileiros podem ter atuado como cofatores na propagação do vírus Zika e incidência de microcefalia no país. Tais fatores podem ser denominados como fatores indiretos na incidência de arboviroses, que se referem àqueles que não afetam diretamente o número e o ciclo de vida dos mosquitos vetores, mas que podem influenciar na distribuição destes e, conseqüentemente, na incidência de Zika e arboviroses em geral (SIRIYASATIEN *et al.*, 2018).

Ali *et al.* (2017) destacam que a pobreza pode criar condições ideais para a transmissão de doenças. Isso ocorre em função da falta de recursos econômicos afetar o acesso da população aos serviços de saneamento, educação e cuidados de saúde. Segundo os autores, a associação existente entre a pobreza e a falta de acesso aos serviços citados levam à existência de correlações negativas entre renda e o risco de Zika. Os autores verificaram que, em todos os estados brasileiros, o PIB *per capita* esteve negativamente correlacionado aos casos confirmados de microcefalia em decorrência do vírus Zika. Tal relação também foi observada no presente estudo, que apontou que o grupo de municípios que possui o menor valor médio de PIB *per capita* apresentou as maiores taxas de incidência de microcefalia no ano 2016.

O estudo desenvolvido por Ali *et al.* (2017) também salientou que a ausência de acesso da população aos serviços públicos de saneamento pode aumentar o contato humano-mosquito e, desse modo, influenciar a exposição humana ao vírus Zika. Para casos de dengue, Siriyasatien *et al.* (2018) constataram que a maioria dos surtos ocorrem em comunidades cujas moradias não possuem infraestrutura dos serviços de saneamento ou onde a população está desnutrida. Esta última situação implica menor imunidade e maior risco de doenças infecciosas. Embora, nesse estudo, dados de desnutrição não tenham sido investigados, o grupo que apresentou maiores taxas de incidência de Zika e microcefalia contém os municípios brasileiros com os maiores percentuais de extrema pobreza no país e, conforme destacado por Siddiqui *et al.* (2020), tais fatores (pobreza *versus* desnutrição) estão intrinsecamente associados.

Aguiar *et al.* (2018) destacam que a forma de destinação dos resíduos sólidos, do abastecimento de água e o tipo de instalação sanitária dos domicílios desempenham um papel importante na expansão de arboviroses, tais como Zika e Chikungunya. A relação entre a incidência de dengue e as condições de esgotamento sanitário também foi investigada por Rodrigues *et al.* (2016). Assim como Siriyasatien *et al.* (2018), Rodrigues *et al.* (2016) relataram haver uma relação inversa entre a o atendimento com redes de esgoto e o risco de dengue.

Embora comumente seja citado que fêmeas de *Aedes aegypti* são atraídas para recipientes que contêm água com baixos índices de poluição, estudos têm demonstrado que quanto maior o grau de poluição da água, maior a atratividade da mesma ao mosquito. De acordo com Chitolina *et al.* (2016), a presença de coliformes termotolerantes em esgotos é capaz de aumentar a atratividade para a oviposição do *Aedes aegypti*. Adicionalmente, Beserra *et al.* (2010) verificaram grande capacidade de desenvolvimento do *Aedes aegypti* em águas com elevados níveis de poluição, tais como esgoto bruto e efluentes de reator UASB, lagoa de polimento e filtro anaeróbio, confirmando a tolerância a poluentes para a referida espécie de mosquito. Segundo os mesmos autores, a existência de altos níveis de turbidez e maior concentração de nutrientes em esgotos são fatores que contribuem na atração para oviposição do mosquito. Além disso, um estudo desenvolvido por MacKay *et al.* (2009) identificou a presença de *Aedes aegypti* em fossas sépticas. Os autores verificaram uma maior taxa de reprodução e emergência de mosquitos de maior tamanho nesse ambiente do que em recipientes de superfície que continham água limpa. Os resultados do referido estudo despertam preocupação, sobretudo, no Brasil, onde grande parcela da população utiliza fossas sépticas e fossas rudimentares como meio de tratamento ou afastamento de esgotos (IBGE, 2010).

A falta de acesso à rede de distribuição de água também é apontada como um fator capaz de influenciar a incidência de doenças transmitidas por mosquitos. Isso ocorre em função da necessidade de armazenamento doméstico de água, o que resulta na criação de locais propícios para a reprodução dos mosquitos vetores (DADA *et al.*, 2013). Além da falta de acesso à rede de distribuição de água, a escassez hídrica contribui fortemente para o aumento do armazenamento doméstico de água. Conforme discutido anteriormente no presente estudo, no nordeste do Brasil, os frequentes períodos de escassez hídrica resultam na utilização de recipientes para armazenamento de água, o que pode contribuir para a proliferação de mosquitos vetores.

Além da forma de abastecimento de água e do tipo de instalação sanitária, a existência de serviços públicos de coleta de resíduos também foi relacionada à existência de criadouros para mosquitos do gênero *Aedes* em áreas urbanas (AGUIAR *et al.*, 2018). Em estudo desenvolvido por MacCormack-Gelles *et al.* (2018), dentre os fatores relacionados ao acesso aos serviços públicos, a existência de coleta regular de resíduos sólidos esteve, consistentemente, associada à incidência de dengue. Os referidos autores destacaram que tal associação é mais pronunciada em períodos de baixa precipitação, onde locais com acúmulo de resíduos sólidos podem promover a preservação dos ovos dos mosquitos, que eclodem após a incidência de chuvas intermitentes.

Vissoci *et al.* (2018) destacam que o acesso simultâneo aos serviços públicos de coleta, direta ou indireta, de resíduos sólidos e de abastecimento de água e esgotamento sanitário por rede geral ainda apresentam disparidades entre as regiões brasileiras. Nesse cenário, municípios que não possuem serviços apropriados de saneamento básico podem estar sujeitos a um maior risco de infestação por mosquitos do gênero *Aedes* e, conseqüentemente, maior risco de ocorrência de surtos de arboviroses. As diferenças entre as taxas de incidência de Zika e microcefalia nos grupos de municípios com os melhores e piores índices de saneamento observadas nesse estudo corroboram com as informações citadas por Vissoci *et al.* (2018).

3.4 Considerações finais

Nesse estudo, foi verificado que as taxas de incidência de Zika e microcefalia nos municípios brasileiros estão associadas à localização geográfica de ocorrência dos eventos. As maiores taxas de incidência ocorreram, majoritariamente, nos agrupamentos formados por municípios que apresentam menores índices socioeconômicos e de saneamento no país. Tal fato sugere que a elevada variabilidade espacial da incidência de Zika e microcefalia no Brasil pode ser explicada a partir da avaliação das diferentes características sanitárias, socioeconômicas e ambientais dos municípios.

Além disso, foi verificado que existem diferenças entre as taxas de incidência de Zika e microcefalia nos municípios que possuem histórico de ocorrência de eventos climáticos extremos, sendo os municípios que possuem eventos de seca aqueles mais atingidos pela epidemia do vírus Zika. Desse modo, é possível que, além dos fatores sanitários e econômicos, a ocorrência de eventos climáticos extremos pode desempenhar um papel importante na propagação de arboviroses.

Diante do exposto, acredita-se que os resultados obtidos nesse estudo, ainda que possuam caráter exploratório, podem fornecer informações valiosas para a tomada de decisão em saúde pública. O planejamento antecipado de ações para áreas com potencial risco de desenvolvimento de doenças é fundamental para a formulação de programas bem-sucedidos de saúde pública e alocação de recursos para prevenção e controle da doença em áreas específicas. Por fim, os resultados obtidos nesse estudo demonstram a necessidade de atenção aos serviços de saneamento básico como estratégia para redução das populações de vetores e, conseqüentemente, das doenças a eles associadas.

3.5 Referências

- ADUGNA, D. Challenges of sanitation in developing countries - Evidenced from a study of fourteen towns, Ethiopia. **Heliyon**, v. 9, n. 1, p. e12932, 2023.
- AGUIAR, B. S. *et al.* Potential risks of Zika and chikungunya outbreaks in Brazil: A modeling study. **International Journal of Infectious Diseases**, v. 70, p. 20–29, 2018.
- AKANDA, A. S. *et al.* Prioritizing Water Security in the Management of Vector-Borne Diseases: Lessons From Oaxaca, Mexico. **GeoHealth**, v. 4, n. 3, p. e2019GH000201, 2020.
- ALI, S. *et al.* Environmental and Social Change Drive the Explosive Emergence of Zika Virus in the Americas. **PLOS Neglected Tropical Diseases**, v. 11, n. 2, p. e0005135, 2017.
- ALMEIDA, L. S.; ARAÚJO, L. A. DE; LÍDIA, A. Confins Distribuição dos casos de infecção por vírus Zika vírus (ZIKV) na Região Metropolitana de Maceió. **Revista franco-brasileira de geografia**, n. 41, 2019.
- ALMEIDA, L. S.; COTA, A. L. S.; RODRIGUES, D. F. Saneamento, Arboviroses e Determinantes Ambientais: impactos na saúde urbana. **Ciência & Saúde Coletiva**, v. 25, n. 10, p. 3857–3868, 2020.
- AMARAL, P. *et al.* Geospatial modeling of microcephaly and zika virus spread patterns in Brazil. **PLOS ONE**, v. 14, n. 9, p. e0222668, 2019.
- BARBEITO-ANDRÉS, J.; SCHULER-FACCINI, L.; GARCEZ, P. P. Why is congenital Zika syndrome asymmetrically distributed among human populations? **PLOS Biology**, v. 16, n. 8, p. e2006592, 2018.
- BERMUDI, P. M. M. *et al.* Padrão espacial da mortalidade por câncer de mama e colo do útero na cidade de São Paulo. **Revista de Saúde Pública**, v. 54, p. 142, 2020.
- BESERRA, E. B. *et al.* Efeito da qualidade da água no ciclo de vida e na atração para oviposição de *Aedes aegypti* (L.) (Diptera: Culicidae). **Neotropical Entomology**, v. 39, n. 6, p. 1016–1023, 2010.
- BOEHMKE, B.; GREENWELL, B. **Hands-on Machine Learning with R**. 1. ed. CRC press, 2020.
- BRASIL. **Introdução à Estatística Espacial para a Saúde Pública**. Brasília: Ministério da Saúde, 2007.
- BRASIL. **Orientações integradas de vigilância e atenção à saúde no âmbito da emergência de saúde pública de importância nacional**. Brasília: Ministério da Saúde, 2007.
- CALVET, G. *et al.* Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study. **The Lancet. Infectious diseases**, v. 16, n. 6, p. 653–660, 2016.
- CARDO, M. V. *et al.* Heterogeneous distribution of *Culex pipiens*, *Culex quinquefasciatus* and their hybrids along the urbanisation gradient. **Acta Tropica**, v. 178, p. 229–235, 2018.

CHARRAD, M. *et al.* Nbclust: An R package for determining the relevant number of clusters in a data set. **Journal of Statistical Software**, v. 61, n. 6, 2014.

CHEN, B. *et al.* Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis. **IEEE Computational Systems Bioinformatics Conference, Workshops and Poster Abstracts**, p. 105–108, 2005.

CHIEN, L. C.; YU, H. L. Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence. **Environment International**, v. 73, p. 46–56, 2014.

CHITOLINA, R. F. *et al.* Raw sewage as breeding site to *Aedes* (*Stegomyia*) *aegypti* (Diptera, culicidae). **Acta Tropica**, v. 164, p. 290–296, 2016.

COUSINEAU, D.; CHARTIER, S. Outliers detection and treatment: a review. **International Journal of Psychological Research**, v. 3, n. 1, p. 58–67, 2010.

DADA, N. *et al.* Relationship between *Aedes aegypti* production and occurrence of *Escherichia coli* in domestic water storage containers in rural and sub-urban villages in Thailand and Laos. **Acta Tropica**, v. 126, n. 3, p. 177–185, 2013.

DESJARDINS, M. R. *et al.* Space-time clusters and co-occurrence of chikungunya and dengue fever in Colombia from 2015 to 2016. **Acta Tropica**, v. 185, p. 77–85, 2018.

DIDERICHSEN, F.; AUGUSTO, L. G. DA S.; PEREZ, B. Understanding social inequalities in Zika infection and its consequences: A model of pathways and policy entry-points. **lobal Public Health**, v. 14, n. 5, p. 675–683, 2018.

DIDERICHSEN, F.; SILVA, A. L. G. Inequality in Zika virus and congenital Zika syndrome. **Zika Virus Biology, Transmission, and Pathways: Volume 1: The Neuroscience of Zika Virus**, p. 43–54, 2021.

DONALISIO, M. R.; FREITAS, A. R. R.; ZUBEN, A. P. B. VON. Arboviroses emergentes no Brasil: desafios para a clínica e implicações para a saúde pública. **Revista de Saúde Pública**, v. 51, 2017.

EAKIN, H. C.; LEMOS, M. C.; NELSON, D. R. Differentiating capacities as a means to sustainable climate change adaptation. **Global Environmental Change**, v. 27, n. 1, p. 1–8, 2014.

FERREIRA, L.; HITCHCOCK, D. B. A Comparison of Hierarchical Methods for Clustering Functional Data. **Communications in Statistics - Simulation and Computation**, v. 38, n. 9, p. 1925–1949, 2009.

GARCIA, L. P. **Epidemia do vírus Zika e microcefalia no Brasil: emergência, evolução e enfrentamento**. Rio de Janeiro: Ipea, 2018.

GASTEYER, S. P. *et al.* BASICS INEQUALITY: Race and Access to Complete Plumbing Facilities in the United States. **Du Bois Review: Social Science Research on Race**, v. 13, n. 2, p. 305–325, 2016.

HAIR, J. J. F.; BLACK, W. C.; SANT'ANNA, A. S. **Análise multivariada de dados**. 6 ed, 2009.

HAN, M. *et al.* Novel integrated systems for controlling and prevention of mosquito-borne diseases caused by poor sanitation and improper water management. **Journal of Environmental Chemical Engineering**, v. 4, n. 4, p. 3718–3723, 2016.

IBGE. **Síntese de indicadores sociais : uma análise das condições de vida da população brasileira : 2020**. Rio de Janeiro: IBGE, 2020.

JESUS, V. DE. Racializando o olhar (sociológico) sobre a saúde ambiental em saneamento da população negra: um continuum colonial chamado racismo ambiental. **Saúde e Sociedade**, v. 29, n. 2, p. e180519, 2020.

KORKMAZ, S.; GOKSULUK, D.; ZARARSIZ, G. MVN: An R package for assessing multivariate normality. **R Journal**, v. 6, n. 2, p. 151–162, 2014.

LEKER, H. G.; GIBSON, J. M. D. Relationship between race and community water and sewer service in North Carolina, USA. **PLOS ONE**, v. 13, n. 3, p. e0193225, 2018.

LEYS, C. *et al.* Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. **Journal of Experimental Social Psychology**, v. 74, n. March 2017, p. 150–156, 2018.

LIPPI, C. A. *et al.* Geographic shifts in *Aedes aegypti* habitat suitability in Ecuador using larval surveillance data and ecological niche modeling: Implications of climate change for public health vector control. **PLoS Neglected Tropical Diseases**, v. 13, n. 4, p. 1–19, 2019.

LOWE, R. *et al.* Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. **PLoS Medicine**, v. 15, n. 7, p. 1–24, 2018.

LOWE, R. *et al.* Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. **The Lancet Planetary Health**, v. 5, n. 4, p. e209–e219, 2021.

MACCORMACK-GELLES, B. *et al.* Epidemiological characteristics and determinants of dengue transmission during epidemic and non-epidemic years in Fortaleza, Brazil: 2011-2015. **PLOS Neglected Tropical Diseases**, v. 12, n. 12, p. e0006990, 2018.

