Aerosol Science and Engineering
https://doi.org/10.1007/s41810-020-00061-7

ORIGINAL PAPER

Identification and Analysis of Nitrogen Dioxide Concentration for Air Quality Prediction Using Seasonal Autoregression Integrated with Moving Average

S. R. Mani Sekhar[1] · G. M. Siddesh[1] · Anjaneya Tiwari[1] · Anay Khator[1] · Robin Singh[1]

Received: 28 November 2019 / Revised: 20 April 2020 / Accepted: 5 May 2020

Abstract
Air quality is getting degraded every day, and this problem has by now caught the eyes of all the concerned authorities and people across the globe. In previous works, the forecasting could only be done for a very short interval of time, and the prediction was only reliable for up to 24 h. This work proposes a method that can be used to predict the Nitrogen dioxide (NO_2) value, which in turn can be used to predict the air quality in the near future and warn the concerned authorities beforehand. The model used here is based on the concepts of autoregression and moving average. The primary reason for using these methods is the fact that the proposed model works on the concept of Time series analysis. The results show that proposed model Seasonal autoregression integrated moving average (SARIMA) performs better in terms of predicting future NO_2 concentration with lower RMSE error when compared with Support vector machine and Linear regression on the dataset taken from a government of India website (data.gov.in) containing NO_2 values in Delhi from the year 1987 to 2015. The results showed that the RMSE value (microgram per meter cube) computed using the proposed work was found to be 7.43 $\mu g/m^3$, whereas for Support vector machine are 8.1 $\mu g/m^3$ and for Linear regression are 8.5 $\mu g/m^3$. Finally, the proposed model is also validated with the real-time one-week Delhi data for March 2020.

Keywords Air quality index (AQI) · Time series analysis · SARIMA · Air quality forecasting · Linear regression · Support vector machine · RMSE · Delhi · NO2 · Nitrogen dioxide

1 Introduction

Urbanized and budding countries across the globe are contributing to the degradation of the air quality. Countries like Saudi Arabia, Qatar, Egypt, Bangaladesh, and Kuwait etc. are one of the countries that are causing significant damages to the air (World Economic Forum 2020). According to the leading scientists around the world, by 2030, the quality of the air would have degraded so much that many people would face breathing issues, and subsequently, the quality of water would also get affected (WHO 2020). Therefore, air quality is a significant issue, and one of the elementary pollution problems in many parts of the world.

There are mainly two sources of air pollutants, the first being the human-based factor such as combustion of petrol and diesel, industrial pollution, mining operations, agricultural activities etc. Series of data related to air pollution have been recorded, and it can be observed that air quality degradation is the principal cause of ill effects on human health. Poor air quality also affects major environmental problems such as global warming and greenhouse gases (Gupta et al. 2016), thus increasing the risk of the earth facing major damages as a long term effect.

There are many factors on which the air quality depends on, but we can narrow it down to six major parameters as per the Clean air act that is (NAAQS 2016) Particulate matter ($PM_{2.5}$), Ozone (O_3), Nitrogen dioxide (NO_2), Sulphur dioxide (SO_2), Lead and carbon monoxide (CO). Mainly, NO_2 contaminants generate the uppermost volume of contaminants in the air and the main threat to human life. Attention to these toxins, especially in cities, has been controlled by the Clean Air Act since 1970 (Cunningham and Cunningham

✉ S. R. Mani Sekhar
manisekharsr@gmail.com

[1] Department of Information Science and Engineering, M.S. Ramaiah Institute of Technology, Bangalore, India

Published online: 16 May 2020

2002). But predicting air quality from all these factors is quite cumbersome, so we instead use the ambient air quality known as Air quality index (AQI). Since all the data is recorded in continuous and uniform time intervals, time series analysis, and forecasting techniques are used in this paper. By using time series forecasting, we can predict the quality of air in the near future, which helps us make plans for the future and to take the necessary actions required to control the pollution levels.

Air Quality Index: The air quality index (Ganesh et al. 2017) is one of the simplest and most generalized ways for describing the air quality status. The way in which the AQI is calculated is different around the world. Every country has its own way of calculating the AQI. It mainly takes into account five main parameters namely Particulate matter ($PM_{2.5}$), Carbon monoxide, Ozone, Sulphur dioxide and Nitrogen dioxide. The daily AQI value is then recorded based on the highest index value of the parameters. Table 1, shows the AQI scale and values associated with it.

Box-Jenkins Modeling (Essallah et al. 2015; Box et al. 2016): This method is a type of linear model that is capable of expressing both motionless and non-stationary time series. Most people make use of this model to anticipate univariate time series data. Box-Jenkins methods are a very practical implementation of time series forecasting. It includes models such as autoregressive (AR) models (Jain et al. 2016), the integrated (I) models and other models such as the moving average (MA) models (Putra et al. 2017).

The paper is organized as follows. Section 2 brief about the related work in air quality prediction. Next Sect. 3 illustrates the proposed Seasonal autoregression integrated moving average model (SARIMA). Subsequently, Sect. 4 focuses on the Results and discussion part. Finally, Sect. 5 provides a brief discussion on Conclusion.

