Research Proposal

Title : ‘Forecasting Probability of Dengue Clusters in Singapore through Data Analysis’

Stuart Lim Yi Xiong 4S3(23) (Leader)

Si Wen Xuan, Terry 4S3(22)

Hwa Chong Institution

(High School)

Table of Contents

1 - Introduction **3**

1.1 - Introduction and Rationale 3

1.2 - Terminology 4

1.3 - Objectives 5

1.4 - Research Questions 5

1.5 - Field of Mathematics Involved 5

2 - Literature Review **6**

2.1 - Mapping Dengue Risk in Singapore using Random Forest 6

2.2 - Patterns of Urban Housing Shape Dengue Distribution in Singapore at Neighbourhood and Country Scales 6

2.3 - Takeaways 7

3 - Methodology / Procedure **8**

3.1 - Methodology 8

3.2 - Data Collection 8

3.3 - Data Analysis 9

4 - Results **11**

4.1 - Research Question 1 11

4.1.1 - Single Variable Regression Models 11

4.1.2 - Multi-Variable Regression Models 17

4.1.3 - Conclusion 19

4.2 - Research Question 2 19

4.2.1 - Single Variable Regression Model 20

4.2.2 - Multivariable Regression Model 21

4.3 - Research Question 3 22

5 - Conclusion **25**

5.1 - Summary and Conclusion 25

5.2 - Limitations and Possible Extensions 25

References **25**

1 - Introduction

1.1 - Introduction and Rationale

Dengue fever is a mosquito-borne disease caused by the dengue virus. It largely affects tropical regions such as Singapore. Bhatt S et al. (2013) estimated that there are 390 million dengue infections every year, with a 95% credible interval of 284 to 528 million. People with dengue infections typically only show mild symptoms, but in about 5% of all dengue cases, dengue fever may develop into dengue haemorrhagic fever or dengue shock syndrome, both of which are life-threatening illnesses (Ranjit & Kissoon, 2011). According to the National Environment Agency (NEA), in the year 2018, 3,285 dengue cases were reported in Singapore. Five people died due to complications resulting from dengue fever.

At the moment, no specific antiviral cures for dengue infections exist. The treatment given to those with dengue infections is supportive, and only seeks to improve the patient's comfort (Simmons et al, 2012). Thus, the most effective way to fight dengue is to prevent it.

In this project, we aim to improve local dengue prevention methods by analysing the factors at play in the formation of "dengue clusters" in Singapore. We believe that by isolating and determining the main causes, NEA can take better action to remove and prevent the formation of dengue clusters in Singapore.

1.2 - Terminology

Term	Definition
Dengue Cluster	A locality deemed to have active transmission of the dengue virus. According to NEA, a dengue cluster is an area in which two or more dengue cases occur within 14 days and are located within 150 metres of each other, with respect to residential and workplace addresses as well as movement history.
Regression Analysis	Regression analysis is a set of statistical processes for estimating the relationships among variables, used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.
Random Forest Algorithm	Random forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of *decision trees* at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
Decision Trees	A predictive model to go from observations about an item to conclusions about the item's target value

1.3 - Objectives

Through this project, we aim to determine the factors that correlate to the risk of dengue in a specific region in Singapore and create a prediction model to determine the risk level of a dengue cluster appearing, so as to improve local dengue prevention methods and increase the efficiency of the removal of dengue clusters.

1.4 - Research Questions

This project will be categorised into 3 main research questions.

Firstly, we will investigate the significance of environmental and physical factors in determining the degree of dengue infestation.

Secondly, we will investigate the impact of nearby dengue clusters on the degree of dengue infestation.

Thirdly, we will build a prediction model to forecast the risk level of dengue clusters in every 500 metre by 500 metre area in Singapore.

1.5 - Field of Mathematics Involved

The main field of mathematics involved will be Data Analysis, as well as Probability (in Research Question 3)

2 - Literature Review

2.1 - Mapping Dengue Risk in Singapore using Random Forest

Ong et al. (2018) predicted the risk of spread of dengue based on locational factors. Singapore was divided into 1 km^2 grids. Within each grid, the number of dengue cases was compared against dengue exposure (dengue burden in the grid and neighbouring grids, as well as number of non-resident cases in previous year), population, entomological (breeding percentage of *Aedes aegypti* mosquitoes) and environmental data (vegetation, connectivity and percentage covered by residential areas). Addresses of dengue cases were geo-coded using the Geographic Information System (GIS). Information such as duration of transmission, dengue serotypes detected and number of dengue cases were recorded for each dengue cluster. The risk of spread of dengue was calculated using Random Forest. Each grid was then assigned colour-coded risk levels and mapped to give a clear graphical representation of risk across the country.

Ong et al. found that the most important factors were population, dengue burden and breeding percentage from the previous year. Ong et al. also determined that risk maps produced by Random Forest were very accurate.

2.2 - Patterns of Urban Housing Shape Dengue Distribution in Singapore at Neighbourhood and Country Scales

Seidahmad et al. (2018) analysed dengue incidence rates between low-rise and high-rise housing residential areas in Geylang. The objective was to uncover the influence of urban drainage on the distribution of dengue presence and outdoor breeding.

Geylang was chosen as the neighbourhood for the study as there was a continuous circulation of all four serotypes of dengue. Geylang could also be roughly divided into two parts, namely the low-rise and high-rise neighbourhoods. Spatial housing data was retrieved from Singapore Land Authority (SLA) and includes breakdown of buildings by type of utility (commercial, industrial, types of buildings, etc). OpenStreetMap was also used to obtain the locations and lengths of roads by category.

Seidahmad et al. used the ArcGIS program to break Geylang into 200 m^2 blocks, and Singapore into 1 km^2 blocks. They also calculated the drainage density of each block.

