ISSN No:-2456-2165

Short-Term Forecasting of PM 2.5 Using HybridAlgorithm

[1st] Rafay Malik
Mechanical Engineering
DepartmentNED University
Karachi, Pakistan

[2nd] Muhammad Mudassir
Department of EPE PN
Engineering College
Karachi, Pakistan

[3rd] Hasnain Aslam
Department of EPE PN
Engineering College
Karachi, Pakistan

Abstract:- At present time, the forecasting of air particles has been a current research topic due to increase in bad air quality because of industrialization and increase in pollution due to vehicle and in COVID-19 when there was lockdown it came in our notice that this thing can be controlled its effects which was highlighted due to COVID 19 lockdown period. It includes a various source of contamination, making it hard to decide the entirety of the contributing meteorological and ecological components. At the point when just the PM 2.5 fixationtime sequences are taken without other external data, precise estimating is significant and productive. For address, this issue this paper presents the ARIMA based model for forecast PM 2.5 data concerning time. In this paper, two methods are proposed. First dividing the data (80 % training) and (20 % testing) then decomposes one-dimensional data through wavelet decompositionof level-2 dB2. Then, it uses ARIMA model method to forecast each divided sequel and reconstruct its predicted results to obtainthe finalize predicting outcomes. Second without decomposes the data, we directly apply the ARIMA model and forecast the results. The ARIMA model has forecast more accurate results concerning predicting the concentration of PM 2.5 as compare to the WAVELET-ARIMA model. The two proposed ARIMA and Wavelet ARIMA can be efficiently applied to forecastingPM 2.5 concentration in short term and can be enhancing the accuracy. Moreover, relating the forecasted results with the policygoverning to control the pollution as shown by implementinglockdown as PM 2.5 value has been reduced up to 50% in different cities during the lock down period which can be seen from study.

Keywords:- *Particulate Matter 2.5, Auto Regressive Integrated Moving Average, Mean Absolute Value, Weather Research and Forecasting, Carbon Matter, Elemental Carbon, Root Mean Square Error.*

I. INTRODUCTION

During recent times, the prediction of particles with a diameterof less than 2.6 μm or a smaller amount.PM 2.5 is very important concern for many reasons because it is very dangerous. It is harmful for lungs and has potential to cause lung cancer. Asthma related diseases difficulty in breathing. It may cause plaque deposits in arteries if exposed to it for Long-term. It may cause hardening of the veins/ arteries which can finally lead to heart attack and.Visibility is reduced haziness is increased when PM 2.5 in outdoorair increase. Due to the said reasons, it is really important topic and the research on the forecasting of PM 2.5 concentration has been the main focus of air-quality research and the accurate prediction of PM 2.5 concentrations is most important and is one of the key goals. “Itis very important to establish a high-precision PM 2.5 concentration prediction model for control and monitoring.” [1]. When during the study of air quality, we are considering the PM 2.5 concentrationtime series/ values without taking in account other parameters which effects the concentration of PM 2.5 exogenous information, accurate prediction is important and should be highly efficient. To addressthis problem, in this study we have proposed two models and we will check the accuracy of the algorithm how well is the prediction accuracy.

PM 2.5 (particles with a diameter of 2.5 micrometres) are made upof carbon matter (CM), elemental carbon (EC), and other inorganic compounds [2] that are harmful to human health [3]. Because of the health concerns, air quality has been a major focus, and it can be better controlled if accurate forecasting is done, which is the main purpose of the Air pollution and control action plan (APPCAP). Be- cause PM 2.5 is detrimental to humans, creating a high-precision PM 2.5 model for control and monitoring is critical [4]. The predictionof particles with a dia of 2.5 μm or smaller (PM 2.5) has becomea important study area in recent years. Water-soluble ions with Elemental Carbon (EC), Organic Carbon Matter(OCM), and other inorganic compounds make up the majority of PM 2.5 with a diameterof 2.5 *mu*m or less which are extremely hazardous to human health.It involves numerous components making it difficult to identify the primary factors that influence metrological and environmental aspects. Due to increased public concern, air-quality research has been focusing on the correct forecast of PM 2.5 conc, which is oneof the primary

