Predicting Dengue Fever Cases with Machine Learning Models

Emily Johncox, Yidan Nie, Yung-Ching Chen

April 2019

1 Introduction

Dengue fever is a mosquito-borne illness most prevalent in tropical and subtropical parts of the world. Its symptoms can be similar to the flu, but in severe cases, dengue fever can result in severe bleeding and death. Approximately two thirds of the world's population live in areas at risk for dengue fever. Because the transmission vector for dengue fever is the mosquito, dengue fever incidences are tied to climate and weather variables such as precipitation and temperature. As global temperatures rise, climate change is increasing the region susceptible to dengue fever. Once more prevalent in southeast Asia and the Pacific islands, the disease is on the rise in new regions, such as Latin America.

The objective of this study is to utilize predictive modeling and machine learning to predict the total cases of dengue fever in two cities in Latin America: San Juan, Puerto Pico and Iquitos, Peru. Predicting dengue outbreaks before they occur can allow for national and international public health authorities to take action to better prevent, minimize, and treat dengue cases. Because dengue fever is tied directly to mosquito populations, which are heavily impacted by weather, climate, and environmental factors, these variables will be included as predictors for incidents of dengue.

A literature review reveals several conclusions from predictive models of incidents of mosquito-borne illnesses in various regions in the world. Siriyasatien et al.[3] divides predictors of dengue fever into direct and indirect factors. Direct factors include climate, mosquito populations, and mosquito life cycles, while indirect factors include population movement, geography, environment, and human immunology. The mosquito population is the most significant factor, and it is dependent on factors such as precipitation and humidity to create the circumstances appropriate for mosquito development. For example, as mosquitos require standing water to complete their life cycle, mosquito populations and, consequently, disease may increase following a rainy season. Wu et al.[4] predicts dengue fever in Taiwan, and finds that maximum temperature, minimum temperature, humidity, and rainfall are the strongest climate-based predictors. This study also finds that a time series effect is in place, and the strongest effects of climate on dengue fever cases occurs with a two month lag. Additionally, Fuller et al.[1] describes the relationship between local vegetation patterns and dengue prevalence, as there is often an association between seasonal vegetation canopy and mosquito breeding, though this may also be attributed to climate variables. The models in existing literature take various forms, though support vector machines (SVMs) and artificial neural networks (ANNs) were among the most common, as these are often among the most effective in human health domains, according to Fuller et al.[2].

The dataset utilized in this study focuses primarily on climate, weather, and environmental factors, which are measured in two locations over time. Effectively, these variables will serve as a proxy for mosquito populations, which are highly dependent upon weather. The dataset does not include human factors, such as population density or human migration. Based on the available variables, the hypothesis for this model is that temperature and precipitation will be the strongest predictors for total cases of dengue fever. This hypothesis is in alignment with the existing literature, which cites these factors as well as humidity as strong drivers. It is also hypothesized that a time series may be in effect, and that the prevalence of dengue fever will be more directly attributable to the weather in the two months prior than than to the weather on that week.

2 Dataset Description

2.1 Data Overview

The dengue fever dataset used for the creation of this model includes features from a number of different sources. Data is provided by the U.S. Centers for Disease Control (CDC), the Department of Defense, and the Armed Forces Health Surveillance Center in collaboration with the Peruvian government and

U.S. universities. Additionally, environmental and climate data is provided by the National Oceanic and Atmospheric Administration (NOAA).

The dataset in total contains 1456 rows and 26 columns, and the response variable is the total cases of dengue fever in a week. One categorical predictor is the city: San Juan, Puerto Rico or Iquitos, Peru. Other variables include date, climate, weather, and vegetation features over a time period from 1990 to 2008 for San Juan and 2000 to 2010 in Iquitos. A detailed description of each feature can be found in Figure 5.

2.2 Response Variable

The response variable is the total number of cases of dengue fever in a given week in a given city. Its descriptive statistics is showing in Table 1. The distribution of total cases of dengue fever is heavily right-skewed, so linear models will require a log transformation of this response variable.

Table 1: Statistics for Response Variable

Name	Min	Median	Mean	Max
Total cases	0	12	24.66	461

2.3 Predictors

City
The city features classify a response as coming from San Juan, Puerto Rico or Iquitos, Peru. As shown in Figure 6, 936 observations in the complete data set refer to San Juan, which 520 refer to Iquitos. Looking at a map shown in Figure 7, it can be seen that San Juan is an island city, while Iquitos is on the mainland, which may result in differences between the impacts of weather and climate factors in the two cities. While the models developed for this report were trained for both cities, an opportunity for improvement may be the development of different models for responses in each city.

Year
The year feature is given as an integer value from 1990 to 2010. This variable is removed from modeling since our interest is focus on predicting the future dengue cases and the effect of climate features.

Week of Year
The week-of-year feature will be utilized in the model to indicate patterned change in the number of dengue fever cases throughout the year. This variable also holds as a proxy of season.

Precipitation Features
Precipitation features include two measures of total precipitation from different sources. These variables are important, as dengue fever cases are strongly related to mosquito populations, and mosquitos require wet environments and standing water to complete their life cycle.

Temperature Features
There are several temperature features included in this dataset, and they come from two different sources. Included variables include mean maximum and minimum temperatures, as well as diurnal temperature variation. Average overall air temperature and dew point temperature are also included. These variables are needed as temperature also has a strong influence on mosquito populations. Excessively warm summers may not allow mosquito populations to survive, while relatively warmer winters encourage a strong resurgence of the mosquito population in the next year.

Humidity Features
Humidity features include mean relative and specific humidity. These features are included for reasons similar to precipitation and temperature features, as they impact the mosquito life cycle. However, including humidity in addition to temperature and precipitation could prove to be redundant.

Vegetation Features
The vegetation index is derived satellite coverage on pixels surrounding the city. Vegetation coverage amounts are approximated for the northeast, southeast, northwest, and southwest sides of the city. This information is included as there are relationships between vegetation canopy and mosquito populations,

though it may prove to be redundant with other weather features. Still, these variables will be initially included.

2.4 Data Transformations

The distribution of the total number of dengue fever cases was analyzed to determine if transformation is needed. From the density plot of the response shown in Figure 8, it is clear that the total number of cases is not normally distributed. Rather, it is heavily skewed to a lower number of cases per year, with a tail that extends to higher dengue incidences in rare situations.

Not all models can facilitate such a distribution and still meet the mandatory assumptions. For this reason, the response variable of total cases of dengue fever can be transformed using the form $ln(1+y)$, where ln is the natural logarithm. As can be seen in Figure 9, this results in a response distribution that appears to be more normally distributed. This distribution will be used on models that require an approximately normal response variable input.

3 Exploratory Data Analysis

3.1 Missing Values and Imputation Methods

Due to the nature of the dataset, some values for some features were missing. The distribution of these missing values can be seen in Figure 10. 82.3% of the rows are complete, and 1.5% of the observations overall are missing. Among the rows with missing values, it can be seen that the ndvi_ne feature, which corresponds to vegetation coverage northeast of the city center, is absent in 13.3% of rows, compared to most features which are missing about 1% of the observations.

These missing features were managed by trimming the dataset and then imputing missing values. First, any rows missing more than 50% of features were removed. This resulted in the deletion of 10 rows. Next, all other NA values were calculated by imputation, which utilizes existing values to fill in missing ones.

Since the values of vegetation, temperature and humidity do not change dramatically from week to week, the missing values were imputed by the front-fill method, which generates imputed values that match the most recent present value prior to the missing value. The imputation was done with the assistance of the `imputeTS` package.

3.2 Distributions of Features

Response Variable vs. Time
The relationship between the total cases of dengue fever and time is given in Figure 11 and Figure 12. The response variable is plotted against the week start date, which gives a spread of cases through several decades, and also against the week of year, which demonstrates the annual variation.

From these plots, it can be seen that there are some years with many more cases than other years, though at first glance there does not appear to be a clear increase in the number of dengue cases over the years. Additionally, it is clear that there is some annual variation in the number of cases of dengue. This intuitively makes sense, as mosquito populations have seasonal variation as well.

Vegetation Features
The distribution of vegetation features can be seen in Figure 13. Based on the histograms, it appears that the distributions of vegetation features are roughly normally distributed.

Precipitation Features
The precipitation features have three different units of measurement as they are sourced from three different organizations. The density plots of the precipitation features in Figure 14 indicate that the ranges of precipitation amounts are relatively large, but that the majority of observations have low amounts of precipitation.

Humidity Features
The density plots of relative humidity in Figure 15 show the distribution of mean relative humidity, which indicates there are two peaks within its distribution. The reason for this pattern may be that the tropical region has annual dry and wet seasons. The density plot of specific humidity shows that the

distribution is skewed slightly to the right, but appears somewhat normally distributed.

