Maria Cresentia Natasha- General Assembly Project 4

Singapore Weekly Dengue Case Prediction

From Weather and Google Trend

Background

Problem Statement

Government has allocated budget for each department every year. Large scale field trial for Wolbachia Aedes technology has been granted. However, as part of NEA, we would like to ensure that we have fogging measure as it is recommended that we immediately do fogging during a mosquito-borne disease outbreak.

Due to limited budget, **our team has been tasked to proposed effective fogging plan**

Goal: Dengue Case Prevention

All these programs aligned with other project from government to **prevent next outbreak**

Breakthrough: Project Wolbachia

A "mosquito factory" in Ang Mo Kio was opened in 2019 with **a whopping $5 million.** This is an exploratory approach, the sentiment towards this project is very positive

Understanding: Root Problem

Although individual prevention methods such as applying mosquito repellent, wearing long covered clothing, or sleeping under mosquito nets definitely encouraged. However, the most important way to prevent the outbreak must start from the root of the problem, **preventing mosquito breeding habitats.**

History: Dengue Outbreak

First outbreak was recorded in 1960 first. The last 2020 outbreak captured 35,315 cases along with **32 deaths** reported

Success Metrics

01

Lowest Deviation

Lowest root-mean-squared error (**RMSE**) on predicting following week dengue case

02

Important Features

Getting to know what the important features that drives Dengue case number

Final Goal: Finding where and when to effectively do our Fogging programs

Segregating this problems to 3 stages:

1. Spatial Analysis
2. Seasonality Analysis
3. Predictive modeling

Project Flow

"Dengue Fever" keywords even has 91% correlation to dengue case number.

This is an interesting finding on how people leverage google trend search term and that shows how people are more interested in this topic which most of the time is when the case is increasing

Data Scraping (2014-2018)

- **Weather data:** Iterating through links months, year and area code from weather gov website
- **Google trend**: getting the % of search interest based on Dengue symptoms
- **Spatial data:** using Google maps API to capture the Latitude and Longitude

Feature Engineering

- **Data Format:** Ensuring all data sources have same data format to enable concatenation
- **Null Values:** Handling null values by getting the mean of the features as we are looking at the weekly data

Spatial Analysis

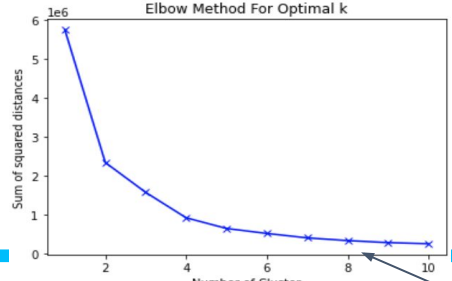
- **K-Mean Clustering:** understanding where the center of dengue case clusters
- **Intent:** to enable focused effort rather than blindly fog the whole singapore (Cost effectiveness)
- **Dengue Case mapping**

Seasonality Analysis

- **Monthly Change:** Understanding if there are months where Dengue case is more prevalent
- **Intent:** to concentrate more effort on the months where there are higher tendency of Dengue case

Final Predictive Model

- **Baseline model:** how if we use only the average number of case assuming all days of the month have same weightage
- **Regression Modeling:** creating a predictive models
- **Comparing it to Deep Learning and Time Series**

