Senior Project

Predicting Dengue-Virus Disease Spread

Pasin Sirirat 593 0351421

Committee Members

1. Asst.Prof Daricha Sutivong, Ph.D (Advisor)
2. Assoc.Prof Angsumalin Senjuntichai, D.Eng
3. Asst.Prof Surapong Sirikulvadhana

Department of Industrial Engineering

Faculty of Engineering,

Chulalongkorn University

Academic Year 2019

Abstract

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. The disease can cause bleeding, low blood pressure, and even death in severe cases. In recent years the fever has been spreading in many parts of the world, especially in Southeast Asia, Latin America, and the Pacific Islands. Because it is carried by mosquitoes, the transmission dynamics of dengue are possibly related to various climate variables such as temperature and precipitation. This study aims to predict the number of dengue cases each week in two locations which are the towns of San Juan, Puerto Rico, and Iquitos, Peru, based on environmental variables describing changes in temperature, precipitation, vegetation index, and more. In this study, three prediction techniques are used, Time-Series Forecasting, Machine Learning, and Neural Network. The results indicate that the Huber Regression and the Gradient Boosting Regression, which are two of the Machine Learning models used in this study, are the best prediction models for San Juan and Iquitos, respectively.

I. Introduction

In recent decades, several disease outbreaks have occurred in all continents over the world, some are already eliminated, but others are not. In the past few years, the Dengue-virus disease outbreak occurs in a lot of countries in South and South East Asia, Africa, Central and South America, which covers nearly a half of the world's population. Each year, up to 400 million people get infected with dengue, approximately 100 million people get sick from infection, and 22,000 die from severe symptoms. According to a number of disease control organizations around the world, the number of Dengue-virus incidences has grown dramatically in the past few decades, and those organizations are currently keeping their eyes on this disease.

Dengue fever is a mosquito-borne disease caused by dengue virus (DENV), which consists of four DENV serotypes, DENV1, DENV2, DENV3, and DENV4, meaning that it is possible to be infected four times in a human life. About half of the world's population in over 100 countries is now at risk. Dengue viruses spread to people through the bites of infected Aedes species mosquitoes (Ae. Aegypti or Ae. Albopictus). These mosquitoes prefer biting people during the day and night, and live both indoors and outdoors. The symptoms of dengue fever vary from the mild level with eye and muscle pain, headache, rash, nausea, and vomiting, to the severe level with bleeding to death.

This study uses the data from the DrivenData online competition, in which competitors try to predict the number of dengue case occurrences in two cities which are in the areas at risk of dengue. The first one is San Juan, the capital city of Puerto Rico, and the other is Iquitos, the largest metropolis in the Amazon rainforest region of Peru. In this paper, a number of methods from Time-Series Forecasting and Machine Learning are implemented, such as the Autoregressive Integrated Moving Average (ARIMA), which is a time-series forecasting model, and ten Machine Learning algorithms. The main objectives of this project are:

1. To compare prediction techniques between the Time-Series Forecasting and the Machine Learning methods.

2. To construct the prediction model to forecast the number of dengue case occurrences.

II. Related Works

Using Machine Learning techniques and Time-Series Forecasting in the medical field such as predicting the number of occurrences or classifying the severity level of a patient has been the interests of many researchers. In the recent years, there have been several studies implementing both mathematical approaches in the medical area.

Relating to dengue fever in particular, various research works applied mathematical procedures to explain the behavior of mosquitoes. Hancock et al. [1] studied the spatial dynamics of Wolbachia infections in Aedes aegypti arbovirus, the arthropod-borne virus, in heterogeneous landscapes. The study resulted in a mathematical model describing mosquito fitness components starting from estimating the duration of each stage of a mosquito. i.e. the egg, larvae, pupae, and mosquito. The study then predicted the rates of spatial spread of Wolbachia and compared the results with the observations from field populations.

Climate data are used in several studies on disease spread prediction along with other data. Sangwon et al. [2] built a number of time-series and Machine Learning models to predict case occurrences of the Malaria, Scarlet fever, and Chickenpox, in South Korea by using climate data along with big data in Twitter. The study used four different techniques, namely the Ordinary Least Square (OLS) Linear Regression, the Autoregressive Integrated Moving Average (ARIMA), the Deep Neural Network (DNN), and the Long-Short Term Memory (LSTM). The time-series data of occurrences of each disease was not the same, the chickenpox and scarlet fever did not have a trend but had a fluctuation of variance, whereas the malaria had an upward trend. The result of this study showed that the LSTM and DNN offered smaller error than ARIMA and OLS in the first two diseases, whereas in the malaria, the LSTM outperformed all other techniques. Anwar et al. [3] implemented the Panel Data Approach in forecasting the spread of dengue using climate and socio-economic variables in South and South-East Asia. In this study, the Panel Data Approach was employed to investigate the effect of climate change, i.e. the fluctuation of temperature, on the number of dengue reported cases. The result indicated that the dengue reported cases were significantly correlated to the climate change. Jia et al. [4] predicted the outbreak of dengue fever using climate factors in the Extreme Gradient Boosting (XGBoost) model. The result showed that the optimized XGBoost model consisted of all climate factors used in the study.

III. Exploratory Data Analysis

A. Problem Description

DrivenData is an organization which acts as a data science solution firm. The company regularly hosts a competition using data from its customers. In the DengAI competition, the data used for prediction comes from two sources. The first one is the National Oceanic and Atmospheric Administration (NOAA), who provides the information about climate such as temperature, precipitation, vegetation index, etc.. The other one is the Centers for Disease Control and Prevention (CDC), who provides the number of dengue case occurrences. The original data in the competition consists of two datasets. The training and testing dataset. The training dataset provides the target variable for competitors to construct their models. The testing dataset requires competitors to predict the target variable, submit the prediction result into the DrivenData website, to receive a score.

