# **Statement of Problem**

For this exercise we were asked to participate in the Driven Data “DengAI” competition. The competition goal is to build predictive models for the spread of dengue fever in the cities of San Juan Puerto Rico and Iquitos Peru. The models predict the number of weekly cases reported in each city during the specified weeks in 2008 – 2013 (San Juan) and 2010 – 2013 (Iquitos).

# **Significance**

Dengue fever is one of the many mosquito-borne illnesses found around the globe. According to the competition website (DrivenData, n.d.), dengue fever transmission is affected by climate variables. With global climate change and related changes to local weather and vegetation patterns, infection patterns of pest-borne illnesses are likely to change too. Rates of dengue fever have increased to the point the World Health Organization estimates about half the world’s population is a risk for it (WHO, n.d.). Given the disease has no specific treatment and that it is a leading cause of serious illness for children in Latin America (WHO, n.d.), predicting outbreaks has the potential to help governments to know when to spend additional resources in vector-management (mosquito eradication), potentially help people know when to seek early treatment which is the best chance to survive the disease, and help medical facilities plan for waves of people seeking treatment.

# **Data – Exploration and processing**

The data are composed of a training set, and a test set used for making submission to the competition ranking tool. There are 24 variables in the data, including a number of meteorological measures, normalized vegetative cover ratings, and date information. There are 936 observations for San Juan and 520 for Iquitos. There is also a file of “labeled” data, which provides the target: weekly cases of reported fever. Training data set was created by joining the reported number of cases to the rest of the data.

There are missing data in the training data file, as seen in the following visualizations. The label file data were complete.

Figure - Missing Data for San Juan

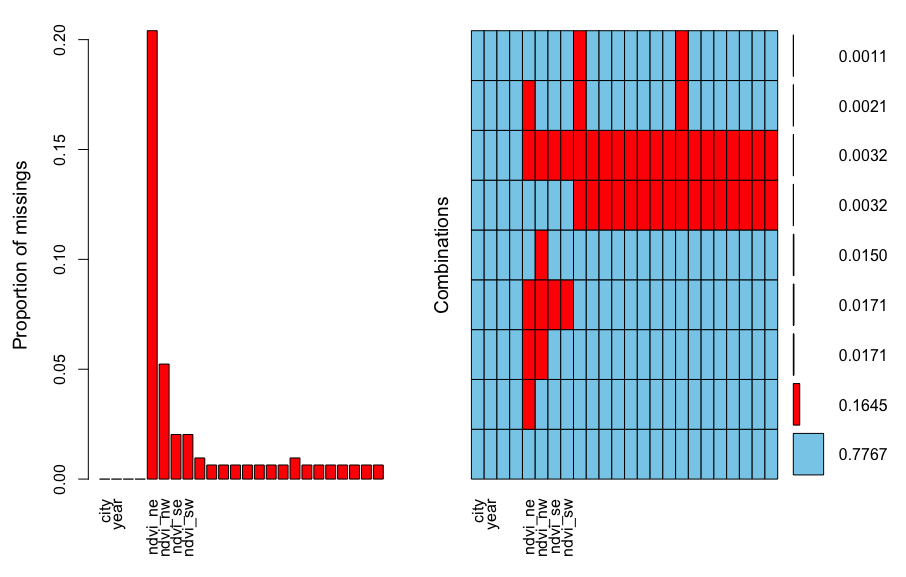
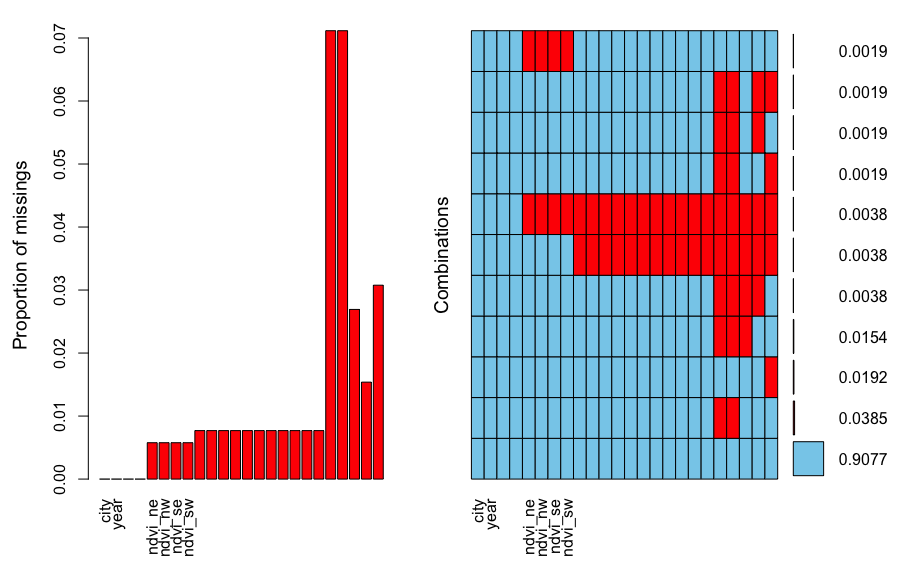


