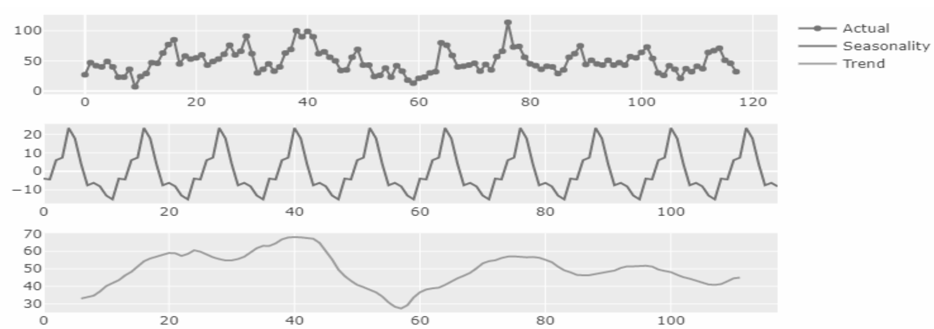
Typhoid Fever is a gastrointestinal infection caused by Salmonella Enterica Typhi bacteria. It is transmitted from person to person through the fecal-oral route where an infected or asymptomatic individual (who does not exhibit symptoms) with poor hand or body hygiene passes the infection to another person when handling food and water. The bacteria multiply in the intestinal tract and can spread to the bloodstream. Paratyphoid fever, a similar illness, is caused by Salmonella Enterica Paratyphi A, B, and C[1]. Typhoid fever is an important cause of avoidable mortality in regions without adequate access to safe water and sanitation. Although contaminated food and water, have been identified as the major risk factors for typhoid prevalence, a range of other factors have been reported in different endemic settings such as poor sanitation, close contact with typhoid cases or carriers, level of education, larger household size, closer location to water bodies, flooding, personal hygiene, poor lifestyle and traveling to endemic areas. In addition, climatic variables such as rainfall, Vapour pressure, and humidity have an important effect on the transmission and distribution of typhoid infections in human populations[2].

Disease Incidence and Climate Interactions

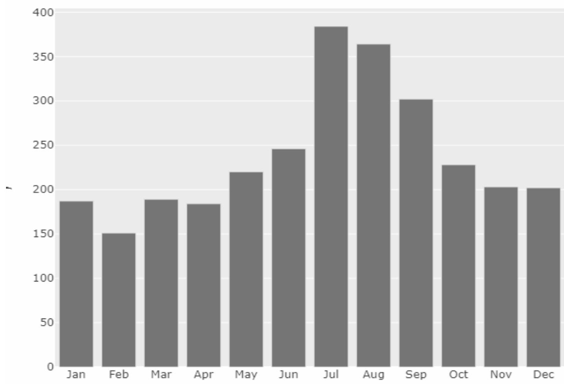
Though typhoid depends on a myriad of other factors, we are going to focus on the climate changes that affect the transmission of the bacterias that spread Typhoid. We are using the meteorological data taken from MOSDAC Meteorological and Oceanographic Satellite Data Archival Centre[14] is a data repository for the mission of the ISRO. We collected Typhoid cases from the site of Surat Municipal Corporation(SMC)[13]. Our meteorological data, including typhoid case data, is month-wise distributed. Let us start by analyzing the typhoid graph.

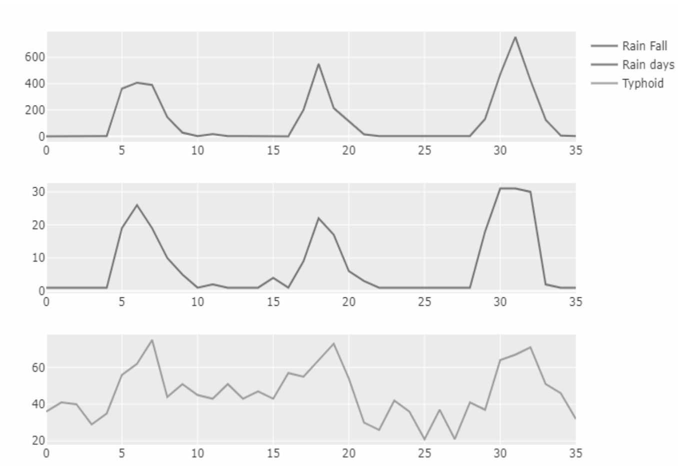
Here, you can depict that a significant pattern can be seen from around 2005. Before that, there was no obvious pattern to train our model on. Let us decompose our data into trend, seasonality, and residuals. We will take data from 2010 onwards.