MACKAY, A. J. *et al.* Dynamics of *Aedes aegypti* and *Culex quinquefasciatus* in Septic Tanks. **Journal of the American Mosquito Control Association**, v. 25, n. 4, p. 409–416, 2009.

MAITRA, A. *et al.* Exploring deeper genetic structures: *Aedes aegypti* in Brazil. **Acta Tropica**, v. 195, p. 68–77, 2019.

MAN, O. M. *et al.* Re-emergence of arbovirus diseases in the State of Rio de Janeiro, Brazil: The role of simultaneous viral circulation between 2014 and 2019. **One Health**, v. 15, p. 100427, 2022.

MARENGO ORSINI, J. A. *et al.* Climatic characteristics of the 2010-2016 drought in the semiarid northeast Brazil region. **Anais da Academia Brasileira de Ciências**, v. 90, n. 2, p. 1973–1985, 2018.

MEASON, B.; PATERSON, R. Chikungunya, Climate Change, and Human Rights . **Health**

and **Human Rights Journal**, v. 16, n. 1, p. 105–112, 2014a.

MESQUITA, T. C. R.; ROSA, A. P.; BORGES, A. C.. Panorama da utilização de tanques sépticos e filtros anaeróbios no tratamento descentralizado de esgotos no Brasil. **Revista DAE**, v. 70, n. 236, p. 34–51, 2022.

OLIVEIRA MELO, A. S. *et al.* Zika virus intrauterine infection causes fetal brain abnormality and microcephaly: tip of the iceberg? **Ultrasound in Obstetrics & Gynecology**, v. 47, n. 1, p. 6–7, 2016.

OLIVER, S. L.; RIBEIRO, H. Zika virus syndrome, lack of environmental policies and risks of worsening by cyanobacteria proliferation in a climate change scenario. **Revista de saude publica**, v. 54, p. 83, 2020.

PEDROSA, C. S. G. *et al.* The cyanobacterial saxitoxin exacerbates neural cell death and brain malformations induced by zika virus. **PLoS Neglected Tropical Diseases**, v. 14, n. 3, p. 1–13, 2020.

ROBERT, M. A.; STEWART-IBARRA, A. M.; ESTALLO, E. L. Climate change and viral emergence: evidence from Aedes-borne arboviruses. **Current Opinion in Virology**, v. 40, p. 41–47, 2020.

RODRIGUES, N. C. P. *et al.* Temporal and Spatial Evolution of Dengue Incidence in Brazil, 2001-2012. **PLOS ONE**, v. 11, n. 11, p. e0165945, 2016.

RODRIGUES, N. C. P. *et al.* Risk factors for arbovirus infections in a low-income community of Rio de Janeiro, Brazil, 2015-2016. **PLOS ONE**, v. 13, n. 6, p. e0198357, 2018.

SARAÇLI, S.; DOĞAN, N.; DOĞAN, I. Comparison of hierarchical cluster analysis methods by cophenetic correlation. **Journal of Inequalities and Applications**, v. 2013, n. 1, p. 1–8, 2013.

SEIDAHMED, O. M. E.; ELTAHIR, E. A. B. A Sequence of Flushing and Drying of Breeding Habitats of *Aedes aegypti* (L .) Prior to the Low Dengue Season in Singapore. **PLoS Neglected Tropical Diseases**. p. 1–13, 2016.

SIDDIQUI, F. *et al.* The Intertwined Relationship Between Malnutrition and Poverty. **Frontiers in Public Health**, v. 8, p. 453, 2020.

SILVA, J. C. B.; MACHADO, C. J. S. Associations between dengue and socio-environmental variables in capitals of the Brazilian northeast by cluster analysis. **Ambiente & Sociedade**, v. 21, p. 1332, 2018.

SILVA, N. S. DA *et al.* Clinical, laboratory, and demographic determinants of hospitalization due to dengue in 7613 patients: A retrospective study based on hierarchical models. **Acta Tropica**, v. 177, p. 25–31, 2018.

SIRIYASATIEN, P. *et al.* Dengue Epidemics Prediction : A Survey of the State-of-the-Art Based on Data Science Processes. **IEEE Access**, v. 6, p. 53757–53795, 2018.

SOUZA, A. I. *et al.* Geography of Microcephaly in the Zika Era: A Study of Newborn Distribution and Socio-environmental Indicators in Recife, Brazil, 2015-2016. **Public Health**

Reports, v. 133, n. 4, p. 461, 2018.

SOUZA, C. D. F. DE *et al.* Estudo espacial da hanseníase na Bahia, 2001-2012: abordagem a partir do modelo bayesiano empírico local. **Epidemiologia e Serviços de Saúde**, v. 27, n. 4, p. e2017479, 2018.

SOUZA, M. L. A. *et al.* Bayesian estimates for the mapping of dengue hotspots and estimation of the risk of disease epidemic in Northeast Brazil. **Urban Climate**, v. 26, n. September, p. 198–211, 2018.

STEWART IBARRA, A. M. *et al.* Dengue vector dynamics (*Aedes aegypti*) influenced by climate and social factors in ecuador: Implications for targeted control. **PLoS ONE**, v. 8, n. 11, 2013.

TEIXEIRA, R. A. *et al.* Mortality due to garbage codes in brazilian municipalities: Differences in rate estimates by the direct and bayesian methods from 2015 to 2017. **Revista Brasileira de Epidemiologia**, v. 24, 2021.

VISSOCI, J. R. N. *et al.* Zika virus infection and microcephaly: Evidence regarding geospatial associations. **PLOS Neglected Tropical Diseases**, v. 12, n. 4, p. e0006392, 2018.

WEN, Z.; SONG, H.; MING, G. L. How does Zika virus cause microcephaly? **Genes & Development**, v. 31, n. 9, p. 849, 2017.

XIMENES, R. A. DE A. *et al.* Risk of adverse outcomes in offspring with RT-PCR confirmed prenatal Zika virus exposure: an individual participant data meta-analysis of 13 cohorts in the Zika Brazilian Cohorts Consortium. **The Lancet Regional Health - Americas**, v. 17, p. 100395, 2023.

CAPITULO 4

Previsão de casos de Zika e microcefalia no Brasil: uma abordagem baseada em notificações de saúde e condições sanitárias, socioeconômicas e climáticas utilizando aprendizado de máquina

4.1 Introdução

A Zika é uma doença transmitida pela picada de mosquitos do gênero *Aedes*, infectados pelo vírus ZIKV. Tal vírus, que foi descrito pela primeira vez por Alexander Haddow em 1947, foi responsável por uma epidemia no Brasil nos anos de 2015 e 2016, que revelou essa doença como uma ameaça mundial. Após a descoberta da circulação do vírus no país, um aumento da incidência de microcefalia em recém-nascidos e posterior identificação da ligação entre Zika e anomalias congênitas (CALVET *et al.*, 2016; MELO *et al.*, 2016), fez com que a Organização mundial da saúde (OMS) emitisse estado de emergência internacional (LOWY, 2019).

A epidemia de Zika e microcefalia no Brasil demonstrou elevada variabilidade espacial na distribuição dos casos, o que indicou a existência de fatores que poderiam estar contribuindo para a sua propagação desigual nas regiões do país (GARCIA, 2018). A identificação de fatores ambientais que seriam capazes de atuar como determinantes na propagação do Zika motivou a realização de alguns estudos nacionais e internacionais a respeito dessa temática. A maior parte desses estudos utilizou fatores climáticos para identificação da adequação de áreas para propagação do Zika, haja vista a conhecida relação entre clima e desenvolvimento de mosquitos (SADEGHIEH *et al.*, 2021; SILVA *et al.*, 2021; XU *et al.*, 2022). Em outros estudos, como os de Morgan, Strode e Salcedo-Sora (2021) e Perrotta *et al.* (2022), foi investigada a influência de fatores socioeconômicos e de mobilidade na dinâmica de propagação do vetor e do vírus.

Apesar da grande importância dos estudos desenvolvidos, informações sanitárias e comportamentais relevantes parecem ter sido negligenciadas, o que pode diminuir a eficácia na identificação de áreas de risco para Zika. Chitolina *et al.* (2016), Han *et al.* (2016) e Silva e Machado (2018) já demonstraram que as condições de saneamento de uma região podem influir na incidência de arboviroses. Desse modo, é possível que a junção de fatores tradicionalmente avaliados em estudos de arboviroses, tais como climáticos e socioeconômicos, com fatores sanitários possa contribuir para uma identificação mais acurada de áreas de risco. De acordo com Dhaka e Singh (2020), a junção de fatores diversos, que influenciam na propagação de

arboviroses, pode auxiliar na criação de um sistema de alerta epidêmico capaz de identificar áreas de risco.

A identificação de áreas de risco e a previsão de casos constituem uma parte essencial dos sistemas de vigilância de vetores e arboviroses (RAHMAN *et al.*, 2021). A OMS, em sua diretriz para prevenção e controle da dengue, destacou a necessidade do desenvolvimento de modelos preditivos capazes de avaliar o risco de surtos de dengue (OMS, 2011). Nesse sentido, a utilização de métodos estatísticos multivariados, estatística espacial e, mais recentemente, de modelos de aprendizado de máquina, tem se mostrado uma ferramenta útil na predição de fenômenos complexos que demandam a avaliação simultânea de múltiplas variáveis (QUEIROZ; MEDRONHO, 2021; ZHAO *et al.*, 2020).

Os métodos de aprendizado de máquina (AM) são considerados abordagens promissoras para a previsão de doenças. Isso se dá em função da sua capacidade de lidar com um grande número de covariáveis preditoras, permitir a modelagem de interações complexas entre estas e não possuir premissas paramétricas, que podem dificultar a utilização de abordagens estatísticas convencionais (CUTLER *et al.*, 2007). A capacidade dos algoritmos de aprendizado de máquina de lidar com dados com as características citadas, indica uma potencialidade para a predição da incidência de Zika e microcefalia, tendo em vista a utilização de diversas covariáveis socioeconômicas e de saneamento, cuja relação com a doença ainda não foi completamente elucidada.

Modelos preditivos baseados em algoritmos de aprendizado de máquina têm sido utilizados em diversos estudos na área da saúde, demonstrando potencial para atuar como uma ferramenta de apoio à tomada de decisão (DHAKA; SINGH, 2020; GALASSO *et al.*, 2022; ISMAIL *et al.*, 2022; LIMA *et al.*, 2021; RAHMAN *et al.*, 2021). No campo de predição de arboviroses, no entanto, o uso de algoritmos de AM ainda é incipiente. Os principais trabalhos desenvolvidos, até o momento, possuem foco na predição de dengue (BARBOZA *et al.*, 2023; GUO *et al.*, 2017; ONG *et al.*, 2018b; ROCHA; GIESBRECHT, 2022; NEJAD; VARATHAN, 2021; ZHAO *et al.*, 2020). Silva *et al.* (2022) investigaram a predição de dengue, Chikungunya e Zika utilizando aprendizado de máquina, contudo, tal estudo possui apenas abrangência municipal. De acordo com Rocha e Giesbrecht (2022), há uma carência de estudos voltados para predição de Zika, o que pode estar relacionado à escassez de dados, uma vez que, as epidemias associadas ao vírus ainda podem ser consideradas recentes.

Adicionalmente, não foram identificados estudos que utilizaram AM para avaliar a capacidade de predição de casos de microcefalia, a partir dos fatores anteriormente citados. Tendo em vista o rápido acúmulo e a gravidade dos casos de microcefalia observados no Brasil,

a identificação de áreas mais susceptíveis à incidência dessa síndrome congênita é de grande relevância para o controle epidemiológico do país.

Diante do exposto, objetivou-se com esse estudo prever quais municípios brasileiros possuem maior propensão de apresentar eventos de Zika e microcefalia, a partir do conhecimento de suas características sanitárias, socioeconômicas e climáticas e da avaliação de seis algoritmos de aprendizado de máquina. Os avanços metodológicos desse estudo abrangem desde a utilização criteriosa de etapas de processamento até a identificação dos algoritmos com melhor capacidade de predição para casos de Zika e microcefalia nos municípios brasileiros. Ademais, no estudo propôs-se a identificação e interpretação de variáveis sanitárias, socioeconômicas e climáticas que possuem maior importância relativa para a construção dos modelos de AM. A abordagem realizada nesse estudo pode fornecer subsídios para a criação de estratégias para melhor monitoramento e prevenção das arboviroses.

4.2 Material e métodos

A identificação dos fatores sanitários, climáticos e socioeconômicos que podem influenciar a ocorrência de epidemias causadas pelo vírus Zika são capazes de fornecer subsídios para a formulação de estratégias para prevenção de eventos futuros. Por esse motivo, nesse estudo, a capacidade de predição da ocorrência de Zika e microcefalia nos municípios brasileiros, a partir dos fatores anteriormente citados, foi investigada. Para isso, uma metodologia baseada em aprendizado de máquina foi utilizada.

Esta seção está dividida em 6 subseções que caracterizam os procedimentos metodológicos adotados nesse estudo. Após a descrição da base de dados utilizada, a metodologia aplicada para partição e preparação dos dados, bem como para seleção de covariáveis, que antecedem o treinamento dos modelos de aprendizado de máquina, é apresentada. Por fim, a aplicação de modelos de aprendizado de máquina – *Random forest* (RF), *Gradient Boosting Machine* (GBM), *Support Vector Machines Radial Sigma* (SVM), *Multivariate Adaptive Regression Spline* (MARS), *K-Nearest Neighbors* (KNN) e Regressão Logística (RL) – e a avaliação do desempenho destes é descrita. Um fluxograma dos procedimentos metodológicos aplicados para a modelagem da ocorrência de Zika e microcefalia nos municípios brasileiros é apresentado na Figura 4.1.