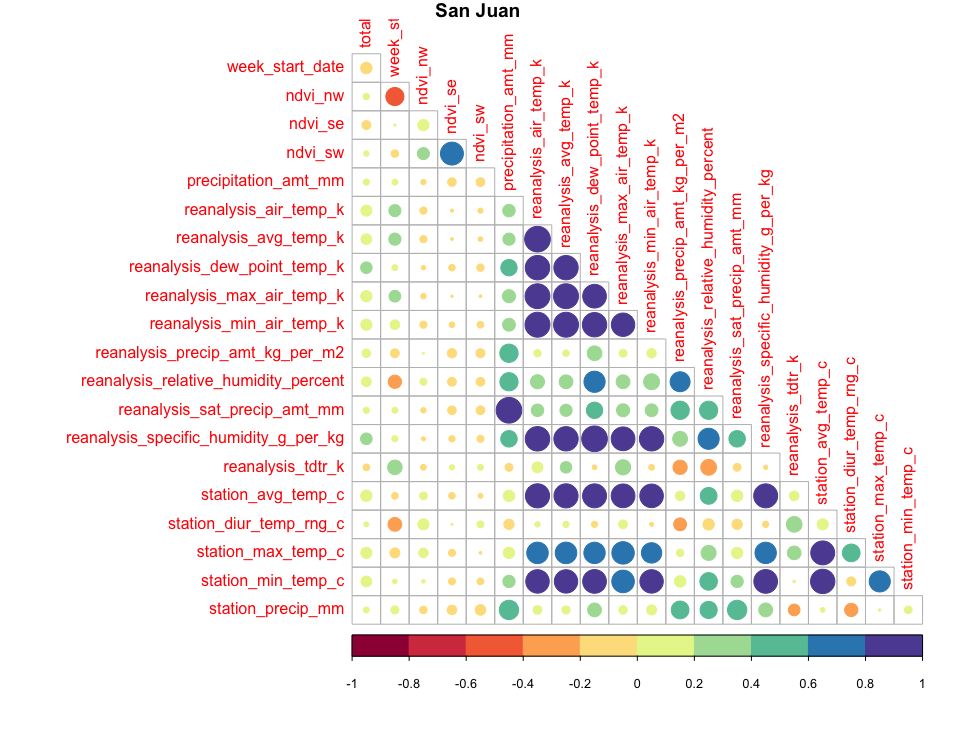
Figure - Missing Data Iquitos



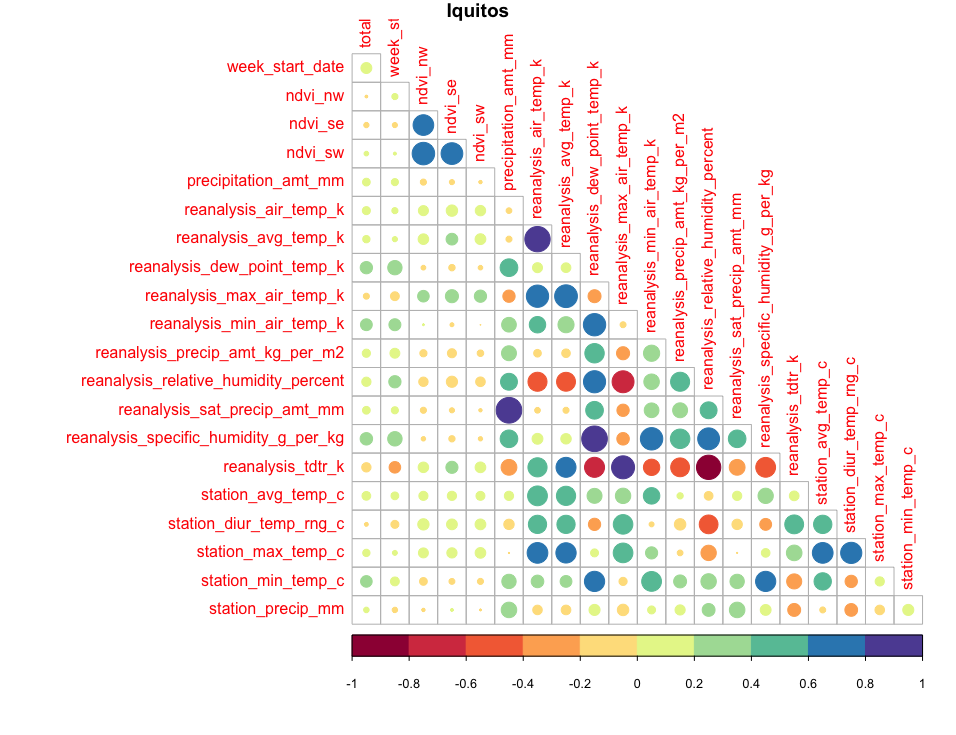
Missing data was imputed by using the *zoo* package’s *na.locf* function to take the last available value and carrying it forward to the missing value. The date for the start of the week was converted to POSIX and divided by 86400 to yields days since 1/1/1970, this allowed for numerical handling of the start-of-week data.

Correlations were computed and are shown below. It is clear that there is a high degree of collinearity between the data for each city.