B. Data Source

The data from the competition website is divided into the training dataset and testing dataset. The training dataset has 1,456 records whereas the testing dataset contains 416. Both datasets consist of observations from two cities. Specifically, the training and testing dataset for the city of San Juan comprises of 936 and 260 records, respectively, whereas the training and testing dataset for the city of Iquitos comprises of 520 and 156 records, respectively. Both datasets contain 24 features which will be used in training models and predictions. The number of dengue case occurrences are weekly values, and all features in the initial datasets is numerical variables. All features are listed in Table I, which came from four sources:

- The Normalized Difference Vegetation Index (NDVI) which measures surface vegetation coverage activity via satellite image processing technique and has a value which ranges from -0.1 to 0.703. Values greater than 0.1 indicates the green color in the image captured by the satellite, values between 0 and 0.1 are commonly characteristics of rocks and bare soil, and values less than 0 shows that there are clouds, rain, or snow in the area. The sample image of the NDVI measurement is shown in Figure I.

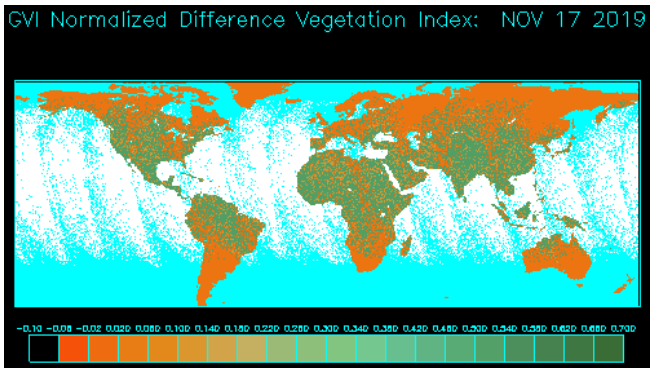

Figure I: Normalized Difference Vegetation Index (NDVI)

(Source: https://www.ospo.noaa.gov/Products/land/gvi/NDVI.html)

- The Precipitation Estimation from Remotely Sensed Information using an Artificial Neural Network (PERSIANN) measurement which is the amount of precipitation calculated from the satellite data. The rain rates are estimated using the GridSat-B1 IRWIN data, which is the infrared-captured image from a satellite, passed into the Artificial Neural Network model which results in a total precipitation. The sample of satellite image and the PERSIANN calculation of precipitation is shown in Figure II.

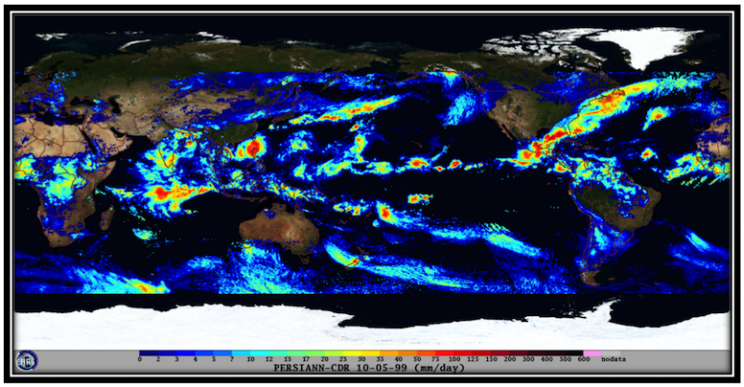

Figure II: NOAA's Satellite Image and PERSIANN Precipitation Calculation

(Source: https://www.ncdc.noaa.gov/cdr/atmospheric/precipitation-persiann-cdr)

- The NOAA's National Center for Environmental Prediction (NCEP)'s Reanalysis method, which is the procedure used to calculate climate indicators by considering how weather and climate are changing over time.
- The NOAA's Global Historical Climatology Network (GHCN) observations which gives some climate data collected in land surface weather stations.

Table I: Features used to build models

Source	Variables
NDVI	The NDVI in four parts of the city, Northeast, Northwest, Southeast, and Southwest
Persiann	Precipitation
NCEP	Specific Humidity, Dew Point Temperature, Average Temperature, Maximum Air Temperature, Minimum Air Temperature, Average Temperature, Diurnal Temperature Range, Relative Humidity, Precipitation from Past Satellite Data, Precipitation from Reanalysis Method
GHCN	Maximum Temperature, Minimum Temperature, Average Temperature, Diurnal Temperature Range, Precipitation

C. Data Exploration

The characteristics of the number of cases in San Juan and Iquitos are investigated with the time-series plot. Figure III shows the time-series plot of dengue case occurrences in both cities. The highest number of cases in San Juan is 461, occurred in October 1994 whereas the highest number of cases in Iquitos is 116, occurred in December 2004.

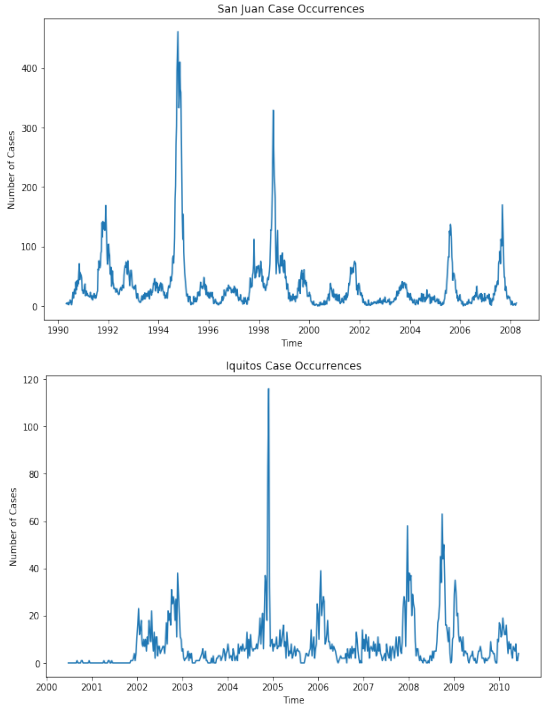

Figure III: Dengue occurrences in San Juan (upper) and Iquitos (lower) over time

The numbers of cases are then plotted as histograms to visualize the distribution of the variables, as shown in Figure IV. We can see that the histograms of the cases in both cities are positively skewed curves. Since some of our models are parametric methods, which means they require a target variable to be normally distributed, we will need to transform number of case variables before incorporating into the model.

