AN EFFICIENT METHOD TO PREDICT DENGUE OUTBREAKS IN KUALA LUMPUR

Duc Nghia Pham¹, Tarique Aziz¹, Ali Kohan¹, Syahrul Nellis², Juraina binti Abd. Jamil³,
Jing Jing Khoo², Dickson Lukose¹, Sazaly bin Abu Bakar² and Abdul Sattar³

¹ MIMOS Berhad, Malaysia
² TIDREC, University of Malaya, Malaysia
³ IIIS, Griffith University, Australia

ABSTRACT

In recent years, there has been a surge in dengue outbreaks in Malaysia. A dengue outbreak can cause severe damages to the society. Hence, it is critical to be able to predict a dengue outbreak in advance to minimize the damage and loss. In this paper, we propose a new machine learning approach to predict the number of dengue cases in Kuala Lumpur, in particular the areas surrounding the University of Malaya (UM) Medical Centre. We identified several different factors that can contribute to a surge in the number of dengue cases that occurred near the UM Medical Centre. Apart from the daily mean temperature and daily rainfall factors that have been frequently used in other studies, we also considered the enhanced vegetation index (EVI) as an input factor to our prediction engine. We trained our linear regression model on these three factors against the number of dengue cases from 2001 to 2010. We then tested our model on the 2011 data. The experimental results showed that our approach was able to predict the number of dengue cases 16 days in advance with high accuracy.

Keywords: Dengue Outbreak Prediction, Linear Regression Model.

1. Introduction

Today, dengue ranks as the most important mosquito-borne viral disease in the world (WHO, 2015). Dengue is transmitted via the bite of infected Aedes mosquitoes, typically Aedes aegypti and Ae. Albopictus. These mosquitoes have adapted to local human habitation with oviposition and larval habitats both in natural (e.g., rock pools, tree holes and leaf axis) and artificial (e.g., water tanks, blocked drains, pot plants and food and beverage containers) uncovered containers holding rain water in the urban and peri-urban environment (Focks, Daniels, Haile & Keesling, 1995).

There are four known serotypes of the virus that causes dengue fever (DF): DENV 1 – 4. After being bitten by an infected Aedes mosquito, a patient normally experiences an abrupt onset of high fever as well as erythema, severe muscle and joint pain, headache, nausea, and vomiting (Rigau-Perez et al., 1998). Furthermore, patients may face the life-threatening dengue hemorrhagic fever (DHF) if they get re-infected with a different viral serotype (Avirutnan et al., 2006). Patients with severe cases of DHF may also experience dengue shock syndrome (DSS) (Rajapakse, 2011). Children with DENV infection are particularly at risk of progressing to DHF or DSS. Mortality in DHF has been reported as high as 10–20% and over 40% if DSS occurs (Gibbons & Vaughn, 2002). Unfortunately, there has been no specific treatment or vaccine for dengue. People are currently rely on dengue vector (i.e. Aedes mosquitoes) control to prevent outbreak of dengue cases.

Thus, dengue fever has been a major public health concern for over half of the world’s population and is a leading cause of hospitalization and death, particularly for children in endemic countries.
Most of the resource-poor countries are particularly vulnerable to transmission of dengue disease, and it is present in urban and suburban areas in the Americas, eastern Mediterranean, western Pacific, South-East Asia and mainly rural areas in Africa (WHO, 2015). In the last 50 years, dengue incidence has increased 30-fold. The World Health Organization (WHO) estimates that there are 2.5 billion people live in over 100 endemic countries and areas where dengue viruses can be transmitted. Every year, there are up to 50 million infections occur with 500,000 cases of DHF and 22,000 deaths mainly among children (WHO, 2015). It is worthy to note that prior to 1970, only 9 countries had experienced cases of DHF. Since then it has increased more than 4-fold and continues to rise.

As of 12 September, it has been reported that there were 85,488 cases of dengue with 234 deaths in Malaysia for 2015 (WHO Western Pacific Region, 2015). This is 21.5% higher compared with the same reporting period of 2014 (Figure 1).

![Figure 1: The number of dengue cases per week in 2014 & 2015 (Source: Department of Health, Malaysia)](image)

Hence, it is very important to be able to effective predict the number of dengue cases and potential dengue outbreaks. This will give public health agencies enough advance time to carefully plan and implement mitigation strategies, especially for the cities or regions where medical treatment facilities may become overwhelmed by an outbreak. Many existing publications have identified and discussed the association of different contribution factors to dengue outbreaks. In particular, many studies – for example (Focks, Daniels, Haile & Keesling, 1995; Barbazan et al., 2010; Shang et al., 2010) – have been focusing on the environmental factors that influence mosquito vector populations. Unfortunately, most of these studies were mainly focused on dengue outbreak detection and surveillance, rather than dengue outbreak prediction (Runge-Ranzinger, Horstick, Marx & Kroeger, 2008). This observation was further supported by a recent literature review of decision support systems for the prediction, prevention, and control of vector-borne diseases (Eisen & Eisen, 2011). They found that although useful risk maps were identified and described in the reviewed studies, neither actual dengue outbreak predictions nor forewarnings of future dengue outbreaks was proposed.

