



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CECILIA CORDEIRO DA SILVA

**Uma metodologia para construção de preditores de doenças baseada em
Aprendizado de Máquina, Computação Bioinspirada e Análise Espaço-Temporal**

Recife

2022

CECILIA CORDEIRO DA SILVA

**Uma metodologia para construção de preditores de doenças baseada em
Aprendizado de Máquina, Computação Bioinspirada e Análise Espaço-Temporal**

Trabalho submetido ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutora em Ciência da Computação.

Área de Concentração: Engenharia da Computação

Orientador: Abel Guilhermino da Silva Filho

Coorientador: Wellington Pinheiro dos Santos

Recife

2022

Catálogo na fonte
Bibliotecária Luiza de Oliveira, CRB4-11316

S586u Silva, Cecília Cordeiro da

Uma metodologia para construção de preditores de doenças baseada em aprendizado de máquina, computação bioinspirada e análise espaço-temporal / Cecília Cordeiro da Silva. – 2022.

193 f.: il., tab.

Orientador: Abel Guilhermino da Silva Filho.

Tese (Doutorado) – Universidade Federal de Pernambuco. CIN, Ciência da Computação, Recife, 2022.

Inclui referências. Inclui apêndice.

1. Arboviroses. 2. Epidemiologia Digital. 3. Endemias. 4. Predição. 5. Aprendizado de Máquina. 6. Computação Bioinspirada I. Silva Filho, Abel Guilhermino (orientador). II. Título.

004

CDD (23. ed.)

UFPE - CCEN 2022-141

Cecilia Cordeiro da Silva

“Uma metodologia para construção de preditores de doenças baseada em Aprendizado de Máquina, Computação Bioinspirada e Análise Espaço-Temporal”

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Engenharia da Computação

Aprovado em: 20/06/2022.

Orientador: Prof. Dr. Abel Guilhermino da Silva Filho

BANCA EXAMINADORA

Prof. Dr. Manoel Eusébio de Lima
Centro de Informática/UFPE

Profa. Dra. Ana Lúcia Bezerra Candeias
Departamento de Engenharia Cartográfica/UFPE

Profa. Dra. Giselle Machado Magalhães Moreno
Instituto de Astronomia, Geofísica e Ciências Atmosféricas/USP

Prof. Dr. Tiago Lima Massoni
Departamento de Sistemas e Computação/UFCG

Prof. Dr. João Fausto Lorenzato de Oliveira
Escola Politécnica de Pernambuco/UPE

À minha família, por todo amor e carinho que me é dedicado.
Sem vocês, nada seria possível.

AGRADECIMENTOS

A Deus.

Ao CNPq, pela bolsa de doutorado que me foi dada, o que possibilitou que eu pudesse me dedicar exclusivamente à minha pesquisa.

Ao meu orientador, Prof. Abel Guilhermino, por ter aceitado me orientar e por todo o apoio que me foi dado.

A Clarisse Lins de Lima, Ana Clara Gomes da Silva e a toda a Equipe do Grupo de Pesquisas em Computação Biomédica, que me ajudaram no desenvolvimento das aplicações e na execução dos experimentos computacionais.

À minha mãe Valdecir, por todo amor e carinho.

À minha tia Cirene (*in memorian*), por todo amor, carinho e paciência.

A meu irmão Emanuel, por sua amizade e companheirismo.

Às minhas peludinhas Lila, Pipoca (*in memorian*) e Flor, pelos momentos de fofura e pelo amor incondicional.

A meu marido Wellington, pelo apoio incondicional e pelo otimismo (às vezes exagerado), principalmente nos momentos de incerteza, muito comuns para quem tenta trilhar novos caminhos na vida.

“Ninguém ignora tudo. Ninguém sabe tudo. Todos nós sabemos alguma coisa. Todos nós ignoramos alguma coisa. Por isso aprendemos sempre.” (FREIRE, 2022).

RESUMO

Em um mundo cada vez mais conectado por meio de vias físicas e virtuais, pessoas e dados circulam com cada vez mais facilidade. As mudanças demográficas e o intenso fluxo migratório das zonas rurais para as regiões urbanas geraram um crescimento desordenado das cidades. Isso combinado à ausência de boas condições de saneamento básico contribui para a proliferação do vetor. A emergência de surtos epidêmicos, como a dengue, a febre chikungunya, a zika e outras doenças tem contribuído para construir um cenário cada vez mais desafiador. A recente pandemia de Covid-19 trouxe grandes mudanças em escala mundial. Nesse cenário, cresceu fortemente o interesse por técnicas para predição espacial e temporal da distribuição de doenças a partir de tecnologias como a Internet das Coisas, aprendizado de máquina e múltiplas bases de dados. Este trabalho tem como objetivo geral propor uma metodologia para construção de preditores capazes de prever a distribuição espaço-temporal de doenças e apontar os fatores mais relevantes para a predição a partir de uma arquitetura baseada no acesso a múltiplas bases de dados. Para validação da proposta, foi adotada como estudo de caso a predição de casos de arboviroses por meio de séries históricas georreferenciadas de informações climáticas e ambientais utilizando técnicas de aprendizado de máquina. Essas informações são coletadas de múltiplos bancos de dados georreferenciados, previamente construídos a partir da coleta de informações por redes de sensores e do Sistema Único de Saúde. Foram utilizadas informações da Cidade do Recife, de casos de arboviroses (dengue, chikungunya e zika) de 2013 a 2016, e informações climáticas e ambientais do mesmo período, da APAC e do INMET. Os sistemas de predição de doenças construídos utilizando a metodologia proposta neste trabalho também devem ser capazes de apontar os fatores mais relevantes para a predição por meio do Comitê de Especialistas Artificiais, proposto neste trabalho e composto de um conjunto de algoritmos de seleção de atributos baseados em métodos de otimização por Computação Evolucionária. O Comitê de Especialistas Artificiais decide por votação. Os melhores resultados de predição de casos foram obtidos com regressão por Random Forest. Os valores do coeficiente de correlação de Pearson foram superiores a 0,99, enquanto o RMSE (%) se manteve inferior a 6%. Os índices de Kendall e Spearman também se mantiveram altos: seus valores foram superiores a 0,99 para Spearman e maiores que 0,90 para Kendall. O desempenho superior da Random Forest mostra que o problema de regressão é de difícil generalização, dado que a Random Forest é baseada em comitês de

árvores de decisão e a regressão é realizada por uma média ponderada dos resultados das diferentes árvores de decisão que compõem o modelo.

Palavras-chaves: Arboviroses. Epidemiologia Digital. Endemias. Predição. Aprendizado de Máquina. Computação Bioinspirada.

ABSTRACT

In a world increasingly connected through physical and virtual pathways, people and data move more and more easily. Demographic changes and the intense migratory flow from rural areas to urban regions have generated disorderly growth in cities. This fact, combined with the absence of good basic sanitation conditions, contributes to the proliferation of disease vectors. The emergence of epidemic outbreaks such as dengue, chikungunya fever, zika and other diseases has contributed to building an increasingly challenging scenario. The recent Covid-19 pandemic has brought about major changes on a world scale. In this scenario, the interest in techniques for spatial and temporal prediction of disease distribution using technologies such as the Internet of Things, machine learning and multiple databases has grown strongly. The general objective of this work is to propose a methodology for building predictors capable of predicting the spatio-temporal distribution of diseases and pointing out the most relevant factors for prediction based on an architecture based on access to multiple databases. To validate the proposal, the prediction of arboviruses cases through georeferenced historical series of climatic and environmental information using machine learning techniques was adopted as a case study. This information is collected from multiple georeferenced databases, previously built from the collection of information by sensor networks and the Unified Health System. We used information from the City of Recife, from cases of arboviruses (dengue, chikungunya and zika) from 2013 to 2016; and climate and environmental information from the same period, from APAC and INMET. Disease prediction systems built using the methodology proposed in this work should also be able to point out the most relevant factors for prediction through the Artificial Experts Committee, proposed in this work and composed of a set of feature selection algorithms based on Evolutionary Computation optimization methods. The Artificial Experts Committee decides by vote. The best case prediction results were obtained with Random Forest regression. Pearson's correlation coefficient values were above 0.99, while the RMSE (%) remained below 6%. The Kendall and Spearman indices also remained high: their values were above 0.99 for Spearman and above 0.90 for Kendall. The superior performance of Random Forest shows that the regression problem is difficult to generalize, given that Random Forest is based on decision tree committees and the regression is performed by a weighted average of the results of the different decision trees that make up the model.

Keywords: Arboviruses. Digital Epidemiology. Endemics. Forecasting. Machine Learning. Bioinspired Computing.

LISTA DE FIGURAS

Figura 1 – Diagrama esquemático do processo de integração de diferentes bases de geodados em um GIS. As transformações geométricas consistem em rotações e translações para que os modelos sejam ajustados e correspondentes às mesmas posições geográficas.	37
Figura 2 – Visualização da tela de trabalho do Quantum GIS, com mapa e tabela de geodados exibidos simultaneamente.	38
Figura 3 – Potenciais criadouros de tipo A1.	47
Figura 4 – Potenciais criadouros de tipo A2.	47
Figura 5 – Potenciais criadouros de tipo B.	48
Figura 6 – Potenciais criadouros de tipo C.	49
Figura 7 – Potenciais criadouros de tipo D1.	49
Figura 8 – Potenciais criadouros de tipo D2.	50
Figura 9 – Potenciais criadouros de tipo E.	51
Figura 10 – Representação gráfica de uma rede neural Perceptron Multicamadas com n entradas e três camadas, ou seja, duas camadas escondidas e uma camada de saída.	53
Figura 11 – Arquitetura geral de uma máquina de vetor de suporte com uma única saída: o cálculo dos kernels com as entradas e os vetores de suporte pode ser visto como uma camada escondida com um tipo especial de neurônio, enquanto o cálculo da saída tem um neurônio usual linear. A camada escondida teria n_S neurônios cujos vetores de pesos seriam os vetores de suporte $\mathbf{x}_i$, para $i = 1, 2, \dots, n_S$, e a saída determinada pelo kernel K	55
Figura 12 – Esquema geral de uma árvore de decisão, indicando o nó raiz, os nós de decisão e os nós folhas.	60
Figura 13 – Esquema geral de uma Random Forest como comitê de árvores de decisão.	62

Figura 14 – Fluxograma de um algoritmo genético. A população inicial é gerada aleatoriamente. Em cada geração, os indivíduos mais aptos, ou seja, os que retornam melhores valores da função objetivo, são selecionados para serem pais. Os pais geram filhos por cruzamento e mutação aleatória dos genes. O processo se repete até ser atingido o máximo de iterações.	65
Figura 15 – Representação geral de um algoritmo de otimização por enxame de partículas: as partículas iniciam com posição e velocidade aleatórias. A cada iteração, a qualidade da posição de cada partícula é avaliada com a função objetivo. Se o problema for de maximização, a partícula com a posição de maior valor da função objetivo é escolhida como a líder global. Cada partícula também observa, na sua vizinhança, qual é a sua líder local. Então são ajustadas a velocidade e a posição de cada partícula. O processo se repete pelo máximo de iterações.	67
Figura 16 – Representação geral de um algoritmo de colônia artificial de abelhas: as abelhas empregadas buscam novas soluções na vizinhança das fontes de alimento; as abelhas observadoras avaliam as fontes de alimentam para decidir quais devem ser mantidas e quais devem ser abandonadas; as abelhas exploradoras encontram aleatoriamente novas fontes de alimento.	69
Figura 17 – Fluxograma geral de um algoritmo de otimização por colônia de formigas.	72
Figura 18 – Comparativo do número de publicações sobre o tema predição da dengue, da febre chikungunya e do vírus zika por ano, de 2000 a 2021, usando o mecanismo de busca Google Acadêmico. Os resultados incluem artigos completos em periódicos, trabalhos completos em anais de eventos, resumos, capítulos de livro e livros.	76
Figura 19 – Perfil e localização geográfica da Cidade do Recife, capital do Estado de Pernambuco, Brasil, latitude -8.053889 e longitude -34.880833. . . .	98
Figura 20 – Interface do Sistema de Informações Geográficas da APAC, SIGH-PE, com a distribuição das Plataformas de Coleta de Dados, PCDs, ao longo do Estado de Pernambuco. Os dados pluviométricos (precipitação em mm) e fluviométricos são monitorados diariamente.	101

Figura 21 – Metodologia proposta aplicada à predição de arboviroses: predição da distribuição espacial de casos de arboviroses a partir de bases de dados de séries históricas de distribuições de informações climáticas e ambientais, de vigilância ambiental e clínicas, com destaque para as contribuições deste trabalho: (1) metodologia para organização das bases de dados e construção dos conjuntos de treino e de teste; (2) modelo de predição espaço-temporal baseado em regressor com Random Forest; e (3) o Comitê de Especialistas Artificiais.	106
Figura 22 – Processo de coleta de dados georreferenciados, montagem dos mapas das variáveis de interesse (temperaturas, pluviometrias, velocidades dos ventos e casos de arboviroses para os seis bimestres), das saídas desejadas, i.e. casos de arboviroses para o bimestre seguinte aos seis anteriores, e registro.	109
Figura 23 – Processo de concatenação dos mapas registrados.	110
Figura 24 – Comitê de Especialistas Artificiais para seleção de atributos, onde cada especialista é baseado em um algoritmo de busca heurística bioinspirado: algoritmo genético (busca genética), algoritmo genético modificado (busca evolucionária), otimização por enxame de partículas, colônia de abelhas artificiais (busca de abelhas), e colônia de formigas artificiais (busca de formigas), para populações iniciais de 20 indivíduos e máximo de 500 gerações ou iterações.	113
Figura 25 – Coeficiente de correlação R para a regressão linear.	119
Figura 26 – Coeficiente de correlação R para redes neurais MLP com uma única camada escondida, com teste de 2, 5, 10 e 20 neurônios.	120
Figura 27 – Coeficiente de correlação R para redes neurais ELM com uma única camada escondida, número de neurônios determinado automaticamente, para <i>kernels</i> polinomiais de grau 1 a 3 e RBF.	121
Figura 28 – Coeficiente de correlação R para redes neurais ESM e Deep-ESM com 2, 5 e 10 camadas escondidas, 10 neurônios por camada.	122
Figura 29 – Coeficiente de correlação R para SVM, <i>kernels</i> polinomiais de grau 1 a 3 e RBF.	122
Figura 30 – Coeficiente de correlação R para: Random Forest, de 10 a 100 árvores, variando de 10 em 10.	123

Figura 31 – RMSE(%) para a regressão linear.	123
Figura 32 – RMSE(%) para redes neurais MLP com uma única camada escondida, com teste de 2, 5, 10 e 20 neurônios.	124
Figura 33 – RMSE(%) para redes neurais ELM com uma única camada escondida, número de neurônios determinado automaticamente, para <i>kernels</i> po- linomiais de grau 1 a 3 e RBF.	125
Figura 34 – RMSE(%) para redes neurais ESM e Deep-ESM com 2, 5 e 10 camadas escondidas, 10 neurônios por camada.	126
Figura 35 – RMSE(%) para SVM, <i>kernels</i> polinomiais de grau 1 a 3 e RBF.	126
Figura 36 – RMSE(%) para: Random Forests, de 10 a 100 árvores, variando de 10 em 10.	127
Figura 37 – Tempo de treinamento (ms) para a regressão linear.	127
Figura 38 – Tempo de treinamento (ms) para redes neurais MLP com uma única camada escondida, com teste de 2, 5, 10 e 20 neurônios.	128
Figura 39 – Tempo de treinamento (ms) para redes neurais ELM com uma única ca- mada escondida, número de neurônios determinado automaticamente, para <i>kernels</i> polinomiais de grau 1 a 3 e RBF.	129
Figura 40 – Tempo de treinamento (ms) para redes neurais ESM e Deep-ESM com 2, 5 e 10 camadas escondidas, 10 neurônios por camada.	130
Figura 41 – Tempo de treinamento (ms) para SVM, <i>kernels</i> polinomiais de grau 1 a 3 e RBF.	130
Figura 42 – Tempo de treinamento (ms) para: Random Forests, de 10 a 100 árvores, variando de 10 em 10.	131
Figura 43 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 43(a)) e 2 (março a abril, cf. Figura 43(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	144

Figura 44 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 44(a)) e 4 (julho a agosto, cf. Figura 44(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	145
Figura 45 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 45(a)) e 6 (novembro a dezembro, cf. Figura 45(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	146
Figura 46 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 46(a)) e 2 (março a abril, cf. Figura 46(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	147
Figura 47 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 47(a)) e 4 (julho a agosto, cf. Figura 47(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	148
Figura 48 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 48(a)) e 6 (novembro a dezembro, cf. Figura 48(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	149

Figura 49 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 49(a)) e 2 (março a abril, cf. Figura 49(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	150
Figura 50 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 50(a)) e 4 (julho a agosto, cf. Figura 50(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	151
Figura 51 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 51(a)) e 6 (novembro a dezembro, cf. Figura 51(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	152
Figura 52 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 52(a)) e 2 (março a abril, cf. Figura 52(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	153
Figura 53 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 53(a)) e 4 (julho a agosto, cf. Figura 53(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	154

- Figura 54 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 54(a)) e 6 (novembro a dezembro, cf. Figura 54(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399. 155
- Figura 55 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 55(a)) e 2 (março a abril, cf. Figura 55(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399. 156
- Figura 56 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 56(a)) e 4 (julho a agosto, cf. Figura 56(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399. 157
- Figura 57 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 57(a)) e 6 (novembro a dezembro, cf. Figura 57(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399. 158
- Figura 58 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 58(a)) e 2 (março a abril, cf. Figura 58(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399. 159

Figura 59 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 59(a)) e 4 (julho a agosto, cf. Figura 59(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	160
Figura 60 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 60(a)) e 6 (novembro a dezembro, cf. Figura 60(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.	161

LISTA DE TABELAS

Tabela 1 – Quadro comparativo dos trabalhos relacionados com esta proposta, de acordo com os critérios propostos por Siriya-satien et al. (2018a) e Bhunia e Shit (2019).	95
Tabela 2 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para regressão linear.	117
Tabela 3 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para rede neural MLP, uma única camada escondida, para 2, 5, 10 e 20 neurônios na camada escondida. Os melhores resultados em média estão em destaque em vermelho.	117
Tabela 4 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para SVM, <i>kernel</i> polinomial grau 1 (linear), 2, 3 e RBF. Os melhores resultados em média estão em destaque em vermelho.	118
Tabela 5 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para rede neural ELM, uma única camada escondida, número de neurônios na camada escondida determinado automaticamente, <i>kernel</i> polinomial grau 1 (linear), 2, 3 e RBF. Os melhores resultados em média estão em destaque em vermelho.	118
Tabela 6 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para redes neurais ESM e Deep-ESM. Os melhores resultados em média estão em destaque em vermelho.	118
Tabela 7 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para Random Forest, de 10 a 100 árvores, variando a cada 10. Os melhores resultados em média estão em destaque em vermelho.	119
Tabela 8 – Resultado da análise de fatores mais relevantes por meio dos algoritmos de Busca Genética, Busca Evolucionária, Otimização por Enxame de Partículas, Busca de Abelhas e Busca de Formigas, para 20 candidatos à solução evoluindo em 500 gerações, para o ano de 2014.	134

Tabela 9 – Resultado da análise de fatores mais relevantes por meio dos algoritmos de Busca Genética, Busca Evolucionária, Otimização por Enxame de Partículas, Busca de Abelhas e Busca de Formigas, para 20 candidatos à solução evoluindo em 500 gerações, para o ano de 2015.	135
Tabela 10 – Resultado da análise de fatores mais relevantes por meio dos algoritmos de Busca Genética, Busca Evolucionária, Otimização por Enxame de Partículas, Busca de Abelhas e Busca de Formigas, para 20 candidatos à solução evoluindo em 500 gerações, para o ano de 2016.	136
Tabela 11 – Resultados de predição para a regressão linear e o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais. Como métricas de qualidade foram considerados o índice de correlação R, o erro RMSE (%) e os índices de correlação de Kendall e Spearman. Os melhores resultados índice de correlação R acima de 0,9 e erro RMSE (%) menor ou igual a 5% estão em vermelho. Já os melhores resultados que não atendem a esse critério estão em azul. . .	140
Tabela 12 – Relação de 46 dos 94 bairros da Cidade do Recife em ordem alfabética, dos Afritos ao Ibura, com seus respectivos códigos e coordenadas de centroide nos sistemas XY e em latitude e longitude. Os códigos e as coordenadas dos centroides foram definidos e padronizados pela Prefeitura do Recife.	141
Tabela 13 – Relação de 48 dos 94 bairros da Cidade do Recife em ordem alfabética, da Ilha do Leite ao Zumbi, com seus respectivos códigos e coordenadas de centroide nos sistemas XY e em latitude e longitude. Os códigos e as coordenadas dos centroides foram definidos e padronizados pela Prefeitura do Recife.	142

SUMÁRIO

1	INTRODUÇÃO	23
1.1	CONTEXTO E MOTIVAÇÃO	23
1.2	OBJETIVOS	29
1.3	BASES DE DADOS E ASPECTOS METODOLÓGICOS	30
1.4	ORGANIZAÇÃO DO TRABALHO	31
2	FUNDAMENTAÇÃO TEÓRICA	33
2.1	PRINCÍPIOS DE GEOGRAFIA DA SAÚDE E EPIDEMIOLOGIA DIGITAL	33
2.1.1	Definições e Conceitos	33
2.1.2	Sistemas de Informações Geográficas em Saúde	36
2.2	ARBOVIROSES: DENGUE, CHIKUNGUNYA E ZIKA	39
2.3	MÉTODOS DE PREDIÇÃO DE ARBOVIROSES	43
2.4	METODOLOGIA DE VIGILÂNCIA SANITÁRIA E AMBIENTAL NO SUS	44
2.5	REGRESSÃO LINEAR	52
2.6	PRINCÍPIOS DE APRENDIZADO DE MÁQUINA	52
2.6.1	Redes Neurais Artificiais	52
2.6.2	Máquinas de Vetor de Suporte	54
2.6.3	Máquinas de Aprendizado Extremo	56
2.6.4	Máquinas de Estado de Eco	57
2.6.5	Árvores de Decisão e Random Forests	58
2.7	ALGORITMOS META-HEURÍSTICOS BIOINSPIRADOS	63
2.7.1	Algoritmos Genéticos	63
2.7.2	Otimização por Enxame de Partículas	64
2.7.3	Colônias de Abelhas Artificiais	68
2.7.4	Colônias de Formigas Artificiais	70
2.8	MÉTRICAS DE REGRESSÃO	73
2.9	CONSIDERAÇÕES	74
3	TRABALHOS RELACIONADOS	75
3.1	ESTADO DA ARTE	75
3.2	CONSIDERAÇÕES	96
4	PREDIÇÃO BASEADA NA ANÁLISE ESPAÇO-TEMPORAL	97

4.1	ÁREA SOB ESTUDO	97
4.2	BASES DE DADOS	99
4.2.1	Sistema de Informações Geográficas da Agência Pernambucana de Águas e Clima - SIGH-PE	99
4.2.2	Banco de Dados Meteorológicos para Ensino e Pesquisa - BDMEP	100
4.2.3	Portal de Dados Abertos da Prefeitura da Cidade do Recife	103
4.2.4	LIRAA Recife	103
4.3	PROPOSTA	105
4.4	ARRANJO EXPERIMENTAL	113
4.5	CONSIDERAÇÕES	114
5	RESULTADOS E DISCUSSÃO	115
5.1	PREDIÇÃO DE CASOS DE ARBOVIROSES	115
5.1.1	Modelo preditor	115
5.1.2	Análise de fatores mais relevantes	132
5.1.3	Análise espaço-temporal	140
5.2	CONSIDERAÇÕES	162
6	CONCLUSÃO	164
6.1	CONCLUSÕES GERAIS	164
6.2	DIFICULDADES APRESENTADAS	166
6.3	CONTRIBUIÇÕES	167
6.4	TRABALHOS FUTUROS	168
	REFERÊNCIAS	170
	ANEXO A – CARTA DE ANUÊNCIA DA SECRETARIA DE SAÚDE DA PREFEITURA DA CIDADE DO RECIFE	193

1 INTRODUÇÃO

Este capítulo apresenta a introdução à Tese de Doutorado, sendo apresentados o contexto, a justificativa e a motivação para a realização da pesquisa relatada, bem como o objetivo geral e os objetivos específicos ou metas. As bases de dados e aspectos metodológicos também são apresentados.

1.1 CONTEXTO E MOTIVAÇÃO

As revoluções industriais são momentos da história humana em que as bases produtivas de diversas nações mudam radicalmente de natureza, refletindo importantes avanços científicos e mudando as relações dentro das diferentes sociedades nacionais e entre as diversas nações. A Primeira Revolução Industrial, iniciada no Reino Unido, no final do Século XIX, foi baseada no domínio das tecnologias de automação, a partir da invenção do tear mecânico a vapor (DEANE; DEANE, 1979; BARHAM, 2013; CRAFTS, 2011; XU et al., 2018). No começo do Século XX inicia-se, centrada nos países europeus e nos Estados Unidos, a Segunda Revolução Industrial, com o domínio da eletricidade e a invenção do motor elétrico, dos sistemas de iluminação e da transformação de diferentes formas de energia em energia elétrica (ROSENBERG, 1998; ZHANG et al., 2017; XU et al., 2018). A Terceira Revolução Industrial é resultado do mundo que surgiu após a Segunda Guerra Mundial, com o advento da microeletrônica, dos avanços nas telecomunicações, dos computadores eletrônicos e das redes de computadores, iniciando a Era da Informação. Como revoluções, as revoluções industriais podem abrir crises sociais profundas, mas também podem ser oportunidades para que as diferentes sociedades estabeleçam novas bases para o seu desenvolvimento (GREENWOOD, 1997; SMITH, 2001; XU et al., 2018).

A Quarta Revolução Industrial é marcada pela emergência de diversas tecnologias, como a Inteligência Artificial, a Internet das Coisas, a Robótica e a Biotecnologia (PRISECARU, 2016; DAVIS, 2016; XU et al., 2018). Como toda revolução industrial, também a Quarta Revolução Industrial, ou Indústria 4.0, abre oportunidades para a construção de novas bases produtivas, tendo destaque as novas tecnologias em saúde, que se utilizam da Inteligência Artificial, da Internet das Coisas e da Biotecnologia para construir novos medicamentos, desenvolver novas terapias, e prever surtos epidemiológicos, epidemias e pandemias, por exemplo (ACETO; PERSICO; PESCAPÉ, 2020; LECK; LIM, 2015; BONACCORSI et al., 2020; ACETO; PERSICO; PESCAPÉ,

2019).

A partir dos anos 2000, ficam cada vez mais evidentes as mudanças climáticas, como resultado do desequilíbrio ambiental resultante da modificação ou mesmo da extinção de diversos biomas, além da extinção de diversas espécies animais e vegetais. Uma das consequências desse processo é a proliferação de novas doenças a partir de novos vírus, que migram de populações animais, onde muitas vezes são inofensivos, para populações humanas (SARBEEN; GHEENA, 2016; THOMSON; GARCIA-HERRERA; BENISTON, 2008; GODLEE, 1991; SCHOENTAL, 1994; MONAGHAN et al., 2018; SÁNCHEZ; GURPEGUI; LUCAS, 2018). Alguns dos exemplos mais importantes são as doenças advindas dos coronavírus, como a SARS, a MERS e, recentemente, a Covid-19. A SARS (Síndrome Respiratória Aguda Severa, *Severe Acute Respiratory Syndrome*), transmitida pelo vírus SARS-CoV, teria sido originada do consumo de morcegos na China em 2002 (LI et al., 2005; BURKI, 2020). A MERS (Síndrome Respiratória do Oriente Médio, *Middle East Respiratory Syndrome*), transmitida pelo vírus MERS-CoV, surge em 2012 na Arábia Saudita a partir do contato entre humanos e camelos (DUDAS et al., 2018; ALSHUKAIRI et al., 2018; AZHAR et al., 2014). A Covid-19, por sua vez, surge em Wuhan, na China, transmitida pelo vírus SARS-CoV-2, presente em morcegos e em pangolins, em dezembro de 2019 e se torna em poucos meses uma pandemia mundial com milhões de pessoas infectadas e mortas (BURKI, 2020).

No caso das doenças transmitidas por coronavírus, embora o vetor inicial tenha sido animais, a partir da infecção do primeiro ser humano, os próprios seres humanos são ao mesmo tempo população a ser infectada e vetores. Assim, passa a ser importante o monitoramento e controle da população humana (MOSER et al., 2020). Nesse caso, os fatores relevantes são aqueles capazes de movimentar indivíduos e populações humanas (MOSER et al., 2020). Já no caso de doenças transmitidas por outros vetores, como as arboviroses, doenças transmitidas por mosquitos, passa a ser importante monitorar fatores climáticos, ambientais e socioeconômicos, por exemplo (GOULD; HIGGS, 2009; MORDECAI et al., 2020; RUSSELL, 1998; LIU-HELMERSSON et al., 2019; LORENZ et al., 2017; WHITEMAN et al., 2020). Assim, é bastante razoável considerar que vetores, populações e casos de doenças podem ser influenciados por diferentes variáveis geográficas climáticas, ambientais e socioeconômicas, além da distribuição geográfica e temporal dos casos.

A análise multimodal consiste na análise de dados de diferentes origens, como diferentes sensores e diversas bases de dados, visando o entendimento de um determinado fenômeno (SOLEYMANI et al., 2017; PORIA et al., 2017a). A análise de dados multimodais tem crescido

em importância na Ciência da Computação, em áreas como diagnóstico clínico e aplicações de apoio ao diagnóstico (JI et al., 2020; CARDOSO; ARBEL, 2017; HAO et al., 2021; KIM et al., 2019) e análise de sentimentos (SOLEYMANI et al., 2017; SHAH; ZIMMERMANN, 2017; CHEN et al., 2017; MAJUMDER et al., 2018; CHATURVEDI et al., 2019; PORIA et al., 2017b; PORIA et al., 2018), por exemplo. A análise multimodal tem crescido bastante como campo de aplicação de aprendizado de máquina e mineração de dados em sistemas com múltiplas bases de dados, com diversas aplicações em saúde (JI et al., 2020; CARDOSO; ARBEL, 2017; HAO et al., 2021; KIM et al., 2019; YAZDAVAR et al., 2020; GARCIA-CEJA et al., 2018). Considerando a predição de doenças e de surtos epidemiológicos, epidemias e pandemias, o problema de predição espaço-temporal de doenças é necessariamente um problema de análise multimodal (SINGH et al., 2019; KUMAR; MAHANTI; WANG, 2019). Assim, a integração entre diferentes bases de dados geográficas referentes a variáveis climáticas, ambientais, socioeconômicas e de casos dos sistemas de saúde podem auxiliar na construção de sistemas de análise multimodal para predição espaço-temporal de doenças (LIMA et al., 2020; SILVA et al., 2021b; SILVA et al., 2021a; LIMA et al., 2021; SILVA et al., 2021d). A Internet das Coisas tem ajudado a construir rapidamente novas bases de dados por meio de sistemas embarcados com múltiplos sensores e utilizando dispositivos móveis como instrumento de sensoriamento, gerando novas bases geográficas (LIN et al., 2017; JAYACHITRA et al., 2021).

A análise multimodal pode ser importante para epidemiologistas, para gestores de saúde e para a Saúde Pública como um todo, podendo auxiliar na tomada de decisão e na elaboração de políticas públicas. Para isso, também é importante encontrar quais seriam as variáveis mais importantes para a predição em cada momento predito. Nesse sentido, os métodos meta-heurísticos bioinspirados, como aqueles da Computação Evolucionária e da Inteligência de Enxames, podem ser importantes para destacar quais seriam os fatores mais relevantes para a predição. Esses algoritmos, no entanto, tanto por sua natureza aleatória quanto por suas diferentes dinâmicas de busca, podem escolher como fatores mais relevantes variáveis diferentes. Assim, pode ser interessante considerar que cada algoritmo se comportaria como um especialista artificial independente, o que pode também tornar interessante algum tipo de combinação desses resultados, como um comitê, por exemplo, onde as variáveis consideradas mais relevantes seriam aquelas de consenso ou mais votadas entre os diferentes especialistas virtuais ou algoritmos.

Dentre as diferentes doenças infecciosas, as arboviroses são bastante interessantes como estudo de caso para a proposta de uma metodologia para construção de arquiteturas de

sistemas de análise multimodal baseados em aprendizado de máquina e algoritmos meta-heurísticos para predição espaço-temporal de doenças infecciosas, pelo fato de o seu vetor, os mosquitos, serem influenciados por diferentes fatores climáticos e ambientais que, por sua vez, também sofrem influência de fatores socioeconômicos, por seu consequente impacto ambiental (GOULD; HIGGS, 2009; MORDECAI et al., 2020; RUSSELL, 1998; LIU-HELMERSSON et al., 2019; LORENZ et al., 2017; WHITEMAN et al., 2020). Assim, as arboviroses fornecem provavelmente o cenário mais rico para um estudo de caso.

A dengue foi considerada por muito tempo como a mais importante das doenças virais transmitidas por mosquitos (arboviroses), sendo também a mais difundida arbovirose no mundo (BRAGA; VALLE, 2007). A dengue se manifesta clinicamente sob duas formas principais: a dengue clássica, chamada simplesmente de febre de dengue, e a forma hemorrágica, conhecida ora como febre hemorrágica de dengue (FHD), ora como síndrome de choque de dengue (FHD/SCD). Desde o início da década de 1970, a Organização Mundial da Saúde (OMS) tem estado ativamente envolvida no desenvolvimento e na promoção de estratégias de tratamento e controle da doença (BRAGA; VALLE, 2007). A dengue é transmitida por mosquitos do gênero *Aedes*, sendo o *Aedes aegypti* seu principal vetor. O *Aedes aegypti* é encontrado, principalmente, nas regiões urbanas, em depósitos de armazenamento de água (BRAGA; VALLE, 2007).

A reemergência de epidemias de dengue clássica e a emergência da febre hemorrágica de dengue são considerados como parte dos maiores problemas de Saúde Pública da segunda metade do século XX e do começo do século XXI (BRAGA; VALLE, 2007). As mudanças demográficas e o intenso fluxo migratório das zonas rurais para as regiões urbanas geraram um crescimento desordenado nas cidades, o que, somado à ausência de boas condições de saneamento básico, acaba por resultar na proliferação do vetor, essencialmente em países tropicais e subtropicais, onde têm sido comuns surtos periódicos da doença (BRAGA; VALLE, 2007; BRAGA et al., 2004).

Devido às alterações climáticas propícias à dispersão de vetores e suas doenças e ao crescente número de voos internacionais, favoráveis à movimentação de pessoas doentes ou infectadas em período de incubação, vive-se no Brasil a introdução e um rápido processo de dispersão rumo a se tornar endêmicos dois novos arbovírus para as Américas: o vírus Chikungunya, introduzido entre julho e agosto de 2014, após ter entrado no Caribe em dezembro de 2013 e, anteriormente, ter causado grandes epidemias no continente africano e na Ásia desde 2004; e o vírus Zika, possivelmente introduzido no mesmo período durante a Copa do Mundo

realizada em 2014 no Brasil (VASCONCELOS, 2015). A prevenção e o combate à proliferação do *Aedes aegypti* são fundamentais para conter eventuais surtos e garantir uma melhora das condições de salubridade das populações do campo e das cidades.

O vírus Zika é um flavivírus (família *Flaviviridae*) transmitido pelo *Aedes aegypti*. O vírus Zika causa febre acompanhada pela ocorrência de outros sintomas gerais, tais como cefaleia, exantema¹, mal estar, edema² e dores articulares intensas. Quadros mais severos, incluindo comprometimento do sistema nervoso central (síndrome de Guillain-Barré, mielite transversa e meningite), associados ao Zika têm sido comumente registrados, o que mostra quão pouco conhecida ainda é essa doença (VASCONCELOS, 2015). Existem evidências da relação entre o zika vírus e a microcefalia, além de outros efeitos sobre os fetos, mas essas questões ainda não estão totalmente esclarecidas, sendo objeto de estudo (BRASIL et al., 2016; MARTINES, 2016).

A prevenção das arboviroses depende da priorização da eliminação do vetor, ou seja, os focos de *Aedes aegypti*, e de estratégias de predição que possam suprir os gestores de saúde de informações adequadas para evitar surtos de arboviroses (KRAEMER et al., 2015; PAUL et al., 2018). Isto requer a cooperação ativa entre governo e agências de saúde na elaboração e execução de estratégias de controle das doenças junto à população geral e o uso de ferramentas que possam extrair informações de forma eficiente e eficaz das diversas bases de dados que os municípios já detêm, que possuem informações sobre infraestrutura, aspectos sócio-econômicos e ambientais, distribuição de serviços de saúde e mapeamento de casos, cruzando essas bases com outras, como aquelas com informações climáticas e ambientais das agências de águas e das secretarias de saúde estaduais, além do mapeamento do índice LIRAa (BRASIL, 2012; PAUL et al., 2018).

O Sistema Único de Saúde (SUS) no Brasil está organizado em três níveis de atenção: atenção básica, secundária e terciária (SILVEIRA, 2010). O Agente Comunitário de Saúde (ACS) é o pilar desse sistema organizativo, responsável pelo contato direto e constante com as famílias e com as entidades locais da sociedade civil organizada (SILVEIRA, 2010). O Agente de Controle de Endemias (ACE), ou Agente de Saúde Ambiental e Controle de Endemias (ASACE) é o profissional de saúde responsável pelo controle de vetores de endemias e por ações educativas junto à população para engajamento nesse controle. Os ACEs estão distribuídos por regiões geográficas dentro dos distritos sanitários em cada município, e têm por obrigação visitar de 25

¹ Exantema é uma irritação generalizada da pele. Pode ser resultante de brotoejas, picadas de insetos, queimaduras de sol ou efeitos colaterais da ingestão de medicamentos.

² Edema é um excesso de fluido preso no tecido mole, provocando inchaço.

a 35 imóveis (originalmente de 20 a 25) diariamente, visando fazer inspeções para controle dos vetores de doenças, como criadouros de mosquitos e outros focos de zoonoses, além de serem responsáveis por ações pedagógicas junto às famílias e em escolas e centros comunitários (BRASIL, 2012). Os ACEs são os principais atores na coleta de informações a respeito dos locais infectados pelo *Aedes aegypti*, por coletarem informações a respeito do vetor por meio de formulários específicos, onde são registradas os endereços e as condições de eventuais criadouros do mosquito e a presença ou não do mosquito ou de larvas (BRASIL, 2012). Esses endereços podem ser geocodificados, agregando informações de latitude e longitude ao índice LIRAa.

Diversos trabalhos mostram que existe uma forte correlação entre a distribuição dos casos de arboviroses e dos focos de *Aedes aegypti* com fatores climáticos, como as séries históricas de umidade, pluviometria e temperaturas (TOSEPU et al., 2018; HAMLET et al., 2018; KRAEMER et al., 2015; JANSEN; BEEBE, 2010; SIRIYASATIEN et al., 2018a). A urbanização e as mudanças no cultivo de determinadas culturas também influenciam fortemente a distribuição do vetor (JANSEN; BEEBE, 2010; KRAEMER et al., 2015). Também é possível integrar informações ambientais coletadas de séries históricas obtidas de imagens de satélites (SCAVUZZO et al., 2017; FAJARDO-HERRERA; VALDELAMAR-VILLEGAS; ARRIETA-PÉREZ, 2017; ALBRIEU-LLINÁS et al., 2018) e de sensores diversos usando IoT (*Internet of Things*, Internet das Coisas) (THAKUR; KAUR, 2017), combinando uma abordagem de geoprocessamento com o uso de ferramentas de aprendizado de máquina, como métodos estatísticos, evolutivos e redes neurais artificiais, para prever a distribuição do vetor (SCAVUZZO et al., 2017).

A combinação de variadas bases de dados e de imagens para predição pode contribuir para se ter grandes volumes de dados, talvez de difícil generalização (SIRIYASATIEN et al., 2018a). Nesse sentido, faz-se necessário investigar também outras abordagens de aprendizagem de máquina para além dos métodos clássicos e das redes neurais artificiais mais usuais. Nesta pesquisa também será investigada a efetividade do uso de redes neurais profundas (*deep nets*), dado que a abordagem de *deep learning* tem se mostrado efetiva na resolução de diversos problemas (SUN; WANG; TANG, 2014; DENIL et al., 2013; LECUN; BENGIO; HINTON, 2015), além de outras arquiteturas, tais como as máquinas de aprendizado extremo (*Extreme Learning Machines*, ELMs) (HUANG; ZHU; SIEW, 2006; HUANG; ZHU; SIEW, 2004; HUANG et al., 2012) e as máquinas de vetor de suporte (*Support Vector Machines*, SVMs) (SUYKENS; VANDEWALLE, 1999; YUSOF; MUSTAFFA, 2011; CH et al., 2014).

Este trabalho é orientado pela seguinte pergunta de pesquisa: como construir preditores

de doenças eficientes, capazes de prever a distribuição espaço-temporal de casos com erro aceitável, alimentados com bases de dados de informações geográficas de diversas fontes e que possam ser utilizados como apoio aos gestores de saúde no planejamento, vigilância e combate à sua disseminação, apontando também os fatores que mais influenciam a distribuição dos casos em determinados momentos?

1.2 OBJETIVOS

Este trabalho tem como objetivo geral propor uma metodologia para construção de preditores capazes de prever a distribuição espaço-temporal de doenças e apontar os fatores mais relevantes para a predição a partir de uma arquitetura baseada no acesso a múltiplas bases de dados. Para validação da proposta, foi adotada como estudo de caso a predição de casos de arboviroses por meio de séries históricas georreferenciadas de informações climáticas e ambientais utilizando técnicas de aprendizado de máquina. Essas informações são coletadas de múltiplos bancos de dados georreferenciados, previamente construídos a partir da coleta de informações por redes de sensores (velocidade dos ventos, temperaturas e pluviometria/precipitação) e do Sistema Único de Saúde (relatórios de casos dos postos de saúde e índices de infestação predial de instâncias de vigilância ambiental). Essas informações são transformadas em mapas e depois agrupadas em séries históricas de mapas para que, a cada seis ciclos (seis bimestres, ou seja, um ano), a distribuição espacial de casos de arboviroses pelo vetor no ciclo (bimestre) consecutivo possa ser predita. Foram utilizadas informações da Cidade do Recife, de casos de arboviroses (dengue, chikungunya e zika) de 2013 a 2016, e informações climáticas e ambientais do mesmo período, da APAC e do INMET. Os sistemas de predição de doenças construídos utilizando a metodologia proposta neste trabalho também devem ser capazes de apontar os fatores mais relevantes para a predição por meio do Comitê de Especialistas Artificiais, proposto neste trabalho e composto de um conjunto de algoritmos de seleção de atributos baseados em métodos de otimização por Computação Evolucionária. O Comitê de Especialistas Artificiais decide por votação.

Como objetivos específicos, têm-se:

1. Construir uma metodologia para completar e organizar os dados georreferenciados obtidos de múltiplas bases de dados em séries históricas de mapas, estimando os dados faltantes;

2. Investigar técnicas de aprendizado de máquina para construir preditores dinâmicos de distribuição espaço-temporal de doenças transmitidas por vetor, ou seja, com relação ao estudo de caso, pelos mosquitos *Aedes aegypti*;
3. Investigar algoritmos bioinspirados que possam auxiliar na avaliação dos fatores de risco, ou seja, dos fatores que mais influenciam cada predição, para construção do Comitê de Especialistas Artificiais;
4. Validar a proposta utilizando informações geográficas ambientais, climáticas, de vigilância ambiental e clínicas, utilizando como estudo de caso a Cidade do Recife, de 2013 a 2016.

1.3 BASES DE DADOS E ASPECTOS METODOLÓGICOS

Neste trabalho é proposta uma metodologia para construir preditores espaço-temporais de doenças e avaliar os fatores mais relevantes para a predição. A metodologia de predição se baseia no uso de aprendizado de máquina sobre múltiplas bases de dados geográficos, para predição espaço-temporal de casos, enquanto os fatores relevantes para predição são levantados por um Comitê de Especialistas Artificiais, proposto neste trabalho. Como estudo de caso, busca-se modelar arquiteturas de preditores de distribuição do vetor *Aedes aegypti* baseadas em sistemas de informações geográficas com informações climáticas e ambientais e aprendizado de máquina, para fornecer informações de epidemiologia digital para especialistas em saúde pública e para gestores de políticas públicas em saúde. A escolha de predição de arboviroses se deu pelo fato de o vetor, ou seja, o mosquito *Aedes aegypti*, sofrer influência de fatores climáticos e ambientais, o que torna esse problema de predição razoavelmente complexo, dada a necessidade de diversas bases de dados diferentes.

São utilizadas séries históricas de dados georreferenciados do ano de 2013 a 2016, do Sistema de Geoinformação Hidrometeorológico de Pernambuco, da Agência Pernambucana de Águas e Clima (APAC)³, e do Instituto Nacional de Meteorologia (INMET)⁴, como umidade, velocidade dos ventos, temperatura e pluviometria/precipitação. Esses dados são combinados aos relatórios consolidados do Sistema LIRAa (Levantamento do Índice Rápido de Infestação

³ Agência Pernambucana de Águas e Clima - APAC, disponível em <<http://www.apac.pe.gov.br/>>, acessado em 30 de julho de 2018.

⁴ Instituto Nacional de Meteorologia - INMET, disponível em <<http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>>, acessado em 30 de julho de 2018.

do *Aedes aegypti*), obtidos do Centro de Vigilância Ambiental da Cidade do Recife, CVA, de 2013 a 2016, e do Portal de Dados Abertos⁵ da Cidade do Recife.

Também serão investigados algoritmos bioinspirados, tais como algoritmos genéticos, algoritmos evolutivos, colônias de abelhas artificiais e colônias de formigas artificiais, que podem auxiliar na identificação de fatores mais relevantes para a predição da distribuição de casos por meio do processo de seleção de atributos (FONG; BIUK-AGHAI; MILLHAM, 2018).

1.4 ORGANIZAÇÃO DO TRABALHO

Esta Tese está organizada da forma que segue:

No Capítulo 2 é apresentada a fundamentação teórica, sendo apresentados os conceitos de epidemiologia digital, as técnicas e modelos clássicos de predição de arboviroses e modelagem do vetor, e breves revisões das técnicas de aprendizado de máquina e de algoritmos bioinspirados para busca e otimização. Também é apresentada uma breve visão da estrutura da vigilância ambiental da Cidade do Recife, o estudo de caso desta pesquisa.

No Capítulo 3 é apresentado o estado da arte dos modelos de preditores, das bases de dados, das técnicas computacionais de predição de arboviroses, o estudo de caso desta Tese, e da difusão do vetor, com ênfase nas técnicas baseadas em aprendizado de máquina, finalizando com uma análise comparativa crítica da metodologia proposta neste trabalho e o estado da arte.

A proposta deste trabalho é apresentada no Capítulo 4 por meio de descrições metodológicas detalhadas e diagramas de blocos. Também são apresentadas as limitações do estudo e as bases de dados utilizadas.

Os resultados são apresentados no Capítulo 5 para predição de casos de arboviroses, análise dos fatores mais relevantes para as predições, e análise espaço-temporal dos resultados de predição.

Por fim, as dificuldades apresentadas, as conclusões gerais, contribuições e trabalhos futuros são apresentados no Capítulo 6.

No Anexo A é apresentada a Carta de Anuência da Prefeitura do Recife, utilizada para a coleta de dados do Índice LIRAA, por meio do projeto “Gamificação para treinamento de profissionais de saúde em vigilância participativa e protocolos associados ao vírus zika” que, por

⁵ Portal de Dados Abertos da Cidade do Recife, disponível em <<http://dados.recife.pe.gov.br/dataset/casos-de-dengue-zika-e-chikungunya>>, acessado em 30 de julho de 2018.

sua vez, é parte do projeto "*A gamified m-training app for health professionals on protocols and participatory surveillance associated with Zika virus*" (2017-2019), financiado pela Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco, FACEPE, e pelo British Council, por meio do Fundo Newton.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentada a fundamentação teórica de princípios de epidemiologia digital; dos métodos de predição de arboviroses baseados em aprendizado de máquina; de princípios de aprendizado de máquina e de algoritmos meta-heurísticos bioinspirados.

2.1 PRINCÍPIOS DE GEOGRAFIA DA SAÚDE E EPIDEMIOLOGIA DIGITAL

2.1.1 Definições e Conceitos

A Epidemiologia é o ramo das ciências da saúde que trata dos fatores genéticos, sociais, ambientais e das condições de exposição que determinam a ocorrência e a distribuição das doenças, deficiências, incapacidades e mortes em populações ou em segmentos dessas populações (SANTANA, 2014; SILVA, 1997; MONKEN; BARCELLOS, 2005; BONFIM; MEDEIROS, 2008; FERREIRA, 1991; FARIA; BORTOLOZZI, 2009; BHUNIA; SHIT, 2019). De forma resumida, a Epidemiologia é o estudo da distribuição das doenças e de seus fatores determinantes.

A Epidemiologia Digital faz uso da Ciência da Computação e de seus métodos para o estudo da distribuição das doenças e de seus fatores determinantes e condicionantes. Assim, a Epidemiologia Digital utiliza intensamente as áreas de Banco de Dados, Processamento Digital de Imagens, Mineração de Dados e Inteligência Computacional na extração de informação e análise de dados epidemiológicos.

A Geografia da Saúde, também conhecida como Geografia Médica, é um ramo da Geografia que estuda a influência dos fatores geográficos, particularmente os aspectos humanos e ambientais terrestres e climáticos, no estudo da proeminência da saúde e da distribuição espacial de doenças (SANTANA, 2014; SILVA, 1997; MONKEN; BARCELLOS, 2005; BONFIM; MEDEIROS, 2008; FERREIRA, 1991; FARIA; BORTOLOZZI, 2009; BHUNIA; SHIT, 2019). Ela está fortemente relacionada com a Epidemiologia. Dessa forma, a Geografia da Saúde fornece informações sobre a relação entre a saúde de indivíduos e de populações e sua distribuição geográfica e sua associação com fatores ambientais (BHUNIA; SHIT, 2019). O conceito de Geografia Médica foi introduzido pela primeira vez por Hipócrates, médico grego que viveu entre os séculos V e IV aC e que criou o Juramento de Hipócrates, que todo profissional da Medicina é obrigado a seguir no exercício de sua profissão (BHUNIA; SHIT, 2019).

A humanidade tem conhecimento do desenvolvimento da disseminação de doenças através de regiões geográficas por eras, mesmo em épocas em que não se conhecia as origens nem as causas de determinadas doenças. Por exemplo, a peste negra, que assolou a Europa entre os anos de 1346 a 1351, difundiu-se pelos territórios ao longo das rotas comerciais da China à Europa, tendo como vetor os ratos (BHUNIA; SHIT, 2019).

A Geografia Médica tem muitas aplicações. O mapeamento desempenha um papel enorme nesse campo. Os mapas são produzidos para estudar epidemias históricas como a gripe de 1918, ou mesmo para estudar a distribuição de doenças como a gripe, a malária, ou a leishmaniose em todo o mundo. Geógrafos, médicos e profissionais de saúde pública têm trabalhado conjuntamente para estudar o comportamento e a distribuição de determinadas doenças no tempo na forma de sinais de doença, como morbidade e mortalidade (BHUNIA; SHIT, 2019). A morbidade é a taxa de portadores de determinada doença em relação a uma determinada população ou segmento populacional em um determinado período de tempo e em uma determinada localização geográfica, enquanto a mortalidade é a taxa de mortos com relação ao mesmo referencial.

A Geografia Médica está focada em três temas principais: ecologia das doenças, prestação de cuidados de saúde e ambiente e saúde. A ecologia da doença engloba a investigação de doenças infecciosas (por exemplo, dengue, chikungunya, zika, febre amarela, malária, filariose, leishmaniose e HIV/AIDS), compreendendo as distribuições geográficas de fenômenos climáticos associados, eventos biológicos e culturais interrelacionados com doenças, juntamente com os fatores demográficos, políticos e econômicos (BHUNIA; SHIT, 2019). A prestação de cuidados de saúde abrange medidas geográficas de transmissão de cuidados de saúde e atividades do paciente (BHUNIA; SHIT, 2019). Meio ambiente e saúde são uma ênfase comparativamente nova para os geógrafos da saúde, que faz uso das abordagens das ciências geográficas na investigação de riscos ambientais, juntamente com a Geografia da Saúde (BHUNIA; SHIT, 2019). Embora a visualização dos dados de distribuição de doenças possa ser relativamente simples, a compreensão dos dados de doenças geograficamente referenciados pode ser complexa, principalmente para doenças altamente infecciosas (BHUNIA; SHIT, 2019). No entanto, o principal problema ainda é a identificação dos fatores que contribuem para a dinâmica da distribuição geográfica da doença, visando identificar sua origem, fatores condicionantes, e prever distribuições futuras.

A gestão da informação de saúde pública atualmente é um campo embrionário que faz uso intensivo da ciência e das tecnologias da informação e da comunicação para a pesquisa

em saúde pública. As diretrizes de saúde pública e as políticas públicas em saúde têm sido baseadas no uso e na exploração de dados espaciais por várias décadas. Em 1854, o Dr. John Snow usou um mapa desenhado à mão para investigar a localização geográfica da epidemia de cólera em Londres. Snow presumiu que a cólera se espalhou através do abastecimento público de água e chegou à conclusão de que determinados chafarizes públicos eram a fonte mais provável da epidemia de cólera, tendo conseguido assim conter a epidemia (HALLIDAY, 2001; TUFTE et al., 1998; NEWSOM, 2006; SNOW, 1857; PANETH et al., 1998; GILBERT, 2004; SNOW, 1855; BRODY et al., 2000). Este foi a primeira aplicação de gestão da informação de saúde pública na contenção de uma epidemia.

Dados geograficamente próximos tendem a serem influenciados por fatores semelhantes. Em 1890, o Dr. Theobald A. Palm realizou um estudo sobre a distribuição geográfica do raquitismo em uma área urbana industrial. Ele percebeu que essa área do norte da Europa continha alta umidade e baixa temperatura. A partir dessas informações, foi percebida a influência da umidade, das baixas temperaturas e da pouca exposição à luz do sol na ocorrência do raquitismo. Assim, foram descobertos os benefícios da exposição à luz solar e da vitamina D para a saúde (MOZOŃOWSKI, 1939; STEPHEN, 1975; RAJAKUMAR, 2003; PALM, 1890).

Ainda no final do século XIX, Florence Nightingale, patrona da moderna Enfermagem, estudou estatísticas de pacientes e conseguiu identificar as razões da mortalidade dos soldados britânicos durante a Guerra da Criméia: doenças relacionadas a contaminações em hospitais. Florence estimulou práticas sanitárias em instalações médicas que têm salvado milhões de vidas, tais como a higienização das mãos e dos instrumentos. A Teoria Ambientalista desenvolvida por Florence Nightingale é focada no meio ambiente, entendido como o conjunto das condições e influências externas que afetam a vida e o desenvolvimento de um organismo, capazes de prevenir, suprimir ou contribuir para a doença e a morte (MEDEIROS; ENDERS; LIRA, 2015; HADDAD; SANTOS, 2011; SELANDERS, 1998; MACEDO et al., 2008). A doença é considerada, nessa teoria, um processo restaurador da saúde (MEDEIROS; ENDERS; LIRA, 2015; HADDAD; SANTOS, 2011). A função da enfermeira seria equilibrar o meio ambiente, com o intuito de conservar a energia vital do paciente para que se recupere da doença, priorizando o fornecimento de um ambiente estimulador do desenvolvimento da saúde. Assim, o ser humano é entendido como um ser integrante da natureza, um indivíduo cujas defesas naturais são influenciadas pelo ambiente. Nightingale acreditava que fornecer um ambiente adequado era o diferencial na recuperação dos doentes (MEDEIROS; ENDERS; LIRA, 2015; HADDAD; SANTOS, 2011; SELANDERS, 1998; MACEDO et al., 2008). Dessa forma, a análise espacial em saúde pú-

blica não se refere apenas à localização geográfica da distribuição de doenças, mas também à estrutura e às condições ambientais de uma população.

2.1.2 Sistemas de Informações Geográficas em Saúde

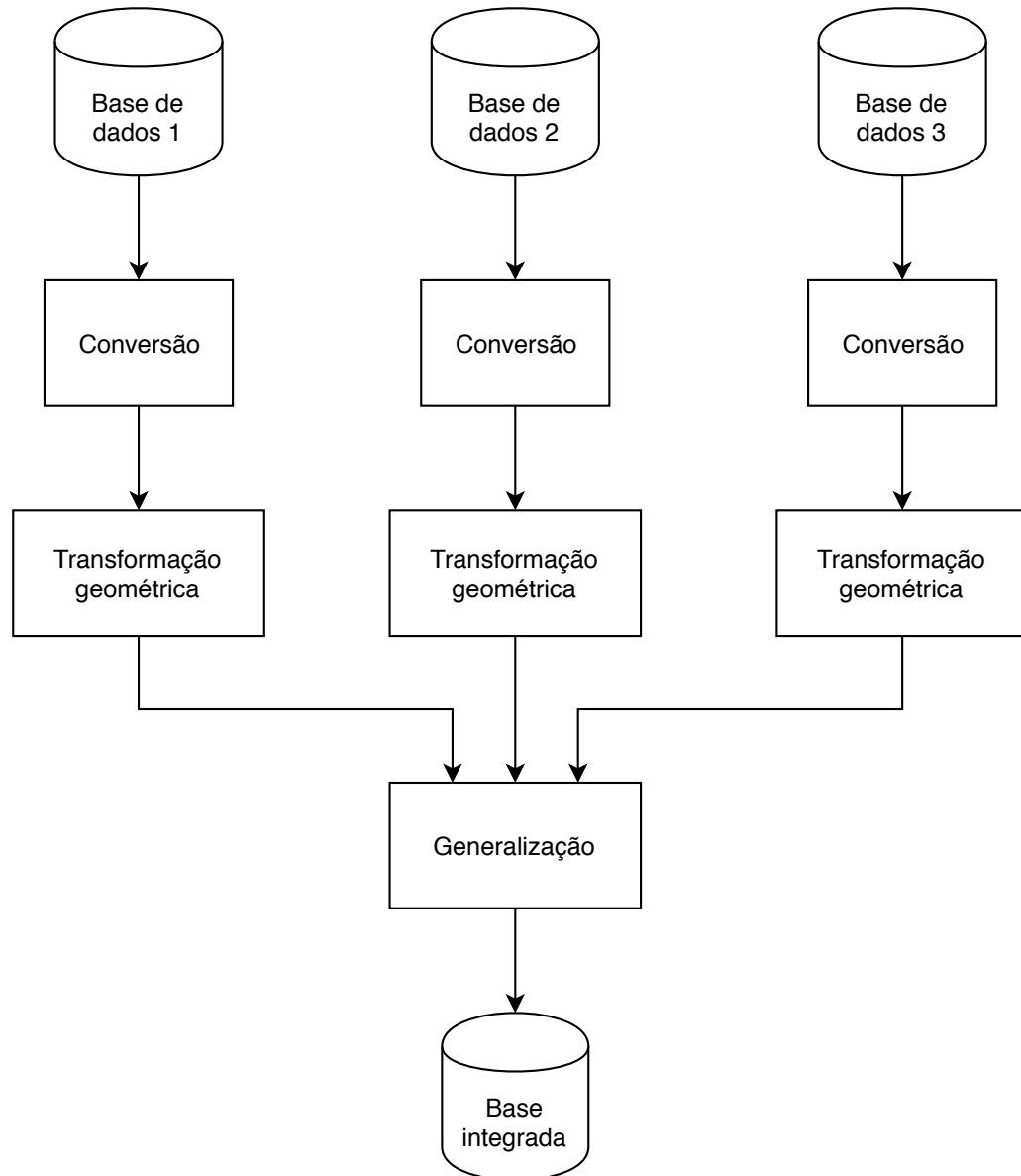
A aplicação de Sistemas de Informações Geográficas (*Geographic Information Systems*, GIS) à saúde pública tem um potencial considerável para ajudar a esclarecer a compreensão de potenciais fatores que podem dar origem a problemas de saúde complexos (BHUNIA; SHIT, 2019). Os GIS podem ser usados no auxílio à administração das políticas de saúde e na avaliação de programas e políticas eficazes centrados na população (BHUNIA; SHIT, 2019). Por meio de GIS, os dados podem ser gravados e editados digitalmente, armazenados e reorganizados, formatados e analisados, bem como visualizados adequadamente, na forma de tabelas ou de mapas. A Organização Mundial da Saúde (OMS) vem afirmando desde 1999 a necessidade de se ter pessoal treinado em tecnologias GIS na gestão da saúde pública (BHUNIA; SHIT, 2019).

Existem três tipos diferentes de estudos geográfico-epidemiológicos: mapeamento de doenças, estudos ecológicos e estudos de migração de populações (BHUNIA; SHIT, 2019). Um GIS pode reduzir vastas extensões de dados de tabelas em mapas visuais convincentes que podem oferecer informações importantes e subsídios para tomada de decisão tanto aos formuladores de políticas públicas em saúde quanto à população. Os GIS têm sido utilizados para mapear e prever a distribuição de doenças e investigar a correlação entre essas doenças e fatores climáticos, ambientais, demográficos e sócio-econômicos geoespaciais, tanto em nível nacional e internacional, quanto regional e local (BHUNIA; SHIT, 2019; PALANIYANDI; ANAND; PAVENDAR, 2017; PALANIYANDI et al., 2016; JOHNSON; JOHNSON, 2001; PALANIYANDI, 2014; EHLERS et al., 2003; KANDWAL; GARG; GARG, 2009; SHEPARD et al., 2011; FLEMING; MERWE; MCFERREN, 2007; CHANG et al., 2009; KHORMI; KUMAR, 2011). Entenda-se por dados geoespaciais ou geodados aqueles dados que estão associados à posição espacial, que pode ser expressa em latitude e longitude ou usando outro sistema de representação.

Nas soluções baseadas em GIS, comumente informações de diversas bases de dados diferentes precisam ser integradas. A Figura 1 ilustra o processo de integração de diferentes bases de geodados. Os dados das diferentes bases podem passar por transformações geométricas (rotações e translações), para que os modelos sejam ajustados e correspondentes às mesmas posições geográficas.

O uso de GIS permite análises espaço-temporais. Nas análises temporais, são visualizadas

Figura 1 – Diagrama esquemático do processo de integração de diferentes bases de geodados em um GIS. As transformações geométricas consistem em rotações e translações para que os modelos sejam ajustados e correspondentes às mesmas posições geográficas.



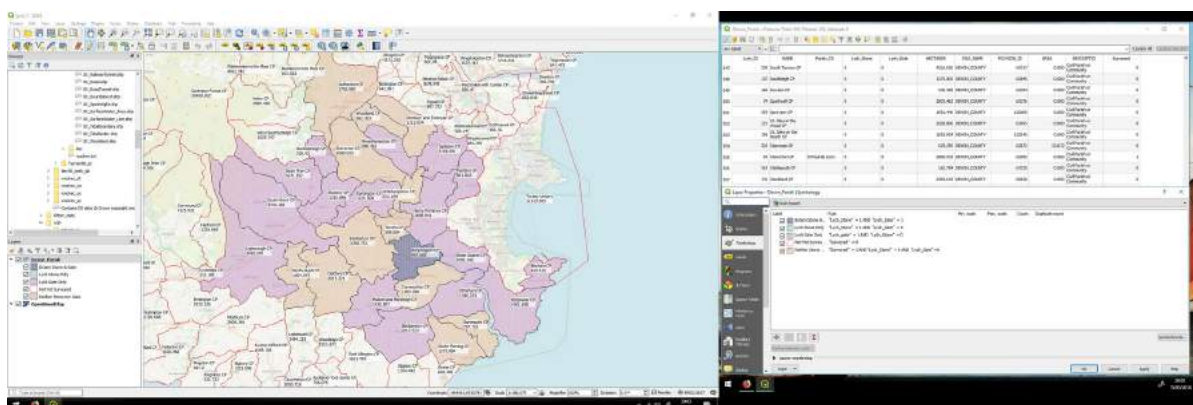
Fonte: A Autora

as séries históricas (também chamadas de séries temporais) dos dados na forma de gráficos em uma única dimensão. Na análise espaço-temporal, os dados são organizados na forma de séries históricas de distribuições espaciais, onde os dados são apresentados em mapas com suas distribuições estimadas correspondentes a cada intervalo de tempo. Dessa forma, tempo e espaço são conjugados em uma mesma análise.

Diversas ferramentas de visualização estão disponíveis para apoiar os usuários na visualização de geodados. O ArcView (<<http://www.esri.com/software/arcgis/arcgisonline>>) é

uma ferramenta comercial para visualizar e analisar dados geográficos dotada de uma série de métodos para fazer mapas temáticos personalizados e analisar dados espaciais. O XGobi (<http://stat-graphics.org/graphics/>), lançado em 1996, é um sistema universal de visualização de dados que também pode se conectar ao *software* ArcView. O Cartographic Data Visualizer (Visualizador de Dados Cartográficos, disponível em <https://www.visualrsoftware.com/cmp/data-visualization-demo/>) agrega mecanismos de mapeamento e visualização de dados em um único aplicativo. Ele lida com mapas dinâmicos, cartogramas populacionais e gráficos estatísticos. O Common GIS (<http://commongis.jrc.it>), desenvolvido pela FraunhoferIAIS, compreende uma grande diversidade de técnicas de geovisualização a partir de mapas dinâmicos, que podem ser combinados com gráficos de barra e gráficos de pizza. O GeoVISTA Studio (<http://www.geovista.in/>) é um *software* de código aberto, usado para geovisualização e estática. O Quantum GIS, ou QGIS, (<https://qgis.org/>) desenvolvido pela Open Source Geospatial Foundation (OSGeo) oferece uma ampla gama de recursos para pesquisa, edição e visualização de dados geoespaciais e funcionalidades adicionais. A Figura 2 apresenta uma visualização da tela de trabalho do QGIS. As ferramentas de geovisualização são compostas de diversas janelas, cada uma apresentando geodados diferentes, podendo ser utilizadas para apoiar a tomada de decisão (BHUNIA; SHIT, 2019). Os usuários podem se beneficiar de tarefas computacionalmente custosas (tarefas que ocupam muita memória ou requerem processamento intensivo), como visualizações em 2D e em 3D em diversas perspectivas.

Figura 2 – Visualização da tela de trabalho do Quantum GIS, com mapa e tabela de geodados exibidos simultaneamente.



Fonte: A Autora.

O Quantum GIS, ou QGIS, possui como vantagem sobre as outras plataformas GIS o fato de ser uma ferramenta poderosa e confiável disponível na forma de *software* livre e gratuito, podendo ser adicionadas extensões construídas na linguagem Python.

2.2 ARBOVIROSES: DENGUE, CHIKUNGUNYA E ZIKA

A dengue (DENV) é uma infecção viral sistêmica transmitida aos seres humanos por mosquitos do tipo *Aedes*. Para alguns pacientes, a dengue é uma doença potencialmente mortal. Atualmente, não existem vacinas licenciadas ou terapias específicas, e esforços substanciais de controle de vetores não impediram seu rápido surgimento e disseminação global. A distribuição mundial do risco de infecção pelo vírus da dengue e seu peso na saúde pública são pouco conhecidos, apesar da ampla quantidade e disponibilidade de informações e do esforço em pesquisa dispendido (DONALÍSIO; GLASSER, 2002; BHATT et al., 2013). A dengue está presente em todos os trópicos, com variações espaciais locais de risco influenciadas fortemente pela chuva, temperatura e grau de urbanização (BHATT et al., 2013). Usando abordagens cartográficas, estima-se que haja 390 milhões de infecções por dengue por ano. Esse total de infecção é mais de três vezes a estimativa de carga de dengue da Organização Mundial da Saúde (DONALÍSIO; GLASSER, 2002; BHATT et al., 2013).

A dengue é uma doença viral sistêmica aguda que se estabeleceu globalmente em ciclos de transmissão endêmica e epidêmica¹ (DONALÍSIO; GLASSER, 2002; BHATT et al., 2013).

A infecção pelo vírus da dengue em humanos muitas vezes não se manifesta, mas pode levar a uma ampla gama de manifestações clínicas, desde febre leve até a síndrome do choque, podendo esta última ser fatal. A imunidade vitalícia só é desenvolvida após a infecção pelo vírus da dengue do tipo 1 (BHATT et al., 2013). Existem quatro tipos. A progressão para doença mais grave é frequentemente, mas não exclusivamente, associada à infecção secundária pelos outros tipos. Ainda não existem agentes antivirais eficazes para tratar a infecção da dengue. Além disso, nenhuma vacina licenciada contra a dengue está disponível, e a vacina candidata de dengue mais avançada não atendeu às expectativas (BHATT et al., 2013). Os esforços atuais para conter a transmissão da dengue se concentram no vetor, usando combinações de alvos químicos e biológicos de mosquitos *Aedes* e manejo de potenciais criadouros (SIMMONS et al., 2012; BHATT et al., 2013; DONALÍSIO; GLASSER, 2002). Apesar de a expansão histórica desta doença estar bem documentada, a quantidade de problemas de saúde atribuídos à dengue em grande parte do mundo tropical e subtropical permanece pouco documentada.

¹ Na transmissão endêmica, a doença é mais contínua e está mais restrita a uma determinada área. Uma doença endêmica é, portanto, uma doença típica de uma determinada região, podendo ocorrer em determinados períodos. Um surto ocorre quando o número de casos de uma determinada doença aumenta repentinamente em uma determinada região. Diz-se que ocorreu uma epidemia quando ocorrem surtos em diversas regiões ao mesmo tempo. Consequentemente, na transmissão epidêmica, a doença se alastra rapidamente por diversas áreas.

O conhecimento da distribuição geográfica e dos potenciais problemas que a dengue pode provocar é essencial para entender sua contribuição para as cargas globais de morbidade e mortalidade, determinar como alocar os recursos limitados disponíveis para o controle da dengue e avaliar o impacto dessas atividades internacionalmente (BHATT et al., 2013; SIMMONS et al., 2012). Além disso, estimativas de distribuições de infecção com e sem sintomas aparentes formam um requisito fundamental para avaliar a vigilância clínica e para definir de forma confiável as futuras demandas de vacina (BHATT et al., 2013; SIMMONS et al., 2012).

Os sintomas da dengue vão desde febre e sintomas constitucionais leves até manifestações hemorrágicas e choque, ou dengue hemorrágico/síndrome do choque associada ao dengue (DH/SCD) (SINGHI; KISSOON; BANSAL, 2007). A dengue clássica é caracterizada por febre, cefaléia, mialgia (dores musculares), e sintomas constitucionais. O DH é um quadro clínico mais sério. Ele surgiu entre crianças no Sudeste Asiático na década de 1950 e desde então se transformou num grave problema de saúde pública mundial, sendo uma grande causa de morbidade e mortalidade pediátrica (SINGHI; KISSOON; BANSAL, 2007). As crianças acometidas pela doença precisam de acompanhamento cuidadoso. Outras complicações, tais como desconforto respiratório devido a intenso derrame pleural, disfunção do miocárdio, sangramento intenso e falência múltipla dos órgãos, inclusive síndrome de desconforto respiratório agudo, falência hepática aguda e falência renal aguda podem representar risco de morte e necessitam de atenção (SINGHI; KISSOON; BANSAL, 2007).