Temperature Features
The distribution of temperature features is shown in Figure 16. The features are from two different sources, and it can be seen that the distributions of the same climate features are quite different between the two sources. This variation could be due to the difference between station-based real-time data versus satellite-estimated data. As it is unclear which source has more accurate data, the first stage of modeling will include all temperature variables.

3.3 Correlation Analysis

From the correlation plots in Figure 17, Figure 18, and Figure 19, it can be seen that many variables are at least slightly correlated to one another. There are also a few instances in which the correlation between some variables was "1". It can be seen that the two precipitation measures are highly correlated, and that the dew point and humidity variables are also related. Furthermore, all vegetation data is correlated. These correlations correspond to intuitive understanding of weather and climate patterns.

While many models are sensitive to correlation, it was determined not to delete highly correlated variables before the model development stages. The reason is that different models may operate better when different sets of variables are included. So, the variable selection stage will occur at the level of individual models, rather than for the entire dataset.

4 Methodology

4.1 Division of Dataset

80% of the dataset is allocated for training the models, and 20% of the dataset is allocated for testing model comparisons. The training and test datasets were both randomly selected from the dataset that had been completed by imputation.

The training dataset is used for model development and training, as well as tuning using k-fold cross-validation (or cross-validation algorithms specific to certain modeling packages). Due to the relatively small number of observations (about 1500), it was determined that k-fold cross validation should be used for tuning, rather than setting aside a proportion of the dataset. This strategy results in better-tuned models while still allowing a relatively large test dataset to be used for model comparison.

The test dataset will remain unseen to all models during the model development stage, and then be used to evaluate final performance of each model. This allows for a fair comparison of each model.

4.2 Model Development

Several models have been developed to best predict the variation in the number of cases of dengue fever relative to the set of predictor variables. Each model is developed using the training dataset, and then using k-fold cross-validation, each model is tuned to create the optimal model of that type. The following models were built, tuned, and compared: null model, linear model, generalized linear model, generalized additive model, random forest model, BART model, GBM model, MARS model, and SVM model.

Preprocessing for Linear Models
The linear models require pre-processing of the data in order to build an effective model. This is especially the case with a dataset such as this, in which different response variables have very different scales. A preprocessing function from the `caret` package was used to standardize the predictor variables. This preprocessing step was not required for other models, such as tree-based models.

Null Model
A null model is used as a basis of comparison for all other models. The null model functions as a linear model, except that it assumes that the response variable is not at all impacted by the predictors. As such, the predictor coefficients are zero, and the response is only dependent upon the response mean as a predictor. The performance of each model will be compared against the null model, and only models that yield significant improvement over the null model are considered successful.

Multiple Linear Regression
The multiple linear regression model has a rigid formula, and therefore model modifications are dependent

mostly on the transformation of the response variable and the inclusion of predictor variables. The original distribution of the response variable was highly skewed and nonlinear. A logarithmic transformation $ln(1+y)$ was used to modify the response to make it more normal.

Given this transformation, the model was first built in its default form, which predicted a response as a function of all predictor variables. This base model included many variables which were not statistically significant. So, a function within the `olsrr` package was used to determine the optimal variables in a stepwise manner. Variable selection was done based on p-value and used the entire training dataset. Variable selection results are given in Table 2.

Table 2: Variable Selection for Linear Model

Variable	Description
reanalysis_tdtr_k	Diurnal temperature range
weekofyear	Week of year
reanalysis_air_temp_k	Mean air temperature
ndvi_se	Vegetation pixel southeast of city centroid
ndvi_ne	Vegetation pixel northeast of city centroid
reanalysis_max_air_temp_k	Maximum air temperature
station_diur_temp_rng_c	Diurnal temperature range

The multiple linear regression model is dependent upon the fulfillment of several assumptions, so the tuned model was checked against these assumptions. As can be seen in Figure 20, the residuals vs. fitted values plot is fairly level and shows little pattern, demonstrating a linear relationship between the response and variables. Additionally, the fit to the normal line on the Q-Q plot is fairly close, with only slight deviation at the tails, demonstrating the normality of the residuals. The scale-location plot is also fairly horizontal with little pattern, demonstrating homoscedasticity. Finally, the residuals vs. leverage plot does not show any outliers that are too extreme.

However, it appears that there are also inconsistencies between the linear plots and what would be expected. For example, in the residuals vs. fitted values plot and in the scale-location plot, there are visible straight-line patterns. It is hypothesized that these patterns may occur due to a cyclical relationship between time and the response variable. This possibility remains to be explored further.

Generalized Linear Model

A common modification to a generalized linear model (GLM) is the tuning parameter *alpha*, which is the elastic-net mixing parameter. The *alpha* parameter differentiates between a ridge regression when $alpha = 0$ and the lasso regression in which $alpha = 1$, but *alpha* can take any value between 0 and 1. Another tuning parameter of a GLM is *lambda*, but the package `glmnet` already has a cross-validation loop for *lambda* built-in.

The GLM for this project was tuned by modifying the *alpha* parameter in cross validation. All predictor variables were included in this model, and the standardized predictors were used. The optimal *alpha* value was determined via 10-fold cross validation of the training dataset.

Table 3: Tuning Parameters for GLM Model

Hyperparameter	Description	Range of Values	Optimal Parameter
alpha	Elastic-net mixing parameter	[0, .01, ..., 1]	0

Generalized Additive Model

The generalized additive model (GAM) is a generalized linear model in which the predictors depend linearly on smooth functions. Because GAM is fundamentally a linear model, the preprocessed and standardized predictors are required for this model.

The application of GAM to this scenario requires some modifications to the model. For instance, the incidence of dengue response variable seems to be overdispersed, so a negative binomial distribution is used for the link function. Additionally, the week-of-year variable is treated as a cyclical term and the knot of its basis function is set at 52. By the likelihood-based methods of REML, the package `mgcv` did the selection of the smoothness level automatically. Based on the results of the optimal GAM model, some degrees of freedom of basis functions are up to 8, which leads to very smooth curves which can be seen on the partial residual plots.

In the diagnostic plots in Figure 21, it can be seen that some plots have some odd patterns that may be attributable to a hidden time series behavior in the data. While the model was already modified to account for some cyclical model, it appears that further modifications may be necessary. However, it appears that some assumptions, such as the normality of residuals, are well met.

Table 4: Tuning Parameters for GAM Model

Hyperparameter	Description	Range of Values	Optimal Parameter
df	Degree of smoothness functions	[1,2,...]	-

Random Forest Model

The random forest algorithm is a powerful non-parametric tool for constructing prediction rules based on various types of predictor variables without making any prior assumption on the form of their association with the response variable. For this reason, the preprocessed predictor variables are not necessary for this model.

Two important hyperparameters in random forest model are *mtry*, the number of variables randomly sampled as candidates at each split and *ntree*, the number of trees to grow. A 10-fold cross-validation is performed to find the optimal combination of these hyperparameters. The tuning details of the random forest model are found in Table 5. The RMSE under different hyperparameter combinations is found in Figure 24.

The optimal random forest model draws 21 candidates per split and has the number of trees equals to 750. The training set is used to rebuild a random forest model with *mtry* equal to 21 and *ntree* equal to 750. Figure 25 shows the percentage of increased MSE and percentage of increased node purity of each predictor.

Table 5: Tuning Parameters for Random Forest Model

Hyperparameter	Description	Range of Values	Optimal Parameter
mtry	Number of variables sampled at each split	[7, 9,...21, 23]	21
ntree	Number of trees to grow	[250, 500, 750, 1000]	750

Bayesian Additive Regression Trees

The Bayesian Additive Regression Tree model (BART) is a Bayesian sum-of-trees model in which each tree is constrained by a prior to be a weak learner.

There are many hyperparameters that can be tuned in BART model. Here only the most important are tuned: number of trees grown, prior probability, degree of freedom, and quantile of the prior on the error variance at which the data-based estimate is placed. Under the scope set in each hyperparameter in Table 6, there are 18 combinations in total, meaning that 18 different BART models are compared. 10-fold cross-validation was performed to determine the optimal BART model.

Table 6: Tuning Parameters for BART Model

Hyperparameter	Description	Range of Values	Optimal Parameter
num_tree	Number of trees to be grown	[50, 200]	50
k	Prior probability	[2, 3, 5]	5
nu	Degrees of freedom	[3, 10]	10
q	Quantile of the prior on the error variance	[0.75, 0,9, 0,99]	.75

Multivariate Adaptive Regression Splines

There are two tuning hyperparameters associated with MARS model: the degree of interactions and the number of retained terms. By performing a cross-validated grid search, the optimal combination of the hyperparameters which minimizes the prediction RMSE was identified. The graph of cross-validation results can be found in Figure 22. It can be seen that the optimal model has up to third degree interactions and retain 23 terms. In Figure 23, it can be seen that the optimal MARS model includes 10 predictors. Week-of-year, satellite reanalysis minimum temperature, year, and satellite reanalysis maximum temperature are the top four predictors that have a large impact on the prediction power. Additionally, 16 of 23 retained terms are third degree interactions of hinge functions, which indicates

strong interaction effect among features.

