National College of Ireland

Exploring Dengue Fever in Malaysia

MSc Research Project
Data Analytics

Qadirah Chan Latif
x16132297

School of Computing
National College of Ireland

Supervisor: Dr. Eugene O'Loughlin

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Qadirah Chan Latif
Student ID:	x16132297
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Eugene O'Loughlin
Submission Due Date:	11/12/2017
Project Title:	Exploring Dengue Fever in Malaysia
Word Count:	5920

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	6th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploring Dengue Fever in Malaysia

Qadirah Chan Latif
x16132297
MSc Research Project in Data Analytics

6th December 2017

Abstract

Dengue fever is a common disease in the countries with tropical climate such as the Southeast Asia region. The intention of this research is to explore the transmission patterns of dengue cases in various climate conditions. The study also forecasts the trend and seasonal pattern for dengue cases considering weather variables in time series forecasting models. Other studies are reviewed to support the models and evaluation in this research. From this study, it is observed that the simplest technique of data mining could lead to a better a accuracy of forecast. The result of this study could be a stepping stone to new strategies or biological development to fight dengue transmission.

1 Introduction

The World Health Organization (WHO) has placed dengue as one of the severe diseases as it affects more than 100 countries with an estimation of 390 million people infected yearly (WHO; 2017). The disease has been spreading geographically since the last 50 years with a rise of reported incidences by 30 times (Mutheneni et al.; 2016). This disease is transmitted through two species of mosquitoes, the *Aedes aegypti* or *Aedes albopictus* that carries the Dengue Virus (DV).

Johansson et al. (2009) indicated that the virus is prone to spread in environments which are suitable for the mosquitoes lifecycle such as in tropical and subtropical regions. Climate was identified by Morin et al. (2013) as one of the vectors that could influence the mosquito population. The DV virus is sensitive to climate which indicates a necessity of the vector being present in order to develop the virus in the mosquito regardless the ambience (Naish et al.; 2014). Hii et al. (2012) mentioned studies have shown these mosquitoes are capable to get as far as the 21st floor of a residence. An infected female mosquito will have its lifecycle of virus replication lengthen in environments containing humidity coupled with some rainfall (Pinto et al.; 2011).

In Southeast Asia countries, the endemic increased greatly overtime as is a health burden in countries of this region (Ooi and Gubler; 2009). 2015 was the highest year observed with the dengue fever endemic in this region (Nguyen; 2015). The capability of Aedes mosquitoes to habituate in various locations and environment conditions has led this disease being widely spread in urban and rural areas. Initiatives to develop strategies such as an early surveillance system or new biological development are required to reduce the transmission of this virus.

The purpose of this paper is to review the influence of climate towards dengue cases in Malaysia by months for 2011 to 2017 and further investigate how the trend of dengue cases look in the future in varying climate conditions. Climate conditions to dengue cases can be identified using clustering or spatial distribution methods as a base to gaining knowledge related to disease transmission patterns. Time series forecasting models capable to estimate the future volume, trend and seasonal pattern of dengue cases were developed. It is hypothesized there is

- A form of relationship between dengue cases and weather conditions when observing the clustering model
- An increased trend in dengue cases overtime with inclusive assessment of weather conditions

2 Related Work

Existing works related to the implementation methods applied in this research is discussed in this section.

2.1 Identification of Dengue Incidences Patterns

Hasan and Md (1998) introduced Spatial Data Distribution technique to discover unknown patterns from a spatial set of data. The authors analyzed Self-Organizing Maps (Kohonen; 1982) (SOM) as an algorithm for data pattern identification or clustering. Bação et al. (2005) conducted research on SOM to determine if is an alternative to the well-known K-Means clustering MacQueen et al. (1967) technique. Performance of both techniques were compared by robustness and SOM outperformed K-Means by error rates and extensively provides the possibility of exploring data patterns. Augustijn and Zurita-Milla (2013) concluded SOM as a suitable alternative technique for analyzing patterns for diseases data as compared to other pattern detection methods.

Using Spatial Distribution pattern, a dengue hot spot location study was conducted by Mutheneni et al. (2016) on Andhra Pradesh, India. Execution of SOM classified the districts in Andhra Pradesh into 3 clusters with 7 sub clusters from high to low dimensional level clustering. The sub clusters were distributed by the number of cases of 3 levels of endemic severity where the district of Khammam was detected as a definite hot spot area. Mutheneni et al. (2016) suggests that population of a neighbouring area may have impact into how fast the infection can be transferred from one location to another.

A systematic study by Naish et al. (2014) on the influence of climate towards dengue cases suggested that areas of research should be developed on the topic of climate to gather a better understanding on the temporal pattern distribution of variables. Liu-Helmersson et al. (2016) detected a potential of the dengue epidemic in Europe due to the changing climate caused by greenhouse gases emission. The research suggests that dengue outbreak is a complex affair and may involve other climate influences rather than just mean temperature and daily temperature differences.

2.2 Models for Time Series Analysis

2.2.1 Count Data Models for Relationship and Prediction

An investigation in West Java, Indonesia by Ahdika and Lusiyana (2017), compared a Poisson model with Markov model to determine the best fitting model for dengue prediction where Markov model was proved to be better. Studies on relationship of dengue with climate by Naish et al. (2014) and Morin et al. (2013) favoured Poisson model in establishing identification of relationships and detected prediction models can be developed. Using Poisson model for a study of dengue cases in Singapore, Pinto et al. (2011) indicated minimum and maximum temperature as the best cause of increased dengue cases. Lu et al. (2009) conducted a time series analysis between dengue cases and weather variables with Poisson model suggested it could be developed into a prediction model.

Lowe et al. (2009) used a Negative Binomial (NB) model as opposed to Poisson model to fit to over dispersion of data that occurs in count data. Him et al. (2012) and Choi et al. (2016) explored dengue data in Malaysia and Cambodia respectively using the NB Model. Him et al. (2012) detected that climate information does influence dengue cases in Malaysia with significance of p-value < 0.5 for the model. Previous 3-month rainfall for a location was found to contribute towards dengue transmission for the current month of the same location. The analysis found by Him et al. (2012) was supported in the study by Lowe et al. (2009) which implemented a NB model into a Generalized Linear model (GLM) over an 88-month period for dengue incidences in Brazil. The authors found a similar pattern on rainfall where past 3-month rainfall contributes to increase dengue incidences.

Choi et al. (2016) studied Cambodian dengue data and concluded that mean, maximum and minimum temperatures does influence the increase of dengue incidences in three provinces of Cambodia. The authors suggests for other studies to consider wind speed, humidity and water evaporation variables in order to investigate how rainfall interacts with other climate variable towards dengue incidences. Fairos et al. (2010) modelled a time series data of dengue outbreak in Malaysia with Poisson and NB models. The NB model in the research was proved as the best where temperature, wind speed and humidity are noteworthy predictors. Association of dengue cases with climate variables coupled with locality variation could lead to generating complex knowledge and understanding of dengue incidences (Choi et al.; 2016).