O chikungunya (CHIKV) é um vírus com origem nas regiões silvestres da África, originalmente circulando entre mosquitos do tipo *Aedes* e primatas não humanos. Ele possui três tipos: África Ocidental, Leste/Central/Sul da África e Asiático (HONÓRIO et al., 2015). Após ser isolado em 1952, na Tanzânia, a primeira emergência documentada do CHIKV ocorreu quando o vírus chegou ao sudeste asiático e à Índia (HONÓRIO et al., 2015; PIALOUX et al., 2007; BURT et al., 2012). De lá, vem se tornando um vírus urbano cujo principal vetor é o mosquito *Aedes aegypti*. A segunda emergência do CHIKV ocorreu no Quênia, em 2004, espalhando-se depois por ilhas do Oceano Índico, pela Índia e pelo sudeste asiático (HONÓRIO et al., 2015; PIALOUX et al., 2007). Em 2006, nas Ilhas da Reunião, ocorreu uma epidemia de CHIKV por meio de mutações do vírus que o fizeram ser transmitido também pelo mosquito *Aedes albopictus* (HONÓRIO et al., 2015). Em outubro de 2013, o CHIKV chegou ao Caribe. No Brasil, o CHIKV foi encontrado em setembro de 2014, na cidade de Oiapoque (Amapá) (HONÓRIO et al., 2015). Ao longo de 2014 foram confirmados 2.772 casos de CHIKV, distribuídos em seis estados: Amapá (1.554 casos), Bahia (1.214), Distrito Federal (2), Mato Grosso do Sul (1), Roraima

(1) e Goiás (1) (HONÓRIO et al., 2015). Em 2015, no período de 4 de janeiro a 28 de março de 2015, foram confirmados 1.513 casos: 735 no Amapá, onde foi identificado o tipo asiático, e 778 casos na Bahia, onde foi identificado o tipo africano (HONÓRIO et al., 2015).

A infecção por CHIKV produz dores articulares repentinas e debilitantes. Devido às fortes dores nas articulações provocadas pela febre, a doença recebeu o nome de Chikungunya que, no idioma africano makonde, significa “andar encurvado” (HONÓRIO et al., 2015). As dores nas articulações estão presentes em 80% dos pacientes, podendo durar meses ou até mesmo anos (HONÓRIO et al., 2015; PIALOUX et al., 2007). Após a cura da chikungunya, pode ocorrer até artrite reumatoide, interferindo na qualidade de vida do indivíduo e reduzindo sua produtividade em diversos aspectos da vida (HONÓRIO et al., 2015; PIALOUX et al., 2007). A doença pode inclusive levar a óbito idosos e crianças. O CHIKV causa doença neurológica em idosos e recém-nascidos, podendo ser fatal (HONÓRIO et al., 2015). A transmissão de mãe para filho é incomum, mas todos os recém-nascidos infectados durante o trabalho de parto apresentaram doença sintomática com manifestações graves (HONÓRIO et al., 2015). A gravidade da chikungunya em recém-nascidos e possibilidade de paralisia cerebral requerem medidas preventivas e terapêuticas.

Apesar da relativa facilidade para se diagnosticar a chikungunya a partir dos sintomas, a maioria dos pacientes suspeitos vive em área endêmica de dengue. Assim, a chance de erros no diagnóstico clínico é grande. Infecções graves como a malária em viajantes que retornam de regiões endêmicas podem ser subestimadas durante um surto de chikungunya, podendo levar à morte (HONÓRIO et al., 2015). A confirmação laboratorial rápida é fundamental para a condução clínico-terapêutica adequada e para o início de respostas de ações de controle.

As possibilidades de ocorrência de epidemias no Brasil são grandes, devido aos seguintes fatores: (1) presença dos dois principais vetores da CHIKV, o mosquito *Aedes aegypti* e os primatas não humanos, em todo o território nacional; (2) circulação simultânea de DENV e CHIKV, dificultando o diagnóstico e o tratamento; (3) adaptação do CHIKV ao *Ae. albopictus*, um mosquito muito mais comum que o *Ae. aegypti*; (4) maior proporção de casos com sintomas comparado ao dengue; (5) maior período de atuação do vírus: até 8 dias depois do início da febre; (6) todos os seres humanos são suscetíveis à doença, favorecendo a disseminação rápida do vírus; (7) abundância de espécies de primatas; (8) e, por fim, a grande extensão territorial do país, com regiões onde o acesso tanto a estratégias de controle do vetor quanto ao atendimento médico é difícil (HONÓRIO et al., 2015; PIALOUX et al., 2007). Entretanto, o desenvolvimento econômico não protege os países de doenças transmitidas por vetores.

Estilos de vida modernos podem amplificar uma epidemia por meio de viagens, envelhecimento da população e produção de resíduos sólidos, gerando criadouros para os mosquitos *Aedes* (HONÓRIO et al., 2015; LEPARC-GOFFART et al., 2014; GRANDADAM et al., 2011; REZZA et al., 2007).

O vírus Zika (ZIKV) é um flavivírus relacionado aos vírus da febre amarela, dengue, vírus do Nilo Ocidental e da encefalite japonesa (HAYES, 2009). Em 2007, o ZIKV causou um surto de doença relativamente leve, caracterizada por erupções cutâneas, artralgia (dores nas articulações) e conjuntivite na Ilha Yap, no sudoeste do Oceano Pacífico (HAYES, 2009). Esta foi a primeira vez que o ZIKV foi detectado fora da África e da Ásia (HAYES, 2009). O ZIKV foi inicialmente isolado de um macaco *rhesus* na floresta Zika, em Uganda, em 1947 (MLAKAR et al., 2016; DICK; KITCHEN; HADDOW, 1952; CAMPOS; BANDEIRA; SARDI, 2015). É transmitido por várias espécies de mosquitos *Aedes*. Após a primeira infecção humana pelo ZIKV, foram notificados alguns poucos casos no Sudeste Asiático e na África (MLAKAR et al., 2016). Em 2007 ocorreu um surto de ZIKV na Ilha Yap da Micronésia (MLAKAR et al., 2016). Também ocorreram epidemias de ZIKV na Polinésia Francesa, Nova Caledônia, Ilhas Cook e Ilha de Páscoa nos anos de 2013 e 2014 (MLAKAR et al., 2016). Em 2015, aconteceu o surto de ZIKV nas Américas. O Brasil foi o país mais afetado: 1,3 milhões de casos de ZIKV registrados até dezembro de 2015 (MLAKAR et al., 2016). Além da transmissão pelos mosquitos *Aedes aegypti* e *Aedes albopictus*, também há suspeitas de transmissão do ZIKV por meio de relação sexual com pessoa doente, mas essa relação ainda não foi comprovada (MUSSO et al., 2015).

O quadro clínico do ZIKV é semelhante ao da dengue e da chikungunya: febre, dor de cabeça, dores nas articulações, dores nos músculos e erupção maculopapular (manchas pequenas na pele) (MLAKAR et al., 2016; PETERSEN et al., 2016). Essa grande variedade de sintomas dificulta o diagnóstico diferencial (MLAKAR et al., 2016; PETERSEN et al., 2016). Além disso, esses sintomas são um tanto parecidos com casos de dengue e chikungunya (MLAKAR et al., 2016; CAMPOS; BANDEIRA; SARDI, 2015). Casos de manifestações neurológicas e de síndrome de Guillain-Barré ocorreram na Polinésia Francesa e no Brasil durante epidemias de ZIKV (MLAKAR et al., 2016; PETERSEN et al., 2016; CAMPOS; BANDEIRA; SARDI, 2015). Os relatórios do Ministério da Saúde apontam um aumento de 20% nos casos de microcefalia em recém-nascidos na região nordeste do país, o que indica uma possível associação entre infecção por ZIKV na gravidez e malformações fetais (MLAKAR et al., 2016; PETERSEN et al., 2016).

2.3 MÉTODOS DE PREDIÇÃO DE ARBOVIROSES

De acordo com Siriyasatien et al. (2018a), na criação de modelos de predição, o conjunto de dados é o elemento mais importante para garantir resultados relevantes e utilizáveis. O processo de predição implica necessariamente em informações históricas, e requer um número elevado de dados o suficiente para treinar modelos de predição. Além disso, deve haver dados suficientes para permitir testes válidos de previsões contra eventos históricos observados. Portanto, o volume e a variedade de dados é importante, assim como a correção dos dados a partir de fontes confiáveis. Tipicamente, os dados vêm de instituições governamentais, como o Ministério da Saúde, Institutos de Meteorologia, os Ministério da Agricultura e outras instituições formais, como hospitais. No entanto, não existe uma única fonte que possa fornecer dados cobrindo todos os aspectos da construção do modelo de predição. Enormes volumes de dados estão disponíveis a partir de novas fontes de dados na Internet e nas redes sociais. Esses volumes de dados têm recebido o nome de Big Data (SIRIYASATIEN et al., 2018a).

As fontes de dados podem ser convencionais ou modernas. Dados convencionais e dados estatísticos incluem dados médicos e epidemiológicos de centros de dados de saúde tradicionais (como hospitais), dados ambientais, meteorológicos (temperatura, humidade, velocidade dos ventos, chuvas etc.) e geográficos de outras bases de dados públicas, podendo-se obter grandes quantidades e variedade de dados, como diagnósticos, exames laboratoriais, medicação e dados clínicos auxiliares (SIRIYASATIEN et al., 2018a). As fontes de dados modernas podem ser as redes sociais (SIRIYASATIEN et al., 2018a), consultas em motores de busca na Internet (ALTHOUSE; NG; CUMMINGS, 2011; SIRIYASATIEN et al., 2018a), chamadas telefônicas e dados obtidos de sensores e biossensores (SIRIYASATIEN et al., 2018a).

A Internet permite a comunicação remota entre pacientes e médicos fora do ambiente hospitalar. As redes sociais podem ser usadas para conectar médicos com pacientes, pacientes com pacientes ou médicos com médicos, permitindo o compartilhamento de dados e de conhecimento. As redes sociais podem ser utilizadas para prever surtos de diversas doenças. A maioria das previsões de arboviroses foi baseada em dados ambientais, dados climáticos e meteorológicos e dados geográficos. Diversas pesquisas têm se dedicado a minerar dados em redes sociais como Facebook e Twitter para levantar dados úteis para predição de doenças, por meio da captura e análise automática de determinadas palavras chaves nos textos das postagens e captura da localização dos usuários (SIRIYASATIEN et al., 2018a; MATTHEWS et al., 2012; QUINCEY; KOSTKOVA, 2009; SZOMSZOR; KOSTKOVA; QUINCEY, 2010; KOSTKOVA, 2015).

Em países onde a difusão da Internet e dos smartphones ainda não é tão grande, pode-se utilizar também pesquisas por telefone para obter dados de predição (SIRIYASATIEN et al., 2018a). Já em realidades onde se dispõe de diversas soluções para Internet das Coisas (*Internet of Things*, IoT), pode-se usar de sensores acoplados a peças de roupas de pacientes (*wearable computing*, computação vestível) para monitorar sinais vitais dos pacientes e, assim, obter mais rapidamente seu diagnóstico e disponibilizar esses dados em bancos de dados hospitalares e até mesmo na Internet (SIRIYASATIEN et al., 2018a). Biossensores são comumente dedicados a testes rápidos de diagnóstico. Eles também podem ser utilizados de forma semelhante aos sensores comuns, utilizando da IoT para disponibilizar rapidamente esses dados (SIRIYASATIEN et al., 2018a).

Uma vez adquiridos os dados, eles devem ser organizados na forma de arquivos de conjunto de treinamento e de teste. O conjunto de treino é utilizado para treinar o modelo preditivo. Já o conjunto de teste é utilizado para testar a capacidade de generalização do modelo, sendo composto por dados que não estão presentes no conjunto de treino. Hoje é mais recomendado utilizar validação cruzada: uma metodologia onde o conjunto total é dividido em partes e cada modelo preditivo é testado variando o conjunto de treino e o de teste. Assim, o melhor modelo preditivo é escolhido de acordo com seu comportamento geral para os diversos treinamentos e testes.

Os modelos preditivos podem ser: métodos estatísticos (exemplos: regressão linear e logística) ou métodos baseados em aprendizado de máquina, como as redes neurais artificiais, as árvores de decisão e outros. Também costumam ser muito utilizados os modelos baseados em equações diferenciais, mas esses modelos estão sendo aos poucos abandonados, dada a sua alta complexidade matemática e resultados não tão precisos para situações reais, pois necessitam da modelagem matemática da população do mosquito e da população humana (SIRIYASATIEN et al., 2018a; YAMASHITA; DAS; CHAPIRO, 2018; LI; CAI; LI, 2017; LU; LI, 2011; WAN; ZHU, 2014; WAN, 2013; YAMASHITA; TAKAHASHI; CHAPIRO, 2016). Assim, os modelos baseados em equações diferenciais não serão abordados neste trabalho. No Capítulo 3 é apresentado o estado da arte das técnicas de previsão de arboviroses.

2.4 METODOLOGIA DE VIGILÂNCIA SANITÁRIA E AMBIENTAL NO SUS

O Departamento de Vigilância das Doenças Transmissíveis, da Secretaria de Vigilância em Saúde do Ministério da Saúde, definiu uma metodologia para, no âmbito do Sistema Único

de Saúde (SUS), levantar rapidamente estatísticas sobre a difusão do vetor da dengue e, hoje, também da chikungunya e da zika: o Levantamento Rápido de Índices para o *Aedes aegypti*, LIRAA (BRASIL, 2012). São métodos simplificados de amostragem para facilitar a obtenção pelos serviços de saúde de informações para o controle do vetor. A adoção desses métodos simplificados para levantamento epidemiológico tem sido estimulada tanto pela Organização Mundial da Saúde (OMS) quanto pela Organização Pan-Americana da Saúde (Opas) (BRASIL, 2012).

De acordo com BRASIL (2012) amostragem é feita da seguinte forma: os municípios devem ser divididos em estratos que apresentem características sócio-ambientais semelhantes; cada estrato deve ter entre 8.100 e 12.000 imóveis, embora o número recomendado seja de 9.000 imóveis; esses imóveis são distribuídos em bairros; os bairros a serem visitados são definidos de forma aleatória; dentro de um bairro a ser visitado, são escolhidos 20% dos imóveis. Se o município for pequeno, os estratos deverão ser menores: de 2.000 a 8.100 imóveis, e deverão ser visitados 50% dos imóveis de cada estrato. Por simplificação, para municípios médios e grandes devem ser visitados 450 imóveis por estrato. Para municípios pequenos existe uma fórmula de correção para calcular o melhor número de imóveis a serem visitados por estrato (BRASIL, 2012).

O LIRAA é composto das seguintes métricas (BRASIL, 2012):

IP: Índice Predial, definido como a razão entre o número de imóveis positivos (ou seja, imóveis onde foi detectada a presença do mosquito *Aedes aegypti*, seja na forma adulta, seja na forma larval) e o número de imóveis visitados:

$$IP = \frac{\#IMP}{\#IMV} \times 100\%, \quad (2.1)$$

onde $\#IMP$ é o número de imóveis positivos e $\#IMV$ é o número de imóveis visitados.

IB: Índice de Breteau, definido como a razão entre o número de recipientes positivos (ou seja, recipientes onde foi detectada a presença do mosquito *Aedes aegypti* na forma larval) e o número de imóveis visitados:

$$IB = \frac{\#RCP}{\#IMV} \times 100\%, \quad (2.2)$$

onde $\#RCP$ é o número de recipientes positivos e $\#IMV$ é o número de imóveis visitados.

ITR: Índice por Tipo de Recipiente, definido como a razão entre o número de recipientes positivos de um determinado tipo (ou seja, recipientes onde foi detectada a presença do mosquito *Aedes aegypti* na forma larval) e o total de recipientes positivos:

$$\text{ITR} = \frac{\#RCX}{\#RCP} \times 100\%, \quad (2.3)$$

onde $\#RCX$ é o número de recipientes positivos de um determinado tipo X e $\#RCP$ é o número total de recipientes positivos.

Os recipientes podem ser dos seguintes tipos: o grupo A trata de grandes depósitos de armazenamento de água; o grupo B trata de depósitos móveis; o grupo C envolve depósitos fixos; o grupo D engloba recipientes passíveis de remoção; já o grupo E trata de recipientes naturais (BRASIL, 2012).

- A1:** Depósito de água elevado, ligado à rede pública ou ao sistema de captação mecânica em poço, cisterna ou mina d'água, caixas d'água, tambores, depósitos de alvenaria (ver Figura 3);
- A2:** Depósitos ao nível do solo para armazenamento doméstico (tonel, tambor, barril, tina), depósitos de barro (filtros, moringas, potes), cisternas, caixas d'água, captação de água em poço, cacimba ou cisterna (ver Figura 4);
- B:** Depósitos móveis: vasos e frascos com água, pratos, garrafas retornáveis, pingadeiras, recipientes de degelo em geladeiras, bebedouros em geral, pequenas fontes ornamentais, materiais em depósitos de construção (sanitários estocados, canos, etc.), objetos religiosos e rituais (ver Figura 5);
- C:** Depósitos fixos: tanques em obras de construção civil, borracharias e hortas, calhas, lajes e toldos em desnível, ralos, sanitários em desuso, piscinas não tratadas, fontes ornamentais; cacos de vidro em muros, outras obras e adornos arquitetônicos, como caixas de inspeção e de passagem (ver Figura 6);
- D1:** Pneus e outros materiais rodantes, como câmaras de ar e manchões (ver Figura 7);
- D2:** Resíduos sólidos: recipientes plásticos, garrafas PET, latas, sucatas, entulhos de construção (ver Figura 8);
- E:** Axilas de folhas (exemplo: bromélias), buracos em árvores e em rochas, restos de animais (cascas, carapaças, etc.), conforme Figura 9.

Figura 3 – Potenciais criadouros de tipo A1.



Fonte: BRASIL (2012)

Figura 4 – Potenciais criadouros de tipo A2.



Fonte: BRASIL (2012)

Para controlar os vetores (mosquitos, ratos e outros potenciais vetores de doenças contagiosas), os municípios organizam os Centros de Vigilância Ambiental, de acordo com uma estrutura geral sugerida pelo Ministério da Saúde. No caso específico do controle de mosquitos *Aedes aegypti*, para operacionalizar o levantamento do LIRAa, os times são organizados de acordo com as seguintes funções com seus respectivos atributos:

Coordenador:

- Buscar apoio e sustentabilidade para a realização do LIRAa;
- Estratificar e calcular o número de imóveis a pesquisar;
- Definir os estratos, considerando-se as características sócio-ambientais;
- Definir os quarteirões a serem trabalhados, utilizando-se do sistema de informação disponível;

Figura 5 – Potenciais criadouros de tipo B.



Fonte: BRASIL (2012)

- Definir os recursos humanos necessários: supervisores, agentes, laboratoristas e digitadores;
- Definir necessidades: veículos, equipamentos e material de campo (componentes da bolsa de trabalho de campo);
- Definir o(s) laboratório(s) de apoio;
- Planejar ações de supervisão durante a realização do LIRAa;
- Analisar dados e elaborar relatório final;
- Planejar as ações necessárias, após o levantamento, para as áreas consideradas mais críticas: operações de campo, ações de informação, educação e comunicação social.

Supervisor:

- Organizar e distribuir os agentes na área de trabalho;
- Abastecer os agentes de controle de endemias com os insumos necessários;

Figura 6 – Potenciais criadouros de tipo C.



Fonte: BRASIL (2012)

Figura 7 – Potenciais criadouros de tipo D1.



Fonte: BRASIL (2012)

Figura 8 – Potenciais criadouros de tipo D2.



Fonte: BRASIL (2012)

- Supervisionar as atividades dos agentes de controle de endemias;
- Receber e conferir o Boletim de Campo e Laboratório (BCL);
- Consolidar os dados no BCL e preencher o Resumo Parcial (RP);
- Encaminhar ao laboratório os BCLs e RPs com as amostras coletadas;
- Enviar ao setor de digitação os BCLs e RPs por estrato.

Agente de Controle de Endemias:

- Visitar de 20 a 25 imóveis por dia;
- Realizar minuciosa pesquisa larvária nos imóveis definidos no estrato;
- Coletar e preencher os rótulos dos tubitos;

Figura 9 – Potenciais criadouros de tipo E.



Fonte: BRASIL (2012)

- Registrar as informações no formulário BCL;
- Repassar, ao final do dia, o BCL devidamente preenchido ao supervisor.

Há alguns anos o Agente de Controle de Endemias (ACE) não existia como um profissional diferente do Agente de Saúde (BRASIL, 2012). Hoje são profissionais diferentes cujas denominações e atribuições podem diferir ligeiramente de acordo com o município. Por exemplo, em Recife o ACE é chamado de Agente de Saúde Ambiental e Controle de Endemias (ASACE). Os ASACEs mais recentes são profissionais de nível superior, o que torna o controle de vetores em Recife diferenciado em relação a outros municípios. A carga de trabalho dos ACEs também pode variar de acordo com o município, não obedecendo ao manual do LIRAa. Em Olinda, por exemplo, os ACEs devem visitar até 35 imóveis por dia.

2.5 REGRESSÃO LINEAR

A regressão linear é o método mais simples de regressão. Neste método, assume-se que os dados apresentam um comportamento linear, e que a variável de predição possa ser representada como uma combinação linear dos atributos com seus pesos pré-determinados. Sendo assim, o modelo geral da regressão linear é representado pela Equação 2.4:

$$y = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n, \quad (2.4)$$

onde y é a variável de predição; $x_1, x_2, \dots, x_n$, representam os valores dos atributos e $w_0, w_1, w_2, \dots, w_n$ representam os pesos de cada atributo. A ideia do algoritmo da regressão linear é encontrar os pesos ótimos que melhor representam o problema. Uma das maneiras de encontrar os pesos ótimos é fazendo com que a soma do quadrado da diferença entre o valor predito e o valor real seja mínimo. A soma do quadrado da diferença é calculada pela Equação 2.5:

$$S = \sum_{i=1}^n \left[y^{(i)} - \sum_{j=0}^k w_j x_j^{(i)} \right]^2. \quad (2.5)$$

2.6 PRINCÍPIOS DE APRENDIZADO DE MÁQUINA

Nesta seção são apresentadas breves revisões sobre os princípios dos métodos de aprendizado de máquina utilizados neste trabalho: redes neurais artificiais, do tipo Perceptron Multicamadas, Máquinas de Aprendizado Extremo, e Máquinas de Estado de Eco rasas e profundas; Máquinas de Vetor de Suporte lineares e baseadas em kernel; e árvores de decisão e Random Forests. As redes neurais convolucionais não estão no escopo deste trabalho.

2.6.1 Redes Neurais Artificiais

Redes Neurais Artificiais (RNA's), consistem em uma técnica de aprendizado de máquina baseada no comportamento do cérebro humano (SIRIYASATIEN et al., 2018b). As redes neurais consistem em redes de unidades menores interconectadas, os neurônios artificiais. A capacidade de aprendizado de uma rede neural artificial depende da forma como os seus neurônios estão organizados e de como são as suas interconexões. Os neurônios artificiais contêm os seguintes elementos: (1) um conjunto de sinapses ou conectores, onde está um sinal x_i na entrada da sinapse j conectado ao k -ésimo neurônio multiplicado pelo peso sináptico $w_{k,j}$; (2) um

somador para somar os sinais de entrada, ponderados pelas respectivas sinapses neuronais; (3) uma função de ativação, para limitar a saída de um neurônio (Haykin, 2001). Matematicamente, um neurônio artificial é representado pela Equação 2.6 e pela Equação 2.7:

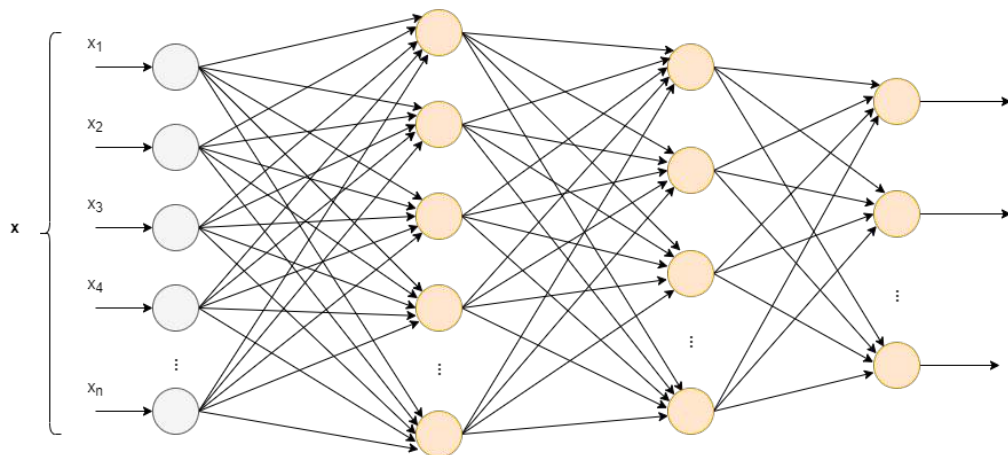
$$u_k = \sum_{j=1}^n w_{k,j} x_j \quad (2.6)$$

$$y_k = \varphi(u_k + b_k), \quad (2.7)$$

em que $x_1, x_2, \dots, x_n$ representam os sinais de entrada; $w_{k,1}, w_{k,2}, \dots, w_{k,n}$ representam os pesos sinápticos dos sinais de entrada x_i para o k -ésimo neurônio; b_k , é o termo bias, e φ é uma função de ativação de neurônios. Em aplicações de regressão, as entradas $x_1, x_2, \dots, x_n$ da camada de entrada correspondem à janela de previsão, no caso de séries temporais. Por exemplo, no caso de previsão temporal, as entradas são a janela de tempo observada da série temporal.

Uma arquitetura de rede neural artificial muito utilizada é o Perceptron Multicamadas (*Multilayer Perceptron*, MLP). Nesta configuração, a rede neural possui uma camada de entrada, duas ou mais camadas ocultas e uma camada de saída (Haykin, 2001). Neste tipo de arquitetura, as entradas dos neurônios são os valores de saída da camada anterior, como pode ser visto na Figura 10.

Figura 10 – Representação gráfica de uma rede neural Perceptron Multicamadas com n entradas e três camadas, ou seja, duas camadas escondidas e uma camada de saída.



Fonte: A autora

Atualmente existe um grande interesse por redes neurais profundas, devido à sua capacidade de modelar funções complexas. Uma rede é considerada profunda quando ela possui

muitas camadas escondidas. O algoritmo clássico para treinar uma MLP é o *backpropagation*, onde os erros das saídas das camadas escondidas são estimados a partir do erro da camada de saída. O problema é que, quando a rede é profunda, as camadas mais distantes da camada de saída têm estimativas de erros muito imprecisas. Isso faz com que não seja possível usar esse algoritmo para treinar uma rede MLP profunda. Existem novas versões desse algoritmo utilizadas para treinar redes profundas, mas cujos neurônios possuem matrizes de pesos, não vetores de pesos, como as MLPs clássicas. Essas redes são chamadas de Redes Neurais Convolucionais (*Convolutional Neural Networks*, CNNs), e são especialmente úteis para trabalhar com imagens como entrada, mas o estudo das CNNs foge ao escopo deste trabalho, focado em regressão (SCHMIDHUBER, 2015; TAVANAEI et al., 2019). Para regressão, outras redes neurais, treinadas com outros algoritmos, profundas ou rasas, podem ser utilizadas, como as Máquinas de Estado de Eco, por exemplo (TCHAKOUCHT; EZZIYYANI, 2018; MALIK; HUSSAIN; WU, 2016; GALLICCHIO; MICHELI; PEDRELLI, 2018; MA; SHEN; COTTRELL, 2020).

2.6.2 Máquinas de Vetor de Suporte

A regressão por Máquina de Vetor de Suporte (SVM-R ou simplesmente SVR) é uma técnica de aprendizado de máquina supervisionada para análise de dados e reconhecimento de padrões. A ideia do algoritmo SVR é encontrar o melhor hiperplano definido pela função de custo de Vapnik. Quando este hiperplano é encontrado, uma regressão linear é aplicada ao hiperplano correspondente. Em situações onde o problema é linearmente separável, o melhor hiperplano é dado pela Equação 2.8:

$$y = \mathbf{w}^T \mathbf{x} + b, \quad (2.8)$$

onde $\mathbf{w} = (w_1, w_2, \dots, w_n)^T$ é o vetor de pesos, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ é o vetor de características ou vetor de atributos, e b é o *bias*.

Em problemas linearmente separáveis, podem existir infinitos hiperplanos. No algoritmo de treinamento da SVM, primeiramente são encontrados os vetores de suporte, ou seja, os vetores que ficam nas fronteiras entre as classes e que minimizam a função de custo. Isso é feito usando o algoritmo de busca Otimização Mínima Sequencial (*Sequential Minimal Optimization*, SMO). Para problemas que não são linearmente separáveis, os dados são mapeados para um hiperplano em uma dimensão maior. Com isso, o algoritmo busca resolver o problema aplicando a regressão linear da Equação 2.8 no hiperplano correspondente. Para problemas

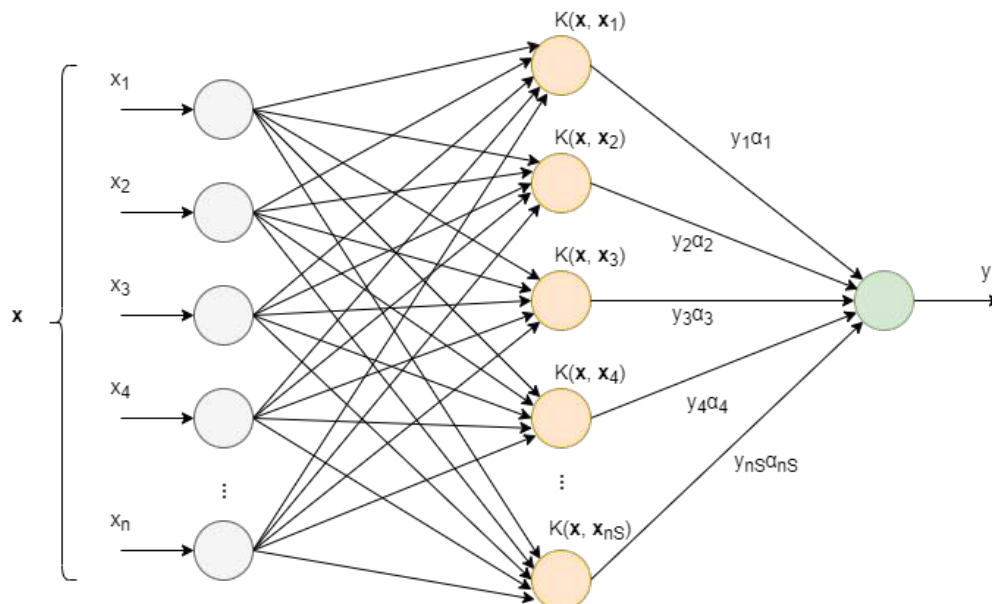
não linearmente separáveis, as máquinas SVR usam funções de kernel, $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$. Então, a saída SVR assume a seguinte expressão:

$$y = \mathbf{w}^T \varphi(\mathbf{x}) + b = \sum_{i=1}^{n_S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b, \quad (2.9)$$

onde n_S é o número de vetores de suporte, $\mathbf{x}_i$ é o i -ésimo vetor de suporte, y_i é sua respectiva saída desejada no conjunto de treinamento, e α_i é o seu peso para o cálculo da saída do SVM. A função do kernel $K(\mathbf{x}_i, \mathbf{x}) = \varphi(\mathbf{x}_i)\varphi(\mathbf{x})$ pode ser polinomial, sigmoideal, Gaussiana, ou mesmo assumir outras expressões matemáticas (DRUCKER et al., 1997; WITTEN; FRANK, 2005; SMOLA; SCHÖLKOPF, 2004).

Matematicamente, uma SVM se comporta como uma MLP com uma única camada escondida, com n_S neurônios e cujos pesos correspondem aos vetores de suporte. Esses neurônios realizariam operações de kernel, enquanto o neurônio de saída é linear. A Figura 11 ilustra como seria uma SVM vista como se fosse uma rede neural.

Figura 11 – Arquitetura geral de uma máquina de vetor de suporte com uma única saída: o cálculo dos kernels com as entradas e os vetores de suporte pode ser visto como uma camada escondida com um tipo especial de neurônio, enquanto o cálculo da saída tem um neurônio usual linear. A camada escondida teria n_S neurônios cujos vetores de pesos seriam os vetores de suporte $\mathbf{x}_i$, para $i = 1, 2, \dots, n_S$, e a saída determinada pelo kernel K .



Fonte: A autora

2.6.3 Máquinas de Aprendizado Extremo

As Máquinas de Aprendizado Extremo (*Extreme Learning Machines*, ELMs) surgiram como um método alternativo de aprendizagem para redes de alimentação de camada única, (*Single-Layer Feed-Forward Networks*, SLFN's), que geralmente dependem do algoritmo de retropropagação para tarefas de classificação ou regressão. Criadas por (HUANG; ZHU; SIEW, 2004), as ELMs consistem em SLFN's, nas quais os pesos e vieses (*biases*) entre a camada de entrada e a camada oculta são definidos aleatoriamente, enquanto apenas os pesos de saída são determinados analiticamente, como um sistema linear, usando a pseudo-inversa generalizada de Moore-Penrose (BARATA; HUSSEIN, 2012). A confiabilidade de ELM's como uma alternativa para SLFN's é encontrada no fato de que os pesos e vieses entre a camada de entrada e a camada oculta de um SLFN não precisam ser ajustados, se garantirmos que suas funções de ativação são infinitamente diferenciáveis (HUANG; ZHU; SIEW, 2006), e que um SLFN com uma camada oculta com N neurônios suficientes e com quase qualquer função de ativação será capaz de aprender N observações distintas (HUANG; BABRI, 1998).

As ELMs também fornecem outras características desejadas para aplicações reais, incluindo uma velocidade de aprendizagem que pode ser maior do que as redes convencionais diretas (*feed-forward*), devido à fase de treinamento *one-shot* (uma única iteração); a tendência de alcançar um melhor desempenho de generalização; menores erros médios de treinamento; e a menor norma dos vetores de pesos (HUANG; ZHU; SIEW, 2006). De acordo com Bartlett (1998), quanto menor a norma de pesos, melhor desempenho de generalização a rede neural tende a atingir.

O algoritmo ELM proposto por Huang, Zhu e Siew (2006) é o seguinte:

Suponha um conjunto de treinamento:

$$\Psi_T = \{(\mathbf{x}_i, \mathbf{t}_i) \mid \mathbf{x}_i \in \mathbb{R}^n, \mathbf{t}_i \in \mathbb{R}^m\}_{i=1}^{N_T},$$

uma função de ativação $g : \mathbb{R} \rightarrow \mathbb{R}$ e um número N de nós ocultos, faça

1. Aleatoriamente, atribua pesos de entrada e polarização, $\mathbf{w}_i$ e b_i , respectivamente, para os neurônios de entrada $i = 1, \dots, N$;
2. Calcula a matriz de saída da camada oculta $\mathbf{H}$, com entradas $H_{i,j} = g(w_j x_i + b_j)$;
3. Calcule o peso de saída β , de $\beta = \mathbf{H}^\dagger \mathbf{O}$, onde $\mathbf{O} = [o_1, \dots, o_N]^T$ é a saída desejada do conjunto de dados de treinamento.

2.6.4 Máquinas de Estado de Eco

Jaeger (2001) e Maass, Natschläger e Markram (2002) marcaram o nascimento da Reservoir Computing (RC). As técnicas propostas consistem em uma nova abordagem para o treinamento e uso de redes neurais recorrentes, e são denominadas Redes de Estado do Eco (ESN) e Máquinas de Estado Líquido (LSM). Posteriormente, uma nova regra de aprendizagem chamada BackPropagation DeCorrelation emergiu de um background completamente diferente, mas apresentando conceitos muito semelhantes se comparados com as duas abordagens anteriores (VERSTRAETEN, 2009).

A facilidade de uso e o excelente desempenho dessas técnicas aumentaram o interesse da academia em usá-las, e as pesquisas nesse campo ganharam impulso. Dessa forma, ESN e LSM foram aplicados em vários problemas, conforme explicado nos parágrafos a seguir.

Jaeger (2001) aplicou ESN à classificação de padrão dinâmico (i.e. padrões com características estocásticas variantes no tempo), testando a rede para aprendizagem de sequência periódica, usando um sinal alvo preparado a partir da melodia “*The House of The Rising Sun*” e aprendizagem de atrator de ponto comutável. A rede, quando devidamente sintonizada, mostrou capacidade suficiente para gerar sequências periódicas estáveis de sons robustos a ruídos.

Ghani et al. (2008) também mostrou o desempenho do ESN quando aplicado ao reconhecimento de fala. O desempenho da rede foi comparado às redes neurais *feed-forward* (diretas, sem realimentação) normais, e os resultados sugeriram que o reservatório, aliado às redes neurais *feed-forward*, apresentaram melhores taxas de classificação do que apenas as redes neurais. Resultados muito bons também foram alcançados quando uma etapa de pré-processamento foi adicionada à tarefa de classificação.

A abordagem proposta de ESN consiste em usar Rede Neural Recorrente (RNN) com topologia e pesos definidos aleatoriamente, onde a rede é dirigida por sua entrada externa e então a resposta obtida é usada para treinar uma regressão linear ou função de classificação (VERSTRAETEN, 2009), onde apenas a camada de leitura é treinada.

As ESN's muitas vezes apresentam bom desempenho, gerenciado pela Propriedade de Estado de Eco (ESP), que afirma que as redes devem esquecer assintoticamente seu estado inicial quando acionadas por um sinal externo (VERSTRAETEN, 2009). Além disso, dado o ESP, as ESN's devem lentamente esquecer as entradas iniciais quando uma nova chega, causando a *memória desbotada*, ou *memória em declínio*. Esta propriedade também permite

a reminiscência de ecos, enriquecendo o conjunto de transformações não lineares e misturas dos sinais atuais e anteriores, dentro de seu reservatório.

Algumas características de desempenho desejadas, alcançadas pela RC, são a extensão não linear dos dados de entrada para outros domínios, onde a tarefa de classificação pode ser realizada com mais precisão. Isso permite o uso de uma estrutura relativamente simples e de classificação linear ou algoritmos de regressão pouco exigentes em termos de computação. Além disso, como os reservatórios são criados aleatoriamente, eles toleram alguma variação interna em seus parâmetros, sem comprometer significativamente seu desempenho. Gallicchio, Micheli e Pedrelli (2018) propôs uma arquitetura de rede profunda para as ESNs, mantendo basicamente o mesmo algoritmo de treinamento.

A ESN proposta inicialmente por Jaeger (2001) e aperfeiçoada por Lukoševičius (2012), pode ser resumido da seguinte forma:

1. Gera um reservatório aleatório de Redes Neurais Recorrentes, com pesos de entrada aleatórios (W^{in}), pesos de matrizes recorrentes aleatórios ($\mathbf{W}$) e uma taxa de vazamento aleatória α ;
2. Execute a topologia usando a entrada de treinamento, $\mathbf{u}(n)$, e coletando os resultados correspondentes $\mathbf{x}(n)$;
3. Calcule os pesos da leitura linear, W^{out} , do reservatório, usando regressão linear e minimizando o erro entre a saída obtida, $y(n)$ e o resultado desejado, $y^{desejado}(n)$;
4. Use a rede treinada em novos dados de entrada, calculando a saída obtida $y(n)$ empregando os pesos de saída W^{out} .

2.6.5 Árvore de Decisão e Random Forests

As árvores de decisão estão entre as técnicas mais antigas da Inteligência Computacional. Árvores de Decisão (*Decision Trees*, DTs) são máquinas de aprendizado que obedecem ao paradigma baseado em regras. Ainda são bastante utilizadas, por serem modelos considerados simples e eficientes (MYLES et al., 2004; NOWOZIN et al., 2011; ZHOU; CHEN, 2002; RAO et al., 2019; LI; XU; DENG, 2019).

Uma árvore de decisão é construída a partir do modelo de grafo. Ela é composta de *nós* (ou *nodos*, como também são chamados), *ramos* e *folhas*. A árvore começa com o *nó raiz*.

Os nós são ligados a outros nós e distribuídos em *níveis*. Cada nó é ligado a outros dois nós localizados no nível imediatamente abaixo deste. Esses dois nós são chamados de *nós filhos*, enquanto o nó que dá acesso a esses dois, localizado no nível imediatamente superior, é o *nó pai*. Em cada nó, o valor de um atributo do vetor de atributos de entrada é comparado com um determinado *valor de referência* do nó. Assim, se o valor do atributo associado a esse nó for maior do que esse valor de referência, é verificado o resultado da comparação no nó filho correspondente. Caso contrário, é verificado o resultado da comparação do outro nó filho. A profundidade de uma árvore é dada pelo número de níveis. O último nó é chamado de nó folha.

O treinamento de uma árvore de decisão consiste em definir a profundidade máxima da árvore e calcular os valores de referência em cada nó. O conjunto de treino completo é apresentado ao nó raiz. Otimizando uma determinada métrica de qualidade, calcula-se o melhor valor de referência que permita dividir o conjunto de treino em duas metades, sendo encaminhada uma metade do conjunto de treino para cada nó filho (MYLES et al., 2004; NOWOZIN et al., 2011; ZHOU; CHEN, 2002; RAO et al., 2019; LI; XU; DENG, 2019). O processo de determinação dos melhores valores de referência e divisão dos subconjuntos para os nós filhos prossegue até que seja atingido o último nível da árvore ou não se possa mais dividir em subconjuntos. Nesse último caso, a árvore ficará com menos níveis que o especificado no início do treinamento (MYLES et al., 2004; NOWOZIN et al., 2011; ZHOU; CHEN, 2002; RAO et al., 2019; LI; XU; DENG, 2019). A Figura 12 apresenta um esquema geral de uma árvore de decisão e definição de nós e folhas.

A *entropia* é uma métrica da teoria da informação que mede a impureza ou incerteza em um grupo de observações. Ele determina como uma árvore de decisão escolhe dividir os dados. Dado um conjunto de J classes $\Omega = \{\omega_1, \omega_2, \dots, \omega_J\}$, se p_i for a fração de instâncias da classe ω_i colocada em um determinado subconjunto do nó, para $i = 1, 2, \dots, J$, a entropia $E(p)$ é dada por:

$$E(p) = - \sum_{i=1}^J p_i \log p_i, \quad (2.10)$$

onde p_i é a probabilidade de selecionar aleatoriamente um exemplo na i -ésima classe.

Pode-se definir *ganho de informação* como uma medida de quanta informação um atributo fornece sobre uma classe. O ganho de informação ajuda a determinar a ordem dos atributos nos nós de uma árvore de decisão. O nó principal é referido como o nó pai, enquanto os sub-nós são conhecidos como nós filhos. Em cada nível da árvore de decisão, calcula-se o ganho de

Figura 12 – Esquema geral de uma árvore de decisão, indicando o nó raiz, os nós de decisão e os nós folhas.



Fonte: A autora

informação para cada nó. O nó com o maior ganho de informação será expandido. O cálculo do ganho de informação $G(p)$ é dado por:

$$G(p) = E_{\text{pai}}(p) - E_{\text{filho}}(p). \quad (2.11)$$

O algoritmo CART (*Classification and Regression Tree*) utiliza como métrica para dividir o conjunto de treinamento em partes a *Impureza de Gini*, também conhecida como *Entropia de Tsallis de grau 2*. No processo de treinamento, o valor de referência de um dado nó é variado e é calculada a Impureza de Gini correspondente. Quando a Impureza de Gini atinge seu valor mínimo, o conjunto de treino no nó pai é dividido igualmente e com características distintas para os dois nós filhos subsequentes (MYLES et al., 2004; NOWOZIN et al., 2011; ZHOU; CHEN, 2002; RAO et al., 2019; LI; XU; DENG, 2019). Matematicamente, dado um conjunto de J classes $\Omega = \{\omega_1, \omega_2, \dots, \omega_J\}$, se p_i for a fração de instâncias da classe ω_i colocada em um determinado subconjunto do nó, para $i = 1, 2, \dots, J$, a Impureza de Gini é calculada como segue:

$$I_G(p) = \sum_{i=1}^J \left(p_i \sum_{k \neq i} p_k \right) = 1 - \sum_{i=1}^J p_i^2. \quad (2.12)$$

A tomada de decisão se dá a partir dos nós folhas: em problemas de classificação, a decisão é pelo número de instâncias do conjunto de treino que ficaram classificadas naquela folha; já

em problemas de regressão, a decisão é pela média dos valores das instâncias associadas àquela folha (MYLES et al., 2004; NOWOZIN et al., 2011; ZHOU; CHEN, 2002; RAO et al., 2019; LI; XU; DENG, 2019).

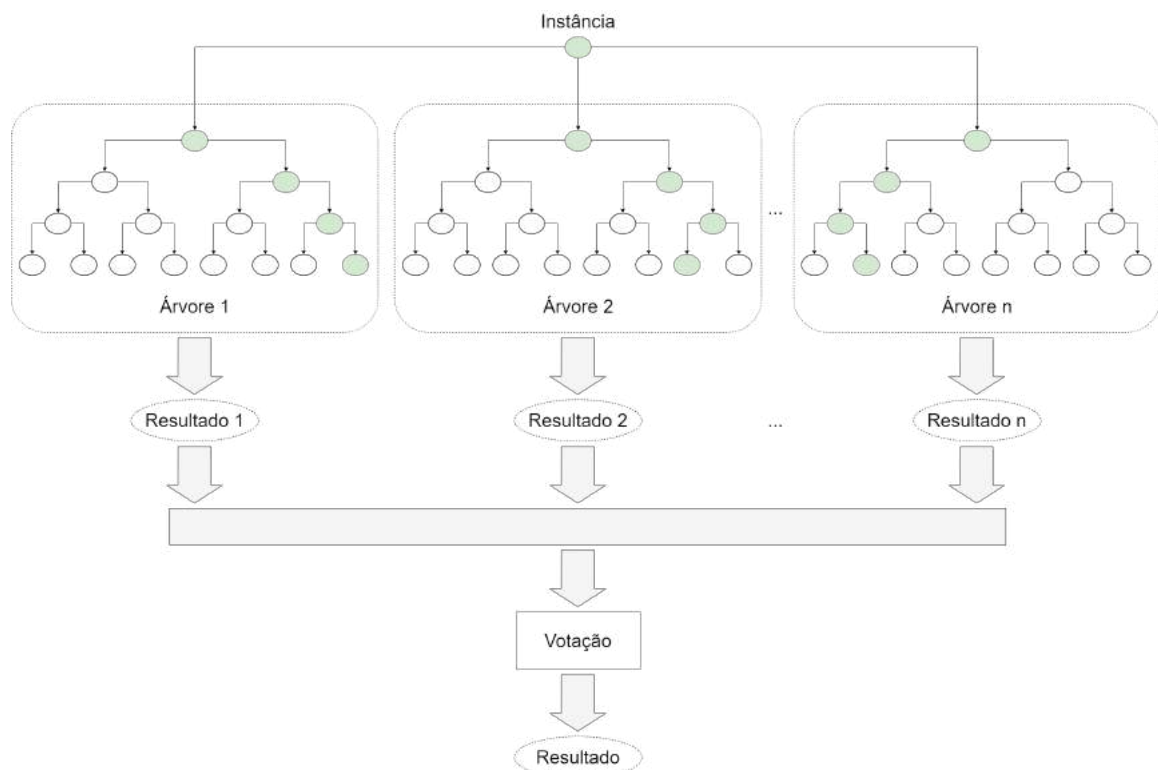
Árvores de decisão são modelos ainda bastante populares nas aplicações de aprendizado de máquina devido aos seguintes fatores (MYLES et al., 2004; NOWOZIN et al., 2011; ZHOU; CHEN, 2002; RAO et al., 2019; LI; XU; DENG, 2019):

- Árvores de decisão possuem execução rápida por serem máquinas de aprendizado baseadas em regras simples;
- São métodos de treinamento rápido: o processo de cálculo dos valores de referência dos nós não demanda muito processamento;
- Árvores de decisão são métodos bastante flexíveis: lidam bem com instâncias com muitos atributos, atingindo desempenho de classificação ou de regressão razoáveis para um treinamento pouco custoso, o que resulta uma boa relação custo-benefício;
- São modelos relativamente fáceis de explicar a pessoas que não são especialistas em Inteligência Computacional, sendo muitas vezes classificados como modelos transparentes e inteligíveis, ao contrário das redes neurais artificiais, que são modelos do tipo caixa preta, que não são fáceis de serem explicados e que os não especialistas muitas vezes precisam aceitar o método em função de eventuais bons resultados;
- Por serem de treinamento rápido, árvores lidam bem também com conjuntos de treino grandes (*big data*);
- As árvores de decisão podem ser utilizadas para determinar os atributos estatisticamente mais relevantes, dada a sua presença na árvore, podendo ser utilizadas como um método relativamente simples para seleção de atributos;
- Possuem como desvantagens a tendência a decorar os dados (*overfitting*), mas esse problema pode ser evitado limitando adequadamente o número máximo de níveis nas árvores.

No entanto, árvores de decisão têm como desvantagem a limitação para lidar com dados complexos, que exijam fronteiras de decisão complexas, o que costuma limitar o desempenho da classificação ou da regressão, a depender do problema (MYLES et al., 2004; NOWOZIN et al., 2011; ZHOU; CHEN, 2002; RAO et al., 2019; LI; XU; DENG, 2019).

Para compensar as limitações das árvores de decisão foram criadas as Random Forests (RF), ou Florestas Aleatórias. Random Forests são comitês de árvores de decisão (ATHEY; TIBSHIRANI; WAGER, 2019; BIAU, 2012; RIGATTI, 2017; IWENDI et al., 2020; ZHONG et al., 2020). Uma Random Forest é definida a partir do número de árvores de decisão que a integram e o número de níveis que cada árvore pode atingir. As árvores de decisão da Random Forest são similares de certa maneira, mas cada árvore é treinada com um conjunto de treino diferente, com instâncias com um número reduzido de atributos. Dessa forma, cada árvore é treinada considerando pontos de vista diferentes das instâncias. A formação de um comitê dessa maneira, onde cada especialista artificial, ou seja, neste caso, cada árvore, recebe uma versão do conjunto de treinamento com vetores de atributos reduzidos, é chamada de *bagging*. Se o problema for de classificação, o comitê decide por classificar a partir da classe mais votada; se o problema for de regressão, o comitê decide pela média das regressões de cada membro do comitê (ATHEY; TIBSHIRANI; WAGER, 2019; BIAU, 2012; RIGATTI, 2017; IWENDI et al., 2020; ZHONG et al., 2020). A Figura 13 ilustra como funciona uma Random Forest.

Figura 13 – Esquema geral de uma Random Forest como comitê de árvores de decisão.



Fonte: A autora

2.7 ALGORITMOS META-HEURÍSTICOS BIOINSPIRADOS

Diversos algoritmos são bioinspirados. As redes neurais artificiais, por exemplo, são inspiradas na biologia do cérebro humano, e depois foram ficando cada vez mais complexas. Os algoritmos de busca heurística são aqueles algoritmos de busca que pesquisam potenciais para um determinado problema usando estratégias de busca inspiradas em heurísticas, ou seja, em verdades circunstanciais. Essa busca é guiada por uma função chamada de função objetivo, função de adaptação ou função de *fitness*, que modela o problema em si. Essas heurísticas podem ser de diversos tipos. Fala-se em algoritmos meta-heurísticos quando essas heurísticas são baseadas em metáforas. Algoritmos meta-heurísticos bioinspirados são algoritmos cuja busca é inspirada em processos da natureza. Nesta seção são apresentados brevemente os algoritmos meta-heurísticos bioinspirados utilizados neste trabalho, para representar os especialistas virtuais no Comitê de Especialistas Virtuais proposto neste trabalho: algoritmos genéticos, otimização por enxame de partículas, colônia de abelhas artificiais e colônia de formigas artificiais.

2.7.1 Algoritmos Genéticos

Os algoritmos genéticos (*Genetic Algorithms*, GA) são algoritmos meta-heurísticos de busca e otimização inspirados no processos de evolução. Os algoritmos genéticos foram criados por Holland (1973) e aperfeiçoados por Holland (1992), Holland (1984), Booker, Goldberg e Holland (1989). A criação dos algoritmos genéticos iniciou a área da Computação Evolucionária, que compreende um amplo conjunto de algoritmos baseados nos princípios evolutivos, como a Programação Evolucionária, a Programação Genética, a Evolução Diferencial e os algoritmos de estratégias evolutivas.

Como todos os métodos meta-heurísticos, algoritmos genéticos, como algoritmos de busca e otimização, servem para encontrar a melhor solução, ou seja, o ponto de mínimo ou de máximo de uma dada função objetivo, se o problema for de minimização ou de maximização, respectivamente. A metáfora expressa pela meta-heurísticas serve para definir como o espaço de busca vai ser explorado em busca da melhor solução. A função objetivo também é chamada de função de adaptação ou função de *fitness*. Os candidatos à solução são modelados como indivíduos representados por vetores, por sua vez chamados de cromossomos. As ordenadas desses vetores são chamadas de genes. A função objetivo modela a avaliação da adaptação de

cada indivíduo ao meio em que está inserido. Portanto, em um problema de maximização, os indivíduos mais aptos são aqueles que retornam os maiores valores da função objetivo.

Em um algoritmo genético, primeiramente, uma população com indivíduos com cromossomos aleatórios é gerada. Começa então o processo evolutivo, em que uma determinada fração de indivíduos mais aptos é selecionada como indivíduos mais prováveis para reprodução, processo também chamado de cruzamento. Assim, pares de indivíduos pais produzem indivíduos filhos a partir da combinação dos genes dos pais. Os cromossomos dos filhos assim gerados passam então por mutações aleatórias.

Ao final dessa etapa, têm-se a população antiga e a nova população, composta pelos filhos gerados pela combinação dos pais selecionados. Em seguida, vem a seleção dos indivíduos para formar a geração seguinte. No método geracional, a nova população é composta pelos filhos apenas, havendo a substituição de toda a população anterior. No método elitista, apenas os mais aptos sobrevivem, independente de serem filhos ou não. Dependendo do tipo de algoritmo genético, o critério para formar a próxima geração pode mudar. O processo de evolução continua até que seja atingido o máximo de gerações. O fluxograma da Figura 14 apresenta o funcionamento de um algoritmo genético.

De forma simplificada, um algoritmo genético é composto pelas seguintes etapas:

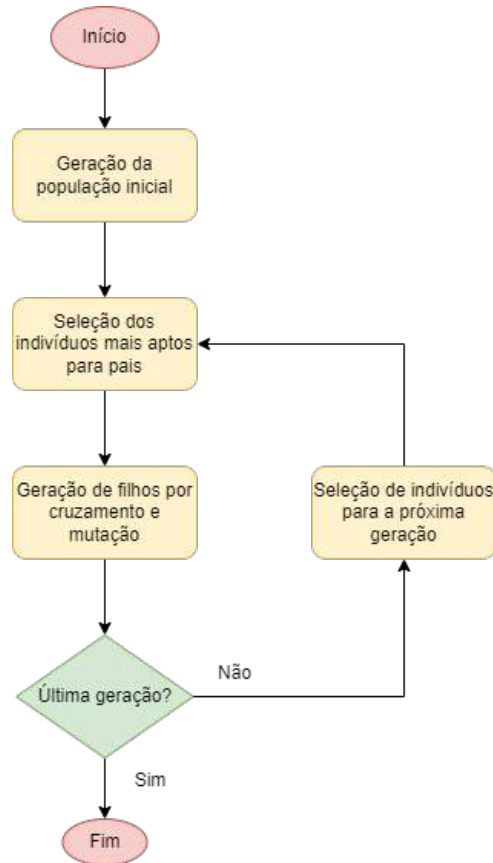
1. Geração de uma população aleatória de indivíduos.
2. Para cada geração:
 - a) Seleção de pais dentre os indivíduos mais aptos;
 - b) Geração de filhos por cruzamento, ou seja, combinação de genes dos pais, e mutação aleatória de determinados genes;
 - c) Seleção de indivíduos para a próxima geração.

Diversas versões de algoritmos genéticos vêm sendo desenvolvidas, para diferentes aplicações (BURATTI; EIJKENS; HAMEIRI, 2020; HILALI-JAGHDAM et al., 2020; LIU et al., 2019; HASSANAT; ALKAFWEEN, 2017; KHALFI; ALAA; GUEDDA, 2018; MEHMOOD et al., 2020).

2.7.2 Otimização por Enxame de Partículas

O algoritmo de Otimização por Enxame de Partículas (*Particle Swarm Optimization*, PSO) é um algoritmo meta-heurístico inspirado no comportamento social de bandos de pássaros,

Figura 14 – Fluxograma de um algoritmo genético. A população inicial é gerada aleatoriamente. Em cada geração, os indivíduos mais aptos, ou seja, os que retornam melhores valores da função objetivo, são selecionados para serem pais. Os pais geram filhos por cruzamento e mutação aleatória dos genes. O processo de repete até ser atingido o máximo de iterações.



Fonte: A autora

cardumes de peixes e coletivos de outros animais sociais. Esse algoritmo foi originalmente criado por Kennedy e Eberhart (1995) e mais tarde aperfeiçoado por Shi e Eberhart (1998). O algoritmo PSO iniciou a área de estudo de Inteligência de Enxames.

Seguindo a inspiração original do PSO, considerando bandos de pássaros, o bando começa sem uma forma definida. Cada pássaro está em uma posição aleatória. Na busca por comida, o pássaro mais habilidoso do bando é considerado o líder global. Cada pássaro, no entanto, modifica sua posição e sua velocidade considerando não somente o líder global do bando, mas também o líder local, ou seja, o pássaro mais habilidoso na sua vizinhança. Dessa forma, o movimento total do bando considera a posição e o movimento do líder global como um todo, mas também considera os movimentos dos líderes locais. A habilidade para encontrar comida é modelada pela função objetivo, que guia o processo de busca e otimização.

Generalizando, no PSO, existe uma população de partículas, ou seja, o enxame. Cada

partícula do enxame possui um vetor de posição e um vetor de velocidade. Os candidatos à solução do problema são representados pelas posições das partículas. Já as velocidades das partículas modelam a capacidade de mudar de uma posição para outra, ou seja, e avaliar outra solução possível para o problema de otimização, pois deve-se buscar a posição que retorne o maior valor da função objetivo. A cada iteração, a qualidade das posições das partículas é avaliada. A partícula que tiver o maior valor da função objetivo é a partícula líder global. Cada partícula individualmente também monitora a qualidade das soluções locais. Assim, para cada partícula é definida uma vizinhança, onde serão verificados qual é a posição o líder local. A posição e a velocidade de cada partícula é ajustada considerando os comportamentos do líder global e do líder local.

O algoritmo simplificado é apresentado no fluxograma da Figura 15.

O algoritmo básico PSO funciona da seguinte maneira, para um problema de maximização com uma função objetivo f_o :

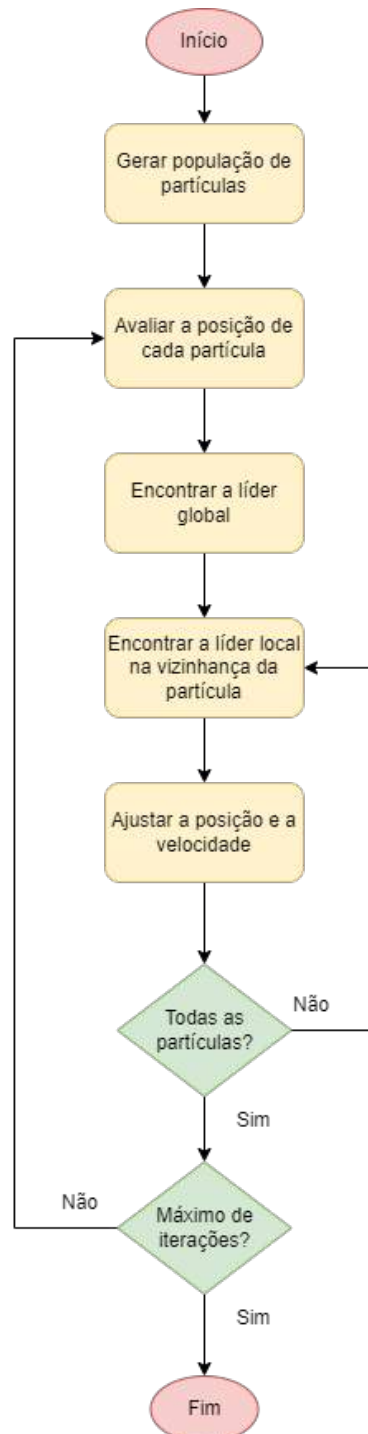
1. Geração da população inicial de partículas: Para cada uma das m partículas, $\mathbf{x}_i \sim U(\mathbf{b}_{\min}, \mathbf{b}_{\max})$, ou seja, o vetor posição $\mathbf{x}_i$ da i -ésima partícula é um vetor de posições aleatórias escolhido entre o vetor de limites inferiores, $\mathbf{b}_{\min}$, e o vetor de limites superiores, $\mathbf{b}_{\max}$, onde $1 \leq i \leq m$.
2. Para cada iteração:
 - a) Definição do líder global: se $f_o(\mathbf{x}_j) = \max\{f_o(\mathbf{x}_i) | i = 1, 2, \dots, m\}$, então $\mathbf{g} \leftarrow \mathbf{x}_j$, onde $\mathbf{g}$ é a posição do máximo global e $1 \leq j \leq m$.
 - b) Para cada i -ésima partícula, onde $i = 1, 2, \dots, m$:
 - i. Definição do líder local: se $f_o(\mathbf{x}_j) = \max\{f_o(\mathbf{x}_k) | \mathbf{x}_k \in N_i\}$, então $\mathbf{p}_i \leftarrow \mathbf{x}_j$, onde $\mathbf{p}_i$ é a posição do máximo local na vizinhança N_i da partícula de posição $\mathbf{x}_i$, onde $\mathbf{x}_j \in N_i$ e $1 \leq j, k \leq m$;
 - ii. Ajuste do vetor de velocidades:

$$\mathbf{v}_i \leftarrow w\mathbf{v}_i + \phi_p r_p (\mathbf{p}_i - \mathbf{x}) + \phi_g r_g (\mathbf{g} - \mathbf{x}),$$
 - iii. Ajuste do vetor de posições:

$$\mathbf{x}_i \leftarrow \mathbf{x}_i + \mathbf{v}_i,$$

onde $0 < w < 1$ é o fator de inércia; $r_g \sim U(0, 1)$ e $r_p \sim U(0, 1)$ são pesos aleatórios entre 0 e 1; ϕ_g é o coeficiente social; e ϕ_p é o coeficiente cognitivo.

Figura 15 – Representação geral de um algoritmo de otimização por enxame de partículas: as partículas iniciam com posição e velocidade aleatórias. A cada iteração, a qualidade da posição de cada partícula é avaliada com a função objetivo. Se o problema for de maximização, a partícula com a posição de maior valor da função objetivo é escolhida como a líder global. Cada partícula também observa, na sua vizinhança, qual é a sua líder local. Então são ajustadas a velocidade e a posição de cada partícula. O processo se repete pelo máximo de iterações.



Fonte: A autora

Diversas versões do PSO, com diferentes definições de vizinhança, formas de atualização de posições e velocidades, e diferentes aplicações vêm sendo desenvolvidas (TIAN; SHI, 2018; ELBES et al., 2019; ELSHEIKH; ELAZIZ, 2019; HAJIHASSANI; ARMAGHANI; KALATEHJARI, 2018; DING et al., 2019; ZEMMAL et al., 2020; ZHENG et al., 2018; GHORPADE et al., 2021).

2.7.3 Colônias de Abelhas Artificiais

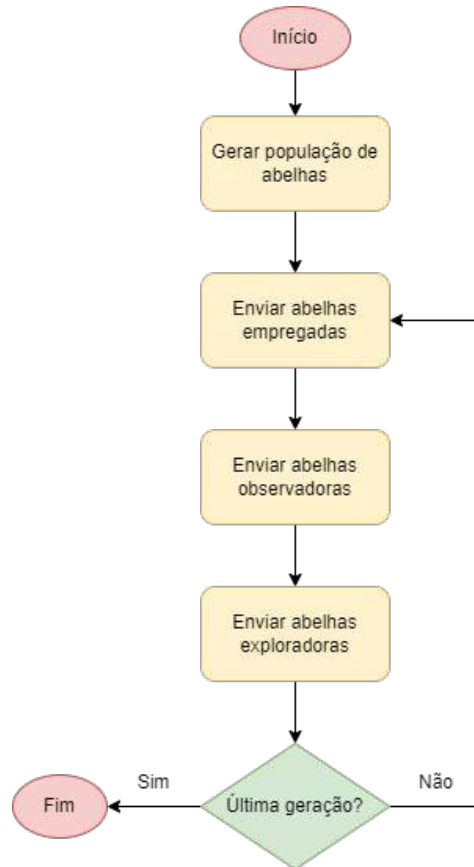
O algoritmo de otimização por Colônia de Abelhas Artificiais, ou Colônia Artificial de Abelhas (*Artificial Bee Colony*, ABC) foi criado por Karaboga (2005) e aperfeiçoado por Karaboga e Akay (2009). Atualmente existem diversas versões do ABC (GAO; LIU, 2012; AKAY; KARABOGA, 2012; ZHU; KWONG, 2010; BANSAL; SHARMA; JADON, 2013; BANHARNSAKUN; ACHALAKUL; SIRINAOVAKUL, 2011; WANG et al., 2020), bem como diversas aplicações de ABC (GAO et al., 2018; GHAMBARI; RAHATI, 2018; XU et al., 2020; LU et al., 2019; PAN; ZHANG; PAN, 2020).

No ABC, existem três tipos de abelhas: empregadas, exploradoras e observadoras. No algoritmo, existe uma única abelha empregada para cada possível fonte de alimento. A posição de uma fonte de alimento representa uma possível solução para o problema de otimização. A quantidade de néctar de uma fonte de alimento corresponde à qualidade ou aptidão da solução associada. Em um problema de otimização de maximização, a quantidade de néctar é dada pelo resultado da função objetivo aplicada à posição de alimento para cada abelha empregada.

Inicialmente, é gerada uma população de abelhas com posições aleatórias de alimento. Então se inicia o processo de busca das abelhas empregadas, exploradoras e observadoras. Uma abelha empregada memoriza uma posição da fonte de alimento. Mas se a abelha empregada descobrir uma fonte de alimento com mais néctar, ela esquece a fonte anterior e armazena a atual. Do contrário, a posição armazenada é mantida. Depois que todas as abelhas empregadas terminam a busca, elas apresentam as informações de posição das fontes para as outras abelhas na área de dança. Cada abelha observadora avalia as informações sobre o néctar obtidas de todas as abelhas empregadas e escolhe uma fonte de alimento e sua vizinhança para visitar e avaliar a quantidade de néctar. As abelhas observadoras avaliam as vizinhanças para determinar se uma fonte deve ser abandonada ou não. Caso seja abandonada, uma nova fonte é produzida aleatoriamente. As abelhas exploradoras são então enviadas. Elas determinam se as fontes produzidas aleatoriamente podem substituir as fontes abandonadas.

O algoritmo simplificado é apresentado no fluxograma da Figura 16.

Figura 16 – Representação geral de um algoritmo de colônia artificial de abelhas: as abelhas empregadas buscam novas soluções na vizinhança das fontes de alimento; as abelhas observadoras avaliam as fontes de alimentam para decidir quais devem ser mantidas e quais devem ser abandonadas; as abelhas exploradoras encontram aleatoriamente novas fontes de alimento.



Fonte: A autora

O algoritmo básico ABC pode ser descrito da seguinte maneira:

1. Geração de uma população aleatória de abelhas empregadas, exploradoras e observadoras. O número de fontes de comida m e o número de abelhas empregadas, exploradoras e observadoras é igual.
2. Enviar abelhas empregadas, para encontrar novas e melhores fontes de comida, ou seja, melhores candidatos à solução do problema de otimização. Cada abelha empregada $\mathbf{x}_i$ gera uma nova candidata à solução, $\mathbf{v}_i$, da forma que segue:

$$\mathbf{v}_i = \mathbf{x}_i + \Phi_i \bullet (\mathbf{x}_i - \mathbf{x}_j), \quad (2.13)$$

onde $j \neq i$ é uma abelha diferente selecionada aleatoriamente, e $-1 < \Phi_i < 1$ é um número aleatório, para $1 \leq i \leq m$. Se a candidata à solução $\mathbf{v}_i$ for melhor do que $\mathbf{x}_i$, a

posição $\mathbf{x}_i$ é atualizada, ou seja:

$$f_o(\mathbf{v}_i) > f_o(\mathbf{x}_i) \Rightarrow \mathbf{x}_i \leftarrow \mathbf{v}_i, \quad (2.14)$$

onde f_o é a função objetivo que guia o processo de otimização de maximização.

3. Enviar as abelhas observadoras. Elas são selecionadas a partir do melhor desempenho, calculando suas probabilidades de sucesso da forma que segue:

$$p_i = \frac{f_o(\mathbf{x}_i)}{\sum_{i=1}^m f_o(\mathbf{x}_i)}. \quad (2.15)$$

Em seguida são selecionadas as mais prováveis de sucesso. As melhores soluções são memorizadas.

4. Enviar as abelhas exploradoras, responsáveis por encontrar novas fontes de comida. As posições das abelhas exploradoras são geradas aleatoriamente, para proporcionar novas posições a serem investigadas pelas abelhas empregadas no próximo ciclo do processo, da forma que segue:

$$x_{i,j} = x_{\min,j} + r \bullet (x_{\max,j} - x_{\min,j}), \quad (2.16)$$

onde $x_{i,j}$ é a j -ésima coordenada da posição $\mathbf{x}_i$ da i -ésima abelha; $x_{\min,j}$ e $x_{\max,j}$ são os valores mínimo e máximo assumidos pela j -ésima coordenada da posição $\mathbf{x}_i$; e $0 < r < 1$ é um número aleatório.

5. Repete-se o ciclo a partir da posição (2) até que seja atingido o máximo de gerações.

2.7.4 Colônias de Formigas Artificiais

Os algoritmos de Otimização por Colônias de Formigas (*Ant Colony Optimization*, ACO) são algoritmos meta-heurísticos inspirados no comportamento natural das formigas. Eles foram criados por Marco Dorigo (DORIGO; GAMBARDILLA, 1997a; DORIGO; GAMBARDILLA, 1997b; COLONI et al., 1992; DORIGO; BIRATTARI, 2010; GAMBARDILLA; DORIGO, 1996; DORIGO et al., 1996; DORIGO; STÜTZLE, 2019) e novas versões vêm sendo desenvolvidas para aplicações gerais e específicas (DORIGO; STÜTZLE, 2019; SEBAYANG et al., 2017; SEMET et al., 2019; BELTRAMO et al., 2016; YU; CHEN; LI, 2015).