They found that low-rise areas generally had a higher drainage density. There was a moderate positive correlation between drainage density and breeding places of *Aedes aegypti*, and between drainage density and dengue cases. In both the Geylang neighbourhood and country contexts, low-rise areas generally had more dengue incidents than high-rise areas.

2.3 - Takeaways

Both research papers carried out analysis by dividing the area into a grid using GIS/ ArcGIS software. This presents a very useful way of analysing data based on location. However, we may not be able to use ArcGIS as an ArcGIS license is very expensive. Nonetheless, indirect methods still exist for analysis of locational variables. The two research papers have also provided us with some software and data sources that we could use in our analysis.

The two research papers have also pointed us towards some potential variables which we could analyse. For instance, Ong et al. analysed residential coverage of a block, while Seidahmad et al. analysed the heights of buildings.

3 - Methodology / Procedure

3.1 - Methodology

We divided the mainland Singapore map into 0.5km by 0.5km grids. Grids that did not contain any residential buildings were identified and removed as dengue cases would not be attributed to such grids. The remaining grids were then represented using their latitude and longitude bounds. Then, dengue cases as well as the potential factors causing them can be analysed based on the grids.

We divided potential factors into two categories, the first being environmental and physical factors and the second being the presence of dengue cases nearby. Environmental and physical factors include total daily rainfall, mean temperature, mean wind speed, as well as the number of public residential buildings and private residential projects, such as condominiums and landed properties. Nearby dengue cases are defined as cases present in the 5 by 5 grid, or a 2.5km by 2.5km area around a given grid.

3.2 - Data Collection

Dengue cases and weather data lasting the entire duration of 2018 were obtained from the website of Singapore's National Environment Agency (NEA). The addresses of all public housing in Singapore were obtained from the Housing and Development Board (HDB), while the

locations of all private residential projects were retrieved from the Urban Redevelopment Authority (URA).

The process of data collection was largely automated using Python 3.

3.3 - Data Analysis

We analysed the significance of the factors mentioned previously using R, a programming language designed for data and statistical analysis. We performed single-variable and multivariable regression to determine how well each factor explained the variability in the number of dengue cases. This helped us to complete Research Questions 1 and 2.

3.4 - Prediction Model

Two popular machine learning methods for prediction include regression analysis and random decision forests. Both methods are built into R.

In regression analysis, multivariable regression is performed like in Research Questions 1 and 2, and prediction values are retrieved by interpolating or extrapolating from the regression line.

In the random forest algorithm, a number of decision trees (specified by the analyst) are built. A decision tree is a tree- or flowchart-like structure in which each internal node represents a condition or a test on attributes, each branch represents the outcome of a condition, and each leaf node represents one of all decisions taken after computing the results of all conditions. The random forest algorithm adds a randomness factor into the construction of decision trees, so multiple different decision trees can be built with the same data. The returned prediction value of

the random forest is the mean of the returned values from each decision tree. A random forest gives more accurate results than individual decision trees as it corrects for a decision tree's tendency to overfit to the training data. For the sake of our research, each random forest is made up of 1000 individual decision trees.

The above two methods used given data for analysed factors to predict the degree of infestation of a grid. The degree of infestation was determined based on the number of dengue cases. The four categories used for the purpose of assessment are: 0 or 1 case, 2 to 9 cases, 10 to 29 cases and 30 or more cases.

Data including the number of dengue cases and information for analysed factors were randomly divided into a training set and a testing set, roughly in a 3:1 ratio. The two machine learning models used the training set to form a prediction model. The prediction models then use the data for factors in the testing set to predict the degrees of infestation. For each test, the score of a prediction model is the number of correct predictions made. We carried out 1000 iterations of this test. Finally, we plotted a frequency-score graph to determine which learning method produces the more accurate prediction model.

4 - Results

4.1 - Research Question 1

For our first research question, we examined the significance of the environmental factors and physical factors, such as temperature, rainfall, wind speed and housing, in determining the degree of dengue infestation in a specific 0.5km by 0.5km grid.

4.1.1 - Single Variable Regression Models

In order to determine the correlation between the aforementioned factors and dengue cases in a specific grid, regression models are utilised in R, to attempt to fit a suitable straight line or a curve onto the set of data points.

Multiple factors are considered, which includes, temperature, wind speed, total rainfall in millimetres, public residential buildings, as well as private residential projects. The environmental factors, namely temperature, wind speed and temperature, are further divided into subcategories. Rainfall data is further processed to acquire the total rainfall in millimetres within the span of 4 to 6 days ago, 7 to 9 days ago, 10 to 13 days ago, and 14 to 17 days ago from a certain day, whereas temperature and wind speed data is further processed to acquire the mean temperature and wind speed 4 days ago, 7 days ago, 10 days ago, as well as 14 days ago from a certain day.

Each of these factors, together with its subcategories, are then pitched against their grids' corresponding number of dengue cases in a regression model, in order to determine if there is a correlation between the factor in question and the degree of dengue infestation. A scatterplot is first plotted, before proceeding on with the analysis of the regression model.

