PREDICTING DENGUE FEVER INCIDENCE IN LATIN AMERICA

The Aedes mosquito, the main vector that transmits the viruses that cause Dengue Fever

GENERAL ASSEMBLY

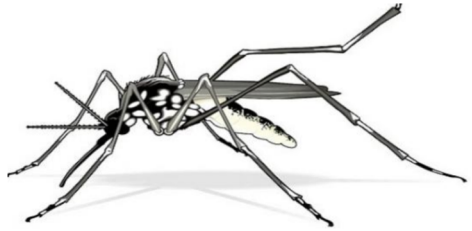
Data Science Immersive
Capstone Project

Matthew Perkins

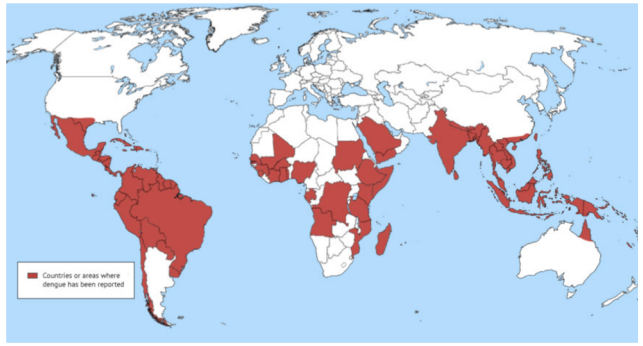
June 2018

PROJECT BACKGROUND

DENGUE FEVER EPIDEMIOLOGY



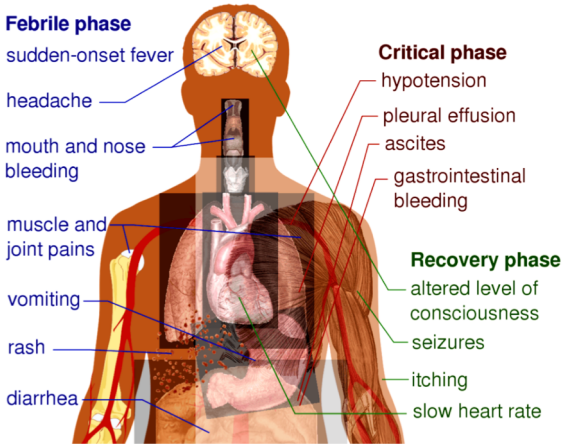
Caused by any one of four dengue viruses transmitted by **Aedes** mosquitos and infected humans (indirectly).



2.5 billion people in 128 countries are at risk of infection.

400 million people infected
500,000 hospitalisations
25,000 deaths annually

Symptoms Severe flu-like. 5% progress to life- threatening complications and/or death.



No vaccine
Most effective prevention is to avoid mosquito bites. Infection doesn't give immunity.

Endemic in 100 countries, vs. 9 in 1970. 30-fold increase between 1960 & 2010.

FOR 2 LOCATIONS IN LATIN AMERICA, THIS PROJECT AIMS TO…

- Understand relationships between weekly dengue case counts and environmental predictors.
- Predict weekly dengue case counts over a multi-year timeframe.

Answer the question…

"Can weekly dengue case counts be accurately predicted 4 weeks in advance?"

AN ACCURATE 4 WEEK PREDICTION ENABLES…

- **Timely interventions**: mosquito population control, distribution of nets, communications to raise awareness and promote individual and household protection.
- **Heightened preparedness** of health agencies.
- More generally, **enhanced understanding** of how global dengue footprint may evolve.

LOCATIONS AND DATA

SAN JUAN, PUERTO RICO

SAN JUAN, PUERTO RICO

- Puerto Rico's capital and most populous city: 2.5m people.
- Located on north-eastern coast. 38% of area is water.
- Tropical monsoon climate: hotter, wetter months from Jun-Nov; cooler drier months from Jan-March. Average annual max temp: 27°C.
- Endemic dengue activity since 1963.

IQUITOS, PERU

- World's largest city unreachable by road: 500,000 people.
- Surrounded on 3 sides by rivers, including the Amazon.
- Equatorial climate, with no distinct dry season. Rainfall throughout year, wetter from Nov-May. Average annual max temp: 27°C.
- Dengue returned in late 80s. City-wide mosquito control programs.

THE DATA

- Sourced from National Oceanographic and Atmospheric Administration
- Outcome CSV files consisting of weekly dengue case counts:
  - San Juan: April, 1990 - April, 2009 (19 years)
  - Iquitos: July, 2000 - June, 2009 (9 years)
- Predictor CSV and excel files consisting of:
  - Daily weather station measurements, dating back to 1973
  - Daily satellite precipitation data, dating back to 1983
  - Daily climate model data, dating back to 1973
  - Weekly satellite vegetation density measurements from 4 centroids around each city, dating back to 1981

IQUITOS, PERU

DATA MUNGING

Acquire competition data — Sourced, imported and reviewed aggregated weekly data from **active competition site**

Parse competition data — Reconciled competition data with raw source data and identified range of **inconsistent aggregation and date alignment issues**

Acquire raw source data — Decided to use raw daily source data provided by the **National Oceanographic and Atmospheric Administration**

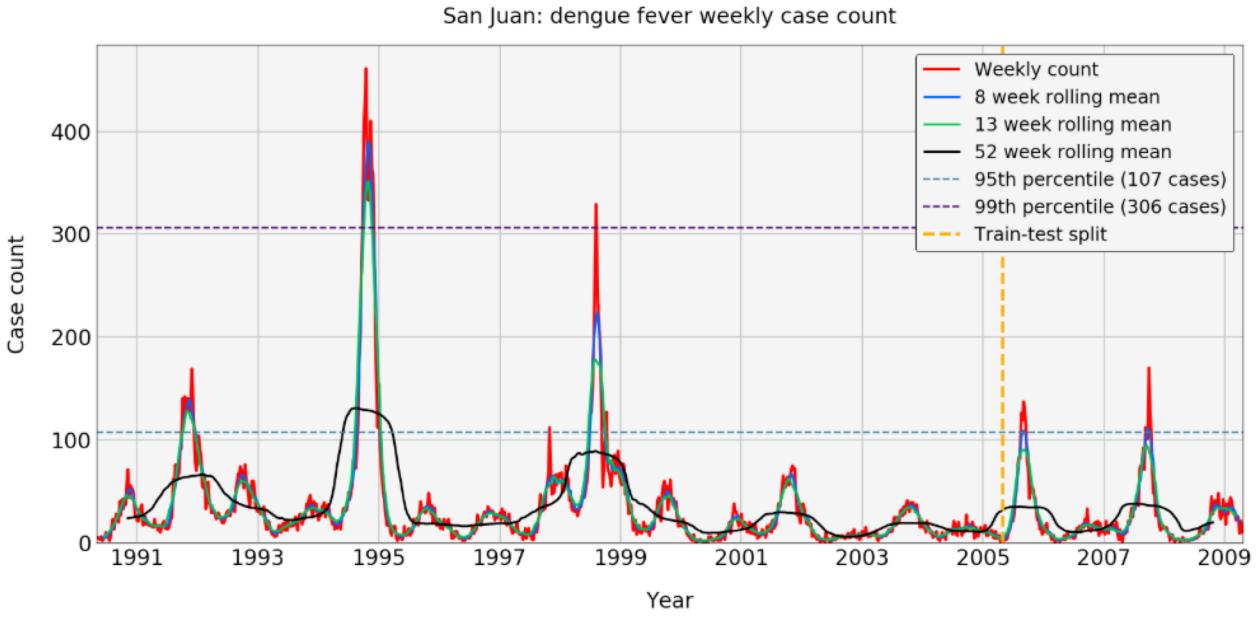
Parse raw source data — Reviewed data and documentation to understand **limitations and quality issues**