Um ACO funciona da seguinte maneira, de forma geral: as formigas inicialmente seguem trajetórias totalmente aleatórias em busca de comida, ou seja, do objetivo da otimização; cada

vez que uma formiga segue um caminho, deposita uma substância chamada feromônio, que é usado para marcar o caminho; se o caminho levar a comida, a formiga tende a repeti-lo, reforçando esse caminho com feromônio, que por sua vez atrai mais formigas, que reforçam o caminho; o processo se repete até um máximo de gerações ou até que não exista mais feromônio. O algoritmo é representado de forma simplificada pelos seguintes passos:

1. Inicialmente é gerado um conjunto de formigas artificiais. A cada formiga está associado um caminho de sua posição até uma outra posição.
2. O caminho de cada formiga é avaliado com relação à proximidade do objetivo.
3. Calculam-se as probabilidades de cada formiga manter o seu caminho.
4. A taxa de feromônio é atualizada, de forma que caminhos sem muito sucesso tem o feromônio rapidamente apagado, enquanto caminhos com mais sucesso são seguidos por mais formigas e têm seu feromônio reforçado.
5. Repete-se o processo a partir do passo (2) até que se atinja o máximo de gerações ou que todo o feromônio seja apagado.

O fluxograma da Figura 17 apresenta como funciona um ACO.

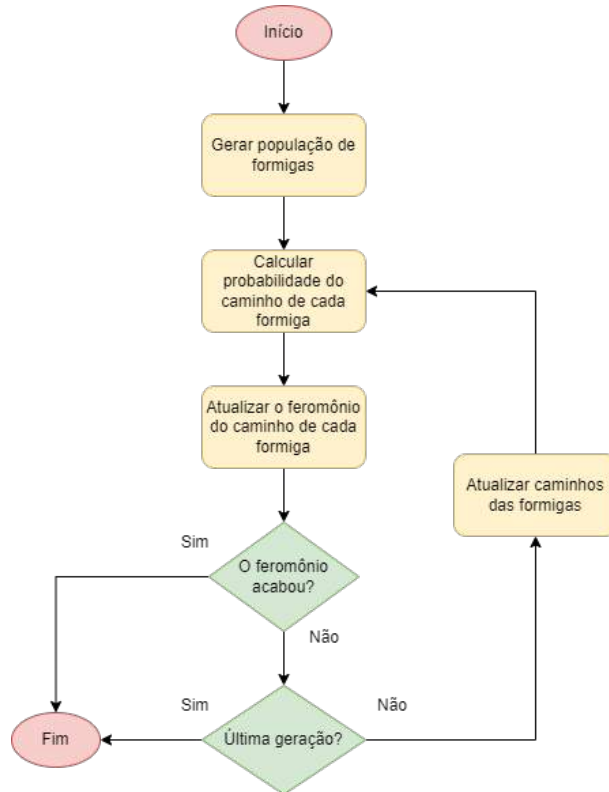
A probabilidade individual de cada formiga não mudar de caminho é dada pela Equação 2.17:

$$p_k = \frac{\tau_k^\alpha \eta_k^\beta}{\sum_{i=1}^m \tau_i^\alpha \eta_i^\beta}, \quad (2.17)$$

onde p_k é a probabilidade da k -ésima formiga de permanecer no caminho, τ_k é a quantidade de feromônio depositada no caminho pela k -ésima formiga, e η_k é a atratividade do caminho traçado pela k -ésima formiga, para um conjunto de m formigas, ou seja, $1 \leq k \leq m$. Em problemas de otimização de maximização, dada uma função objetivo $f_o(\mathbf{x})$, onde o vetor $\mathbf{x}$ modela um caminho, a atratividade do caminho da k -ésima formiga pode ser dada por $\eta_k = f_o(\mathbf{x}_k)$, onde $\mathbf{x}_k$ é o caminho da k -ésima formiga. Os parâmetros $\alpha \geq 0$ e $\beta \geq 1$ são parâmetros de controle da influência da quantidade de feromônio e da atratividade do caminho, respectivamente. A quantidade de feromônio para a k -ésima formiga é atualizada ao final de cada geração pela expressão da Equação 2.18:

$$\tau_k \leftarrow (1 - \rho)\tau_k + \sum_{i=1}^m \Delta\tau_k, \quad (2.18)$$

Figura 17 – Fluxograma geral de um algoritmo de otimização por colônia de formigas.



Fonte: A autora

onde $0 < \rho < 1$ é o coeficiente de evaporação de feromônio e

$$\Delta\tau_k = f_o(\mathbf{x}_k), \quad (2.19)$$

para $1 \leq k \leq m$.

Um caminho de uma formiga pode ser representado como um vetor. A atualização do caminho pode ser feita seguindo um procedimento similar àquele de mutação em um algoritmo genético.

Em problemas de seleção de atributos, os caminhos de cada formiga podem ser modelados como vetores binários, onde 1 representa a presença do respectivo atributo, enquanto 0 modela a exclusão do atributo. A avaliação da qualidade do caminho costuma ser feita usando uma função objetivo baseada em uma máquina de aprendizado simples e rápida, como uma árvore de decisão, que é rapidamente treinada usando versões dos conjuntos de treinamento e teste com os atributos reduzidos e retornando o coeficiente de correlação ou outra medida de erro de regressão, por exemplo.

2.8 MÉTRICAS DE REGRESSÃO

As principais métricas que adotamos para avaliar os modelos são as seguintes: o coeficiente de correlação e o Erro Médio Quadrático Relativo (RMSE percentual). O coeficiente de correlação é uma medida estatística entre os valores esperados e previstos. Este valor varia de -1 a 1. Quando se aproxima de 1, indica uma forte correlação positiva. Por outro lado, quando o coeficiente de correlação é próximo a -1, indica que as variáveis possuem uma correlação negativa forte. Quando o coeficiente de correlação é próximo a zero, indica que não há correlação entre as variáveis (WITTEN; FRANK, 2005). O valor do coeficiente de correlação serve como avaliador global para o modelo. Portanto, é possível obter um alto coeficiente de correlação e ao mesmo tempo obter altos valores para erros locais. Por esse motivo, não pode ser a única métrica para avaliar o desempenho do modelo. Para evitar uma avaliação superficial dos regressores, optamos, portanto, pelo RMSE (%) como métrica de avaliação. A Equação 2.20 mostra a expressão do cálculo do erro quadrático relativo, onde p_i é o valor previsto e a_i é o valor real, para $i = 1, 2, \dots, n$.

$$\text{RMSE}(\%) = \sqrt{\frac{\sum_{i=1}^n (p_i - a_i)^2}{\sum_{i=1}^n a_i^2}} \times 100\%. \quad (2.20)$$

Além do RMSE (%), também calculamos o erro quadrático médio (RMSE), o erro médio absoluto (MAE), o erro médio percentual absoluto (MAPE) e o erro médio percentual (MPE) (Equações 2.21 - 2.24):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2}, \quad (2.21)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |e_i|, \quad (2.22)$$

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{e_i}{a_i} \right|, \quad (2.23)$$

$$\text{MPE} = \frac{100\%}{n} \sum_{i=1}^n \left(\frac{p_i - a_i}{p_i} \right) = \frac{100\%}{n} \sum_{i=1}^n \left(\frac{-e_i}{p_i} \right), \quad (2.24)$$

onde, p_i é o valor previsto, a_i é o valor real e $e_i = a_i - p_i$ é a diferença entre o valor real e o valor previsto.

O coeficiente de correlação de Pearson R é definido da seguinte forma:

$$R = \frac{\sum_{i=1}^n (p_i - \bar{p})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^n (p_i - \bar{p})^2 \cdot \sum_{i=1}^n (a_i - \bar{a})^2}}, \quad (2.25)$$

onde $\bar{p}$ e $\bar{a}$ são os valores médios amostrais para os conjuntos de valores previstos e reais, respectivamente.

Da mesma forma, o coeficiente de correlação de classificação de Spearman ρ é definido da seguinte forma:

$$\rho = \frac{\sum_{i=1}^n (R(p_i) - \bar{R}(p))(R(a_i) - \bar{R}(a))}{\sqrt{\sum_{i=1}^n (R(p_i) - \bar{R}(p))^2 \cdot \sum_{i=1}^n (R(a_i) - \bar{R}(a))^2}}, \quad (2.26)$$

onde $R(p_i)$ e $R(a_i)$ são os ranks de p_i e a_i , enquanto $\bar{R}(p)$ e $\bar{R}(a)$ são as médias amostrais das classificações de p_i e a_i , respectivamente.

A correlação de classificação de Kendall τ é dada da seguinte forma:

$$\tau = \frac{2}{n(n-1)} \sum_{j=1}^n \sum_{i=1}^{j-1} \text{sign}(p_i - p_j) \cdot \text{sign}(a_i - a_j), \quad (2.27)$$

onde n é o número de observações e $1 \leq i, j \leq n$. A função do sinal, sign , é definida da seguinte forma:

$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases},$$

para $x \in \mathbb{R}$.

2.9 CONSIDERAÇÕES

Neste capítulo foi apresentada a fundamentação teórica deste trabalho. Foram apresentados os conceitos básicos de Geografia da Saúde e Epidemiologia Digital, discutindo como o sensoriamento remoto e os sistemas de informação geográfica podem ser utilizados no controle da difusão de doenças. Foram também apresentadas as definições das arboviroses dengue, chikungunya e zika, métodos de predição de arboviroses e a metodologia de vigilância ambiental adotada no Sistema Único de Saúde no controle dessas doenças. Foram ainda apresentados brevemente conceitos fundamentais de aprendizado de máquina e breves descrições dos métodos utilizados neste trabalho como modelos preditivos. Por fim, foi apresentado brevemente o conceito de algoritmo bioinspirado de busca heurística e os métodos bioinspirados utilizados no comitê de especialistas para seleção de fatores mais relevantes adotado neste trabalho. No Capítulo 3 é apresentado o estado da arte de modelos preditivos para predição de arboviroses, o estudo de caso apresentado neste trabalho.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados os trabalhos que foram considerados mais relacionados a esta pesquisa de doutorado, dentre as publicações mais recentes (2014-2021). Foram estudados trabalhos que lidam com predição de arboviroses por meio de modelos estatísticos e de aprendizado de máquina, focando aqueles que fazem uso de dados geoespaciais e combinam várias bases de dados. Ao final, suas principais características são comparadas com esta pesquisa, tanto em relação à metodologia utilizada quanto aos resultados.

3.1 ESTADO DA ARTE

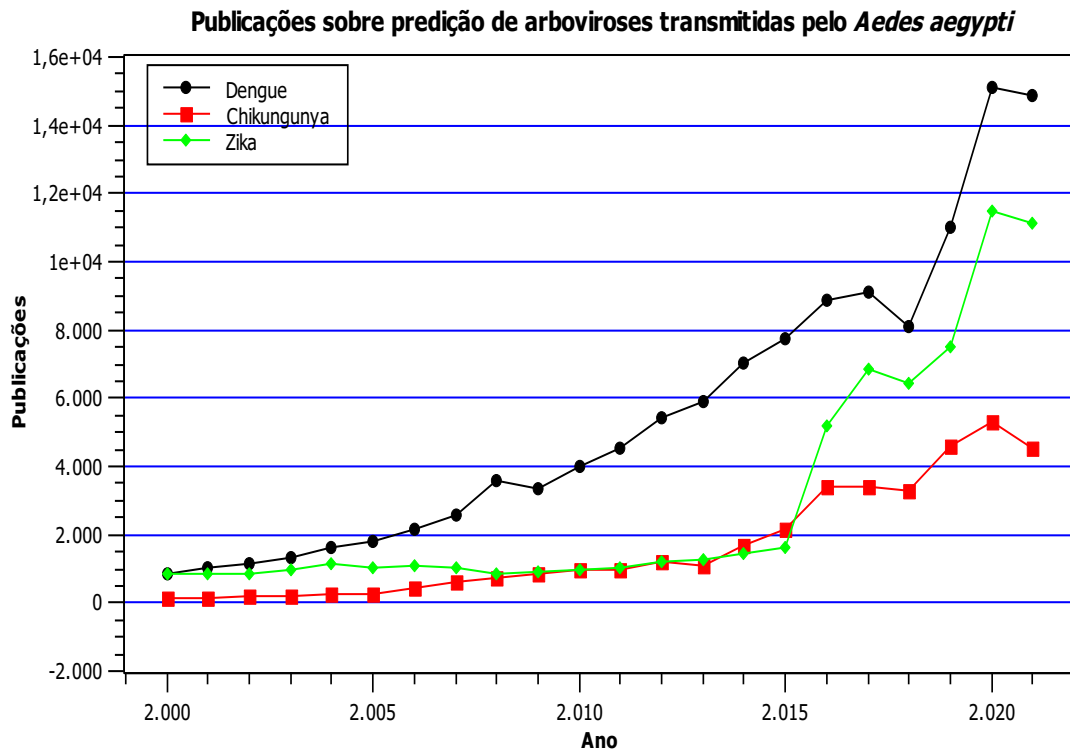
O número de publicações sobre predição de dengue por ano, de 2000 a 2021, foi levantado por meio de buscas no Google Acadêmico. A Figura 18 mostra um crescente interesse na construção e no uso de sistemas de predição de casos de dengue, com leve queda em 2008, desaceleração a partir de 2016 e queda em 2018, para nova subida em 2019, 2020 e estabilidade em 2021. Acreditamos que o aumento das publicações nos anos de 2019 a 2021 estejam relacionados ao maior desenvolvimento das pesquisas relacionadas à epidemiologia digital por causa da pandemia de Covid-19. A busca foi feita usando as seguintes chaves: “dengue AND (forecasting OR prediction OR surveillance)”, ou seja, dengue, predição (*prediction* e *forecasting*) e vigilância (*surveillance*). De forma similar foi feito para chikungunya e zika, sendo realizadas buscas com as seguintes chaves: “chikungunya AND (forecasting OR prediction OR surveillance)” e “zika AND (forecasting OR prediction OR surveillance)”, respectivamente. O Google Acadêmico não diferencia na busca entre os tipos de publicação. Logo, os resultados incluem artigos completos em periódicos, trabalhos completos em anais de eventos, resumos, capítulos de livro e livros.

De forma semelhante, o gráfico da Figura 18 exibe a evolução do número de publicações sobre predição das doenças febre chikungunya e zika para o período de 2000 a 2018, respectivamente. É notável o aumento de publicações sobre chikungunya a partir de 2007, devido ao surto que aconteceu na Europa nesse ano. A partir daí o interesse de pesquisa na predição da febre chikungunya é crescente, à medida em que a doença se espalha para a Ásia. Com relação à zika, percebe-se que o número de publicações cresce de maneira lenta até 2015.

Em 2016 ocorre um salto, devido ao surto de zika vírus em 2015 e suas sérias consequências, como a microcefalia em recém-nascidos, começam a afligir a região Nordeste do Brasil. Há um crescimento do número anual de trabalhos para os anos de 2019 e 2020, de forma semelhante ao que acontece com as pesquisas sobre predição de dengue, com estabilização em 2021, o que também atribuímos ao aumento do interesse em pesquisas em epidemiologia digital por conta da pandemia de Covid-19.

Para fins de comparação, o gráfico da Figura 18 mostra os resultados para as buscas por predição de dengue, chikungunya e zika. A partir de 2015, as publicações sobre predição dessas três arboviroses acabam por coincidir, pelo fato de terem o mesmo vetor de transmissão: o mosquito *Aedes aegypti*. Assim, a prevenção dessas doenças acaba passando pelo controle do vetor, o que faz com que os preditores devam ser voltados à predição de arboviroses transmitidas pelo *Aedes aegypti*.

Figura 18 – Comparativo do número de publicações sobre o tema predição da dengue, da febre chikungunya e do vírus zika por ano, de 2000 a 2021, usando o mecanismo de busca Google Acadêmico. Os resultados incluem artigos completos em periódicos, trabalhos completos em anais de eventos, resumos, capítulos de livro e livros.



Fonte: A Autora

A queda no número de publicações a partir de 2017 sobre predição das três arboviroses pode ter várias razões. Uma delas pode ser o efeito dos investimentos dos poderes públicos no

controle do vetor logo após os surtos de febre chikungunya, causando a síndrome de Guillain-Barré, uma doença do sistema nervoso responsável por perda temporária movimentos dos membros superiores e inferiores e até por paralisia temporária ou permanente; e de zika vírus, causando microcefalia em recém-nascidos em 2015. Uma outra hipótese pode ser a diminuição do financiamento de pesquisas nessas áreas, com o controle desses surtos. Por fim, uma outra hipótese pode estar relacionada à natureza endêmica das arboviroses transmitidas pelo *Aedes aegypti* em muitas regiões de baixa renda e de difícil controle do vetor, o que pode estar levando tanto o poder público quanto a população a se habituar com essas doenças. Uma vez que as publicações de predição de dengue, chikungunya e zika passam a praticamente coincidir a partir de 2015, as buscas se limitaram a sistemas de predição de dengue no período de 2014 a 2021.

De toda forma, acredita-se neste trabalho que essa variação pode indicar uma janela de oportunidade para propor uma inovação na predição dessas doenças. Além disso, mesmo que predominem as piores hipóteses, o patamar do número de publicações em predição de arboviroses pode ser considerado alto, mas pelo que foi observado de uma amostra dos trabalhos de 2018, a área ainda está longe de ficar estagnada. Nesse sentido, Sistemas de Informações Geográficas (GIS, *Geographic Information Systems* e técnicas de Aprendizado de Máquina podem contribuir bastante para a construção de uma metodologia de predição inovadora.

De acordo com Siriyasatien et al. (2018a), novos dados sobre dengue estão sendo constantemente gerados e devem ser incorporados aos dados existentes para garantir que modelos preditivos tenham um conjunto completo de novos dados a partir dos quais aprendam, tornando os modelos preditivos atuais e relevantes. Isso vale também para a zika e a chikungunya. No entanto, garantir que novas observações sejam incorporadas ao corpo de dados existente é essencialmente uma tarefa manual, que é demorada e inconveniente, e pode não ser abrangente, resultando em modelos de previsão ineficazes. Essa é uma questão fundamental, para desenvolver mecanismos de atualização automática dos dados em uma base contínua. Isso otimizaria a eficácia e a eficiência do modelo de previsão. Esse problema é agravado pela necessidade de atualizar os modelos de previsão de forma frequente, o que importaria despesas gerais muito altas (importação manual dos dados das bases). A atualização pouco frequente pode diminuir a eficácia e a eficiência do modelo de previsão e faz com que o planejamento e o gerenciamento das políticas de controle do vetor sejam ineficazes. Como a coleta manual de dados é uma tarefa trabalhosa, é desejável coletar automaticamente informações usando aplicativos móveis e Internet das Coisas (*Internet of Things*, IoT).

A IoT é a rede de dispositivos móveis, dispositivos portáteis e sensores de todos os tipos (SIRIYASATIEN et al., 2018a). A existência de milhões de dispositivos conectados, que reúnem e transmitem dados, torna a IoT uma importante nova tecnologia que deve ser de grande interesse para os pesquisadores epidemiológicos. Uma aplicação significativa dessa tecnologia é na área de medicina, usando o que hoje é chamado de Internet das Coisas Médicas (*Internet of Medical Things*, IoMT). O IoMT é uma fusão de dispositivos médicos e aplicativos que podem se conectar a sistemas de tecnologia da informação em saúde usando tecnologias de rede (SIRIYASATIEN et al., 2018a). É possível, por exemplo, construir biossensores e dispositivos conectados para identificação de dengue e monitoramento de pacientes para medir vários parâmetros do paciente para identificar e diagnosticar infecção por dengue, níveis de imunidade do paciente, pressão arterial, frequência cardíaca e respiração (MANOGARAN et al., 2017; MANOGARAN et al., 2018; WANG; RANJAN, 2015).

Para Siriyasatien et al. (2018a), uma limitação dos modelos de previsão de epidemias é que a interpretação dos resultados da previsão pode não ser fácil de entender por usuários que não têm experiência, conhecimento ou experiência profissional. Muitas vezes, a informação médica é difícil de entender mesmo por profissionais médicos. As técnicas de Aprendizado de Máquina são difíceis de entender pela maioria das pessoas sem o treinamento necessário. Quanto mais complexo o modelo de previsão, mais essencial é ter os resultados de previsão exibidos de uma maneira facilmente interpretada e compreendida pelos usuários, incluindo uma necessária customização (ABREU et al., 2016). Em qualquer caso, a apresentação de dados deve ser facilmente entendida como de máxima utilidade para a tomada de decisões.

De acordo com Bhunia e Shit (2019), o uso de sistemas de informações geográficas é muito útil como apoio ao controle de epidemias e endemias, organizando as variáveis de interesse na forma de planos de informação, fornecendo dados também qualitativos para gestores de saúde pública.

Gomes et al. (2012) construíram um sistema de informações geográficas para estudar como a migração de áreas rurais para zonas urbanas e a ocupação desordenada do território afetaram a distribuição geográfica da esquistossomose, uma doença endêmica, na localidade de Porto de Galinhas, Pernambuco.

Leal-Neto et al. (2014) propuseram o Schisto Track, uma ferramenta para automatizar as atividades de cadastramento de criadouros e residências, organizar e padronizar os dados e facilitar a recuperação das informações. Possibilitava interface com outros sistemas e possuía arquitetura totalmente modular. Foi integrado ao Google Maps, e a API ficou disponível para

todos os sites que pudessem ser acessados, gratuitamente, por usuários e desenvolvedores, facilitando a visualização dos dados distribuídos espacialmente. Outros trabalhos seguiram o mesmo princípio: utilizar informações geográficas para fins epidemiológicos utilizando dispositivos móveis (LEAL-NETO et al., 2015; LEAL et al., 2017a; LEAL et al., 2017b).

Souza et al. (2021) destacam que, para prever epidemias, as abordagens clássicas geralmente precisam de estimativas de parâmetros, como a taxa de contágio ou o número básico de reprodução. Assim, Souza et al. (2021) propuseram uma abordagem geométrica baseada em dados, sem parâmetros, para acessar o surgimento de um estado pandêmico, estudando as curvaturas da rede Forman-Ricci e Ollivier-Ricci. A curvatura discreta de Ollivier-Ricci tem sido usada com sucesso para prever o risco em redes financeiras. Souza et al. (2021) sugerem que esses resultados podem fornecer resultados análogos para séries temporais de epidemia de Covid-19. Primeiramente, Souza et al. (2021) construíram ambas as curvaturas em um modelo sintético de séries temporais epidêmicas com atrasos, permitindo criar redes epidêmicas. Esses resultados foram comparados com métricas de rede clássicas. Os autores puderam verificar que as curvaturas de Ollivier-Ricci e Forman-Ricci podem ser uma estimativa sem parâmetros para identificar um estado pandêmico na epidemia simulada. Com base nisso, foram calculadas as curvaturas Forman-Ricci e Ollivier-Ricci para redes epidêmicas reais construídas a partir de séries temporais de casos de Covid-19 disponíveis na Organização Mundial da Saúde (OMS). Essa abordagem permitiu detectar os primeiros sinais de alerta do surgimento da pandemia. De acordo com os autores, a vantagem do método proposto está em fornecer um marcador de dados geométricos precoce para o estado de pandemia, independentemente da estimativa de parâmetros e da modelagem estocástica, permitindo usar geometria discreta para estudar redes epidêmicas (SOUZA et al., 2021; SOUZA et al., 2020).

Kesorn et al. (2015) propuseram um sistema para predição da taxa de mortalidade da dengue hemorrágica baseado em máquinas de vetor de suporte. Na proposta de Kesorn et al. (2015), ao invés de modelar o problema usando regressão e, assim, buscar prever diretamente a taxa de mortalidade, o problema foi transformado em um problema de classificação. Assim, o problema consiste em classificar se a taxa de mortalidade é alta ou baixa. Como métricas de avaliação, são utilizadas a sensibilidade, a especificidade e a acurácia. Como áreas de estudo, foi considerada a região central da Tailândia, composta por três províncias: Nakhon Pathom, Ratchaburi e Samut Sakhon. Os dados foram coletados de 2007 a 2013. Essa região foi escolhida pelas seguintes razões: 1) alta taxa de mortalidade por dengue hemorrágica; 2) alta densidade de mosquitos; 3) fatores climáticos similares. Para predição foram utilizadas quatro

bases de dados: distribuição de densidade populacional, fornecida pelo Ministério do Interior da Tailândia; taxa de infecção por *Aedes aegypti* (larvas, população de mosquitos machos e de fêmeas, em percentual), fornecida pela Universidade de Chulalongkorn; fatores climáticos (temperatura, umidade, velocidade dos ventos e pluviometria), fornecida pelo Departamento de Meteorologia Tailandês; e distribuição espacio-temporal dos casos de dengue, fornecida pelo Birô de Epidemiologia. A partir da distribuição dos casos de dengue, foi calculada a taxa de mortalidade e definido o limiar a partir do qual uma taxa deve ser considerada alta. Cada uma dessas variáveis de interesse foi organizada na forma de série temporal, sem considerar informação de localização geográfica. A base de dados resultante possui duas classes e 648 instâncias com 9 atributos. Os dados de saúde são de notificações mensais.

Kesorn et al. (2015) investigaram os seguintes classificadores e configurações: árvore de decisão (*Decision Tree*, DT), *k* vizinhos mais próximos (*k-Nearest Neighbours*, kNN), rede neural perceptron multicamadas (MLP) e máquinas de vetor de suporte (SVMs). Não há informações a cerca de que rede neural foi utilizada no estudo. Como foi utilizada a biblioteca Weka e a rede neural supervisionada padrão dessa biblioteca é a MLP com configuração padrão de uma única camada escondida com número de neurônios igual à soma do número de entradas e do número de saídas, acredita-se que tenham sido utilizados 11 neurônios na única camada escondida. Já com relação aos SVMs, foram investigadas as seguintes configurações de *kernel*: linear, polinomial (não foi especificado o grau do polinômio) e de função de base radial (*Radial Basis Function*, RBF). Todos os testes foram feitos com validação cruzada com 10 partes. Os melhores resultados foram obtidos com SVM-RBF: sensibilidade de 0,882, especificidade de 0,789, e acurácia de 90,740%.

Apesar dos bons resultados de predição, a transformação de um problema onde caberia melhor uma regressão em um problema de classificação por Kesorn et al. (2015) carrega imprecisões. Por exemplo: por que classificar apenas entre taxa de mortalidade alta ou não alta? O valor específico da taxa não é importante? Conhecer a taxa aproximada não poderia ser mais útil aos gestores de saúde pública e não teria um interesse epidemiológico maior? Além disso, tendo em vista que as informações de posicionamento geográfico não são consideradas, o método proposto não retorna uma distribuição espacial aproximada da taxa de mortalidade e nem do critério (alta ou não), o que dificulta o uso do resultado no planejamento local, no caso, no nível das províncias e de suas cidades. Kesorn et al. (2015) também não analisam o peso de cada fator na predição.

Apesar de Ch et al. (2014) apresentarem uma pesquisa sobre predição de malária e não

de dengue, optou-se por apresentar aqui um resumo crítico desse trabalho pela sua relevância e potencial aplicação também na predição de arboviroses transmitidas pelo *Aedes aegypti*, já que a prevenção da malária também se dá pelo controle do vetor, e o vetor da malária também é a distribuição de mosquitos, no caso, os mosquitos da espécie *Anopheles*. Ch et al. (2014) propuseram um método para prever a transmissão da malária baseado na otimização de um regressor SVM utilizando o algoritmo de vagalumes artificiais (*Firefly Algorithm*, FFA). Como área de estudo, foram investigadas as regiões de Jodhpur e Bikaner, no Estado do Rajastão, Índia. Foram utilizadas bases de dados obtidas de notificações mensais, de janeiro de 1998 a dezembro de 2002, correspondentes a pluviometria, temperatura, umidade e casos de malária. O FFA foi utilizado para encontrar os melhores valores do parâmetro de regularização C e dos parâmetros γ e ϵ de um SVM-RBF para o maior valor possível de erro médio quadrático normalizado (*Normalized Mean Square Error*, NMSE). Os resultados foram comparados com métodos clássicos, como Modelos Auto-Regressivos (*Auto-Regressive Model*, AR), Média Móvel Auto-Regressiva (*Auto-Regressive Moving Average*, ARMA) e Média Móvel Integrada Auto-Regressiva (*Auto-Regressive Integrated Moving Average*, ARIMA). Os resultados também foram comparados com MLP (camada única, 19 e 17 neurônios na camada escondida para regressores de Bikaner e Jodhpur, respectivamente) e SVM com *kernel* linear. Os experimentos foram feitos usando divisão percentual, com 80% do conjunto de dados para treino e 20% para teste. O modelo proposto, SVM-FFA, se mostrou muito superior aos outros, por ter erro médio muito menor e ser mais estável, ou seja, variar menos, o que foi verdade para ambas as regiões. Como as informações de posicionamento geográfico não foram consideradas, o método proposto não retorna uma distribuição espacial aproximada dos casos de malária, o que dificulta o uso do resultado no planejamento local, no caso, no nível das províncias e de suas cidades. Como Kesorn et al. (2015), Ch et al. (2014) também não analisam o peso de cada fator na predição.

Shaukat et al. (2015) propõem um método para predição de surtos de dengue por meio da mineração de dados de informações clínicas fornecidas por unidades de saúde. As informações utilizadas são os sintomas de dengue mais conhecidos, ou seja, febre, fadiga, mialgia (dor muscular), gripe muito forte, hemorragia, outros sintomas, e o diagnóstico de presença ou não de dengue. A base possui assim vetores de 7 atributos e uma única classe. As informações são binárias e o problema é modelado como classificação. A base de dados utilizada foi fornecida pelo Sistema de Saúde Pública do Paquistão e corresponde a dados do município de Jhelum, e corresponde a registros de 95 pacientes. Foram avaliados o classificador ingênuo de Bayes,

árvores de decisão (J48, Random Tree e Rep Tree) e SVM linear. Como métricas de avaliação, foram consideradas a acurácia, taxa de falsos positivos, taxa de falsos negativos, precisão, *recall*, medida F e área da curva ROC. Neste trabalho não são consideradas as informações de localização geográfica dos pacientes nem os fatores climáticos e ambientais. É muito mais um sistema de apoio ao diagnóstico do que um sistema para predição de surtos. Os melhores resultados foram obtidos com o classificador ingênuo de Bayes, o que mostra que a modelagem pode ser simples demais, dado que esse classificador considera que os atributos são estatisticamente independentes, o que não é verdade para a maior parte dos problemas reais.

Laureano-Rosario et al. (2018a) aplicaram redes neurais artificiais para prever ocorrências de surto de dengue em San Juan, Porto Rico, e em vários municípios do Estado de Yucatán, México. Os modelos foram treinados com 19 anos de dados de dengue para Porto Rico e seis anos para o México. Os dados climáticos e ambientais e as informações demográficas incluídas nos modelos preditivos foram: temperatura da superfície do mar (*Sea Surface Temperature*, SST), precipitação, temperatura do ar (mínima, máxima e média), umidade, casos anteriores de dengue e tamanho da população. Foram construídos dois modelos para predição de ocorrência de dengue: um voltado para a população sob risco, ou seja, jovens com menos de 24 anos; e outro modelo, voltado para a população mais vulnerável, ou seja, crianças com menos de seis anos e idosos com mais de 65 anos. Os dados foram coletados semanalmente. Para Porto Rico, foram coletados 986 casos de população sob risco e 986 casos de população mais vulnerável a cada semana. Para o México, foram coletados 310 casos de população sob risco e 310 casos de população mais vulnerável a cada semana. O problema foi modelado como problema de classificação, sendo assumidos limiares:

- Porto Rico: população sob risco:
 - Pré-Epidêmico: 2 casos por semana;
 - Epidêmico: 6 casos por semana;
 - Pós-Epidêmico: 5 casos por semana.
- Porto Rico: população mais vulnerável:
 - Pré-Epidêmico: 1 caso por semana;
 - Epidêmico: 2 casos por semana;
 - Pós-Epidêmico: 1 caso por semana.

- México: população sob risco:
 - Pré-Epidêmico: 2 casos por semana;
 - Epidêmico: 6 casos por semana;
 - Pós-Epidêmico: 5 casos por semana.
- México: população mais vulnerável:
 - Pré-Epidêmico: 1 caso por semana;
 - Epidêmico: 6 casos por semana;
 - Pós-Epidêmico: 4 casos por semana.

Laureano-Rosario et al. (2018a) treinaram redes MLP utilizando o algoritmo de otimização NSGA-II (*Non-dominated Sorting Genetic Algorithm II*, Algoritmo Genético de Sorteamento Não Dominado) (DEB et al., 2002; DEB et al., 2000; JENSEN, 2003), visando minimizar ao mesmo tempo os falsos positivos e os falsos negativos. A probabilidade de cruzamento do NSGA foi de 0,1. Já a probabilidade de mutação foi de 0,2. Esses valores foram escolhidos baixos para evitar a diversidade destrutiva, ou seja, a destruição de boas soluções por conta de cruzamentos excessivos ou mutações muito agressivas. Laureano-Rosario et al. (2018a) organizaram o treinamento da seguinte maneira: os dados de Porto Rico do número de casos de população sob risco (N_{PR24}) e de população mais vulnerável (N_{PR5-65}), foram separados em treinamento ($N_{PR24} = 1950$ e $N_{PR5-65} = 437$), validação ($N_{PR24} = 609$ e $N_{PR5-65} = 140$) e teste ($N_{PR24} = 1950$ e $N_{PR5-65} = 437$); para o México, os dados do número de casos da população sob risco (N_{MX24}) e mais vulnerável (N_{MX5-65}) foram organizados também em treinamento ($N_{MX24} = 2169$ e $N_{MX5-65} = 187$), validação ($N_{MX24} = 1643$ e $N_{MX5-65} = 89$) e teste ($N_{MX24} = 4096$ e $N_{MX5-65} = 459$).

Apesar da sofisticação do modelo de Laureano-Rosario et al. (2018a), ou seja, o uso de uma MLP otimizada por um algoritmo genético multiobjetivo, e de uma base dados bastante grande, os resultados não podem ser considerados bons: para Porto Rico, foram obtidas acurácias de 47% e 58% para predição de população sob risco e mais vulnerável, respectivamente; para o México, as acurácias foram de 51% e 66% para predição de população sob risco e mais vulnerável, nesta ordem. Foram ainda utilizadas como métricas de avaliação dos resultados a área sob a curva ROC e a medida F.

Como as informações de posicionamento geográfico não foram consideradas, o método proposto por Laureano-Rosario et al. (2018a) não retorna uma distribuição espacial aproximada

dos casos, o que dificulta o uso do resultado no planejamento local. No entanto, diferente de Kesorn et al. (2015) e Ch et al. (2014), os fatores ambientais relevantes para a predição foram avaliados utilizando o algoritmo NPSFS (*Neural Pathway Strength Feature Selection*, Seleção de Atributos por Força do Caminho Neural) (KESHTEGAR; HEDDAM; HOSSEINABADI, 2019; LAUREANO-ROSARIO et al., 2018b; DUNCAN, 2014). Foi verificada uma influência considerável do tamanho da população e da ocorrência de casos anteriores de dengue. Dos fatores climáticos e ambientais, o mais influente foi a temperatura do ar.

Lee, Chung e Hwang (2016) propuseram dois modelos para predição do número de mosquitos: um é baseado em rede neural artificial, e o outro em múltiplas regressões lineares, MLR (*Multiple Linear Regression*, Regressão Linear Múltipla). Como área de estudo foi escolhido o bairro de Yeongdeungpo-gu, da cidade de Seoul, Coréia do Sul. Foram utilizados dados diários de número de mosquitos coletados nos anos de 2011 e 2012. Os mosquitos são contados a partir de um sistema que realiza emissões de CO₂ para atrair-los. As contagens são então realizadas e disponibilizadas no Sistema de Monitoramento Digital de Mosquito (*Digital Mosquito Monitoring System*, DSM). Como variáveis climáticas e ambientais, foram utilizadas a temperatura, a velocidade dos ventos, a umidade e a pluviometria. O objetivo foi prever a população de mosquitos nos meses de maio a outubro de 2011 e de 2012. A base consiste em 317 instâncias. Dessas, 220 instâncias foram utilizadas para treinamento e 97 para teste. Para construção do modelo baseado em rede neural artificial, foi utilizada uma MLP com camada única. Foram investigadas configurações da MLP variando o número de neurônios da camada escondida de 2 a 23. Ambos os modelos são voltados a regressão. Como métricas de avaliação da qualidade, foram utilizados o Erro Médio Quadrático (*Root Mean Square Error*, RMSE), o Índice de Aceitação (*Index of Agreement*, IA), e o coeficiente de correlação (R). Os melhores resultados foram atingidos com o modelo baseado em rede neural MLP, com $RMSE = 14,38$, $IA = 0,75$ e $R = 0,61$. O peso dos fatores climáticos foi também analisado usando análise de relevância por seleção de atributos. A relevância da umidade foi de apenas 20%, enquanto temperatura, velocidade dos ventos e pluviometria tiveram uma relevância de 65% cada. Como as informações de posicionamento geográfico não foram consideradas, o método proposto não retorna uma distribuição espacial aproximada dos casos, o que dificulta o uso do resultado no planejamento local.

Iqbal e Islam (2017) fizeram uma pesquisa sobre trabalhos sobre métodos de predição de dengue, no período de 1995 a 2013. Foram pesquisados 30 artigos. Iqbal e Islam (2017) concluíram que, para prever surtos de dengue, é preciso considerar séries históricas de sintomas

(febre, dor de cabeça severa, náuseas, vômito, diarreia, hipotensão, hemorragia gastrointestinal, excesso de apetite ou anorexia, dor por trás dos olhos, dor abdominal, efusão pleural, sangramento pela boca ou pelo nariz, dor abdominal etc.) (RISSINO; LAMBERT-TORRES, 2009; BUTT et al., 2008; BRASIER et al., 2012), fatores climáticos e ambientais (temperatura, pluviometria, velocidade dos ventos e umidade) (LU et al., 2009; CHOUDHURY; BANU; ISLAM, 2008; WONGKON; JAROENSUTASINEE; JAROENSUTASINEE, 2012), além de fatores genéticos (BRASIER et al., 2012; BUTT et al., 2008; KRUIF et al., 2008; TANNER et al., 2008; NASIRUDEEN et al., 2011) e proteínas (REY, 2003; GUZMAN et al., 2010). Iqbal e Islam (2017) concluem que os melhores modelos para predição de surtos de dengue devem ser baseados em: árvores de decisão; redes neurais artificiais; classificadores evolucionários otimizados por algoritmos meta-heurísticos de otimização, como algoritmos genéticos; e conjuntos de classificadores (*ensemble classifiers*). Como as informações de posicionamento geográfico não foram consideradas, os métodos estudados e propostos por Iqbal e Islam (2017) não retornam uma distribuição espacial aproximada, o que dificulta o uso do resultado no planejamento local, no caso, no nível das províncias e de suas cidades. Iqbal e Islam (2017) também não analisam o peso de cada fator na predição.

Hamlet et al. (2018) descreveram uma maneira de parametrizar o efeito da temperatura na transmissão da febre amarela utilizando modelos estatísticos para prever as heterogeneidades geográficas e temporais na distribuição da doença, enquanto demonstraram sua robustez em comparação com os modelos que simplesmente predizem a distribuição geográfica. Hamlet et al. (2018) afirmam que essa quantificação da sazonalidade poderia levar a aplicações mais precisas de campanhas de vacinação e programas de controle de vetores. Por sua vez, isso ajudaria a maximizar o impacto dessas campanhas e racionalizar recursos, além de poder contribuir para diminuir o risco de surtos em larga escala. Os autores acreditam que os métodos descritos podem ser aplicados a outras doenças transmitidas pelos mosquitos *Aedes*, fornecendo uma ferramenta útil para o entendimento e o combate de várias outras doenças importantes, como a dengue e a zika. Hamlet et al. (2018) utilizaram modelos de regressão linear sobre as séries históricas de cada uma das 167 províncias selecionadas de países do continente africano, para aproximar probabilidades de surto de febre amarela. Os dados foram coletados semanalmente, de 1971 a 2015. Foram adotados limiares de probabilidade para avaliar os modelos de regressão como classificação. A métrica de qualidade utilizada foi a área da curva ROC, que variou de 0,79 a 0,84. Diferente dos outros trabalhos, Hamlet et al. (2018) combinam informações climáticas e ambientais (temperatura e índice de vegetação), de probabilidade de

surto de dengue e de localização geográfica, sendo os dados organizados como séries históricas de distribuições geoespaciais, permitindo também avaliação qualitativa dos resultados. Apesar das evidentes vantagens da análise espaço-temporal adotada, mais métricas para avaliação da qualidade da regressão e da classificação deveriam ter sido adotadas, pois a área sob a curva ROC tende a superestimar a qualidade da classificação.

Cortes et al. (2018) utilizaram dados de 2001 a 2014 para prever 2015, das cidades de Recife e Goiânia. Os dados foram coletados mensalmente e correspondem ao número de casos de dengue informados pelo Sistema Nacional de Notificações (SINAN), do Sistema Único de Saúde. O problema de predição é abordado como o ajuste de duas séries temporais usando os métodos de regressão ARIMA e SARIMA (*Season Auto-Regressive Integrated Moving Average*, Média Móvel Integrada Auto-Regressiva Sazonal). Os resultados não foram considerados adequados para a predição dos casos do Recife. A pesquisa careceu de considerar outros fatores importantes para a predição que não o histórico de casos, como as variáveis climáticas e ambientais e outras informações de interesse. Além disso, não houve nenhum tipo de refinamento que permitisse a predição no nível dos bairros e distritos, já que informações de posicionamento geográfico dos casos não foram coletadas.

Buczak et al. (2018) construíram dois modelos para prever surtos de dengue nas cidades de Iquitos, no Peru, e San Juan, em Porto Rico, baseados em conjuntos de regressores ARIMA e SARIMA. Os dados dos casos de dengue foram coletados semanalmente a partir dos respectivos sistemas de saúde pública dos dois países, no período de 2009 a 2010 e de 2012 a 2013. Foram utilizados dados de pluviometria, temperatura e vegetação, coletados diariamente por satélite. As informações de posicionamento geográfico não foram consideradas. Logo, os métodos propostos por Buczak et al. (2018) não retornam uma distribuição espacial aproximada, o que dificulta o uso do resultado no planejamento local, no caso, no nível dos bairros de cada cidade. Buczak et al. (2018) também não analisam o peso de cada fator na predição.

Para Albrieu-Llinás et al. (2018), sistemas de sensoriamento remoto e informações geográficas oferecem ferramentas valiosas para mapear a distribuição de espécies em uma determinada área. No entanto, a previsão de ocorrências de espécies por meio de mapas de distribuição de probabilidade baseados em pesquisas entomológicas¹ transversais tem utilidade limitada para as autoridades locais. Albrieu-Llinás et al. (2018) tiveram como objetivo examinar a evolução temporal do número de casas infestadas com estágios imaturos de *Aedes aegypti* em cada

¹ Entende-se como pesquisa entomológica a pesquisa que tem como objeto insetos e sua relação com os seres humanos, com outros seres vivos e com o ambiente.

bairro individual e investigar os agrupamentos (*clusters*) ambientais gerados com informações fornecidas por variáveis de sensoriamento remoto para explicar o comportamento observado ao longo do tempo. Pesquisas entomológicas foram realizadas entre 2011 e 2013 na cidade de Clorinda, Argentina, registrando o número de casas com criadouros com larvas de *Aedes aegypti*. Foram visitadas 10.981 casas, escolhidas aleatoriamente. Os dados foram organizados por bairro e coletados mensalmente. Clorinda possui 32 bairros. Uma imagem de satélite SPOT 5 foi usada para obter sete variáveis de cobertura da terra: solo descoberto, águas superficiais, zonas húmidas, vegetação baixa (erva), vegetação alta (arbustos e árvores), edifícios urbanos e pastagens ou culturas. Essas variáveis foram submetidas ao particionamento usando o algoritmo de k-médias para agrupamento de vizinhanças em quatro agrupamentos ambientais. O problema de predição de locais infectados pelo *Aedes aegypti* foi modelado como problema de regressão. Foi utilizado um regressor baseado em modelo linear generalizado. Os resultados foram apresentados também na forma de distribuições geoespaciais, usando mapas de calor, ou seja, mapas em pseudocor, montados após interpolação dos resultados. Como ferramenta para visualização da distribuição espacial e análise qualitativa, foi utilizado o sistema de informações geográficas Quantum GIS, ou QGIS. O método proposto por Albrieu-Llinás et al. (2018) mostrou um grande potencial para o planejamento local. No entanto, outros métodos de regressão poderiam ter sido testados.

De forma semelhante a Albrieu-Llinás et al. (2018), Scavuzzo et al. (2018) propuseram uma abordagem para predição de locais infectados por ovos e larvas do *Aedes aegypti* utilizando imagens de satélite e dados de posição de criadouros infectados. A área de estudos foi a cidade de Tartagal, da província de Salta, noroeste da Argentina, próxima à fronteira com a Bolívia. Os dados dos criadouros foram coletados semanalmente, de agosto de 2012 a julho de 2016, abrangendo sempre 50 imóveis escolhidos aleatoriamente. Para coletar os ovos do mosquito, foram utilizadas ovitrampas, armadilhas feitas de plástico que atraem as fêmeas para deposição dos ovos e os deixam presos (BRASIL, 2012). As variáveis climáticas e ambientais foram obtidas por meio de sensoriamento remoto: índice de vegetação (*Normalized Difference Vegetation Index*, NDVI) e índice de água e umidade (*Normalized Difference Water Index*, NDWI), do satélite MODIS MOD13Q1; distribuição de temperatura, obtida do satélite TRRM (*Tropical Rainfall Measuring Mission*) da JAXA (*NASA/Japan Aerospace Exploration Agency*). O problema de predição foi modelado como predição de séries temporais usando regressão. Foram experimentados métodos de regressão linear, SVM com *kernel* RBF, MLP com três camadas escondidas de três neurônios cada, k vizinhos mais próximos (kNN) e árvores de

decisão. Como índices de qualidade foram utilizados o índice de correlação e o MSE. Os melhores resultados foram obtidos com kNN, MLP e SVM, com correlação 0,888, 0,875 e 0,837, nesta ordem, contra regressão linear e árvore de decisão, com 0,774 e 0,679, respectivamente. Os resultados não foram apresentados na forma de distribuições geoespaciais, uma vez que a abordagem não usou as informações de posicionamento de cada criadouro. A abordagem de Scavuzzo et al. (2018) é semelhante à de Scavuzzo et al. (2017), que focam em redes neurais artificiais.

Ahmad et al. (2018) estudaram os sinais de alerta da Diretriz de Prática Clínica da Malásia na previsão de dengue severa e seus fatores associados entre os casos confirmados apresentados em um hospital escola no nordeste da Malásia em 2014. Os autores realizaram um estudo transversal em fevereiro de 2015, utilizando dados obtidos nos prontuários hospitalares. Foram 2607 casos confirmados de dengue apresentados ao Hospital da Universidade Sains Malaysia (HUSM) em 2014. 700 pacientes foram selecionados após amostragem aleatória estratificada proporcional (ou seja, respeitando a proporcionalidade de gênero, de etnia, de origem geográfica do paciente e de faixa etária da população) realizada de acordo com o número de casos em 12 meses diferentes em 2014. A base consistiu de 700 pacientes, correspondendo a 700 instâncias, com as seguintes características: 75,4% com mais de 15 anos, 24,6% com idade menor ou igual a 15 anos; 49,7% de homens e 50,3% de mulheres; 94,6% de malaios e 5,4% de outras etnias; 83,7% da região de Kota, 8,9% de Bachok e 7,4% de outras regiões. Os resultados de dengue severa representaram 4,9% dos casos. A prevalência de qualquer um dos sinais de alerta na dengue severa foi de 91,2%. Os sinais de alerta mais comuns antes da dengue severa eram vômitos persistentes (55,9%) e dor ou sensibilidade abdominal (52,9%). O sinal de alerta mais sensível na detecção da dengue severa foi dor abdominal (59%). Os sinais de alerta individuais se mostraram muito relevantes para a predição de dengue severa, especialmente de acúmulo de líquido clínico (99%), hepatomegalia (98%) e sangramento da mucosa (93%). Outros fatores associados à dengue severa foram vômitos persistentes (2,41%), sangramento da mucosa (4,73%) e aumento do hematócrito com queda rápida de plaquetas (2,74%). O problema foi modelado como classificação. Assim, foi usado um classificador baseado em uma única classe, detectando presença ou não de dengue severa. O classificador foi baseado em regressão logística. O resultado atingiu uma acurácia aproximada de 80% e uma área sob a curva ROC de 70%. Diferentemente das outras propostas, Ahmad et al. (2018) propõem muito mais um sistema de apoio ao diagnóstico, com avaliação da relevância dos atributos e considerando informações geográficas em certo nível, do que um sistema preditor para apoio ao controle da

distribuição da doença. Não há análise espacio-temporal. Além disso, outros métodos poderiam ter sido utilizados na classificação, tais como redes neurais artificiais e máquinas de vetor de suporte, que poderiam eventualmente contribuir para um maior desempenho da classificação.

Para Beltrán et al. (2018), as consequências devastadoras dos recém-nascidos infectados com o vírus zika tornam necessário combater e deter a propagação desse vírus e de seus vetores: os mosquitos *Aedes aegypti*. Uma parte essencial da luta contra os mosquitos é o uso da tecnologia móvel para apoiar a vigilância de rotina e a avaliação de risco pelos agentes de controle de endemias (ACEs). Além disso, para melhorar os sistemas de alerta precoce, as autoridades de saúde pública precisam prever com mais precisão onde é provável que ocorra um surto do vírus e seu vetor. O sistema ZIKA, proposto por Beltrán et al. (2018), visa desenvolver uma estrutura abrangente que combine o *e-learning* para capacitar ACEs, e proporcionar vigilância participativa baseada na comunidade e previsão de ocorrências e distribuição do vírus zika e seus vetores em tempo real. Atualmente, esse sistema está sendo implementado no Brasil, nas cidades de Campina Grande, Recife, Jaboatão dos Guararapes e Olinda, no Estado de Pernambuco e na Paraíba, com a maior prevalência da doença do zika vírus. O sistema ZIKA também tem como objetivo ajudar os ACEs a aprender novas técnicas e boas práticas para melhorar a vigilância do vírus e oferecer uma previsão de distribuição em tempo real do vírus e do vetor. O modelo de previsão proposto pode ser recalibrado em tempo real com informações provenientes dos ACEs, de instituições governamentais e de estações meteorológicas para prever as áreas com maior risco de surto de vírus zika e de outras arboviroses transmitidas pelo *Aedes aegypti* em um mapa interativo. Esse sistema de mapeamento e alerta tem o potencial de ajudar as instituições governamentais a tomar decisões rápidas e usar seus recursos de maneira mais eficiente para impedir a disseminação do vírus Zika. Embora proponham o uso de regressores Random Forest para fazer previsões, Beltrán et al. (2018) não realizaram experimentos e se concentraram na proposta do aplicativo móvel para apoio à vigilância participativa, sendo a proposta apenas um modelo teórico.

Zhao et al. (2020) desenvolveram um modelo nacional agrupado para prever contagens de casos de dengue em diferentes departamentos da Colômbia. Os autores assumiram que precipitação, temperatura do ar e tipo de cobertura da terra mostraram ser três importantes determinantes da abundância do mosquito *Aedes* e são frequentemente usados como preditores na previsão da dengue (GHARBI et al., 2011; DOM et al., 2013; SCAVUZZO et al., 2018; BEKETOV et al., 2014). Os dados de precipitação foram obtidos a partir do conjunto de dados de precipitação estimada diária CMORPH (*Climate Prediction Center Morphing Method*)

(JOYCE et al., 2004). As temperaturas da superfície terrestre foram extraídas dos produtos de imagem MODIS *Terra Land Surface Temperature 8-day* (MOD11C2.006). As estimativas do índice de vegetação aprimorado (EVI) foram obtidas a partir dos produtos de imagem de 16 dias do MODIS *Terra Vegetation Indices* (MOD13C1.006). Considerando o papel da injustiça social nas epidemias, os autores também incluíram a população, a cobertura educacional e o Índice de Gini (uma medida de desigualdade de renda) como preditores potenciais, obtidos do Departamento Administrativo Nacional de Estatística da Colômbia. Os dados de vigilância de casos de dengue foram extraídos de uma plataforma eletrônica, SIVIGILA, criada pelo programa nacional de vigilância da Colômbia e estava disponível em nível de departamento. O programa nacional de vigilância recebe relatórios semanais de todas as unidades públicas de saúde que atendem casos de dengue. Os casos de dengue notificados foram uma mistura de casos prováveis e confirmados por laboratório, sem distinção entre as duas definições de caso diferentes.

Zhao et al. (2020) descobriram que, para a maioria dos departamentos da Colômbia, o modelo nacional previu com mais precisão os futuros casos de dengue no nível do departamento em comparação com o modelo local. Isso indica a semelhança em importância dos vetores da dengue em diferentes regiões administrativas da Colômbia. O agrupamento de dados de departamentos individuais cria um conjunto de dados de treinamento com faixas maiores de variáveis, aumentando a capacidade de extrapolação dos modelos Random Forest. Os resultados com Random Forests foram superiores aos obtidos com MLPs e redes neurais convolucionais profundas. O modelo nacional agrupado treinado por um conjunto de dados maior teve maior precisão de predição em comparação com os modelos locais. Os autores também descobriram que as variáveis meteorológicas e ambientais eram mais importantes para a precisão da previsão em horizontes de previsão menores em comparação com as variáveis sociodemográficas, com os dados sociodemográficos sendo mais importantes em horizontes de previsão maiores. Moradia de má qualidade e gestão de saneamento com alta densidade populacional são os principais fatores de risco para a transmissão da dengue, intimamente relacionada à educação e à pobreza. Esses resultados demonstram a natureza complementar desses diferentes grupos de variáveis preditoras e a importância de sua inclusão nos modelos de previsão da dengue.

De acordo com Chakraborty, Chattopadhyay e Ghosh (2019), os conjuntos de dados da Dengue não são puramente lineares nem não lineares. Eles geralmente contêm padrões lineares e não lineares. Se for esse o caso, então nem o ARIMA nem uma Rede Neural Artificial

(ANN) individualmente seriam adequadas para modelar essas situações. Consequentemente, a combinação de modelos lineares e não lineares pode ser bem adequada para modelar com precisão essas estruturas de autocorrelação complexas. Modelos híbridos ARIMA-ANN se tornaram mais populares devido à sua capacidade de prever séries temporais complexas com precisão (ZHANG, 2003). *Neural Network Auto-Regression* (NNAR) corresponde a um modelo de rede neural *feed-forward* com apenas uma camada oculta com uma série temporal com valores defasados da série como entradas. Diferentemente das RNAs, SVMs e LSTMs puras, o NNAR é um modelo autorregressivo não linear. Modelos híbridos populares são o híbrido ARIMA-ANN (KHASHEI; BIJARI, 2011; ZHANG; PATUWO; HU, 1998), o modelo híbrido ARIMA-SVM (PAI; LIN, 2005) e o híbrido ARIMA-LSTM (CHOI, 2018). Esses modelos tentam ajustar os padrões lineares e não lineares dos dados da série temporal.

Chakraborty, Chattopadhyay e Ghosh (2019) propõem um modelo híbrido ARIMA-NNAR para capturar estruturas de dados complexas e comportamento linear e não linear de conjuntos de dados da dengue. Na primeira fase, o ARIMA captura os padrões lineares do conjunto de dados. Em seguida, o modelo NNAR é empregado para capturar os padrões não lineares nos dados usando valores residuais obtidos do modelo ARIMA base. Três conjuntos de dados populares de acesso aberto sobre dengue, a saber, dados de San Juan, Iquitos e as Filipinas, são usados para determinar a eficácia do modelo proposto. Diferentes modelos lineares e não lineares foram estudados nesses conjuntos de dados que mostram padrões altamente não lineares nessas regiões. Erro médio absoluto (MAE); erro quadrático médio (RMSE) e erro percentual médio absoluto simétrico (SMAPE) são usados como métricas de avaliação. Para as regiões endêmicas de San Juan e Iquitos, casos confirmados por laboratório semanais para os períodos de maio de 1990 a outubro de 2011 e de julho de 2000 a dezembro de 2011, respectivamente, são considerados neste estudo. O conjunto de dados das Filipinas contém os casos registrados mensalmente de dengue por 100.000 habitantes nas Filipinas. A incidência mensal de dengue está disponível para o período de janeiro de 2008 a dezembro de 2016. O conjunto de dados mensais das Filipinas contém um total de 108 observações mensais e usamos o total de casos relatados de todas as regiões das Filipinas neste estudo. O conjunto de dados semanais de San Juan contém um total de 1144 observações, enquanto o conjunto de dados de Iquitos contém apenas 520 observações. Os autores organizaram os três conjuntos de dados da dengue em conjuntos de treinamento e teste. Eles estudaram o modelo ARIMA, ANN, SVM, LSTM, NNAR para esses dados. O conjunto de dados é dividido em duas amostras de treinamento e teste para avaliar o desempenho de previsão do modelo proposto. O modelo

híbrido ARIMA-NNAR proposto foi capaz de prever tendências não lineares em comparação com os outros modelos, ou seja, ARIMA, ANN, SVM, NNAR, LSTM, ARIMA-SVM, ARIMA-ANN e ARIMA-LSTM.

Lima e Laporta (2020) tentaram avaliar a precisão dos modelos estatísticos determinísticos e estocásticos usando um protocolo para análise de séries temporais mensais da incidência de casos confirmados de dengue nas unidades federativas e distrito federal do Brasil de janeiro de 2000 a dezembro de 2017. A abordagem estatística e computacional foi aplicada para avaliar e validar a precisão de 10 modelos estatísticos para prever a série temporal de novos casos de dengue: ARIMA, *Exponential Smoothing Model* (ETS) (LIVERA; HYNDMAN; SNYDER, 2011), BATS (LIVERA; HYNDMAN; SNYDER, 2011), TBATS (NAIM; MAHARA; IDRISI, 2018; GOS et al., 2020), *Long-Short Term Memory* (LSTM), *Structural Model* (StructTS) (LENDEK et al., 2011), *Neural Network Auto-Regression* (NNETAR), *Extreme Learning Machine* (ELM) (YASEEN et al., 2019; ZHANG et al., 2019; CHRISTOU et al., 2019; TANG; DENG; HUANG, 2015; HUANG; WANG; LAN, 2011; HUANG; ZHU; SIEW, 2006; LIMA; SILVA-FILHO; SANTOS, 2016; SANTANA et al., 2018) e Multilayer Perceptron (MLP). Em modelos BATS, B significa transformação Box-Cox, A - resíduos ARMA, T - componente de tendência, S - componente sazonal, o que implica a transformação anterior de Box-Cox dos dados originais (GOS et al., 2020). Livera, Hyndman e Snyder (2011) introduziram uma representação trigonométrica dos componentes sazonais com base na série de Fourier, fornecendo o modelo TBATS. Os dados oficiais sobre os casos mensais de dengue de janeiro de 2000 a dezembro de 2016 foram usados para treinar os métodos estatísticos, enquanto os do período de janeiro a dezembro de 2017 foram utilizados para testar a capacidade preditiva de cada modelo, considerando três horizontes de previsão (12, 6, e 3 meses). Os resultados foram avaliados por meio do Erro Percentual Médio Absoluto (MAPE). Segundo os autores, a não consideração dos fatores externos (climáticos, ambientais, sociais ou imunológicos) e do índice de infestação predial pode ser visto como uma possível limitação deste estudo. Segundo os autores, os modelos de previsão baseados em ARIMA, TBATS e ELM poderiam ser usados para prever dengue nos estados brasileiros.

Stolerman, Maia e Kutz (2019) desenvolveram algoritmos de aprendizado de máquina para analisar séries temporais de clima e sua conexão com a ocorrência de anos de epidemia de dengue em sete capitais brasileiras. Os autores enfocaram o impacto de duas variáveis-chave: frequência de precipitação e temperatura média durante um determinado intervalo de janelas de tempo no ciclo anual. Os autores utilizaram os conjuntos de dados disponíveis publicamente do Sistema de Informação de Agravos de Notificação (SINAN), especialmente o número total

de casos de dengue por ano, de 2002 a 2017, para todas as capitais brasileiras. Os autores presumiram que os números relatados eram suficientes para identificar anos de epidemia de dengue. Um ano é considerado epidemia se, para uma determinada cidade, a incidência de dengue for superior a 100 casos por 100.000 habitantes no período de janeiro a dezembro. Para encontrar assinaturas climáticas críticas, a pesquisa se restringiu a sete capitais com pelo menos 3 anos epidêmicos e 3 anos não epidêmicos no período de 2002-2012. Os dados climáticos utilizados neste trabalho foram adquiridos do Instituto Nacional de Meteorologia (INMET), incluindo séries temporais de temperatura média e precipitação para as seguintes cidades: Aracajú, Belo Horizonte, Manaus, Recife, Salvador e São Luís (a partir de 01/01/2001 a 31/12/2012) e para o Rio de Janeiro (de 01/01/2002 a 31/12/2013). Em vez de regressão, os autores reduziram o problema de previsão da dengue à classificação de séries temporais. Eles usaram RBF e SVMs de kernel linear. Para avaliar seus resultados, eles empregaram precisão, uma métrica consistente com uma abordagem de classificação. Os autores usaram o modelo treinado com dados dos anos anteriores de 2002-2012 para prever os resultados da dengue de 2013-2017. A capital do estado de São Luís apresentou a maior precisão (100% correspondendo a 3 previsões corretas de um total de 3 anos de teste), seguida por Manaus e Salvador (80% de precisão correspondendo a 4 previsões corretas de um total de 5 anos de teste) Para Rio de Janeiro, Aracajú, Belo Horizonte e Recife, os autores alcançaram acertos abaixo de 70%. Os autores obtiveram uma precisão geral de 74% considerando todas as 7 capitais. Portanto, o método proposto previu corretamente o resultado de 23 dos 31 experimentos. Seus resultados indicam que cada capital brasileira considerada tem suas próprias assinaturas climáticas que se correlacionam com o número total de casos humanos de dengue. O inverno imediato antes de um ano epidêmico é um fator importante nas previsões de anos epidêmicos. No entanto, os autores reconhecem que sua abordagem para reduzir a previsão de dengue à classificação pode ser considerada um tanto arbitrária, uma vez que não é a forma canônica de prever arbovírus de acordo com o Ministério da Saúde do Brasil.

Na Tabela 1 é exibido um quadro comparativo dos trabalhos relacionados com esta proposta, de acordo com os critérios propostos por Siriyasatien et al. (2018a) e Bhunia e Shit (2019).

O trabalho proposto nesta Tese se destaca com relação aos demais apresentados no quadro comparativo da Tabela 1 da seguinte maneira: quanto à natureza dos dados, são utilizados dados demográficos, fatores climáticos e casos de arboviroses; com relação à natureza da análise, é uma análise espaço-temporal; diferente da maioria dos trabalhos do estado da arte escolhi-

dos, ele toma proveito da avaliação dos fatores mais relevantes para a predição; o problema está organizado como problema de regressão; com relação aos modelos de predição utilizados, são utilizados regressão linear, máquinas de vetor de suporte, redes neurais perceptron multicamadas, máquinas de aprendizado extremo, máquinas de estado de eco rasas e profundas, árvores de decisão individuais, e Random Forests; por fim, também diferente da maioria dos trabalhos do estado da arte, neste trabalho são geradas visualizações espaço-temporais dos resultados.

Tabela 1 – Quadro comparativo dos trabalhos relacionados com esta proposta, de acordo com os critérios propostos por Siriya et al. (2018a) e Bhunia e Shit (2019).

Trabalho	Natureza dos dados	Natureza da análise	Avaliação de fatores mais relevantes	Tipo de predição	Métodos de predição	Visualização espaço-temporal
Kesorn et al (2015)	Dados demográficos, fatores climáticos, casos de dengue	Temporal	Não	Regressão	SVM, MLP e árvore de decisão	Não
Ch et al (2014)	Dados demográficos, fatores climáticos, criadouros	Temporal	Não	Regressão	SVM-FFA, SVM, MLP, ARMA e ARIMA	Não
Shaukat et al (2015)	Dados clínicos de sintomas	Apoio ao diagnóstico precoce	Não	Classificação	Classificador ingênuo de Bayes, SVM e árvores de decisão	Não
Laureano-Rosario et al (2018)	Dados demográficos, fatores climáticos, casos de dengue	Temporal	Sim	Classificação	MLP e MLP otimizada por NSGA-II	Não
Lee, Chung e Hwang (2016)	Dados demográficos, fatores climáticos, criadouros	Temporal	Sim	Regressão	MLP e regressão linear múltipla	Não
Iqbal e Islam (2017)	Dados demográficos, fatores climáticos, dados clínicos de sintomas, casos de dengue	Temporal	Não	Classificação e regressão	MLP, árvores de decisão e MLP otimizadas por métodos evolutivos	Não
Hamlet et al (2018)	Dados demográficos, fatores climáticos, casos de dengue	Espaço-Temporal	Sim	Classificação e regressão	Regressor linear	Sim
Cortes et al (2018)	Casos de dengue	Temporal	Não	Regressão	ARIMA e SARIMA	Não
Buczak et al (2018)	Dados demográficos, fatores climáticos, casos de dengue	Temporal	Não	Regressão	ARIMA e SARIMA	Não
Albrieu-Llinás et al (2018)	Dados ambientais de superfície, criadouros	Espaço-Temporal	Não	Regressão	Regressor linear	Sim
Ahmad et al (2018)	Dados demográficos e clínicos de sintomas	Temporal	Sim	Classificação	Regressão logística	Não
Beltrán et al (2018)	Dados demográficos, fatores climáticos, criadouros	Espaço-Temporal	Não	Regressão	Random Forest	Não
Scavuzzo et al (2018)	Dados ambientais de superfície, fatores climáticos, criadouros	Temporal	Não	Regressão	Regressor linear, SVM, MLP, e árvores de decisão	Não
Zhao et al (2020)	Dados demográficos, fatores climáticos, casos de dengue	Espaço-Temporal	Não	Regressão	Random Forest	Sim
Chakraborty, Chattopadhyay e Gosh (2019)	Dados demográficos, casos de dengue	Temporal	Não	Regressão	ARIMA-ANN, SVM, LSTM, ARIMA-SVM, ARIMA-LSTM, e NNAR	Não
Lima e Laporta (2019)	Dados demográficos, criadouros	Espaço-Temporal	Não	Regressão	ARIMA, ETS, BATS, TBATS, LSTM, StrucTS, NNETAR, ELM, e MLP	Sim
Stolerman, Maia e Kutz (2019)	Dados demográficos, fatores climáticos, casos de arbovirose, criadouros	Espaço-Temporal	Sim	Classificação	RBF e SVM	Sim
Proposta	Dados demográficos, fatores climáticos, casos de arbovirose	Espaço-Temporal	Sim	Regressão	Regressor linear, SVM, MLP, ELM, ESN, Deep-ESN, Random Forest e árvores de decisão	Sim

3.2 CONSIDERAÇÕES

Neste Capítulo foi apresentado o estado da arte dos métodos de predição de locais infectados e de casos de arboviroses, levando em consideração a natureza dos dados, da análise (temporal, espaço-temporal ou de apoio ao diagnóstico), do tipo de predição (regressão ou classificação), dos métodos utilizados, se são avaliados os fatores mais relevantes na predição (seleção de atributos), e se é feita análise qualitativa dos resultados por meio de visualização espaço-temporal das distribuições geoespaciais aproximadas. No Capítulo 4 é apresentada a proposta de pesquisa deste trabalho.

4 PREDIÇÃO BASEADA NA ANÁLISE ESPAÇO-TEMPORAL

Neste capítulo é apresentada a metodologia proposta para a pesquisa, para predição espaço-temporal de casos de doenças infecciosas e avaliação dos fatores mais relevantes para a predição por meio de um Comitê de Especialistas Artificiais. Como estudo de caso, considera-se a predição espaço-temporal de casos de arboviroses tendo como vetor o mosquito *Aedes aegypti*. A análise de fatores mais relevantes para a predição, realizada pelo Comitê de Especialistas Artificiais proposto neste trabalho, é feita por meio de algoritmos bioinspirados de otimização para seleção de atributos. Seus resultados são comparados com os métodos do estado da arte, como a Análise de Componentes Principais. Também são apresentadas as bases de dados utilizadas nesta pesquisa: a base de casos de arboviroses (2013-2016), cedida pela Prefeitura da Cidade do Recife, e as bases de dados climáticos e ambientais (temperaturas velocidades dos ventos e pluviometrias), do Instituto Nacional de Meteorologia e da Agência Pernambucana de Águas e Clima.

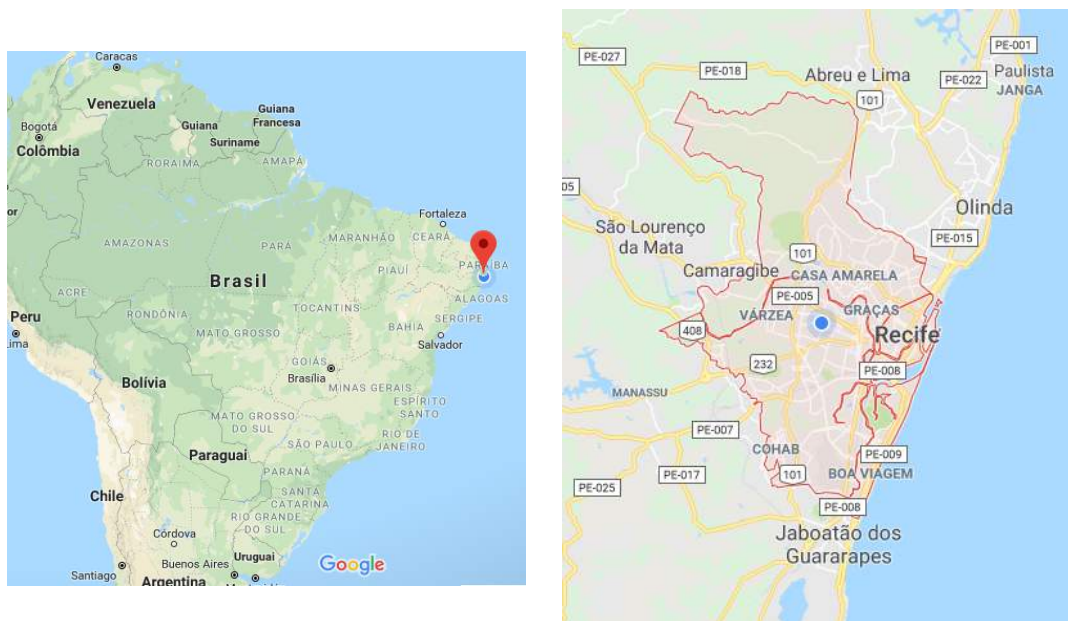
4.1 ÁREA SOB ESTUDO

A Cidade do Recife é a capital do Estado de Pernambuco, localizado na Região Nordeste do Brasil, latitude e longitude -8.053889 e -34.880833, respectivamente (ver Figura 19). Com área territorial de aproximadamente 218km², ocupa um território majoritariamente plano, formado por montes, ilhas, penínsulas e manguezais. De acordo com o Instituto Brasileiro de Geografia e Estatística - IBGE¹, Recife, é a cidade nordestina com o melhor Índice de Desenvolvimento Humano (IDH-M): 0,772, o que é considerado alto. É a quarta capital brasileira mais importante, após Brasília, Rio de Janeiro e São Paulo, e possui o quarto aglomerado urbano mais populoso do Brasil, com 4 milhões de habitantes em 2017, superado apenas pelas regiões metropolitanas de São Paulo, Rio de Janeiro e Belo Horizonte. Recife é sede do aglomerado urbano mais rico do Norte-Nordeste e oitavo mais rico do Brasil, além de ter o décimo quarto maior PIB do país e o maior PIB per capita entre as capitais nordestinas. A cidade é a nona mais populosa do país, com 1.637.834 habitantes, e sua região metropolitana

¹ Instituto Brasileiro de Geografia e Estatística, IBGE, Indicadores da Cidade do Recife, disponível em <<https://cidades.ibge.gov.br/brasil/pe/recife/panorama>>, acessado em 7 de janeiro de 2019.