Figure

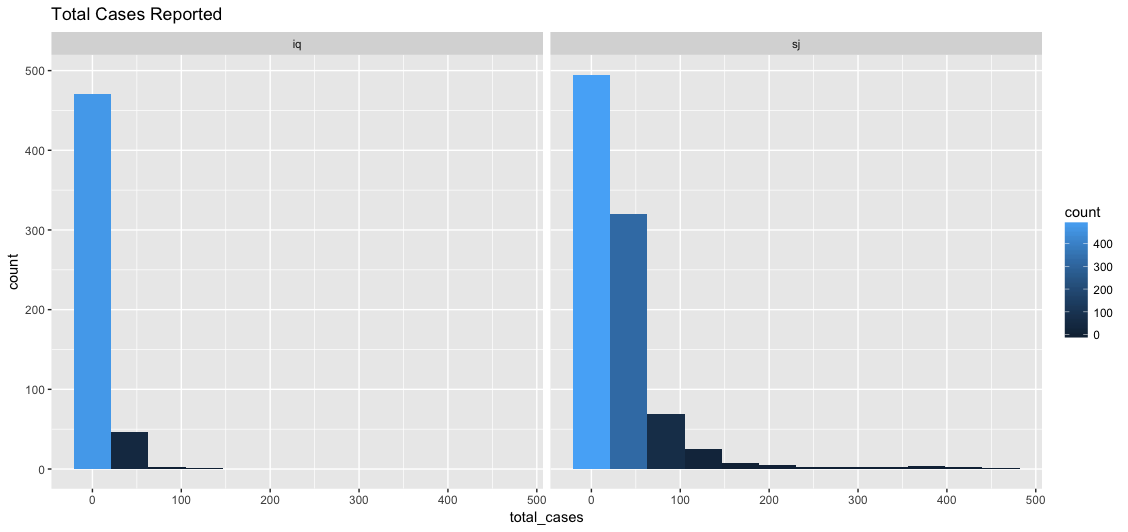


Figure



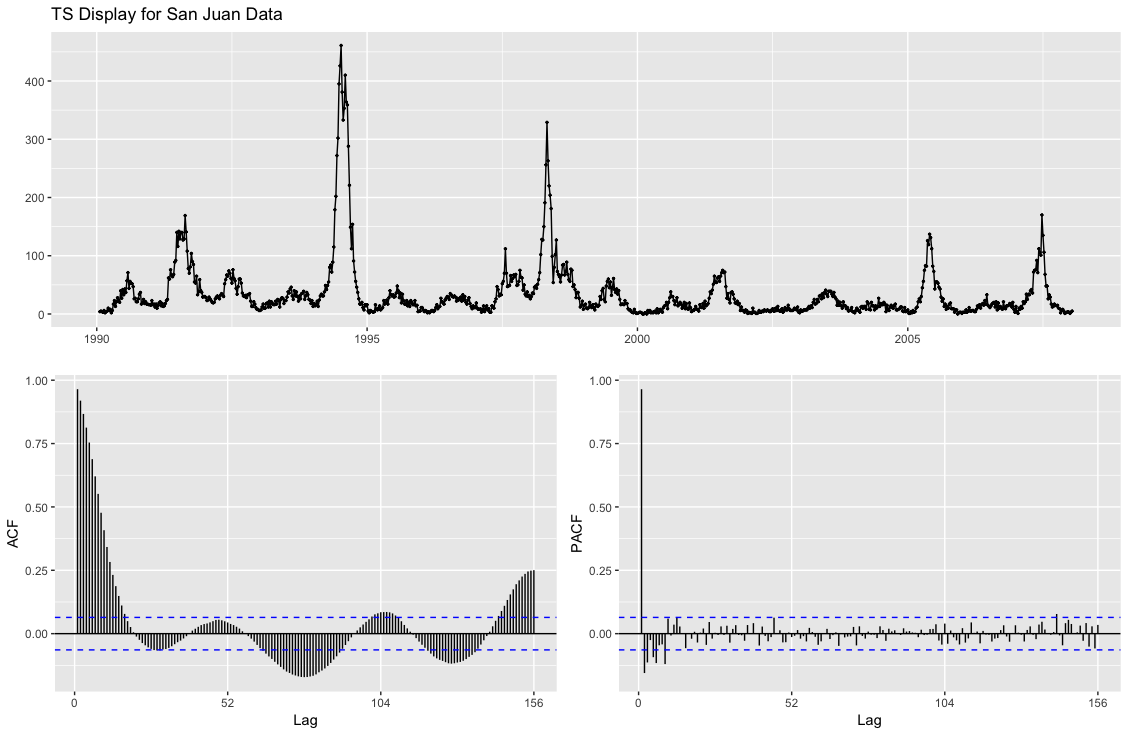
The distribution of the cases reported appears to be a Poisson distribution, with a small lambda.

