Weekly Dengue Cases forecasting at Iquitos and San Juan with Machine Learning Methods

Machine Learning for Cities

Jianqi Tang (jt2900@nyu.edu)

Rongjian Yang (ry1121@nyu.edu)

Seunggyun Han (sh5498@nyu.edu)

Siqi Huang (sh5688@nyu.edu)

Link to the code and datasets

https://drive.google.com/open?id=1e7Ry_VkcIw8MGqRkkeCvFTzjzuShCDKQ

Introduction

Project Goals

The purpose of this study was to develop a machine learning model to forecast the weekly total cases of Dengue fever in Iquitos and San Juan, which could provide early warning of a dengue outbreak several months in advance to allow sufficient time for effective control to be implemented.

What specific question does this project solve? How might the results be used in practice?

Public health is one of the top topics of public concern, people hope the information is transparent and timely. In the project, the team aims to build a weekly predictive model for total cases of Dengue fever, one of the world's fast-spreading mosquito-borne diseases and serious public health issues in tropical and subtropical areas. It's estimated that about four hundred million are infected by the virus every year, and one hundred million people gradually get sick. However, there is no effective vaccine to prevent the disease, thus the immediate prediction is crucial in terms of personal prevention and the environmental management for the prevention of this epidemic disease.

In this context, for tropical and subtropical areas, the prediction of dengue disease could be included in their urban informatics system, help the authority monitor the spread in real-time, give the residents more information about the disease, and warn them to take correspondent actions if needed. According to WHO, the dengue disease has shown temporal patterns especially related to during and after rainy seasons (WHO, n.d). For this reason, the team believes predictive machine learning models with temporal factors like weather or historical records of the disease cases can help for active intervention for the prevention of the disease. As a part of the project, the team joined a competition (DengAI: Predicting Disease Spread). The accuracy of the models was measured by Mean Absolute Error.

Related Work

Many researches focused on identifying the significant factors for prediction, since new factors are desired to enhance the accuracy of prediction. Statistical based-analysis showed positive correlation between female mosquitoes infection rates and season with the number of the cases (Siriyasatien et al., 2016). Season variables include seasonal temperature, rainfall, humidity, and wind had indicated significantly high correlation coefficients.

Beyond traditional data, a group of scholars (Liu et al., 2019) introduced Internet search data (such as Google, Wikipedia) into the monitoring data. They select related keywords from a website of keyword mining, and filter them with correlation coefficient, and finally calculate the weights of remaining terms. A low cost prediction model developed by Hii, Zhu, et al. (2012) just used weekly mean temperature and cumulative rainfall, allowing warning 16 weeks in advance. The model suggests that mean temperature, rainfall, past dengue cases, season and trend explained 84% of the variance of weekly dengue distribution.

Learning from previous work, the approach we would like to address including different machine learning methods with different feature selection. We set three different approaches to preprocess a historical dataset as input variables. In each approach, we use different lag values and apply 2 ~ 4 different models [Figure 2.A]. By comparing the results of the approaches and models, we identify limitations of each method and suggest further steps of the project to get better results.

Data

In the project, we used a dataset from the competition DengAi: Predicting Disease Spread, which generates data of 24 climate features (input) and weekly case number of the disease (label) for two cities, San Juan and Iquitos [Table 1.A]. Input and labeled datasets were given as a training dataset, and a different input dataset was given for testing. The input dataset for training consisted of 1456 entities (936 for San Juan, 520 for Iquitos) with a spatial feature (city), temporal features (year & week), and climate features (temperature, humidity, dew point, etc.). There were 548 missing values in the given dataset. Most of the missing values were in the 'NDVI' related columns [Figure 1.A].

To understanding the dataset, exploratory data analysis was conducted. First, we checked correlations between the features and the number of total cases through a heatmap [Figure 1.C]. However, there were no strong correlations between the input features and a label feature even though the temperature features were strongly correlated. Second, we looked at data distributions in each feature. Histograms in [Figure 1.D & E] showed the distributions in each feature. Regarding San Juan, features related to precipitation showed left-skewed distribution, and features regarding temperature and dew points showed right-skewed distribution. Third, we checked autocorrelation. It shows how a weekly incidence value correlates with the past 100 week lag values. We observe that weekly incidence values have over 0.4 correlation with its past 10 week values and only -0.2 correlation with value at the same time last year (52 weeks ago) [Figure 1.B]. This observation inspires us to create an autoregressive model that utilizes lag values from the previous few weeks. Finally, we also applied seasonal decomposition in label data (weekly total cases of the dengue patients) [Figure 1.F & G]. The seasonal pattern was clear, and several outbreaks also existed.

Methods

All of our methods are compared under three schemes of time series prediction [Figure 2.A.].

- **Forecast of the number of the cases based solely on environmental factors in the past**
- **Recursive Forecast of one week ahead cases based on environmental factors and lags constructed from predicted cases**
- **Recursive forecast of one week ahead cases in next week based on environmental factors and lags constructed from environmental factors based predicted cases**

Processes of those approaches would be discussed in detail at the end of this chapter. Accuracies of the models in each method were measured as ‘Mean Absolute Error’. Each approach was tried with 3 ~ 4 different machine learning algorithms. For optimizing the results, each approach chose slightly different features and data preprocessing strategies. Below were common things about the team preprocessed data and models that were applied in the approaches.

Data Preprocessing:

To measure the performance of our results and tune the parameters, we split our training data into 70% training and 30% validation based on its time order. Normalization, minmax scaling, log transform of features and cases were also performed, but none of the preprocessing steps were found to have a large impact on the result.

Models:

We investigated four popular machine learning methods for time series forecasting: Gaussian Process, random forest, support vector regression, and xgboost.

The Gaussian process is a nonparametric Bayesian approach for regression that works well for small dataset and provides uncertainty. It basically remembers the "trajectory" of the past incidence and produces predictions from the same distribution as past data. Previously the Gaussian process has been successfully used to make Dengue case prediction for peak week,peak incidence and total peak incidences (Johnson,L.R, 2017). Though this task has different objectives of minimizing mean absolute error. We think it's still worth exploring Gaussian process's power to capture the nonstationary dynamics of the data.

Random forest is considered a general-purpose non-linear prediction model that could fit any given dataset. In the DengAI case, the team believes the unique quality of decision trees can provide a different perspective of how data is being processed by algorithms. By determining the information gain of each decision node in the tree structure, we might find a good fit that the model could perform well by predicting the number of possible dengue patients in the future given proper environmental features in the dataset. Random forest uses bagging strategy to build multiple trees parallelly and result in a balanced output from trees.

SVM regression (Support vector regression) is considered a nonparametric technique which relies on kernel functions. SVR desires the distance of the farest point to the hyperplane is the smallest. There is a previous Dengue forecast model using SVR reaching a good result among other regression models (Guo et al., 2017). So we want to try whether this model can work well with our dataset.

XGBoost (Extreme gradient boosting) is an optimized distributed gradient boosting algorithm (XGBoost, n.d.). As one of the popular models in data competitions and machine learning competitions, it also has been applied to various research about time-series prediction problems from electricity consumption to store sales (Wang, Shi, Lyu, and Deng, 2017; Pavlyshenko, 2016). For this reason, the team also expects XGBoost would be fit for the DengAI case. XGBoost also provides approximate interpretability with the importance of each input feature.

Approaches:

1) Forecast of the number of the cases based solely on environmental factors in the past (Figure 2.B)

In this part, we use the environmental features of three weeks ago to predict the number of dengue patient cases. We split the given dataset into two datasets by cities. The models are trained with each dataset separately, so 8 different models were built (2 cities, 4 algorithms) as a result of this phase. We use only six features to avoid overfitting. For feature selection, we consider factors like 'temperature', or 'humidity' that are known as related to the mosquito population, which is directly related to the dengue disease (WHO, n.d). We also briefly check the feature importances using XGBoost. Based on this, the team selects the features which have larger importances among the weather features. Also, We decide to use the 'ndvi_se' feature, which has the most substantial feature importance [Figure 2.A].

2) Recursive Forecast of one week ahead case value based on environmental factors and number of cases in the past (Figure 2.D)

Another method we attempted is building an autoregressive model. **We make predictions of one week ahead case value based on several previous week case values as well as previous weeks environmental factors.** In doing so we hope to track the nonstationary dynamics of the case values. In our validation set, we do not have the actual case values available. However since our training and validation set are continuous in time, we can first make predictions of the first few values in the validation set from the last few values in the training set. Then we can recursively construct new lags from predicted values. We choose lags as the previous two weeks because we want to use the absolute value of the recent past as well as the gradient information as predictors.

3) Recursive forecast based on environmental factors, week of year, and previous 2 week prediction lags (Figure 2.E)

This method combines method 1 and method 2. When making predictions in validation set, instead of constructing new lags from recursively predicted values, we use “guess” from predictions made by a model trained with environmental factors only. This feature selection had been applied to Random Forest and XGBoost only. Two RF/XGBoost models were built on the dataset to complete the task, and all models’ parameters were tuned using comprehensive GridSearch to achieve the best MAE score.

To build the first RF/XGBoost model, all environmental features were selected along with ‘week of year’ as a time variable, feature ‘city’ became a binary feature to differentiate San Juan, Costa Rica and Iquitos, Peru. Then, the result of the first RF/XGBoost was transformed to a new feature ‘Previous_2Week_Patients’, the sum of previous two weeks Dengue patients from the first RF/XGBoost model prediction. The purpose of this step is to hope the second RF/XGboost model could pick up the past two-week patients developing trend to achieve better performance.

After adding the ‘previous 2week patients’ feature, the second RF model will be applied on that dataset to generate final prediction. Feature importance between two models is in [Figure 3.P].

Using a lag feature based on prediction value could be considered risky in machine learning. Because when the second model picks up that featrue’s pattern, it could either amplify the final prediction to correctly locate the peak of a Dengue epidemic. Or, it could downplay the prediction result by missing the peak of an incoming Dengue epidemic. So, it is a dilemma, but worth a try.