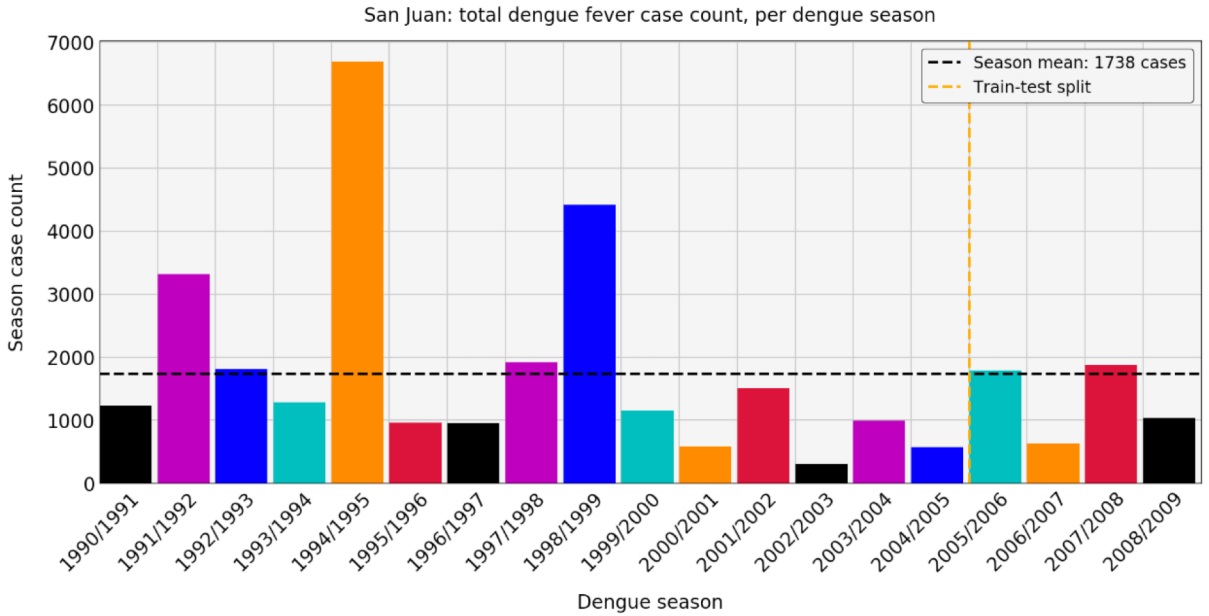
Mine raw data — **Created my own aggregated dataset** from daily data and engineered a range of new features

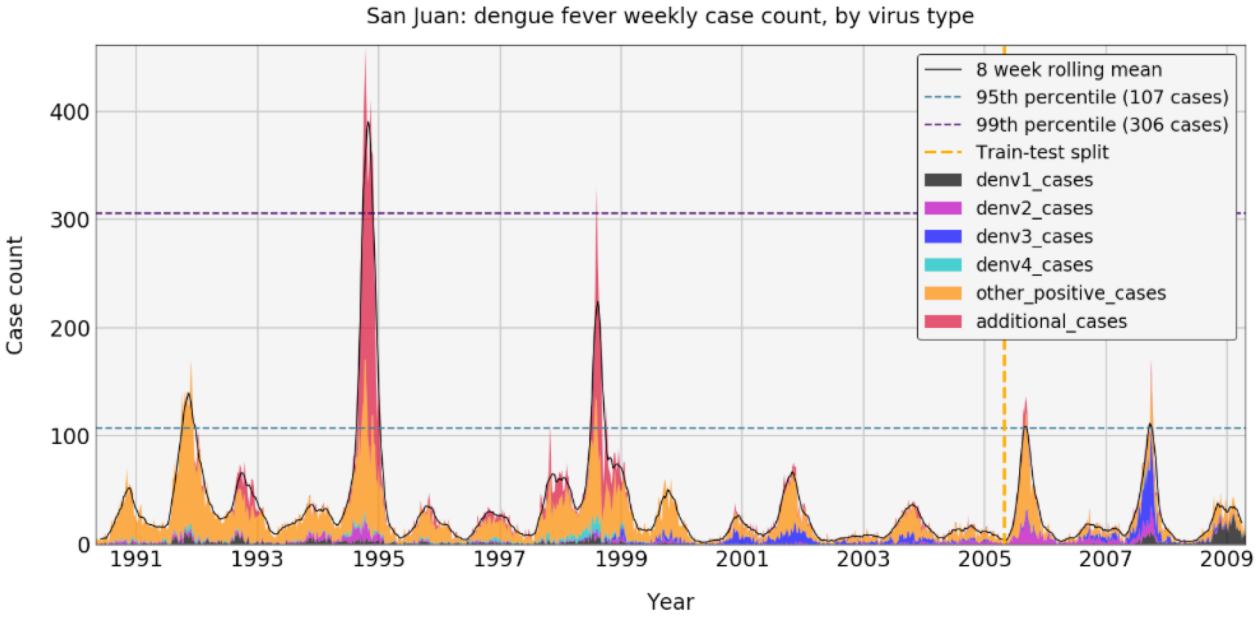
Refine new data set — **Conducted EDA,** imputed missing values and transformed data with lagged, cross correlation analysis

- ✓ It's worth taking time to **reconcile 'pre-packaged' data with the source**
- ✓ Don't underestimate **leap year and 52.17-week year challenges(!!!)** when working with aggregated time series data
- ✓ If the data doesn't 'feel right', it probably isn't. **Make an early call** and avoid sinking time into work-arounds.
- ✓ **Effort rewarded** with greater trust in data and more flexible feature engineering opportunities