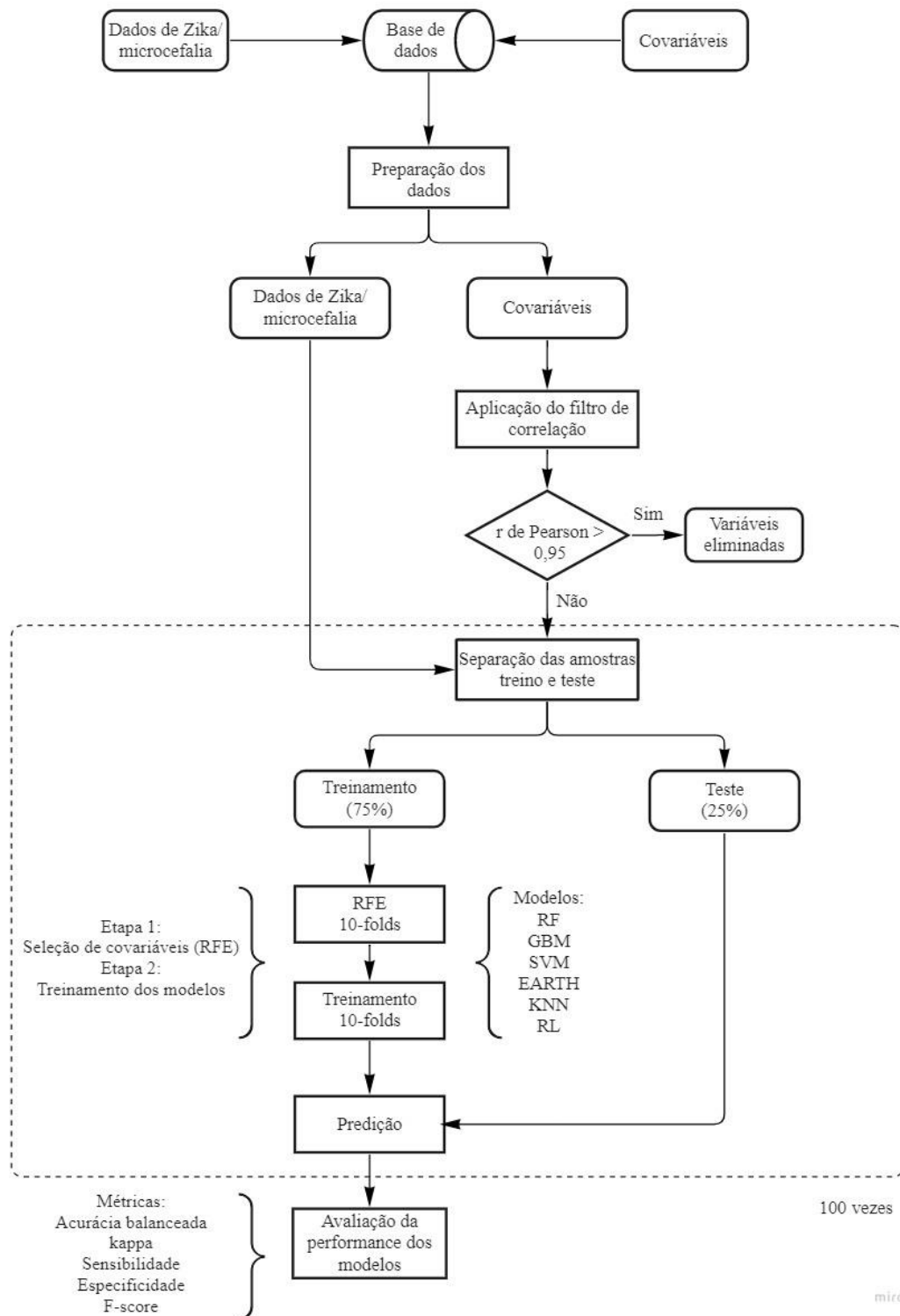


Figura 4.1. Fluxograma do procedimento metodológico aplicado para a modelagem da ocorrência de Zika e microcefalia.

4.2.1 Base de dados

Os dados de ocorrência de Zika e microcefalia (em decorrência da infecção congênita pelo vírus ZIKV) nos municípios brasileiros foram obtidos junto à Secretaria de Vigilância em Saúde do Ministério da Saúde, por meio de solicitação no Serviço de Informação ao Cidadão (SIC). Os dados brutos, anonimizados, continham a listagem de todas as notificações de Zika - presentes no Sistema de Informações de agravo de notificações (SINAN) - e microcefalia - extraídas do Painel de Monitoramento de Malformações Congênitas, Deformidades e Anomalias Cromossômicas (D180 e Q00-Q99) - ao longo do ano de 2016. Tal ano foi selecionado em função da ocorrência, nos anos de 2015 e 2016, de uma epidemia associada ao vírus Zika no Brasil. Os dados do ano de 2015 não foram selecionados para esse estudo devido à doença aguda pelo vírus Zika só ter sido considerada de notificação compulsória nacional no ano 2016.

As notificações foram, então, agrupadas em função do município provável da fonte de infecção e da classificação final dos casos notificados. Foram excluídos os casos que, embora tenham cumprido os critérios para notificação, após investigação, não se enquadraram nas definições de classificação de caso confirmado, provável ou inconclusivo (BRASIL, 2017).

Após a referida avaliação preliminar dos dados, os municípios foram categorizados em i) municípios que apresentaram pelo menos 1 caso de Zika (1) e ii) municípios que não apresentaram casos de Zika (0). A mesma categorização foi utilizada para identificar os municípios que apresentaram ou não casos de microcefalia no período avaliado. Desse modo, as ocorrências de Zika/microcefalia nos 5.571 municípios brasileiros foram utilizadas como variáveis resposta/dependentes nesse estudo.

Para a predição da ocorrência de Zika e microcefalia foram utilizadas diversas covariáveis sanitárias, socioeconômicas e climáticas que, juntas, perfizeram um total de 141 e 149 covariáveis, respectivamente. As referidas covariáveis foram selecionadas com base em sua relação, direta ou indireta, com a transmissão de arboviroses, sendo realizada uma revisão em diversos estudos relativos a conhecimentos a respeito da dinâmica da doença.

Um resumo das covariáveis utilizadas é indicado na Tabela 4.1. Para a predição da microcefalia, além das 141 covariáveis citadas na Tabela 4.1, foram incluídas 8 covariáveis, totalizando 149 covariáveis, sendo estas: taxa de prevalência de Zika em gestantes, taxa de prevalência de Zika em gestantes no 1º, 2º e 3º trimestre, ocorrência de casos de Zika no ano de 2016, taxa de natalidade, ocorrência de casos de microcefalia no ano de 2015 e taxa de internação de doenças gastrointestinais.

As principais fontes de dados utilizadas nesse estudo foram o banco de dados do Cadastro Único para Programas Sociais (CadÚnico), Secretaria Nacional de Proteção e Defesa Civil (SEDEC), Índice FIRJAN de Desenvolvimento Municipal (IFDM) e WorldClim - *Global Climate Data*.

A maior parte dos dados sanitários e socioeconômicos utilizados foram provenientes do CadÚnico. O CadÚnico, ligado ao Ministério dos Direitos Humanos e da Cidadania, é um instrumento de coleta de dados para identificação e caracterização das famílias brasileiras que possuem renda per capita inferior a meio salário mínimo ou renda familiar total de até 3 salários mínimos. Famílias com renda superior à citada também podem ser incluídas no CadÚnico, contanto que sua inclusão esteja relacionada à participação em programas sociais ou serviços implementados em nível federal, estadual ou municipal. Tal possibilidade visa o atendimento, em programas sociais, de famílias cuja situação de vulnerabilidade não está necessariamente vinculada à renda. De acordo com Brasil (2017), o CadÚnico conta com informações de cerca de 40% da população brasileira. Por esse motivo, tal banco de dados é considerado um mapa representativo das famílias vulneráveis do país.

Partindo do princípio que existem fortes indícios de que houve uma desigualdade social na ocorrência da infecção pelo ZIKV (DIDERICHSEN; AUGUSTO, 2021), considerar essa exposição diferencial ao vírus pode auxiliar no entendimento da interação existente entre o Zika e as condições ambientais e socioeconômicas da população. Para isso, é necessário avaliar informações que retratam a população vulnerável e, potencialmente, mais exposta ao vírus. Nesse sentido, o banco de dados do CadÚnico pode ser apontado como um meio eficaz de obtenção de dados que retratam a população mencionada.

Além do CadÚnico, os dados do sistema FIRJAN foram utilizados para caracterizar os municípios quanto às suas características sociais e econômicas. Optou-se pela utilização do Índice FIRJAN de Desenvolvimento Municipal (IFDM), relativo aos anos de 2015 e 2016, em detrimento do Índice de Desenvolvimento Humano dos Municípios (IDH-M). Tal escolha baseou-se na possibilidade de obtenção de informações referentes aos anos de ocorrência de epidemia do vírus Zika e microcefalia no Brasil. O IDH-M, embora publicado no ano de 2013, baseia-se em dados do censo demográfico, cuja última edição foi publicada no ano 2010.

O IFDM é um estudo do sistema FIRJAN que fornece informações anuais a respeito do desenvolvimento socioeconômico de todos os municípios brasileiros, sendo baseado em três áreas: emprego & renda, educação e saúde. Criado em 2008, o IFDM é baseado, exclusivamente, em estatísticas públicas oficiais provenientes dos Ministérios do Trabalho, Educação e Saúde. O referido índice foi criado com o objetivo de suprir a inexistência de um

parâmetro capaz de medir, periodicamente, o desenvolvimento socioeconômico dos municípios brasileiros. De modo geral, o cálculo do IFDM e do IDH-M baseia-se nas mesmas áreas: saúde, educação e renda. No entanto, para a avaliação das áreas citadas, os índices utilizam indicadores distintos.

Ao se comparar o IDH-M com o IFDM, este apresenta avanços no sentido de incluir variáveis que captam aspectos estruturais do mercado de trabalho e da renda, tais como o crescimento de empregos formais e o crescimento e distribuição da renda. Na dimensão da educação, há um avanço do IFDM ao representar aspectos qualitativos do ensino no país com a inclusão de variáveis como: taxa de distorção da idade-série, percentual de docentes com ensino superior, número médio diário de horas-aula, taxa de abandono escolar e o resultado médio no Índice de Desenvolvimento da Educação Básica (IDEB). Por fim, em relação aos indicadores de saúde, o IFDM avança por incluir aspectos qualitativos da atenção básica à saúde, que incluem a contabilização de óbitos e internações por causas mal definidas ou evitáveis e o acompanhamento médico de gestantes.

Em relação às covariáveis climáticas, foram utilizados dados provenientes da SEDEC e WorldClim. O WorldClim é um banco de dados que disponibiliza dados de alta resolução espacial referentes às condições meteorológicas e climáticas globais. Tal banco de dados é disponibilizado gratuitamente, sendo mantido por pesquisadores da Universidade de Berkeley na Califórnia.

A SEDEC, por sua vez, é vinculada ao Ministério do Desenvolvimento Regional e os dados foram obtidos por meio do Sistema Integrado de Informações sobre Desastres - S2ID. Tais registros são baseados nos decretos de declaração de Situação de Emergência (SE) e de Estado de Calamidade Pública (ECP) expedidos pelos municípios, que são posteriormente homologados pelos estados e reconhecidos pela União. Foram avaliados registros relacionados a eventos de cheias (inundações, alagamentos e enxurradas) e secas (estiagens e secas).

Tabela 4.1. Covariáveis utilizadas para predição da ocorrência de Zika e microcefalia.

Classe	Variável	Abreviatura	Fonte
<i>Sanitária</i>	Percentual de domicílios atendidos por rede geral de distribuição de água	agua.rede	CadÚnico
	Percentual de domicílios com abastecimento de água através de poço	agua.poco	
	Percentual de domicílios com abastecimento de água através de nascente	agua.nasc	
	Percentual de domicílios com abastecimento de água através de cisterna	agua.cist	
	Percentual de domicílios com abastecimento de água por outras fontes	agua.dem	
	Percentual de domicílios atendidos por rede coletora de esgotos ou pluvial	esg.rede	
	Percentual de domicílios que utilizam fossa séptica	esg.foss	
	Percentual de domicílios que utilizam fossa rudimentar	esg.foss.rud	
	Percentual de domicílios que dispõem o esgoto em vala a céu aberto	esg.ceu	
	Percentual de domicílios que dispõem o esgoto direto em um rio, lago ou mar	esg.rio	
	Percentual de domicílios que utilizam outras formas de tratamento/disposição do esgoto sanitário	esg.dem	
	Percentual de domicílios atendidos com coleta direta ou indireta de resíduos	col.dir.ind	
	Percentual de domicílios que possuem resíduos sólidos queimados	col.quei	
	Percentual de domicílios que possuem resíduos sólidos enterrados na propriedade	col.ent	
	Percentual de domicílios que possuem resíduos sólidos jogados em terreno baldio ou logradouro	col.terr	
	Percentual de domicílios que possuem resíduos jogados em rio ou mar	col.rio	
Percentual de domicílios que dão outras formas de destinação aos resíduos sólidos	col.dem		
<i>Climática e bioclimática</i>	Incidência cumulativa de eventos extremos de cheias nos municípios brasileiros	cheia.hist	ANA/SEDEC
	Incidência de eventos extremos de cheias nos municípios brasileiros no ano de 2016	cheia.2016	ANA/SEDEC
	Incidência cumulativa de eventos extremos de secas nos municípios brasileiros	seca.hist	ANA/SEDEC
	Incidência de eventos extremos de secas nos municípios brasileiros no ano de 2016	seca.2016	ANA/SEDEC
	Temperatura média anual	bio1	WorldClim
	Variação média de temperatura diurna	bio2	WorldClim
	Isotermalidade	bio3	WorldClim
	Temperatura sazonal	bio4	WorldClim
	Temperatura máxima do mês mais quente	bio5	WorldClim
Temperatura mínima do mês mais frio	bio6	WorldClim	
Faixa de temperatura anual	bio7	WorldClim	

	Temperatura média do trimestre mais chuvoso	bio8	WorldClim
	Temperatura média do trimestre mais seco	bio9	WorldClim
	Temperatura média do trimestre mais quente	bio10	WorldClim
	Temperatura média do trimestre mais frio	bio11	WorldClim
	Precipitação anual	bio12	WorldClim
	Precipitação do mês mais chuvoso	bio13	WorldClim
	Precipitação do mês mais seco	bio14	WorldClim
	Sazonalidade de precipitação	bio15	WorldClim
	Precipitação do trimestre mais úmido	bio16	WorldClim
	Precipitação do trimestre mais seco	bio17	WorldClim
	Precipitação do trimestre mais quente	bio18	WorldClim
	Precipitação do trimestre mais frio	bio19	WorldClim
	Temperatura mínima (mensal) ²	tmin	WorldClim
	Temperatura média (mensal) ²	tavg	WorldClim
	Temperatura máxima (mensal) ²	tmax	WorldClim
	Precipitação (mensal) ²	prec	WorldClim
	Velocidade do vento (mensal) ²	wind	WorldClim
	Pressão de vapor de água (mensal) ²	vapr	WorldClim
	Radiação solar (mensal) ²	srad	WorldClim
	Percentual da população inscrita no CadÚnico	pop.cad	CadÚnico/IBGE
	Percentual de famílias em situação de extrema pobreza ¹	ext.pob	CadÚnico
	Percentual de responsáveis familiares sem instrução	sem.inst	CadÚnico
	Percentual de responsáveis familiares com nível superior completo de educação	sup.comp	CadÚnico
	Percentual de pessoas inscritas no CadÚnico com cor/raça preta ou parda	pret.e.pard	CadÚnico
<i>Social e econômica</i>	Percentual de responsáveis familiares do sexo feminino	sex.fem	CadÚnico
	Localização do domicílio – Relação domicílio urbano/domicílio rural	rura.urb	CadÚnico
	Índice FIRJAN de Desenvolvimento Municipal - Consolidado	firjan	FIRJAN
	Índice FIRJAN de Desenvolvimento Municipal – Emprego e renda	fir.emp	FIRJAN
	Índice FIRJAN de Desenvolvimento Municipal – Saúde	fir.sau	FIRJAN
	Índice FIRJAN de Desenvolvimento Municipal – Educação	fir.edu	FIRJAN
	Produto interno bruto <i>per capita</i> do município	pib	IBGE
<i>Outras</i>	Macrorregião de localização do município	macro	IBGE

Unidade federativa ao qual o município pertence	est	IBGE
Densidade demográfica	dens	IBGE
Taxa de incidência e de internação de dengue no ano de 2016	tx.int.dengue	Ministério da saúde
Ocorrência de Zika no ano de 2015	zika.2015	Ministério da saúde

Nota: ¹São consideradas famílias em situação de extrema pobreza aquelas que possuem renda mensal per capita de até R\$ 89,00 (BRASIL, 2018). ²Inclui uma variável para cada mês do ano.

Obs.: Todas as covariáveis sanitárias, sociais e econômicas são referentes ao ano de 2016.

4.2.2 Preparação dos dados e eliminação de covariáveis correlacionadas

A preparação ou pré-processamento dos dados envolve as etapas de identificação e correção de inconsistências em bases de dados, devendo ser realizada anteriormente ao processo de predição. Tais etapas são necessárias tendo em vista que os dados provenientes de diferentes fontes podem apresentar falhas que resultam no mau desempenho das análises preditivas. Diante disso, a execução de procedimentos capazes de corrigir inconsistências na base de dados deve ser realizada previamente à utilização de algoritmos de aprendizado de máquina (SILVA *et al.*, 2021). Nesse estudo, o pré-processamento dos dados consistiu nas seguintes etapas: identificação e eliminação de valores ausentes (*missing values*), identificação de covariáveis com variância próxima a zero, padronização dos dados e eliminação de covariáveis altamente correlacionadas (DIAS *et al.*, 2021; PECHT; KANG, 2018).

