Predicting Dengue Fever

Technical Presentation

Alex Freeman

September 26, 2017

Email: alexjf12@gmail.com

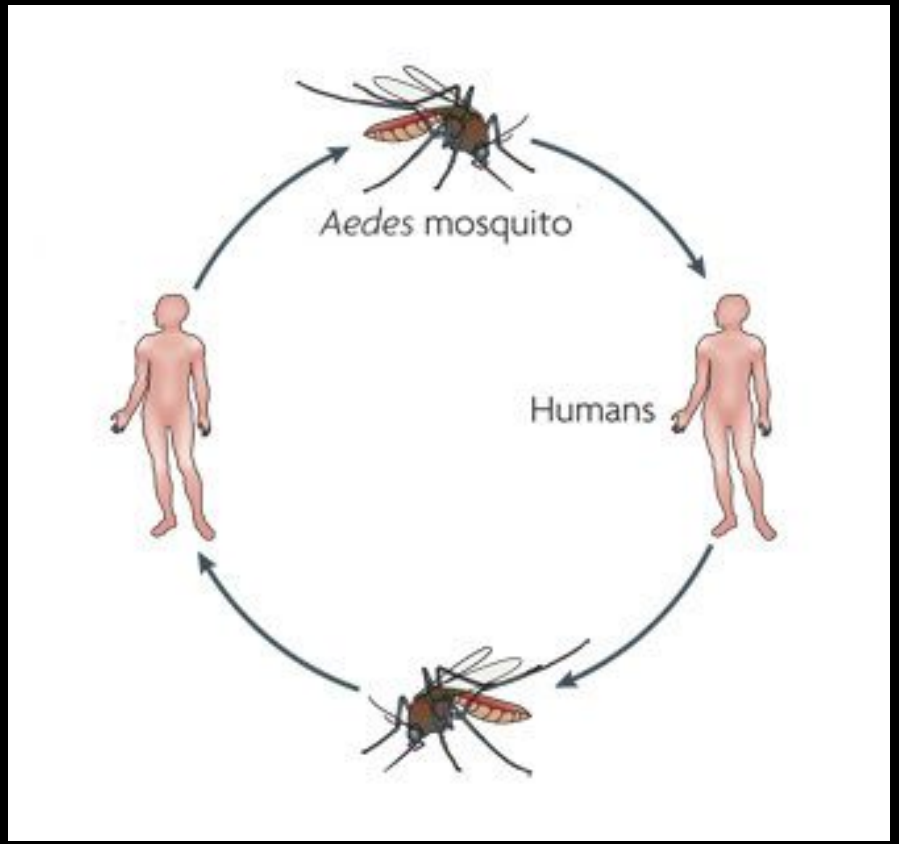
Can weather predict the weekly number of dengue fever cases?

The Competition
Hosted by Driven Data

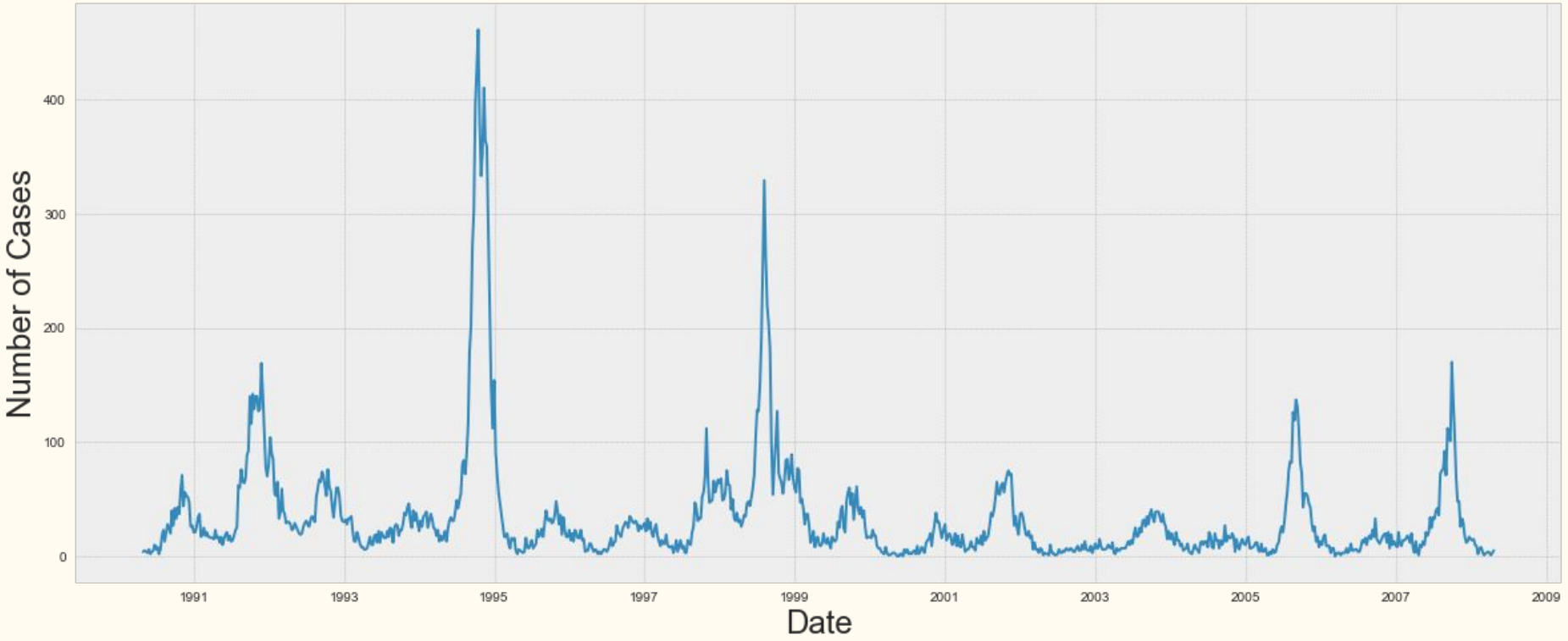
- Goal:
  - Predict number of cases in San Juan, Puerto Rico and Iquitos, Peru
- Scoring Metric:
  - Mean Absolute Error
- Duration:
  - Ends Dec. 31, 2017
- Prize:
  - Glory!