ISSN No:-2456-2165

aims of the 2017 Air Pollution Prevention and Control Action Plan (APPCAP) [5]. Because PM 2.5 is dangerous to humans, a high-precision PM 2.5 conc prediction model for control and monitoring is critical. Furthermore, government agencies can issue more scientific warnings based on the outcomes of forecasting severe air pollution situations. For prediction and analysis, the Weather Research and Forecasting WRF-CMAQ air quality modelling system is often used [6]. Furthermore, PM 2.5 is predicted using principle component analysis (PCA). Another technique is the Autoregressive Integrated Moving Average (ARIMA) model, which may be used to anticipate PM 2.5 concentrations and is well suited to these types of time series. China improved its PM 2.5 prediction using this model [9]. Other models, such as Artificial Neural Networks (ANN), can forecast the Pm 2.5 values. It has proven a useful technique for predicting nonlinear events in the environment, such as PM 2.5, ozone prediction, SO2 concentration, and other factors. Furthermore, the support vector machine has a high prediction capability as well as the ability to adapt to changes and incorporate them into the model. GIS with statistical techniques is the best technique to find trend in pollution level [9]. Although ARIMA, ANN and SVM have all showed best results for predicting the data but out of these ARIMA has the least error in short term forecasting [10]. Thus the focus of this paper is to do forecasting of the PM 2.5 data and then comparing that data with the actual values obtained. Once that thing is done the accuracy of the system can be found using residue error square. Now we will forecast using 2 techniques firstly we will use directly ARIMA on model and then find the results and secondly Wavelet decomposition will be used to decompose 1-D data into several segments and then individually ARIMA will be applied and results will be forecasted this technique is called Wavelet-ARIMA Model. Moreover, the paper is focusing on the policy making by using the forecasted data.

II. AREA DESCRIPTION

Air quality is basically the function of population the higher population the more the pollution, In Pakistan the urban area is taken to the most seriously poisonous pollution issues. Pakistan cities are in top 10 worst air quality of the world so this shows it is serious concern for Pakistan to work on and Pakistan come on 2nd position in world in the list of deaths cause by bad air quality (World economic forum). This paper have selected four representative cities of Pakistan which are populated the most and are economic hub of Pakistan and collected the data for those cities and worked on them for forecasting of the result of PM2.5 values. The cities includes Karachi the industrial hub of Pakistan; Islamabad the capital of Pakistan; Lahore the provincial capital of biggest province in terms of population of Pakistan; and Peshawar the major city of Pakistan the geographical location is mentioned in Fig 1.

Fig. 1. Map of PM 2.5 Value in Pakistan.

III. DATA-SET COLLECTION

The PM 2.5 data implemented was taken from website [**?**]. The PM2.5 were recorded using RP1400a auto sample which can automatically realize real time data and sample the data in real time. The calibration of the system can be done using the monitoring analysis method. The period for the measurement is on hourly basis for one day it takes 24 values of PM2.5 and the data we have used for different cities are different minimum of 1 year has been used for all cities.