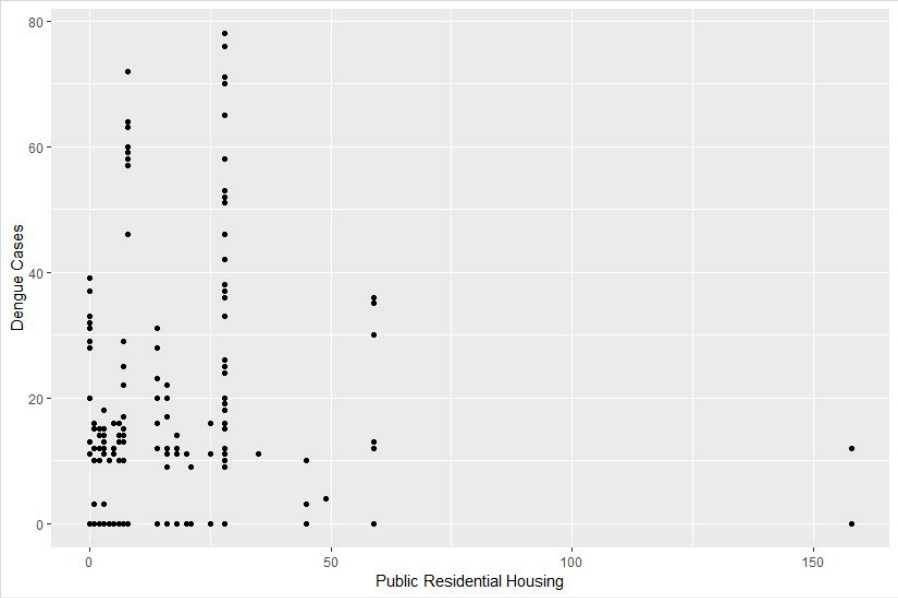


Fig 1.1 Scatterplot of dengue incidences against number of public residential housing

```
Call:
lm(formula = dc ~ public, data = copy)

Residuals:
    Min      1Q  Median      3Q     Max
-11.440  -7.823  -7.211   3.177  70.177

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.04421    0.95855   7.349 8.57e-13 ***
public       0.02782    0.03660   0.760    0.448
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.62 on 484 degrees of freedom
Multiple R-squared:  0.001193   Adjusted R-squared:  -0.0008711
F-statistic: 0.5779 on 1 and 484 DF,  p-value: 0.4475
```

Fig 1.2 Regression analysis of public residential housing and dengue incidences

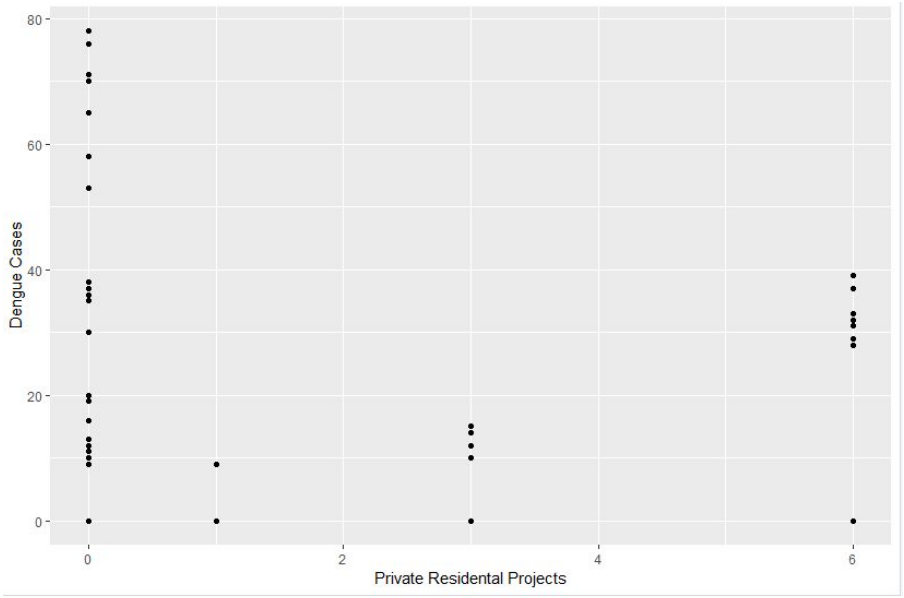


Fig 1.3 Scatterplot of dengue incidences against number of private residential projects

```
Call:
lm(formula = dc ~ private, data = copy)

Residuals:
    Min      1Q  Median      3Q     Max
-10.446 -10.446  -8.536  -0.446  67.554

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4461     2.0145   5.185 7.97e-07 ***
private      -0.6366     0.6810  -0.935    0.352
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.18 on 131 degrees of freedom
Multiple R-squared:  0.006628   Adjusted R-squared:  -0.0009552
F-statistic: 0.874 on 1 and 131 DF,  p-value: 0.3516
```

Fig 1.4 Regression analysis of private residential projects and dengue incidences

Based on the scatterplots *(Fig 1.1 and Fig 1.3)*, there is no clear distinguishable trend between the density of housing and dengue incidences in an area. The R^2 value of both regression models for public and private housing are roughly 0.001 and 0.006

respectively, which indicate that the regression models are only able to account for <1% of the values in the given dataset.

By the same procedure, each subcategory of the environmental factors are then analysed systematically to determine its correlation to dengue incidences.