Results

1) Forecast of the number of the cases based solely on environmental factors in the past (Figure 3. A~F)

From the result we see that parametric method xgboost and random forest can learn the yearly cyclic patterns in the training data and make predictions based on the average cycle amplitude [Figure 3.C]. Thus it tends to overestimate the validation data, since in the validation dataset there are less frequent cycles. In contrast, the nonparametric method SVM and Gaussian process learns mean value where the most data are centered around. Since most of the time the cases remain at a low level. On average, SVM and Gaussian processes have a better performance in MAE with out-of-sample dataset [Figure 3.A]. Regarding feature importances in the random forest and xgboost, 'ndvi' feature shows the largest number in both of the models [Figure 3.F].

2) Recursive forecast of number of cases in next week based on environment factors and predicted cases in the past (Figure 3.G~L)

In this case we see that MAE performance improves for all methods except for SVR, and the overestimation problem of RF is mitigated [Figure 3.G]. All methods started with good performance initially but the results diverge as prediction moves further into the future [Figure 3.I]. Random forest and Gausian processes pick up the upward trend but could not follow the abrupt change. SVR can predict the upward trend but could not predict the downward trend. XGB could not capture any of the peak points, it returned values around 10 [Figure 3.I]. In addition to this, with feature importances from the random forest and xgboost model, we figure out that the predictions of both models mostly depend on the total number of dengue cases of one week ago [Figure 3. L].

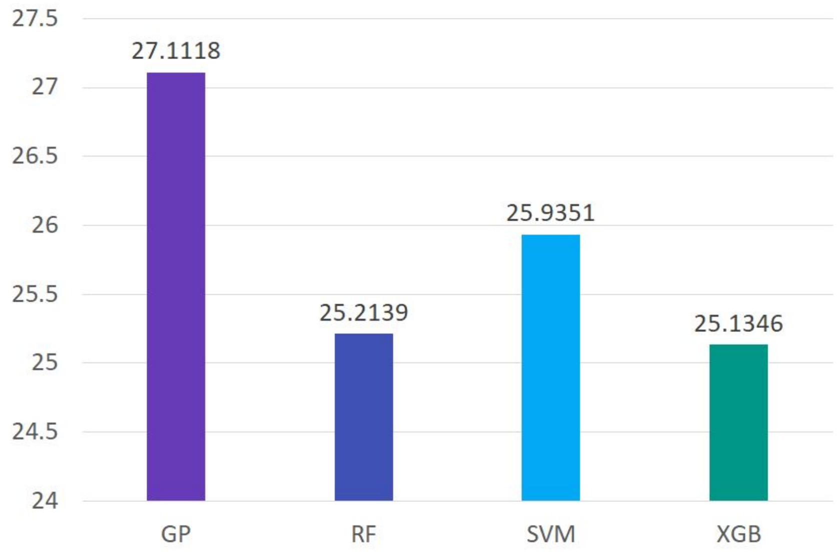
3) Recursive forecast of one week ahead cases in next week based on environmental factors and lags constructed from environmental factors based predicted cases (Figure 3.M~P)

From this result, we can see both XGBoost and Random Forest successfully traced the trend and peaks of predicting Dengue patients. MAE achieved 9.54 and 11.14 for RF and XGBoost, a score we considered well performed [Figure 3.M]. In out-sample validation, both XGBoost and

Random Forest successfully predict the time when there is a peak of Dengue patients. But compared to in-sample performance, both models generated low prediction values (most less than 20, and some equals to 0) that failed to measure the scale of a peak but accurately predicted when the peak would take place [Figure 3.N & G].

Conclusion

Figure. MAE score of four models performance over testset.



After each team member completed a comprehensive research on the models, we used the model we constructed to make predictions on the Test dataset the competition provided. Although overall, models performed pretty well in training, all of them successfully picked up the trend and peak dengue patients, and also in short term future validation. But in the test set, the result is somewhat off the chart because none of the four models has a MAE lower than 25 compared to 8-15 MAE in the training set. The reason to explain this situation could be complicated, because an Dengue disease outbreak is not entirely triggered by environmental change but also many factors of human activity interference. In our case, we did not include any other feature except time and environmental indicators, and that could be the cause of long term underprediction.

Future improvements:

The result of approach 3 gives us good prediction of time of peak occurrences. So it is possible to use this information as a feature for another model to forecast once again. Since disease transmission is a complicated problem that involves human and societal factors, using external data sources besides environmental factors should account for some of the nonstationarity in time series.

Contribution of each team member to the project

All of the team members made contributions to the project, attended all of the group meetings to discuss the project, and built and optimized one model.

Reference

Chen, W. (2018). Dengue outbreaks and the geographic distribution of dengue vectors in Taiwan: A 20-year epidemiological analysis. *Biomedical Journal*, 41(5), 283-289. doi: 10.1016/j.bj.2018.06.002

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y. and Ma, W., (2017). Developing a dengue forecast model using machine learning: A case study in China. *PLOS Neglected Tropical Diseases*, 11(10), p.e0005973.

Hii, Y., Zhu, H., Ng, N., Ng, L. and Rocklöv, J., (2012). Forecast of Dengue Incidence Using Temperature and Rainfall. *PLoS Neglected Tropical Diseases*, 6(11), p.e1908. doi: 10.1371/journal.pntd.0001908

Liu, D., Guo, S., Zou, M., Chen, C., Deng, F., Xie, Z., Hu, S. and Wu, L. (2019). A dengue fever predicting model based on Baidu search index data and climate data in South China. *PLOS ONE*, 14(12), p.e0226841. doi: 10.1371/journal.pone.0226841

Johnson, L. R., R. B., Cohen, J., Mordecai, E., Murdock, C., Rohr, J., Ryan, S. J., Sadie J., Stewart-Ibarra, A. M., Weikel, D. (2017). Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: a dengue case study, arXiv:1702.00261.

Pavlyshenko, B., M. (2016). Linear, machine learning and probabilistic approaches for time series analysis, *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP),* 377-381, doi: 10.1109/DSMP.2016.7583582

Siriyasatien, P., Phumee, A., Ongruk, P., Jampachaisri, K. and Kesorn, K., (2016). Analysis of significant factors for dengue fever incidence prediction. *BMC Bioinformatics*, 17(1). doi: 10.1186/s12859-016-1034-5

Wang, W., Shi, Y., Lyu, G., and Deng, W. (2017). Electricity Consumption Prediction Using XGBoost Based on Discrete Wavelet Transform, *2017 2nd International Conference on Artificial Intelligence and Engineering Applications*, 716-729, doi: 10.12783/dtcse/aiea2017/15003

WHO. (n.d). *Dengue and severe dengue*, Retrieved May, 5th, 2020, from https://www.who.int/health-topics/dengue-and-severe-dengue#tab=tab_1

XGBoost. (n.d.). *XGBoost Documentation*, Retrieved May, 5th, 2020, from https://xgboost.readthedocs.io/en/latest/

Appendix

1. Data

Table 1 A. Features in the given dataset

	Feature	Data Type
Input (X)	city	string
	week of year	int
	Maximum temperature	float
	Minimum temperature	float
	Average temperature	float
	Total precipitation	float
	Total precipitation (reanalysis)	float
	Diurnal temperature range	float
	Total precipitation	float
	Mean dew point temperature	float
	Mean air temperature	float
	Mean relative humidity	float
	Mean specific humidity	float
	Total precipitation	float
	Maximum air temperature	float
	Minimum air temperature	float
	Average air temperature	float
	Diurnal temperature range	float
Output (y)	Total cases	int

