Oxford Journal of Intelligent Decision and Data Science 2018 SI. 1 (2018) 1-13

Available online at www.ispacs.com/ojids
Volume 2018, Special Issue 1, Year 2018 Article ID ojids-00017, 13 Pages
doi:10.5899/2018/ojids-00017
Research Article

Forecasting Dengue Incidences: Statistical and Dynamic Models

Tru Hoang Cao[1,2*], Anh Duy Nguyen[2], Tuan Quang Dinh[3], Quang Chan Luong[4], Hai Thanh Diep[4]

(1) School of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam.
(2) John von Neumann Institute, Vietnam National University at Ho Chi Minh City, Vietnam.
(3) Department of Computer Sciences, University of Wisconsin at Madison, USA.
(4) Department of Disease Control and Prevention, Pasteur Institute at Ho Chi Minh City, Vietnam.

Abstract
Infectious disease forecasting, especially for dengue fever, has attracted many scientists and much of research effort for the past decade. Firstly, this paper surveys and presents typical approaches used for prediction of dengue fever. Secondly, we propose two methods for this task. One is a state space model recursively estimating the hidden states and making prediction one-step ahead. The other is a two-phase model simulating the disease transmission process that includes the local outbreak and then province transmission. Experiments are conducted on the two real datasets in Southern Vietnam and Singapore. They show improvement in prediction performance of our proposed methods in comparison with the typical existing ones for dengue disease.

Keywords: Artificial neural network, Hidden Markov model, Locality model, Transmission model, Mosquito characteristic, Epidemic.

1 Introduction

Dengue disease [1] is still a major concern of governments and scientists, especially in Asian countries like Vietnam, China and India. Predicting dengue incidences timely can help the authorities to prevent and suppress the disease, and save resources for intervention actions.

Production and hosting by ISPACS GmbH.

* Corresponding Author. Email address: tru.cao@jvn.edu.vn

There has been an increasing number of research works conducted for predicting infectious diseases, in particular dengue and malaria, for the past decade. Conventional methods, such as autoregressive integrated moving average (ARIMA) (e.g. [2]) or regression (e.g. [3]), which try to find a linear relationship between underlying factors and disease incidences, still obtain limited accuracy. Meanwhile, modern approaches have been applied in recent years to investigate dynamic characteristics of diseases, such as support vector machines (SVM) (e.g. [4]), artificial neural networks (ANN) (e.g. [5], [6]), and hidden Markov models (HMM) (e.g. [7]).

While there are works on long-term prediction, i.e., more than one month ahead, which could give us a picture of a disease development in next four months or a half-year (e.g. [8], [9]), in this paper we consider short-term periods ranging from one week to one month (e.g. [10]). In fact, during an epidemic period, the authority would take intervention to reduce and restrict the disease weekly or monthly. Therefore, a system that monitors and gives timely prediction every week could help a lot in real life.

The transmission of dengue fever is determined by many factors such as retrospective cases, climate, mosquito characteristics or human mobility. Each factor contributes differently to the spreading of the disease and many models solely focused on the few factors that strongly affect the transmission. In [2], [11], [3], and [5] the authors used reported incidences and climatic factors to build prediction models for dengue and malaria diseases. However, mosquito biological characteristics also play an important role in the transmission mechanism, which is the reason why mosquito factors were incorporated in the models in [12], [13], [6], and [7]. Human mobility also affects the transmission, and thus was taken into account in [14] and [6].

A survey to different approaches can give new insights to modeling of mosquito-borne disease transmission. In [16], the authors surveyed research on both long-term and short-term malaria forecasting. The authors also looked into models that forecast outbreaks and numbers of cases. For dengue fever, [13] compared different forecasting techniques for dengue fever. However, that work aimed to predict outbreaks of dengue, and thus considered only classification methods. In contrast, our work surveys and compares regression methods that predict actual numbers of dengue incidences.

In this paper, we focus on prediction of dengue fever in one-week period that takes into account the characteristics of the mosquito type causing the disease as well as climatic factors. The paper's purpose is two-fold. First, we propose two new methods for forecasting dengue incidences. Second, we survey different approaches to prediction of mosquito-infected incidences, in particular for both of dengue and malaria, and compare their performances with those of our methods when applied to dengue incidences on the real datasets in Southern Vietnam and Singapore. Here, to the best of our knowledge, our work is the first to predict for one week ahead the magnitude of dengue in Vietnam.

For the first method, we adapt the state space model for malaria in [7] to capture dengue transmission. The model is based on a hidden Markov chain, which can incorporate many features that affect the transmission, such as retrospective cases and environmental factors. We fit a time series into the model and learn the transmission properties through the hidden states. The hidden states are recursively calculated given the features, from which future values of the series can be forecasted.