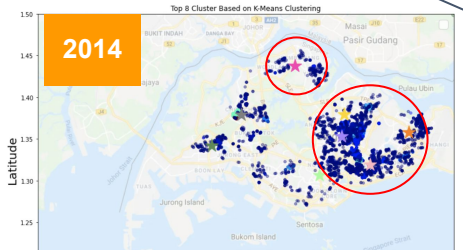


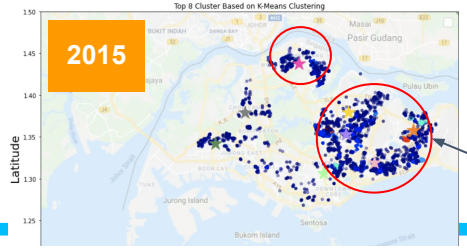
Spatial Analysis

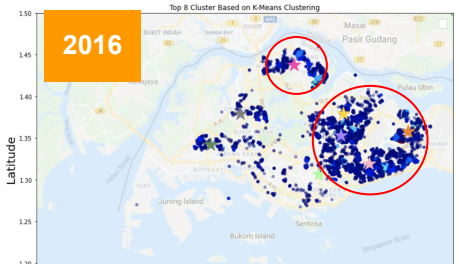
K-Means Clustering

Using the elbow method and weighing by the total number of Dengue case, we capture that **8 centroids** are enough to capture the distribution of our data.

	Latitude	Longitude	Address
Centroids 1	1.357561	103.946310	**[Top5]** 70 Tampines Avenue 4, Singapore, 529681
Centroids 2	1.379080	103.868603	**[Top5]** 5006 Ang Mo Kio Avenue 5, Singapore, 569873
Centroids 3	1.306131	103.838893	[8] 8 Cairnhill Circle, Singapore, 229814 (Newton)
Centroids 4	1.341684	103.709469	[6] 25 Boon Lay Drive, Singapore, 649922 (Jurong West)
Centroids 5	1.437334	103.809485	**[Top5]** Woodlands Avenue 12, Singapore
Centroids 6	1.352729	103.865674	**[Top5]** 250 Lorong Chuan, Singapore, 556748 (Serangoon)
Centroids 7	1.318858	103.899753	**[Top5]** 410 Eunos Road 5, Singapore, 400410
Centroids 8	1.378721	103.745003	[7] 251A Choa Chu Kang Avenue 2, Singapore, 681251

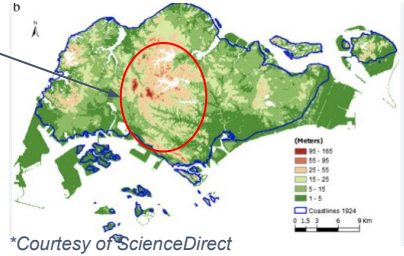






Stagnant water = breeding

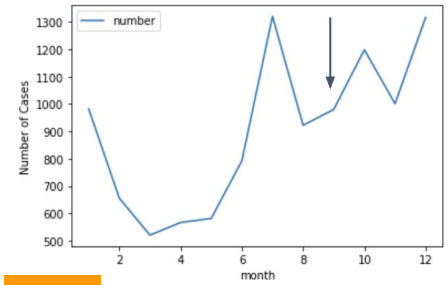
Year-on-year data shows that there is higher number of cases in the North-East of Singapore, this is due to the **most dense real estate** is located in this area. Aside from that **higher elevation**, leads to lower number of case.