Figure 1.A. Number of missing values in each features

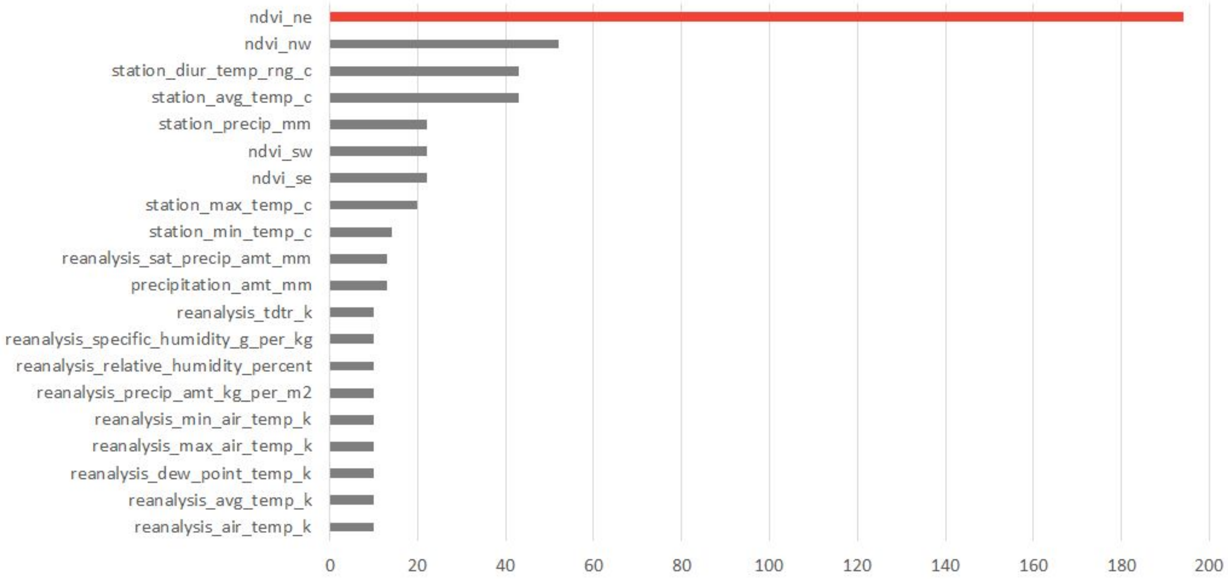


Figure 1.B. Autocorrelation

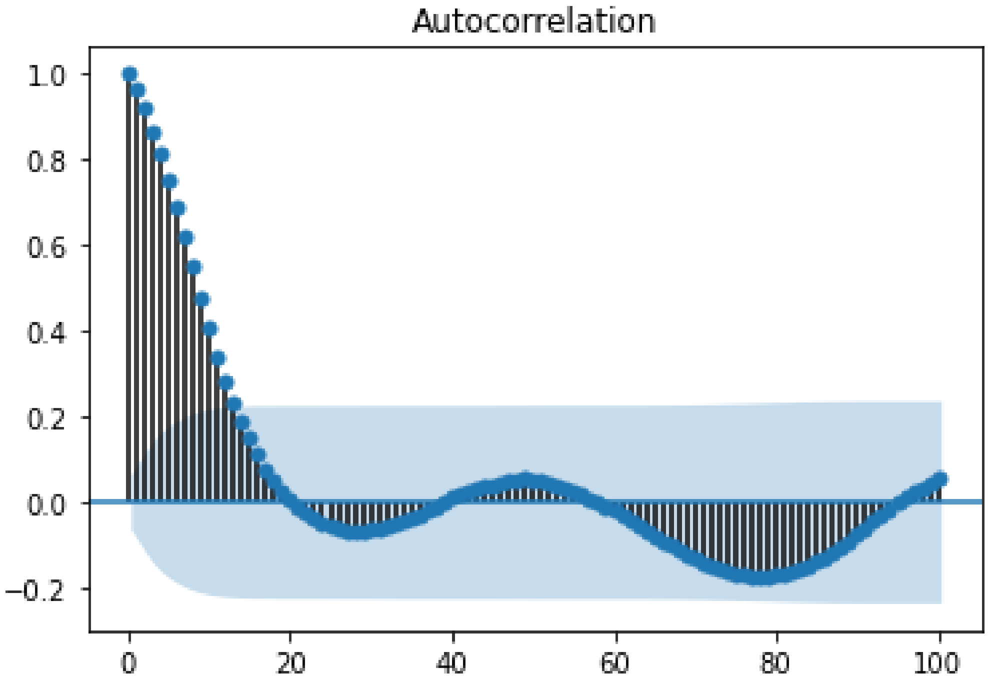


Figure 1.C. Heatmap of correlations between variables

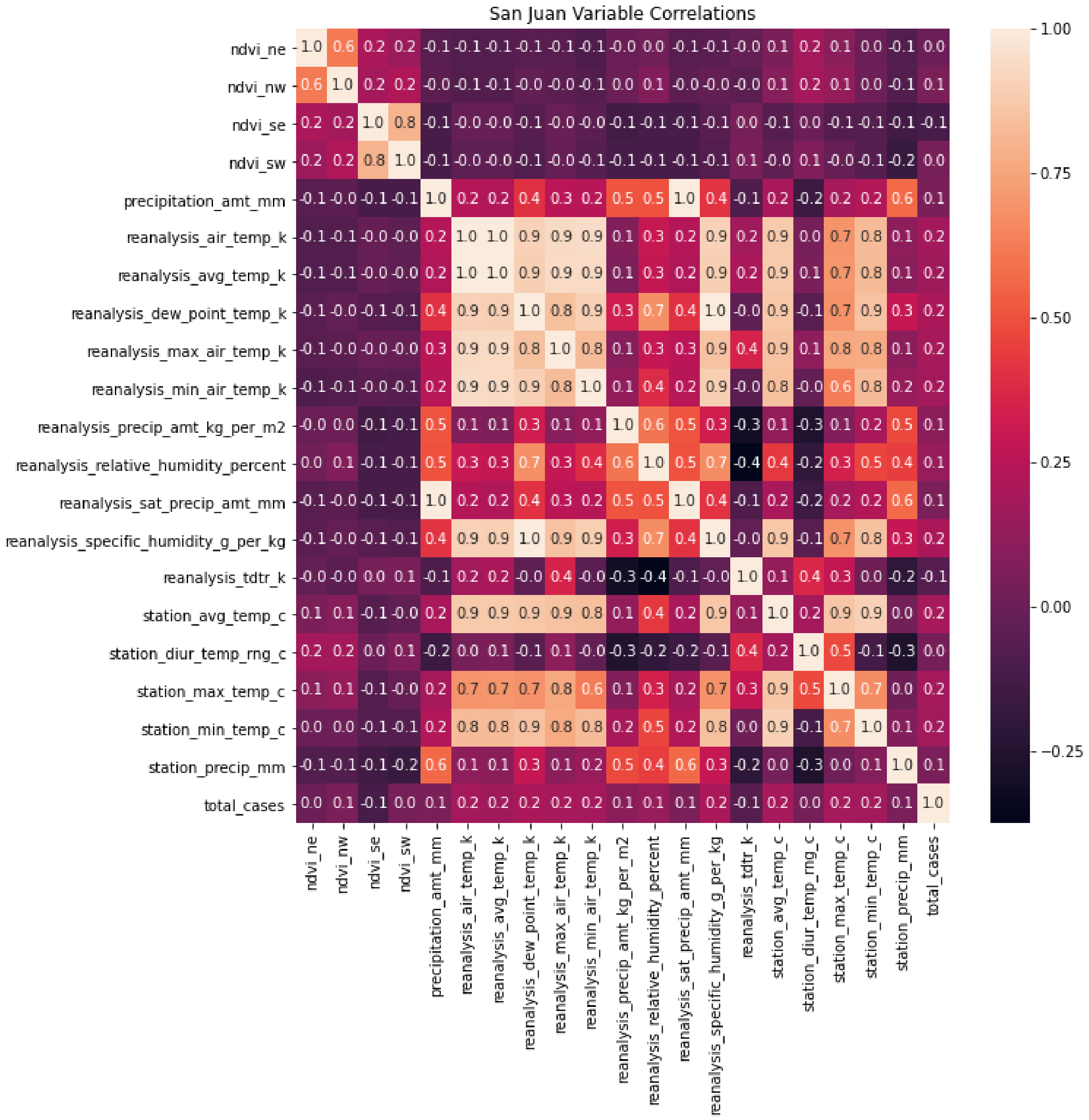


Figure 1.D Histograms about the features (San Juan)

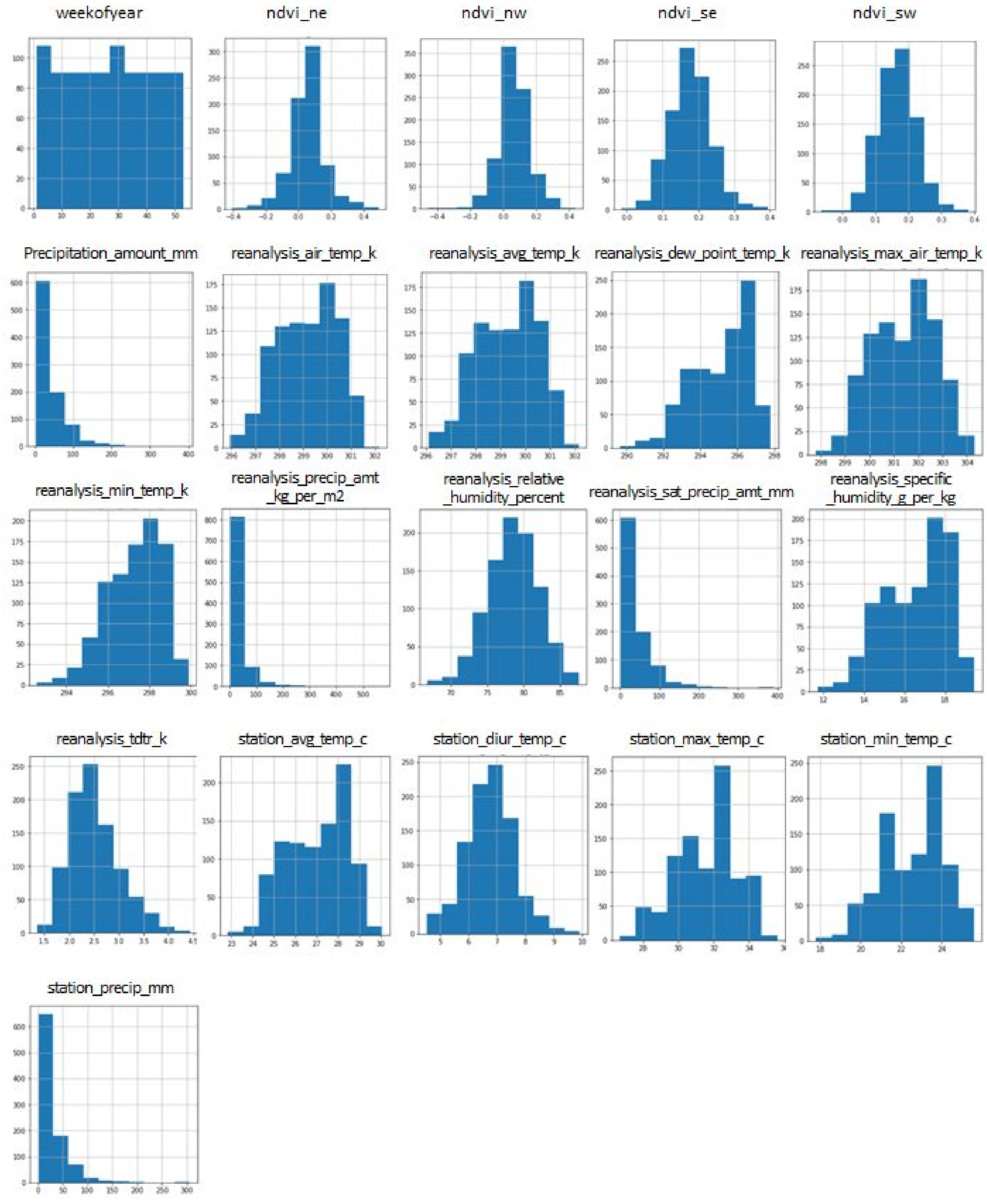


Figure 1.E Histograms about the features (Iquitos)