DRIVENDATA

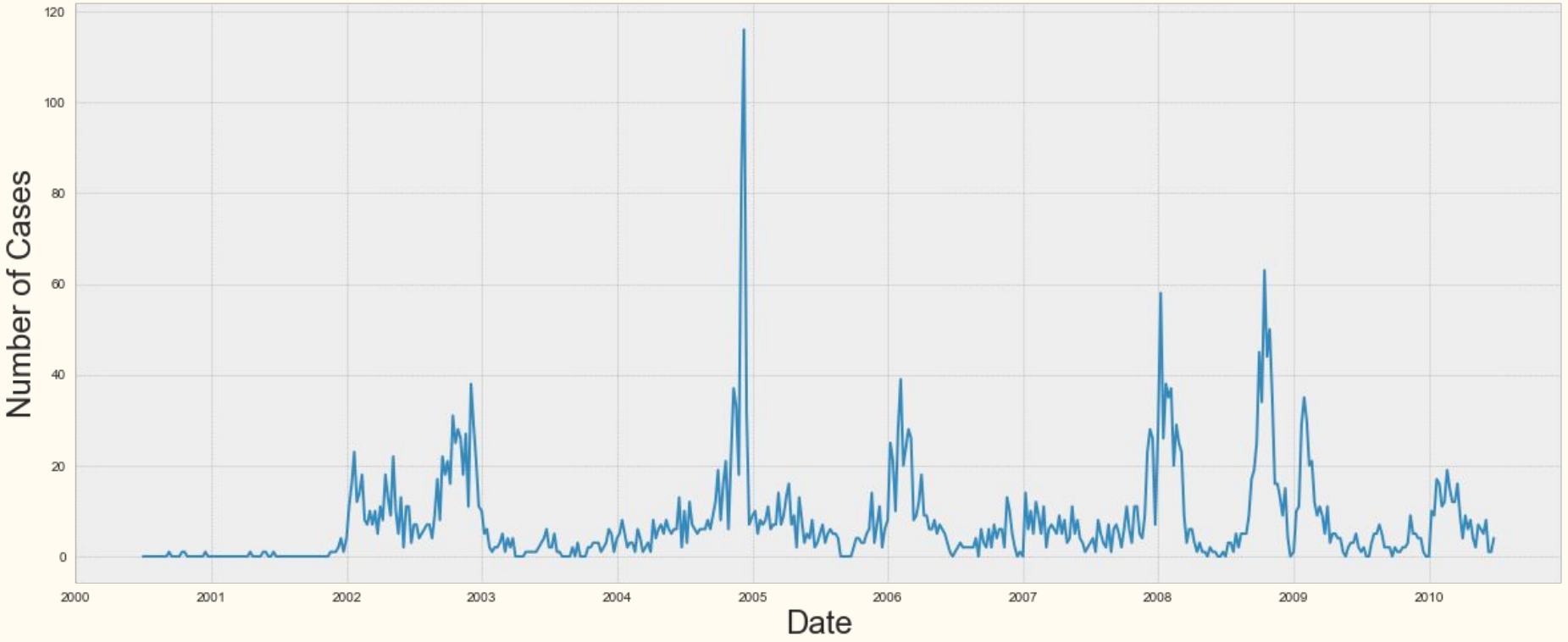
Dengue: A Primer



Number of Cases per week in San Juan, Puerto Rico



Number of Cases per week in Iquitos, Peru



Weather Features

- Temperature
  - Max, min, average, diurnal range, dew point
- Precipitation
  - Total rainfall
- Humidity
  - Mean relative and mean specific
- Vegetation
  - Level of vegetation in NW, NE, SW and SE quadrants of city as measured by satellite image

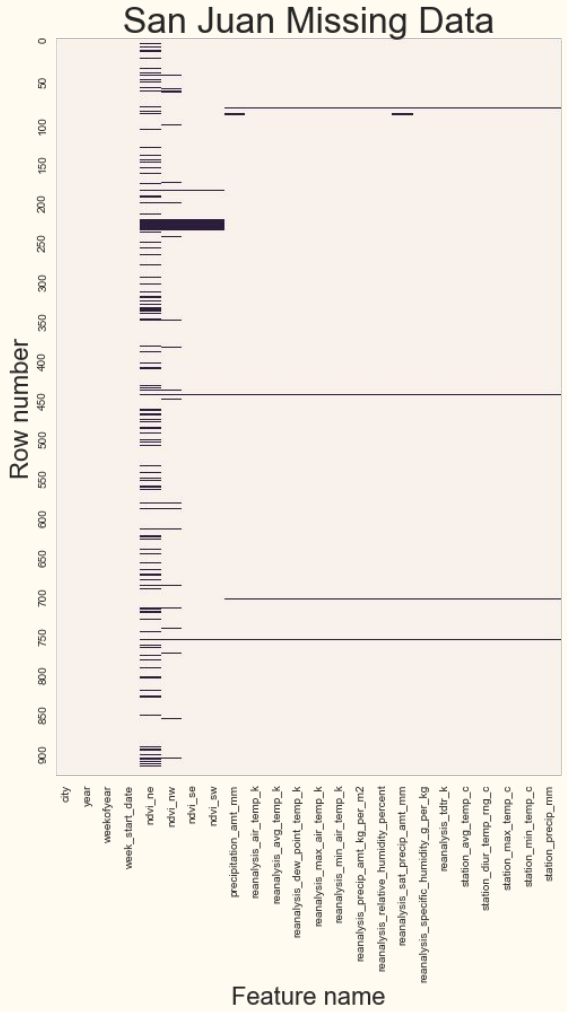
Front Fill Missing Data

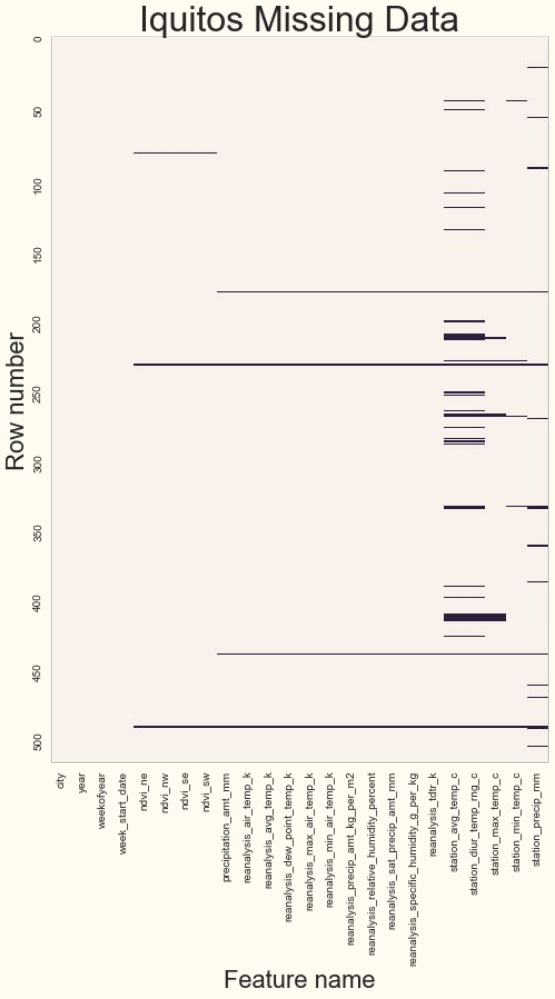
San Juan

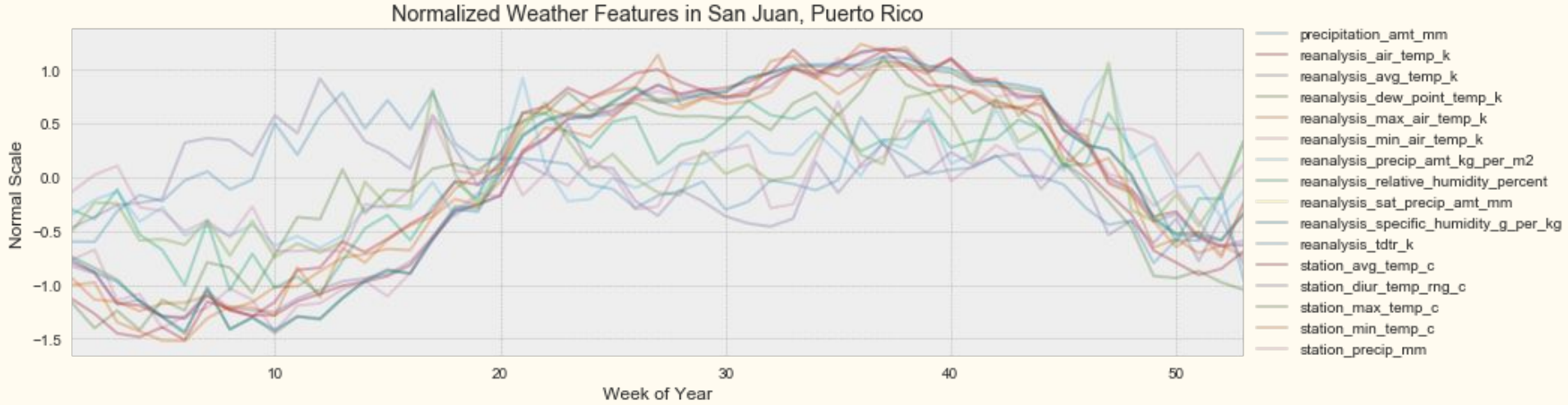
- ndvi_ne is often missing through the entirety of the dataset
- Data for NDVI is missing for October -December, 1994

