Time Series Prediction of Temperature in Pune Using Seasonal ARIMA Model

Aarati Gangshetty and Gurpreet Kaur

September 26, 2021

Time Series Prediction of Temperature in Pune using Seasonal ARIMA model

Aarati Gangshetty[1] AND Gurpreet Kaur[2]

[1]Department of Computer Engineering,JRF,R&DE(Engrs),Pune,India

aaratigangshetty7@gmail.com

[2]Scientist E,R&DE(Engrs),Pune,India

gurpreet.drdo@gmail.com

Abstract: In this paper, an attempt has been made to develop a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to predict temperature using past data of Pune, Maharashtra. The dataset from 2009 to 2020 has been taken for analysis. When trend and seasonality is present in a time series, instead of decomposing it manually to fit an ARIMA model, another very popular method is to use the seasonal autoregressive integrated moving average (SARIMA) model which is a generalization of an ARIMA model. Time series lately is becoming very popular, a reason for that is decreasing hardware's cost and capability of processing. The model can be used to calculate what Patterns should be in the coming year. Quantify the effects of sudden changes or disruptions in the system. The seasonal ARIMA model is implemented by running Python 3.7.4 on Jupyter Notebook and using the package matplotlib 3.2.1 for data visualization. The goodness of fit of the model was tested against standardized residuals, the autocorrelation function, and the partial autocorrelation function. We discover that SARIMA (1,1,1)(1,1,1)12 can represent very well in the data behavior. We obtained MAE of 0.60850 and RMSE of 0.76233 for SARIMA model. According to the model diagnostics, the model was reliable for predicting temperature.

***Keywords* —- SARIMA, Prediction, ARIMA, temperature.**

I. INTRODUCTION

The main goal of time series model is to collect and analyze past values to develop appropriate models that describe the inherent structure and characteristics of the series. On the other hand, time series forecasting model observes different values predict future values. Regression analysis often tests theories that the current data of one or more time series has the impact on the current data of another time series [1]. Time series data occurs in many areas like financial analysis, sensor monitoring of network, analysis of medical issues, and mining of social activity. More modern fields focus on the topic and refer to it as time series forecasting. Forecasting involves taking models fit on historical data and using them to predict future observations. Descriptive models can borrow for the future (i.e., to smooth or remove noise), they only seek to best describe the data. An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what has already happened.

Time series forecasting is the use of certain model to forecast future values based on past observed values, and thus can be understood as a method for predicting future values by understanding past values [2]. Unlike numerical weather prediction, time series forecasting uses a model to predict future values based on past values. Owing to the importance of time series forecasting in countless practical fields, researchers should pay proper attention to fitting an appropriate model to the time series. Over the year, many intelligent time series models have been developed in the literature to improve the accuracy and efficiency of time series forecasting. One of the most widely used and recognized statistical forecasting time series models is the Autoregressive Integrated Moving Average (ARIMA) model. The ARIMA model is well-known for notable forecasting accuracy and efficiency in representing various types of time series [3] with simplicity as well as the associated, Box–Jenkins's methodology for optimal model construction. The basic assumption made in implementing this model is to assume the time series is linear and follows a statistical distribution, such as the normal distribution. For seasonal time series forecasting Box Jenkins[4] proposed a quite successful variation of the ARIMA model called the Seasonal ARIMA (SARIMA) model. The primary objectives of this paper are as follows:

1) Plotting the data as a time series plot
2) Checking the data, if it has any trend or seasonality
3) Predicting values of SARIMA (p, d, q) (P, D, Q)s
4) Applying SARIMA (p, d, q) (P, D, Q)s to predict future values.

II. LITERATURE SURVEY

Rios-Moreno et al. [5] used outside air temperature, relative humidity, air velocity, and global solar radiation flux as external variables to an autoregressive (AR) and an autoregressive moving average (ARMA) model. They successfully predicted the room temperature in a university classroom in Mexico. The results showed that the external variable older than 20 minutes did not improve the performance of the model. Felice et al. [6] used a non-seasonal time- series method to predict electricity demand at the national and regional level in Italy. It was demonstrated that using temperature as an external variable improved the prediction results. Mahmudur Rahman, A.H.M. Saiful Islam, Sahah Yaser Maqnoon Nadvi, Rashedur M Rahman (2013) consider Arima and Anfis Model and explained how ARIMA Model can more efficiently capture the dynamic behavior of the weather property, say, Minimum Temperature, Maximum Temperature, Humidity and Air pressure which must be compared by different performance metrics, for example, with Root Mean Square Error (RMSE), R-Square Error and the Sum of the Square Error(SSE) [7] and author can prove that ARIMA would give the more efficient result than other modeling techniques like ANFIS.