Tendo em vista que a existência de valores ausentes pode inviabilizar a utilização de alguns algoritmos, tais valores foram removidos da base de dados utilizada nesse estudo. Tal procedimento foi adotado seguindo as recomendações de Harrell Jr (2015), que cita que, quando a proporção de dados faltantes em alguma das covariáveis é inferior a 3%, a análise de dados completos (eliminação de valores ausentes) pode ser realizada.

Além da existência de valores ausentes, as bases de dados podem conter covariáveis que não contribuem para a predição, tais como aquelas que possuem variância próxima a zero, ou seja, que apresentam uma frequência elevada de um único valor. Para identificação e, posterior, eliminação de tais covariáveis foi utilizada a ferramenta *nearZeroVar* do pacote *caret* (KUHN *et al.*, 2020).

Tendo em vista que as covariáveis avaliadas nesse estudo possuem diferentes unidades de medida (dimensões), foi realizada a padronização dos dados com transformação para escala Z (média zero e desvio padrão 1), conforme indicado na Equação 4.1.

$$Z = \frac{(X-\mu)}{\sigma} \quad \text{Equação 4.1}$$

Onde,

Z: variável padronizada;

X: valor original da variável;

μ : média da variável; e

σ : desvio padrão da variável

Após a padronização dos dados foi realizada a eliminação de covariáveis altamente correlacionadas. Nesse procedimento, a correlação de Pearson foi utilizada para identificar o nível de correlação entre as variáveis e permitir a posterior eliminação daquelas que continham informações redundantes. A eliminação das variáveis foi realizada conforme procedimento citado por Dias *et al.* (2021), Santos *et al.* (2022) e Souza *et al.* (2018), em que, para pares de variáveis que apresentam coeficiente de correlação superior a 0,95, é realizada a eliminação de uma das variáveis do par, sendo esta a que possui a maior média do valor absoluto da correlação com as demais variáveis.

4.2.3 Partição dos dados

De acordo com Silva *et al.* (2021), a avaliação do desempenho de um modelo utilizando os dados empregados em seu treinamento não é apropriada. Tal afirmação deriva do fato que modelos podem apresentar um bom desempenho preditivo nos dados utilizados no processo de treinamento e, no entanto, não serem capazes de fornecer respostas apropriadas para novos dados. A fim de evitar esse problema e obter modelos que apresentam bom desempenho preditivo em novos dados, estratégias de partição do conjunto de dados original podem ser adotadas.

Diante do exposto, o conjunto de dados utilizado nesse estudo foi dividido em dois subconjuntos. O primeiro subconjunto, correspondente a 75% dos dados, foi utilizado para o treinamento dos modelos preditivos. O segundo subconjunto, utilizado para testar o desempenho preditivo em novos dados, foi composto pelos 25% dos dados restantes. As amostras alocadas em cada um dos subconjuntos citados foram definidas aleatoriamente. A estratégia citada foi aplicada 100 vezes, com o intuito de variar a composição dos subconjuntos de treinamento e de teste, ou seja, foram geradas 100 amostras diferentes para treinamento e teste dos modelos (DIAS *et al.*, 2021). A realização da partição dos dados N vezes é fundamental para maximizar as chances de gerar avaliações estatisticamente confiáveis e avaliar a variabilidade das predições obtidas, uma vez que, diferentes subconjuntos de treinamento e teste podem gerar diferentes resultados (KUHN; JOHNSON, 2013).

4.2.4 Seleção das covariáveis por importância

Segundo Baquero *et al.* (2018), comumente, as técnicas de predição de arboviroses utilizam um grande número de variáveis que podem tornar as predições mais eficazes. No entanto, apenas a inclusão de variáveis não garante um melhor desempenho nos modelos de predição, uma vez que, a adição de variáveis irrelevantes pode diminuir a precisão preditiva dos modelos e exigir maior processamento computacional. Além disso, a inclusão de variáveis irrelevantes torna os modelos mais complexos, difíceis de serem explicados e aumenta a chance de ocorrência de superajuste. Por esse motivo, antes da aplicação de modelos de predição baseados em aprendizado de máquina é comum a realização de uma seleção de variáveis para remoção de dados ruidosos e redundantes (KUBAT, 2017).

Nesse estudo, após a utilização do filtro de correlação, que selecionou um subconjunto das variáveis originais, foi aplicado o algoritmo *Recursive Feature Elimination* (RFE) do pacote *caret* (KUHN *et al.*, 2020). O objetivo do RFE foi realizar uma seleção recursiva de variáveis, eliminando iterativamente aquelas menos promissoras no modelo avaliado. Inicialmente, o algoritmo ajustou o modelo a todas as variáveis e realizou uma classificação destas a partir de sua importância para o modelo. Em seguida, as variáveis mais importantes para o modelo foram retidas e as menos importantes foram eliminadas. Nesse ponto, o modelo foi reajustado e o seu desempenho avaliado. O procedimento descrito foi repetido recursivamente, considerando diferentes tamanhos de subconjuntos de variáveis e, por fim, as principais variáveis foram utilizadas para o ajuste do modelo final (CHEN; MANONGGA; DEWI, 2022).

O RFE foi executado com o conjunto completo de variáveis (que permaneceram após a utilização do filtro de correlação) e com subconjuntos de 5, 6, 7, ..., 20, 30, 40 e 50 variáveis. Em cada um dos subconjuntos citados, foram testadas diferentes combinações de variáveis, sendo selecionadas aquelas que, para cada um dos modelos de aprendizado de máquina avaliados (RF, GBM, SVM, MARS, KNN e RL) apresentaram a melhor desempenho. Para a otimização dos subconjuntos ideais de variáveis, para cada modelo, foi utilizada a validação cruzada *k-fold* repetida, adotando-se 10 *folds*. O RFE foi executado com os dados de treinamento e o subconjunto ótimo de variáveis foi utilizado para o treinamento de cada modelo testado, conforme detalhado no tópico 4.2.5 (DIAS *et al.*, 2021). Nesse passo, também foram identificadas as covariáveis que possuíam maior importância relativa para o desempenho dos modelos, sendo gerado, posteriormente, uma listagem com a ordem de classificação das variáveis mais "importantes" nos modelos ajustados (GOLDSTEIN; NAVAR; CARTER, 2017).

4.2.5 Treinamento dos modelos

Para o treinamento dos modelos avaliados foram utilizados os subconjuntos ideais de variáveis obtidos na etapa de seleção das covariáveis por importância. Visando a avaliação de diferentes métodos de classificação, seis tipos diferentes de algoritmos de aprendizado de máquina foram testados, sendo estes: *Random forest* (RF) (BREIMAN, 2001), *Gradient Boosting Machine* (GBM) (FRIEDMAN, 2001), *Support Vector Machines Radial Sigma* (SVM) (GUNN, 1998), *Multivariate Adaptive Regression Spline* (MARS) (FRIEDMAN, 1991), *K-Nearest Neighbors* (KNN) (FUKUNAGA; NARENDRA, 1975) e Regressão Logística (RL) – LogitBoost (FRIEDMAN; HASTIE; TIBSHIRANI, 2000).

Após a seleção dos modelos, foi realizada a otimização dos hiperparâmetros de cada modelo. Tais hiperparâmetros podem ser definidos como um conjunto de configurações, ou parâmetros internos, específicos de cada tipo de modelo. A realização do ajuste dos hiperparâmetros visa melhorar o desempenho dos modelos, aumentando seus níveis de assertividade e precisão. Os hiperparâmetros que foram otimizados em cada um dos modelos são indicados na Tabela 4.2. Uma descrição de cada um destes pode ser obtida em Kuhn *et al.* (2020).

Tabela 4.2. Hiperparâmetros otimizados em cada um dos modelos de aprendizado de máquina avaliados.

Modelo	Hiperparâmetros
RF	Número de preditores selecionados aleatoriamente (mtry)
GBM	Profundidade máxima da árvore (interaction.depth) e número de iterações boosting (n.trees)
SVM	Sigma e custo (C)
MARS	Número de termos (nprune) e degree
KNN	Número de vizinhos considerados (k)
RL	Número de iterações boosting (nIter)

A otimização dos hiperparâmetros dos modelos foi realizada utilizando-se a validação cruzada *k-fold* repetida, com 10 *folds* e 5 repetições. Após o procedimento descrito, os modelos foram treinados por 100 vezes. Todos os modelos citados foram ajustados utilizando-se o pacote *caret* no *software* R (KUHNS *et al.*, 2020). Após o ajuste com a estratégia de validação cruzada *k-fold* repetida, foram selecionados os modelos com melhor desempenho para predição das

variáveis resposta/dependente avaliadas (Zika e microcefalia). Para isso, diferentes métricas de desempenho foram avaliadas, conforme indicado na subseção 4.2.6.

4.2.6 Avaliação do desempenho dos modelos

Para a avaliação de um modelo, não é apropriado que a sua qualidade seja atestada em relação aos dados apresentados no processo de treinamento, uma vez que, a capacidade de generalização dos modelos deve ser avaliada (SANTOS *et al.*, 2019). Os modelos ajustados na etapa anterior foram utilizados para fazer a predição do conjunto de teste (25%). Esse conjunto não foi utilizado no treinamento do modelo, uma vez que, o seu objetivo é avaliar a capacidade preditiva do modelo em dados ainda não vistos, ou seja, mensurar a capacidade de generalização dos modelos (SILVA *et al.*, 2021).

Para a avaliação do desempenho dos modelos, foram utilizadas como métricas: acurácia balanceada, sensibilidade (*recall*), especificidade, *F1-score* e índice Kappa. Tais métricas foram selecionadas visando obter resultados mais consistentes para a avaliação de bases de dados desbalanceadas, ou seja, que possuem mais registros de uma determinada classe em detrimento de outra.

As etapas de treinamento e teste dos modelos foram realizadas por 100 vezes, tendo em vista a necessidade da adoção de repetições para a determinação da variabilidade das previsões. Após a obtenção das métricas descritas, para cada modelo avaliado, foram calculadas medidas de estatística descritiva e, em seguida, o teste de Kruskal-Wallis ($\alpha = 5\%$) foi utilizado para identificar a existência de diferenças estatísticas significativas entre os modelos avaliados.

4.3 Resultados e discussão

4.3.1 Visão geral dos dados

Os dados levantados nesse estudo indicaram que, no ano de 2016, cerca de 40% dos municípios brasileiros registraram casos de infecção pelo vírus Zika. A síndrome congênita associada à infecção pelo vírus Zika (microcefalia) em recém-nascidos, por sua vez, foi notificada em 15% dos municípios do país (Figura 4.2).

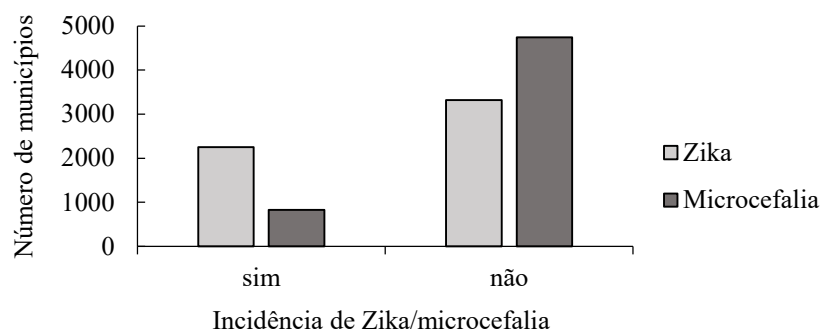


Figura 4.2. Municípios com registros de casos de Zika e microcefalia no Brasil no ano de 2016.

Embora o número absoluto de casos por município não tenha sido objeto de análise nas modelagens realizadas nesse estudo, foi verificado que, no ano de 2016, houve a notificação de 281.464 casos de Zika no Brasil. Destes, 219.266 se enquadraram nas definições de classificação de caso confirmado, provável ou inconclusivo. Os casos apresentaram maior concentração nas regiões Nordeste e Centro-Oeste do Brasil, onde foram verificadas as maiores taxas de incidência de Zika por 100.000 habitantes (Figura 4.3).

Em relação à incidência de microcefalia, o painel de monitoramento de malformações congênitas, deformidades e anomalias cromossômicas identificou 2.276 casos no ano de 2016. Tais casos foram predominantes no nordeste do Brasil (Figura 4.4), onde residiam cerca de 47% dos nascidos vivos registrados, evidenciando a concentração dos casos nessa região. Adicionalmente, Garcia (2018) destaca que a incidência de microcefalia na região Nordeste foi muito superior à observada nos demais países onde ocorreram epidemias associadas ao vírus Zika. Apesar disso, os casos foram relatados em todas as regiões brasileiras, o que revela uma disseminação nacional do vírus Zika.

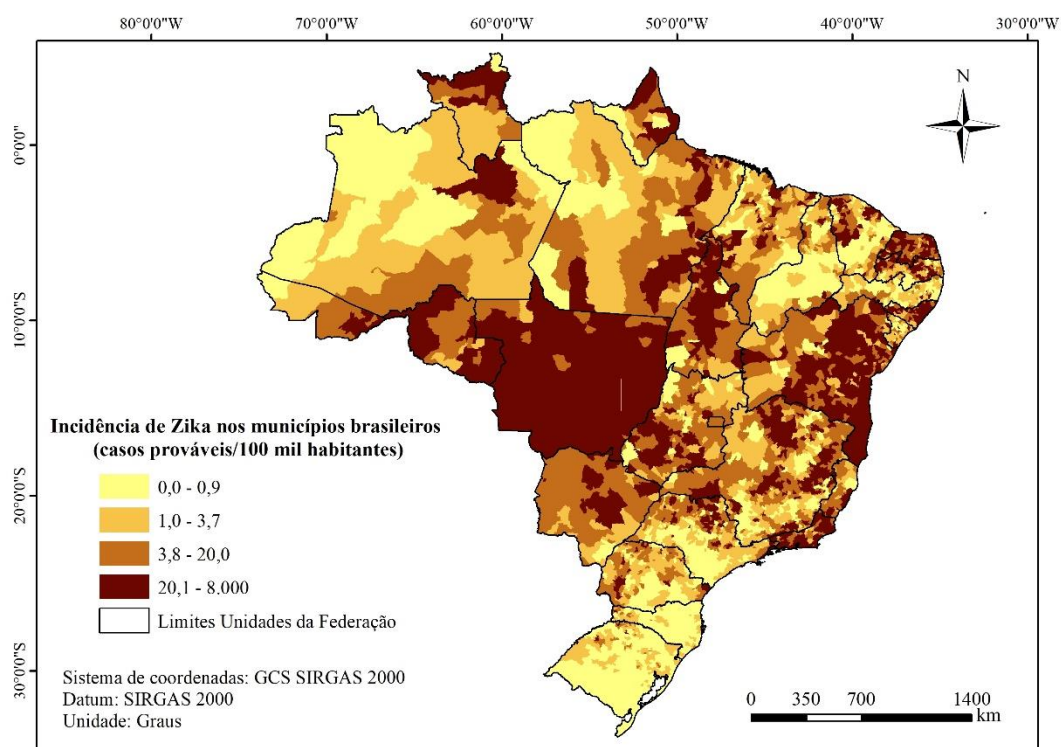


Figura 4.3. Taxa de incidência de Zika nos municípios brasileiros no ano de 2016.