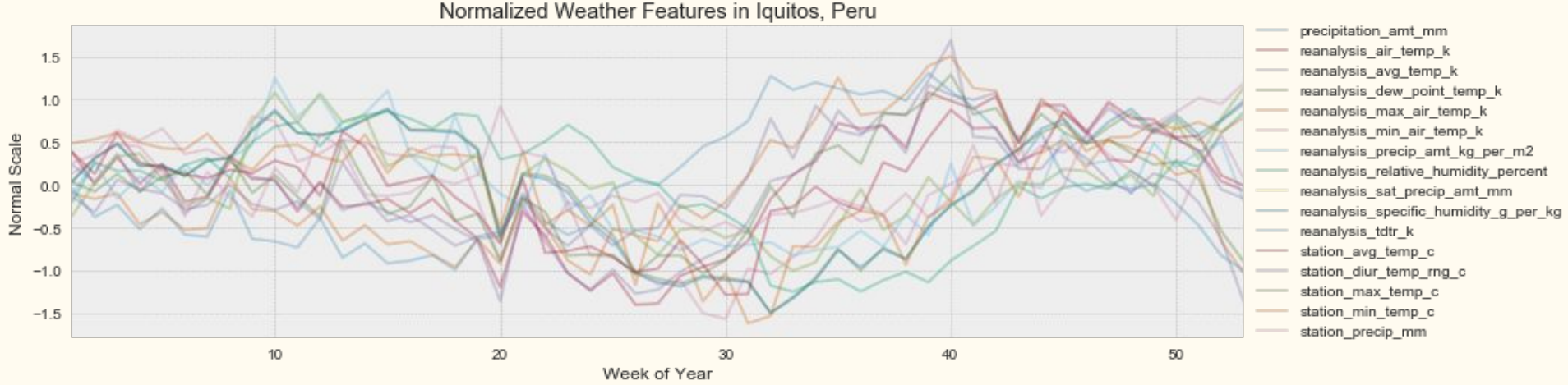
Iquitos

- station_avg_temp_c is missing the most

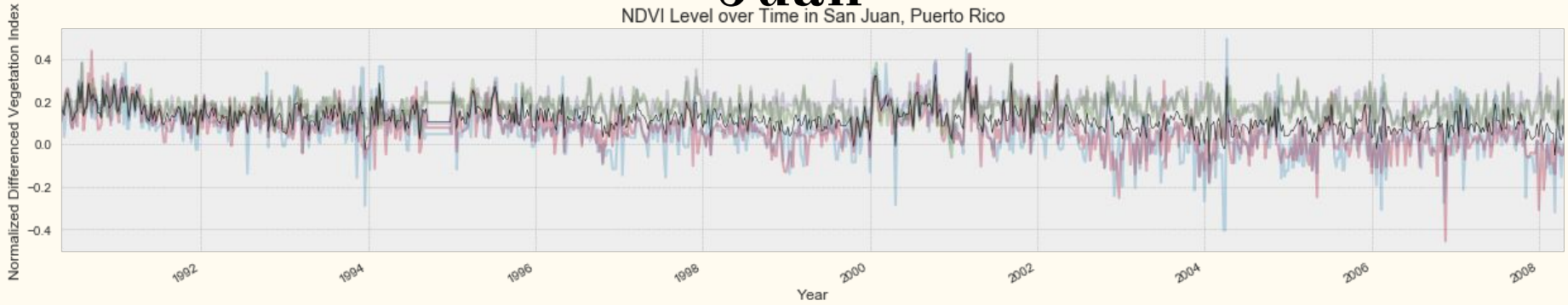


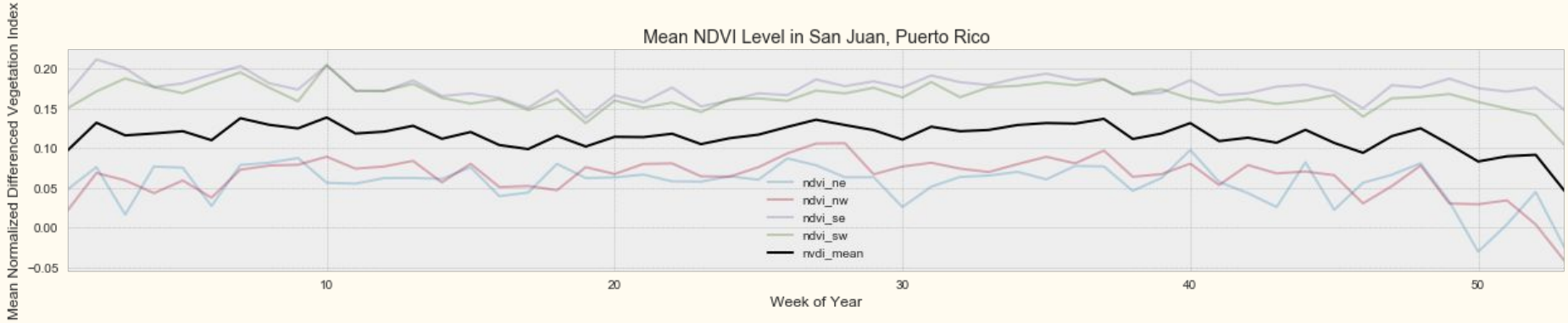



Normalized Weather Features in San Juan, Puerto Rico
Normal Scale
1.0
0.5
0.0
−0.5
−1.0
−1.5
10
20
30
40
50
Week of Year
precipitation_amt_mm
reanalysis_air_temp_k
reanalysis_avg_temp_k
reanalysis_dew_point_temp_k
reanalysis_max_air_temp_k
reanalysis_min_air_temp_k
reanalysis_precip_amt_kg_per_m2
reanalysis_relative_humidity_percent
reanalysis_sat_precip_amt_mm
reanalysis_specific_humidity_g_per_kg
reanalysis_tdtr_k
station_avg_temp_c
station_diur_temp_rng_c
station_max_temp_c
station_min_temp_c
station_precip_mm