ISSN No:-2456-2165

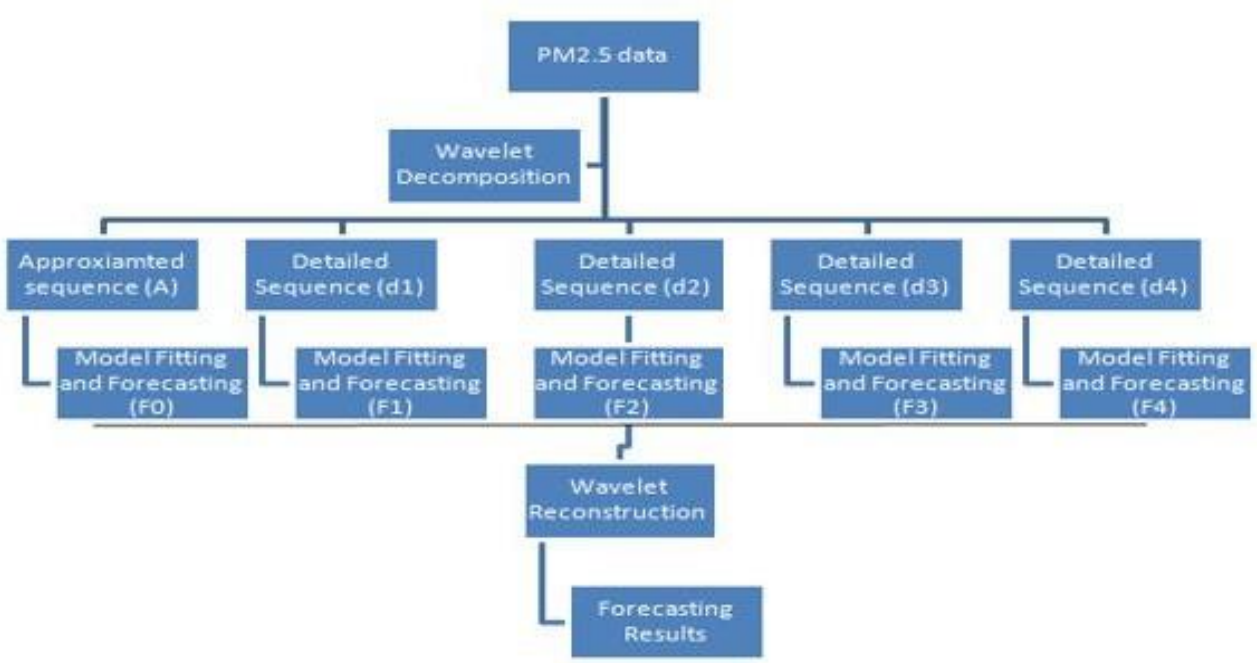

Fig. 2. Different Cities to be consider.

IV. FORECASTING ALGORITHM AND METHODOLOGY

ARIMA Model has been used in this study with two approaches once it is use directly on data and once with wavelet decomposition was proposed to prediction accuracy of 1D time series with both the approaches. The assumptions with ARIMA model is that the data is non-stationary and that can be tested using Augmented Dickey fuller (ADF) to give explanation for the stationary. Matlab is a programminglanguage which we are using for the construction of models. The key steps are to collection of data then breaking data into sub time series by wavelet decomposition. Following are the steps are as follows:

Input: PM 2.5concentration time series.
Output: Forecasting of the average concentrations of PM 2.5data on hourly basis and estimate of the correctness of the models.

Step 1: Standardize the actual data i.e. the data should be in form to apply ARIMA(the time series of PM 2.5daily average concentration data).
Step 2: Selection of best possible wavelet function for input.
Step 3: Decompose into approximation and detail sequences original1D data.
Step 4: Build and forecast data set for each sequence independentlyby using ARIMA to fit which is a traditional prediction model.
Step 5: Now for composing the final forecasting outcome, recreatethe predicted results of each data sequence.
Step 6: Weigh up the prediction accuracies of the both models