Our second method, as preliminarily proposed in [15], adapts the malaria transmission phase in [6], considering inputs as both of the statistics of historical cases and the biological factors affecting the dengue virus, including the temperature, population, and mosquito density. We propose a two-phase model simulating the disease transmission process, which are the local outbreak and then province transmission. The locality phase estimates the number of potential cases in each province independently in the following week. Then, in the transmission phase, an artificial neural network is used to predict the mobility of the dengue virus across provinces.

The performance of our proposed methods is evaluated in comparison with the other surveyed methods on the same datasets of dengue fever in Singapore and six provinces of Southern Vietnam. The Root Mean Square Error (RMSE) is used as the benchmark measure. It shows that our forecasting methods, which take

into account environmental, biological, and demographic features, have smaller errors than the compared ones on the Southern Vietnam dataset, and come second on the Singapore dataset.
The rest of this paper is organized as follows. Section 2 reviews related works. Sections 3 and 4 present details of the proposed state space model and two-phase model, respectively. Performance evaluation in comparison with other methods is presented in Section 5. Finally, Section 6 draws some concluding remarks.

2 Related Works

The nature of forecasting methodology is constructing a relationship between forecasted values and underlying factors from the past data, and then extending it for future data. For the forecasting problem of mosquito borne diseases, there are two main approaches, namely, statistical methods and dynamic modeling [7].

2.1. Statistical Methods

Many statistical methods have been proposed to predict disease incidences based on time series of historical surveillance data, such as ARIMA and Poisson multivariate regression. ARIMA comprises two parts, namely, autoregressive (AR) and moving average (MR). The AR part estimates a future value as a weighted sum of past values, while the MR part is to smooth noises. In [17] and [18], the authors used malaria past cases and meteorological data (temperature, rainfall) to predict malaria incidences in Bhutan and Uganda.
Furthermore, ARIMA can also model a wide range of seasonal data. The Seasonal ARIMA model (SARIMA) adds seasonal terms to the ARIMA model. Dengue and malaria are seasonal fevers, and thus SARIMA is more effective than non-seasonal ARIMA for predicting them. The work [2] modeled monthly malaria cases by using SARIMA on malaria data and rainfall data in Sri Lanka. Meanwhile, [19] applied the model to forecast weekly dengue incidences in Guadeloupe. Since ARIMA alone does not bring high accuracy, its combination with other methods have been experimented. Recently, [11] employed ANN to adjust residuals of prediction results obtained by SARIMA, which gave higher accuracy in monthly forecasting of hand-foot-mouth disease cases in China.
While ARIMA tries to model a variable based only on the past values of the same variable, other regression models also use values of other variables. The principal part of this approach is modeling disease distribution patterns based on retrospective data, and then using the best one for forecasting. As a recent work predicting dengue incidences in Singapore, [3] calculated this distribution by Poisson multivariate regression with five independent variables, namely, the serial correlation of dengue cases, meteorological data cycles, seasons, epidemic cycles, and trends. Similarly, [20] predicted malaria incidences in China, while [21], [22], and [23] predicted dengue incidences in Bangladesh, India, and Taiwan, respectively.
However, as remarked in [7], those statistical methods only study the statistical patterns of past data, which prevent them from understanding the dynamic characteristics of the disease transmission.

2.2. Dynamic Modeling

Recent works in forecasting of disease incidences have focused on individual-based models. That is to model transmission of the disease among individuals in a population and across provinces, based on surveillance. Machine learning methods, such as ANN, SVM, and HMM, could be employed to learn such a model, which is then used to forecast infected cases.
The ANN-Time Series method (ANN-TS) has been successfully applied to forecast time series data. It can approximate any nonlinear and continuous function without prior information about the properties of the data series. Input data are the numbers of infected cases at some time before the prediction period, whereas the output is the forecasted number of infected cases in the prediction period. ANN-TS was employed in [5] for prediction of dengue incidences in Thailand, using a data set of historical dengue cases through 30 years.

SVM is one of the most used machine learning methods for various problems in recent years. In [13], the authors applied pure SVM to predict dengue incidences in central Thailand. Previously, [4] used a combination of SVM and Firefly Algorithm (FFA) for forecasting malaria transmission in India, where FFA as a robust meta-heuristic search algorithm was used to estimate and optimize SVM parameters.