Table 7: Tuning Parameters for MARS Model

Hyperparameter	Description	Range of Values	Optimal Parameter
degree	Degree of interactions	[1,2,3]	1
nprune	Number of retained terms	[2,4,...30]	18

Gradient Boosting Method
The gradient boosting model builds an ensemble of shallow and weak tree-based learners. Each base-learning model is added into the ensemble sequentially and learns from the previous models to slightly improve the remaining error.

Instead of performing 10-fold cross validation, the initial training data was randomized at each hyperparameter combination and then split into 75% of the training set and 25% of the validation set. The optimal values were determined using a grid search. The hyperparameter tuning reveals that the optimal model has relatively deep trees and small incremental steps to make gradient descent faster.

Table 8: Tuning Parameters for GBM Model

Hyperparameter	Description	Range of Values	Optimal Parameter
num.trees	Number of trees to fit	[0, 5000]	4986
depth	Maximum depth of each tree	[1,3,5,7,9]	7
shrinkage	Learning rate	[.01, .1, .3]	.02
bag.fraction	Subsampling	[.65, .8, .1]	.8
min.nodes	Minimum observations in terminal nodes	[3,5,10]	5

Support Vector Machine
The support vector machine (SVM) model can be used for either classification or regression. For the case of regression, the SVM aims to find a hyperplane which maximizes the number of observations which fall within the margin region, given the size of the margin.

A key tuning parameter for an SVM model is *cost*. The *cost* parameter represents the cost of a constraints violation, and is the constant regularization term in the Lagrange formulation. The default cost value is 1, but this value was tuned via 10-fold cross-validation of the training to determine the optimal *cost* parameter. All standardized predictor variables were used in the creation of the model.

Table 9: Tuning Parameters for SVM Model

Hyperparameter	Description	Range of Values	Optimal Parameter
cost	Cost of constraints violation	[0.001, 0.01, 0.1, 1, 5, 10, 100]	0.001

4.3 Model Evaluation

Candidate models are compared against one another by using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics. Each tuned model is evaluated on the same held-out test dataset, which has not been seen during any model development process. Each model predicts the response using this dataset, and the MAE and RMSE are calculated by comparing the predicted values with the actual values. While the out-of-sample error metrics are used for model comparison, in-sample error metrics were also calculated to identify overfitting. The metrics resulting from the evaluation of each model are given in Table 10.

It can be seen that the random forest model yields the lowest RMSE, while the GBM model yields the lowest MAE value. Despite this slight variation, the values for both metrics were still very similar to one another. However, a random forest model is generally more interpretable than a gradient boosting model, and given the low difference in performance, the random forest model was selected as the optimal model to predict dengue fever cases in San Juan and Iquitos.

This optimal random forest model draws 21 candidates per split and has 750 trees. This model yields a 35% improvement over the null model.

Table 10: Model Evaluation Results Based on MAE and RMSE

Model	In-Sample Error		Out-of-Sample Error	
	MAE	RMSE	MAE	RMSE
Null	23.87	45.67	21.84	34.65
Linear	17.94	44.03	14.84	32.13
GLM	22.18	41.36	20.05	31.81
GAM	23.26	50.67	18.03	38.16
MARS	17.05	29.64	16.85	27.29
RF	6.98	13.09	14.86	25.96
GBM	7.53	10.42	14.14	26.14
BART	11.88	15.57	18.98	26.61
SVM	19.78	46.63	15.87	34.27

4.4 Model Interpretation

The random forest model functions largely as a black box, so much of the model interpretation must be done by utilizing variable importance plots and partial dependence plots.

Table 11: Important Variables for Random Forest Model

Variable	Description
ndvi_nw	Vegetation pixel northwest of city centroid
weekofyear	Week of year
reanalysis_min_air_temp_k	Minimum air temperature
ndvi_ne	Vegetation pixel northeast of city centroid
reanalysis_specific_humidity_g_per_kg	Specific Humidity
reanalysis_dew_point_temp_k	Mean dew point temperature
ndvi_sw	Vegetation pixel southwest of city centroid
ndvi_se	Vegetation pixel southeast of city centroid
station_ave_temp_c	Average temperature
station_max_temp_c	Maximum temperature

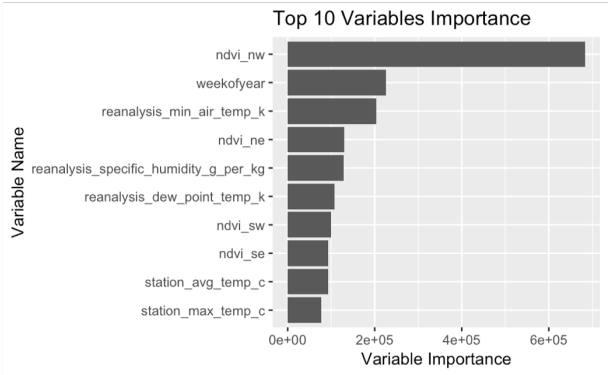

Figure 1: Variable Importance

The variable importance plot is given in Figure 1, and the descriptions of these variables can be seen in Table 11. From this analysis, it can be seen that the most important variable pertains to vegetation, and all four vegetation variables appear in the top 10 predictors. Additionally, it can be seen that week of year, which is a time variable, is another important predictor of the total cases of dengue fever. This speaks to the previously mentioned seasonal variation in dengue cases. Finally, temperature variables prove to be important predictors. This likely ties to the dependence of mosquito populations on climate. The impact of these predictors on the total cases of dengue fever will be explained in more detail below as their partial-dependence plots are analyzed.

Time

First, the dependence of the response variable on time will be examined. Figure 2 shows dependence of total dengue cases on time, and it can be seen that this figure closely resembles Figure 11.

This plot demonstrates a seasonal variation in the total cases of dengue fever, which aligns with the seasonal variation in mosquito populations. It is known that mosquito populations generally are dormant in the winter months, and that they begin development in the late winter and early spring. Then, as the weather begins to warm, mosquito populations increase, and with it, incidences of dengue fever.

However, there are some challenges in developing a clear seasonal interpretation. This is because San Juan is in the northern hemisphere, while Iquitos is in the southern hemisphere. The warmer summer months for San Juan are between weeks 25 and 40, while the warmer months for Iquito are in weeks

40-10. Additionally, because Iquitos is much closer to the equator than San Juan, Iquitos experiences much less seasonal variation than San Juan does. By not creating separate models for San Juan and Iquitos, the seasonal variation impact is shared. As a result, rather than seeing a clear trend based on season in one city, the partial dependence plot could show a trend that is an average of both cities.

Because closer examination is required on the city level, Figure 26 shows the difference in dengue incidences in San Juan and Iquitos. From these plots, it can be seen that the seasonal peak in dengue cases corresponds with early summer in each region. However, for the full dataset, the pattern appears more closely skewed toward the seasons in San Juan, likely because there are more dengue incidences in total in this city.

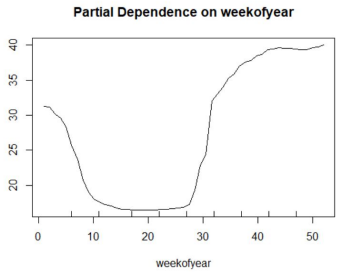

Figure 2: Partial Dependence Plot on Week of Year

Vegetation

Dependence of the response variable on vegetation is also examined. These interactions can be seen in the partial dependence plots in Figure 3.

The random forest model highlighted the importance of vegetation features, though this was not an interaction which was hypothesized. Several interactions can be seen from the partial dependence plots.

First, for each of the vegetation plots, it can be seen that at zero or negative vegetation index, there is lack of dependence. However, the geography of San Juan and Iquitos may play a role in this interaction. For example, because San Juan is on the northern part of the island of Puerto Rico, it is likely that the pixels northwest and northeast of the city are in the ocean, and therefore vegetation is unlikely to be found. Exploration into the vegetation distributions in San Juan confirmed this. Second, it can be seen that for most vegetation regions, an increase in vegetation corresponds to an increase in dengue incidences. Curiously, the exact opposite behavior is exhibited in the southeast region. It is unclear why this relationship exists. However, because of the different geographies of the two regions and the frequent seasonal variation of vegetation levels, it is possible that dependence on the vegetation index is an indicator of dependence on seasons and geography more so than on vegetation itself. However, it is very possible that vegetation does play a large role, especially as a location for mosquito breeding. Because of possible confounding variables, further research in this area is recommended.