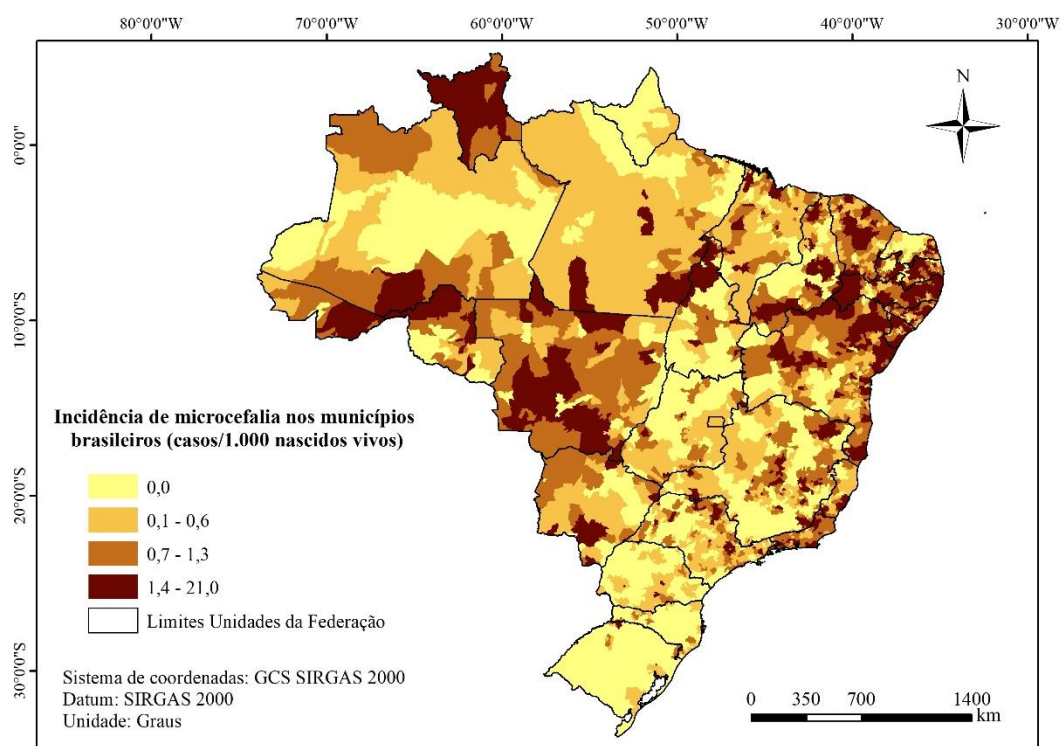


Figura 4.4. Taxa de incidência de microcefalia nos municípios brasileiros no ano de 2016.

Ainda que tenha sido identificada uma diminuição dos casos de infecção por Zika após o ano epidêmico de 2016, dados do Ministério da Saúde indicam que, entre os anos de 2017 e 2021, foram notificados 123.315 casos no país. Comportamento semelhante foi verificado para os casos de microcefalia. De acordo com a Secretaria de Vigilância em Saúde (BRASIL, 2022), embora o período de emergência tenha sido, oficialmente, encerrado em 30 de julho de 2017, atualmente, ainda são registrados novos casos de microcefalia no país.

Entre os anos de 2015 e 2022, dos casos confirmados de microcefalia em decorrência do vírus Zika, 13,8% foram a óbito. Tal informação reforça a necessidade de investigação e vigilância das intercorrências provocadas pelo vírus Zika no Brasil. Ademais, compreender as causas e fatores ambientais que podem influenciar nos processos de propagação e infecção pelo referido vírus pode auxiliar a vigilância para novos casos no país. A compreensão de tais fatores se faz ainda mais importante considerando um país com as características do Brasil, que possui uma ampla circulação do vetor de transmissão do vírus Zika (*Aedes aegypti*), o que torna possível a ocorrência de novos surtos e epidemias associados a este. Ainda de acordo com Brasil (2022), a manutenção da vigilância das infecções pelo vírus Zika é essencial para a adoção de medidas de controle e prevenção, além da implementação de políticas públicas que sejam baseadas em evidências.

4.3.2 Pré-processamento dos dados e seleção de covariáveis

A partir da investigação da existência de dados faltantes, foi verificada a ausência de 0,3% e 0,2% das informações das covariáveis pertencentes aos bancos de dados de Zika e microcefalia, respectivamente. Partindo do pressuposto que a existência de valores ausentes no banco de dados pode inviabilizar a utilização de alguns algoritmos, optou-se pela eliminação das observações correspondentes aos dados faltantes. A eliminação foi realizada considerando casos completos, ou seja, na ausência da informação de uma covariável, o município seria eliminado da análise. A realização de tal procedimento foi possível em virtude do baixo número de dados ausentes identificados. De acordo com Harrell Jr (2015), em situações onde a proporção de dados faltantes em alguma das variáveis é inferior a 0,03 (3%), pode-se realizar a análise de dados completos.

Além dos dados faltantes, foi investigada a existência de covariáveis que possuíam variância próxima zero e, em função disso, poderiam prejudicar o desempenho dos modelos preditivos. Para o banco de dados referente à Zika, apenas uma variável apresentou variância próxima a zero. No segundo banco de dados, referente à microcefalia, por sua vez, foram

identificadas 4 covariáveis com a característica anteriormente citada. Tendo em vista que a fração de covariáveis com valores únicos (ou variância próxima a zero) sobre o tamanho da amostra foi baixa, foi realizada a exclusão das covariáveis citadas. Segundo Kuhn e Johnson (2013), a eliminação de preditores com variância zero, ou próxima a zero, pode resultar em um melhor desempenho da modelagem preditiva.

Na etapa de seleção das covariáveis, utilizando o filtro de correlação de Pearson, foram eliminadas as variáveis altamente correlacionadas, resultando na diminuição do número de covariáveis de 141 para 84 no banco de dados de Zika e de 149 para 90 no conjunto de dados relativo à microcefalia. Em geral, as técnicas de predição utilizam um grande número de covariáveis, a fim de abarcar uma variedade de dimensões que possuem o potencial de tornar as previsões mais eficazes. No entanto, diversos autores apontam que a utilização de preditores altamente correlacionados, e que, portanto, possuem a mesma informação, pode interferir no processo de aprendizado e resultar em modelos instáveis ou que apresentam erros numéricos. Além disso, a remoção de determinados preditores, que garante que as correlações de pares de variáveis estejam abaixo de um determinado limiar, resulta em menores custos computacionais nas análises e pode minimizar a ocorrência de superajuste por meio da simplificação e generalização do modelo (BAQUERO *et al.*, 2018; KUHN; JOHNSON, 2013; SARKER, 2021).

4.3.3 Desempenhos dos modelos preditivos

Na tabela 4.3 são apresentados os resultados da etapa de avaliação (teste) dos modelos utilizados para predição de Zika nos municípios brasileiros. De modo geral, é possível verificar que a maior parte dos modelos apresentou acurácia balanceada superior a 0,7, demonstrando um bom desempenho de predição das classes investigadas. Para o algoritmo *Random Forest*, que apresentou o melhor desempenho ($\alpha = 5\%$) quanto à referida métrica, o resultado obtido foi de 0,75. Tal resultado foi semelhante ao obtido por Sarma *et al.* (2020), que identificaram uma acurácia de 0,74 ao utilizar *Random Forest* para a predição de casos de dengue. Vale ressaltar que os resultados obtidos pelos referidos autores se referem à acurácia, que tende a apresentar valores superiores à acurácia balanceada em estudos epidemiológicos, conforme será relatado a seguir.

Tabela 4.3. Desempenho dos modelos avaliados para predição de Zika nos municípios brasileiros.

Modelos	Métricas de avaliação de desempenho				
	Acurácia balanceada	Sensibilidade	Especificidade	F1-score	Kappa
GBM	0,7470 b	0,8367 c	0,6574 a	0,8088 c	0,5023 b
RF	0,7473 a	0,8496 b	0,6449 a	0,8130 a	0,5054 a
SVM	0,7299 c	0,8675 a	0,5924 b	0,8094 b	0,4760 c
MARS	0,7301 c	0,8357 c	0,6245 a	0,7997 d	0,4703 c
KNN	0,7031 d	0,8171 d	0,5890 c	0,7798 d	0,4155 d
RL	0,6709 d	0,7837 e	0,5582 c	0,7520 d	0,3486 d

Nota: *Random forest* (RF), *Gradient Boosting Machine* (GBM), *Support Vector Machines Radial Sigma* (SVM), *Multivariate Adaptive Regression Spline* (MARS), *K-Nearest Neighbors* (KNN) e Regressão Logística (RL). Medianas seguidas pela mesma letra na coluna não apresentam diferença estatística significativa pelo teste de Kruskal-Wallis ($\alpha = 5\%$). Os valores indicados são relativos a medianas de 100 amostras.

Dentre os algoritmos avaliados, a regressão logística apresentou o pior desempenho, com os índices mais baixos em todas as métricas de avaliação utilizadas. Tal fato pode ser justificado por se tratar de um modelo com menor complexidade, quando comparado aos demais modelos avaliados. Os melhores resultados para predição de Zika foram observados para o algoritmo *Random forest*, que apresentou o melhor desempenho ($\alpha = 5\%$) em quatro dos cinco índices avaliados. Em função disso, as discussões que seguem, a respeito da predição de Zika, terão como foco o desempenho do referido algoritmo.

Em relação à especificidade, verifica-se que o *Random forest* apresentou especificidade de 0,6449, sendo este superior ao obtido por Rocha e Giesbrecht (2022) (0,6053) ao prever casos de dengue. Esse índice, que varia de 0 a 1, indica a capacidade do modelo em prever, corretamente, as classes negativas que, nesse estudo, representam a existência de casos de Zika no município. Desse modo, os resultados demonstram que o modelo construído foi capaz de classificar, corretamente, cerca de 65% das classes “sim”. Tal resultado indica que, a partir das covariáveis utilizadas nesse estudo, o modelo conseguiu prever corretamente a incidência de Zika em 65% dos municípios contidos nos dados de teste.

A sensibilidade, por sua vez, mede a capacidade do modelo de classificar corretamente as classes positivas, que foram representadas, nesse estudo, pela ausência de registros de casos de Zika nos municípios. Tendo em vista que a sensibilidade, no modelo *Random forest*, assumiu o valor de 0,8496, pode-se inferir que o modelo construído possui capacidade de classificar, corretamente, a ausência de casos de Zika em cerca de 85% dos dados de teste. O valor citado foi superior aos encontrados por Rocha e Giesbrecht (2022) e Sarma *et al.* (2020) ao utilizar

Random Forest para predição de casos de dengue. Os autores obtiveram valores de sensibilidade de 0,75 e 0,67 respectivamente.

A partir dos resultados obtidos na avaliação do desempenho dos modelos, verifica-se que o modelo *Random forest* apresenta indícios de ser o modelo mais apropriado para predição de Zika em dados futuros. O referido algoritmo já demonstrou sua potencialidade em pesquisas epidemiológicas realizadas anteriormente para predição da COVID-19 (GALASSO; CAO; HOCHBERG, 2022; LIMA *et al.*, 2021), dengue (ISMAIL *et al.*, 2022; RAHMAN *et al.*, 2021), Chikungunya (DHAKA; SINGH, 2020).

De acordo com Biau (2012), dentre os diversos algoritmos de aprendizado de máquina existentes, o *Random forest* destaca-se em função da sua facilidade de implementação, precisão nas previsões e à sua capacidade de superar o problema comum de ajuste excessivo (*overfitting*). Além disso, Prasad *et al.* (2016) destacam que a seleção do preditor aleatório, realizado em *Random forest*, mantém os desvios baixos e com melhores estimativas do modelo quando comparado a outros modelos que simulam árvores de classificação.

A transmissão de arboviroses é um processo multifatorial, desse modo, por vezes, fatores de risco são correlacionados, o que dificulta a utilização de métodos convencionais de modelagem. A capacidade do algoritmo *Random forest* de lidar com dados com as características citadas, além de ser pouco afetado por dados ruidosos e *outliers* (STAVROPOULOS *et al.*, 2020), pode ser um indicativo que tal algoritmo possui potencialidade para a avaliação da incidência de Zika.

O *Random forest* foi utilizado para predição de dengue por Ong *et al.* (2018) e Zhao *et al.* (2020) e demonstrou forte capacidade preditiva, sendo apontado como ferramenta efetiva de vigilância para orientar as operações de controle de vetores. O algoritmo também foi utilizado, com sucesso, para o mapeamento do risco de transmissão do vírus Zika por Jiang *et al.* (2018). Contudo, de modo geral, os estudos que investigaram a incidência de arboviroses por meio do algoritmo *Random forest* não incluíram fatores relacionados às condições de saneamento e outros fatores sociais, econômicos e demográficos, que melhoram a precisão das previsões nos modelos epidêmicos.

Em relação aos resultados referentes à predição de microcefalia, são apresentados na Tabela 4.4 os resultados da etapa de avaliação (teste) dos modelos utilizados. De modo geral, é possível verificar que os modelos apresentaram resultados inferiores aos observados para predição de Zika. Para o algoritmo GBM, que apresentou o melhor desempenho ($\alpha = 5\%$) em quatro das cinco métricas avaliadas, foi verificada uma acurácia balanceada de 0,6708. O mesmo algoritmo apresentou especificidade de 0,3694, demonstrando que o modelo foi capaz

de prever, corretamente, a incidência de microcefalia em cerca de 37% dos municípios contidos nos dados de teste. Em relação ao coeficiente Kappa, o resultado obtido apresenta uma concordância moderada entre as classes originais e as classes preditas, segundo a classificação de Landis e Koch (1977).

Tabela 4.4. Desempenho dos modelos avaliados para predição de microcefalia nos municípios brasileiros.

Modelos	Métricas de avaliação de desempenho				
	Acurácia balanceada	Sensibilidade	Especificidade	F1-score	Kappa
GBM	0,6708 a	0,9722 c	0,3694 a	0,9334 a	0,4234 a
RF	0,6604 b	0,9756 b	0,3452 bc	0,9333 a	0,4067 b
SVM	0,6153 c	0,9855 a	0,2450 c	0,9305 bc	0,3179 c
MARS	0,6661 a	0,9700 c	0,3623 a	0,9317 b	0,4108 a
KNN	0,6173 bc	0,9750 b	0,2596 bc	0,9263 cd	0,3108 c
RL	0,6558 b	0,9546 d	0,3569 b	0,9233 d	0,3668 c

Nota: *Random forest* (RF), *Gradient Boosting Machine* (GBM), *Support Vector Machines Radial Sigma* (SVM), *Multivariate Adaptive Regression Spline* (MARS), *K-Nearest Neighbors* (KNN) e Regressão Logística (RL). Medianas seguidas pela mesma letra na coluna não apresentam diferença estatística significativa pelo teste de Kruskal-Wallis ($\alpha = 5\%$). Os valores indicados são relativos a medianas de 100 amostras.

O menor desempenho dos modelos preditivos para microcefalia pode estar associado à maior complexidade de se modelar uma síndrome congênita, uma vez que, sua incidência não está associada apenas à presença da doença Zika. Sabe-se que o desempenho de um modelo preditivo está diretamente associado à disponibilidade de covariáveis informativas que possuam poder discriminatório em relação à variável resposta (SILVA *et al.*, 2021).

Embora tenham sido levantadas hipóteses sobre a existência de condições ambientais, sociais, sanitárias e econômicas propícias para a infecção de gestantes pelo vírus Zika (LOWY, 2019; OSORIO-DE-CASTRO *et al.*, 2017), a incidência de microcefalia em recém nascidos pode estar associada a fatores de difícil modelagem, tais como fatores genéticos, influência de infecção prévia por outros vírus (BRASIL, 2022; OSORIO-DE-CASTRO *et al.*, 2017).