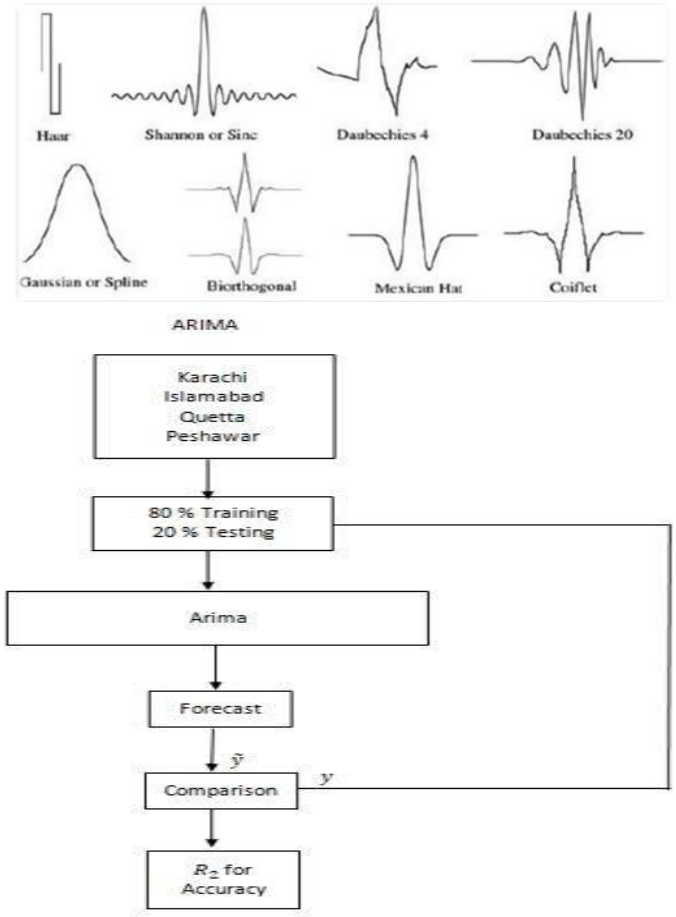

Fig. 3. ARIMA Model Work Methodology.

ISSN No:-2456-2165

V. SELECTION OF WAVELET

There are different wavelets such as Daubechies, Symlet, Meyer, Morlet, etc and the choice of the mother wavelets depends the data available. The Daubechies wavelet transforms is the closed which matches with our data provided. In this paper we have used db2 levelfour. This gives the least error among all wavelets (db3, db4, db5, db7) tested for our data series. There are different wavelets that can be used for the decomposition of signal but the selection of appropriate depends on the signature of actual signal we have to compare the dataoutput signature compare that with the type of signature available as shown in figure below for our signal the closest signal that matches is Daubcchies signal so for wavelet decomposition Db2 is used and the order to which it has to be decomposed can be found by formula given below we have decomposed till 4th order. The higher levels of Jmax can be decomposed, can be evaluated using the equation below:

$$Jmax = fix(log2(N/Nw)) – 1) \quad (1)$$

Where N represents the length of signal Nw represents the length of the decomposition filter associated with the chosen mother wavelet.

Wavelet decomposition results are shown in figure below whichdecomposition 4th level using DB2 type wavelet.

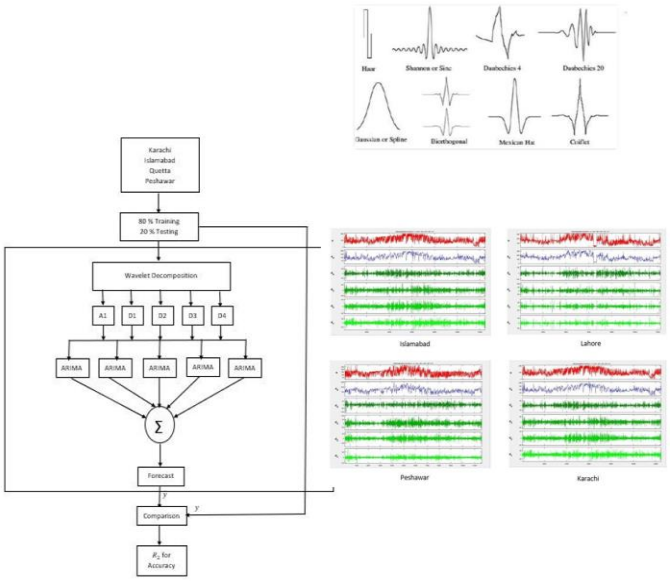

Fig. 4. ARIMA-WAVELET Model Work Methodology.

Wavelet decomposition has been applied on all data sets of four cities (Karachi, Islamabad, Peshawar, Lahore) the data was decomposed into one approximation sequel (A4) which tells the overall trend of the original input signal and detailed series (d1, d2, d3, d4) tells noise of time series. Statistics includes the mean, median, Standard deviation and skewness of PM 2.5 concentration ofeach decomposed series. The mean and median gives us information regarding the central tendencies of data similarly standard deviation dictates degree of dispersion.