Figure

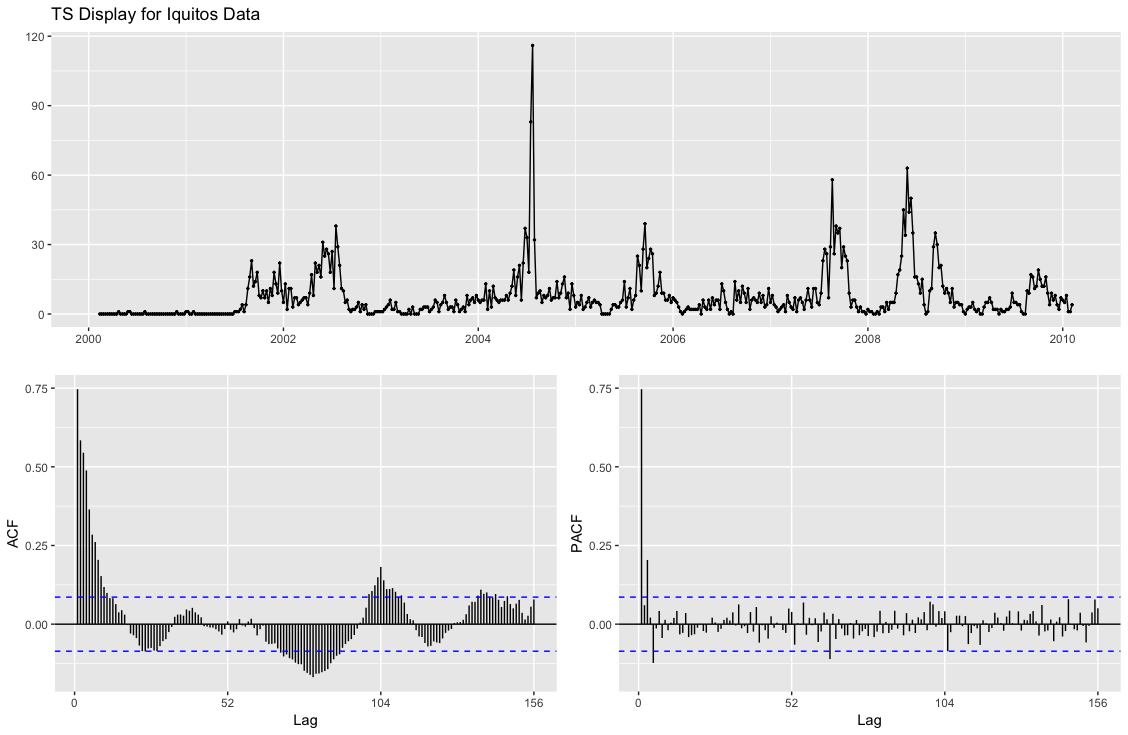


The *ggtsdisplay* function shows that San Juan has had several outbreaks of dengue, as seen in the spikes of the cases. The ACF indicates a non-stationary series, and the PACF suggests starting with a lag of two.

Figure



Figure



Similarly, Iquitos shows non-stationary series and suggests a moving average lag of three might be an appropriate place to start. The ACF is less smooth than for San Juan, and there appear to be more frequent outbreaks.

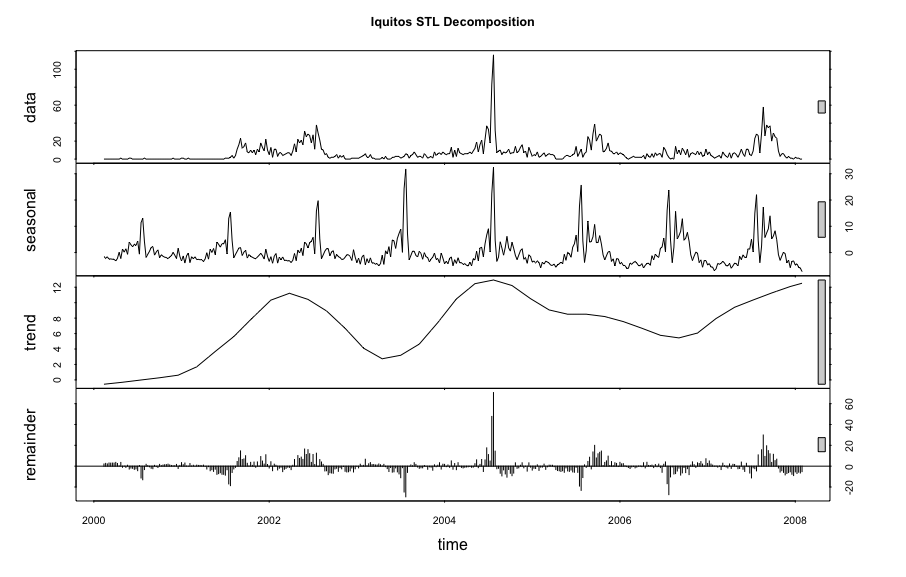
Our target, the total number of cases reported, has a mean of 34.18 and a standard deviation of 51.38 in San Juan and a mean of 7.57 and a standard deviation of 10.77 in Iquitos.

Creating an STL decomposition of the data, and plotting revealed the presence of periodicity and trend information for both San Juan (Figure 8) and Iquitos (Figure9). The seasonal component for both cities appear to change over time. For San Juan, the seasonality appears to be damping down, while for Iquitos the seasonality seems to have an increasing magnitude. The trend component for San Juan is decreasing over time, while the Iquitos trend is increasing.

Figure

# 

Figure



Both cities show outbreaks which result in outliers in the data. I elected to leave the outliers in place. My thinking was that the outliers are legitimate data points, and indicative of the severity of an outbreak; having that data as part of the model seemed important. Although I did try a few variations of models with the outliers removed. My final model was built using the data containing outliers. I also elected to use the full data set for many of my models. Retrospectively, given the multicollinearity in the data, this may not have been the best choice. However, some models were built using reduced data sets based on my selection of non-correlated features. My ultimate “hero” model is an ensemble of a model built on all data and a model built on selected variables.

The full training data set was split into a training set and a test set, so I could look at evaluation statistics for each model built. The split was approximately 80% training, 20% testing, San Juan had 186 test cases, Iquitos 105. Evaluation was done by running the *accuracy* function on the forecast objects and appending the results to a tracking dataframe. Predictions were created using the entire training data set. The MAE score produced by the Driven Data site on their test-scoring data acted as the final arbiter of accuracy.

# **Literature Review**

There are a wide range of articles in the topic area of disease prediction. Many focus on the pure mathematics of the problem, while others approach it from a biological systems and migratory patterns perspective. There was also a body of work focusing on the types of modeling we’ve been doing in this course. Campylobacter prediction was evaluated using regression, decomposition and ARIMA models at the University of Tennessee (Weisent, 2010). While campylobacter is not a mosquito-borne disease, it shares the traits of having autoregressive and moving average aspects to the data, making those results informative to dengue model building.