Adicionalmente, Lessler *et al.* (2016) destacam que a quantificação do risco de microcefalia é complexa em função da incerteza do número de gestantes afetadas pelo vírus Zika. Isso ocorre devido ao grande número de casos assintomáticos, à falta de consenso médico a respeito da definição de microcefalia, além da existência de outras causas infecciosas de microcefalia, tal como a rubéola. Além disso, de acordo com Silva Neto *et al.* (2022), devido à circulação concomitante de diversos arbovírus que apresentam sintomas semelhantes, o diagnóstico clínico de arboviroses é complexo. Tal complexidade resultou em uma grande

parcela de casos de microcefalia que carece de confirmação laboratorial quanto à sua relação com o vírus Zika (OSORIO-DE-CASTRO *et al.*, 2017).

Diante do exposto, verifica-se que há uma dificuldade para lidar com certas lacunas da informação associadas à incidência de microcefalia, uma vez que, trata-se de um cenário recente cujas causas ainda são objeto de estudos. Adicionalmente, por se tratar de um tema ainda considerado recente, não existem estudos prévios que investigaram a influência de fatores sanitários, socioeconômicos e climáticos na incidência de microcefalia. Os estudos existentes, que utilizam modelos de aprendizado de máquina para predição de microcefalia, estão focados na classificação diagnóstica com base em dados clínicos, tais como os estudos de Assis *et al.* (2021) e Veiga *et al.* (2021).

Apesar das diferenças observadas entre os resultados dos modelos de predição para Zika e microcefalia, é possível verificar que, de modo geral, os modelos que melhor se ajustaram aos dados foram aqueles baseados em árvores de decisão (*Random forest* e GBM). Silva Neto *et al.* (2022) também verificaram a predominância de algoritmos de aprendizado de máquina baseados em árvore para a classificação binária de casos de dengue, o que reforça a aplicabilidade de tais algoritmos para a predição de arboviroses. Os autores ressaltam que por meio da utilização de modelos não lineares é possível solucionar problemas de maior complexidade e obter resultados mais precisos que outros modelos simplificados de predição, o que corrobora com os resultados obtidos nesse estudo.

As diferenças no desempenho dos modelos preditivos podem estar associadas a diversos fatores, tais como: gerenciamento dos dados de entrada (covariáveis selecionadas para compor o modelo), métodos utilizados para construir os modelos e refinamento de hiperparâmetros. Para dados epidemiológicos, outra característica que pode influenciar no desempenho dos modelos é a distribuição desbalanceada da variável resposta. Isso se dá em função desses dados apresentarem maior frequência para a classe de amostras com desfecho ausente, o que ocorre, por exemplo, quando o número de indivíduos que não possuem uma determinada doença é maior do que aqueles que a possuem (IZBICKI; SANTOS, 2020). Situações como essa também foram descritas por Castro e Braga (2011) e Santos *et al.* (2019).

Conforme indicado na Figura 4.2, é possível verificar que, no banco de dados avaliado nesse estudo, há uma distribuição desbalanceada das classes das variáveis resposta, sobretudo, para os dados referentes à microcefalia. Nesses dados, há a predominância da classe “não”, ou seja, a maior parte dos municípios brasileiros não apresentou casos de microcefalia em decorrência do vírus Zika. A característica desses dados foi considerada no momento de

selecionar as melhores métricas para avaliação do desempenho dos modelos, a fim de não mascarar os baixos acertos na predição da classe “sim”.

Para diminuir o impacto do desbalanceamento dos dados nesse estudo, optou-se pela utilização da acurácia balanceada ao invés da acurácia como métrica de avaliação (GARCÍA; MOLLINEDA; SÁNCHEZ, 2009). Como indicado na Figura 4.3, a acurácia tende a apresentar valores maiores que a acurácia balanceada, o que ocorre em função de tal métrica, basicamente, considerar todos os acertos realizados pelo modelo, independente da classe da variável resposta. Na acurácia balanceada, por outro lado, os cálculos são realizados considerando a frequência de acertos em cada classe, o que faz com este índice seja menos influenciado pelo desbalanceamento dos dados e, conseqüentemente, forneça resultados mais confiáveis.

As diferenças entre os índices obtidos para acurácia e acurácia balanceada foram ainda mais visíveis ao se avaliar os resultados de predição da microcefalia, que apresentou um maior desbalanceamento dos dados. Ao se comparar os resultados indicados na Tabela 4.4 e na Figura 4.3 é possível verificar que, para o algoritmo GBM, os valores da acurácia e da acurácia balanceada foram de 0,88 e 0,67, respectivamente. Esse resultado reforça a necessidade de utilização de métricas apropriadas para cada banco de dados, a fim de garantir que os modelos sejam desenvolvidos de forma rigorosa e transparente.

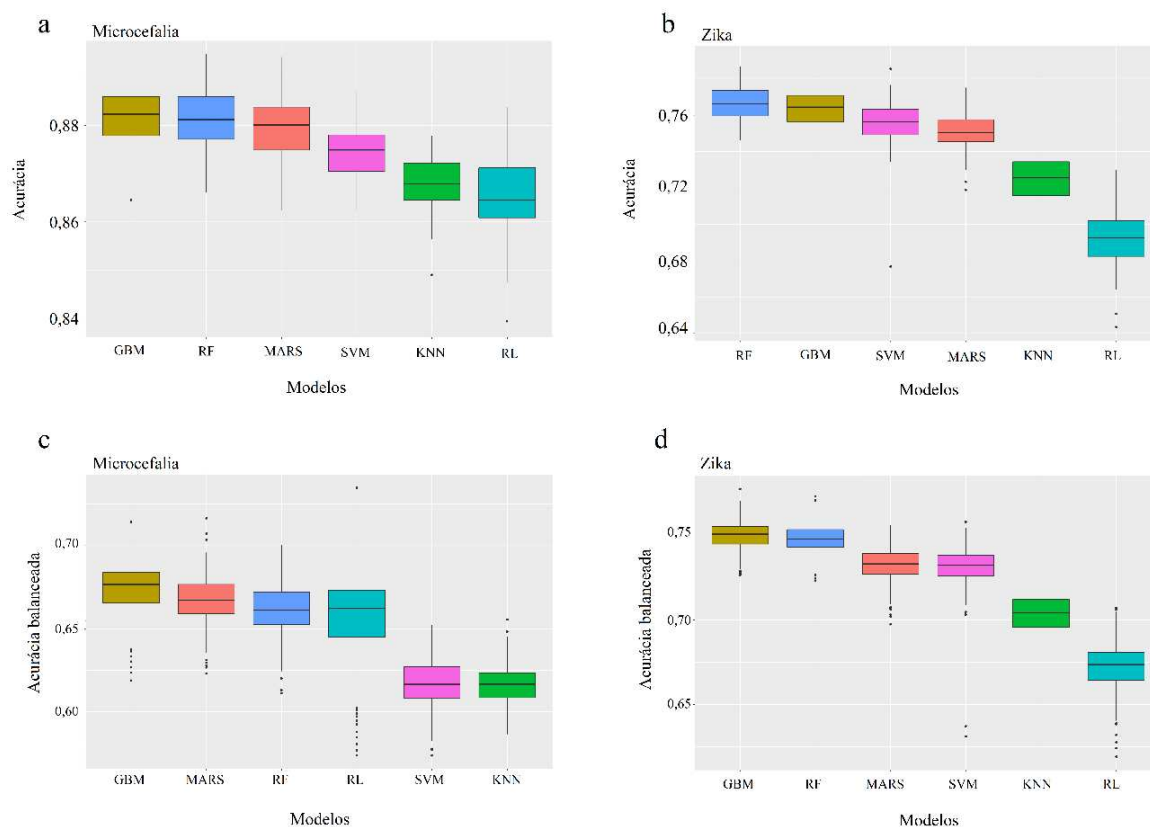


Figura 4.3. Boxplot da acurácia e acurácia balanceada dos modelos avaliados para predição de microcefalia (a e c) e Zika (b e d) nos municípios brasileiros.

4.3.4 Importância das covariáveis em cada modelo

Após a utilização dos diferentes métodos para seleção de covariáveis, o modelo RF, identificado como o mais apropriado para a predição de Zika, utilizou 82 das 141 covariáveis contidas no banco de dados inicial. Para microcefalia, o modelo GBM apresentou melhor desempenho ao utilizar 54 das 149 covariáveis iniciais. Todas as covariáveis incluídas nos modelos finais podem ser consultadas no Apêndice A.

Normalmente, as técnicas de previsão se concentram no maior número possível de covariáveis, adicionando uma variedade maior de dimensões que potencialmente tornam as previsões mais eficazes. A inclusão dessa variedade de fatores, no entanto, não garante um melhor desempenho de previsão. Fatores que não estão correlacionados com as variáveis dependentes diminuirão a precisão preditiva, pois podem ser apenas ruídos e irrelevantes para o modelo de predição. Nesse estudo, as variáveis selecionadas para avaliação da incidência de Zika e microcefalia foram identificadas por meio de consultas a estudos anteriores que

investigaram preditores que melhor explicaram a variabilidade da taxa de incidência de arboviroses.

Além da seleção de covariáveis, a utilização de algoritmos de aprendizado de máquina permite resumir o efeito de cada uma destas covariáveis por meio de uma métrica específica, denominada importância de variáveis. Após a constatação que os modelos RF e GBM apresentaram os melhores desempenhos na predição de Zika e microcefalia, respectivamente, foram avaliadas quais covariáveis contribuíram de forma mais expressiva para o desempenho dos modelos. Na Figura 4.4 são apresentados os rankings de importância das covariáveis, considerando as 30 covariáveis consideradas mais importantes para os modelos.

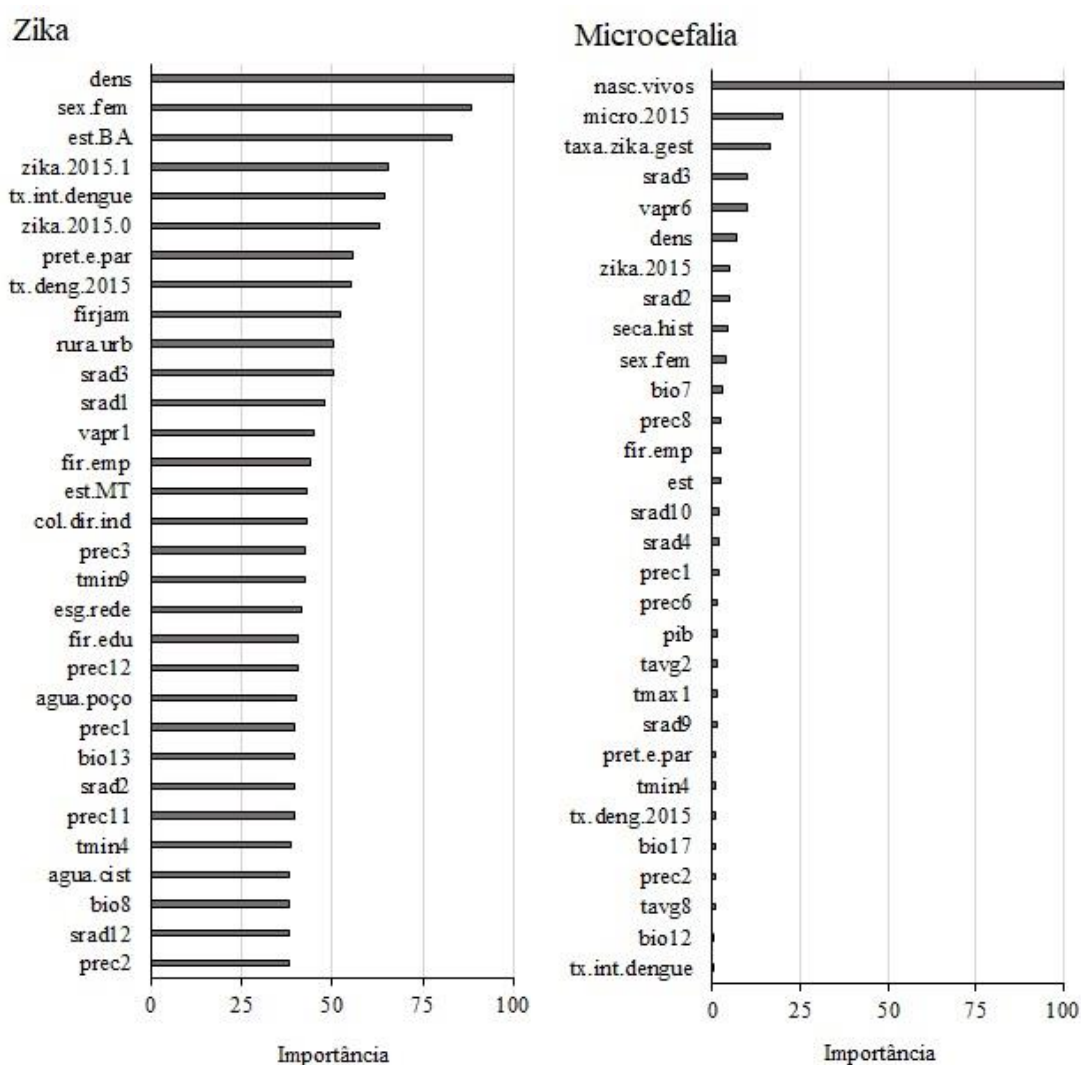


Figura 4.4. Importância das covariáveis na predição de Zika e microcefalia utilizando os modelos RF (Zika) e GBM (microcefalia).

Para o modelo de predição de Zika, covariáveis de todas as classes avaliadas estiveram presentes no ranking de importância, destacando-se a densidade demográfica do município e o percentual de responsáveis familiares do sexo feminino. Em relação à primeira, sabe-se que diversos fatores podem afetar a transmissão de Zika, incluindo fatores ambientais que determinam a adequabilidade de um local para a proliferação de mosquitos vetores. Nesse grupo de fatores está incluída a densidade demográfica de um determinado local. De acordo com Marcondes e Ximenes (2015), infestações de mosquitos vetores de arboviroses são mais recorrentes em áreas densamente povoadas, uma vez que, em tais áreas há uma alta probabilidade de contato humano-vetor. A referida informação também foi confirmada no estudo de Rodrigues *et al.* (2018), que identificaram que a densidade populacional humana afeta, diretamente, no número de *Aedes aegypti* em uma residência, o que aumenta a possibilidade de transmissão de arboviroses. A densidade também foi uma covariável presente no ranking de predição de microcefalia, o que pode ser justificado pelos mesmos motivos anteriormente citados.

A segunda covariável que apresentou destaque no ranking de importância dos preditores de Zika, e também esteve presente no ranking para microcefalia, foi o percentual de responsáveis familiares do sexo feminino. De modo geral, os casos notificados de Zika no Brasil ocorreram, predominante, em mulheres. Dados do ministério da saúde indicam que, no ano de 2016, do total de notificações de Zika no país, cerca de 67% se referiram ao sexo feminino. Para Coelho *et al.* (2016), existem duas hipóteses que podem explicar essa predominância dos casos de Zika em mulheres: a primeira está associada ao fato de que, de modo geral, as mulheres possuem o hábito de visitar médicos com mais frequência do que os homens; a segunda hipótese considera a existência de transmissão sexual do vírus Zika como a causa mais provável. Segundo os autores citados, os homens infectados com o vírus Zika podem transmitir a doença para mulheres por meio do sêmen. Paz-Bailay *et al.* (2017) também citam que a maior parte das transmissões de Zika por contato sexual se originaram de homens com infecção sintomática para suas parceiras. Considerando essa fonte “extra” de transmissão, seria justificável presumir uma maior incidência de Zika em mulheres.