DENGUE FEVER IN SAN JUAN: Inter-annual variability





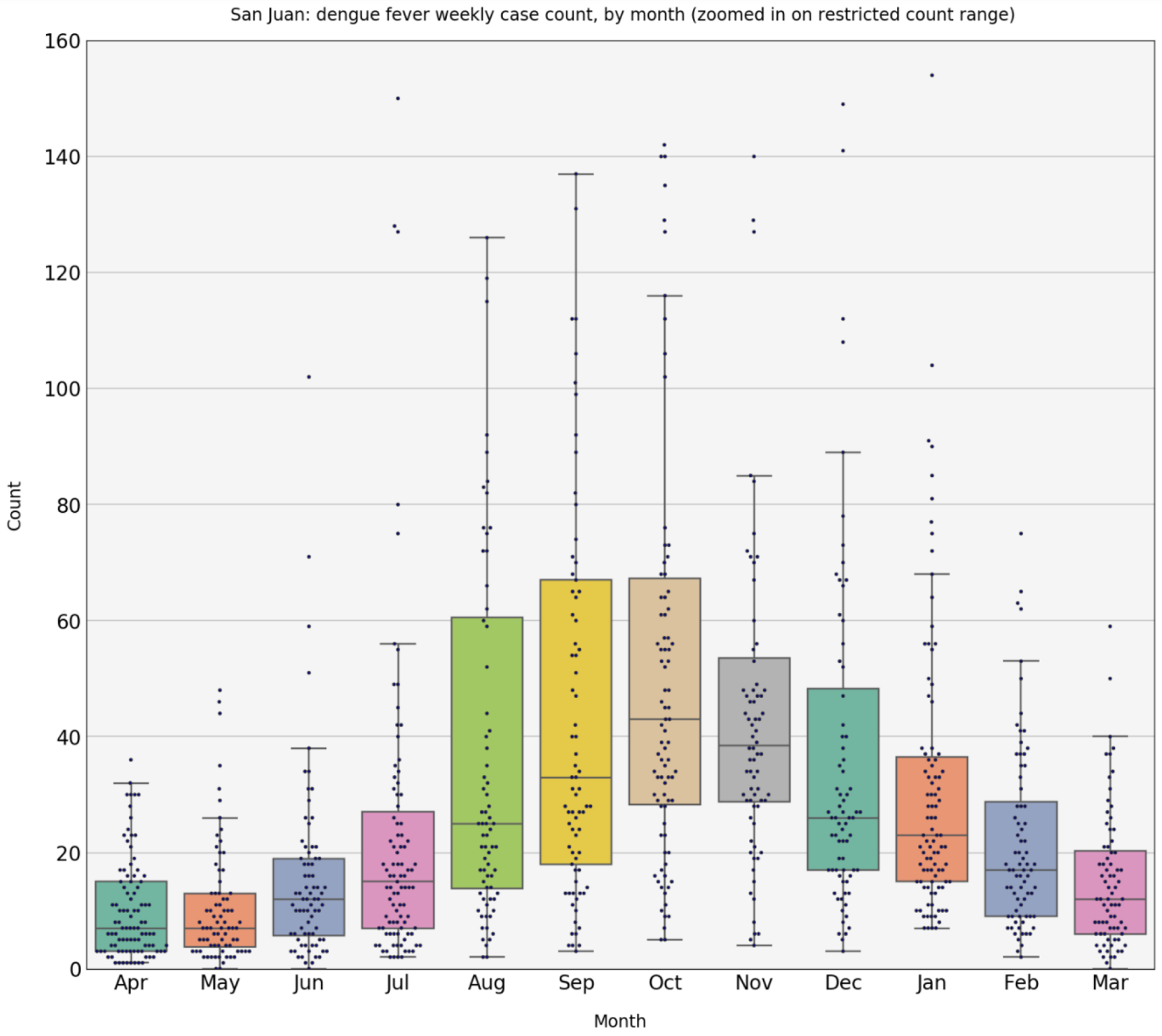


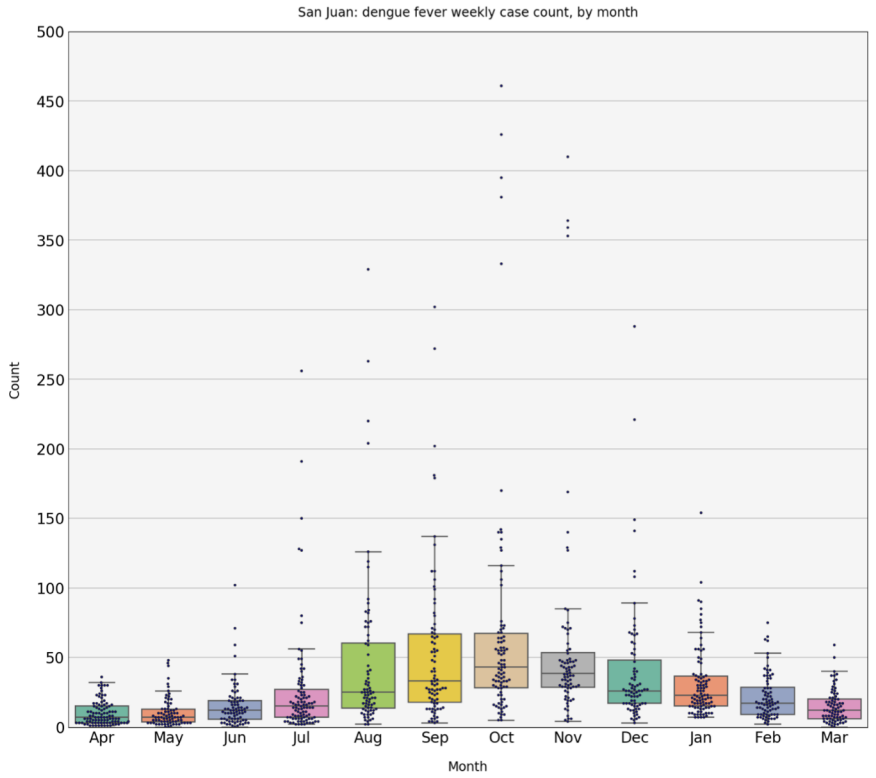
KEY OBSERVATIONS

19 dengue seasons. Train: 1990-2005. Test: 2005-2009.

- 33,028 dengue cases over time period, averaging 33.5 per week.
- Weekly count range: 0-461 (1994 epidemic).
- 6 seasons with counts > 95th percentile (107): 4 in train & 2 in test data.
- Peaks are short & sharp, occur every 3-4 years (apart from test data ☺).
- Slight downward trend punctuated by epidemic events.
- Majority of cases unclassified or estimated up until 2006.
- Seasonal mean: 1,738 cases; Train set: 1,847; Test set: 1,332 in test set.
- Seasonal mean significantly influenced by 1994/95 season (6,700 cases)
- Seasonal std: 1,553 cases; Train set: 1,721; Test set: 602 in test set.
- Difficult to model with relatively consistent environmental predictors.

