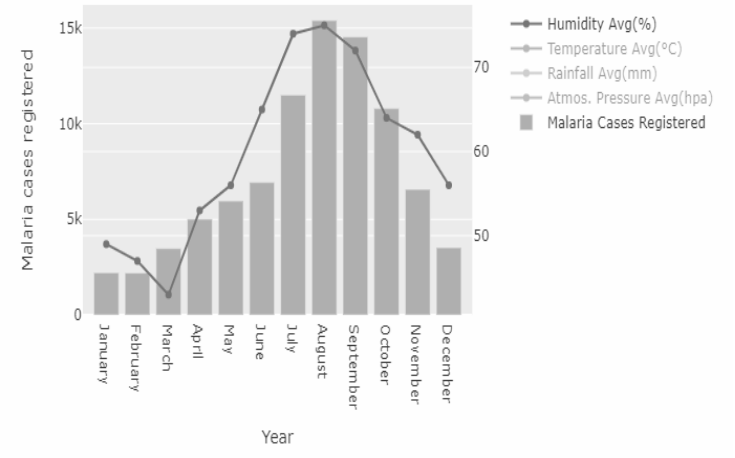
The rising burden of disease counts as one of the most prominent concerns of a warming climate. These risks are particularly severe in rapidly growing cities. Surat is located on the banks of River Tapi that has temperature and humidity patterns that can be described as ideal mosquito genetic conditions. Surat has a long history of river floods and usual water logging during the peak rainy season. It makes Surat prone to endemic vector-borne diseases and morbidity. In the past, most cases in Gujarat were reported in Surat but due to the preventive actions taken by SMC, that number has started to deteriorate. This decline has been reported despite an increase in population over time.

Climate change causes a possible increase in relative humidity and rainfall would increase Malaria risk in the city. We tried to develop an urban climate impact assessment model with public health as our focal point. We are using past data of the number of Malaria cases registered and meteorological data(rainfall, relative humidity) to predict Malaria risk. This helps SMC health and hospital organizations to take preventive steps that can be beneficial from an economic point of view.

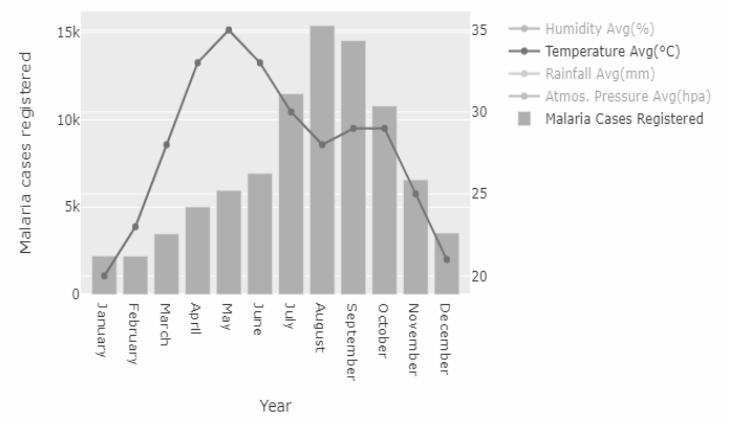
Disease Incidence and Climate Interactions

Since Mosquito breeding and their disease transmission efficiency is highly influenced by climate patterns like rainfall, temperature, and relative humidity, most vector-borne cases have a seasonal trend. Thus, predicting the future spread of disease provides an opportunity for health officials to be prepared for a possible outbreak. According to the annual report by SMC Vector-borne diseases control department[1], relative humidity above 60%, mean temperature around 25-30° along with continuous and dense rainfall is an ideal climate for the rise in Malaria
Cases.