A team from India (Chatterjee & R., 2009) used multi-step polynomial regression to forecast malaria. Given malaria and dengue share the mosquito transmission vector, their work seemed relevant to our task. The authors chose features based on correlation to the target variable and performed non-linear curve fitting. Like with dengue, temperature, humidity, and rainfall were part of the features used in the model. Another malaria study (Teklehaimanot, 2004) used lagged polynomials to model disease spread relative to weather patterns in Ethiopia. Looking more at the vector than at a specific disease, Lee et. al. (Lee, Chung, & Hwang, 2016) evaluated the use of neural networks in predicting the mosquitos themselves. The clear implication being, more mosquitos more mosquito-borne illness. Finally, looking specifically at dengue, there is a study from Rajasthan (Bhatnagar, 2012)used ARIMA modeling to predict dengue outbreaks. And a study using neural networks (Aburas, Cetiner, & Sari, 2010) to predict confirmed cases of dengue.

# **The Models – Formulation, Performance/Accuracy, Limitations**

# Formulation

The vast majority of the model formulation and all data exploration was done in R. The “combos” and ensembles I tried were created using Excel, simply because it was the fastest solution to cut/paste/merge results which were already in CSV format; all other model work was coded in R. I began with a basic default *ETS* model. This got me an MAE of 33.7548 on the evaluation set. From there I moved on to trying *STL* and *STLF* models. *STLF* did relatively well for me, with an MAE of 27.1635 on Driven Data.

Based on my reading during the literature search, I had high hopes of the ARIMA results performing strongly. When the auto-ARIMA results were not as high as I hoped, I also tried tweaking the parameters by hand. I tried 6 and 12 as possible lag values based on the PACF charts, I used 1 for the differences based on the values from *ndiffs*, and I tried different numbers of AR terms based on my reading of the ACF charts. None of these were particular improvements, but the ARIMA (2,1,12)(0,0,0) model for San Juan and ARIMA (2,1,12)(0,0,1) for Iquitos did the best of the bunch.

I also tried *auto-ARIMA* with an assortment of external regressor terms. These performed better than the *auto-ARIMA* models without external regression but did not do as well as the *STLF* model. I did use decomposition, and GARCH to build a model in which I tried to use GARCH to predict residuals, and added those to the ARIMA seasonal predictions. I must not have done it properly since that model only yielded and MAE of 27.1923 which was not as high as I expected from discussions in class.

I also tried traditional regression techniques and machine learning models. Of the regression techniques, Poisson regression outperformed ARIMA for me, and Generalized Linear Model Negative Binomial (GLM.NB) outperformed Poisson. Regular GLM did not do as well as GLM.NB.

I tried *XGBoost* with a few different sets of parameters, as well as *NNetar* models. All of the machine learning models beat the ARIMA models, and one of the *XGBoost* models outperformed the *STLF* model.

My final “best” model was an ensemble, created from an *XGBoost* model and my best GLM Negative Binomial model. My best score was an MAE of 25.6659 which put me at 467 in the Driven Data rankings (see Figure 14 in the appendix).

# Performance and Accuracy

The full set of results for my various models are given in the Appendix. However, I’ll discuss the best model here.

My best results, an MAE of 25.6659, came from an ensemble model of 2 other models: XGB3 + GLM.NB4. The performance of these models, on the test split of the training data set is:

Table

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **ME** | **RMSE** | **MAE** |
| XGB3 SJ | -10.6806 | 55.31922 | 42.6855 |
| XGB3 IQ | 7.071686 | 13.41154 | 7.75637 |
| GLM.NB 4 SJ | -9.39171 | 37.57017 | 30.42294 |
| GLM.NB 4 IQ | 3.17879 | 12.28525 | 7.433272 |

For all of my models, the Iquitos results (RMSE, MAE, ME) are much lower than the results for San Juan. Holding back a larger portion of the data of model evaluation might have helped reduce the difference between my results on the full training set and the models’ performance on the site’s scoring test set.

