Phillip Jauregui
DengAI: Predicting Disease Spread Write-Up
Aug. 18, 2023

Problem

In this paper, I will discuss the construction and performance of 3 models (ETS, ARIMA, and neural network) used to predict the number of cases of Dengue in two South American cities: San Juan, Puerto Rico and Iquitos, Peru.

Background

Dengue is a viral infection caused by a virus transmitted to humans primarily by infected mosquitoes. Dengue is prevalent in tropical and subtropical regions, particularly within urban and semi-urban zones. Dengue's effects on humans range from mild flu-like symptoms to severe conditions like dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS). Mild cases include fever, headache, joint pain, and rash. Meanwhile, DHF involves severe bleeding and plasma leakage, and DSS can lead to organ failure. There is no specific antiviral cure for dengue, but the disease is manageable and most individuals recover with proper medical care (WHO, 2023)

Significance

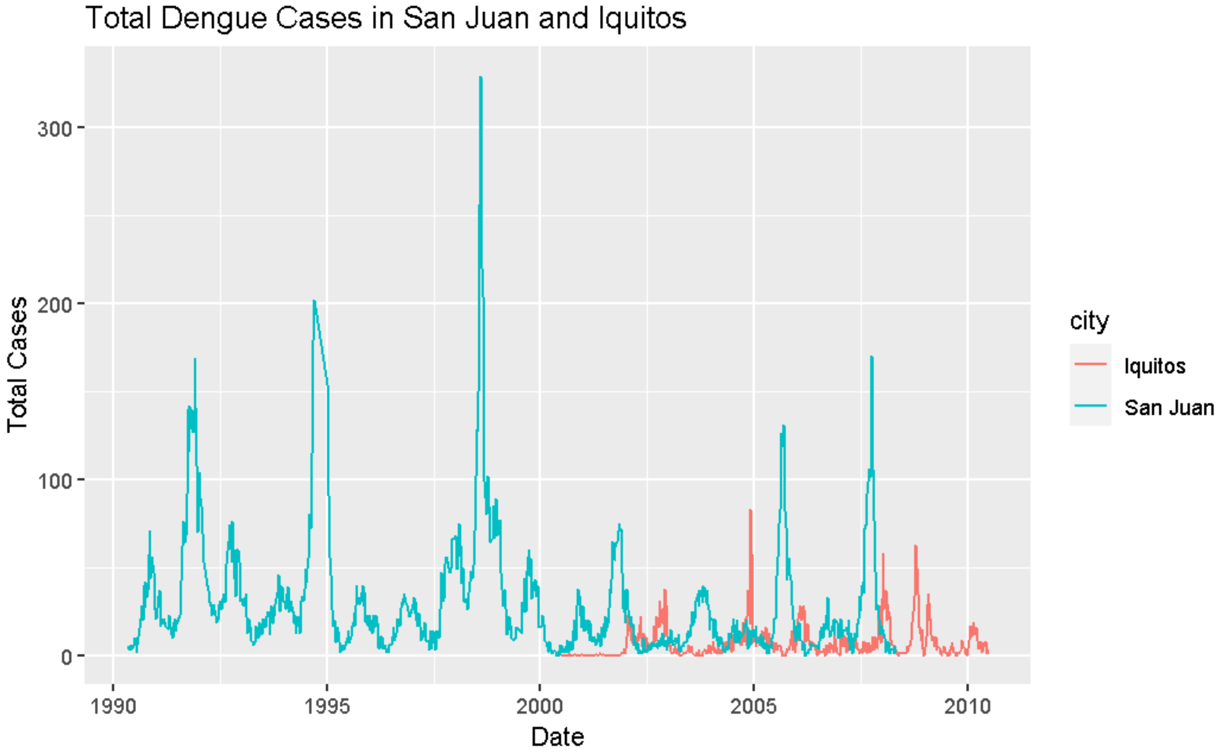
Changes in temperature and humidity and overall climate change can expand the range of mosquitoes as well as increase mosquito reproductive rates. Consequently, climate change poses a threat to public health via the proliferation of dengue by enabling mosquitoes to expand to new areas, particularly urban areas with inadequate sanitation and water management systems where stagnant water can accumulate. Indeed, dengue fever has increased recently. While it was historically concentrated in Southeast Asia and the Pacific islands, nearly half a billion cases now appear in Latin America (Lenharo, 2023). Accordingly, modeling the spread of dengue cases poses an important challenge for data scientists to aid global health (e.g., Patil & Pandya, 2021).

Introduction

Historical data for these two cities was provided, showing the weekly count of reported dengue cases spanning from 1990 to 2010. With this data, I made a training set to train 3 predictive models: ETS, ARIMA, and neural network.