2 Related Work

Gaganjor Kaur Kang (Kang et al. 2018) provides a brief comparison of various machine learning methods like deep learning, decision tree and artificial intelligence etc. for air quality assessments such as NO_2 and SO_2 etc., subsequently they highlighted the various issues and challenges. Ke Gu and Junfei Qiao (Gu et al. 2018) used Recurrent air quality prediction based on meteorology- and pollution-related factors. They further used a support vector machine (SVM) integrated with the repressor. The main problem with this method was that the correlations decline with the time interval enlarged.

In (Soh et al. 2018), authors have used adaptive deep learning concepts for prediction of AQI upto 48 h using an arrangement of multiple neural networks. The neural networks (NN) model used in this paper was similar to an ANN model with decent accuracy but the main problem with this paper was that the features used for prediction did not involve some of the key factors like taking trends or seasonality in the data into consideration for predictions and giving more weightage to recent values than older values (moving average) used for the prediction of AQI and had some chemical features which on analyses with the help of a correlation matrix were found to be irrelevant thus decreasing its accuracy.

Deep learning: Interpolation prediction and feature analysis of fine-grained air quality (Qi et al. 2018). The main focus of this paper was to predict the air quality and feature analysis of fine gained air quality. The difficulties faced in this paper was that the features used were taking in only the historical data stored over time and did not take into account the present values of the parameters, thus reducing the accuracy.

The proposed work tries to overcome the various problems like seasonality and having more weightage of recent values than older values. The proposed SARIMA model can predict accurately for up to 3 years without reducing accuracy, even if the time interval is increased the accuracy is not compromised. Because of the fact that the SARIMA model takes into account the seasonality factor, it produces a better correlation matrix, thus increasing accuracy. Since the model uses both historical data and present values, a better prediction model is achieved.

Table 1 Air quality index (AQI) (Ganesh et al. 2017)

Air quality index (AQI)	Air quality
0–50	Satisfactory
51–100	Centrist
101–200	Unhealthful
201–300	Poor
301–400	Very poor
401–500	Precarious

3 Proposed Solution

Identification and analysis of air quality prediction is one of the most time consuming and challenging tasks. In the proposed authors have developed SARIMA model which is both accurate and that can predict not only the near future values of AQI but also for the upcoming years thus providing the concerned authorities ample of time to take necessary actions and thus warning them well beforehand (Plaia and Ruggieri 2011). The model implemented here is a Time series forecasting (Yu 2016) model which uses autoregression along with moving average giving it the name ARIMA. Now, the very basic of air quality prediction is

that it changes very frequently and depends on seasonality (Brownlee 2019). We are using a stack model library that has mainly taken care of the seasonality factor and is used for trend analysis models and statistical modeling. Statistical modeling involves all of the statistical influential mathematics which includes inference mathematics and theoretical design implementations. Now in air quality the seasonality factor is huge, to get the most accurate prediction, we must take into account the current trend of the data or rather current trend of the seasonality.

To better explain the seasonality factor we can say that every season of the year has an adverse effect on the air quality, like in summers the amount of air conditioning needed and the climate forces more combustion thus resulting in higher AQI value whereas the rainy season mellows it down by some factor as it increases the amount of water vapour in the air. Seasonality illustrates the trend variation, how it is increasing or decreasing and by what factor.

Figure 1, illustrate the proposed SARIMA architecture, the data set being used here has data for multiple cities across India, it also has data related to some irrelevant factors thus some amount of data preprocessing is also required. In this phase, we take in only the data frame related to the city Delhi and fill in all the empty data values with either Nan or the mean values for that parameter. Further, the resampling of the data is performed so that the trend taken in takes into account the entire data at random and does not take in only the past near values.

After this phase, the time series model is applied to the data set and the seasonality factors are set accordingly to get the future prediction of up to 12 months. The model takes in the date/time parameter and maps it with the AQI related parameter to get the correct time series analysis.

The SARIMA model uses this methodology for prediction and presents a comprehensive way for dealing with the data set. The model uses the below prediction method as given in Eq. 1.

$$Q(t) = B(1) * Q(t-1) + C(t) \tag{1}$$

Here '$Q(t)$' shows the investigated time series value. Next '$B(1)$' parameter contains the order 1 autoregression value. Similarly, $Q(t-1)$ value lagged by 1 in terms of time series. Finally computed error is stored in $C(t)$ parameter.

$$Q(t) = B(1) * Q(t-1) + B(2) * Q(t-2) + C(t) \tag{2}$$

Equation 2 shows us that for some given assessment $Q(t)$ can be explained by some function of its earlier value i.e. $Q(t-1)$. This also contains some undefined random error $C(t)$ which is added in it. Usually, the series is related to only one time period but can also be used for more than one which is shown in the last equation.

3.1 Proposed SARIMA Model

The steps illustrated below give brief information about the various steps involved in the proposed SARIMA model.