Further, [8] carried out a study for analyzing the trend and forecast maximum monthly temperature over the South Eastern Nigeria using SARIMA model. Depending on the best suited SARIMA model, the forecasted five years maximum temperature reflects to be slightly stable from that of the reference period. In another study, [9] fitted SARIMA model to average temperature for the period of 1980-2010 of Dibrugarh using automatic arima function i.e., autoarima() in R software. Keeping these points in mind, an attempt has been made to develop a SARIMA model on historical temperature data of Dibrugarh for the period of 1966-2015. The model is developed for both minimum and maximum temperature readings.

III. METHODOLOGY

Temperature data recorded from 2009 to 2020 were obtained for Pune city, from the meteorology department at one-hour intervals [12]. The longitude and latitude of the automatic weather station is 73.856255 and 18.516726, respectively. The data collected has different parameters, such as date time, temperature, humidity, moonrise, wind speed, wind direction, pressure. From this, we have eliminated features that have large amounts of missing data and we have considered temperature as an input parameter. The seasonal ARIMA model is implemented by running Python 3.7.4 on Jupyter Notebook and using the package matplotlib 3.2.1 for data visualization. Time series plot of temperature for the year 2018 was shown in Figure1. The hourly temperature data during 2009–2018 is used as the training set, while that during 2019–2020 is used as the testing set. To evaluate the forecast accuracy, as well as to compare the results obtained from different models, the mean-square error (MSE) is calculated.

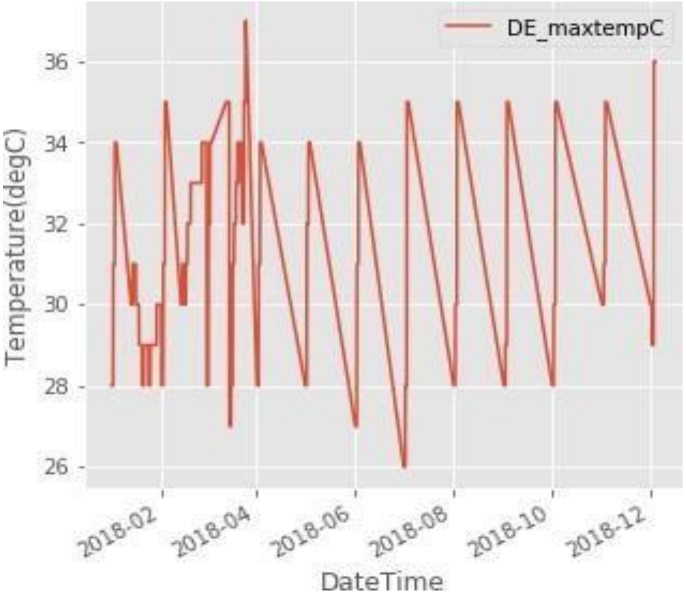


Figure 1 Time series plot of Temperature

in Pune (year-2018)

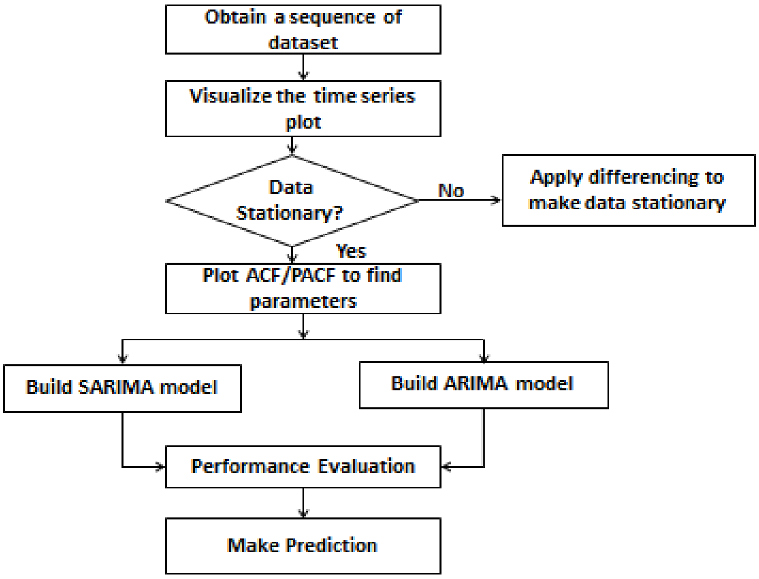

Obtain a sequence of dataset
Visualize the time series plot
Data Stationary?
No
Apply differencing to make data stationary
Yes
Plot ACF/PACF to find parameters
Build SARIMA model
Build ARIMA model
Performance Evaluation
Make Prediction