Method

As a best practice, it is always helpful to first plot the data we are working with:



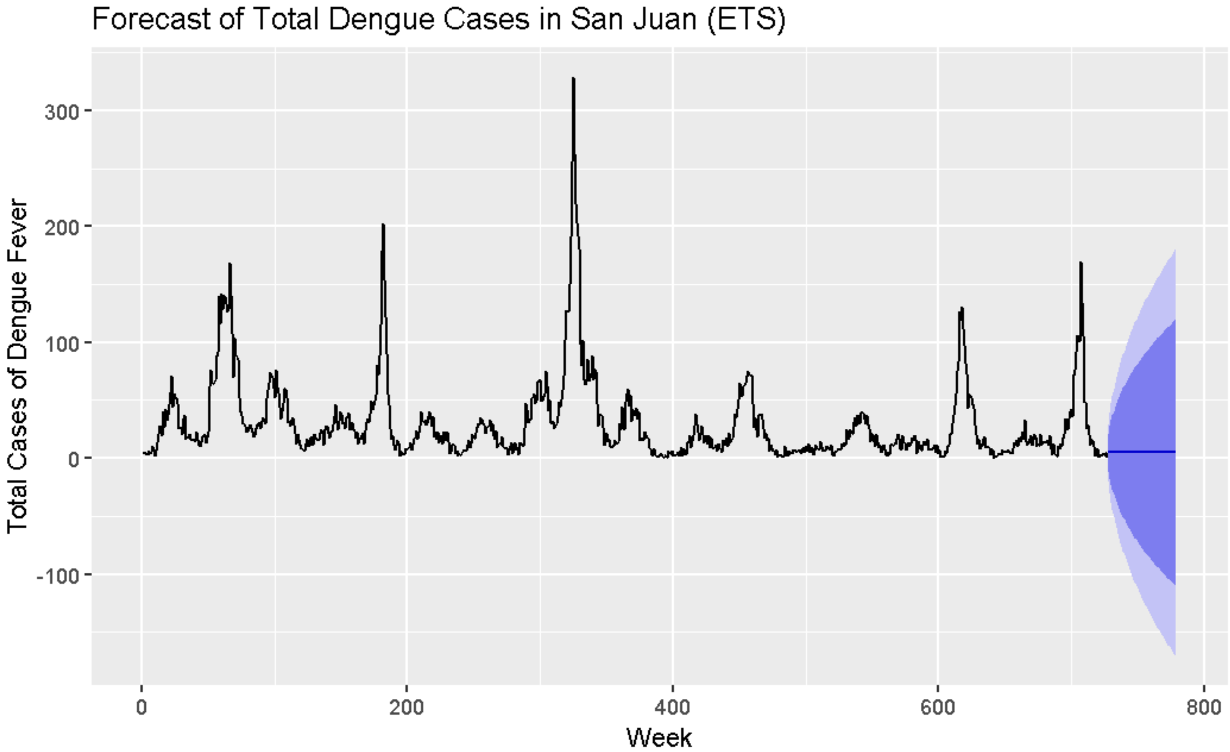
Here, we can see that San Juan has more dengue cases, despite Iquitos being more populous: San Juan boasted 326,953 residents in 2020, while Iquitos reported a population of 491,000 in 2023. Additionally, we can see some clear seasonality to the data indicating outbreaks of dengue at the peaks.