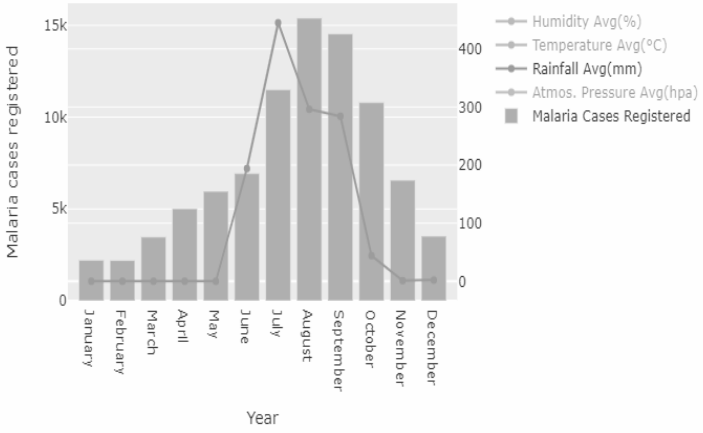
We can see that with the increase in Humidity, the significant increase in Malaria cases occurs. Rainfall follows almost the same seasonal pattern as relative humidity. We have taken all cases registered from 2010 to 2019 and added monthly cases together. For relative humidity data, we have taken the average. The highest increase is between July-October and humidity is greater than 60% in those months.



Here, the decrease in air temperature plays a vital role in the increase in malaria cases. In Summer, when the temperature is at the highest, there is not a significant increase in malaria cases which suggests that air temperature is in the negative correlation of malaria cases reported. Though, change in malaria cases is continuous throughout the year, the relative decrease in temperature suggests that it contributes to the increase in malaria cases. As you see in the picture, the increase in malaria cases occurs during June-October and the temperature is between 25-30°.



Here, the increase in rainfall correlates with the increase in malaria cases[17]. Though rainfall around non-monsoon months is around zero, when it rains, the cases are increased as there are potholes and spots with still water in which Anopheles mosquito lays eggs. Heavy rainfall can have a diverse range of effects on disease. In tropical and subtropical regions with crowding and poverty, heavy rainfall and flooding may trigger behavioral changes such as an increase in the morbidity of malaria.



Pearson Correlation between Meteorological elements and Malaria cases:				
	Relative Humidity(%)	Rainfall	Temperature	Atmospheric Pressure
Malaria Cases	0.23	0.5	0.62	-0.75

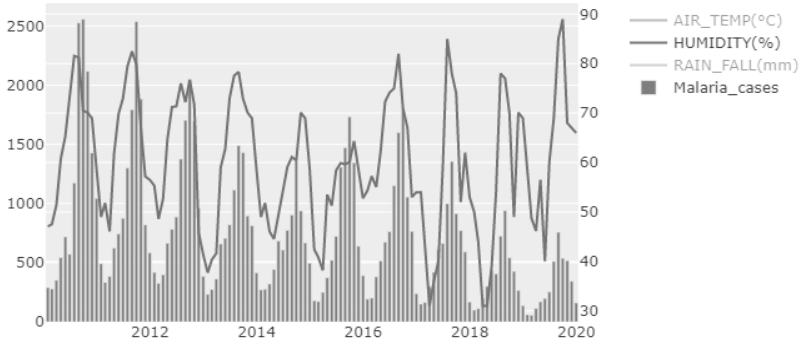
Correlations are calculated between [-1,1]. A correlation of -1 indicates that data points are negatively correlated which means that if one variable increases, other decreases. A correlation of +1 indicates that data points are positively correlated which means that if one variable increases, another 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