Meanwhile, [24] showed that the percentage of current infected mosquitos in the population has a strong impact to the development of a malaria epidemic later on. The work also considered the vectorial capacity, which is computed by the current status of the mosquitos, population, temperature, and some biological factors of both viruses and mosquitos, as a strong clue for determining how bad the epidemic could be in the next period. Additionally, it also introduced the entomological inoculation rate, which refers to the probability a person could be infected. By estimating that number, one can predict the potential of an epidemic in the following days.

Temperature is one of the most effective factors in a prediction model for dengue magnitude. In [25], the authors showed that a moderate fluctuation of temperature in daylight (diurnal temperature range - DTR) could result in a higher probability of mosquito survival, hence increasing the chance of infection to the community; otherwise a large temperature change could reduce the impact of the virus-carrying mosquito Aedes Aegypti, which causes dengue. Similarly, [26] proved a strong correlation between the temperature and the effect of mosquitos in dengue. It revealed that the best condition for mosquitos to grow was approximately 29 °C by the mean temperature, and the lower DTR, the better the condition was for mosquitos to spread the disease. That range of temperature is usual in a tropical country like Vietnam.

In [6], for determining the transmission rate of malaria among towns in Yunnan, China, the authors proposed a two-phase model, including the locality and the transmission ones. The locality phase used a modified Ross MacDonald model [27] to estimate a potential number of malaria cases in each province, based on the temperature, mosquito density, human population, and historical cases. Then the transmission phase used ANN to model a possible network of moving people that could reasonably find the hidden pattern of malaria transmission. Although focusing on simulation of a disease transmission network, the model could be adapted to predict the magnitude of a disease using biological and demographic factors.

Meanwhile, [7] proposed a state space model that used retrospective cases and temperature to estimate hidden states of a Markov chain, i.e., proportions of weekly infected cases. After inferring the hidden states, prediction of the next week's number of malaria incidences was calculated. This model can not only predict potential disease incidences, but also help to understand the transmission dynamics through parameter learning.

3 Proposed State Space Model

3.1. Weekly Infected Proportion Function

In order to establish a relation between future incidences, reported cases, and mosquito characteristics, we first employ the following function proposed in [7]:

$$x_k = f(x_{k-1}) = -bcV_{k-1}x_{k-1}^2 + (1 - r + bcV_{k-1})x_{k-1} + \varepsilon \quad (3.1)$$

where x_k denotes the proportion of infected population at time k, b and c are probabilistic values related to the infection mechanism, r is the human recovery rate, ε is the prediction noise, and V_{k-1} is the vectorial capacity value at time $k-1$.

Vectorial capacity (VCAP), denoted by V_k, is the number of newly infected cases caused by one single case in a day, which represents the infectious ability of the disease. For malaria, [6] adapted Ross-MacDonald model [27] with specific parameters for Anopheles mosquito to compute VCAP. For dengue, whose main vector is the Aedes mosquito, Ross-Macdonald model in [26] provides the following equation to compute the infectious ability:

$$V_k = m_k.a_k^2.e^{-\mu_k n_k} / \mu_k \quad (3.2)$$

where m_k is the ratio of female mosquitos to human population, a_k is the mosquito biting rate, n_k is the virus incubation period, and μ_k is the mosquito mortal rate.

The work [26] also provided formulas to compute other parameters as functions of temperature T_k at time k as follows:

- Mosquito biting rate:

$$a_k = 0.0043T_k + 0.0943 \quad (21 \le T_k \le 32) \tag{3.3}$$

- Incubation period of dengue virus:

$$n_k = 4 + e^{5.15 - 0.123T_k} \quad (12 \le T_k \le 36) \tag{3.4}$$

- Mosquito mortal rate:

$$\mu_k = 0.8692 - 0.1590T_k + 0.01116T_k^2 - 3.408*10^{-4}T_k^3 + 3.809*10^{-6}T_k^4 \quad (10.54 \le T_k \le 33.41) \tag{3.5}$$

The parameters b, c and r in the equations above are estimated as presented in Section 3.3. Furthermore, we note that (3.1) calculates the infected proportion on a daily basis due to the used daily VCAP, where the remaining infections, newly infected and recovered ones are accounted for. Therefore, the proportion of infections of a whole week is equal to the proportion of the last day of that week.

Let $z_{[i,k]}$ denote the infected proportion of the ith day of week k, the infected proportion of week k is:

$$x_k = z_{[7,k]} \tag{3.6}$$

Hence:

$$x_{k+1} = z_{[7,k+1]} = f^7(z_{[7,k]}) = f^7(x_k) \tag{3.7}$$

Thus, for the purpose of one-week ahead prediction, we derive a new function called WIP (Weekly Infected Proportion) to calculate weekly proportion of infections as follows:

$$x_k = F(x_{k-1}) = f^7(x_{k-1}) \tag{3.8}$$

This function is used for both of the two proposed models.