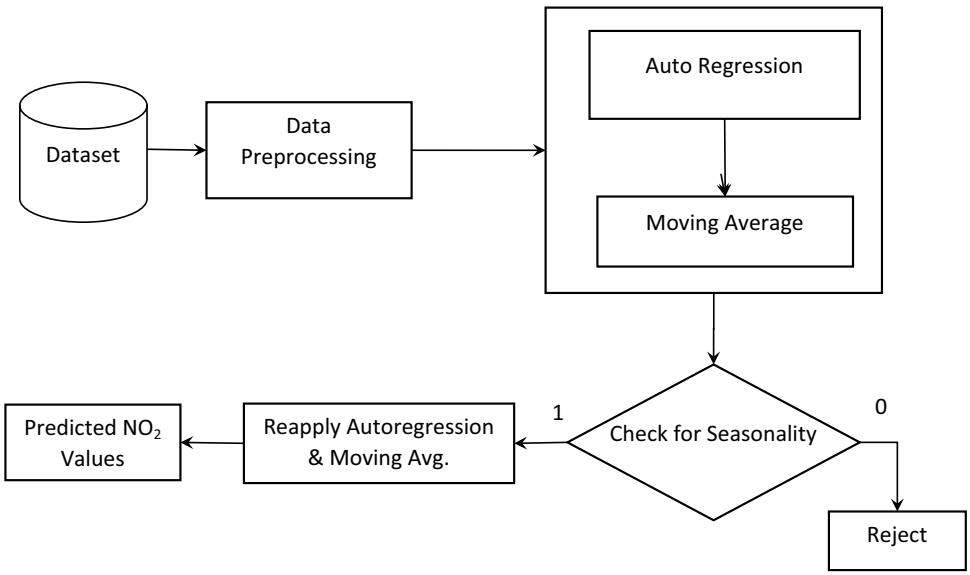

Fig. 1 Proposed SARIMA architecture for AQI prediction

Step 1: Check for stationarity, if it has check current trend component, it need to be stationary before using SARIMA to predict.
Step 2: If it does not have stationary value it needs to be made into one by applying differencing.
Step 3: Remove out a validated sample, used to validate how accurate the model is.
Step 4: Build the model and select the parameter accordingly.
Step 5: Finally validate the model by comparing the predicted values with the actual values.

ARIMA Pseudocode

Input: Air Quality dataset
Output: Predicted Values
begin
for p 0 to 1 **do**
Compute autoregression for past terms
for d 0 to 1 **do**
Compute non-seasonal difference with previous value
for q 0 to 1 **do**
Compute moving average with previous value
Final= fit(arima(p, d, q, allowglide= True))
dataset_curr = compute(model)
model_opt= model
return model_opt
end

In the ARIMA pseudocode the general outline working is presented. The input provided is the dataset and the final output is the predicted parameter values. The p, d, q values can be computed in order to get the expected SARIMA model to be applied where p is the number of auto-regression terms which allows to incorporate the effect of past values into this model, d is the number of non-seasonal differences and q is the number of moving average terms which allows to set the error of this model as a linear combination of the errors values observed at previous time points in the past.

3.1.1 Validation of Proposed SARIMA Model

The accuracy measure for this type of models can be decided by the following criteria—Mean absolute error (MAE), Mean absolute percentage error (MAPE), Mean square error (MSE) and the Root mean square error (RMSE), the. The below Eq. 3, 4, 5 and 6 illustrate the same.

$$\mathrm{MAE} = \frac{\sum_{t=1}^{n} |Z_t - Y_t|}{n} \tag{3}$$

$$\mathrm{MAPE} = \frac{\sum_{t=1}^{n} \left|(Z_t - Y_t)/Z_t\right|}{n} * 10 \tag{4}$$

$$\mathrm{MSE} = \frac{\sum_{t=1}^{n} \left(Z_t - Y_t\right)^2}{n} \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} \left(Z_t - Y_t\right)^2}{n}} \tag{6}$$

Here, Z_t and Y_t describe the actual value and fitted value at time one. Similarly, 'n' represents the observation count. The smallest values of MAE, MAPE, MSE and RMSE are chosen as the best model to be used in forecasting.

4 Results and Discussion

The dataset used has 13 features and 435,743 rows as shown in Fig. 2. The dataset contains monthly recorded values of different air pollutants at different states in India provided by government bodies related to air pollution. The dataset has recorded data from 1995 till 2015 which is used as training set for the proposed model. The main air pollutants present in the dataset are SO_2 and NO_2 concentration in µg/m^3 (microgram per metre cube). In this dataset, we extract the required features for the city Delhi to predict the NO_2