Cities	Statistic	S	A4	D1	D2	D3	D4
Karachi	Mean	111.4	111	0.0001	-0.00019	0.0003	0.0018
	Median	100.84	100.8	0.004	0.0193	-0.016	0.133
	Standard Deviation	38.70	38.7	3.2	5.52	9.009	12.6
	Skewness	0.553	0.55	0.6	0.06	0.066	-0.38
Islamabad	Mean	106.0	106	-7.4e-4	7.6e-4	0.0029	0.0067
	Median	96	99.8	0.02	-0.0129	0.1767	-0.4464
	Standard Deviation	39.8	34.7	4.5	7.9	10.56	13.5
	Skewness	0.46	0.42	0.063	0.1811	-0.06	0.233
Lahore	Mean	155.59	155.5	8.8e-5	9.4e-5	-1.4e-4	6.3e-4
	Median	155	152.3	9e-6	0.0056	-0.025	0.1078
	Standard Deviation	66.6	60.38	6.04	9.52	14.01	21.7
	Skewness	0.53	0.46	0.26	0.23	0.31	0.0086
Peshawar	Mean	134.6	134.5	8.4e-4	-8.6e-4	-8.7e-4	-0.0044
	Median	139	132.4	9.6e-8	-1.5e-15	-0.26	-0.1044
	Standard Deviation	49.4	41.4	5.94	10.4	16.82	17.24
	Skewness	0.35	0.51	0.3	-0.13	0.2134	-0.036

TABLE I STATISTICAL RESULTS OF DECOMPOSED WAVELET

ISSN No:-2456-2165

Cities	parameters	A4	D1	D2	D3	D4
Karachi	P	P	5	4	4	5
	D	D	0	1	1	1
	Q	Q	0	4	4	5
Lahore	P	P	5	4	4	5
	D	D	0	1	1	1
	Q	Q	0	4	4	5
Islamabad	P	P	5	4	4	5
	D	D	0	1	1	1
	Q	Q	0	4	4	5
Peshawar	P	P	5	4	4	5
	D	D	0	1	1	1
	Q	Q	0	4	4	5

TABLE II OPTIMIZED P,Q AND D VALUES.

VI. ACCURACY EVALUATION

The ”Absolute error average (MAE), Root mean square error (RMSE), and coefficient of determination (R2)” were used to de- termine the model’s quality. The equations are shown in Equation2,3 below. Where yi denotes the measured value at sample I *hat yi* denotes the sample I predicted value, y bar is the measured value’s average value, and n denotes the number of samples. The smaller the error, the lower the MAE and RMSE number, and the higher the R2 value, the better the data fitness.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}|y_i - \hat{y}_i| \qquad (2)$$

Equation 4 below shows the entire prediction. ”Following the convention established by Box and Jenkins,” the transference of ”average parameters is (*theta*’s)” is supplied here such that their symptoms are negative in the equation. As an alternative, some writers and software (like the R programming language) use plus marks. While there is no confusion when actual data are entered into the calculation, knowing which convention your software uses when analysing the outcome is crucial. AR(1), AR(2), and MA(1), MA(2) are common abbreviations for the parameters (2). [12] ”You beginby determining the order of differencing (d) required to stationarily the series and remove the gross features of seasonality, maybe in conjunction with a variance-stabilizing treatment such as logging or deflating” [13]. You have fitted a random walk, which is not fitted, ifyou stop here and anticipate that the differences series would remain constant.

$$\hat{y_t} = \mu + \emptyset \times y_{t-1} - \mu_p \times y_{t-p} - \theta_1 \times e_{t-1} - - - - - \theta_q \times e_{t-q}$$

(4)

The Wavelet ARIMA and ARIMA are used to forecasting of data of PM 2.5 concentrations and Comparison of forecasted data of both models. 80% data are used for training of both the model and 20 percent data is used for testing. The sample was decomposed by wavelet function into four detail sequence and one approximation sequence. The ADF unit root test method is for testing the Non- Stationarity of data. The result showed that the data is non stationary (p ¡ 0.001). LM test was carried out to test whether the noise values consist of white noise. After getting all results positive the time seriesis forecasted and final forecasted results are obtained.