Figure 2 Flowchart of the proposed model

Check stationarity: If the time series is not stationary, it needs to be stationarized through differencing. Take the first difference then, determining stationarity with an augmented Dickey-Fuller test. The order of differencing (d) is selected such that it minimizes the standard deviation. This is done by fitting different ARIMA models having various orders of differencing, but a constant coefficient is selected. An already differenced series which is now a stationery series might still have some auto-correlated errors which can be removed by adding AR terms ($p \geq 1$) and MA terms ($q \geq 1$) in the forecasting equation. To compensate for any mild 'under-differencing', AR terms are added to the model, while to compensate any mild 'over-differencing', MA terms are added instead.

Plot ACF and PACF: In this step, the ACF and PACF of the data are plotted. Autocorrelation functions (ACF) and partial autocorrelation functions (PACF) are used to identify potential models. If the ACF and PACF have large values (positive) that decrease very slowly with time, this means that d is bigger than zero, i.e., differencing should be done. The autocorrelation function ACF and partial autocorrelation function (PACF) are often used for the choice of p, d, and q. If there's a pointy cutoff within the PACF of the differenced series and therefore the series shows mild 'under-differencing', an AR term is added to the model. If there is a sharp cutoff in the ACF of the differenced series and the series shows mild 'over-differenced', an MA term is added to the model.

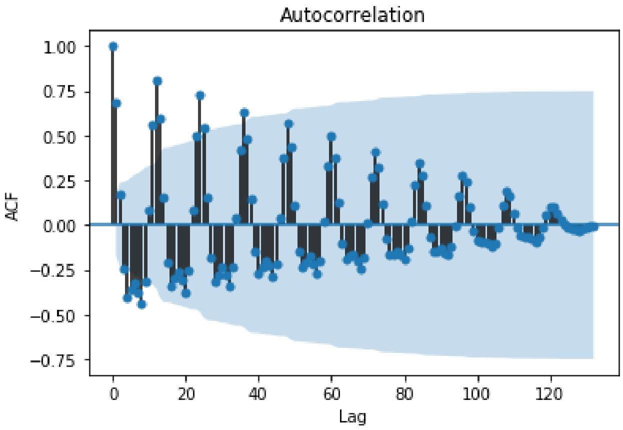


Figure 3 Autocorrelation Function

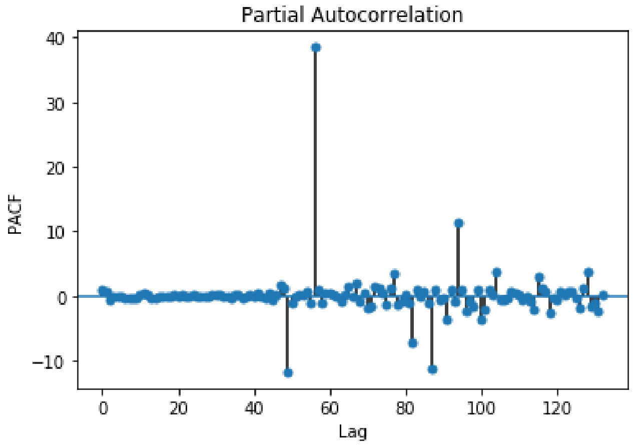


Figure 4 Partial autocorrelation Function

Auto Regressive Integrated Moving Average Model: For ARIMA models, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as

- p: the number of lag observations in the model; also known as the lag order.
- d: the number of times that the raw observations are differenced; also known as the degree of differencing.
- q: the size of the moving average window; also known as the order of the moving average.

According to Box Jenkins methodology an ARIMA model is usually written as ARIMA (p,d,q) [10].