Variáveis como taxa de incidência/internação de dengue, ocorrência de Zika e microcefalia no ano de 2015 e taxa de prevalência de Zika em gestantes, que se encontram nos rankings dos modelos, foram incluídas visando identificar municípios que já haviam declarado a existência de casos de dengue e de Zika. Tal informação atua como um indicativo que, naquele local, havia a circulação prévia do vetor e do agente infeccioso. Essa consideração é importante, uma vez que, o fato de uma área possuir adequabilidade para a distribuição de vetores não

significa, necessariamente, que a transmissão do vírus ocorrerá (COLÓN-GONZÁLEZ *et al.*, 2017; MESSINA *et al.*, 2016). Desse modo, a inclusão dessas variáveis nos modelos visou evitar a superestimação do número de municípios com maior possibilidade de apresentar casos de Zika ou microcefalia.

Dentre as variáveis socioeconômicas avaliadas nesse estudo, a localização do domicílio (urbano/ rural), o índice FIRJAN de desenvolvimento municipal e o percentual de pessoas com cor/raça preta ou parda foram observadas nos rankings de importância dos modelos. O índice FIRJAN fornece informações anuais a respeito do desenvolvimento socioeconômico de todos os municípios brasileiros. Tal índice tem sido utilizado como indicador socioeconômico em estudos que avaliaram determinantes para a incidência de doenças como dengue (CABRAL; FREITAS, 2012).

Em relação à localização dos domicílios, assim como verificado para a Zika nesse estudo, Teixeira e Medronho (2002), ao avaliar a relação entre indicadores sociodemográficos e a epidemia de dengue, verificaram que a proporção da população urbana/rural apresentou correlação significativa com a incidência de dengue. Segundo Silva Neto *et al.* (2022), as doenças tropicais negligenciadas, das quais fazem parte as arboviroses, afetam, sobretudo, as populações mais pobres que, por vezes, vivem em áreas rurais remotas, favelas urbanas ou zonas de conflito.

Diferente do observado no presente estudo, a variável porcentagem da população por raça e cor não demonstrou ser um fator de risco para a contração de arboviroses no estudo realizado por Rodrigues *et al.* (2018a). Por outro lado, Marinho *et al.* (2016) verificaram uma maior prevalência de nascidos vivos com microcefalia em filhos de mães pertencentes à raça/cor da pele preta ou parda.

É possível verificar que diversas condições ambientais e socioeconômicas podem afetar a transmissão do vírus Zika. Para Tunali *et al.* (2021), fatores socioeconômicos como pobreza, desigualdade social e habitação de baixa qualidade limitam o acesso da população a instalações apropriadas que, quando ausentes, podem promover a transmissão viral. A ausência de água canalizada na residência pode ser citada como um exemplo de infraestrutura que, quando ausente, pode contribuir para a promoção de criadouros ambientais favoráveis para mosquitos.

No modelo preditivo de Zika, duas variáveis relacionadas à forma de abastecimento de água estiveram presentes no ranking de importância, sendo elas: percentual de domicílios com abastecimento de água por meio de cisterna e de poço. De acordo com a definição de criadouros prováveis de mosquitos (BRASIL, 2001), o depósito de água em caixas d'água, tanques, poços e cisternas é considerado como possível criadouro ou foco de mosquitos. Tal afirmação

corroborar com os estudos de Dada *et al.* (2013), que citam a falta de acesso à rede de distribuição de água como um fator capaz de influenciar a incidência de arboviroses, uma vez que, resulta na necessidade de armazenamento doméstico de água, criando assim, locais propícios para a reprodução dos mosquitos.

As práticas de armazenamento de água em recipientes inapropriados podem ainda estar relacionadas a ocorrência de eventos extremos de seca. A incidência desses eventos foi considerada importante no modelo de predição de microcefalia. Sabe-se que a referida síndrome congênita apresentou maior incidência na região Nordeste do Brasil. Nessa mesma região, foi registrada uma seca prolongada nos quatro anos que antecederam a epidemia de microcefalia, o que pode ter contribuído para o aumento da densidade de vetores na região (ALMEIDA *et al.*, 2019).

Ainda em relação à maior incidência de microcefalia na região Nordeste, além dos fatores associados à proliferação do *Aedes aegypti*, um estudo indicou que a qualidade da água de abastecimento pode ter influenciado na predominância de microcefalia na região (PEDROSA *et al.*, 2020). Os referidos autores avaliaram a influência do consumo de água contaminada com cianotoxinas em cobaias contaminadas com o vírus Zika durante o período gestacional. Os resultados obtidos indicaram que o consumo da referida água potencializou anormalidades cerebrais derivadas do vírus Zika em camundongos recém-nascidos. Segundo os mesmos autores, o aumento da concentração de cianotoxinas nos reservatórios de abastecimento está, também, associado à ocorrência de eventos de seca e estiagem no Brasil, que acarretam a redução dos reservatórios a volumes críticos e favorecem a proliferação de cianobactérias. Desse modo, verifica-se que a incidência de eventos climáticos extremos, associada às condições inapropriadas de saneamento, também pode ter atuado como um cofator na incidência de microcefalia em municípios brasileiros.

Estudos indicam que a qualidade da água também desempenha um papel crucial na produtividade dos habitats de reprodução para mosquitos. Embora comumente seja citado que fêmeas de *Aedes aegypti* são atraídas para recipientes que contêm água com baixos índices de poluição, estudos recentes têm demonstrado que quanto maior o grau de poluição da água, maior a atratividade da mesma ao mosquito (TUNALI *et al.*, 2021). No presente estudo, as formas de coleta de esgoto e de lixo, que podem influenciar diretamente na qualidade da água de uma região, foram incluídas no ranking de importância no modelo de predição de Zika. As referidas informações possuem expressiva relevância para a orientação de ações de controle do *Aedes aegypti*, tendo em vista que estas apontam para a necessidade de inspeção de todos os

tipos de coleções de água, sobretudo, em áreas que apresentam baixos índices de coleta e tratamento de esgotos.

De acordo com Chitolina *et al.* (2016), a presença de coliformes termotolerantes em esgotos brutos é capaz de aumentar a atratividade para a oviposição do *Aedes aegypti*. Adicionalmente, segundo Beserra *et al.* (2010), a existência de altos níveis de turbidez e maior concentração de nutrientes em esgotos são fatores que contribuem na atração para oviposição do mosquito. Além disso, um estudo desenvolvido por MacKay *et al.* (2009), identificou a presença de *Aedes aegypti* em fossas sépticas. Os resultados do referido estudo despertam preocupação, sobretudo, no Brasil, onde grande parcela da população utiliza fossas sépticas e rudimentares como meio de tratamento ou afastamento de esgotos (IBGE, 2010).

Vale ressaltar que o banco de dados do CadÚnico, fonte dos dados sanitários utilizados no processo de predição, não possui indicadores de abastecimento de água, esgotamento sanitário e manejo de resíduos sólidos que possibilitem a realização de uma caracterização qualitativa dos serviços prestados. Embora o CadÚnico disponibilize dados que caracterizam o saneamento domiciliar da população de baixa renda no Brasil, tais dados não evidenciam se a solução é adequada, uma vez que, não revelam aspectos relativos à potabilidade da água e à existência de intermitências prolongadas ou racionamentos, bem como à existência de tratamento para os esgotos e destinação final ambientalmente adequada dos resíduos sólidos. Em virtude do exposto, a partir do referido banco de dados, não é possível realizar a caracterização do atendimento e déficit de acesso domiciliar em saneamento básico estabelecida pelo Plansab (BRASIL, 2013).

Por fim, as covariáveis climáticas avaliadas nesse estudo demonstraram uma forte influência no desempenho dos modelos de predição para Zika e microcefalia. Os modelos indicaram importância das covariáveis temperatura, precipitação, pressão de vapor de água e radiação solar. Baquero *et al.* (2018) também verificaram que a incorporação de variáveis meteorológicas melhorou o desempenho de modelos preditivos para dengue, cujo mosquito vetor é o mesmo da Zika. Tunali *et al.* (2021) também ressaltam que condições ambientais, sobretudo precipitação e umidade possuem um importante papel na propagação dos vetores e transmissão de arboviroses. A influência da precipitação sobre o crescimento populacional do *Aedes aegypti*, e possível propagação de arboviroses, também é relatada em diversos estudos, tais como Messina *et al.* (2016), Santos *et al.* (2020) e Souza *et al.* (2018).

Todas essas covariáveis climáticas desempenham uma função na incidência de Zika, seja pela influência na atividade e densidade dos vetores ou pelo oferecimento de condições propícias para a transmissão do vírus. Por exemplo, o estudo de Tesla *et al.* (2018) identificou

que a transmissão do vírus Zika é otimizada a 29 °C e possui uma faixa térmica de 22,7–34,7 °C. Desse modo, à medida que a temperatura de uma determinada região aumenta, o desenvolvimento de ovos e larvas dos mosquitos também poderá aumentar. Contudo conforme destacado por Mesquita, Rosa e Borges (2021), é importante ressaltar que os efeitos do clima na transmissão de arboviroses não são lineares. Desse modo, o efeito das mudanças climáticas pode aumentar o potencial de transmissão de doenças em alguns contextos, mas, também, diminuí-lo em outros.

4.4 Considerações finais

Nesse estudo, modelos preditivos foram utilizados para estimar a ocorrência de agravos relacionados à saúde pública (Zika e microcefalia). Foi verificado que os algoritmos de aprendizado de máquina apresentaram potencial para identificação de relações complexas e não-lineares nos dados avaliados.

O modelo de aprendizado de máquina *Random forest* apresentou os melhores resultados na predição de Zika e mostrou possuir capacidade de identificar os municípios com incidência dessa arboviroses a partir dos preditores avaliados. Por outro lado, o modelo para predição de microcefalia apresentou resultados inferiores aos observados para o modelo de Zika, o que pode estar associado à maior complexidade de se modelar uma síndrome congênita, uma vez que, sua incidência não está associada apenas à presença da doença Zika.

Os modelos avaliados indicaram forte influência das condições demográficas, socioeconômicas, climáticas e sanitárias dos municípios na predição da ocorrência de Zika. Diante disso, a avaliação de tais fatores mostra-se relevante para o monitoramento epidemiológico de infecções associadas ao vírus Zika.

O uso de modelos de aprendizado de máquina associado a dados municipais, disponibilizados publicamente e de forma gratuita, demonstrou possuir um forte potencial para identificar onde é provável que o vírus Zika se espalhe. Tal informação pode auxiliar na prevenção de complicações envolvendo gestantes e recém-nascidos. Desse modo, acredita-se que os resultados obtidos nesse estudo podem ser utilizados como uma ferramenta de apoio a ações, estratégias e políticas de prevenção e monitoramento pela vigilância em saúde.

4.5 Referências

- ALMEIDA, I. M. L. M. *et al.* Clinical and epidemiological aspects of microcephaly in the state of Piauí, northeastern Brazil, 2015–2016. **Jornal de Pediatria**, v. 95, n. 4, p. 466–474, 2019.
- ASSIS, É. G. DE; ZÁRATE, L. E.; NOBRE, C. N. Classification of Newborns with Congenital Syndrome Associated with Zika Virus Infection Using Machine Learning. **Anais do Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)**, p. 57–64, 2021.
- BAQUERO, O. S.; SANTANA, L. M. R.; CHIARAVALLOTI-NETO, F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. **PLOS ONE**, v. 13, n. 4, p. e0195065, 2018.
- BARBOZA, L. A. *et al.* Assessing dengue fever risk in Costa Rica by using climate variables and machine learning techniques. **PLOS Neglected Tropical Diseases**, v. 17, n. 1, p. e0011047, 2023.
- BESERRA, E. B. *et al.* Efeito da qualidade da água no ciclo de vida e na atração para oviposição de *Aedes aegypti* (L.) (Diptera: Culicidae). **Neotropical Entomology**, v. 39, n. 6, p. 1016–1023, 2010.
- BIAU, G. Analysis of a Random Forests Model. **Journal of Machine Learning Research**, v. 13, p. 1063–1095, 2012.
- BRASIL. **Orientações integradas de vigilância e atenção à saúde no âmbito da emergência de saúde pública de importância nacional**. Ministério da Saúde, 2017, 158p.
- BRASIL. **Situação epidemiológica da síndrome congênita associada à infecção pelo vírus Zika: Brasil, 2015 a 2022, até a SE 31**. Ministério da Saúde, v. 53, n. 35, p. 1–14, 2022.
- BREIMAN, L. ST4_Method_Random_Forest. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.
- CABRAL, J. A.; FREITAS, M. V. Distribuição espacial e determinantes socioeconômicos e demográficos da dengue nos municípios brasileiros. **Revista Brasileira de Estudos Regionais e Urbanos**, v. 06, n. 1, p. 81–95, 2012.
- CALVET, G. *et al.* Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study. **The Lancet. Infectious diseases**, v. 16, n. 6, p. 653–660, 2016.
- CASTRO, C. L.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, v. 22, n. 5, p. 441–466, 2011.
- CHEN, R. C.; MANONGGA, W. E.; DEWI, C. Recursive Feature Elimination for Improving Learning Points on Hand-Sign Recognition. **Future Internet 2022**, Vol. 14, Page 352, v. 14, n. 12, p. 352, 2022.

CHITOLINA, R. F. *et al.* Raw sewage as breeding site to *Aedes* (*Stegomyia*) *aegypti* (Diptera, culicidae). **Acta Tropica**, v. 164, p. 290–296, 2016.

COELHO, F. C. *et al.* Higher incidence of Zika in adult women than adult men in Rio de Janeiro suggests a significant contribution of sexual transmission from men to women. **International Journal of Infectious Diseases**, v. 51, p. 128–132, 2016.

COLÓN-GONZÁLEZ, F. J. *et al.* After the epidemic: Zika virus projections for Latin America and the Caribbean. **PLoS Neglected Tropical Diseases**, v. 11, n. 11, p. 1–19, 2017.

CUTLER, D. R. *et al.* Random forests for classification in ecology. **Ecology**, v. 88, n. 11, p. 2783–2792, 2007.

DADA, N. *et al.* Relationship between *Aedes aegypti* production and occurrence of *Escherichia coli* in domestic water storage containers in rural and sub-urban villages in Thailand and Laos. **Acta Tropica**, v. 126, n. 3, p. 177–185, 2013.

DHAKA, A.; SINGH, P. Comparative analysis of epidemic alert system using machine learning for dengue and chikungunya. Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, **Data Science and Engineering**, p. 798–804, 2020.

DIAS, R. L. S. *et al.* Machine learning models applied to TSS estimation in a reservoir using multispectral sensor onboard to RPA. **Ecological Informatics**, v. 65, p. 101414, 2021.

FRIEDMAN, J. H. Multivariate Adaptive Regression Splines. **The Annals of Statistics**, v. 19, n. 1, p. 1–67, 1991.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). **Annals of Statistics**, v. 28, n. 2, p. 337–407, 2000.

FUKUNAGA, K.; NARENDRA, P. M. A Branch and Bound Algorithm for Computing k-Nearest Neighbors. **IEEE Transactions on Computers**, v. C-24, n. 7, p. 750–753, 1975.

GALASSO, J.; CAO, D. M.; HOCHBERG, R. A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. **Chaos, Solitons & Fractals**, v. 156, p. 111779, 2022.

GARCIA, L. P. **Epidemia do vírus Zika e microcefalia no Brasil: emergência, evolução e enfrentamento**. Rio de Janeiro: Ipea, 2018.

GARCÍA, V.; MOLLINEDA, R. A.; SÁNCHEZ, J. S. *abn.* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in **Bioinformatics**), v. 5524 LNCS, p. 441–448, 2009.