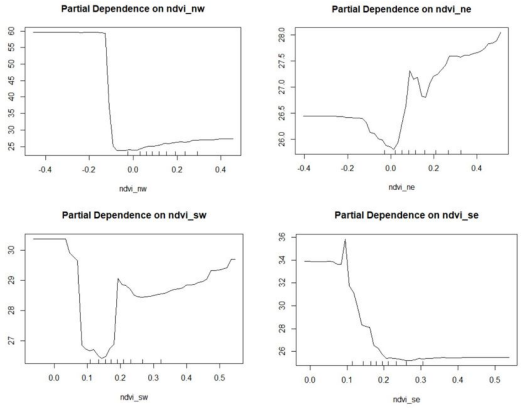

Figure 3: Partial Dependence Plot on Vegetation

Temperature and Humidity

Finally, the dependence of dengue fever on temperature and humidity predictors was analyzed. These dependencies can be found in Figure 4.

First, the dependence on maximum and minimum temperature should be examined. It can be seen that as both maximum and minimum temperatures increase, dengue cases also increase. The cause ties back to the mosquito life cycle. Generally, mosquitoes are in developmental stages in the winter and late spring. They become more active and take their adult form after the weather hits a certain temperature threshold. When the warmer weather results in an increase in mosquito populations, an increase in dengue also occurs.

Next,the diurnal temperature variation should be examined. It can also be seen here that dengue incidences are highest when there is a low diurnal temperature variation, meaning that the temperature is fairly stable between night and day. Again, this often occurs during the warmer seasons and is often accompanied by high humidity, which is what allows temperature to be consistent throughout the day and night.

While discussing climate variables, it can be recalled that strong dependence upon precipitation was hypothesized, yet this dependence was not identified by the model as important. Closer examination of annual precipitation revealed that are not strong precipitation trends (i.e. a rainy and dry season), so this could have minimized dependence.

As with the vegetation, it can still be difficult to take the interpretation of temperature and humidity plots at face value due to the differing geographies and latitudes of San Juan and Iquitos. Additionally, a time-series effect could be especially impactful for climate variables, as certain conditions lead to better mosquito breeding.

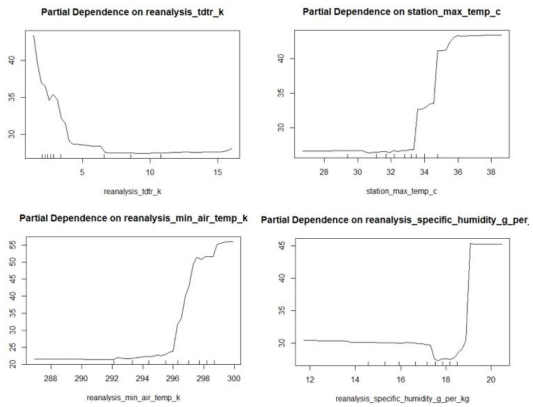

Figure 4: Partial Dependence Plot on Temperature and Humidity

4.5 Conclusion

Dengue fever is a severe and sometimes deadly mosquito-transmitted disease that is most prevalent in the tropics and sub-tropics. Because there is no vaccination, it is imperative that dengue outbreaks be forecasted as early as possible to prepare for mitigation and treatment. Eight models were developed and tuned to predict dengue fever outbreaks in San Juan, Puerto Rico and Iquitos, Peru. Models were evaluated based on their MAE and RMSE performance in an out-of-sample dataset. The optimal model for dengue prediction in these regions was found to be a random forest model, which yields 35% improvement over the null model.

The most important factors predicting dengue fever pertain to time, temperature, and vegetation. The dependence on time and temperature agree with the hypothesis that there will be seasonal variation in dengue fever cases and that climate factors will also impact dengue outbreaks. However, the hypothesis did not predict the strong ability of vegetation to predict cases of dengue fever. The exact reason for this relationship is unclear, but it could pertain to mosquito breeding in jungle regions, to seasonal variation in vegetation, or to differences in geography in San Juan and Iquitos. This is an area requiring future investigation.

The top predictors in this model are all also tied directly to mosquito populations, which is valid considering that dengue is transmitted by mosquitoes. So, it is possible that this model is not predicting dengue outbreaks, but rather that it is predicting mosquito population booms. But, without available information on mosquito population sizes, this variable simply remains internal to the black box of the random forest model.

4.6 Future Work

This work has sparked several questions that may warrant future investigation. First, because seasonal variation is so important to cases of dengue fever, a time-series effect could be in place. For example, mosquito populations in the spring are probably dependent upon breeding conditions in the winter, so some seasonal adjustment could be performed to investigate this relationship in depth. Additionally, it could be useful to create independent models for San Juan and Iquito, especially considering that vegetation variables are so important, while the the geography of these two cities are quite different. Furthermore, as this interaction between vegetation and dengue was not identified from prior literature, further research into this relationship could be useful. Finally, it is possible that by predicting dengue fever as a function of only climate and vegetation variables, the response variable in this project serves mostly as a proxy for mosquito populations. So, it could be useful to add social predictors such as urban cleanliness, health care, and population movement to this analysis to reduce this redundancy.

References

[1] Douglas O Fuller, A Troyo, and John C Beier. El nino southern oscillation and vegetation dynamics as predictors of dengue fever cases in costa rica. *Environmental Research Letters*, 4(1):014011, 2009.

[2] Vijeta Sharma, Ajai Kumar, Dr Lakshmi Panat, Ganesh Karajkhede, et al. Malaria outbreak prediction model using machine learning. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 4(12), 2015.

[3] P Siriyasatien, S Chadsuthi, K Jampachaisri, and K Kesorn. Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes. *IEEE Access*, 6:53757–53795, 2018.

[4] Pei-Chih Wu, How-Ran Guo, Shih-Chun Lung, Chuan-Yao Lin, and Huey-Jen Su. Weather as an effective predictor for occurrence of dengue fever in taiwan. *Acta tropica*, 103(1):50–57, 2007.

A Appendix

City and date indicators

- `city` – City abbreviations: `sj` for San Juan and `iq` for Iquitos
- `week_start_date` – Date given in yyyy-mm-dd format

NOAA's GHCN daily climate data weather station measurements

- `station_max_temp_c` – Maximum temperature
- `station_min_temp_c` – Minimum temperature
- `station_avg_temp_c` – Average temperature
- `station_precip_mm` – Total precipitation
- `station_diur_temp_rng_c` – Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

- `precipitation_amt_mm` – Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- `reanalysis_sat_precip_amt_mm` – Total precipitation
- `reanalysis_dew_point_temp_k` – Mean dew point temperature
- `reanalysis_air_temp_k` – Mean air temperature
- `reanalysis_relative_humidity_percent` – Mean relative humidity
- `reanalysis_specific_humidity_g_per_kg` – Mean specific humidity
- `reanalysis_precip_amt_kg_per_m2` – Total precipitation
- `reanalysis_max_air_temp_k` – Maximum air temperature
- `reanalysis_min_air_temp_k` – Minimum air temperature
- `reanalysis_avg_temp_k` – Average air temperature
- `reanalysis_tdtr_k` – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements

- `ndvi_se` – Pixel southeast of city centroid
- `ndvi_sw` – Pixel southwest of city centroid
- `ndvi_ne` – Pixel northeast of city centroid
- `ndvi_nw` – Pixel northwest of city centroid