DENGUE FEVER IN SAN JUAN: Annual variability

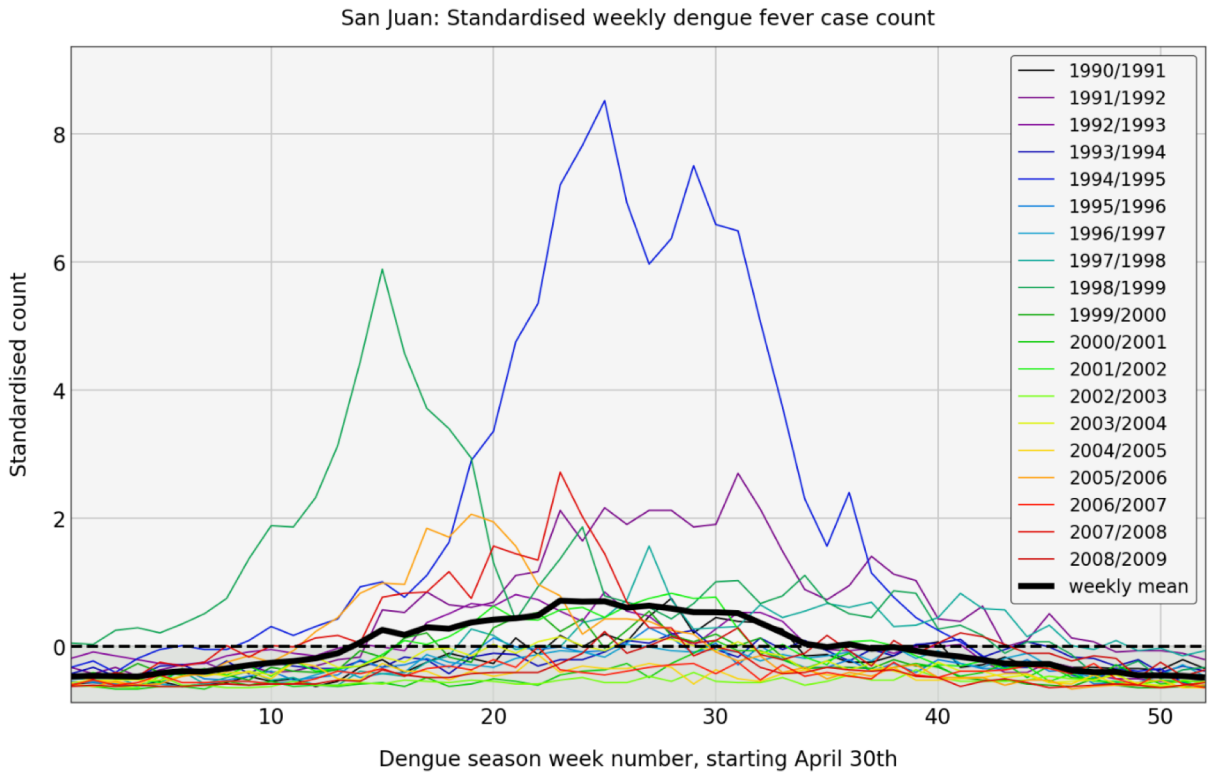


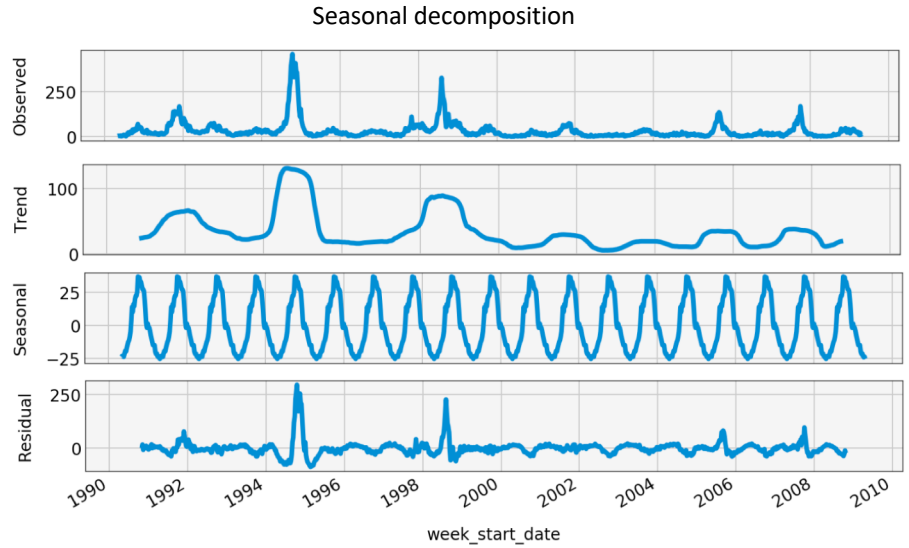


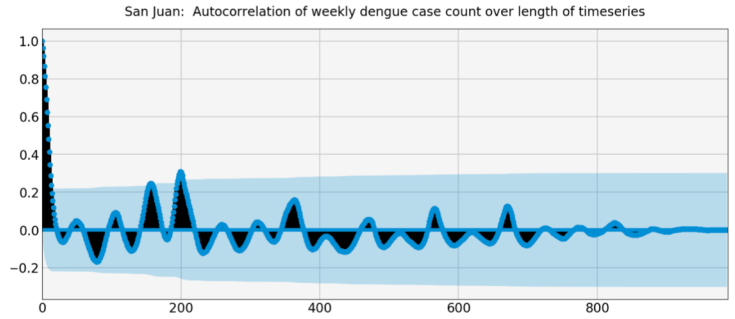
KEY OBSERVATIONS 19 dengue seasons. Train: 1990-2005. Test: 2005-2009.

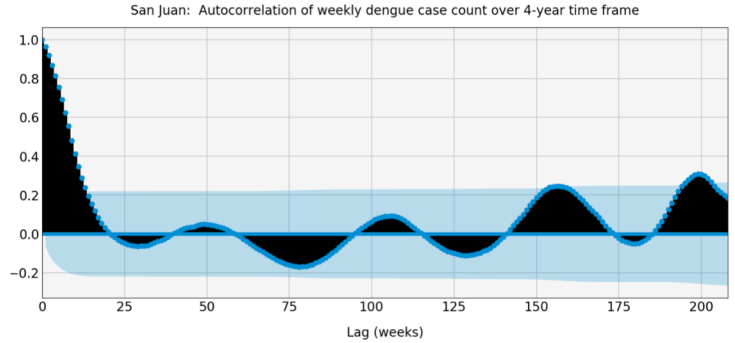
- Median weekly counts peak in latter half of year and are lowest from March to June.
- Variability increases with the median: larger IQR and greater incidence of outliers.
- Variability within time series evidenced by frequency and magnitude of outliers.
- Amplitude of seasonal cycle relatively small.

