Forecasting Dengue, Chikungunya and Zika cases in Recife, Brazil: a spatio-temporal approach based on climate conditions, health notifications and machine learning

Predição de casos de Dengue, Chikungunya e Zika em Recife, Brasil: uma abordagem espaço-temporal com base em condições climáticas, notificações de saúde e aprendizado de máquina

Pronóstico de casos de Dengue, Chikungunya y Zika en Recife, Brasil: un enfoque espacio-temporal basado en las condiciones climáticas, notificaciones de salud y aprendizaje de máquina

Received: 09/19/2021 | Reviewed: 09/24/2021 | Accept: 09/25/2021| Published: 09/26/2021

Cecilia Cordeiro da Silva
ORCID: https://orcid.org/0000-0002-7061-8477
Universidade Federal de Pernambuco, Brazil
E-mail: ccs2@cin.ufpe.br
Clarisse Lins de Lima
ORCID: https://orcid.org/0000-0003-1198-8627
Universidade de Pernambuco, Brazil
E-mail: cll@ecomp.poli.br
Ana Clara Gomes da Silva
ORCID: https://orcid.org/0000-0002-2823-5763
Universidade Federal de Pernambuco, Brazil
E-mail: clara.gomes@ufpe.br
Giselle Machado Magalhães Moreno
ORCID: https://orcid.org/0000-0003-4076-3494
Universidade de São Paulo, Brazil
E-mail: gisellemoreno@usp.br
Anwar Musah
ORCID: https://orcid.org/0000-0001-7978-1871
University College London, United Kingdom
E-mail: a.musah@ucl.ac.uk
Aisha Aldosery
ORCID: https://orcid.org/0000-0003-3287-0986
University College London, United Kingdom
E-mail: a.aldosery@ucl.ac.uk
Livia Dutra
ORCID: https://orcid.org/0000-0002-1349-7138
Universidade de São Paulo, Brazil
E-mail: livia.dutra@iag.usp.br
Tercio Ambrizzi
ORCID: https://orcid.org/0000-0001-8796-7326
Universidade de São Paulo, Brazil
E-mail: tercio.ambrizzi@iag.usp.br
Iuri Valério Graciano Borges
ORCID: https://orcid.org/0000-0001-9274-6309
Universidade de São Paulo, Brazil
E-mail: iurivalerio@usp.br
Merve Tunali
ORCID: https://orcid.org/0000-0003-1612-4705
Bogaziçi University, Turkey
E-mail: merve.tunali@boun.edu.tr
Selma Basibuyuk
ORCID: https://orcid.org/0000-0001-6790-1522
Bogaziçi University, Turkey
E-mail: selmabasibuyuk@gmail.com
Orhan Yenigün
ORCID: https://orcid.org/0000-0002-5904-9832
Bogaziçi University, Turkey
E-mail: yeniguno@boun.edu.tr

Kate Jones
ORCID: https://orcid.org/0000-0001-5231-3293
University College London, United Kingdom
E-mail: kate.e.jones@ucl.ac.uk
Luiza Campos
ORCID: https://orcid.org/0000-0002-2714-7358
University College London, United Kingdom
E-mail: l.campos@ucl.ac.uk
Tiago Lima Massoni
ORCID: https://orcid.org/0000-0002-9423-7556
Universidade Federal de Campina Grande, Brazil
E-mail: massoni@dsc.ufcg.edu.br
Abel Guilhermino da Silva Filho
ORCID: https://orcid.org/0000-0002-7876-2756
Universidade Federal de Pernambuco, Brazil
E-mail: agsf@cin.ufpe.br
Patty Kostkova
ORCID: https://orcid.org/0000-0002-2281-3972
University College London, United Kingdom
E-mail: p.kostkova@ucl.ac.uk
Wellington Pinheiro dos Santos
ORCID: https://orcid.org/0000-0003-2558-6602
Universidade Federal de Pernambuco, Brazil
E-mail: wellington.santos@ufpe.br

Abstract

Dengue has become a challenge for many countries. Arboviruses transmitted by Aedes aegypti spread rapidly over the last decades. The emergence chikungunya fever and zika in South America poses new challenges to vector monitoring and control. This situation got worse from 2015 and 2016, with the rapid spread of chikungunya, causing fever and muscle weakness, and Zika virus, related to cases of microcephaly in newborns and the occurrence of Guillain-Barret syndrome, an autoimmune disease that affects the nervous system. The objective of this work was to construct a tool to forecast the distribution of arboviruses transmitted by the mosquito Aedes aegypti by implementing dengue, zika and chikungunya transmission predictors based on machine learning, focused on multilayer perceptrons neural networks, support vector machines and linear regression models. As a case study, we investigated forecasting models to predict the spatio-temporal distribution of cases from primary health notification data and climate variables (wind velocity, temperature and pluviometry) from Recife, Brazil, from 2013 to 2016, including 2015's outbreak. The use of spatio-temporal analysis over multilayer perceptrons and support vector machines results proved to be very effective in predicting the distribution of arbovirus cases. The models indicate that the southern and western regions of Recife were very susceptible to outbreaks in the period under investigation. The proposed approach could be useful to support health managers and epidemiologists to prevent outbreaks of arboviruses transmitted by Aedes aegypti and promote public policies for health promotion and sanitation.
Keywords: Dengue forecasting; Chikungunya forecasting; Zika forecasting; Arboviruses forecasting; Machine learning; Arboviruses prediction.

Resumo

A dengue se tornou um desafio para muitos países. Os arbovírus transmitidos por Aedes aegypti se espalharam rapidamente nas últimas décadas. A emergência de febre chikungunya e zika na América do Sul apresenta novos desafios para o monitoramento e controle de vetores. Essa situação piorou a partir de 2015 e 2016, com a rápida disseminação da chikungunya, causando febre e fraqueza muscular, e do Zika vírus, relacionado a casos de microcefalia em recém-nascidos e a ocorrência da síndrome de Guillain-Barret, doença autoimune que afeta o sistema nervoso. O objetivo deste trabalho foi construir uma ferramenta para previsão da distribuição de arbovírus transmitidos pelo mosquito Aedes aegypti por meio da implementação de preditores de transmissão de dengue, zika e chikungunya baseados em aprendizado de máquina, com foco em redes neurais perceptrons multicamadas, máquinas de vetores de suporte e modelos de regressão linear. Como um estudo de caso, investigamos modelos de previsão para prever a distribuição espaço-temporal de casos a partir de dados de notificação de saúde primária e variáveis climáticas (velocidade do vento, temperatura e pluviometria) de Recife, Brasil, de 2013 a 2016, incluindo o surto de 2015. O uso de análises espaçotemporais sobre perceptrons multicamadas e resultados de máquinas de vetores de suporte mostraram-se bastante eficazes na previsão da distribuição de casos de arbovírus. Os modelos indicam que as regiões sul e oeste do Recife foram muito suscetíveis a surtos no período investigado. A abordagem proposta pode ser útil para apoiar gestores de saúde e epidemiologistas na prevenção de surtos de arbovírus transmitidos pelo Aedes aegypti e na promoção de políticas públicas de promoção da saúde e saneamento.
Palavras-chave: Previsão da dengue; Previsão de Chikungunya; Previsão do Zika; Previsão de arbovírus; Aprendizado de máquina; Predição de arbovírus.

Resumen

El dengue se ha convertido en un desafío para muchos países. Los arbovirus transmitidos por Aedes aegypti se han propagado rápidamente en las últimas décadas. La aparición de la fiebre chikungunya y Zika en América del Sur presenta nuevos desafíos para el monitoreo y control de vectores. Esta situación se agravó a partir de 2015 y 2016, con la rápida propagación del chikungunya, que provoca fiebre y debilidad muscular, y el virus Zika, relacionado con casos de microcefalia en recién nacidos y la aparición del síndrome de Guillain-Barret, una enfermedad autoinmune que afecta al sistema nervioso. El objetivo de este trabajo fue construir una herramienta para predecir la distribución de arbovirus transmitidos por el mosquito Aedesaegypti mediante la implementación de predictores de transmisión de dengue, zika y chikungunya basados en aprendizaje de máquina, con foco en redes neuronales de perceptrones multicamadas, máquinas de vector de soporte y modelos de regresión lineal. Como estudio de caso, investigamos modelos de predicción para predecir la distribución espacio-temporal de casos a partir de datos de notificación de salud primaria y variables climáticas (velocidad del viento, temperatura y lluvia) de Recife, Brasil, 2013 a 2016, incluido el brote de 2015. El uso de análises espacio-temporal por medio de perceptrones multicamadas y los resultados de las máquinas de vectores de soporte demostraron ser muy eficaces para predecir la distribución de los casos de arbovirus. Los modelos indican que las regiones sur y oeste de Recife fueron muy susceptibles a brotes en el período investigado. El enfoque propuesto puede ser útil para apoyar a los administradores de salud y epidemiólogos en la prevención de brotes de arbovirus transmitidos por Aedes aegypti y en la promoción de políticas públicas para promover la salud y el saneamiento.

Palabras clave: Pronóstico del dengue; Pronóstico de Chikungunya; Pronóstico del Zika; Predicción de arbovirus; Aprendizaje de máquina; Predicción de arbovírus.