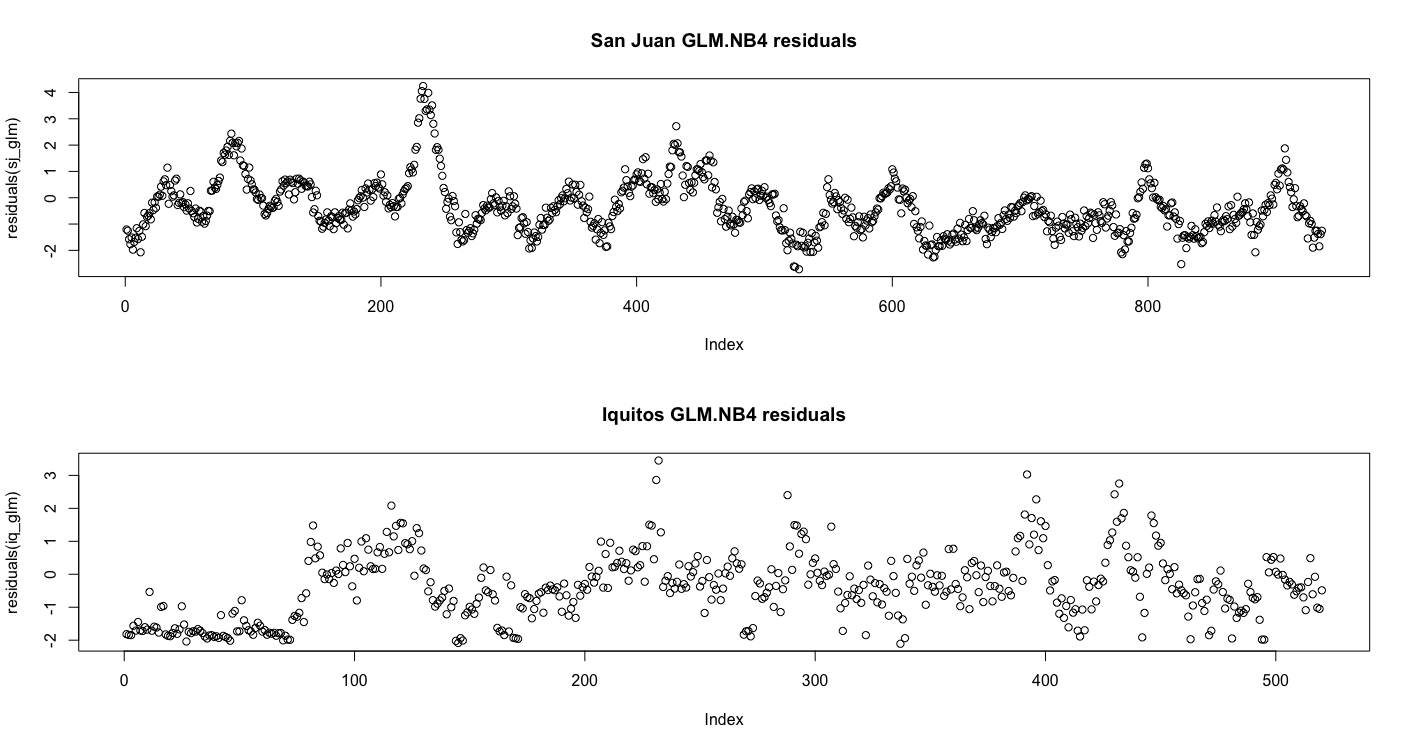
The results of the models on the Driven Data Leaderboard are:

Table

|  |  |
| --- | --- |
| **MAE** | **Model** |
| 27.0337 | XGB3 |
| 26.3582 | GLM NB4 |
| **25.6659** | **XGB3 + GLM.NB4** |

Typically, I would want to look at the residuals for a model, but I’m unclear how you do that when working with an ensemble. Also, XGB doesn’t have a function for residuals that I was able to find, so just looking at the residuals for the GLM.NB model resulted in the following:

Figure



The San Juan residuals look like there is still “signal” in the series, I can see why modeling the residuals with GARCH could have been successful. The Iquitos residual series looks closer to a random distribution, but I’d say there is still a pattern and that more work needs to be done.

My best overall single-technique model was the GLM.NB4, using precipitation\_amt\_mm, reanalysis\_specific\_humidity\_g\_per\_kg, ndvi\_sw and station\_avg\_temp\_c as my regression variables for both San Juan and Iquitos. This resulted in an MAE of 26.3582 from Driven Data.

# Limitations

Based on the residuals from the models in my ensemble, I’d say there is still “signal” to be extracted from the data. This limits the overall reliability of the model. Also, as weather patterns shift, the conditions which led to the cases in the training data could similarly shift and alter the reliability of the model.

# Future Work

To improve these models, I’d start by enhancing the data to make use of information on mosquito life-cycles. I’d do this by adding flags or a category for optimal mosquito conditions. An article from Brazil (Costa, 2010) indicates that it might be possible to divide the data into finer-grained categories based on the weather’s impact to mosquito populations.

Also, I’d like to take into account the differences between Iquitos and San Juan. Iquitos is at a higher elevation than San Juan, and it gets roughly twice the amount of rain that San Juan does on an annual basis; 55 inches annual on average for San Juan, 112 for Iquitos (Puerto Rico; Peru, n.d.). All of my submitted competition results used the same modeling techniques for both cities. They are different enough that an exploration of separate modeling could yield better predictions.

# **Learnings**

Modelling disease vectors is not easy. Given the vast amount of research into modelling malaria, as well as the mosquito life-cycle, it is clear to me these biological systems are not perfectly understood. Add in weather, a notoriously difficult thing to predict accurately, and the problem complexity increases yet more.

I was mildly surprised that the not-time-series specific techniques, like *XGBoost*, and *GLM* performed so well on the data. My take away from this is something Dr. Fulton mentioned the first week, really, time series are just specialized linear regressions. It was nice to see that in action.

I also learned that you can’t always trust your train/test split in these sorts of competitions. According to my test set MAE accuracy results, a model combining the Hand ARIMA 4 results for Iquitos with the *NNetars* results for San Juan should have been the best. It wasn’t; that combination only got a 33.1587 from the Driven Data scoring system. This might have been due to my using only ~20% of the original data for testing. In the future a larger evaluation set might be in order.

XGBoost handled the full data set well, as did the STLF model, however, many of my full-data models were lack-luster. Looking back, I regret not thinking more clearly about the collinearity in the data and reducing the feature set to have less collinearity. I believe that feature reduction would have many very useful in the case of the ARIMA/auto-ARIMA models.

Appendix

# Works Cited

Aburas, H. M., Cetiner, B. G., & Sari, M. (2010). Dengue confirmed-cases prediction: A neural network model. *Expert Systems with Applications*, 4256-4260. https://doi.org/10.1016/j.eswa.2009.11.077.

Bhatnagar, S. L. (2012). Forecasting incidence of dengue in Rajasthan, using time series analyses. *Indian Journal of Public Health*, 281. .