é a sétima do Brasil em população.

Figura 19 – Perfil e localização geográfica da Cidade do Recife, capital do Estado de Pernambuco, Brasil, latitude -8.053889 e longitude -34.880833.



Fonte: A Autora

Em 2016, o salário médio mensal era de 3.1 salários mínimos. A proporção de pessoas ocupadas em relação à população total era de 44,8%. Na comparação com os outros municípios do estado, ocupava as posições 2 de 185 e 3 de 185, respectivamente. Já na comparação com cidades do país todo, ficava na posição 135 de 5570 e 93 de 5570, respectivamente. Considerando domicílios com rendimentos mensais de até meio salário mínimo por pessoa, tinha 38,1% da população nessas condições, o que o colocava na posição 180 de 185 dentre as cidades do estado e na posição 2991 de 5570 dentre as cidades do Brasil.

Em 2015, os alunos dos anos iniciais da rede pública da cidade tiveram nota média de 4,6 no IDEB. Para os alunos dos anos finais, essa nota foi de 3,9. Na comparação com cidades do mesmo estado, a nota dos alunos dos anos iniciais colocava esta cidade na posição 87 de 185. Considerando a nota dos alunos dos anos finais, a posição passava a 67 de 185. A taxa de escolarização (para pessoas de 6 a 14 anos) foi de 97,1 em 2010. Isso posicionava o município na posição 66 de 185 dentre as cidades do estado e na posição 3514 de 5570 dentre as cidades do Brasil.

A taxa de mortalidade infantil média na cidade é de 11,96 para 1.000 nascidos vivos. As internações devido a diarreias são de 0,6 para cada 1.000 habitantes. Comparado com todos os municípios do estado, fica nas posições 107 de 185 e 78 de 185, respectivamente. Quando comparado a cidades do Brasil todo, essas posições são de 2766 de 5570 e 3103 de 5570,

nesta ordem.

Apresenta 69,2% de domicílios com esgotamento sanitário adequado, 60,5% de domicílios urbanos em vias públicas com arborização, e 49,6% de domicílios urbanos em vias públicas com urbanização adequada (presença de bueiro, calçada, pavimentação e meio-fio). Quando comparado com os outros municípios do estado, fica na posição 20 de 185, 107 de 185 e 1 de 185, respectivamente. Já quando comparado a outras cidades do Brasil, sua posição é 1415 de 5570, 3654 de 5570 e 444 de 5570, nesta ordem.

A Cidade do Recife foi escolhida para o estudo de caso deste trabalho não somente pela disponibilidade de dados, mas principalmente pelos diversos surtos de arboviroses que já aconteceram: de forma mais endêmica, embora menos grave, a dengue, em suas diversas formas, inclusive hemorrágica; e os surtos de febre chikungunya e do vírus zika, em 2015, este último responsável pela ocorrência de milhares de casos de microcefalia em recém-nascidos nos anos de 2015 e 2016 (BRAGA et al., 2010; FALBO; FILHO, 2016; CORTES et al., 2018; MAGALHAES et al., 2017). Isso atesta o impacto social da predição dinâmica de arboviroses.

4.2 BASES DE DADOS

4.2.1 Sistema de Informações Geográficas da Agência Pernambucana de Águas e Clima - SIGH-PE

A Agência Pernambucana de Águas e Clima, APAC, é uma agência pública do Estado de Pernambuco, criada pela Lei Estadual 14.028 de 26 de março de 2010 para executar a Política Estadual de Recursos Hídricos, complementar o Sistema Integrado de Gerenciamento dos Recursos Hídricos em Pernambuco, SIGRH-PE, e fortalecer o planejamento e regulação dos usos múltiplos dos recursos hídricos no Estado. O SIGRH-PE é gerido de forma participativa e composto pelas seguintes entidades, além da APAC:

CRH: Conselho Estadual de Recursos Hídricos é o órgão superior deliberativo e consultivo do Sistema Integrado de Gerenciamento dos Recursos Hídricos.

COBHs: Comitês de Bacia Hidrográfica são parlamentos nos quais a sociedade manifesta seus interesses, define as prioridades para cada bacia hidrográfica. Encontram soluções negociadas para os conflitos e acompanham o desempenho da gestão pública.

CONSUS: Conselhos Gestores de Reservatórios - São colegiados formados por representan-

tes do poder público, dos usuários de água e da sociedade civil para atuar na área de influência de um açude.

SRHE: Secretaria de Recursos Hídricos e Energéticos é responsável pela formulação e gestão integrada das Políticas de Recursos Hídricos e de Saneamento.

A APAC realiza o monitoramento hidrometeorológico em tempo real de estações fluviométricas e pluviométricas através das PCDs (Plataformas de Coleta de Dados) distribuídas na região do Estado de Pernambuco. As estações pluviométricas monitoram a precipitação de chuvas em milímetros (mm). As estações fluviométricas monitoram o estado das barragens, a qualidade da água dos reservatórios, e as bacias hidrográficas do Estado. As informações são armazenadas no Sistema de Informações Geográficas SIGH-PE², cuja interface pode ser visualizada na Figura 20, com a distribuição das Plataformas de Coleta de Dados, PCDs, ao longo do Estado de Pernambuco. Os dados pluviométricos e fluviométricos são monitorados diariamente. Os dados podem ser exportados para planilhas, selecionando manualmente a PCD de interesse e o intervalo de tempo, com data de início e fim da coleta de dados. Os dados são distribuídos por dia ou por acumulado mensal.

Dessa base optou-se por usar as pluviometrias acumuladas mensais de 2009 a 2017, dado que a predição é feita usando ciclos de seis bimestres. As coordenadas utilizadas são a latitude e a longitude das estações de monitoramento: Codecipe / Santo Amaro, Várzea e Alto da Brasileira.

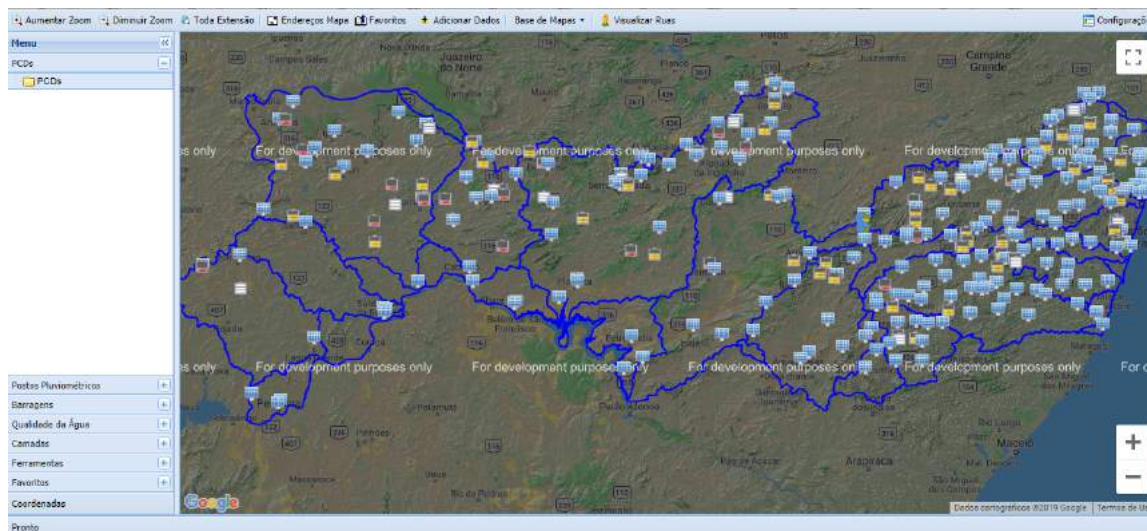
4.2.2 Banco de Dados Meteorológicos para Ensino e Pesquisa - BDMEP

O Instituto Nacional de Meteorologia (INMET) foi criado pelo Decreto 7.672 do presidente Nilo Peçanha, em 18 de novembro de 1909, com o nome de Diretoria de Meteorologia e Astronomia, órgão do Observatório Nacional, vinculado ao Ministério da Agricultura, Indústria e Comércio. O Instituto recebeu várias denominações ao longo de sua história até se chamar Instituto Nacional de Meteorologia (Lei 8.490, de 19 de novembro de 1992), órgão da administração direta do Ministério da Agricultura, Pecuária e Abastecimento.

A missão do INMET, ligado ao Ministério da Agricultura, Pecuária e Abastecimento, é prover informações meteorológicas à sociedade brasileira e influir construtivamente no processo

² Sistema de Informações Geográficas da Agência Pernambucana de Águas e Clima - SIGH-PE, disponível em <<http://www.apac.pe.gov.br/sighpe>>

Figura 20 – Interface do Sistema de Informações Geográficas da APAC, SIGH-PE, com a distribuição das Plataformas de Coleta de Dados, PCDs, ao longo do Estado de Pernambuco. Os dados pluviométricos (precipitação em mm) e fluviométricos são monitorados diariamente.



Fonte: A Autora

de tomada de decisão, contribuindo para o desenvolvimento sustentável do País. Assim, o INMET realiza monitoramento, análise e previsão de tempo e de clima.

São atribuições do INMET: elaborar e divulgar, diariamente, em nível nacional, a previsão do tempo, avisos e boletins meteorológicos especiais; promover a execução de estudos e levantamentos meteorológicos e climatológicos aplicados à agricultura e outras atividades correlatas; coordenar, elaborar e executar programas e projetos de pesquisas agrometeorológicas e de acompanhamento das modificações climáticas e ambientais; estabelecer, coordenar e operar as redes de observações meteorológicas e de transmissão de dados, inclusive aquelas integradas à rede internacional; propor a programação e acompanhar a implementação de capacitação e treinamento de recursos humanos, em atendimento a demandas técnicas específicas.

O INMET representa o Brasil junto à Organização Meteorológica Mundial (OMM), sendo responsável pelo tráfego das mensagens coletadas pela rede de observação meteorológica da América do Sul e os demais centros meteorológicos que compõem o Sistema de Vigilância Meteorológica Mundial. O INMET sedia um Centro de Sistema de Informação Mundial (*Geographic Information System Center, GISC*), integrante do principal núcleo do novo Sistema de Informação da OMM (*World Meteorological Organization Information System, WIS*). O WIS é resultado da evolução do Sistema Mundial de Telecomunicações (*Global Telecommunication System, GTS*).

O Sistema de Coleta e Distribuição de Dados Meteorológicos do Instituto (temperatura,

umidade relativa do ar, direção e velocidade do vento, pressão atmosférica, precipitação, entre outras variáveis) é dotado de estações de sondagem de ar superior (radiossonda); estações meteorológicas de superfície, operadas manualmente; e a maior rede de estações automáticas da América do Sul.

O Banco de Dados Meteorológicos do INMET já incorporou, em forma digital, em seu acervo, informações diárias coletadas desde 1961. Encontra-se em plena atividade um Projeto de Recuperação Digital de Dados Históricos que agregará à base de dados meteorológicos aproximadamente 12 milhões de documentos, patrimônio do clima observado desde tempos do Império (antes de 1900).

O Banco de Dados Meteorológicos para Ensino e Pesquisa³, BDMEP, é um banco de dados para apoiar as atividades de ensino e pesquisa e outras aplicações em meteorologia, hidrologia, recursos hídricos, saúde pública e meio ambiente. O Banco abriga dados meteorológicos diários em forma digital, de séries históricas das várias estações meteorológicas convencionais da rede de estações do INMET com milhões de informações, referentes às medições diárias, de acordo com as normas técnicas internacionais da Organização Meteorológica Mundial.

No BDMEP estão acessíveis os dados diários a partir de 1961 das estações para as quais se disponha, em forma digital, de pelo menos 80% dos dados que foram registrados naquele período. Os dados históricos referentes a períodos anteriores a 1961 ainda não estão em forma digital e, portanto, estão indisponíveis no BDMEP.

As variáveis atmosféricas disponibilizadas para consultas no BDMEP são: precipitação ocorrida nas últimas 24 horas; temperatura do bulbo seco; temperatura do bulbo úmido; temperatura máxima; temperatura mínima; umidade relativa do ar; pressão atmosférica ao nível da estação; insolação; direção e velocidade do vento. Dessa base optou-se por usar neste trabalho as temperaturas e velocidades dos ventos médias por mês. As coordenadas utilizadas são a latitude e a longitude das estações de monitoramento: Codecipe / Santo Amaro, Várzea e Alto da Brasileira.

³ Instituto Nacional de Meteorologia - INMET, disponível em <<http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>>, acessado em 30 de julho de 2018.

4.2.3 Portal de Dados Abertos da Prefeitura da Cidade do Recife

O Portal de Dados Abertos da Prefeitura da Cidade do Recife⁴ foi desenvolvido pela EMPREL, Empresa Municipal de Informática. Ele tem o objetivo de disponibilizar de forma pública o acesso e a busca de dados governamentais gerados por secretarias e órgãos da gestão municipal. A publicação dos dados em formato aberto permite que sejam desenvolvidas aplicações ou visualizações, buscando facilitar a análise dos dados, promover a melhoria de serviços por meio da inovação e da criatividade, e contribuir para uma maior participação da sociedade junto ao governo municipal. Os dados estão disponíveis nos formatos CSV e PDF. Também é possível fazer buscas diretas nas bases de dados usando buscas SQL e o protocolo JSON.

A Secretaria Municipal de Saúde disponibiliza 11 (onze) conjuntos de dados, dentre eles dados sobre os Distritos Sanitários, Vigilância Sanitária, Serviço de Atendimento Móvel de Urgência - SAMU, Unidades de Saúde, Academias da Cidade, e registros de casos de dengue, zika e chikungunya registrados nas unidades de saúde públicas e privadas. Os dados de registros de casos são diários e vão de 2013 a 2016.

Os dados contêm informações incompletas de endereços dos pacientes, para efeitos de anonimização. Essas informações podem ser geocodificadas para se obter latitude e longitude. A geocodificação foi feita parcialmente, utilizando um *plugin* do Google Maps. Alguns pontos não puderam ser geocodificados, por erros e informações incompletas nos cadastros. No entanto, como a resolução das variáveis climáticas e ambientais tem baixa resolução, devido à baixa densidade de pontos (apenas três pontos de temperatura, pluviometria e velocidade dos ventos por mês), optou-se por agrupar essas informações de casos por bairro e utilizar as coordenadas de latitude e longitude do ponto central do bairro. Uma vez que a predição de arboviroses é feita por meio de coletas bimestrais, foram utilizados os totais de casos em cada bimestre, começando em janeiro.

4.2.4 LIRAA Recife

O LIRAA Recife é a base de locais infectados obtida a partir do levantamento do índice LIRAA (Levantamento Rápido de Índices do *Aedes aegypti*), que por sua vez é uma metodologia

⁴ Portal de Dados Abertos da Cidade do Recife, disponível em <<http://dados.recife.pe.gov.br/dataset/casos-de-dengue-zika-e-chikungunya>>, acessado em 4 de janeiro de 2019.

para vigilância entomológica do mosquito *Aedes aegypti* para avaliação dos índices de Breteau e de Infestação Predial (BRASIL, 2012). O Índice de Infestação Predial, ou simplesmente Índice Predial, I_P , é a taxa percentual de imóveis positivos (onde foi encontrado o *Aedes aegypti*) pelo número de imóveis pesquisados. O Índice de Breteau, I_B , é a taxa percentual de recipientes positivos pelo número de imóveis pesquisados.

A base LIRAA Recife está organizada em planilhas por bimestre, de 2009 a 2017, começando em janeiro. Cada bimestre possui uma tabela por distrito sanitário. A Cidade do Recife possui seis distritos sanitários com os seguintes bairros ou distritos:

DS I: Santo Antônio, São José, Recife, Cabanga, Coelho, Ilha do Leite, Paissandú, Ilha Joana Bezerra, Boa Vista, Soledade e Santo Amaro.

DS II: Dois Unidos, Beberibe, Linha do Tiro, Água Fria, Porto da Madeira, Cajueiro, Campina do Barreto, Peixinhos, Arruda, Encruzilhada, Ponto de Parada, Fundão, Hipódromo, Rosarinho, Torreão, Campo Grande, Alto Santa Terezinha, Bomba do Hemetério.

DS III: Derby, Graças, Espinheiro, Aflitos, Jaqueira, Tamarineira, Santana, Parnamirim, Casa Forte, Poço da Panela, Casa Amarela, Alto José Bonifácio, Morro da Conceição, Alto José do Pinho, Vasco da Gama, Monteiro, Apipucos, Alto do Mandú, Sítio dos Pintos, Córrego do Jenipapo, Dois Irmãos, Nova Descoberta, Macaxeira, Brejo de Beberibe, Guabiraba, Brejo de Guabiraba, Passarinho, Mangabeira.

DS IV: Várzea, UR7 Várzea, Caxangá, Iputinga, Engenho do Meio, Torrões, Cordeiro, Zumbi, Torre, Prado, Madalena, Ilha do Retiro, Várzea Brasil, Cidade Universitária.

DS V: Afogados, Bongü, San Martín, Mustardinha, Mangueira, Estância, Jiquiá, Caçote, Areias, Tejió, Sancho, Barro, Coqueiral, Curado, Totó, Jardim São Paulo.

DS VI: Brasília Teimosa, Pina, Imbiribeira, Boa Viagem, Ipsep, Ibura, Jordão, Cohab, UR1, UR2, UR3, UR4, UR5, UR10.

A cada bimestre os distritos sanitários são divididos em estratos. Cada estrato engloba um ou mais bairros ou distritos. Cada linha de uma tabela bimestral corresponde a um estrato. Nas colunas da tabela estão o número de imóveis programados para visita, o número de imóveis efetivamente visitados, a quantidade de recipientes infectados por tipo de recipiente e total. Os tipos de recipiente correspondem aos códigos A1, A2, B, C, D1, D2 e E (BRASIL, 2012). Essas diversas tabelas são montadas a partir dos relatórios consolidados, que por sua

vez são gerados a partir da consolidação dos resultados obtidos pelas visitas dos Agentes de Saúde Ambiental e Controle de Endemias, ASACEs, no bimestre correspondente.

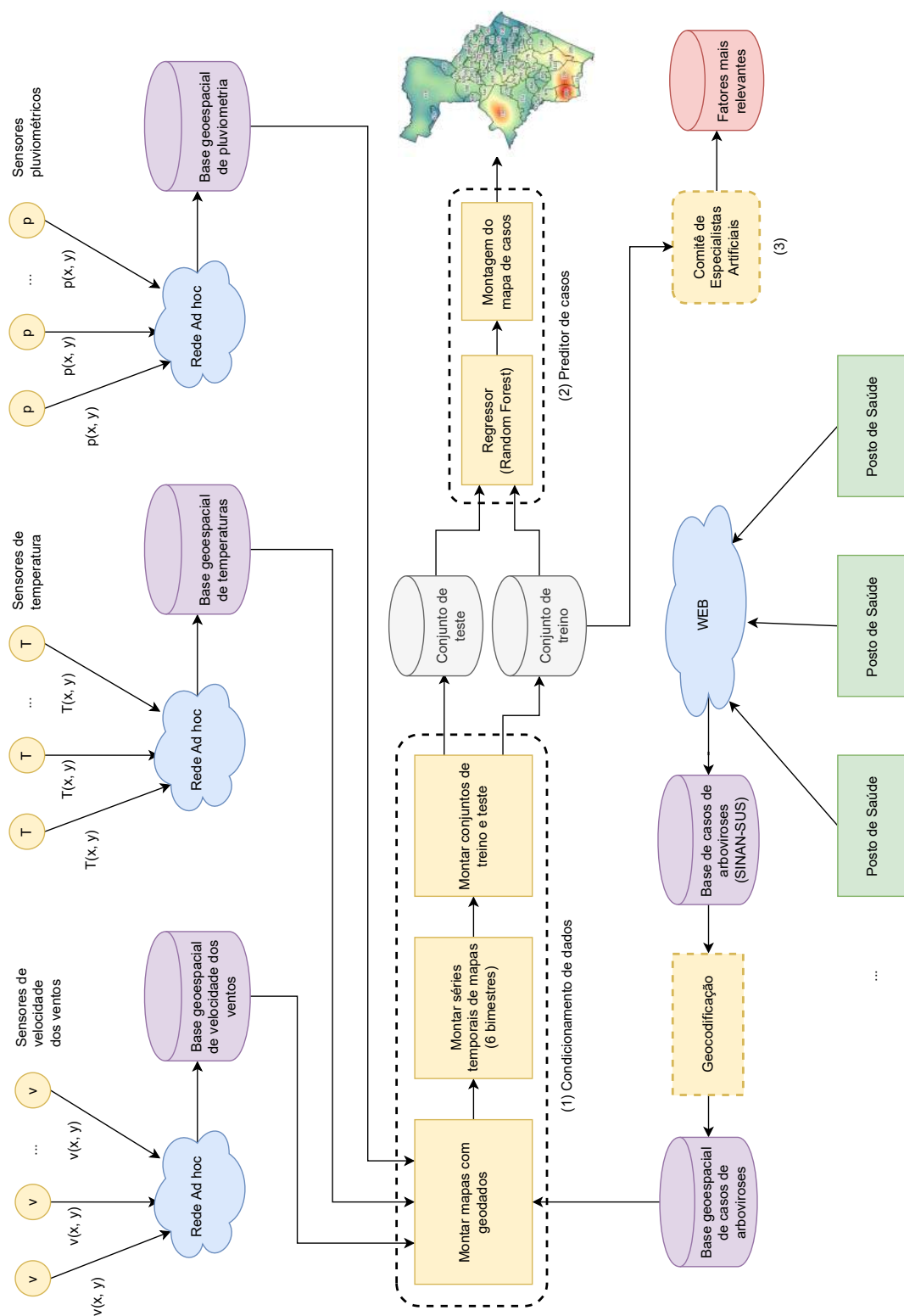
Cada ASACE (Agente de Saúde Ambiental e Controle de Endemias, termo com o qual o Agente de Controle de Endemias, ACE, é denominado na Cidade do Recife) preenche formulários de visita com os endereços completos dos imóveis visitados. Esses endereços podem ser geocodificados para serem transformados em coordenadas de latitude e longitude, gerando assim distribuições bastante precisas de locais infectados. No entanto, embora essa tarefa possa ser realizada pelos ASACEs usando aplicativos específicos e a Prefeitura da Cidade do Recife esteja desenvolvendo um aplicativo para apoio aos ASACEs junto com a EMPREL, hoje esses dados se encontram em formulários em papel ou em arquivos PDF, o que tomaria muito tempo para coletá-los e transformá-los em planilhas pesquisáveis. Assim, neste trabalho se preferiu aqui expandir os estratos em bairros, para se ter uma distribuição dos dados por bairro e, a partir dessa distribuição, posteriormente estimar a distribuição espacial dos locais infectados para cada bimestre por interpolação. No entanto, o foco deste trabalho foi a predição da distribuição de casos.

4.3 PROPOSTA

Neste trabalho é proposta uma metodologia para predição espaço-temporal de casos de doenças infecciosas e de fatores de risco, baseada em séries históricas de distribuições espaciais de casos e de informações climáticas e ambientais, a saber, a distribuição da velocidade dos ventos, das temperaturas e da pluviometria. Como estudo de caso, a proposta é aplicada à predição de casos de arboviroses por meio de técnicas de aprendizado de máquina e análise de fatores de risco usando algoritmos bioinspirados meta-heurísticos para seleção de atributos. A metodologia proposta neste trabalho, aplicada à predição de arboviroses, é ilustrada no diagrama da Figura 21.

O modelo proposto pode ser utilizado para predição dinâmica em tempo real. Idealmente, seriam coletadas informações georreferenciadas de velocidade dos ventos, temperatura e pluviometria a partir de redes de sensores distribuídos pela cidade. Esses sensores poderiam estar conectados por meio de redes *ad hoc* (redes para aplicações específicas) ou ligados diretamente à Internet, numa abordagem de Internet das Coisas (*Internet of Things*, IoT), alimentando bancos de dados específicos com as grandezas medidas (velocidade dos ventos, temperatura ou pluviometria) e as posições de latitude e longitude de cada sensor coletadas por GPS (*Global*

Figura 21 – Metodologia proposta aplicada à predição de arboviroses: predição da distribuição espacial de casos de arboviroses a partir de bases de dados de séries históricas de distribuições de informações climáticas e ambientais, de vigilância ambiental e clínicas, com destaque para as contribuições deste trabalho: (1) metodologia para organização das bases de dados e construção dos conjuntos de treino e de teste; (2) modelo de predição espaço-temporal baseado em regressor com Random Forest; e (3) o Comitê de Especialistas Artificiais.



Fonte: A Autora

Positioning System, Sistema de Posicionamento Global). Como esta ainda não é a realidade da Cidade do Recife, como estudo de caso, para validação do modelo proposto, foram utilizados dados climáticos e ambientais de 2009 a 2017, da Agência Pernambucana de Águas e Clima (APAC) e do Instituto Nacional de Meteorologia (INMET).

Os dados de casos de arboviroses poderiam também ser coletados diretamente da base de dados do sistema eSUS (Acesso Eletrônico ao Sistema Único de Saúde), onde são cadastrados dados de ocorrência de arboviroses coletados tanto das unidades de saúde do sistema público quanto de unidades privadas. No entanto, atualmente, o acesso a esses dados em tempo real é restrito, sendo fornecido por meio de plataformas de dados abertos de tempos em tempos para fins de pesquisa e desenvolvimento. Para um acesso maior a esses dados, é necessário solicitar acesso diretamente à administração do eSUS a partir de um projeto aprovado por um Comitê de Ética em Pesquisa. Como estudo de caso, para validação do modelo proposto, foram utilizados dados cedidos pela Secretaria de Saúde da Prefeitura da Cidade do Recife, de 2013 a 2016, por meio do Portal de Dados Abertos da Prefeitura da Cidade do Recife. Esses dados são organizados por bairro e, por meio de geocodificação, recebem as coordenadas de latitude e longitude do centro do bairro. Contudo, também se poderia usar a geocodificação do endereço simplificado do paciente, já que existem informações incompletas de endereço e suficientemente anonimizadas, ou seja, sem informações que identifiquem os pacientes. Contudo, para validar o modelo proposto e tendo em vista que as outras bases de dados não são tão completas quanto à distribuição espacial, optou-se por manter a representação por bairro.

Os dados dos locais infectados podem ser coletados pelos Agentes de Controle de Endemias (ACEs), por meio de aplicativos móveis onde são registradas as visitas aos imóveis designados e são armazenadas as coordenadas geográficas do imóvel e o seu estado: se há potenciais criadouros de mosquitos e, havendo, se há infestação de mosquitos *Aedes aegypti*, seja na forma de mosquito adulto, larva ou mesmo ovos. Em verdade, há esforços em conjunto para tornar essa uma realidade: o projeto "*A gamified m-training app for health professionals on protocols and participatory surveillance associated with Zika virus*" (2017-2019), financiado pela Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco, FACEPE, e pelo British Council, por meio do Fundo Newton. O projeto resultou de uma parceria entre a Universidade Federal de Pernambuco, a Universidade Federal de Campina Grande, a University College London, e as prefeituras de Recife e Olinda, justamente para construir um aplicativo gamificado para essa função (BELTRÁN et al., 2018). Posteriormente, em junho de 2018, a Prefeitura da Cidade do Recife, junto com a Empresa Municipal de Informática, EMPREL,

desenvolveu uma versão simplificada da proposta inicial para uso próprio. No entanto, como o uso desses aplicativos ainda não é uma realidade, priorizando a validação da proposta, optou-se pela base de relatórios consolidados LIRAA Recife, usando geocodificação, ou seja, transformando as informações textuais de bairros e distritos em coordenadas geográficas.

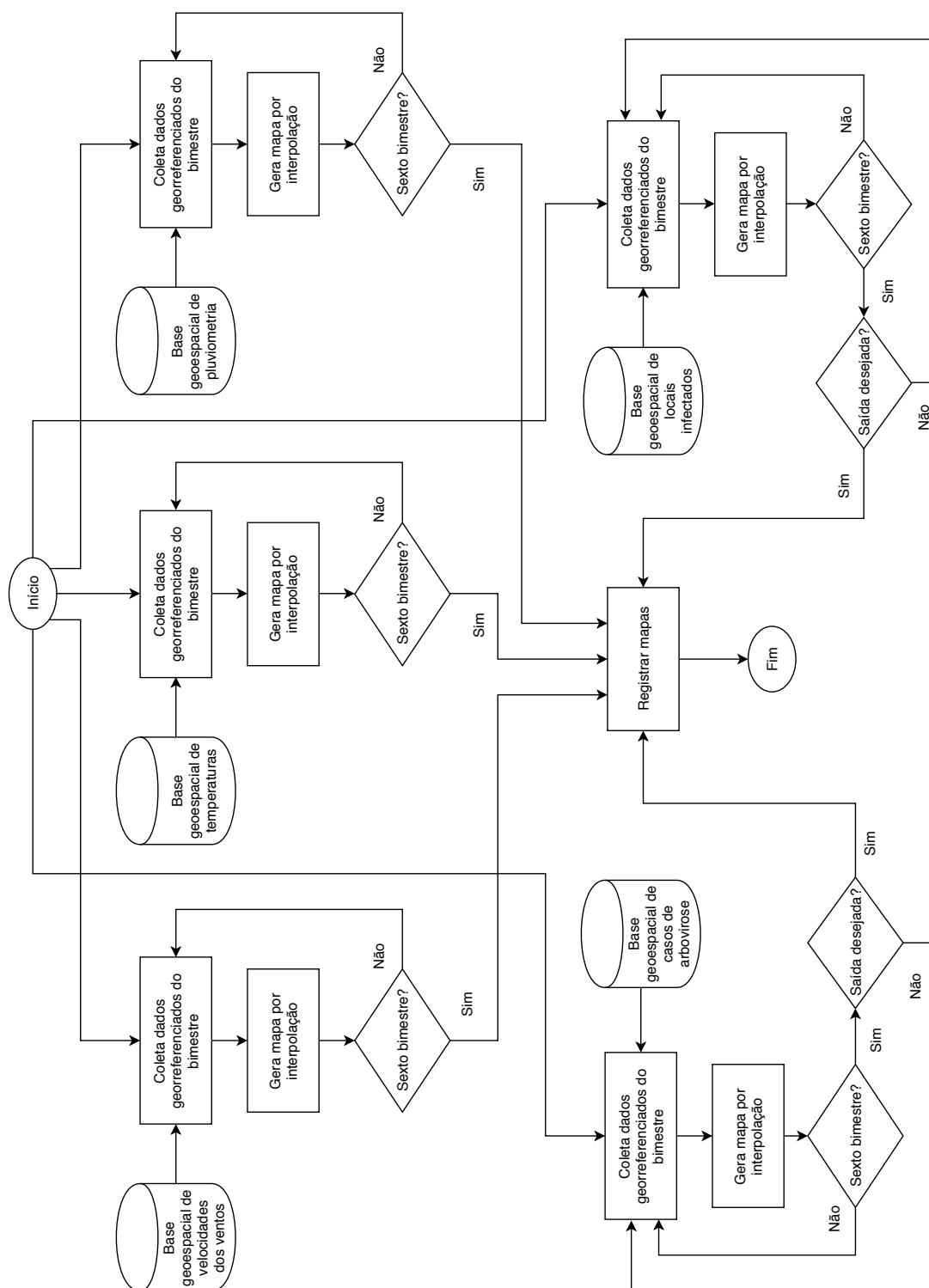
Adotando a perspectiva de utilizar fontes de dados modernas, na metodologia proposta também se poderia incluir dados de locais infectados coletados em redes sociais, como o Twitter, Facebook e Instagram, como também se poderia desenvolver aplicativos móveis exclusivos para vigilância sanitária participativa. Diversos trabalhos mostram que é possível fazer previsões epidemiológicas usando as redes sociais (QUINCEY; KOSTKOVA, 2009; SZOMSZOR; KOSTKOVA; QUINCEY, 2010; KOSTKOVA, 2015; STEIGER; RESCH; ZIPF, 2016), inclusive de dengue (SOUSA et al., 2018).

Para fazer uma previsão, são coletados os “acumulados” mensais (i.e. médias mensais) de cada uma das variáveis climáticas e ambientais, ou seja, temperatura, velocidade dos ventos e pluviometria, correspondentes a 12 meses. Para cada variável, para cada mês, é montado um mapa com a distribuição espacial daquela variável, ou seja, um plano de informação (BHUNIA; SHIT, 2019), naquele mês. Para os dados de casos de arboviroses é montado um mapa com a distribuição espacial dessa variável no bimestre, não no mês, dado que, no âmbito do SUS, o planejamento de ações de combate a surtos de arboviroses é feito considerando ciclos de bimestres. Nessa abordagem não se separa os modelos de previsão por doença, pois o foco da prevenção é no vetor, no caso, o mosquito *Aedes aegypti* (CARDOSO et al., 2015; ROTH et al., 2014; VILLAMIL-GÓMEZ et al., 2016; ZANLUCA et al., 2015; MAYER; TESH; VASILAKIS, 2017). São previstas as distribuições espaciais de casos de arboviroses do bimestre imediatamente seguinte. Os mapas gerados devem ser registrados, ou seja, os pontos da grade de cada mapa devem corresponder exatamente à mesma posição espacial. O processo de montagem e registro dos mapas das variáveis e das previsões desejadas está detalhado na Figura 22. Para montagem dos mapas por interpolação, foi utilizado o sistema de informações geográficas livre e aberto QGIS⁵. Os dados são passados para o QGIS por meio de planilhas, onde cada posição geográfica ocupa uma linha e, nas colunas, estão dispostas as informações de latitude, longitude, e as variáveis de interesse. Os mapas são então concatenados como na Figura 23.

A partir dos dados georreferenciados coletados, é montada uma grade de pontos por interpolação. Os conjuntos de treinamento são montados percorrendo os pontos da grade. A

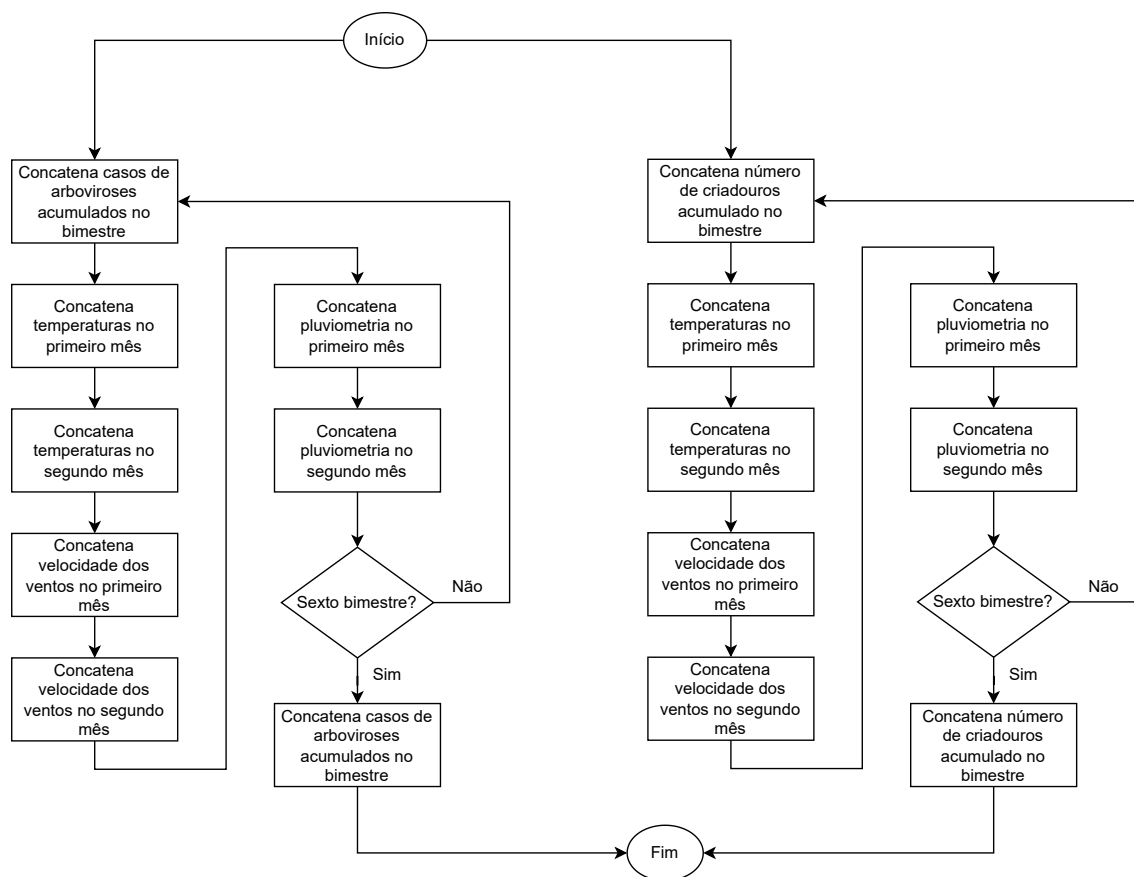
⁵ QGIS, um Sistema de Informações Geográficas Livre e Aberto, disponível em <https://www.qgis.org/pt_BR/site/>, acessado em 8 de janeiro de 2019.

Figura 22 – Processo de coleta de dados georreferenciados, montagem dos mapas das variáveis de interesse (temperaturas, pluviometrias, velocidades dos ventos e casos de arboviroses para os seis bimestres), das saídas desejadas, i.e. casos de arboviroses para o bimestre seguinte aos seis anteriores, e registro.



Fonte: A Autora

Figura 23 – Processo de concatenação dos mapas registrados.



Fonte: A Autora

cada ponto corresponde uma instância, modelada como um vetor de atributos. Os vetores de atributos são montados varrendo os mapas concatenados simultaneamente, concatenando latitude, longitude, e as seguintes informações para cada um dos seis bimestres, nesta ordem: número de casos, temperaturas, pluviometrias e velocidades dos ventos, para o primeiro e o segundo mês do bimestre, cada. A saída desejada é o número de casos na coordenada correspondente.

Com os conjuntos de treinamento, são investigadas as melhores arquiteturas de regressores para predição de casos de arbovirose. São candidatas as seguintes arquiteturas:

- Máquina de Vetor de Suporte (SVM);
- Random Forest (RF);
- Perceptron Multicamadas (MLP);
- Máquina de Aprendizado Extremo (ELM);
- Máquina de Estado de Eco (ESM);

- Rede de Estado de Eco Profunda (Deep-ESN);
- Regressão Linear (LR).

Cada regressor foi avaliado em 30 rodadas usando validação cruzada com 10 *folds*, totalizando 300 experimentos de treinamento e teste. Para avaliação quantitativa, foram calculadas as seguintes métricas de qualidade de predição: coeficiente de correlação (R), τ (tau) de Kendall (KE), ρ (rô) de Spearman (SP), o erro absoluto médio (MAE), o erro quadrático médio (RMSE), o erro absoluto relativo (MAE percentual) e o erro quadrático relativo (RMSE percentual). Os resultados também foram avaliados qualitativamente por meio da extrapolação dos modelos treinados para um dado conjunto de teste, sendo os resultados interpolados para visualização na forma de mapa e comparados à distribuição real. Para validação quantitativa, foi utilizada a biblioteca Weka, para predição por meio de regressão; o sistema de informações geográficas QGIS, para visualização dos mapas de entrada e resultantes; e programas específicos implementados em Python e em Octave, para geração de grade de pontos a partir de interpolação e para implementação dos regressores ELM, ESM e Deep-ESN, respectivamente. Para validação qualitativa, para geração de mapas com os resultados, foi utilizado o QGIS.

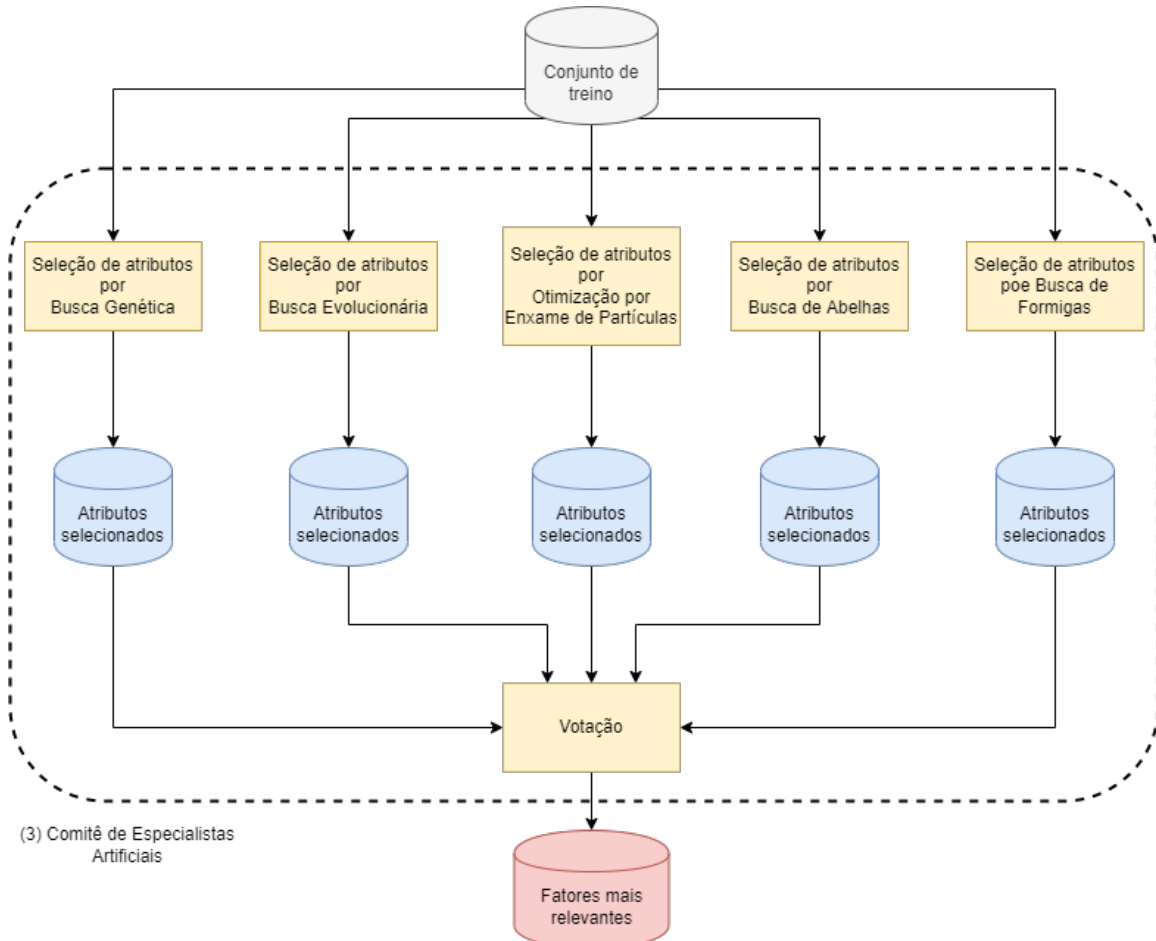
Os fatores mais relevantes para cada predição foram avaliados utilizando como estratégia métodos de seleção de atributos hibridizados com métodos bioinspirados de otimização. A seleção de atributos retorna a relevância de cada atributo para a predição e, assim, destaca as variáveis mais relevantes. Foi hibridizado o método de ranqueamento simples, que por sua vez se baseia em uma árvore de decisão como classificador/regressor de referência. Foram investigados os seguintes métodos:

- Algoritmo Genético (GA);
- Busca Evolucionária (Algoritmo Genético Modificado, ES);
- Otimização por Enxame de Partículas (PSO);
- Colônia de Abelhas Artificiais (ABC);
- Colônia de Formigas Artificiais (AAC).

Todos os métodos meta-heurísticos foram rodados considerando um máximo de 500 gerações e populações iniciais de tamanho 20. Esses parâmetros foram definidos empiricamente, observando a convergência para a solução, considerando a função objetivo árvore de decisão

Random Tree, adotada em comum a todos os métodos. Os experimentos com métodos bioinspirados de otimização aconteceram antes da predição. Esses métodos integram um Comitê de Especialistas Artificiais, proposto neste trabalho para seleção dos fatores mais relevantes para a predição. A arquitetura proposta é mostrada no diagrama da Figura 24. Apesar de se poder argumentar que árvores de decisão, pela sua própria definição, já pode ser utilizadas como modelos para seleção de atributos, e que Random Forests já poderiam ser utilizadas como comitês de seleção de atributos, neste trabalho o principal diferencial é selecionar o conjunto de atributos que otimiza o desempenho de regressão por uma árvore de decisão. Assim, a seleção de conjuntos de atributos realizada por métodos meta-heurísticos é global, em função de uma métrica global de qualidade de regressão, enquanto a seleção feita por árvores de decisão considera métricas locais, como o ganho de informação.

Figura 24 – Comitê de Especialistas Artificiais para seleção de atributos, onde cada especialista é baseado em um algoritmo de busca heurística bioinspirado: algoritmo genético (busca genética), algoritmo genético modificado (busca evolucionária), otimização por enxame de partículas, colônia de abelhas artificiais (busca de abelhas), e colônia de formigas artificiais (busca de formigas), para populações iniciais de 20 indivíduos e máximo de 500 gerações ou iterações.



Fonte: A Autora

4.4 ARRANJO EXPERIMENTAL

Os experimentos foram realizados em computador de uso pessoal, processador Intel I7, 16GB de RAM, sistema operacional Windows 10. A fusão das bases de dados e construção das bases de conhecimento foi feita utilizando um programa construído em Python especialmente para isso (OLIPHANT, 2007). As bases de conhecimento foram salvas em arquivo texto, no formato ARFF. Os experimentos de regressão foram, na sua maioria, realizados no ambiente Weka de aprendizado de máquina e mineração de dados, exceto para os métodos ELM, ESM e Deep-ESM. O ambiente Weka foi construído usando a linguagem Java, e pode ser estendido pela adição de outras bibliotecas e *plugins* (HALL et al., 2009; HOLMES;

DONKIN; WITTEN, 1994; WITTEN et al., 2016). Os métodos ELM, ESM e Deep-ESM foram testados usando programas construídos na linguagem Matlab/Octave, no ambiente GNU Octave. Os resultados brutos foram armazenados em planilhas MS Excel. Os gráficos estatísticos em boxplots foram construídos usando o programa SciDavis, voltado para construção de gráficos científicos (STANDISH, 2016). Já os mapas foram construídos usando o sistema de informações geográficas Quantum GIS (QUANTUM, 2013; HUGENTOBLE, 2008).

4.5 CONSIDERAÇÕES

Neste capítulo foi descrita a metodologia proposta neste trabalho; as bases de dados utilizadas para obtenção das variáveis climáticas e ambientais e suas distribuições geográficas; e a base de dados de casos de arboviroses transmitidas pelo mosquito *Aedes aegypti*, clinicamente atestados, obtida do Sistema Nacional de Notificações, SINAN, do SUS. Também foi apresentada a área geográfica estudada, no caso, a Cidade do Recife em todos os seus bairros. Foram ainda apresentados os métodos de regressão que serão investigados tanto para as predições de casos de arboviroses, com especial destaque para aqueles baseados em redes neurais artificiais. Para avaliação dos fatores de risco, apresentamos uma proposta de Comitê de Especialistas Artificiais baseado em múltiplos algoritmos meta-heurísticos de otimização baseados em meta-heurísticas. No Capítulo 5 serão apresentados os resultados experimentais e a discussão dos resultados qualitativos e quantitativos desta pesquisa.

5 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados quantitativos e qualitativos. São também feitas discussões e tecidas conclusões sobre a metodologia de predição espaço-temporal proposta, aplicada a arboviroses, os modelos espaço-temporais de predição de casos propostos e avaliados, e os fatores mais relevantes levantados pelo Comitê de Especialistas Artificiais.

5.1 PREDIÇÃO DE CASOS DE ARBOVIROSES

5.1.1 Modelo preditor

Conforme apresentado no Capítulo 4, o modelo preditor de casos de arboviroses proposto neste trabalho é baseado na análise espaço-temporal, combinando informações de variáveis climáticas e ambientais dos acumulados mensais, especificamente temperatura, pluviometria e velocidades dos ventos, com os acumulados bimestrais do número de casos, considerando um ciclo de seis bimestres, ou seja, um ano, que é o ciclo do mosquito *Aedes aegypti*. Todas essas informações são relacionadas a coordenadas de latitude e longitude, sendo o resultado da predição também associado ao posicionamento geográfico. Assim, os vetores de atributos possuem um total de 44 atributos, conforme definido no Capítulo 4. Os dados foram coletados das respectivas bases de dados considerando uma distribuição simplificada por bairro, considerando todos os 94 bairros da Cidade do Recife. A cada bairro é associada uma posição geográfica central, ou seja, um par latitude e longitude. Para fins de investigação das melhores arquiteturas de regressores, foram geradas distribuições aproximadas para cada variável de interesse (atributo) usando uma grade irregular, gerando bases de 1786 instâncias de vetores de 44 atributos. Uma vez que a predição de arboviroses é feita por meio de coletas bimestrais, foram utilizados os totais de casos em cada bimestre, começando em janeiro. Utilizando os dados dos anos de 2013, 2014, 2015 e 2016 e considerando que o ciclo considerado para predição é de seis bimestres, ou seja, um ano, são preditos cada um dos seis bimestres de 2014, 2015 e 2016 considerando as informações geoespaciais dos respectivos anos anteriores. Isso tudo resulta em 18 bases.

Cada regressor foi avaliado em 30 rodadas usando validação cruzada com 10 *folds*. Assim, para avaliação geral da predição, considerando que são um total de 18 bases, cada regressor foi

avaliado em $30 \times 10 \times 18 = 5400$ treinamentos e testes. Para avaliação quantitativa, foram calculadas as seguintes métricas de qualidade de predição: Coeficiente de Correlação (R), τ (tau) de Kendall (KE), ρ (rô) de Spearman (SP), o erro absoluto médio (MAE), o erro quadrático médio (RMSE), o erro absoluto médio relativo (MAE percentual) e o erro quadrático médio relativo (RMSE percentual). No entanto, na análise dos dados, foram considerados apenas R, como métrica de qualidade global, e RMSE(%), como métrica de qualidade local. Foram avaliadas as seguintes configurações de regressores, definidas empiricamente, onde a Regressão Linear (*Linear Regression*, LR) é considerada o regressor padrão em aplicações ambientais de análise espaço-temporal, ou seja, como referência (*baseline*):

- Máquina de Vetor de Suporte (SVM): $C = 0, 1$, *kernels* linear (ou grau 1), polinomial de graus 2 e 3, e RBF;
- Random Forest (RF): de 10 a 100 árvores, incrementando as árvores de 10 em 10;
- Perceptron Multicamadas (MLP) com uma única camada escondida com 2, 5, 10 e 20 neurônios;
- Máquina de Aprendizado Extremo (ELM): uma única camada escondida, 50 e 100 neurônios, testando *kernels* linear, polinomial de graus 2 e 3, e RBF;
- Máquina de Estado de Eco (ESM): com uma única camada escondida e versão profunda com 2, 5 e 10 camadas, 10 neurônios por camada;
- Regressão Linear (LR).

O Coeficiente de Correlação R, uma grandeza que vai de 0 (total desconexão) a 1 (total correlação), foi considerado para analisar o comportamento global de cada predição, ou seja, como determinado método regressor se comporta no plano mais geral. A métrica R é considerada uma métrica otimista: o resultado pode parecer muito bom, algo maior de 0,8, por exemplo, e os erros locais, pontuais, podem ser altos. Portanto, o R não pode ser considerado sozinho. Para compensar isso, foi escolhido também observar a métrica RMSE(%), que serve como uma medida de erro local. Quanto mais próximo de 0%, melhor. Assim, um regressor é considerado bom se tiver um R considerado alto e um RMSE(%) considerado baixo. Neste trabalho se está considerando um R alto como sendo acima de 0,9, e um RMSE(%) baixo como sendo abaixo de 5%. O tempo de treinamento foi medido em milissegundos (milésimos de segundos). Tempos de treinamento menores que 1 milissegundo são considerados 0. Muito

embora o tempo de treinamento não seja crítico para esta proposta, essa grande ainda assim foi medida, para ajudar a decidir que configurações de regressores são melhores dentro de uma mesma família de métodos de regressão (exemplo: redes neurais MLP com diferentes quantidades de neurônios na camada escondida). As estatísticas para os tempos de execução ou teste não puderam ser obtidas, já que, para todos os modelos investigados, o tempo de execução foi abaixo de 1 ms, aparecendo, portanto, como zero tanto na interface do Weka (regressores SVM, RF, MLP e LR) quanto nas bibliotecas em Octave utilizadas para implementar ELM e ESM.

Os resultados detalhados de R, RMSE(%) e tempo de treinamento (ms) são exibidos nas Tabelas 2, 3, 4, 5, 6 e 7. Na Tabela 2 são apresentados os resultados para regressão linear. Na Tabela 3 são apresentados os resultados para rede neural MLP, uma única camada escondida, para 2, 5, 10 e 20 neurônios na camada escondida. Na Tabela 4 são mostrados os resultados para SVM, *kernel* polinomial grau 1 (linear), 2, 3 e RBF. Na Tabela 5 são mostrados os resultados para rede neural ELM, uma única camada escondida com número de neurônios determinado automaticamente, *kernel* polinomial grau 1 (linear), 2, 3 e RBF. Na Tabela 6 são exibidos os resultados para redes neurais ESM (uma única camada escondida com 10 neurônios) e Deep-ESM (5 e 10 camadas escondidas, 10 neurônios por camada). Por fim, na Tabela 7 são apresentados os resultados para Random Forest, de 10 a 100 árvores, variando a cada 10. Os melhores resultados em média estão em destaque em vermelho.

Tabela 2 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para regressão linear.

Método de regressão	Configuração	R		RMSE (%)		Tempo de treinamento (ms)	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
Regressão linear	Única	0,8	0,1	51	23	1	3

Tabela 3 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para rede neural MLP, uma única camada escondida, para 2, 5, 10 e 20 neurônios na camada escondida. Os melhores resultados em média estão em destaque em vermelho.

Método de regressão	Configuração	R		RMSE (%)		Tempo de treinamento (ms)	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
MLP, uma camada escondida	2 neurônios	0,99	0,01	6	8	97	84
	5 neurônios	0,9999	0,0002	0,1	0,4	271	150
	10 neurônios	1	4,00E-07	0,09	0,03	989	360
	20 neurônios	1	7,00E-07	0,09	0,04	4127	1456

As Figuras 25, 26, 27, 28, 29 e 30; 31, 32, 33, 34, 35 e 36; e 37, 38, 39, 40, 41 e 42 mostram as distribuições estatísticas em boxplots do Coeficiente de Correlação (R), do RMSE(%), e do tempo de treinamento do modelo em milissegundos, respectivamente, para a regressão

Tabela 4 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para SVM, *kernel* polinomial grau 1 (linear), 2, 3 e RBF. Os melhores resultados em média estão em destaque em vermelho.

Método de regressão	Configuração	R		RMSE (%)		Tempo de treinamento (ms)	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
SVM	kernel polinomial, $p = 1$	0,89	0,07	42	21	1622	703
	kernel polinomial, $p = 2$	1	4,00E-05	0,6	0,2	1518	599
	kernel polinomial, $p = 3$	1	4,00E-05	0,6	0,2	899	351
	kernel RBF	1	9,00E-05	0,7	0,3	490	124

Tabela 5 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para rede neural ELM, uma única camada escondida, número de neurônios na camada escondida determinado automaticamente, *kernel* polinomial grau 1 (linear), 2, 3 e RBF. Os melhores resultados em média estão em destaque em vermelho.

Método de regressão	Configuração	R		RMSE (%)		Tempo de treinamento (ms)	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
ELM	kernel linear	0,999	0,0001	0,33	0,03	235	16
	kernel polinomial, grau 2	1	2,00E-15	2,90E-07	1,90E-07	523	660
	kernel polinomial, grau 3	1	3,30E-16	1,80E-09	2,10E-09	264	38
	Kernel RBF	0,6598	0,0004	37,85	0,06	697	188

Tabela 6 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para redes neurais ESM e Deep-ESM. Os melhores resultados em média estão em destaque em vermelho.

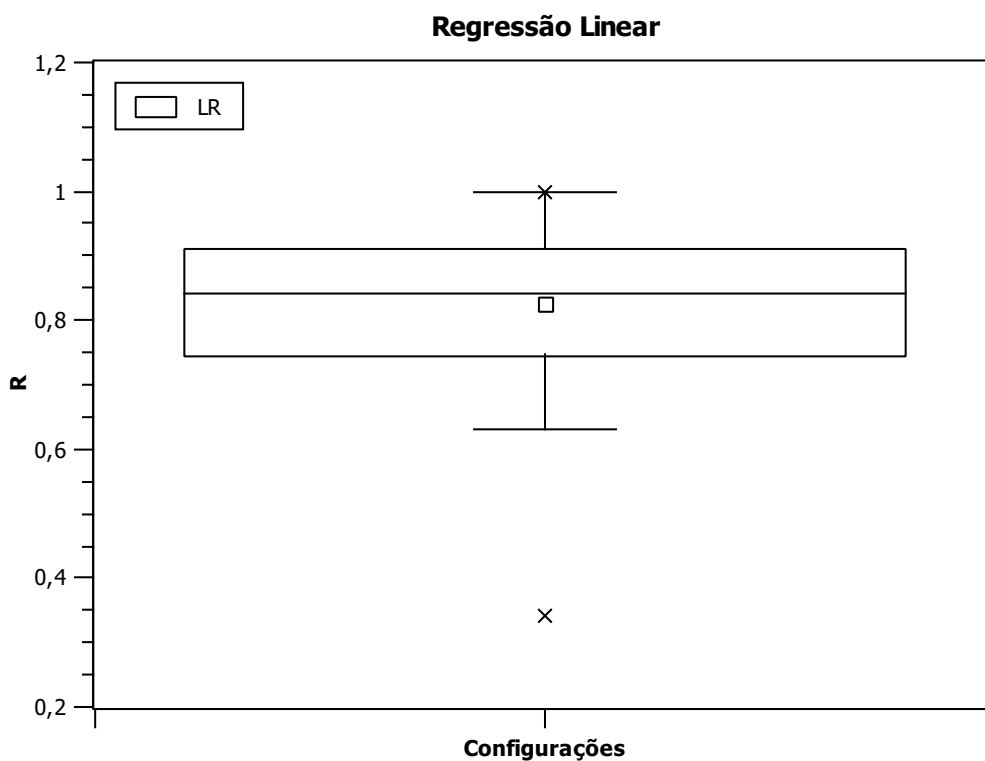
Método de regressão	Configuração	R		RMSE (%)		Tempo de treinamento (ms)	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
Deep-ESM	1 camada escondida, $p = 0,1$	0,5	0,3	7,58E+07	3,22E+09	152	54
	1 camada escondida, $p = 0,5$	0,5	0,3	9,19E+07	4,16E+09	212	30
	1 camada escondida, $p = 0,9$	0,5	0,3	2,80E+07	1,05E+09	206	23
	2 camadas escondidas, $p = 0,1$	0,5	0,3	3,35E+07	1,03E+09	2028	270
	2 camadas escondidas, $p = 0,5$	0,5	0,3	2,02E+07	4,16E+08	1868	222
	2 camadas escondidas, $p = 0,9$	0,5	0,3	2,93E+07	1,09E+09	3101	227
	5 camadas escondidas, $p = 0,1$	0,5	0,3	4,69E+07	1,43E+09	4676	668
	5 camadas escondidas, $p = 0,5$	0,5	0,3	2,21E+07	8,06E+08	7279	51517
	5 camadas escondidas, $p = 0,9$	0,5	0,3	7,47E+07	3,23E+09	7287	51512
	10 camadas escondidas, $p = 0,1$	0,4	0,3	3,84E+07	6,34E+08	11285	51643
	10 camadas escondidas, $p = 0,5$	0,5	0,3	3,25E+07	1,04E+09	13212	159842
	10 camadas escondidas, $p = 0,9$	0,5	0,3	3,64E+07	1,05E+09	13452	159844

Tabela 7 – Resultados de R, RMSE(%) e tempo de treinamento (ms) para Random Forest, de 10 a 100 árvores, variando a cada 10. Os melhores resultados em média estão em destaque em vermelho.

Método de regressão	Configuração	R		RMSE (%)		Tempo de treinamento (ms)	
		Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
Random Forest	10 árvores	0,9999	0,0001	0,6	0,6	30	9
	20 árvores	0,9999	0,0001	0,6	0,5	60	21
	30 árvores	0,9999	0,0001	0,5	0,5	101	23
	40 árvores	0,9999	0,0001	0,5	0,5	114	21
	50 árvores	0,9999	0,0001	0,5	0,5	143	26
	60 árvores	0,9999	0,0001	0,5	0,5	175	33
	70 árvores	0,9999	0,0001	0,5	0,5	200	36
	80 árvores	0,9999	0,0001	0,5	0,5	231	43
	90 árvores	0,9999	0,0001	0,5	0,5	271	50
	100 árvores	0,9999	0,0001	0,5	0,5	305	61

linear, redes neurais artificiais perceptron multicamadas (MLP), máquinas de aprendizado extremo (ELM), máquinas de estado de eco de camada única (ESM) e profundas (Deep ESM), máquinas de vetor de suporte (SVM) e Random Forests, nesta ordem.

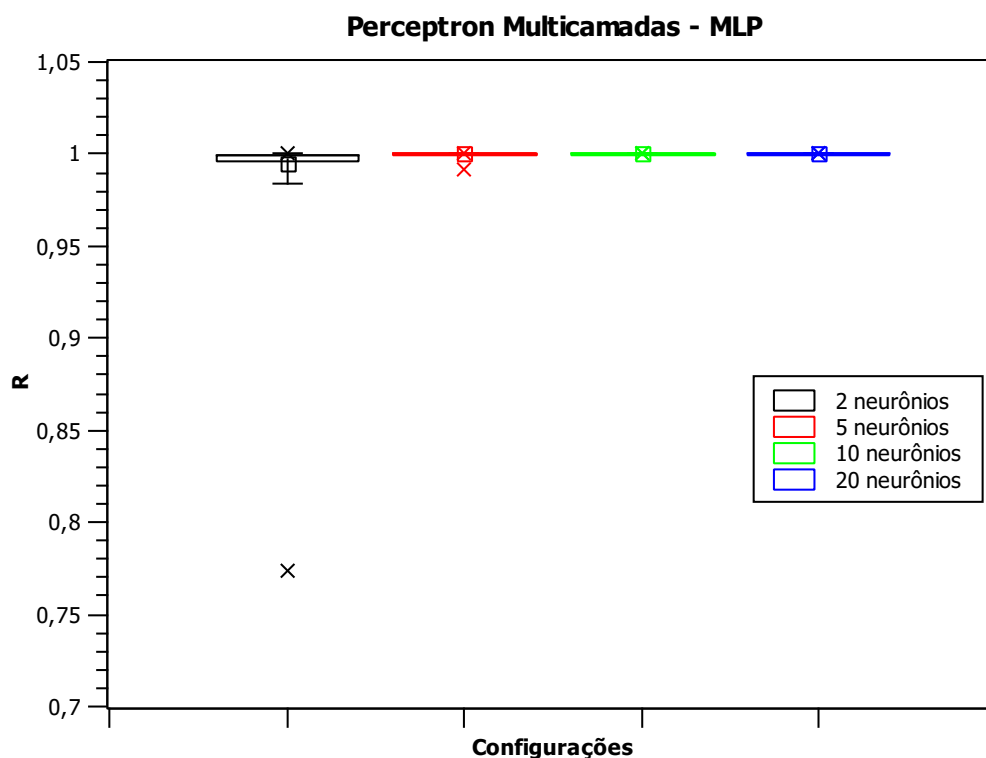
Figura 25 – Coeficiente de correlação R para a regressão linear.



Fonte: A Autora

A Tabela 2 mostra que os resultados com regressão linear são razoáveis com relação ao índice de correlação R, com média 0,8 e desvio padrão 0,1, e que o aprendizado é muito rápido, com tempo de treinamento médio de 1ms e desvio padrão de 3ms. Contudo, o RMSE (%) é alto: 51 em média com desvio padrão de 23. O gráfico boxplot da Figura 25 mostra

Figura 26 – Coeficiente de correlação R para redes neurais MLP com uma única camada escondida, com teste de 2, 5, 10 e 20 neurônios.

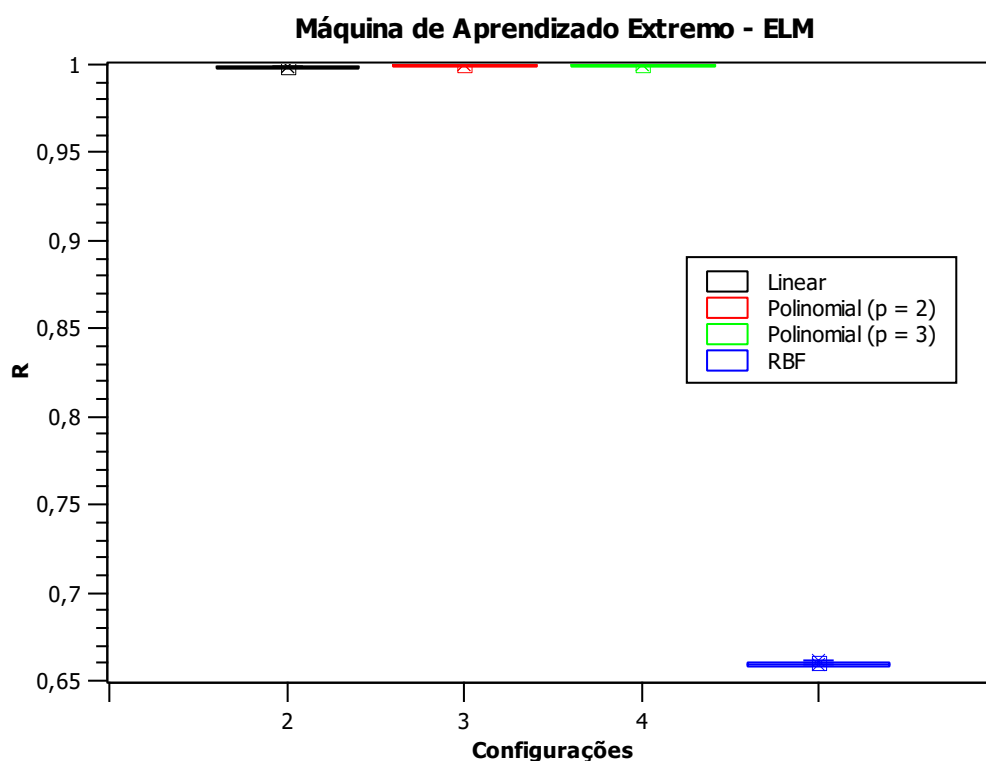


Fonte: A Autora

a distribuição do índice de correlação R, com metade dos resultados (a caixa) variando de 0,75 a 0,9. Considerando a distribuição do bigode do gráfico, a variação é maior: de 0,65 a 1,00. No entanto, existe um ponto fora que mostra que pelo menos um resultado de R foi de apenas 0,35. Assim, o índice de correlação R varia consideravelmente. A Figura 31 mostra o boxplot da distribuição do RMSE (%), com concentração em torno de 50%, o que é um erro consideravelmente grande. Existe ainda um ponto fora próximo de 200, o que mostra que pelo menos um resultado possui um erro de quase 200%. A Figura 37 mostra o boxplot da distribuição do tempo de treinamento, evidenciando que a regressão linear é um método de regressão muito rápido, com a maioria absoluta dos resultados concentrados em menos de 1ms, apesar de um ponto fora próximo de 16ms, o que ainda é bastante baixo.

Na Tabela 3 são apresentados os resultados para redes neurais perceptron multicamadas (MLP), com 2, 5, 10 e 20 neurônios na camada escondida. Os resultados se mostraram já interessantes com 2 neurônios na camada escondida: o índice de correlação R se mostrou bastante alto, com média de 0,99, e estável, pois o desvio padrão foi de somente 0,01. O RMSE (%) também se mostrou razoavelmente baixo: média de 6% com desvio padrão de 8, mas ainda um pouco acima do limite que consideramos ótimo neste trabalho, de 5%. O tempo

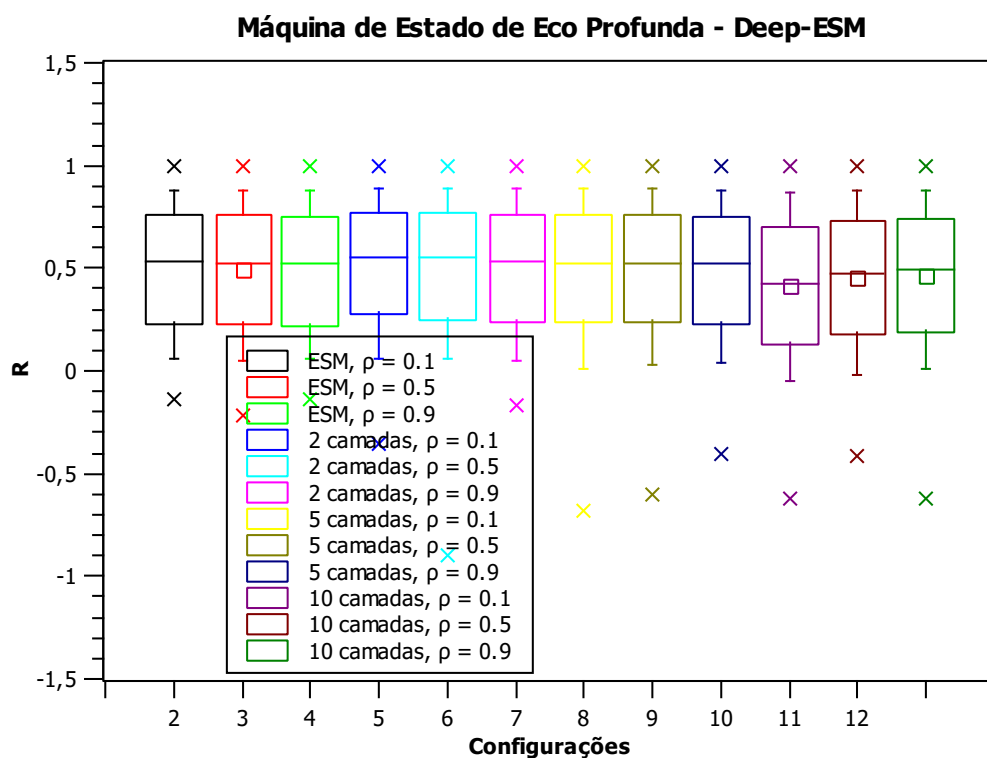
Figura 27 – Coeficiente de correlação R para redes neurais ELM com uma única camada escondida, número de neurônios determinado automaticamente, para *kernels* polinomiais de grau 1 a 3 e RBF.