Results / Model Performance

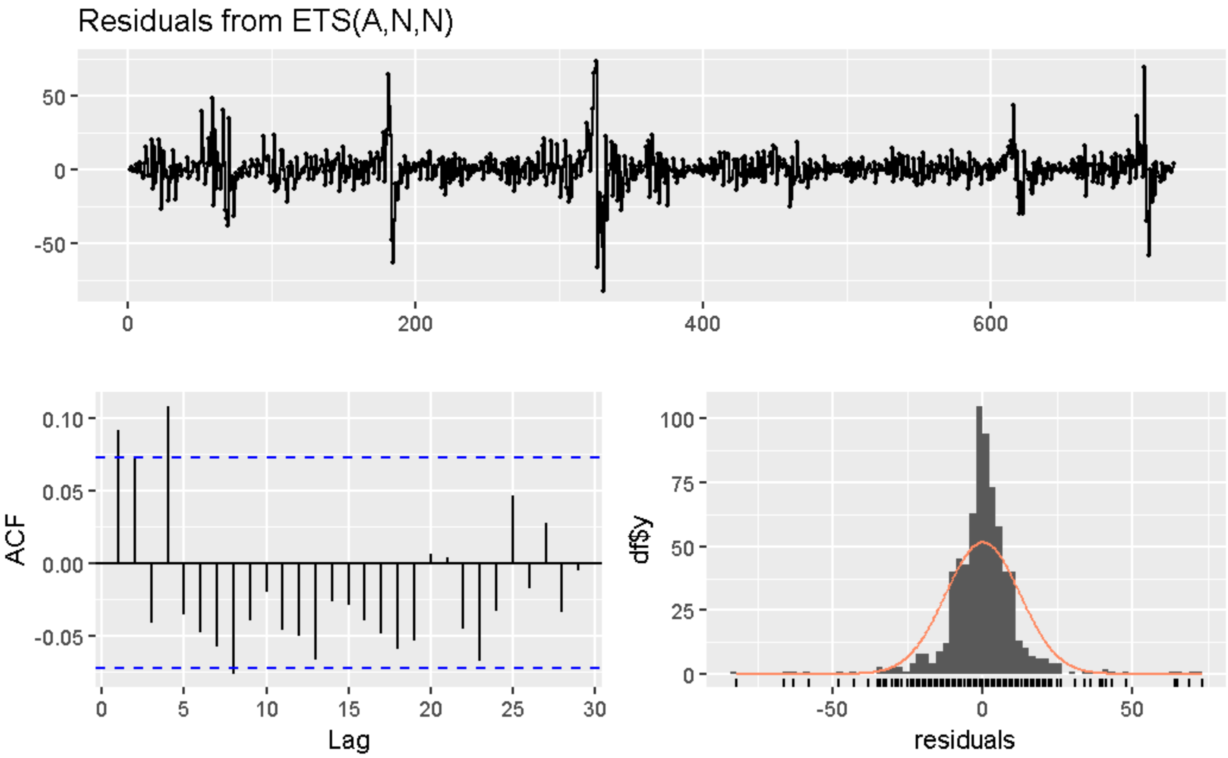
ETS

To forecast future dengue cases, I first constructed an ETS model using the “ets” function in R. Exponential Smoothing (ETS) forecasting is a method used to predict future values in a time series by incorporating the weighted average of past observations. It's particularly useful for time series data with patterns such as trend and seasonality. ETS models are based on the idea that future values are a combination of three components: error, trend, and seasonality. Error refers to the random noise or irregular fluctuations in the time series data that cannot be explained by the other components. Trend refers to the underlying long-term direction in the data. Lastly, seasonality regers to recurring fluctuations in the data at specific intervals, often due to external factors like time of year (e.g., ice cream sales going up in the summer and down in the winter; Jain & Mallick, 2017).

For all model methods, I produce 2 models (for 6 total [3x2]): one for San Juan and another for Iquitos.

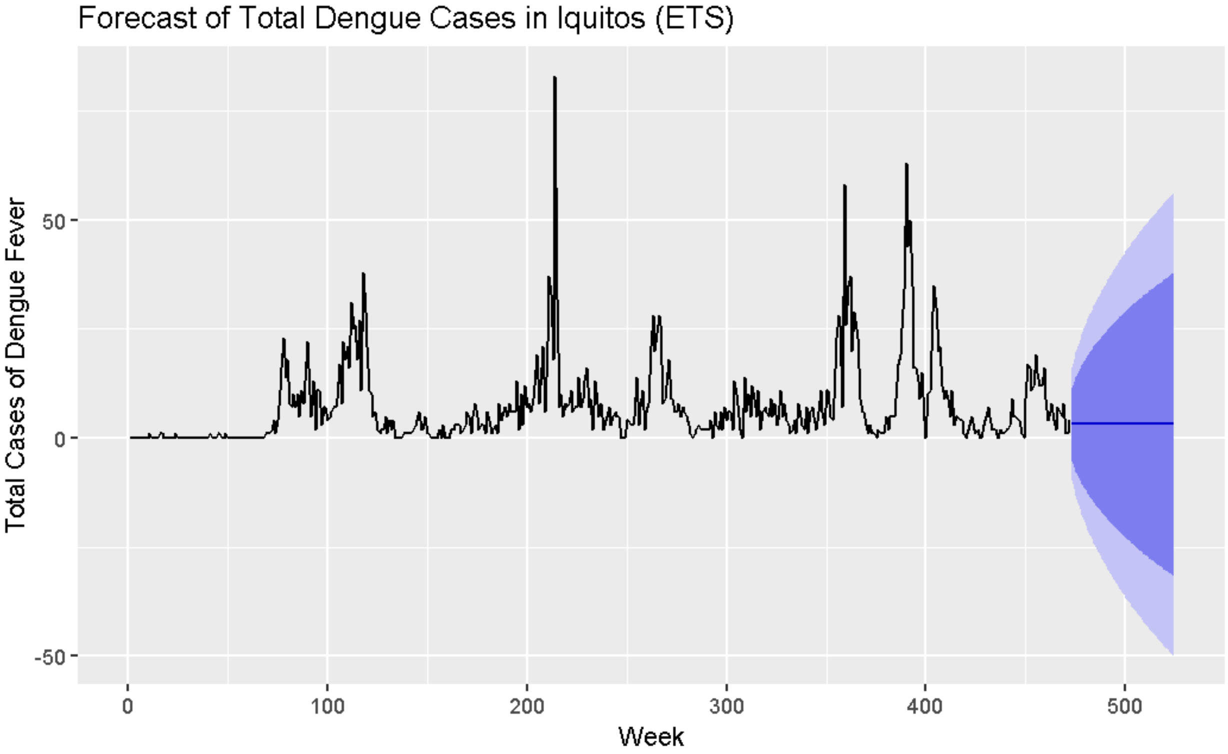


Looking at the ETS model for dengue in San Juan, we see that the model produces quite a low estimate for future dengue cases that seems more akin to a naive method of simply using the most recent value as a prediction for future values.