2.2.2 Seasonal Time Series Analysis Model

Forecast of disease outbreak commonly involves time series models as the technique is suitable for cyclic or repeating observations (Kane et al.; 2014). Ho and Ting (2015) described time series models capable to model forecasts for infectious disease outbreaks with the facilitation of meteorological variables. Nury et al. (2013) mentioned time series analysis as a vital equipment of analysis for study areas related to climate or environmental variables.

Naish et al. (2014) studied 16 papers on modelling approaches for dengue cases and indicated 6 papers where Seasonal Autoregressive Integrated Moving Average (SARIMA) and Wavelet time series were used in analysis of dengue cases time series forecasting. Research on dengue data over a time series involving weather variable was performed by Martinez et al. (2011) using SARIMA model. Johansson et al. (2009) constructed a

Wavelet time series model to define dengue and climate variability associations. Martinez and Silva (2011) stated Autoregressive Integrated Moving Average (Box et al.; 1976) (ARIMA) model to be well received in articles related to epidemiological topics such as dengue.

A research on Northern Thailand by Silawan et al. (2008) used SARIMA to forecast monthly rates of dengue incidences modelled over 8 years detected a biyearly case peak in June and July each year. Gharbi et al. (2011) conducted a Guadeloupe dengue time series investigation with SARIMA model by including climate as seasonal factors that could increase the reliability of forecasting power. The study found models fitted with the external factors to have a good level of forecasting power. Naish et al. (2014) supported the inclusion of external factors into SARIMA as it is capable to produce a more robust forecasting model. The investigation by Gharbi et al. (2011) detected a positive correlation between dengue cases with humidity, minimum and relative temperature.

Johansson et al. (2016) chooses the best fitted SARIMA model by designating specific weather variable over time. From the analysis, it was found increase in dengue cases is substantial over an increased seasonal period. The investigation by Martinez and Silva (2011) conducted on Ribeiro Preto, Southeast of Brazil concluded SARIMA models to be reliable as the forecast value is reasonable when compared to the actual observations. Azam et al. (2016) supported the reliability and robustness of SARIMA to forecast dengue cases and act as base model to develop an epidemic surveillance tool.

2.2.3 Other Time Series Model

Dayama and Kameshwaran (2013) used the Holt-Winters (Holt; 1957) and (Winters; 1960) (HW) method to compare with SARIMA as they both work well for time series forecast. A study forecasting the arrival of tourists in India by Sood and Jain (2017) executed a comparative analysis between HW and SARIMA. In both cases, the authors described HW model as an exponential smoothing model which is appropriate for data with trend and seasonality as these elements are considered in the models algorithm. Ho and Ting (2015) explains that the model uses the past weighted averages with reduced weights over specific time leading to a higher forecast value. The model is fast and inexpensive which allows it to accommodate as a frequently used model (Sood and Jain; 2017).

An investigation on dengue cases forecasting by Ho and Ting (2015) compared the exponential smoothing model with SARIMA using the same time lag and open dataset. The authors describe the models algorithm to differ as past observations are weighted heavily into HW and a seasonal fit is weighted heavily for SARIMA. Results of the research evaluation show both models return almost identical forecast values which could be used in comparison with other research of similar topic. Also discovered was an increase in dengue cases in the fourth quarter of a year.

Briët et al. (2008) mentioned in their work using HW on Malaria forecasting that external variables such as weather may not improve the robustness of the model but rather reduce in the models forecasting accuracy. On that account, the dengue forecasting research by Ho and Ting (2015) does not involve external variables into the forecast model.

2.2.4 Time Series Performance Evaluation

Performance evaluation of incidence forecasts is executed by comparing the actual values of incidence against the forecasted incidences values (Johansson et al.; 2016). Dayama

and Kameshwaran (2013) and Song et al. (2015) executed the suggested method of performance validation of forecast models for a research on disease incidences in Singapore and China respectively. Dayama and Kameshwaran (2013), Song et al. (2015) and Ho and Ting (2015) suggested accuracy matrices as performance evaluation. Accuracy matrices mentioned by these authors are the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). A study of diseases by Ramadona et al. (2016) and Song et al. (2015) both respectively used RMSE to forecast model fitting. This metric evaluates by calculating the difference between fitted values with observed values where smaller the value produced indicates a strong forecast (Song et al.; 2015). Ramadona et al. (2016) found that using RMSE evaluation was reasonable to detect the forecast strength of the observed year 2013.

Johansson et al. (2016) advocated the use of MAE for the research of dengue in Mexico. The authors extended the fit of the models by calculating the relative MAE (relMAE) to compare forecasts by locations. A value of relMAE higher than 1 indicates a forecast model which is further from the observed values which then means a less fitting model. Both Dayama and Kameshwaran (2013) and Ho and Ting (2015) stated MAE in the research results as well. The MAE value was included only as additional information for readers without detailed explanation of MAE in their research.

Armstrong (1985) recommended the use of MAPE as a performance metrics (Shi et al.; 2016). Earnest et al. (2012) used MAPE to measure accuracy of ARIMA forecast models in the research on dengue. MAPE was compared to Mean Squared Error (MSE) as a comparison metric. It was mentioned MAPE can be easily interpreted and hence a more intuitive method of evaluation as the evaluation is relative to the magnitude of actual observations.

3 Methodology

In this section, the methodologies selected to execute the research is explained.

3.1 Data Mining Methodology

This research was developed based upon the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. The CRISP-DM methodology enables a project to be well structured and document for easy revision (Azevedo and Santos; 2008). The six stages of CRISP-DM are defined as Figure 1.

3.2 Data Collection

Malaysia was the chosen country for investigation for this research with data collected for the course of 7 years, from 2010 to 2017. As 2017 has not ended when the research was ongoing, the data collected for 2017 was up to week 39 which is equivalent to the month end of September.

Dengue data was collected manually from weekly dengue reports documented in pdf at MOH (2017). The dengue data was reported weekly across 15 states in Malaysia. Weather data was downloaded from NCDC (2017) for weather stations across Malaysia whereby contains daily observation of weather condition. The weather dataset collected does not include the rainfall amount for the states in Malaysia. The data on rainfall was

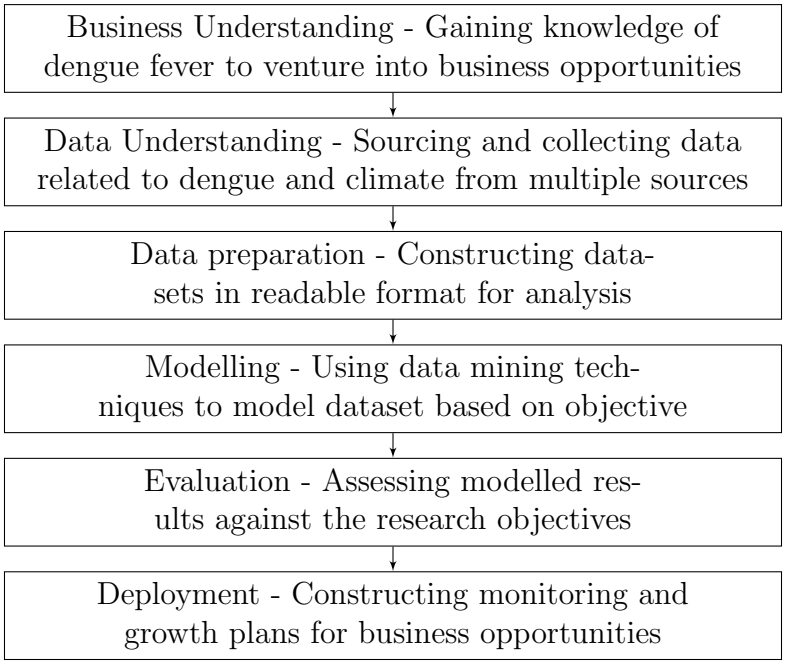


Figure 1: CRISP-DM methodology developed for research based on model described by Azevedo and Santos (2008)

collected separately from CCKR (2015) which contains observations for the years 1991 to 2015 on monthly basis.