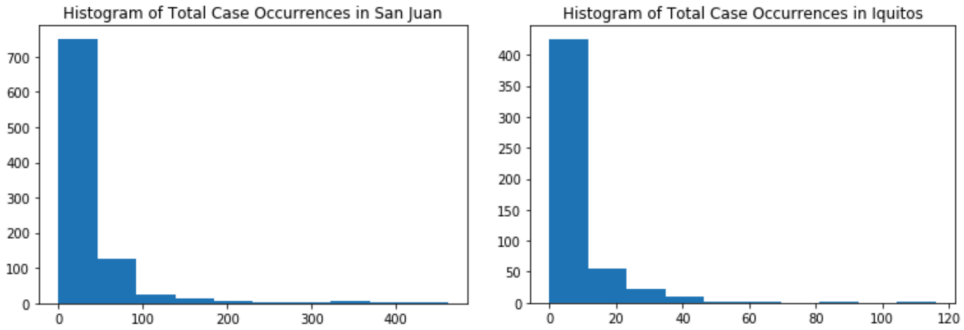

Figure IV: Histograms of number of cases in San Juan (left) and Iquitos (right)

D. Data Preprocessing

a. Missing Values

To address missing values in some independent variables, we analyze the average temperatures from the Reanalysis method and from the weather station in San Juan, as shown in Figure V. Note the seasonality over the year that is repeated every year. Therefore, we have decided to replace the missing value for each variable is replaced with its mean for the week.

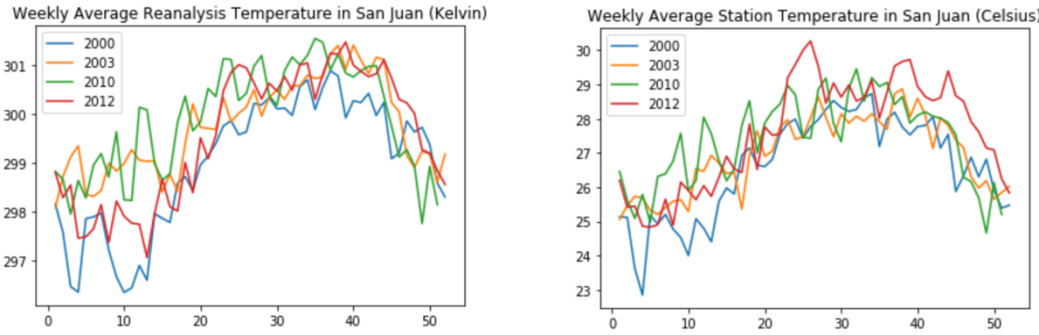

Figure V: Monthly average temperature from the weather station in San Juan

b. Data Transformation

As shown in Figure VI, the case occurrences variables need to be transformed. For the San Juan cases, we use the log1p-transformation method, whose the transformation equation is shown in Equation 1, because the variable has a number of

records in which the value is zero. For the Iquitos cases, after applying the log-1p transformation, the data is still not normalized. Hence, the log-transformed number of cases is then divided with the population data of the city in each year then multiplied by 100000, as shown in Equation 2. The histograms of both transformed variables are illustrated in Figure VI.

$$y'_{SanJuan} = \log\left(y_{SanJuan} + 1\right) \tag{1}$$

$$y'_{Iquitos} = \frac{100{,}000 * \log\left(y_{Iquitos} + 1\right)}{population} \tag{2}$$

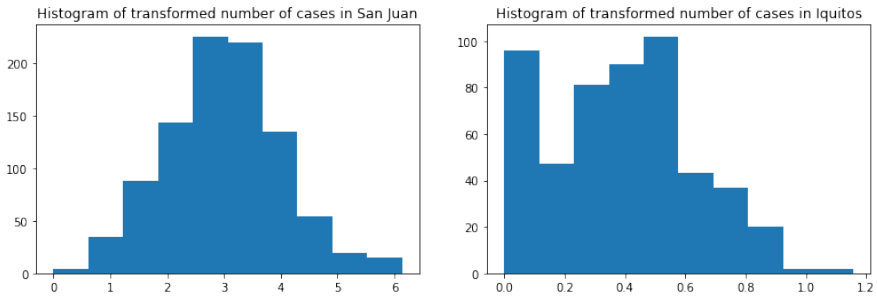

Figure VI: Histograms of transformed number of cases in San Juan (left) and Iquitos (right)

E. Performance Indicator

This study implements a number of Machine Learning and time-series techniques in order to forecast number of dengue case occurrences. To compare the results across methods, the Mean Absolute Error (MAE) is used. The MAE indicates how much a prediction from a model differs from an actual in terms of the average of the differences between the actual value and the respective prediction values. As shown in Equation 3, if a higher MAE means that the prediction is further from the actual, and the model is inaccurate.

$$MAE = \frac{\sum_{n=1}^{N} \left| x_{predict} - x_{observed} \right|}{N} \tag{3}$$

IV. Training Models

1. Lag Selection

Prediction is the process by which data in the past are used to predict the future, one of the important questions is how far we need to go back to the past and use the data then to predict the future. In this study, we implement the lag selection method to all Machine Learning and Deep Learning models to find the best lag.

2. Feature Selection Method

Using all features to build a model may not be beneficial since some factors are statistically uncorrelated to our target variable, which means that they will reduce the performance and increase the computation time of the process. Thus, we may remove some features that are not useful for our model to improve our efficiency.

Stepwise elimination is one of the popular methods to select only the features that are correlated to the target variable. The method implements Linear Regression, it starts from a general linear model with every factor, finds the factor which has the lowest significance of correlation with the target variable, and removes it if its p-value exceeding the Type-I error (which is set to 0.05 in this study). The removing process is iterated until no more factor has its p-value exceeds the Type-I error. In this study, we compare the result of Machine Learning models with and without feature selection.

3. Modeling

A. Time-Series Model

a. Autoregressive Integrated Moving Average

The Autoregressive Integrated Moving Average (ARIMA) is the time-series model used to forecast a time-series in which the random noise is not generated through independent shocks. That is, successive records in time-series show dependence. The ARIMA model comprises of two time-series model. The Moving-Average model with parameter q, denoted as MA(q), is the moving average model of the error. The other one is the Autoregressive model with parameter p, denoted as AR(p). The Autoregressive model considers all values from the past. The model use weights with value between zero and one as the coefficient of a past term in the equation, presumably the value of the coefficient gets smaller when the term is older.