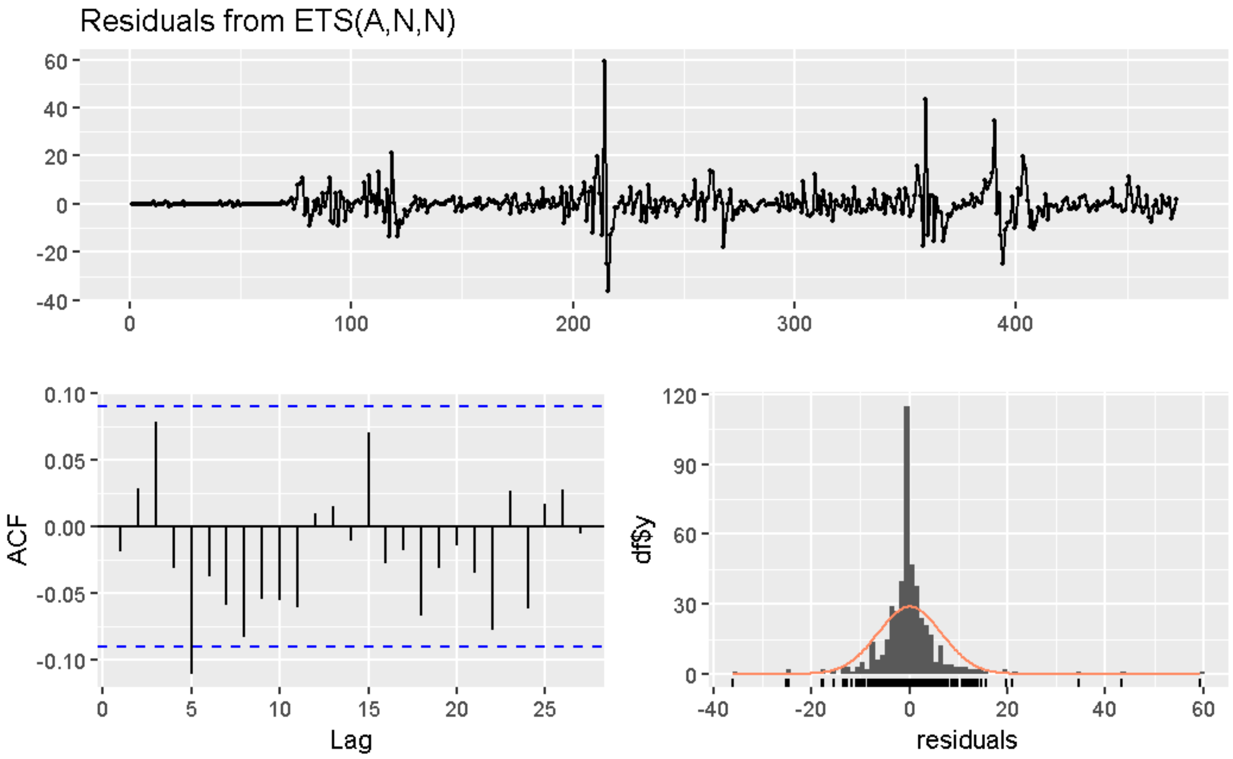


Examining the residuals of the model, we see a few significant autocorrelations but nothing out of the ordinary. As such moving forward with this model is suitable and no assumptions appear severely violated.

Looking at the ETS model for Iquitos, we see more of the same:

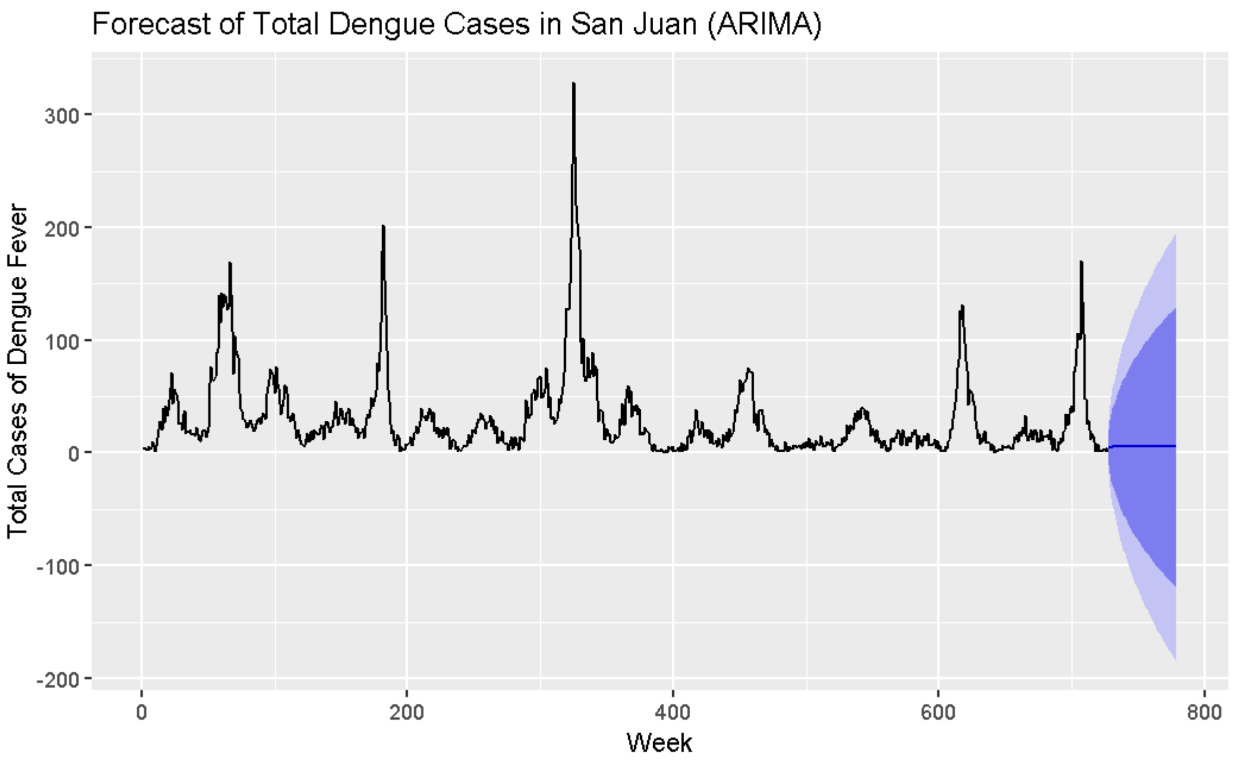


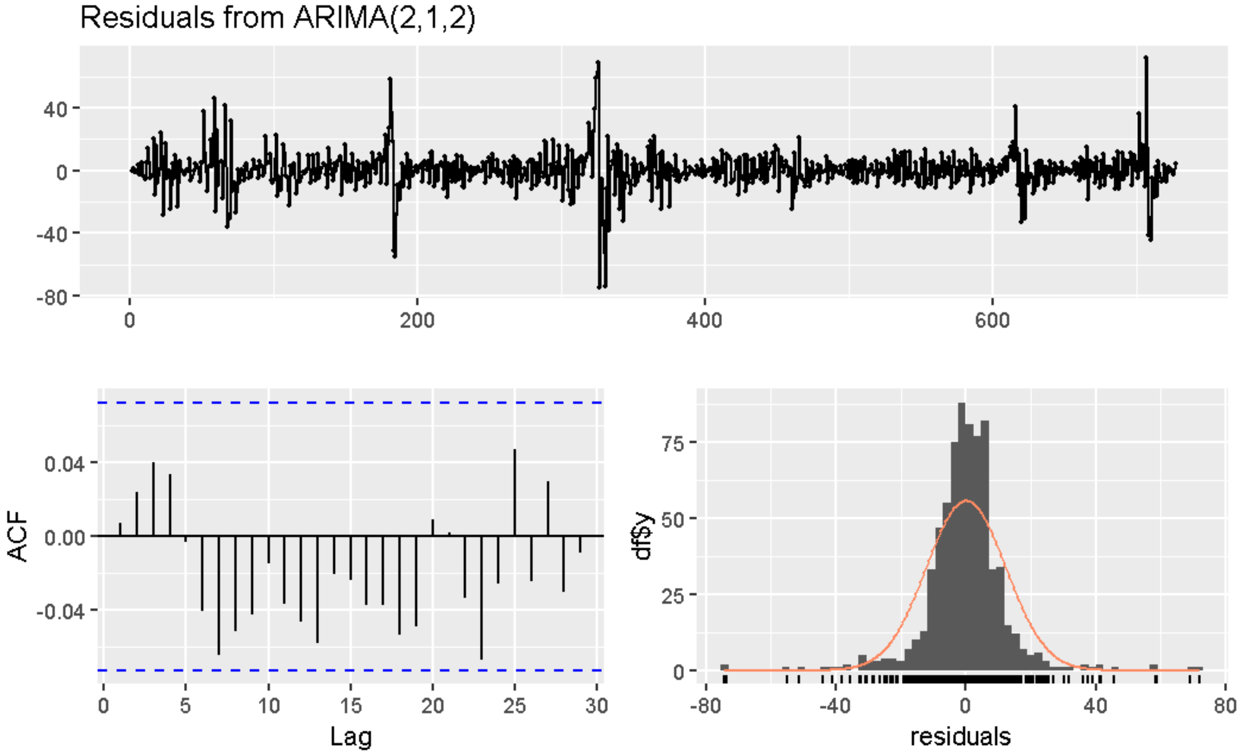
The model appears to be operating as a naive model, predicting the most recent value as the value for all future values. Despite the overall maximum number of cases being lower in Iquitos, the ETS model apparently specifies a higher rate of dengue cases in Iquitos. Also repeated, the residuals appear acceptable for the Iquitos ETS model:



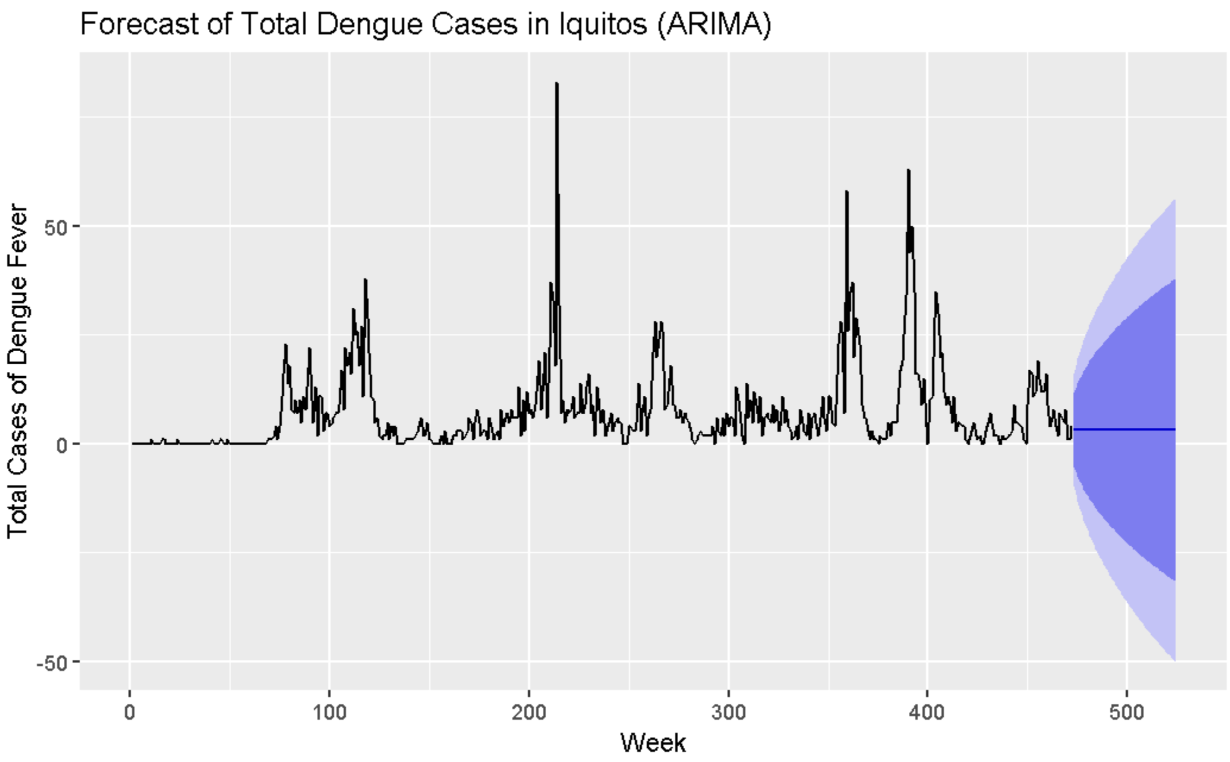
ARIMA

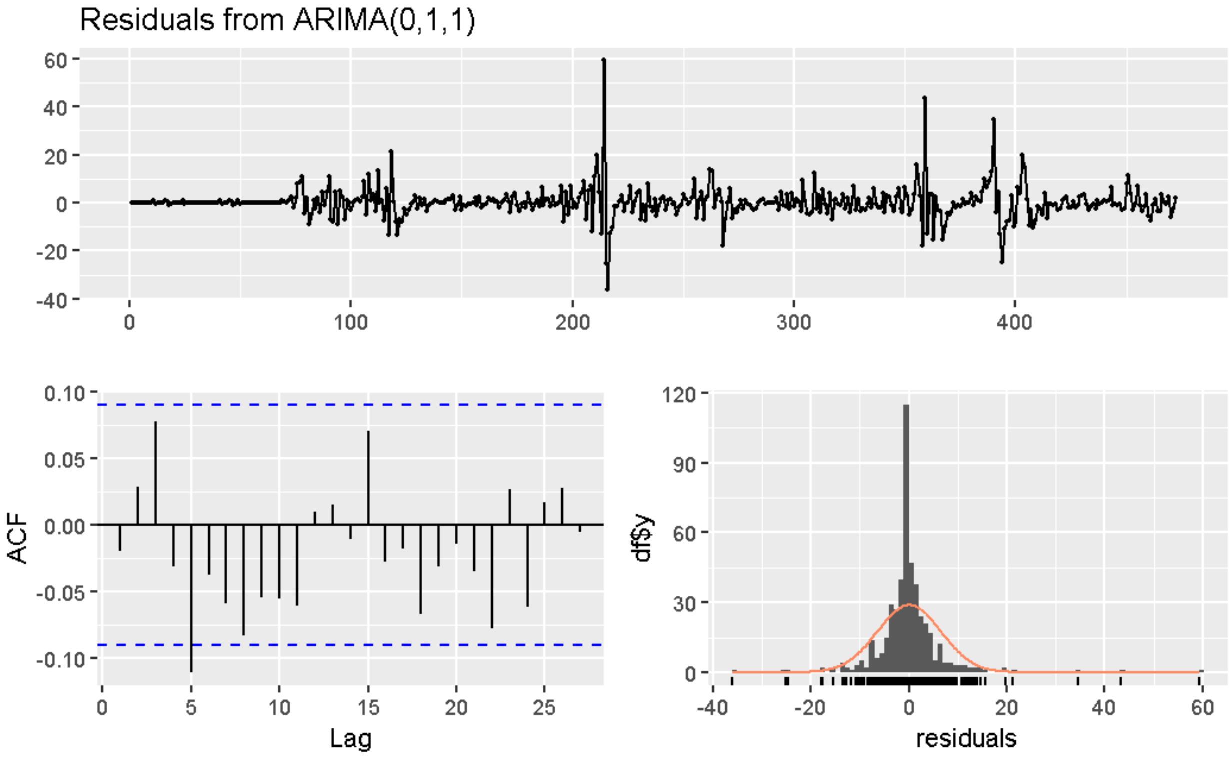
Second, using the “auto.arima” function in R, I constructed ARIMA models to forecast the total number of dengue cases in the two cities. ARIMA or "AutoRegressive Integrated Moving Average," models predict future values based on past observations. The autoregressive component (AR) models the relationship between the current value and previous values in the time series, capturing the relationship between a data point and its own past values. The integrated component (I) involves differencing the time series to make it stationary wherein statistical properties of a time series, like mean and variance, remain constant over time. Lastly, the moving average component (MA) models the relationship between the current value and past forecast errors (i.e., residuals). It helps capture short-term fluctuations in the data that are not explained by the auto-regressive and differencing components (Box et al., 2015).