DENGUE FEVER IN SAN JUAN: Seasonal variability





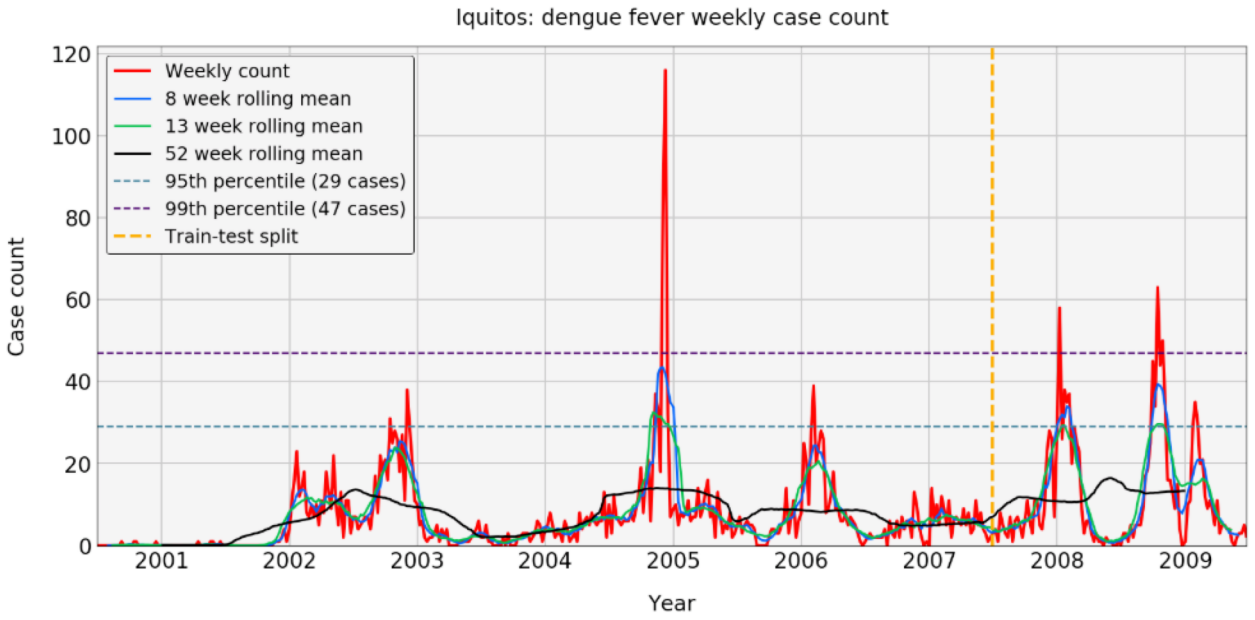


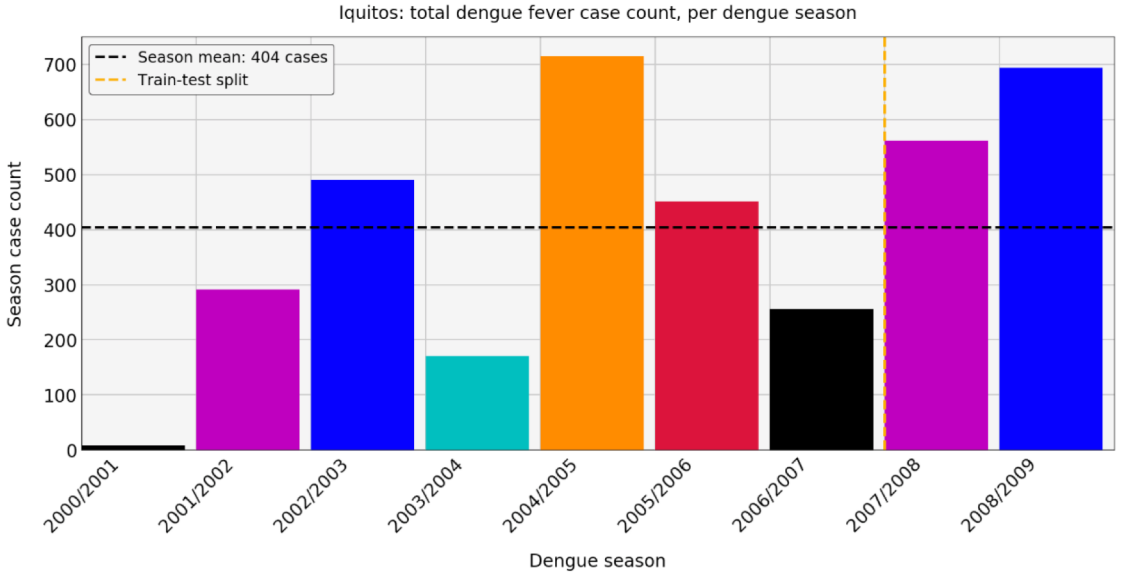


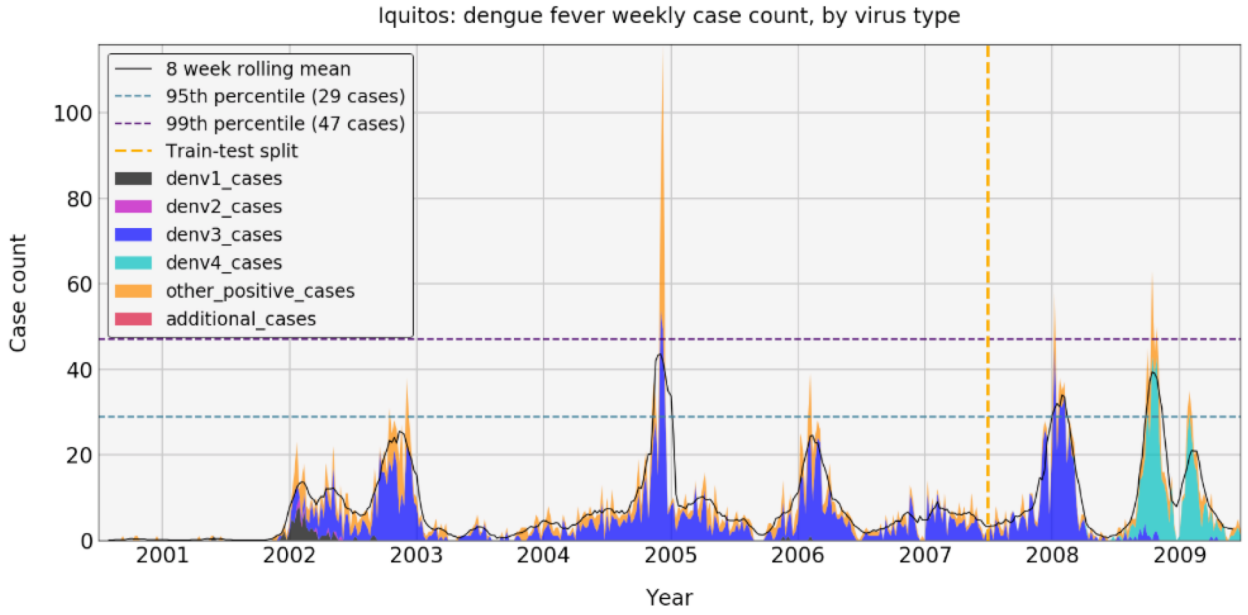
KEY OBSERVATIONS 19 dengue seasons. Train: 1990-2005. Test: 2005-2009.

- Case counts generally exceed mean between weeks 15 and 35. All seasons below mean are at beginning and end of season.
- Peak of the dengue season varies over a wide time range within the season.
- A single season may have multiple peaks.
- Seasonal cycle explaining range of ~6o counts per week. Dominated by slightly downward trend, which peaks with epidemics.
- Statistically significant correlations within time series for lags up to 12 weeks and at 3 and 4 year lags. Surprised to not see higher seasonal correlation.

DENGUE FEVER IN IQUITOS: Inter-annual variability





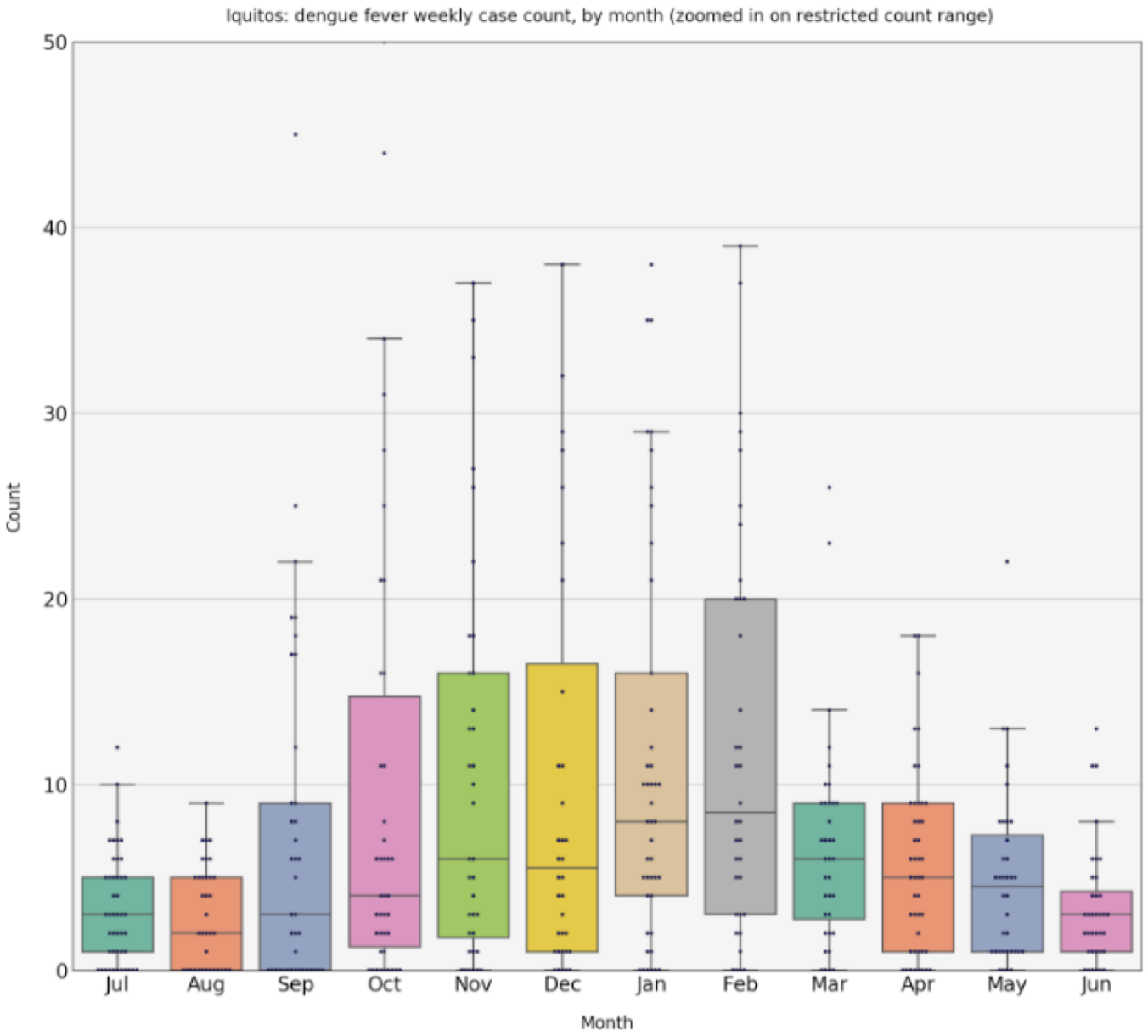


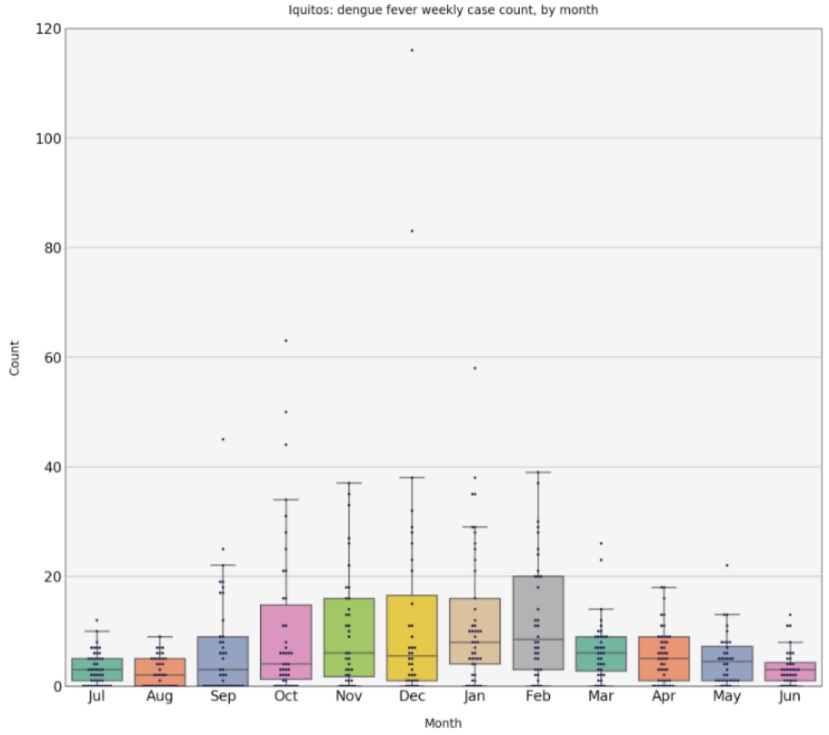
KEY OBSERVATIONS

9 dengue seasons. Train: 2000-07. Test: 2007-2009.