Figure 5: Data Overview and Feature List

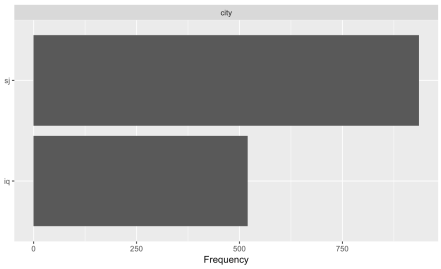

Figure 6: Histogram Demonstrating City Frequency

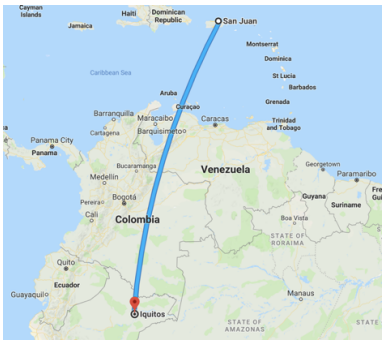

Figure 7: Map

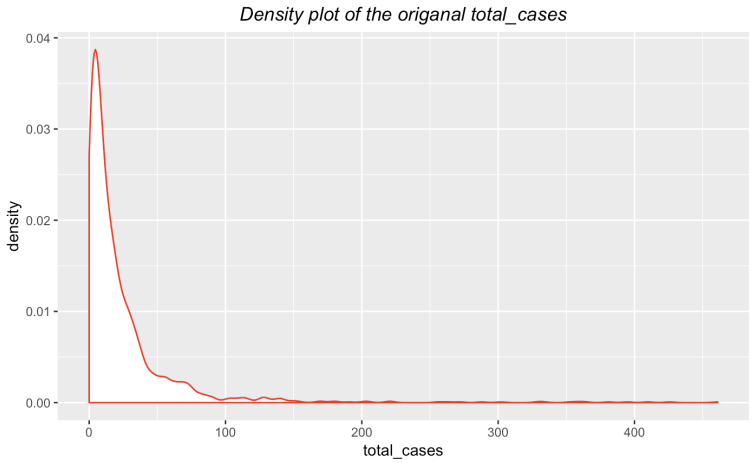

Figure 8: Density Plot of the Original Response

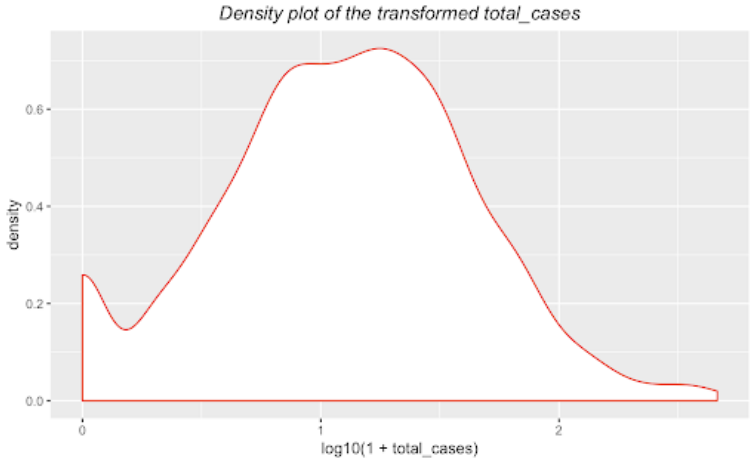

Figure 9: Density Plot of the Transformed Response

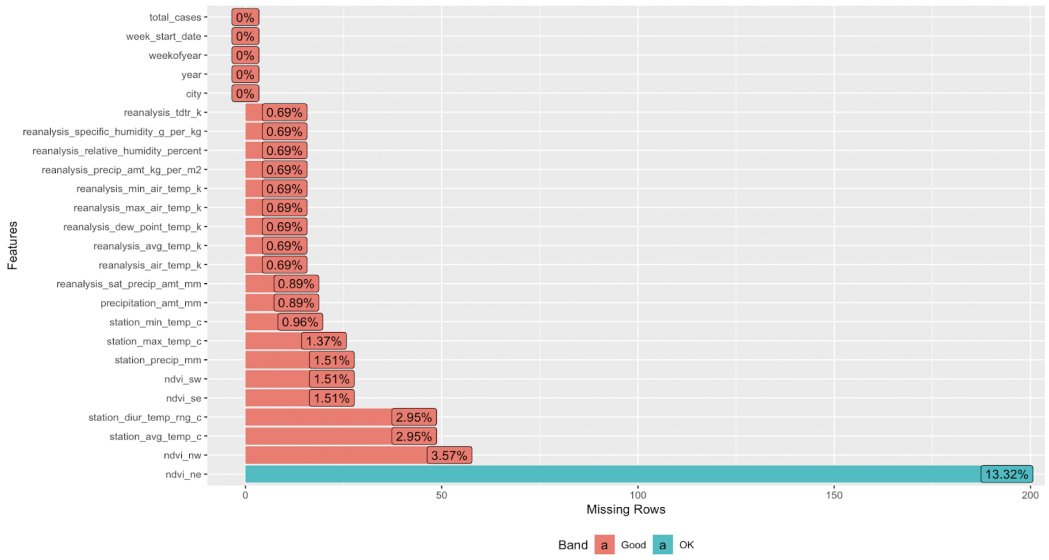

Figure 10: Plot of Missing Value Distribution

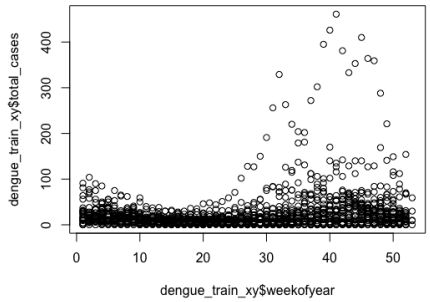

Figure 11: Dengue Cases within Year