The AR and MA are then combined with another parameter d which is called the order of differentiation. This parameter is added to make the new time-series model robust to the non-stationary model.

B. Machine Learning Models

a. Neural Network

The Neural Network consists of two main components, a neuron and a linkage between two neurons. A Neural Network composes of three or more neuron layers. The first one is the input layer in which the number of neurons is equal to the number of features used in a Machine Learning problem. The second one is a set of one or more hidden layers. The last one is the output layer in which the number of neurons is equal to one if the network is used in a regression problem or equal to the number of classes if the network is used in a classification problem.

In this study, we implement the Neural Network with the Rectifier Linear Unit (ReLU) as an activation function in each layer, we also vary the number of layers and the number of nodes in a layer to optimize the network structure and its performance

b. Regularization Models

A regularization model is a regression model that has a penalty term to shrink the coefficient estimations towards zero. Regularization models are designed to penalize the complexity of the equation to avoid the risk of overfitting. In this study, four Regularization methods are explored, the Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Bayesian Ridge Regression, Kernel Ridge Regression, and Elastic-Net Regression.

c. Ensemble Models

Ensemble model is a Machine Learning model in which two or more other Machine Learning models are combined, especially the ones based on Decision Tree. This method gains its popularity in the recent years because it usually offers more accuracy and has less overfitting effect while keeping the easy interpretation characteristic of a single Decision Tree model. In this study, three ensemble models are investigated, Random Forests, Gradient Boosting, and Extreme Gradient Boosting (XGBoost).

d. Robust Regression Models

Robust Regression is one form of the regression analyses which is designed to overcome the limitations of traditional parametric regression methods. It is built to avoid pitfalls from certain data characteristics. For example, the ordinary least-square regression methods are highly sensitive to outliers. Instead, robust regression methods reduce the effect of outliers

in the prediction. In this study, two forms of Robust Regression methods are used, namely Huber Regression and Random Sample Consensus (RANSAC) Regression.

4. Hyperparameters Tuning (GridSearch)

In order to optimize a set of hyperparameters for each model, the GridSearch method is implemented. The method takes a list of hyperparameter values of interest and then generates combinations of every possible outcome of hyperparameter sets. Next, it feeds a hyperparameter set into a Machine Learning model, fine tunes it, and then returns a tuned model. The method does this for every combination created and then selects the best combination as a result.

V. Results

A. Time-Series Model

Figure VII shows the results from running ARIMA for both cities with different sets of hyperparameters. As a result, the best models we have got from the time-series model is ARIMA(4,1,1) for San Juan, and AR(5) from Iquitos, with the corresponding MAEs of 17.18 and 8.29, respectively.

In San Juan, we can see if the differencing order (d) is set to one, the error is lower than setting the differencing order to zero. In Iquitos, the models with the differencing order set to zero perform better than those with the differencing order equals to one. On the other hand, changing the parameters p and q of an ARIMA model barely affect the model's performance. This illustrates that there may be other factors driving the change in the magnitude of case occurrences rather than the past cases themselves in the past. Moreover, due to the fact that we need to convert the output value of our model to be an integer value (to tell us how many people get infected), and ARIMA itself has no constraint to force its output to be an integer, the error of the model increases, and the model does not perform well.

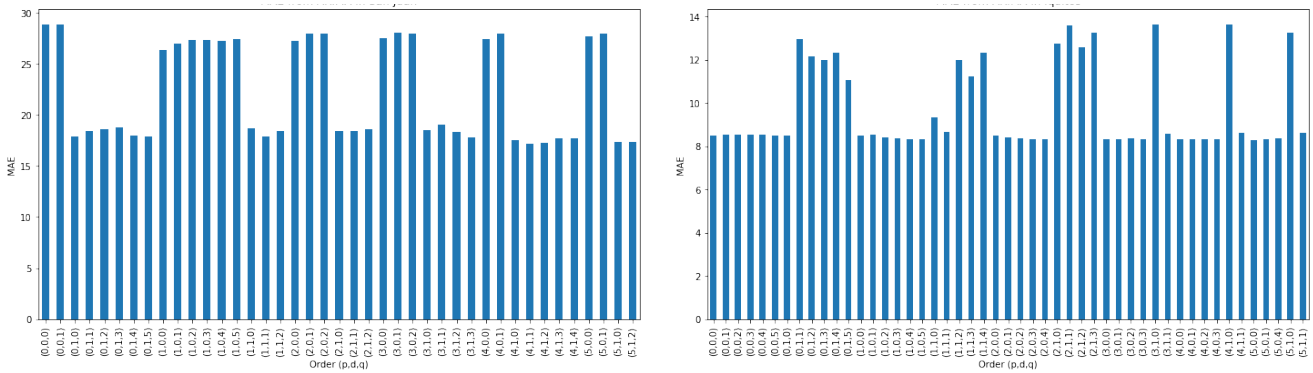

Figure VII: Mean Absolute Error of forecasting case occurrences using ARIMA in San Juan (left) and Iquitos (right).

B. Machine Learning Models

a. Neural Network

Figure VIII illustrates the Neural Network results. Three Neural Networks are trained using a lag from one to twenty for each city. The models differ in the number of hidden layers and the number of nodes in each layer. The first model, Model 1, is constructed with one hidden layer of 128 nodes. The second model, Model 2, uses three hidden layers with 64, 128, and 64 nodes, respectively. The last one, Model 3, contains five hidden layers with 64, 128. 256, 128, and 64 nodes, respectively. The best Neural Network model for San Juan uses the data at the lag of 1, yielding the MAE of 19.91, whereas the best model for Iquitos also uses the lag of 1, yielding the MAE of 16.38.

From the results, we can see that changing the number of hidden layers and the number of nodes in each hidden layer barely affect the performance of the Neural Network. A Neural Network generally requires substantial training data to achieve high accuracy; therefore, it may not be effective in this study.