The **AR(p)** model is defined by the equation:

$$X_t = \alpha + \phi 1X_{t-1} + \phi 2X_{t-2} + \cdots + \phi pX_{t-p} + \omega t \qquad (1)$$

Where,

- Xt = response variable at time t
- Xt-1, Xt-2, …., Xt-p = response variable at time t-1, t-2 and t-p respectively.
- α=constant term
- ϕ1, ϕ2 and ϕp = coefficients to be estimated
- ωt = error term at time t

The **MA(q)** model is defined by the equation:

$$X_t = \alpha + \theta 1w_{t-1} + \theta 2w_{t-2} + \cdots + \theta qw_{t-q} + \omega t \qquad (2)$$

Where,

- Xt=response variable at time t
- α =constant term
- wt-1, wt-2, …, wt-q = forecast errors at time series lags t-1, t-2 and t-q
- θ1, θ2 and θq=coefficients to be estimated
- ωt = error terms at time t

By combining equation (1) and (2) Autoregressive integrated moving average model ARIMA (p,d,q) can be written mathematically as

$$X_t = \alpha + \phi 1X_{t-1} + \phi 2X_{t-2} + \cdots + \phi pX_{t-p} + \theta 1w_{t-1} + \theta 2w_{t-2} + \cdots + \theta qw_{t-q} + \omega t \qquad (3)$$

Seasonal ARIMA model: In addition to trend, stationary series quite commonly display seasonal behavior where a certain basic pattern tends to be repeated at regular seasonal intervals. Seasonal ARIMA model (SARIMA) is formed by adding seasonal terms in the ARIMA models listed above**.** SARIMA models are written as,

ARIMA (p, d, q) (P, D, Q) m (4)

Where (p, d, q) and (P, D, Q) m are the non-seasonal and seasonal part of the model, respectively. The d parameter tells how many differencing orders are going to be used to make the series stationary. The parameter m is the number of periods per season. The value of m is set with a period of 12.

IV. EXPERIMENTAL RESULTS

Akaike's Information Criterion (AIC) is the most commonly used model selection criterion [10]. AIC essentially measures the goodness of fit of a model. AIC is calculated as [10]:
AIC = -2 ln (maximum likelihood) + 2p
Where, p denotes the number of independent parameters estimated. Therefore, when comparing models, the one with the least AIC value is chosen. According to Table 1, SARIMA (1, 1, 1) × (1, 1, 1) 12 shows the lowest AIC value (AIC=196085.724). Thus, this model should be considered as the best forecasting model.

Table1 AIC values of SARIMA models.

SARIMA (p, d, q) (P, D, Q)s	**AIC Values**
SARIMA(0, 0, 0)x(0, 0, 0)12	AIC:204680.074
SARIMA(0, 0, 0)x(0, 0, 1)12	AIC:217984.649
SARIMA(0, 0, 0)x(0, 1, 0)12	AIC:200841.533
SARIMA(0, 0, 0)x(0, 1, 1)12	AIC:215339.952
SARIMA(0, 0, 0)x(1, 0, 0)12	AIC:197085.724
SARIMA(0, 0, 0)x(1, 0, 1)12	AIC:247760.254
SARIMA(0, 0, 0)x(1, 1, 0)12	AIC:196650.122
SARIMA(0, 0, 0)x(1, 1, 1)12	AIC:205623.637
SARIMA(0, 0, 1)x(0, 0, 0)12	AIC:245623.637
SARIMA(0, 0, 1)x(0, 0, 1)12	AIC:225623.637
SARIMA(0, 0, 1)x(0, 1, 0)12	AIC:235623.637
SARIMA(0, 0, 1)x(0, 1, 1)12	AIC:197623.637
SARIMA(0, 0, 1)x(1, 0, 0)12	AIC:208688.838
…..	…..
SARIMA(1, 1, 1)x(1, 1, 1)12	**AIC:196085.724**