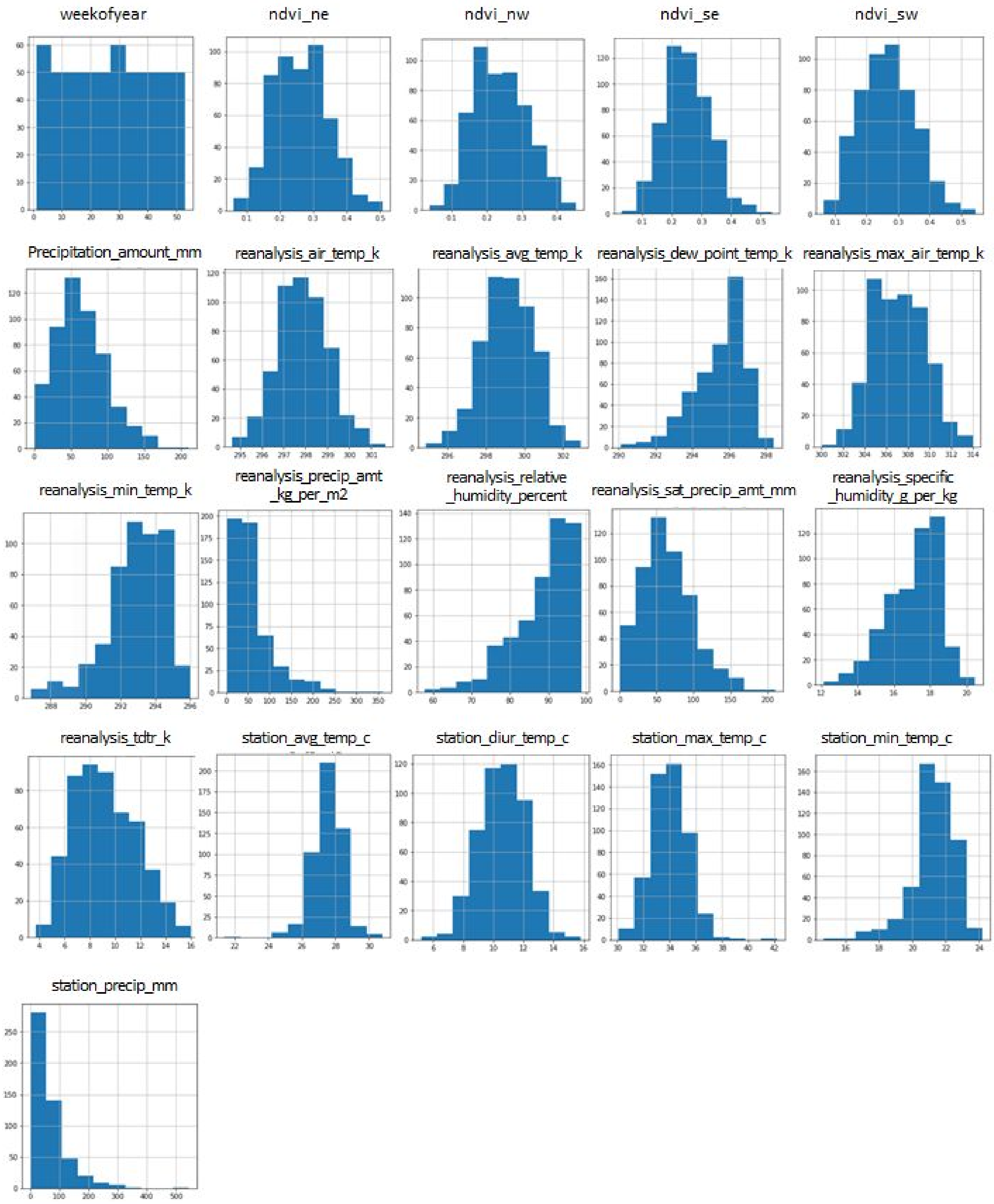


Figure 1.F Seasonal decomposition of the weekly total cases (San Juan)

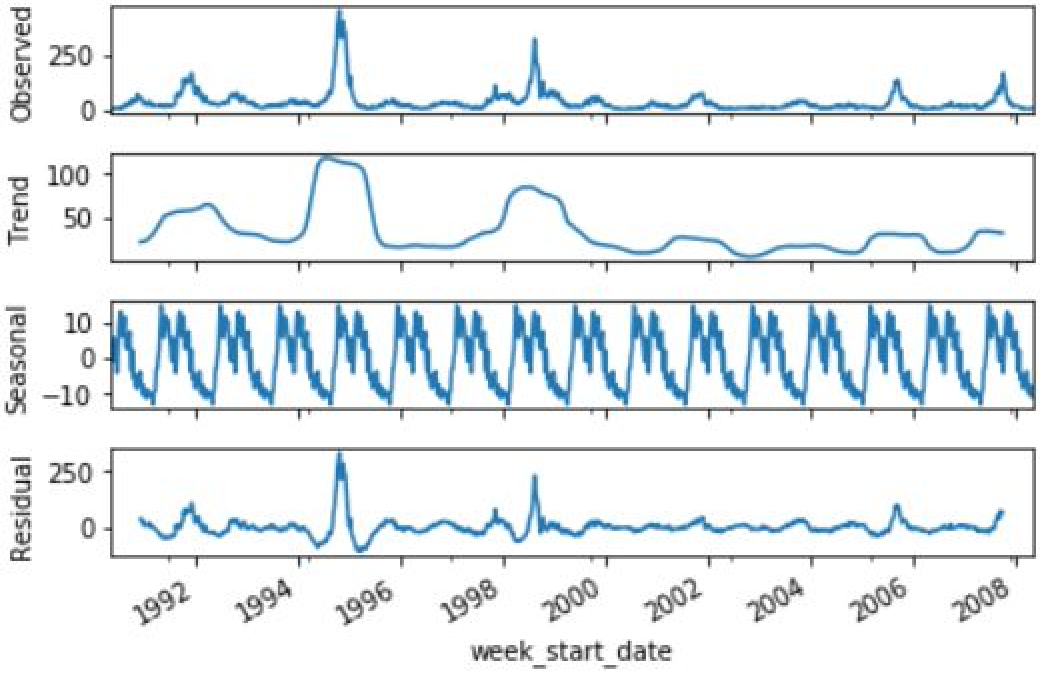
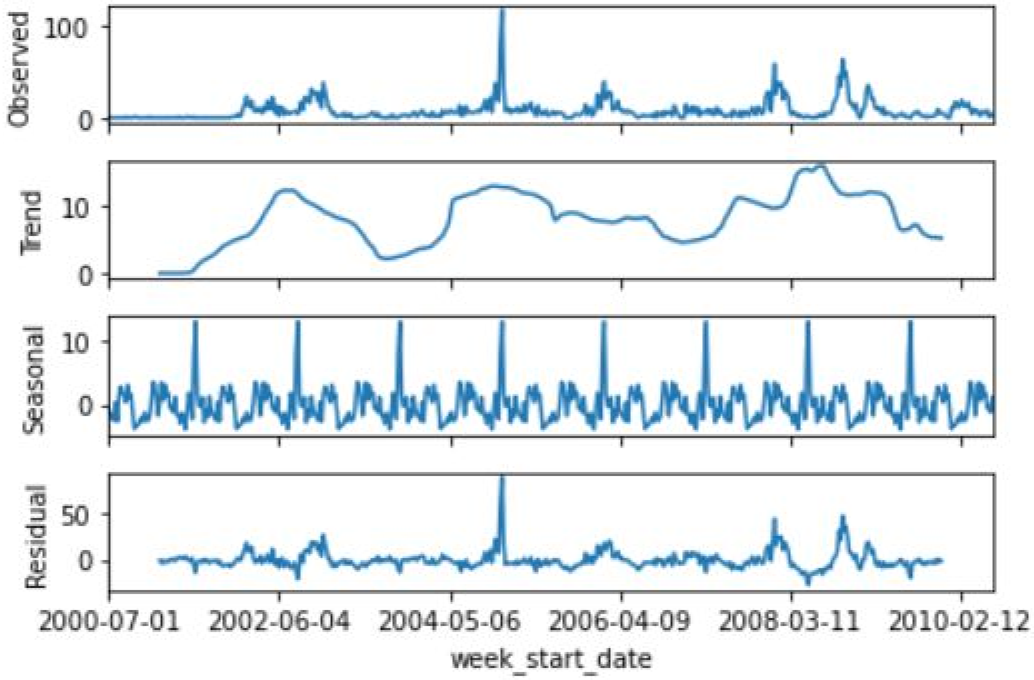


Figure 1.G Seasonal decomposition of the weekly total cases (Iquitos)