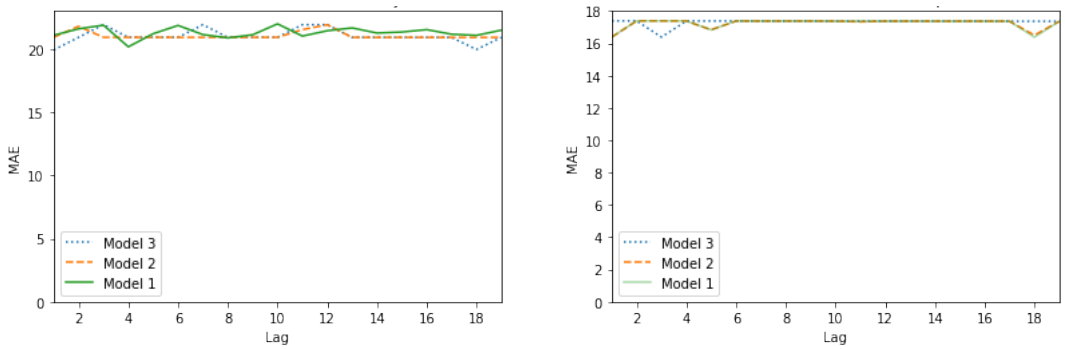

Figure VIII: Mean Absolute Error of forecasting case occurrences using Neural Network in San Juan (left) and Iquitos (right).

b. Regularization Models

Figure IX and X show the mean-absolute error from Regularization models in both two cities. For San Juan, if we do not use the stepwise elimination, the Elastic Net Regression with lag 5 is the best model with MAE of 15.7, but if we do the stepwise elimination, the Kernel Ridge Regression with lag 9 is the best model with MAE of 13.9. In Iquitos, if we do not use the stepwise elimination, the Kernel Ridge Regression with lag 10 is the best model with MAE of 8.74, but if we do the stepwise elimination, the Elastic Net Regression with lag 4 is the best model with MAE of 9.01.

We can see that the MAE of all methods are similar for both two cities. We further investigate the effect of the stepwise elimination on the performance of the models. Table II and III shows the difference of the mean absolute error of models with and without stepwise elimination for each algorithm and lag for both cities. If a number is positive, it means that the mean absolute error of a model with stepwise elimination is less than training such a model without stepwise elimination. We can see that most of the models perform better when using the stepwise elimination.

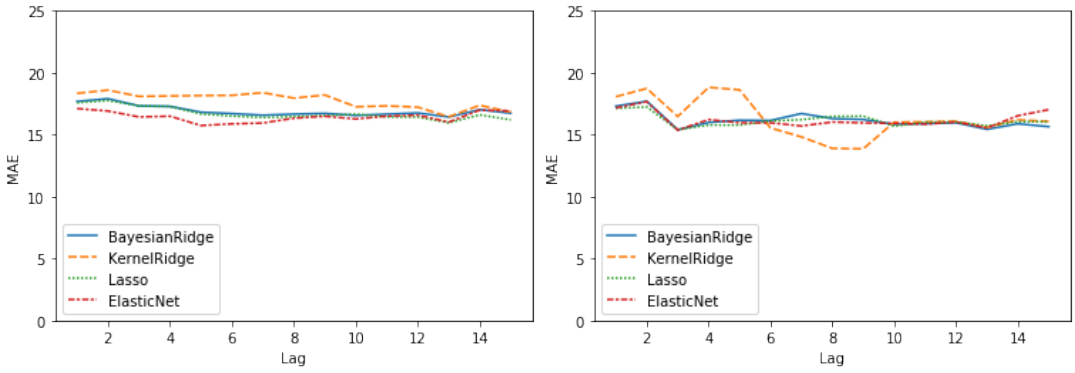

Figure IX: Mean Absolute Error of Regularization models by Lag in San Juan without feature selection (left) and with feature selection (right).

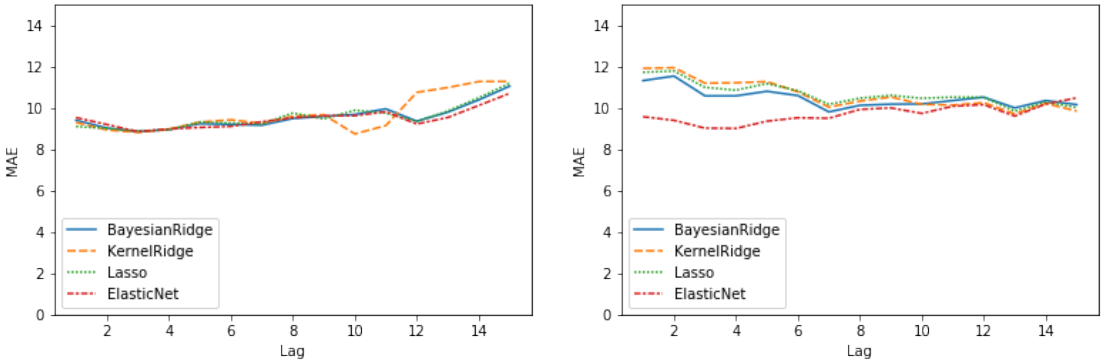

Figure X: Mean Absolute Error of Regularization models by Lag in Iquitos without feature selection (left) and with feature selection (right).

Table II: Difference of mean absolute error of Regularization models with and without stepwise elimination in San Juan

Algorithm	KernelRidge	BayesianRidge	Lasso	ElasticNet
Lag				
1	0.264	0.378	0.45	-0.057
2	-0.135	0.208	0.521	-0.757
3	1.623	1.924	1.925	1.062
4	-0.688	1.296	1.495	0.275
5	-0.461	0.65	0.898	-0.227
6	2.631	0.559	0.394	-0.09
7	3.57	-0.136	0.179	0.24
8	4.057	0.358	-0.018	0.308
9	4.344	0.491	0.043	0.53
10	1.245	0.716	0.971	0.349
11	1.284	0.755	0.439	0.655
12	1.14	0.788	0.403	0.547
13	0.856	0.993	0.242	0.473
14	1.173	1.137	0.542	0.458
15	0.751	1.087	0.177	-0.108