Fonte: A Autora

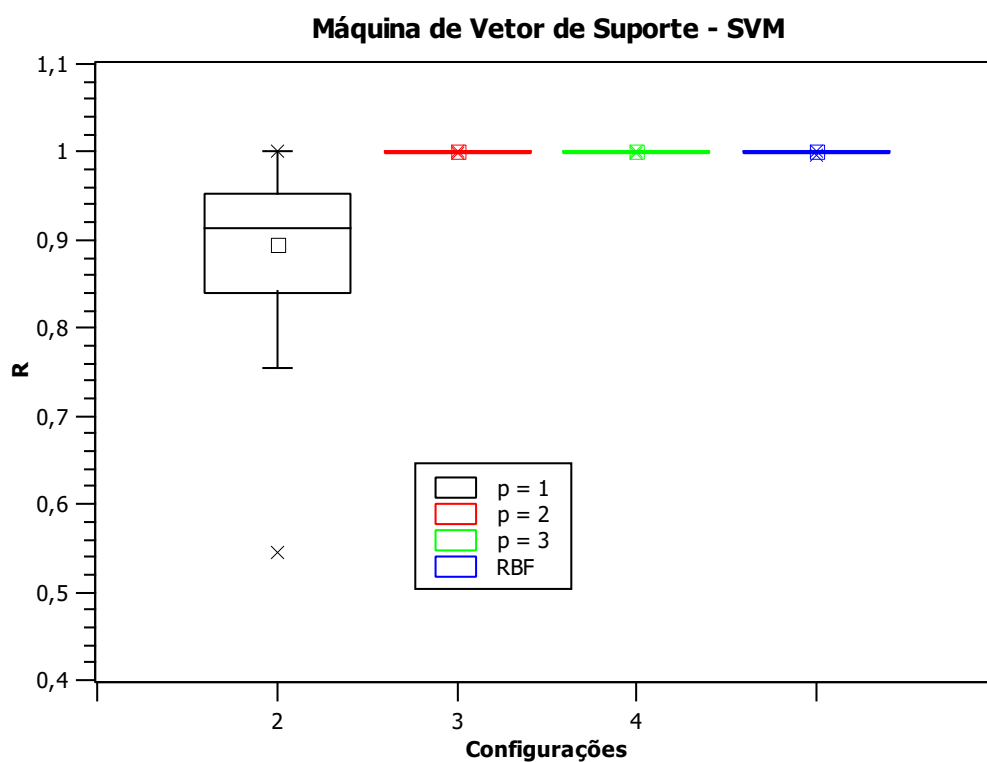
de treinamento pode ser considerado baixo: 97ms com desvio padrão de 84ms. A Figura 26 mostra o boxplot da distribuição do índice de correlação R, mostrando que os resultados de R para as quatro configurações estudadas são muito similares e consideravelmente altos: todas as distribuições são muito concentradas em 1,00. Esse comportamento é um pouco diferente apenas para configuração com 2 neurônios na camada escondida, com um ponto fora em 0,77, mas no geral também essa configuração se assemelha com as outras três. Já o comportamento para o RMSE (%) é um tanto diferente, conforme boxplot da Figura 32: embora a configuração com 2 neurônios na camada escondida se concentre em torno de 5%, o bigode da distribuição atinge 20%. Também há um ponto fora em 65%. Já as configurações com 5, 10 e 20 neurônios na camada escondida são muito similares entre si, exceto pelo fato de que a configuração com 5 neurônios tem um ponto fora em aproximadamente 15%. As configurações com 10 e 20 neurônios podem ser consideradas equivalentes. A Figura 38 mostra a distribuição do tempo de treinamento em ms. As configurações com 2 e 5 neurônios na camada escondida são sensivelmente mais rápidas, enquanto a configuração com 20 neurônios é não somente de treinamento mais lento em mediana, mas também esse tempo de treinamento se espalha

Figura 28 – Coeficiente de correlação R para redes neurais ESM e Deep-ESM com 2, 5 e 10 camadas escondidas, 10 neurônios por camada.



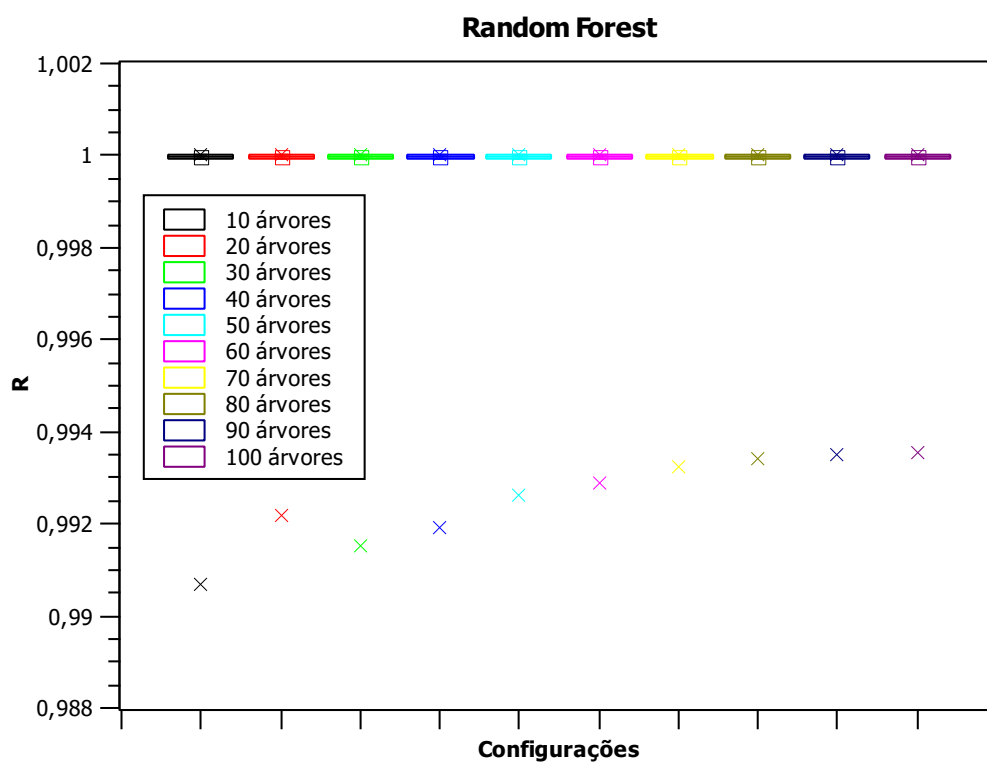
Fonte: A Autora

Figura 29 – Coeficiente de correlação R para SVM, *kernels* polinomiais de grau 1 a 3 e RBF.



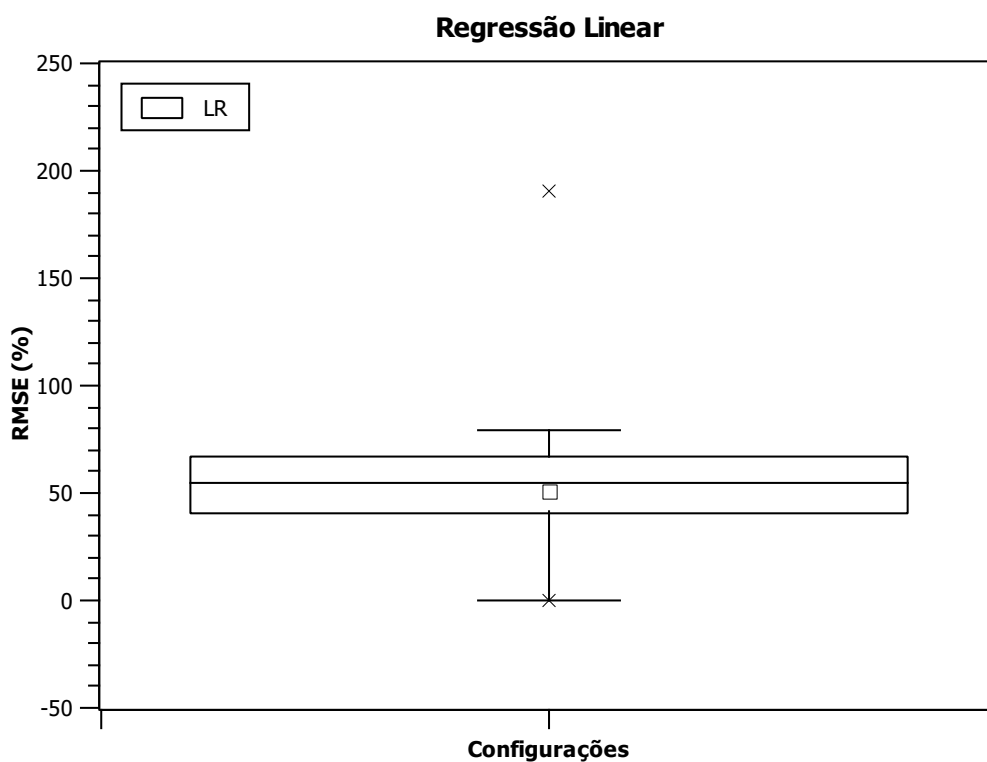
Fonte: A Autora

Figura 30 – Coeficiente de correlação R para: Random Forest, de 10 a 100 árvores, variando de 10 em 10.



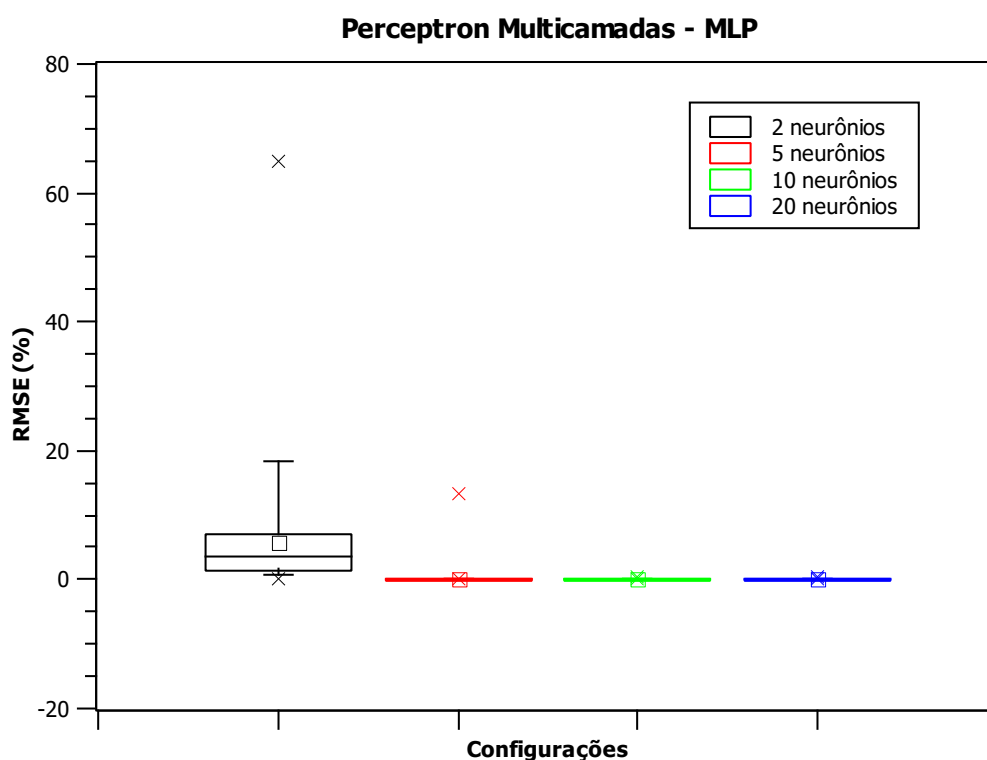
Fonte: A Autora

Figura 31 – RMSE(%) para a regressão linear.



Fonte: A Autora

Figura 32 – RMSE(%) para redes neurais MLP com uma única camada escondida, com teste de 2, 5, 10 e 20 neurônios.

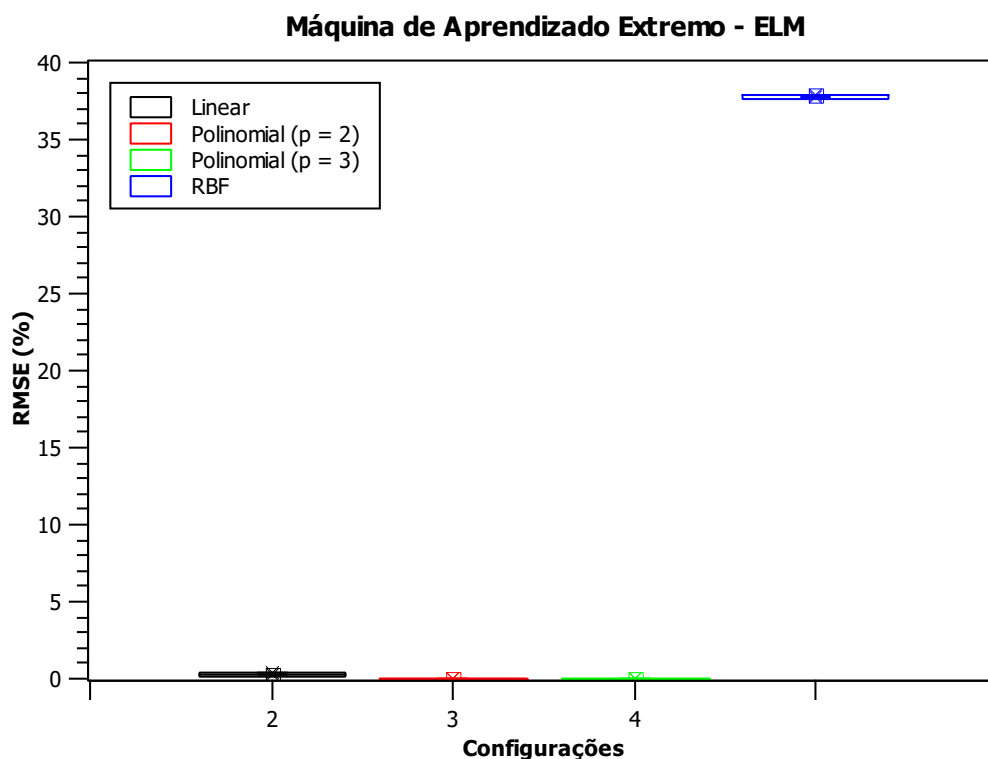


Fonte: A Autora

mais, podendo variar de 1000ms a quase 10000ms, ou seja, de 1 a 10 segundos. Apesar de esse tempo não ser crítico, pode-se considerar como a melhor configuração aquela com 10 neurônios na camada escondida, por ter um alto índice de correlação R, um baixo RMSE (%), e consumir menos memória (menos neurônios) e ter treinamento mais rápido.

Na Tabela 5 são mostrados os resultados para máquinas de aprendizado extremo (ELM), com *kernel* linear, polinomial graus 2 e 3, e de função de base radial (RBF). Exceto pelo *kernel* RBF, todas as configurações apresentaram índice de correlação R médio altíssimo (de 0,999 a 1,000) e desvio padrão muito baixo. Da mesma forma, todas as configurações polinomiais alcançaram erros RMSE (%) muito abaixo dos 5% considerados neste trabalho como padrão. O tempo de treinamento médio também variou, nessas configurações polinomiais, de 200ms a 500ms. A Figura 27 mostra o boxplot da distribuição do índice de correlação R, com resultados praticamente similares e consideravelmente altos para os *kernels* linear, polinomial grau 2 e 3, tocando 1,00 e com dispersão quase nula. O *kernel* RBF também se mostrou estável, mas com distribuição de R concentrada em 0,66. A Figura 33 apresenta um cenário semelhante: erros RMSE (%) bastante concentrados entre 0,00% e 0,20%, exceto para o *kernel* RBF, novamente. Na Figura 39 são apresentadas as distribuições do tempo de treinamento

Figura 33 – RMSE(%) para redes neurais ELM com uma única camada escondida, número de neurônios determinado automaticamente, para *kernels* polinomiais de grau 1 a 3 e RBF.

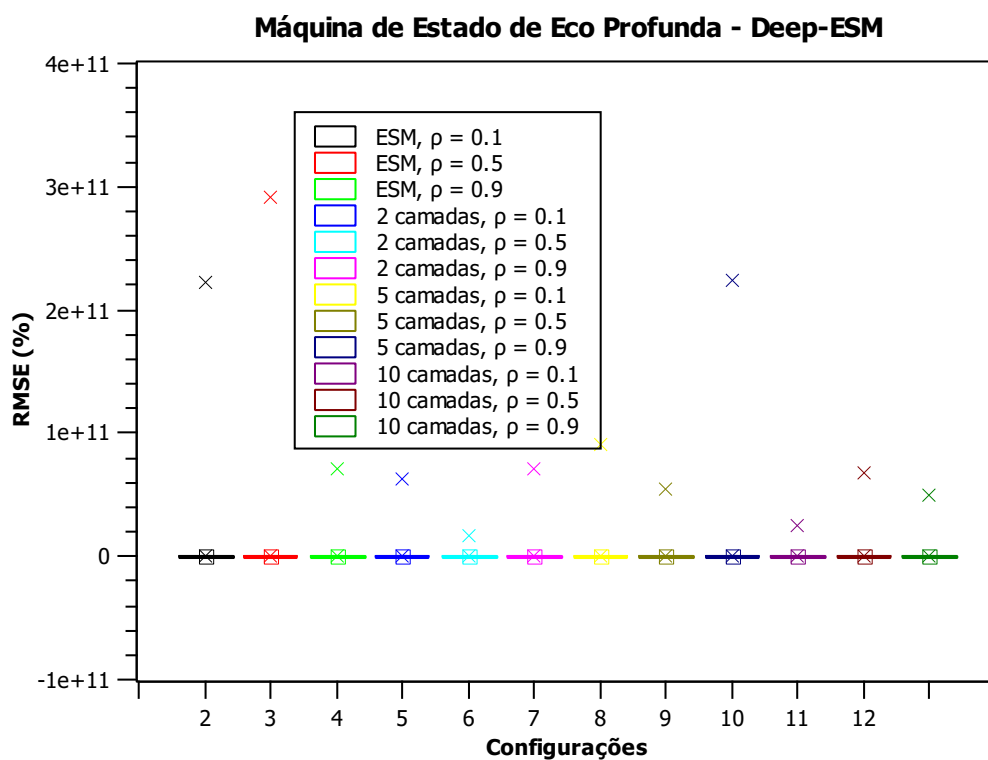


Fonte: A Autora

das arquiteturas utilizadas. Os comportamentos dos *kernels* linear e polinomial grau 2 e 3 poderiam ser considerados similares se a configuração de *kernel* polinomial grau 2 não tivesse um espalhamento inesperado, com bigode atingindo 2400ms e ponto fora de 3500ms. Novamente os resultados com RBF não se mostraram interessantes. Assim, o ELM linear pode ser considerado a configuração ELM mais adequada à resolução do problema.

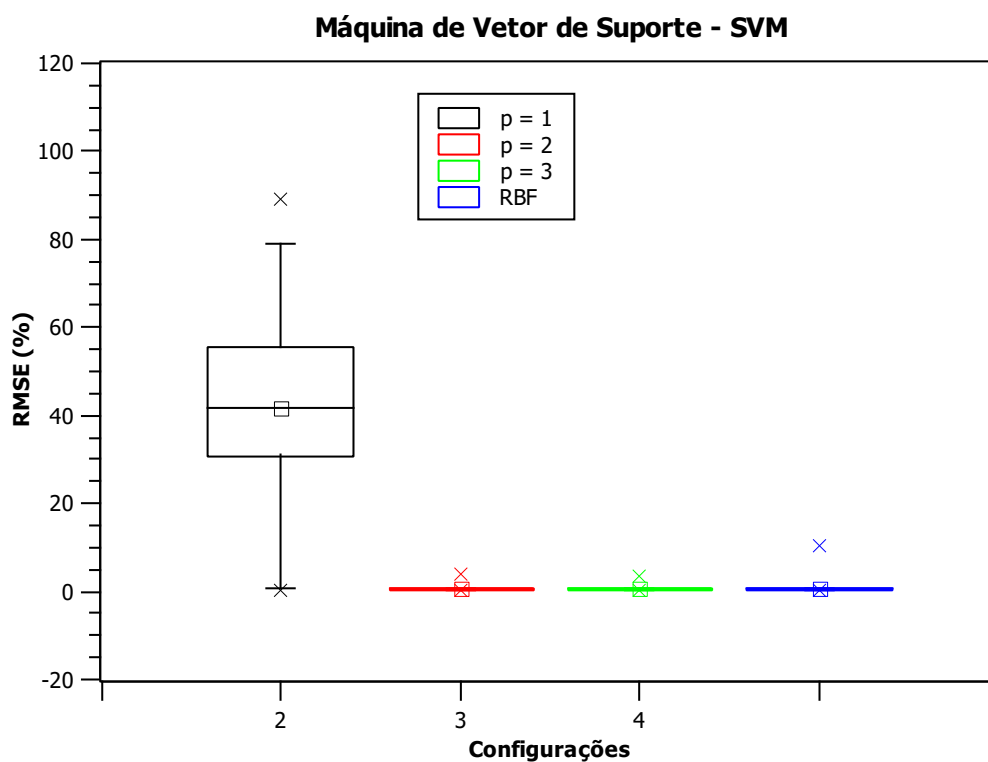
A Tabela 4 mostra os resultados para máquinas de vetor de suporte (SVM), com *kernel* linear (ou polinomial grau 1), polinomial graus 2 e 3 e RBF. Os resultados com *kernel* linear lembram um pouco os resultados com regressão linear, embora um pouco melhores e mais estáveis (baixo desvio padrão): índice de correlação R de 0,87, com desvio padrão de 0,07, e RMSE (%) de 42% com desvio padrão de 21%. O tempo de treinamento para o *kernel* linear, contudo, foi muito variável e relativamente longo: 1622ms com desvio padrão de 703ms. O *kernel* linear foi a configuração de SVM mais lenta de todas as avaliadas. Já os resultados com *kernels* polinomiais graus 2 e 3 e RBF foram considerados muito bons: índices de correlação R baixíssimos, praticamente 1,00, com desvios padrão muito baixos ($4,00 \times 10^{-5}$ para *kernels* polinomiais graus 2 e 3, e $9,00 \times 10^{-5}$ para o *kernel* RBF). Os resultados do RMSE (%) também

Figura 34 – RMSE(%) para redes neurais ESM e Deep-ESM com 2, 5 e 10 camadas escondidas, 10 neurônios por camada.



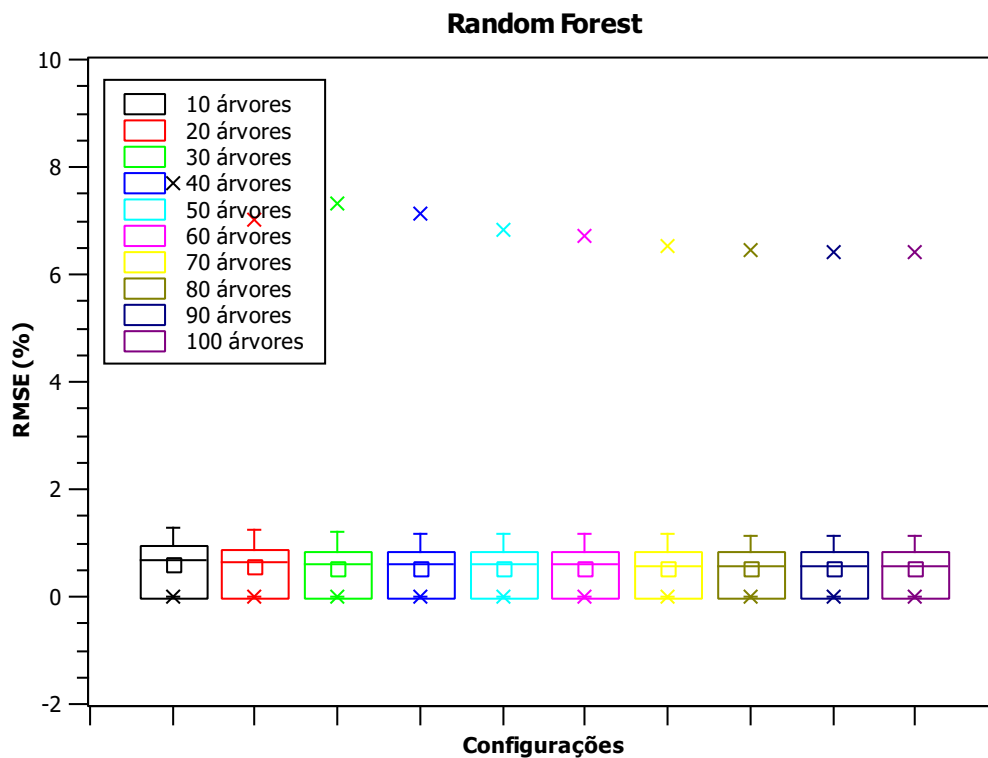
Fonte: A Autora

Figura 35 – RMSE(%) para SVM, *kernels* polinomiais de grau 1 a 3 e RBF.



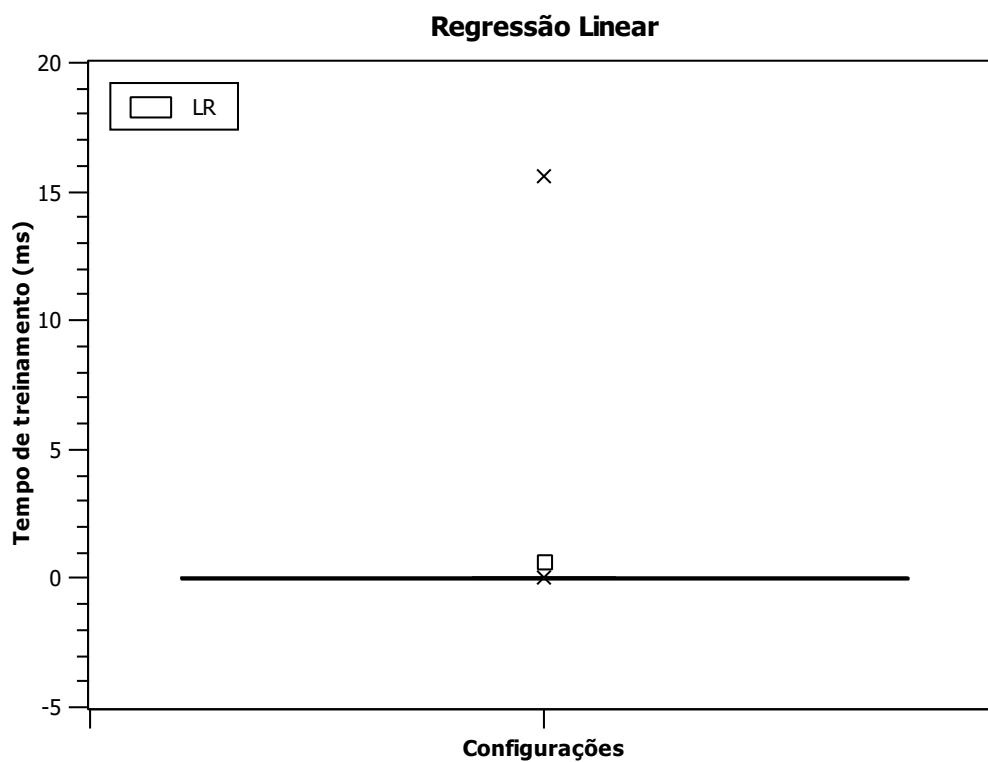
Fonte: A Autora

Figura 36 – RMSE(%) para: Random Forests, de 10 a 100 árvores, variando de 10 em 10.



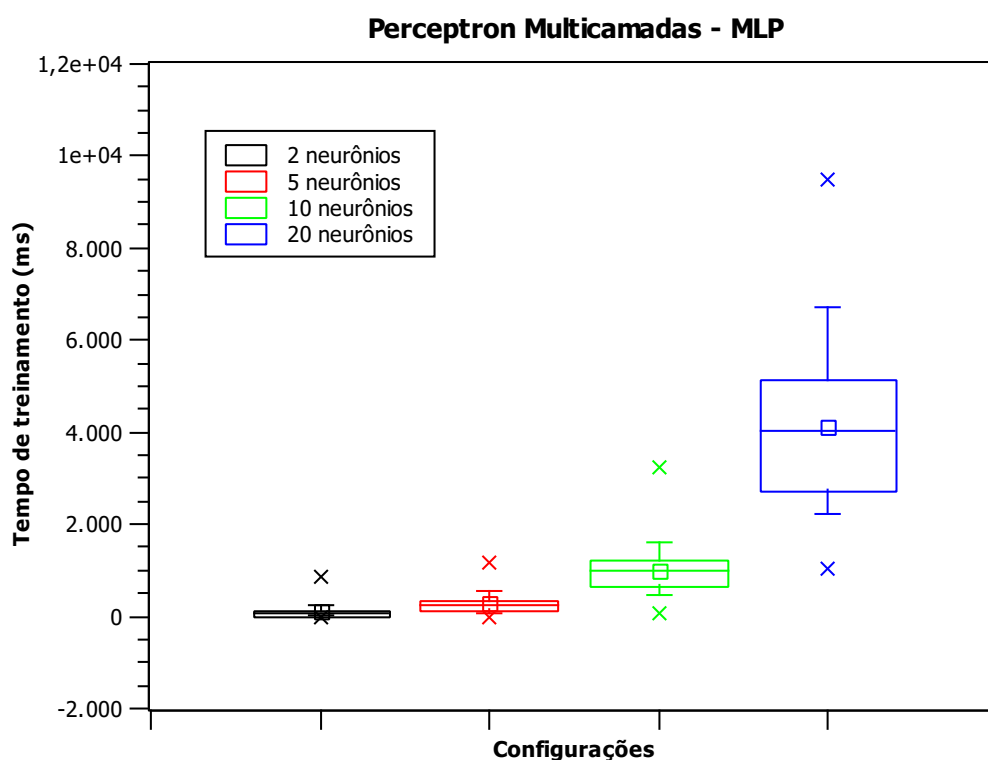
Fonte: A Autora

Figura 37 – Tempo de treinamento (ms) para a regressão linear.



Fonte: A Autora

Figura 38 – Tempo de treinamento (ms) para redes neurais MLP com uma única camada escondida, com teste de 2, 5, 10 e 20 neurônios.

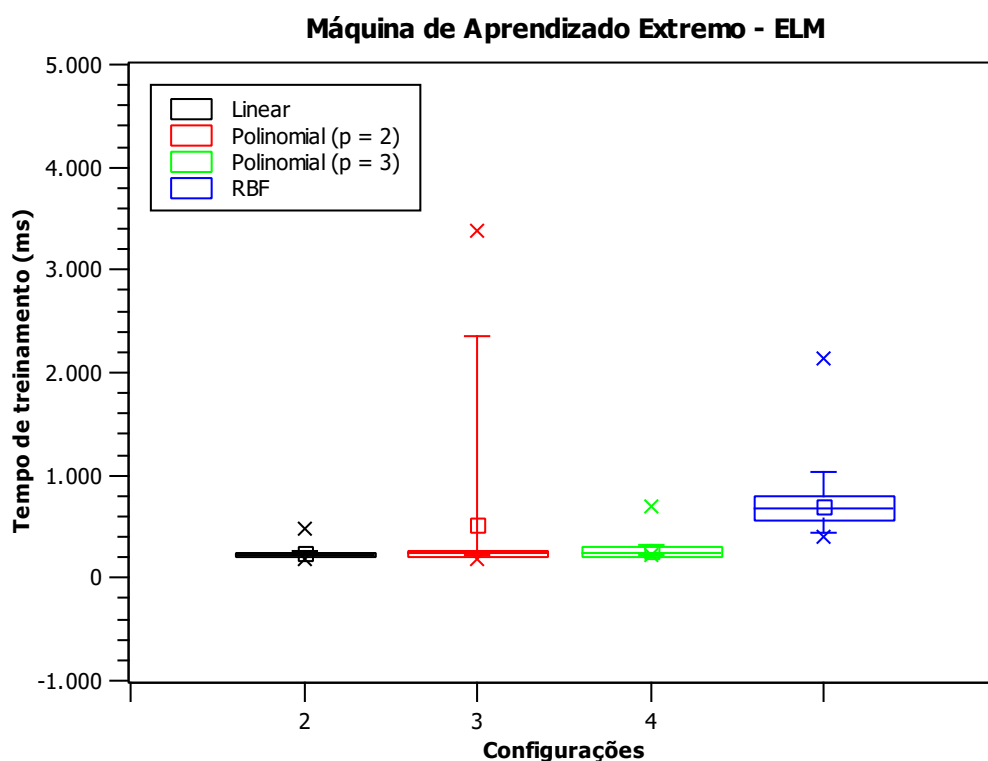


Fonte: A Autora

ficaram bem abaixo dos 5%: 0,6% com desvio padrão 0,2% para os *kernels* polinomiais de grau 2 e 3, e 0,7% com desvio padrão 0,3% para o *kernel* RBF. O menor tempo de treinamento médio foi obtido com o *kernel* RBF: 490ms com desvio padrão de 124ms. A Figura 29 mostra o boxplot da distribuição do índice de correlação *R*. Pode-se notar que há uma concentração entre 0,84 e 0,96 para o *kernel* linear (polinomial grau 1). Também fica evidente que os *kernels* polinomiais de grau 2 e 3 e RBF atingiram resultados excelentes: concentração em 1,00 com desvio praticamente nulo. Do ponto de vista estatístico, os resultados obtidos para o índice de correlação *R* para os *kernels* polinomiais de grau 2 e 3 e RBF são equivalentes. O mesmo se dá para o erro RMSE (%), conforme o boxplot da Figura 35, com leve desvantagem para o *kernel* RBF, que apresenta um ponto fora em 10%. Analisando o boxplot da Figura 41, apesar da equivalência estatística entre os métodos com *kernels* polinomiais de grau 2 e 3 e RBF, percebe-se claramente que o tempo de treinamento é mais baixo e mais estável para o *kernel* RBF, sendo esta a configuração de SVM mais adequada à resolução do problema.

Na Tabela 6 são exibidos os resultados para máquinas de estado de eco de camada única e profundas (ESM e Deep ESM, respectivamente), com 1, 2, 5 e 10 camadas escondidas. As ESM e Deep ESM são redes neurais treinadas da mesma forma que as ELM, mas seus

Figura 39 – Tempo de treinamento (ms) para redes neurais ELM com uma única camada escondida, número de neurônios determinado automaticamente, para *kernels* polinomiais de grau 1 a 3 e RBF.

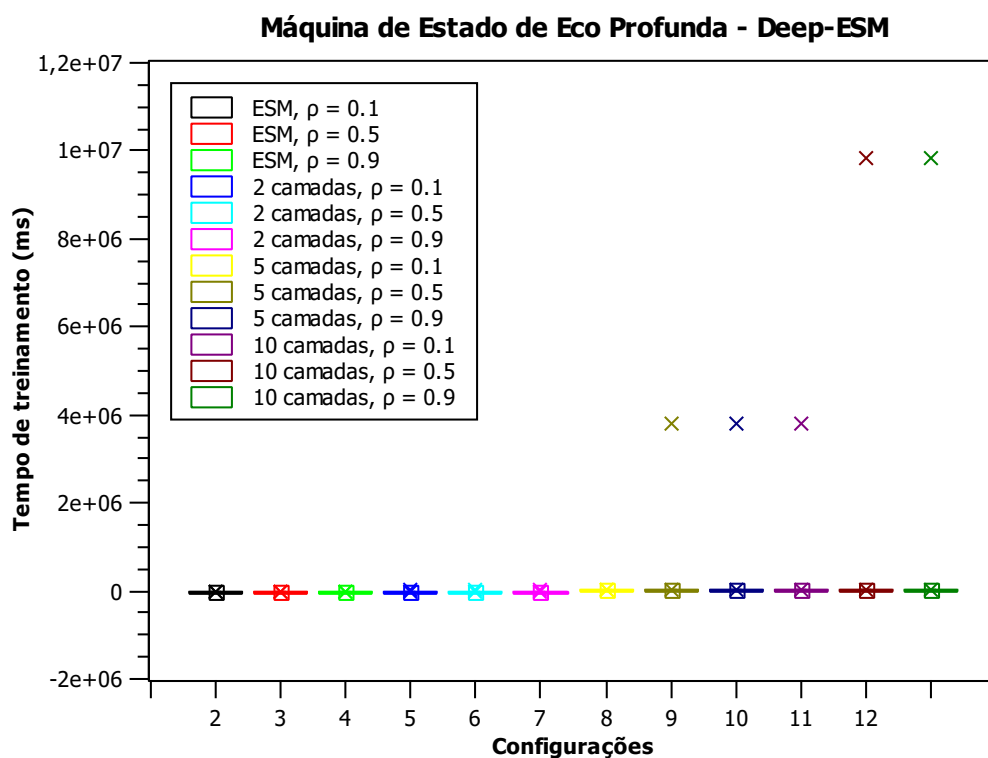


Fonte: A Autora

neurônios são diferentes: são chamados de *reservatórios* e são influenciados não somente pelas entradas, mas também por valores passados das saídas (realimentação). Assim, esperava-se que seus resultados fossem iguais ou melhores àqueles obtidos com as ELM, mas foram muito inferiores, tornando essa abordagem realmente inadequada ao problema: os valores médios do índice de correlação R estão praticamente estagnados em 0,5, com desvio 0,3, enquanto o erro RMSE (%) é da ordem de 10^7 com desvios ainda maiores, da ordem de 10^9 , enquanto o tempo de treinamento aumenta muito com a inclusão de mais camadas, como também mostra o boxplot da Figura 40. Os boxplots das Figuras 28 e 34 confirmam não somente a inadequação da ESM e da Deep ESM à solução do problema, como também indicam que esses métodos, quanto ao índice de correlação R e o erro RMSE (%), são estatisticamente equivalentes.

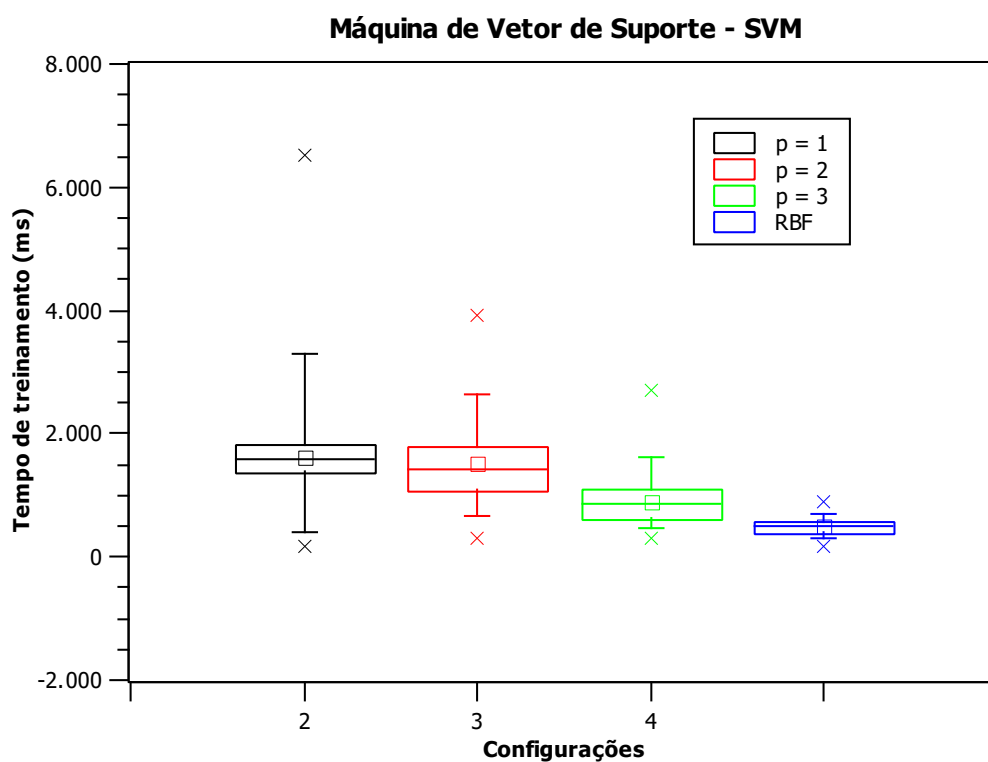
A Tabela 7 ilustra os resultados para Random Forest, para configurações de 10 a 100 árvores, variando de 10 em 10. Os resultados são muito bons e se repetem para todas as configurações testadas: índice de correlação R médio de 0,9999 com desvio padrão 0,0001, ou seja, praticamente 1,00 em todos os experimentos; RMSE (%) médio de 0,6% com desvio

Figura 40 – Tempo de treinamento (ms) para redes neurais ESM e Deep-ESM com 2, 5 e 10 camadas escondidas, 10 neurônios por camada.



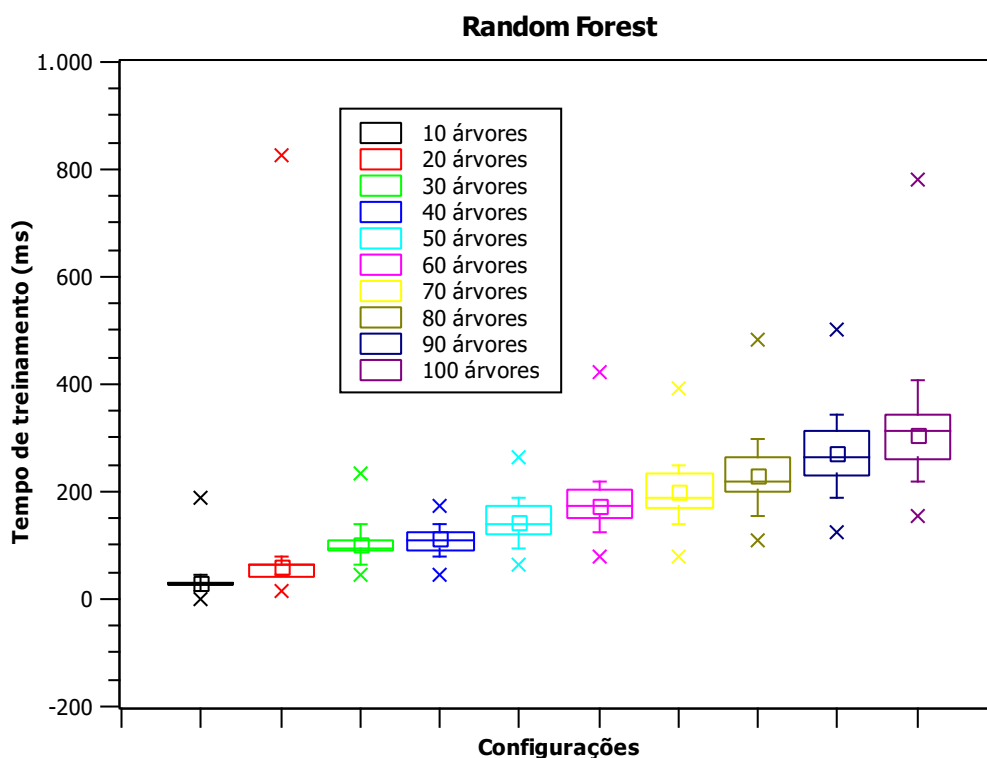
Fonte: A Autora

Figura 41 – Tempo de treinamento (ms) para SVM, *kernels* polinomiais de grau 1 a 3 e RBF.



Fonte: A Autora

Figura 42 – Tempo de treinamento (ms) para: Random Forests, de 10 a 100 árvores, variando de 10 em 10.



Fonte: A Autora

padrão 0,6% para 10 árvores, 0,6% com desvio padrão 0,5% para 20 árvores, e 0,5% com desvio 0,5% de 30 árvores em diante. O boxplot da Figura 30 mostra que, para o índice de correlação R, todas as configurações testadas são equivalentes e podem ser muito boas, com resultados praticamente concentrados em 1,00 e com pontos fora ainda aceitáveis, entre 0,990 e 0,994. O boxplot da Figura 36 mostra a distribuição dos resultados do erro RMSE (%), mostrando que os resultados estão concentrados entre 0 e 1%, portanto abaixo do limite de 5% que foi adotado neste trabalho, com pontos fora entre 7% e 8%. Os resultados também mostram que as configurações de Random Forest são estatisticamente equivalentes. Assim, o tempo de treinamento pode ser considerado como critério de desempate. A Tabela 7 mostra que a configuração com 10 árvores possui o treinamento mais rápido: média de 30ms com desvio padrão de 9ms. O boxplot da Figura 42 mostra que o tempo de treinamento aumenta linearmente com o aumento do número de árvores, assim como os tempos passam a se dispersar mais. Na configuração com 10 árvores, os tempos de treinamento estão concentrados em torno de 25ms com muito pouca dispersão, com um ponto fora em 200ms. Desta forma, a configuração de Random Forest com 10 árvores é a mais adequada para o problema por ser a de mais rápido treinamento, embora todas as configurações testadas sejam estatisticamente

equivalentes e igualmente boas de acordo com o índice de correlação R e o erro RMSE (%).

5.1.2 Análise de fatores mais relevantes

Para analisar os fatores mais relevantes para a predição, foram utilizados cinco métodos de busca heurística bioinspirados para seleção de atributos. Esses métodos utilizaram como função objetivo um regressor baseado na árvore de decisão Random Tree, avaliada usando validação cruzada com 10 *folds*. Cada método retornou os atributos mais relevantes e seu grau de relevância em percentual, de 0% a 100%. Um atributo foi considerado relevante quando seu grau de relevância foi diferente de 0%, tipicamente no mínimo 10% nos experimentos realizados neste trabalho. Esses cinco métodos formaram um comitê de especialistas artificiais, sendo o resultado final de relevância decidido pela votação. Assim, um atributo é considerado relevante se ele foi considerado relevante pela maioria simples dos membros do comitê de especialistas artificiais. Todos os algoritmos meta-heurísticos foram rodados considerando populações iniciais de candidatos à solução de 20 indivíduos e um total de 500 iterações ou gerações. As configurações utilizadas foram as que seguem:

- Algoritmo Genético (ou Busca Genética, *Genetic Search*, GS): probabilidade de cruzamento de 0,6, probabilidade de mutação de 0,033, frequência de comunicação de 20, 20 indivíduos, 500 gerações;
- Algoritmo Genético Modificado (Busca Evolucionária, *Evolutionary Search*, ES): probabilidade de cruzamento de 0,6, probabilidade de mutação de 0,1, mutação do tipo *bit-flip*, substituição geracional (os filhos substituem os pais na geração seguinte), seleção de indivíduos por torneio, frequência de comunicação de 20, 20 indivíduos, 500 gerações;
- Otimização por Enxame de Partículas (*Particle Swarm Optimization Search*, *PSO Search*, PSO): peso individual de 0,34, fator de inércia de 0,33, peso social de 0,33, probabilidade de mutação de 0,01, frequência de comunicação de 20, 20 partículas, 500 iterações;
- Colônia de Abelhas Artificiais (ou Busca de Abelhas, *Bee Search*, BS): coeficiente caótico de 4,0, tipo caótico mapa logístico, probabilidade de mutação de 0,01, mutação do tipo

bit-flip, raio de humidade de 0,98, raio de mutação de 0,8, frequência de comunicação de 20, 20 abelhas, 500 iterações;

- Colônia de Formigas Artificiais (ou Busca de Formigas, *Ant Search*, AS): coeficiente caótico de 4,0, tipo caótico mapa logístico, evaporação de 0,9, heurística de 0,7, probabilidade de mutação de 0,01, mutação do tipo *bit-flip*, méritos como objetivo, feromônio de 2,0, frequência de comunicação de 20, 20 formigas, 500 iterações.

Os resultados foram gerados considerando as bases iniciais de dados distribuídos por bairro, uma base para cada um dos seis bimestres, para os anos de 2014, 2015 e 2016, tendo sido utilizados dados de 2013 a 2016. Cada base possui, portanto, 94 instâncias, associadas aos 94 bairros e distritos da Cidade do Recife, com 44 atributos, que incluem latitude e longitude do ponto central do bairro, número de casos por bimestre e temperaturas, pluviometrias e velocidades dos ventos por mês, para cada um dos seis bimestres do ano. As Tabelas 8, 9 e 10 apresentam os resultados da análise automática de fatores mais relevantes para os bimestres dos anos de 2014, 2015 e 2016, nesta ordem.

A Tabela 8 mostra os resultados de predição para os seis bimestres de 2014 considerando como entrada o ano de 2013. Analisando o resultado do comitê de analistas artificiais, pode-se perceber que, para a predição do bimestre janeiro e fevereiro de 2014 (2014.1), foram considerados mais relevantes os seguintes fatores: posição geográfica (latitude e longitude); temperatura (t2b1) e velocidade dos ventos (v2b1) em fevereiro de 2013; número de casos de arboviroses no bimestre março e abril de 2013 (cb2); pluviometria em abril de 2013 (p2b2); velocidade dos ventos em março (v1b2) e maio (v1b3) de 2013; número de casos em julho e agosto de 2013 (cb4); temperatura em agosto de 2013 (t2b4); velocidade dos ventos em julho (v1b4) e em agosto (v2b4) de 2013; número de casos em setembro e outubro de 2013 (cb5); temperatura (t2b5) e velocidade dos ventos (v2b5) em outubro de 2013; número de casos em novembro e dezembro de 2013 (cb6) e temperatura (t2b6) em dezembro de 2013. É evidente a importância da posição geográfica e das velocidades, temperaturas e número de casos de março a dezembro de 2013 na predição do número de casos nos meses de janeiro e fevereiro de 2014.

Na predição do bimestre de março a abril de 2014 (2014.2), conforme Tabela 8, fica evidente a importância dos seguintes fatores: posição geográfica (longitude); temperaturas nos bimestres 2 (março a abril de 2013), 4 (julho a agosto de 2013) e 5 (setembro e outubro de 2013); pressão nos bimestres 2 a 5 (de março a outubro de 2013); velocidade dos ventos nos

Tabela 8 – Resultado da análise de fatores mais relevantes por meio dos algoritmos de Busca Genética, Busca Evolucionária, Otimização por Enxame de Partículas, Busca de Abelhas e Busca de Formigas, para 20 candidatos à solução evoluindo em 500 gerações, para o ano de 2014.

Método	2014.1	2014.2	2014.3	2014.4	2014.5	2014.6
Genetic Search (Algoritmo Genético)	latitude (10%), longitude (100%), t2b1 (50%), v2b1 (100%), cb2 (90%), p2b2 (10%), v1b2 (20%), v1b3 (100%), cb4 (100%), t2b4 (10%), v1b4 (60%), v2b4 (20%), cb5 (80%), t2b5 (90%), v2b5 (20%), cb6 (70%), t2b6 (80%)	longitude (30%), t2b2 (10%), p1b2 (10%), v2b2 (40%), p2b3 (20%), v1b3 (30%), v2b3 (10%), cb4 (90%), t2b4 (30%), p2b4 (10%), v2b4 (10%), cb5 (10%), t1b5 (30%), t2b5 (60%), p1b5 (20%), p2b5 (50%), v1b5 (40%), cb6 (100%), v2b6 (60%), cb1 (100%), v1b1 (80%)	latitude (30%), longitude (10%), cb3 (70%), t2b3 (100%), cb4 (70%), t2b4 (90%), v1b4 (20%), v2b4 (60%), p1b5 (10%), v1b5 (10%), v2b5 (20%), cb6 (100%), t1b6 (90%), t2b6 (30%), v1b6 (10%), cb1 (90%), t1b1 (10%), t2b1 (60%), v1b1 (30%), t2b1 (20%), cb2 (90%), t1b2 (70%), t2b2 (60%), v1b2 (60%), v2b2 (70%)	latitude (100%), longitude (20%), t2b4 (30%), p1b4 (80%), p2b4 (10%), t1b5 (80%), p1b5 (20%), v1b5 (20%), v2b5 (10%), p2b6 (10%), v2b6 (80%), cb1 (10%), t1b1 (30%), v1b1 (40%), t1b2 (10%), v1b2 (50%), v2b2 (50%), cb3 (90%), t1b3 (90%), t2b3 (10%), p1b3 (30%), v1b3 (20%), v2b3 (20%)	latitude (100%), t1b5 (40%), p1b5 (30%), v1b5 (20%), v2b6 (40%), cb1 (10%), t1b1 (20%), v1b1 (20%), t1b2 (10%), v1b2 (10%), v2b2 (10%), cb3 (100%), t1b3 (90%), p1b3 (10%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (30%), v1b4 (10%), v2b4 (100%)	latitude (90%), p2b6 (10%), v2b6 (10%), cb1 (10%), t1b2 (10%), cb3 (90%), t1b3 (80%), p1b3 (20%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (70%), v1b4 (10%), v2b4 (100%), cb5 (100%), t1b5 (40%), t2b5 (80%)
Evolutionary Search (Algoritmo Genético Modificado)	latitude (80%), longitude (100%), t2b1 (90%), v2b1 (100%), cb4 (100%), v1b3 (100%), cb4 (100%), t2b4 (80%), v2b4 (80%), cb5 (100%), t1b5 (80%), t2b5 (100%), cb6 (100%), t2b6 (30%)	longitude (100%), p1b2 (40%), v2b2 (100%), p2b3 (30%), v1b3 (100%), v2b3 (100%), cb4 (100%), p2b4 (10%), t2b4 (10%), p1b5 (20%), v1b5 (100%), cb6 (100%), v2b6 (20%), cb1 (100%), v1b1 (100%)	longitude (10%), cb3 (100%), t2b3 (100%), cb4 (40%), t2b4 (100%), v1b4 (10%), v2b4 (40%), t1b5 (30%), cb6 (100%), t1b6 (100%), t2b6 (50%), v1b6 (90%), cb1 (60%), t2b1 (100%), v2b1 (30%), cb2 (100%), t1b2 (70%), t2b2 (40%), v1b2 (100%), v2b2 (40%)	latitude (100%), longitude (100%), t2b4 (10%), p1b4 (80%), t1b5 (20%), p2b6 (10%), v2b6 (100%), t1b1 (90%), v1b1 (90%), t1b2 (90%), v1b2 (100%), v2b2 (20%), cb3 (100%), t1b3 (100%), p1b3 (10%), v1b3 (10%)	latitude (90%), t1b5 (30%), p1b5 (20%), v1b5 (10%), v2b6 (90%), v1b1 (80%), t1b2 (10%), p1b2 (10%), v1b2 (10%), v2b2 (80%), cb3 (100%), t1b3 (90%), p1b3 (10%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (30%), v1b4 (10%), v2b4 (100%)	latitude (90%), p2b6 (10%), v2b6 (10%), cb1 (10%), t1b2 (10%), cb3 (90%), t1b3 (80%), p1b3 (20%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (70%), v1b4 (10%), v2b4 (100%), cb5 (100%), t1b5 (40%), t2b5 (80%)
PSO Search (Otimização por Enxame de Partículas)	latitude (10%), longitude (100%), t2b1 (50%), v2b1 (100%), cb2 (90%), p2b2 (10%), v1b2 (20%), v1b3 (100%), cb4 (100%), t2b4 (10%), v1b4 (60%), v2b4 (20%), cb5 (80%), t2b5 (90%), v2b5 (20%), cb6 (70%), t2b6 (80%)	longitude (40%), t2b2 (10%), p1b2 (20%), v2b2 (40%), p2b3 (20%), v1b3 (30%), v2b3 (10%), cb4 (90%), t2b4 (20%), p2b4 (10%), v2b4 (10%), cb5 (10%), t1b5 (20%), t2b5 (60%), p1b5 (20%), p2b5 (40%), v1b5 (50%), cb6 (100%), v2b6 (60%), cb1 (100%), v1b1 (80%)	latitude (30%), longitude (10%), cb3 (70%), t2b3 (100%), cb4 (70%), t2b4 (90%), v1b4 (20%), v2b4 (60%), p1b5 (10%), v1b5 (10%), v2b5 (20%), cb6 (100%), t1b6 (90%), t2b6 (30%), v1b6 (10%), cb1 (90%), t1b1 (10%), t2b1 (60%), v1b1 (30%), t2b1 (20%), cb2 (90%), t1b2 (70%), t2b2 (60%), v1b2 (60%), v2b2 (70%)	latitude (100%), longitude (30%), cb4 (10%), t2b4 (30%), p1b4 (80%), p2b4 (10%), cb5 (10%), t1b5 (80%), p1b5 (20%), p2b5 (10%), v1b5 (20%), v2b5 (10%), cb6 (100%), p1b6 (100%), p2b6 (10%), v2b6 (80%), cb1 (10%), t1b1 (40%), t2b1 (10%), p1b1 (10%), v1b1 (40%), v2b1 (10%), cb2 (100%), t1b2 (10%), v1b2 (50%), v2b2 (50%), cb3 (100%), t1b3 (20%), p1b3 (30%), p2b3 (10%), v1b3 (20%), v2b3 (20%)	latitude (100%), t1b5 (50%), p1b5 (30%), v1b5 (20%), v2b6 (50%), cb1 (10%), t1b1 (20%), v1b1 (30%), t1b2 (10%), v1b2 (20%), v2b2 (40%), cb3 (90%), t1b3 (90%), p1b3 (30%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (40%), v1b4 (10%), v2b4 (100%)	latitude (90%), p2b6 (10%), v2b6 (10%), cb1 (10%), t1b2 (10%), cb3 (90%), t1b3 (80%), p1b3 (20%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (70%), v1b4 (10%), v2b4 (100%), cb5 (100%), t1b5 (40%), t2b5 (80%)
Bee Search (Colônia de Abelhas Artificiais)	longitude (60%), t2b1 (60%), v2b1 (60%), cb2 (90%), p2b2 (10%), v1b3 (90%), cb4 (100%), t2b4 (10%), v1b4 (60%), v2b4 (20%), cb5 (80%), t2b5 (90%), v1b5 (20%), v2b5 (10%), cb6 (70%), t2b6 (80%)	latitude (10%), longitude (20%), p1b2 (20%), v2b2 (10%), p2b3 (20%), v1b3 (20%), v2b3 (10%), cb4 (80%), t1b4 (20%), t2b4 (20%), p2b4 (10%), v2b4 (10%), cb5 (10%), t1b5 (20%), t2b5 (40%), p1b5 (20%), p2b5 (40%), v1b5 (40%), cb6 (100%), p2b6 (10%), v1b6 (10%), v2b6 (40%), cb1 (100%), v1b1 (50%)	latitude (80%), longitude (10%), cb3 (40%), t2b3 (90%), v1b3 (10%), cb4 (40%), t2b4 (90%), p2b4 (10%), v2b4 (50%), t1b5 (10%), t2b5 (10%), p1b5 (10%), v1b5 (10%), v2b5 (40%), cb6 (100%), t1b6 (100%), t2b6 (30%), v1b6 (10%), cb1 (90%), t1b1 (20%), t2b1 (30%), v1b1 (40%), v2b1 (10%), cb2 (70%), t1b2 (30%), t2b2 (30%), v1b2 (20%), v2b2 (50%)	latitude (80%), longitude (10%), t2b4 (40%), p1b4 (80%), t1b5 (80%), p1b5 (20%), v1b5 (20%), v2b5 (10%), p2b6 (10%), v2b6 (90%), cb1 (10%), v1b1 (40%), t1b2 (10%), v1b2 (60%), v2b2 (30%), cb3 (90%), t1b3 (90%), t2b3 (20%), p1b3 (30%), v1b3 (20%), v2b3 (20%)	latitude (90%), t1b5 (40%), p1b5 (30%), v1b5 (10%), p2b6 (10%), v2b6 (20%), cb1 (10%), t1b2 (10%), v1b2 (20%), cb3 (90%), t1b3 (90%), p1b3 (20%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (70%), v1b4 (10%), v2b4 (100%)	latitude (90%), p2b6 (10%), v2b6 (10%), cb1 (10%), t1b2 (10%), cb3 (90%), t1b3 (80%), p1b3 (20%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (70%), v1b4 (10%), v2b4 (100%), cb5 (100%), t1b5 (40%), t2b5 (80%)
Ant Search (Colônia de Formigas Artificiais)	latitude (10%), longitude (100%), t2b1 (50%), v2b1 (100%), cb2 (90%), p2b2 (10%), v1b2 (20%), v1b3 (100%), cb4 (100%), t2b4 (10%), v1b4 (60%), v2b4 (20%), cb5 (80%), t2b5 (90%), v2b5 (20%), cb6 (70%), t2b6 (80%)	longitude (20%), t2b2 (10%), p1b2 (10%), v2b2 (30%), p2b3 (30%), v1b3 (30%), v2b3 (10%), cb4 (80%), t2b4 (30%), p2b4 (10%), v2b4 (20%), cb5 (10%), t1b5 (30%), t2b5 (50%), p1b5 (20%), p2b5 (50%), v1b5 (40%), cb6 (100%), v2b6 (60%), cb1 (100%), v1b1 (80%)	latitude (60%), longitude (10%), cb3 (40%), t2b3 (100%), cb4 (40%), t2b4 (90%), v1b4 (20%), v2b4 (40%), p1b5 (10%), v1b5 (10%), v2b5 (20%), cb6 (100%), t1b6 (90%), t2b6 (20%), v1b6 (10%), cb1 (90%), t1b1 (20%), t2b1 (50%), v1b1 (30%), t2b1 (10%), cb2 (70%), t1b2 (40%), v1b2 (40%), v2b2 (40%)	latitude (100%), longitude (20%), t2b4 (30%), p1b4 (80%), p2b4 (10%), t1b5 (80%), p1b5 (20%), v1b5 (20%), v2b5 (10%), p2b6 (10%), v2b6 (80%), cb1 (10%), t1b1 (30%), v1b1 (40%), t1b2 (10%), v1b2 (50%), v2b2 (50%), cb3 (90%), t1b3 (90%), t2b3 (10%), p1b3 (30%), v1b3 (20%), v2b3 (20%)	latitude (100%), t1b5 (30%), p1b5 (30%), v1b5 (20%), v2b6 (60%), cb1 (10%), t1b1 (10%), v1b1 (20%), t1b2 (10%), v1b2 (10%), v2b2 (40%), cb3 (90%), t1b3 (90%), p1b3 (30%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (40%), p1b4 (40%), v1b4 (10%), v2b4 (100%)	latitude (90%), p2b6 (10%), v2b6 (10%), cb1 (10%), t1b2 (10%), cb3 (90%), t1b3 (80%), p1b3 (20%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (70%), v1b4 (10%), v2b4 (100%), cb5 (100%), t1b5 (40%), t2b5 (80%)
Votação ou Consenso	latitude (10%), longitude (100%), t2b1 (50%), v2b1 (100%), cb2 (90%), p2b2 (10%), v1b2 (20%), v1b3 (100%), cb4 (100%), t2b4 (10%), v1b4 (60%), v2b4 (20%), cb5 (80%), t2b5 (90%), v2b5 (20%), cb6 (70%), t2b6 (80%)	longitude, t2b2, p1b2, v2b2, p2b3, v1b3, v2b3, cb4, t2b4, p2b4, v2b4, cb5, t1b5, t2b5, p1b5, p2b5, v1b5, cb6, v2b6, cb1, v1b1	latitude, longitude, cb3, t2b3, cb4, t2b4, v1b4, v2b4, p1b5, v1b5, v2b5, cb6, t1b6, t1b5, v1b6, cb1, t1b1, t2b1, p1b1, p2b1, v2b1, cb2, t1b2, t2b2, v1b2, v2b2	latitude, longitude, t2b4, p1b4, p2b4, t1b5, p1b5, v1b5, v2b5, p2b6, v2b6, cb1, t1b1, v1b1, t1b2, v1b2, v2b2, cb3, t1b3, t2b3, p1b3, v1b3, v2b3	latitude, t1b5, p1b5, v1b5, v2b6, cb1, t1b1, v1b1, t1b2, v1b2, v2b2, cb3, t1b3, p1b3, p2b3, v1b3, v2b3, cb4, t2b4, p1b4, v1b4, v2b4	latitude (90%), p2b6 (10%), v2b6 (10%), cb1 (10%), t1b2 (10%), cb3 (90%), t1b3 (80%), p1b3 (20%), p2b3 (10%), v2b3 (10%), cb4 (100%), t2b4 (30%), p1b4 (70%), v1b4 (10%), v2b4 (100%), cb5 (100%), t1b5 (40%), t2b5 (80%)
* População de 20 indivíduos. 500 gerações/iterações. validação cruzada com 10 folds						

Tabela 9 – Resultado da análise de fatores mais relevantes por meio dos algoritmos de Busca Genética, Busca Evolucionária, Otimização por Enxame de Partículas, Busca de Abelhas e Busca de Formigas, para 20 candidatos à solução evoluindo em 500 gerações, para o ano de 2015.

