Predict local epidemics of dengue fever

Group Members: Lu Cao, Elva Gao, Kinya Wang

I. Problem Specification

Dengue fever is a mosquito-borne disease that occurs in tropical and subtropical regions of the world. Due to the transmission characteristics of dengue fever, more and more scientists believe that climate factors have a complex and non-neglectable relationship with dengue fever's spread. In our project, we want to get a better understanding of the relationship between climate and dengue dynamics to improve research initiatives and resource allocation to help fight life-threatening pandemics.

The purpose of this project is to train a model to predict the number of dengue fever cases reported each week in regions including Puerto Rico, Peru, and San Juan based on environmental variables including temperature, precipitation, vegetation, etc. Being able to predict the infected cases would be impactful with providing a deeper understanding of the root and cause of dengue fever spread, along with helping with research and resource allocation to prevent and fight severe pandemics.

II. Data Description

For our project, since we are working on the relationship between climate and dengue dynamics, the dataset provides us with comprehensive information about geographical data, climates, and disease in quantitative measurements in a (year, week of year) timescale. Specifically, for each entry, the variables are mainly in 5 categories: City and date indicators, NOAA's GHCN daily climate data weather station measurements, PERSIANN satellite precipitation measurements, NOAA's NCEP Climate Forecast System Reanalysis measurements, and Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index measurements.

Our dataset consists of information of one of the two cities, date of the week, temperature distribution and precipitation measured by different climate forecast systems, and vegetation index measured on a weekly basis over the timespan of around 20 years. There are a total of 22 features, which includes 1 ordinal variable, 1 discrete variable, and 20 continuous variables. We could explore the relationship between the number of dengue fever cases and features in this dataset using different graphing and statistical measure methods. Those underlying correlations will enable us to model the number of increased cases of dengue fever every week in Puerto Rico and San Juan.