Table III: Difference of mean absolute error of Regularization models with and without stepwise elimination in Iquitos

Algorithm	KernelRidge	BayesianRidge	Lasso	ElasticNet
Lag				
1	-1.619	-0.922	-1.623	0.957
2	-2.014	-1.52	-1.779	0.797
3	-1.382	-0.732	-1.107	0.829
4	-1.25	-0.625	-0.925	0.971
5	-0.961	-0.568	-0.861	0.696
6	-0.355	-0.398	-0.566	0.584
7	0.211	0.358	0.05	0.832
8	0.272	0.366	0.28	0.595
9	0.143	0.409	-0.143	0.616
10	-0.432	0.489	0.421	0.888
11	0.007	0.597	0.23	0.737
12	1.496	-0.162	-0.165	0.061
13	2.278	0.791	1.004	0.935
14	2.058	1.032	1.206	0.928
15	2.437	1.895	2.177	1.22

c. Ensemble Models

Figure XI and XII show the mean-absolute error from Ensemble models in both two cities. For San Juan, if we do not use the stepwise elimination, the Gradient Boosting Regression with lag 15 is the best model with MAE of 14.9. On the other hand, if we do the stepwise elimination, the Gradient Boosting Regression with lag 15 is still the best model, but with MAE of 15.1. In Iquitos, if we do not use the stepwise elimination, the Random Forest Regression with lag 6 is the best model with MAE of 9.39, but if we do the stepwise elimination, the Gradient Boosting Regression with lag 15 is the best model with MAE of 8.32.

We can see that in San Juan, the Random Forest tends to perform better than the other two, although the difference of error decreases if we increase the number of lags. On the other hand, in Iquitos, the Random Forest and the Gradient Boosting do not perform differently from each other. In addition, we can see that the Extreme Gradient Boosting (XGB) generally have higher MAE than the other two models. This is because the XGB is a Machine Learning model which was built with more complex mathematical equations which requires a lot of data to capture complex patterns in the data. In this case, the amount of data used to train the model is not high enough to make the XGB learned and make a good prediction.

We also investigate the effect of the stepwise elimination on the performance of the models, similar to what we have done in the Regularization models. Table IV and V shows the difference of the mean absolute error of models with and without stepwise elimination for each algorithm and lag for both cities. We can see that most of the models perform better when using the stepwise elimination.

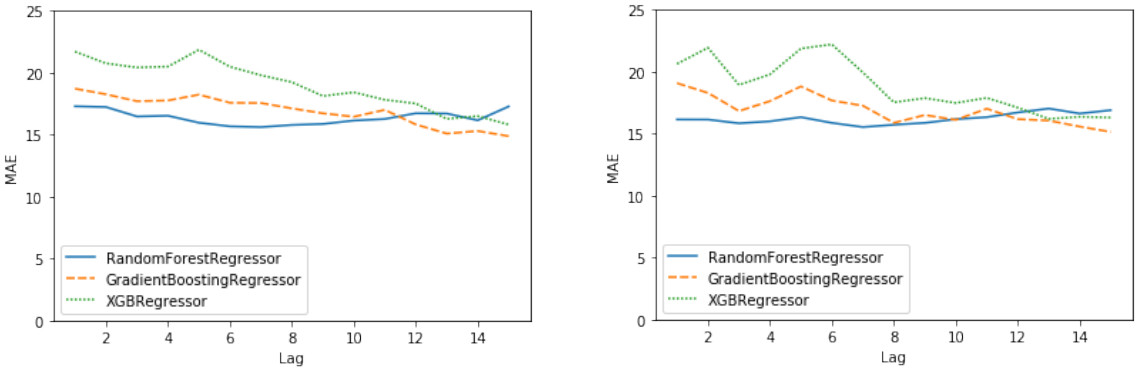

Figure XI: Mean Absolute Error of Ensemble models by Lag in San Juan without feature selection (left) and with feature selection (right).

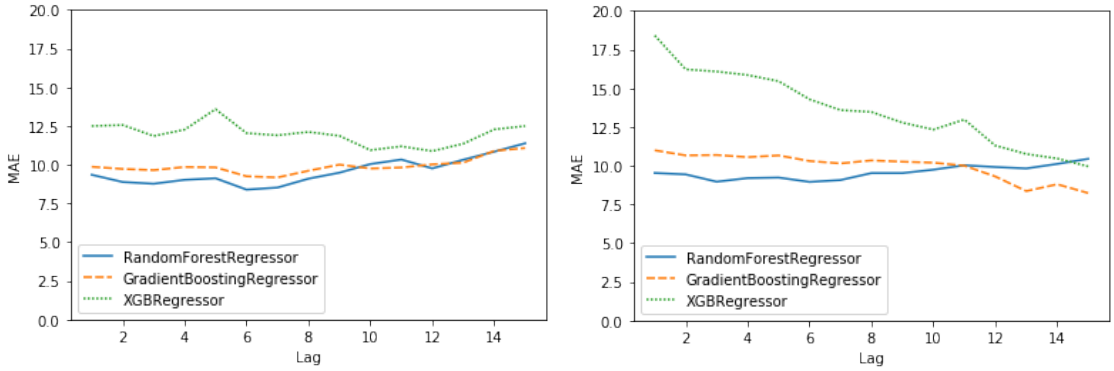

Figure XII: Mean Absolute Error of Ensemble models by Lag in Iquitos without feature selection (left) and with feature selection (right)

Table IV: Difference of mean absolute error of Ensemble models with and without stepwise elimination in San Juan

Algorithm	RandomForestRegressor	GradientBoostingRegressor	XGBRegressor
Lag			
1	1.147	-0.364	1.069
2	1.094	-0.028	-1.181
3	0.631	0.876	1.49
4	0.534	0.131	0.71
5	-0.368	-0.582	-0.018
6	-0.197	-0.097	-1.724
7	0.079	0.297	-0.14
8	0.065	1.247	1.71
9	-0.007	0.219	0.251
10	-0.029	0.335	0.942
11	-0.068	-0.014	-0.072
12	0.014	-0.349	0.414
13	-0.321	-0.968	0.079
14	-0.469	-0.278	0.134
15	0.383	-0.26	-0.491