Método	2015.1	2015.2	2015.3	2015.4	2015.5	2015.6
Genetic Search (Algoritmo Genético)	latitude (20%), cb1 (80%), t1b1 (10%), p1b1 (10%), cb2 (100%), t2b2 (90%), p1b2 (10%), v1b2 (30%), v2b2 (10%), cb3 (10%), v1b3 (50%), t2b5 (100%), v1b5 (10%), v2b5 (10%), t1b6 (10%), t2b6 (90%), v1b6 (60%)	latitude (100%), cb2 (100%), t1b2 (10%), p2b2 (30%), cb3 (100%), t2b3 (30%), p1b3 (10%), v1b3 (100%), cb4 (40%), t1b4 (10%), t2b4 (10%), p2b4 (10%), cb5 (40%), t1b5 (20%), p1b5 (40%), v2b5 (50%), cb6 (30%), t2b6 (20%), p1b6 (30%), v1b6 (10%), v2b6 (100%), t2b1 (10%), p1b1 (10%), p2b1 (70%), v1b1 (30%)	latitude (100%), longitude (100%), cb3 (100%), t1b3 (90%), t2b3 (10%), v1b3 (60%), v2b3 (60%), cb4 (100%), t1b4 (20%), t2b4 (100%), p2b4 (90%), v1b4 (60%), v2b4 (70%), cb5 (100%), t1b5 (100%), t2b5 (10%), p1b5 (50%), p2b5 (20%), v1b5 (40%), cb6 (100%), t1b6 (30%), t2b6 (10%), p1b1 (30%), t2b1 (10%), p1b1 (20%), p2b1 (100%), v1b1 (10%), v2b1 (100%), cb2 (100%), t1b2 (10%), p2b2 (50%), v1b2 (60%), v2b2 (100%)	latitude (100%), longitude (80%), cb4 (10%), t1b4 (10%), p2b4 (90%), v2b4 (10%), cb5 (10%), t2b5 (10%), v1b5 (20%), v2b5 (10%), cb6 (10%), t2b6 (10%), p1b6 (10%), v1b6 (60%), t1b1 (100%), t2b1 (90%), p1b1 (90%), v1b1 (90%), v2b1 (10%), t1b2 (90%), t2b2 (10%), p2b2 (10%), v1b2 (90%), v2b2 (80%), cb3 (100%), t1b3 (80%), t2b3 (40%), p1b3 (80%), v1b3 (100%), v2b3 (90%)	latitude (60%), longitude (10%), cb5 (10%), t2b5 (40%), p2b5 (30%), v1b5 (100%), v2b5 (80%), cb6 (10%), t1b6 (50%), t2b6 (30%), p1b6 (90%), p2b6 (100%), v1b6 (90%), t1b1 (100%), p1b1 (30%), p2b1 (100%), v1b1 (60%), v2b1 (90%), cb2 (100%), t1b2 (30%), t2b2 (10%), p1b2 (10%), p2b2 (60%), v1b2 (50%), v2b2 (70%), t1b3 (10%), t2b3 (100%), p1b3 (60%), p2b3 (20%), v1b3 (100%), t2b5 (80%), cb4 (100%), t1b4 (90%), p1b4 (100%), v2b4 (90%)	latitude (80%), longitude (10%), cb6 (10%), t1b6 (70%), t2b6 (100%), p2b6 (30%), v1b6 (100%), v2b6 (90%), cb1 (100%), t1b1 (100%), p1b1 (80%), p2b1 (100%), v1b1 (100%), v2b1 (100%), cb2 (100%), p2b2 (70%), v1b2 (10%), v2b2 (10%), t1b3 (20%), p2b3 (20%), v2b3 (50%), t1b4 (50%), t2b4 (20%), p1b4 (100%), v1b4 (100%), v2b4 (80%), cb5 (100%), t1b5 (20%), t2b5 (90%), p1b5 (10%), p2b5 (60%), v1b5 (40%), v2b5 (100%)
Evolutionary Search (Algoritmo Genético Modificado)	latitude (80%), cb1 (100%), t1b1 (80%), p1b1 (10%), v1b1 (80%), cb2 (100%), t2b2 (100%), v1b2 (20%), v1b3 (80%), t2b5 (100%), v1b5 (80%), t1b6 (10%), t2b6 (80%), v1b6 (30%)	latitude (100%), longitude (10%), cb2 (100%), p2b2 (30%), cb3 (100%), p1b3 (10%), v1b3 (100%), t2b4 (10%), p2b4 (100%), t1b5 (40%), p1b5 (40%), v2b5 (60%), cb6 (100%), t2b6 (60%), v2b6 (100%), p2b1 (100%), v1b1 (10%)	latitude (100%), longitude (100%), cb3 (100%), v1b3 (100%), t2b4 (100%), p2b4 (20%), cb5 (100%), t1b5 (100%), p1b5 (90%), cb6 (100%), t1b6 (100%), t2b6 (100%), p2b6 (100%), t1b1 (100%), p1b1 (100%), p2b1 (100%), v1b1 (100%), v2b1 (100%), cb2 (100%), v1b2 (100%), v2b2 (100%)	latitude (30%), longitude (90%), cb4 (10%), p2b4 (100%), t2b6 (100%), v1b6 (90%), t1b1 (100%), t2b1 (10%), p1b1 (100%), v1b1 (100%), t1b2 (100%), v1b2 (10%), v2b2 (100%), cb3 (100%), t1b3 (10%), t2b3 (100%), p1b3 (30%), v1b3 (100%), v2b3 (90%)	latitude (60%), longitude (10%), t2b5 (70%), p2b5 (30%), v1b5 (100%), v2b5 (70%), t1b6 (60%), t2b6 (30%), p1b6 (100%), p2b6 (100%), v1b6 (60%), v2b6 (70%), t1b1 (100%), p1b1 (30%), p2b1 (100%), v1b1 (60%), v2b1 (90%), cb2 (100%), t1b2 (30%), t2b2 (30%), p2b2 (60%), v1b2 (10%), v2b2 (70%), t1b3 (30%), t2b3 (100%), p1b3 (50%), p2b3 (90%), cb4 (100%), t1b4 (70%), p1b4 (100%), v2b4 (70%)	latitude (50%), longitude (50%), cb6 (50%), t1b6 (50%), t2b6 (100%), p2b6 (30%), v1b6 (50%), v2b6 (60%), cb1 (100%), t1b1 (100%), p1b1 (50%), p2b1 (100%), v1b1 (100%), v2b1 (100%), cb2 (100%), t1b2 (50%), t2b2 (50%), p2b2 (70%), v1b2 (50%), v2b2 (50%), p2b3 (50%), v2b3 (20%), t1b4 (50%), t2b4 (30%), p1b4 (100%), v1b4 (100%), v2b4 (80%), cb5 (100%), t1b5 (50%), p2b5 (80%), v1b5 (30%), v2b5 (100%)
PSO Search (Otimização por Enxame de Partículas)	latitude (20%), cb1 (80%), t1b1 (10%), p1b1 (10%), cb2 (100%), t2b2 (90%), p1b2 (10%), v1b2 (30%), v2b2 (10%), cb3 (10%), v1b3 (50%), t2b5 (100%), v1b5 (10%), v2b5 (10%), t1b6 (10%), t2b6 (90%), v1b6 (60%)	latitude (100%), cb2 (100%), t1b2 (10%), p2b2 (30%), v2b2 (10%), cb3 (100%), t2b3 (30%), p1b3 (10%), v1b3 (100%), cb4 (10%), t1b4 (10%), t2b4 (20%), p2b4 (10%), v2b4 (10%), cb5 (50%), t1b5 (30%), p1b5 (40%), p2b5 (10%), v1b5 (40%), cb6 (70%), t1b6 (10%), t2b6 (10%), p1b6 (30%), p2b6 (10%), v1b6 (10%), v2b6 (100%), cb1 (10%), t2b1 (10%), p1b1 (10%), p2b1 (70%), v1b1 (30%)	latitude (100%), longitude (100%), cb3 (100%), t1b3 (90%), t2b3 (10%), v1b3 (60%), v2b3 (60%), cb4 (100%), t1b4 (20%), t2b4 (100%), p2b4 (90%), v1b4 (70%), v2b4 (70%), cb5 (100%), t1b5 (100%), t2b5 (10%), p1b5 (50%), p2b5 (20%), v1b5 (40%), cb6 (100%), t1b6 (30%), t2b6 (10%), p1b1 (30%), t2b1 (10%), p1b1 (20%), p2b1 (100%), v1b1 (10%), v2b1 (100%), cb2 (100%), t1b2 (10%), p2b2 (50%), v1b2 (60%), v2b2 (100%)	latitude (100%), longitude (80%), cb4 (10%), t1b4 (10%), p2b4 (90%), v2b4 (10%), cb5 (10%), t2b5 (10%), v1b5 (20%), v2b5 (10%), cb6 (10%), t2b6 (10%), p1b6 (10%), v1b6 (60%), t1b1 (100%), t2b1 (90%), p1b1 (90%), v1b1 (90%), v2b1 (10%), t1b2 (90%), t2b2 (10%), p2b2 (10%), v1b2 (90%), v2b2 (80%), cb3 (100%), t1b3 (80%), t2b3 (40%), p1b3 (80%), v1b3 (100%), v2b3 (90%)	latitude (60%), longitude (10%), cb5 (10%), t2b5 (40%), p2b5 (30%), v1b5 (100%), v2b5 (80%), cb6 (10%), t1b6 (50%), t2b6 (30%), p1b6 (90%), p2b6 (100%), v1b6 (90%), t1b1 (100%), p1b1 (30%), p2b1 (100%), v1b1 (60%), v2b1 (90%), cb2 (100%), t1b2 (30%), t2b2 (10%), p1b2 (10%), p2b2 (60%), v1b2 (50%), v2b2 (70%), t1b3 (10%), t2b3 (100%), p1b3 (60%), p2b3 (20%), v1b3 (100%), t2b5 (80%), cb4 (100%), t1b4 (90%), p1b4 (100%), v2b4 (90%)	latitude (90%), longitude (10%), cb6 (10%), t1b6 (70%), t2b6 (100%), p2b6 (30%), v1b6 (100%), v2b6 (90%), cb1 (100%), t1b1 (100%), p1b1 (80%), p2b1 (100%), v1b1 (100%), v2b1 (100%), cb2 (100%), p2b2 (70%), v1b2 (10%), v2b2 (10%), t1b3 (20%), p2b3 (20%), v2b3 (50%), t1b4 (50%), t2b4 (20%), p1b4 (100%), v1b4 (100%), v2b4 (80%), cb5 (100%), t1b5 (20%), t2b5 (90%), p1b5 (10%), p2b5 (60%), v1b5 (40%), v2b5 (100%)
Bee Search (Colônia de Abelhas Artificiais)	latitude (10%), longitude (10%), cb1 (80%), t1b1 (10%), p1b1 (40%), cb2 (100%), t2b2 (90%), v1b2 (20%), v2b2 (10%), cb3 (10%), t1b3 (10%), v1b3 (20%), v1b4 (10%), t2b5 (60%), v1b5 (10%), v2b5 (10%), t1b6 (10%), t2b6 (60%), p1b6 (10%), v1b6 (60%)	latitude (50%), longitude (10%), cb2 (100%), t1b2 (10%), p1b2 (10%), p2b2 (50%), v2b2 (10%), cb3 (100%), p1b3 (10%), v1b3 (70%), cb4 (30%), t2b4 (10%), p2b4 (10%), cb5 (70%), t1b5 (10%), p1b5 (50%), v2b5 (40%), cb6 (10%), t2b6 (30%), p1b6 (30%), v1b6 (10%), v2b6 (100%), p1b1 (10%), p2b1 (60%), v1b1 (40%)	latitude (90%), longitude (60%), cb3 (100%), t1b3 (70%), t2b3 (30%), v2b3 (70%), cb4 (100%), t2b4 (10%), p2b4 (90%), v1b4 (50%), v2b4 (60%), cb5 (100%), t1b5 (100%), t2b5 (20%), p1b5 (60%), p2b5 (20%), v1b5 (40%), v2b5 (10%), cb6 (100%), t1b6 (10%), t2b6 (20%), p2b6 (30%), v1b6 (10%), v2b6 (10%), t1b1 (40%), t2b1 (10%), p1b1 (20%), p2b1 (100%), v1b1 (10%), v2b1 (10%), cb2 (100%), p2b2 (50%), v1b2 (20%), v2b2 (90%)	latitude (70%), longitude (20%), cb4 (10%), p2b4 (100%), cb5 (10%), t2b5 (20%), v1b5 (70%), t2b6 (90%), v1b6 (60%), t1b1 (80%), t2b1 (10%), p1b1 (90%), p2b1 (30%), v1b1 (90%), v2b1 (30%), t1b2 (70%), v1b2 (90%), v2b2 (30%), cb3 (100%), t1b3 (100%), t2b3 (80%), p1b3 (70%), v1b3 (90%), v2b3 (80%)	latitude (60%), longitude (10%), cb5 (10%), t2b5 (30%), p2b5 (30%), v1b5 (100%), v2b5 (40%), cb6 (30%), t1b6 (50%), t2b6 (20%), p1b6 (90%), p2b6 (100%), v1b6 (30%), p2b6 (70%), t1b1 (80%), p1b1 (30%), p2b1 (40%), v1b1 (40%), v2b1 (80%), cb2 (100%), t1b2 (40%), p1b2 (100%), v1b2 (50%), v2b2 (40%), v2b2 (60%), t1b3 (30%), t2b3 (100%), p1b3 (60%), p2b3 (20%), v1b3 (100%), v2b3 (50%), cb4 (100%), t1b4 (80%), t2b4 (30%), p1b4 (100%), v2b4 (90%)	latitude (20%), longitude (40%), cb6 (10%), t1b6 (50%), t2b6 (100%), p2b6 (30%), v1b6 (20%), v2b6 (90%), cb1 (100%), t1b1 (100%), p1b1 (70%), p2b1 (40%), v1b1 (100%), v2b1 (80%), cb2 (100%), t1b2 (90%), t2b2 (50%), p1b2 (10%), v2b2 (10%), t1b3 (40%), t2b3 (20%), v1b3 (100%), t2b5 (60%), p1b5 (10%), p2b5 (50%), v1b5 (40%), v2b5 (100%)
Ant Search (Colônia de Formigas Artificiais)	latitude (20%), cb1 (80%), t1b1 (10%), p1b1 (10%), cb2 (100%), t2b2 (90%), p1b2 (10%), v1b2 (30%), v2b2 (10%), cb3 (10%), v1b3 (50%), t2b5 (100%), v1b5 (10%), v2b5 (10%), t1b6 (10%), t2b6 (90%), v1b6 (60%)	latitude (100%), cb2 (100%), t1b2 (10%), p2b2 (30%), v2b2 (10%), cb3 (100%), t2b3 (30%), p1b3 (10%), v1b3 (100%), cb4 (20%), t1b4 (10%), t2b4 (20%), p2b4 (10%), v2b4 (10%), cb5 (40%), t1b5 (30%), p1b5 (40%), p2b5 (10%), v1b5 (40%), cb6 (10%), t2b6 (30%), p1b6 (30%), v1b6 (10%), v2b6 (100%), cb1 (10%), t2b1 (10%), p1b1 (10%), p2b1 (70%), v1b1 (30%)	latitude (100%), longitude (100%), cb3 (100%), t1b3 (90%), t2b3 (10%), v1b3 (60%), v2b3 (60%), cb4 (100%), t1b4 (20%), t2b4 (100%), p2b4 (90%), v1b4 (70%), v2b4 (70%), cb5 (100%), t1b5 (100%), t2b5 (10%), p1b5 (50%), p2b5 (20%), v1b5 (40%), cb6 (100%), t1b6 (30%), t2b6 (10%), p1b1 (30%), t2b1 (10%), p1b1 (20%), p2b1 (100%), v1b1 (10%), v2b1 (100%), cb2 (100%), t1b2 (10%), p2b2 (50%), v1b2 (60%), v2b2 (100%)	latitude (90%), longitude (80%), cb4 (10%), t1b4 (10%), p2b4 (90%), v2b4 (10%), cb5 (10%), t2b5 (10%), v1b5 (20%), v2b5 (10%), cb6 (10%), t2b6 (10%), p1b6 (10%), v1b6 (60%), t1b1 (100%), t2b1 (90%), p1b1 (90%), v1b1 (90%), v2b1 (10%), t1b2 (90%), t2b2 (10%), p2b2 (10%), v1b2 (90%), v2b2 (80%), cb3 (100%), t1b3 (80%), t2b3 (40%), p1b3 (80%), v1b3 (100%), v2b3 (90%)	latitude (40%), longitude (10%), cb5 (10%), t2b5 (40%), p2b5 (30%), v1b5 (100%), v2b5 (80%), cb6 (10%), t1b6 (50%), t2b6 (30%), p1b6 (90%), p2b6 (100%), v1b6 (90%), t1b1 (100%), p1b1 (30%), p2b1 (100%), v1b1 (60%), v2b1 (90%), cb2 (100%), t1b2 (30%), t2b2 (10%), p1b2 (10%), p2b2 (60%), v1b2 (50%), v2b2 (70%), t1b3 (10%), t2b3 (100%), p1b3 (60%), p2b3 (20%), v1b3 (100%), t2b5 (80%), cb4 (100%), t1b4 (90%), p1b4 (100%), v2b4 (90%)	latitude (80%), longitude (10%), cb6 (10%), t1b6 (70%), t2b6 (100%), p2b6 (40%), v1b6 (100%), v2b6 (90%), cb1 (100%), t1b1 (100%), p1b1 (80%), p2b1 (100%), v1b1 (100%), v2b1 (100%), cb2 (100%), p2b2 (60%), v1b2 (10%), v2b2 (10%), t1b3 (20%), p2b3 (20%), v2b3 (50%), t1b4 (100%), v1b4 (100%), v2b4 (80%), cb5 (100%), t1b5 (20%), t2b5 (90%), p1b5 (10%), p2b5 (60%), v1b5 (40%), v2b5 (100%)
Votação ou Consenso	latitude, cb1, t1b1, p1b1, cb2, p1b2, v1b2, v2b2, cb3, v1b3, t2b5, v1b5, v2b5, t1b6, t2b6, v1b6	latitude, cb2, t1b2, p2b2, v2b2, cb3, t2b3, p1b3, v1b3, cb4, t1b4, t2b4, p2b4, cb5, t1b5, p1b5, v2b5, cb6, t2b6, p2b6, v1b6, v2b6, t2b1, p1b1, p2b1, v1b1	latitude, longitude, cb3, t1b3, t2b3, v1b3, v2b3, cb4, t1b4, t2b4, p2b4, v1b4, v2b4, cb5, t1b5, t2b5, p1b5, v2b5, v1b6, t2b6, p2b6, v1b6, v2b6, t1b1, t2b1, p1b1, p2b1, v1b1, cb2, t1b2, p1b2, v1b2, v2b2	latitude, longitude, cb4, t1b4, p2b4, v2b4, cb5, t2b5, v1b5, v2b5, cb6, t2b6, p1b6, v1b6, t1b1, t2b1, p1b1, v1b1, t1b2, t2b2, p2b2, v1b2, v2b2, cb3, t1b3, p1b3, v1b3, v2b3	latitude, longitude, cb5, t2b5, p2b5, v1b5, v2b5, cb6, t1b6, v1b6, t2b6, p1b6, p2b6, v1b6, v2b6, t1b1, p1b1, v1b1, t1b2, t2b2, p1b2, p2b2, v1b2, v2b2, t1b3, p1b3, cb4, t1b4, p1b4, v2b4	latitude, longitude, cb6, t1b6, t2b6, p2b6, v1b6, v2b6, cb1, t1b1, p1b1, v1b1, v2b1, cb2, t1b2, t2b2, p2b2, v1b2, v2b2, t1b3, p1b3, t1b4, t2b4, p1b4, v1b4, v2b4, cb5, t1b5, t2b5, p1b5, p2b5, v1b5, v2b5
* População de 20 indivíduos. 500 gerações/iterações. validação cruzada com 10 folds						

Tabela 10 – Resultado da análise de fatores mais relevantes por meio dos algoritmos de Busca Genética, Busca Evolucionária, Otimização por Enxame de Partículas, Busca de Abelhas e Busca de Formigas, para 20 candidatos à solução evoluindo em 500 gerações, para o ano de 2016.

Método	2016.1	2016.2	2016.3	2016.4	2016.5	2016.6
Genetic Search (Algoritmo Genético)	longitude (100%), cb1 (90%), p1b1 (90%), p2b1 (10%), vb1 (100%), vb21 (10%), cb2 (100%), t2b2 (70%), t1b2 (80%), p2b2 (20%), vb12 (100%), cb3 (100%), t2b3 (10%), p1b3 (10%), p2b3 (70%), vb23 (90%), t1b4 (100%), t2b4 (10%), vb4 (100%), t1b5 (90%), t2b5 (100%), p1b5 (10%), p2b5 (10%), vb5 (40%), vb25 (100%), cb6 (100%), t2b6 (90%), p1b6 (100%), p2b6 (90%), vb16 (30%), vb26 (70%)	t1b2 (40%), p1b2 (20%), vb12 (70%), t1b3 (30%), t2b3 (40%), p1b3 (50%), p2b3 (100%), vb23 (60%), t1b4 (50%), p1b4 (30%), vb4 (10%), vb24 (10%), cb5 (100%), t1b5 (90%), p1b5 (50%), p2b5 (80%), t2b6 (30%), p1b6 (70%), p2b6 (30%), vb16 (90%), vb26 (100%), cb1 (100%), p1b1 (20%), p2b1 (100%)	latitude (90%), longitude (50%), cb3 (10%), t1b3 (10%), t2b3 (10%), p2b3 (50%), vb13 (60%), vb23 (10%), cb4 (10%), t2b4 (10%), vb24 (60%), cb5 (100%), t1b5 (40%), t2b5 (40%), p1b5 (70%), p2b5 (80%), vb15 (30%), vb25 (10%), cb6 (10%), p1b6 (60%), vb16 (100%), vb26 (30%), cb1 (100%), t1b1 (10%), t2b1 (80%), p1b1 (80%), p2b1 (80%), vb11 (100%), vb21 (100%), cb2 (100%), t1b2 (90%), t2b2 (50%), p1b2 (90%), p2b2 (40%), vb12 (20%), vb22 (40%),	latitude (100%), t1b4 (70%), t2b4 (90%), p1b4 (30%), vb4 (10%), t2b5 (90%), t1b5 (40%), t2b5 (100%), p1b5 (30%), p2b5 (80%), cb6 (70%), vb6 (90%), vb26 (10%), cb1 (70%), t1b1 (80%), t2b1 (10%), p1b1 (10%), vb1 (90%), vb21 (100%), vb11 (90%), vb21 (100%), cb2 (10%), t1b2 (10%), t2b2 (40%), p1b2 (40%), p2b2 (80%), vb2 (100%), vb22 (50%), t2b3 (10%), p1b3 (10%), p2b3 (80%), vb13 (100%), vb23 (50%)	latitude (100%), longitude (70%), t1b5 (40%), t2b5 (90%), p1b5 (70%), vb15 (10%), t1b6 (20%), cb1 (30%), t2b1 (100%), p1b1 (80%), vb1 (100%), t1b2 (20%), p2b2 (20%), vb2 (70%), vb22 (10%), cb3 (100%), t1b3 (40%), t2b3 (10%), vb13 (60%), vb23 (30%), cb4 (100%), p1b4 (20%), vb14 (10%)	latitude (10%), cb6 (90%), t2b6 (60%), vb26 (80%), cb1 (100%), t1b1 (60%), vb21 (70%), cb2 (10%), t2b2 (90%), p2b2 (60%), vb12 (80%), vb22 (100%), cb3 (80%), t1b3 (90%), t2b3 (10%), vb13 (20%), t1b4 (10%), p1b4 (100%), vb14 (10%), cb5 (20%), t1b5 (70%), t2b5 (20%), p1b5 (40%), p2b5 (30%), vb15 (90%), vb25 (10%)
Evolutionary Search (Algoritmo Genético Modificado)	latitude (20%), longitude (100%), cb1 (100%), p1b1 (100%), p2b1 (10%), vb1 (100%), vb21 (10%), cb2 (100%), t1b2 (10%), t2b2 (80%), p1b2 (90%), p2b2 (10%), vb12 (100%), t1b3 (40%), p1b3 (100%), t2b3 (10%), p1b5 (30%), t2b5 (90%), t1b4 (10%), vb4 (100%), t1b5 (90%), t2b5 (10%), p2b5 (10%), vb15 (50%), vb25 (100%), cb6 (100%), t2b6 (90%), p1b6 (100%), p2b6 (90%), vb16 (60%), vb26 (40%)	t1b2 (100%), t1b3 (100%), p1b3 (100%), p2b3 (100%), vb23 (100%), t1b4 (100%), cb3 (100%), t1b5 (100%), vb15 (100%), p1b5 (100%), p2b5 (100%), vb16 (100%), cb1 (100%), p1b1 (10%), p2b1 (100%)	latitude (90%), longitude (10%), cb3 (10%), t1b3 (30%), p2b3 (60%), vb13 (80%), cb4 (100%), t1b4 (30%), vb24 (20%), cb5 (100%), t1b5 (40%), t2b5 (50%), p1b5 (90%), p2b5 (60%), vb15 (10%), vb25 (30%), cb6 (10%), p1b6 (60%), vb16 (100%), vb26 (10%), cb1 (100%), t1b1 (30%), t2b1 (70%), p1b1 (100%), p2b1 (100%), vb11 (100%), vb21 (100%), cb2 (100%), t1b2 (100%), t2b2 (30%), p1b2 (100%), p2b2 (70%)	latitude (90%), vb24 (100%), t1b5 (10%), p1b5 (100%), p2b5 (100%), cb6 (100%), p2b6 (100%), vb26 (100%), cb1 (100%), t1b1 (100%), vb1 (100%), t2b1 (100%), p2b1 (100%), vb11 (100%), t2b3 (100%), p2b3 (100%), vb13 (100%), vb23 (100%)	latitude (100%), longitude (70%), t2b5 (100%), p1b5 (90%), vb15 (20%), t1b6 (10%), vb6 (10%), cb1 (30%), t2b1 (100%), p1b1 (90%), vb1 (100%), vb2 (100%), cb3 (100%), t1b3 (100%), p2b3 (70%), vb13 (100%), cb4 (100%)	cb6 (90%), t2b6 (100%), vb26 (100%), cb1 (100%), t1b1 (100%), p1b1 (100%), cb2 (10%), t2b2 (100%), vb22 (100%), p2b2 (100%), t1b3 (100%), p1b4 (100%), t1b5 (100%), vb15 (100%)
PSO Search (Otimização por Enxame de Partículas)	longitude (100%), cb1 (90%), p1b1 (90%), p2b1 (10%), vb1 (100%), vb21 (10%), cb2 (100%), t2b2 (70%), t1b2 (80%), p2b2 (20%), vb12 (100%), cb3 (100%), t2b3 (10%), p1b3 (10%), p2b3 (70%), vb23 (90%), t1b4 (100%), t2b4 (10%), vb4 (100%), t1b5 (90%), t2b5 (100%), p1b5 (10%), p2b5 (10%), vb5 (40%), vb25 (100%), cb6 (100%), t2b6 (90%), p1b6 (100%), p2b6 (90%), vb16 (30%), vb26 (70%)	t1b2 (40%), p1b2 (20%), vb12 (70%), t1b3 (30%), t2b3 (40%), p1b3 (50%), p2b3 (100%), vb23 (60%), t1b4 (50%), p1b4 (30%), vb4 (10%), vb24 (10%), cb5 (100%), t1b5 (90%), p1b5 (50%), p2b5 (80%), t2b6 (30%), p1b6 (70%), p2b6 (30%), vb16 (90%), vb26 (100%), cb1 (100%), p1b1 (20%), p2b1 (100%)	latitude (90%), longitude (50%), cb3 (10%), t1b3 (10%), t2b3 (10%), p2b3 (50%), vb13 (60%), vb23 (10%), cb4 (10%), t2b4 (10%), vb24 (60%), cb5 (100%), t1b5 (40%), t2b5 (40%), p1b5 (70%), p2b5 (80%), vb15 (30%), vb25 (10%), cb6 (10%), p1b6 (60%), vb16 (100%), vb26 (30%), cb1 (100%), t1b1 (10%), t2b1 (80%), p1b1 (80%), p2b1 (80%), vb11 (100%), vb21 (100%), cb2 (100%), t1b2 (90%), t2b2 (50%), p1b2 (90%), p2b2 (40%), vb12 (20%), vb22 (40%),	latitude (100%), t1b4 (70%), t2b4 (90%), p1b4 (30%), vb4 (10%), t2b5 (90%), t1b5 (40%), t2b5 (100%), p1b5 (30%), p2b5 (80%), cb6 (70%), vb6 (90%), vb26 (10%), cb1 (70%), t1b1 (80%), t2b1 (10%), p1b1 (10%), vb1 (90%), vb21 (100%), vb11 (90%), vb21 (100%), cb2 (10%), t1b2 (10%), t2b2 (40%), p1b2 (60%), p2b2 (80%), vb2 (100%), vb22 (50%), t2b3 (10%), p1b3 (10%), p2b3 (90%), vb13 (100%), vb23 (50%)	latitude (100%), longitude (70%), t1b5 (40%), t2b5 (90%), p1b5 (70%), vb15 (10%), t1b6 (20%), cb1 (30%), t2b1 (100%), p1b1 (80%), vb1 (100%), t1b2 (20%), p2b2 (20%), vb2 (70%), vb22 (10%), cb3 (100%), t1b3 (40%), t2b3 (10%), vb13 (60%), vb23 (30%), cb4 (100%), p1b4 (20%), vb14 (10%)	latitude (10%), cb6 (90%), t2b6 (60%), vb26 (80%), cb1 (100%), t1b1 (60%), vb21 (70%), cb2 (10%), t2b2 (90%), p2b2 (50%), vb12 (80%), vb22 (100%), cb3 (80%), t1b3 (90%), t2b3 (20%), vb13 (20%), t1b4 (10%), p1b4 (100%), vb14 (10%), cb5 (20%), t1b5 (70%), t2b5 (20%), p1b5 (40%), p2b5 (30%), vb15 (90%), vb25 (10%)
Bee Search (Colônia de Abelhas Artificiais)	latitude (30%), longitude (90%), cb1 (90%), t2b1 (40%), p1b1 (90%), p2b1 (40%), vb1 (70%), vb21 (10%), cb2 (100%), t1b2 (30%), t2b2 (60%), p1b2 (70%), p2b2 (20%), vb12 (100%), cb3 (100%), t2b3 (10%), p1b3 (10%), p2b3 (60%), vb23 (50%), t1b4 (30%), t2b4 (50%), vb4 (90%), t1b5 (90%), t2b5 (100%), p1b5 (50%), p2b5 (10%), vb5 (40%), vb25 (100%), cb6 (100%), t2b6 (90%), p1b6 (90%), p2b6 (80%), vb16 (30%), vb26 (20%)	latitude (20%), longitude (40%), t1b2 (30%), p1b2 (20%), vb12 (20%), vb22 (10%), t1b3 (40%), t2b3 (40%), p1b3 (50%), p2b3 (90%), t1b3 (10%), vb13 (100%), t1b4 (30%), p1b4 (60%), t1b4 (30%), p1b3 (50%), vb14 (10%), vb24 (20%), cb5 (100%), t1b5 (80%), p1b5 (40%), p2b5 (10%), vb15 (40%), vb25 (100%), cb6 (100%), t2b6 (100%), p1b6 (90%), p2b6 (80%), vb16 (30%), vb26 (20%)	latitude (40%), longitude (70%), cb3 (10%), t1b3 (10%), p2b3 (50%), vb13 (50%), vb23 (10%), cb4 (10%), t1b4 (10%), t2b4 (10%), vb24 (30%), cb5 (100%), t1b5 (30%), t2b5 (30%), p1b5 (70%), p2b5 (70%), vb15 (30%), vb25 (10%), cb6 (10%), p1b6 (60%), vb16 (100%), cb1 (20%), t1b1 (50%), p1b1 (60%), vb21 (80%), vb1 (90%), vb21 (80%), cb2 (100%), t1b2 (90%), t2b2 (60%), p1b2 (90%), p2b2 (40%), vb12 (20%), vb22 (40%)	latitude (90%), longitude (40%), t1b4 (10%), t2b4 (40%), p1b4 (30%), vb4 (70%), t1b5 (30%), t2b5 (10%), p1b5 (20%), p2b5 (30%), cb6 (70%), p2b6 (20%), vb16 (80%), vb26 (20%), cb1 (70%), t1b1 (50%), t2b1 (10%), p1b1 (10%), vb1 (20%), vb21 (90%), vb21 (80%), cb2 (100%), t1b2 (20%), t2b2 (20%), p1b2 (30%), p2b2 (80%), vb2 (40%), vb22 (30%), t1b3 (100%), t1b3 (60%), p1b3 (20%), p2b3 (40%), vb13 (40%), vb23 (50%)	latitude (100%), longitude (20%), t1b5 (30%), t2b5 (90%), p1b5 (70%), vb15 (10%), t1b6 (80%), p1b1 (70%), t1b1 (80%), vb1 (80%), t1b2 (20%), vb12 (70%), cb3 (100%), t1b3 (10%), t2b3 (10%), vb13 (60%), vb23 (30%), cb4 (100%), p1b4 (20%), vb14 (10%)	latitude (10%), longitude (30%), cb6 (90%), t2b6 (50%), vb26 (80%), cb1 (100%), t1b1 (60%), vb1 (10%), vb21 (80%), cb2 (10%), t2b2 (90%), p2b2 (80%), vb12 (60%), vb22 (90%), cb3 (80%), t1b3 (90%), t2b3 (20%), p2b3 (10%), vb13 (30%), t1b4 (10%), p1b4 (80%), vb14 (10%), cb5 (20%), t1b5 (60%), vb25 (20%), p1b5 (40%), vb25 (20%)
Ant Search (Colônia de Formigas Artificiais)	longitude (100%), cb1 (90%), p1b1 (90%), p2b1 (10%), vb1 (100%), vb21 (10%), cb2 (100%), t2b2 (70%), t1b2 (80%), p2b2 (20%), vb12 (100%), cb3 (100%), t2b3 (10%), p1b3 (10%), p2b3 (70%), vb23 (90%), t1b4 (100%), t2b4 (10%), vb4 (100%), t1b5 (90%), t2b5 (100%), p1b5 (10%), p2b5 (10%), vb5 (40%), vb25 (100%), cb6 (100%), t2b6 (90%), p1b6 (100%), p2b6 (90%), vb16 (30%), vb26 (70%)	t1b2 (40%), p1b2 (20%), vb12 (70%), t1b3 (30%), t2b3 (40%), p1b3 (50%), p2b3 (100%), vb23 (60%), t1b4 (50%), p1b4 (30%), vb4 (10%), vb24 (10%), cb5 (100%), t1b5 (90%), p1b5 (50%), p2b5 (80%), t2b6 (30%), p1b6 (70%), p2b6 (30%), vb16 (90%), vb26 (100%), cb1 (100%), p1b1 (20%), p2b1 (100%)	latitude (90%), longitude (50%), cb3 (10%), t1b3 (10%), t2b3 (10%), p2b3 (50%), vb13 (60%), vb23 (10%), cb4 (10%), t2b4 (10%), vb24 (60%), cb5 (100%), t1b5 (40%), t2b5 (40%), p1b5 (70%), p2b5 (80%), vb15 (30%), vb25 (10%), cb6 (10%), p1b6 (60%), vb16 (100%), vb26 (30%), cb1 (100%), t1b1 (10%), t2b1 (80%), p1b1 (80%), p2b1 (80%), vb11 (100%), vb21 (100%), cb2 (100%), t1b2 (90%), t2b2 (50%), p1b2 (90%), p2b2 (40%), vb12 (20%), vb22 (40%)	latitude (100%), t1b4 (70%), t2b4 (90%), p1b4 (30%), vb4 (20%), vb24 (90%), t1b5 (40%), p1b5 (20%), vb25 (70%), cb6 (70%), t2b6 (10%), p2b6 (10%), vb16 (80%), t1b1 (60%), t2b1 (10%), p1b1 (10%), vb1 (90%), vb21 (100%), vb11 (90%), vb21 (100%), cb2 (10%), t1b2 (10%), t2b2 (40%), p1b2 (30%), p2b2 (80%), vb2 (100%), vb22 (30%), t1b3 (30%), p1b3 (10%), p2b3 (50%)	latitude (100%), longitude (70%), t1b5 (40%), t2b5 (90%), p1b5 (70%), vb15 (10%), t1b6 (20%), cb1 (30%), t2b1 (100%), p1b1 (80%), vb1 (100%), t1b2 (20%), p2b2 (20%), vb2 (70%), vb22 (10%), cb3 (100%), t1b3 (40%), t2b3 (10%), vb13 (60%), vb23 (30%), cb4 (100%), p1b4 (20%), vb14 (10%)	latitude (10%), cb6 (90%), t2b6 (60%), vb26 (80%), cb1 (100%), t1b1 (60%), vb21 (70%), cb2 (10%), t2b2 (90%), p2b2 (60%), vb12 (80%), vb22 (100%), cb3 (80%), t1b3 (90%), t2b3 (10%), vb13 (20%), t1b4 (10%), p1b4 (100%), vb14 (10%), cb5 (20%), t1b5 (70%), t2b5 (20%), p1b5 (40%), p2b5 (30%), vb15 (90%), vb25 (10%)
Votação ou Consenso	longitude, cb1, p1b1, p2b1, vb1, vb21, cb2, t2b2, p1b2, p2b2, vb2, cb3, t1b3, t2b3, p1b3, p2b3, vb3, vb23, t1b4, t2b4, vb4, t1b5, t2b5, p1b5, p2b5, vb5, vb25, cb6, t2b6, p1b6, p2b6, vb6, vb26	t1b2, p1b2, vb12, t1b3, t2b3, p1b3, p2b3, t1b4, p1b4, vb4, vb24, cb5, t1b5, p1b5, p2b5, p1b6, p2b6, vb6, vb26, cb1, p1b1, p2b1	latitude, longitude, cb3, t1b3, t2b3, p2b3, vb13, vb23, cb4, t2b4, vb24, cb5, t1b5, t2b5, p1b5, p2b5, vb15, vb25, cb6, p1b6, vb16, vb26, cb1, t1b1, t2b1, p1b1, p2b1, vb1, vb21, cb2, t1b2, p1b2, p2b2, vb2, vb22	latitude, t1b4, t2b4, p1b4, vb4, t1b5, t2b5, p1b5, p2b5, cb6, p2b6, vb16, vb26, cb1, t1b1, p1b1, p2b1, vb1, vb21, cb2, t1b2, t2b2, p1b2, p2b2, vb2, vb22, cb3, t1b3, t2b3, p1b3, p2b3, vb13, vb23	latitude, longitude, t1b5, t2b5, p1b5, vb15, t1b6, cb1, t2b1, p1b1, vb1, vb21, t2b2, p2b2, vb12, vb22, cb3, t1b3, t2b3, vb13, vb23, cb4, p1b4, vb14	latitude, cb6, t2b6, vb26, cb1, t1b1, vb21, cb2, t2b2, p1b2, vb12, vb22, cb3, t1b3, t2b3, vb13, t1b4, p1b4, vb14, cb5, t1b5, t2b5, p1b5, p2b5, vb15, vb25
* População de 20 indivíduos, 500 operações/iterações, validação cruzada com 10 folds						

bimestres de 2 a 4 (março a julho de 2013), 6 (novembro e dezembro de 2013) e 1 (janeiro e fevereiro de 2014); e número de casos do bimestre 4 ao 1 (julho de 2013 a fevereiro de 2014).

Na predição do bimestre de maio a junho de 2014 (2014.3), conforme Tabela 8, apareceram os seguintes fatores como mais relevantes: posição geográfica (latitude e longitude); temperatura nos bimestres 3 a 4 (maio a agosto de 2013) e de 6 a 2 (novembro de 2013 a abril de 2014); velocidade dos ventos nos bimestres de 4 a 2 (julho de 2013 a abril de 2014); pluviometria nos bimestres 5 (setembro a outubro de 2013) e 1 (janeiro a fevereiro de 2014); e número de casos em todo o período, exceto o bimestre 5 (setembro a outubro de 2013).

Na predição do bimestre de julho a agosto de 2014 (2014.4), conforme Tabela 8, os seguintes fatores se destacaram: posição geográfica (latitude e longitude); temperatura nos bimestres 4 a 5 (julho a outubro de 2013), e 1 a 3 (janeiro a junho de 2014); pluviometria nos bimestres 4 a 6 (julho a dezembro de 2013) e 3 (maio a junho de 2014); velocidade dos ventos nos bimestres de 5 a 3 (setembro de 2013 a junho de 2014); e número de casos nos bimestres 1 (janeiro a fevereiro de 2014) e 3 (maio a junho de 2014).

Na predição do bimestre de setembro a outubro de 2014 (2014.5), conforme Tabela 8, destacaram-se os seguintes fatores: posição geográfica (latitude); temperatura nos bimestres 5 (setembro a outubro de 2013), e de 1 a 4 (janeiro a agosto de 2014); pluviometria nos bimestres 5 (setembro a outubro de 2013) e de 3 a 4 (maio a agosto de 2014); velocidade dos ventos em todos os seis bimestres; e número de casos nos bimestres 1 (janeiro a fevereiro de 2014) e de 3 a 4 (maio a agosto de 2014).

Na predição do bimestre de novembro a dezembro de 2014 (2014.6), conforme Tabela 8, os seguintes fatores tiveram destaque: posição geográfica (latitude); pluviometria nos bimestres 6 (novembro a dezembro de 2013) e de 3 a 4 (maio a agosto de 2014); velocidade dos ventos nos bimestres 6 (novembro a dezembro de 2013) e de 3 a 4 (maio a agosto de 2014); temperatura nos bimestres 2 (março a abril de 2014) e de 4 a 5 (julho a outubro de 2014); e número de casos de arboviroses nos bimestres 1 (janeiro a fevereiro de 2014) e de 3 a 5 (maio a agosto de 2014).

A Tabela 9 mostra os resultados de predição para os seis bimestres de 2015 considerando o ano de 2014. Analisando o resultado do consenso ou da votação no comitê de analistas artificiais têm-se, para a predição do bimestre de janeiro a fevereiro de 2015 (2015.1), os seguintes fatores considerados mais relevantes: posição geográfica (latitude); número de casos de arboviroses nos bimestres de 1 a 3 (janeiro a junho de 2014); temperatura nos bimestres 1 (janeiro de 2014), e de 5 a 6 (setembro a dezembro de 2014); pluviometria nos bimestres

de 1 a 2 (janeiro a abril de 2014); e velocidade dos ventos nos bimestres de 2 a 3 (março a junho de 2014) e de 5 a 6 (setembro a dezembro de 2014).

Na predição do bimestre de março a abril de 2015 (2015.2), conforme Tabela 9, os seguintes fatores tiveram destaque: posição geográfica (latitude); número de casos de arboviroses nos bimestres de 2 a 6 (março a dezembro de 2014); temperatura e pluviometria em praticamente todo o período; e velocidade dos ventos nos bimestres de 2 a 3 (março a abril de 2014) e de 5 a 1 (setembro de 2014 a fevereiro de 2015).

Ao prever-se o número de casos no bimestre de maio a junho de 2015 (2015.3), conforme Tabela 9, os seguintes fatores foram considerados mais relevantes: posição geográfica (latitude e longitude); número de casos de arboviroses nos bimestres de 3 a 6 (maio a dezembro de 2014) e 2 (março a abril de 2015); temperatura e velocidade dos ventos em todo o período; e pluviometria dos bimestres 4 a 2 (julho de 2014 a abril de 2015).

A predição do número de casos de arboviroses de julho a agosto de 2015 (2015.4), conforme Tabela 9, dependeu predominantemente dos seguintes fatores: posição geográfica (latitude e longitude); número de casos de arboviroses dos bimestres 4 a 6 (julho a dezembro de 2014) e 3 (maio a junho de 2015); temperatura e velocidade dos ventos em todo o período; e pluviometria nos bimestres 4 (julho a agosto de 2014) e de 6 a 3 (dezembro de 2014 a junho de 2015).

A predição do número de casos de arboviroses de setembro a outubro de 2015 (2015.5), conforme Tabela 9, dependeu predominantemente dos seguintes fatores: posição geográfica (latitude e longitude); número de casos de arboviroses nos bimestres de 5 a 6 (setembro a dezembro de 2014), 2 (março a abril de 2015) e 4 (julho a agosto de 2015); e pluviometria, temperatura e velocidade dos ventos em todo o período.

A predição do número de casos de arboviroses de novembro a dezembro de 2015 (2015.6), conforme Tabela 9, dependeu majoritariamente dos seguintes fatores: posição geográfica (latitude e longitude); número de casos de arboviroses nos bimestres de 6 a 2 (dezembro de 2014 a abril de 2015) e 5 (setembro a outubro de 2015); e pluviometria, temperatura e velocidade dos ventos em todo o período.

A Tabela 10 mostra os resultados de predição para os seis bimestres de 2016 considerando o ano de 2015. Analisando o resultado do consenso ou da votação no comitê de analistas artificiais têm-se, para a predição do bimestre de janeiro a fevereiro de 2016 (2016.1), os seguintes fatores considerados mais relevantes: posição geográfica (longitude); número de casos de arboviroses nos bimestres de 1 a 3 (janeiro a junho de 2015) e 6 (novembro a dezembro

de 2015); temperatura nos bimestres de 2 a 6 (março a dezembro de 2015); pluviometria nos bimestres de 1 a 3 (janeiro a junho de 2015) e de 5 a 6 (setembro a dezembro de 2015); e velocidade dos ventos em todo o período considerado para predição.

Na predição do bimestre de março a abril de 2016 (2016.2), conforme Tabela 10, fica evidente a importância dos seguintes fatores: temperatura nos bimestres de 2 a 5 (março a outubro de 2015); velocidade dos ventos nos bimestres de 2 a 4 (março a agosto de 2015) e 6 (novembro a dezembro de 2015); e pluviometria em todo o período. Curiosamente, a posição geográfica não foi considerada relevante.

Na predição do bimestre de maio a junho de 2016 (2016.3), conforme Tabela 10, os seguintes fatores foram considerados mais relevantes: posição geográfica (latitude e longitude); temperatura nos bimestres de 3 a 5 (maio a outubro de 2015) e de 1 a 2 (janeiro a abril de 2016); e pluviometria nos bimestres 3 (maio a junho de 2015) e de 5 a 2 (setembro de 2015 a abril de 2016). A velocidade dos ventos e o número de casos de arboviroses foram considerados importantes em todo o período considerado para predição.

Na predição do bimestre de julho a agosto de 2016 (2016.4), conforme Tabela 10, os seguintes fatores foram considerados mais relevantes: posição geográfica (latitude); temperatura nos bimestres de 4 a 5 (julho a outubro de 2015) e de 1 a 3 (janeiro a junho de 2016); velocidade dos ventos nos bimestres 4 (julho a agosto de 2015) e de 6 a 3 (novembro de 2015 a junho de 2016); e número de casos de arboviroses nos bimestres de 6 a 3 (novembro de 2015 a junho de 2016). A pluviometria foi considerada relevante para todo o período de predição.

Na predição do bimestre de setembro a outubro de 2016 (2016.5), conforme Tabela 10, os seguintes fatores foram considerados mais relevantes: posição geográfica (latitude e longitude); temperatura nos bimestres de 5 a 6 (setembro a dezembro de 2015) e de 2 a 3 (março a junho de 2016); velocidade dos ventos nos bimestres 5 (setembro a outubro de 2015) e de 1 a 4 (janeiro a agosto de 2016); e o número de casos nos bimestres 1 (janeiro a fevereiro de 2016) e de 3 a 4 (maio a agosto de 2016). A pluviometria foi considerada importante para todo o período de predição.

A predição do número de casos de arboviroses de novembro a dezembro de 2016 (2016.6), conforme Tabela 10, dependeu majoritariamente dos seguintes fatores: posição geográfica (latitude); número de casos de arboviroses nos bimestres de 6 a 3 (novembro de 2015 a junho de 2016) e 5 (setembro a outubro de 2016); pluviometria nos bimestres 2 (março a abril de 2016) e de 4 a 5 (julho a outubro de 2016). A temperatura e a velocidade dos ventos foram consideradas relevantes para todo o período de predição, o que mostra que a distribuição das

populações de mosquitos são fortemente influenciadas por esses fatores.

5.1.3 Análise espaço-temporal

A Tabela 11 mostra os resultados de predição para a regressão linear e o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais. Os melhores resultados que atendem ao critério de qualidade estabelecido neste trabalho, ou seja, índice de correlação R acima de 0,9 e erro RMSE (%) menor ou igual a 5%, estão destacados em vermelho, enquanto os melhores resultados que não atendem a esses critérios estão em azul. Como métricas de qualidade foram considerados o índice de correlação R, o erro RMSE (%) e os índices de correlação de Kendall e Spearman. Com relação aos resultados qualitativos, todos os mapas foram gerados com resolução de 120 dpi, escala 1:164399.

Tabela 11 – Resultados de predição para a regressão linear e o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais. Como métricas de qualidade foram considerados o índice de correlação R, o erro RMSE (%) e os índices de correlação de Kendall e Spearman. Os melhores resultados índice de correlação R acima de 0,9 e erro RMSE (%) menor ou igual a 5% estão em vermelho. Já os melhores resultados que não atendem a esse critério estão em azul.

		Regressão Linear				Random Forest 10 árvores			
		R	RMSE (%)	Kendall	Spearman	R	RMSE (%)	Kendall	Spearman
2014	1	0,9786	20,5835	0,8476	0,9577	0,9989	4,6734	0,9749	0,9985
	2	0,9726	23,2557	0,8046	0,9249	0,9985	5,5724	0,9654	0,9965
	3	0,9811	19,4476	0,7991	0,9240	0,9993	3,7901	0,9698	0,9982
	4	0,9466	32,2455	0,6494	0,7990	0,9990	4,5063	0,9589	0,9952
	5	1,0000	0,0000	1,0000	1,0000	0,9995	3,2371	0,9716	0,9980
	6	1,0000	0,0000	1,0000	1,0000	0,9997	2,3431	0,9758	0,9987
2015	1	0,8829	46,9592	0,7365	0,8923	0,9959	9,1897	0,9678	0,9951
	2	0,9599	28,1220	0,7825	0,9126	0,9990	4,4446	0,9668	0,9971
	3	0,9849	17,3334	0,8008	0,9229	0,9996	2,9670	0,9678	0,9974
	4	0,9345	35,6996	0,5560	0,7052	0,9986	5,2136	0,9080	0,9802
	5	0,9587	28,4837	0,7188	0,8717	0,9987	5,1249	0,9631	0,9955
	6	0,9436	33,1233	0,7998	0,9140	0,9983	5,9983	0,9688	0,9967
2016	1	0,9832	18,2522	0,8619	0,9624	0,9992	4,1462	0,9740	0,9982
	2	0,9834	18,1346	0,8608	0,9610	0,9993	3,8790	0,9752	0,9984
	3	0,9821	18,8461	0,8606	0,9624	0,9991	4,3817	0,9718	0,9973
	4	0,9867	16,2483	0,8761	0,9649	0,9995	3,1539	0,9772	0,9988
	5	0,9907	13,6027	0,9013	0,9816	0,9992	3,9360	0,9779	0,9999
	6	0,9909	13,4588	0,8869	0,9727	0,9995	3,3026	0,9748	0,9984

As Tabelas 12 e 13 apresentam a relação dos 94 bairros da Cidade do Recife com seus respectivos códigos e coordenadas de centroide nos sistemas XY e em latitude e longitude. Os códigos e as coordenadas dos centroides foram definidos e padronizados pela Prefeitura do Recife, sendo utilizados em todas os mapas digitais e sistemas de informações geográficas oficiais da Cidade do Recife. Os mapas temáticos utilizados neste trabalho utilizam esses códigos e coordenadas de centroides oficialmente definidos.

Tabela 12 – Relação de 46 dos 94 bairros da Cidade do Recife em ordem alfabética, dos Aflitos ao Ibura, com seus respectivos códigos e coordenadas de centroide nos sistemas XY e em latitude e longitude. Os códigos e as coordenadas dos centroides foram definidos e padronizados pela Prefeitura do Recife.

Bairro	Código	Y	X	latitude WGS84	longitude WGS84
Aflitos	132	9110884	290992	-8,03921	-34,89648
Afogados	779	9106661,8	289910,65	-8,07733	-34,90647
Água Fria	337	9113343	290787,7	-8,01697	-34,89823
Alto do Mandu	493	9112712,9	287498	-8,02252	-34,9281
Alto José Bonifácio	540	9113775	289181	-8,01299	-34,91279
Alto José do Pinho	523	9112846	289698	-8,02141	-34,90814
Alto Santa Terezinha	949	9113502	289671	-8,01548	-34,90836
Apipucos	507	9113095	286390	-8,01902	-34,93813
Areias	850	9104782	287566,7	-8,09423	-34,92781
Arruda	329	9112449,1	291553,7	-8,02508	-34,89132
Barro	876	9104647,9	285112,8	-8,09533	-34,95008
Beberibe	370	9114684	290898	-8,00485	-34,89718
Boa Viagem	205	9101556	290274	-8,12351	-34,90339
Boa Vista	86	9108394	292017	-8,06176	-34,88729
Bomba do Hemetério	426	9112839,1	290413	-8,02151	-34,90165
Bongi	809	9107910,4	288238,4	-8,06597	-34,92159
Brasília Teimosa	256	9106073	292947	-8,08278	-34,87895
Brejo da Guabiraba	612	9116097	287054	-7,99191	-34,93198
Brejo de Beberibe	574	9115386,1	287902,9	-7,99837	-34,92431
Cabanga	51	9106392	290894,3	-8,07981	-34,89756
Caçote	841	9104023,9	287032,9	-8,10106	-34,93269
Cajueiro	353	9113888	292172	-8,0121	-34,88566
Campina do Barreto	310	9113399	292574	-8,01653	-34,88203
Campo Grande	280	9111728	292874	-8,03165	-34,87938
Casa Amarela	434	9112482	288643	-8,02466	-34,91772
Casa Forte	442	9111470	288586	-8,03381	-34,91828
Caxangá	744	9111832	285184	-8,03039	-34,94912
Cidade Universitária	728	9109172	284920	-8,05442	-34,95163
Coelhos	60	9107816	291882	-8,06698	-34,88854
Cohab	884	9101433	284767	-8,12438	-34,95336
Coqueiral	922	9105465	283108	-8,08786	-34,96823
Cordeiro	680	9109547	287467	-8,05114	-34,92851
Córrego do Jenipapo	620	9114488	286627	-8,00644	-34,93592
Curado	752	9107651	283800	-8,06813	-34,96186
Derby	140	9108825	290723	-8,05781	-34,89901
Dois Irmãos	590	9114707	284762	-8,00438	-34,95283
Dois Unidos	396	9115967	289395	-7,99319	-34,91076
Encruzilhada	175	9111158	291613	-8,03675	-34,89084
Engenho do Meio	710	9108934	285940	-8,05662	-34,94239
Espinheiro	124	9110478	291571	-8,0429	-34,89125
Estância	833	9105647	287423	-8,0864	-34,92908
Fundão	345	9113575,6	291651,2	-8,0149	-34,89039
Graças	167	9110179,5	290528	-8,04556	-34,90072
Guabiraba	418	9120358	284635	-7,95329	-34,95374
Hipódromo	302	9111621,2	292022,4	-8,03258	-34,88711
Ibura	230	9101668,1	287104,4	-8,12236	-34,93214

Tabela 13 – Relação de 48 dos 94 bairros da Cidade do Recife em ordem alfabética, da Ilha do Leite ao Zumbi, com seus respectivos códigos e coordenadas de centroide nos sistemas XY e em latitude e longitude. Os códigos e as coordenadas dos centroides foram definidos e padronizados pela Prefeitura do Recife.

Bairro	Código	Y	X	latitude WGS84	longitude WGS84
Ilha do Leite	78	9107953,6	291260	-8,06571	-34,89418
Ilha do Retiro	760	9108164,9	290198,2	-8,06375	-34,9038
Ilha Joana Bezerra	43	9107504	290753	-8,06975	-34,8988
Imbiribeira	264	9103837,8	289440,3	-8,10284	-34,91086
Ipsep	213	9103015,3	288448	-8,11023	-34,9199
Iputinga	698	9110794,2	286487,9	-8,03983	-34,93734
Jaqueira	159	9111235,3	290316,6	-8,036	-34,9026
Jardim São Paulo	868	9106260,1	285989	-8,0808	-34,94206
Jiquiá	825	9105794	288086	-8,0851	-34,92306
Jordão	221	9100464	286239	-8,1332	-34,94005
Linha do Tiro	388	9114147	290008	-8,00967	-34,90527
Macaxeira	582	9113886	287156	-8,0119	-34,93115
Madalena	647	9109267	289691	-8,05377	-34,90835
Mangabeira	515	9112518	290270	-8,0244	-34,90296
Mangueira	795	9106721	288365	-8,07673	-34,92049
Monteiro	485	9112291	287425	-8,02634	-34,92878
Morro da Conceição	531	9113078	288997	-8,01929	-34,91449
Mustardinha	787	9107205	288581	-8,07236	-34,91851
Nova Descoberta	566	9115154	287233	-8,00044	-34,9304
Paissandu	94	9108269	290754	-8,06284	-34,89875
Parnamirim	469	9111464	289457	-8,0339	-34,91038
Passarinho	400	9117218	288021	-7,98182	-34,92316
Pau Ferro	639	9118853	282335	-7,96679	-34,97466
Paixinhos	299	9113220	293141	-8,01818	-34,8769
Pina	248	9104566	291222	-8,09633	-34,89466
Poço da Panela	477	9111320	287974	-8,03514	-34,92384
Ponto de Parada	930	9111936	291363	-8,02971	-34,89308
Porto da Madeira	361	9114446	291875	-8,00704	-34,88833
Prado	663	9108586	289083	-8,0599	-34,9139
Recife	19	9108963,5	293824,9	-8,05669	-34,87087
Rosarinho	183	9111653	290839	-8,03225	-34,89784
San Martin	817	9107439	287447	-8,0702	-34,92879
Sancho	906	9105834	283826	-8,08455	-34,9617
Santana	450	9110875	288843	-8,0392	-34,91598
Santo Amaro	108	9109594	292657	-8,05094	-34,88144
Santo Antônio	27	9108074	293047	-8,0647	-34,87796
São José	35	9107439	292735	-8,07042	-34,88082
Sítio dos Pintos	604	9113960	284132	-8,01111	-34,95857
Soledade	116	9109031	291653	-8,05599	-34,89057
Tamarineira	191	9111623	290489	-8,0325	-34,90102
Tejipió	892	9105574	284554	-8,08694	-34,95511
Torre	655	9110512	289784	-8,04252	-34,90746
Torreão	272	9110768	292379	-8,04031	-34,88391
Torrões	701	9108525	286617	-8,06035	-34,93627
Totó	914	9106121	283026	-8,08192	-34,96895
Várzea	736	9110202	282982	-8,04503	-34,96917
Vasco da Gama	558	9113906	288489	-8,01178	-34,91906
Zumbi	671	9109277	288621	-8,05363	-34,91806

As Figuras 43, 44 e 45; 46, 47 e 48; e 49, 50 e 51 mostram os resultados de predição sobre o mapa de Recife para os anos de 2014, 2015 e 2016, considerando as predições por regressão linear, respectivamente. Já as Figuras 52, 53 e 54; 55, 56 e 57; e 58, 59 e 60 mostram os resultados de predição sobre o mapa de Recife para os anos de 2014, 2015 e 2016, considerando as predições por Random Forest com 10 árvores, nesta ordem.

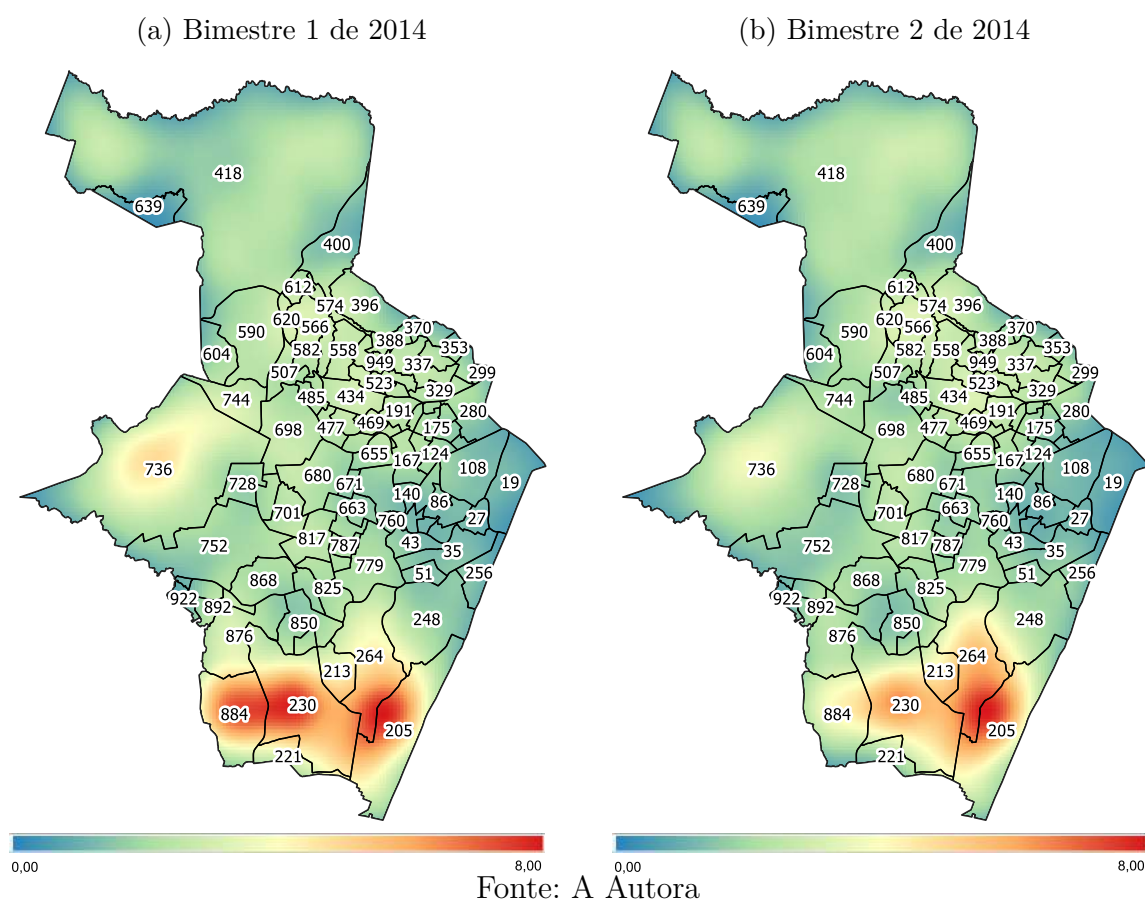
Considerando a predição usando regressão linear, a Figura 43(a) mostra uma distribuição razoável de casos de arboviroses no bimestre 1 de 2014, do sul ao norte e ao oeste, com concentração maior em bairros do sul, especificamente Ibura, Boa Viagem e Cohab, e no oeste, na Várzea. No bimestre 2, cf. Figura 43(b), o número de casos na Várzea, no Ibura e na Cohab diminuem um pouco, mas aumenta em Boa Viagem na direção do Ipsep e do Pina. No bimestre 3, de acordo com a Figura 44(a), o número de casos diminui bastante, mas se mantém concentrado no Ibura, espalhando-se um pouco entre Cohab, Jordão e Boa Viagem. No bimestre 4 o número de casos diminui nos bairros vizinhos, mas segue concentrado no Ibura, como mostrado na Figura 44(b). A situação permanece aproximadamente constante nos bimestres 5 e 6, cf. Figuras 45(a) e 45(b), respectivamente.

A predição do bimestre 1 de 2015 usando regressão linear, conforme a Figura 46(a), mostra que o número de casos aumentou razoavelmente em relação ao mesmo período de 2014: aumentam os casos na Guabiraba e na Várzea e surge uma grande concentração de casos de arboviroses na região centro-norte, passando por Casa Amarela, Vasco da Gama, Brejo da Guabiraba e Dois Unidos e se espalhando por Parnamirim, Arruda, Cajueiro, Beberibe, Iputinga, Sítio dos Pintos e Caxangá. Também há uma grande concentração de casos nos bairros do sul, especialmente Cohab, Ibura, Boa Viagem e Ipsep. A situação fica mais controlada no bimestre 2, conforme a Figura 46(b), com um pico de casos concentrado no bairro do Ibura e se espalhando de forma bem mais suave entre Cohab, Ibura, Ipsep e Boa Viagem. Nos bimestres 3 e 4, a situação fica um pouco mais controlada, com um recuo no número de casos, agora concentrados no Ibura e na Cohab, de acordo com as Figuras 47(a) e 47(b), nesta ordem. A situação do bimestre 5 é muito similar à do bimestre 4, de acordo com a Figura 48(a), mas no bimestre 6 passam a ocorrer alguns casos se espalhando por Ipsep, Pina e Afogados, como visto na Figura 48(b).

A predição do bimestre 1 de 2016 usando regressão linear, conforme a Figura 49(a), ilustra uma situação um pouco diferente de 2014 e 2015: apesar de ainda se ter, para o começo do ano, uma concentração de casos nos bairros do sul de Recife, especificamente Cohab, Ibura, Ipsep e Boa Viagem, espalhando-se nas direções do Jordão e do Pina, desta vez aumentam

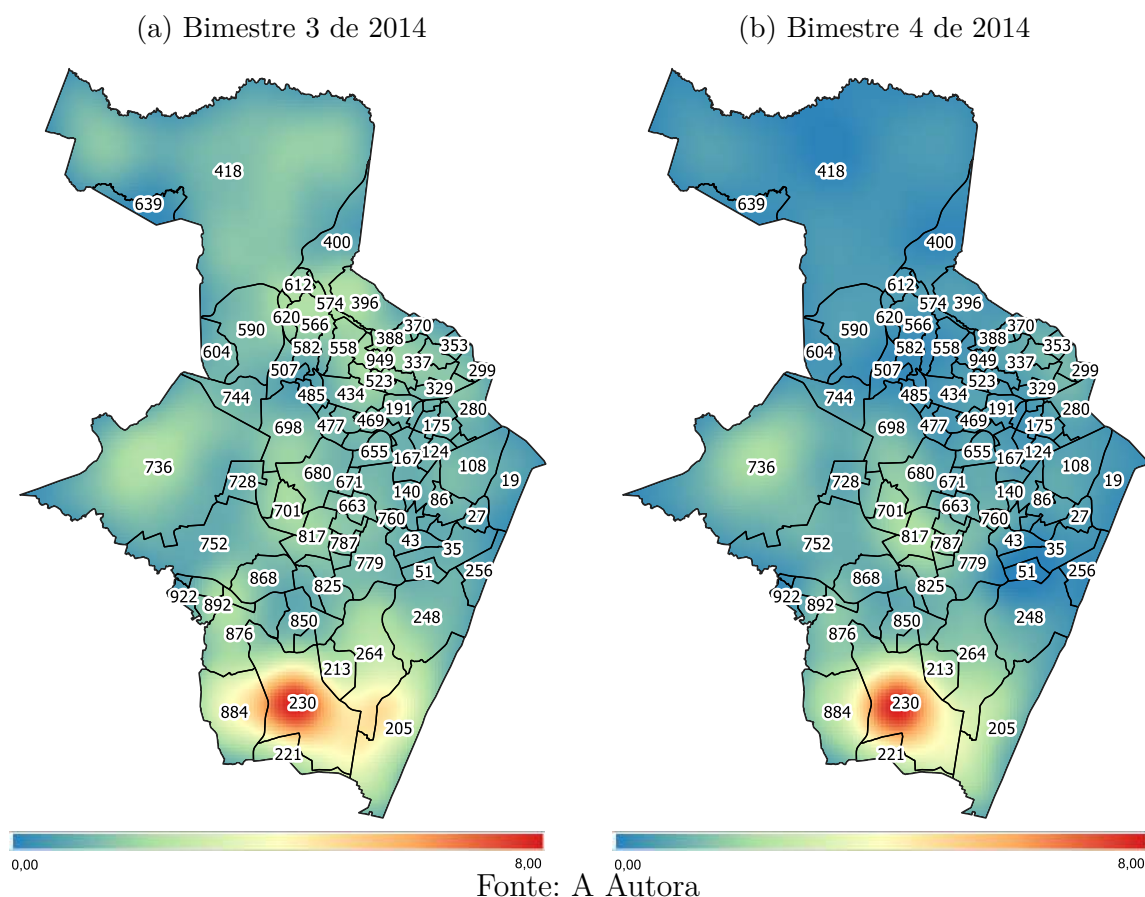
muito os casos no oeste, particularmente na Várzea. No bimestre 2, a situação no sul do município melhora um pouco, com a diminuição do número de casos, mas aumentam os casos na Várzea, conforme Figura 49(b). No bimestre 3, o número de casos na Várzea diminui, mas aumentam os casos no Ibura e na Cohab e surgem mais casos no centro-oeste, nos bairros de Cordeiro, Torrões e Iputinga, conforme a Figura 50(a). No bimestre 4, aumenta o número de casos nos bairros do centro-oeste e no sul, nos bairros de Boa Viagem, Ipsep e Pina. As Figuras 51(a) e 51(b) mostram um maior controle dos casos nos bimestres 5 e 6, respectivamente, terminando o ano de forma similar ao final de 2015 e 2014, exceto por uma quantidade maior de casos na Várzea.

Figura 43 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 43(a)) e 2 (março a abril, cf. Figura 43(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



Os resultados de predição usando Random Forest tendem a concordar qualitativamente com os resultados de regressão linear, mas mostram áreas mais bem concentradas e com fronteiras mais suaves. Como as métricas da Tabela 11, especialmente o erro RMSE (%),

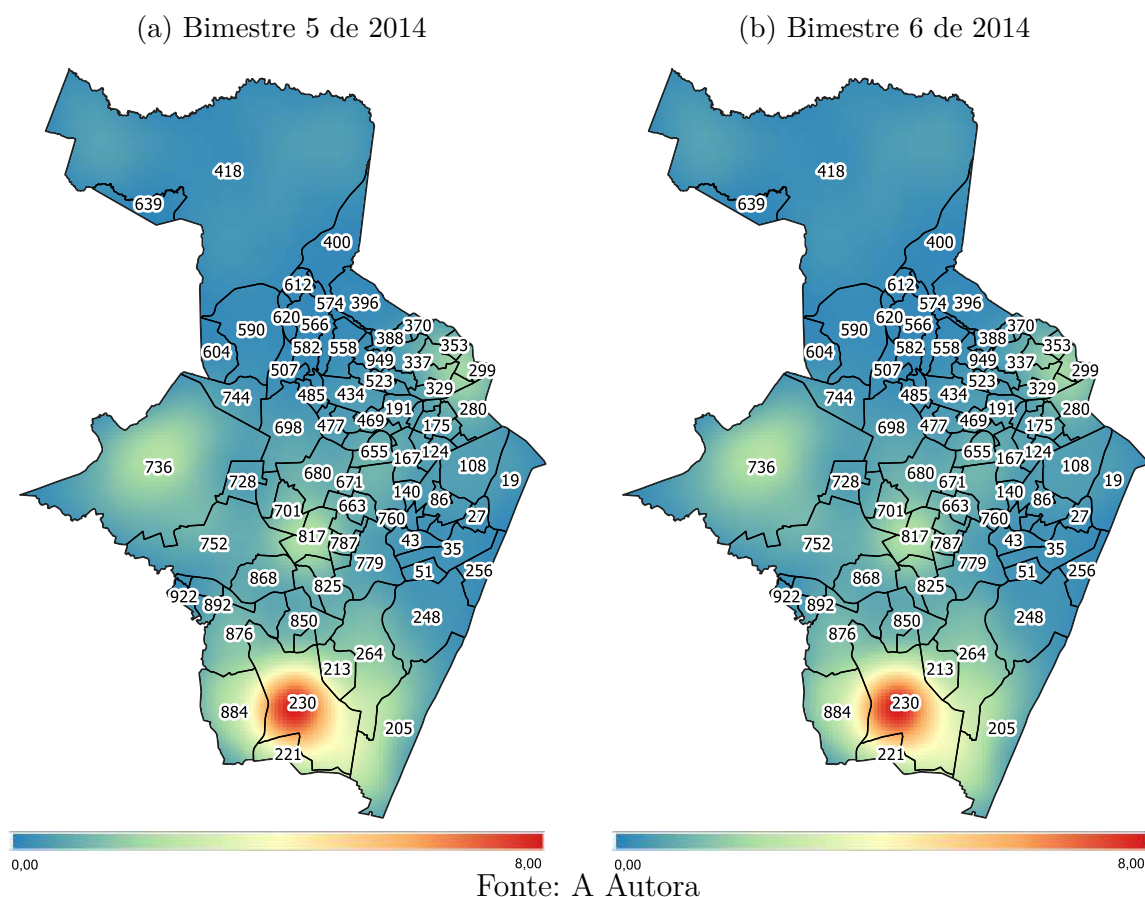
Figura 44 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 44(a)) e 4 (julho a agosto, cf. Figura 44(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



apontam para a superioridade do Random Forest sobre a regressão linear para este problema, é de se esperar que os resultados das predições com regressão linear apontem algumas regiões um pouco melhores e outras um pouco piores do que realmente são. A predição do bimestre 1 de 2014, mostrada na Figura 52(a), concorda no geral com a predição da Figura 43(a), com concentração de casos na região sul da cidade, mas a situação da Várzea aparece bem mais claramente, ficando evidente que há uma quantidade de casos razoável nesse bairro. As predições com Random Forest para os demais bimestres de 2014 tendem a coincidir de maneira qualitativa com os resultados com regressão linear, conforme Figuras 52(b), 53(a), 53(b), 54(a) e 54(b) quando comparadas com as Figuras 43(b), 44(a), 44(b), 45(a) e 45(b), nesta ordem.