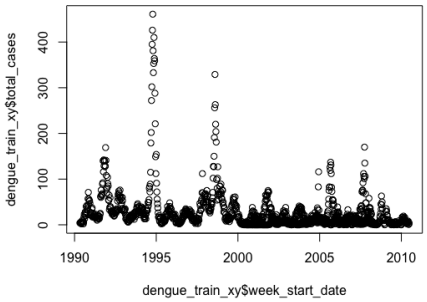

Figure 12: Dengue Cases over Time

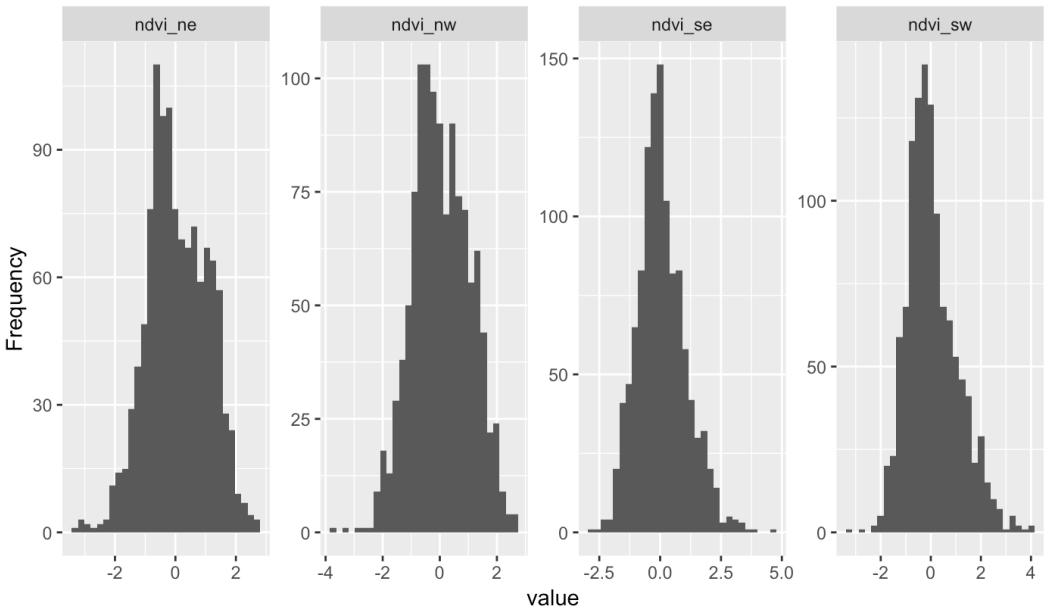

Figure 13: Density Plot of Vegetation Features

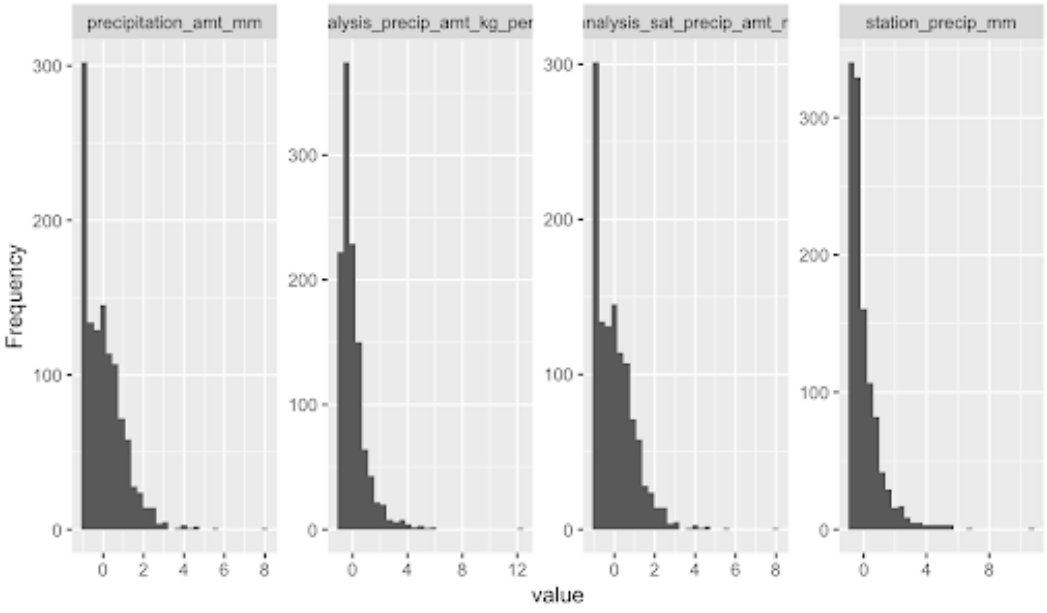

Figure 14: Density Plot of Precipitation Features

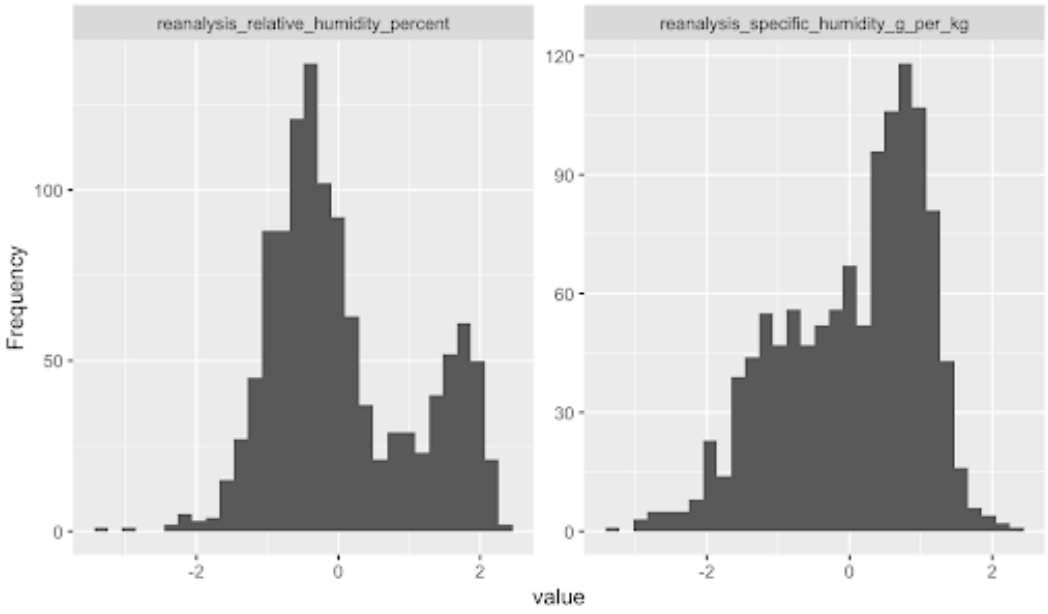

Figure 15: Density Plot of Humidity Features

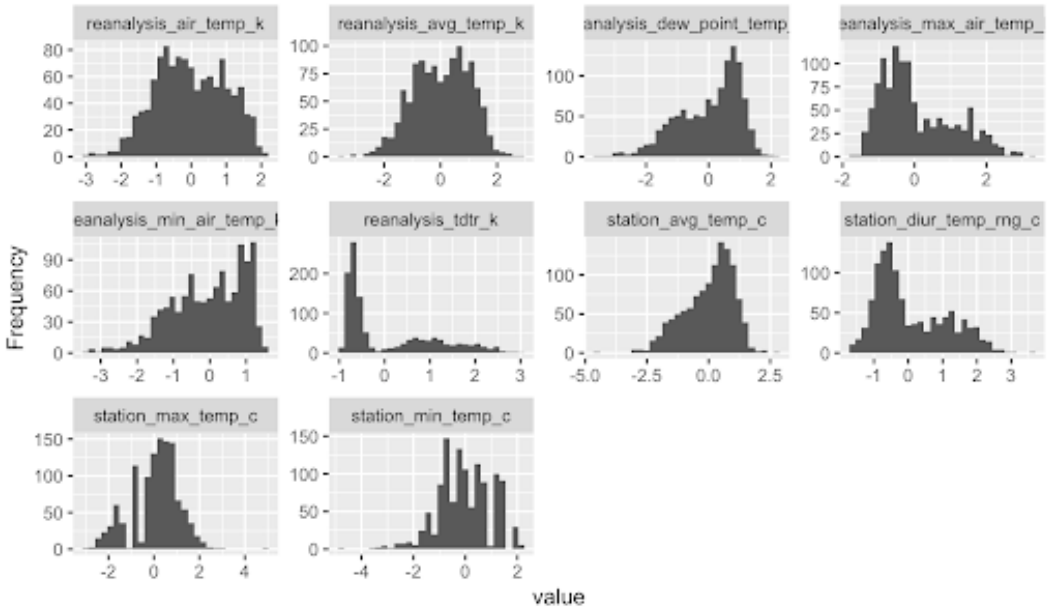

Figure 16: Density Plot of Temperature Features

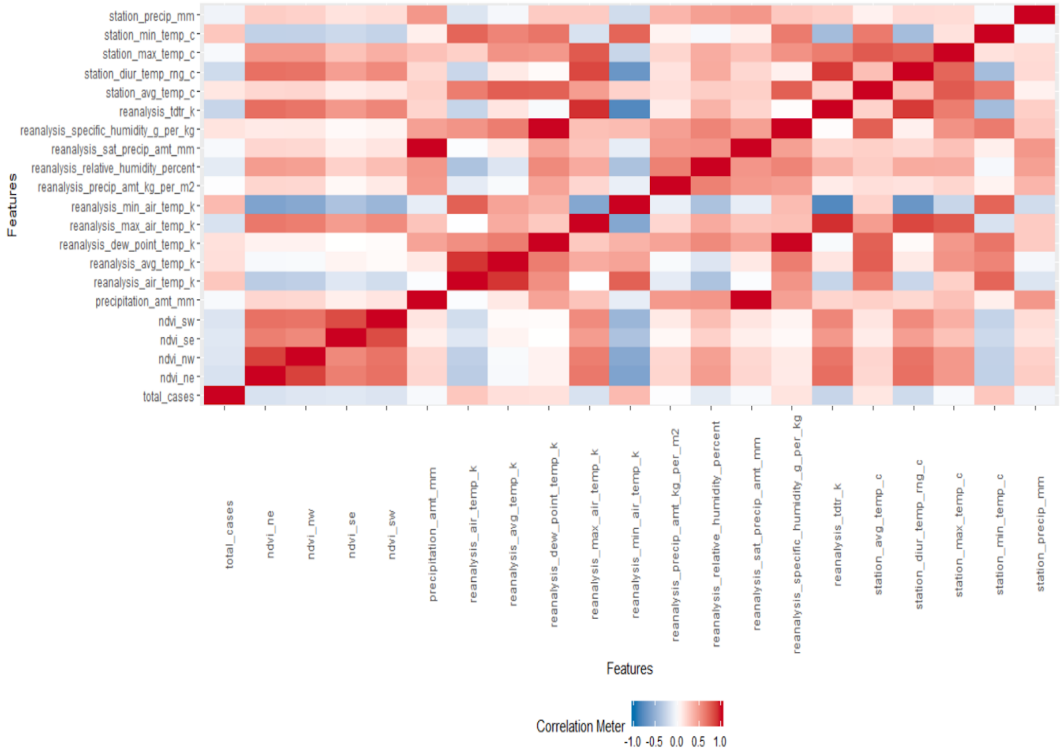

Figure 17: Correlation Plot of All Continuous Variables