1. Introduction

Prevention and control of dengue fever, chikungunya fever and zika has been a major public health challenge for many countries. Since 2015 other arboviruses have interacted with the dengue virus, which has spread rapidly over the past two decades (de Lima et al., 2016; Bhatt et al., 2013). It is estimated that around 390 million new cases of dengue occur each year. However, problems such as misdiagnosis and inaccurate reporting or absence of case reporting in many regions can contribute to the underestimation of the impact of dengue and other arboviruses transmitted by the mosquito *Aedes aegypti* (de Lima et al., 2016). The emergence of other arboviruses, such as chikungunya fever and zika, especially in South America, poses new challenges to vector monitoring and control. This situation worsens from 2015 and 2016, with the rapid spread of chikungunya, causing fever and muscle weakness, among other symptoms, and the emergence of Zika virus, partially related to cases of microcephaly in newborns and directly related to the occurrence of Guillain-Barret syndrome, an autoimmune disease that affects the nervous system, ranging from muscle weakness to paralysis (Cao-Lormeau et al., 2016).

Dengue is a viral infection transmitted to humans through mosquitoes, and is spreading rapidly around the world. Its primary vector is the mosquito *Aedes aegypti*, a species well adapted to urban areas and distributed mainly in tropical and subtropical regions, but also operating in North America and Europe. Evidence indicates that a secondary vector, the mosquito *Aedes albopictus*, has also been expanding its geographic range (de Lima et al., 2016; Bhatt et al., 2013). The risk of arbovirus outbreaks and their endemic presence is higher in tropical and subtropical regions, but is also increasingly present in North America and Europe, due to the presence of mosquitoes *Aedes* and the introduction of viruses (de Lima et al., 2016; Bhatt et al., 2013).

The transmission of arboviruses is a complex process that involves the interaction of multiple agents: human populations, mosquitoes and viruses conditioned by climatic and environmental factors in a very heterogeneous space. The space in which these interactions take place is complex enough that the study of arboviral transmission is fraught with challenges. Arborovirus pandemics have been favored by a combination of several factors: the global mobility of human populations and mosquito circulation; the swelling of overcrowded urban areas; the difficulty of access by urban populations, especially the economically disadvantaged sectors, to basic sanitation, regular water supply, and the public health system; environmental and climatic factors, such as temperature and rainfall, which measure rainfall density and occurrence; and, finally, the inefficiency of vector control strategies (de Lima et al., 2016; Gubler, 2011; Mohammed & Chadee, 2011).

Several research groups have been dedicated to building risk maps and estimating the global distribution of arboviruses and their correlation with environmental data. Despite the importance of these efforts to map the distribution of these diseases, it is also important to understand the dynamics of arboviruses on a local scale, which is done through mathematical and computational models (Padmanabhan et al., 2017; Jindal & Rao, 2017; de Lima et al., 2016). Local climatic conditions, such as temperature, rainfall and humidity, interfere with vector development, from hatching to mosquito life and dispersal, and other aspects of arboviral transmission (de Lima et al., 2016; Gubler, 2011). The advancement of Digital Epidemiology and geoprocessing technologies, coupled with the development of Data Mining and Machine Learning techniques, have provided rapid monitoring, control and simulation of disease spread, assisting public health systems in controlling epidemics and of the environmental and behavioral factors that favor the vectors of these diseases (Salathe et al., 2012; Beltrán et al., 2018; Musah et al., 2019; Rubio-Solis et al., 2019; Kostkova et al., 2019).

In Brazil, arboviruses have received special attention from the Unified Health System through public health policies and campaigns (Pessanha et al., 2009). In Recife, the Recife Municipal Health Secretariat, through its Open Data Portal, distributes the mapping of diseases and symptoms by health unit and the patient’s neighborhood of origin since 2015. The Pernambuco Water and Climate Agency, APAC also provides a geographic information system where the daily and monthly rainfall series are published since 2006, by city and, in the case of Recife, by neighborhood.

Machine learning techniques have been shown to be useful to support the diagnosis and prediction of prognosis of different diseases based on biomedical signs and images and different clinical parameters (Commowick et al., 2018; S. M. de Lima et al., 2016; Santana et al, 2018; Cordeiro et al., 2016; Barbosa et al., 2021; de Souza et al., 2021; Pereira et al., 2021). Additionally, machine learning-based regression techniques have been successfully used for temporal and spatiotemporal prediction of contagious diseases such as Covid-19 (da Silva et al., 2021; de Lima et al., 2020). We believe that this set of techniques can achieve good accuracy results when adapted to arboviruses, including not only the spatial and temporal windows of the number of cases, but also climatic and environmental variables.

The objective of this work was to construct a tool to forecast the distribution of arboviruses transmitted by the mosquito *Aedes aegypti* by implementing dengue, zika and chikungunya transmission predictors based on machine learning, focused on multilayer perceptrons neural networks, support vector machines and linear regression models. As a case study, we investigated forecasting models to predict the spatio-temporal distribution of cases from primary health notification data and climate variables (wind velocity, temperature and pluviometry) from Recife, Brazil, from 2013 to 2016, including 2015’s outbreak. Multiplayer perceptrons demonstrated to be the most adequate models, reaching considerable high correlation coefficient values and percentual errors lower than 5%.

2. Methodology

2.1 Proposed Method

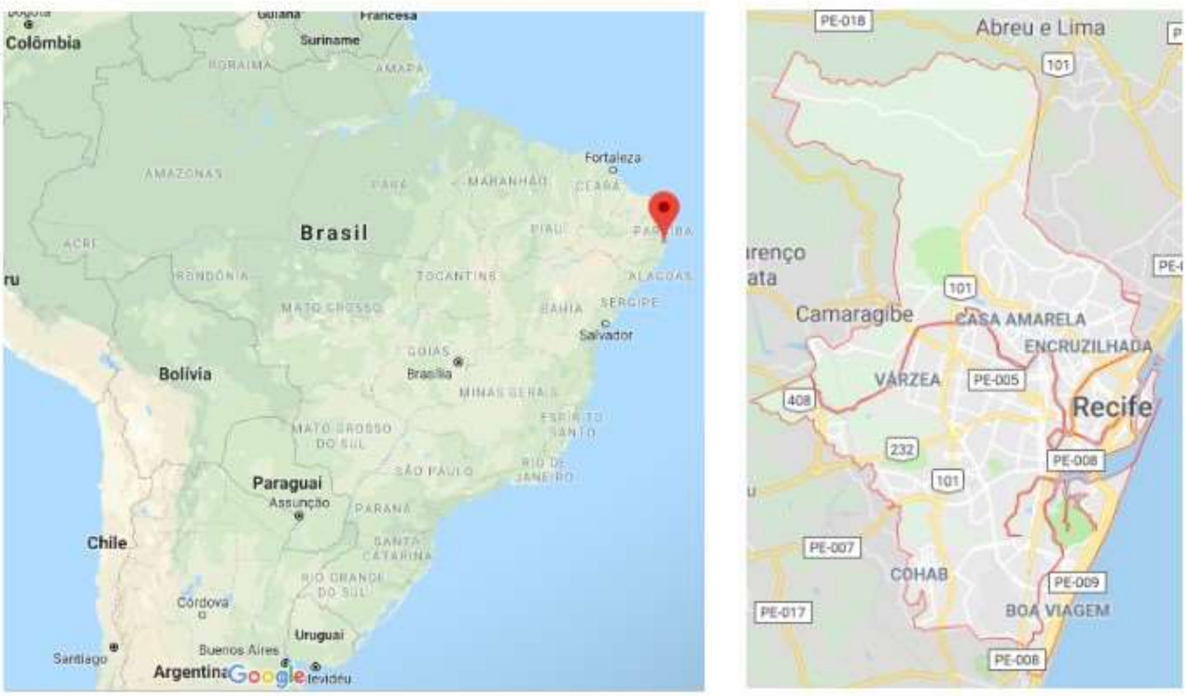
In this work, we propose a prototype of a system for spatio-temporal prediction of the distribution of cases of arboviruses, i.e. dengue, chikungunya and zika. The main hypothesis of this work is that the monthly average measurements of temperature and wind speed, and the number of arbovirus cases per two months by geographic location, considering a 12-month prediction window, can be used to predict the spatial and temporal distribution of dengue, chikungunya and Zika cases. As a case study, we used the climatic variables obtained from the national meteorological systems, and the case information by neighborhood of the City of Recife, available in the National Notification System and in the Open Data Portal of the City of Recife, from 2014 to 2016. Considering the predictive models, we also start from the hypothesis that machine learning methods can be used successfully to predict the spatiotemporal distribution of arboviruses cases in this context (Koche, 2011; A. S. Pereira et al., 2018; Ludke & André, 2013; Yin, 2015).

Therefore, this research is characterized as a quali-quanti case study, ie a case study that combines qualitative aspects (visual analysis from the spatial distributions generated by the geographic information system) and quantitative aspects (regression evaluation indices and other statistics of interest) in your analysis. The following subsections present in detail the databases, data preparation and pre-processing, machine learning methods used to build the predictive models, the geographic information system, and quality indices.

2.2 Area under study

The area delimited for this study was the City of Recife (8° 03’14” S, 34° 52’51” W), capital of the State of Pernambuco which is located in the northeast region of Brazil (Figure 1). Recife, according to the Brazilian Institute of Geography and Statistics (IBGE) has a territorial extension of approximately 218km^2 and about 1,637,834 million inhabitants, besides being the city Northeast with the highest Human Development Index (HDI). The climate of the city of Recife is characterized as tropical humid, with average monthly temperatures above 18° C, high relative humidity and high rainfall throughout the year (INMET).

Figure 1: Localization of the City of Recife.