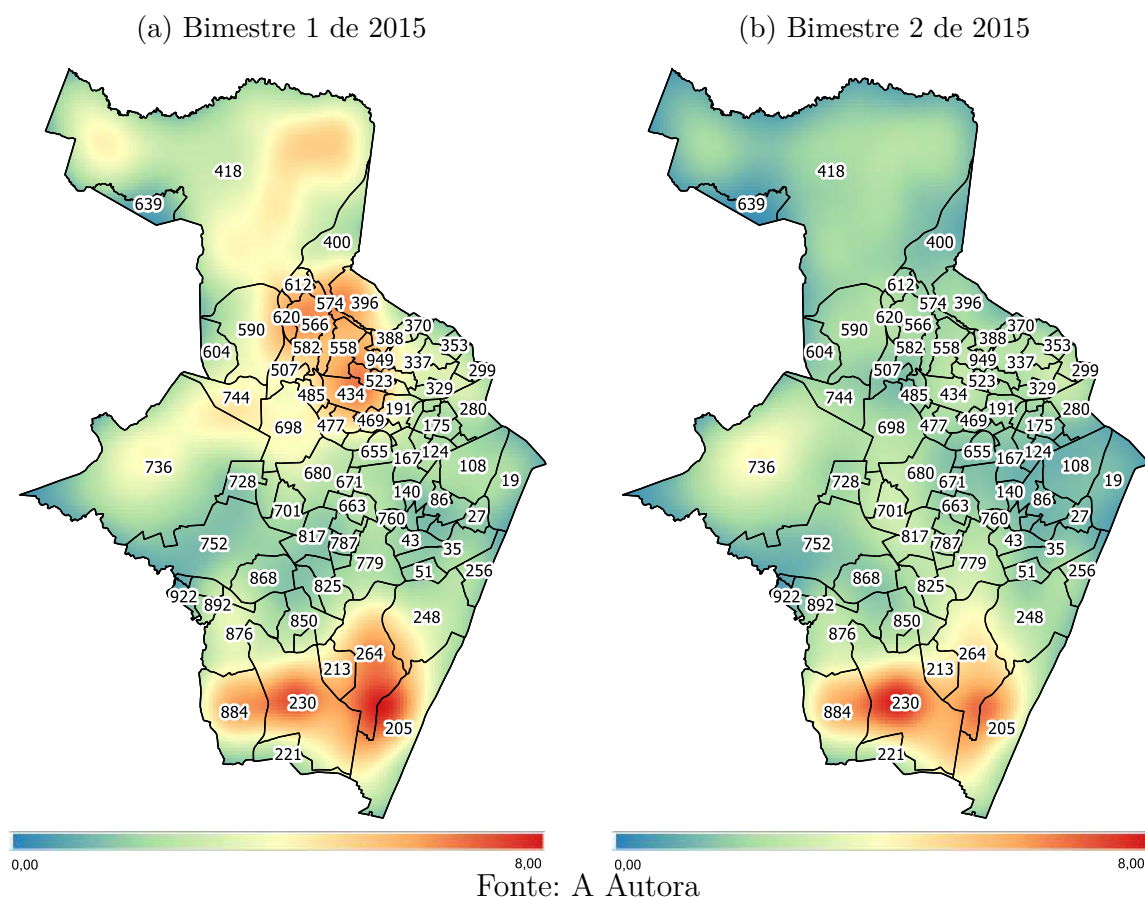
A predição feita com Random Forest para o bimestre 1 de 2015 mostra com mais clareza o surgimento de casos de arboviroses nos bairros do centro-norte com relação à predição

Figura 45 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 45(a)) e 6 (novembro a dezembro, cf. Figura 45(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



com regressão linear, conforme comparação entre as Figuras 55(a) e 46(a), que ilustram os resultados para Random Forest e regressão linear, respectivamente. Além disso, o resultado com Random Forest ilustra melhor a existência de dois picos diferentes nos bairros do centro-norte: um entre Vasco da Gama e Casa Amarela, e outro entre Dois Irmãos, Dois Unidos e Brejo da Guabiraba. A predição para o bimestre 2 mostra um maior controle dos casos de arboviroses, com diminuição geral dos casos, embora ainda persista uma concentração de casos no Ibura, espalhando-se entre Cohab, Jordão, Ipsep, Boa Viagem e na direção do Pina, cf. Figura 55(b). A Figura 56(a), em relação com a Figura 55(b), mostra que o preditor com Random Forest conseguiu mostrar diminuição da área de casos em Boa Viagem no bimestre 3 quando comparado com o 2, enquanto com a regressão linear isso não fica claro, conforme a Figura 47(a) comparada com a 46(b). A Figura 56(b) mostra que, para o bimestre 4, o número de casos na Cohab superou o do Ibura. Essa situação se mantém constante no bimestre 5,

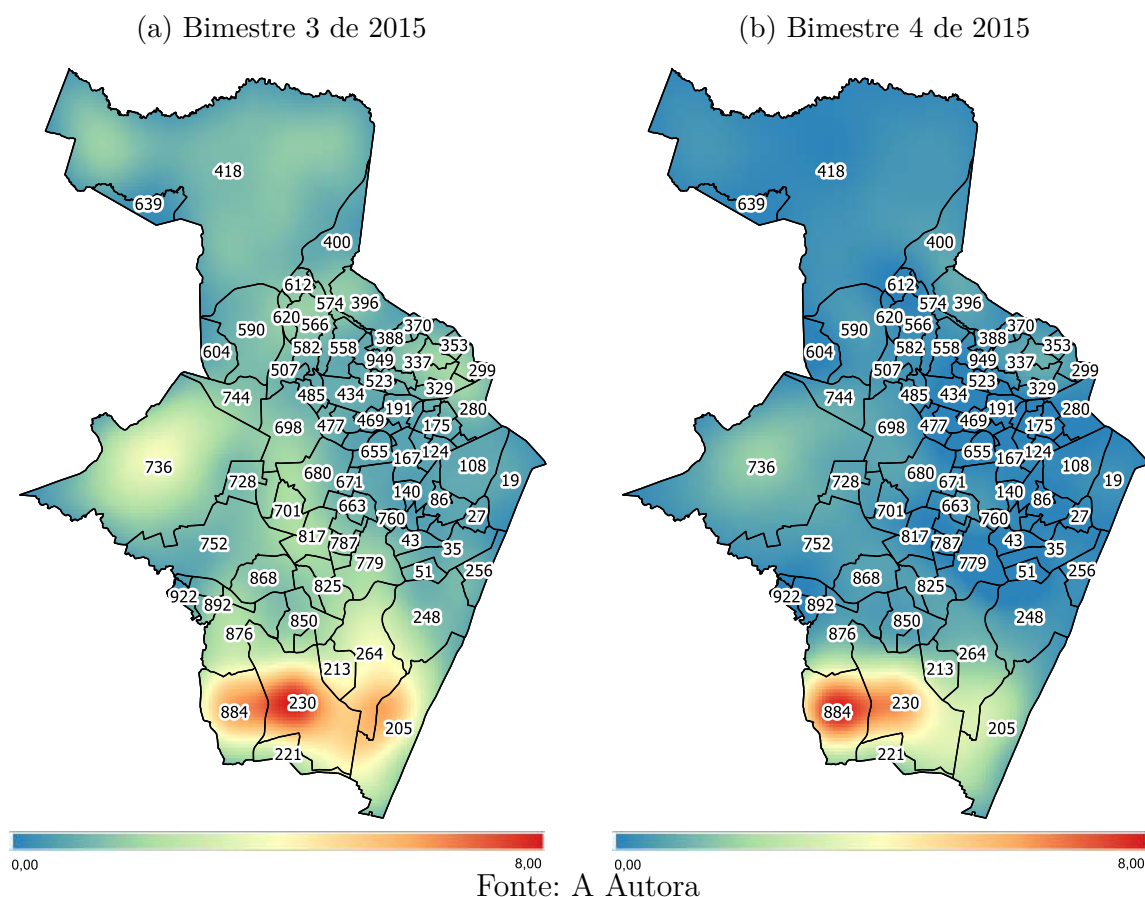
Figura 46 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 46(a)) e 2 (março a abril, cf. Figura 46(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



como ilustra a Figura 57(a). A predição do bimestre 6 mostra que surgiram mais casos se espalhando do sul ao centro, do Ibura, que volta a ter o maior número de casos, até Afogados, Mustardinha, Bongü, Brasília Teimosa e São José, passando por Ipsep, Boa Viagem e Pina, conforme mostrado na Figura 57(b). Nesse caso também a situação foi melhor ilustrada pela predição com Random Forest, conforme a Figura 57(b), do que com a regressão linear, como ilustrado na Figura 48(b).

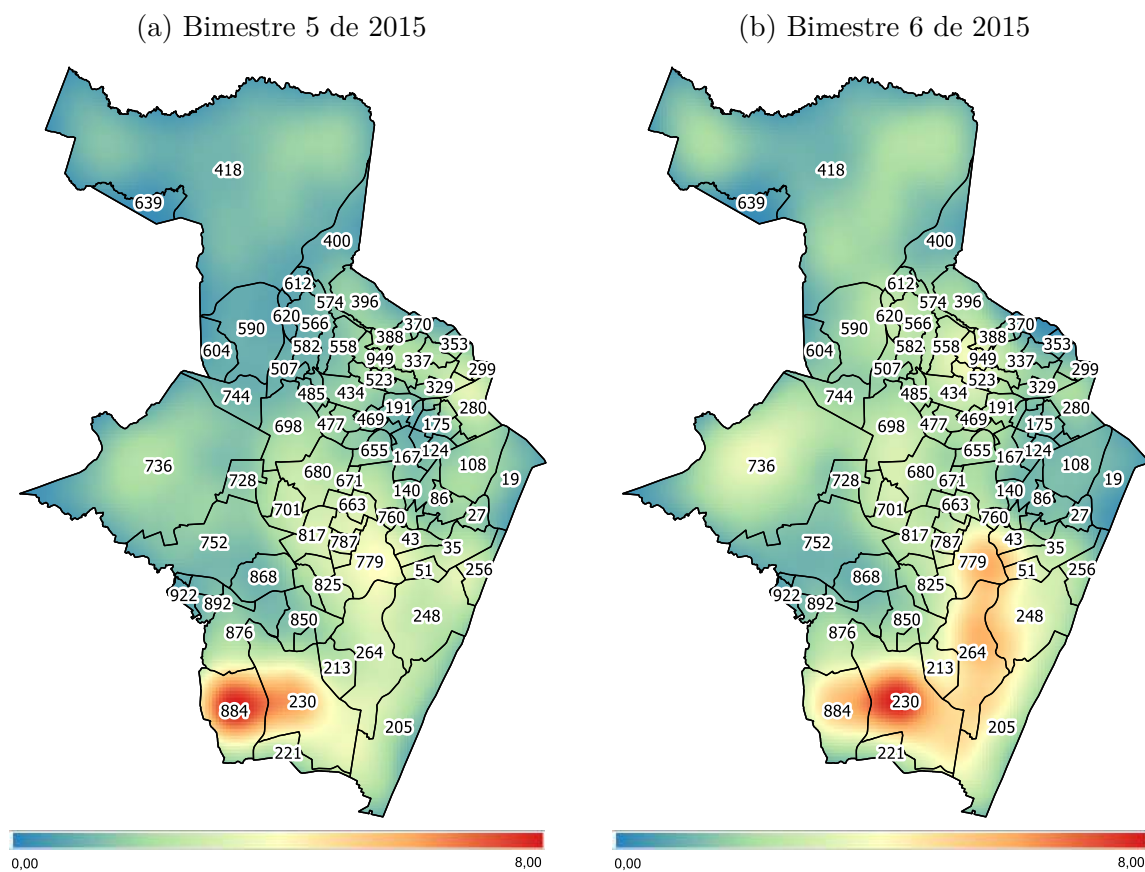
A predição feita com Random Forest para o bimestre 1 de 2016 mostra muito claramente a Várzea como um pico de casos de arboviroses, junto com o Ibura, a Cohab e, de forma mais fraca, Boa Viagem, conforme Figura 58(a). A predição do bimestre 2, ilustrada na Figura 58(b), mostra um maior controle no número de casos nos bairros da zona sul, mas a Várzea persiste praticamente da mesma forma que no bimestre 1, com alguns poucos casos espalhados nos bairros do centro-oeste. A predição do bimestre 3 mostra novamente picos de casos na

Figura 47 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 47(a)) e 4 (julho a agosto, cf. Figura 47(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



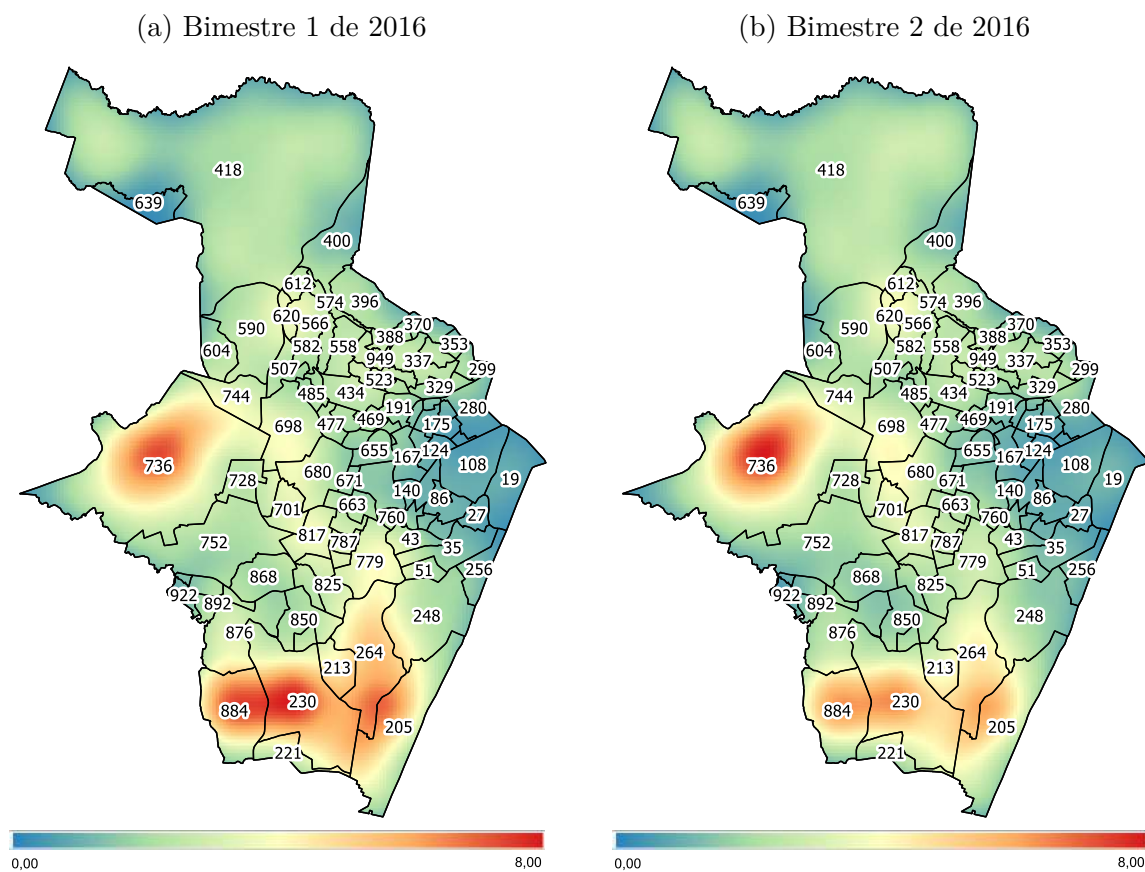
Cohab e no Ibura e um espalhamento leve de casos do Ibura em direção a Boa Viagem, Ipsep e Pina, mas com um maior controle de casos na Várzea e nos bairros da zona centro-oeste, conforme a Figura 59(a). A predição do bimestre 4 ilustra o aumento de casos na direção de Boa Viagem, Ipsep e Pina, confirmando a tendência observada no bimestre 3, de acordo com a Figura 59(b). A predição para o bimestre 5 mostra uma ligeira melhora em relação ao bimestre 4, conforme a Figura 60(a). A predição para o bimestre 6 mostra uma diminuição geral dos casos, mas persiste uma concentração de casos na Cohab e no Ibura, e o número de casos da Várzea aumenta um pouco, conforme Figura 60(b).

Figura 48 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 48(a)) e 6 (novembro a dezembro, cf. Figura 48(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



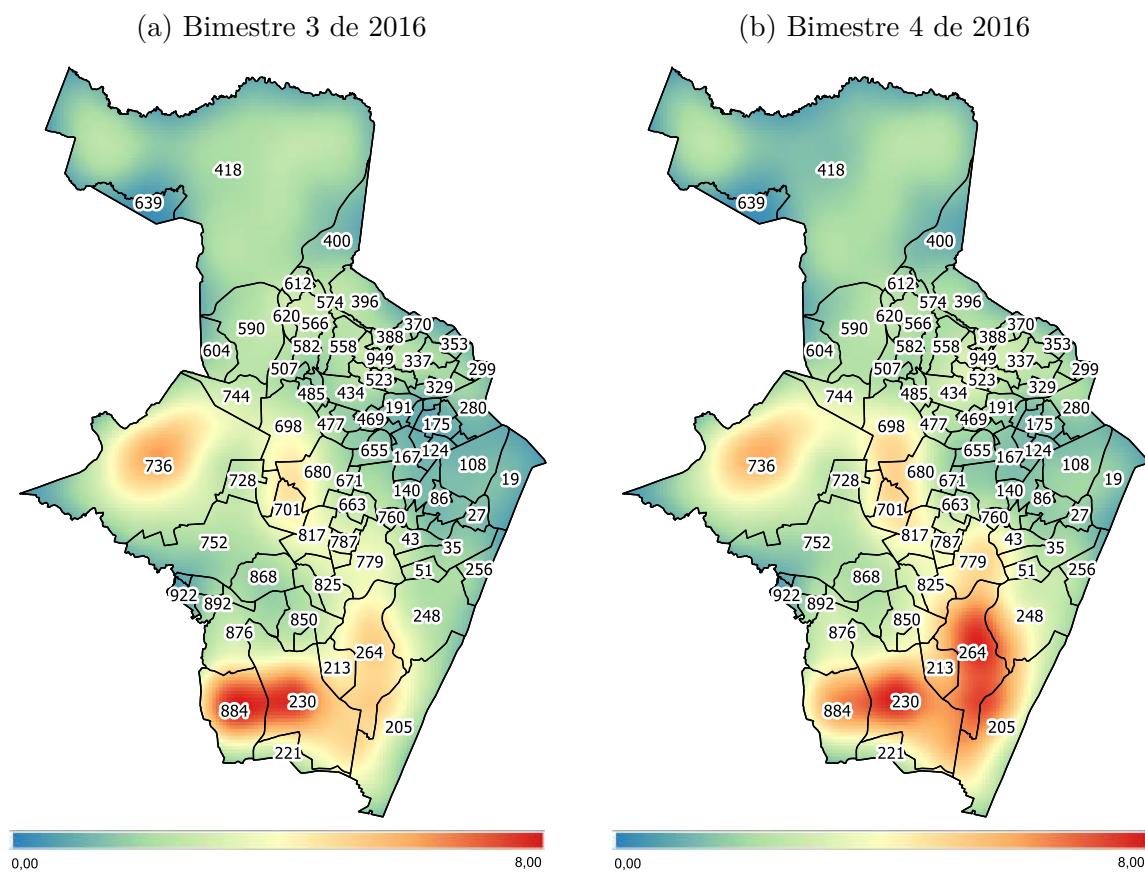
Fonte: A Autora

Figura 49 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 49(a)) e 2 (março a abril, cf. Figura 49(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



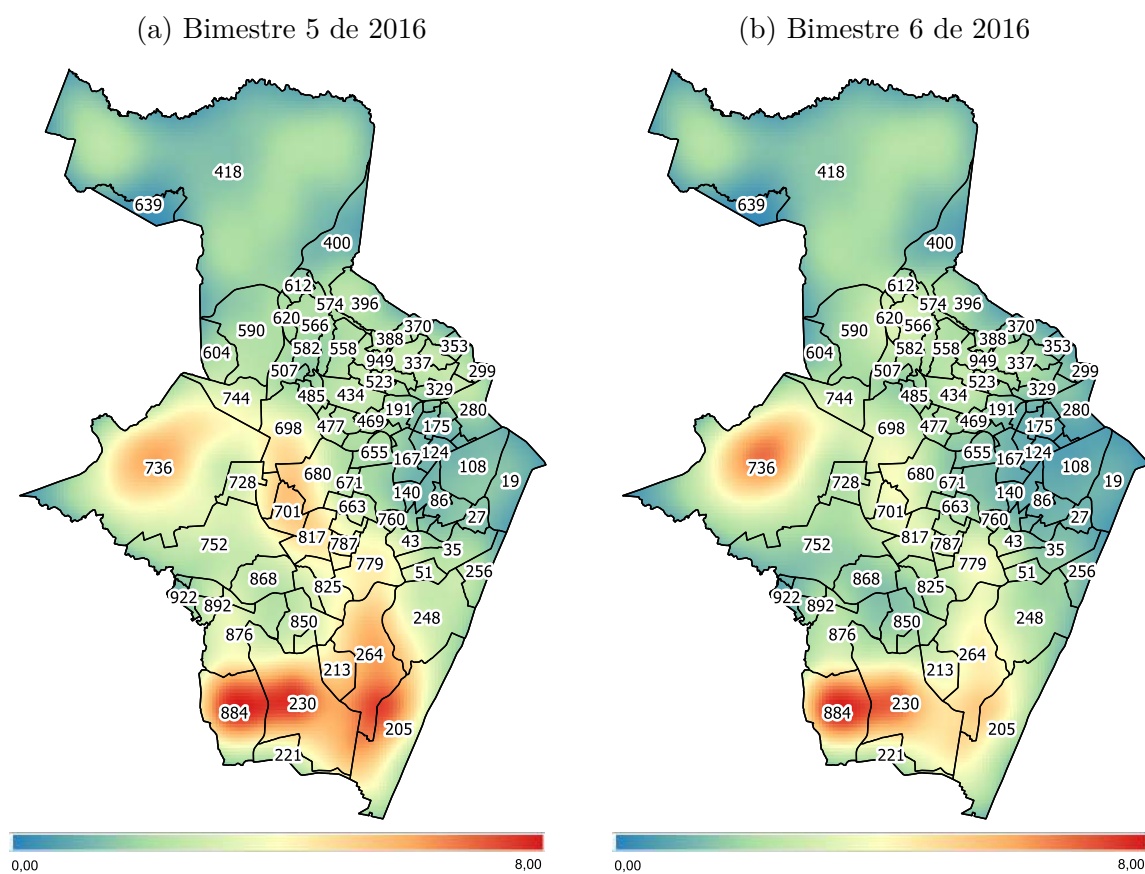
Fonte: A Autora

Figura 50 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 50(a)) e 4 (julho a agosto, cf. Figura 50(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



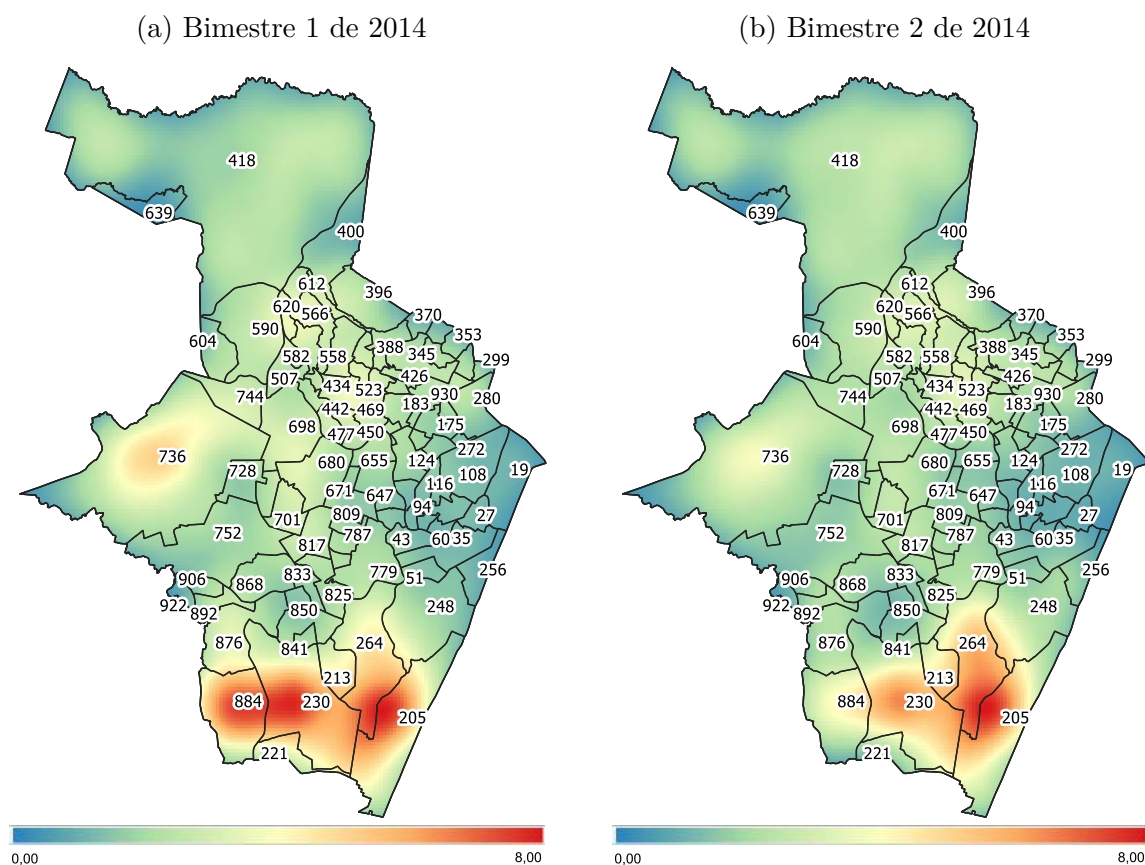
Fonte: A Autora

Figura 51 – Resultados de predição para a regressão linear, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 51(a)) e 6 (novembro a dezembro, cf. Figura 51(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



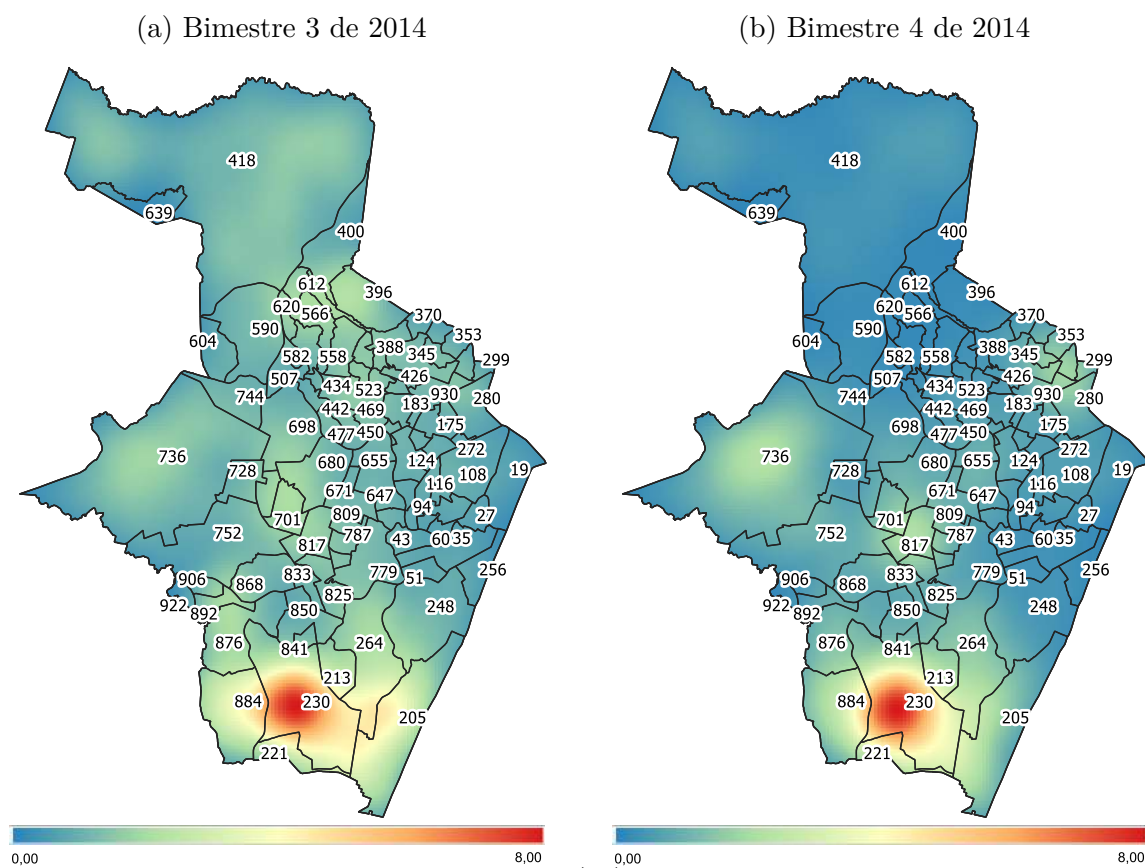
Fonte: A Autora

Figura 52 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 52(a)) e 2 (março a abril, cf. Figura 52(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



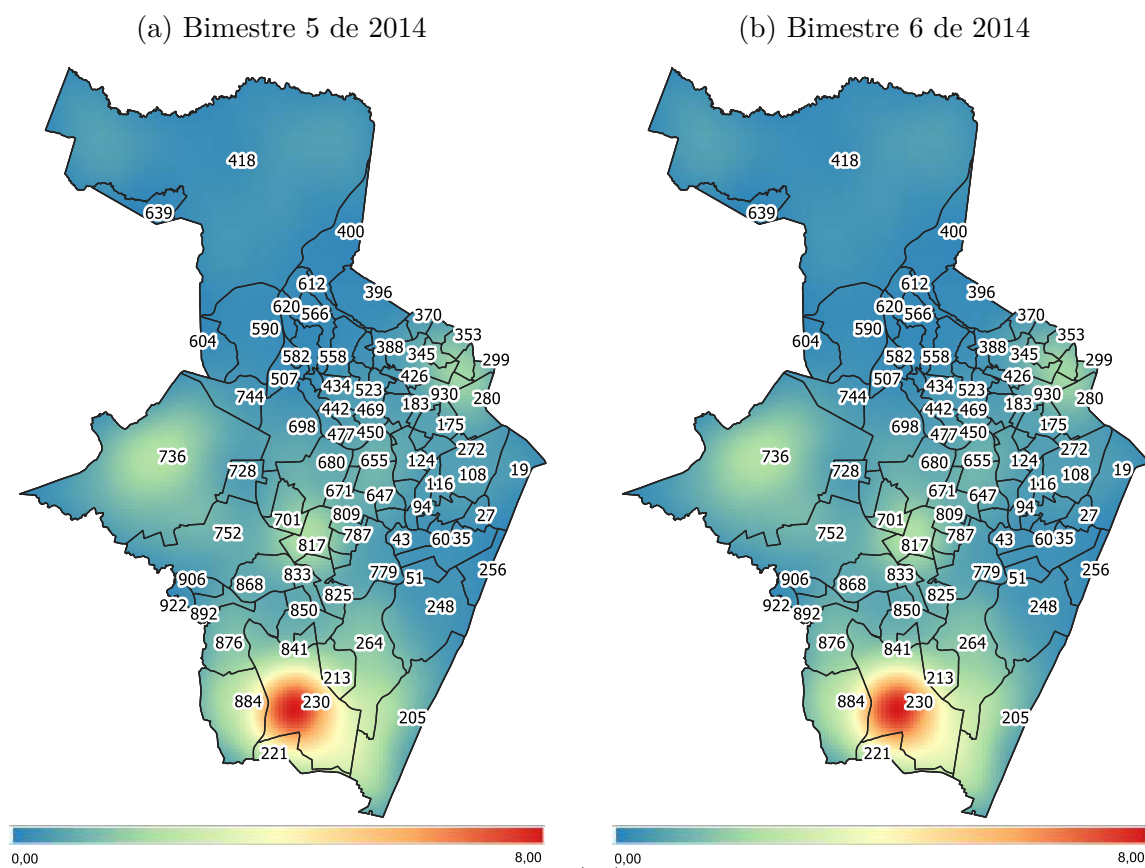
Fonte: A Autora

Figura 53 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 53(a)) e 4 (julho a agosto, cf. Figura 53(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



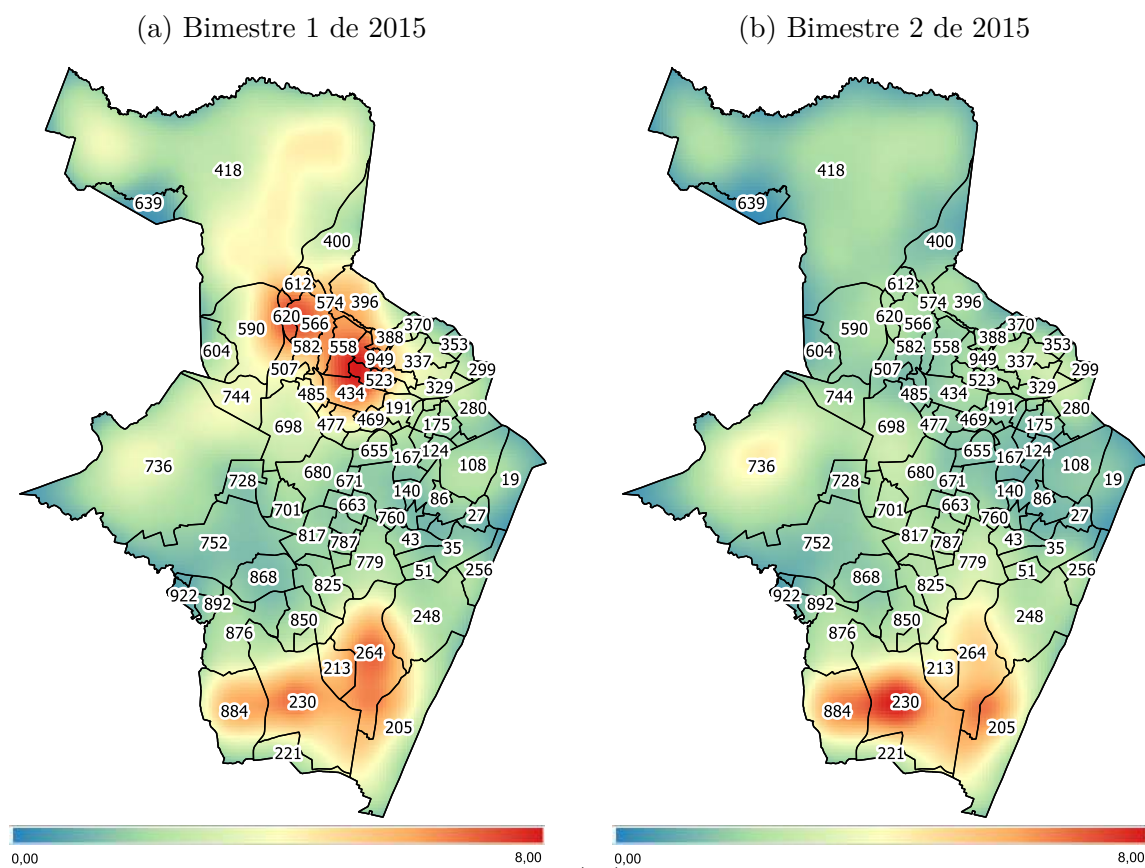
Fonte: A Autora

Figura 54 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 54(a)) e 6 (novembro a dezembro, cf. Figura 54(b)), para o ano de 2014. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



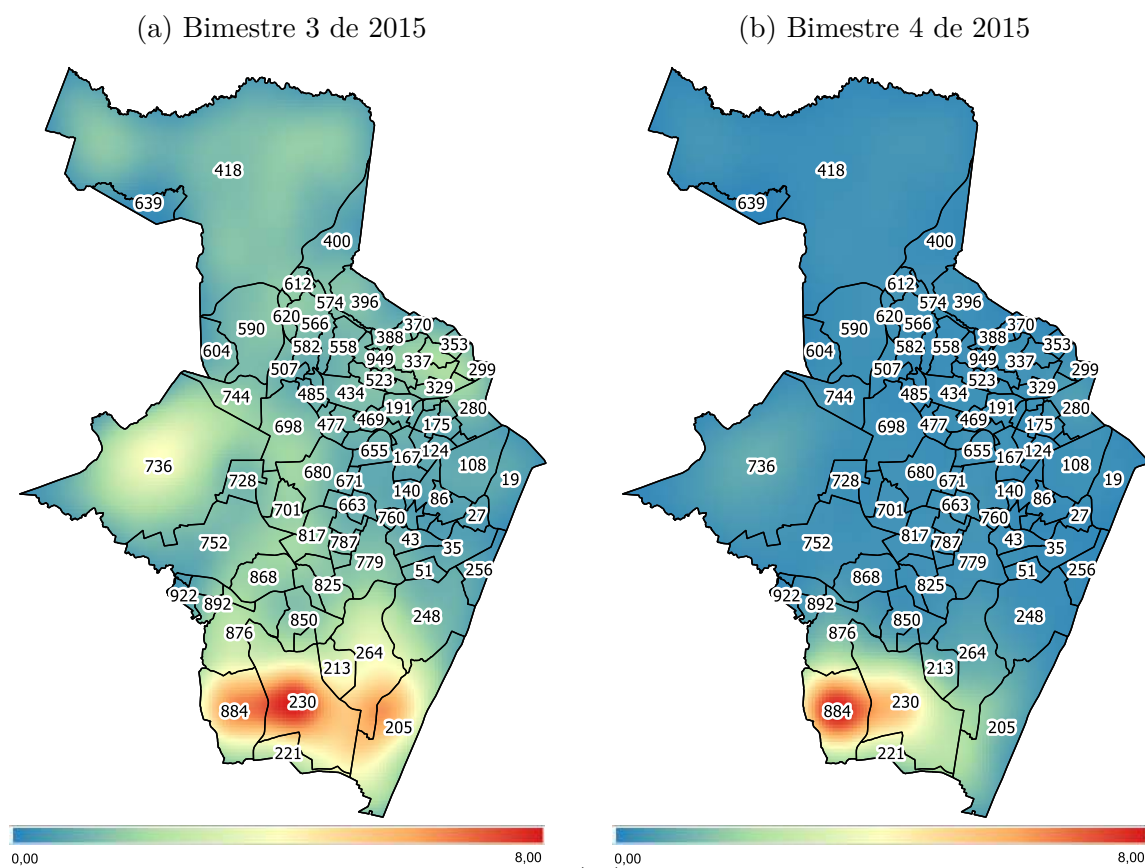
Fonte: A Autora

Figura 55 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 55(a)) e 2 (março a abril, cf. Figura 55(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



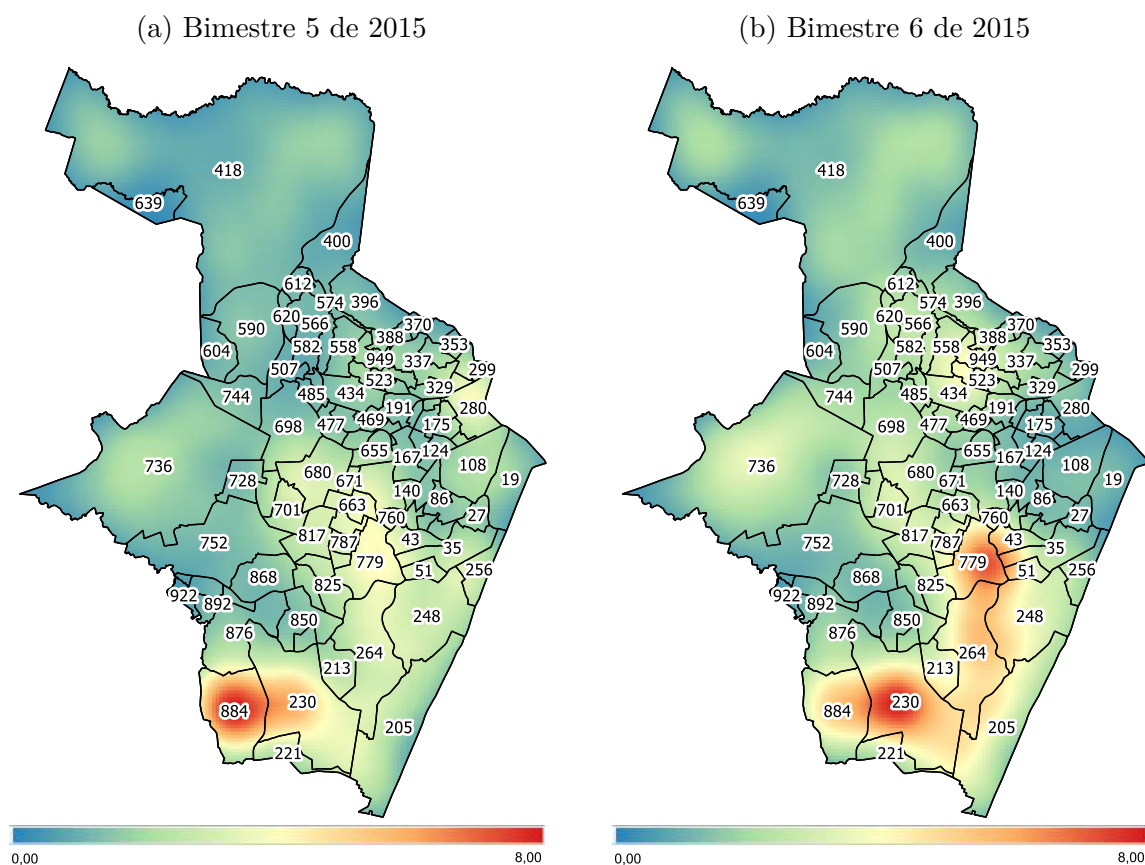
Fonte: A Autora

Figura 56 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 56(a)) e 4 (julho a agosto, cf. Figura 56(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



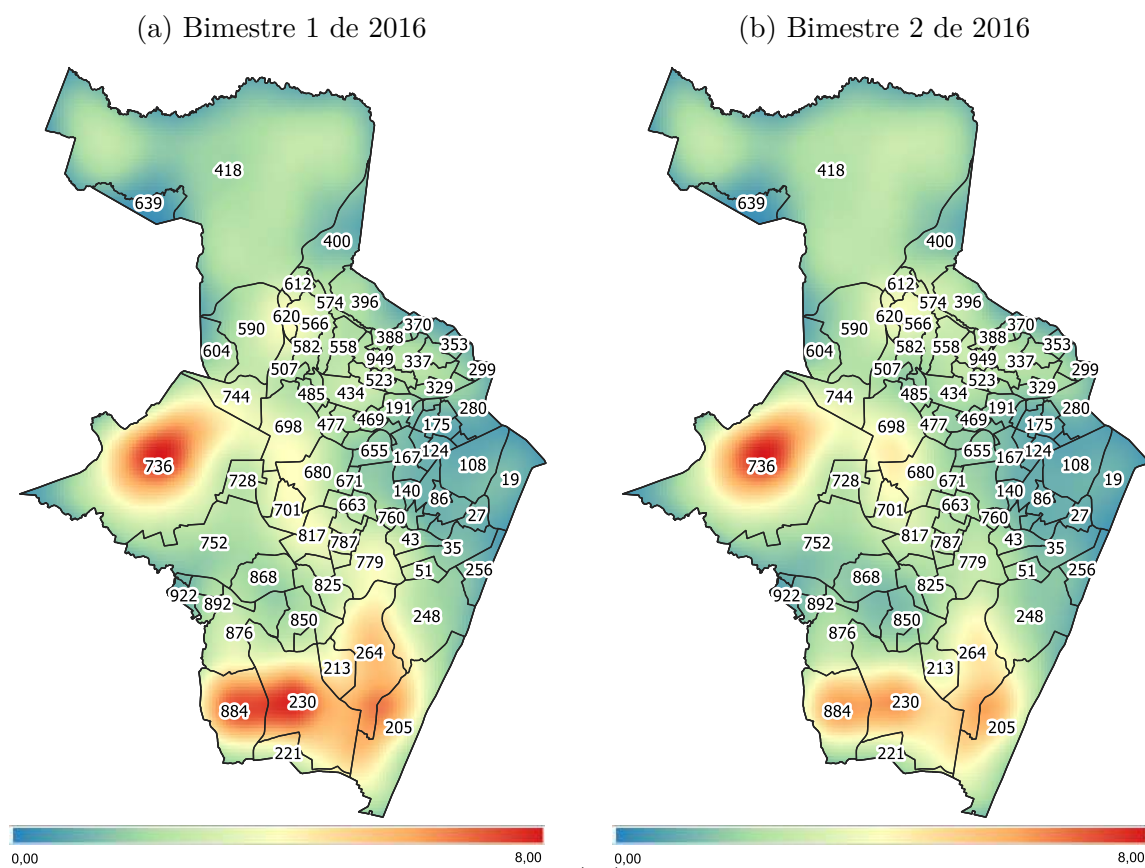
Fonte: A Autora

Figura 57 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 57(a)) e 6 (novembro a dezembro, cf. Figura 57(b)), para o ano de 2015. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



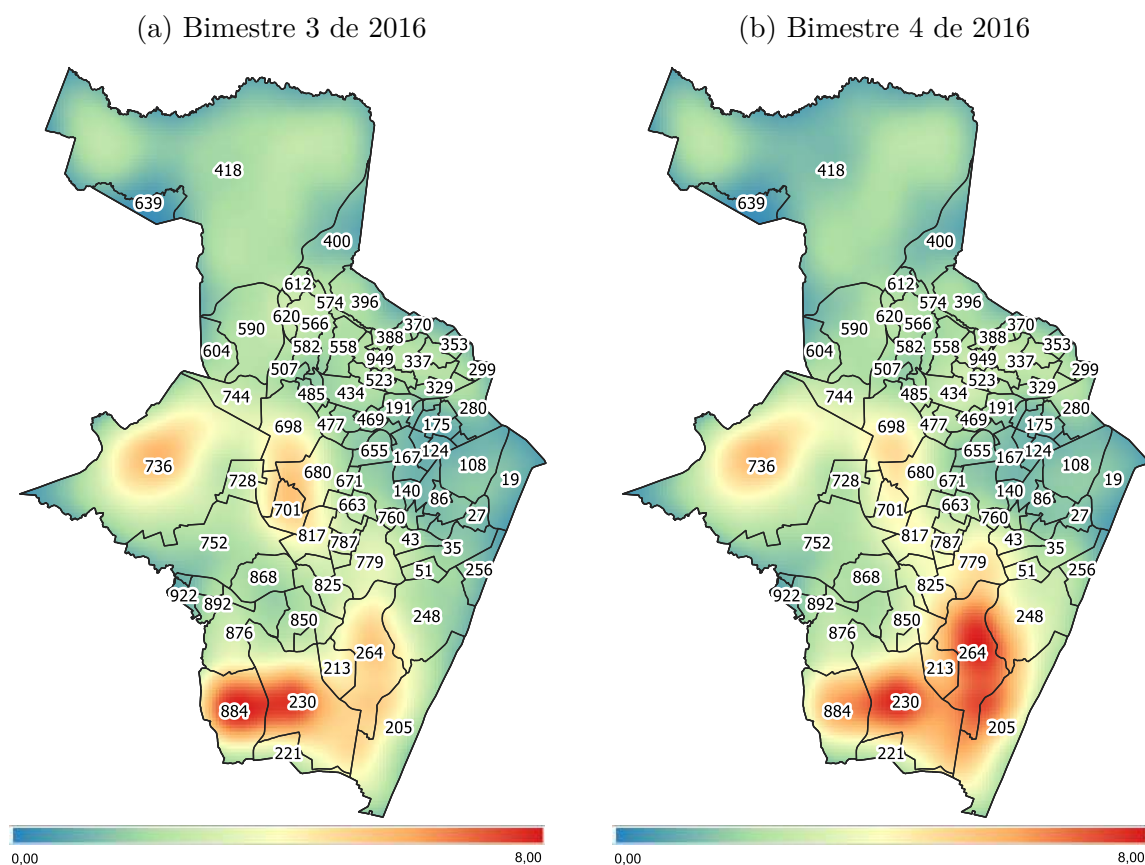
Fonte: A Autora

Figura 58 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 1 (janeiro a fevereiro, cf. Figura 58(a)) e 2 (março a abril, cf. Figura 58(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



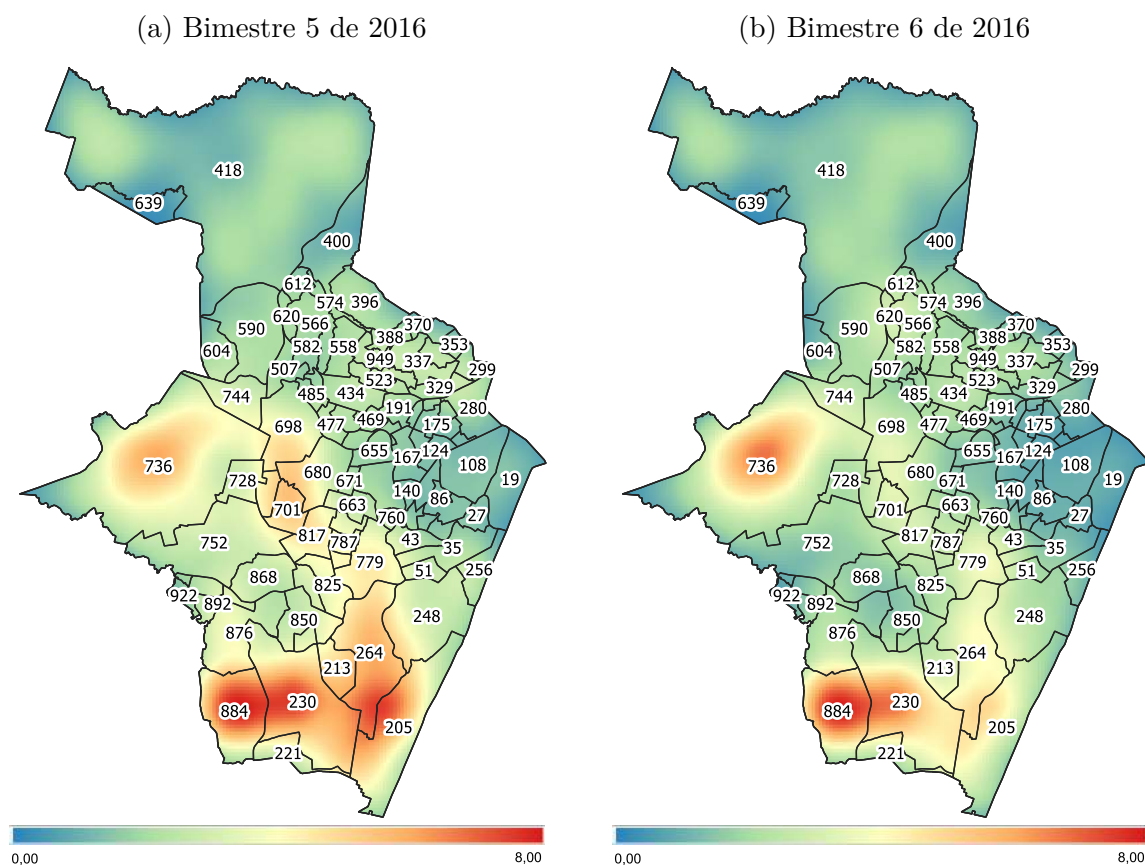
Fonte: A Autora

Figura 59 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 3 (maio a junho, cf. Figura 59(a)) e 4 (julho a agosto, cf. Figura 59(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



Fonte: A Autora

Figura 60 – Resultados de predição para o Random Forest com 10 árvores, considerando conjuntos de treino de 4665 instâncias e de teste de 1555 instâncias, obtidos a partir de interpolação dos dados reais, para os bimestres 5 (setembro a outubro, cf. Figura 60(a)) e 6 (novembro a dezembro, cf. Figura 60(b)), para o ano de 2016. Os códigos dos bairros estão descritos nas Tabelas 12 e 13. A resolução é de 120 dpi, escala 1:164399.



Fonte: A Autora

5.2 CONSIDERAÇÕES

Neste capítulo foram apresentados os resultados parciais da pesquisa de doutorado, correspondendo à validação da metodologia proposta de predição de casos de arboviroses a partir de informações georreferenciadas de pluviometria, temperatura, velocidade dos ventos e casos passados de arboviroses, para ciclos de seis bimestres, ou seja, um ano. Os experimentos evidenciam que bons resultados de predição podem ser conseguidos com redes neurais perceptron multi-camadas, máquinas de vetor de suporte, máquinas de aprendizado extremo e Random Forests. Considerando as métricas de qualidade de regressão e o tempo de treinamento do modelo preditivo, o Random Forest foi considerado o mais efetivo. Esses resultados são superiores ao método padrão de regressão, com regressor linear.

Os resultados qualitativos de predição em imagens de mapa de calor usando Random Forest tendem a concordar qualitativamente com os resultados de regressão linear, mas mostram áreas mais bem concentradas e com fronteiras mais suaves. Uma vez que as métricas apontam para a superioridade do Random Forest sobre a regressão linear para este problema, é esperado que os resultados das predições com regressão linear apontem algumas regiões um pouco melhores e outras um pouco piores do que realmente são.

A abordagem utilizada também se mostrou mais rica em informações do que as abordagens consideradas padrão por meio da adição da análise espaço-temporal, que pode contribuir qualitativamente para apoiar gestores de saúde e epidemiologistas no controle do *Aedes aegypti* e das doenças transmitidas por esse mosquito.

Este trabalho teve como principais limitações: a) o período estudado (2013 a 2016), limitado pela disponibilidade de dados no momento dos experimentos computacionais; b) a baixa granularidade dos dados com relação aos pontos, uma vez que não havia recursos para a geocodificação de versões reduzidas dos endereços dos pacientes, que precisaram ser agrupados por bairro. Contudo, considerando o estado da arte, este foi o primeiro estudo até onde se conhece onde a predição espaço-temporal e a seleção dos fatores mais relevantes foram combinadas de forma a não somente prever a distribuição geográfica de casos com uma janela temporal de um ano, mas também apontar os fatores climáticos e ambientais que mais exercem influência sobre o vetor, i.e. os mosquitos *Aedes aegypti*. Outra limitação foi a ausência de informações atualizadas sobre a distribuição geográfica da renda ou do acesso à água potável, tendo em vista que o fornecimento não regular de água potável nas regiões de mais baixa renda exerce pressão sobre a necessidade de manter depósitos alternativos de água, que podem vir a se

tornar criadouros de mosquitos.

6 CONCLUSÃO

6.1 CONCLUSÕES GERAIS

Neste trabalho, nós propomos uma metodologia para construir modelos preditivos, baseados em aprendizado de máquina, para predição da distribuição espaço-temporal de doenças a partir de bases de dados de casos notificados e de bases de informações geográficas. Como estudo de caso, aplicamos a metodologia à predição de arboviroses, especificamente da dengue, da chikungunya e da zika, cujo vetor é o mosquito *Aedes aegypti*, com dados da Cidade do Recife, Brazil, do período de 2013 a 2016, o que incluiu o surto de zika vírus de 2015, associado à ocorrência de más formações em recém-nascidos. Utilizamos dados abertos e anônimos, disponíveis para pesquisa no Portal de Dados Abertos da Cidade do Recife e obtidos do Sistema Nacional de Notificações, SINAN, do Sistema Único de Saúde (SUS) do Brasil. Também empregamos as seguintes informações geográficas climáticas e ambientais: velocidade dos ventos, temperaturas e precipitação, obtidas de sistemas de informações meteorológicas. Uma vez que a densidade de estações ambientais é muito pequena, a distribuição dessas variáveis climáticas e ambientais foi estimada em uma grade regular utilizando interpolação. A predição dos casos foi realizada a cada bimestre, considerando os dados dos últimos 12 meses.

Os melhores resultados de predição de casos foram obtidos com regressão por Random Forest. Devido ao baixo custo computacional e à estabilidade dos resultados, escolhemos a Random Forest com 10 árvores. Os valores do coeficiente de correlação de Pearson foram superiores a 0,99, enquanto o RMSE (%) se manteve inferior a 6%. Os índices de Kendall e Spearman também se mantiveram altos: seus valores foram superiores a 0,99 para Spearman (próximos ao coeficiente de Pearson) e maiores que 0,90 para Kendall (um índice mais rigoroso do que o de Pearson e o de Spearman). O desempenho superior da Random Forest, quando comparada aos outros modelos de regressão, mostra que o problema de regressão é de difícil generalização, dado que a Random Forest é baseada em comitês de árvores de decisão organizadas como *bagging* e a regressão é realizada por uma média ponderada dos resultados das diferentes árvores de decisão que compõem o modelo. Esses resultados permitem que se possa utilizar modelos baseados em Random Forest para construir sistemas de predição da distribuição de casos de arboviroses. Esses sistemas podem auxiliar os gestores do Sistema Único de Saúde e a sociedade civil organizada na elaboração, execução e fiscalização de políticas públicas e de ações de controle de doenças infecciosas.

Os resultados qualitativos de predição espaço-temporal também mostram que é possível observar a dinâmica das arboviroses transmitidas pelo mosquito *Aedes aegypti* com relativa precisão. As distribuições de casos obtidas são suaves: os números de casos obtidos foram concentrados nos bairros e depois interpolados, o que resulta em distribuições com comportamento local mais suavizado, embora o comportamento global tenda a se aproximar bastante do esperado para dados com maior granularidade. No entanto, os resultados qualitativos sugerem que distribuições mais precisas podem ser obtidas caso seja adotada uma abordagem baseada em IoT para medição das velocidades dos ventos, das temperaturas e de pluviometria em diferentes locais dos bairros, de modo a aumentar a densidade de dados. Similarmente, poderiam ser utilizadas informações aproximadas sobre a região da ocorrência do caso. No entanto, isso deve ser feito com cuidado para evitar a exposição e a localização do local de moradia do paciente.

Neste trabalho, também propomos um Comitê de Especialistas Artificiais, para selecionar os fatores mais relevantes para a predição. Esse comitê é integrado por métodos meta-heurísticos de busca e otimização, a saber: algoritmos genéticos, otimização por enxame de partículas, otimização por colônia de abelhas, e otimização por colônia de formigas. Esses especialistas artificiais foram treinados utilizando como função objetivo uma árvore de decisão, utilizando populações de mesmo tamanho evoluindo em uma mesma quantidade de gerações. Os fatores mais relevantes para a predição são definidos por votação. Dessa maneira, as especificidades de cada especialista virtual acabam sendo compensadas e os fatores mais relevantes para a predição são melhor definidos, o que pode auxiliar gestores de saúde pública na definição dos fatores mais importantes para o controle de um vetor. No caso da dengue, da chikungunya e da zika, como o vetor é o *Aedes aegypti*, a grande influência de fatores como a velocidade dos ventos, temperaturas e pluviometria apontam para a forte sazonalidade dessas doenças o que, por sua vez, indica sua dependência do vetor, i.e. o mosquito e seu ciclo de vida. Esses e outros fatores, como a influência de números passados de casos, podem ser bastante úteis para o planejamento e a execução de políticas públicas voltadas à melhoria da infraestrutura sanitária e ao planejamento e controle do vetor. De um ponto de vista prático, a implementação do Comitê de Especialistas Artificiais envolve um custo computacional relativamente alto, por envolver diversos métodos meta-heurísticos de busca e otimização. Contudo, a seleção dos fatores mais relevantes não necessita ser feita em tempo real. Além disso, a implementação do Comitê de Especialistas Artificiais é altamente paralelizável, tanto com relação à execução paralela dos algoritmos meta-heurísticos quanto à paralelização de cada algoritmo em si, no

que concerne à avaliação da função objetivo para cada candidato à solução.

6.2 DIFICULDADES APRESENTADAS

A primeira dificuldade apresentada foi a geração da base de treinamento. Inicialmente foi necessário baixar os dados manualmente do Portal de Dados Abertos da Prefeitura do Recife, da APAC e do INMET e, em seguida, organizar os dados em planilhas, para gerar as distribuições geográficas no Sistema de Informações Geográficas QGIS. Como as variáveis climáticas utilizadas não se distribuíam adequadamente pelo território da Cidade do Recife por existirem apenas três estações climáticas, utilizamos um programa Python desenvolvido por estudantes de mestrado do Laboratório de Computação Biomédica para estimar os valores das variáveis climáticas nos bairros onde não existiam. Esse programa foi utilizado para fazer essas estimativas e também interpolar os dados, para gerar uma distribuição mais contínua dos dados.

Outra dificuldade razoável foi a organização e integração das etapas da pesquisa, que teve que ser feita utilizando diversos *softwares* diferentes, além de procedimentos manuais: os algoritmos de aprendizado de máquina foram avaliados majoritariamente no ambiente de mineração de dados Weka; os especialistas virtuais foram todos implementados no ambiente Weka, utilizando os algoritmos de computação evolucionária implementados em *plugins* específicos, mas rodando em momentos diferentes. Os resultados foram então organizados em planilhas, onde a operação de votação dos comitês também foi feita por meio de fórmulas organizadas nessas planilhas.

É importante destacar que o uso de vários *softwares*, como o Weka, o QGIS e programas específicos para manipulação de dados, e de procedimentos manuais se deveu a basicamente três razões: (1) a ausência de bibliotecas específicas para construir a solução que envolvessem aprendizado de máquina, computação evolucionária e sistemas de informações geográficas utilizando uma única linguagem de programação nos anos de 2019 e 2020, quando foi realizada a maioria dos experimentos; (2) a relativa praticidade de integrar soluções e ambientes prontos e validados, diminuindo o tempo para obtenção de resultados e focando na implementação e validação da proposta de pesquisa em si; e (3) a pouca experiência da pesquisadora em desenvolvimento.

6.3 CONTRIBUIÇÕES

Neste trabalho, tivemos como principais contribuições: a formalização de uma metodologia para predição espaço-temporal de doenças infecciosas baseada em aprendizado de máquina e integração de bases de dados georreferenciados de variáveis climáticas e ambientais e de saúde; a proposta de um Comitê de Especialistas Artificiais baseado em agentes inteligentes compostos por métodos meta-heurísticos de busca, como os algoritmos genéticos, a otimização por enxame de partículas, as colônias de formigas e colônias de abelhas artificiais. Como estudo de caso, aplicamos e validamos a metodologia proposta e o Comitê de Especialistas Artificiais à predição de arboviroses transmitidas pelo mosquito *Aedes aegypti* na Cidade do Recife, de 2013 a 2016. Os resultados desta pesquisa foram comunicados a partir das seguintes publicações científicas:

1. Artigo “Spatio-temporal forecasting for dengue, chikungunya fever and zika using machine learning and artificial expert committees based on meta-heuristics”, publicado em 2022 no periódico *Research on Biomedical Engineering*, Qualis A4, da Springer-Nature, onde são apresentados os resultados mais gerais desta Tese, incluindo o Comitê de Especialistas Artificiais (SILVA et al., 2022);
2. Artigo “Forecasting Dengue, Chikungunya and Zika cases in Recife, Brazil: a spatio-temporal approach based on climate conditions, health notifications and machine learning”, onde são apresentados os primeiros resultados de predição espaço-temporal baseados em aprendizado de máquina, publicado no periódico *Research, Society and Development*, Qualis A3, em 2021 (SILVA et al., 2021d);
3. Capítulo de revisão bibliográfica “Intelligent systems for dengue, chikungunya and zika temporal and spatio-temporal forecasting: a contribution and a brief review”, onde é apresentada uma revisão crítica dos trabalhos relacionados a esta Tese, parte do livro “Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis”, publicado pela editora Springer em 2021 (LIMA et al., 2021).

A metodologia de predição espaço-temporal proposta neste trabalho foi adaptada para a predição de casos de Covid-19 no contexto de um projeto emergencial para combate à pandemia de Covid-19, tendo resultado nos seguintes trabalhos publicados:

1. Artigo “Covid-19 Dynamic Monitoring and Real-Time Spatio-Temporal Forecasting”, que apresenta uma aplicação detalhada da abordagem desta Tese para predição espaço-temporal de doenças aplicada à Covid-19, publicado no periódico *Frontiers in Public Health*, Qualis A3, da editora Frontiers, em 2021 (SILVA et al., 2021c);
2. Artigo “COVID-SGIS: A smart tool for dynamic monitoring and temporal forecasting of Covid-19”, que apresenta uma aplicação detalhada da abordagem desta Tese para predição temporal e espaço-temporal de doenças aplicada à Covid-19, publicado no periódico *Frontiers in Public Health*, Qualis A3, da editora Frontiers, em 2020 (LIMA et al., 2020);
3. Capítulo de revisão bibliográfica “Machine learning approaches for temporal and spatio-temporal Covid-19 forecasting: a brief review and a contribution”, parte do livro “Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis”, publicado pela editora Springer em 2021 (SILVA et al., 2021b);
4. Artigo “COVID-SGIS: Uma ferramenta inteligente para monitoramento dinâmico previsão temporal da Covid-19”, que apresenta uma aplicação resumida da abordagem desta Tese para predição espaço-temporal de doenças aplicada à Covid-19, publicado no *IV Simpósio de Inovação em Engenharia Biomédica - SABIO 2020*, em 2020 (SILVA et al., 2021a).

6.4 TRABALHOS FUTUROS

Como trabalhos futuros, para predição temporal e espaço-temporal eficiente de arboviroses e de outras doenças infecciosas cujo vetor seja influenciado por condições climáticas e ambientais, sugerimos a construção de uma rede de pequenas estações climáticas georreferenciadas com sensores para aferição da velocidade dos ventos, da temperatura e da umidade, em complementação às atuais estações climáticas, numa estratégia de Internet das Coisas (IoT). Isso permitiria predições a partir de uma grade de dados mais precisa. Também sugerimos como trabalhos futuros a integração dos dados georreferenciados tanto de número de casos quanto dos potenciais criadouros a partir da integração das bases de dados do SINAN e dos aplicativos móveis em uso pelos agentes de saúde ambiental e controle de endemias. Acreditamos firmemente que a integração de bases de dados de naturezas diferentes e de redes de sensores

em IoT pode contribuir para a melhoria da vigilância sanitária, permitindo o monitoramento e o controle rápido e eficiente de doenças infecciosas e de seus vetores, o que pode fazer uma grande diferença no controle de surtos epidêmicos, epidemias e pandemias em um contexto onde a Internet das Coisas, a Inteligência Artificial e as tecnologias da IV Revolução Industrial assumem uma importância cada vez maior no presente século nas mais diversas dimensões da vida cotidiana.

REFERÊNCIAS

- ABREU, P. H.; SANTOS, M. S.; ABREU, M. H.; ANDRADE, B.; SILVA, D. C. Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Computing Surveys (CSUR)*, ACM, v. 49, n. 3, p. 52, 2016.
- ACETO, G.; PERSICO, V.; PESCAPÉ, A. A survey on information and communication technologies for industry 4.0: State-of-the-art, taxonomies, perspectives, and challenges. *IEEE Communications Surveys & Tutorials*, IEEE, v. 21, n. 4, p. 3467–3501, 2019.
- ACETO, G.; PERSICO, V.; PESCAPÉ, A. Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. *Journal of Industrial Information Integration*, Elsevier, v. 18, p. 100129, 2020.
- AHMAD, M.; IBRAHIM, M.; MOHAMED, Z.; ISMAIL, N.; ABDULLAH, M.; SHUEB, R.; SHAFEL, M. The Sensitivity, Specificity and Accuracy of Warning Signs in Predicting Severe Dengue, the Severe Dengue Prevalence and Its Associated Factors. *International Journal of Environmental Research and Public Health*, Multidisciplinary Digital Publishing Institute, v. 15, n. 9, 2018.
- AKAY, B.; KARABOGA, D. A modified artificial bee colony algorithm for real-parameter optimization. *Information Sciences*, Elsevier, v. 192, p. 120–142, 2012.
- ALBRIEU-LLINÁS, G.; ESPINOSA, M. O.; QUAGLIA, A.; ABRIL, M.; SCAVUZZO, C. M. Urban environmental clustering to assess the spatial dynamics of *Aedes aegypti* breeding sites. *Geospatial Health*, v. 13, n. 1, 2018.
- ALSHUKAIRI, A. N.; ZHENG, J.; ZHAO, J.; NEHDI, A.; BAHAROON, S. A.; LAYQAH, L.; BOKHARI, A.; JOHANI, S. M. A.; SAMMAN, N.; BOUDJELAL, M. et al. High prevalence of MERS-CoV infection in camel workers in Saudi Arabia. *MBio*, Am Soc Microbiol, v. 9, n. 5, p. e01985–18, 2018.
- ALTHOUSE, B. M.; NG, Y. Y.; CUMMINGS, D. A. Prediction of dengue incidence using search query surveillance. *PLoS Neglected Tropical Diseases*, Public Library of Science, v. 5, n. 8, p. e1258, 2011.
- ATHEY, S.; TIBSHIRANI, J.; WAGER, S. Generalized random forests. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 47, n. 2, p. 1148–1178, 2019.
- AZHAR, E. I.; EL-KAFRAWY, S. A.; FARRAJ, S. A.; HASSAN, A. M.; AL-SAEED, M. S.; HASHEM, A. M.; MADANI, T. A. Evidence for camel-to-human transmission of mers coronavirus. *New England Journal of Medicine*, Mass Medical Soc, v. 370, n. 26, p. 2499–2505, 2014.
- BANHARNSAKUN, A.; ACHALAKUL, T.; SIRINAOVAKUL, B. The best-so-far selection in artificial bee colony algorithm. *Applied Soft Computing*, Elsevier, v. 11, n. 2, p. 2888–2901, 2011.
- BANSAL, J. C.; SHARMA, H.; JADON, S. S. Artificial bee colony algorithm: a survey. *International Journal of Advanced Intelligence Paradigms*, Inderscience, v. 5, n. 1-2, p. 123–159, 2013.

BARATA, J. C. A.; HUSSEIN, M. S. The Moore–Penrose Pseudoinverse: A Tutorial Review of the Theory. *Brazilian Journal of Physics*, v. 42, n. 1, p. 146–165, 2012. ISSN 1678-4448. Disponível em: <<http://dx.doi.org/10.1007/s13538-011-0052-z>>.

BARHAM, L. *From hand to handle: the first industrial revolution*. [S.l.]: Oxford University Press, 2013.

BARTLETT, P. L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, IEEE, v. 44, n. 2, p. 525–536, 1998.

BEKETOV, M. A.; YURCHENKO, Y. A.; BELEVICH, O. E.; LIESS, M. What environmental factors are important determinants of structure, species richness, and abundance of mosquito assemblages? *Journal of Medical Entomology*, Entomological Society of America, v. 47, n. 2, p. 129–139, 2014.

BELTRAMO, T.; RANZAN, C.; HINRICHS, J.; HITZMANN, B. Artificial neural network prediction of the biogas flow rate optimised with an ant colony algorithm. *Biosystems Engineering*, Elsevier, v. 143, p. 68–78, 2016.

BELTRÁN, J. D.; BOSCOR, A.; SANTOS, W. P. dos; MASSONI, T.; KOSTKOVA, P. ZIKA: A New System to Empower Health Workers and Local Communities to Improve Surveillance Protocols by E-learning and to Forecast Zika Virus in Real Time in Brazil. In: ACM. *Proceedings of the 2018 International Conference on Digital Health*. [S.l.], 2018. p. 90–94.

BHATT, S.; GETHING, P. W.; BRADY, O. J.; MESSINA, J. P.; FARLOW, A. W.; MOYES, C. L.; DRAKE, J. M.; BROWNSTEIN, J. S.; HOEN, A. G.; SANKOH, O. et al. The global distribution and burden of dengue. *Nature*, Nature Publishing Group, v. 496, n. 7446, p. 504, 2013.

BHUNIA, G. S.; SHIT, P. K. *Geospatial Analysis of Public Health*. [S.l.]: Springer International Publishing, 2019.

BIAU, G. Analysis of a random forests model. *The Journal of Machine Learning Research*, JMLR, v. 13, p. 1063–1095, 2012.

BONACCORSI, A.; CHIARELLO, F.; FANTONI, G.; KAMMERING, H. Emerging technologies and industrial leadership. a wikipedia-based strategic analysis of industry 4.0. *Expert Systems with Applications*, Elsevier, v. 160, p. 113645, 2020.

BONFIM, C.; MEDEIROS, Z. Epidemiologia e geografia: dos primórdios ao geoprocessamento. *Espac Saude*, v. 10, n. 1, p. 53–62, 2008.

BOOKER, L. B.; GOLDBERG, D. E.; HOLLAND, J. H. Classifier systems and genetic algorithms. *Artificial Intelligence*, Elsevier, v. 40, n. 1-3, p. 235–282, 1989.