B65863 | 09-01-1987

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
65863	60	########	Delhi	Delhi	Central Pollution Control Board	Residential	8.6	16.2	NA	323	NA	NA	######	
65864	59	########	Delhi	Delhi	Central Pollution Control Board	Residential	6.5	9.3	NA	488	NA	NA	######	
65865	58	########	Delhi	Delhi	Central Pollution Control Board	Industrial	25.8	10.4	NA	487	NA	NA	######	
65866	57	########	Delhi	Delhi	Central Pollution Control Board	Industrial	0.5	2.8	NA	528	NA	NA	######	
65867	56	########	Delhi	Delhi	Central Pollution Control Board	Industrial	2.5	12.8	NA	506	NA	NA	######	
65868	55	########	Delhi	Delhi	Central Pollution Control Board	Industrial	16.4	17.2	NA	371	NA	NA	######	
65869	60	########	Delhi	Delhi	Central Pollution Control Board	Residential	3.1	14.9	NA	414	NA	NA	######	
65870	59	########	Delhi	Delhi	Central Pollution Control Board	Residential	14.2	14.6	NA	337	NA	NA	######	
65871	58	########	Delhi	Delhi	Central Pollution Control Board	Industrial	27.3	13.2	NA	372	NA	NA	######	
65872	57	########	Delhi	Delhi	Central Pollution Control Board	Industrial	40.5	22.5	NA	638	NA	NA	######	
65873	56	########	Delhi	Delhi	Central Pollution Control Board	Industrial	NA	12.9	NA	505	NA	NA	######	
65874	55	########	Delhi	Delhi	Central Pollution Control Board	Industrial	27	20.6	NA	396	NA	NA	######	
65875	60	########	Delhi	Delhi	Central Pollution Control Board	Residential	10.5	14.2	NA	435	NA	NA	######	
65876	59	########	Delhi	Delhi	Central Pollution Control Board	Residential	3.9	14.3	NA	571	NA	NA	######	
65877	58	########	Delhi	Delhi	Central Pollution Control Board	Industrial	7.5	8.2	NA	601	NA	NA	######	
65878	57	########	Delhi	Delhi	Central Pollution Control Board	Industrial	46.2	33.4	NA	777	NA	NA	######	
65879	56	########	Delhi	Delhi	Central Pollution Control Board	Industrial	2.5	20.5	NA	894	NA	NA	######	
65880	55	########	Delhi	Delhi	Central Pollution Control Board	Industrial	9.4	24.5	NA	522	NA	NA	######	
65881	60	########	Delhi	Delhi	Central Pollution Control Board	Residential	3.5	30.8	NA	714	NA	NA	######	
65882	59	########	Delhi	Delhi	Central Pollution Control Board	Residential	20.7	26.7	NA	440	NA	NA	######	
65883	58	########	Delhi	Delhi	Central Pollution Control Board	Industrial	12.9	22.1	NA	566	NA	NA	######	
65884	57	########	Delhi	Delhi	Central Pollution Control Board	Industrial	60.4	35.4	NA	895	NA	NA	######	
65885	56	########	Delhi	Delhi	Central Pollution Control Board	Industrial	3.7	20.7	NA	909	NA	NA	######	
65886	55	########	Delhi	Delhi	Central Pollution Control Board	Industrial	14.2	30.6	NA	484	NA	NA	######	
65887	59	########	Delhi	Delhi	Central Pollution Control Board	Residential, Rura	NA	NA	NA	114	NA	NA	######	

data

Fig. 2 Sample air pollutant dataset (DataGov 2019)

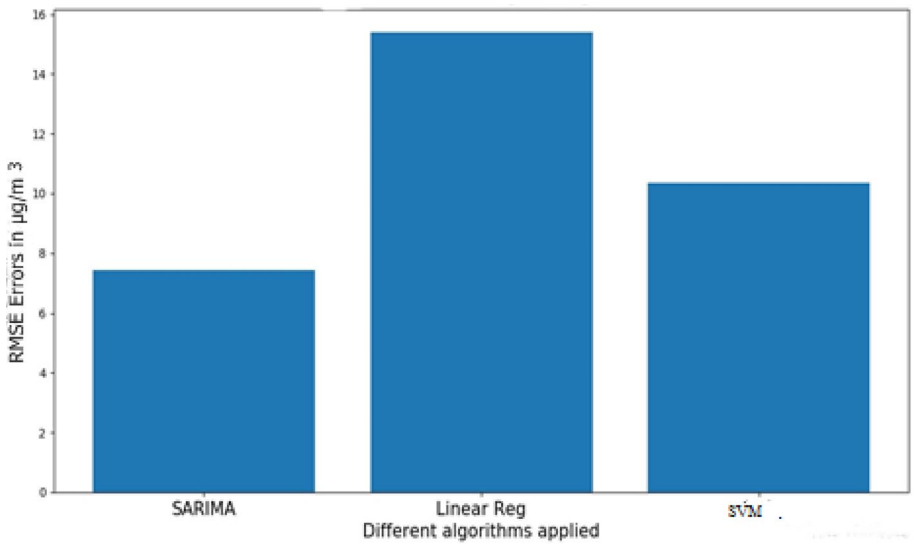

Fig. 3 Comparison of RMSE errors for proposed SARIMA with Linear regression and SVM models

concentration. The dataset has been changed into a structured dataset for better visualization of it.

All the computation performed in this work has been performed by using Python version 3.6. Before applying the SARIMA model the data set has been a plot and checks the forecasting performances based on models built.

In Fig. 3, provides a comparison of RMSE error with proposed SARIMA, Linear regression and Support vector machine (SVM) model. The different RMSE values are shown with respect to SVM, Linear regression algorithms when applied on the same dataset which is used in the present study. The best value was achieved by the proposed SARIMA model. Whereas, Linear regression was the least acceptable because of the fact that it does not take into consideration the seasonality factor. Later Support vector machine model showed better results than linear regression, but the RMSE was further reduced by SARIMA and brought down to 7.43 µg/m^3 as compared to 9.8 µg/m^3 and 15.2 µg/m^3 which were achieved by SVM and Linear regression respectively.