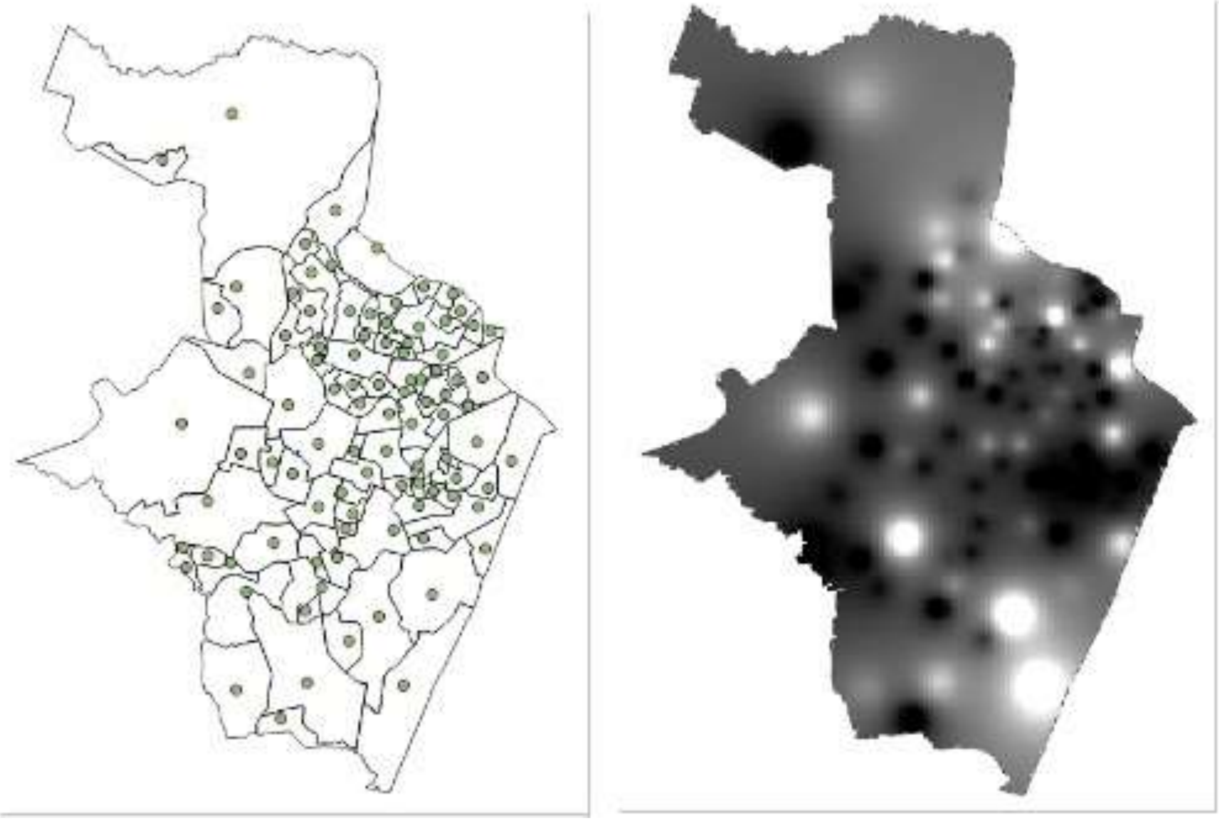
Source: Authors.

2.3 Mapping of arbovirus cases

Data on arbovirus cases were obtained through the Open Data Portal of Recife City (http://dados.recife.pe.gov.br/), which contains the records of the number of cases of Dengue, Zika. and *chinkugunya* from 2013 to 2016. For each two months of each year, the number of arboviral cases in each of the 94 districts of Recife was counted separately. From the information on the number of cases in each neighborhood, a vector layer of points was generated, *shapefile* (.shp), geographically locating the number of cases to each neighborhood of the city geographically, as can be observed in the map on the right in Figure 2. In order to estimate the distribution of arboviral cases throughout the municipality, the QGIS interpolation tool was used, in which the interpolation method selected was the inverse distance interpolation. As a result of the interpolation, we obtained a raster image

(.tif) that can be observed in the map on the left in Figure 2. Rasters were generated for each quarter from 2013 to 2016, where each raster represents the distribution map of the cases of arboviruses.

Figure 2: On the left is the dotted vector layer of arbovirus cases. On the right is the arboviruses distribution map of the first two months of 2013.



Source: Authors.

2.4 Mapping of climate variables

Climatic factors such as rainfall and temperature are among the causes of an increase in arboviruses. Mosquito behavior is determined by weather conditions. This is because rainfall, temperature, and humidity affect the interaction of biological and viral vectors throughout life, mating age, spread, feeding, and faster viral replication (Morin et al., 2013; LaDeau et al., 2015).

The monitoring of climate variables in Brazil is performed by the National Institute of Meteorology (http://www.inmet .gov.br), INMET. This monitoring is performed through stations distributed throughout the country, one of which is located in Recife City. Temperature and wind speed data were collected from the INMET database, where the historical series of daily measurements of weather stations from 1961 are found. From this database, the historical series of the years were collected from 2013 to 2016.

The Pernambuco Water and Climate Agency (http://www.apac.pe.gov.br), in Recife, monitors hydrometeorological indices through the Pernambuco Hydrometeorological Geoinformation System, the SIGHPE. Data related to the rainfall indexes of the city of Recife were collected in the SIGHPE database, which contains the historical series of hydrometeorological indexes, since 2006, of the rainfall stations distributed in the city. For this work, only the accumulated rainfall from 2013 to 2016 were collected. In the case of Recife, wind temperature and speed records are monitored by a single station, while hydrometeorological records are carried out by three stations in different neighborhoods. The temperature and wind velocity records in the other districts of Recife were estimated using the Gaussian distribution.

Sample standard deviation values were calculated from Equation 1, where x_{max} represents the maximum value and μ represents the monthly average of wind temperature and velocity values.

$$\sigma = \frac{x_{max} - \mu}{4}. \tag{1}$$

For the monthly accumulated rainfall, the maximum value considered was the maximum value recorded between the three monitoring stations, while the average considered was the average of the accumulated rainfall between the three monitoring stations. With information on climate variables in all neighborhoods of Recife, the *shapefiles* were generated for each of the variables for each month from 2013 to 2016. Finally, the inverse distance interpolation tool was used to estimate the spatial distribution of climatic variables throughout the Recife territory.

2.5 Regression models

Linear Regression

The linear regression is the simplest method to predict numeric values. In this method, it is assumed that the data has a linear behavior, and that the prediction variable can be represented as a linear combination of the attributes with their pre-determined weights (Witten & Frank, 2005). Thus, the general model of linear regression is represented by the Equation 2.

$$y = w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n, \tag{2}$$

where y is the prediction variable; $x_1,x_2,...,x_n$, represent the values of the attributes and $w_0,w_1,w_2,...,w_n$ represent the weights of each attribute. The idea of the linear regression algorithm is, then, to find the optimal weights that best represent the problem. One of the ways to find the optimal weights is to minimize the sum of the squared difference between the predicted value and the actual value (Witten & Frank, 2005). The sum of the squared difference is calculated by Equation 3:

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

Artificial Neural Networks

Artificial neural networks (ANN), consists in a machine learning technique based on the behavior of the human brain (Siriyasatien et al., 2018). The neural networks consist of smaller units, artificial neurons, which are fundamental to their functioning. The artificial neurons contains the following elements: (1.) a set of synapses or connectors where a signal x_i at the entrance to the synapse j connected to the k neuron is multiplied by the synaptic weight $w_{k,j}$ (2.) an adder to add the input signals, weighted by the respective neuron synapses; (3.) an activation function to limit the output of a neuron (Haykin, 2001). Mathematically, an artificial neuron is represented by the Equation 4 and by the Equation 5:

$$u_k = \sum_{j=1}^{n} w_{k,j} x_i, \tag{4}$$

$$y_k = \varphi(u_k + b_k), \tag{5}$$

wherein $x_1,x_2,...,x_n$ represent the input signals; $w_{k,1},w_{k,2},...,w_{k,n}$ represent the synaptic weights of the input signals x_i for the k-th neuron; b_k, is the term bias and φ is a neuron activation function. In regression applications, the inputs $x_1,x_2,...,x_n$ of the input layer correspond to the forecasting window. For instance, in case of temporal forecasting, the inputs are observed time window of the time series.

The network architecture used in this work was the Multilayer Perceptron (MLP). In this configuration, the neural network has an input layer, two or more hidden layers and an output layer (Haykin, 2001). ANNs have also been widely used to predict disease cases. For example, in the prediction of dengue cases in the city of São Paulo, Brazil (Baquero et al., 2018). They were also used to predict dengue outbreaks in the northeastern coast of Yucatán, Mexico, and in San Juan, Puerto Rico (Laureano-Rosario et al., 2018). Moreover, the ANNs were applied to model cases of infection by Salmonella in the state of Mississippi, USA (Akil and Ahmad, 2016).

Support Vector Regression

The support vector regression is a supervised machine learning technique for data analysis and pattern recognition. The idea of the SVR algorithm is to find the best hyperplane defined by Vapnik's ε-insensitivity loss function. When this hyperplane is found, a linear regression is applied to the corresponding hyperplane. In situations where the problem is linearly separable, the best hyperplane is given by the equation:

$$y = \mathbf{w}^T\mathbf{x} + b, \tag{6}$$

where $\mathbf{w} = (w_1,w_2,...,w_n)^T$ is the vector of weights, $\mathbf{x} = (x_1,x_2,...,x_n)^T$ is the feature vector, and b is the bias. For problems that are not linearly separable, the data is mapped to a hyperplane in a larger dimension. Thereupon, the algorithm seeks to solve the problem by applying the linear regression of the equation 6 in the corresponding hyperplane. For nonlinearly separable problems, SVR machines use kernel functions, $K : \mathrm{R}\times\mathrm{R} \rightarrow \mathrm{R}$. Then, the SVR output assumes the following expression:

$$y = K(\mathbf{w},\mathbf{x}), \tag{7}$$

where the kernel function can be polynomial, sigmoidal, Gaussian, or even assume other mathematical expressions (Drucker et al., 1997; Witten and Frank, 2005; Smola and Schölkopf, 2004).