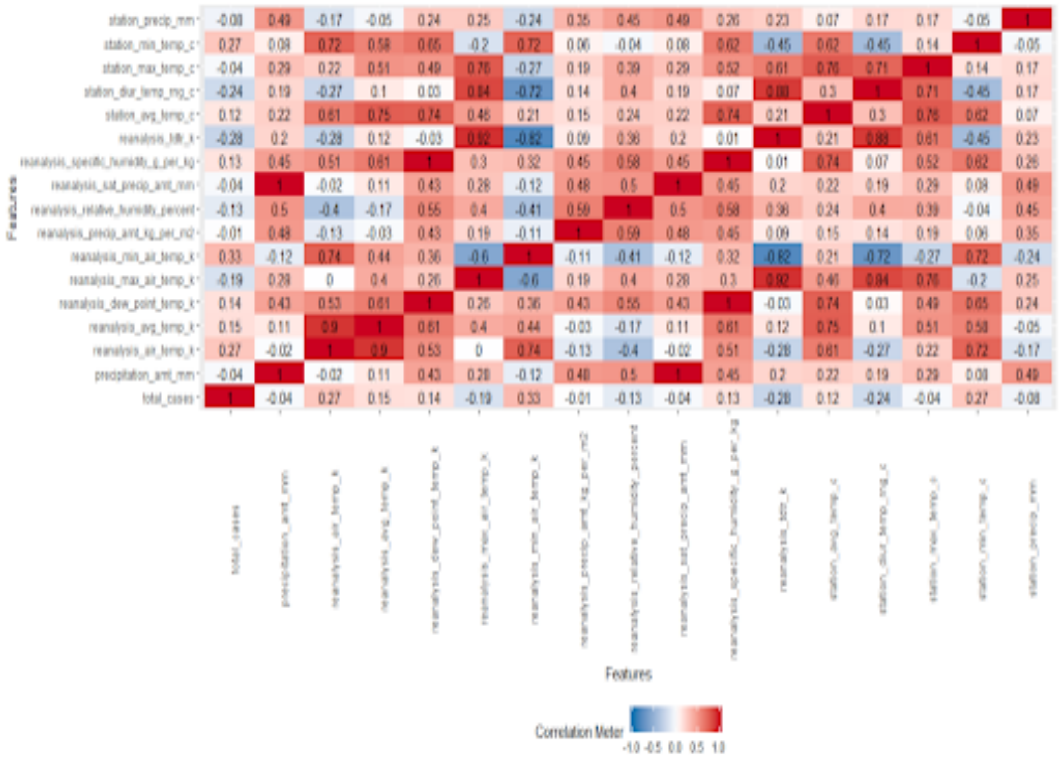

Figure 18: Correlation Plot of Climate Variables

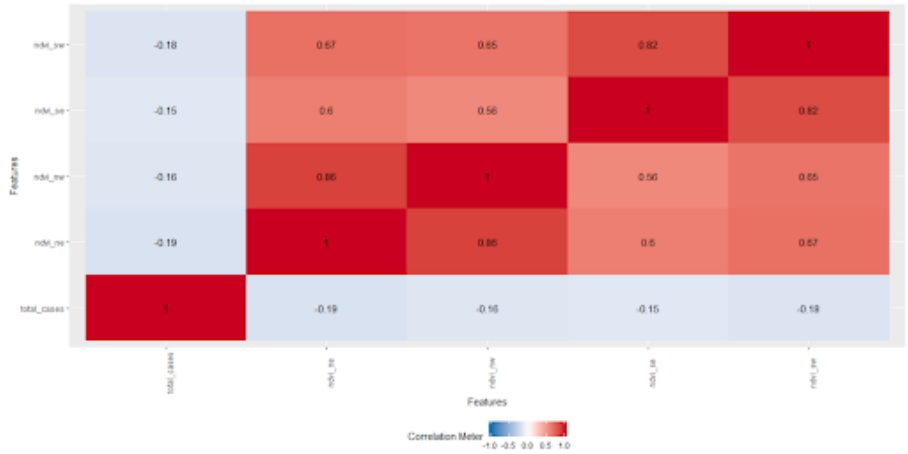

Figure 19: Correlation Plot of Vegetation Variables

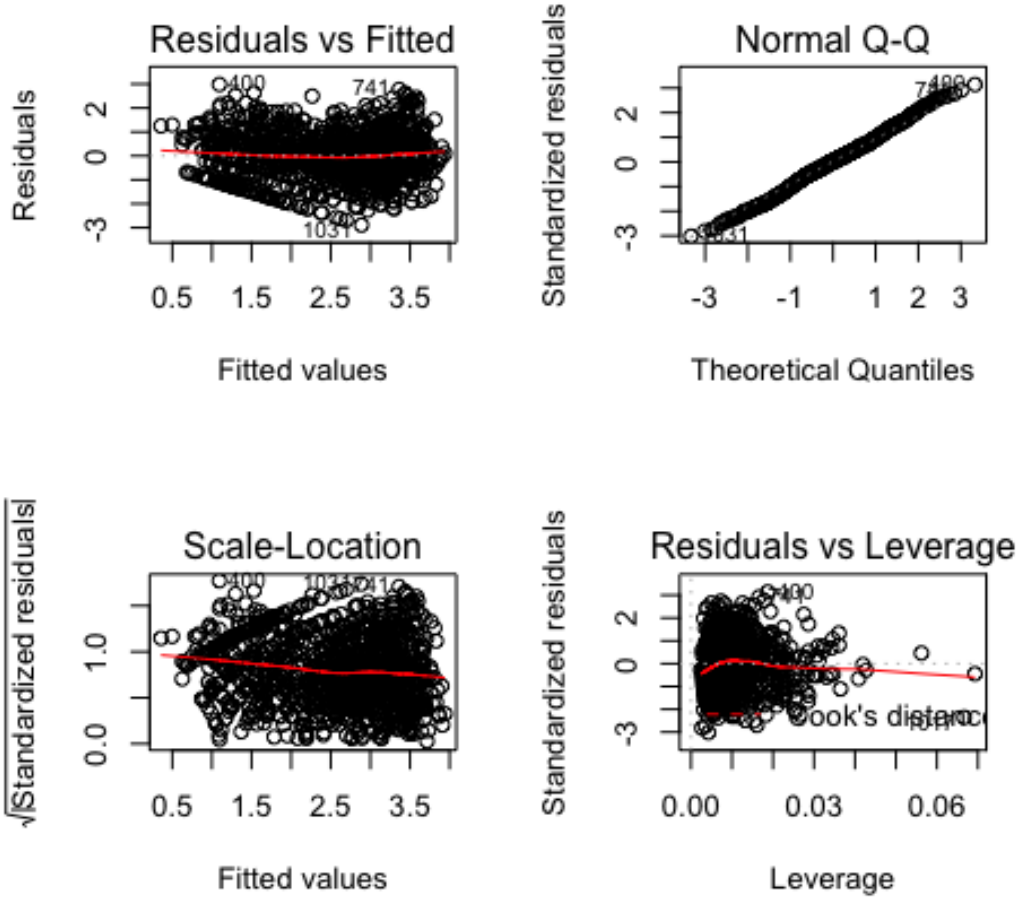

Figure 20: Linear Model Diagnostic Plots

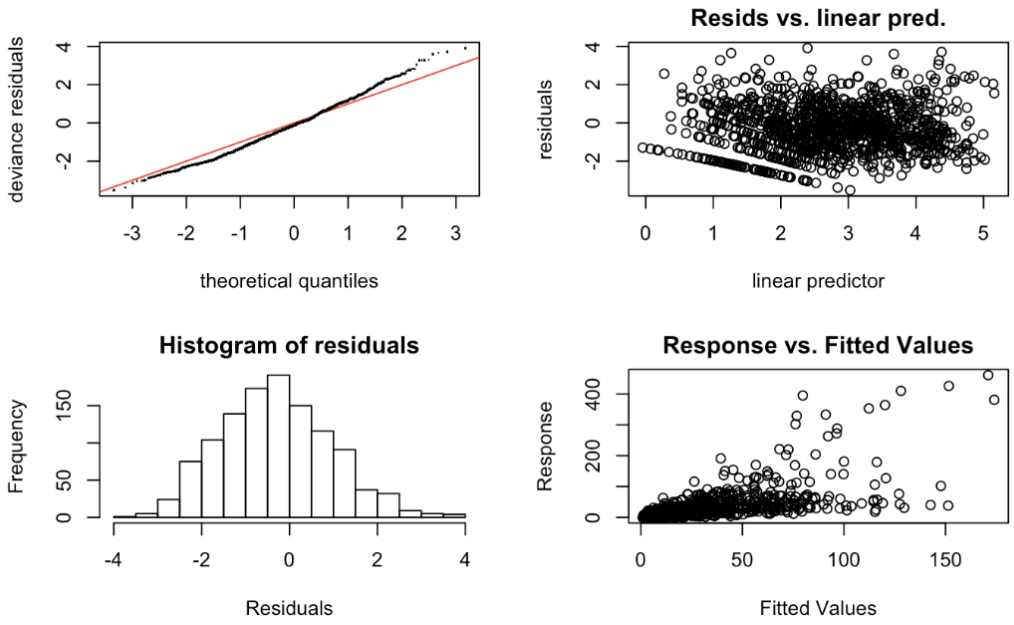

Figure 21: Generalized Additive Model Diagnostic Plots

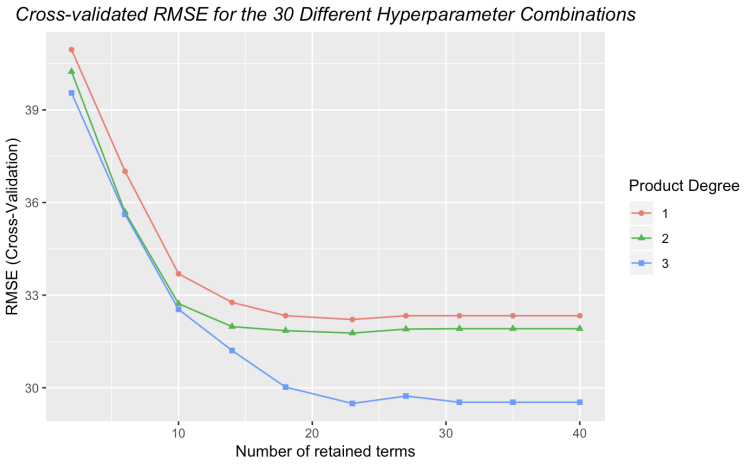

Figure 22: Cross-validated RMSE for the 30 Different Hyperparameter Combinations for MARS