The latest research available related to dengue prediction in Malaysia was by Ho and Ting (2015) where the authors used an existing dataset available with from 2010 up to 2015. Azam et al. (2016) used a dataset with longer period whereby evaluated years 1981 to 1999 and used validation periods from 2000 to 2014. Mathulamuthu et al. (2016) and Latif and Mohamad (2015) used observation for a year for Selangor state in Malaysia for their investigation of dengue. It is worth mentioning that the data used in this research is currently the most updated in this topic area.

3.3 Data Preprocessing

Microsoft Excel and R were used to prepare data into its final form before analysis is executed. The main task used in Excel was used to group data from daily observations to weekly observations. R is not suitable to be used as for this task as R does not provide an accurate grouping for weeks which have dates with overlapping years. Imputing missing values in dengue dataset, column selection for analysis and aggregation of data are the main data processing tasks in R. Executing these tasks in R is more efficient and effective as it is automated and flexible to accommodate to new change in the dataset. The process to format data for analysis is described in detailed the configuration manual.

3.4 Analysis Tool

The chosen analysis tool for this research is the open source programming language, R. It is capable to accommodate to the implementation and evaluation through packages that can be installed prior to the analysis. The packages used are described in detail in the configuration manual.

3.5 Evaluation method

In a study of accuracy performance metrics, (Li; 2017) introduced the coefficients r and r^2 as well-known evaluation methods for predictive models which have been implemented in various topic of studies. However, a downside of these coefficients is that it does not measure the accuracy but rather measures for correlation. The author mentioned that in the topic of environmental sciences, the RMSE and MAE are other commonly used matrices even though they are limited to be used to models using the same dataset in evaluation.

As mentioned in Section 2, Dayama and Kameshwaran (2013) used RMSE and MAE as predictive model accuracy. The authors found the SARIMA model to perform better in predictions than HW model by comparing the RMSE and MAE values. The research by Johansson et al. (2016) was also mentioned in Section 2 where the author used MAE as the metric of model assessment and found that models with year and month produces lower error by the MAE. For a research of dengue in Indonesia, Ahdika and Lusiyana (2017) used the measurements MSE, MAE and MAPE to compare forecast models. Based on the successful use of RMSE, MAE and MAPE in other research related to dengue, these measures will be used as the evaluation metrics for this research.

4 Design and Implementation

In this section, the exploratory analysis of dataset and implementation techniques are discussed.

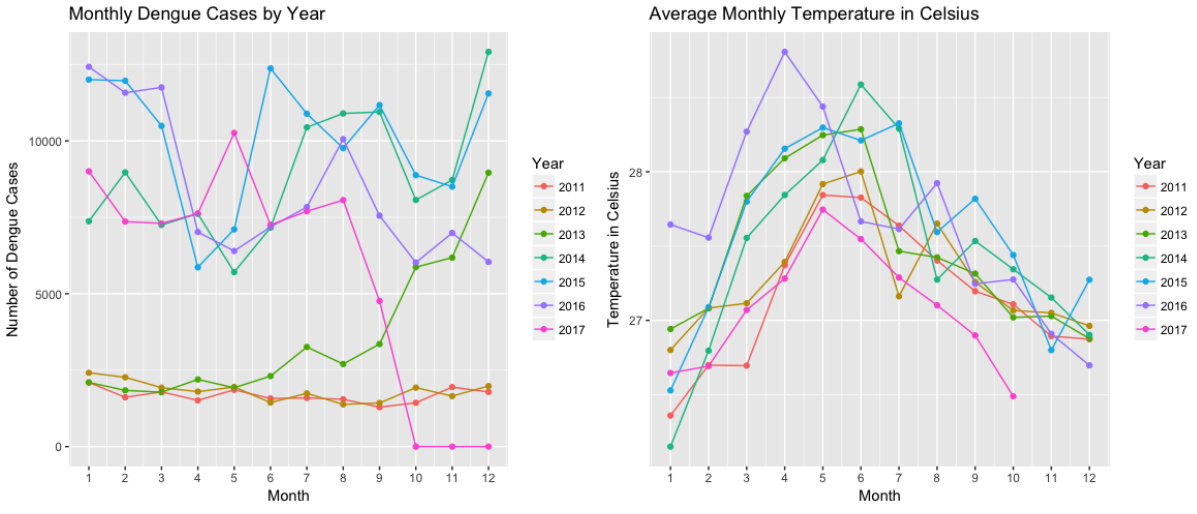
4.1 Exploratory Analysis

Datasets are explored to understand the type of data that is being analyzed where the charts are generated using the ggplot2 package. Visuals assist in analyzing data better and sets an expectation for future prediction evaluation on dataset.

4.1.1 Anomaly detection

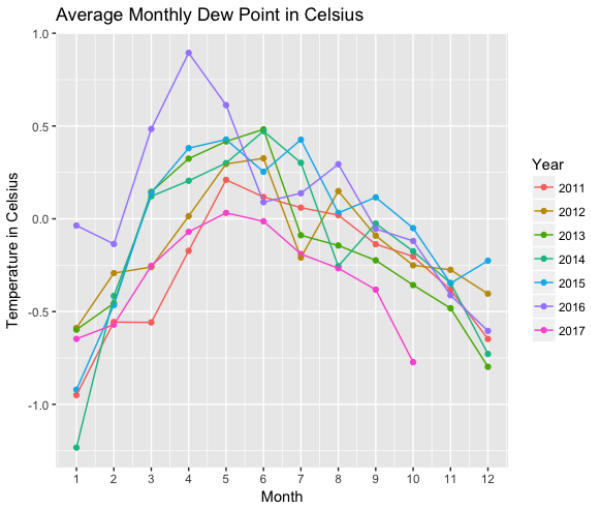
As seen in Figure 2a, there are some notable spikes of dengue cases throughout the years such as at December 2014, June 2015, January 2016 and May 2017 which are identified as peaks. It can also be observed that the number of cases multiplied between 5 to 10 times for 2014 to 2017 as compared to 2011 to 2013. From observation, 2015 has generally high number of cases spread across the months in the year as compared to other years which indicates a total high number of cases. This assumption is evident as reported in *Dengue most contagious disease since 2015* (Carvalho et al.; 2017), the number of cases in 2015 topped the highest so far in the country. Further detailed observation on dengue cases in Malaysia can be viewed in the configuration manual.