2.6 Metrics

The main metrics we adopted to evaluate the models are the following: the correlation coefficient and the Relative Quadratic Error (RMSE percentage). The correlation coefficient is a statistical measure between expected and forecasted values. This value varies from -1 to 1. When it approaches 1, it indicates a strong positive correlation. Conversely, when the correlation coefficient is close to -1, it indicates that the variables have a strong negative correlation. When the correlation coefficient is close to zero, it indicates that there is no correlation between the variables (Witten and Frank, 2005). The value of the correlation coefficient serves as the global evaluator for the model. Therefore, it is possible to obtain a high correlation coefficient as well as at the same time obtain high values for local errors. For this reason, it cannot be the only metric for assessing model performance. In order to avoid a superficial evaluation of the regressors, we therefore chose the RMSE (%) as an evaluation metric. The Equation 8 shows the expression of the calculation of the relative quadratic error, where p_i is the predicted value and a_i is the actual value, for $i = 1,2,...,n$.

In addition to the RMSE (%), we also calculated the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Mean Percentage Error (MPE) (Equations 9-12):

$$\mathrm{RMSE}(\%) = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{\sum_{i=1}^{n} a_i^2}} \times 100\%, \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}, \tag{9}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i|, \tag{10}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{e_i}{a_i}\right|, \tag{11}$$

$$\text{MPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left(\frac{p_i - a_i}{p_i}\right), \tag{12}$$

where, p_i is the forecasted value, a_i is the actual value and $e_i = a_i - p_i$ is the difference between the actual value and the forecasted value.

The Pearson's Correlation Coefficient R is defined as follows:

$$R = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2 \cdot \sum_{i=1}^{n}(a_i - \bar{a})^2}}, \tag{13}$$

where p^{-} and a^{-} are the sample average values for the sets of predicted and actual values, respectively. Similarly, the Spearman's Rank Correlation Coefficient ρ is defined as following:

$$\rho = \frac{\sum_{i=1}^{n}(R(p_i) - \bar{R}(p))(R(a_i) - \bar{R}(a))}{\sqrt{\sum_{i=1}^{n}(R(p_i) - \bar{R}(p))^2 \cdot \sum_{i=1}^{n}(R(a_i) - \bar{R}(a))^2}}, \tag{14}$$

where $R(p_i)$ and $R(a_i)$ are the ranks of p_i and a_i, whilst $R^{-}(p)$ and $R^{-}(a)$ are the sample averages of the ranks of p_i and a_i, respectively.

The Kendall's Rank Correlation τ is given as follows:

$$\tau = \frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{i=1}^{j-1}\text{sign}(p_i - p_j)\cdot\text{sign}(a_i - a_j), \tag{15}$$

where n is the number of observations and $1 \leq i,j \leq n$. The signal function, sign, is defined as following:

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

for $x \in$ R.

2.7 Forecasting Set

The prediction sets were assembled from the distribution maps of arboviruses cases and climatic variables for each two months. The bimonthly prediction model was chosen due to the fact that the Brazilian Unified Health System (SUS) is planning to combat arbovirus outbreaks considering the bimester cycle. The attribute vectors of the prediction sets were assembled by simultaneously scanning the spatial distribution maps pixel by pixel and concatenating latitude and longitude with the following information, in the following order: distribution of arbovirus cases, and for each month of the bimester, the temperature distribution, rainfall and wind speed. Each prediction vector contains information from the six quarters preceding the prediction quarter. Therefore, 18 prediction sets were assembled, each with 15,553 instances and 44 attributes, in which the output of each prediction set is the pixel value of the arbovirus case distribution at the corresponding coordinate. The 15,553 instance sets were established as test sets.

The Weka machine learning environment (Frank et al., 2004; Hall et al., 2009), version 3.8, was used to assemble the training set from the resample tool. This tool allows a new database to be created with random values for instances, but with the same statistical characteristics as the original database. The number of instances of the new base must be specified. In this case, the training sets were generated by applying resample to each of the prediction sets with the number of instances equivalent to 30% the original set. Sets containing 15,553 instances were used to test the models created by the best regressor.

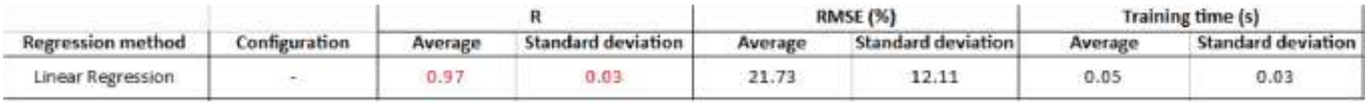
From the training set, we investigated the best regression architectures for predicting the distribution of arbovirus cases, namely: Linear Regression, Support Vector Machine (SVM) and Multilayer Perceptron (MLP) with a single hidden layer. For the SVM regressor, evaluations were performed with the following configurations: C = 0.1 and linear (or degree 1), 2 and 3-degree polynomial kernels, and RBF kernel. Regarding single layer MLP, we investigated architectures with 10, 20, 30 and 40 neurons in the hidden layer.

3. Results and Discussion

We evaluated each of the regressors in 30 rounds using 10-fold cross-validation. For the quantitative evaluation, we calculated the Correlation Coefficient (R), the Absolute Mean Error (MAE), the Mean Square Error (RMSE) and the Percent Relative Quadratic Error (RMSE percentage). However, the data were analyzed considering only the correlation coefficient, as global quality, and relative quadratic error as local quality metric. The detailed results of R, RMSE% and training time of each regressor are shown in Tables 1, 2 and 3. In this paper, we consider a high correlation coefficient to be above 0.9 and a low relative squared error to be below 5%. Best results are highlighted in red.

In Table 1, the results show that the linear regression presents satisfactory values for the correlation coefficient R, with average 0.97 and standard deviation of 0.03, and for the training time, as average of 0.05 and standard deviation. 0.03 and is therefore considered a very fast prediction method. On the other hand, the relative square error RMSE% presents a considerably high value, with an average of 21.23% and standard deviation of 12.11%.

Table 1: Correlation coefficient, relative square error and training time results for Linear Regression.

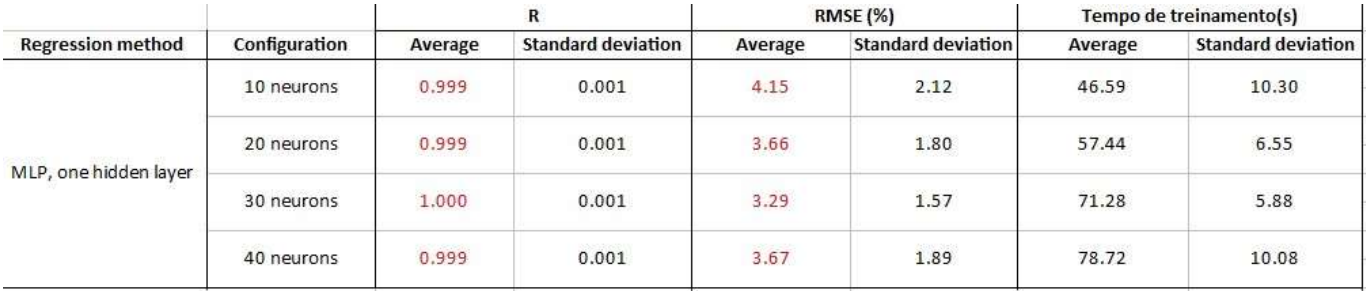


		R		RMSE (%)		Training time (s)	
Regression method	Configuration	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Linear Regression	-	0.97	0.03	21.73	12.11	0.05	0.03

Source: Authors.

Table 2 presents the results for multilayer neural networks with a single hidden layer in the configurations of 10, 20, 30 and 40 neurons. The results indicate that for all architectures evaluated, the correlation coefficients R presented very high values with averages around 0.999 and 1, and standard deviation of 0.001. Regarding the relative quadratic error, we observed that the 10-neuron configuration has a fairly low RMSE% with an average of 4.15% and this value decreases as the number of neurons in the hidden layer increases, reaching a minimum value of 3.29% in the 30-neuron configuration, followed by a considerable increase to 3.67% in the 40-neuron configuration. The behavior of training time shows an increase as the number of neurons in the hidden layer increases. Thus, considering the evaluation metrics, the best network configuration among the ones evaluated was the network with 30 neurons because of its high correlation coefficient, relative square error satisfactorily below the established limit of 5% and having a training time reasonably low.

Table 2: Correlation coefficient, relative squared error and training time results for multilayer perceptron, MLP, with 10, 20, 30 and 40 neurons in the hidden layer.