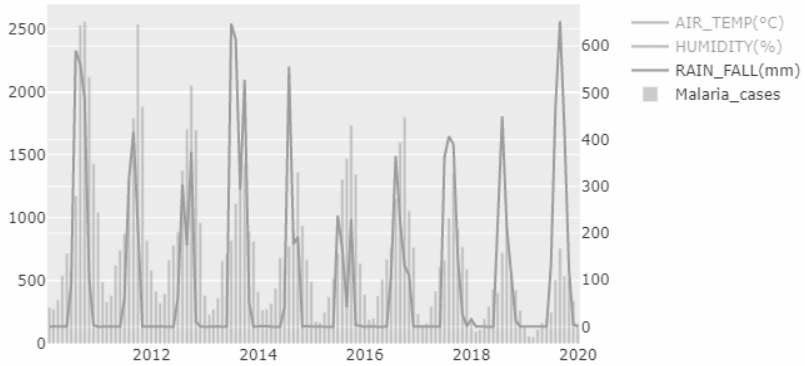
$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

nean of X
nean of Y
umber of variables

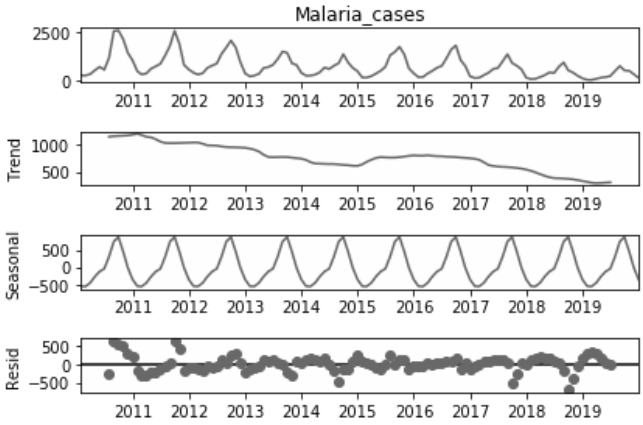
Prediction

We started predicting with conventional machine learning models and though the results were adequate, we were looking for models that could help us with the time series forecasting. Conventional machine learning models do not train models in a sequential manner. Thus, we needed models that could perform satisfactorily on sequential data. We decided to use the SARIMA model which is Seasonal Auto-regressive Integrated Moving Average[3].

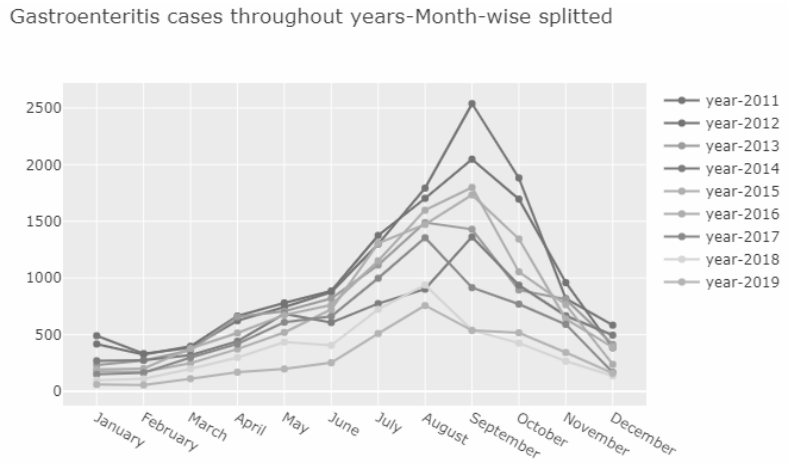




From the above Images, we can depict that a recent drop in the number of malaria cases is not especially related to Exogenous variables like Rainfall and Relative Humidity. There are no significant changes in either Humidity or Rainfall in the last few years that we can correlate with the recent changes in the trend of a number of cases registered. In Rainfall, the last year 2019 was an exceptional year but no changes in malaria cases occurred which was surprising considering the rainfall effect on malaria cases mentioned earlier in the report. Though Rainfall might affect malaria cases registered on a seasonal basis, it certainly is not helpful in predicting the trend of the graph. The same case can be made for Relative Humidity as well. The graph below can show you the decomposition of the time series which is malaria cases registered from 2010-2019.



