Prediction of Aedes aegypti breeding distribution through spatiotemporal analysis and machine learning: A case study in Recife, Pernambuco

Clarisse Lins de Lima
Universidade de Pernambuco
Cecilia Cordeiro da Silva
Universidade Federal de Pernambuco
Ana Clara Gomes da Silva
Universidade Federal de Pernambuco
Aracely Andrade da Silva
Universidade Federal de Pernambuco
Felipe Rodrigues de Almeida
Universidade Federal de Pernambuco
Cristine Martins Gomes de Gusmão
Universidade Federal de Pernambuco
Giselle Machado Magalhães Moreno
Universidade de Sao Paulo
Anwar Musah
University College London
Aisha Aldosery
University College London
Ella Browning
University College London
Livia Dutra
Universidade de Sao Paulo
Tercio Ambrizzi
Universidade de Sao Paulo
Iuri V. G. Borges
Universidade de Sao Paulo
Merve Tunali
Bogazici Universitesi
Selma Basibuyuk
Bogazici Universitesi
Orhan Yenigün

Bogazici Universitesi
Tiago Lima Massoni
Universidade Federal de Campina Grande
Kate Jones
University College London
Luiza C. Campos
University College London
Patty Kostkova
University College London
Abel Guilhermino da Silva Filho
Universidade Federal de Pernambuco
Wellington Pinheiro dos Santos (wellington.santos@ufpe.br)
Universidade Federal de Pernambuco https://orcid.org/0000-0003-2558-6602

Research Article

Keywords: Forecasting, Arbovirus infections, Machine learning, Ovoposition, Mosquito control, Aedes aegypti

Posted Date: March 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1200442/v1

Prediction of Aedes aegypti breeding distribution through spatiotemporal analysis and machine learning: A case study in Recife, Pernambuco

Clarisse Lins de Lima[1], Cecilia Cordeiro da Silva[2], Ana Clara Gomes da Silva[3], Aracely Andrade da Silva[3], Felipe Rodrigues de Almeida[4], Cristine Martins Gomes de Gusmão[3], Giselle Machado Magalhães Moreno[5], Anwar Musah[7], Aisha Aldosery[6], Ella Browning[10], Livia Dutra[5], Tercio Ambrizzi[5], Iuri V. G. Borges[5], Merve Tunali[8], Selma Basibuyuk[8], Orhan Yenigün[8], Tiago Lima Massoni[9], Kate Jones[10], Luiza C. Campos[11], Patty Kostkova[6], Abel Guilhermino da Silva Filho[2], Wellington Pinheiro dos Santos[3*]

[1]Escola Politécnica da Universidade de Pernambuco, Recife, Brazil, E-mail: cll@ecomp.poli.br

[2]Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil, E-mails: ccs2@cin.ufpe.br, agsf@cin.ufpe.br

[3]Departamento de Engenharia Biomédica, Universidade Federal de Pernambuco, Recife, Brazil, E-mails: clara.gomes@ufpe.br, aracely_andrade@hotmail.com, cristine.gusmao@ufpe.br, wellington.santos@ufpe.br

[4] Programa de Pós-Graduação em Odontologia, Universidade Federal de Pernambuco, Recife, Brazil, E-mails: almeidabiomed@gmail.com

[5]Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, Brazil, E-mail: gisellemoreno@usp.br, livia.dutra@iag.usp.br, terzio.ambrizzi@iag.usp.br, iurivalerio@usp.br

[6]Centre for Digital Public Health and Emergencies, Institute for Risk and Disaster Reduction, University College London, United Kingdom, E-mails: a.aldosery@ucl.ac.uk, p.kostkova@ucl.ac.uk

[7]Department of Geography, University College London, London, United Kingdom, Email: a.musah@ucl.ac.uk

[8]Boğaziçi Üniversitesi, Institute of Environmental Sciences, Istanbul, Turkey, E-mails: merve.tunali@boun.edu.tr, selmabasibuyuk@gmail.com, yeniguno@boun.edu.tr

[9]Departamento de Sistemas e Computação, Universidade Federal de Campina Grande, Campina Grande, Brazil, E-mail: massoni@dsc.ufcg.edu.br

[10]Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, United Kingdom, E-mail: ella.browning.14@ucl.ac.uk, kate.e.jones@ucl.ac.uk

[11]Department of Civil Environmental & Geomatic Engineering, University College London, United Kingdom, E-mail: l.campos@ucl.ac.uk

*Corresponding author: Wellington Pinheiro dos Santos, E-mail: wellington.santos@ufpe.br,

ORCID 0000-0003-2558-6602

Abstract

Purpose: *Aedes aegypti* is a mosquito responsible for transmitting mainly dengue, zika, and chikungunya. In low- and middle-income countries, controlling the spread of this mosquito poses a major public health challenge. Currently, *Aedes aegypti* control policies are extremely important for lowering the risk of potential arbovirus outbreaks. One of the effective strategies for combating the burden of mosquito-borne arboviruses are the pre-emptive predictions and forecasts for future outbreaks. In this sense, we, therefore, apply machine learning using a spatiotemporal approach to build distribution maps of *Aedes aegypti* breeding sites in Recife.

Methods: We obtained data from *Aedes aegypti* breeding sites and climatic factors in Recife City during 2013-2014. From the information of the breeding sites, bimonthly spatial distribution maps of the breeding sites were generated using the Inverse Distance Interpolation (IDW). We generated monthly spatial distribution maps

of climatic variables (temperature, rainfall, and wind speed) using the same method. From the generated distribution maps, several models were evaluated, among them: support vector regressor, multilayer perceptron, random forest, and linear regression. The model performance was evaluated according to Pearson's correlation coefficient and percentage root relative squared error (RRSE%) metrics.

Results: Among the evaluated regressors, the 3-degree polynomial-kernel support vector regressor showed superior performance compared to the other regressors evaluated. For this regressor, the correlation coefficient was on average 0.9875 (and standard deviation of 0.01) while the RRSE% metric was on average 14.60%.

Conclusion: Machine learning proved to be a promising tool in predicting the *Aedes aegypti* breeding sites' spatial distribution in the city of Recife. The spatiotemporal predictions pointed out that neighborhoods with low income and lack of water supply are presented with elevated concentrations of mosquito breeding sites. The findings of this work can support health authorities in decision-making linked to policies for reducing the burden of mosquito infestation, while at the same time allowing authorities to optimize their limited resources in low-resource settings.

Keywords: Forecasting • Arbovirus infections • Machine learning • Ovoposition • Mosquito control • *Aedes aegypti*

1. Introduction

The *Aedes aegypti* mosquito is the main vector responsible for the transmission of dengue viruses to humans. Dengue is an endemic disease in several tropical regions. However, its rapid global spread raises new challenges for the scientific community and government entities (Wilder-Smith et al. 2017). It is currently estimated that approximately half of the world's population is at risk of being infected by the dengue virus. In the Americas, for example, only Chile and Canada do not have *Aedes aegypti* occurrence records (PAHO 2021).

Dengue fever is a fatal mosquito-borne disease caused by four different virus serotypes (DENV-1, DENV-2, DENV-3, and DENV-4) that sometimes circulate simultaneously in the human body. The immunity acquired in response to infection is lifelong, however, cross-immunity increases the risks for the development of the most severe form of the disease. This is probably the most limiting aspect of the development of vaccines against dengue. *Aedes aegypti* is also responsible for the transmission of other arboviruses such as Zika and Chikungunya. Chikungunya fever is a disease caused by an alphavirus, wherein the main symptoms are fever, muscle weakness, and arthralgias in peripheral joints (ankle, knees, and wrists). In most cases, the disease has a benign course, but in more severe cases individuals can develop neurological complications such as Guillain-Barre Syndrome. Zika, in turn, is a disease caused by a virus belonging to the Flaviviridae family. The most common symptoms of Zika are fever, rash, myalgia, and arthralgia, which are less intense than in Chikungunya cases. Zika virus is also partially related to cases of microcephaly in newborns (Cao-Lormeau et al. 2016).

The transmission of Dengue, Zika, and Chikungunya occurs mainly through the bites of infected *Aedes aegypti* female mosquitoes. These mosquitoes deposit their eggs in water reservoirs and prefer to feed on human

blood. The *Ae. aegypti* occurrence is, therefore, directly related to population density and lack of infrastructure. These factors lead to water storage in inadequate reservoirs due to a lack of water supply. Other important factors influencing the life cycle of mosquitoes are temperature, humidity, and rainfall. The temperature acts by accelerating the mosquito's life cycle, whereas the rainfall increases the availability of breeding sites for laying eggs (Fered et al. 2018; Tosepu et al. 2018; Siriyasatien et al. 2018).

Monitoring the *Aedes aegypti* occurrence is the main strategy for controlling these diseases. By monitoring the mosquito, it is possible to interfere in its life cycle with relatively simple treatment measures and/or breeding sites elimination. In Brazil's Northeast region, where such arboviruses are endemic, there are special teams of public health agents: the environmental health agents (EHA). These agents are responsible for identifying, treating, or eliminating potential breeding sites. They are also responsible for educating the population and reporting suspected cases of Dengue, Zika, and Chikungunya. The environmental health agents are distributed across cities' territories according to their population densities. In the event of an arbovirus outbreak, the agents are located from less-affected regions to more affected regions, wherein they carry out task forces to eliminate and treat the breeding sites (Ministério da Saúde 2009).

Despite its importance in containing the advance of arboviruses outbreaks, the environmental surveillance system is not capable to anticipate outbreaks throughout the year. Hence, it is crucial to explore different strategies that can support health managers to prevent future outbreaks. Therefore, spatiotemporal distribution models for *Aedes aegypti* breeding sites can be an alternative to put public policies in an advantageous position in the fight against Dengue, Zika, and Chikungunya. In this sense, the present work was guided by the following research question: how to build low-cost and effective spatiotemporal model for predicting the locations and abundance of *Aedes aegypti* breeding sites, using geographic information and climate variables databases?

Therefore, in this work, we proposed a method for the spatiotemporal prediction of *Aedes aegypti* mosquito breeding sites in Recife. In our proposal, we collected climate data from the National Institute of Meteorology database, and the Pernambuco Agency of Waters and Climates database. We also obtained information regarding the *Aedes aegypti* breeding sites collected in the Rapid Survey for *Aedes aegypti* Indices for the City of Recife. Machine learning algorithms such as linear regression, support vector for regression, and artificial neural networks were used to build spatiotemporal models. The goal is to contribute to the development of a tool to support public health managers in planning, health surveillance, and decision-making to prevent the vector.

2. Related Works

Among the existing prediction models, the use of geospatial technologies is essential. Some central aspects are considered: rainfall (Rahman et al. 2020), temperature variations (Peña-García et al. 2016), number of reported arbovirus cases (Freitas et al. 2018), infrastructure data and environmental conditions (Espinosa et al. 2016), and mosquito vector population analysis (Fong-Shue et al. 2015). One of the main methods used for population analysis of mosquitoes is the use of ovitraps i.e., oviposition traps. These traps are regularly

inspected. They are an effective method to provide useful data on the spatial distribution of *A. aegypti* and other mosquito species, which allows for a better understanding of their population and activity (Scavuzzo et al. 2017).

Prediction models that use mathematical-computational modeling and simulation constitute an interesting tool for the study of complex phenomena, thus being able to support planning, scenario evaluation, and decision-making activities to for example identify spatiotemporal patterns of transmission of arboviruses, both in relation to viral transmission and the ecology of their vector, which is influenced by a wide range of environmental and social factors.

Several studies have been carried out by the academic community to explore different techniques for the reduction and control of *A. aegypti* since it has a social and environmental impact. Verísimo et al. (2016) proposed dengue control policies in each season of the year, through the genetic algorithm (GA), in order to minimize public management expenses related to *A.aegypti* control. In this algorithm, the search follows the direction of a trend with information akin to that provided by the gradient, which is not done by any conventional genetic operator. This maintains the advantage of AG of evaluating only the objective function, without having to calculate any derivatives. In this work, three dengue vector control policies were proposed, considering the year with 360 days, and divided into three periods: favorable, intermediate, and unfavorable. The actual polarized genetic algorithm was implemented in the MATLAB software, being simulated 30 times. The study found that the control performed in the summer season was more efficient when compared to the control performed in the other seasons of the year.

Freitas et al. (2018) in their research proposed statistical models of time series capable of predicting possible outbreaks. The data used to refer to the number of monthly notifications of dengue in five municipalities in Pernambuco (Serra Talhada, Ipojuca, Vitória de Santo Antão, Recife and Petrolina) that were made available by the National Meteorology Institute (INMET), in the period of January 2000 to December 2016. The Box & Jenkins (Hyndman and Athanasopoulos, 2018) methodology was used in their study, that is, the SARIMA models. Predictions and residual analyses were made from the proposed models. It was found that the residues do not behave like a sequence of independent random variables, identically distributed with zero mean and constant variance. These models can capture the level and slope (central values), such models are good at predicting observations over long periods of time. They concluded that the SARIMA model presents better results, as it manages to capture seasonal behaviors, in addition to presenting with good accuracy the cases of dengue in such municipalities. As an advantage, this model allows predicting the number of cases in periods after the series studied. The use of SARIMA (Seasonal Auto-Regressive Integrated Moving Average) is useful in situations where time-series data exhibit periodic seasonal fluctuations that are repeated with approximately the same intensity each year.

