Abstract

Operationally, the National Environmental Agency of Singapore (NEA) defines a dengue cluster as a locality in which at least 2 dengue cases occur within 150m and 14 days of each other. Dengue clusters remain active until no new cases occur for the last 14 days. In order to better understand the behavioral patterns of dengue clusters in Singapore, publicly available dengue-related information was collected from NEA and other sources. These included the daily locations of dengue clusters, dengue cases and breeding habitats detected by NEA. Other pieces of information include population data in Singapore subzones, and geospatial data on various landmarks such as roads and waterbodies.

I have created and implemented a work flow structure to automate the process of handling daily incoming streams of raw data from NEA. This consisted of downloading raw dengue-related data on a daily basis, cleaning and processing the raw data in the R environment, as well as organizing, storing and updating the processed data. The R scripts that were used to implement the aforementioned processes are original. I have also investigated the limitations in usefulness and reliability of the consolidated data.

A preliminary exploratory data analysis was conducted on the observed dengue clusters in order to investigate the spatial distribution of dengue clusters in Singapore, as well as how dengue clusters differ in various regions in Singapore. The analysis then moved on to attempting to predict 4 severity outcome variables associated to dengue clusters: total number of dengue cases, duration of dengue cluster lifetime, average occurrence rate of new cases, and whether the dengue cluster grew in area. The predictive methods that were used included k nearest neighbours and random forests, which incorporated the data collected for various other variables as covariates. I ran simulations to examine the effectiveness of training in-sample dengue cluster data using these predictive methods in order to predict the values of the outcome variables in out-of-sample dengue clusters.