Examining the plot of the ARIMA forecast we once again see a seemingly naive model that uses the most recent value as the estimate for all future values. This is true for both the San Juan and Iquitos ARIMA models:

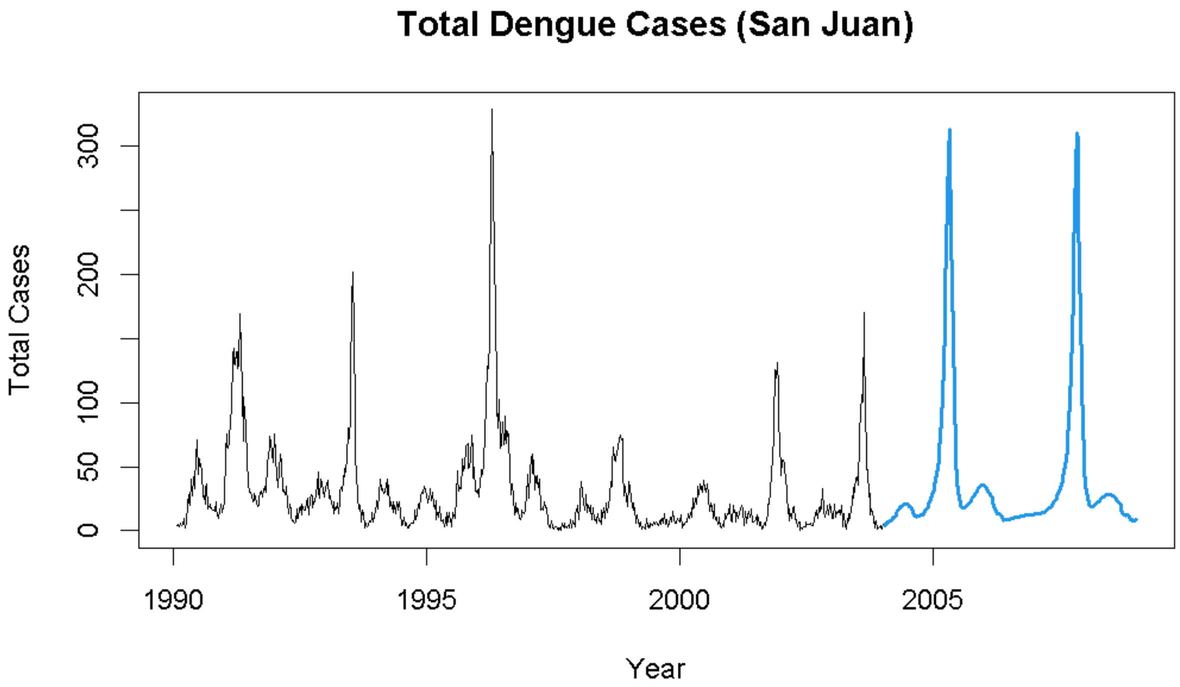


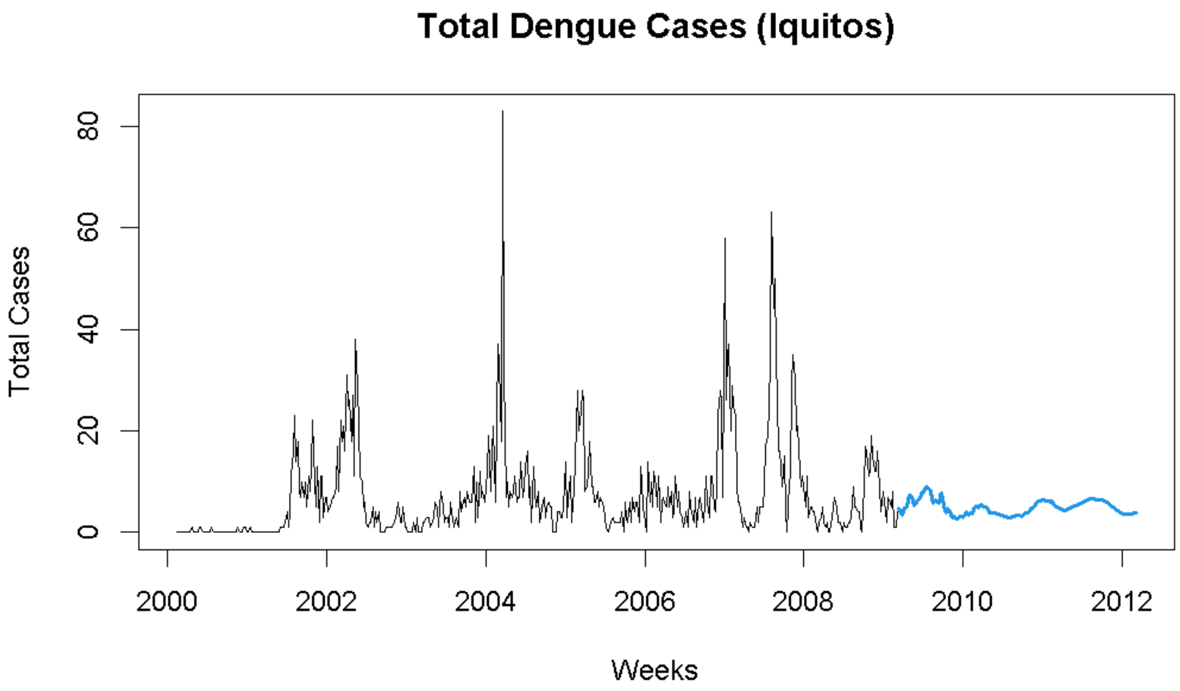


Neural Network

Third, using the “nnetar” function in R, I constructed neural network models to forecast the total number of dengue cases in the two cities. Neural network models are a class of machine learning algorithms, drawing inspiration from the connections between neurons in the human brain. These models are designed to tackle complex problems by discerning patterns and relationships from data (Zhang & Qi, 2005)