- 3,638 dengue cases over time period, averaging 7.8 per week.
- Weekly count range: 0-116 (2004 epidemic). 0 cases in 19.5% of weeks.
- 5 seasons with counts > 95th percentile (29): 3 in train & 2 in test data.
- Peaks are short (often a single week) , sharp, irregular and can occur multiple times per year.
- Upward trend punctuated by epidemic events.
- 3 different dengue viruses dominating at different times.
- Seasonal mean: 404 cases; Train set: 340; Test set: 628 in test set.
- Seasonal std: 240 cases; Train set: 232; Test set: 93 in test set.
- Difficult to model with relatively consistent environmental predictors.

DENGUE FEVER IN IQUITOS: Annual variability

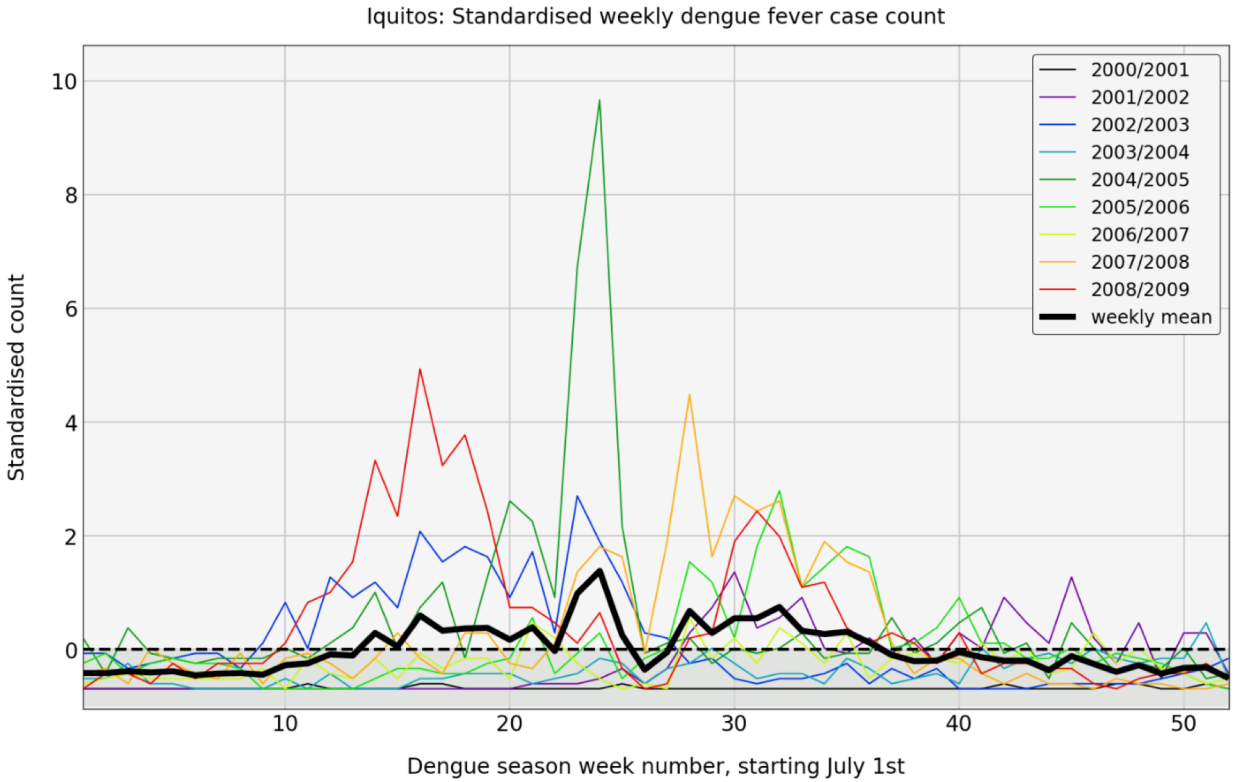


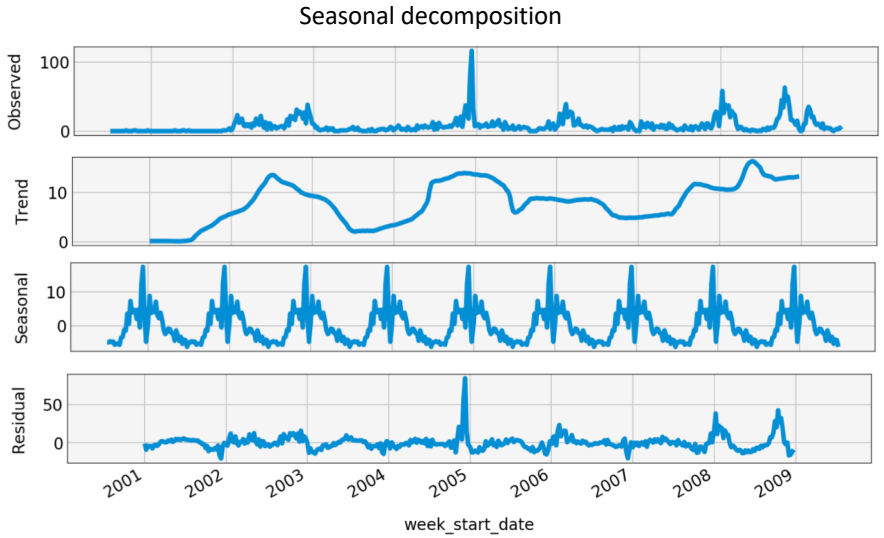


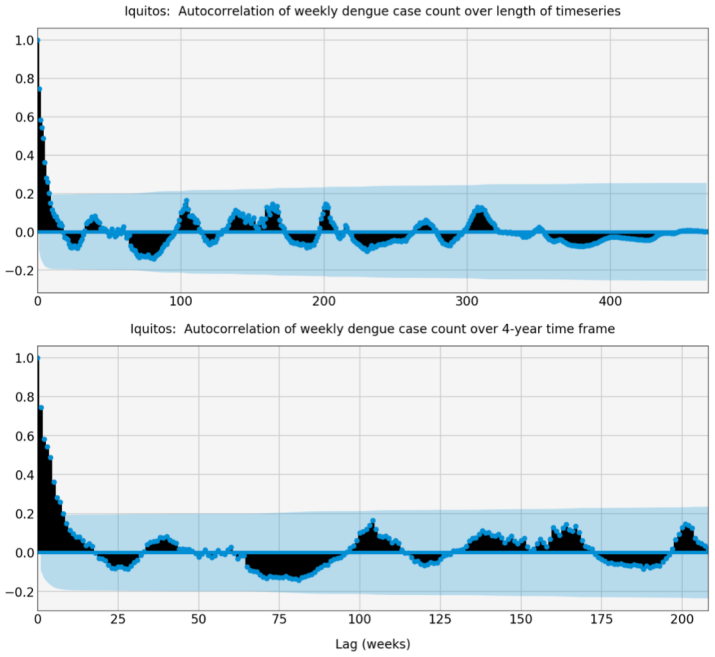
KEY OBSERVATIONS 9 dengue seasons. Train: 2000-07. Test: 2007-2009.

- Small and weaker annual signal than San Juan.
- While all median weekly cases counts < 10, they peak between Nov and Mar, with lower counts between Jun and Aug.
- Like San Juan, variability increases with the median: larger IQR.

DENGUE FEVER IN IQUITOS: Seasonal variability







KEY OBSERVATIONS 9 dengue seasons. Train: 2000-07. Test: 2007-2009.

- Case counts generally exceed mean between weeks 15 and 35. Most seasons below mean at beginning and end of season. Anomalous mid-season dip, coinciding with timing of mass fumigation efforts.
- Peak of the dengue season varies over a wide time range within the season.
- A single season may have multiple peaks.
- Irregular seasonal cycle explaining range of ~15 counts per week. Exists alongside an upward trend.
- Statistically significant correlations within time series for lags up to 8 weeks.