Mittelmann & Soares (2017) proposed to develop a model for predicting Dengue cases with Artificial Neural Networks (ANNs) for the city of Guarulhos - São Paulo. This work compares the use of two distinct RNAs, the Multilayer Perceptron networks (MLPs) and the Auto-Regressive Neural Networks (ARNNs) with exogenous inputs. The MLPs and NARXs networks were compared to choose the network with the best performance in dengue prediction in the studied area. The best performance was obtained by MLP modeled with 10 neurons in the hidden layer. The input set that obtained this result consisted of the number of Dengue cases in

the two months prior to the month in which the forecast was made. It is noted that the climatic variables were not the most adequate to estimate the number of Dengue cases, as the network trained with information on past cases performed better. The data samples of the city studied are made up of five input attributes (total precipitation, maximum temperature, minimum temperature, average air temperature and average relative humidity) and one output (number of Dengue cases). The meteorological database consists of monthly data from January 2009 to November 2014 for total precipitation (mm), average maximum temperature, average compensated temperature, average minimum temperature, and average relative humidity (%). Only quantitative data related to indigenous dengue cases in the city were selected to achieve a greater correlation with the region's climatic variables. 70% of the total set of available data was randomly chosen for the training subset. In turn, the remaining data (30%) integrated the subset of testing and validation. Networks with different initial parameters were modeled and trained. The results presented by the networks show that it is possible to forecast Dengue cases in the study area, with acceptable error and advance, using neural networks with meteorological and/or historical data on the number of dengue cases from previous months. However, all mathematical modeling, the generalization of the results obtained in a specific case study cannot be used directly for other cases, as it must be considered that the characteristics of each region are unique, being the use of the models built in this work restricted to the area and selected study.

Mattioli, Andrade & Estevez (2017) presented a prediction model for Dengue cases through the application of multi-layered Perceptron neural network training. They performed tests with different configurations, and for each configuration 10 repetitions were performed, calculating the average error and the average execution time. The neural network was fed with data provided by the State Department of Health and SINAM (Notifiable Diseases Information System). It was observed that the resulting neural network was very close to the result obtained by a real series that served as a comparison parameter to verify the performance of the developed neural network. This work used an artificial multi-layered Perceptron (MLP) neural network (Haykin, 2001). Being indicated for data prediction since its training can be supervised with backpropagation algorithms. MLPs have been applied, through their supervised training with an error backpropagation algorithm. They concluded that the neural network developed in the work carried out could serve as a method of prediction and could suggest future acts of prevention against dengue.

Melo & Moraes (2018) aimed at the development of a decision support system based on spatial information, which considers different factors related to an uncertain situation through the application of a system based on fuzzy rules that allow the generation of maps to detect priority areas for combating Dengue. This architecture has the advantage of being flexible and allows only a few modules to be used to solve problems, according to its specificity and need. The maps aim to guide health managers in making decisions, such as the intervention or not about certain diseases. The system was applied to the study in the state of Paraíba between the years 2009 to 2013. The results obtained are used to generate choropleth maps (colored) to better visualize the geographic region. The space for decisions using fuzzy rules is given by the set: Non-priority, Possibly non-priority, Possibly priority, and Priority, which represents all priority levels in which a municipality can be classified. The model may inform the condition of the municipalities that should be given priority. 1115

applications were made to cover the 223 municipalities of Paraíba in the five years covered, so that in the end the five decision maps were generated, one for each year.

Hamlet et al. (2018) sought to assess the influence of seasonal variations on climatic factors in association with the seasonality of *A. aegypti* reports. To this end, a temperature adequacy index for the transmission of yellow fever was constructed, capturing the temperature dependence of the vector's behavior and the viral replication within it. Also, a series of multilevel logistic regression models were adjusted to a set of artificial intelligence reporting data across the African continent, considering the location and seasonality of the occurrence of seasonal models, as to the temperature adequacy index, precipitation, and improved vegetation index as covariates, associated with other demographic indicators. The adjustment made in the model was evaluated by the area under the curve and these models were classified according to the Akaike information criterion, which was used to weight the model's outputs to create combined model forecasts. The seasonal model was able to accurately capture the geographic and temporal heterogeneities in the transmission of yellow fever, with no significantly worse performance than the annual model, which only captured the geographical distribution. The validity of the forecasts was assessed through the cross-validation of single exclusion, where the data set was divided by randomly assigning countries to subsets, the models being adjusted to the data set, aiming to create out-of-sample forecasts, being repeated 10 times, which resulted in 10 different provincial allocations. For each of these, the average between the 10 achievements was calculated. The relationship between adequacy of temperature and precipitation was considered responsible for most of the occurrence of the disease, thus offering a statistical explanation for the spatiotemporal variability in its transmission.

Salami et al. (2020) predicted the importation of Dengue cases in Europe in 21 countries, through machine learning and independent model methods. Four classification algorithms were trained: partial least squares (pls), generalized linearized models of loop and elastic network (glmnet), random forest (randomForest), extreme gradient impulse (xgboost), using historical data from 06 years (2010-2015). The data set was randomly divided into two sets, 70% for training and 30% for testing.

The performance of each classifier was evaluated using the area under the characteristic curve of receiving operation (AUC), measures of sensitivity (true positive rate), and specificity (false positive rate). All four models performed well. However, when evaluating the AUC score, the randomForest and xgboost predictions were better adapted to the data set, with almost insignificant differences. The xgboost model outperforms randomForest, with a true positive rate of 0.88 and a false positive rate of 0.12, xgboost was chosen as the ideal model for the data set. Being able to predict 88% of dengue import cases in our test data set.

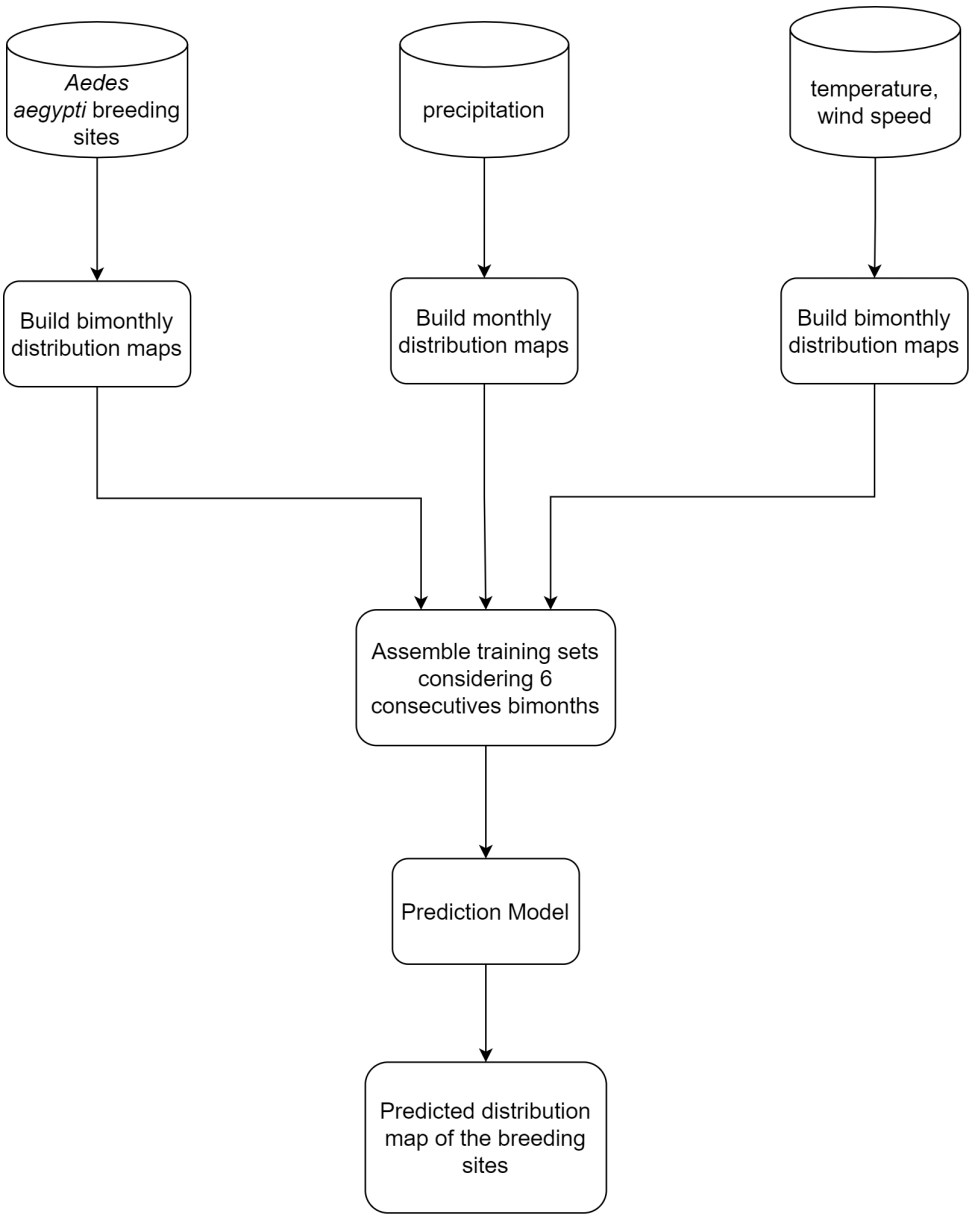
The most important predictor variables were the Dengue incidence rate in the country of origin, population size, and volume of air passengers. Air transport network centrality measures, describing the position of European countries in the air travel network, also influenced the predictions. It was concluded that the study had high predictive capacity. With the predictive model and the interpretability tools, it can be applied at the regional or national level to develop a prediction tool aiming to avoid the importation of Dengue.

3. Methods

3.1 Proposed method

In this work, we proposed a system to predict the spatiotemporal distribution of *Aedes aegypti* mosquitoes' breeding sites in the city of Recife. Then, we collected data related to breeding sites (number of mosquitoes breeding sites in each neighborhood) for each neighborhood of the city, as well as data associated with climatic variables (i.e., temperature, rain, and wind speed) from 2013 to 2016. From this information, we generated shapefile files (.shp format), to geographically locate breeding sites and climatic variables in each neighborhood in the city. In the next step, we generated spatial distribution maps using the Inverse Distance Weighted (IDW) Interpolation with QGIS software. In the case of *A. aegypti* mosquito breeding sites, two-month period maps were generated, whereas, for climatic variables, monthly distribution maps were generated. Then, we used a Python script to group the maps resulting from the previous process to store information in the following order: latitude, longitude, distribution of breeding sites, and - for each month of the two months - the distribution temperature, precipitation, and wind speed. Thus, we assembled the prediction sets so that the distribution maps of six consecutive bimesters are used to predict the breeding sites' distribution map for the following two months. Then, the Weka software is used to evaluate the regressors and predict the distribution maps of the breeding sites. Finally, the prediction maps are visualized using the QGIS software. Figure 1 illustrates the proposed method described above.

Figure 1. Proposed method: we obtained data on the distribution of *Aedes aegypti* breeding sites and climatic variables (rainfall, temperature and wind speed). From the data regarding the breeding sites, we generated bimonthly distribution maps using the Inverse Distance Weighted (IDW) interpolation. For the climatic variables (temperature, rain, and wind speed), we generated monthly distribution maps using the same interpolation method. Then, the prediction sets were assembled by concatenating the following information, in the following order: latitude, longitude, two-month period breeding sites, distribution of temperature, rainfall, and wind speed for the months of the respective bimester. From the prediction set, we generate models using linear regression, support vectors for regression, and artificial neural networks. Finally, we generated the prediction maps of *Aedes aegypti* breeding sites in Recife City using the model which showed the best performance.

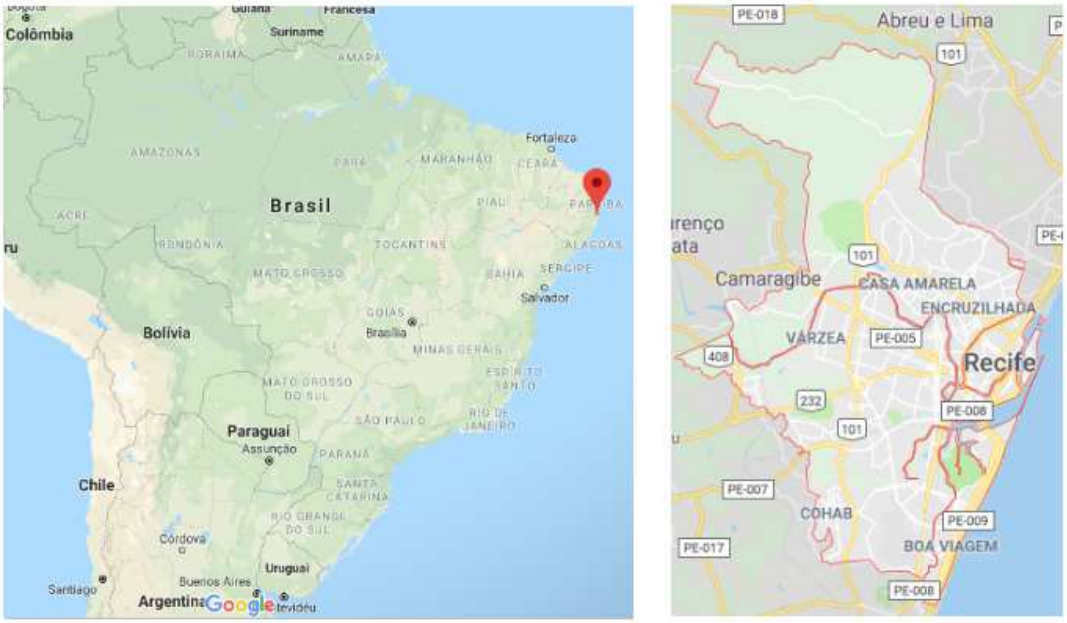


Source: the authors.