2. Method

Figure 2.A Overview

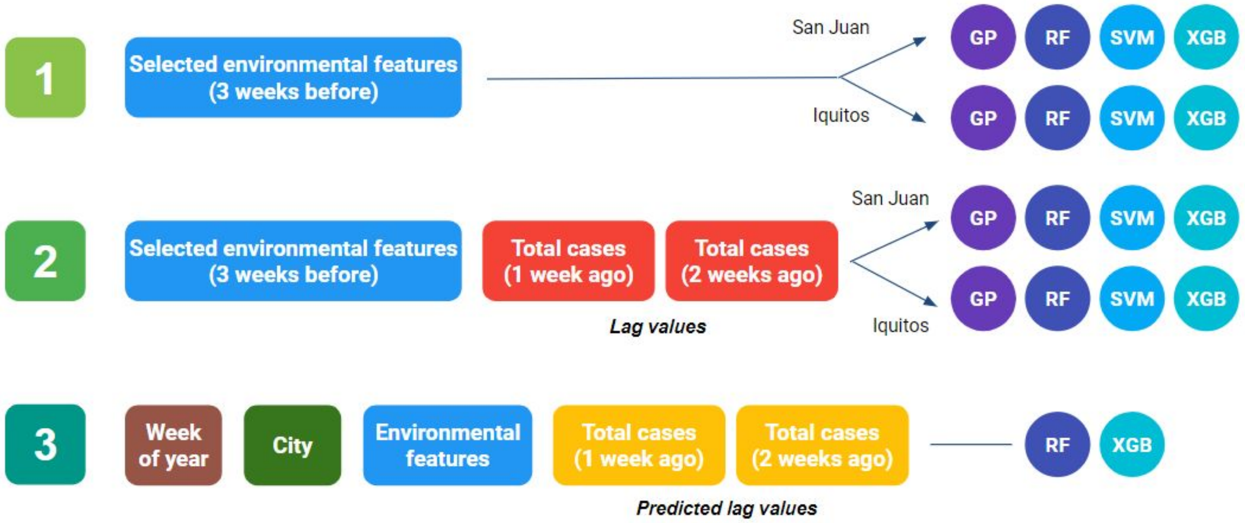


Figure 2.B. Diagram - Forecast of the number of the cases based solely on environmental factors in the past

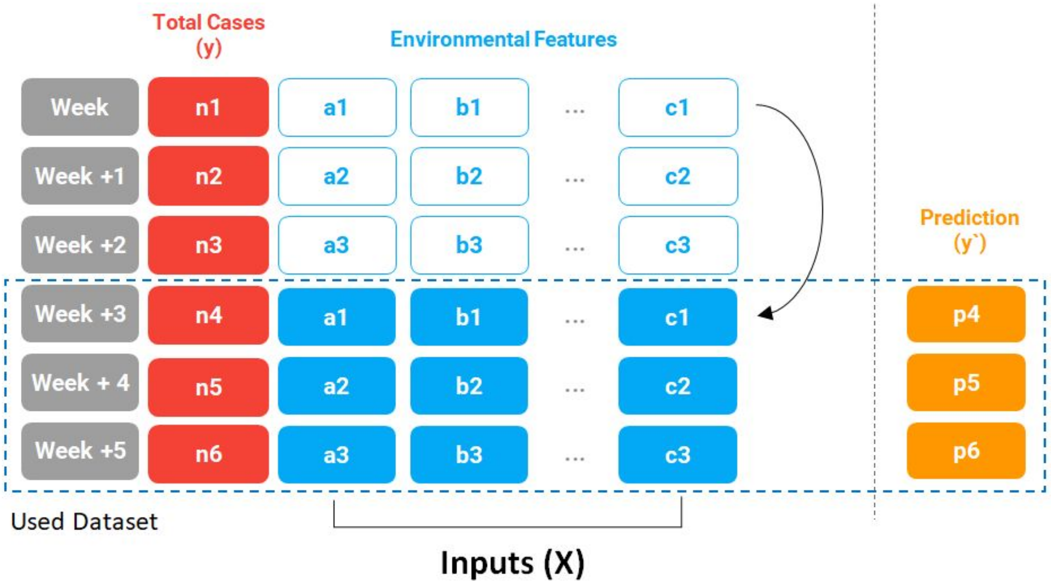


Figure 2.C. Feature Importances in XGBoost

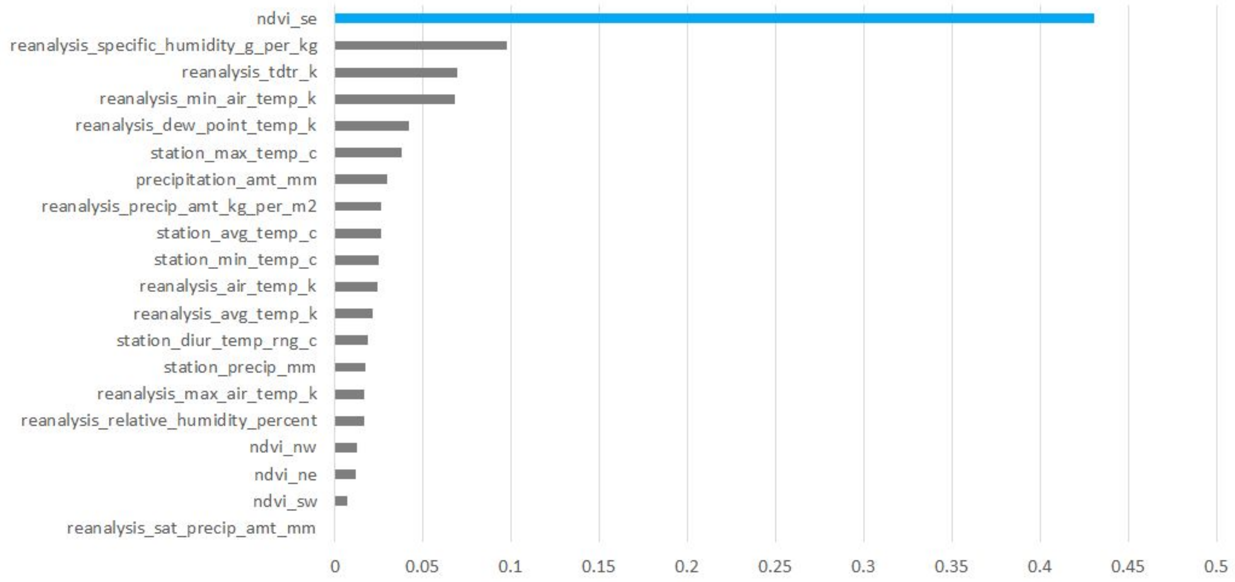
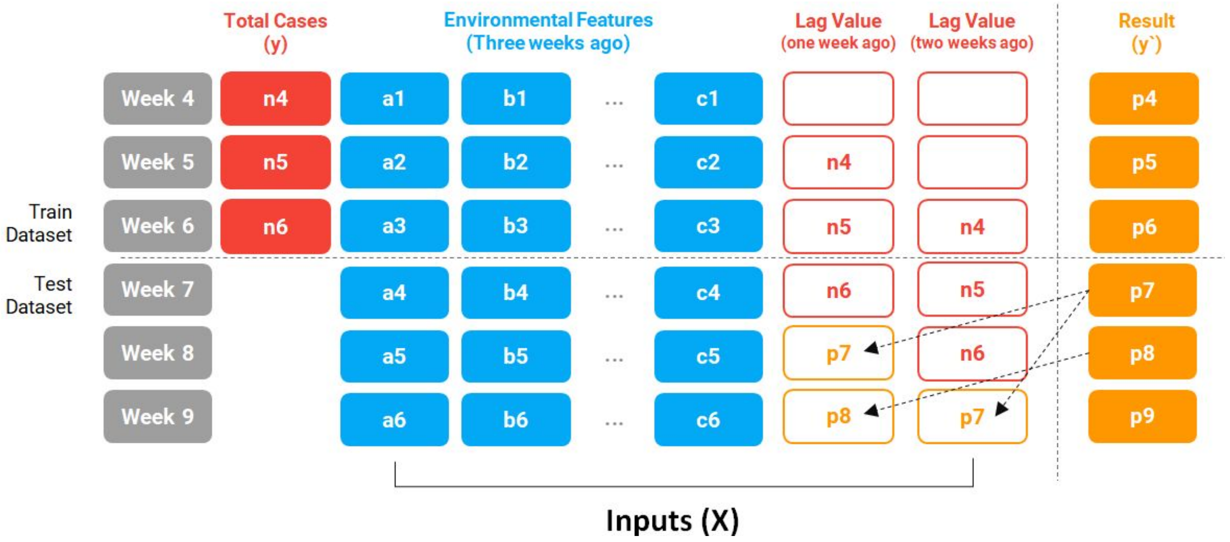
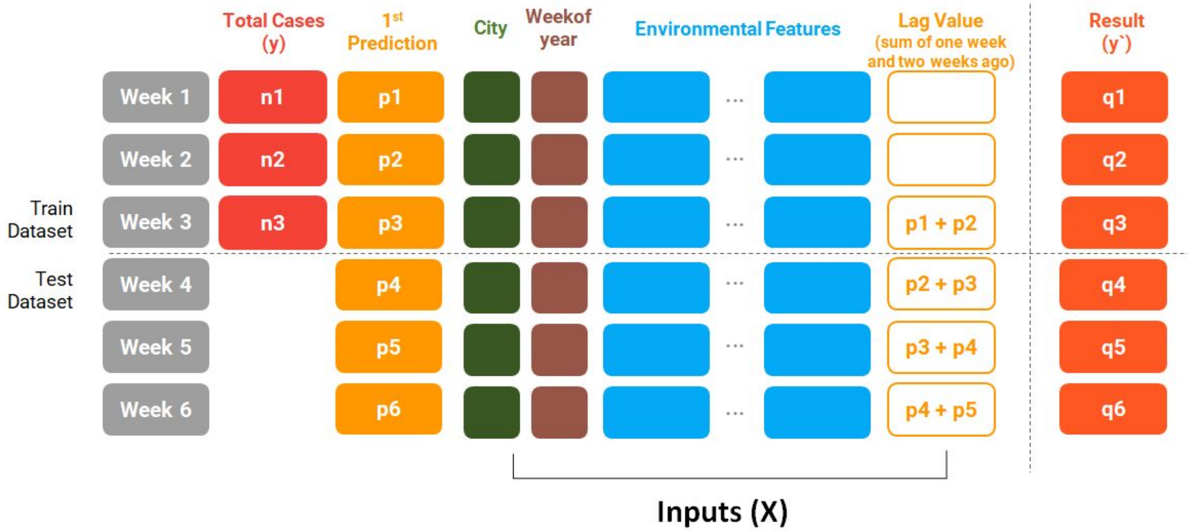