MODELLING

Lagged predictor analysis

Conducted **lagged correlation analysis** between outcome and predictors over a time range of 26 weeks

Identified optimum time lag

Identified **lag to maximise each predictor's correlation** across range e.g. 4-26 wks, 8-26 weeks; honouring prediction lead-time objective. Filtered features by correlation threshold.

Reviewed variation over time

Reviewed variations in correlations over lagged timeframes to ensure **local maximum identified by optimum time lag**

Created new predictor dataset

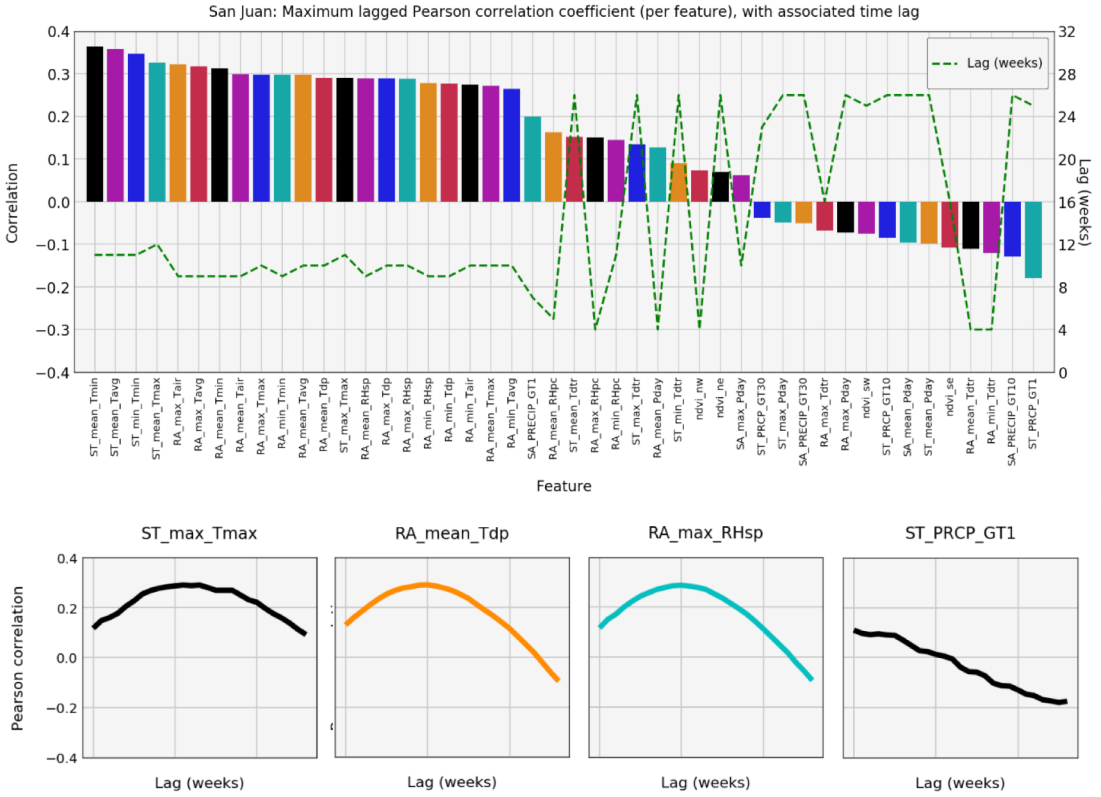
Lagged all predictors by optimum time lag for maximum correlation with outcome throughout the time series

Added historical dengue cases

Added 4, 8 and 12-wk **lagged dengue counts, seasonal cycle** (train) and lagged 4-year counts (for San Juan) as predictors

Ready to model…

…after a lot of **verification of predictor transformations**!!!



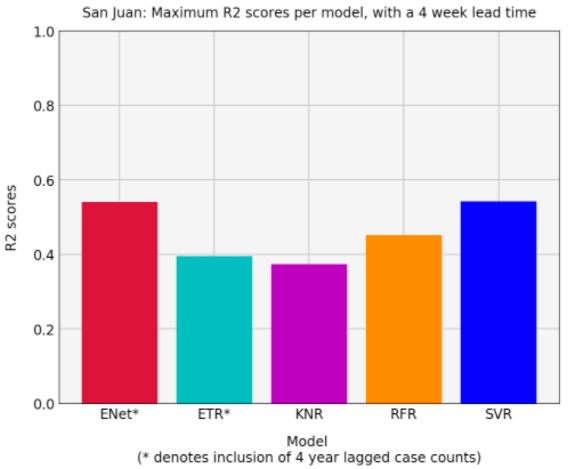
KEY OBSERVATIONS

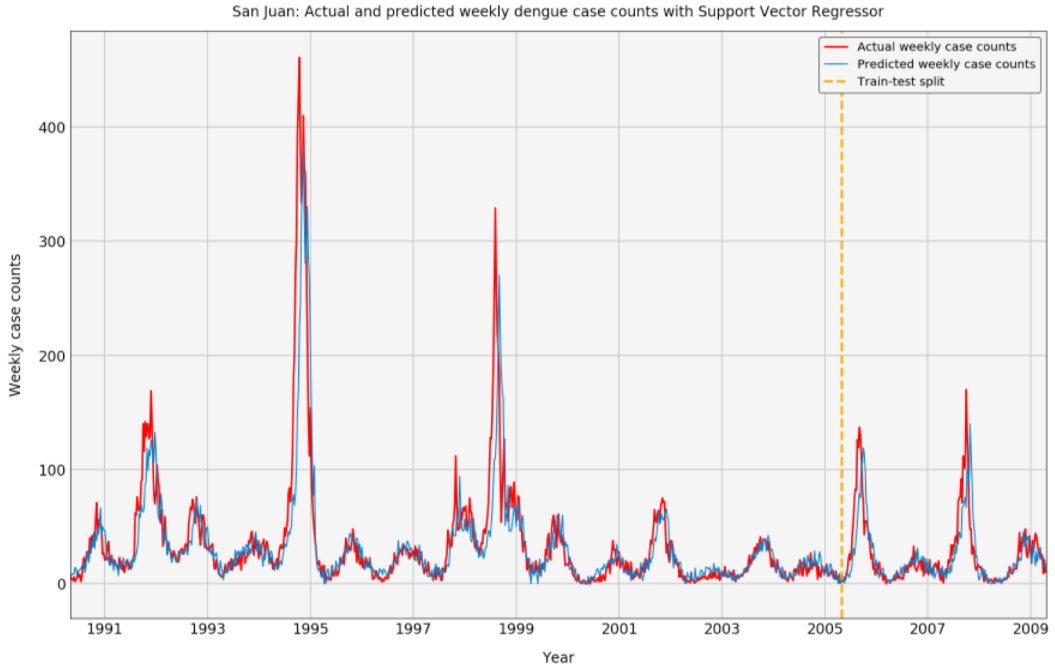
- Surprisingly large number of lagged predictors with correlations exceeding +/- 0.2 at some point over lag.
- Positive correlations tend to peak between weeks 6 and 13, while negative correlations peak around weeks 24-26.
- Short lags: positive correlations are temperate and humidity; negative correlations are rainfall related.
- Correlation heatmap reversed for second half of the year.