In Fig. 4 the plot shows the data values of NO_2 concentration over the years for the city Delhi in India. The *x*-axis

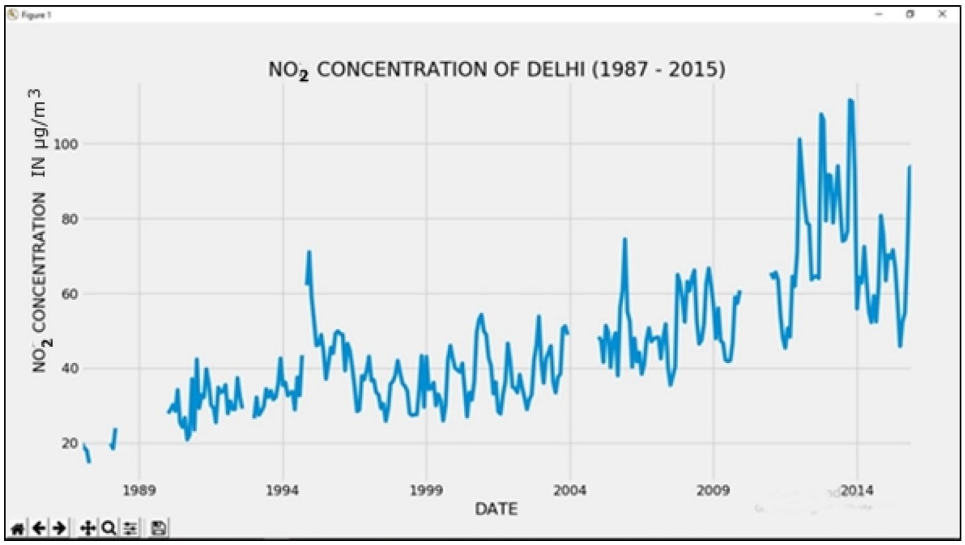

Fig. 4 NO_2 concentration visualization of air pollutant dataset for Delhi from 1987 to 2015

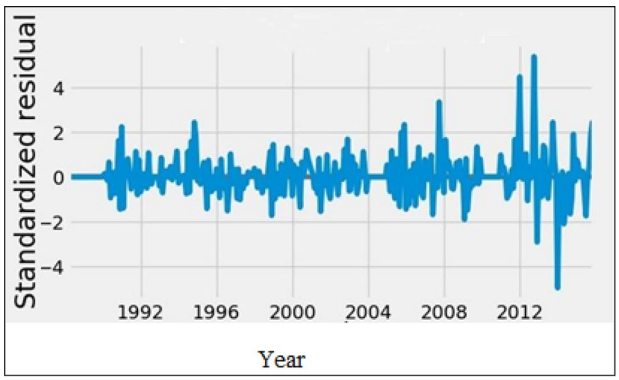

Fig. 5 Year wise standardized residual of SARIMA model

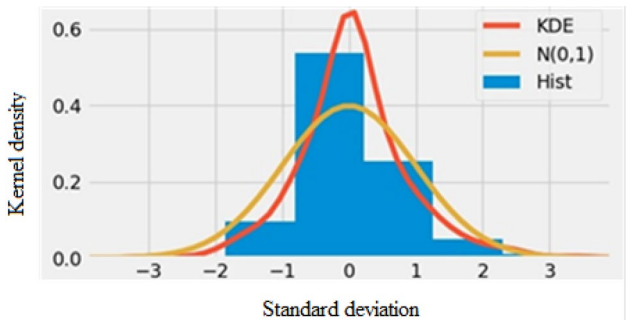

Fig. 6 Histogram plus estimated density of SARIMA

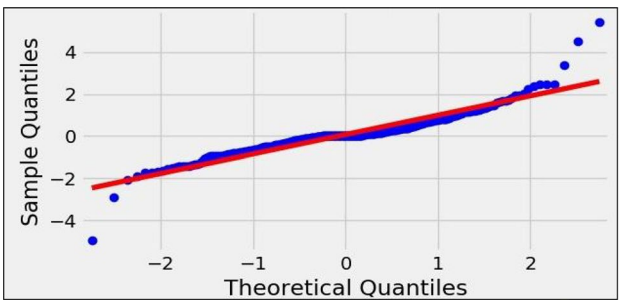

Fig. 7 Normal quantile–quantile plot for proposed SARIMA model

shows the year while the *y*-axis is associated with the NO_2 concentration in µg/m^3. The visualization shows that in 1999 the NO_2 concentration was 43 µg/m^3, meanwhile in 2004 NO_2 concentration was 50 µg/m^3, whereas in 2009 NO_2 concentration is increased to 65 µg/m^3, later in 2014 NO_2 concentration was 112 µg/m^3, finally in 2016 NO_2 concentration was reduced to 92 µg/m^3.

Proposed SARIMA model is visualized and validated using the following techniques (Deviant 2011):

- Standardized residual
- Kernel density estimation
- Normal quantile–quantile plot
- Correlogram

4.1 Standardized Residual

Standardized residual is used to compute the difference between the experimental value and predictable value. It also tells the importance of the value of the Chi-square measure. Subsequently, it tells which part of value contributes more and which less is.

The standardized residual of SARIMA shows that the data which has been chosen follows a particular trend and thus is a valid and well build data which can be used for future prediction. The main reason for doing this is that it helps the reader in understanding the trend followed through the years. In Fig. 5 the *x*-axis shows us the year while the *y*-axis represents the standard deviation of the residuals.

This is calculated by getting the mean and dividing it by the standard deviation. The graph show that residual value was less than 2 from the year 1987, 1988, 1989, 1991 to 1994, 1996–2005, 2007, 2009, 2010, 2011, 2015 except for the year 1990, 1995, 2006, 2008, 2012, 2013, 2014, 2016.