Buczak, A. L. (2012). BMC Medical InforA data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data (matics and Decision Making. *BMC Medical Informatics and Decision Making*, 124-144. https://doi.org/10.1186/1472-6947-12-124.

Chatterjee, C., & R., S. R. (2009). Multi-Step Polynomial Regression Method to Model and Forecast Malaria Incidence. *PLoS ONE* , https://doi.org/10.1371/journal.pone.0004726.

Costa, E. A. (2010). Impact of small variations in temperature and humidity on the reproductive activity and survival of Aedes aegypti (Diptera, Culicidae). . *Revista Brasileira de Entomologia*, 488-493. Available from: http://www.scielo.br/scielo.php?script=sci\_arttext&pid=S0085-56262010000300021&lng=en. http://dx.doi.org/10.1590/S0085-56262010000300021.

DrivenData. (n.d.). *DengAI: Predicting Disease Spread*. Retrieved from DrivenData: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

Lee, K. Y., Chung, N., & Hwang, S. (2016). Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. *Ecological Informatics*, 172-180. https://doi.org/10.1016/j.ecoinf.2015.08.011.

*Puerto Rico; Peru*. (n.d.). Retrieved from Climates to Travel: https://www.climatestotravel.com/climate/

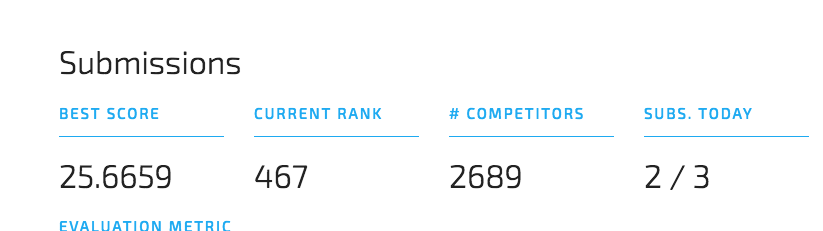
Teklehaimanot, H. D. (2004). Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. *Malaria Journal*, 3 - 41. doi:10.1186/1475-2875-3-41.

Weisent, J. S. (2010). omparison of three time-series models for predicting campylobacteriosis risk. *Epidemiology and Infection*, 898-906. Retrieved from http://www.jstor.org.turing.library.northwestern.edu/stable/40603351.

WHO. (n.d.). *Dengue and severe dengue*. Retrieved from World Health Organization: http://www.who.int/mediacentre/factsheets/fs117/en/

# Leaderboard position

Figure – Rank as of March 17, 2018, 12:30 PDT



# Full accuracy and leaderboard metrics on all models

Table – Leaderboard MAE scores

|  |  |
| --- | --- |
| **MAE** | **Model** |
| 25.6659 | GLMNB4 + XBG3 |
| 25.7813 | STLF+GLM.NB2+GLM+XGB3+XGB4+XGB5 |
| 25.8774 | STLF+GLM.NB2+GLM+XGB3 |
| 26.0024 | STLF+XGB3+XGB4 Ensemble |
| 26.3582 | GLM NB4 |
| 26.4111 | GLM.NB 2 |
| 26.5025 | STLF+GLM.NB2+GLM+XGB3+poisson+nnetars3 |
| 27.0337 | XGB3 |
| 27.0793 | XGB5 |
| 27.1635 | STLF |
| 27.1923 | STLF GARCH |
| 27.4183 | XGB4 (XGB3 using MAE instead of RMSE in the metrics) |
| 27.6418 | XGB grid |
| 28.7404 | GLM |
| 29.6779 | STLF+GLM.NB2+XGB5\_alt |
| 29.6779 | GLM.NB no outliers |
| 30.0986 | Combo 4 (Auto ARIMA IQ XREG1 and Nnetar SJ) |
| 30.1466 | Possion GLM |
| 30.9111 | NNETAR3 |
| 30.9952 | NNETARS no outliers |
| 31.9567 | NNETARS 12 nodes |
| 32.1995 | NNETAR 2 |
| 32.9135 | NNETAR |
| 33.1587 | Combo3 |
| 33.2716 | NNETAR XREG2 |
| 33.3708 | XGB1 |
| 33.7548 | ETS |
| 33.7933 | Auto ARIMA w/ xreg |
| 34.7788 | NNETAR XREG |
| 35.1827 | Nnetar Corr |
| 36.0913 | Auto ARIMA |