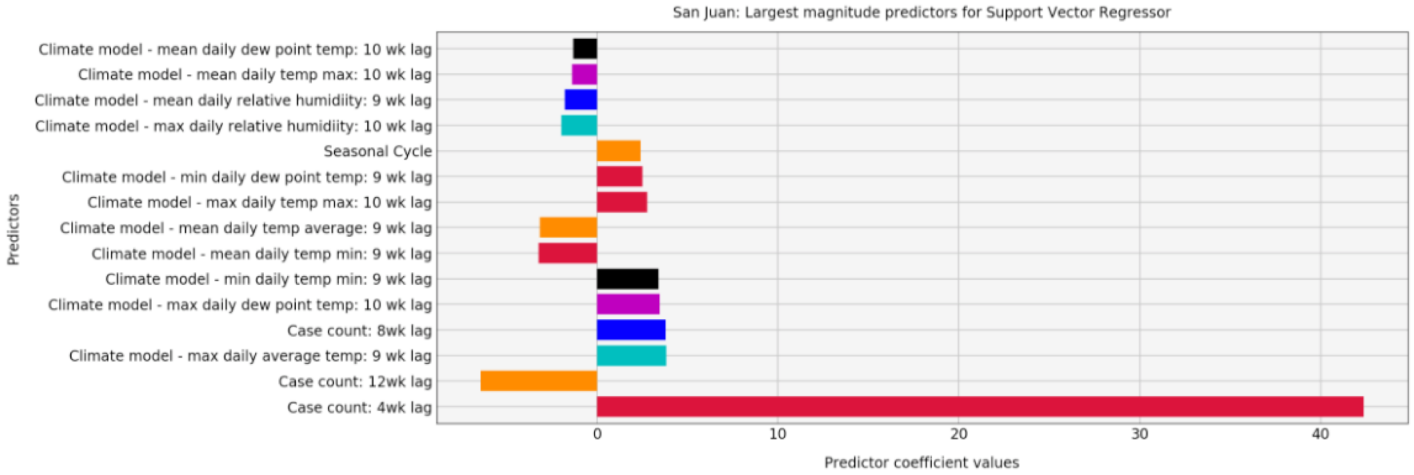
MODEL SUMMARY: San Juan

APPROACH AND OUTCOMES

- Grid searched 5 regressors: Elastic Net, Random Forest, Extra Trees, Support Vector Machines, KNeighbors.
- Model experimentation: range of prediction lead times, environmental vs. case count predictors, PCA, seasonal averages, 4-year lag (for San Juan only).
- Best models: SVR (R2 = 0.543 and RMSE = 20) and Elastic Net (R2 = 0.541 and RMSE = 20).
- Optimum SVR predicts shape of test time series (with lag) and approximates peaks and troughs in the test series. Model performance degrades as actual counts increase. Largest coefficients: lagged weekly dengue case counts and climate model temperature and relative humidity predictors with lags of 9-10 weeks.



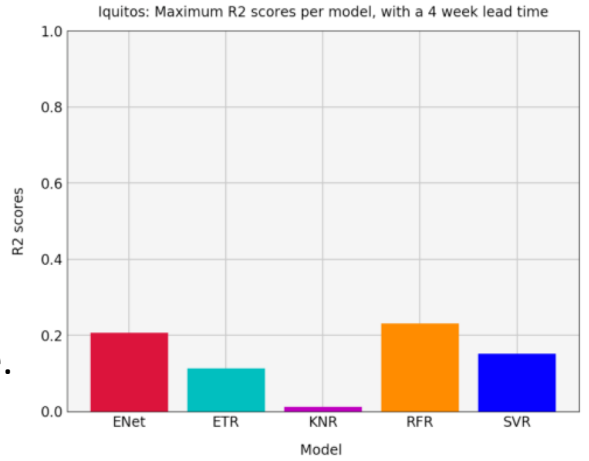


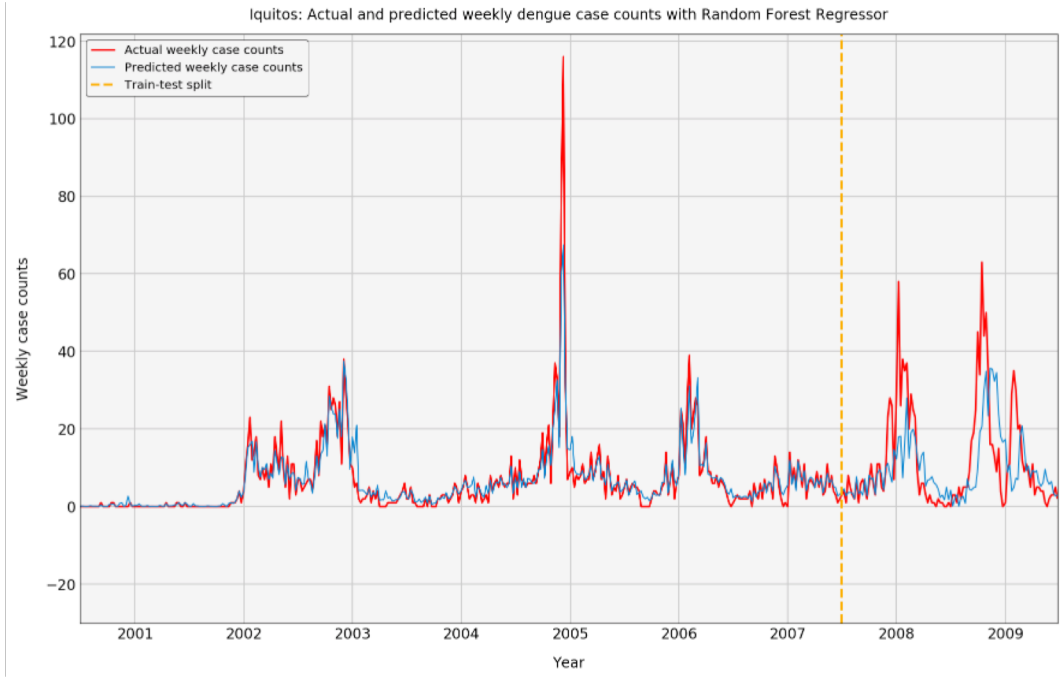


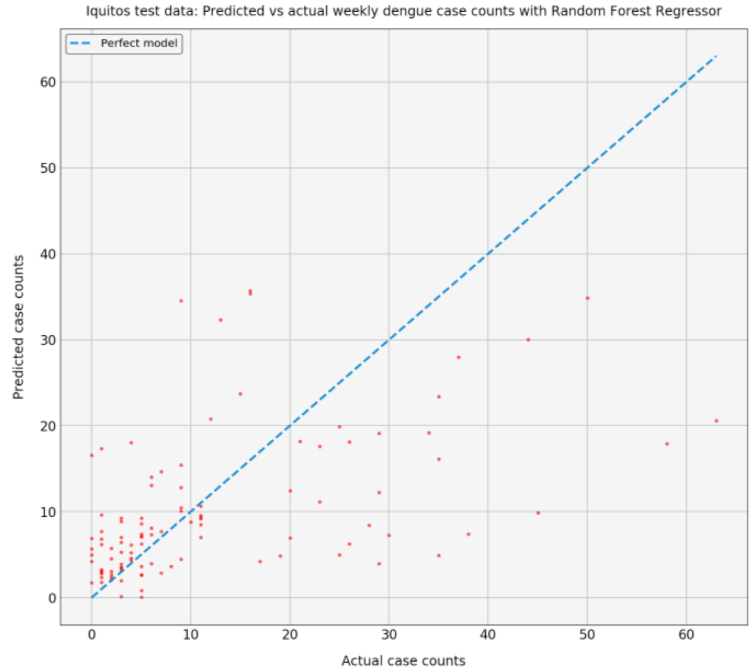
MODEL SUMMARY: Iquitos

APPROACH AND OUTCOMES

- Grid searched 5 regressors: Elastic Net, Random Forest, Extra Trees, Support Vector Machines, KNeighbors.
- Model experimentation: range of prediction lead times, environmental vs. case count predictors, PCA and seasonal averages. Not surprisingly, model performance much weaker than San Juan.
- Best models: RFR (R2 = 0.231 and RMSE = 11) and Elastic Net (R2 = 0.207 and RMSE = 12).
- Optimum RFR predicts approximate shape of test time series (with lag) but unable to replicate variance - underestimating peaks and overestimating troughs. Model performance degrades as actual counts increase.







CONCLUSION AND NEXT STEPS

"Can weekly dengue case counts be accurately predicted 4 weeks in advance?"

YES! We can do a reasonably good job*!

(in San Juan)*

- Expansion of model selection and more intentional tuning of hyperparameters, including time series modelling (ARIMA, SARIMAX, seasonally differenced).
- Enhance understanding of impact of PCA on model performance.
- Modelling of aggregated / smoothed dengue fever counts (e.g. monthly, seasonal).
- Experimenting with additional smoothed predictors (e.g. historical rolling means vs. lagged weekly counts).
- Classification modelling: probability of a weekly case count exceeding the 80th percentile.
- Iquitos: engineering a feature to represent mosquito control interventions and seeking a longer time series.