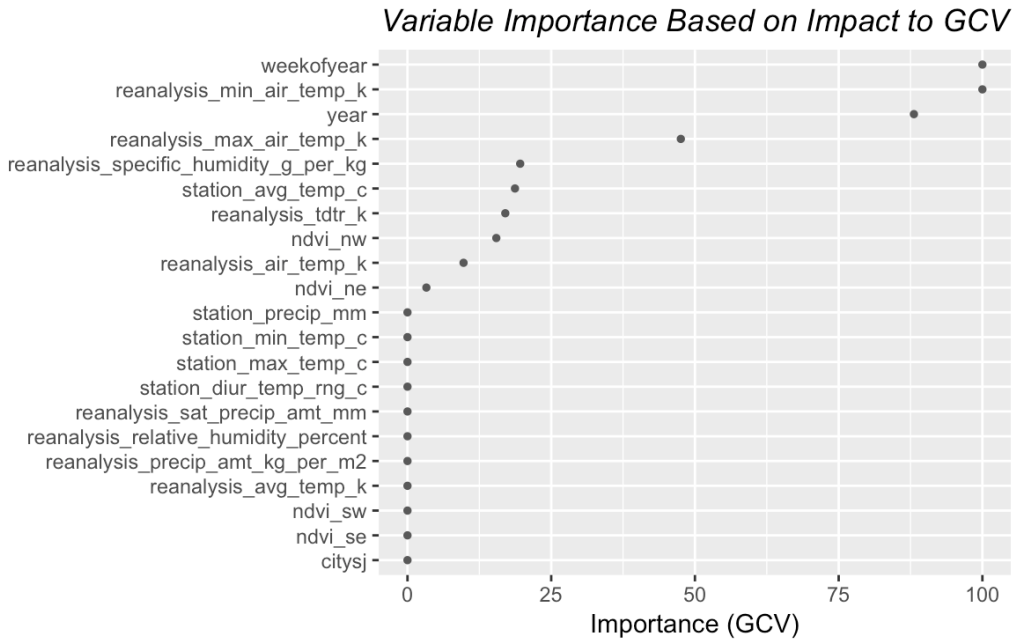

Figure 23: Variable Importance in MARS

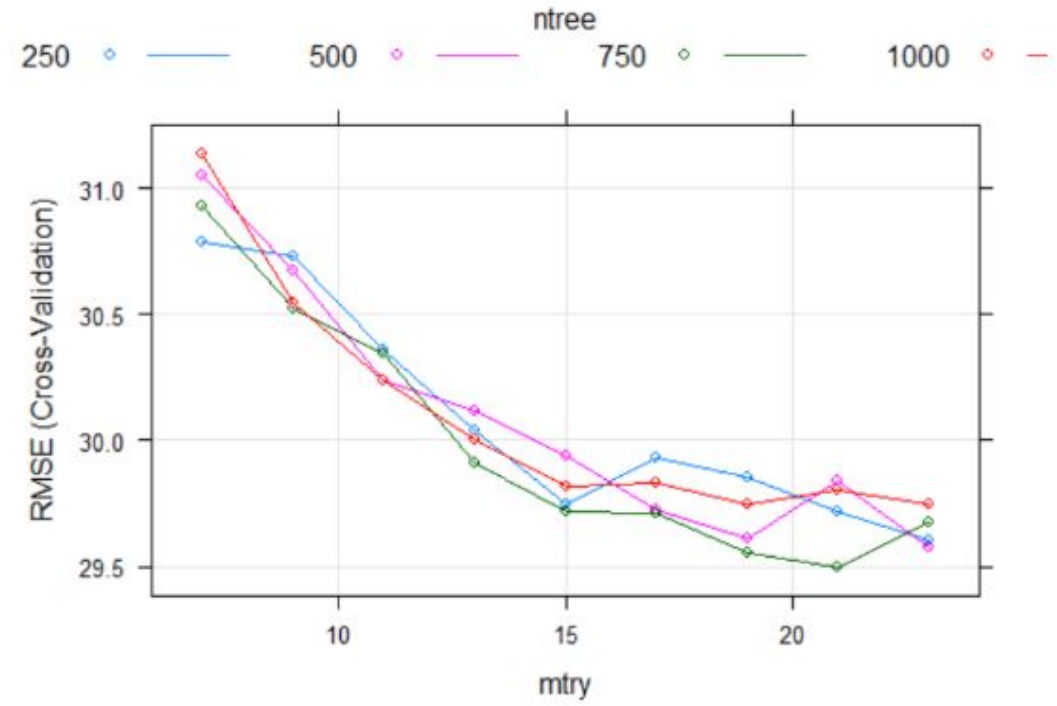

Figure 24: Tuning of Random Forest Model

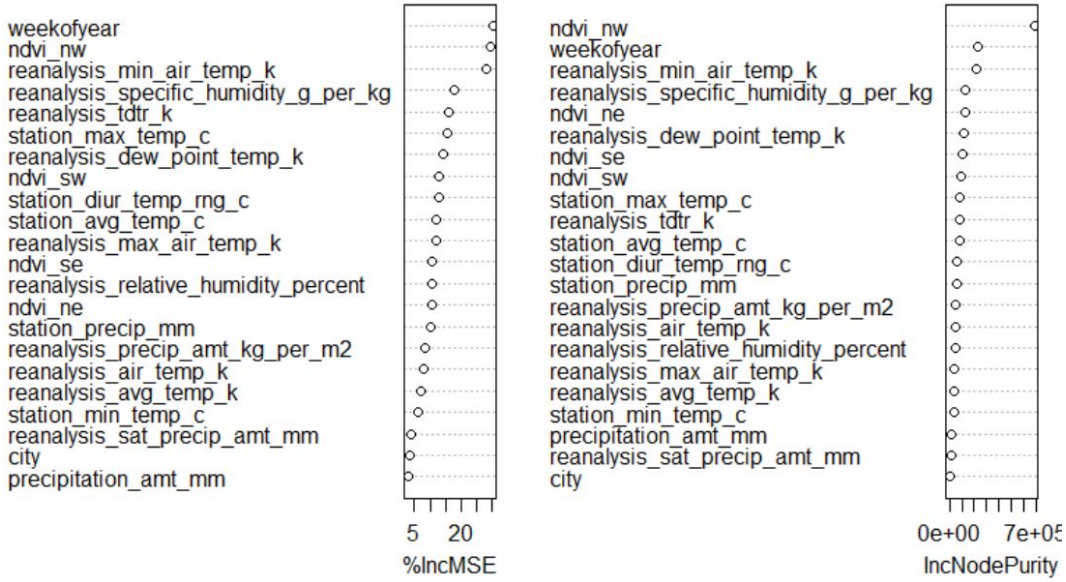

Figure 25: Importance Plot of Random Forest

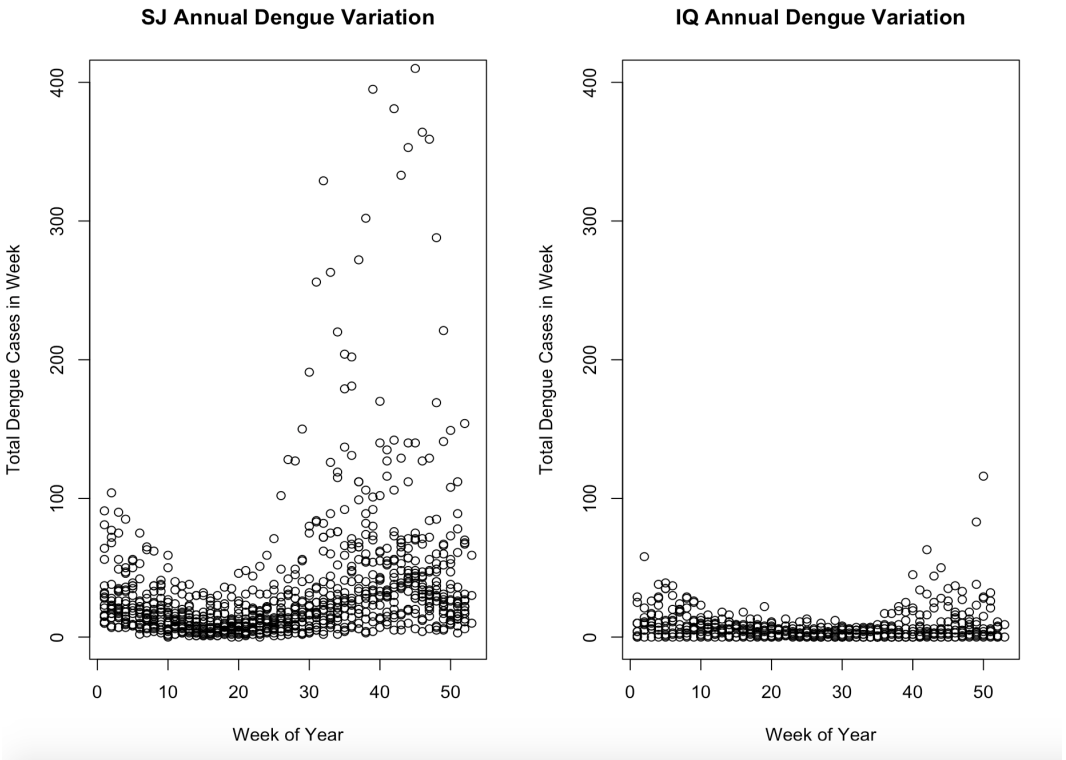

Figure 26: Different Seasonal Patterns Between Two Cities