3.2 Study area

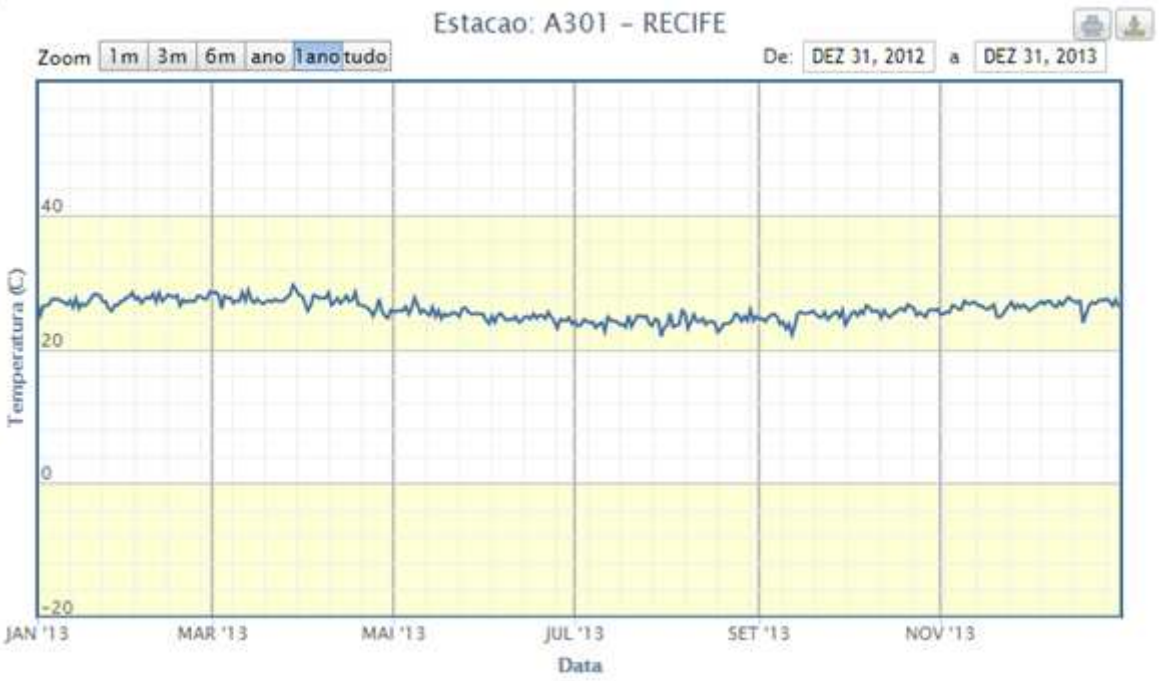
The delimited area for this study was the capital of the state of Pernambuco, Recife (8° 03'14" S, 34° 52'51" W), which is in the Northeast Region of Brazil (Figure 2). According to the Brazilian Institute of Geography and Statistics (IBGE), it has a territorial extension of approximately 218 km² and approximately 1,637,834 million inhabitants. According to the last census, Recife is the capital with the highest Human Development Index (HDI). Its climate is characterized as tropical humid, with an average monthly temperature above 18°C, high relative humidity, and high rainfall throughout the year, as presented in the Brazilian Institute of Meteorology charts (Figures 4 and 5, respectively).

Figure 2. Localization of the City of Recife.



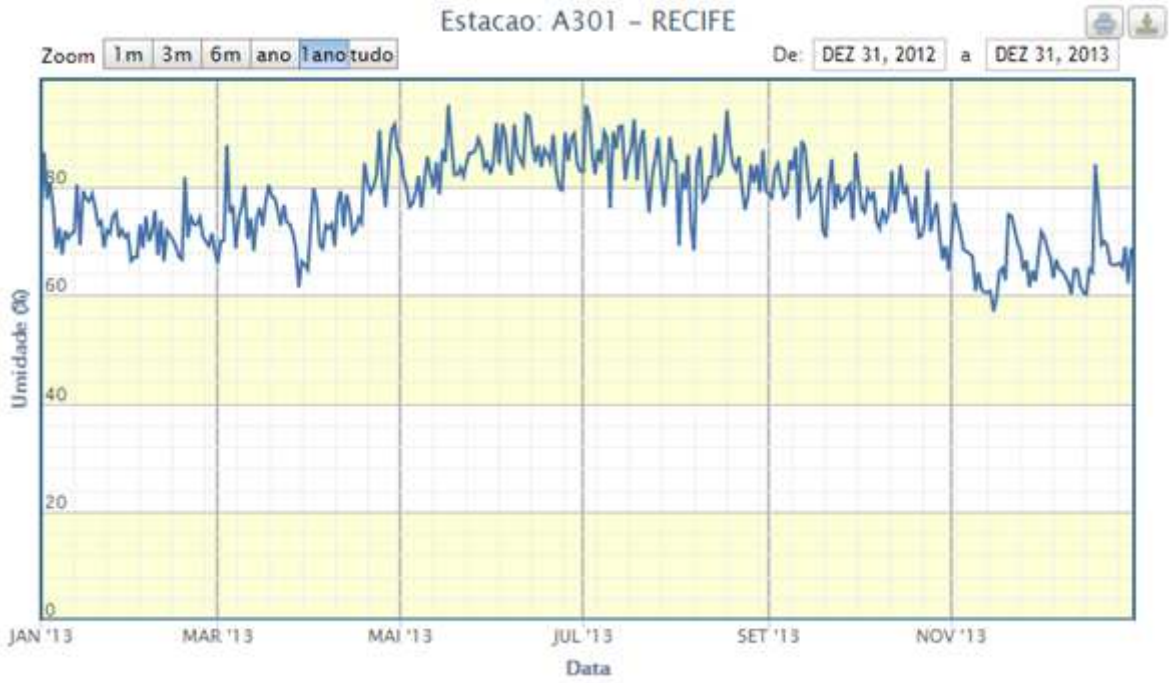
Source: Google Maps.

Figure 3. Temperature distribution in Recife (Brazil) corresponding to station A301.



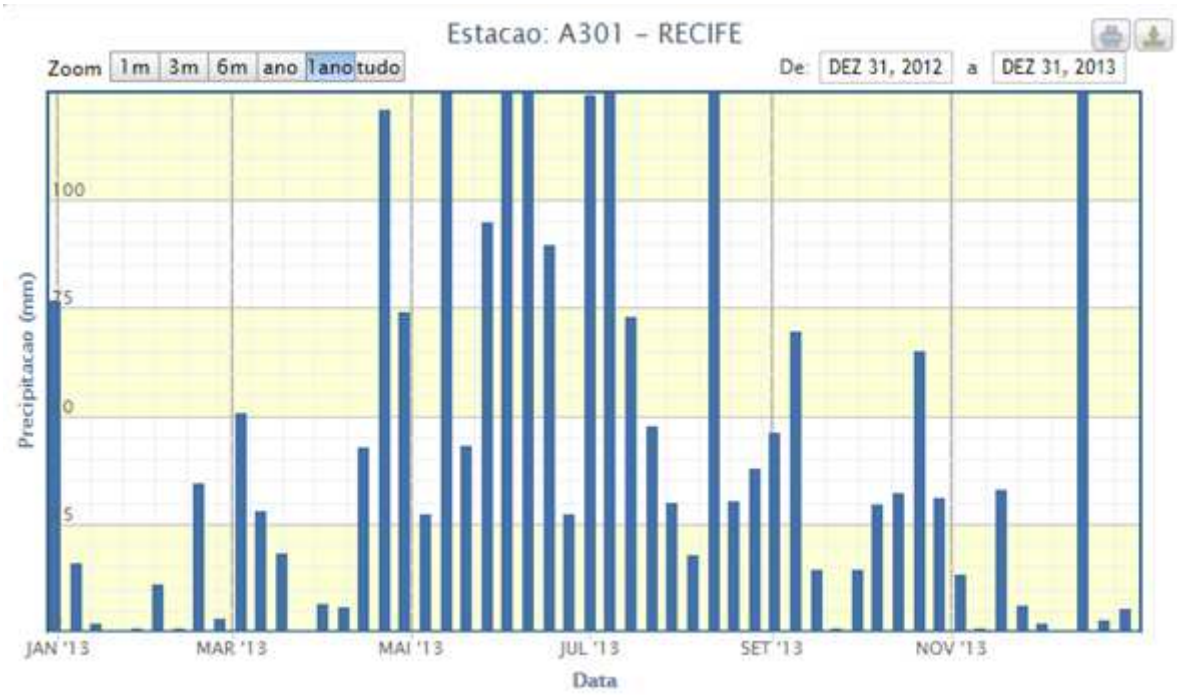
Source: National Institute of Meteorology.

Figure 4. Relative air humidity distribution in Recife (Brazil) corresponding to station A301.



Source: National Institute of Meteorology.

Figure 5. Rainfall distribution in Recife (Brazil) corresponds to station A301.



Source: National Institute of Meteorology

3.3 Databases

3.3.1 Rapid Index Survey for *Aedes aegypti*

The Rapid Index Survey for *Aedes aegypti* (LIRAa) is a tool to track and monitor the *Aedes aegypti* outbreak occurrence. It is an important apparatus that composes the epidemiological surveillance of the National Plan to Combat Dengue (PNCD). LIRAa has a simplified sampling method capable of providing systematic and

periodic information, quickly, with low cost and acceptable errors. In LIRAa's method, they use a stratified probability sampling of the total properties in the municipality's urban area. Therefore, the visits occur only in strategic points of the municipalities (properties more likely to develop breeding sites and that have a specific surveillance approach) (Brazil, 2013).

Thus, according to LIRA, the municipalities are divided into Health Districts and each health district is subdivided into strata. Each stratum contains several properties ranging from 8,000 to 12,000, however, the number considered ideal is around 9,000 properties. Sampling is carried out in two steps. The first one refers to the blocks that will be investigated. The second step is the selection of houses that will be visited in each neighborhood (Brazil, 2013).

The two main measures of the LIRAa are the Breteau Index (IB = Positive Recipients x 100 /Researched Property) and the Building Infestation Index (PI = Positive Property x 100 /Researched Property). These indices are relevant to classify the monitored areas into three levels according to the percentage of houses with mosquito breeding sites. The levels are classified as follows: satisfactory (<1% of the properties inspected); alert (between 1% and 3.9% of inspected properties) and risk (>4% of inspected properties). In addition to infestation rates, mosquito breeding sites found in each stratum are also computed. The information collected is stored by strata, which are divided considering the socioeconomic characteristics of the neighborhoods (Brazil, 2013). Therefore, in this work, we obtained data referring to the total number of breeding sites of the *Aedes aegypti* mosquito from the years 2013 to 2016.

3.3.2 Pernambuco Water and Climate Agency

The Pernambuco Water and Climate Agency (APAC) is a state agency that supports water resources management in the state of Pernambuco, Brazil. APAC monitors fluviometric and precipitation levels through Data Collection Platforms (PCDs) distributed throughout the territory of the state of Pernambuco. The PCDs provide information related to dams, reservoir water quality, as well as the State's hydrographic basins (APAC, 2017). The information collected is stored in the Pernambuco Hydrometeorological Geoinformation System (SIGHPE) database. It contains information on daily measurements and monthly accumulated data measured by each rainfall station. In Recife, there are three rainfall stations: Alto da Brasileira, Várzea and Codecipe/Santo Amaro (http://www.sirh.srh.pe.gov.br/apac/sighpe/). The data referring to rainfall were, thereby, obtained through the SIGHPE database. We collected the monthly accumulated rainfall from 2013 to 2016.

3.3.3 National Institute of Meteorology

The National Institute of Meteorology (INMET) is responsible for preparing and disclosing daily weather forecasts, warnings, and special weather bulletins. INMET collects meteorological data through sounding stations (radiosonde), automatic stations, and stations that are manually operated (National Institute of Meteorology [INMET], n.d). The data monitored by meteorological stations distributed throughout the territory are stored in the Meteorological Database for Teaching and Research (BDMEP). In this work, we collected the monthly records of wind speed and temperature data from 2013 to 2016 relative to station A301 which is in Recife.

3.4 Software

3.4.1 QGIS

QGIS (Quantum GIS) is an open-source geographic information system under the GPL (General Public License) and LGPL (Lesser General Public License) terms. It was initially developed by Gary Sherman in 2002, however, it is currently an official project of the Source Geospatial Foundation (OSGeo). The software was developed using the Qt (https://www.qt.io) and C++ tools and, since it is open-source software, their collaborators can inspect and/or modify the source code. The system runs on several platforms (Unix, Windows, macOS, and Android) and supports various vector, matrix, raster, and database formats and functionalities (QGIS, 2021). Additionally, the data is visualized in a thematic way, which can be customized by the user. Furthermore, its functionality can be expanded using plugins written in Python or C++ (QGIS, 2021). In this work, we used version 2.18.25 to create shapefiles, distribution maps of *Aedes aegypti* mosquito breeding sites, and climatic variables.

3.4.2 Weka

WEKA (Waikato Environment for Knowledge Analysis) is an open-source software developed by a group of researchers at the University of Waikato, New Zealand, in 1993. Weka supports different types of files (such as ".arff" and ".csv" formats) and presents a collection of machine learning algorithms and data pre-processing tools. It is a tool developed in Java and, since it is open-source software, it is under the terms of the GNU (General Public License). In addition, Weka is a cross-platform application that can be run on operating systems such as Windows, Linux, and Macintosh (Witten and Frank, 2005). In this work, we use Weka, version 3.8, to generate the training databases, evaluate the regressors, and validate the generated models.