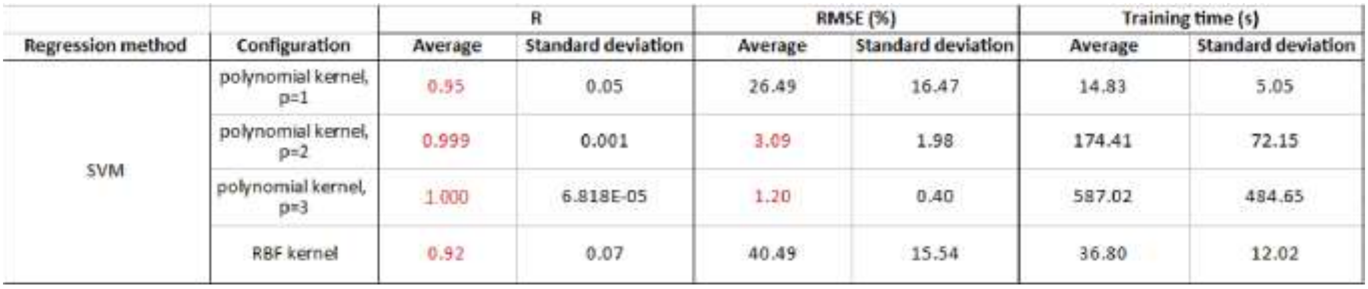


		R		RMSE (%)		Tempo de treinamento(s)	
Regression method	Configuration	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
MLP, one hidden layer	10 neurons	0.999	0.001	4.15	2.12	46.59	10.30
	20 neurons	0.999	0.001	3.66	1.80	57.44	6.55
	30 neurons	1.000	0.001	3.29	1.57	71.28	5.88
	40 neurons	0.999	0.001	3.67	1.89	78.72	10.08

Source: Authors.

Table 3 shows the results for SVM, with 1-degree (linear), 2- and 3-degree polynomial kernels, and RBF kernel. Correlation coefficient R values for all kernel configurations were considerably high and quite stable (low standard deviation), with emphasis on polynomial kernels of degrees 2 and 3, with averages of 0.999 and 1, respectively, and standard deviation of 0.001 and 6.818×10^{-5}, in that order. Regarding RMSE%, the 2- and 3-degree polynomial kernels obtained satisfactory values of 3.09% and 1.20%, respectively. Linear and RBF kernels obtained very high values of RMSE% (with average of 26.49% and 40.49%), considerably higher than the boundary of 5% established for this type of error. Regarding training time, SVM configurations were quite slow for 2- and 3-degree polynomial kernels, with a major disadvantage of 3-degree polynomials in its speed and stability. In contrast, for the linear and RBF kernels, the results showed that they are relatively fast compared to the training time of neural networks. However, they are considerably slower compared to linear regression. Taking into consideration the evaluation metrics, the 2-degree polynomial kernel is the best SVM configuration due to its high correlation coefficient R, low RMSE% and shorter training time among configurations that meet the requirements of R and RMSE% set for this task.

Table 3: Correlation coefficient, relative squared error and training time results for SVM, with linear (or grade 1), 2- and 3-degree polynomials, and RBF kernels.



		R		RMSE (%)		Training time (s)	
Regression method	Configuration	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
SVM	polynomial kernel, p=1	0.95	0.05	26.49	16.47	14.83	5.05
	polynomial kernel, p=2	0.999	0.001	3.09	1.98	174.41	72.15
	polynomial kernel, p=3	1.000	6.818E-05	1.20	0.40	587.02	484.65
	RBF kernel	0.92	0.07	40.49	15.54	36.80	12.02

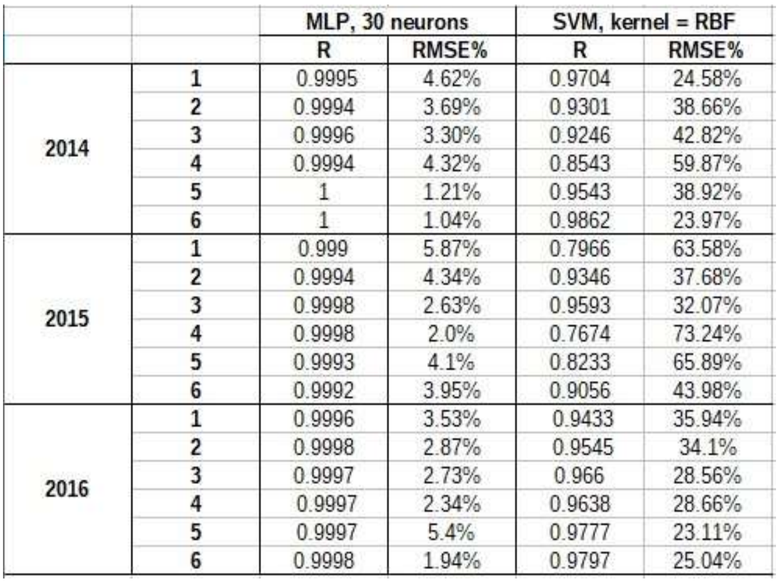
Source: Authors.

Overall, by evaluating all tested architectures, we can observe that the multilayer perceptron is a regressor that meets the needs of the prediction problem in question. As mentioned in this section, the configuration with 30 hidden layer neurons reached the best evaluation because it has a high correlation coefficient R, RMSE% below 5%, and reasonably short training time when compared to other regressors. Although training time is not critical in this type of problem, it was adopted as an important criterion to select the best regressor. After all, despite having very good correlation coefficients and very low RMSE%, 2- and 3-degree polynomial kernels SVM settings achieved very high values for training time. On the other hand, configuring SVM with RBF kernel has proved to be quite unsuitable for solving this problem. The training time for this configuration was considerably shorter than the MLP training time with 30 hidden layer neurons. However, as seen in Table 3, the RMSE% reached very high values, far above the established value as adequate.

Table 4 presents the results for MLP with 30 neurons in the hidden layer, considering the training set with 4,665 instances and the test set with 15,553 instances. This table also presents results for RBF-kernel SVM, considering these same sets. The metrics used to quantitatively evaluate the models are the correlation coefficient R and the relative square error (RMSE%). The qualitative evaluation of the models generated the distribution maps of arboviruses cases on the Recife map for each two months of 2014, 2015 and 2016.

For qualitative analysis, we generated the prediction images from the results obtained in the model validations. Figures 3, 4 and 5 correspond to the bimonthly predictions using MLP with 30 hidden layer neurons in 2014, 2015 and 2016, respectively.

Table 4: Validation results of the prediction models created by the multilayer perceptron, with a single layer and 30 neurons in the hidden layer.

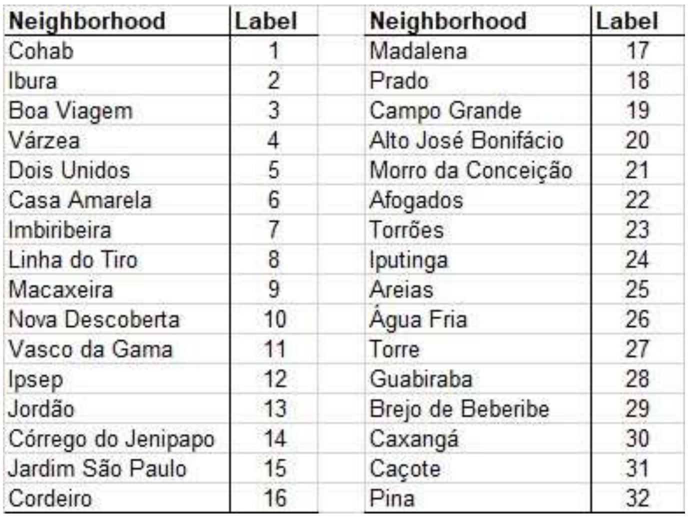


		MLP, 30 neurons		SVM, kernel = RBF	
		R	RMSE%	R	RMSE%
2014	1	0.9995	4.62%	0.9704	24.58%
	2	0.9994	3.69%	0.9301	38.66%
	3	0.9996	3.30%	0.9246	42.82%
	4	0.9994	4.32%	0.8543	59.87%
	5	1	1.21%	0.9543	38.92%
	6	1	1.04%	0.9862	23.97%
2015	1	0.999	5.87%	0.7966	63.58%
	2	0.9994	4.34%	0.9346	37.68%
	3	0.9998	2.63%	0.9593	32.07%
	4	0.9998	2.0%	0.7674	73.24%
	5	0.9993	4.1%	0.8233	65.89%
	6	0.9992	3.95%	0.9056	43.98%
2016	1	0.9996	3.53%	0.9433	35.94%
	2	0.9998	2.87%	0.9545	34.1%
	3	0.9997	2.73%	0.966	28.56%
	4	0.9997	2.34%	0.9638	28.66%
	5	0.9997	5.4%	0.9777	23.11%
	6	0.9998	1.94%	0.9797	25.04%

Source: Authors.

The warmer regions represent the areas with the highest concentrations of arbovirus cases, while the colder areas represent low case rates. The numeric labels of Recife's neighborhoods are shown in Table 5.

Table 5: Numeric labels of Recife's neighborhoods.