3.2. Constructing the Model

We use a first order Markov chain to represent the disease transmission. Let y_k be the observed number of weekly infections at time k, x_k be the corresponding hidden state representing the proportion of infections at time k. Given the observed cases $y_1, y_2, \dots, y_T$, the model aims to predict the number of incidences y_{T+1} through inferring the hidden states.

To make one-step ahead prediction, the hidden states $x_1, x_2, \dots, x_T, x_{T+1}$ are calculated recursively. Then the prediction y_{T+1} can be obtained based on x_{T+1}.

In [7] the authors established the state transition function as presented in (3.1), and the observation function as in (3.9) below:

$$y_k = g(x_k) = hx_k + \eta \tag{3.9}$$

where x_k and y_k respectively denote the hidden state and observed value at time k, and h reflects the linear relationship between hidden states and observed cases. The prediction and observation noises are assumed to be Gaussian, i.e., $\varepsilon \sim N(0, Q)$ and $\eta \sim N(0, R)$.

However, those equations in [14] were to predict daily infected proportions only. To adapt the model for weekly prediction of dengue incidences, we replace (3.1) with our derived WIP function in (3.8).

The state space model operates in three main steps: (1) inferring the hidden states (filtering); (2) smoothing the hidden states; and (3) estimating the parameters (Fig. 1). Firstly, the data points of T consecutive weeks are received as inputs to compute the hidden states using a filtering method. Each data point consists of the number of reported cases and the average temperature of the corresponding week. The hidden states are then

smoothed with a smoother. Finally, the model parameters are estimated and the output of the model is the predicted incidences at week $T + 1$.

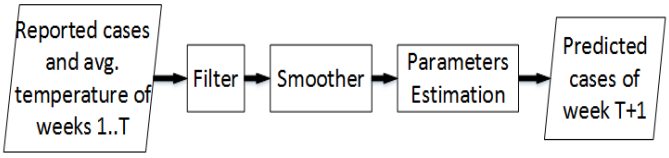

Figure 1: The main steps of the proposed state space model

3.3. Implementing the Model

a) Filtering and Smoothing

We assume the hidden states have the Gaussian distribution $x \sim N(\mu, P)$. The goal of this step is to calculate the expectation and variance of each hidden state. To estimate the states, we first use a filtering method to calculate the expectations of x_k, and then a smoother is applied to refine the estimation. We use the Extended Kalman Filter (EKF) algorithm [28] for the first step. The EKF receives the weekly incidences $y_1, y_2, \dots, y_T$ as inputs and calculates the expectations μ_k and variances P_k of the hidden states recursively, for $k = 1,2, \dots, T$. The hidden state at time $k = 0$ is initialized, i.e., $x_0 \sim N(\mu_0, P_0)$, and is part of the paramaters to be estimated.

The outputs of the filtering step $x_k \sim N(\mu_k, P_k)$ then become the inputs of the smoothing step. We employ the Rauch-Tung-Striebel (RTS) algorithm [29] as the smoother. The outputs of the smoothing step are the smoothed estimations of the hidden states, denoted by $x_{k|T} \sim N(\mu_{k|T}, P_{k|T})$.

b) Estimating the Parameters

The parameters that determine the disease transmission are not known. So we have to estimate them before we can use the model for prediction. Let $\theta = \{\mu_0, P_0, bc, r, h, Q, R\}$ denote the set of parameters. The values of θ are selected such that the likelihood $p(y_{1..T}|\theta)$ is maximized. With that objective, we use the Expectation Maximization (EM) algorithm to estimate θ.

In the E-step, [29] suggests that the objective function Q is estimated by:

$$\begin{aligned} Q(\theta, \theta_n) \approx & -0.5 \log 2\pi P_0 - 0.5T \log 2\pi Q - 0.5T \log 2\pi R \\ & - 0.5 tr\{P_0^{-1}(P_{0|T} + (\mu_{0|T} - \mu_0)(\mu_{0|T} - \mu_0)^T)\} \\ & - 0.5 \sum_{k=1}^{T} tr\{Q^{-1}[(x_{k|T} - F(x_{k-1|T}, \theta))(x_{k|T} - F(x_{k-1|T}, \theta))^T]\} \\ & - 0.5 \sum_{k=1}^{T} tr\{R^{-1}[(y_k - g(x_{k|T}, \theta))(y_k - g(x_{k|T}, \theta))^T]\} \end{aligned} \quad (3.10)$$

where the smoothed hidden states $x_{k|T}$ are calculated according to θ_n and F is the WIP function (3.8). In the M-step, the objective function is maximized according to θ. Since the parameters are constrained, we use L-BFGS [29] as an optimization method.