Using the “nnetar” function I generated forecast models for dengue cases in San Juan and Iquitos:





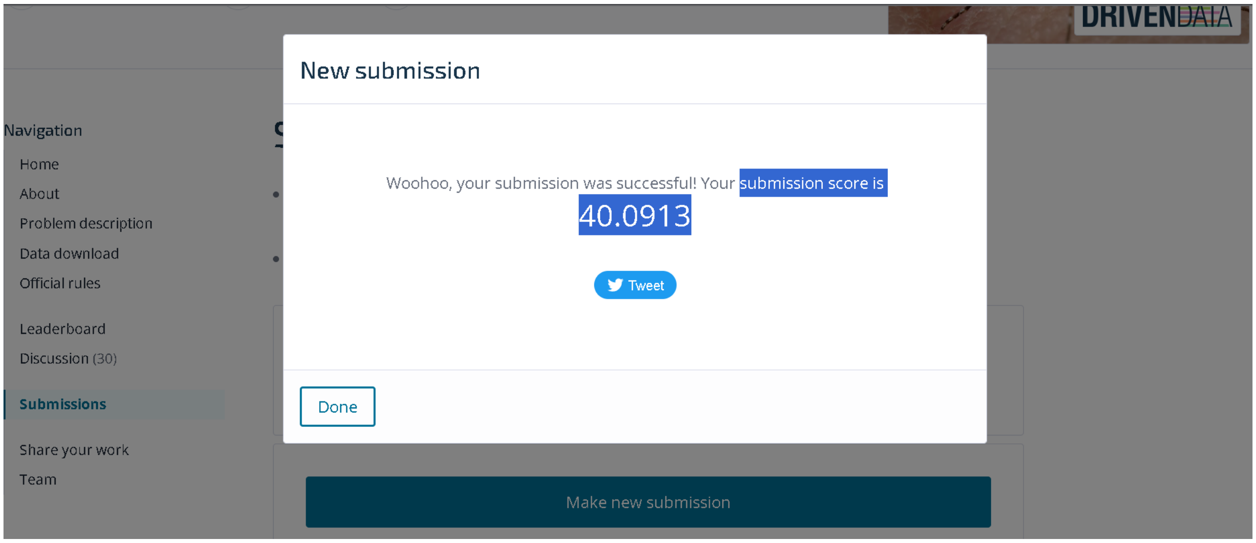
Here we can see the San Juan model follows the extreme peaks more closely while the Iquitos model smooths out its estimated predictions more so.

Evaluating the Models

Examining accuracy metrics for the models, we see that the neural network models outperformed the other models, displaying the lowest RMSE and MAE values (8.7 and 6.1 for San Juan and 3.36 and 2.44 for Iquitos).

	SJ			IQ		
	ETS	**ARIMA**	**Nnet**	**ETS**	**ARIMA**	**Nnet**
RMSE	12.44	12.26	8.7*	6.53	6.53	3.36*
MAE	7.75	7.88	6.1*	3.59	3.59	2.44*

Accordingly, I moved forward with the neural network model and used that forecasting model as my submission to DrivenData. My DrivenData submission score is 40.0913. A screenshot is included below for verification (user ID: pgjauregui):



New submission

Woohoo, your submission was successful! Your submission score is 40.0913

Conclusion

The neural network outperformed both an ETS and ARIMA model in forecasting the number of weekly dengue cases in the two cities, exhibiting an MAE of 6.1 and 2.44 for San Juan and Iquitos, respectively. There may be some reasons for this. For example, a neural network model can outperform ARIMA and ETS models when dealing with complex nonlinear patterns, changing data patterns, interactions between variables, and the ability to incorporate external information such as the environmental variables we were given in the “features” dataset. We see most if not all of these elements in the current dataset: it’s complex, nonlinear, and has changing data patterns.

Limitations

We must acknowledge the potential limitations of the current models. For example, neural networks can also be prone to overfitting and incur greater computational burden. We must bear these potential drawbacks in mind when selecting and constructing forecasting models. Some limitations of the ETS and ARIMA models are on display in the current project as you can specify ETS and ARIMA models in R but end up with naive forecasts which are not that helpful in predicting dengue cases. Moreover, ARIMA is not suitable for time series that lack stationarity, even after differencing.

Future Directions

Future work could examine additional cities to better nail down the nature of the spread of dengue. Perhaps the two cities in the current dataset are outliers regarding how the spread of dengue typically plays out. Additionally, future work could utilize different modeling techniques that properly leverage the environmental variables (e.g., ARIMA with external regressors).

References

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed). John Wiley & Sons.

Jain, G., & Mallick, B. (2017). A study of time series models ARIMA and ETS. *Available at SSRN 2898968*.

Lenharo M. (2023). Dengue is breaking records in the Americas - what's behind the surge?. *Nature*, 10.1038/d41586-023-02423-w. Advance online publication. https://doi.org/10.1038/d41586-023-02423-w

Patil, S., & Pandya, S. (2021). Forecasting dengue hotspots associated with variation in meteorological parameters using regression and time series models. *Frontiers in Public Health*, *9*, 798034.

World Health Organization: WHO. (2023). Dengue and severe dengue. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#:~:text=Dengue%20(break%2Dbone%20fever),body%20aches%2C%20nausea%20and%20rash.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European journal of operational research*, *160*(2), 501-514.