## Full set of test metrics

Table – Test data metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ME | RMSE | MAE | MASE | ACF1 | Model |
| -13.2911 | 32.32109 | 27.95881 | 0.702923 | 0.950476 | ETS ZZZ San Juan |
| 9.28497 | 14.73154 | 9.316572 | 1.011036 | 0.842687 | ETS ZZZ Iquitos |
| -0.78605 | 39.70175 | 29.6286 | 0.744904 | 0.950399 | STLF SJ |
| 8.155448 | 14.37328 | 8.889227 | 0.964661 | 0.808526 | STLF IQ |
| -0.78605 | 39.70175 | 29.6286 | 0.744904 | 0.950399 | Forecast STLF SJ |
| 8.155448 | 14.37328 | 8.889227 | 0.964661 | 0.808526 | Forecast STLF IQ |
| -4.8136 | 32.95337 | 25.54176 | 0.642155 | 0.933269 | Auto ARIMA SJ |
| 9.108426 | 14.58818 | 9.160504 | 0.9941 | 0.844296 | Auto ARIMA IQ |
| -10.2429 | 34.33527 | 28.64779 | 0.720245 | 0.934386 | Auto ARIMA SJ XREG1 |
| 2.185026 | 11.43606 | 6.998169 | 0.759443 | 0.84389 | Auto ARIMA IQ XREG1 |
| -10.2429 | 34.33527 | 28.64779 | 0.720245 | 0.934386 | Auto ARIMA SJ XREG2 |
| 2.185026 | 11.43606 | 6.998169 | 0.759443 | 0.84389 | Auto ARIMA IQ XREG2 |
| -10.2429 | 34.33527 | 28.64779 | 0.720245 | 0.934386 | Auto ARIMA SJ XREG2 |
| 2.185026 | 11.43606 | 6.998169 | 0.759443 | 0.84389 | Auto ARIMA IQ XREG2 |
| -10.2429 | 34.33527 | 28.64779 | 0.720245 | 0.934386 | Auto ARIMA SJ XREG3 |
| 2.185026 | 11.43606 | 6.998169 | 0.759443 | 0.84389 | Auto ARIMA IQ XREG3 |
| -10.2429 | 34.33527 | 28.64779 | 0.720245 | 0.934386 | Hand ARIMA SJ |
| 2.185026 | 11.43606 | 6.998169 | 0.759443 | 0.84389 | Hand ARIMA IQ |
| -10.2044 | 34.32202 | 28.62416 | 0.719651 | 0.934393 | Hand ARIMA SJ 4 |
| 2.331985 | 11.47404 | 6.967502 | 0.756115 | 0.844253 | Hand ARIMA IQ 4 |
| -10.4663 | 34.34471 | 28.74204 | 0.722614 | 0.934222 | Hand ARIMA SJ 5 |
| 2.393724 | 11.5326 | 7.017732 | 0.761566 | 0.838941 | Hand ARIMA IQ 5 |
| -8.61022 | 39.49491 | 31.44546 | NA | NA | GLM.NB SJ |
| 3.650161 | 12.40988 | 7.239657 | NA | NA | GLM.NB IQ |
| -9.60129 | 36.37813 | 29.7793 | NA | NA | GLM.NB short SJ |
| 3.296716 | 11.91026 | 7.178016 | NA | NA | GLM.NB short IQ |
| 24.12232 | 39.53228 | 24.12983 | NA | NA | GLM.NB diffed SJ |
| 9.41885 | 14.78403 | 9.421243 | NA | NA | GLM.NB diffed IQ |
| -24.3869 | 49.50258 | 42.59011 | NA | NA | GLM all SJ |
| 4.90431 | 12.88343 | 7.465889 | NA | NA | GLM all IQ |
| -14.9223 | 43.82736 | 36.74467 | NA | NA | GLM reduced SJ |
| 3.289415 | 12.12602 | 7.35114 | NA | NA | GLM reduced IQ |
| -10.6806 | 55.31922 | 42.6855 | NA | NA | XGB3 SJ |
| 7.071686 | 13.41154 | 7.75637 | NA | NA | XGB3 IQ |
| 1.763477 | 30.89987 | 20.15828 | 0.506807 | 0.926636 | Nnetar SJ |
| 5.071104 | 12.5978 | 7.192274 | 0.780507 | 0.849972 | Nnetar IQ |
| -3.80544 | 32.56291 | 24.79345 | 0.623341 | 0.93259 | Nnetar3 lagged SJ |
| 4.270299 | 12.23661 | 7.088239 | 0.769217 | 0.847601 | Nnetar3 lagged IQ |
| -10.4454 | 13.36141 | 10.44539 | 1.307596 | NA | Nnetar Scaled SJ |
| -14.0948 | 18.39066 | 14.09479 | 3.610659 | NA | Nnetar Scaled IQ |
| 11.87902 | 37.79827 | 23.25696 | 0.584712 | 0.940191 | Nnetar Xreg SJ |
| 0.631555 | 12.66948 | 7.356229 | 0.798299 | 0.85229 | Nnetar Xreg IQ |
| 14.74998 | 38.64309 | 23.47824 | 0.590275 | 0.938541 | Nnetar Xreg2 SJ |
| -0.01209 | 13.62117 | 8.196108 | 0.889443 | 0.858959 | Nnetar Xreg2 IQ |
| 24.5979 | 42.10469 | 25.6067 | 2.003802 | 0.93932 | Nnetar4 diff SJ |
| 9.281995 | 14.7728 | 9.72498 | 1.487302 | 0.797482 | Nnetar4 diff IQ |
| -16.5639 | 20.93472 | 19.23887 | 1.030443 | 0.752281 | Nnetar no outliers SJ |
| 5.421526 | 12.6907 | 8.071704 | 0.976351 | 0.754754 | Nnetar no outliers IQ |
| -39.1573 | 53.28982 | 49.46264 | NA | NA | XGB5 alt param SJ |
| 4.413784 | 12.1153 | 7.21722 | NA | NA | XGB5 alt param IQ |
| -19.2956 | 40.5066 | 34.88123 | NA | NA | XGB grid param SJ |
| 1.072648 | 11.30702 | 7.419892 | NA | NA | XGB grid param IQ |
| -24.2811 | 39.13476 | 24.28115 | *NA* | *NA* | NeuralNet SJ |
| 1.072648 | 11.30702 | 7.419892 | NA | NA | NeuralNet IQ |
| -9.39171 | 37.57017 | 30.42294 | NA | NA | GLM.NB 4 SJ |
| 3.17879 | 12.28525 | 7.433272 | NA | NA | GLM.NB 4 IQ |