4 Proposed Two-Phase Model

Based on the model of malaria transmission in [6], we propose a new modified dengue-based version. Figure 2 below describes the architecture of our model:

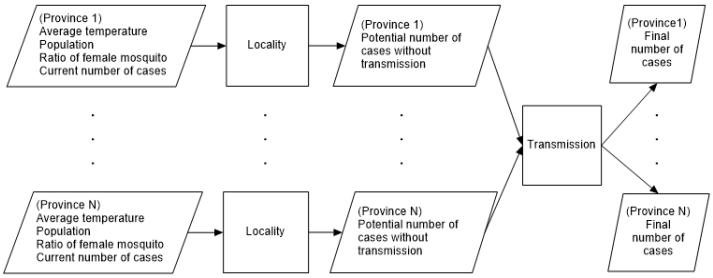

Figure 2: Proposed two-phase model for forecasting dengue incidences

As seen from above, our model consists of two consecutive phases: locality and transmission. At the beginning, the locality model (the first left rectangles) is applied separately to each province. The goal of this phase is to predict the number of cases happening in each assumed-isolated region using its own factor, which includes four inputs in the left parallelograms. The output of this phase (the middle parallelograms) is the number of disease cases estimated in each province. We make use of the result obtained from recent dengue research in [26] to extend the original locality model of malaria to a new one embedding dengue's characteristics. Using all of these outputs, the transmission model (the right rectangle), in the following phase, then calculates the final number of cases (the right parallelograms) in each province in the following week. For this purpose of prediction, a new neural network is proposed in our transmission phase.

4.1. The Locality Phase

The locality phase aims to predict the number of cases locally. It receives the average temperature, population, and current number of infections in a week as inputs, and then outputs the predicted number of incidences for the next week. In [6], the authors used the Entomological Inoculation Rate (EIR) for this step, which is defined as the number of infectious bites each person receives per day. With β being the intervention rate, P_{k-1} the population, and h_{k-1} the probability of an infected mosquito transmitting the virus to an uninfected person, [6] calculated the number of infected cases as follows:

$$\delta_k = \beta.P_{k-1}.h_{k-1}.EIR_{k-1} \tag{4.11}$$

However, since our work performs prediction for one week ahead and uses weekly aggregated data, we replace EIR with our derived WIP function (3.8). The potential number of infected cases at time k is thus calculated as follows:

$$\delta_k = \beta.P_k.x_k = \beta.P_k.F(x_{k-1}) \tag{4.12}$$

where δ_k is the predicted number of infected cases at time k, and x_k and x_{k-1} are the infected proportions of the population at time k and $k-1$, respectively.

4.2. The Transmission Phase

We build a 2-layer neural network to imitate the disease spreading process with respect to the number of provinces. Fig. 3 shows the architecture of our transmission network where each node represents a province of 6 provinces in a same order for all layers. Inputs are locally estimated values transferred directly from the locality phase. The output layer gives us the final predicted numbers of disease after spreading in the cycle of one week. An important modification is the intervention factor of the authority which is weighted by coefficient β of connections between the input and the first hidden layer. This parameter is motivated from the intervention of the local authority to tackle a dengue outbreak when it occurs. As a significant difference from [6] where connections are established only between adjacent provinces, out network has two fully connected hidden layers simulating the spreading of dengue virus among provinces. Weight matrix W, therefore, represents transmission rates between every pair of provinces. This modification comes from the fact that infected humans can carry dengue virus during their travel.

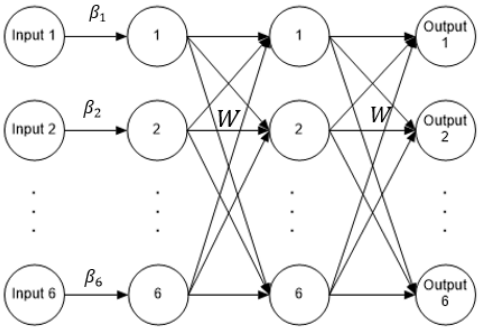

Figure 2: Structure of the proposed neural network

In addition, we use sigmoid as activation function instead of linear function in [6] due to our prediction purpose:

$$g(x) = 1/(1+e^{-x}) \tag{4.13}$$

Predicted output vector y then has the formula:

$$y = g((W')^d .diag(\beta).x) \tag{4.14}$$

where W' denotes transpose of matrix W, $diag(\beta)$ is diagonal matrix of β, and x is input vector.
Popular Back-Propagation algorithm [30] and loss function RMSE are used in training:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \tag{4.15}$$

where y_i is a predicted value, y'_i is the corresponding actual value, and n is the dataset size.

5 Evaluation

5.1. Data

We acquire dengue surveillance data of six provinces in Southern Vietnam, namely, Binh Duong, Ba Ria-Vung Tau, Dong Nai, Long An, Tay Ninh, and Ho Chi Minh City. Relative positions of these provinces can be seen in Fig. 4. The data are provided by the Dengue Center at Pasteur Institute in Ho Chi Minh City. The dengue incidences are collected weekly, spanning a period of 10 years, from 2004 to 2013.
We also collect dengue surveillance data in Singapore to see if the model can be adapted to a dengue epidemic in a different country. The data range from 2011 to 2015, have weekly intervals, and are made available to public on the official website of Singapore Ministry of Health. Since the area of Singapore is small as well as its population, we do not divide the country into smaller regions.

Figure 3: Administrative boundaries of the six provinces in Southern Vietnam

To collect the surface air temperatures for the above locations, we use Moderate Resolution Imaging Spectroradiometer (MODIS) [31]. After preprocessing, we calculate weekly average mean temperatures from the 8-day interval temperature data collected from MODIS.

The dengue surveillance and air temperature data are combined to create weekly data points (522 weeks for Southern Vietnam, 261 weeks for Singapore). For the proposed two-phase model, we employ 5-fold cross validation method to train and assess the model. For the proposed state space model, we use the first 80% of the data as the initial training set and use the last 20% for testing.

5.2. Experiments

We use the Root Mean Squared Error (RMSE) to measure the prediction accuracy and compare our proposed methods with six methods mentioned in Section 2, which are:

a) Seasonal ARIMA (SARIMA) [2]

This model takes inputs as dengue cases and meteorological variables (rainfall, temperature and relative humidity). Using clinical suspected cases, we feed the SARIMA model with dengue surveillance data, and then use it to calculate dengue incidences.

b) Seasonal ARIMA-Artificial Neural Network (SARIMA-ANN) [11]

The model receives retrospective cases as inputs, applies the ARIMA model to predict incidences, and finally uses the ANN to adjust residuals series.

c) Poisson Regression [3]

This model takes inputs as mean temperatures and cumulative rainfalls, and is built through 3 steps, namely, model construction, model selection, and forecasting. Distributions are calculated by the Poison multivariate regression from independent variables.

d) Artificial Neural Network-Time Series (ANN-TS) [5]

The model receives the number of infected cases in 4 weeks prior to the prediction week as inputs. A multi-layer feed forward back propagation neural network with the 4-17-1 architecture is implemented.

e) Support Vector Machine-Grid Search (SVM-GS) [4]

The inputs of this model are the dengue cases, precipitation, temperature, and humidity. SVM is implemented using the Radial Basis as a kernel function to solve the regression problem, and the Grid Search method is employed to optimize the SVM parameters.

f) Shi et al.'s Method [6]

This network model is re-implemented by replacing the malaria locality phase therein by the dengue one.

All the experiments are conducted on the two datasets for Southern Vietnam and Singapore. For the Southern Vietnam dataset, we calculate average RMSEs of the six provinces. The Shi et al.'s and two-phase methods cannot be applied to the Singapore dataset, because these methods require multiple regions for network training. The comparison of the results is shown in Table I.

Table I: RMSEs of the compared methods

Method	Average of Southern VN provinces	Singapore
SARIMA	0.067	0.072
SARIMA – ANN	0.060	0.061
Poisson regression	0.035	0.066
ANN – TS	0.024	0.040
SVM – GS	0.027	0.054
Shi et al.'s	0.024	N/A
Proposed two-phase	0.021	N/A
Proposed state space	0.020	0.043

As shown in the table, for the Southern Vietnam dataset, our proposed state space model achieves the best result, and the proposed two-phase model comes as the second. Meanwhile, for the Singapore dataset, our proposed state space method outperforms all other methods except for the ANN Time Series, which is slightly better. Overall, Seasonal ARIMA and Seasonal ARIMA - ANN have the lowest accuracies. Also, the recent approaches that use machine learning perform better than the traditional ones.

6 Conclusion

In this work, we apply several models to make one-step ahead prediction of dengue incidences, evaluated on the datasets in Southern Vietnam's six provinces and Singapore. Some of the selected methods are traditional and the rest are more recent ones that use machine learning.

We also propose two models, namely, the two-phase and the state space ones, to capture the disease transmission and make predictions for short-term dengue incidences. For our proposed models, we develop the WIP function, which uses Aedes mosquito characteristics to calculate weekly dengue-infected proportions.