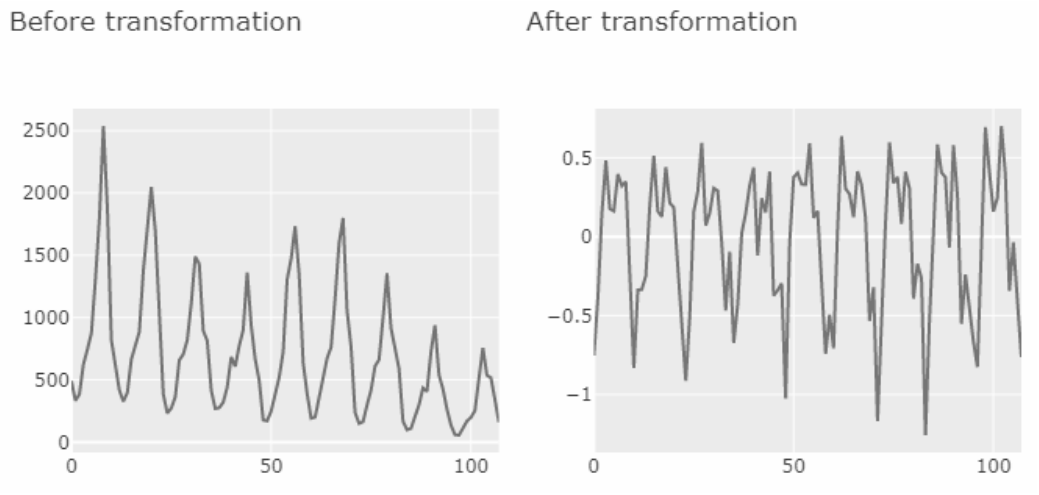
Time series decomposition involves thinking of time series as a combination of Trend, Seasonality, and Residual which is known as noise as well. Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting[4]. Here, our seasonal component is considerably high, which means that there are more similarities between the seasonal values of the time series. If we follow the trend graph, we can see that it is moving downward that infers that the number of malaria cases being recorded is decreasing due to the SMC's initiative. Starting from 2016, the graph has been in a downward trend.



If we analyze the graph beside, we can depict that graph is continuous through the start of the year to the end. Cases start to rise from January to September and start to decrease after September. Almost every line which represents a whole year suggests the same. Thus, we are using the SARIMA model without using any external variables. SARIMA is one of the most widely used forecasting methods for univariate time series data forecasting[5]. This method can handle the trend as well as the seasonality of the time series. Before we start training our model, we need to make sure that our data is stationary. There are a myriad of ways to check for time series stationarity but the easiest way is to check whether the mean and variance are constant over certain time periods. In our case, the time series we are dealing with does not have a constant mean over the years. One other way is through Unit test roots such as the Augmented Dickey-Fuller test[6]. There are numerous ways to make your time series stationary. We have achieved Stationarity through Log Transformation[7] of the time series and then, taking a difference of the time series[8]. The change can be seen in the table below.

Results of Dickey-Fuller Test:		
	Before applying any transformations	After applying log transformation and difference
Test Statistic	-0.830413	-3.556707
p-value	0.810025	0.006645
Number of Observations Used	106	106
Critical Value (1%)	-3.493602	-3.493602
Critical Value (5%)	-2.889217	-2.889217
Critical Value (10%)	-2.581533	-2.581533

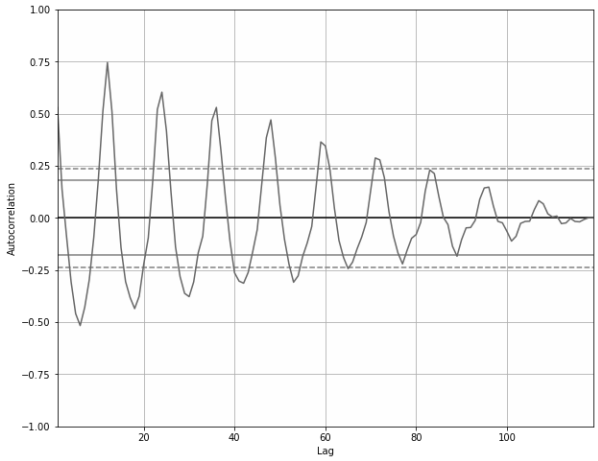
From the table above, we can analyze that Test Statistic in the latter, is smaller than Critical Value(1%) and the p-value is smaller than the significance level of 0.05, which fundamentally means that we are rejecting the null hypothesis with the confidence level of 99%. Consequently, we can say that our time series is now stationary. The difference can be seen in the picture below.