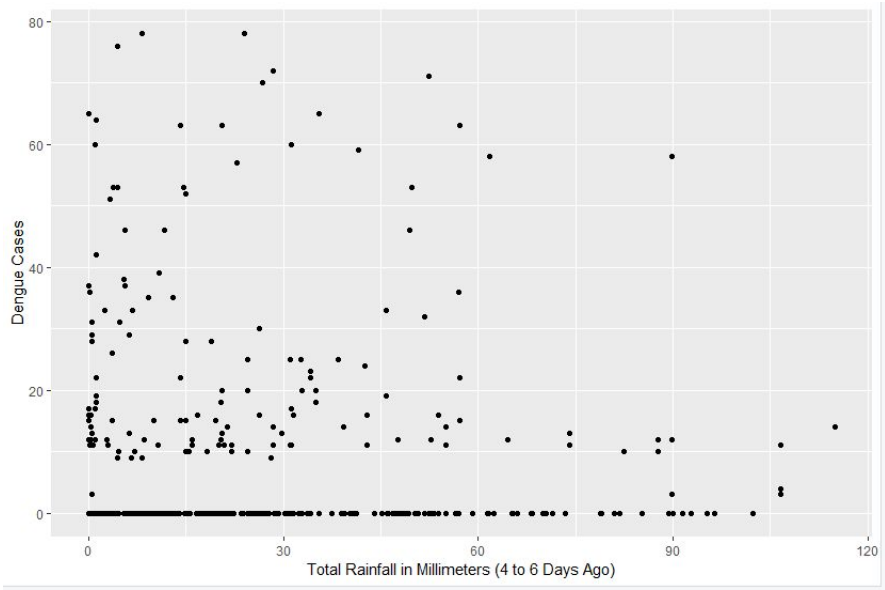


Fig 1.5 Scatterplot of dengue incidences against total rainfall 4 to 6 days ago

```
Call:
lm(formula = dc ~ rf4to6, data = copy)

Residuals:
   Min     1Q Median     3Q    Max
-8.286 -8.007 -7.428  3.246 70.346

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.28590    1.04332   7.942 1.44e-14 ***
rf4to6      -0.02632    0.02937  -0.896    0.371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.69 on 478 degrees of freedom
Multiple R-squared:  0.001677,  Adjusted R-squared:  -0.0004118
F-statistic: 0.8028 on 1 and 478 DF,  p-value: 0.3707
```

Fig 1.6 Regression analysis of total rainfall from 4 to 6 days ago and dengue incidences

Considering the relationship between rainfall from 4 to 6 days ago and dengue cases, there is no distinguishable trend between total rainfall and dengue incidences as observed in the scatterplot *(Fig 1.5)*. The R^2 value of the regression model is also negligibly low at 0.001677, which indicates that the model can only account for no more than 2% of the given data.

Generally, repeating the same procedure with other subcategories under rainfall will yield similar results, no distinguishable trend in the scatterplot, and a negligibly low R^2 value. The results are as follows:

Factor	**Multiple R^2 Value**	**Adjusted R^2 Value**
Rainfall 4 to 6 days ago	0.001677	-0.0004118
Rainfall 7 to 9 days ago	0.0003623	-0.001742
Rainfall 10 to 13 days ago	$1.683 * 10^{-6}$	-0.002112
Rainfall 14 to 17 days ago	$1.838 * 10^{-5}$	-0.002137

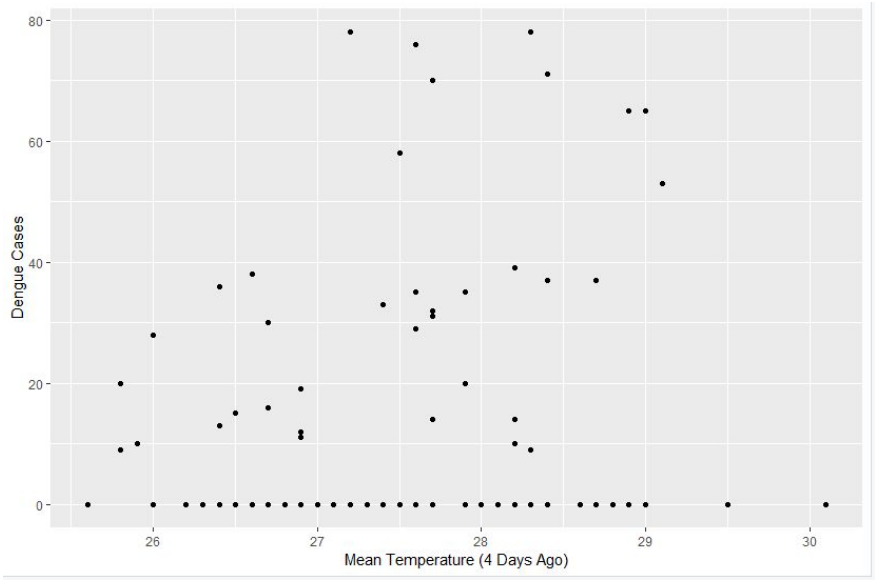


Fig 1.7 Scatterplot of dengue incidences against mean temperature 4 days ago

Next, considering the relationship between mean temperature and dengue cases, a range of values of mean temperature can be observed to yield results where the number of dengue cases is not 0, and the number of dengue incidences would generally peak somewhere in the middle. Thus, it was decided that a best-fit quartic curve would be used to attempt to explain the trend.

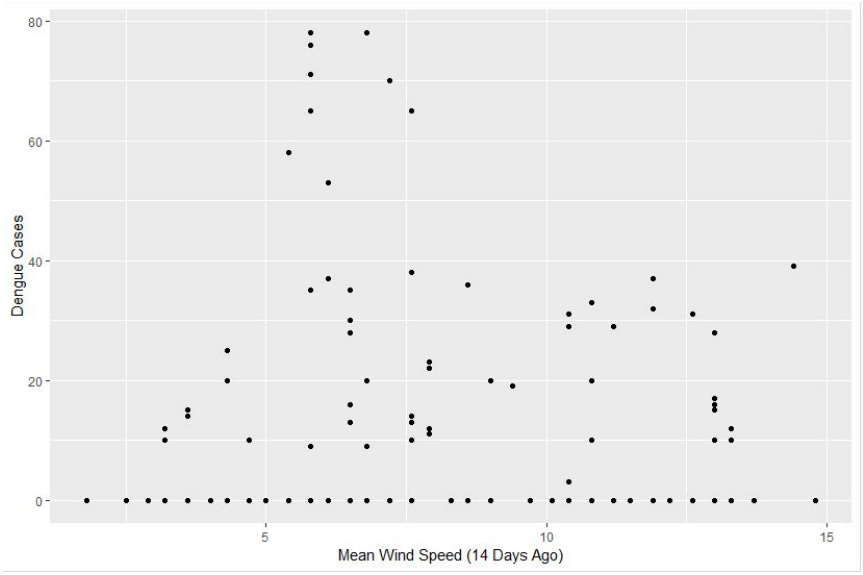


Fig 1.8 Scatterplot of dengue incidences against mean wind speed 14 days ago

A similar situation can be observed in the scatterplot of dengue incidences against mean wind speed, and thus it was also decided that a best-fit quartic curve would be used to attempt to explain the trend.