The conducted experiments have shown that our proposed models perform the best of all compared methods for the Southern Vietnam provinces, while the proposed state space model is just slightly behind the top ANN-TS for Singapore. The results have also shown that machine learning methods generally perform better than traditional statistical ones, and mosquito related features help improve the prediction accuracy. For future work, we suggest to incorporate more features into the models to perform prediction more accurately and further in time.

References

[1] R. L. Felissa, D. D. Jerry, Emerging Infectious Diseases: Trends and Issues, 2nd ed, Springer, (2007), Part II, Chapter 7.
https://doi.org/10.3201/eid1402.071424

[2] O. J. T. Briët, P. Vounatsou, D. M. Gunawardena, G. N. L. Galappaththy, P. H. Amerasinghe, Models for short term malaria prediction in Sri Lanka, Malaria Journal, 7 (76) (2008).
https://doi.org/10.1186/1475-2875-7-76

[3] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, J. Rocklöv, Forecast of dengue incidence using temperature and rainfall, PLoS Neglected Tropical Diseases, 6 (11) (2012).
https://doi.org/10.1371/journal.pntd.0001908

[4] C. Sudheer, S. K. Sohania, D. Kumara, A. Malikc, B. R. Chaharc, A. K. Nemac, B. K. Panigrahib, R. C. Dhiman, A support vector machine-Firefly algorithm based forecasting model to determine malaria transmission, Neurocomputing, 129 (2014) 279-288.
https://doi.org/10.1016/j.neucom.2013.09.030

[5] S. Wongkoon, M. Jaroensutasinee, K. Jaroensutasinee, Development of temporal modeling for prediction of dengue infection in Northeastern Thailand, Asian Pacific Journal of Tropical Medicine, 5 (3) (2012) 249-252.
https://doi.org/10.1016/S1995-7645(12)60034-0

[6] B. Shi, J. Liu, X-N. Zhou, G-J. Yang, Inferring Plasmodium vivax transmission networks from tempo-spatial surveillance data, PLoS Neglected Tropical Diseases, 8 (2) (2014).
https://doi.org/10.1371/journal.pntd.0002682

[7] F. Liang, B. Shi, J. Liu, L. Chen, Forecasting disease incidences using state space model, in Proceedings of the 2nd International Workshop on Pattern Recognition for Healthcare Analytics, ICPR, Stockholm, Sweden, (2014).

[8] M. C. Wimberly, T. W. Chuang, G. M. Henebry, Y. Liu, A. Midekisa, P. Semuniguse, G. Senay, A computer system for forecasting malaria epidemic risk using remotely sensed environmental data, in Proceedings of iEMSs, Leipzig, Germany, (2012) 482-489.

[9] Y. L. Hii, J. Rocklöv, S. Wall, L. C. Ng, C. S. Tang, N. Ng, Optimal lead time for dengue forecast, PLoS Neglected Tropical Diseases, 6 (10) (2012).
https://doi.org/10.1371/journal.pntd.0001848

[10] P. Dayama, S. Kameshwaran, Predicting the dengue incidence in Singapore using univariate time series models, in Proceedings of AMIA Annual Symposium, Washington DC, (2013) 285-292.
https://www.ncbi.nlm.nih.gov/pubmed/24551338

[11] L. Yu, L. Zhou, L. Tan, H. Jiang, Y. Wang, S. Wei, S. Nie, Application of a new hybrid model with seasonal auto-regressive integrated moving average (ARIMA) and nonlinear auto-regressive neural network (NARNN) in forecasting incidence cases of HFMD in Shenzhen, China, PLoS ONE, 9 (6) (2014).
https://doi.org/10.1371/journal.pone.0098241

[12] R. C. P. K. Srimath-Tirumula-Peddinti, N. R. R. Neelapu, N. Sidagam, Association of climatic variability, vector population and malarial disease in district of Visakhapatnam, India: A modeling and prediction analysis, PLoS ONE, 10 (6) (2015).
https://doi.org/10.1371/journal.pone.0128377

[13] K. Kesorn, P. Ongruk, J. Chompoosri, A. Phumee, U. Thavara, A. Tawatsin, P. Siriyasatien, Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas, PLoS ONE, 10 (5) (2015).
https://doi.org/10.1371/journal.pone.0125049

[14] S. Sang, S. Gu, P. Bi, W. Yang, Z. Yang, L. Xu, J. Yang, X. Liu, T. Jiang, H. Wu, C. Chu, Q. Liu, Predicting unprecedented dengue outbreak using imported cases and climatic factors in Guangzhou, PLoS Neglected Tropical Diseases, 9 (5) (2015).
https://doi.org/10.1371/journal.pntd.0003808