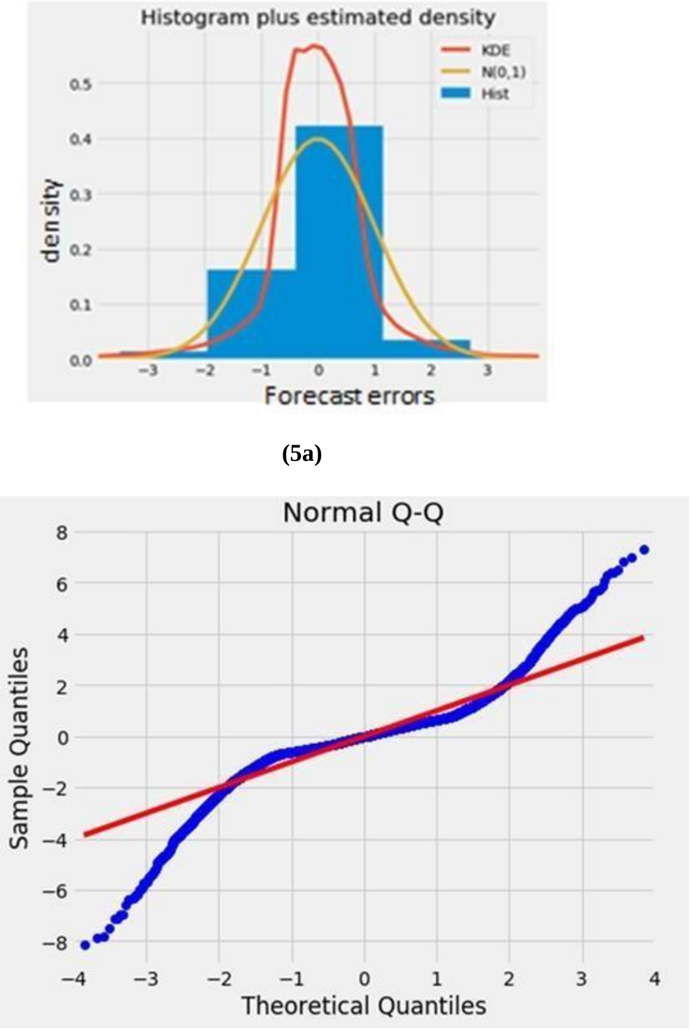
Diagnostic Test: The forecast accuracy of the selected model is validated by applying a Dickey-Fuller test. According to Table2 the AIC value of SARIMA (1, 1, 1) × (1, 1, 1)12 is the lowest. Table2 summarizes the results of the diagnostics test of the SARIMA (1, 1, 1) × (1, 1, 1,) 12 model.

The results of the diagnostic test on SARIMA (1,1,1) × (1,1,1) 12 are shown in Figure 5. According to Figure 5a, the results imply that the residual follows a normal distribution, with mean equal to 0 and standard deviation equal to 1. In Figure 5b, the Q-Q plot of the residuals implies that the residuals follow a linear trend. Thus, the residuals are normally distributed. Table3 show the comparison between actual and predicted value of temperature in℃.

Table2 Summary of the diagnostics test of the SARIMA (1,1, 1) × (1, 1, 1,) 12 model.

	Coef	*std err*	z	P>|z|	[0.025	0.0975]
Const	28.1213	0.461	60.976	0.000	27.217	29.025
AR.L1	0.7704	0.003	222.732	0.000	0.764	0.777
MA.L1	-1.0000	0.006	-155.452	0.000	-1.013	-0.987
AR.S.L12	-0.6651	0.005	-147.274	0.000	-0.674	-0.656
MA.S.L12	-0.8680	0.002	-381.308	0.000	-0.873	-0.864

The second column is the weight of the coefficients. The "Coef" column shows the weighting (i.e., importance) of each feature and how each one impacts the time series. Since all values of P> |z| are less than 0.05, the results are statistically significant.



(5a)

(5b)

Figure 5 Diagnostic tests on the residuals of the model
(5a) Distribution of standardize residuals
(5b) Normal Q-Q plot

Table3 Actual value vs predicted values of temperature (℃)

DateTime	Actual Values	Predicted Values
2019-01-31 02:00:00	30.193548	28.765794
2019-02-28 02:00:00	32.642857	31.115332
2019-03-31 02:00:00	35.032258	35.282232
2019-04-30 02:00:00	35.466667	35.745039
2019-05-31 02:00:00	35.677419	35.375808
2019-06-30 02:00:00	30.033333	33.774278
2019-07-31 02:00:00	28.258065	27.734503
2019-08-31 02:00:00	29.129032	28.324608
2019-09-30 02:00:00	29.166667	29.434541
2019-10-31 02:00:00	28.580645	29.058096
2019-11-30 02:00:00	29.533333	28.366377
2019-12-31 02:00:00	29.516129	29.968087

Figure 6, shows the time series plot of actual value and predicted values of the temperature using SARIMA model. To evaluate the quality of the model, we will first compare the predicted values with the actual values. We can also see some kind of variations in the plot. These types of seasonal variations may cause by climate condition and any other external factors. From this figure, we can observe that the prediction results are almost equal to the actual data. We can say that the seasonal ARIMA model is performing better. Figure 7, shows the time series plot of actual value and predicted values of the temperature using ARIMA model. From the results we can say that the model is not fitted well as compare to SARIMA model. Figure 8, shows the future prediction of temperature using SARIMA model.