In this paper, we propose a new machine learning approach to predict the number of dengue cases in Kuala Lumpur, in particular the areas surrounding the University of Malaya (UM) Medical Centre.
We identified several different factors that can contribute to a surge in the number of dengue cases that occurred near the UM Medical Centre. Apart from the daily mean temperature and daily rainfall factors that have been frequently used in other studies, we also considered the enhanced vegetation index (EVI) as an input factor to our prediction engine. We trained our linear regression model on these three factors against the number of dengue cases from 2001 to 2010. We then tested our model on the 2011 data. The experimental results showed that our approach was able to predict the number of dengue cases 16 days in advance with high accuracy. In addition, our results confirmed that the EVI was the best predictor among the three factors used in this study.

2. Related Work

In general, existing methods described in the literature for predicting a dengue outbreak can be classified in the following categories: Statistical Analysis (), Neural Network (), Bayesian Network () and Support Vector Machine (). Among these streams, statistical analysis, in particular regression model, is the most popular choice and has been used in many studies. For example, Hii, Zhu, Ng, Ng & Rocklöv (2012) proposed a time series Poisson multivariate regression model to predict weekly dengue cases in Singapore. Their model was developed and trained using weekly mean temperature, cumulative rainfall, and previous dengue incidence during 2000–2010. They then validated their model based on observed weather data and dengue incidences for the period from week 1 of 2011 to week 16 of 2012. Another example that applied Statistical Analysis for dengue outbreak prediction is the work by Eastin, Delmelle, Casas, Wexler, & Self (2014). They developed two environment-based multivariate, autoregressive forecast models to anticipate the dengue outbreaks from 2 weeks to 6 months in advance using epidemiological and meteorological data for the urban environment of Cali, Colombia from January of 2000 to December of 2011. They found that significant dengue outbreaks generally occurred during warm-dry periods with extreme daily temperatures confined between 18°C and 32°C. Many studies also used autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) models to predict dengue outbreaks (e.g. Gharbi et al. 2011; Bhatnagar, Lal, Gupta & Gupta, 2012; Dom, Hassan, Latif & Ismail, 2012). Those models are well established in the financial industry and take into account the seasonality of financial data for prediction. However, it may be hard to regularise the environmental data for dengue outbreak prediction due to climate change.

Yu, Yang, Yen & Christakos (2011) developed their Bayesian Maximum Entropy model on a more variety of data, including temperature, rainfall, Southern Oscillation Index, mosquito larvae abundance, and human health data. They trained their model on Taiwan data from 2002-2006 and tested it prediction for 1 week in advance against Taiwan data in 2007. They found that the probabilities that a dengue outbreak actually occurred given a positive prediction and given a negative prediction were 0.5541 and 0.031, respectively. Another example of Bayesian based approach is the work by Lowe et al. (2011). They analysed different types of data from 2001-2008 to develop a Bayesian generalized linear mixed model to predict monthly dengue incidence in Brazil. Their approach combined a spatio-temporal hierarchical model (or generalised linear mixed model) with a Bayesian framework using Markov Chain Monte Carlo. They found that this approach improved the prediction both spatially and temporally.

In recent years, researchers have applied data mining techniques to further support the predictive modelling of dengue outbreaks (Bakar, Kefli, Abdullah, & Sahani, 2011; Buczak, Koshute, Babin, Feighner & Lewis, 2012) in the cases of nonlinear models and high dimensional data containing many potentially interacting predictor variables. These techniques automatically generate association rules that represent the potential relationships between input data parameters and the predict outcomes. These approaches require the users to set a pre-determined criteria of
significance to filter out association rules whose metrics fail to meet these criteria. This is to ensure that variables that are not strongly correlated are excluded. A rule classifier built with the final rules will be used to predict dengue outbreaks. It is worthy to note that Bakar, Kefli, Abdullah, & Sahani (2011) used medical records of dengue patients from the Selangor area in Malaysia as their main data source. Unlike other studies, their classifiers predicted whether or not a patient was a repeat dengue case, rather than the dengue incidence. One drawback of these data mining techniques is they are good at predicting periods of high and low dengue cases but they cannot predict the actual dengue incidence.