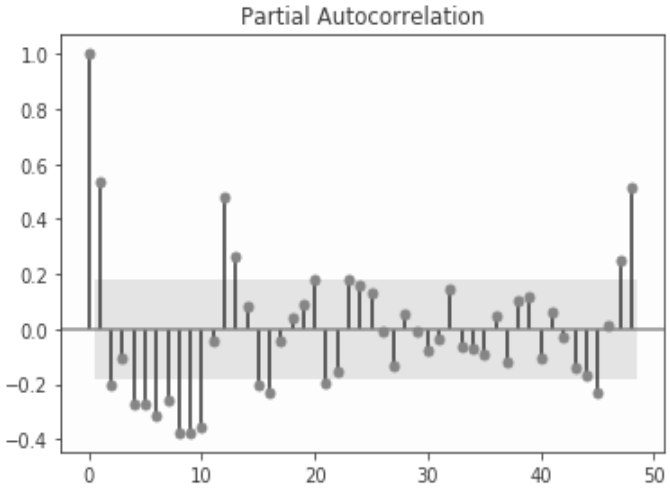


We can see in the figure above that the latter image has a constant mean and variance over different periods. Now, we can start predicting using the modified time series. Now, let us understand the SARIMA model letter by letter. SARIMA(p,d,q)(P,D,Q,s):

- AR(p) - Autoregression model i.e. regression of the time series onto itself. The basic assumption is that the current series values depend on its previous values with some lag. It can be determined from the PACF plot.
- MA(q) - Moving average model. Without going into detail, it helps the model to analyze the error of the time series. This can be determined using the ACF plot.
- I(d) - Order of Integration. This is a number of nonseasonal differences needed to make the time series stationary.
- S(s) - This is responsible for seasonality and it gives season period length of the given time series.
- P - Order of Autoregression for the seasonal component of the model. It can be determined from the PACF plot.
- Q - Same as P, but using ACF plot.
- D - Order of seasonal integration.

Configuring the SARIMA model requires selecting the above-mentioned hyperparameters for both trend and seasonal elements of the time series. These hyperparameters can be analyzed through ACF and PACF plots[9]. ACF stands for Autocorrelation function and PACF stands for Partial Autocorrelation function. These are plots that graphically summarize the strength of a relationship with an observation in a time series with observations at prior time steps.

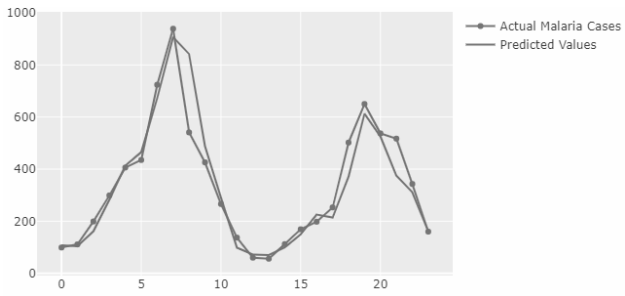




Let us analyze the hyperparameters from ACF and PACF graphs:

- p is probably 1 because it is the most significant lag in the PACF graph.
- q is 1 as well because it is the most significant lag before it starts decreasing in the ACF graph.
- d should be 1 because we are taking differencing once.
- s should be 12 as we have a monthly data of malaria cases reported.
- P should be 1 as we get the second significant lag at lag number 12 in the PACF graph.
- Q should be 3 or 4 as we get significant lags at lag number 12,24,36 and 48 in the ACF plot.
- D is 1 because we are taking differencing once.
- Thus, we get the SARIMA(1,1,1)(1,1,3,12).