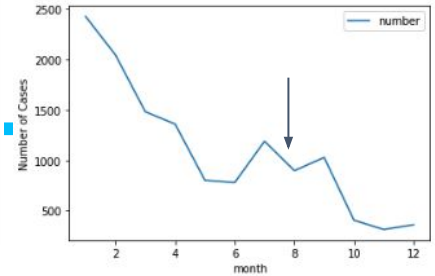
Courtesy of ScienceDirect

Seasonality Analysis

2015



2016

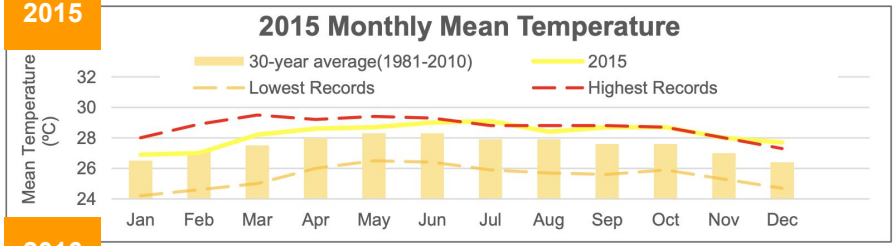


Dengue Case Number Month-by-Month Observation

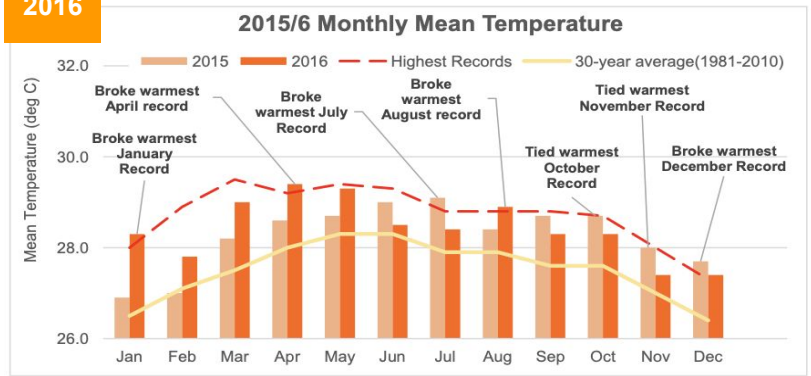
Observing a higher number of case from July to October, one of the reason may be due to the **warm climate** which is suitable for mosquito breeding. (25-30C)

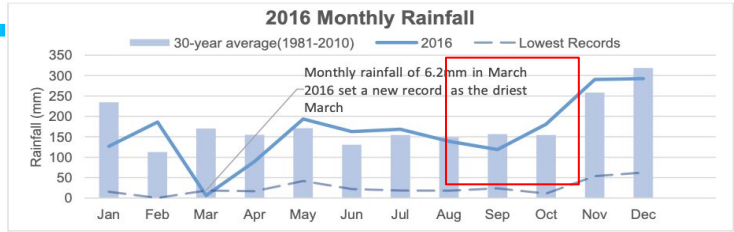
Aside from that the **high rainfall** may cause the **potential stagnant water** creation, which is the optimal locations for mosquito breeding.

2015



2016





*Courtesy of ScienceDirect

Predictive Model

	Baseline	Linear Reg	Linear Reg Square Rooted Features	XGBoost	XGBoost Square Rooted Features	LSTM Deep Learning	VAR Time Series Analysis	SARIMA Time Series Analysis
RMSE	157.35	25.03	16.12	17.15	16.08	61.75	16.82	18.05

Baseline

Taking the mean of the last 3 years

We use this as a baseline assuming that every day of the year has the same weightage without any spatial or seasonality

Linear Regression

R2 score is very low

We used MinMaxScaler to scale our features as these features (temp, rainfall, google trend interest score) are well defined

XGBoost

XGBoost is a black box model

Team will have a hard time to understand what is the important feature that drives dengue case

LSTM

Iterated the 4 weeks data to predict the next day

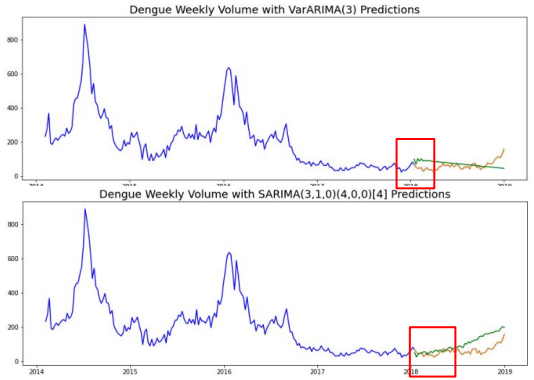
We use moving window so that our machine able to learn and remember previous data by giving higher weightage for newer input (Long Short Term Memory).

Deep Learning generally is a black box model but usually it generates quite good score. However, it seems not the case.