4.2 Kernel Density Estimation

Kernel density helps in the computation of probability density function for a given random number, it belongs to the categories of non-parametric density. Meanwhile, to overcome the issues of histogram such as un-even, bins width and endpoint dependency. kernel estimator is used. In Fig. 6 the plot is showing the kernel density and the bell curve that represents or rather shows the values which are following the current trend and thus help to visualize the data easily. The *x*-axis here shows the standard deviation while *y*-axis gives the kernel density value associated with it.

4.3 Quantile–Quantile Plot

Quantile–quantile plot (Q–Q) helps in comparing different probability distribution by assigning quantiles values against each other by graphical method. A quantile is a portion where definite values fall under that quantile. In Fig. 7, the Quantile–quantile plot helps us to know if the data has come from some theoretical distribution or not and thus help in assessing the efficiency of the prediction model. As long as the values are closely joined to the line it represents a good prediction model, hence the computed Q–Q values show that proposed model SARIMA performs better as the predicted values are close to the line. The *x*-axis here represents the standard deviation while the *y*-axis gives specific quantile values associated with it. Quantile can be calculated by dividing them into equal parts and then further dividing it by the standard deviation.

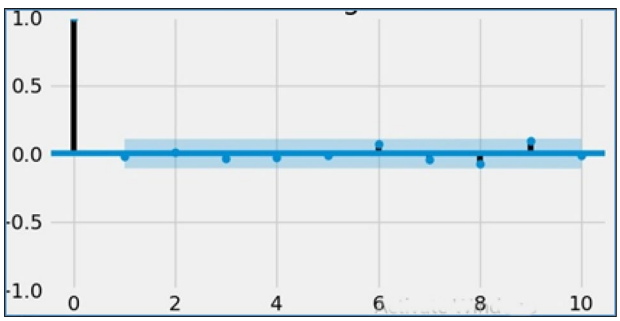

Fig. 8 Dependency computation using correlogram for proposed SARIMA model

4.4 Correlogram

A correlogram is a way to visualize data change over a certain time. Subsequently, helps to validate whether the group of records show autocorrelation or not. The correlogram shown in Fig. 8 is a correlation graph which tells us how much the attributes are dependent on each other. A correlogram is said to be good if all the data values lie close to the given range as near-zero values indicate very less dependency between attributes thus preventing redundancy and bias in the dataset. As the computed values are equal and near to zero, it shows that the proposed SARIMA model performs well. Meanwhile, if points lie outside this range we must scrap that value to get the best possible correlogram. The *x*-axis shows us the standard deviation while the *y*-axis gives the associated deviation from the origin with respect to the *y*-axis.

The final result here which is needed is a prediction of NO_2 concentration for future years; thus, getting the above plots play an important role in getting the right standard output. If the data values are not supporting the above plots then the prediction would be very vague and out of context.

Table 2 and Fig. 9 show the forecasted NO_2 concentration from Jan 2017 to Dec 2019 on monthly intervals for the proposed SARIMA, Linear Regression and SVM model. The NO_2 concentration is $\mu g/m^3$ (micrograms per metre cube). The result shows that the proposed SARIMA model outperforms in the prediction of NO_2 concentration in all years.

To further validate the proposed SARIMA model we have taken a week data (9/3/2020–15/3/2020) of Delhi from Real-time Air Quality Index, India government and compared it with the proposed model results. Table 3 shows the comparison between the predicted (SARIMA model) value and the actual observed value. The result shows that the SARIMA model is able to predict approximate similar result when compared with the observed values from 9/3/2020 to 15/3/2020.

In Fig. 10 the plot visualizes the actual prediction (represented by the blue line) and shows the one step ahead forecast (represented by the orange line) and helps in predicting the future concentration values from the year 2017 to 2019 (represented by the green line). This graph shows the time series analysis used mainly because of the continuous variable we had in our data set. The *x*-axis here represents the year while the *y*-axis gives the associated NO_2 concentration value in $\mu g/m^3$. The RMSE of the predicted value calculated as specified in Eq. 3 is 7.43 $\mu g/m^3$ for the proposed SARIMA model. This graph shows the complete time series analysis showing the plot between the observed value of NO_2 concentration and predicted values of NO_2 concentration using SARIMA model.

Table 2 Predicted NO_2 concentration in μg/m^3—for three years using proposed model SARIMA, LR and SVM

Month/year	Predicted NO_2 concentration using proposed SARIMA model	Predicted NO_2 concentration using Linear regression	Predicted NO_2 concentration using SVM
Jan-17	87.648266	81.234322	80.342422
Feb-17	85.068222	83.142321	82.123111
Mar-17	81.135937	78.131144	76.141467
Apr-17	78.922184	80.342444	80.653666
May-17	76.929068	75.345888	76.909999
June-17	72.202758	75.789788	71.6889
Jul-17	65.956946	60.352344	61.897699
Aug-17	67.491892	60.234242	61.423444
Sep-17	69.540163	65.897978	79.667868
Oct-17	76.353035	75.786888	73.786866
Nov-17	84.3807	81.780233	82.345111
Dec-17	84.95519	83.545255	81.423444
Jan-18	83.678987	81.657675	80.411678
Feb-18	82.123467	81.786999	79.087823
Mar-18	89.454221	88.546666	83.675785
Apr-18	90.783921	78.536356	80.345354
May-18	78.009422	75.682422	76.98
Jun-18	62.981131	60.127835	58.997777
Jul-18	68.78419	69.342588	70.356952
Aug-18	74.435229	72.907666	73.899796
Sep-18	75.234246	80.574799	89.687567
Oct-18	86.798734	85.345346	82.786734
Nov-18	85.248222	83.892346	82.657567
Dec-18	69.321345	65.767865	69.798798
Jan-19	70.468234	75.478234	76.23482
Feb-19	61.834561	60.872342	65.234233
Mar-19	85.780022	80.423468	81.388222
Apr-19	90.235263	98.242422	91.459645
May-19	83.856336	80.234984	79.097384
Jun-19	78.234294	79.893634	75.289634
Jul-19	80.374608	79.652342	76.497722
Aug-19	86.234904	85.234343	83.786244
Sep-19	89.234244	87.529895	86.235844
Oct-19	72.60795	70.674353	78.34955
Nov-19	78.038743	76.442344	75.897978
Dec-19	82.675345	80.732267	88.997442