We can say that our trend is not constant but, we have serious seasonality explaining ¼ of the actual time series. Which is a positive sign because we have a significant seasonal pattern in our time series. Time series with significant seasonality are easier to model because it follows the same pattern as prior records. We still have to worry about the volatile trend, but we can deal with it later. Now that we have seen that our time series has significant seasonality, we can plot the monthly sum of the last 5 years to check seasonality throughout the year.

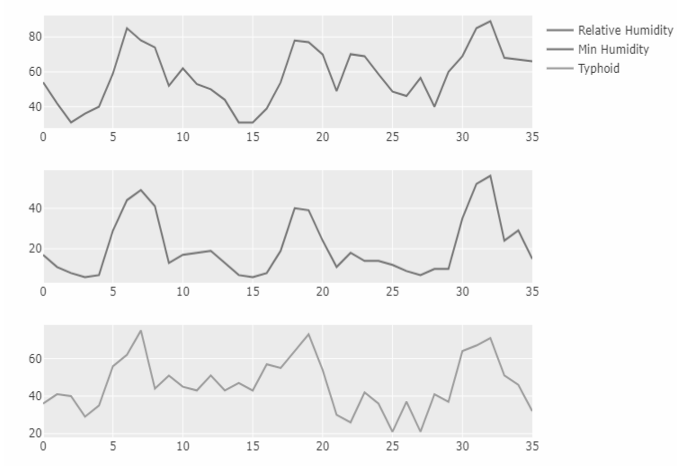
It can be seen from the figure beside that most cases are registered in the monsoon season. We can say that apart from the monsoon season, the graph is pretty balanced. Therefore, we can categorize annual typhoid data into two distinct sections. We will consider June to September as one section and other months as another. Now that we have seen the annual distribution of typhoid cases, we can start analyzing the change in climate affecting Typhoid cases reported.