Exploring the weather dataset, it was found that the highest points of temperature as Figure 2b is in the month of April 2016 at 28.8°C with the lowest point of temperature is months of January and December for all observed years. It was identified the highest dew point incidence as Figure 2c at April 2016 for overall years of observation, at a point of 0.8941. As per Figure 2d, the months with highest amount of rainfall are in January and December with a seasonal observation found for the rainfall pattern in Malaysia.

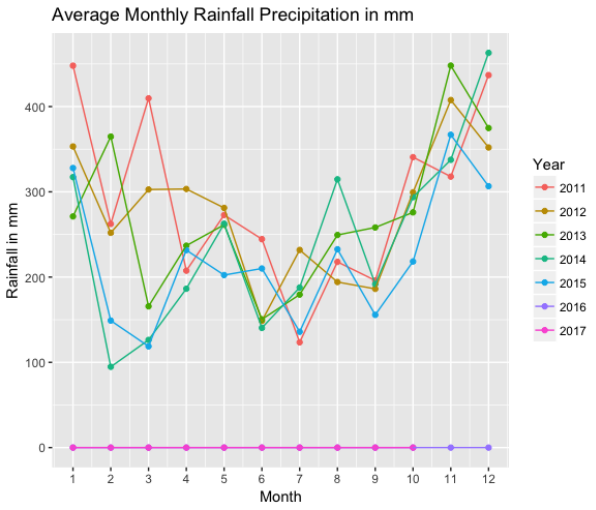


(a) Dengue cases

(b) Actual Temperature



(c) Dew point



(d) Rainfall

Figure 2: Plots of dengue cases and weather condition observations over months of 7 years

4.1.2 Time series data exploration

By observing Figure 2a, it can be assumed that there is an upward trend of dengue cases across time. The time series is decomposed to explore the actual trend and seasonal pattern for dengue cases. From the observation of Figure 3, it can be agreed that an upward trend does exist for dengue cases over the years with a cyclical seasonal pattern overtime. The first assumption can be set where there would be an increased number of dengue cases over the years.

4.2 Implementation

The implementation methods for the research will be discussed in this section where the topics covered are i) Self Organizing Maps ii) Count Data Model iii) SARIMA iv) HW.

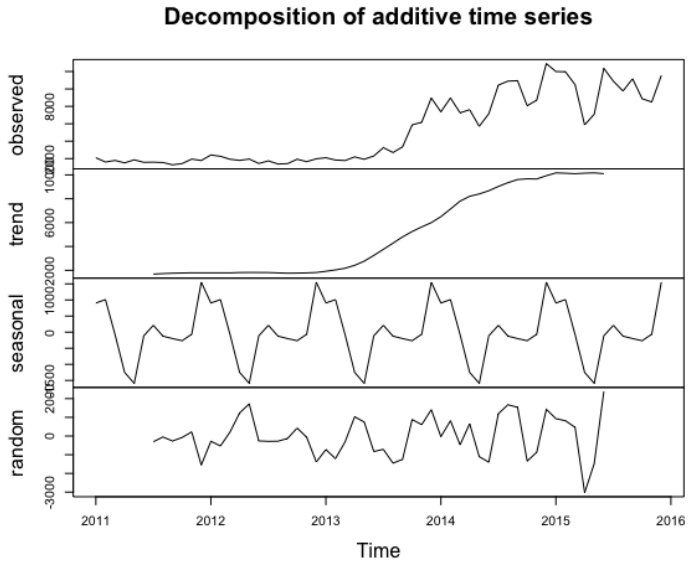


Figure 3: Decomposition of Time Series

4.2.1 Self-Organizing Maps (SOM)

It is a clustering technique that groups observations that contain similar properties by a reiterating learning process. The kohonen package in R was used to train an SOM model that is later used to plot heat maps that allows identification of node patterns that explains relationships between properties in the dataset. The dataset was split to a 70:30 ratio as a scale over the probability of variants available. For SOM model to work, a sample input size of model needs to be provided based on the size of dataset matrix that is being used to generate a SOM grid size for arrangement of nodes. The formula below recommended by Kohonen (1982) was used in determining the sample size of the dengue weather dataset.

$$5 \times \sqrt{numberofrows} \tag{1}$$

The model is iterated at a selected 1000 times before convergence. At each iteration, data is being supplied into the model to match the closest neuron or Best Matching Unit into the training vector. The new weights of neuron and its corresponding neighbours are then being updated. Once the iteration reaches 1000 times, the iteration reaches stability which indicates there is no further matching that could take place.

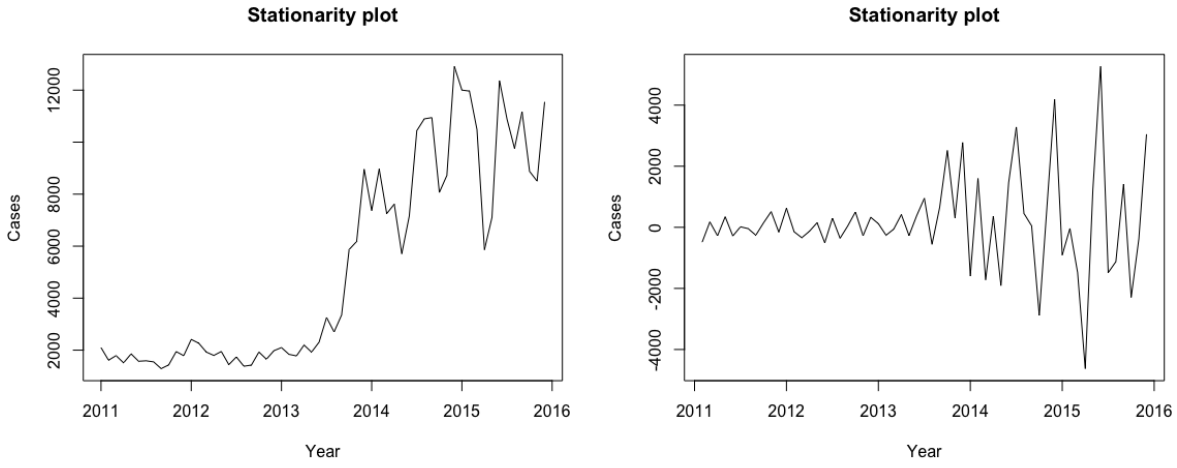
4.2.2 Count Data Model

This regression technique is suitable for count data to unfold relationship between variables and can be incorporated into a prediction model. The package MASS in R was used to inspect the suitable Poisson model for the dataset. A Poisson GLM was tested for over dispersion with function dispersiontest from the AER package. Over dispersion occurs when the variance is higher value than the mean in the dataset. From the tests, it was identified an over dispersion in the Poisson model by 620.164 indicating that a NB model would be appropriate for the data. A NB GLM was created and the model was found a better fit through a goodness of fit of p-value > 0.05 at 0.08562568 while Poisson is at 0.