Neighborhood	Label	Neighborhood	Label
Cohab	1	Madalena	17
Ibura	2	Prado	18
Boa Viagem	3	Campo Grande	19
Várzea	4	Alto José Bonifácio	20
Dois Unidos	5	Morro da Conceição	21
Casa Amarela	6	Afogados	22
Imbiribeira	7	Torrões	23
Linha do Tiro	8	Iputinga	24
Macaxeira	9	Areias	25
Nova Descoberta	10	Água Fria	26
Vasco da Gama	11	Torre	27
Ipsep	12	Guabiraba	28
Jordão	13	Brejo de Beberibe	29
Córrego do Jenipapo	14	Caxangá	30
Jardim São Paulo	15	Caçote	31
Cordeiro	16	Pina	32

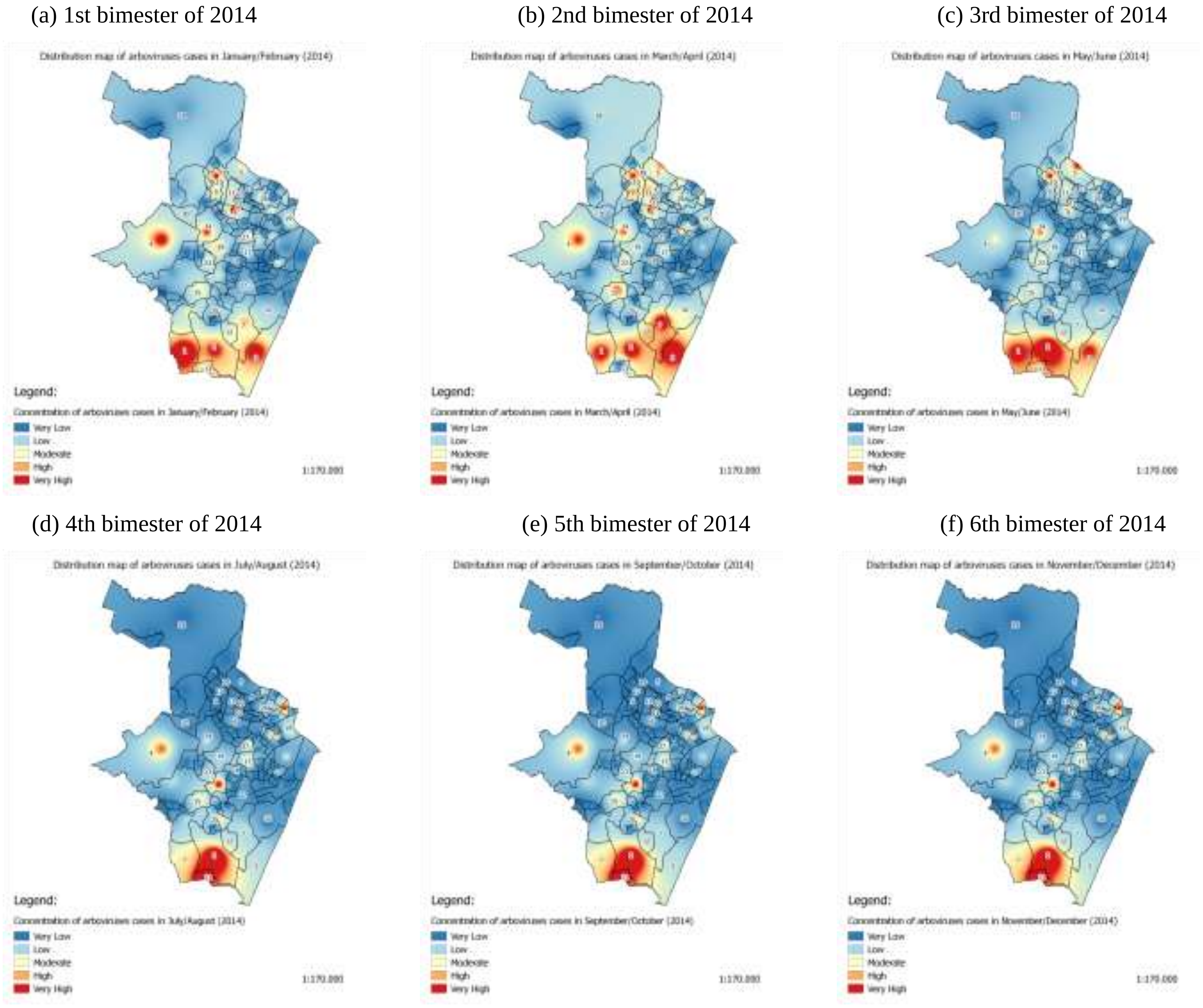
Source: Authors.

In Figure 3a, we can see that there was a higher concentration of arbovirus cases in the southern region of Recife. The most affected neighborhood in this area was Cohab, followed by Ibura and Boa Viagem. In the west of the city, the neighborhood with the highest rate of cases was Várzea. In the northernmost region of the city, the most affected neighborhoods were Dois Unidos and Casa Amarela. In the second quarter of the same year, Figure 3b, we can observe that the southern region of the city remains the most affected region. However, there is an increase in arbovirus cases in the Boa Viagem neighborhood and a decrease in cases in Cohab. We can also perceive an increase of cases in the neighborhood of Imbiribeira towards Pina. In the northern part of the city, it is possible to identify a significant increase in the neighborhoods of Dois Unidos, Linha do Tiro, Macaxeira, Nova Descoberta, and Vasco da Gama. In the third quarter of 2014 (Figure 3c), in the south zone, there is a considerable decrease in cases in Boa Viagem, Ipsep and Imbiribeira. However, the opposite occurs in the neighborhood of Ibura and Cohab. In the west of the city, Várzea also shows a remarkable reduction of cases.

The remaining two-month periods of 2014, Figures 3d, 3e and 3f, showed very similar behaviors. The southern zone remains the place with the highest rate of cases. From the fourth quarter onwards, the neighborhoods with the highest number of cases were Ibura and Jordão. And in the west, there is an increase in arboviruses from the third quarter of 2014 to the fourth quarter of that same year. In the following two months, for this neighborhood, cases of arboviruses are constant, although they are still considered very high.

From the last two months of 2014 (Figure3f) to the first two months of 2015 (Figure4a), prediction using MLP with 30 neurons in the hidden layer showed a significant increase in arboviruses in the northern region. Recife. The affected neighborhoods were mainly Córrego do Jenipapo, Casa Amarela, Vasco da Gama, Nova Descoberta, and Dois Unidos. The southern zone showed a considerable decrease in arboviruses cases, having a very high incidence only in the neighborhood of Cohab. In the second and third bimesters, Figures 4b and 4c, respectively, the points with the highest incidence are the neighborhoods of the southern zone, especially Cohab, Ibura and Boa Viagem. In the west, there are also cases in the Várzea neighborhood and, more to the southwest of the city, in Jardim São Paulo as well. In the fourth bimester of 2015, the situation is more controlled, where cases were concentrated only in Ibura and Cohab (see Figure 4d).

Figure 3: Prediction results for regression using MLP with 30 hidden layer neurons for the year 2014.



(a) 1st bimester of 2014 (b) 2nd bimester of 2014 (c) 3rd bimester of 2014

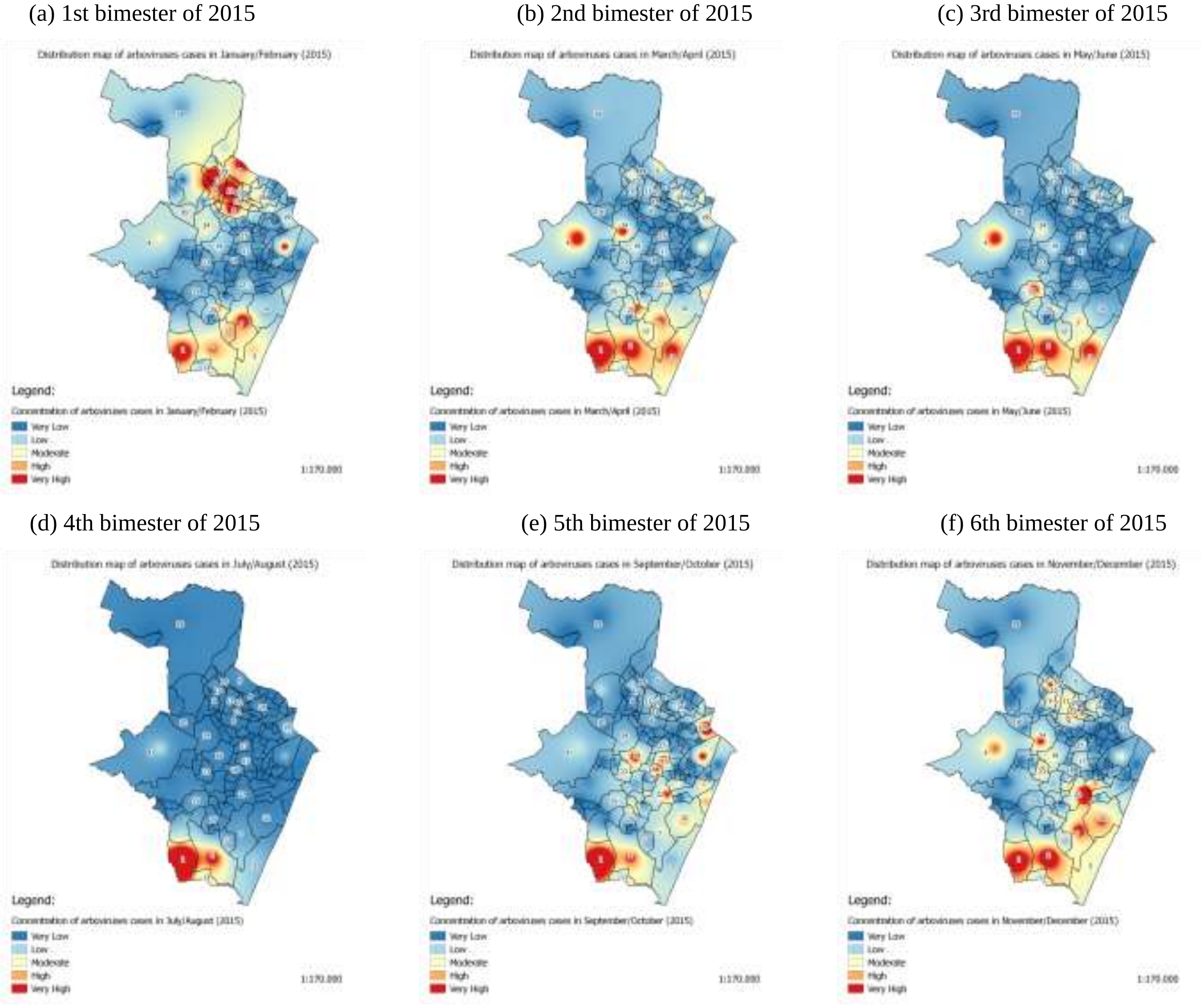
(d) 4th bimester of 2014 (e) 5th bimester of 2014 (f) 6th bimester of 2014

Distribution map of arboviruses cases in January/February (2014)
Distribution map of arboviruses cases in March/April (2014)
Distribution map of arboviruses cases in May/June (2014)
Distribution map of arboviruses cases in July/August (2014)
Distribution map of arboviruses cases in September/October (2014)
Distribution map of arboviruses cases in November/December (2014)
Legend:
Concentration of arboviruses cases in January/February (2014)
Very Low
Low
Moderate
High
Very High
1:170.000

Source: Authors.