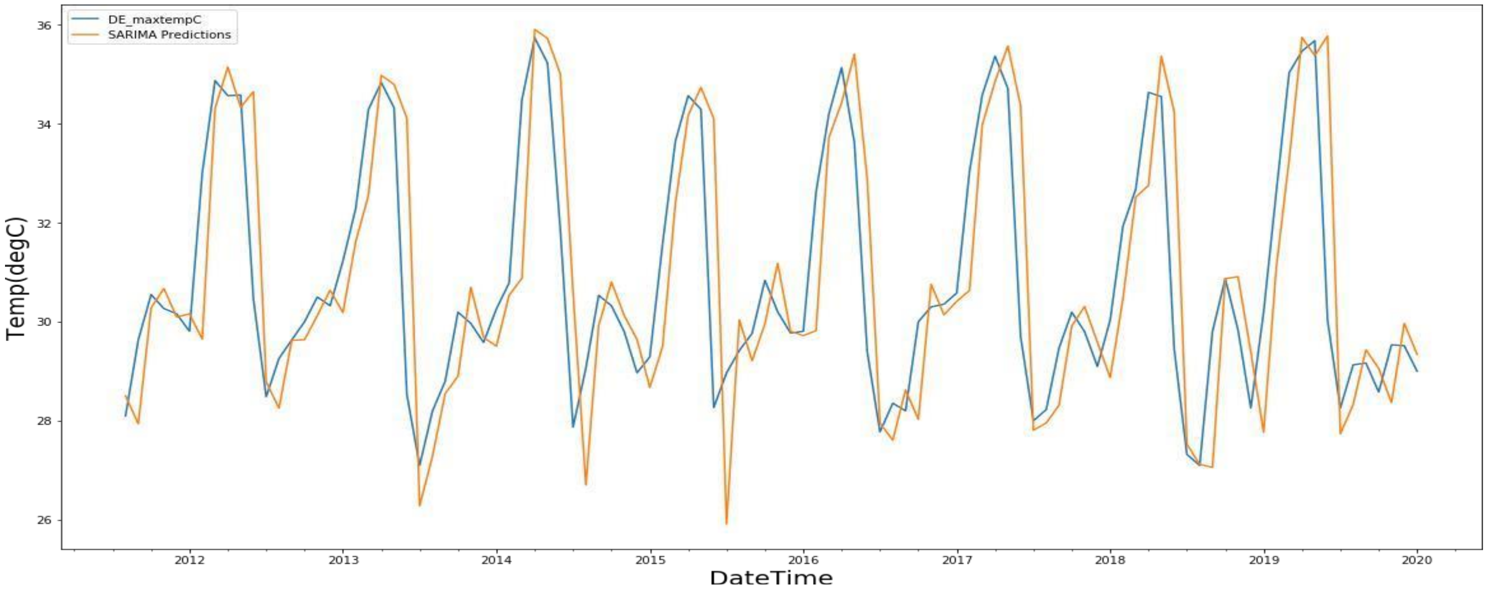


Figure 6Actual value v/s predicted value using SARIMA model

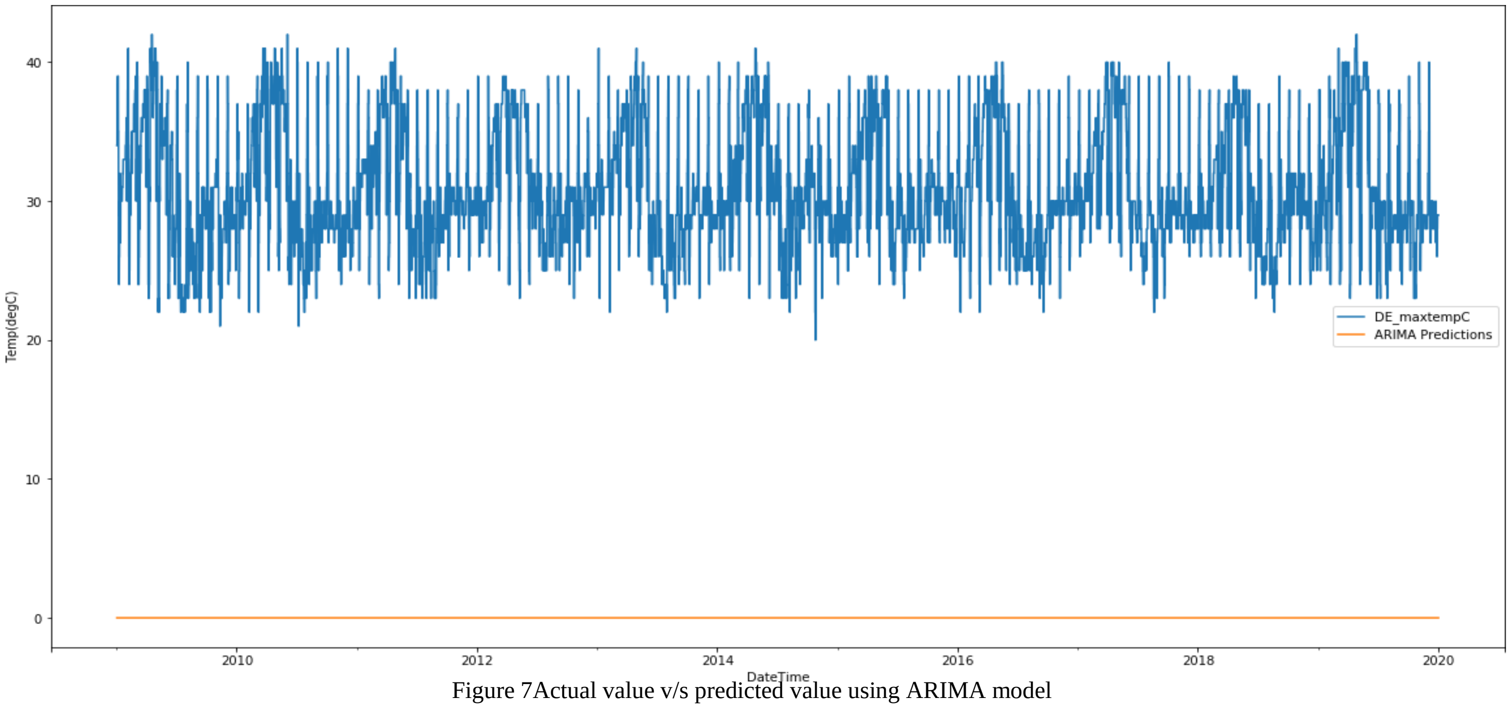


Figure 7Actual value v/s predicted value using ARIMA model

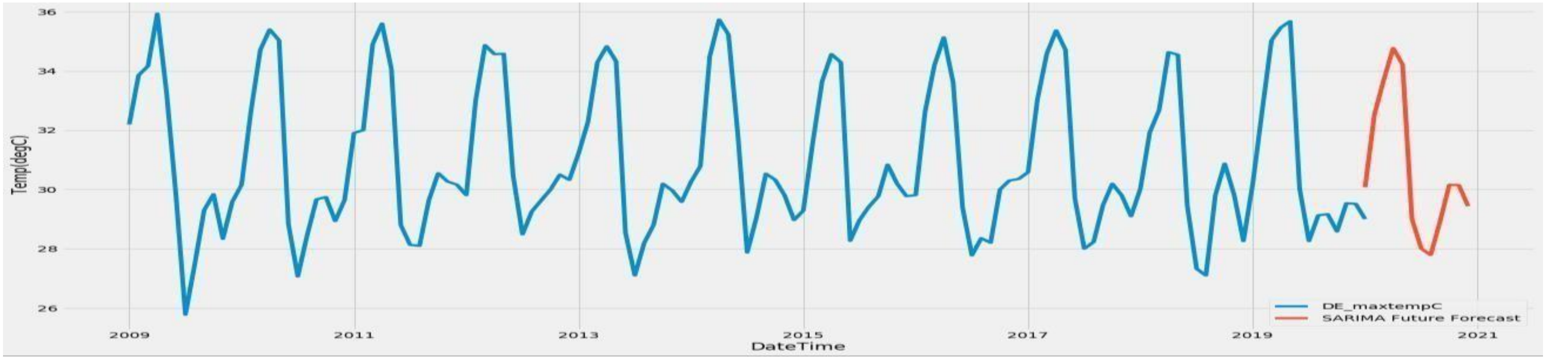


Figure 8 Future Prediction of temperature using SARIMA model

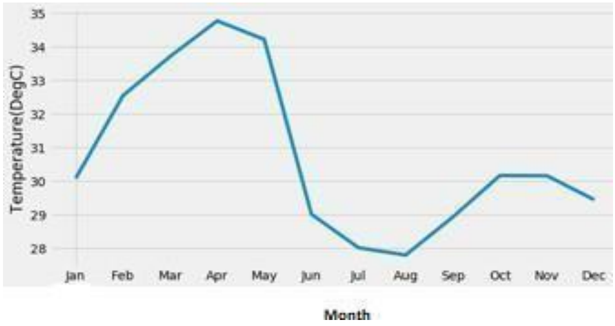


Figure 9 Time series plot of the future prediction (year-2021)

The above figure shows the time series plot of the temperature prediction. During the month from June-August we can see the sudden decrease in temperature, we can assume that this will be due to rainy season.