BRAGA, C.; LUNA, C. F.; MARTELLI, C. M.; SOUZA, W. V. d.; CORDEIRO, M. T.; ALEXANDER, N.; ALBUQUERQUE, M. d. F. P. M. d.; SILVEIRA-JR., J. C.; MARQUES, E. T. Seroprevalence and risk factors for dengue infection in socio-economically distinct areas of Recife, Brazil. *Acta Tropica*, Elsevier, v. 113, n. 3, p. 234–240, 2010.

BRAGA, I. A.; LIMA, J. B. P.; SOARES, S. d. S.; VALLE, D. *Aedes aegypti* resistance to temephos during 2001 in several municipalities in the states of Rio de Janeiro, Sergipe, and Alagoas, Brazil. *Memórias do Instituto Oswaldo Cruz*, SciELO Brasil, v. 99, n. 2, p. 199–203, 2004.

BRAGA, I. A.; VALLE, D. *Aedes aegypti*: histórico do controle no Brasil. *Epidemiologia e Serviços de Saúde*, Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços/Secretaria de Vigilância em Saúde/Ministério da Saúde, v. 16, n. 2, p. 113–118, 2007.

BRASIER, A. R.; JU, H.; GARCIA, J.; SPRATT, H. M.; VICTOR, S. S.; FORSHEY, B. M.; HALSEY, E. S.; COMACH, G.; SIERRA, G.; BLAIR, P. J.; ROCHA, C.; MORRISON, A. C.; SCOTT, T. W.; BAZAN, I.; KOCHER, T. J. A three-component biomarker panel for prediction of dengue hemorrhagic fever. *The American Journal of Tropical Medicine and Hygiene*, ASTMH, v. 86, n. 2, p. 341–348, 2012.

BRASIL, M. d. S. *Levantamento rápido de índices para Aedes aegypti LIRaA para vigilância entomológica do Aedes aegypti no Brasil: Metodologia para avaliação dos índices de Breteau e predial e tipos de recipientes*. 1. ed. Brasília: BRASIL, Ministério da Saúde, 2012.

BRASIL, P.; PEREIRA-JR, J. P.; RAJA-GABAGLIA, C.; DAMASCENO, L.; WAKIMOTO, M.; RIBEIRO-NOGUEIRA, R. M.; SEQUEIRA, P. C. d.; MACHADO-SIQUEIRA, A.; CARVALHO, L. M. A. d.; CUNHA, D. C. d. Zika virus infection in pregnant women in Rio de Janeiro: preliminary report. *New England Journal of Medicine*, Mass Medical Soc, 2016.

BRODY, H.; RIP, M. R.; VINTEN-JOHANSEN, P.; PANETH, N.; RACHMAN, S. Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet*, Elsevier, v. 356, n. 9223, p. 64–68, 2000.

BUCZAK, A. L.; BAUGHER, B.; MONIZ, L. J.; BAGLEY, T.; BABIN, S. M.; GUVEN, E. Ensemble method for dengue prediction. *PloS One*, Public Library of Science, v. 13, n. 1, p. e0189988, 2018.

BURATTI, Y.; EIJKENS, C.; HAMEIRI, Z. Optimization of solar cell production lines using neural networks and genetic algorithms. *ACS Applied Energy Materials*, ACS Publications, v. 3, n. 11, p. 10317–10322, 2020.

BURKI, T. The origin of SARS-CoV-2. *The Lancet Infectious Diseases*, Elsevier, v. 20, n. 9, p. 1018–1019, 2020.

BURT, F. J.; ROLPH, M. S.; RULLI, N. E.; MAHALINGAM, S.; HEISE, M. T. Chikungunya: a re-emerging virus. *The Lancet*, Elsevier, v. 379, n. 9816, p. 662–671, 2012.

BUTT, N.; ABBASSI, A.; MUNIR, S.; AHMAD, S. M.; SHEIKH, Q. H. Haematological and biochemical indicators for the early diagnosis of dengue viral infection. *J Coll Physicians Surg Pak*, v. 18, n. 5, p. 282–5, 2008.

CAMPOS, G. S.; BANDEIRA, A. C.; SARDI, S. I. Zika virus outbreak, bahia, brazil. *Emerging Infectious Diseases*, Centers for Disease Control and Prevention, v. 21, n. 10, p. 1885, 2015.

CARDOSO, C. W.; PAPLOSKI, I. A.; KIKUTI, M.; RODRIGUES, M. S.; SILVA, M. M.; CAMPOS, G. S.; SARDI, S. I.; KITRON, U.; REIS, M. G.; RIBEIRO, G. S. Outbreak of exanthematous illness associated with Zika, chikungunya, and dengue viruses, Salvador,

Brazil. *Emerging Infectious Diseases*, Centers for Disease Control and Prevention, v. 21, n. 12, p. 2274, 2015.

CARDOSO, M. J.; ARBEL, T. *Deep learning in medical image analysis and multimodal learning for clinical decision support*. [S.l.]: Springer, 2017. 7th International Workshop on Multimodal Learning for Clinical Decision Support, ML-CDS 2017.

CH, S.; SOHANI, S.; KUMAR, D.; MALIK, A.; CHAHAR, B.; NEMA, A.; PANIGRAHI, B. K.; DHIMAN, R. A support vector machine-firefly algorithm based forecasting model to determine malaria transmission. *Neurocomputing*, Elsevier, v. 129, p. 279–288, 2014.

CHAKRABORTY, T.; CHATTOPADHYAY, S.; GHOSH, I. Forecasting dengue epidemics using a hybrid methodology. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 527, p. 121266, 2019.

CHANG, A. Y.; PARRALES, M. E.; JIMENEZ, J.; SOBIESZCZYK, M. E.; HAMMER, S. M.; COPENHAVER, D. J.; KULKARNI, R. P. Combining Google Earth and GIS mapping technologies in a dengue surveillance system for developing countries. *International Journal of Health Geographics*, BioMed Central, v. 8, n. 1, p. 49, 2009.

CHATURVEDI, I.; SATAPATHY, R.; CAVALLARI, S.; CAMBRIA, E. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters*, Elsevier, v. 125, p. 264–270, 2019.

CHEN, M.; WANG, S.; LIANG, P. P.; BALTRUŠAITIS, T.; ZADEH, A.; MORENCY, L.-P. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. [S.l.: s.n.], 2017. p. 163–171.

CHOI, H. K. Stock price correlation coefficient prediction with ARIMA-LSTM hybrid model. *arXiv preprint arXiv:1808.01560*, 2018.

CHOUDHURY, Z. M.; BANU, S.; ISLAM, A. M. Forecasting dengue incidence in Dhaka, Bangladesh: A time series analysis. WHO Regional Office for South-East Asia., 2008.

CHRISTOU, V.; TSIPOURAS, M. G.; GIANNAKEAS, N.; TZALLAS, A. T.; BROWN, G. Hybrid extreme learning machine approach for heterogeneous neural networks. *Neurocomputing*, Elsevier, v. 361, p. 137–150, 2019.

COLORNI, A.; DORIGO, M.; MANIEZZO, V. et al. Distributed optimization by ant colonies. In: CAMBRIDGE, MA. *Proceedings of the First European Conference on Artificial Life*. [S.l.], 1992. v. 142, p. 134–142.

CORTES, F.; MARTELLI, C. M. T.; XIMENES, R. A. de A.; MONTARROYOS, U. R.; JUNIOR, J. B. S.; CRUZ, O. G.; ALEXANDER, N.; SOUZA, W. V. de. Time series analysis of dengue surveillance data in two Brazilian cities. *Acta Tropica*, Elsevier, v. 182, p. 190–197, 2018.

CRAFTS, N. Explaining the first industrial revolution: two views. *European Review of Economic History*, Cambridge University Press, v. 15, n. 1, p. 153–168, 2011.

DAVIS, N. What is the fourth industrial revolution. In: *World Economic Forum*. [S.l.: s.n.], 2016. v. 19.

- DEANE, P. M.; DEANE, P. M. *The first industrial revolution*. [S.l.]: Cambridge University Press, 1979.
- DEB, K.; AGRAWAL, S.; PRATAP, A.; MEYARIVAN, T. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: SPRINGER. *International Conference on Parallel Problem Solving From Nature*. [S.l.], 2000. p. 849–858.
- DEB, K.; PRATAP, A.; AGARWAL, S.; MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, IEEE, v. 6, n. 2, p. 182–197, 2002.
- DENIL, M.; SHAKIBI, B.; DINH, L.; FREITAS, N. D. et al. Predicting parameters in deep learning. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 2148–2156.
- DICK, G.; KITCHEN, S.; HADDOW, A. Zika virus (i). isolations and serological specificity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Oxford University Press, v. 46, n. 5, p. 509–520, 1952.
- DING, J.; WANG, Q.; ZHANG, Q.; YE, Q.; MA, Y. A hybrid particle swarm optimization-cuckoo search algorithm and its engineering applications. *Mathematical Problems in Engineering*, Hindawi, v. 2019, 2019.
- DOM, N. C.; HASSAN, A. A.; LATIF, Z. A.; ISMAIL, R. Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia. *Asian Pacific Journal of Tropical Disease*, Elsevier, v. 3, n. 5, p. 352–361, 2013.
- DONALÍSIO, M. R.; GLASSER, C. M. Vigilância entomológica e controle de vetores do dengue. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 5, p. 259–279, 2002.
- DORIGO, M.; BIRATTARI, M. *Ant colony optimization*. [S.l.]: Springer, 2010.
- DORIGO, M.; GAMBARDELLA, L. M. Ant colonies for the travelling salesman problem. *Biosystems*, Elsevier, v. 43, n. 2, p. 73–81, 1997.
- DORIGO, M.; GAMBARDELLA, L. M. Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, IEEE, v. 1, n. 1, p. 53–66, 1997.
- DORIGO, M.; MANIEZZO, V.; COLORNI, A. et al. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, v. 26, n. 1, p. 29–41, 1996.
- DORIGO, M.; STÜTZLE, T. Ant colony optimization: overview and recent advances. In: *Handbook of Metaheuristics*. [S.l.]: Springer, 2019. p. 311–351.
- DRUCKER, H.; BURGESS, C. J.; KAUFMAN, L.; SMOLA, A.; VAPNIK, V. et al. Support vector regression machines. *Advances in Neural Information Processing Systems*, Morgan Kaufmann Publishers, v. 9, p. 155–161, 1997.
- DUDAS, G.; CARVALHO, L. M.; RAMBAUT, A.; BEDFORD, T. MERS-CoV spillover at the camel-human interface. *elife*, eLife Sciences Publications Limited, v. 7, p. e31257, 2018.

DUNCAN, A. P. *The Analysis and Application of Artificial Neural Networks for Early Warning Systems in Hydrology and the Environment*. Tese (Doutorado) — University of Exeter, 2014.

EHLERS, M.; MÖLLER, M.; MARANGON, S.; FERRE, N. The use of Geographic Information System (GIS) in the frame of the contingency plan implemented during the 1999–2001 Avian Influenza (AI) epidemic in Italy. *Avian Diseases*, v. 47, n. s3, p. 1010–1014, 2003.

ELBES, M.; ALZUBI, S.; KANAN, T.; AL-FUQAHA, A.; HAWASHIN, B. A survey on particle swarm optimization with emphasis on engineering and network applications. *Evolutionary Intelligence*, Springer, v. 12, n. 2, p. 113–129, 2019.

ELSHEIKH, A.; ELAZIZ, M. A. Review on applications of particle swarm optimization in solar energy systems. *International Journal of Environmental Science and Technology*, Springer, v. 16, n. 2, p. 1159–1170, 2019.

FAJARDO-HERRERA, R. J.; VALDELAMAR-VILLEGAS, J.-C.; ARRIETA-PÉREZ, D. Prediction of the potential establishment of the mosquito *Aedes aegypti* in non-residential urban spaces in Colombia using eco-urban and landscape variables. *Gestión y Ambiente*, Universidad Nacional de Colombia, v. 20, n. 1, p. 95, 2017.

FALBO, G.; FILHO, J. E. C. Facing a severe epidemic outbreak: a fight against arboviruses. *Revista Brasileira de Saúde Materno Infantil*, Scielo, v. 16, p. S3 – S4, 11 2016. ISSN 1519-3829. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-38292016000800001&nrm=iso>.

FARIA, R. M.; BORTOLOZZI, A. Espaço, território e saúde: contribuições de Milton Santos para o tema da geografia da saúde no Brasil. *Raega-O Espaço Geográfico em Análise*, v. 17, 2009.

FERREIRA, M. U. Epidemiologia e geografia: o complexo patogênico de Max. Sorre. *Cadernos de Saúde Pública*, SciELO Public Health, v. 7, p. 301–309, 1991.

FLEMING, G.; MERWE, M. Van der; MCFERREN, G. Fuzzy expert systems and GIS for cholera health risk prediction in southern Africa. *Environmental Modelling & Software*, Elsevier, v. 22, n. 4, p. 442–448, 2007.

FONG, S.; BIUK-AGHAI, R. P.; MILLHAM, R. C. Swarm Search Methods in Weka for Data Mining. In: *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*. New York, NY, USA: ACM, 2018. (ICMLC 2018), p. 122–127. ISBN 978-1-4503-6353-2. Disponível em: <<http://doi.acm.org/10.1145/3195106.3195167>>.

FREIRE, P. *A importância do ato de ler: em três artigos que se completam*. [S.l.]: Cortez, 2022.

GALLICCHIO, C.; MICHELI, A.; PEDRELLI, L. Design of deep echo state networks. *Neural Networks*, Elsevier, v. 108, p. 33–47, 2018.

GAMBARDELLA, L. M.; DORIGO, M. Solving symmetric and asymmetric tsps by ant colonies. In: IEEE. *Proceedings of IEEE international conference on evolutionary computation*. [S.l.], 1996. p. 622–627.

GAO, H.; SHI, Y.; PUN, C.-M.; KWONG, S. An improved artificial bee colony algorithm with its application. *IEEE Transactions on Industrial Informatics*, IEEE, v. 15, n. 4, p. 1853–1865, 2018.

- GAO, W.-f.; LIU, S.-y. A modified artificial bee colony algorithm. *Computers & Operations Research*, Elsevier, v. 39, n. 3, p. 687–697, 2012.
- GARCIA-CEJA, E.; RIEGLER, M.; NORDGREEN, T.; JAKOBSEN, P.; OEDEGAARD, K. J.; TØRRESEN, J. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, Elsevier, v. 51, p. 1–26, 2018.
- GHAMBARI, S.; RAHATI, A. An improved artificial bee colony algorithm and its application to reliability optimization problems. *Applied Soft Computing*, Elsevier, v. 62, p. 736–767, 2018.
- GHANI, A.; MCGINNITY, T. M.; MAGUIRE, L. P.; HARKIN, J. Neuro-inspired speech recognition with recurrent spiking neurons. In: SPRINGER. *International Conference on Artificial Neural Networks*. [S.l.], 2008. p. 513–522.
- GHARBI, M.; QUENEL, P.; GUSTAVE, J.; CASSADOU, S.; RUCHE, G. L.; GIRDARY, L.; MARRAMA, L. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC Infectious Diseases*, BioMed Central, v. 11, n. 1, p. 1–13, 2011.
- GHORPADE, S. N.; ZENNARO, M.; CHAUDHARI, B. S.; SAEED, R. A.; ALHUMYANI, H.; ABDEL-KHALEK, S. Enhanced differential crossover and quantum particle swarm optimization for iot applications. *IEEE Access*, IEEE, v. 9, p. 93831–93846, 2021.
- GILBERT, P. K. *Mapping the Victorian social body*. [S.l.]: SUNY Press, 2004.
- GODLEE, F. Health implications of climatic change. *BMJ: British Medical Journal*, BMJ Publishing Group, v. 303, n. 6812, p. 1254, 1991.
- GOMES, E. C. d. S.; LEAL-NETO, O. B.; ALBUQUERQUE, J.; SILVA, H. P. da; BARBOSA, C. S. Schistosomiasis transmission and environmental change: a spatio-temporal analysis in porto de galinhas, pernambuco-brazil. *International Journal of Health Geographics*, BioMed Central, v. 11, n. 1, p. 1–11, 2012.
- GOS, M.; KRZYSZCZAK, J.; BARANOWSKI, P.; MURAT, M.; MALINOWSKA, I. Combined TBATS and SVM model of minimum and maximum air temperatures applied to wheat yield prediction at different locations in Europe. *Agricultural and Forest Meteorology*, Elsevier, v. 281, p. 107827, 2020.
- GOULD, E. A.; HIGGS, S. Impact of climate change and other factors on emerging arbovirus diseases. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Royal Society of Tropical Medicine and Hygiene, v. 103, n. 2, p. 109–121, 2009.
- GRANDADAM, M.; CARO, V.; PLUMET, S.; THIBERGE, J.-M.; SOUARES, Y.; FAILLOUX, A.-B.; TOLOU, H. J.; BUDELOT, M.; COSSERAT, D.; LEPARC-GOFFART, I. et al. Chikungunya virus, southeastern france. *Emerging Infectious Diseases*, Centers for Disease Control and Prevention, v. 17, n. 5, p. 910, 2011.
- GREENWOOD, J. *The third industrial revolution: Technology, productivity, and income inequality*. [S.l.]: American Enterprise Institute, 1997.

- GUZMAN, M. G.; HALSTEAD, S. B.; ARTSOB, H.; BUCHY, P.; FARRAR, J.; GUBLER, D. J.; HUNSPERGER, E.; KROEGER, A.; MARGOLIS, H. S.; MARTÍNEZ, E. et al. Dengue: a continuing global threat. *Nature Reviews Microbiology*, Nature Publishing Group, v. 8, n. 12supp, p. S7, 2010.
- HADDAD, V. C. d. N.; SANTOS, T. C. F. The environmental theory by florence nightingale in the teaching of the nursing school Anna Nery (1962-1968). *Escola Anna Nery*, SciELO Brasil, v. 15, n. 4, p. 755–761, 2011.
- HAJIHASSANI, M.; ARMAGHANI, D. J.; KALATEHJARI, R. Applications of particle swarm optimization in geotechnical engineering: a comprehensive review. *Geotechnical and Geological Engineering*, Springer, v. 36, n. 2, p. 705–722, 2018.
- HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.
- HALLIDAY, S. Death and miasma in Victorian London: an obstinate belief. *BMJ: British Medical Journal*, BMJ Publishing Group, v. 323, n. 7327, p. 1469, 2001.
- HAMLET, A.; JEAN, K.; PEREA, W.; YACTAYO, S.; BIEY, J.; KERKHOVE, M. V.; FERGUSON, N.; GARSKE, T. The seasonal influence of climate and environment on yellow fever transmission across Africa. *PLoS Neglected Tropical Diseases*, Public Library of Science, v. 12, n. 3, p. e0006284, 2018.
- HAO, Y.; HAO, S.; ANDERSEN-NISSEN, E.; III, W. M. M.; ZHENG, S.; BUTLER, A.; LEE6, M. J.; WILK aron J.; DARBY, C.; ZAGER, M.; HOFFMAN, P.; STOECKIUS, M.; PAPALEXI, E.; MIMITOU, E. P.; JAIN, J.; SRIVASTAVA, A.; STUART, T.; FLEMING, L. M.; YEUNG, B.; ROGERS, A. J.; MCEL RATH, J. M.; BLISH, C. A.; GOTTARDO, R.; SMIBERT, P.; SATIJA, R. Integrated analysis of multimodal single-cell data. *Cell*, Elsevier, v. 184, n. 13, p. 3573–3587.e29, 2021.
- HASSANAT, A. B.; ALKAFaweEN, E. On enhancing genetic algorithms using new crossovers. *International Journal of Computer Applications in Technology*, Inderscience, v. 55, n. 3, p. 202–212, 2017.
- HAYES, E. B. Zika virus outside africa. *Emerging Infectious Diseases*, Centers for Disease Control and Prevention, v. 15, n. 9, p. 1347, 2009.
- Haykin, S. *Redes Neurais: Princípios e Prática*. Porto Alegre: Bookman, 2001.
- HILALI-JAGHDAM, I.; ISHAK, A. B.; ABDEL-KHALEK, S.; JAMAL, A. Quantum and classical genetic algorithms for multilevel segmentation of medical images: A comparative study. *Computer Communications*, Elsevier, v. 162, p. 83–93, 2020.
- HOLLAND, J. H. Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, SIAM, v. 2, n. 2, p. 88–105, 1973.
- HOLLAND, J. H. Genetic algorithms and adaptation. In: *Adaptive Control of Ill-Defined Systems*. [S.l.]: Springer, 1984. p. 317–333.
- HOLLAND, J. H. Genetic algorithms. *Scientific American*, JSTOR, v. 267, n. 1, p. 66–73, 1992.

- HOLMES, G.; DONKIN, A.; WITTEN, I. H. Weka: A machine learning workbench. In: IEEE. *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*. [S.l.], 1994. p. 357–361.
- HONÓRIO, N. A.; CÂMARA, D. C. P.; CALVET, G. A.; BRASIL, P. Chikungunya: uma arbovirose em estabelecimento e expansão no brasil. *Cadernos de Saúde Pública*, SciELO Public Health, v. 31, p. 906–908, 2015.
- HUANG, G.-B.; BABRI, H. A. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE Transactions on Neural Networks*, IEEE, v. 9, n. 1, p. 224–229, 1998.
- HUANG, G.-B.; WANG, D. H.; LAN, Y. Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, Springer, v. 2, n. 2, p. 107–122, 2011.
- HUANG, G.-B.; ZHOU, H.; DING, X.; ZHANG, R. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 42, n. 2, p. 513–529, 2012.
- HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: a new learning scheme of feedforward neural networks. In: IEEE. *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*. [S.l.], 2004. v. 2, p. 985–990.
- HUANG, G.-B.; ZHU, Q.-Y.; SIEW, C.-K. Extreme learning machine: theory and applications. *Neurocomputing*, Elsevier, v. 70, n. 1-3, p. 489–501, 2006.
- HUGENTOBLE, M. Quantum GIS. In: *Encyclopedia of GIS*. [S.l.]: Springer, 2008. p. 935–939.
- IQBAL, N.; ISLAM, M. Machine learning for dengue outbreak prediction: An outlook. *International Journal of Advanced Research in Computer Science*, v. 8, n. 1, 2017.
- IWENDI, C.; BASHIR, A. K.; PESHKAR, A.; SUJATHA, R.; CHATTERJEE, J. M.; PASUPULETI, S.; MISHRA, R.; PILLAI, S.; JO, O. Covid-19 patient health prediction using boosted random forest algorithm. *Frontiers in Public Health*, Frontiers, v. 8, p. 357, 2020.
- JAEGGER, H. The echo state approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, v. 148, n. 34, p. 13, 2001.
- JANSEN, C. C.; BEEBE, N. W. The dengue vector *Aedes aegypti*: what comes next. *Microbes and Infection*, Elsevier, v. 12, n. 4, p. 272–279, 2010.
- JAYACHITRA, V.; NIVETHA, S.; NIVETHA, R.; HARINI, R. A cognitive iot-based framework for effective diagnosis of covid-19 using multimodal data. *Biomedical Signal Processing and Control*, Elsevier, v. 70, p. 102960, 2021.
- JENSEN, M. T. Reducing the run-time complexity of multiobjective EAs: The NSGA-II and other algorithms. *IEEE Transactions on Evolutionary Computation*, IEEE, v. 7, n. 5, p. 503–515, 2003.
- JI, A. L.; RUBIN, A. J.; THRANE, K.; JIANG, S.; REYNOLDS, D. L.; MEYERS, R. M.; GUO, M. G.; GEORGE, B. M.; MOLLBRINK, A.; BERGENSTRÄHLE, J. et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, Elsevier, v. 182, n. 2, p. 497–514, 2020.

- JOHNSON, C.; JOHNSON, J. GIS: a tool for monitoring and management of epidemics. *Proceedings of Map India*, Citeseer, 2001.
- JOYCE, R. J.; JANOWIAK, J. E.; ARKIN, P. A.; XIE, P. CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology*, v. 5, n. 3, p. 487–503, 2004.
- KANDWAL, R.; GARG, P.; GARG, R. Health GIS and HIV/AIDS studies: Perspective and retrospective. *Journal of Biomedical Informatics*, Elsevier, v. 42, n. 4, p. 748–755, 2009.
- KARABOGA, D. *An idea based on honey bee swarm for numerical optimization*. [S.l.], 2005.
- KARABOGA, D.; AKAY, B. A comparative study of artificial bee colony algorithm. *Applied Mathematics and Computation*, Elsevier, v. 214, n. 1, p. 108–132, 2009.
- KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: IEEE. *Proceedings of ICNN'95 - International Conference on Neural Networks*. [S.l.], 1995. v. 4, p. 1942–1948.
- KESHTEGAR, B.; HEDDAM, S.; HOSSEINABADI, H. The employment of polynomial chaos expansion approach for modeling dissolved oxygen concentration in river. *Environmental Earth Sciences*, Springer, v. 78, n. 1, p. 34, 2019.
- KESORN, K.; ONGRUK, P.; CHOMPOOSRI, J.; PHUMEE, A.; THAVARA, U.; TAWATSIN, A.; SIRIYASATIEN, P. Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas. *PloS One*, Public Library of Science, v. 10, n. 5, p. e0125049, 2015.
- KHALFI, H.; ALAA, N. E.; GUEDDA, M. Period steady-state identification for a nonlinear front evolution equation using genetic algorithms. *International Journal of Bio-Inspired Computation*, Inderscience, v. 12, n. 3, p. 196–202, 2018.
- KHASHEI, M.; BIJARI, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, Elsevier, v. 11, n. 2, p. 2664–2675, 2011.
- KHORMI, H. M.; KUMAR, L. Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study. *Science of the Total Environment*, Elsevier, v. 409, n. 22, p. 4713–4719, 2011.
- KIM, D.-W.; YAO, Z.; GRAYBUCK, L. T.; KIM, T. K.; NGUYEN, T. N.; SMITH, K. A.; FONG, O.; YI, L.; KOULENA, N.; PIERSON, N.; SHAH, S.; LO, L.; POOL, A.-H.; OKA, Y.; PACHTER, L.; CAI, L.; TASIC, B.; ZENG, H.; ANDERSON, D. J. Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell*, Elsevier, v. 179, n. 3, p. 713–728, 2019.
- KOSTKOVA, P. Grand challenges in digital health. *Frontiers in Public Health*, Frontiers, v. 3, p. 134, 2015.
- KRAEMER, M. U.; SINKA, M. E.; DUDA, K. A.; MYLNE, A. Q.; SHEARER, F. M.; BARKER, C. M.; MOORE, C. G.; CARVALHO, R. G.; COELHO, G. E.; BORTEL, W. V. et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife*, eLife Sciences Publications Limited, v. 4, p. e08347, 2015.

- KRUIF, M. D. D.; SETIATI, T. E.; MAIRUHU, A. T.; KORAKA, P.; ABERSON, H. A.; SPEK, C. A.; OSTERHAUS, A. D.; REITSMA, P. H.; BRANDJES, D. P.; SOEMANTRI, A. et al. Differential gene expression changes in children with severe dengue virus infections. *PLoS Neglected Tropical Diseases*, Public Library of Science, v. 2, n. 4, p. e215, 2008.
- KUMAR, S.; MAHANTI, P.; WANG, S.-J. Intelligent computational techniques for multimodal data. *Multimedia Tools and Applications*, Springer, v. 78, n. 17, p. 23809–23814, 2019.
- LAUREANO-ROSARIO, A. E.; DUNCAN, A. P.; MENDEZ-LAZARO, P. A.; GARCIA-REJON, J. E.; GOMEZ-CARRO, S.; FARFAN-ALE, J.; SAVIC, D. A.; MULLER-KARGER, F. E. Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico. *Tropical Medicine and Infectious Disease*, Multidisciplinary Digital Publishing Institute, v. 3, n. 1, p. 5, 2018.
- LAUREANO-ROSARIO, A. E.; DUNCAN, A. P.; SYMONDS, E. M.; SAVIC, D. A.; MULLER-KARGER, F. E. Predicting culturable enterococci exceedances at Escambron Beach, San Juan, Puerto Rico using satellite remote sensing and artificial neural networks. *Journal of Water and Health*, 2018.
- LEAL-NETO, O. B.; ALBUQUERQUE, C. M.; ALBUQUERQUE, J. O.; BARBOSA, C. S. The schisto track: a system for gathering and monitoring epidemiological surveys by connecting geographical information systems in real time. *JMIR mHealth and uHealth*, JMIR Publications, v. 2, n. 1, p. e2859, 2014.
- LEAL-NETO, O. B.; LOYO, R.; ALBUQUERQUE, J.; PERAZZO, J.; BARBOSA, V.; BARBOSA, C. S. Using mobile technology to conduct epidemiological investigations. *Revista da Sociedade Brasileira de Medicina Tropical*, SciELO Brasil, v. 48, p. 105–107, 2015.
- LEAL, O. B.; ALBUQUERQUE, J.; CRUZ, O. G.; CESSE, E.; SOUZA, W. V. d. Perspectivas otimistas para a saúde do futuro. *Cadernos de Saúde Pública*, SciELO Brasil, v. 33, 2017.
- LEAL, O. B.; ALBUQUERQUE, J.; SOUZA, W. V.; CESSE, E.; CRUZ, O. G. Inovações disruptivas e as transformações da saúde pública na era digital. *Cadernos de Saúde Pública*, SciELO Brasil, v. 33, 2017.
- LECK, A.; LIM, R. J. A peek into the health tech boom. *Managing Intell. Prop.*, HeinOnline, v. 251, p. 33, 2015.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015.
- LEE, K. Y.; CHUNG, N.; HWANG, S. Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. *Ecological Informatics*, Elsevier, v. 36, p. 172–180, 2016.
- LENDEK, Z.; GUERRA, T. M.; BABUŠKA, R.; SCHUTTER, B. *Stability analysis and nonlinear observer design using Takagi-Sugeno fuzzy models*. [S.l.]: Springer, 2011.
- LEPARC-GOFFART, I.; NOUGAIREDE, A.; CASSADOU, S.; PRAT, C.; LAMBALLERIE, X. D. Chikungunya in the americas. *The Lancet*, Elsevier, v. 383, n. 9916, p. 514, 2014.
- LI, J.; CAI, L.; LI, Y. Stage-structured wild and sterile mosquito population models and their dynamics. *Journal of Biological Dynamics*, Taylor & Francis, v. 11, n. sup1, p. 79–101, 2017.

- LI, M.; XU, H.; DENG, Y. Evidential decision tree based on belief entropy. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 21, n. 9, p. 897, 2019.
- LI, W.; SHI, Z.; YU, M.; REN, W.; SMITH, C.; EPSTEIN, J. H.; WANG, H.; CRAMERI, G.; HU, Z.; ZHANG, H. et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science*, American Association for the Advancement of Science, v. 310, n. 5748, p. 676–679, 2005.
- LIMA, C. L. de; SILVA, A. C. G. da; SILVA, C. C. da; MORENO, G. M. M.; SILVA-FILHO, A. G. da; MUSAH, A.; ALDOSERY, A.; DUTRA, L.; AMBRIZZI, T.; BORGES, I. V. G.; TUNALI, M.; BASIBUYUK, S.; YENIGUN, O.; MASSONI, T. L.; JONES, K.; CAMPOS, L.; KOSTKOVA, P.; SANTOS, W. P. dos. Intelligent systems for dengue, chikungunya and zika temporal and spatio-temporal forecasting: a contribution and a brief review. In: PANI, S.; DASH, S.; SANTOS, W. P. dos; BUKHARI, S. A. C.; FLAMMINI, F. (Ed.). *Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis*. 1. ed. Geneve: Springer, 2021. v. 1, p. 301–334. ISBN 9783030797522.
- LIMA, C. L. de; SILVA, C. C. da; SILVA, A. C. G. da; SILVA, E. L.; MARQUES, G. S.; ARAÚJO, L. J. B. de; ALBUQUERQUE-JÚNIOR, L. A.; SOUZA, S. B. J. de; SANTANA, M. A. de; GOMES, J. C.; BARBOSA, V. A. de F.; MUSAH, A.; KOSTKOVA, P.; SANTOS, W. P. dos; SILVA-FILHO, A. G. da. COVID-SGIS: A Smart Tool for Dynamic Monitoring and Temporal Forecasting of Covid-19. *Frontiers in Public Health*, Frontiers, v. 8, p. 761, 2020. Disponível em: <<https://www.frontiersin.org/articles/10.3389/fpubh.2020.580815>>.
- LIMA, M. V. M. d.; LAPORTA, G. Z. Evaluation of the Models for Forecasting Dengue in Brazil from 2000 to 2017: An Ecological Time-Series Study. *Insects*, Multidisciplinary Digital Publishing Institute, v. 11, n. 11, p. 794, 2020.
- LIMA, S. M. de; SILVA-FILHO, A. G. da; SANTOS, W. P. dos. Detection and classification of masses in mammographic images in a multi-kernel approach. *Computer Methods and Programs in Biomedicine*, Elsevier, v. 134, p. 11–29, 2016.
- LIN, K.; WANG, D.; XIA, F.; GE, H. Device clustering algorithm based on multimodal data correlation in cognitive internet of things. *IEEE Internet of Things Journal*, IEEE, v. 5, n. 4, p. 2263–2271, 2017.
- LIU-HELMERSSON, J.; BRÄNNSTRÖM, Å.; SEWE, M. O.; SEMENZA, J. C.; ROCKLÖV, J. Estimating past, present, and future trends in the global distribution and abundance of the arbovirus vector aedes aegypti under climate change scenarios. *Frontiers in Public Health*, Frontiers, v. 7, p. 148, 2019.
- LIU, P.; BASHA, M. D. E.; LI, Y.; XIAO, Y.; SANELLI, P. C.; FANG, R. Deep evolutionary networks with expedited genetic algorithms for medical image denoising. *Medical image analysis*, Elsevier, v. 54, p. 306–315, 2019.
- LIVERA, A. M. D.; HYNDMAN, R. J.; SNYDER, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, Taylor & Francis, v. 106, n. 496, p. 1513–1527, 2011.
- LORENZ, C.; AZEVEDO, T. S.; VIRGINIO, F.; AGUIAR, B. S.; CHIARAVALLOTI-NETO, F.; SUESDEK, L. Impact of environmental factors on neglected emerging arboviral diseases. *PLoS Neglected Tropical Diseases*, Public Library of Science San Francisco, v. 11, n. 9, p. e0005959, 2017.

LU, J.; LI, J. Dynamics of stage-structured discrete mosquito population models. *Journal of Applied Analysis and Computation*, v. 1, p. 53–67, 2011.

LU, L.; LIN, H.; TIAN, L.; YANG, W.; SUN, J.; LIU, Q. Time series analysis of dengue fever and weather in Guangzhou, China. *BMC Public Health*, BioMed Central, v. 9, n. 1, p. 395, 2009.

LU, R.; HU, H.; XI, M.; GAO, H.; PUN, C.-M. An improved artificial bee colony algorithm with fast strategy, and its application. *Computers & Electrical Engineering*, Elsevier, v. 78, p. 79–88, 2019.

LUKOŠEVIČIUS, M. A practical guide to applying echo state networks. In: *Neural networks: tricks of the trade*. [S.l.]: Springer, 2012. p. 659–686.

MA, Q.; SHEN, L.; COTTRELL, G. W. DeePr-ESN: A deep projection-encoding echo-state network. *Information Sciences*, Elsevier, v. 511, p. 152–171, 2020.

MAASS, W.; NATSCHLÄGER, T.; MARKRAM, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, MIT Press, v. 14, n. 11, p. 2531–2560, 2002.

MACEDO, P. de O.; QUITETE, J. B.; LIMA, E. C.; SANTOS, I. dos; VARGENS, O. M. da C. As tecnologias de cuidado de enfermagem obstétrica fundamentadas pela teoria ambientalista de Florence Nightingale. *Escola Anna Nery*, Universidade Federal do Rio de Janeiro, v. 12, n. 2, p. 341–347, 2008.

MAGALHAES, T.; BRAGA, C.; CORDEIRO, M. T.; OLIVEIRA, A. L. S.; CASTANHA, P. M. S.; MACIEL, A. P. R.; AMANCIO, N. M. L.; GOUVEIA, P. N.; SILVA-JR., V. J. P. da; PEIXOTO, T. F. L.; BRITTO, H.; LIMA, P. V.; LIMA, A. R. S.; ROSENBERGER, K. D.; JAENISCH, T.; MARQUES, E. T. A. Zika virus displacement by a chikungunya outbreak in Recife, Brazil. *PLoS Neglected Tropical Diseases*, Public Library of Science, v. 11, n. 11, p. e0006055, 2017.

MAJUMDER, N.; HAZARIKA, D.; GELBUKH, A.; CAMBRIA, E.; PORIA, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, Elsevier, v. 161, p. 124–133, 2018.

MALIK, Z. K.; HUSSAIN, A.; WU, Q. J. Multilayered echo state machine: A novel architecture and algorithm. *IEEE Transactions on Cybernetics*, IEEE, v. 47, n. 4, p. 946–959, 2016.

MANOGARAN, G.; LOPEZ, D.; THOTA, C.; ABBAS, K. M.; PYNE, S.; SUNDARASEKAR, R. Big data analytics in healthcare Internet of Things. In: *Innovative healthcare systems for the 21st century*. [S.l.]: Springer, 2017. p. 263–284.

MANOGARAN, G.; VARATHARAJAN, R.; LOPEZ, D.; KUMAR, P. M.; SUNDARASEKAR, R.; THOTA, C. A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. *Future Generation Computer Systems*, Elsevier, v. 82, p. 375–387, 2018.

MARTINES, R. B. Notes from the field: evidence of Zika virus infection in brain and placental tissues from two congenitally infected newborns and two fetal losses Brazil, 2015. *MMWR. Morbidity and Mortality Weekly Report*, v. 65, 2016.

- MATTHEWS, J.; KULKARNI, R.; GERLA, M.; MASSEY, T. Rapid dengue and outbreak detection with mobile systems and social networks. *Mobile Networks and Applications*, Springer-Verlag New York, Inc., v. 17, n. 2, p. 178–191, 2012.
- MAYER, S. V.; TESH, R. B.; VASILAKIS, N. The emergence of arthropod-borne viral diseases: A global prospective on dengue, chikungunya and zika fevers. *Acta Tropica*, Elsevier, v. 166, p. 155–163, 2017.
- MEDEIROS, A. B. d. A.; ENDERS, B. C.; LIRA, A. L. B. D. C. The Florence Nightingales environmental theory: A critical analysis. *Escola Anna Nery*, SciELO Brasil, v. 19, n. 3, p. 518–524, 2015.
- MEHMOOD, A.; ZAMEER, A.; LING, S. H.; REHMAN, A. ur; RAJA, M. A. Z. Integrated computational intelligent paradigm for nonlinear electric circuit models using neural networks, genetic algorithms and sequential quadratic programming. *Neural Computing and Applications*, Springer, v. 32, n. 14, p. 10337–10357, 2020.
- MLAKAR, J.; KORVA, M.; TUL, N.; POPOVIĆ, M.; POLJŠAK-PRIJATELJ, M.; MRAZ, J.; KOLENC, M.; RUS, K. R.; VIPOTNIK, T. V.; VODUŠEK, V. F. et al. Zika virus associated with microcephaly. *New England Journal of Medicine*, Mass Medical Soc, v. 374, n. 10, p. 951–958, 2016.
- MONAGHAN, A. J.; SAMPSON, K.; STEINHOFF, D.; ERNST, K.; EBI, K.; JONES, B.; HAYDEN, M. The potential impacts of 21st century climatic and population changes on human exposure to the virus vector mosquito *Aedes aegypti*. *Climatic Change*, Springer, v. 146, n. 3, p. 487–500, 2018.
- MONKEN, M.; BARCELLOS, C. Vigilância em saúde e território utilizado: possibilidades teóricas e metodológicas. *Cadernos de Saúde Pública*, SciELO Public Health, v. 21, p. 898–906, 2005.
- MORDECAI, E. A.; RYAN, S. J.; CALDWELL, J. M.; SHAH, M. M.; LABEAUD, A. D. Climate change could shift disease burden from malaria to arboviruses in africa. *The Lancet Planetary Health*, Elsevier, v. 4, n. 9, p. e416–e423, 2020.
- MOSER, A.; CARLANDER, M.; WIESER, S.; HÄMMIG, O.; PUHAN, M. A.; HÖGLINGER, M. The COVID-19 Social Monitor longitudinal online panel: Real-time monitoring of social and public health consequences of the COVID-19 emergency in Switzerland. *PloS one*, Public Library of Science, v. 15, n. 11, p. e0242129, 2020.
- MOZOŃOWSKI, W. Jäccaron;drzej Sniadecki (1768–1838) on the cure of rickets. *Nature*, Nature Publishing Group, v. 143, n. 3612, p. 121, 1939.
- MUSSO, D.; ROCHE, C.; ROBIN, E.; NHAN, T.; TEISSIER, A.; CAO-LORMEAU, V.-M. Potential sexual transmission of zika virus. *Emerging Infectious Diseases*, Centers for Disease Control and Prevention, v. 21, n. 2, p. 359, 2015.
- MYLES, A. J.; FEUDALE, R. N.; LIU, Y.; WOODY, N. A.; BROWN, S. D. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, Wiley Online Library, v. 18, n. 6, p. 275–285, 2004.
- NAIM, I.; MAHARA, T.; IDRISI, A. R. Effective short-term forecasting for daily time series with complex seasonal patterns. *Procedia Computer Science*, Elsevier, v. 132, p. 1832–1841, 2018.

NASIRUDEEN, A.; WONG, H. H.; THIEN, P.; XU, S.; LAM, K.-P.; LIU, D. X. RIG-I, MDA5 and TLR3 synergistically play an important role in restriction of dengue virus infection. *PLoS Neglected Tropical Diseases*, Public Library of Science, v. 5, n. 1, p. e926, 2011.

NEWSOM, S. Pioneers in infection control: John Snow, Henry Whitehead, the Broad Street pump, and the beginnings of geographical epidemiology. *Journal of Hospital Infection*, Elsevier, v. 64, n. 3, p. 210–216, 2006.

NOWOZIN, S.; ROTHER, C.; BAGON, S.; SHARP, T.; YAO, B.; KOHLI, P. Decision tree fields. In: IEEE. *2011 International Conference on Computer Vision*. [S.l.], 2011. p. 1668–1675.

OLIPHANT, T. E. Python for scientific computing. *Computing in Science & Engineering*, IEEE, v. 9, n. 3, 2007.

PAI, P.-F.; LIN, C.-S. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, Elsevier, v. 33, n. 6, p. 497–505, 2005.

PALANIYANDI, M. The environmental aspects of dengue and chikungunya outbreaks in India: GIS for epidemic control. *International Journal of Mosquito Research*, Directory of Open Access Journals, v. 1, n. 2, p. 38–44, 2014.

PALANIYANDI, M.; ANAND, P.; MANIYOSAI, R.; MARIAPPAN, T.; DAS, P. The integrated remote sensing and GIS for mapping of potential vector breeding habitats, and the Internet GIS surveillance for epidemic transmission control, and management. *Environment*, v. 7, n. 12, p. 14–16, 2016.

PALANIYANDI, M.; ANAND, P.; PAVENDAR, T. Environmental risk factors in relation to occurrence of vector borne disease epidemics: Remote sensing and GIS for rapid assessment, picturesque, and monitoring towards sustainable health. *International Journal of Mosquito Research*, Akinik, v. 4, n. 3, p. 09–20, 2017.

PALM, T. The geographic distribution of rickets. *The Practitioner*, 1890.

PAN, X.; ZHANG, Q.; PAN, H. Improved artificial bee colony algorithm and its application to fundus retinal blood vessel image binarization. *IEEE Access*, IEEE, v. 8, p. 123726–123734, 2020.

PANETH, N.; VINTEN-JOHANSEN, P.; BRODY, H.; RIP, M. A rivalry of foulness: official and unofficial investigations of the London cholera epidemic of 1854. *American Journal of Public Health*, American Public Health Association, v. 88, n. 10, p. 1545–1553, 1998.

PAUL, K. K.; DHAR-CHOWDHURY, P.; HAQUE, C. E.; AL-AMIN, H. M.; GOSWAMI, D. R.; KAFI, M. A. H.; DREBOT, M. A.; LINDSAY, L. R.; AHSAN, G. U.; BROOKS, W. A. Risk factors for the presence of dengue vector mosquitoes, and determinants of their prevalence and larval site selection in Dhaka, Bangladesh. *PloS One*, Public Library of Science, v. 13, n. 6, p. e0199457, 2018.

PETERSEN, L. R.; JAMIESON, D. J.; POWERS, A. M.; HONEIN, M. A. Zika virus. *New England Journal of Medicine*, Mass Medical Soc, v. 374, n. 16, p. 1552–1563, 2016.

PIALOUX, G.; GAÜZÈRE, B.-A.; JAURÉGUIBERRY, S.; STROBEL, M. Chikungunya, an epidemic arbovirosis. *The Lancet Infectious Diseases*, Elsevier, v. 7, n. 5, p. 319–327, 2007.

- PORIA, S.; CAMBRIA, E.; BAJPAI, R.; HUSSAIN, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, Elsevier, v. 37, p. 98–125, 2017.
- PORIA, S.; CAMBRIA, E.; HAZARIKA, D.; MAZUMDER, N.; ZADEH, A.; MORENCY, L.-P. Multi-level multiple attentions for contextual multimodal sentiment analysis. In: IEEE. *2017 IEEE International Conference on Data Mining (ICDM)*. [S.l.], 2017. p. 1033–1038.
- PORIA, S.; MAJUMDER, N.; HAZARIKA, D.; CAMBRIA, E.; GELBUKH, A.; HUSSAIN, A. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, IEEE, v. 33, n. 6, p. 17–25, 2018.
- PRISECARU, P. Challenges of the fourth industrial revolution. *Knowledge Horizons. Economics*, Dimitrie Cantemir Christian University, v. 8, n. 1, p. 57, 2016.
- QUANTUM, G. *Development Team.(2013). Quantum GIS geographic information system. Open Source Geospatial Foundation Project*. 2013.
- QUINCEY, E. D.; KOSTKOVA, P. Early warning and outbreak detection using social networking websites: The potential of Twitter. In: SPRINGER. *International Conference on Electronic Healthcare*. [S.l.], 2009. p. 21–24.
- RAJAKUMAR, K. Vitamin D, cod-liver oil, sunlight, and rickets: a historical perspective. *Pediatrics*, Am Acad Pediatrics, v. 112, n. 2, p. e132–e135, 2003.
- RAO, H.; SHI, X.; RODRIGUE, A. K.; FENG, J.; XIA, Y.; ELHOSENY, M.; YUAN, X.; GU, L. Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing*, Elsevier, v. 74, p. 634–642, 2019.
- REY, F. A. Dengue virus envelope glycoprotein structure: new insight into its interactions during viral entry. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 100, n. 12, p. 6899–6901, 2003.
- REZZA, G.; NICOLETTI, L.; ANGELINI, R.; ROMI, R.; FINARELLI, A.; PANNING, M.; CORDIOLI, P.; FORTUNA, C.; BOROS, S.; MAGURANO, F. et al. Infection with chikungunya virus in Italy: an outbreak in a temperate region. *The Lancet*, Elsevier, v. 370, n. 9602, p. 1840–1846, 2007.
- RIGATTI, S. J. Random forest. *Journal of Insurance Medicine*, American Academy of Insurance Medicine, v. 47, n. 1, p. 31–39, 2017.
- RISSINO, S.; LAMBERT-TORRES, G. Rough set theoryfundamental concepts, principals, data extraction, and applications. In: *Data mining and knowledge discovery in real life applications*. [S.l.]: InTech, 2009.
- ROSENBERG, N. The role of electricity in industrial development. *The Energy Journal*, International Association for Energy Economics, v. 19, n. 2, 1998.
- ROTH, A.; MERCIER, A.; LEPERS, C.; HOY, D.; DUITUTURAGA, S.; BENYON, E.; GUILLAUMOT, L.; SQUARES, Y. Concurrent outbreaks of dengue, chikungunya and Zika virus infections—an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012–2014. *Eurosurveillance*, European Centre for Disease Prevention and Control, v. 19, n. 41, p. 20929, 2014.

RUSSELL, R. C. Mosquito-borne arboviruses in australia: the current scene and implications of climate change for human health. *International Journal for Parasitology*, Elsevier, v. 28, n. 6, p. 955–969, 1998.

SÁNCHEZ, L. E.; GURPEGUI, I. S.; LUCAS, S. Z. Climatic change and emerging virus dissemination. In: *VIII International Veterinary Congress*. [S.l.: s.n.], 2018. v. 2018.

SANTANA, M. A. de; PEREIRA, J. M. S.; SILVA, F. L. da; LIMA, N. M. de; SOUSA, F. N. de; ARRUDA, G. M. S. de; LIMA, R. d. C. F. de; SILVA, W. W. A. da; SANTOS, W. P. dos. Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Research on Biomedical Engineering*, Brazilian Society of Biomedical Engineering, v. 34, n. 1, p. 45–53, 2018.

SANTANA, P. *Introdução à geografia da saúde: território, saúde e bem-estar*. [S.l.]: Imprensa da Universidade de Coimbra/Coimbra University Press, 2014.

SARBEEN, J. I.; GHEENA, S. Microbial variation in climatic change and its effect on human health. *Research Journal of Pharmacy and Technology*, A & V Publications, v. 9, n. 10, p. 1777–1781, 2016.

SCAVUZZO, J. M.; TRUCCO, F.; ESPINOSA, M.; TAURO, C. B.; ABRIL, M.; SCAVUZZO, C. M.; FRERY, A. C. Modeling Dengue vector population using remotely sensed data and machine learning. *Acta Tropica*, Elsevier, v. 185, p. 167–175, 2018.

SCAVUZZO, J. M.; TRUCCO, F. C.; TAURO, C. B.; GERMAN, A.; ESPINOSA, M.; ABRIL, M. Modeling the temporal pattern of Dengue, Chicungunya and Zika vector using satellite data and neural networks. In: IEEE. *Information Processing and Control (RPIC), 2017 XVII Workshop on*. [S.l.], 2017. p. 1–6.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural Networks*, Elsevier, v. 61, p. 85–117, 2015.

SCHOENTAL, R. Climatic change and human health. *Journal of the Royal Society of Medicine*, Royal Society of Medicine Press, v. 87, n. 8, p. 495, 1994.

SEBAYANG, A.; MASJUKI, H.; ONG, H. C.; DHARMA, S.; SILITONGA, A.; KUSUMO, F.; MILANO, J. Optimization of bioethanol production from sorghum grains using artificial neural networks integrated with ant colony. *Industrial Crops and Products*, Elsevier, v. 97, p. 146–155, 2017.

SELANDERS, L. C. The power of environmental adaptation: Florence Nightingale's original theory for nursing practice. *Journal of Holistic Nursing*, Sage Publications Sage CA: Thousand Oaks, CA, v. 16, n. 2, p. 247–263, 1998.

SEMET, Y.; MARCON, B.; DEMESTICHAS, K.; KOUTSOURIS, N.; ASCOLESE, A. Artificial ant colonies for adaptive rewards in serious games. In: MIT PRESS. *The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*. [S.l.], 2019. p. 533–540.

SHAH, R.; ZIMMERMANN, R. *Multimodal analysis of user-generated multimedia content*. [S.l.]: Springer, 2017.

SHAUKAT, K.; MASOOD, N.; MEHREEN, S.; AZMEEN, U. Dengue fever prediction: A data mining problem. *Journal of Data Mining in Genomics & Proteomics*, v. 2015, 2015.

SHEPARD, C. W.; GORTAKOWSKI, H. W.; NASRALLAH, H.; CUTLER, B. H.; BEGIER, E. M. Using GIS-based density maps of HIV surveillance data to identify previously unrecognized geographic foci of HIV burden in an urban epidemic. *Public Health Reports*, SAGE Publications, v. 126, n. 5, p. 741, 2011.

SHI, Y.; EBERHART, R. A modified particle swarm optimizer. In: IEEE. *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence*. [S.l.], 1998. p. 69–73.

SILVA, A. C. G. da; LIMA, C. L. de; SILVA, C. C. da; SILVA, E. L.; MARQUES, G. S.; ARAÚJO, L. J. B. de; ALBUQUERQUE-JÚNIOR, L. A.; SOUZA, S. B. J. de; SANTANA, M. A. de; GOMES, J. C.; BARBOSA, V. A. de F.; MUSAH, A.; KOSTKOVA, P.; SANTOS, W. P. dos; SILVA-FILHO, A. G. da. COVID-SGIS: Uma ferramenta inteligente para monitoramento dinâmico previsão temporal da Covid-19. In: SANTOS, W. P. dos; GUSMÃO, C. M. G. de; MORENO, G. M. M. (Ed.). *Anais do IV Simpósio de Inovação em Engenharia Biomédica - SABIO 2020*. Recife: Editora Universitária da UFPE, 2021. v. 1, n. 2020, p. 116–123.

SILVA, A. C. G. da; LIMA, C. L. de; SILVA, C. C. da; MORENO, G. M. M.; SILVA, E. L.; MARQUES, G. S.; ARAÚJO, L. J. B. de; ALBUQUERQUE-JÚNIOR, L. A.; SOUZA, S. B. J. de; SANTANA, M. A. de; GOMES, J. C.; BARBOSA, V. A. de F.; MUSAH, A.; KOSTKOVA, P.; SILVA-FILHO, A. G. da; SANTOS, W. P. dos. Machine learning approaches for temporal and spatio-temporal Covid-19 forecasting: a brief review and a contribution. In: PANI, S.; DASH, S.; SANTOS, W. P. dos; BUKHARI, S. A. C.; FLAMMINI, F. (Ed.). *Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis*. 1. ed. Geneve: Springer, 2021. v. 1, p. 335–358. ISBN 9783030797522.

SILVA, C. C. da; LIMA, C. L. de; SILVA, A. C. G. da; SILVA, E. L.; MARQUES, G. S.; ARAÚJO, L. J. B. de; ALBUQUERQUE-JÚNIOR, L. A.; SOUZA, S. B. J. de; SANTANA, M. A. de; GOMES, J. C.; BARBOSA, V. A. de F.; MUSAH, A.; KOSTKOVA, P.; SANTOS, W. P. dos; SILVA-FILHO, A. G. da. Covid-19 dynamic monitoring and real-time spatio-temporal forecasting. *Frontiers in Public Health*, Frontiers, v. 9, p. 314, 2021. Disponível em: <<https://www.frontiersin.org/article/10.3389/fpubh.2021.641253>>.

SILVA, C. C. da; LIMA, C. L. de; SILVA, A. C. G. da; MORENO, G. M. M.; MUSAH, A.; ALDOSERY, A.; DUTRA, L.; AMBRIZZI, T.; BORGES, I. V. G.; TUNALI, M.; BASIBUYUK, S.; YENIGUN, O.; JONES, K.; CAMPOS, L.; MASSONI, T. L.; SILVA-FILHO, A. G. da; KOSTKOVA, P.; SANTOS, W. P. dos. 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*, CDRR Editors, v. 10, n. 12, p. e452101220804, 2021. Disponível em: <<https://rsdjournal.org/index.php/rsd/article/view/20804>>.

SILVA, C. C. da; LIMA, C. L. de; SILVA, A. C. G. da; MORENO, G. M. M.; MUSAH, A.; ALDOSERY, A.; DUTRA, L.; AMBRIZZI, T.; BORGES, I. V. G.; TUNALI, M.; BASIBUYUK, S.; YENIGÜN, O.; MASSONI, T. L.; JONES, K.; CAMPOS, L.; KOSTKOVA, P.; FILHO, A. G. da S.; SANTOS, W. P. dos. Spatiotemporal forecasting for dengue, chikungunya fever and zika using machine learning and artificial expert committees based on meta-heuristics.

Research on Biomedical Engineering, Springer-Nature, v. 2022, p. s42600–022–00202–6, 2022.

SILVA, L. J. d. O conceito de espaço na epidemiologia das doenças infecciosas. *Cadernos de Saúde Pública*, SciELO Public Health, v. 13, p. 585–593, 1997.

SILVEIRA, P. H. M. *Programa Saúde da Família: Uma discussão sobre o modelo de atenção básica à saúde*. Recife: Masangana, 2010.

SIMMONS, C. P.; FARRAR, J. J.; CHAU, N. van V.; WILLS, B. Dengue. *New England Journal of Medicine*, Mass Medical Soc, v. 366, n. 15, p. 1423–1432, 2012.

SINGH, P.; SINGH, A. K.; SINGH, P.; KUMARI, S.; SANGAIAH, A. K. Multimodal data modeling for efficiency assessment of social priority based urban bus route transportation system using gis and data envelopment analysis. *Multimedia Tools and Applications*, Springer, v. 78, n. 17, p. 23897–23915, 2019.

SINGHI, S.; KISSOON, N.; BANSAL, A. Dengue and dengue hemorrhagic fever: management issues in an intensive care unit. *Jornal de pediatria*, SciELO Brasil, v. 83, n. 2, p. S22–S35, 2007.

SIRIYASATIEN, P.; CHADSUTHI, S.; JAMPACHAISRI, K.; KESORN, K. Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes. *IEEE Access*, IEEE, v. 6, p. 53757–53795, 2018.

SIRIYASATIEN, P.; CHADSUTHI, S.; JAMPACHAISRI, K.; KESORN, K. Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes. *IEEE Access*, v. 6, p. 53757–53795, 2018. ISSN 2169-3536.

SMITH, B. L. The third industrial revolution: Policymaking for the internet. *Colum. Sci. & Tech. L. Rev.*, HeinOnline, v. 3, p. 1, 2001.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and Computing*, Springer, v. 14, n. 3, p. 199–222, 2004.

SNOW, J. *On the mode of communication of cholera*. [S.l.]: John Churchill, 1855.

SNOW, J. Cholera, and the water supply in the south districts of London. *British Medical Journal*, BMJ Publishing Group, v. 1, n. 42, p. 864, 1857.

SOLEYMANI, M.; GARCIA, D.; JOU, B.; SCHULLER, B.; CHANG, S.-F.; PANTIC, M. A survey of multimodal sentiment analysis. *Image and Vision Computing*, Elsevier, v. 65, p. 3–14, 2017.

SOUSA, L.; MELLO, R. de; CEDRIM, D.; GARCIA, A.; MISSIER, P.; UCHÔA, A.; OLIVEIRA, A.; ROMANOVSKY, A. VazaDengue: An information system for preventing and combating mosquito-borne diseases with social networks. *Information Systems*, Elsevier, v. 75, p. 26–42, 2018.

SOUZA, D. B. de; CUNHA, J. T. da; FIGUEIROA, E.; CORREIA, J. B.; SILVA, H. P. da; FILHO, J. L. de L.; ALBUQUERQUE, J.; SANTOS, F. A. Using curvature to infer covid-19 fractal epidemic network fragility and systemic risk. <https://www.medrxiv.org/content/10.1101/2020.04.01.20047225.v1>, 2020.

SOUZA, D. B. de; CUNHA, J. T. da; SANTOS, E. F. dos; CORREIA, J. B.; SILVA, H. P. da; FILHO, J. L. de L.; ALBUQUERQUE, J.; SANTOS, F. A. Using discrete Ricci curvatures to infer COVID-19 epidemic network fragility and systemic risk. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2021, n. 5, p. 053501, 2021.

STANDISH, R. SciDAVis. *Free Application for Scientific Data Analysis and Visualization*, 2016.

STEIGER, E.; RESCH, B.; ZIPF, A. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, Taylor & Francis, v. 30, n. 9, p. 1694–1716, 2016.

STEPHEN, J. M. Epidemiological and dietary aspects of rickets and osteomalacia. *Proceedings of the Nutrition Society*, Cambridge University Press, v. 34, n. 2, p. 131–138, 1975.

STOLERMAN, L. M.; MAIA, P. D.; KUTZ, J. N. Forecasting dengue fever in brazil: An assessment of climate conditions. *PLoS One*, Public Library of Science San Francisco, v. 14, n. 8, p. e0220106, 2019.

SUN, Y.; WANG, X.; TANG, X. Deep learning face representation from predicting 10,000 classes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1891–1898.

SUYKENS, J. A.; VANDEWALLE, J. Least squares support vector machine classifiers. *Neural Processing Letters*, Springer, v. 9, n. 3, p. 293–300, 1999.

SZOMSZOR, M.; KOSTKOVA, P.; QUINCEY, E. D. #Swineflu: Twitter predicts swine flu outbreak in 2009. In: SPRINGER. *International Conference on Electronic Healthcare*. [S.l.], 2010. p. 18–26.

TANG, J.; DENG, C.; HUANG, G.-B. Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning systems*, IEEE, v. 27, n. 4, p. 809–821, 2015.

TANNER, L.; SCHREIBER, M.; LOW, J. G.; ONG, A.; TOLFVENSTAM, T.; LAI, Y. L.; NG, L. C.; LEO, Y. S.; PUONG, L. T.; VASUDEVAN, S. G. et al. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS Neglected Tropical Diseases*, Public Library of Science, v. 2, n. 3, p. e196, 2008.

TAVANAEI, A.; GHODRATI, M.; KHERADPISHEH, S. R.; MASQUELIER, T.; MAIDA, A. Deep learning in spiking neural networks. *Neural Networks*, Elsevier, v. 111, p. 47–63, 2019.

TCHAKOUCHE, T. A.; EZZIYYANI, M. Multilayered echo-state machine: A novel architecture for efficient intrusion detection. *IEEE Access*, IEEE, v. 6, p. 72458–72468, 2018.

THAKUR, P.; KAUR, S. An intelligent system for predicting and preventing Chikungunya virus. In: IEEE. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. [S.l.], 2017. p. 3483–3492.

THOMSON, M. C.; GARCIA-HERRERA, R.; BENISTON, M. *Seasonal Forecasts, Climatic Change and Human Health*. [S.l.]: Springer, 2008.

- TIAN, D.; SHI, Z. MPSO: Modified particle swarm optimization and its applications. *Swarm and Evolutionary Computation*, Elsevier, v. 41, p. 49–68, 2018.
- TOSEPU, R.; TANTRAKARNAPA, K.; WORAKHUNPISET, S.; NAKHAPAKORN, K. Climatic Factors Influencing Dengue Hemorrhagic Fever in Kolaka District, Indonesia. *Environment and Natural Resources Journal*, v. 16, n. 2, p. 1–10, 2018.
- TUFTE, E. R.; MCKAY, S. R.; CHRISTIAN, W.; MATEY, J. R. *Visual explanations: images and quantities, evidence and narrative*. [S.l.]: AIP, 1998.
- VASCONCELOS, P. F. d. C. Doença pelo vírus Zika: um novo problema emergente nas Américas? *Revista Pan-Amazônica de Saúde*, Instituto Evandro Chagas/Secretaria de Vigilância em Saúde / Ministério da Saúde, v. 6, n. 2, p. 9–10, 2015.
- VERSTRAETEN, D. *Reservoir computing: computation with dynamical systems*. Tese (Doutorado) — Ghent University, 2009.
- VILLAMIL-GÓMEZ, W. E.; GONZÁLEZ-CAMARGO, O.; RODRIGUEZ-AYUBI, J.; ZAPATA-SERPA, D.; RODRIGUEZ-MORALES, A. J. Dengue, chikungunya and Zika co-infection in a patient from Colombia. *Journal of Infection and Public Health*, Elsevier, v. 9, n. 5, p. 684–686, 2016.
- WAN, H. Modelling mosquito population dynamics: The impact of resource and temperature. In: TRANS TECH PUBL. *Advanced Materials Research*. [S.l.], 2013. v. 726, p. 156–159.
- WAN, H.; ZHU, H. A new model with delay for mosquito population dynamics. *Mathematical Biosciences and Engineering*, v. 11, n. 6, p. 1395–1410, 2014.
- WANG, H.; WANG, W.; XIAO, S.; CUI, Z.; XU, M.; ZHOU, X. Improving artificial bee colony algorithm using a new neighborhood selection mechanism. *Information Sciences*, Elsevier, v. 527, p. 227–240, 2020.
- WANG, L.; RANJAN, R. Processing distributed internet of things data in clouds. *IEEE Cloud Computing*, IEEE, v. 2, n. 1, p. 76–80, 2015.
- WHITEMAN, A.; LOAIZA, J. R.; YEE, D. A.; POH, K. C.; WATKINS, A. S.; LUCAS, K. J.; RAPP, T. J.; KLINE, L.; AHMED, A.; CHEN, S. et al. Do socioeconomic factors drive aedes mosquito vectors and their arboviral diseases? a systematic review of dengue, chikungunya, yellow fever, and zika virus. *One Health*, Elsevier, p. 100188, 2020.
- WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Technique*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2005.
- WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016.
- WONGKOON, S.; JAROENSUTASINEE, M.; JAROENSUTASINEE, K. Development of temporal modeling for prediction of dengue infection in Northeastern Thailand. *Asian Pacific Journal of Tropical Medicine*, Elsevier, v. 5, n. 3, p. 249–252, 2012.
- XU, F.; LI, H.; PUN, C.-M.; HU, H.; LI, Y.; SONG, Y.; GAO, H. A new global best guided artificial bee colony algorithm with application in robot path planning. *Applied Soft Computing*, Elsevier, v. 88, p. 106037, 2020.

- XU, M.; DAVID, J. M.; KIM, S. H. et al. The fourth industrial revolution: Opportunities and challenges. *International Journal of Financial Research*, Sciedu Press, v. 9, n. 2, p. 90–95, 2018.
- YAMASHITA, W.; TAKAHASHI, L.; CHAPIRO, G. Dispersive models describing mosquitoes' population dynamics. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2016. v. 738, n. 1, p. 012065.
- YAMASHITA, W. M.; DAS, S. S.; CHAPIRO, G. Numerical modeling of mosquito population dynamics of aedes aegypti. *Parasites & Vectors*, BioMed Central, v. 11, n. 1, p. 245, 2018.
- YASEEN, Z. M.; SULAIMAN, S. O.; DEO, R. C.; CHAU, K.-W. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology*, Elsevier, v. 569, p. 387–408, 2019.
- YAZDAVAR, A. H.; MAHDAVINEJAD, M. S.; BAJAJ, G.; ROMINE, W.; SHETH, A.; MONADJEMI, A. H.; THIRUNARAYAN, K.; MEDDAR, J. M.; MYERS, A.; PATHAK, J. et al. Multimodal mental health analysis in social media. *Plos One*, Public Library of Science San Francisco, CA USA, v. 15, n. 4, p. e0226248, 2020.
- YU, Q.; CHEN, L.; LI, B. Ant colony optimization applied to web service compositions in cloud computing. *Computers & Electrical Engineering*, Elsevier, v. 41, p. 18–27, 2015.
- YUSOF, Y.; MUSTAFFA, Z. Dengue outbreak prediction: A least squares support vector machines approach. *International Journal of Computer Theory and Engineering*, IACSIT Press, v. 3, n. 4, p. 489, 2011.
- ZANLUCA, C.; MELO, V. C. A. d.; MOSIMANN, A. L. P.; SANTOS, G. I. V. d.; SANTOS, C. N. D. d.; LUZ, K. First report of autochthonous transmission of Zika virus in Brazil. *Memórias do Instituto Oswaldo Cruz*, SciELO Brasil, v. 110, n. 4, p. 569–572, 2015.
- ZEMMAL, N.; AZIZI, N.; SELLAMI, M.; CHERIGUENE, S.; ZIANI, A.; ALDWAIRI, M.; DENDANI, N. Particle swarm optimization based swarm intelligence for active learning improvement: Application on medical data classification. *Cognitive Computation*, Springer, v. 12, n. 5, p. 991–1010, 2020.
- ZHANG, C.; ZHOU, K.; YANG, S.; SHAO, Z. On electricity consumption and economic growth in china. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 76, p. 353–368, 2017.
- ZHANG, D.; PENG, X.; PAN, K.; LIU, Y. A novel wind speed forecasting based on hybrid decomposition and online sequential outlier robust extreme learning machine. *Energy Conversion and Management*, Elsevier, v. 180, p. 338–357, 2019.
- ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, Elsevier, v. 14, n. 1, p. 35–62, 1998.
- ZHANG, G. P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, Elsevier, v. 50, p. 159–175, 2003.

- ZHAO, N.; CHARLAND, K.; CARABALI, M.; NSOESIE, E. O.; MAHEU-GIROUX, M.; REES, E.; YUAN, M.; BALAGUERA, C. G.; RAMIREZ, G. J.; ZINSZER, K. 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*, Public Library of Science, v. 14, n. 9, p. e0008056, 2020.
- ZHENG, Z.; SAXENA, N.; MISHRA, K.; SANGAIAH, A. K. Guided dynamic particle swarm optimization for optimizing digital image watermarking in industry applications. *Future Generation Computer Systems*, Elsevier, v. 88, p. 92–106, 2018.
- ZHONG, Y.; YANG, H.; ZHANG, Y.; LI, P. Online random forests regression with memories. *Knowledge-Based Systems*, Elsevier, v. 201, p. 106058, 2020.
- ZHOU, Z.-H.; CHEN, Z.-Q. Hybrid decision tree. *Knowledge-based systems*, Elsevier, v. 15, n. 8, p. 515–528, 2002.
- ZHU, G.; KWONG, S. Gbest-guided artificial bee colony algorithm for numerical function optimization. *Applied Mathematics and Computation*, Elsevier, v. 217, n. 7, p. 3166–3173, 2010.

ANEXO A – CARTA DE ANUÊNCIA DA SECRETARIA DE SAÚDE DA PREFEITURA DA CIDADE DO RECIFE



CARTA DE ANUÊNCIA

Autorizo **Cecilia Cordeiro da Silva**, pesquisadora da Universidade Federal de Pernambuco, a desenvolver pesquisa na Diretoria Executiva de Vigilância à Saúde, da Secretaria de Saúde do Recife, sob o título: **"Gamificação para treinamento de profissionais de saúde em vigilância participativa e protocolos associados ao vírus Zika"**, sendo orientada por Abel Guilhermino da Silva Filho e coorientação de Wellington Pinheiro dos Santos.

Estarei ciente que me são resguardados e abaixo listados:

- O cumprimento das determinações éticas das resoluções 466/12 e 510/16 do Conselho Nacional de Saúde.
- A garantia de solicitar e receber esclarecimentos, antes e durante o curso da pesquisa;
- A liberdade de recusar a participar ou retirar minha anuência, em qualquer fase da pesquisa, sem penalização alguma;
- A garantia de que nenhuma das pessoas envolvidas será identificada e terá assegurado privacidade quanto aos dados envolvidos na pesquisa;
- Não haverá nenhuma despesa para a Secretaria de Saúde do Recife decorrente da participação na pesquisa.

O(s) pesquisador(es) comprometem-se a trazer para esta diretoria o relatório final da pesquisa através de cópia em *Compact Disk* (CD), uma vez que só serão autorizadas novas pesquisas se não houver pendências de devolutiva do serviço.

Tenho ciência do exposto e concordo em fornecer subsídios para a pesquisa.

Recife, 12 de maio de 2017.

Atenciosamente,

Valderiz Ribeiro

Gestora da Gerência de Formação e Educação na Saúde

Valderiz Ribeiro de Andrade
Gerente de Unidade de Gestão de
Educação na Saúde (REGE-SUS-SA)
Matrícula 103.755.2