The results are tabulated as follows:

Factor	**Multiple R^2 Value**	**Adjusted R^2 Value**
Mean temperature 4 days ago	0.04046	0.01048
Mean temperature 7 days ago	0.03566	0.005528
Mean temperature 10 days ago	0.01168	-0.0192
Mean temperature 14 days ago	0.0261	-0.00457

Mean wind speed 4 days ago	$1.931 * 10^{-5}$	-0.005445
Mean wind speed 7 days ago	0.02017	-0.002486
Mean wind speed 10 days ago	0.03014	0.009625
Mean wind speed 14 days ago	0.02543	0.003773

4.1.2 - Multi-Variable Regression Models

After all the factors' correlation to dengue incidences have been analysed one by one, all of the factors are then brought together to formulate a multi-variable regression model to determine each factor's overall significance in influencing the number of dengue incidences within a 0.5km by 0.5km grid.

A total of 4 more regression models are formulated, with the physical factors, both public and private housing, as well as the environmental factors, temperature, wind speed and rainfall from 4, 7, 10, 14 days prior.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.02194    6.56724   1.983   0.0497 *
poly(wind14, 4)1  -4.61537   40.05626  -0.115   0.9085
poly(wind14, 4)2 -12.90028   38.94777  -0.331   0.7411
poly(wind14, 4)3  38.57626   27.18245   1.419   0.1584
poly(wind14, 4)4   4.03432   21.25887   0.190   0.8498
poly(temp14, 4)1  21.09385   21.71096   0.972   0.3332
poly(temp14, 4)2   5.49472   21.00947   0.262   0.7941
poly(temp14, 4)3 -16.71977   20.97000  -0.797   0.4268
poly(temp14, 4)4 -13.84430   19.99763  -0.692   0.4901
rf4to6            -0.05901    0.08007  -0.737   0.4625
public            -0.07027    0.15103  -0.465   0.6426
private           -0.30103    2.13421  -0.141   0.8881
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.42 on 120 degrees of freedom
Multiple R-squared:  0.06092,   Adjusted R-squared:  -0.02516
F-statistic: 0.7077 on 11 and 120 DF,  p-value: 0.7292
```

Fig 1.9 Multi-variable regression analysis using data values corresponding to 14 days prior

Repeating the same procedure, R^2 values of the models still fall well below 0.1, which indicates it still cannot even explain 10% of the trend in the given data points even with all the environmental and physical factors brought into a single regression model. The other 3 multi-variable regression models generally yield similar results, with negligibly low R^2 values and reveal most factors to be rather insignificant in determining the degree of dengue infestation in an area. Results of the regression models are as follows:

Factors	**Multiple R^2 Value**	**Adjusted R^2 Value**
Housing + Environment Data 4 days ago	0.05846	-0.02714
Housing + Environment Data 7 days ago	0.0622	-0.02305
Housing + Environment Data 10 days ago	0.05176	-0.03444
Housing + Environment Data 14 days ago	0.06092	-0.02516

Although, it is worth noting that the mean temperature, wind speed, and total rainfall from 7 days ago provide a relatively more accurate model as compared to 4 days ago, 10 days ago or 14 days ago. As such, it can be deduced that dengue incidences are possibly or most likely influenced by surrounding conditions from 7 days prior.

4.1.3 - Conclusion

Overall, the R^2 values of all the regression models formulated in the whole of Research Question 1 all fall well below 0.1, which indicates that all these physical and environmental factors, whether alone or together, are unable to explain any variability in dengue incidences at all. As such, it is highly unlikely that any physical and environmental factors are, in any way, related to daily dengue cases.

However, this does not imply that these physical or environmental factors are totally unrelated to the degree of dengue infestation in an area. These factors may act as supporting roles for other potential major factors that have yet to be considered.

4.2 - Research Question 2

For our second research question, we examined the significance of nearby dengue cases in determining the degree of dengue infestation. We look at the potential of continuity of trends in daily dengue cases by considering the cases present in the 5 by 5 grid (2.5 km by 2.5 km area) around any specific grid square. Based on results from Research Question 1, factors from 7 days prior appear to have a stronger relationship with the number of dengue cases in an area. Thus, for Research Question 2, we considered the additional factor of nearby dengue cases from 7 days prior.

4.2.1 - Single Variable Regression Model

Compared to the environmental factors listed in Research Question 1, the number of nearby dengue cases from 7 days prior appears to show a stronger relationship with the number of dengue cases.

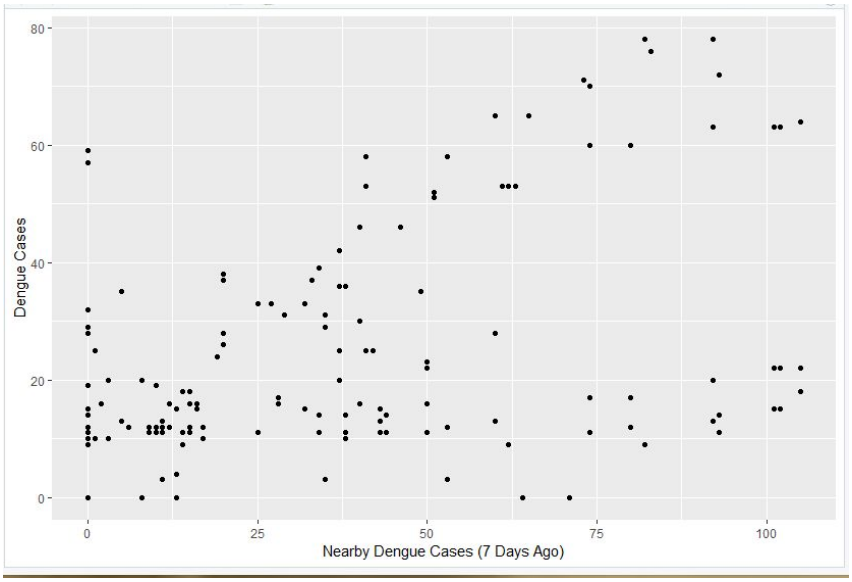


Fig 2.1: Scatter plot of dengue cases against nearby dengue cases from 7 days ago

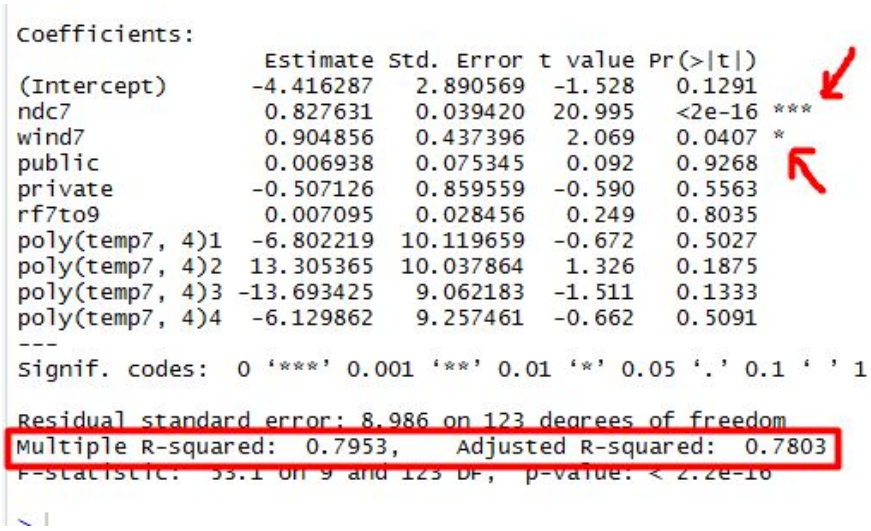
```
Call:
lm(formula = dc ~ ndc7, data = copy)