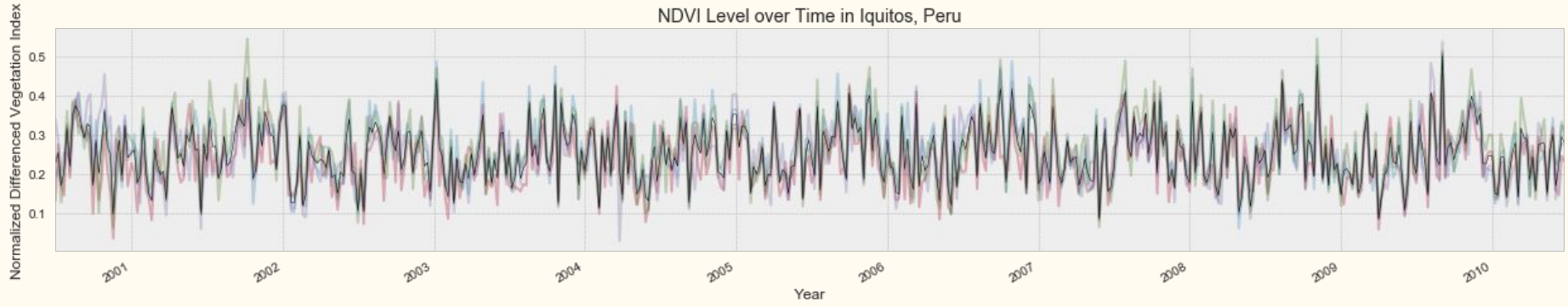

Normalized Weather Features in Iquitos, Peru
Normal Scale
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
10
20
30
40
50
Week of Year
precipitation_amt_mm
reanalysis_air_temp_k
reanalysis_avg_temp_k
reanalysis_dew_point_temp_k
reanalysis_max_air_temp_k
reanalysis_min_air_temp_k
reanalysis_precip_amt_kg_per_m2
reanalysis_relative_humidity_percent
reanalysis_sat_precip_amt_mm
reanalysis_specific_humidity_g_per_kg
reanalysis_tdtr_k
station_avg_temp_c
station_diur_temp_rng_c
station_max_temp_c
station_min_temp_c
station_precip_mm

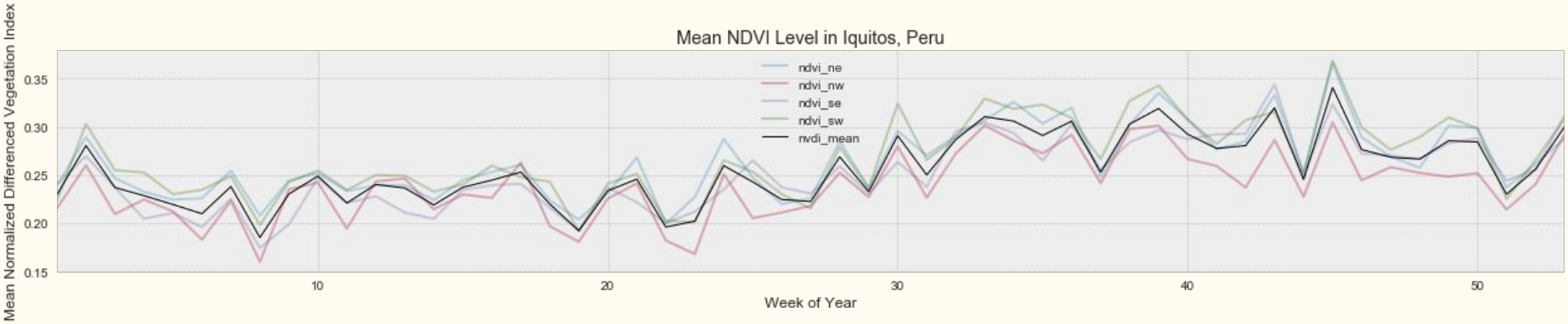
Normalized Difference Vegetation Index - San Juan