A count time series Poisson and NB model were created using the tsglm function in tscount package where the regression models. The years 2011 to 2015 were used as the training set and 2016 to 2017 for evaluation. Stepwise regression was used to identify independent variables that contributes to the dependent variable. Poisson and NB models were created to compare the best fitting model and later the model was used to predict values for 2016 to 2018.

4.2.3 SARIMA

SARIMA is widely applied for time series analysis for data containing seasonality over a specific period. The tseries package was used during the implementation of SARIMA to check the form stationarity and autocorrelation of data. From the Augmented Dickey-Fuller Test (Dickey and Fuller; 1979) (ADF), the dataset was not stationary as in Figure 4a and differencing was applied to the dataset to convert it into a stationary form per Figure 4b so that the dataset would fit into the SARIMA model. Autocorrelations of the differenced dataset is used to determine the Autoregressive (AR) and Moving Average (MA) terms required for the model.



(a) Before (b) After

Figure 4: Stationarity plot before and after differencing

The function auto.arima was used to check if weather variables would influence the SARIMA model containing data from 2011 to 2015. Akaike Information Criterion (Akaike; 1998) (AIC) value was used as a model selector. It is an unbiased estimator that informs the amount of information loss where smaller AIC value is better (Posada and Buckley; 2004). With AIC value 834.92, the SARIMA model without weather variables are better fitted as compared to the model with weather variables. SARIMA models are indicated with $(p, d, q)(P, D, Q)$ where p is the order of AR, d is the degree of differencing, q is the order of MA and P, D, Q are denoted as the p, d, q of seasonal terms.

The model ARIMA(0,1,2)(0,1,1) determined by the auto.arima model without variables is then compared with the model ARIMA(1,1,1)(1,1,0) which was selected manually through the ACF and PACF plots which can be found in the configuration manual. Models were fitted with hold-out data from 2016 to 2017 into the forecast model of 36 months. Best fitted model was determined by comparing the AIC value and the closest predicted value against actual values of 2016 to 2018.

4.2.4 Holt-Winters

The HoltWinters function from the R package forecast was used to compute a HW filtering model for data from 2011 to 2015. Time series and seasonal components from the time series dataset are considered in the model automatically. The alpha α, beta β and gamma γ of the model were observed and interpreted. α at level 0.4733555 indicates the estimates of model are based upon a combination of previous and recent data. β at 0.01093996 indicates some trend being identified and high γ at 0.8146861 indicates the seasonal component to consider very recent seasonal observations. The model is then fitted into a forecast function with a period of 36 months representing years 2016 to 2018.

5 Evaluation

The evaluation of models implemented based on methods mentioned in Section 4 is explained in this section.

5.1 Pattern Identification

The patterns distribution generated by SOM model is defined in a clusters plot where each node contains the codes or scaled variable segmentation involved in a node. The optimum number of clusters for the SOM model is 5 as identified through the elbow point of within sum of squares plot. It can be observed in Figure 5 the nodes are clustered based on different proportion of dengue cases with various weather conditions. From this point onwards, temperature variables refer to actual, minimum and maximum temperature otherwise mentioned.

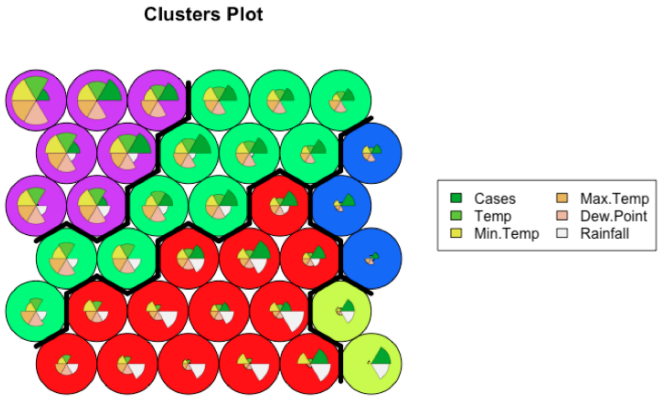


Figure 5: Clustering of data based on SOM Model

By analyzing Figure 5, there are 5 conditions where dengue cases could occur in Malaysia. For the purple cluster, low to high proportion of cases occur when there is a combination of moderate to high temperature with minimal to no rainfall. In green where temperature and dew point are in average scale with none to minimal rainfall leads to diverse number of cases. The red cluster signifies rain as a source of cases even at below average temperature and dew conditions. The lime green cluster shows cases to occur during rainfall with slight influence of temperature. It can be assumed that other factors are causing cases within the blue cluster as there is almost distinct occurrence of weather conditions.

From observation, dengue cases occurs in Malaysia in all types of weather condition. This may be due to the fluctuating weather conditions that occurs in the tropical country. It is agreed with the research by Liu-Helmersson et al. (2016) that dengue outbreak observation is complex and may involve other external factors as it can be observed that there are dengue occurrences with slight influence of weather conditions as in some clusters. A detailed heatmap of the pattern distribution over variables and a K-Means clustering plot designed to validate the results of clustering from SOM can be found in the configuration manual to support findings from SOM.

5.2 Time Series Analysis

5.2.1 Count Data Model

A proper scoring rules assesses the sharpness of models where the small values of overall scoring indicate a better model (Liboschik et al.; 2015). The NB model highlighted in bold in Table 1 has smaller overall scoring as compared to the Poisson model. The AIC value of models was used as a model selector where the NB model has an AIC value of 1003.166 as compared to the AIC value of Poisson model at 12787.76. The NB model has been detected as the better model by proper scoring rules and AIC value.

Model	*log*	*qdrtc*	*sphrc*	*rnkd*	*dw*	*nsqe*	*sqe*
Poisson	Inf	-0.0012	NaN	4.4325	210.7582	202.4995	1150773
NegBin	**8.1931**	**-0.0009**	**-598698533**	**2.9116**	**14.5678**	**0.8500**	**1150773**

Table 1: Proper scoring rules calculated for models. Scoring rules compared are logarithmic, quadratic, spherical, ranked probability, Dawid-Sebastiani score, normalized squared error and squared error

Probability Integral Transform (PIT) histogram was also observed and from the observation the NB model is better fitted. From the example by Christou and Fokianos (2015) on Measles dataset the authors analyzed the shape of the PIT histogram where they mentioned a uniform shape histogram indicates a well calibrated model. The PIT histogram can be found in the configuration manual.

The outcome of NB model as a better fitting model over Poisson model is in line with the authors in Section 2.2.1 whom advocated the use of NB model for over dispersed data. The NB model is fitted into a prediction model with weather variables where the overall predicted values are notably higher than the actual dengue cases for 2016 and 2017.

5.2.2 SARIMA

From comparison of the AIC value for the automatically modelled ARIMA(0,1,2)(0,1,1) and manual modelled ARIMA(1,1,1)(1,1,0), it was observed that the first model has a lower AIC value of 837.64. This is an indicator that the first model would be a better SARIMA model. However, the second criteria that will be taken into consideration to select a better SARIMA model is by the predicted values of the model.