Figure 2.D. Diagram - Recursive Forecast of one week ahead cases based on environmental factors and lags constructed from predicted cases



		Total Cases (y)	Environmental Features (Three weeks ago)				Lag Value (one week ago)	Lag Value (two weeks ago)	Result (y`)
	Week 4	n4	a1	b1	...	c1			p4
	Week 5	n5	a2	b2	...	c2	n4		p5
Train Dataset	Week 6	n6	a3	b3	...	c3	n5	n4	p6
Test Dataset	Week 7		a4	b4	...	c4	n6	n5	p7
	Week 8		a5	b5	...	c5	p7	n6	p8
	Week 9		a6	b6	...	c6	p8	p7	p9

Figure 2.E. Diagram - Recursive forecast of one week ahead cases in next week based on environmental factors and lags constructed from environmental factors based predicted cases



3. Results

Figure 3.A ~ F : Results from Approach 1

Figure 3. A In-sample and out-of-sample score based on selected environmental factors (3 weeks ago)

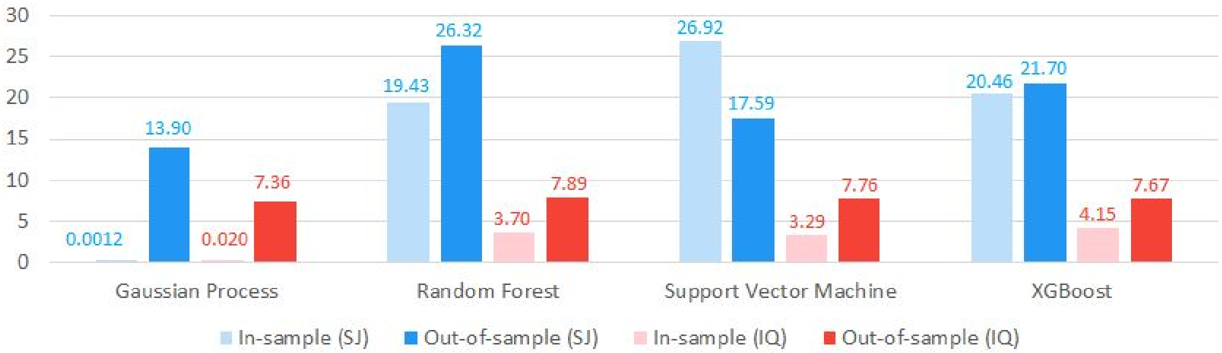


Figure 3. B Prediction for in-sample (San Juan) based on selected environmental factors (3 weeks ago)

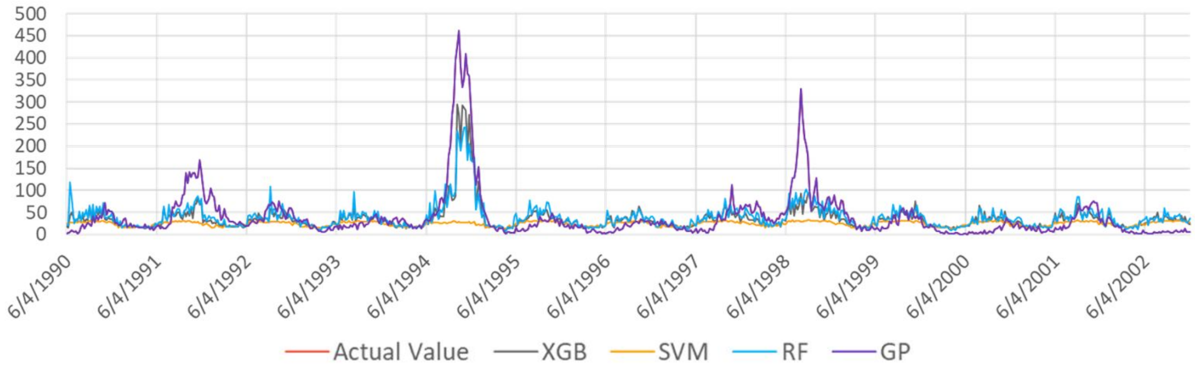


Figure 3. C Prediction for out-of-sample (San Juan) based on selected environmental factors (3 weeks ago)

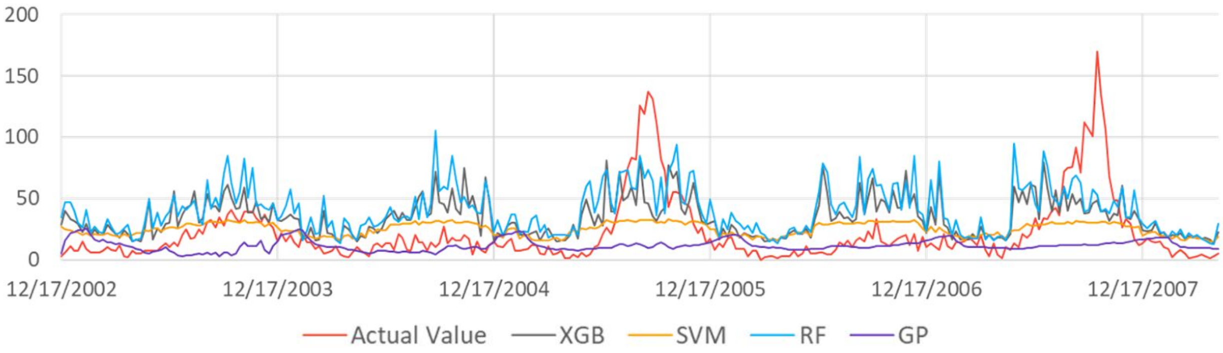


Figure 3. D Prediction for in-sample (Iquitos) based on selected environmental factors (3 weeks ago)

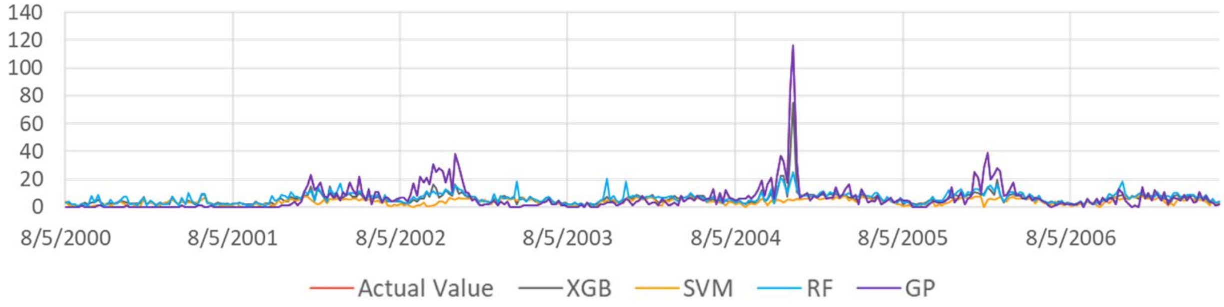


Figure 3. E Prediction for out-of-sample (Iquitos) based on selected environmental factors (3 weeks ago)

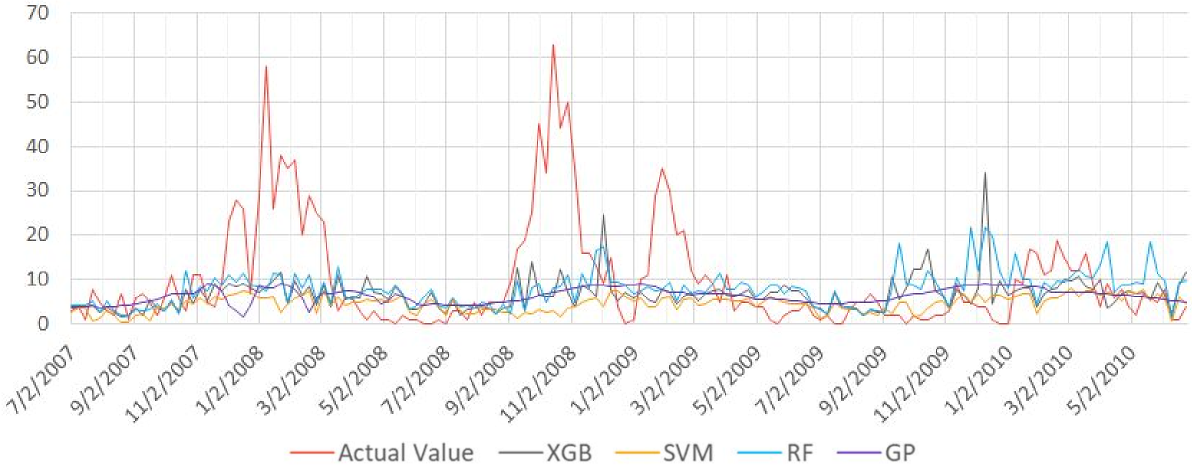


Figure 3. F Feature importances between Random Forest and XGBoost (with San Juan Training dataset)