Normalized Difference Vegetation Index - Iquitos

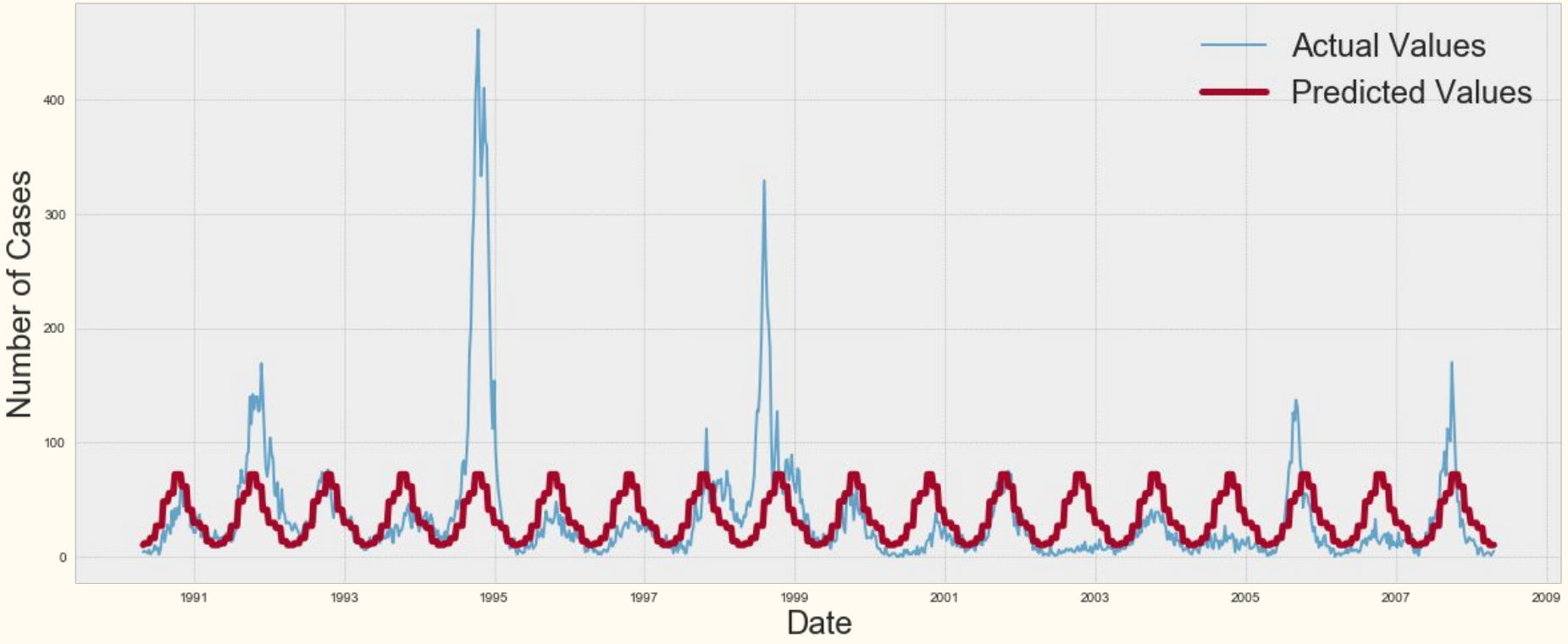




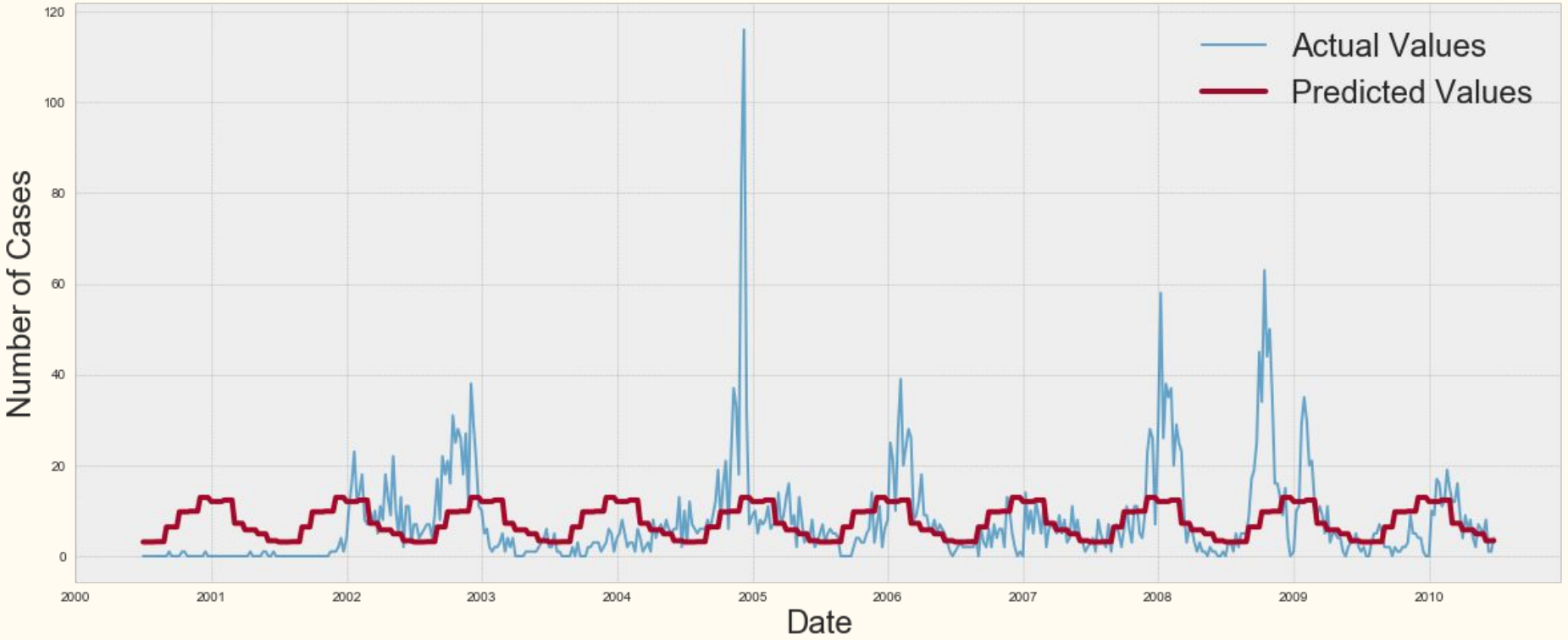
Modelling Methods

1. Find monthly trend
2. Use weather variables to predict residuals of the monthly trend
3. Combine monthly trend and weather predictions for total cases.

Monthly Trend of Dengue Fever in San Juan



Monthly Trend of Dengue Fever in Iquitos



Using the Monthly Trend to find the Residuals

Monthly Trend

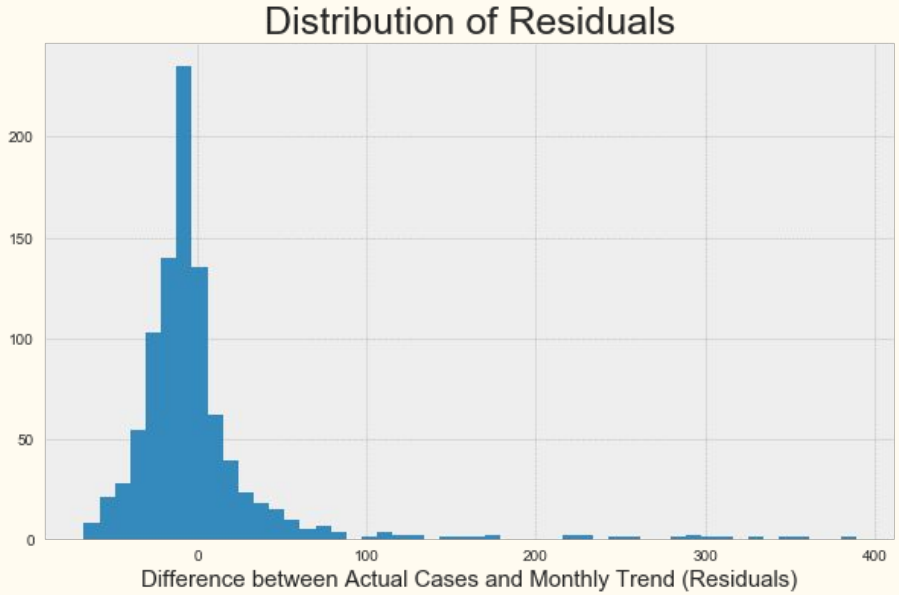
1. Create a dummy variable of the month of each observation
2. Fit a linear regression model on the month dummies to predict caseload

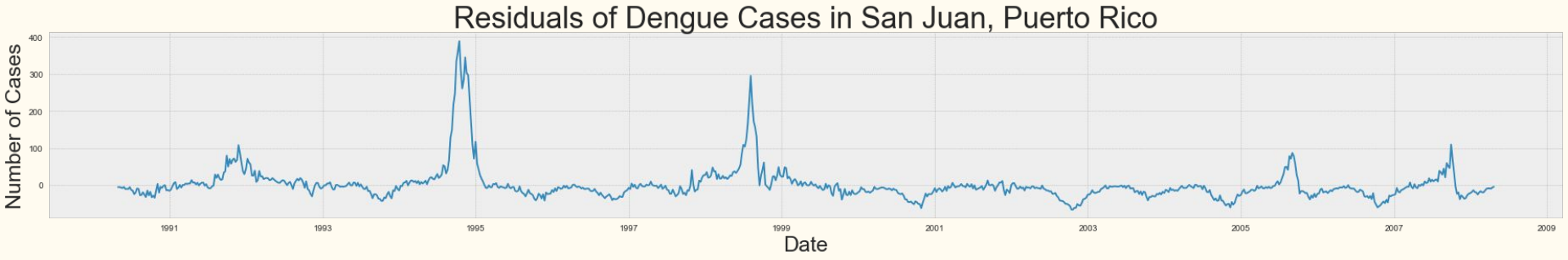
Residuals

1. ‘Detrend’ the cases by subtracting the monthly trend from the actual cases per week
2. Find a way to predict these residuals with the weather features

The Residuals

- Skewed to the right when there are outbreaks
- Fairly stable around 0 in normal periods
- Can weather predict an outbreak?





This week's weather doesn't matter...
... it's the weeks and months before

Mosquitos need the right weather to reproduce. Abundant vegetation can provide pools of stillwater for spawn points.

Over what time period should we study the weather to predict a single week's worth of cases?

- A month (4 weeks)?
- A quarter (16 weeks)?
- A year (52 weeks)?