3.5 Forecasting datasets

The forecasting datasets were assembled using the data described in Section 3.3. Then, firstly, we stored the information regarding the *Aedes aegypti* breeding sites in comma-separated files (.csv). This information was stored along with the centroid coordinates of the corresponding neighborhood, and the data were organized considering a two-month period from 2013 to 2016. As mentioned in Section 3.3.1, the Health Districts are divided into strata, and each stratum contains neighborhoods that have similar sociodemographic characteristics. However, the number of strata for each health district is not fixed and may vary according to the needs of health authorities for epidemiological control. Therefore, in this work, to store the breeding information for each neighborhood, we consider that the neighborhoods belonging to the same stratum contained the same amount of breeding sites.

To estimate the temperature and wind speed in the other neighborhoods of the city, we used a Gaussian distribution. The samples' standard deviations (σ) were calculated using Equation 1, considering the maximum values (x_{max}) and the monthly mean (μ) of the temperature, and wind speed samples, separately. Since the rainfall is monitored by more than one station, the maximum value was the maximum value registered among the stations. On the other hand, the monthly average considered was the average among the monitoring stations.

$$\sigma = \frac{x_{max} - \mu}{4} \quad (1)$$

Therefore, we stored the information regarding the meteorological data for each neighborhood - along with its coordinates - in .csv files. From the .csv files, we generated the vector point files (.shp format) to geolocate the total breeding sites and meteorological data to their respective neighborhood. So, for the breeding sites, we generate bimonthly shapefiles from 2013 to 2016. However, for the climatic factors, we generate monthly shapefiles for the same time interval. In this context, to estimate the breeding sites distribution, temperature, rainfall, and wind speeds throughout the territory under study, we generated spatial distribution maps of each variable. The method chosen for this process was the IDW interpolation, using its IDW interpolation tool from the QGIS software.

The forecasting sets were elaborated using the distribution maps generated in the previous step. The bimonthly model was chosen due to the methodology adopted by the Unified Health System (SUS) for planning policies to combat arboviruses, which consider a bimonthly cycle. Then, the prediction vectors were assembled by performing a simultaneous scan, pixel-by-pixel, in the distribution maps, and concatenating the following information: (1.) latitude; (2.) longitude; (3.) two-month period distribution of breeding sites. (4.) distribution of temperature, rainfall, and wind speed for the months of the respective bimester. Therefore, 18 prediction sets were assembled with 15,446 instances each. All forecast sets contain 44 attributes wherein the set output is the pixel value of the breeding site's distribution map in the corresponding coordinate.

3.6 Regressors

3.6.1 Linear Regression

Proposed by Sir Francis Galton, in the nineteenth century, regression analysis is one of the most used statistical concepts in machine learning (Senn, 2011). Galton initially proposed a mathematical model to study the relationship between the heights of children and their parents (Senn, 2011). In this model, he analyzed the association between the dependent variable and the independent variable. This model was later known as the simple linear regression model. In other words, this technique originates from the linear correlation, from the analysis of the existence of a relationship between two variables and assesses the range of variation between these variables (Montgomery, 2013). Thus, this technique uses the data points to find the best fit line (from the relationship of a scatter diagram) to model the relationship. The linear regression model is represented by Equation 2. Where $x_1, x_2, \ldots, x_n$ represents the input attributes of the model and $w_{k,1}, w_{k,2}, \ldots, w_{k,n}$ represents the weights related to the input attributes.

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \quad (2)$$

Thus, the idea of the linear regression algorithm is to find the optimal weights that satisfy Equation 2, so that the error obtained between the input value and the output value is minimal. Optimal weights can be

calculated using the least-squares method. The least-squares method approach is to minimize the sum of squares of the differences between the estimated value and the observed data (Equation 3) (Witten and Frank, 2005).

$$S = \sum_{i=1}^{n} \left[y^{(i)} - \sum_{j=0}^{k} w_{ij} x_j^{(i)} \right]^2 \quad (3)$$

3.6.2 Support Vector Regression

Proposed by Vladmir Vapnik in 1979, the support vector machine is a supervised learning algorithm that has been widely used to solve different classification and regression problems (de Freitas Barbosa et al., 2021; Barbosa et al., 2021; Guo et al., 2017, da Silva et al., 2021). The idea of SVR is to find an optimal hyperplane, out of a range of hyperplanes, in which the data can be represented as a linearly separable problem. The optimal hyperplane is the one for which the margin of separation between the support vectors is maximum (Haykin, 2001). When the problem is linearly separable, the mathematical expression which defines the optimal hyperplane is represented by Equation 4. Where $w = (w_1, w_2, \ldots, w_n)^T$ represents the weights vector, $x = (x_1, x_2, \ldots, x_n)$ is the attributes vector, and b represents the bias (Witten & Frank, 2005).

$$y = w^T x + b \quad (4)$$

In cases where the problem is not linearly separable, the data is mapped to the optimal hyperplane through a mathematical function, known as Kernel. Thus, the output of the SVM is given by Equation 5. The kernel can be a polynomial, Gaussian, sigmoidal, or other mathematical function. Hence, changing the kernel type changes the learning machine type since it alters how the dot product between the attribute vector and the weight vector is generated (Drucker et al., 1997; Witten & Frank, 2005; Smola & Schölkopf, 2004).

$$y = K(w, x) \quad (5)$$

This algorithm is robust in large dimensions and has some advantages. It contains only a global minimum since its objective function is convex. Additionally, it has good generalizability, which avoids the possibility of overfitting. Furthermore, it is sensitive to the choice of parameters and is ideal for two-class problems. However, the disadvantage of this technique is the computational cost associated (Haykin, 2001).

3.6.3 Multilayer Perceptron

Multilayer perceptrons (MLP) are a type of artificial neural network, which are machine learning algorithms inspired by the behavior of the human brain (Siriyasatien, 2018; Haykin, 2001). The perceptron was proposed by Frank Rosenblatt in 1958, and it consists of the following elements: connectors or synapses, adders, and an activation function. The synapses (or connectors) are connected to a neuron k, which receive the input signal x_i which is multiplied by the synaptic weight w_{ij}. The adder adds the signals weighted by the synaptic

weights and the activation function limits the perceptron output value (Rosenblatt, 1958; Haykin, 2001). Mathematically, the artificial neuron is represented by Equations 6 and 7 (Haykin, 2001), where $x_1, x_2, \ldots, x_n$ represent the input signals; $w_{k,1}, w_{k,2}, \ldots, w_{k,n}$ represent the synaptic weights of the input signal x_i referring to the kth neuron; b_k refers to bias, and φ is the function of activating the neuron.

$$u_k = \sum_{j=1}^{n} w_{kj} x_i \quad (6)$$

$$y_k = \varphi(u_k + b_k) \quad (7)$$

In MLP, the weights are adjusted through the learning rule which consists of minimizing the error between the neuron output value and the expected value. To achieve this feat, the backpropagation algorithm is generally used to adjust these weights. The backpropagation algorithm is based on the descending gradient and contains two steps: the propagation phase and backpropagation. In the propagation phase, an output value is obtained for a given input pattern. In the backpropagation phase, an error is calculated using the desired value, and the output value from the previous phase. In this way, the synaptic weights are iteratively updated until the error between the output value and the expected value is minimal (Haykin, 2001). Neural Networks have been widely used in several biomedical applications (Laureano-Rosario et al., 2018; Baquero et al., 2018; Akil & Ahmad, 2016; da Silva et al., 2021).

3.6.4 Random Forest

Random forest algorithms are based on decision tree committees organized in "bagging". The decision trees have three nodes - wherein the most common are root, leaves, parent, and child -, which are structures that store information. The root is the starting point and has the highest hierarchical level. The nodes can connect with other nodes, establishing a parent-child relationship, where a parent node generates a child node. The leaves, in turn, are terminal nodes that do not generate child nodes and represent a decision. The Random Forest, therefore, is a set of decision trees that hierarchically divide the data. Then, each tree votes for a class of the problem in question. The class most voted for by the algorithm is chosen as the regressor prediction (Breiman, 2001).

3.7 Regressors evaluation

The experiments to generate and validate the models were carried out in the Weka environment (version 3.8.3). The training set was assembled from the set with 15,446 instances through Weka's "resample" tool. This tool produces a new database with random values for the instances, but with the same statistical characteristics as the original database. The number of instances of the new database must be specified by the user. Considering this, the training set was generated by applying the "resample" tool to each prediction set, with the number of instances equivalent to 30% of the original set. The sets with 15,446 were used to validate the models generated by the regressors.

We evaluated four types of regression algorithms (Table 1): linear regression, random forest, support vector regression (SVR), Multilayer Perceptron (MLP). For the random forest, we evaluated configurations from 10-40 trees. Regarding the SVR algorithm, we investigated the configurations with polynomial kernels (degrees 1, 2, and 3), and RBF kernels. The parameter C was set to 0.1 for all SVR settings described. Finally, the MLP algorithm, with a single hidden layer, was evaluated using 10, 20, 30, and 40 neurons in the hidden layer. Each regressor was evaluated 30 times with 10-fold cross-validation.

Table 1: Configuration of the regressors used to generate prediction models for the *Aedes aegypti* mosquito's breeding sites distribution maps in Recife.

Regressor	Parameters
Linear Regressor	-
SVR	Linear kernel, C=0.1 Polynomial kernel, degree 1, C=0.1 Polynomial kernel, degree 2, C=0.1 Polynomial kernel, degree 3, C=0.1 RBF kernel
Random Forest	Number of trees: 10, 20, 30, 40.
MLP	Number of neurons in the hidden layer: 10, 20, 30

3.8 Metrics

We evaluated the models generated by the regressors using two metrics: the Pearson correlation coefficient and the Root Relative Squared Error (RRSE %). Pearson correlation coefficient is a measure that assesses the statistical relationship between two continuous variables. In this case, the statistical relationship between the actual values and the predicted values by the model. The value of this coefficient varies between -1 and 1. When the correlation coefficient is close to 1, it indicates that there is a strong correlation between the variables. When the value of the correlation coefficient is close to -1, it indicates that there is a strong negative correlation between the evaluated variables. However, when the coefficient approaches zero, it means that the variables have a weak correlation or no correlation (Witten & Frank, 2005). In this work, we use the correlation coefficient as an overall metric for evaluating prediction models. However, it is possible to obtain high values for this metric and at the same time obtain high values for local errors. For this reason, to avoid a superficial analysis of the method, it is important to use another metric to evaluate the model's performance. In this sense, we chose RRSE (%) as the second evaluation metric. Equation 8 shows the expression to calculate the relative quadratic error, where p_i represents the predicted value and a_i is the real value, for $i = 1, 2, \ldots, n$. Also, we consider a high correlation coefficient a coefficient above 0.9, and a low RRSE% a value below 5%. Therefore, we consider as good regressors those with a high correlation coefficient and a low RRSE%.

$$\text{RRSE (\%)} = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n(a_1^2 + \ldots + a_n^2)}} \quad (8)$$

4. Results

The models were quantitatively evaluated according to the metrics described in Section 3.7. The correlation coefficient was used as a global assessment metric, whereas the RRSE (%) was used as a local assessment metric. In this work, we consider high correlation coefficient values above 0.9, and low RRSE (%) values below 5%. In addition to the metrics mentioned, we also computed the training time of the models to assess the associated computational cost.

Table 2 shows the findings of the experiments with the regressors presented in Table 1. It shows the average and standard deviation of the evaluation metrics for each regressor investigated. Thus, for the models generated by the linear regression algorithm, the linear correlation coefficient obtained, on average, was 0.8488 (with a standard deviation of 0.05). On the other hand, the RRSE% associated with the regressor was considerably high, reaching an average above 50%. Although this method obtained a high correlation coefficient, its error was higher than the value established for this work. The findings in Table 2 also show the results obtained with the random forest experiments. For this regressor, the values of the correlation coefficient were 0.9687, 0.9745, 0.9764, and 0.9773 for the configurations with 10, 20, 30, and 40 trees, respectively. The RRSE (%) values did not show a great variation among the evaluated configurations. The configuration with the lowest value was the configuration with 40 trees, in which the RRSE (%) reached a value of 22.53% and a standard deviation of 4.91%.

The experiments with support vector regression showed only two configurations presented with a coefficient above 0.9. The best performances were obtained with a polynomial kernel of degrees 2 and 3. For these two configurations, the correlation coefficients achieved were 0.9488, and 0.9875, respectively. As for the RRSE (%), the configuration with grade 3 polynomial kernel presented the lowest value (14.60%) among the evaluated configurations. On the other hand, the configuration with the RBF kernel showed the worst performance, reaching RRSE (%) above 60%. Finally, the MLP experiments showed errors of 23.29%, 22.63%, and 20.61% for the configurations with 10, 20, and 30 neurons in the hidden layer, respectively. However, all correlation coefficients were above 0.9.

Table 2: Results of the performance of linear regression, support vector regression (SVR), and multilayer perceptron (MLP). For each regressor, we computed the following evaluation metrics: correlation coefficient (R), root relative squared error, RSSE (%), and the training time (s).