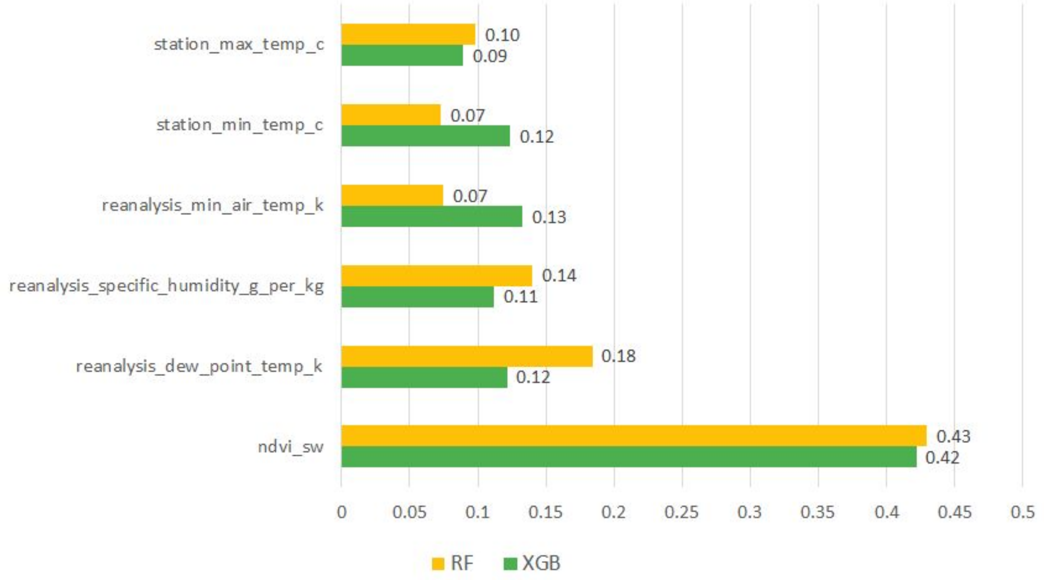


Figure 3.G ~ L : Results from Approach 2

Figure 3. G In-sample and out-of-sample score based on environmental factors & the number of cases in the past

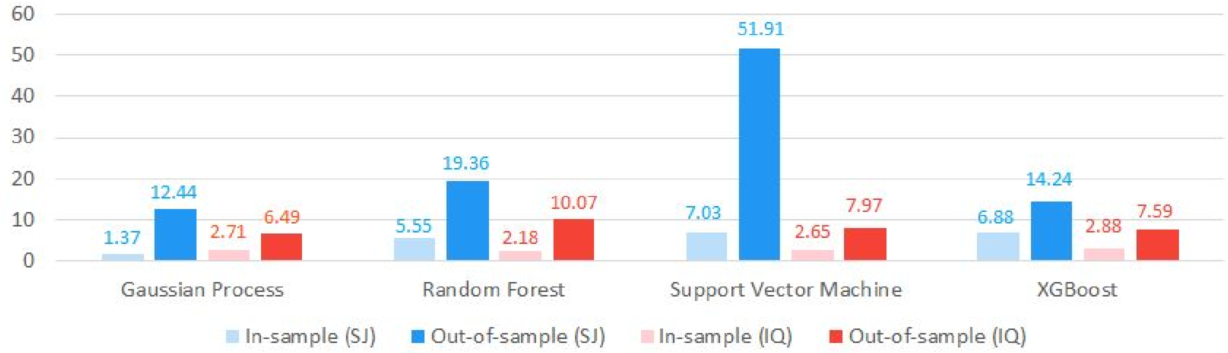


Figure 3. H Prediction for in-sample (San Juan) based on environmental factors & the number of cases in the past

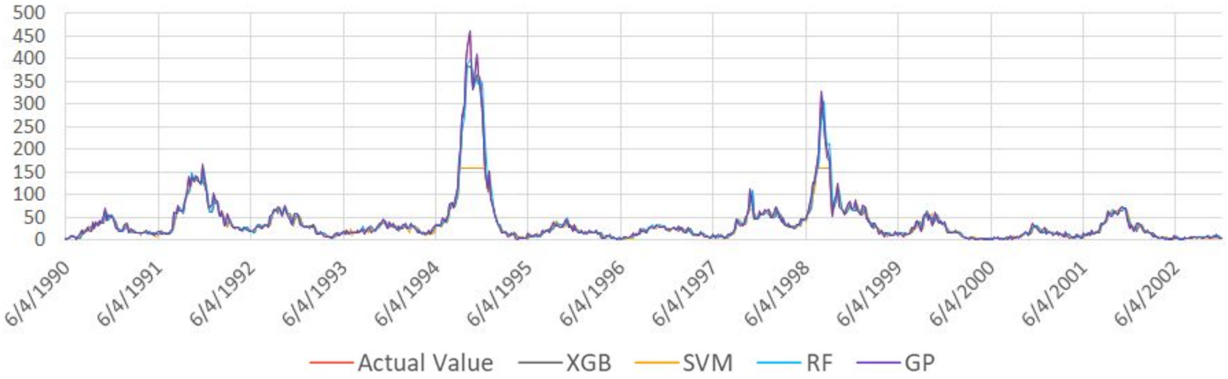


Figure 3. I Prediction for out-of-sample (San Juan) based on environmental factors & the number of cases in the past

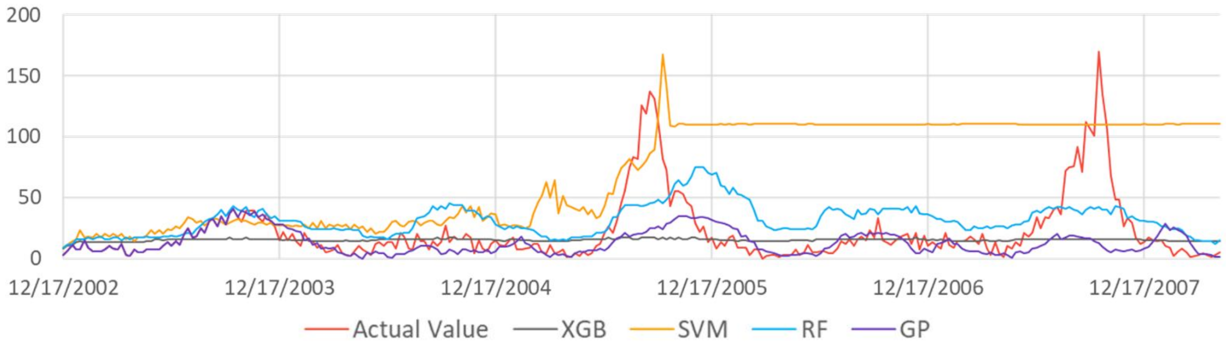


Figure 3. J Prediction for in-sample (Iquitos) based on environmental factors & the number of cases in the past

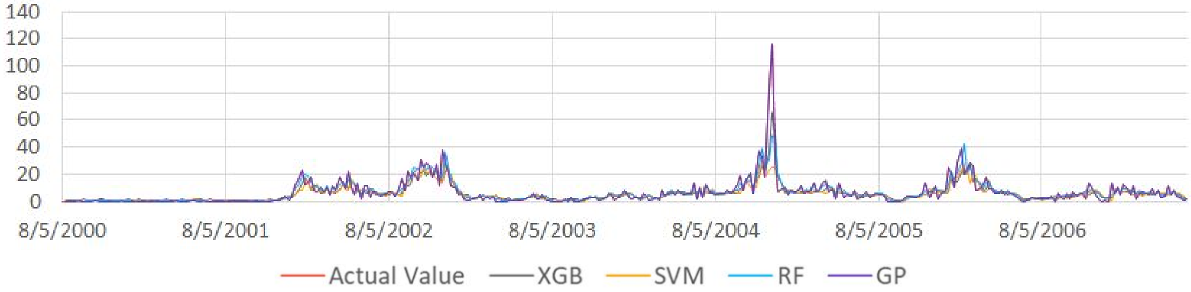


Figure 3. K Prediction for out-of-sample (Iquitos) based on environmental factors & the number of cases in the past

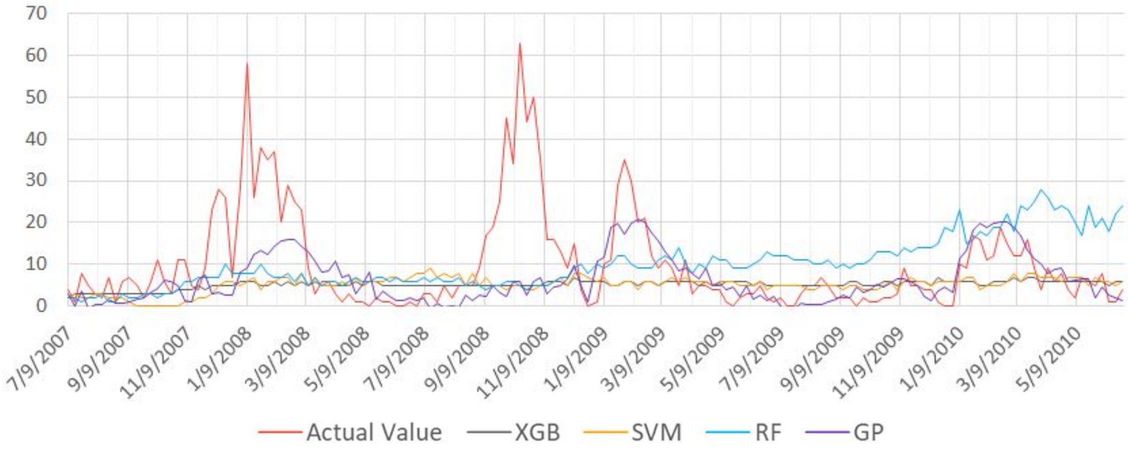


Figure 3. L Feature importances between Random Forest and XGBoost (with San Juan Training dataset)

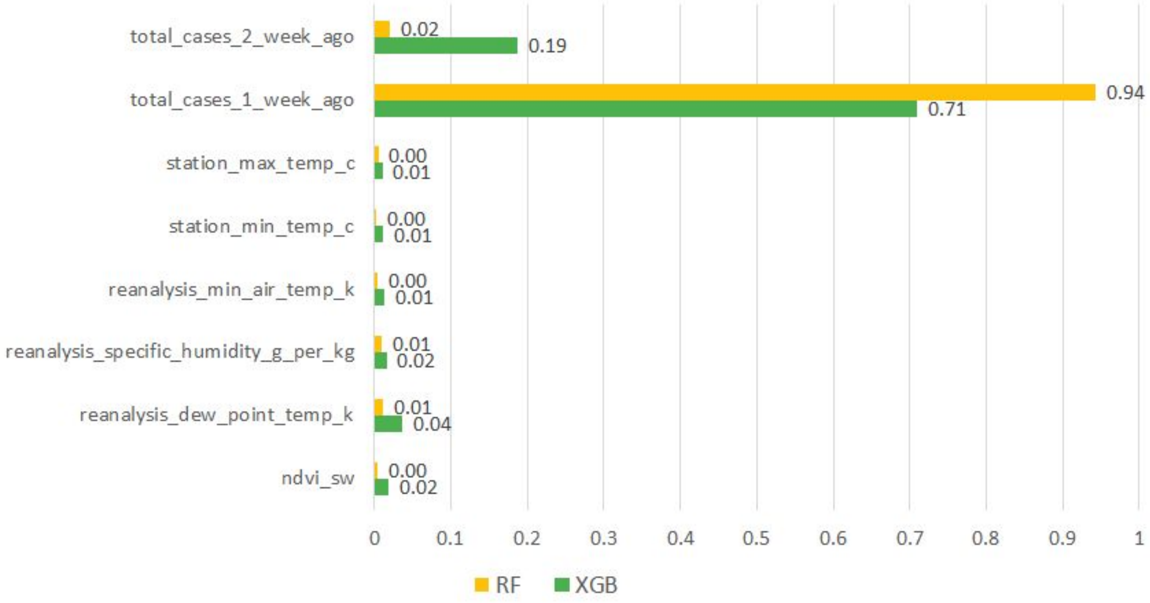


Figure 3.M ~ P : Results from Approach 3

Figure 3. M In-sample and out-of-sample score based on environmental factors and lags constructed from environmental factors based predicted cases

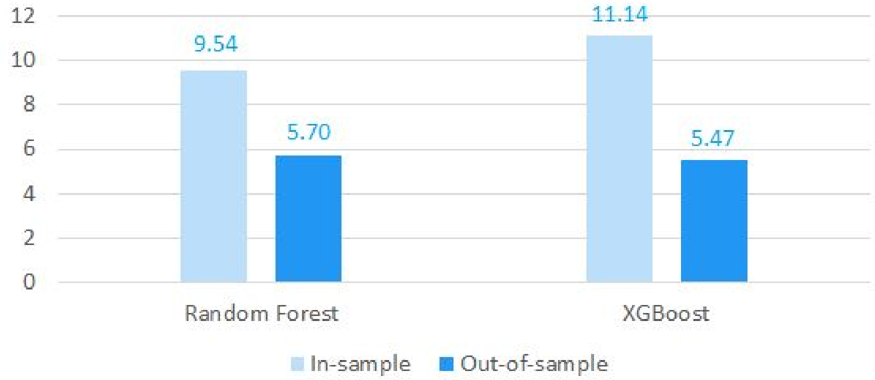


Figure 3. N. In-sample prediction

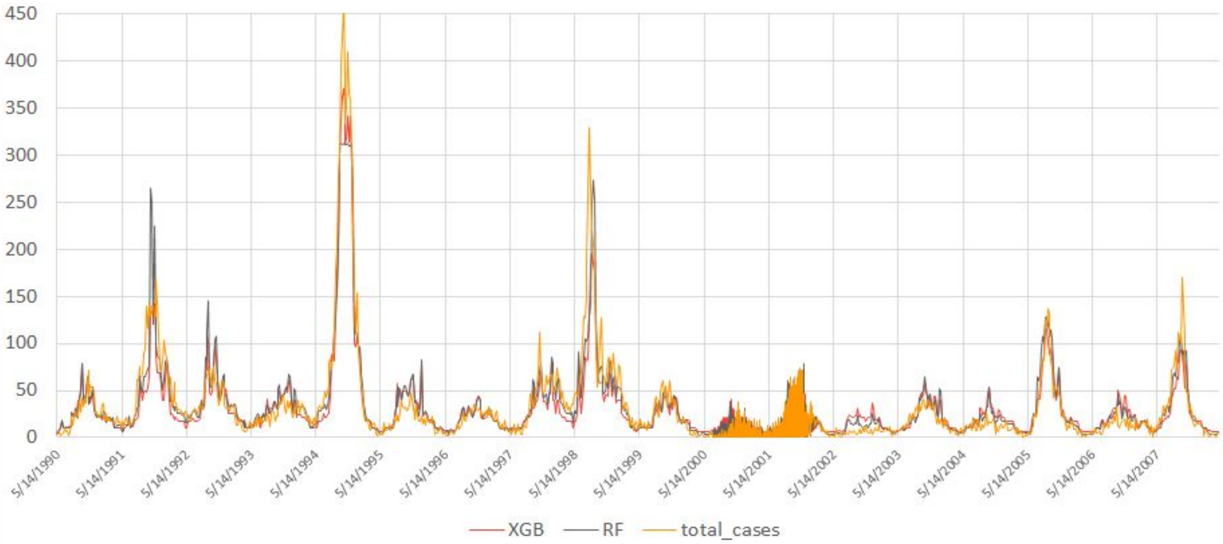


Figure 3. O. Out-of-sample prediction

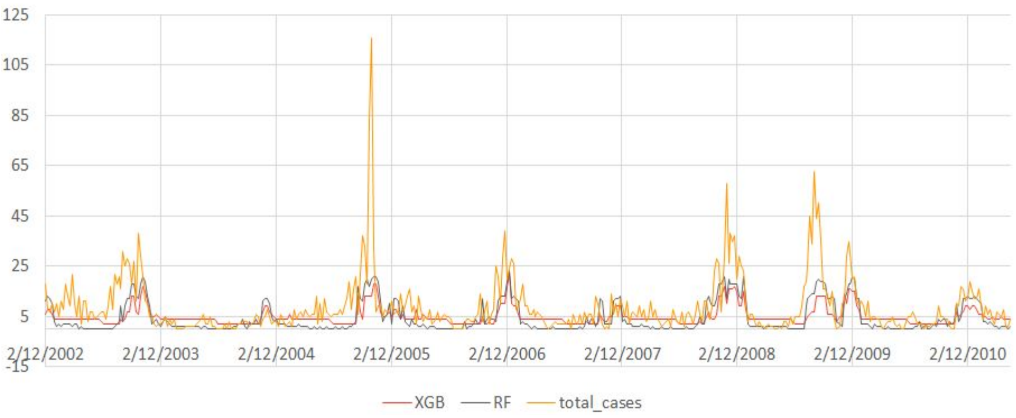
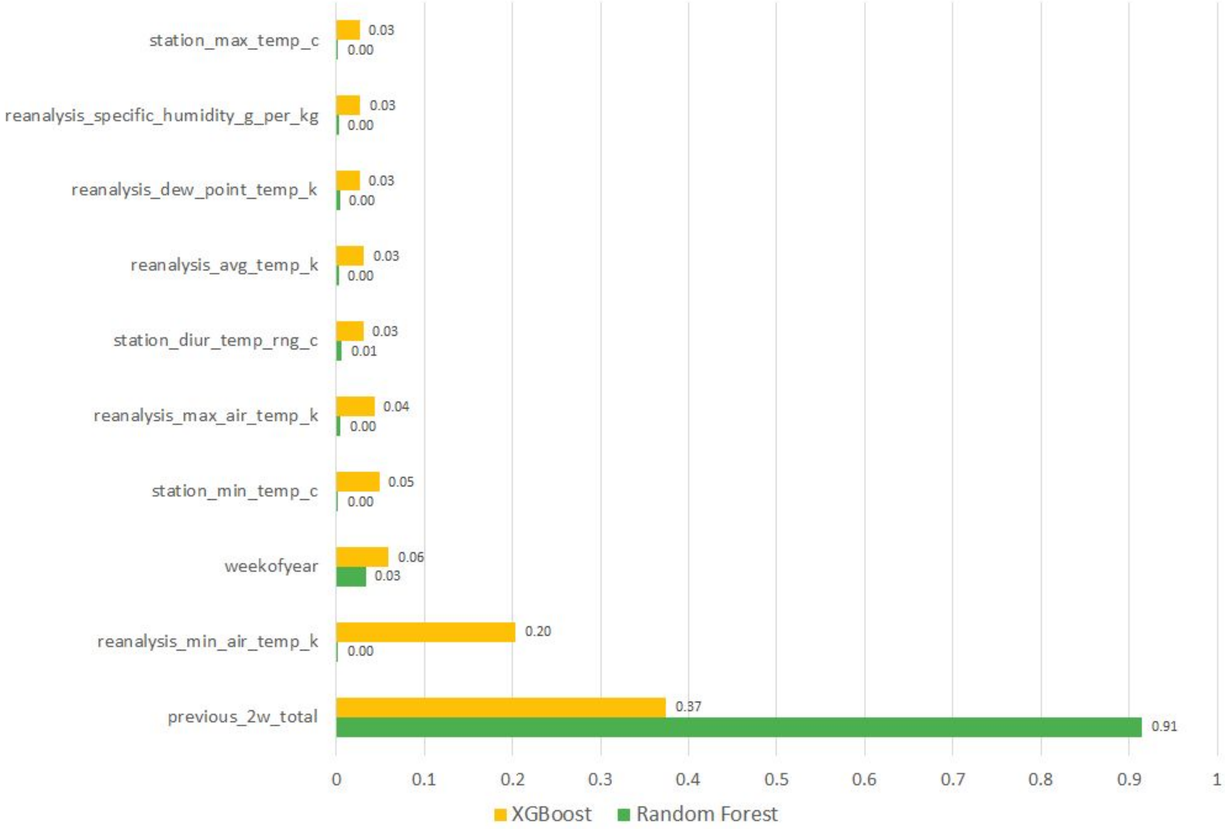


Figure 3. P Top 10 Feature importances between Random Forest and XGBoost (with Training dataset)


station_max_temp_c
reanalysis_specific_humidity_g_per_kg
reanalysis_dew_point_temp_k
reanalysis_avg_temp_k
station_diur_temp_rng_c
reanalysis_max_air_temp_k
station_min_temp_c
weekofyear
reanalysis_min_air_temp_k
previous_2w_total
XGBoost
Random Forest