5 Conclusion

Taking good care of the environment in which we live is one of the most challenging responsibilities we have as humans. The air which we breathe is constantly deteriorating to protect future generations from its harmful effects we must ensure that things do not go out of our hands. In this study, the time series forecasting is used to predict the air quality index value for Delhi in India. This work proves that the proposed SARIMA model can be used efficiently to predict the AQI value and proves its steel in this research field. Moreover, this study concludes that the SARIMA model has proved its efficiency in forecasting and has become very popular in this research field. As can be seen in Fig. 3, the value predicted has the least amount of RMSE error (7.43 μg/m^3) because of the fact that this work takes into account the current trend as well and thus helps for a better prediction model. Here SARIMA is used in the prediction of air quality depending upon the NO_2 concentration values associated with it with the dataset of 435,743 rows, achieved an RMSE of

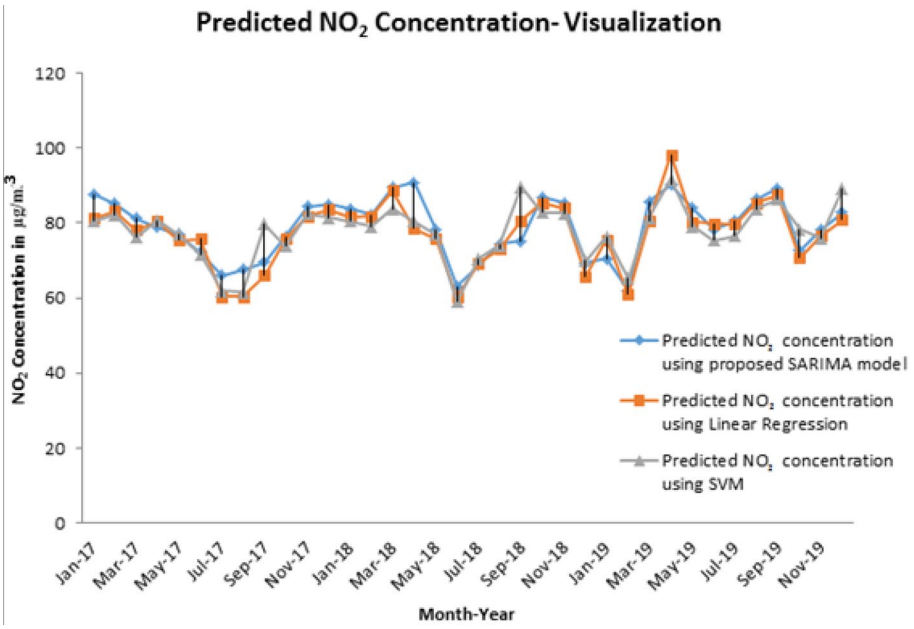

Fig. 9 Predicted NO_2 concentration using proposed SARIMA, LR and SVM model

Table 3 Predicted NO_2 concentration vs. observed values in µg/m^3

Date	Predicted NO_2 concentration using proposed SARIMA model	Actual NO_2 concentration observed
09/03/2020	76.342598	79.428398
10/03/2020	79.561276	81.137397
11/03/2020	75.129634	78.137935
12/03/2020	81.248561	80.185239
13/03/2020	77.126512	80.547239
14/03/2020	85.874521	82.912241
15/03/2020	83.592718	79.439211

7.43 µg/m^3. The result shows that SARIMA performs better than previous models such as SVM and Linear regression, which gave the RMSE of 9.8 µg/m^3 and 15.2 µg/m^3, respectively. Later, NO_2 predicted values shows that the proposed SARIMA model out performs when compared with linear regression and SVM for the year Jan 2017 to Dec 2019. Finally, the proposed model result is compared with the one-week real-time data (9/3/2020–15/3/2020), and it has proven that the achieved result are approximately similar to the actual one. In summary, the time series analysis model is an important tool used in the field of forecasting and helps in controlling and monitoring the