		RRSE (%)		**R**		**Training time (s)**	
Regression method	**Configuration**	**Average**	**standard deviation**	**Average**	**standard deviation**	**Average**	**standard deviation**
Linear	-	51.09	13.86	0.8488	0.05	0.09	0.03

Regression							
Random Forest	10 trees	25.47	5.07	0.9687	0.01	0.31	0.06
	20 trees	23.54	4.93	0.9745	0.01	0.62	0.11
	30 trees	22.87	4.91	0.9764	0.01	0.93	0.16
	40 trees	22.53	4.91	0.9773	0.01	1.27	0.20
Support Vector Regressor	polynomial kernel, p=1	53.34	14.50	0.8343	0.06	18.72	6.23
	polynomial kernel, p=2	30.30	9.98	0.9488	0.03	191.17	65.32
	polynomial kernel, p=3	14.60	6.41	0.9875	0.01	3121.20	1083.96
	RBF kernel	63.84	7.44	0.8002	0.06	19.49	4.48
MLP, one hidden layer	10 neurons	23.29	9.44	0.9750	0.02	62.68	7.95
	20 neurons	22.63	8.44	0.9770	0.01	74.50	9.50
	30 neurons	20.61	7.61	0.9809	0.01	86.76	10.49

Source: The authors.

Table 3 shows the results obtained by validating the models generated by the best and worst regressors, considering the RRSE (%) and R metrics. We validated the models generated for each of the two-month periods from 2014 to 2016 using the test database with 15,446 instances. According to the results in Table 3, the correlation coefficients obtained by validating each of the 6 bi-monthly periods from 2014 to 2016 were above 0.9 and in line with the findings in Table 2. For the SVR with RBF kernel, most values were below the limit established by this work. However, for the last two months of 2016, the correlation coefficient reached a value of 0.9424.

Regarding the RRSE (%), the models generated by the SVR with the RBF kernel reached values above 55%, except for the sixth quarter of 2016, where the value obtained was 41.47%. These findings are consistent with what was shown in Table 2. For the models generated with the SVR with grade 3 polynomial kernel, the RRSE (%) ranged from 1.04% to 25.19%.

Table 3: Comparison of the results obtained for linear regression and SVM 3-degree polynomial kernel. For each regressor, we computed the following evaluation metrics: correlation coefficient (R), root relative squared error, RSSE (%), and the training time (s).

		SVR, kernel = RBF		**SVR – 3-degree poly-kernel**	
Year	**Bimester**	**R**	**RRSE%**	**R**	**RRSE%**
	1	0.7671	68.26	0.9953	9.70

	2	0.7664	68.23	0.9968	8.03
	3	0.7649	68.35	0.9964	8.53
2014	4	0.8296	58.73	0.9987	5.05
	5	0.8142	61.46	0.9978	6.69
	6	0.7935	64.33	0.9971	7.61
	1	0.8278	60.86	0.9898	14.25
	2	0.8131	61.82	0.9917	12.99
	3	0.8492	56.54	0.9903	14.06
2015	4	0.7781	66.44	0.9781	21.40
	5	0.8044	61.69	0.9944	10.55
	6	0.7739	67.70	0.9768	21.62
	1	0.7675	66.32	0.9684	25.19
	2	0.8271	59.05	0.9907	13.64
	3	0.7522	74.75	0.9903	14.13
2016	4	0.8352	58.23	0.9962	8.76
	5	0.8210	63.53	0.9901	14.22
	6	0.9424	41.47	0.9999	1.04

Source: The authors, 2020.

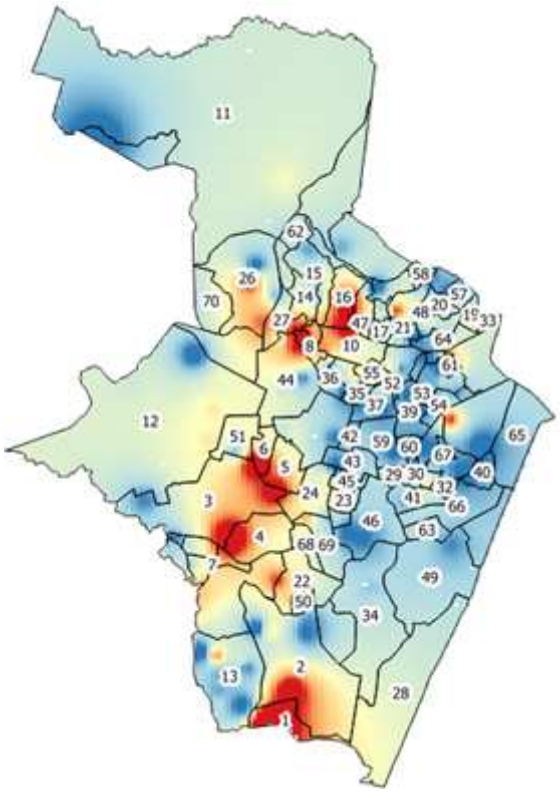
5. Discussion

In this study, we incorporated spatiotemporal predictive models to generate distribution maps for *Aedes aegypti* breeding sites in Recife, Brazil. We used machine learning algorithms to generate the models which, in turn, made evaluations for each bimester from 2014 to 2016. Moreover, the bimonthly model was chosen due to the methodology adopted by the Unified Health System to combat *Aedes aegypti*. In general, the algorithm which presented the best performance was the 3-degree polynomial-kernel SVR, despite being the algorithm with the longest training time among the evaluated regressors. For this regressor, the correlation coefficient achieved was 0.9875, and the RRSE (%) reached 14.60%. The linear regression and RBF-kernel SVR algorithms, in turn, presented an undermost performance considering both correlation coefficient, and RRSE (%). For both regressors configurations, the errors achieved were superior to 50%. Furthermore, the Random Forest and Multilayer Perceptron algorithms presented similar performances regarding the associated error (around 20%). However, considering the model's training time, the Random Forest had a superior performance than the MLP.

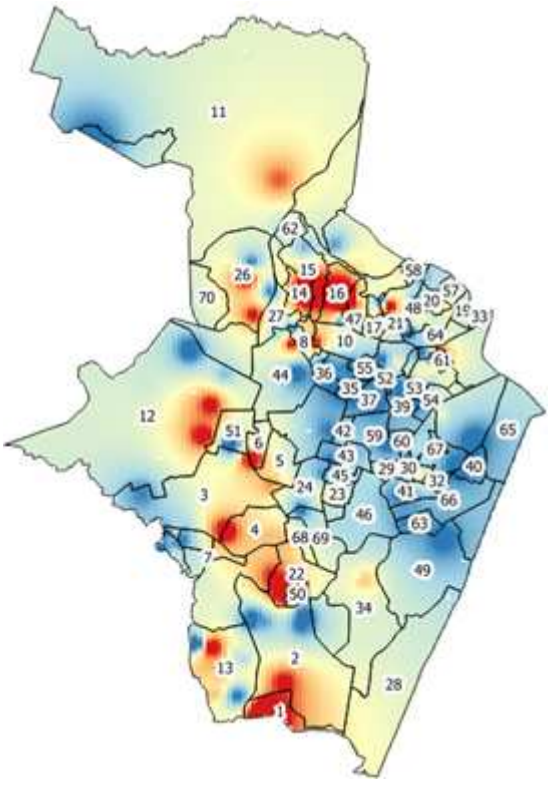
The SVR predictions (kernel, p=3) for the first two-month period of 2014 are shown in Figure 6a. According to the map, the neighborhoods which presented the highest concentrations of mosquito breeding sites in the southern region were: Jordão (region 1) and Ibura (region 2). In the city's southwest region, the most affected neighborhoods were Curado (region 3), Jardim São Paulo (region 4). In addition, Torrões (region 5), Engenho do Meio (region 6), and part of the Tejipió neighborhood (region 7) were also highly affected. In the city's northern region, according to the SVR model, the neighborhoods of Monteiro (region 8), Alto do Mandu (region 9) and Casa Amarela (region 10), stand out. From the January/February of 2014 to March/April of the same year (Figure 6b), there was a slight increase in the concentration of breeding sites in most parts of the city. Especially for the neighborhoods of Guabiraba (region 11), Várzea (region 12), Cohab (region 13), Macaxeira (region 14), Nova Descoberta (region 15), Vasco da Gama (region 16) and Alto José do Pinho (region 17). From the second to the third bimester of 2014, there was a massive increase in the breeding site concentration in most parts of Recife. Nonetheless, in May/June and July/August of 2014 (Figure 6c and Figure 6d, respectively), the distribution maps showed similar behavior. The predictions indicated that for a significant part of the city, the breeding sites concentration varied from moderate to very high. In both bimesters, there was a notable increase of breeding sites in Recife's northeast and western region. From the northeast region, the following neighborhoods stand out: Água Fria (region 48), Alto José do Pinho (region 17), Mangabeira (region 18), Campina do Barreto (region 19). Moreover, Fundão and Bomba do Hemetério (regions 20 and 21, respectively) also presented high concentrations of the mosquito breeding sites. In the city's western region, the most affected neighborhoods were Várzea (region 12), Areias (region 22), Engenho do Meio (region 6), and part of the neighborhoods of Mustardinha, San Martin, and Mangueira (regions 23, 24, and 25, respectively).

From July/August (Figure 6d) to September/October of 2014, there was a considerable decrease in the breeding sites in most areas of the city. Nevertheless, for the remaining two-month periods of 2014 (Figures 6e and 6f, respectively) the predictions showed that the *Aedes aegypti*'s breeding sites were mainly concentrated in the northern and southern neighborhoods. In the northern area, we highlight Dois Irmãos (region 26), Macaxeira (region 14), Alto José do Pinho (region 17), Mangabeira (region 18). But Apipucos, Monteiro (region 8), Vasco da Gama (region 16) and Alto do Mandu (region 9) also presented a high concentration of breeding sites. In the southernmost areas, the neighborhoods which stand out are Cohab (region 13), Ibura (region 2), and Jordão (region 1).

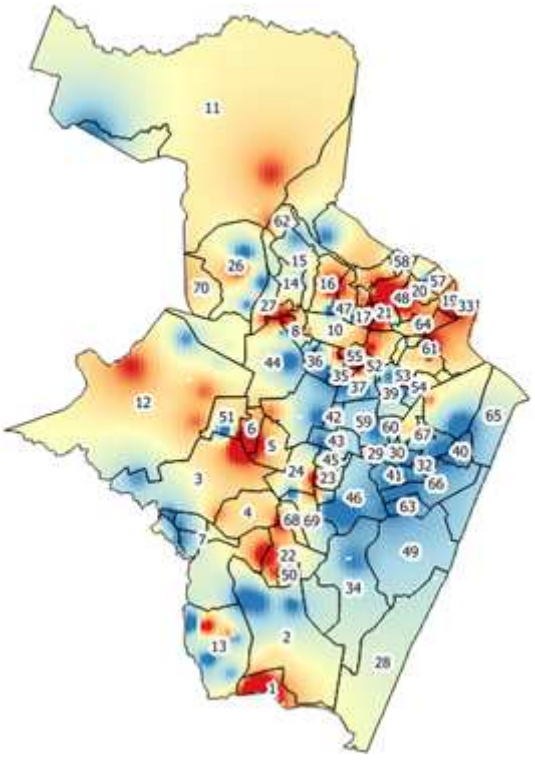
Figure 6: Prediction maps of *Aedes aegypti*'s breeding sites in the City of Recife for Jan/Feb (a), Mar/Apr (b), May/Jun (c), Jul/Aug (d), Sep/Oct, (e) and Nov/Dez (f) of 2014. The results were generated using a 3-degree polynomial kernel support vector regressor. To generate the models, we used the training datasets with 3861 instances and tested with the datasets with 15,446 instances. The warmer regions (red) of the map indicate a high concentration of breeding sites, whereas the cooler regions (blue) indicate low concentrations of *Aedes aegypti*'s breeding sites.



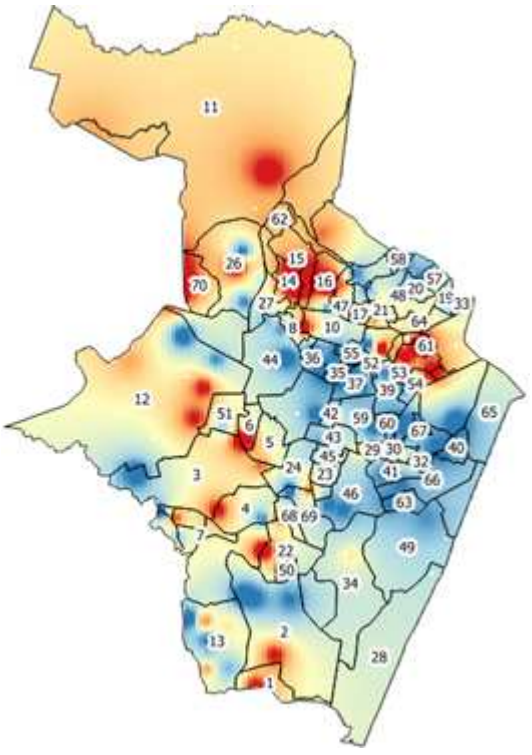
(a)