Table V: Difference of mean absolute error of Ensemble models with and without stepwise elimination in Iquitos

Algorithm	RandomForestRegressor	GradientBoostingRegressor	XGBRegressor
Lag			
1	0.819	-0.128	-3.512
2	0.445	0.06	-1.256
3	0.804	-0.043	-1.829
4	0.825	0.3	-1.189
5	0.889	0.164	0.529
6	0.437	-0.05	0.151
7	0.452	0.029	0.71
8	0.573	0.269	1.047
9	0.961	0.746	1.484
10	1.299	0.558	1.007
11	1.306	0.817	0.608
12	0.845	1.719	1.975
13	1.498	2.769	3
14	1.726	3.09	4.213
15	1.939	3.841	4.942

d. Robust Regression Models

Figure XIII and XIV show the mean-absolute error from Robust Regression models in both two cities. For San Juan, if we do not use the stepwise elimination, the Huber Regression with lag 16 is the best model with MAE of 16.0, but if we do the stepwise elimination, the Huber Regression with lag 9 is the best model with MAE of 13.67. In Iquitos, if we do not use the stepwise elimination, the RANSAC Regression with lag 15 is the best model with MAE of 9.12, but if we do the stepwise elimination, the Huber Regression with lag 10 is the best model with MAE of 10.06.

We can see that the Huber Regression model outperforms the RANSAC Regression. This issue is due to the fact that the RANSAC Regression neglects all outlier data found in the training data, while the Huber Regression still keeps the outliers but give them less weight of calculation than non-outlier data.

We also investigate the effect of the stepwise elimination on the performance of the models, as what we have done previously. Table VI and VII shows the difference of the mean absolute error of models with and without stepwise elimination for each algorithm and lag for both cities. We can see that most of the models perform better when using the stepwise elimination.

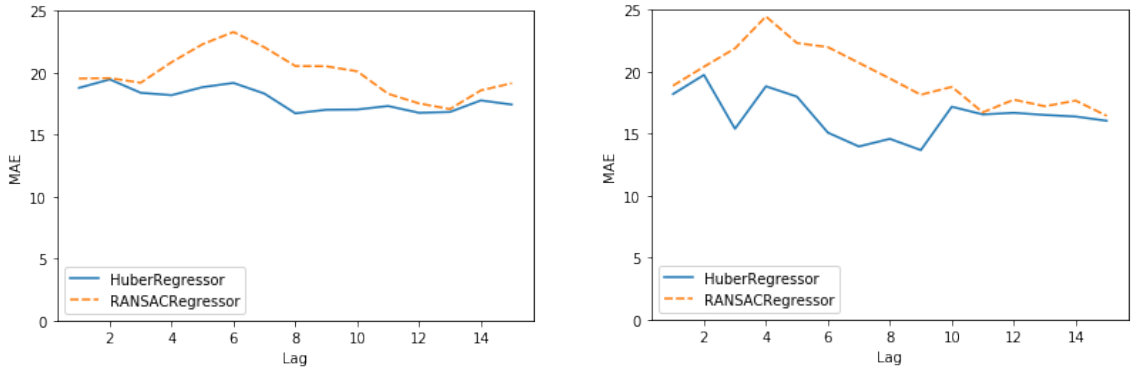

Figure XIII: Mean Absolute Error of Robust Regression models by Lag in San Juan without feature selection (left) and with feature selection (right).

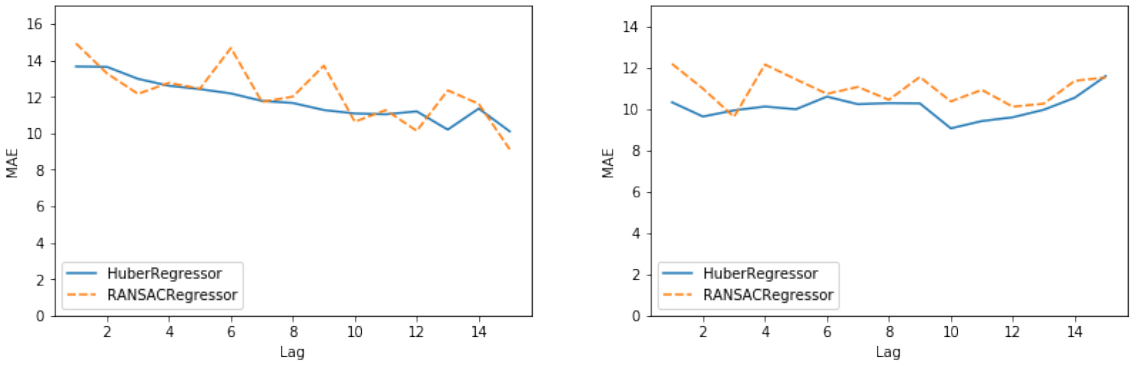

Figure XIV: Mean Absolute Error of Robust Regression models by Lag in Iquitos without feature selection (left) and with feature selection (right).

Table VI: Difference of mean absolute error of Robust Regression models with and without stepwise elimination in San Juan

Algorithm	HuberRegressor	RANSACRegressor
Lag		
1	0.575	0.661
2	-0.279	-0.837
3	2.987	-2.69
4	-0.637	-3.603
5	0.846	-0.018
6	4.093	1.301
7	4.348	1.337
8	2.14	1.09
9	3.337	2.38
10	-0.144	1.335
11	0.759	1.583
12	0.083	-0.223
13	0.321	-0.144
14	1.383	0.906
15	1.39	2.708

Table VII: Difference of mean absolute error of Robust Regression models with and without stepwise elimination in Iquitos

Algorithm	HuberRegressor	RANSACRegressor
Lag		
1	2.342	1.74
2	3.011	1.302
3	2.054	1.543
4	1.489	-0.382
5	1.436	0.014
6	0.588	2.946
7	0.545	-0.344
8	0.376	0.566
9	0.004	1.161
10	1.025	-0.734
11	0.629	-0.651
12	0.604	-0.978
13	-0.762	1.101
14	-0.188	-0.747
15	-2.487	-3.404

C. Models summary

Figure XV shows the performance of each algorithm which has the best performance. For San Juan, the best prediction model is the Huber Regression using the lag number equals to 9, which means the climate data in the 9^{th} previous week are used. The best model implements stepwise elimination method, which means using only features which are statistically correlated to the case occurrences is better than using all of the features, the model gives us the Mean Absolute Error of 13.67.