Time Series Analysis

Time Series tend to follow the mean. In SARIMA, we have tried to predict the delta instead, but Time Series Analysis is indeed very complex

It only able to predict for **short term**



Predictive Model

	Baseline	Linear Reg	Linear Reg Square Rooted Features	XGBoost	XGBoost Square Rooted Features	LSTM Deep Learning	VAR Time Series Analysis	SARIMA Time Series Analysis
RMSE	157.35	25.03	16.12	17.15	16.08	61.75	16.82	18.05

Linear Regression (Square Rooted Features)

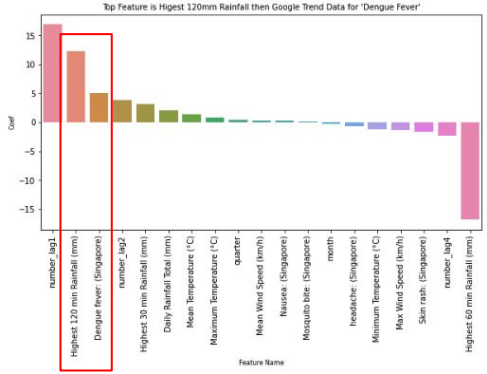
This is chosen as final model

- Prior to scaling the features, we square-rooted all the features except the target variable (weekly dengue case number). This is a **normalization process that help model to perform better.**
- Comparing to non-square rooted features, normalization process actually helps overfitting problem
- We added lag 1,2,4 week as features
- We observe that there is not significant difference of RMSE between the black box machines and easily explainable Linear Regression.

Rainfall and Google Trend Search of "Dengue Fever" are important features

Based on our model features' coefficient, we observe these 2 to be highly related to dengue case number. It seems rainfall has more effect compared to temperature.

It gives us indication to increase our fogging activity when the season of heavy rainfall comes



Recommendation

	Baseline	Linear Reg	Linear Reg Square Rooted Features	XGBoost	XGBoost Square Rooted Features	LSTM Deep Learning	VAR Time Series Analysis	SARIMA Time Series Analysis
RMSE	157.35	25.03	16.12	17.15	16.08	61.75	16.82	18.05

Linear Regression (Square Rooted Features)

However, model's accuracy is at staggering 57% R2 score

This model is far from perfect to be deployed as the sole reference to predict dengue.

As part of data science team, we would like to propose 2 options to NEA budgeting team:

1. We can allocate budget to obtain **more relevant features** such as how far the reservoir to the current dengue case location, HDB age, how humidity changes in Singapore in **weekly base**.
2. If we need to really come up with a model, we will rather leverage on the Exploratory Data Analysis on the spatial and dengue case distribution throughout the months. We suggest that **NEA allocated more fogging sessions on these 5 locations** (4 in the east and 1 in the northwest) then followed by Jurong West, Choa Chu Kang and then Newton subsequently.

Aside from the location, we suggest on the timing as well. Looking at the past 5 years data, there is a tendency for the case to go up and peak around July. This shows how we can put our resources efficiently by **putting more fogging sessions on June (due to 2 weeks incubation, mosquitoes might breed during this timing) to October.**.

	Latitude	Longitude	Address
Centroids 1	1.357561	103.946310	**[Top5]** 70 Tampines Avenue 4, Singapore, 529681
Centroids 2	1.379080	103.868603	**[Top5]** 5006 Ang Mo Kio Avenue 5, Singapore, 569873
Centroids 3	1.306131	103.838893	[8] 8 Cairnhill Circle, Singapore, 229814 (Newton)
Centroids 4	1.341684	103.709469	[6] 25 Boon Lay Drive, Singapore, 649922 (Jurong West)
Centroids 5	1.437334	103.809485	**[Top5]** Woodlands Avenue 12, Singapore
Centroids 6	1.352729	103.865674	**[Top5]** 250 Lorong Chuan, Singapore, 556748 (Serangoon)
Centroids 7	1.318858	103.899753	**[Top5]** 410 Eunos Road 5, Singapore, 400410
Centroids 8	1.378721	103.745003	[7] 251A Choa Chu Kang Avenue 2, Singapore, 681251