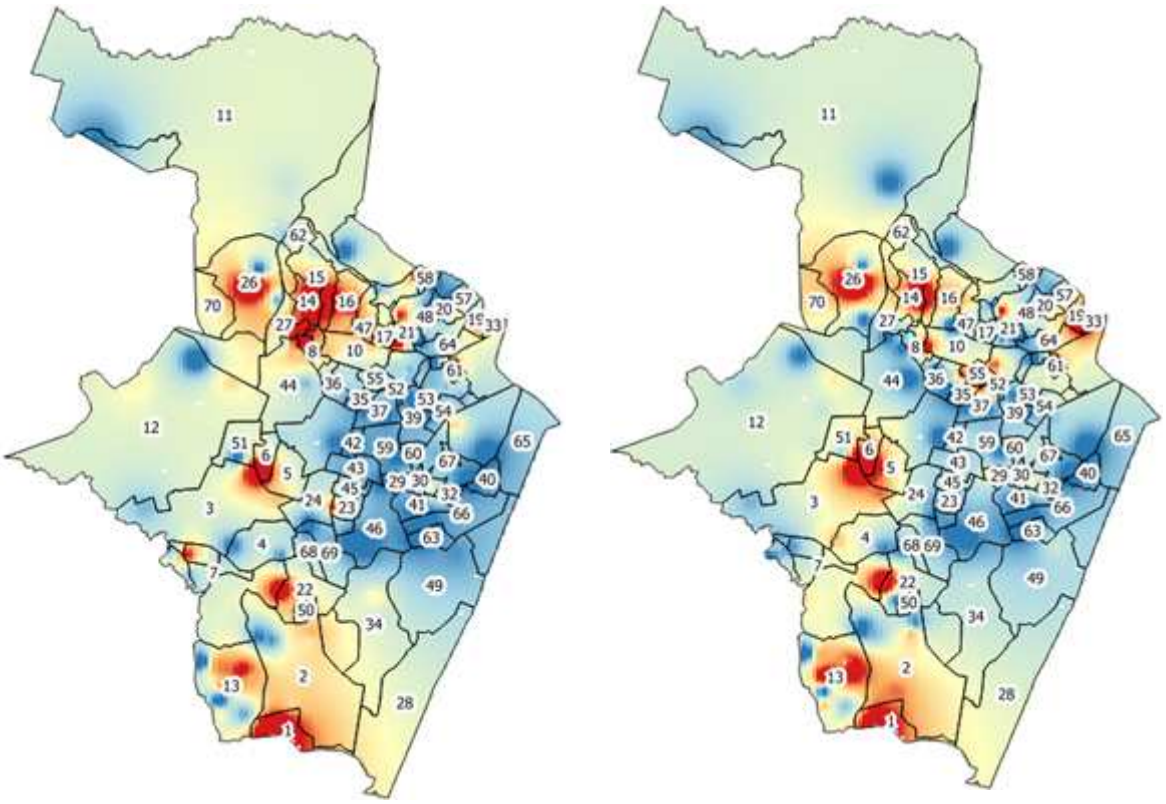
(b)



(c)



(d)



(e) (f)

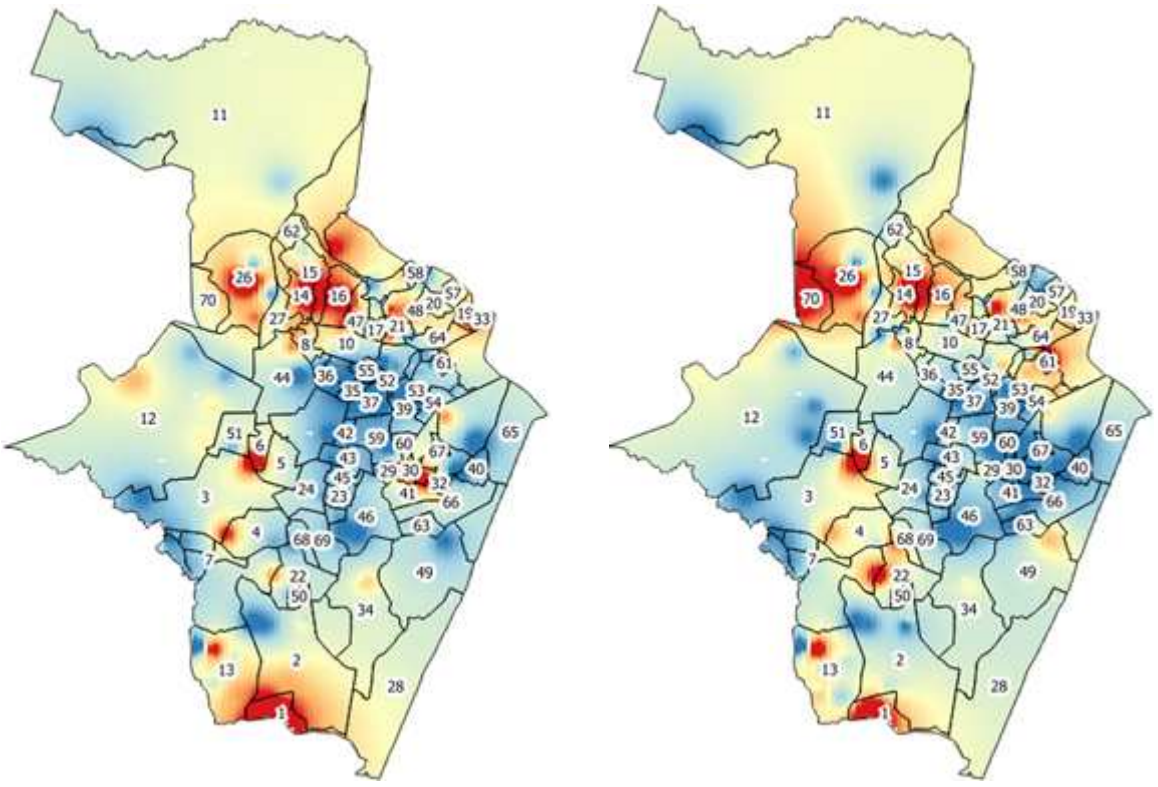
Figure 7a shows the predicted distribution map of the mosquitoes' breeding sites in the first bimester of 2015. According to the SVR (kernel p=3) algorithm, the breeding sites concentration decreased considerably in most of the western part of the city. The same behavior was detected in some neighborhoods in the southernmost region, such as (1.) Cohab - region 13; (2.) Boa Viagem - region 28; and (3.) Ibura - region 2. The opposite behavior occurred in the neighborhood of Jordão (region 1), which remained with a high concentration of *Aedes aegypti* breeding sites. The breeding sites concentration also increased significantly in some neighborhoods in the city's northeast, for example, Campina do Barreto (region 19), and Peixinhos (region 33). In the following bimester (Figure 7b), the SVR (kernel p=3) found a similar behavior to the previous bimester. In the northernmost part of the city, we stand out the neighborhood of Dois Irmãos (region 26), and Sítio dos Pintos (region 70). In the southernmost region of Recife, the algorithm identified an intensification in the breeding sites concentration in Cohab, and Ibura (regions 13 and 2, respectively).

Figure 7c shows the predicted distribution map of the mosquitoes' breeding sites in May-June of 2015. According to the map, the breeding sites concentration, overall, reduced significantly in most parts of the city. However, in the southernmost region of Recife, the predictions show that the breeding sites concentration greatly enhanced in two neighborhoods: Cohab (region 13), and Ibura (region 2). Another highlight is the Várzea neighborhood (region 12), which also had a high concentration of breeding sites when compared to the prediction maps for the first and second bimesters. In July-August of 2015 (Figure 7d), in the northern area, we point out the neighborhoods of (1.) Nova Descoberta; (2.) Vasco da Gama; (3.) Macaxeira; and (4.) Alto José Bonifácio had

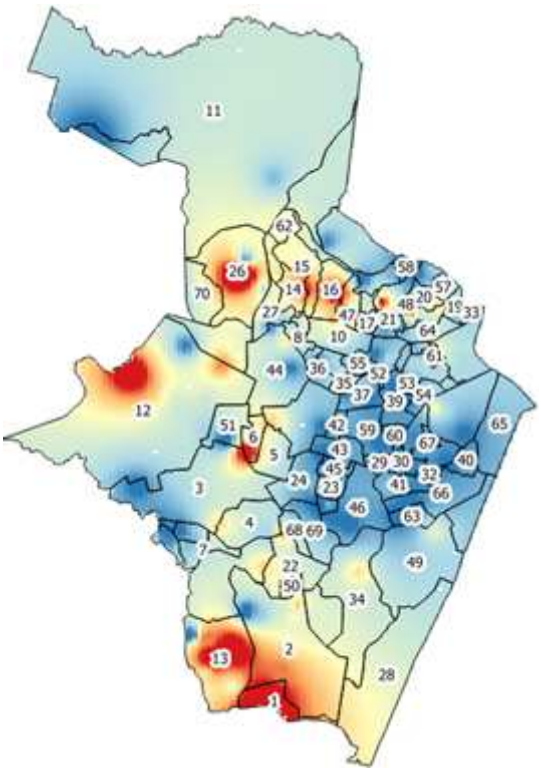
high concentrations of breeding sites. Similarly, in the southernmost neighborhoods, Jordão (region 1), and Imbiribeira (region 13) stand out.

From July/August (Figure 7d) to September/October (Figure 7e), the concentration of breeding sites increased again in several districts of the city, according to the predictions. In some neighborhoods such as Pina (region 49), Cohab (region 13), and Dois Irmãos (region 26), this increase was quite intense. However, in other neighborhoods of the city, the decrease was quite notable, although these neighborhoods continue with high/very high or moderate concentrations of breeding sites. In the city's northern and northeastern region, the neighborhoods of Nova Descoberta (region 15), Vasco da Gama (region 16) and Campina do Barreto (region 19) stand out. In the southern region, predictions showed a slight decrease in the neighborhood of Imbiribeira (region 34). In the last two months of 2015, in general, the concentration of breeding sites remained constant. Yet, there was an intensification in the neighborhoods of Várzea (region 12), Afogados (region 46), 69. In the city's southernmost region, there was a notable decrease in the concentration of breeding sites mainly in Imbiribeira (region and 34) and Pina (region 49).

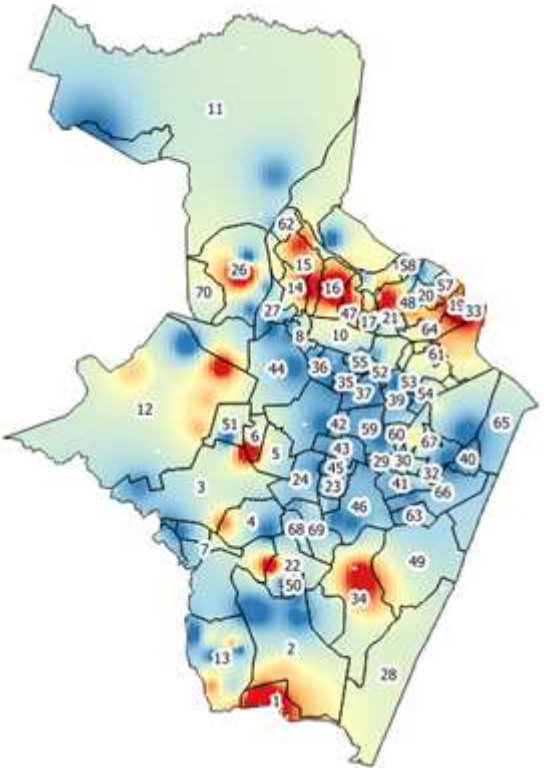
Figure 7: Prediction maps of *Aedes aegypti*'s breeding sites in the City of Recife for Jan/Feb (a), Mar/Apr (b), May/Jun (c), Jul/Aug (d), Sep/Out (e) and Nov/Dez (f) of 2015. The results were generated using a 3-degree polynomial kernel support vector regressor. To generate the models we used the training datasets with 3861 instances and tested with the datasets with 15,446 instances. The warmer regions (red) of the map indicate a high concentration of breeding sites, whereas the cooler regions (blue) indicate low concentrations of *Aedes aegypti*'s breeding sites.



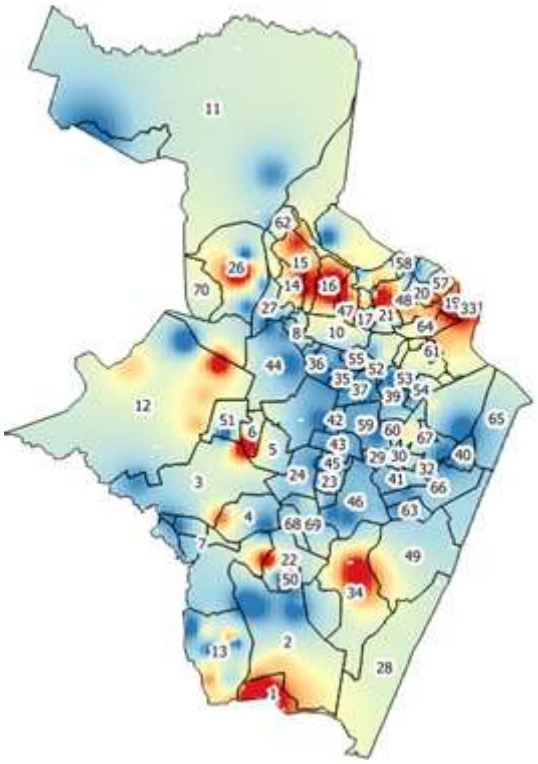
(a) (b)



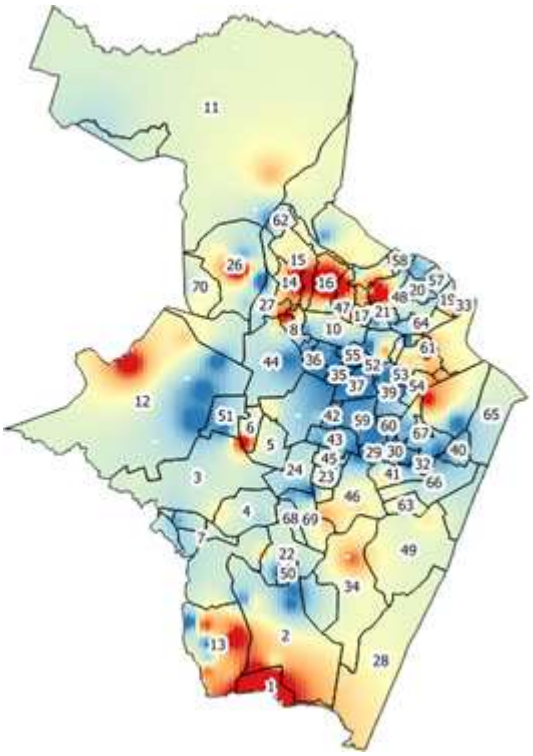
(c)



(d)



(e)



(f)

The year 2016 started with high rates of mosquito presence in various regions of the city according to SVR predictions (Figure 8a). In the city's westernmost region, there is an intense increase in the breeding sites - compared to the previous two months. We observe this increase on the border between region Curado (region 3) and Jardim São Paulo (region 4), and in Várzea (region 12). The neighborhoods of Pina (region 34) and Imbiribeira (region 49) also obtained a considerable increase in the presence of the mosquito according to the predictions. In the period March-April 2016, the algorithm detected an enhancement in the city's northernmost neighborhoods. The most intense increases were in the neighborhoods of Dois Irmãos (region 26), Vasco da Gama (region 16), and on the border between Guabiraba (region 11) and region 62. In the west part of the city, the prediction detected a relative decrease in the presence of mosquitoes in some neighborhoods, but mainly in the Várzea (region 12) and on the border between Cidade Universitária, Engenho do Meio, and 5 (regions 51, 6 and 5, respectively). In the southern region of the city, there was also a decrease in the breeding sites concentration. The decrease was observed mainly in the neighborhoods of Cohab (region 13) and Pina (region 49).