VII. SIMULATION RESULTS

More over in Figure 5 to 12, graphs are shown for all four cites in which first 250 forecasted values have been compared with the actualvalue for both the models that is ARIMA and Wavelet-ARIMA of respective cities and it can be seen that it is following the data invery well manner. The comparison or both model is shown belowin Figures 13 to 16 for ARIMA model and Figure 17 to 20 for Wavelet- ARIMA. The correlation has been shown in the figures. Scatter plot given below shows the potency of a linear association between the modeled value (y) and the observed value (x) which is in good agreement with the data.

$$RMSE = \sqrt{\frac{1}{n}}\sqrt{\sum_{t=1}^{n}\left(\frac{y_i - y_i}{\sigma_i}\right)^2} \qquad (3)$$

VIII. WAVELET-ARIMA MODEL

Once the wavelet decomposition it done the next step is to apply ARIMA model on the decomposed training data. After decomposition5 sub- time series are obtained including one approximation sequence(A4) and 4 detailed sequences (D1), (D2), (D3) and (D4) in orderto make the forecasting individual ARIMA is applied on each five sequence and then the output is summed to get one signal that signal isused to forecast the output. In ARIMA there are 3 input parameters autoregressive order (P), degree (D) and Moving Average (Q) by changing the values of these parameters the output changes and goodness of fit changes and it can be regulated by checking the AIC and BIC values the lower these values the better is the model fitting by testing best fit values of p, q and d are obtained and given in Table 2.

ISSN No:-2456-2165

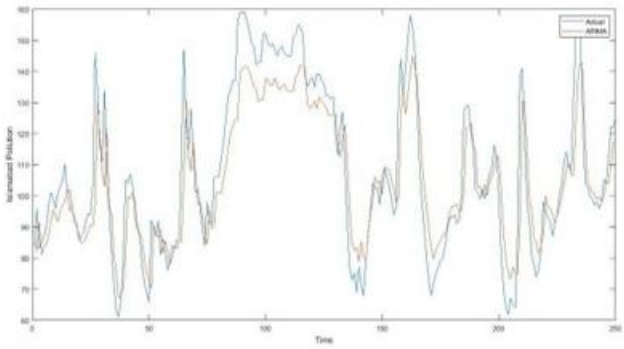

Fig. 5. Forecasted Results for Islamabad using ARIMA.

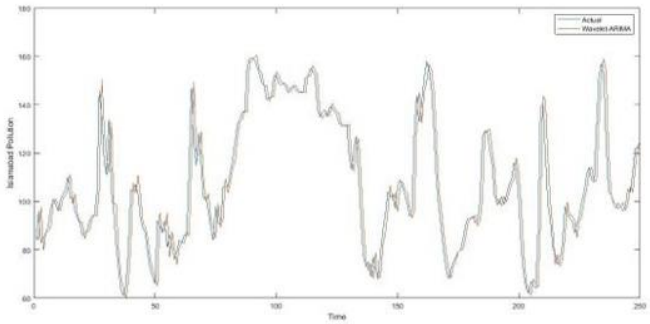

Fig. 6. Forecasted Results for Islamabad using Wavelet-ARIMA.

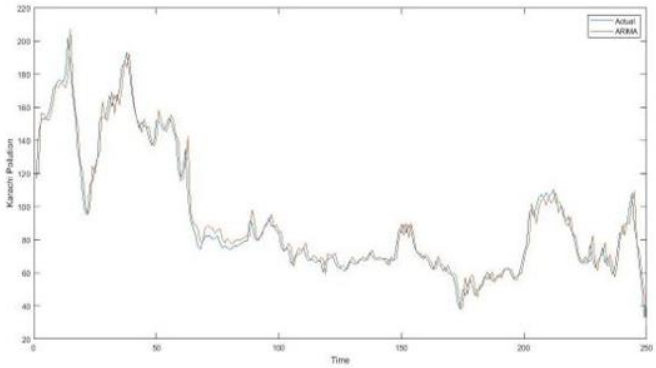

Fig. 7. Forecasted Results for Karachi using ARIMA.