Feature Engineering the Weather

- Rolling Mean
- Rolling Standard Deviation
- Exponentially Weighted Mean
- Shifted Values

Feature Engineering and Selection

Engineered Statistics:

- Rolling mean
- Exponentially weighted mean
- Rolling Standard deviation
- Shifted dates

Total Engineered Features:

- 21 Original Features
- 4 Statistics
- 99 Lengths to ‘look back’

Feature Selection:

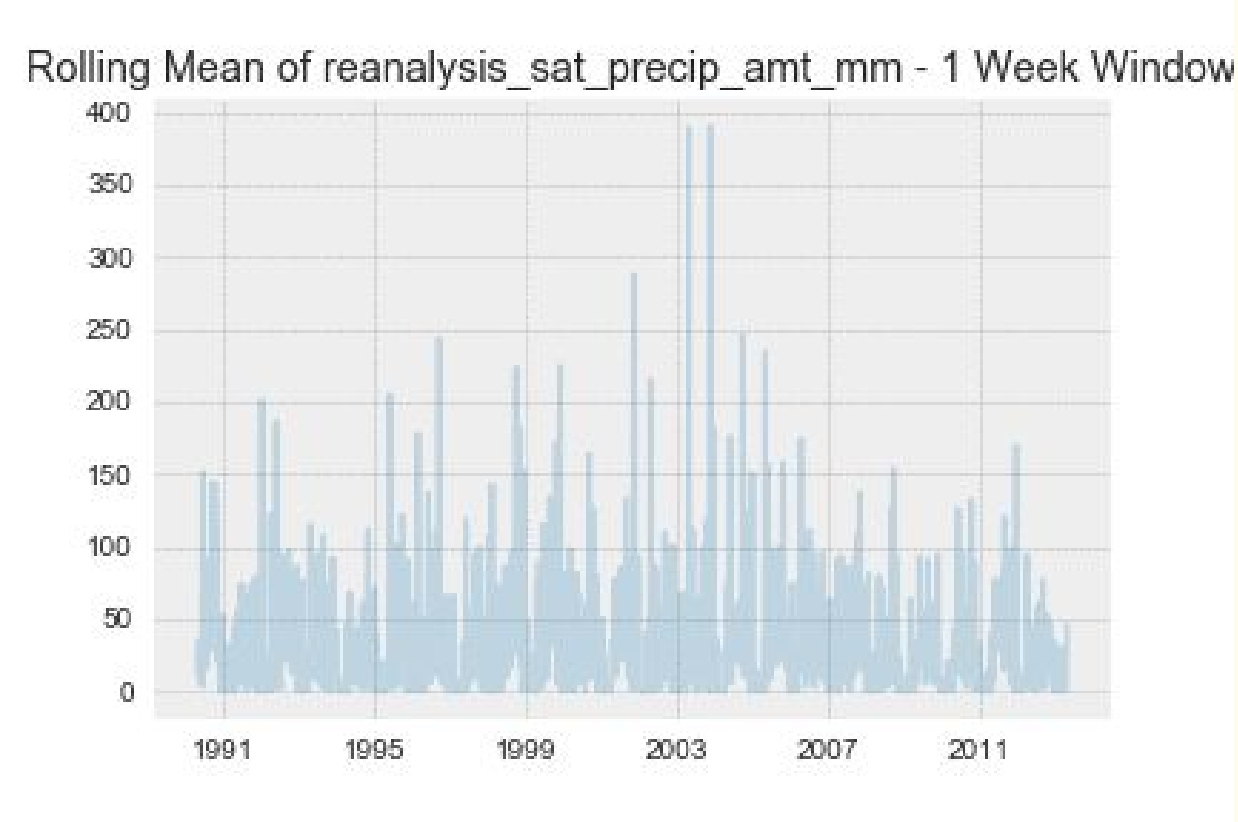
- Find correlation between residual and engineered feature looking back 1 week to 100 weeks

21 x 4 x 99 = 8,316 Possible Features

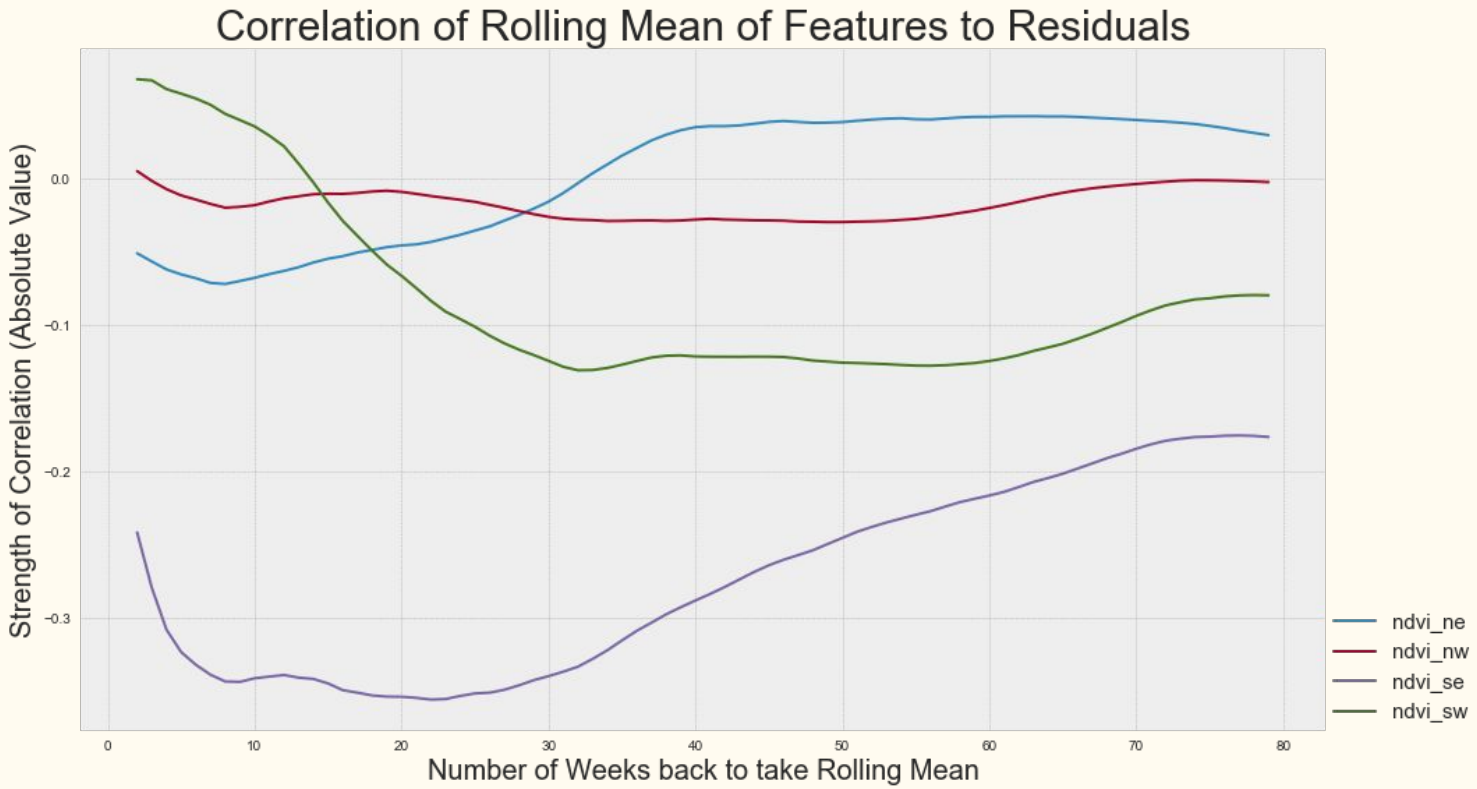
Visualizing a Rolling Mean

The larger the week window gets, the more historical information is used in determining the mean.