[15] T. Q. Dinh , H. V. Le, T. H. Cao, Q. C. Luong, H. T. Diep, Forecasting the magnitude of dengue in Southern Vietnam, in Proceedings of the 8th Asian Conference on Intelligent Information and Database Systems, Da Nang, Vietnam, Springer, LNCS 9621, (2016) 554-563.
https://doi.org/10.1007/978-3-662-49381-6_53

[16] K. Zinszer, A. D. Verma, K. Charland, T. F. Brewer, J. S. Brownstein, Z. Sun, D. L. Buckeridge, A scoping review of malaria forecasting: past work and future directions, BMJ Open, 2 (6) (2012).
https://doi.org/10.1136/bmjopen-2012-001992

[17] K. Zinszer, R. Kigozi, K. Charland, G. Dorsey, T. F. Brewer, J. S. Brownstein, M. R. Kamya, D. L. Buckeridge, Forecasting malaria in a highly endemic country using environmental and clinical predictors, Malaria Journal, 14 (245) (2015).
https://doi.org/10.1186/s12936-015-0758-4

[18] K. Wangdi, P. Singhasivanon, T. Silawan, S. Lawpoolsri, N. J. White, J. Kaewkungwal, Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan, Malaria Journal, 9 (251) (2010).
https://doi.org/10.1186/1475-2875-9-251

[19] M. Gharbi, P. Quenel, J. Gustave, S. Cassadou, G. L. Ruche, L. Girdary, L. Marrama, Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors, BMC Infectious Diseases, 11 (166) (2011).
https://doi.org/10.1186/1471-2334-11-166

[20] C. Chatterjee, R. R. Sarkar, Multi-step polynomial regression method to model and forecast malaria incidence, PLoS ONE, 4 (3) (2009).
https://doi.org/10.1371/journal.pone.0004726

[21] M. N. Karim, S. U. Munshi, N. Anwar, M. S. Alam, Climatic factors influencing dengue cases in Dhaka City: a model for dengue prediction, Indian Journal of Medical Research, 136 (1) (2012) 32-39.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3461715/

[22] F-S. Chang, Y-T. Tseng, P-S. Hsu, C-D. Chen, I-B. Lian, D-Y. Chao, Re-assess vector indices threshold as an early warning tool for predicting dengue epidemic in a dengue non-endemic country, PLoS Neglected Tropical Diseases, 9 (9) (2015).
https://doi.org/10.1371/journal.pntd.0004043

[23] Y. Zhang, T. Wang, K. Liu, Y. Xia, Y. Lu, Q. Jing, Z. Yang, W. Hu, J. Lu, Developing a time series predictive model for dengue in Zhongshan, China based on weather and Guangzhou dengue surveillance data, PLoS Neglected Tropical Diseases, 10 (2) (2016).
https://doi.org/10.1371/journal.pntd.0004473

[24] L. S. David, F. E. McKenzie, Statics and dynamics of malaria infection in Anopheles mosquitos, Malaria Journal, 3 (13) (2004).
https://doi.org/10.1186/1475-2875-3-13

[25] L. Louis, K. P. Paaijmans, T. Fansiri, L. B. Carrington, L. D. Kramer, M. B. Thomas, T. W. Scott, Impact of daily temperature fluctuations on dengue virus transmission by Aedes Aegypti, Proceedings of the National Academy of Sciences of the United States of America, 108 (18) (2011) 7460-7465.
https://doi.org/10.1073/pnas.1101377108

[26] J. Liu-Helmersson, H. Stenlund, A. Wilder-Smith, J. Rocklöv, Vectorial capacity of Aedes Aegypti: effects of temperature and implications for global dengue epidemic potential, PLoS ONE, 9 (3) (2014).
https://doi.org/10.1371/journal.pone.0089783

[27] L. S. David, K. E. Battle, S. I. Hay, C. M. Barker, T. W. Scott, F. E. McKenzie, Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens, PLoS Pathogens, 8 (4) (2012).
https://doi.org/10.1371/journal.ppat.1002588

[28] M. Briers, A. Doucet, S. Maskell, Smoothing algorithms for state-space models, Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.498, (2004).

[29] J. Kokkala, A. Solin, S. Särkkä, Expectation maximization based parameter estimation by sigma-point and particle smoothing, in Proceedings of Information Fusion, Salamanca, (2014) 1-8.
http://ieeexplore.ieee.org/document/6916073/

[30] N. Michael, Neural Networks and Deep Learning, Determination Press, (2015).

[31] Z. Wan, Temperature, Emissivity, Aqua MODIS, (2015).