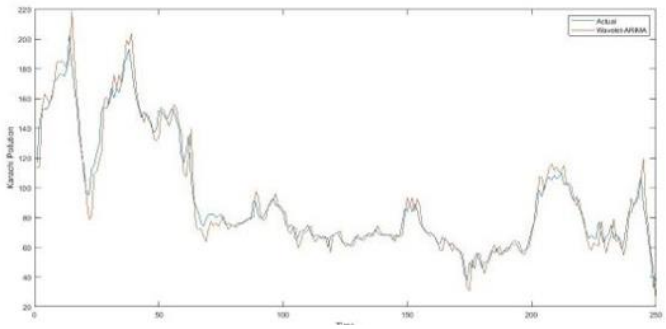

Fig. 8. Forecasted Results for Karachi using Wavelet-ARIMA.

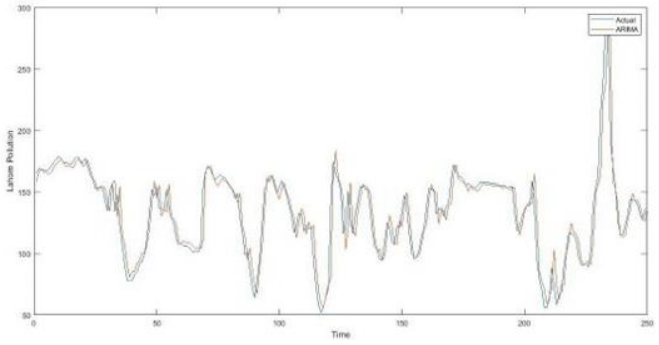

Fig. 9. Forecasted Results for Lahore using ARIMA.

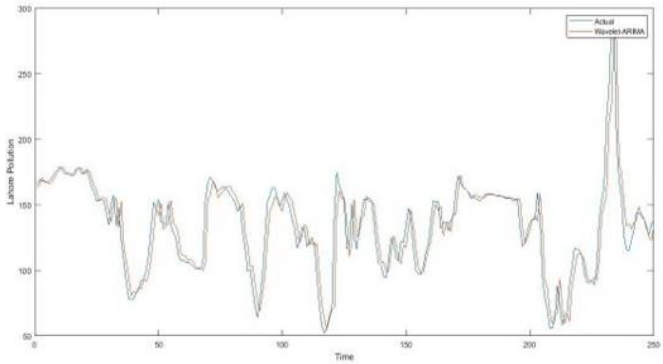

Fig. 10. Forecasted Results for Lahore using Wavelet-ARIMA.

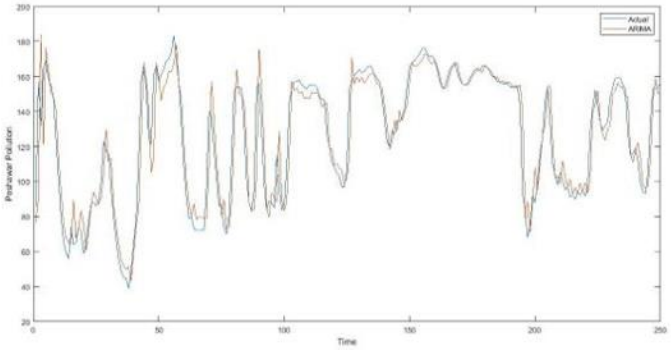

Fig. 11. Forecasted Results for Peshawar using ARIMA.