During the 5th and 6th bimesters (Figures 4e and 4f, respectively), the situation worsens again. In the case of the 5th bimester, the neighborhoods located more in the center of the city had a significant increase in cases, especially the districts of Cordeiro, Madalena, and Prado. In the northeast of the city, the largest concentration of cases was in the neighborhood of Campo Grande. In the last two months of 2015, there was a considerable increase in cases of the northern region in neighborhoods such as Casa Amarela, Alto José Bonifácio, Nova Descoberta, Macaxeira, Vasco da Gama, and Morro da Conceição. In the southern zone, the neighborhoods of Ibura, Ipsep, Imbiribeira and Pina stand out towards Afogados. In the western neighborhoods of the city, the highest concentrations of cases occurred in the neighborhoods of Torrões and Iputinga.

Figure 4: Prediction results for regression using MLP with 30 hidden layer neurons for the year 2015.


(a) 1st bimester of 2015

(b) 2nd bimester of 2015

(c) 3rd bimester of 2015

(d) 4th bimester of 2015

(e) 5th bimester of 2015

(f) 6th bimester of 2015

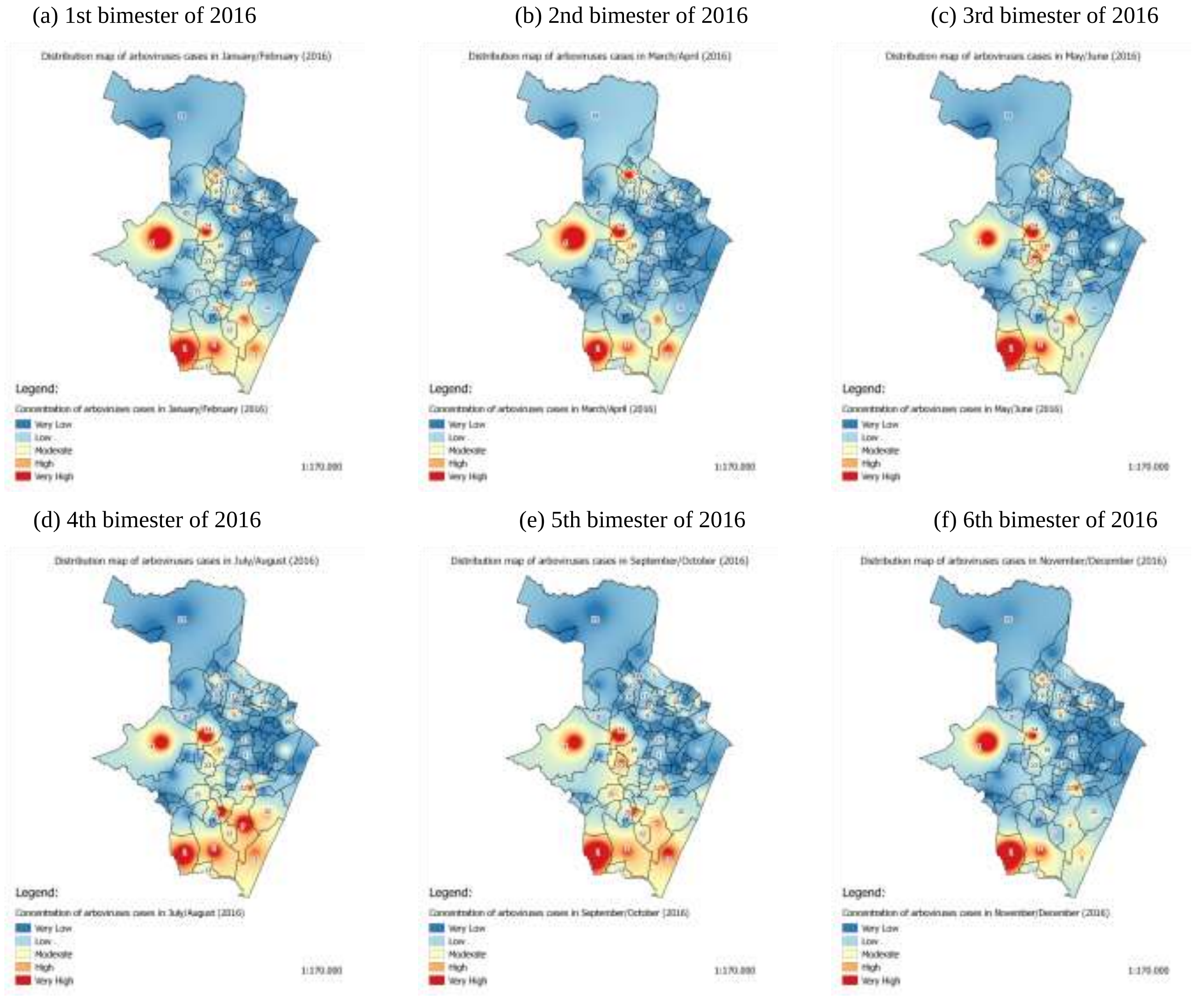
Distribution map of arboviruses cases in January/February (2015)
Legend:
Concentration of arboviruses cases in January/February (2015)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in March/April (2015)
Legend:
Concentration of arboviruses cases in March/April (2015)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in May/June (2015)
Legend:
Concentration of arboviruses cases in May/June (2015)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in July/August (2015)
Legend:
Concentration of arboviruses cases in July/August (2015)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in September/October (2015)
Legend:
Concentration of arboviruses cases in September/October (2015)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in November/December (2015)
Legend:
Concentration of arboviruses cases in November/December (2015)
Very Low
Low
Moderate
High
Very High
1:170.000

Source: Authors.

The predictions of the first two bimesters of 2016 (Figure 5) using MLP with 30 hidden layer neurons showed very similar behaviors. According to the images, the most affected regions in these two quarters were the south and west of the city. In the southern zone, the highest concentrations of cases occurred in Cohab and Ibura, with less intensity in the Ipsep, Boa Viagem and Imbiribeira neighborhoods. In the west of the city, the most affected neighborhoods were Várzea and Iputinga. In the third bimester of 2015 (Figure 5c), there was a considerable decrease in cases in the southern region of the city, focusing only on the neighborhoods of Cohab and Ibura. In the west of the city, it is also possible to notice a reduction of cases in the Várzea neighborhood. However, cases in the neighborhoods of Iputinga, Cordeiro and Torrões intensified. In the next two months (Figure 5d), the situation gets even worse. In the 4th bimester, the cases were mainly concentrated in the south towards the west of the city. The most affected neighborhoods were Cohab, Imbiribeira, Boa Viagem, Areias, Varzea, Iputinga, and Cordeiro.

In the fifth bimester of 2015 (Figure 5e), the behavior of the distribution of arbovirus cases was quite similar to the previous bimester. There is a decrease in cases in some neighborhoods of the southern region such as Ibura and Imbiribeira. In contrast, Boa Viagem showed a significant increase in the number of cases from the fourth to the fifth bimester. Comparing the map of Figure 5e with the map of Figure 5f, we can see that, from the southern region of the city towards the western region, the neighborhoods located in this region showed an increase in arbovirus cases. Finally, in the last two months of 2016 (Figure 5f), the prediction showed an improvement in the situation. The southern zone showed a significant reduction in arboviruses, except for the Cohab neighborhood. In the western region of the city, we can also notice a significant decrease in cases, except for Várzea and Iputinga.

Figure 5: Prediction results for regression using MLP with 30 hidden layer neurons for the year 2016.



(a) 1st bimester of 2016

(b) 2nd bimester of 2016

(c) 3rd bimester of 2016

(d) 4th bimester of 2016

(e) 5th bimester of 2016

(f) 6th bimester of 2016

Distribution map of arboviruses cases in January/February (2016)
Legend:
Concentration of arboviruses cases in January/February (2016)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in March/April (2016)
Legend:
Concentration of arboviruses cases in March/April (2016)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in May/June (2016)
Legend:
Concentration of arboviruses cases in May/June (2016)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in July/August (2016)
Legend:
Concentration of arboviruses cases in July/August (2016)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in September/October (2016)
Legend:
Concentration of arboviruses cases in September/October (2016)
Very Low
Low
Moderate
High
Very High
1:170.000
Distribution map of arboviruses cases in November/December (2016)
Legend:
Concentration of arboviruses cases in November/December (2016)
Very Low
Low
Moderate
High
Very High
1:170.000

Source: Authors.