Observing Figure 7, the manual SARIMA model in red line fits closer to the actual observation line for the years 2016 to 2017. A third inspection by comparing the MAPE accuracy value of the trained SARIMA model against the testing model. Based on the

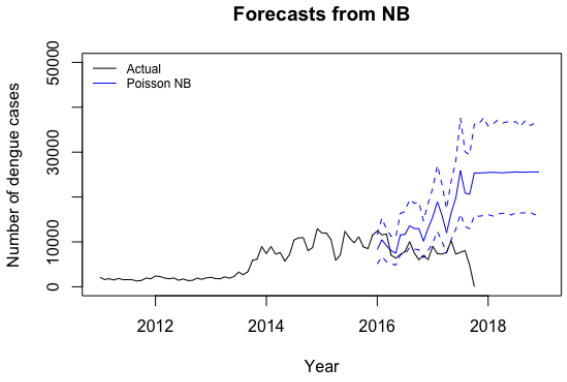


Figure 6: Forecasted dengue cases by NB model

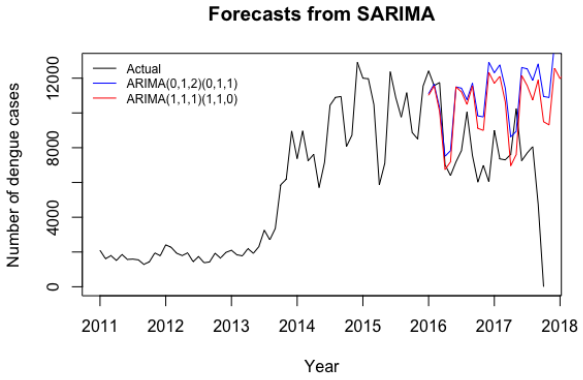


Figure 7: Forecasted dengue cases by SARIMA model

forecast accuracy measures, the observed is on the MAPE where ARIMA(1,1,1)(1,1,0) has a lower value at 14.8974 as compared to the other model at 14.93497. The findings are in line with the findings by Johansson et al. (2016) where values forecasted by a SARIMA increases substantially over a seasonal period.

5.2.3 Holt-Winters

A univariate time series data for dengue cases from 2011 to 2015 was fitted into the HW model. As mentioned in Section 4.2.4 there is some trend and seasonal observation detected in the fitted model. The model fitted was forecasted into a period of 36 months and produced the chart as in Figure 8. It can be observed that the model produced an 80% and 95% confidence level of forecast values in dark and light grey respectively. There is an upward trend over the years with a seasonal pattern of forecast observations. It is observed that there is some similarity of the seasonal pattern between the actual and forecasted lines.

The observation of HW made by Ho and Ting (2015) on the weightage of components for this algorithm is agreed by the α at 0.4733555 generated by the model. The observation made by the authors on increased dengue cases in fourth quarter of the year could also be acknowledged. The forecasted value from the HW model indicates that the highest point of dengue observation is towards the fourth quarter of years.

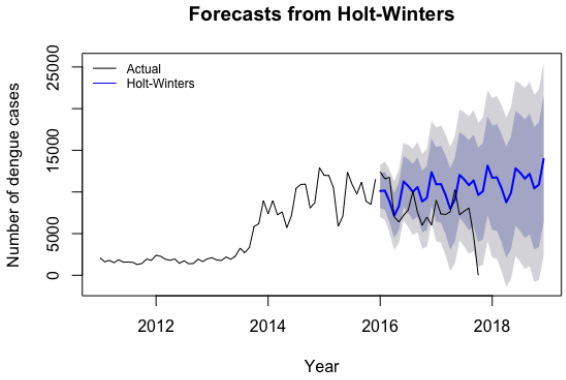


Figure 8: Forecasted dengue cases by HW model

5.3 Evaluation of Time Series Models

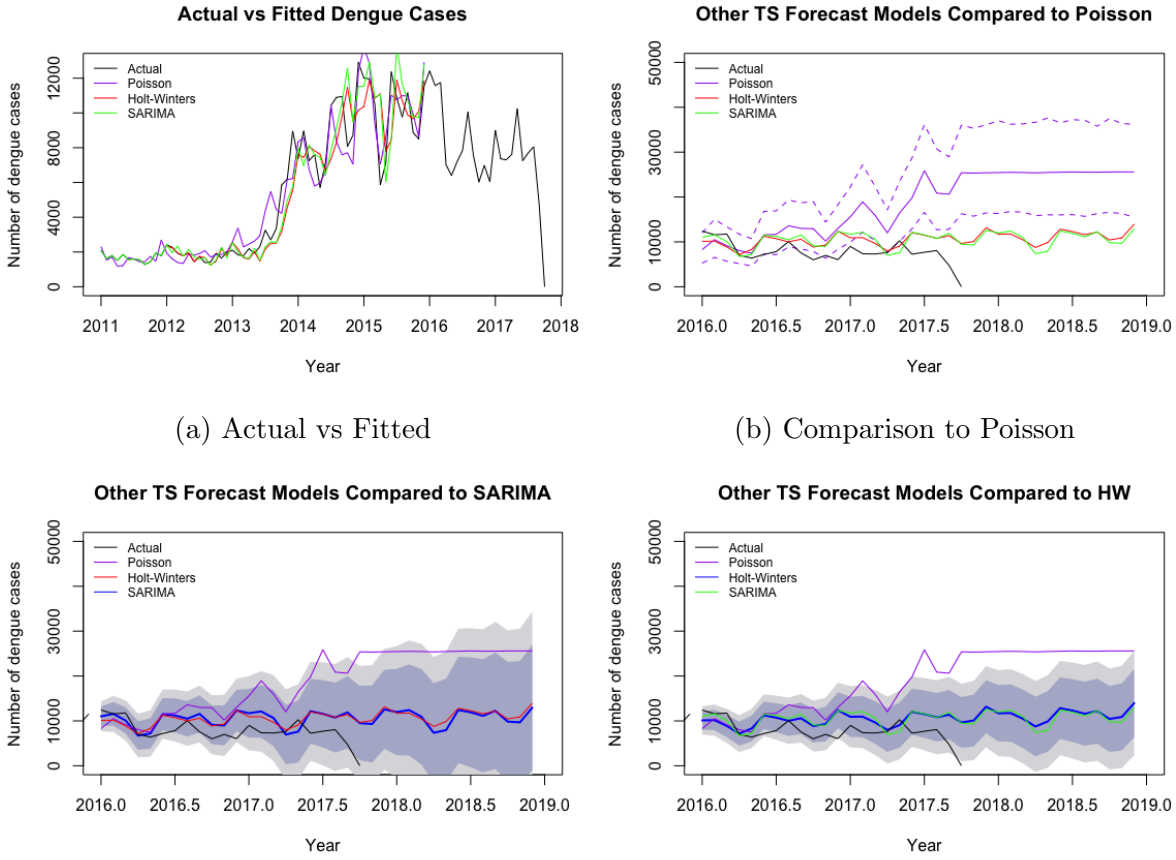
Observing Figure 9, the models fitted value has a similar line pattern as the actual observations. HW and SARIMA has an actual fit of lines for 2011 to mid 2012. Poisson model fits at relatively high value from 2011 to beginning of 2014 and fitted reasonably accurately from 2014 to 2016. Figure 9b, Figure 9c and Figure 9d are comparisons of time series models. It can be observed that Poisson model in figure has the highest upper forecast value as compared to other models. The Poisson model also forecasted a quite stagnant occurrence of dengue case in 2018 as compared to SARIMA and HW model. SARIMA and HW models both forecasted a seasonal pattern like the seasonal plot Figure 3. SARIMA has a wider forecast range for confidence level 80% and 95% while the 80% confidence level for HW fitted closely to the actual observations.