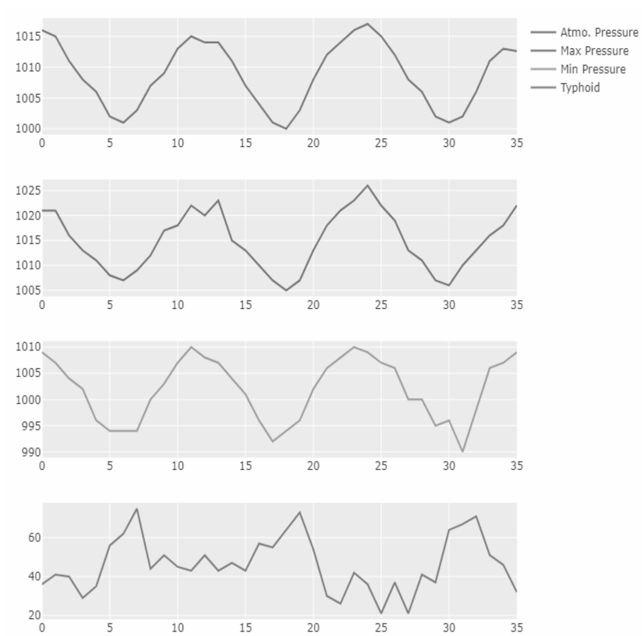




In figure beside, Typhoid cases of the last 3 years along with rainfalls and rainy days can be seen. We can depict that rainfall and rain days have significant associations during the monsoon season(Jun-Sep). Furthermore, it can be said that a lag of 1 month in rainfall and rain days is more correlated with our time series compared to the same month's rainfall and rain days. Though we require daily data to further analyze that supposition as that lag could be of three weeks or possibly 2 weeks as well. It should be added that the incubation period for typhoid is around 1-2 weeks[3]. We also analyzed that typhoid cases registered post-monsoon have little to no association with rainfall registered that year. There does not seem any variation in the number of cases registered annually with the change in annual rainfall registered. This means we can use rainfall and rain days to predict seasonality but it won't be helpful to analyze the trend of our data. Monsoon-related flooding increases the risk of waterborne diseases, including typhoid fever and related diseases such as paratyphoid and invasive non-typhoidal Salmonella disease. Clearly, the monsoon season creates the ideal climate for the transmission of typhoid: plentiful waters facilitate the growth of typhoid bacteria; intense rains can damage water and sewage lines, releasing sewage into the environment and contaminating water sources used for cleaning, cooking and drinking. The monsoon particularly exacerbates the risk of cross-contamination and causes open drains in the community to overflow with sewage, potentially contaminating otherwise clean water[4].

The figure beside shows the average relative humidity graph along with relative humidity and min humidity of the last 36 months. We can see that Typhoid cases are in apparent positive correlation with relative humidity. Though there have not been any significant studies that suggest how humidity has a direct effect on typhoid cases, one report can be found here[5]. Humidity is relatively higher during monsoon because the longer it rains, the air becomes more humid. That is why we can use relative humidity to train our model. Plus, humidity is lowest during winters, which helps because typhoid cases are relatively low during winter.





The figure beside shows the average atmospheric pressure, minimum pressure, maximum pressure, and typhoid cases of the last 36 months. It is evident that typhoid cases are negatively correlated with pressure. Rain comes down in varying intensities, so long, steady rain isn't always what you'll see. When a long, steady rain does happen, it's because of the location of the low-pressure system in relation to a warm front. Warm, moist air enters the area of low pressure and is pulled up and over the mass of cool air ahead of the warm front. This results in longer, steadier periods of rain. Average atmospheric pressure produces better results when we take a lag of 1 month. For which, we require daily data to analyze further.

Correlations are calculated between [-1,1]. A correlation of -1 indicates that data points are negatively correlated which means that if one variable increases, other decreases. A correlation of +1 indicates that data points are positively correlated which means that if one variable increases, another increase as well. A correlation value 0 indicates that there is no correlation between two variables[2]. A Pearson correlation between two variables X and Y is calculated by

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

$\overline{X}$ - mean of X
$\overline{Y}$ - mean of Y
n - number of variables

Pearson Correlation between Meteorological elements and Typhoid cases:					
	Relative Humidity(%)	Rainfall	Temperature	Atmospheric Pressure	Rain Days
Typhoid Cases	0.47	0.4	0.32	-0.53	0.49

Prediction

We are using the prophet package developed by Facebook. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. It provides us with the ability to make time-series predictions with good accuracy using simple intuitive parameters and has support for including the impact of custom seasonality. We use a decomposable time series model with three main model components: trend, seasonality, and holidays. They are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- g(t): piecewise linear or logistic growth curve for modeling non-periodic changes in time series
- s(t): periodic changes (e.g. weekly/yearly seasonality)
- h(t): effects of holidays (user-provided) with irregular schedules
- ε_t: error term accounts for any unusual changes not accommodated by the model