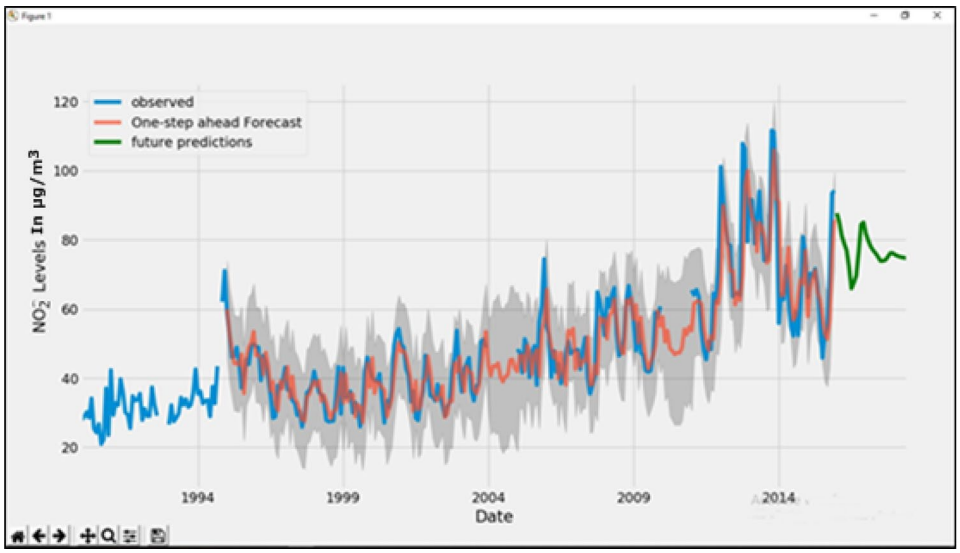

Fig. 10 Predicted NO_2 concentration for 2017, 2018, 2019 years using proposed SARIMA model

air quality conditions. It is useful to take quick action well before the situation worsens in the long run.

Funding None.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2016) Time series analysis: forecasting and control. Wiley, New Jersey

Brownlee J (2019) How to identify and remove seasonality from time series data with Python. Machine learning mastery. https://machinelearningmastery.com/time-series-seasonality-with-python. Accessed 20 Oct 2019.

Cunningham WP, Cunningham MA (2002) Principles of environmental science inquiry and applications. McGraw Hill Company, New York

DataGov (2019) Air quality. https://data.gov.in/dataset-group-name/air-quality. Accessed 10 Oct 2019.

Deviant S (2011) The practically cheating statistics handbook. Lulu.com.

Essallah S, Bouallegue A, Khedher A (2015) Performance evaluation of Box-Jenkins and linear-regressions methods versus the study-period's variations: Tunisian grid case. In 2015 IEEE 12th International multi-conference on systems, signals and devices (SSD15) IEEE. https://doi.org/10.1109/SSD.2015.7348153

Ganesh SS, Modali SH, Palreddy SR, Arulmozhivarman P (2017) Forecasting air quality index using regression models: a case study on Delhi and Houston. In 2017 International conference on trends in electronics and informatics (ICEI) IEEE. https://doi.org/10.1109/ICOEI.2017.8300926

Gu K, Qiao J, Lin W (2018) Recurrent air quality predictor based on meteorology-and pollution-related factors. IEEE Trans Industr Inf. https://doi.org/10.1109/TII.2018.2793950

Gupta P, Kumar R, Singh SP, Jangid A (2016) A study on monitoring of air quality and modeling of pollution control. In 2016 IEEE Region 10 humanitarian technology conference (R10-HTC) IEEE. https://doi.org/10.1109/R10-HTC.2016.7906800

Yu C (2016) Research of time series air quality data based on exploratory data analysis and representation. In 2016 Fifth International conference on agro-geoinformatics (Agro-Geoinformatics) IEEE. https://doi.org/10.1109/Agro-Geoinformatics.2016.7577697

Jain A, Abbas B, Farooq O, Garg SK (2016) Fatigue detection and estimation using auto-regression analysis in EEG. In 2016 International conference on advances in computing, communications and informatics (ICACCI) (pp. 1092–1095). IEEE. https://doi.org/10.1109/ICACCI.2016.7732190

Kang GK, Gao JZ, Chiao S, Lu S, Xie G (2018) Air quality prediction: big data and machine learning approaches. Int J Environ Sci Dev 9:8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066

Plaia A, Ruggieri M (2011) Air quality indices: a review. Rev Environ Sci Biotechnol 10:165–179. https://doi.org/10.1007/s11157-010-9227-2

Putra BU, Nhita F, Saepudin D, Wisesty UN (2017) An implementation of weighted moving average and genetic programming for rainfall forecasting in Bandung Regency. In 2017 International conference on control, electronics, renewable energy and communications (ICCREC) IEEE. https://doi.org/10.1109/ICCEREC.2017.8226674

Qi Z, Wang T, Song G, Hu W, Li X, Zhang Z (2018) Deep air learning: interpolation, prediction, and feature analysis of fine-grained air quality. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2018.2823740

Soh PW, Chang JW, Huang JW (2018) Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access 6:38186–38199. https://doi.org/10.1109/ACCESS.2018.2849820

NAAQS table (2016). United states environment protection agency. https://www.epa.gov/criteria-air-pollutants/naaqs-table. Accessed 15 Oct 2019

WHO (2020). Air pollution. https://www.who.int/health-topics/air-pollution#tab=tab_1. Accessed 20 Oct 2019.

World Economic Forum (2020) The best and worst—countries for air pollution and electricity use. https://www.weforum.org/agenda/2017/02/the-best-and-worst-countries-for-air-pollution-and-electricity-use. Accessed 20 Oct 2019.