V. PERFORMANCE EVALUATION

MSE, RMSE and MAE were used as performance evaluation metrics given in Table 4. By taking the square of the errors, MSE is calculated as [11]:

$$MSE = \frac{1}{n}\sum_{1}^{n} e_t^2 \tag{5}$$

RMSE takes the root of the MSE. Thus, it has the same unit of measurement as the data. It is calculated as [11]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n} e_t^2} \tag{6}$$

Mean absolute error is the average of the absolute values of the deviation.

$$\boldsymbol{MAE = mean(|e_t|)} \tag{7}$$

Table4 Results of the performance evaluation of the model

Method	MAE	MSE	RMSE
ARIMA	**6.052**	**56.187**	**7.496**
SARIMA	**0.60850**	**0.58114**	**0.762325**

The predicted temperature values are compared with actual values for accuracy based on error metrics. We obtained MAE of 0.60850 and RMSE of 0.76233for SARIMA model and MAE of 6.052 and RMSE of 7.496 for ARIMA model. From the above table, we concluded that SARIMA model forecasts yielded least error in prediction of temperature as output.

CONCLUSION

In this paper, temperature data were collected from the year 2009-2020 at one-hour intervals in Pune. The estimation and diagnostic analysis results revealed that the model adequately fitted to the historical data. A grid search was conducted to find the best model from different combinations of seasonal (P, D, Q) and non-seasonal parameters (p, d, q). Each parameter was set to take a value of either zero or one. The model with the least AIC was selected. Finally, the predicted values were compared with the actual values of both using ARIMA and SARIMA model. From the results we can say that SARIMA model is working well. Prediction is very poor with ARIMA and forecast accuracy measures, including MAE, MSE, and RMSE were calculated. Thus, this model was used to predict values in 2021using temperature as an input variable.

References

[1] Imdadullah. "Time Series Analysis". Basic Statistics and Data Analysis. itfeature.com. Retrieved 2 January 2014.

[2] Raicharoen, t., lursinsap, c., & sanguanbhokai, p. (2018). Application of critical support vector machine to time series prediction. International symposium on circuits and systems (vol.5, pp.v-741- v-744 vol.5). IEEE

[3] Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid arima and ann models based on DWT Decomposition. In Procedia Computer Science (Vol. 48, pp. 173–179) Elsevier B.V. https://doi.org/10.1016/j.procs.2015.04.167

[4] Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis forecasting and control - rev. ed. Oakland, California, Holden-Day, 1976, 37 (2), 238 - 242.

[5] G. J. Rios-Moreno, M. Trejo-Perea, R. Castaneda-Miranda, V. M. Hernandez-Guzman, and G. Herrera-Ruiz, "Modelling temperature in intelligent buildings by means of autoregressive models," vol. 16, pp. 713–722, 2014.

[6] M. De Felice, A. Alessandri, and P. M. Ruti, "Electricity demand forecasting over Italy, Potential benefits using numerical weather prediction models," Electr. Power Syst. Res., vol. 104, pp. 71–79, 2013.

[7] Mahmudur Rahman, A.H.M. Saiful Islam, Sahah Yaser Maqnoon Nadvi, Rashedur M Rahman (2013): Comparative Study of ANFIS and ARIMA Model for weather forecasting in Dhaka" IEEE

[8] Chisimkwuo, J., Uchechukwu, G. And Okezie S.C. 2014 Time series analysis and forecasting of monthly maximum temperatures in south eastern Nigeria. International Journal of Innovative Research and Development. ISSN 2278 – 0211. 3(1), pp. 165- 171

[9] Roy, T. D. and Das K. K. 2012 Time series analysis of Dibrugarh air temperature. Journal of Atmospheric and Earth Environment. 1(1), pp. 30- 34.

[10] Box-Jenkins models, NIST handbook of statistical method

[11] C. Chatfield, Time-series Forecasting. Chapman & Hall/CRC, 2015.

[12] www.kaggle.com