From March-April (Figure 8b) to May-June (Figure 8c) of 2016, there was a drastic decrease in the presence of mosquito breeding sites in the city, in general. However, the concentrations of breeding sites increased considerably in some neighborhoods. For example, in Várzea (region 12) and Nova Descoberta (region 34), there was an intensification in the presence of *Aedes aegypti*. In the following two months (Figure 8d), according to the predictions, there was an explosion in the presence of the mosquito, mainly in the north and northeast regions of the city. Most neighborhoods in these areas had either high or very high concentrations of *Aedes aegypti* mosquito breeding sites. The maps of the last two-month period of 2016 showed similar behavior, as shown in Figure 8e and Figure 8f. For both predictions, the northern region was the most affected by the intense presence of the mosquito, especially in the neighborhoods of Vasco da Gama (region 16), Nova Descoberta (region 15), and Dois Irmãos (region 26). In the city's southern region, only Jordão (region 1) and part of Pina (region 49) presented a high concentration of breeding sites. In the western region of the city, there were also high levels of mosquito presence in specific points of some neighborhoods. They are: Curado, Jardim São Paulo and Engenho do Meio – regions 3,4 and 6, respectively.

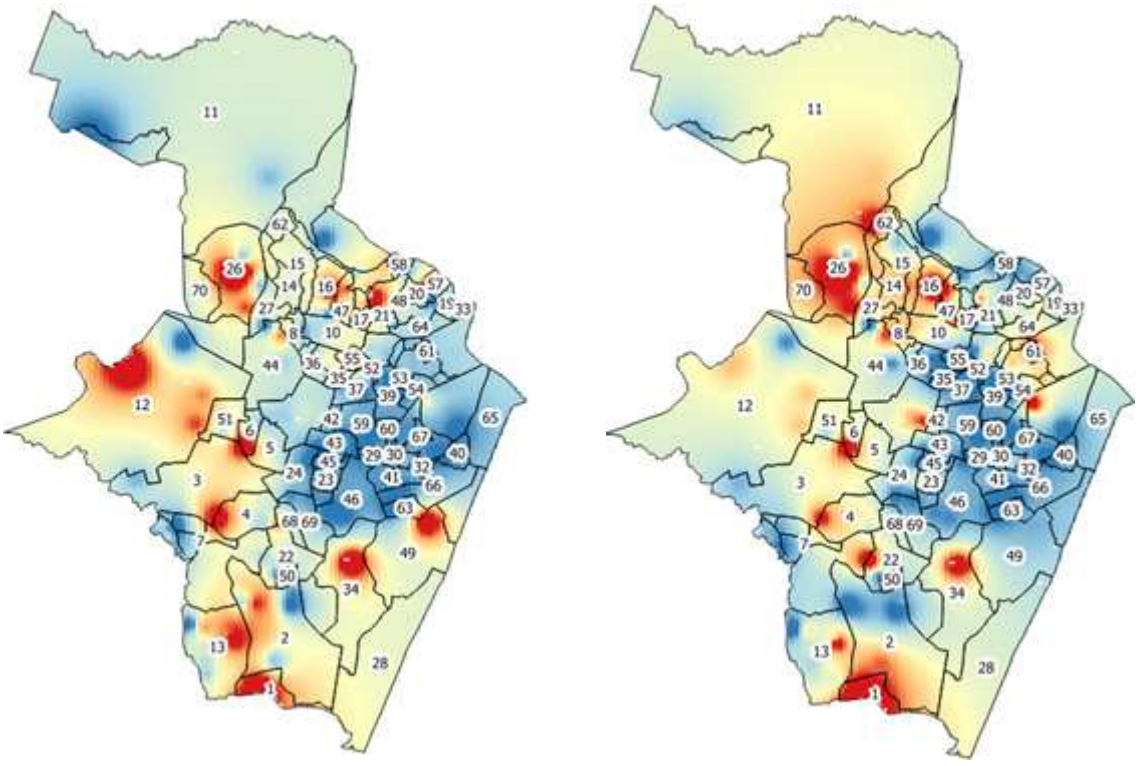
It is important to highlight that the neighborhood of Jordão (region 1), in most cases, presented a very high concentration of mosquito breeding sites throughout 2014-2016. According to the predictions, in this neighborhood, the concentration was low only in July-August of 2016. Another area of the city that showed a similar behavior was the area comprising the neighborhoods of Engenho do Meio and (region 5). In these areas, the mosquito presence were usually intense throughout the evaluated two-month period.

We made an additional observation of areas which appeared more frequently with high breeding sites' concentrations are neighborhoods with lower incomes. Except for some neighborhoods such as Recife neighborhood (region 65). Although this neighborhood has an average monthly income below two times the minimum wage (Figure 9), it is not a residential neighborhood. The fact that neighborhoods with lower incomes have a higher concentration of mosquito presence may indicate a strong relationship with infrastructure problems. That is problematic given the fact that these areas are known to lack decent sanitation standards, as well as have irregular water distribution in these regions. In this context, local inhabitants tend to store water in containers that increase the risk of infestation and proliferation of *Aedes aegypti* species.

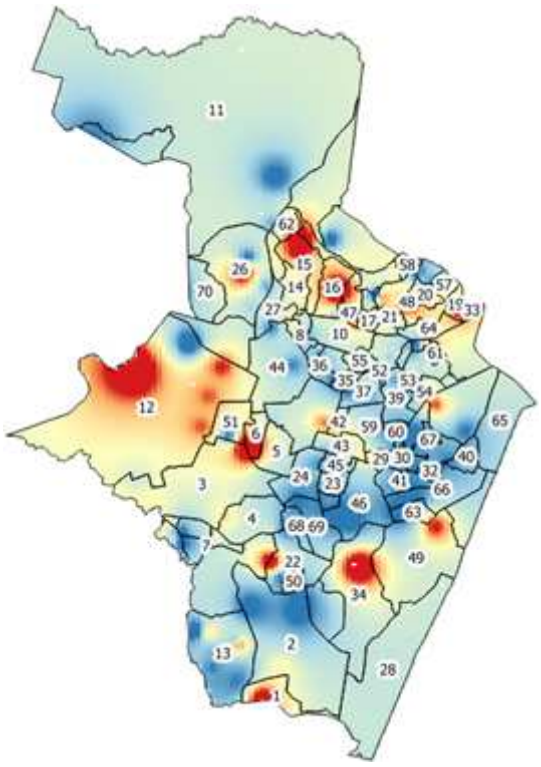
Finally, we must highlight some points. The first is that we estimate the same number of breeding sites per neighborhood within the same stratum of a health district. The strata represent areas with similar characteristics, according to criteria established by teams of epidemiologists from the City of Recife. However, as the definition of strata varies according to the characteristics of epidemiological interest, these strata can change, and they can bring together different neighborhoods within the same health district. This condition may explain the fact that the Monteiro neighborhood has a higher concentration of breeding sites, an exception concerning the behavior of middle and high-income neighborhoods.

Another point is the fact that there are only three weather stations in Recife, we had to build an estimator to compensate for this data paucity, and this was based on a Gaussian distribution so as to estimate the temperature and wind speed distribution for each neighborhood. A solution to this problem may be to build a network of low-cost devices for measuring temperature, rainfall, and wind speeds from an Internet of Things (IoT) perspective. The INMET data is also collected from one station, and a Gaussian distribution is also used when building the shapefile. As it is not information from each neighborhood, there may be an associated error which the authors concede and acknowledge as a limitation.

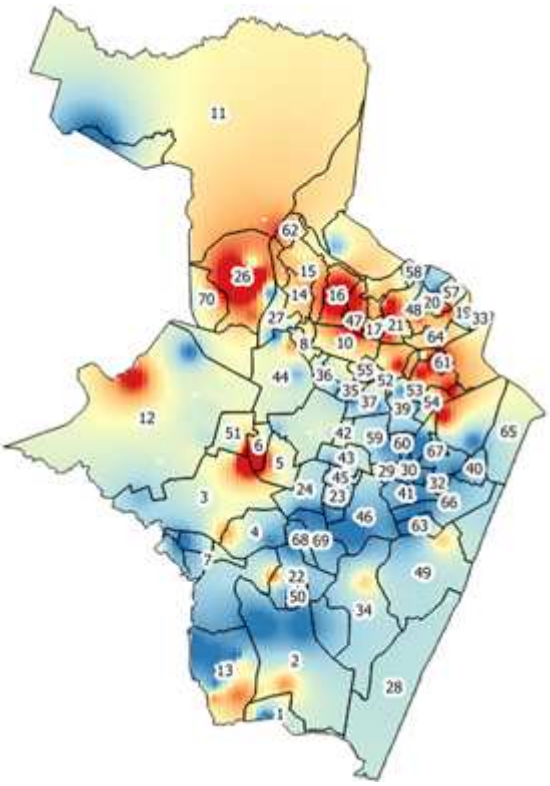
Figure 8: Prediction maps of *Aedes aegypti*'s breeding sites in the City of Recife for Jan/Feb (a), Mar/Apr (b), May/Jun (c), Jul/Aug (d), Sep/Out (e) and Nov/Dez (f) of 2016. The results were generated using a 3-degree polynomial kernel support vector regressor. To generate the models we used the training datasets with 3861 instances and tested with the datasets with 15,446 instances. The warmer regions (red) of the map indicate a high concentration of breeding sites, whereas the cooler regions (blue) indicate low concentrations of *Aedes aegypti*'s breeding sites.



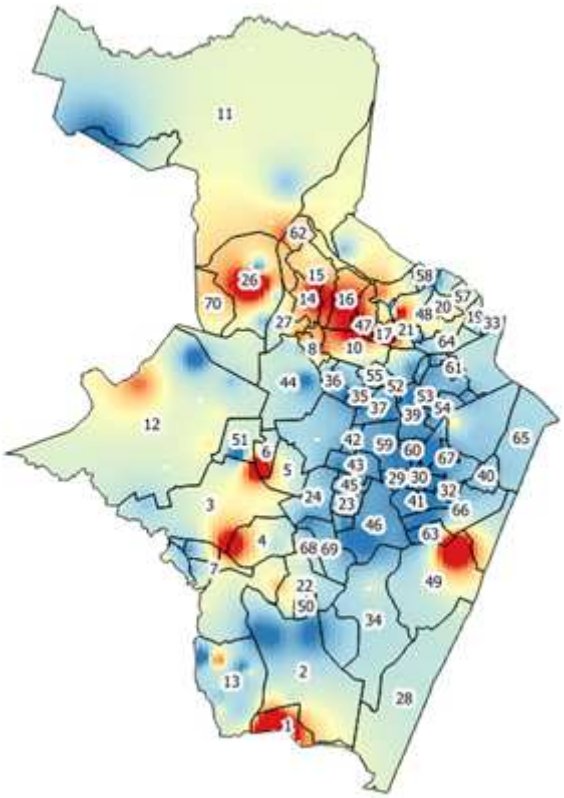
(a)



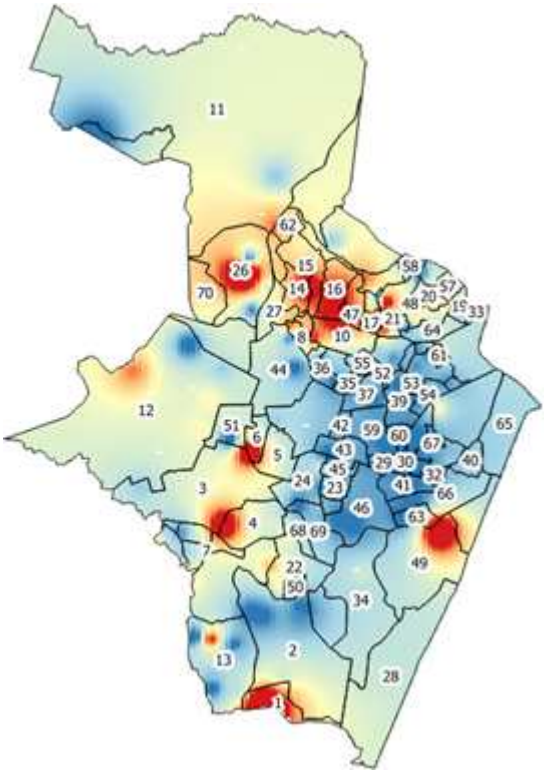
(b)