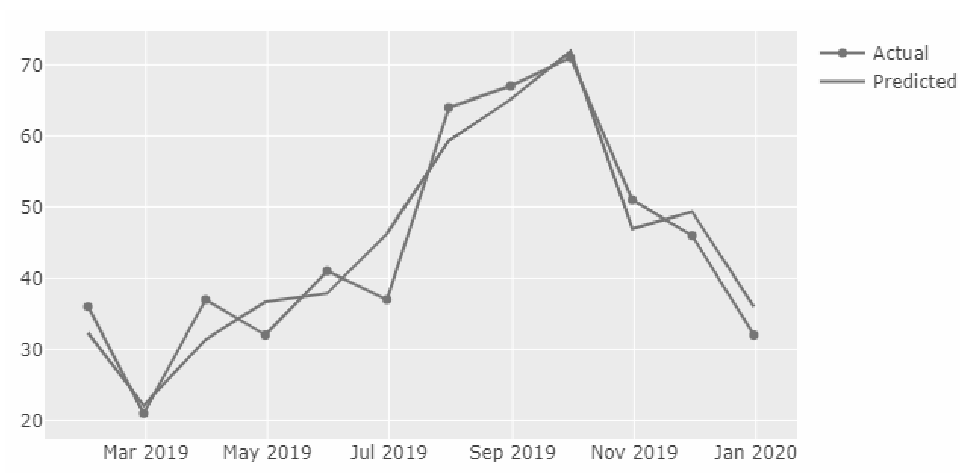
Using time as a regressor, Prophet is trying to fit several linear and nonlinear functions of time as components. The reason why we are using the prophet package and not SARIMA as we did in Malaria prediction is because the prophet allows us to configure each external regressor we use such as rainfall, rain days, etc. It allows us to configure how each regressor will affect our time series by setting their prior scales[6]. We are adding a yearly seasonality to our model and also specifying the Fourier order for added seasonality. Fourier order specifies how well we want our model to fit with our time series. A Fourier series is an expansion of a periodic function. in terms of an infinite sum of sines and cosines. Fourier series makes use of the orthogonality relationships of the sine and cosine functions[7]. We need to be careful while choosing the Fourier order because we want to avoid overfitting or underfitting our model[8]. We can also add more parameters to make our model more accurate in which we categorize our time series into two parts: Months with the monsoon season and other months. These added regressors should be in True or False form. We can also specify each month-wise parameter and their Fourier order as well. After we add these regressors, our dataset has large dimensions. Prophet traditionally uses Monte Carlo simulation to calculate the uncertainty interval[9]. When we predict a value of something using Prophet we get not only the estimated value but also the lower and upper bound of the uncertainty interval. The problem with Monte Carlo is it does not perform well with high dimension dataset. We are using Markov Chain Monte Carlo(MCMC)[10] samples to train and predict. We are using Bayesian inference using MCMC samples which takes a long time to run but provides a better result in our case. Here are the actual vs values our model predicted.

After the prediction, now we have to check the accuracy of our model. We are using the Mean Absolute Percentage Error(MAPE). Mean Absolute Percentage Error[10] is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

A_t = Actual Value
F_t = Forecasted Value

We got the MAPE of 9.7594 which means that our model was wrong by9.75% on average which is quite an impressive result considering we are working on a relatively small time series and each



value causes a significant change in our results.

Conclusion

Though we were able to achieve a pretty good result, we still haven't considered all factors that affect Typhoid like level of education, personal hygiene, and poor lifestyle. We have only done our analysis of Surat. We were not able to get the disease data from any other cities. Our research would have been more accurate if we had more data available for other cities as well. Though we might need more accurate meteorological data from the Ahmedabad Meteorological department under which the Surat Meteorological department comes. We had In Situ data from MOSDAC but some data was still missing so we had to do some preprocessing.

Primary strategies for typhoid fever control involve safe water supply, adequate sanitation facilities, and proper hygienic practices. However, these require sustainable investments, huge financial outlays, and long-term political commitment. The introduction of the WASH program has helped to improve the WASH situation in India although there still remains a significant gap in achievements. To make the best use of the control measures in such resource-poor endemic settings, there is a need to develop disease burden extrapolation models to choose the sites that need to be prioritized for routine intervention.

There might be shortcomings or errors of fact or interpretation in our report but we hope that it helps the government in handling possible epidemic in the future.

References

1. IAMAT - India recommended vaccination - Typhoid fever[link]
2. Ashraf M. Dewan, Rober Corner, Masahiro Hashizume, Emmanuel T. Ongee (2013)- Typhoid Fever and Its Association with Environmental Factors in the Dhaka Metropolitan Area of Bangladesh: A Spatial and Time-Series Approach[link]
3. HealthVic - Typhoid and paratyphoid[link]
4. Dipika Sur, Mohammad Ali - Comparisons of predictors for typhoid and paratyphoid fever in Kolkata, India[link]
5. Times of India - Cases of typhoid on the rise as weather gets humid[link]
6. Prophet - Seasonality, Holiday Effects, And Regressors[link]
7. Mathworld - Fourier Series[link]
8. Machine learning mastery - Overfitting and Underfitting With Machine Learning Algorithms[link]
9. Bartosz Mikulski - Understanding uncertainty intervals generated by Prophet[link]
10. Machine Learning Mastery - A Gentle Introduction to Markov Chain Monte Carlo for Probability[link]