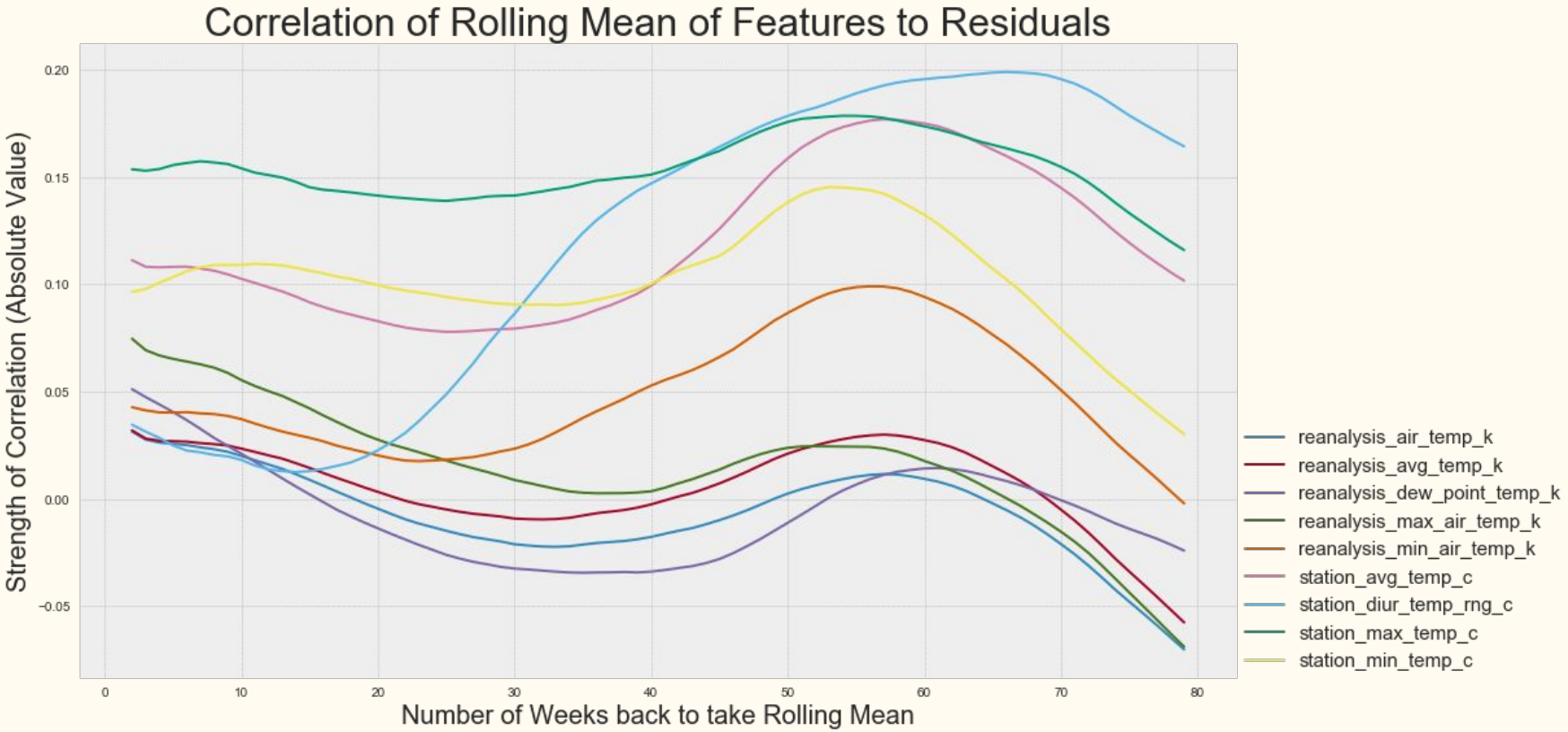
A rolling mean with a week window of 52 means the average of the previous year's data.