3. Contribution Factors to a Dengue Outbreak

There are many factors that can contribute to a dengue outbreak. Existing data parameters that were used in the literature for predicting a dengue outbreak can broadly be grouped into three main categories: environmental data, epidemiological and socioeconomic data.

3.1. Environmental Data

Dengue is transmitted via the bite of infected Aedes mosquitoes. It has been proved that the life cycle of Aedes mosquitoes is heavily impacted by the environmental conditions. Therefore, it is important for the prediction model to be able to predict whether the future climate conditions will be in the optimal range for Aedes mosquito’s survival and viral transmission. Table 1 summarises environmental data that have shown significant correlation with dengue incidence in the literature.

<table>
<thead>
<tr>
<th>Rainfall</th>
<th>Temperature</th>
<th>Altitude</th>
<th>Biome</th>
<th>Land surface temperature</th>
<th>Remotely sensed vegetation</th>
<th>Sea surface temp anomalies</th>
<th>Southern Oscillation Index</th>
<th>Humidity</th>
<th>Mosquito population index</th>
<th>Larvae population index</th>
</tr>
</thead>
</table>

3.2. Epidemiological and Socioeconomic Data

Aedes mosquitoes acquires the dengue virus from biting an infected patient and passes it to another person. Therefore, it is important to consider historical human incidence rates in the prediction model as it reflect the presence of the virus in the local human population.

In addition, lots of socioeconomic data can also be good predictors for dengue outbreaks (Åström, Rocklöv, Hales, Béguin, Louis & Sauerborn, 2012). For example, the current population and people movement can help predict the fluctuation of the virus presence. Other socioeconomic data like poverty index, electricity access and drinking water access are also good predictors as they indicate
the living conditions of people in the interested area. Table 2 summarises the epidemiological and socioeconomic data that have been found useful for dengue outbreak prediction.

Table 2: List of epidemiological and socioeconomic data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dengue incidence</td>
<td></td>
</tr>
<tr>
<td>Census data (population)</td>
<td></td>
</tr>
<tr>
<td>Administrative boundaries</td>
<td></td>
</tr>
<tr>
<td>Poverty index</td>
<td></td>
</tr>
<tr>
<td>Electricity access</td>
<td></td>
</tr>
<tr>
<td>Drinking water access</td>
<td></td>
</tr>
<tr>
<td>Sanitation index</td>
<td></td>
</tr>
<tr>
<td>Child health index</td>
<td></td>
</tr>
<tr>
<td>Child education index</td>
<td></td>
</tr>
<tr>
<td>Child quality of life index</td>
<td></td>
</tr>
<tr>
<td>Human population movement</td>
<td></td>
</tr>
<tr>
<td>Vector control</td>
<td></td>
</tr>
</tbody>
</table>

4. Our Approach

4.1. Data Collection and Processing

In this study, we focused on building a model to predict the dengue incidence in the area within 15km radius from the UM Medical Centre. The daily dengue incidence records for years 2001-2011 were provided by our collaborator TIDREC, University of Malaya. Each daily dengue incidence rate represents the number of confirmed dengue cases reported to the UM Medical Centre. We trained our model with these daily incidence. However, we found that the data was too noisy. For example for a particular month the total number of dengue cases may be high but still for some days within that month the turnout may be low due to various social and economic factors. People may tend to visit hospital on certain days even if they are sick, for convenience. We decided to use the averages over a period of time as input data to overcome this issue. We found that an interval of 15 days yields the best results. This is also coincidence with the enhanced vegetation index interval.

In addition, we considered the mean daily temperature, daily rainfall and enhanced vegetation index as the environmental factors for our prediction model. We was not able to obtain other data factors in time for this study.

Our mean daily temperature and daily rainfall data were provided by the Malaysian Meteorological Department. They were collected by the two main stations in Subang Jaya closed to the UM Medical Centre. We obtained the enhanced vegetation index from NASA. The data was provided for each 500 m² of land area in an interval of 16 days. The EVI data was not available for 48 days in our studied period of years 2001-2011. For those days, we used cubic spline interpolation to estimate the missing data.
4.2. Prediction Model

In this study, we found that there was no regular seasonal pattern for our input data. Therefore, linear regression model turned out to be a good choice for our prediction model as it can deal with irregular observable variables and produces model with low variances. We started with a simple model with only the previous number of dengue cases as input, and then gradually incorporated other data factors (i.e. daily temperature, daily rainfall and enhanced vegetation index) to improve the performance of our model.