Residuals:
    Min      1Q  Median      3Q     Max
-34.020  -2.425  -2.425  -2.425  56.575

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.42472    0.55888   4.339 1.75e-05 ***
ndc7         0.44500    0.02063  21.567  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.11 on 479 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.4927,    Adjusted R-squared:  0.4916
F-statistic: 465.2 on 1 and 479 DF,  p-value: < 2.2e-16
```

Fig 2.2: Single variable regression analysis of dengue cases against nearby dengue cases from 7 days ago

The multiple R^2 value for nearby dengue cases alone is much higher than those of the environmental factors explored previously, indicating that the number of nearby dengue cases from 7 days prior is a much better indicator of future dengue infestation.

4.2.2 - Multivariable Regression Model

Taking into account all environmental and surrounding factors in addition to the number of nearby dengue cases from 7 days prior, we constructed a final regression model.



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.416287   2.890569  -1.528   0.1291
ndc7             0.827631   0.039420  20.995   <2e-16 ***
wind7            0.904856   0.437396   2.069   0.0407 *
public           0.006938   0.075345   0.092   0.9268
private         -0.507126   0.859559  -0.590   0.5563
rf7to9           0.007095   0.028456   0.249   0.8035
poly(temp7, 4)1 -6.802219  10.119659  -0.672   0.5027
poly(temp7, 4)2 13.305365  10.037864   1.326   0.1875
poly(temp7, 4)3 -13.693425   9.062183  -1.511   0.1333
poly(temp7, 4)4 -6.129862   9.257461  -0.662   0.5091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.986 on 123 degrees of freedom
Multiple R-squared:  0.7953,    Adjusted R-squared:  0.7803
F-statistic:  53.1 on 9 and 123 DF,  p-value: < 2.2e-16
```

Fig 2.3: Regression analysis of dengue cases against all factors covered

This regression model explains more than 75% of the trend of the data points, which is more accurate than all of the other regression models constructed previously.

Considering numerous factors has the potential to produce regression lines which fit better to the data points.

Overall, the number of nearby dengue cases 7 days prior has the most significant impact on the number of recorded dengue cases on a given day. The second most significant variable appears to be the mean wind speed from 7 days prior, which at best is only moderately correlated with the number of recorded dengue cases.

4.2.3 - Conclusion

The addition of the condition of nearby recorded dengue cases 7 days prior has led to a much more accurate regression model, with an R^2 value exceeding 0.75, a much higher value as compared to if conditions from 4 days prior, 10 days prior or 14 days prior were used instead.

Furthermore, the potential for continuity of dengue incidences in a certain area proves to be the most significant influence on the number of recorded dengue cases on current day in that same area.

4.3 - Research Question 3

For our third research question, we examined the accuracy of various machine learning methods in forecasting the risk level of dengue clusters in each 500 m by 500 m grid. The two methods we investigated were multivariable regression analysis and random forests.

Excess noise in the data could negatively impact the accuracy of prediction models. Thus, it was important for us to remove certain data points that we deemed unhelpful in detecting

trends. We found that clusters with less than 3 dengue cases were usually sudden occurrences and disappeared within a few days. These data points are erratic and do not match up with the factors identified. So, we only considered clusters with at least 3 dengue cases as relevant in the context of this research and worth considering in our prediction models. Further removing more data points due to missingness, we are left with 133 data points to train and test our prediction models.