(c)



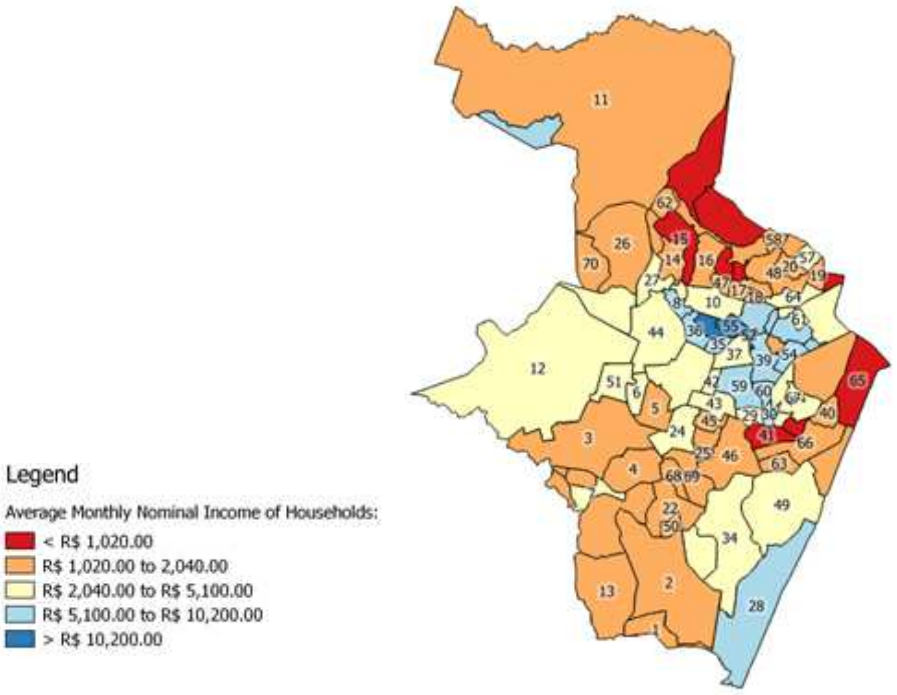
(d)



(e)

(f)

Figure 9: Distribution of the average monthly salary income in Recife's neighborhoods according to the 2010 census.



6. Conclusion

Mosquito-borne diseases pose a significant challenge for public health in Brazil. In Recife, the most common arboviruses transmitted are Dengue, Zika, and chikungunya fever, whose main vector is the *Aedes aegypti* mosquito. The main strategy adopted by health systems to control these diseases is monitoring the occurrence of mosquitoes. Therefore, the mosquito's reproduction cycle is interrupted and, once the mosquito occurrence is monitored, the health authorities adopt simple measures. For example, the treatment and/or elimination of the breeding sites. Consequently, the creation of spatiotemporal models for the geospatial prediction of mosquito breeding sites can assist health managers to develop effective strategies to combat mosquitoes and prevent their emergence of new outbreaks.

The forecasting strategy based on machine learning was revealed to be highly advantageous in creating spatiotemporal distribution models for *Aedes aegypti* mosquito breeding sites. For the city of Recife, the algorithm which presented the best performance was the 3-degree polynomial-kernel SVR. For this regressor, the RRSE (%) achieved 14.60%, whereas the correlation coefficient obtained was 0.9875. In contrast, the models generated with the RBF-kernel SVR presented the worst performance among the evaluated regressors. For this regressor, the correlation coefficient and RRSE (%) achieved were 63.84% and 0.8002, respectively. According to the qualitative results, the northern, western, and southern regions of Recife appear more frequently with large concentrations of mosquito breeding sites. In the city's southernmost region, Jordão stands out as the neighborhood which usually presented very high concentrations from 2014-2016. In the city's northern part, the

neighborhoods of Nova Descoberta, Vasco da Gama, and Dois Irmãos are the most affected by the mosquitos' presence.

Thus, the spatiotemporal approach provides the agents better assessment of areas that are greatly affected by the mosquito's presence. In addition, the heatmaps assist the identification of regions that are in transition areas, which can become areas with high/very high concentrations of mosquito breeding sites. Finally, the spatiotemporal approach can support health managers in decision-making, in directing human and financial resources. In addition, it can contribute to improving the intervention strategies to combat the mosquito, as well as create new ones.

Conflicts of Interest

The authors do not have any conflicts of interest to declare.

Acknowledgment

This research was conducted under the project titled: Mosquito populations modelling for early warning system & rapid public health response (MEWAR). This project is funded by the Belmont Forum, which was supported in the United Kingdom by UKRI NERC under the grant NE/T013664/1, and in Turkey by TÜBİTAK under the grant 119N373. This work was supported in Brazil by FAPESP under the grants 2019/23553-1 and 2020/11567-5, and this is financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001, and Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq Brazil, grand CNPq-DT2-2018.

References

Akil, L., & Ahmad, H. A. (2016). Salmonella infections modelling in Mississippi using neural network and geographical information system (GIS). *BMJ open*, *6*(3), e009255. https://doi.org/10.1136/bmjopen-2015-009255.

Baquero OS, Santana LMR, Chiaravalloti-Neto F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. PLoS ONE. 2018; https://doi.org/10.1371/journal.pone.0195065.

Barbosa VADF, Gomes JC, de Santana MA, de Lima CL, Calado RB, Bertoldo Júnior CR, Dos Santos WP. Covid-19 rapid test by combining a Random Forest-based web system and blood tests. Journal of Biomolecular Structure and Dynamics. 2021; https://doi.org/10.1101/2020.06.12.20129866 .

Brazil. Health Ministry. Health Surveillance Department. Communicable Disease Surveillance Department. Larval Index Rapid Assay for *Aedes aegypti* (LIRAa) for entomological surveillance of *Aedes aegypti* in Brazil: methodology for assessment of Breteau Index and Building's indexes and type of containers. Brasília. 2013.

Brazil. Health Ministry. O Agente Comunitário de Saúde no controle da dengue. 2009; ISBN 978-85-334-1548-5.

Breiman, L. Random forests. Machine Learning. 2001;https://doi.org/10.1023/A:1010933404324. .

Cao-Lormeau VM, Blake A, Mons S, Lastère S, Roche C, Vanhomwegen J, Dub T, Baudouin L, Teisser A, Larre P et al. Guillain Barré syndrome outbreak associated with zika virus infection in French Polynesia: a case control study. *The Lancet*, Elsevier. 2016;https://doi.org/10.1016/S0140-6736(16)00562-6.

da Silva CC, de Lima CL, da Silva, ACG, Silva EL, Marques GS, de Araújo LJB da Silva Filho AG. Covid-19 dynamic monitoring and real-time spatio-temporal forecasting. Frontiers in Public Health. 2021; https://doi.org/10.3389/fpubh.2021.641253.

da Silva, CC, de Lima CL, da Silva ACG, Moreno GMM, Musah A, Aldosery A, ... dos Santos, WP. 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, 2021; https://doi.org/10.33448/rsd-v10i12.19984.

de Freitas Barbosa VA, Gomes JC, de Santana MA, Jeniffer EDA, de Souza RG, de Souza RE, dos Santos, WP . Heg. IA: Um sistema inteligente para apoiar o diagnóstico de Covid-19 com base em exames de sangue. Research on Biomedical Engineering . 2021;

Donateli CP, Avelar PS, Einloft ABN, Cotta RMM, Costa GD. Evaluation of Health Surveillance in the Zona da Mata Mineira: from standards to practice. Ciência & Saúde Coletiva. 2017; https://doi.org/10.1590/1413-812320172210.18252017.

Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Advances in neural information processing systems. 1997.

Espinosa M, Weinberg D, Rotela CH, Polop F, Abril M, Scavuzzo CM. Temporal Dynamics and Spatial Patterns of *Aedes aegypti* Breeding Sites, in the Context of a Dengue Control Program in Tartagal (Salta Province, Argentina). PLoS Neglected Tropical Diseases. 2016; https://doi.org/10.1371/journal.pntd.0004621.

Fered G, Tiruneh M, Abate E, Kassa WJ , Wondimeneh Y, Damtie D, Tessema B. Distribution and larval breeding habitats of Aedes mosquito species in residential areas of northwest Ethiopia. Epidemiol Health. 2018; https://doi.org/10.4178/epih.e2018015.

Freitas JR, Santos ALP, Piscoya VC, Cunha ALX Filho MC. Modelagem em Séries Temporais Aplicados a Números de Notificações Mensais de Dengue em Pernambuco. Comunicação oral. In the III Congresso internacional das ciências agrárias. 2018; https://doi.org/10.31692/2526-7701.IIICOINTERPDVAGRO.2018.00166.

Fong-Shue C, Yao-Ting T, Pi-Shan H, Chaur-Dong C, Ie-Bin L, Day-Yu C. Re-assess Vector Indices Threshold as an Early Warning Tool for Predicting Dengue Epidemic in a Dengue Non-endemic Country. 2015; https://doi.org/10.1371/journal.pntd.0004043.

Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. PLoS Negl Trop Dis. 2017; https://doi.org/10.1371/journal.pntd.0005973.

Hamlet A, Jean K, Perea W, Yactayo S, Biey J, Van Kerkhove M, Ferguson N, Garske T. The seasonal influence of climate and environment on yellow fever transmission across Africa. PLoS Neglected Tropical Diseases. 2018; https://doi.org/10.1371/journal.pntd.0006284

Haykin, S. (2001). Neural networks: principles and practice. *Bookman*, *11*, 900.

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 10 Nov 2021.

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. (2018). 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*, *3*(1), 5. https://doi.org/10.3390/tropicalmed3010005

Mattioli, F.E.R., Andrade, R.B. & Estevez E.T. (2017). Predição de Casos de Dengue Utilizando Redes Neurais Artificiais. *Jornal de Engenharia, Tecnologia e Meio Ambiente*, 1(2), 8-12.

Melo, J.C.S. & Moraes, R.M. (2018). Sistema Espacial de Suporte à decisão para gestão do combate ao dengue usando lógica fuzzy. *Tendências em Matemática aplicada e computacional*, 19(3), 405-421. https://doi.org/10.5540/tema.2018.019.03.0405

Mittelmann M, Soares DG. Previsão de Casos de Dengue no Município de Guarulhos com Redes Neurais Artificiais Multicamadas e Recorrentes. Revista de Informática Aplicada. 2017; https://doi.org/10.13037/ria.vol13n2.200.

Montgomery DC, Runger GC. Estatística aplicada e probabilidade para engenheiros, 2ª. Ed. Rio de Janeiro: Editora LTC. 2003.

National Institute of Meteorology. (n.d). Sobre o INMET. Retrieved from https://portal.inmet.gov.br/sobre

PAHO. Dengue topics. 2021. Retrieve from https://www.paho.org/pt/topicos/dengue. Accessed 25 Set 2021.

Peña-García VH, Triana-Chávez O, Mejía-Jaramillo AM, Díaz FJ, Gómez-Palacio A, Arboleda-Sánchez S. Infection Rates by Dengue Virus in Mosquitoes and the Influence of Temperature May Be Related to Different Endemicity Patterns in Three Colombian Cities. Int J Environ Res Public Health. 2016; https://doi.org/10.3390/ijerph13070734.

Pernambuco Water and Climate Agency. Institutional. 2017. Retrieved from https://www.apac.pe.gov.br/intitucional

QGIS, A Free and Open Source Geographic Information System. Retrieved June , 2021, https://qgis.org/en/site/

Rosenblatt F.. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 1958.

Salami D, Sousa CA, Martins MRO, Capinha C. Predicting dengue importation into Europe, using machine learning and model-agnostic methods. Scientific Reports. 2020; https://doi.org/10.1038/s41598-020-66650-1

Scavuzzo JM, Trucco FC, Tauro CB, German A, Espinosa M, Abril M. Modeling the temporal pattern of Dengue, Chikungunya, and Zika vector using satellite data and neural networks. In the XVII Workshop on Information Processing and Control (RPIC). 2017.

Senn S. Francis Galton e regressão à média. *Significância.* 2011.

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. 2018; DOI: 10.1109/ACCESS.2018.2871241.

Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing. 2004.

Rahman KM, Sharker Y, Rumi RA, Khan MI, Shomik MS, Rahman MW, Billah M, Rahman M, Streatfield PK, Harley D, Luby SP. An Association between Rainy Days with Clinical Dengue Fever in Dhaka, Bangladesh: Findings from a Hospital Based Study. Int J Environ Res Public Health. 2020; https://doi.org/10.3390/ijerph17249506.

Tosepu R, Tantrakarnapa K, Worakhunpiset S, Nakhapakorn K. Climatic Factors Influencing Dengue Hemorrhagic Fever in Kolaka District, Indonesia. Environment and Natural Resources Journal. 2018; https://doi.org/10.14456/ennrj.2018.10

Veríssimo FS, Barsante LS, Acebal JL, Cardoso RTN. Modelagem e controle do *Aedes aegypti* durante as estações do ano através do Algoritmo Genético. *Proceeding Series of the Brazilian Society of Applied and Computational Mathematics*. 2016; https://doi.org/10.5540/03.2016.004.01.0062.

Wilder-Smith A, Gubler DJ, Weaver SC, Monath TP, Heymann DL, Scott TW. Epidemic arboviral diseases: priorities for research and public health. *The Lancet of Infection Diseases*. 2017; https://doi.org/10.1016/S1473-3099(16)30518-7

Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Technique (San Francisco, CA, USA: Morgan Kaufmann Publishers), 2005.