GOLDSTEIN, B. A.; NAVAR, A. M.; CARTER, R. E. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. **European Heart Journal**, v. 38, n. 23, p. 1805–1814, 2017.

GUNN, S. R. Support Vector Machines for Classification and Regression. **ISIS technical report**, v. 14, n. 1, p. 5–16, 1998.

GUO, P. *et al.* Developing a dengue forecast model using machine learning: A case study in China. **PLOS Neglected Tropical Diseases**, v. 11, n. 10, p. e0005973, 2017.

HAN, M. *et al.* Novel integrated systems for controlling and prevention of mosquito-borne diseases caused by poor sanitation and improper water management. **Journal of Environmental Chemical Engineering**, v. 4, n. 4, p. 3718–3723, 2016.

HARRELL JR, F. E. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2a ed. Cham: **Springer International Publishing**, 2015.

ISMAIL, S. *et al.* The practicality of Malaysia dengue outbreak forecasting model as an early warning system. **Infectious Disease Modelling**, v. 7, n. 3, p. 510–525, 2022.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. 1. ed. São Carlos: Rafael Izbicki, 2020.

JIANG, D. *et al.* Mapping the transmission risk of Zika virus using machine learning models. **Acta Tropica**, v. 185, n. June, p. 391–399, 2018.

KUBAT, M. **An Introduction to Machine Learning**. Cham, Switzerland: Springer International Publishing, 2017.

KUHN, M. *et al.* **Package “caret”: Classification and regression training**. Disponível em: <<https://topepo.github.io/caret/>>. Acesso em: 30 maio. 2022.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. 1. ed. New York: Springer New York, 2013.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data - **Biometrics**, v. 33, n. 1, p. 159–174, 1977.

LESSLER, J. *et al.* Assessing the global threat from Zika virus. **Science**, v. 353, n. 6300, 2016.

LIMA, T. P. F. *et al.* Previsão de óbito e importância de características clínicas em idosos com COVID-19 utilizando o Algoritmo Random Forest. **Revista Brasileira de Saúde Materno Infantil**, v. 21, n. suppl 2, p. 445–451, 2021.

LOWY, I. **Zika no Brasil: história recente de uma epidemia**. Rio de Janeiro: Editora Fiocruz, 2019.

MACKAY, A. J. *et al.* Dynamics of *Aedes aegypti* and *Culex quinquefasciatus* in Septic Tanks. **Journal of the American Mosquito Control Association**, v. 25, n. 4, p. 409–416, 2009.

MARCONDES, C. B.; XIMENES, M. DE F. F. DE M. Zika virus in Brazil and the danger of infestation by *Aedes* (*Stegomyia*) mosquitoes. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 49, n. 1, p. 4–10, 2015.

MARINHO, F. *et al.* Microcefalia no Brasil: prevalência e caracterização dos casos a partir do Sistema de Informações sobre Nascidos Vivos (Sinasc), 2000-2015. **Epidemiologia e serviços de saúde : revista do Sistema Unico de Saude do Brasil**, v. 25, n. 4, p. 701–712, 2016.

MESQUITA, T. C. R.; ROSA, A. P.; BORGES, A. C. Mudanças Climáticas e seu Impacto na Incidência de Arboviroses: Uma Revisão Sistemática de Estudos Recentes. **Revista Brasileira de Geografia Física**, v. 14, n. 6, p. 3361–3377, 2021.

MESSINA, J. P. *et al.* Mapping global environmental suitability for Zika virus. **eLife**, v. 5, n. APRIL2016, 2016.

MORGAN, J.; STRODE, C.; SALCEDO-SORA, J. E. Climatic and socio-economic factors supporting the co-circulation of dengue, Zika and chikungunya in three different ecosystems in Colombia. **PLoS Neglected Tropical Diseases**, v. 15, n. 3, 2021.

SILVA NETO, S. R. D. *et al.* Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review. **PLoS Neglected Tropical Diseases**, v. 16, n. 1, p. e0010061, 2022.

OLIVEIRA MELO, A. S. *et al.* Zika virus intrauterine infection causes fetal brain abnormality and microcephaly: tip of the iceberg? **Ultrasound in Obstetrics & Gynecology**, v. 47, n. 1, p. 6–7, 2016.

OMS, O. M. DA S. Comprehensive Guidelines for Prevention and Control of Dengue and Dengue Haemorrhagic Fever. Edição revisada. **WHO Regional Office for South-East Asia**, 2011.

ONG, J. *et al.* Mapping dengue risk in Singapore using Random Forest. **PLoS Neglected Tropical Diseases**, v. 12, n. 6, p. 1–12, 2018a.

ONG, J. *et al.* Mapping dengue risk in Singapore using Random Forest. **PLoS Neglected Tropical Diseases**, v. 12, n. 6, p. e0006587, 2018b.

OSORIO-DE-CASTRO, C. G. S. *et al.* The Zika Virus Outbreak in Brazil: Knowledge Gaps and Challenges for Risk Reduction. **American Journal of Public Health**, v. 107, n. 6, p. 960, 2017.

PAZ-BAILEY G. *et al.* Persistence of Zika Virus in Body Fluids - Final Report. **The New England Journal of Medicine**, v. 379, p. 1234-1243, 2017.

PECHT, M. G.; KANG, M. **Prognostics and Health Management of Electronics**. Chichester, UK: Wiley, 2018.

PEDROSA, C. S. G. *et al.* The cyanobacterial saxitoxin exacerbates neural cell death and brain malformations induced by Zika virus. **PLoS Neglected Tropical Diseases**, p. 1–13, 2020.

PERROTTA, D. *et al.* Comparing sources of mobility for modelling the epidemic spread of Zika virus in Colombia. **PLoS Neglected Tropical Diseases**, v. 16, n. 7, p. e0010565, 2022.

PRASAD, S. *et al.* ADM1 modeling of UASB treating domestic wastewater in Nepal. **Renewable Energy**, v. 95, p. 263–268, 2016.

QUEIROZ, E. R. S.; MEDRONHO, A. M. Spatial analysis of the incidence of Dengue, Zika and Chikungunya and socioeconomic determinants in the city of Rio de Janeiro, **Brazil**. **Epidemiology & Infection**, v. 149, p. e188, 2021.

RAHMAN, M. S. *et al.* Mapping the spatial distribution of the dengue vector *Aedes aegypti* and predicting its abundance in northeastern Thailand using machine-learning approach. **One Health**, v. 13, p. 100358, 2021.

ROCHA, F. P.; GIESBRECHT, M. Machine learning algorithms for dengue risk assessment: a case study for São Luís do Maranhão. **Computational and Applied Mathematics**, v. 41, n. 8, p. 1–27, 2022.

RODRIGUES, N. C. P. *et al.* Risk factors for arbovirus infections in a low-income community of Rio de Janeiro, Brazil, 2015-2016. **PLoS ONE**, v. 13, n. 6, p. 2015–2016, 2018a.

RODRIGUES, N. C. P. *et al.* Risk factors for arbovirus infections in a low-income community of Rio de Janeiro, Brazil, 2015-2016. **PLOS ONE**, v. 13, n. 6, p. e0198357, 2018b.

SADEGHIEH, T. *et al.* Zika virus outbreak in Brazil under current and future climate. **Epidemics**, v. 37, p. 100491, 2021.

SANTOS, E. P. *et al.* A Machine Learning approach to reconstruct cloudy affected vegetation indices imagery via data fusion from Sentinel-1 and Landsat 8. **Computers and Electronics in Agriculture**, v. 194, p. 106753, 2022.

SANTOS, H. G. *et al.* Machine learning para análises preditivas em saúde: exemplo de aplicação para prever óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**, v. 35, n. 7, p. e00050818, 2019.

SANTOS, I. C. DA S. *et al.* The influence of meteorological variables on the oviposition dynamics of *Aedes aegypti* (Diptera: Culicidae) in four environmentally distinct areas in northeast Brazil. **Memórias do Instituto Oswaldo Cruz**, v. 115, n. 6, p. 1–10, 2020.

SARKER, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**, v. 2, n. 3, p. 1–21, 2021.

SARMA, D. *et al.* Dengue Prediction using Machine Learning Algorithms. **IEEE Region 10 Humanitarian Technology Conference**, R10-HTC, 2020.

SILVA, C. C. *et al.* Forecasting Dengue, Chikungunya and Zika cases in Recife, Brazil: a spatio-temporal approach based on climate conditions, health notifications and machine learning. **Research, Society and Development**, v. 10, n. 12, p. e452101220804–e452101220804, 2021.

SILVA, C. C. *et al.* Spatiotemporal forecasting for dengue, chikungunya fever and Zika using machine learning and artificial expert committees based on meta-heuristics. **Research on Biomedical Engineering**, v. 38, n. 2, p. 499–537, 2022.

SILVA, J. C. B.; MACHADO, C. J. S. Associations between dengue and socio-environmental variables in capitals of the Brazilian northeast by cluster analysis. **Ambiente & Sociedade**, v. 21, p. 1332, 2018.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados - Com Aplicações em R**. 1. ed. Rio de Janeiro: LTC, 2021.

SOUZA, C. M. P. DE *et al.* Multivariate Analysis and Machine Learning in Properties of Ultisols (Argissolos) of Brazilian Amazon. **Revista Brasileira de Ciência do Solo**, v. 42, n. 0, 2018.

SOUZA, M. L. A. *et al.* Bayesian estimates for the mapping of dengue hotspots and estimation of the risk of disease epidemic in Northeast Brazil. **Urban Climate**, v. 26, n. September, p. 198–211, 2018.

STAVROPOULOS, G. *et al.* **Random Forest and Ensemble Methods**. 2. ed. Elsevier Inc., 2020.

TESLA, B. *et al.* Temperature drives Zika virus transmission: evidence from empirical and mathematical models. **Proceedings of the Royal Society B**, v. 285, n. 1884, 2018.

TUNALI, M. *et al.* A review exploring the overarching burden of Zika virus with emphasis on epidemiological case studies from Brazil. **Environmental Science and Pollution Research**, v. 28, n. 40, p. 55952–55966, 2021.

VEIGA, R. V. *et al.* Classification algorithm for congenital Zika Syndrome: characterizations, diagnosis and validation. **Scientific Reports** 2021 11:1, v. 11, n. 1, p. 1–7, 2021.

XU, Y. *et al.* Assessing the risk of spread of Zika virus under current and future climate scenarios. **Biosafety and Health**, v. 4, n. 3, p. 193–204, 2022.

YAVARI NEJAD, F.; VARATHAN, K. D. Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction. **BMC Medical Informatics and Decision Making**, v. 21, n. 1, p. 1–12, 2021.

ZHAO, N.; CHARLAND, K.; CARABALI, M.; NSOESIE, E. O.; *et al.* Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. **PLOS Neglected Tropical Diseases**, v. 14, n. 9, p. e0008056, 2020.

5 CONSIDERAÇÕES GERAIS

O estudo desenvolvido na presente tese baseou-se na realização de uma análise multi-abordagem acerca das epidemias do vírus Zika e microcefalia no Brasil. Por meio dos resultados obtidos, foi possível estabelecer uma relação entre a incidência de Zika e microcefalia e as características sanitárias, climáticas, ambientais e socioeconômicas dos municípios brasileiros. Desse modo, constatou-se que a junção de fatores tradicionalmente avaliados em estudos de arboviroses, tais como climáticos e socioeconômicos, com fatores sanitários e ambientais pode contribuir para uma identificação mais acurada de áreas de risco epidêmico.

As maiores taxas de incidência de Zika e microcefalia ocorreram, majoritariamente, nos municípios que apresentam menores índices socioeconômicos e de saneamento no Brasil. Tal informação evidencia a necessidade de uma revisão das estratégias de controle de arboviroses adotadas no país. Atualmente, as ações governamentais de prevenção às arboviroses baseiam-se, predominantemente, no controle vetorial, por meio da atuação de agentes de saúde municipais e da mobilização popular. No entanto, a forte influência dos fatores sanitários na incidência de Zika e microcefalia demonstram a necessidade de se garantir, antes de tudo, serviços públicos de saneamento básico de qualidade, sendo estes fundamentais para o controle de arboviroses.

O uso de modelos de aprendizado de máquina associado a dados municipais, disponibilizados publicamente e de forma gratuita, demonstrou possuir um forte potencial para identificar onde é provável que o vírus Zika se espalhe. Tal informação pode auxiliar na prevenção de complicações envolvendo gestantes e recém-nascidos. Além disso, o processamento da maior parte das análises realizadas não apresentou elevado custo computacional, sendo utilizado um microcomputador pessoal¹ para o processamento e armazenamento dos dados. No entanto, as análises que possuíam caráter mais complexo demandaram a utilização do Laboratório de Geoprocessamento da UFV, a fim de obter os resultados de forma mais ágil.

Desse modo, acredita-se que os resultados obtidos nesse estudo possam fornecer informações valiosas para a tomada de decisão em saúde pública e definição de estratégias para prevenção de epidemias associadas ao vírus Zika. O planejamento antecipado de ações para áreas com potencial risco de desenvolvimento de doenças é fundamental para a formulação de programas bem-sucedidos de saúde pública e alocação de recursos para prevenção e controle da doença em áreas específicas.

Vale ressaltar que o trabalho desenvolvido também apresenta limitações que precisam ser consideradas. A principal delas está associada à qualidade dos dados relativos aos casos de Zika e microcefalia, contidos nos sistemas de informação brasileiros, e às falhas no processo de vigilância em saúde. Ademais, além do fato de as epidemias associadas ao vírus ainda serem consideradas recentes, a Zika integra a lista de doenças tropicais negligenciadas, o que também resulta em obstáculos para o desenvolvimento de estudos acerca da referida temática.

¹Foi utilizado um microcomputador pessoal que possui SSD de 480GB, memória RAM de 16GB e processador Intel core i5, 7th Gen.

Apêndice A

Covariáveis incluídas nos modelos RF e GBM para a predição de Zika e microcefalia nos municípios brasileiros, respectivamente.

Classe da variável	Zika	Microcefalia
Sanitária	col.dir.ind, esg.rede, agua.poço, agua.cist, col.terre, esg.rio, esg.fos.rud, esg.vala, agua.rede, esg.foss, col.rio,	agua.cist, col.terre, esg.foss, esg.rede
Socioeconômica	sex.fem, pret.e.par, firjam, rura.urb, fir.emp, fir.edu, sem.inst, ext.pob, pop.cad, fir.sau, sup.comp, pib	sex.fem, fir.emp, pib, pret.e.par, nasc.vivos, pop.cad, fir.sal, ext.pob, fir.edu, sup.comp
Climáticas	srad3, srad1, vapr1, prec3, tmin9, prec12, prec1, bio13, srad2, prec11, tmin4, bio8, srad12, prec2, vapr6, bio7, tavg8, bio2, bio12, tavg9, srad4, bio4, srad11, srad9, bio9, tavg10, srad7, bio19, srad6, srad5, srad10, tmax1, bio3, tmin1, bio15, tavg2, bio5, prec5, prec6, prec10, bio17, wind1, seca.hist, prec4, prec8, tavg1, bio18, wind11, wind7, bio14, seca2016, cheia.hist, prec9, umidade	srad3, vapr6, srad2, seca.hist, bio7, prec8, srad10, srad4, prec1, prec6, tavg2, tmax1, srad9, tmin4, bio17, prec2, tavg8, bio12, prec4, tavg9, bio2, tmin1, prec10, tmin9, bio3, prec12, bio4, bio5, bio9, srad5, tavg1, tavg10
Outras	dens, est, Zika.2015, tx.int.dengue, tx.deng.2015, macro	micro.2015, taxa.Zika.gest, dens, Zika.2015, est, tx.deng.2015, tx.int.dengue, macro