To ensure that there are sufficient data points for both training and testing purposes, we randomly selected 30 data points for the testing set, and allocated the remaining 103 data points to the training set. The training and testing procedure was run 1000 times and both prediction models were scored based on their accuracy of predicting the dengue risk. The results are

summarised in *Fig 3.1* and *Fig 3.2.*

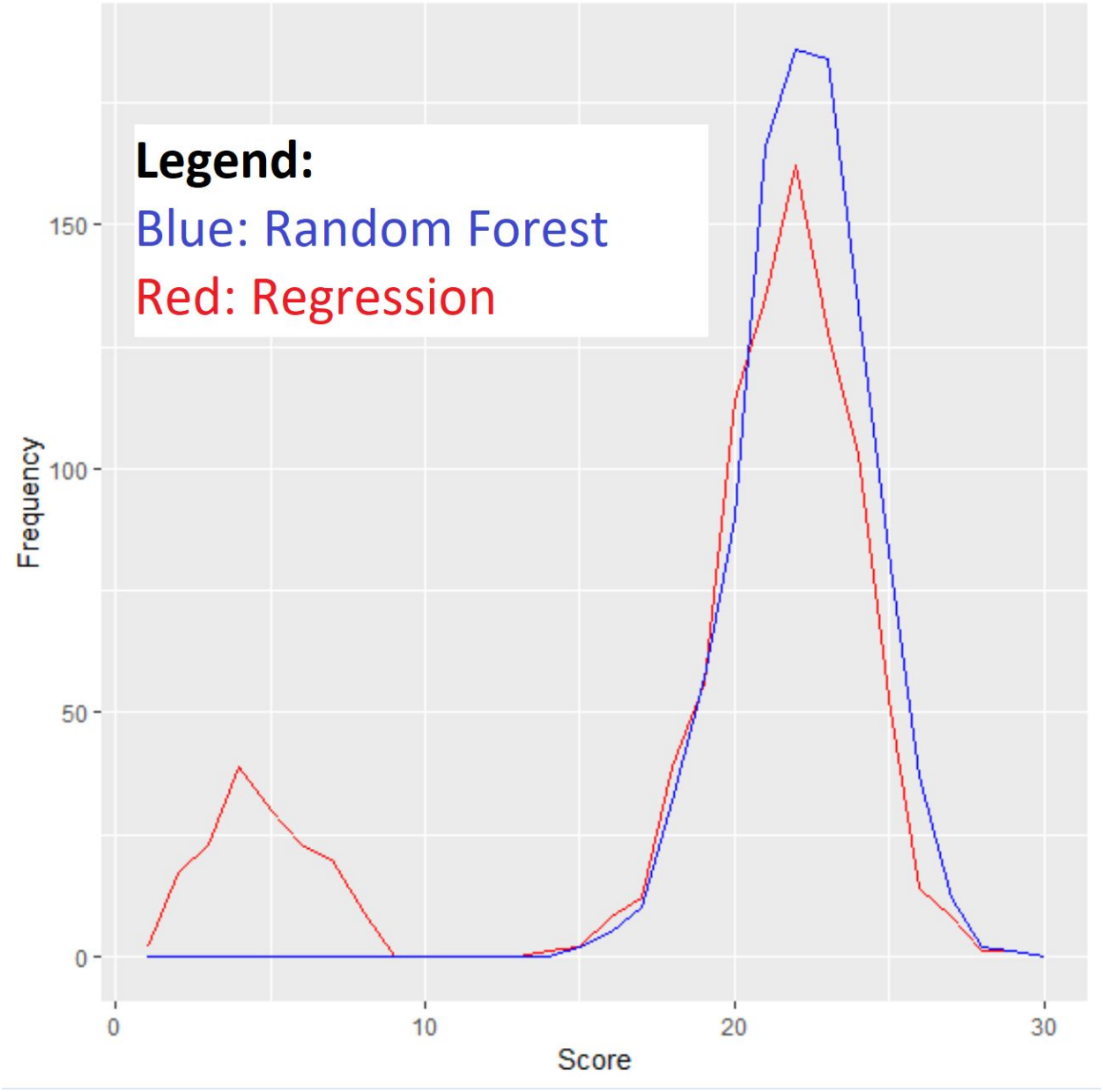


Fig 3.1: Graph depicting the frequency of each score for the two prediction models