Accuracy measures RMSE, MAE and MAPE were calculated to compare the 3 time series models to conform the best forecasting model. Comparing the accuracy measures in Table 2, it can be seen that the RMSE for NB is the highest at 9695.453 as compared to SARIMA and HW at 3767.15 and 3933.765 which indicates that the model produced average highly deviated predictions from actual values. This is the first indication where the NB model is not suitable as a predictor as less deviation from actual observation is ideal. The MAE for NB is the highest among all models explains the model has the highest absolute difference between predicted and actual as based on the analysis in the previous paragraph.

Model	*RMSE*	*MAE*	*MAPE*	*MASE*
NB	9695.453	7564.093	Inf	NA
SARIMA	3933.765	3174.7137	Inf	1.3552794
HW	3767.15	3051.305	Inf	1.3025965

Table 2: Calculated accuracy for time series models

From RMSE and MAE evaluations it is deduced the NB model is not robust as a predicting model. SARIMA and HA models both has Inf value for MAPE which indicates there is an error in the accuracy measurement. Hyndman and Koehler (2006) recommends to use the Mean Absolute Scaled Error (MASE) to evaluate models when Inf appears in MAPE. Slightly lower MASE of HW at 1.3025965 indicates that HW is a better fitting model than SARIMA where MASE is 1.3552794. The low observation



(a) Actual vs Fitted

(b) Comparison to Poisson

(c) Comparison to SARIMA

(d) Comparison to HW

Figure 9: Plots to compare actual with fitted or forecast values of time series models

values in the accuracy measures for HW demonstrated it as an overall better time series forecasting model without the inclusion of external variables of weather.

6 Conclusion and Future Work

This research compared 3 time series models - Poisson, SARIMA and HW for prediction and forecasting of dengue cases in Malaysia with inclusive assessment of weather towards the models. The models generated a similar upward trend and seasonal pattern for prediction over the course of 3 years from 2016 to 2018 albeit producing vast differences of predicted values. HW and SARIMA were better fit as compared to the NB model that was fitted with weather variables. From the observation of model fits, it could indicate that there may be an indirect relationship between climate and dengue cases. It is efficient to develop the HW model as the algorithm considers seasonal smoothing components automatically like explained by Sood and Jain (2017). The efficiency of implementing the HW model is evident that the model could be used in implementing a dengue surveillance system as it would be cost effective.

SOM is helpful to create visualizations for understanding how variables influence each other. The clusters generated from the SOM model acts as a base of knowledge towards the patterns of dengue cases occurrences in Malaysia. It can be learned that dengue cases could happen in various weather conditions and due to other reasons. The findings

from Him et al. (2012) on influence of previous 3 months rainfall towards dengue could support the reasoning of breeding of mosquitoes in water containers or stagnant water found by Nalongsack et al. (2009). The findings from the pattern analysis could assist in creating new strategies in fighting dengue fever transmission and Aedes mosquito breeding.

The vast difference in accuracy measures for the models could be due to the insufficient data for the hold-out dataset. It is recommended to collect a larger dataset that would include weather and rainfall data than sourcing them separately as it would assist in effectively aggregating the data where required. This topic area can further be explored such as using the Seasonal Naïve method for time series forecasting as simpler models may generate a more accurate result. This can be seen as HW was predicting more accurate values as compared to the other models.

References

Ahdika, A. and Lusiyana, N. (2017). Comparison of inar (1)-poisson model and markov prediction model in forecasting the number of dhf patients in west java indonesia, *Journal of Physics: Conference Series*, Vol. 814, IOP Publishing, p. 012002.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle, *Selected Papers of Hirotugu Akaike*, Springer, pp. 199–213.

Armstrong, J. S. (1985). From crystal ball to computer, *New York* .

Augustijn, E.-W. and Zurita-Milla, R. (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns, *International journal of health geographics* **12**(1): 60.

Azam, M. N., Yeasmin, M., Ahmed, N. U. and Chakraborty, H. (2016). Modeling occurrence of dengue cases in malaysia, *Iranian journal of public health* **45**(11): 1511.

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview, *IADS-DM* .

Bação, F., Lobo, V. and Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering, *Computational Science–ICCS 2005* pp. 9–28.

Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M. (1976). *Time series analysis: forecasting and control*, John Wiley & Sons.

Briët, O. J., Vounatsou, P., Gunawardena, D. M., Galappaththy, G. N. and Amerasinghe, P. H. (2008). Models for short term malaria prediction in sri lanka, *Malaria Journal* **7**(1): 76.

Carvalho, m., Sivanandam, H., Shagar, L. K. and Ghazali, R. (2017). Dengue most contagious disease since 2015, `https://www.thestar.com.my/news/nation/2017/11/28/dengue-most-contagious-disease-since-2015/`. Accessed: 2017-11-29.

CCKR, C. C. K. P. (2015). Rainfall dataset. Accessed: 2017-10-20.
URL: *http://sdwebx.worldbank.org*

Choi, Y., Tang, C. S., McIver, L., Hashizume, M., Chan, V., Abeyasinghe, R. R., Iddings, S. and Huy, R. (2016). Effects of weather factors on dengue fever incidence and implications for interventions in cambodia, *BMC public health* **16**(1): 241.

Christou, V. and Fokianos, K. (2015). On count time series prediction, *Journal of Statistical Computation and Simulation* **85**(2): 357–373.

Dayama, P. and Kameshwaran, S. (2013). Predicting the dengue incidence in singapore using univariate time series models, *AMIA Annual Symposium Proceedings*, Vol. 2013, American Medical Informatics Association, p. 285.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American statistical association* **74**(366a): 427–431.

Earnest, A., Tan, S. B., Wilder-Smith, A. and Machin, D. (2012). Comparing statistical models to predict dengue fever notifications, *Computational and mathematical methods in medicine* **2012**.

Fairos, W. W., Azaki, W. W., Alias, L. M. and Wah, Y. B. (2010). Modelling dengue fever (df) and dengue haemorrhagic fever (dhf) outbreak using poisson and negative binomial model, *Int J Math, Comput, Phys Electr Comput Eng* **4**: 46–51.

Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., La Ruche, G., Girdary, L. and Marrama, L. (2011). Time series analysis of dengue incidence in guadeloupe, french west indies: forecasting models using climate variables as predictors, *BMC infectious diseases* **11**(1): 166.

Hasan, S. and Md, M. N. (1998). Spatial data clustering based on self organizing map.

Hii, Y. L., Zhu, H., Ng, N., Ng, L. C. and Rocklöv, J. (2012). Forecast of dengue incidence using temperature and rainfall, *PLoS neglected tropical diseases* **6**(11): e1908.

Him, N. C., Bailey, T. C. and Stephenson, D. B. (2012). Climate variability and dengue incidence in malaysia.

Ho, C. C. and Ting, C.-Y. (2015). Time series analysis and forecasting of dengue using open data, *International Visual Informatics Conference*, Springer, pp. 51–63.

Holt, C. (1957). Forecasting trends and seasonals by exponentially weighted averages. carnegie institute of technology, *Technical report*, Pittsburgh ONR memorandum.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy, *International journal of forecasting* **22**(4): 679–688.

Johansson, M. A., Cummings, D. A. and Glass, G. E. (2009). Multiyear climate variability and dengueel nino southern oscillation, weather, and dengue incidence in puerto rico, mexico, and thailand: a longitudinal data analysis, *PLoS medicine* **6**(11): e1000168.

Johansson, M. A., Reich, N. G., Hota, A., Brownstein, J. S. and Santillana, M. (2016). Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for mexico, *Scientific reports* **6**.

Kane, M. J., Price, N., Scotch, M. and Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks, *BMC bioinformatics* **15**(1): 276.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological cybernetics* **43**(1): 59–69.

Latif, Z. A. and Mohamad, M. H. (2015). Mapping of dengue outbreak distribution using spatial statistics and geographical information system, *Information Science and Security (ICISS), 2015 2nd International Conference on*, IEEE, pp. 1–6.

Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? then what?, *PloS one* **12**(8): e0183250.

Liboschik, T., Fokianos, K. and Fried, R. (2015). *tscount: An R package for analysis of count time series following generalized linear models*, Universitätsbibliothek Dortmund.

Liu-Helmersson, J., Quam, M., Wilder-Smith, A., Stenlund, H., Ebi, K., Massad, E. and Rocklöv, J. (2016). Climate change and aedes vectors: 21st century projections for dengue transmission in europe, *EBioMedicine* **7**: 267–277.

Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R., Coelho, C. A., Carvalho, M. S. and Barcellos, C. (2009). Climate-based dengue predictions for brazil.

Lu, L., Lin, H., Tian, L., Yang, W., Sun, J. and Liu, Q. (2009). Time series analysis of dengue fever and weather in guangzhou, china, *BMC Public Health* **9**(1): 395.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA., pp. 281–297.

Martinez, E. Z. and Silva, E. A. S. d. (2011). Predicting the number of cases of dengue infection in ribeirão preto, são paulo state, brazil, using a sarima model, *Cadernos de saude publica* **27**(9): 1809–1818.

Martinez, E. Z., Silva, E. A. S. d. and Fabbro, A. L. D. (2011). A sarima forecasting model to predict the number of cases of dengue in campinas, state of são paulo, brazil, *Revista da Sociedade Brasileira de Medicina Tropical* **44**(4): 436–440.

Mathulamuthu, S. S., Asirvadam, V. S., Dass, S. C., Gill, B. S. and Loshini, T. (2016). Predicting dengue incidences using cluster based regression on climate data, *Control System, Computing and Engineering (ICCSCE), 2016 6th IEEE International Conference on*, IEEE, pp. 245–250.

MOH, M. o. H. M. (2017). Dengue dataset. Accessed: 2017-09-18.
URL: *http://jkt.kpkt.gov.my*

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

Mutheneni, S. R., Mopuri, R., Naish, S., Gunti, D. and Upadhyayula, S. M. (2016). Spatial distribution and cluster analysis of dengue using self organizing maps in andhra pradesh, india, 2011–2013, *Parasite Epidemiology and Control* .

Naish, S., Dale, P., Mackenzie, J. S., McBride, J., Mengersen, K. and Tong, S. (2014). Climate change and dengue: a critical and systematic review of quantitative modelling approaches, *BMC infectious diseases* **14**(1): 167.

Nalongsack, S., Yoshida, Y., Morita, S., Sosouphanh, K. and Sakamoto, J. (2009). Knowledge, attitude and practice regarding dengue among people in pakse, laos.

NCDC, N. C. D. C. (2017). Weather dataset. Accessed: 2017-10-07.
URL: *https://www7.ncdc.noaa.gov/CDO/dataproduct*

Nguyen, C. (2015). Dengue escalates in southeast asia, `http://www.healthmap.org/site/diseasedaily/article/dengue-escalates-southeast-asia-62015`. Accessed: 2017-11-29.

Nury, A., Koch, M. and Alam, M. (2013). Time series analysis and forecasting of temperatures in the sylhet division of bangladesh, *4th International Conference on Environmental Aspects of Bangladesh (ICEAB), August*, pp. 24–26.

Ooi, E.-E. and Gubler, D. J. (2009). Dengue in southeast asia: epidemiological characteristics and strategic challenges in disease prevention, *Cadernos de saude publica* **25**: S115–S124.

Pinto, E., Coelho, M., Oliver, L. and Massad, E. (2011). The influence of climate variables on dengue in singapore, *International journal of environmental health research* **21**(6): 415–426.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests, *Systematic biology* **53**(5): 793–808.

Ramadona, A. L., Lazuardi, L., Hii, Y. L., Holmner, Å., Kusnanto, H. and Rocklöv, J. (2016). Prediction of dengue outbreaks based on disease surveillance and meteorological data, *PloS one* **11**(3): e0152688.

Shi, Y., Liu, X., Kok, S.-Y., Rajarethinam, J., Liang, S., Yap, G., Chong, C.-S., Lee, K.-S., Tan, S. S., Chin, C. K. Y. et al. (2016). Three-month real-time dengue forecast models: an early warning system for outbreak alerts and policy decision support in singapore, *Environmental health perspectives* **124**(9): 1369.

Silawan, T., Singhasivanon, P., Kaewkungwal, J., Nimmanitya, S. and Suwonkerd, W. (2008). Temporal patterns and forecast of dengue infection in northeastern thailand, *Southeast Asian Journal of Tropical Medicine and Public Health* **39**(1): 90.

Song, Y., Wang, F., Wang, B., Tao, S., Zhang, H., Liu, S., Ramirez, O. and Zeng, Q. (2015). Time series analyses of hand, foot and mouth disease integrating weather variables, *PloS one* **10**(3): e0117296.

Sood, S. and Jain, K. (2017). Comparative analysis of techniques for forecasting tourists arrival, *J Tourism Hospit* **6**(285): 2167–0269.

WHO (2017). Dengue and severe dengue, `http://www.who.int/mediacentre/factsheets/fs117/en/`. Accessed: 2017-11-29.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages, *Management science* **6**(3): 324–342.