Formally, our linear regression model with n data factors \( f_1, f_2, ..., f_n \) can be described as:

\[
F(d) = C + a_1 * f_1 + a_2 * f_2 + ... + a_n * f_n
\]

where:
- \( F \): a transformation function for dengue incidence
- \( d \): the number of dengue cases
- \( C \): a constant factor
- \( a_1, a_2, ..., a_n \): the coefficients (i.e. weights) for these factors \( f_1, f_2, ..., f_n \)

5. Results and Discussion

In our experiment, we divided the 2001-2011 dataset into three subsets for training, validation and testing. We ran two different models on datasets with different ratios of training, validation and testing data subsets: one model with only the previous number of dengue cases and mean temperature as inputs, and the other one with all four data factors as inputs. After each model was trained, we validate the model to fit the best regularization parameter and avoid overfitting. Tables 3 and 4 show the mean test errors for these two different models on different runs.
As shown in Tables 3 and 4, the first model that used only the previous dengue cases and mean temperature performed very robust on different splits of data for training, validation and testing. With as small as 50% of the whole dataset available for training, the first model returned the mean error of 9.3 cases per 15 days. On the other hand, the second model with all input data factors required at least 80% of the dataset available for training to produce good prediction. The performance of both models increased when more data was used for training. When there was at least 80% of the dataset available for training, our second model outperformed the other one. It produced the best mean error of 7.2 cases per 15 days on the 90-5-5 data split.

Figures 2 and 3 graph the predicted dengue cases (red line) 15 days in advance that were generated by our second model against the actual dengue cases (blue line) on a daily basis. In these graphs, the x-axis represents the days in the timeline and the y-axis measures the number of dengue cases. The two green vertical lines in each graph indicate where the dataset was split into the training, validation and testing subsets.

As shown in these two figures, the accuracy of the predicted dengue cases improved largely when trained with more data. Indeed the predicted trend is more accurate and the gap between the predicted cases and the actual cases is much smaller in figure 3 than in figure 2, especially after the first 2000 days.

### Table 3: Dengue cases and mean temperature

<table>
<thead>
<tr>
<th>Training Data (%)</th>
<th>Validation Data (%)</th>
<th>Test Data (%)</th>
<th>Mean Test Error (number of dengue cases per 15 days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>35</td>
<td>35</td>
<td>14.5</td>
</tr>
<tr>
<td>40</td>
<td>30</td>
<td>30</td>
<td>11</td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>25</td>
<td>9.3</td>
</tr>
<tr>
<td>60</td>
<td>20</td>
<td>20</td>
<td>9.4</td>
</tr>
<tr>
<td>70</td>
<td>15</td>
<td>15</td>
<td>8.7</td>
</tr>
<tr>
<td>80</td>
<td>10</td>
<td>10</td>
<td>9.2</td>
</tr>
<tr>
<td>90</td>
<td>5</td>
<td>5</td>
<td>8.3</td>
</tr>
</tbody>
</table>

### Table 4: Dengue cases, mean temperature, rainfall and EVI

<table>
<thead>
<tr>
<th>Training Data (%)</th>
<th>Validation Data (%)</th>
<th>Test Data (%)</th>
<th>Mean Test Error (number of dengue cases per 15 days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>35</td>
<td>35</td>
<td>19.9</td>
</tr>
<tr>
<td>40</td>
<td>30</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>25</td>
<td>11.4</td>
</tr>
<tr>
<td>60</td>
<td>20</td>
<td>20</td>
<td>10.8</td>
</tr>
<tr>
<td>70</td>
<td>15</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>80</td>
<td>10</td>
<td>10</td>
<td>8.8</td>
</tr>
<tr>
<td>90</td>
<td>5</td>
<td>5</td>
<td>7.2</td>
</tr>
</tbody>
</table>
Figure 2: The results of the second model with all input data on the 60-20-20 split for training-validation-testing. The vertical green lines indicate the split.

Figure 3: The results of the second model with all input data on the 90-5-5 split for training-validation-testing. The vertical green lines indicate the split.
6. Conclusion

In conclusion, we proposed a new linear regression model to predict the number of dengue cases in the areas surrounding the University of Malaya Medical Centre. Our model used the previous dengue incidence, the daily mean temperature, the daily rainfall and the enhanced vegetation index (EVI) as input factors to our prediction engine. The experimental results showed that our approach was able to predict the number of dengue cases 16 days in advance with high accuracy. In future work, we plan to incorporate extra dengue outbreak contribution factors into our model and identify the best combination of predictive factors.

Acknowledgement

This research was supported by the Science Fund Grant 01-03-04-SF0061 by the Ministry of Science, Technology & Innovation Malaysia (MOSTI). The dengue incidence data was collected and processed with the support of the Malaysia Genome Institute Initiative Grant 07-05-MGI-GMB015 by the Ministry of Science, Technology & Innovation Malaysia (MOSTI).

References


Shang, C.S. et al. (2010). The role of imported cases and favorable meteorological conditions in the onset of dengue epidemics. PLoS Negl Trop Dis 4(8), e775.