For Iquitos, the best prediction model is the Gradient Boosting Regression using the lag number of 15 and also using the stepwise elimination method. The model gives us the Mean Absolute Error of 8.23. The Huber Regression, which is the best Machine Learning algorithm in San Juan, performs better than the Time-Series model and Neural Network in its own city. In Iquitos, the Gradient Boosting Regression, which is the best Machine Learning model, outperforms the Neural Network but generates minimally lower Mean Absolute Error than the Time-Series model.

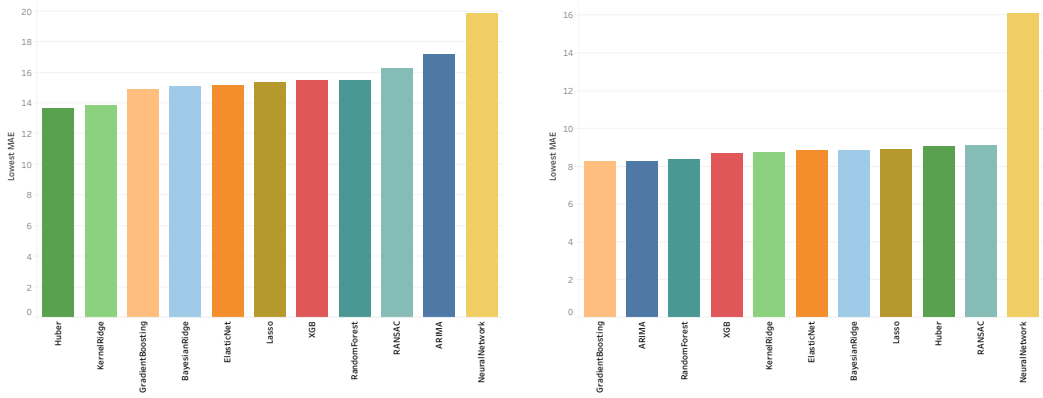

Figure XV: Bar graphs comparing the lowest Mean Absolute Error from each algorithm in San Juan (left) and Iquitos (right)

VI. Conclusion

Since this paper aims to construct the predictive models to forecast the number of Dengue Virus case occurrences in the cities of San Juan, Puerto Rico, and Iquitos, Peru. We use different prediction methods, namely Time-Series Forecasting and ten Machine Learning algorithms, along with data preprocessing techniques like log1p-transformation in order to transform the data into an appropriate feature for our prediction models. The results indicate that the more advanced Machine Learning algorithms outperform the Time-Series Forecasting, providing less error in both cities. Within the Machine Learning methods, for the city of San Juan, the Huber Regression with a lag of 9 and the stepwise elimination is shown to be the best model with the MAE of 13.67. For the city of Iquitos, the Gradient Boosting Regression with a lag of 15 and the stepwise elimination is the best model with the MAE of 8.23. From the best models of the two cities, we have found that the stepwise elimination method consistently improves the performance of the Machine Learning model. Moreover, a more advanced Machine Learning models tend to perform better than the Neural Network when the number of observations is limited.

References

[1] P. A. Hancock, S. A. Ritchie, C. J. M. Koenraadt, T. W. Scott, A. A. Hoffmann, H. C. J. Godfray, "Predicting the spatial dynamics of Wolbachia infections in Aedes aegypti arbovirus vector populations in heterogeneous landscapes, Journal of Applied Ecology, 56(7), pp. 1674-1686.

[2] S. Chae, S. Kwon, D. Lee, "Prediction infectious disease using deep learning and big data", International Journal of Environmental Research and Public Health 2018, 15, 1596.

[3] A. Anwar, N. Khan, M. Ayub, F. Nawaz, A. Shah, A. Flahault, "Modeling and predicting dengue incidence in highly vulnerable countries using panel data approach", International Journal of Environmental Research and Public Health, 16(13), 2296.

[4] N. Jia, X. Liao, J. Chen, X. Chen, J. Chen, G. Dong, G. Hu, "Using climate factors to predict the outbreak of dengue fever", 2018 7th International Conference on Digital Home (ICDH), pp.213-218.

[5] E. Pelaez, "A Fuzzy Cognitive Map (FCM) as a learning model for early prognosis of seasonal related virus diseases in tropical regions", IEEE 2019, pp. 150-156.

[6] S. Molaei, M. Khansari, H. Veisi, M. Salehi, "Predicting the spread of influenza epidemics by analyzing twitter messages", Health and Technology, 9(4), pp. 517-532.

[7] J. Kim, I. Ahn, "Weekly ILI patient ratio change prediction using news articles with support vector machine", BMC Bioinformatics, 20(1), 259.

[8] S. Chen, J. Xu, Y. Wu, X. Wang, S. Fang, J. Cheng, H. Ma, R. Zheng, Y. Liu, L. Zhang, X. Zhang, L. Chen, X. Liu, "Predicting temporal propagation of seasonal influenza using improved gaussian process model", Journal of Biomedical Informatics, 93, 103144.

[9] M. Gharbi, P. Quenel, J. Gustave, S. Cassadou, G. La Ruche, L. Girdary, L. Marrama, "Time series analysis of dengue incidence in Guadelopue, French West Indies: Forecasting models using climate variables as predictors", BMC Infectious Disease, 11, 166.

[10] Y. Zheng, L. Zhang, X. Zhang, K. Wang, Y. Zheng, "Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China", PLoS One, 10(3), e0116832.

[11] A. Mohammadinia, B. Saeidian, B. Pradhan, Z. Ghaemi, "Prediction mapping of human leptospirosis using ANN, GWR, SVM, and GLM approaches", BMC Infectious Diseases, 19(1), 971.

[12] J. Gao, H. Zhang, P. Lu, Z. Wang, "An effective LSTM recurrent network to detect arrythmia on imbalanced ECG dataset", Journal of Healthcare Engineering, 2019, 6320651.