Using these hyperparameters in the SARIMA model, we got quite an accurate result. A comparison between actual and predicted lines of the last two years can be seen in the figure beside. We kept the trend constant as our time series had no significant trend after its transformation. After the prediction, it is time to check the accuracy of our model. We are using two measures to check the prediction accuracy of our model:



1) 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. We got the MAPE of 14.0215 which means that our model was wrong by 14% 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.

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

A_t = Actual Value
F_t = Forecasted Value

2) AIC - Akaike Information Criterion[11] can be used to determine the quality of the model. It is an estimator of the out-of-sample prediction error; a lower prediction score indicates a more predictive model. We got the AIC score of -5.398 which is a relative measure which indicates that we have taken the accurate hyperparameters to train our model.

$$AIC = -2/N * LL + 2 * k/N$$

N- Number of examples in dataset
LL- Log-likelihood of the

Conclusions

In this report, we have only taken the weather effect into consideration. Supplementary data is required to comprehend the effect of malaria on the human immune system. There are few non-climatic factors that affect the Malaria transmission. The type of vector, the type of parasite, environmental development and urbanisation, population movement and migration, the level of immunity to malaria in the human hosts, insecticide resistance in mosquitoes, and drug resistance in parasites, all have a role in affecting the severity and incidence of malaria[18].

Malaria-free Gujarat 2022 campaign is the initiative started by the Health and family welfare department, Government of Gujarat[12]. Malaria is a major public health problem in Gujarat but is preventable and curable. Malaria interventions are highly cost-effective and demonstrate one of the highest returns on investment in public health. In regions where the disease is endemic, efforts to control and eliminate malaria are increasingly viewed as high-impact strategic investments that generate significant returns for public health, help to alleviate poverty, improve equity and contribute to overall development. Gujarat state has made significant achievements in malaria control as the state could keep the Annual Parasitic Incidence (API) less than 1.0 during the last three years. The lowest overall state API was recorded in 2015 and 2016 since 1961. Govt. of Gujarat is committed to making the state free from the burden of malaria. This commitment is reinforced by the National Framework for Malaria Elimination and also the target to be achieved in the health sector under Sustainable Development Goals. Gujarat State with good infrastructure and resources can take rapid strides in the plan to achieve malaria elimination by 2022.

We hope that our project helps the government in achieving its target. Our SARIMA model is capable of detecting a rise in the reported number of cases and can give pretty accurate forecasts that can help the government to take decisive precautionary actions in the future.

Reference

1. SMC Vector-borne diseases control Annual Report - 2018[link]
2. SPSS-tutorials - Pearson Correlations – Quick Introduction[link]
3. Statsmodels - SARIMA[link]
4. Machine learning mastery - How to Decompose Time Series Data into Trend and Seasonality[link]
5. TowardsDataScience - How to forecast sales with Python using the SARIMA model[link]
6. Statisticshowto - what is the Augmented Dickey-Fuller test?[link]
7. NCBI - Log Transformation[link].
8. Otexts - Stationarity and differencing[link]
9. Machine learning mastery - A Gentle Introduction to Autocorrelation and Partial Autocorrelation[link]
10. StatisticsHowto- Mean absolute percentage error (MAPE)[link]
11. Machine learning mastery - Probabilistic Model Selection with AIC[link]
12. Malaria-free Gujarat 2022[link]
13. SMC - Disease records[link]
14. MOSDAC - Providing the meteorological data[link]
15. Timeanddate - Providing with additional Meteorological data[link]
16. Healthline - About Malaria[link]
17. Researchgate - Characteristics and trends of malaria in Surat district of Gujarat: a hospital-based study[link]
18. Open.edu - Communicable Diseases Module: 6. Factors that Affect Malaria Transmission[link]