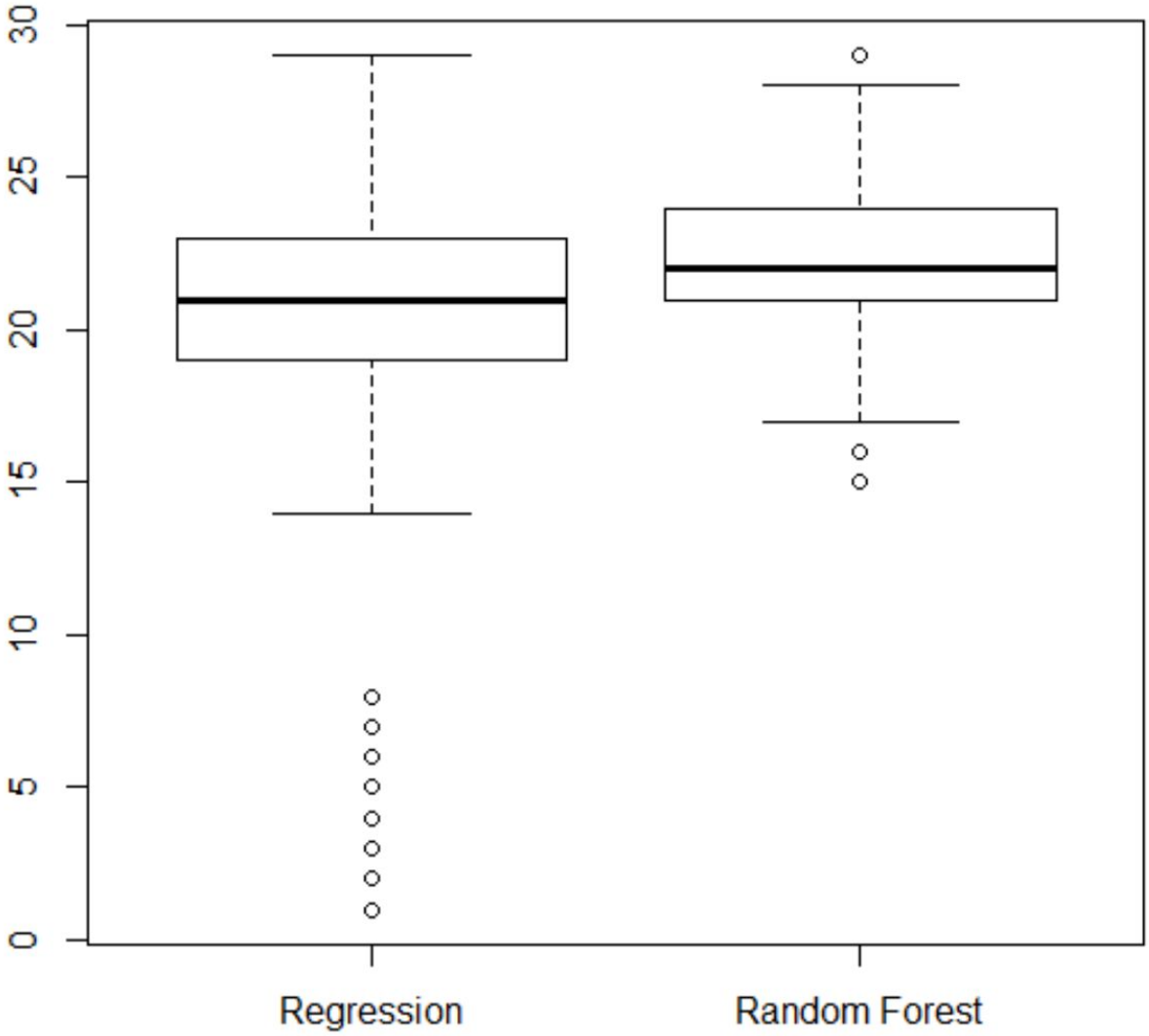


Fig 3.2: Box-and-whisker plot of scores obtained

As seen in *Fig 3.1* and *Fig 3.2,* the random forest prediction model is more likely to predict the number of dengue cases with greater accuracy than the regression model. The median and third quartile scores of the random forest model are larger than those of the regression model. Within the tests we ran, the random forest prediction model is consistently able to predict

accurately at least half of the time. On the other hand, the regression model gets a score of less than 10 about 16.2% of the time, making it unreliable as a prediction model. Thus, the random forest prediction model is more suitable to be used in the prediction of dengue risk.

5 - Conclusion

5.1 - Summary and Conclusion

As an entrance into our research, we discovered that weather data which includes total rainfall, mean temperature and mean wind speed, the number of public residential buildings as well as the number of private residential projects are insufficient in explaining the trend of recorded dengue cases. The number of nearby dengue cases appears to be much more important as a factor. Moreover, data from 7 days prior was found to be more significant than any other duration. Taking residential data, weather data and nearby dengue cases from 7 days prior, we can construct a multivariable regression model that explains more than 75% of the trend in dengue cases. While the multivariable regression model was a suitable prediction model in determining future dengue incidences, there were times where the regression model failed to accurately predict 30% of the values in the testing set. The Random Forest Algorithm was then determined to be much more reliable and accurate as a prediction model of dengue risk, with it able to consistently predict more than 70% of results accurately.

5.2 - Limitations and Possible Extensions

The lack of data proved to be a major limitation of our project. Data on dengue cases is only available if the cases are part of a cluster. Weather data was unavailable for some regions as well. This greatly limited the amount and quality of data we had to work with. More complete data would likely allow for much more accurate prediction models.

Possible extensions include considering even more factors, such as road density. Considering more factors could allow us to construct a regression line which fits even better with the data points. Ong et al. (2018) and Seidahmad et al. (2018) have analysed different potential factors of dengue cases in their respective papers. It would be useful to study all of the variables at once to determine the most important causes of dengue and identify trends.

References

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., . . . Hay, S. I. (2013, April 07). The global distribution and burden of dengue. Retrieved April 1, 2019, from https://www.nature.com/articles/nature12060

Ranjit, S., & Kissoon, N. (2011, January). Dengue hemorrhagic fever and shock syndromes* : Pediatric Critical Care Medicine. Retrieved April 1, 2019, from

https://journals.lww.com/pccmjournal/Abstract/2011/01000/Dengue_hemorrhagic_fever_and_shock_syndromes_.17.aspx

Simmons, C. P., Ph.D., Farrar, J. J., M.D., Ph.D., Van Vinh Chau, N., M.D., Ph.D., & Wills, B., M.D., D.M. (2012, April 12). Dengue | NEJM. Retrieved April 1, 2019, from https://www.nejm.org/doi/full/10.1056/NEJMra1110265

News Releases. (2019, January 9). Retrieved April 1, 2019, from https://www.nea.gov.sg/media/news/news/index/nea-urges-continued-vigilance-to-avoid-surge-in-dengue-cases-in-2019

National Environment Agency. (n.d.). Retrieved April 1, 2019, from https://www.nea.gov.sg/dengue-zika/dengue/dengue-clusters

Ong, J., Liu, X., Rajarethinam, J., Kok, S. Y., Liang, S., Tang, C. S., . . . Yap, G. (2018, June 18). Mapping dengue risk in Singapore using Random Forest. Retrieved April 1, 2019, from https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006587

Seidahmed, O. M., Lu, D., Chong, C. S., Ng, L. C., & Eltahir, E. A. (2018, January 26). Patterns of Urban Housing Shape Dengue Distribution in Singapore at Neighborhood and Country Scales. Retrieved April 1, 2019, from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017GH000080