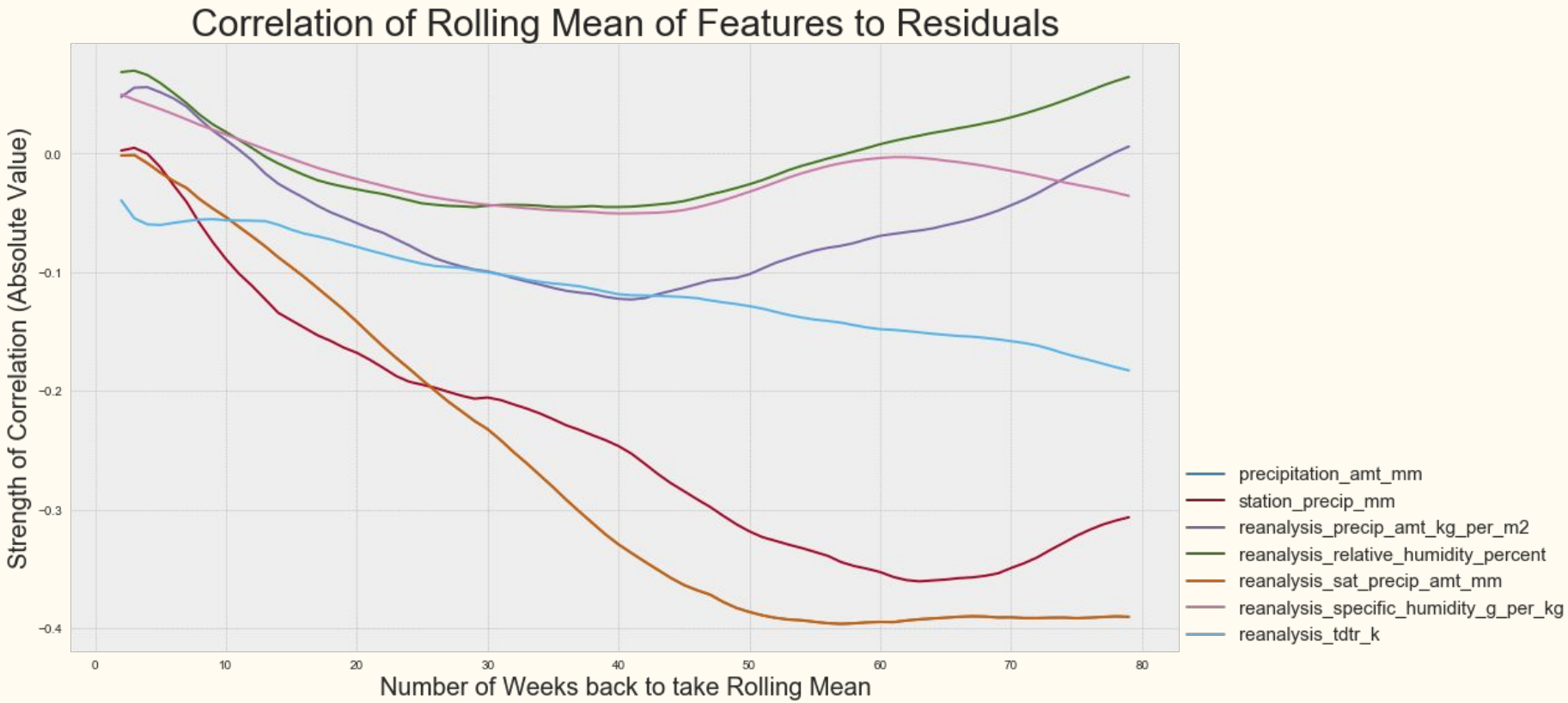
Rolling Mean Correlations - NDVI Index



Rolling Mean Correlations - Temperature



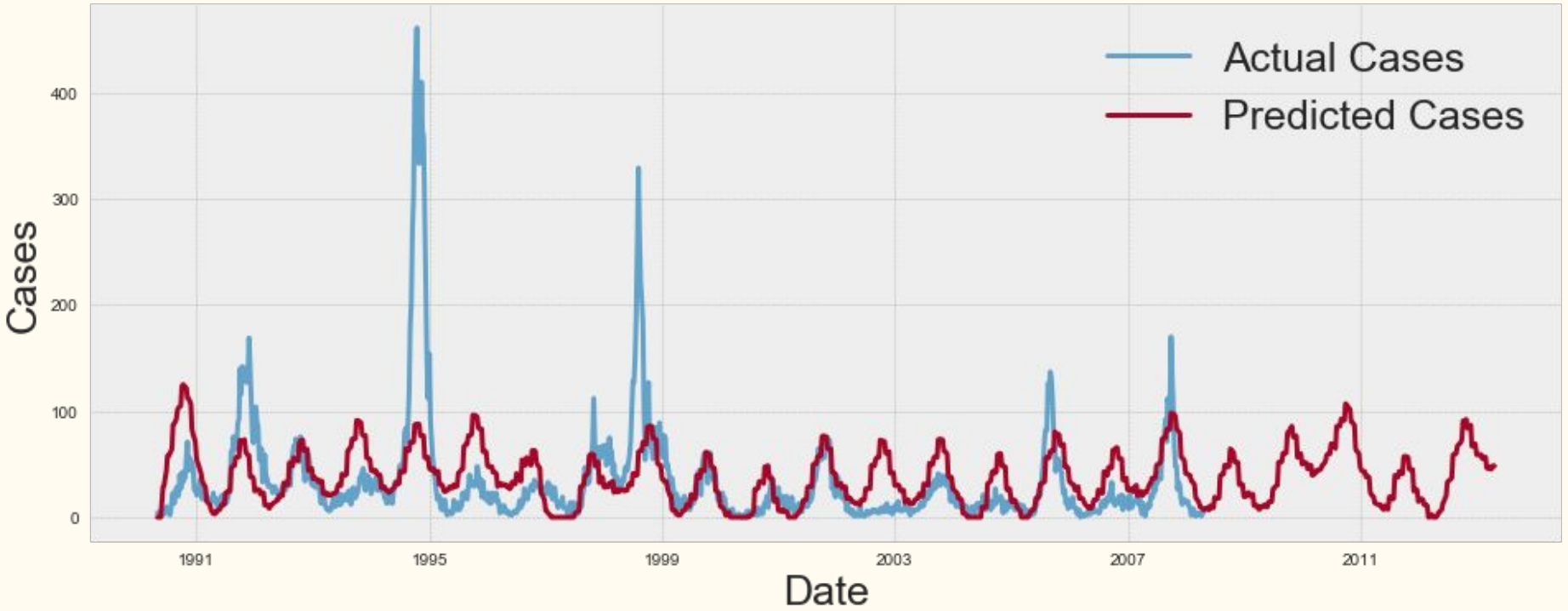
Rolling Mean Correlations - Rain and Humidity



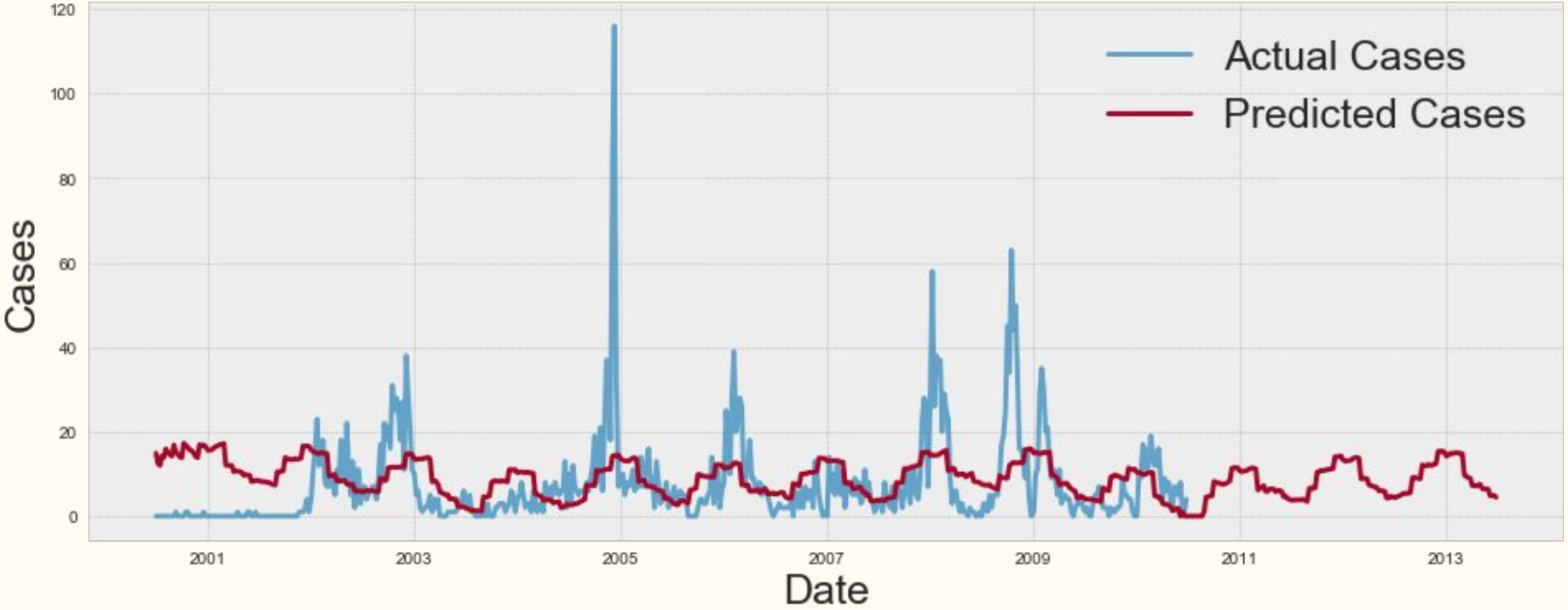
Cases ~ Monthly Trend + Rolling Mean of Temperature 55 Weeks back

Sometimes, a simple model is the best model

Best Predictions of Dengue Fever in San Juan



Best Predictions of Dengue Fever in Iquitos



RESULTS

Mean Absolute Error between my predicted values and the true values only known to DrivenData.

Model Description	Mean Absolute Error
Monthly Trend	26.5144
XGBoost with current weather and rolling mean 52 weeks back	27.3774
Monthly Trend + Rolling mean of Temp. and rainfall (52 weeks back) and rolling std of Temp. and rainfall (8 weeks back)	24.1274
Monthly Trend + Custom rolling means, stds, and shifts for San Juan and Iquitos	22.9351
Monthly Trend + Rolling temp of 52 and NDVI_SE of 10 for San Juan and temp of 52 for Iquitos	21.3317
Monthly Trend + Rolling mean of Temp. (52 weeks back) and rolling std of Temp. (8 weeks back)	20.7981
Monthly Trend + Rolling Mean of Temperature (55 weeks back) for both	20.7764

My best performing model is 61st of 1922 submissions

Problems

- Overfitting the data. Received better validation scores, but worse test scores after Submitting to DrivenData.
- Model cannot not predict outbreaks, just increased caseloads.
- San Juan and Iquitos need unique models to reflect unique weather predictors.

Further Information

Code and analysis can be found in my GitHub repo here:

- https://github.com/AlexJF12/predicting-dengue

Non-technical Presentation can be found here:

- https://github.com/AlexJF12/predicting-dengue/blob/master/Predicting%20Dengue%20-%20Non-Technical%20Presentation.pdf

Competition

- https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread