Dan Cusick December 2022

I. Problem

The objective here is to build forecasting models that predict weekly dengue fever case totals in San Juan, Puerto Rico and Iquitos, Peru. The project originates from a 2015 "Predict the Next Pandemic" federal government initiative. Drivendata.org retains the challenge as a data-modeling exercise.

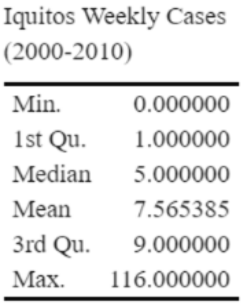
II. Significance

[Removed]

III. Data

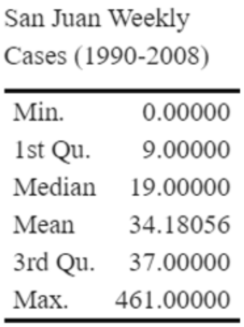
DrivenData's files include weekly dengue fever cases for San Juan, Puerto Rico from 1990 to 2008 and Iquitos, Peru from 2000 to 2010. The target forecast timeframe ranges from 2008 through 2013. Fever case data was collected by the U.S. Centers for Disease Control, Department of Defense, Peruvian government and U.S. Universities. Drivendata.org also provides environmental and climate data from the U.S. National Oceanic and Atmospheric Administration (NOAA).

Here is a summary of dengue fever cases from both cities:



Iquitos Weekly Cases (2000-2010)

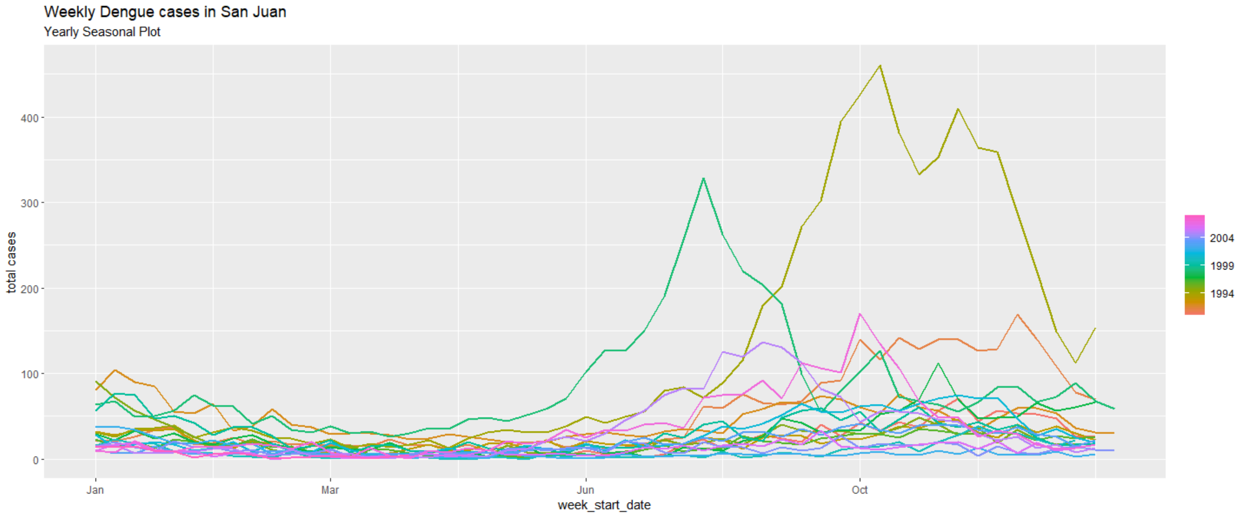
Min.	0.000000
1st Qu.	1.000000
Median	5.000000
Mean	7.565385
3rd Qu.	9.000000
Max.	116.000000

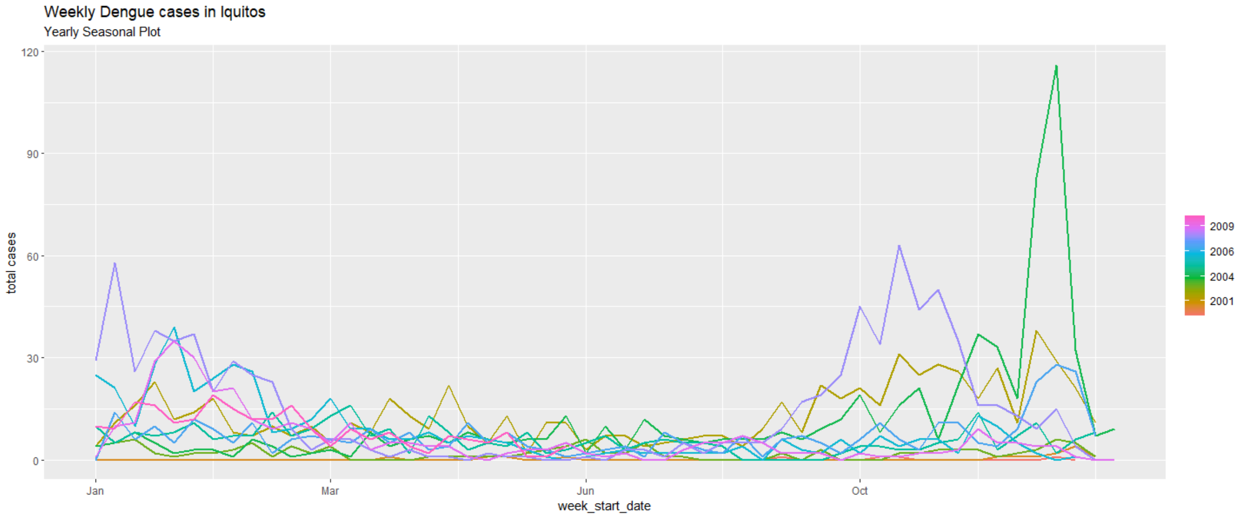


San Juan Weekly Cases (1990-2008)

Min.	0.00000
1st Qu.	9.00000
Median	19.00000
Mean	34.18056
3rd Qu.	37.00000
Max.	461.00000

Both cities generally exhibit a seasonal pattern with a spike towards the end of each year:





IV. Data Processing

The data was separated with a roughly 70%-30% split. The training set covered 1990 through 2004 and the test set covered 2005 through 2010. Next a correlation matrix of the training set climate variables was examined to narrow down the number of features included in the model. I manually reviewed each variable’s column and chose to eliminate many highly-correlated predictors. Then I ran a linear regression model for the remaining variables against total cases and reviewed the Variance Inflation Factor (VIF). Instead of showing individual pairwise relationships, the VIF measures multicollinearity by comparing the variance of each predictor’s linear fit to the variance of the full model. None of the climate variables had a VIF higher than 3.5, so I elected to keep them all moving forward. Here is the final list of climate predictions and their definition, per DrivenData:

Satellite vegetation, NOAA's CDR Normalized Difference Vegetation Index
ndvi_ne: Pixel northeast of city centroid
ndvi_se: Pixel southeast of city centroid

PERSIANN satellite precipitation
precipitation_amt_mm: total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements
reanalysis_precip_amt_kg_per_m2: Total precipitation
reanalysis_specific_humidity_g_per_kg: Mean specific humidity

NOAA's GHCN daily climate data weather station measurements
station_avg_temp_c: Average temperature
station_precip_mm: Total precipitation

V. Literature

[Removed]

VI. Models & Model Formulations

Three models were selected and submitted to drivendata.org for this project. Two of them include Fourier Terms, which mimic seasonal fluctuations using a series of sine and cosine functions.

Model 1: Linear regression of lagged precipitation, lagged vegetation, humidity and temperature with Auto AR errors
This model was built based on Wongkoon's findings that lagged rainfall was a successful factor in predicting cases. I used a 4 week lag for the previous month's precipitation. The model also performed slightly better with a lag of the vegetation index. An AR(2) framework was applied to the Iquitos errors and an AR(1) was applied to San Juan's errors. The submission for this model included missing total_cases forecasts, so I imputed the previous week's forecast to remove the NAs.

Model 2: Dynamic Harmonic Regression with climate predictors and Auto ARIMA errors
A dynamic harmonic regression outperformed an Auto ARIMA and Seasonal ARIMA. K=5 terms outperformed K=10 and K=2. The regression was extended here to also include climate factors. Iquitos errors were finetuned with an ARIMA(4,0,1) while San Juan's errors were modeled with an MA(2) technique. The submission for this model included missing total_cases forecasts, so I imputed the previous week's forecast to remove the NAs.

Model 3: Dynamic Harmonic Regression with Auto AR errors
Given an improved submission with dynamic harmonic regression in Model 2, I tried a straight-up Dynamic Harmonic Regression for the third trial. K=5 fourier terms were used and errors were modeled as AR(1) for Iquitos and AR(3) for San Juan.

VII. Performance

The following shows the test set MAE scores for these three models. The Test MAEs are the score on *my own* test set, which was a subset of DrivenData's full training set. I allowed the automatic ARIMA function to generate new parameters on the full data set. The DrivenData score is the MAE across both cities on the true test set. All MASE scores came out well below 1, so all three models cleared the benchmark set by the naive forecast.

	Test set MAE	MASE on full Driven Data Train set	Driven Data MAE score
Model 1: Lagged Regression with ARIMA errors	Iquitos : 11.1439093 San Juan: 16.6576091	Iquitos : 0.376 San Juan: 0.190	26.4111
Model 2: Dynamic Harmonic Regression with climate predictors	Iquitos : 9.0891673 San Juan: 28.1986353	Iquitos : 0.377 San Juan: 0.267	25.5505 1386 out of 4585 appx 69th percentile
Model 3: Dynamic Harmonic Regression with ARIMA errors	Iquitos : 18.3011511 San Juan: 19.2223004	Iquitos : 0.418 San Juan: 0.218	25.8846

The DrivenData.org submission of the Model 2 forecasts resulted in a score of 25.5505, which ranks around the 69th percentile of preexisting submissions (1386 out of 4585).

VIII. Limitations

The climate variables used in Models 1 and 2 contained missing values, which resulted in missing total_cases forecasts. Model 1 produced 94 missing cases forecasts, while Model 2 had 57 missing forecasts. Missing predictions were filled in with the previous week's forecast. Also, not every climate factor provided by DrivenData was put into use.

IX. Future Work

Missing forecast values were imputed with the previous week's forecasts using, essentially, a naive approach. Other avenues to avoid or impute missing data should be explored.

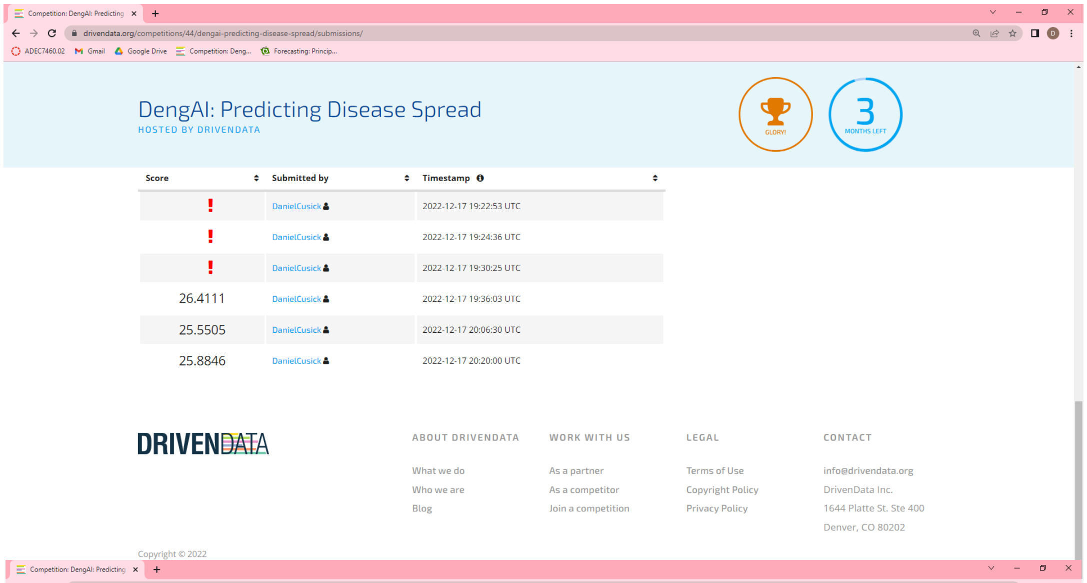
Filling in the climate variables themselves, instead of the forecasts, is one option. More climate variables could also be considered. While Roster et. al. had some success with reducing collinearity, that is not necessarily a universal principle. More information could very well produce more accurate predictions. Different lags of existing and new climate variables should also be assessed. A greater review of scientific literature could also be conducted to make more informed decisions about what factors to include and how much to lag them. While the models used were justified, more ARIMA and Fournier term parameters could be vetted to increase forecast accuracy. Ensemble models with a combination of approaches could also be reviewed in future work.

X. Learning

While much of the literature regarding dengue fever focused on ARIMA methods, linear regression, and advanced machine learning, this work shows Dynamic Harmonic Regression should also be included in the public health toolkit.

Appendix

Driven Data submission


Competition: DengAI: Predicting
drivendata.org/competitions/44/dengai-predicting-disease-spread/submissions/
ADEC7460.02
Gmail
Google Drive
Competition: Deng...
Forecasting: Princip...
GLORY!
3
MONTHS LEFT
DRIVENDATA

DengAI: Predicting Disease Spread
HOSTED BY DRIVENDATA

Score	Submitted by	Timestamp
!	DanielCusick	2022-12-17 19:22:53 UTC
!	DanielCusick	2022-12-17 19:24:36 UTC
!	DanielCusick	2022-12-17 19:30:25 UTC
26.4111	DanielCusick	2022-12-17 19:36:03 UTC
25.5505	DanielCusick	2022-12-17 20:06:30 UTC
25.8846	DanielCusick	2022-12-17 20:20:00 UTC

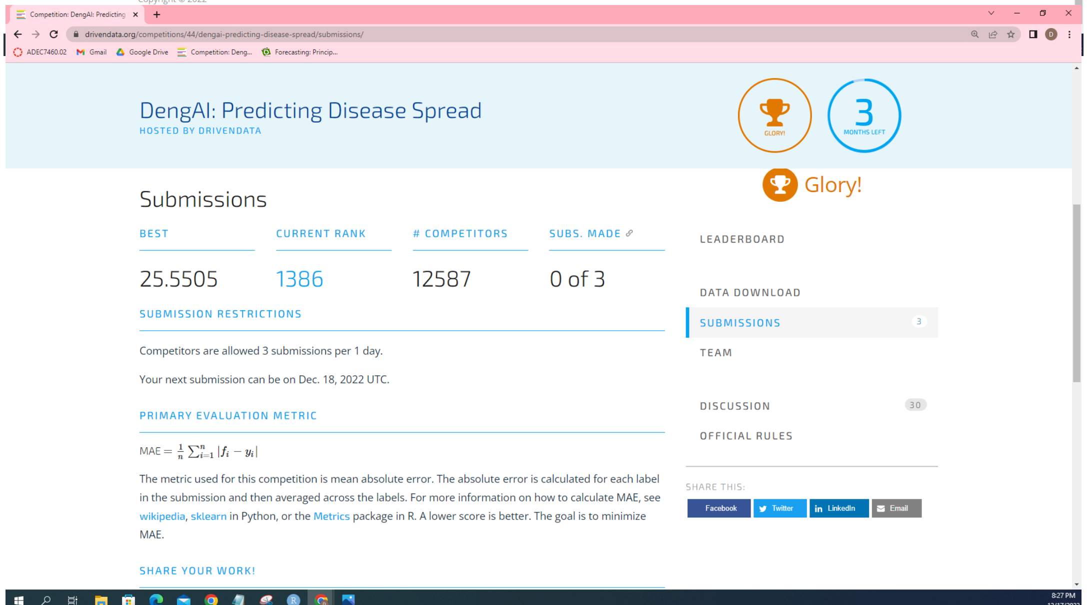
ABOUT DRIVENDATA
What we do
Who we are
Blog

WORK WITH US
As a partner
As a competitor
Join a competition

LEGAL
Terms of Use
Copyright Policy
Privacy Policy

CONTACT
info@drivendata.org
DrivenData Inc.
1644 Platte St. Ste 400
Denver, CO 80202

Copyright © 2022



DengAI: Predicting Disease Spread
HOSTED BY DRIVENDATA

Submissions

BEST	CURRENT RANK	# COMPETITORS	SUBS. MADE
25.5505	1386	12587	0 of 3

SUBMISSION RESTRICTIONS

Competitors are allowed 3 submissions per 1 day.

Your next submission can be on Dec. 18, 2022 UTC.

PRIMARY EVALUATION METRIC

$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i|$

The metric used for this competition is mean absolute error. The absolute error is calculated for each label in the submission and then averaged across the labels. For more information on how to calculate MAE, see wikipedia, sklearn in Python, or the Metrics package in R. A lower score is better. The goal is to minimize MAE.

SHARE YOUR WORK!