Overall, the prediction maps using MLP with 30 neurons showed that the main regions of Recife with high concentration cases are the west and south regions. In the western region, the neighborhood that appears most frequently with regard to the highest concentration of cases is the Várzea neighborhood. In the southern region, the neighborhoods that appear most frequently at the highest concentration of cases are the neighborhoods of Cohab, Ibura, Imbiribeira, and Boa Viagem. The prediction maps presented are similar to the actual distribution maps of arbovirus cases. This corroborates the quantitative analysis metrics of the chosen regression method.

4. Conclusion

The use of machine learning predictors proved to be very effective in predicting the distribution of arbovirus cases. According to the qualitative results presented in the section 3, the regions in which arbovirus outbreaks transmitted by *Aedes aegypti* predominate are the southern and western regions of Recife. In the western region, the neighborhood that appears most frequently with regard to the highest concentration of cases is the Várzea neighborhood. In the southern region, the neighborhoods that appear most frequently at the highest concentration of cases are the neighborhoods of Cohab, Ibura, Imbiribeira and Boa Viagem. Already the northern region of the city appears with a high concentration of cases in the first two quarters of the year. Although there are cases throughout the year, it was also observed that arbovirus cases usually occur predominantly in the warmer months of the year (October to March).

Finally, the approach using spatio-temporal analysis provided a broader assessment of those regions where more or less arboviral outbreaks occur. From the qualitative results it was possible to differentiate in the heat maps the regions with very high concentration of cases from the regions with low concentration and the regions that are in the transition range. This type of approach is very relevant in supporting health managers and epidemiologists in the planning of short and medium term actions to prevent outbreaks of arboviruses transmitted by *Aedes aegypti*, and may also support the development of public policies for health promotion and sanitation.

As future work, we intend to evaluate new learning machines: statistical learning methods, random forests, classifier committees, meta-classifiers, and approaches based on hybrid architectures combining deep learning and linear regression methods. We also intend to expand the period observed until 2020, even considering that, in 2020, the data are underreported due to the Covid19 pandemic and the overload of the public health system in Recife. Additionally, we intend to use the proposed methodology to also predict in time and space the location of potential breeding sites for the *Aedes aegypti* mosquito.

Acknowledgments

We are grateful to the British Council and the Brazilian research agencies CNPq, CAPES and Facepe, for the partial support of this work.

References

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

Baquero, O. S., Santana, L. M. R., & Chiaravalloti-Neto, F. (2018, 04). Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. *PLOS ONE*, *13*(4), 1-12. https://doi.org/10.1371/journal.pone.0195065 10.1371/journal.pone.0195065

Barbosa, V. A. d. F., Gomes, J. C., de Santana, M. A., de Lima, C. L., Calado, R. B., Bertoldo Júnior, C. R., et al (2021). Covid-19 rapid test by combining a random forest-based web system and blood tests. *Journal of Biomolecular Structure and Dynamics*, *65*, 1–20.

Beltrán, J. D., Boscor, A., dos Santos, W. P., Massoni, T., & Kostkova, P. (2018). 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 *Proceedings of the 2018 international conference on digital health* (pp. 90–94).

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., et al (2013). The global distribution and burden of dengue. *Nature*, *496*(7446), 504–507.
Cao-Lormeau, V.-M., Blake, A., Mons, S., Lastère, S., Roche, C., Vanhomwegen, J., et al (2016). Guillain-Barré Syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study. *The Lancet*, *387*(10027), 1531–1539.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, *8*(1), 1–17.

Cordeiro, F. R., Santos, W. P., & Silva-Filho, A. G. (2016). A semi-supervised fuzzy growcut algorithm to segment and classify regions of interest of mammographic images. *Expert Systems with Applications*, *65*, 116–126.

da Silva, C. C., de Lima, C. L., da Silva, A. C. G., Silva, E. L., Marques, G. S., de Araújo, L. J. B., et al (2021). Covid-19 dynamic monitoring and real-time spatio-temporal forecasting. *Frontiers in Public Health*, *9*.

de Lima, C. L., da Silva, C. C., da Silva, A. C. G., Luiz Silva, E., Marques, G. S., de Araújo, L. J. B., et al (2020). COVID-SGIS: a smart tool for dynamic monitoring and temporal forecasting of Covid-19. *Frontiers in Public Health*, *8*, 761.

de Lima, S. M., da Silva-Filho, A. G., & dos Santos, W. P. (2016). Detection and classification of masses in mammographic images in a multi-kernel approach. *Computer methods and programs in biomedicine*, *134*, 11–29.

de Lima, T. F. M., Lana, R. M., de Senna Carneiro, T. G., Codeço, C. T., Machado, G. S., Ferreira, L. S., & Davis Junior, C. A. (2016). DengueME: A Tool for the Modeling and Simulation of Dengue Spatiotemporal Dynamics. *International Journal of Environmental Research and Public Health*, *13*(9), 920.

de Souza, R. G., dos Santos Lucas e Silva, G., dos Santos, W. P., & de Lima, M. E. (2021). Computer-aided diagnosis of Alzheimer’s disease by MRI analysis and evolutionary computing. *Research on Biomedical Engineering*, *37*, 455–-483.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, *20*(15), 2479–2481.

Gubler, D. J. (2011). Dengue, urbanization and globalization: the unholy trinity of the 21st century. *Tropical Medicine and Health*, *39*(4SUPPLEMENT), S3–S11.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Haykin, S. (2001). *Redes Neurais: Princípios e Prática*. Bookman.

Jindal, A., & Rao, S. (2017). Agent-based modeling and simulation of mosquito-borne disease transmission. In *Proceedings of the 16th conference on autonomous agents and multiagent systems* (pp. 426–435).

Koche, J. C. (2011). *Fundamentos de metodologia científica: teoria da ciência e iniciação à pesquisa*. Vozes.

Kostkova, P., dos Santos, W. P., & Massoni, T. L. (2019). ZIKA: improved surveillance and forecast of Zika virus in Brazil. *European Journal of Public Health*, *29*(Supplement 4), 414-415, ckz186.085.

LaDeau, S. L., Allan, B. F., Leisnham, P. T., & Levy, M. Z. (2015). The ecological foundations of transmission potential and vector-borne disease in urban landscapes. *Functional Ecology*, *29*(7), 889–901.

Laureano-Rosario, A. E., Duncan, A. P., Mendez-Lazaro, P. A., Garcia-Rejon, J. E., Gomez-Carro, S., Farfan-Ale, J., & MullerKarger, 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.

Ludke, M., & André, M. E. D. A. (2013). *Pesquisas em educação: uma abordagem qualitativa*. São Paulo, Brasil: EPU Editora Pedagógica e Universitária.

Mohammed, A., & Chadee, D. D. (2011). Effects of different temperature regimens on the development of aedes aegypti (l.)(diptera: Culicidae) mosquitoes. *Acta Tropica*, *119*(1), 38–43.

Morin, C. W., Comrie, A. C., & Ernst, K. (2013). Climate and dengue transmission: evidence and implications. *Environmental health perspectives*, *121*(11-12), 1264–1272.

Musah, A., Rubio-Solis, A., Birjovanu, G., dos Santos, W. P., Massoni, T., & Kostkova, P. (2019). Assessing the Relationship between various Climatic Risk Factors & Mosquito Abundance in Recife, Brazil. In *Proceedings of the 9th international conference on digital public health* (pp. 97–100).

Padmanabhan, P., Seshaiyer, P., & Castillo-Chavez, C. (2017). Mathematical modeling, analysis and simulation of the spread of zika with influence of sexual transmission and preventive measures. *Letters in Biomathematics*, *4*(1), 148–166.

Pereira, A. S., Shitsuka, D. M., Parreira, F. J., & Shitsuka, R. (2018). *Metodologia da pesquisa científica*. Santa Maria, Rio Grande do Sul, Brasil: Universidade Federal de Santa Maria.

Pereira, J., Santana, M. A., Gomes, J. C., de Freitas Barbosa, V. A., Valença, M. J. S., de Lima, S. M. L., & dos Santos, W. P. (2021). Feature selection based on dialectics to support breast cancer diagnosis using thermographic images. *Research on Biomedical Engineering*, *37*, 485–-506.

Pessanha, J. E. M., Caiaffa, W. T., César, C. C., & Proietti, F. A. (2009). Avaliação do plano nacional de controle da dengue. *Cad. Saúde Pública*, *25*(7), 1637–1641.

Rubio-Solis, A., Musah, A., dos Santos, W. P., Massoni, T., Birjovanu, G., & Kostkova, P. (2019). ZIKA Virus: Prediction of Aedes Mosquito Larvae Occurrence in Recife (Brazil) using Online Extreme Learning Machine and Neural Networks. In *Proceedings of the 9th international conference on digital public health* (pp. 101–110).

Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., et al (2012). Digital epidemiology. *PLoS Computational Biology*, *8*(7), e1002616.

Santana, M. A. d., Pereira, J. M. S., Silva, F. L. d., Lima, N. M. d., Sousa, F. N. d., Arruda, G. M. S. d., & Santos, W. P. d. (2018). Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Research on Biomedical Engineering*, *34*, 45–53.

Siriyasatien, P., Chadsuthi, S., Jampachaisri, K., & Kesorn, K. (2018). Dengue epidemics prediction: A survey of the state-ofthe-art based on data science processes. *IEEE Access*, *6*, 53757-53795.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Witten, I. H., & Frank, E. (2005). *Data mining: Pratical machine learning tools and technique*. Morgan Kaufmann Publishers.

Yin, R. K. (2015). *Estudo de caso: Planejamento e métodos*. Bookman.