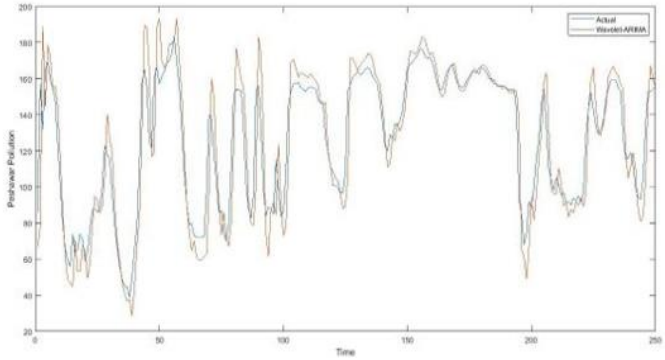

Fig. 12. Forecasted Results for Peshawar using Wavelet-ARIMA.

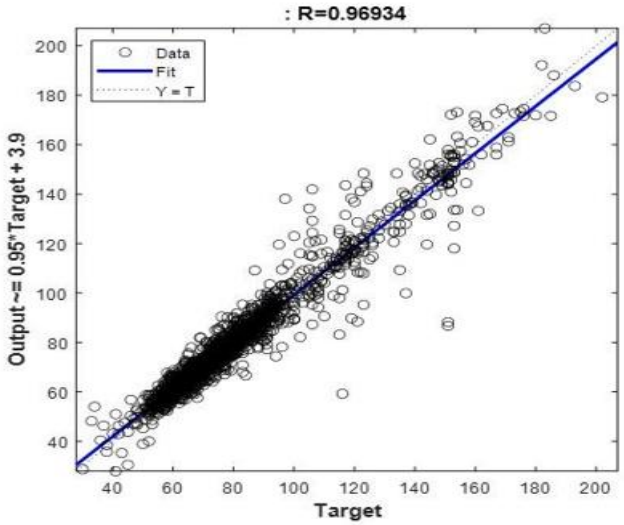

Fig. 13. Scattered Plot ARIMA Karachi.

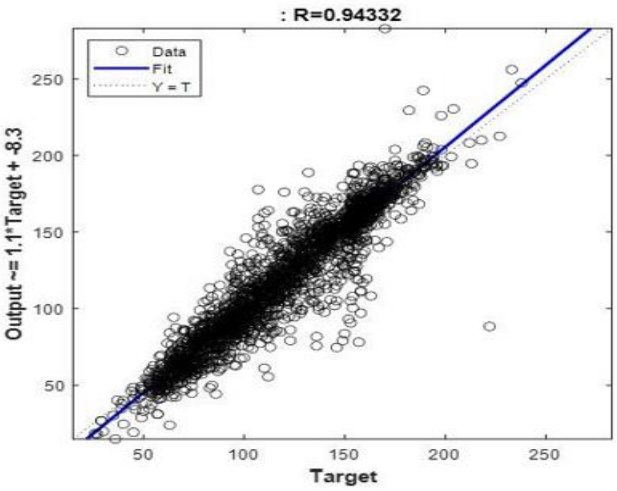

Fig. 16. Scattered Plot Peshawar Karachi.

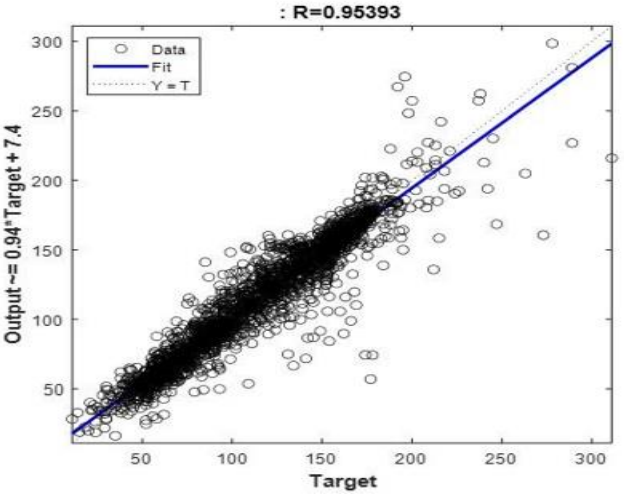

Fig. 14. Scattered Plot ARIMA Islamabad.

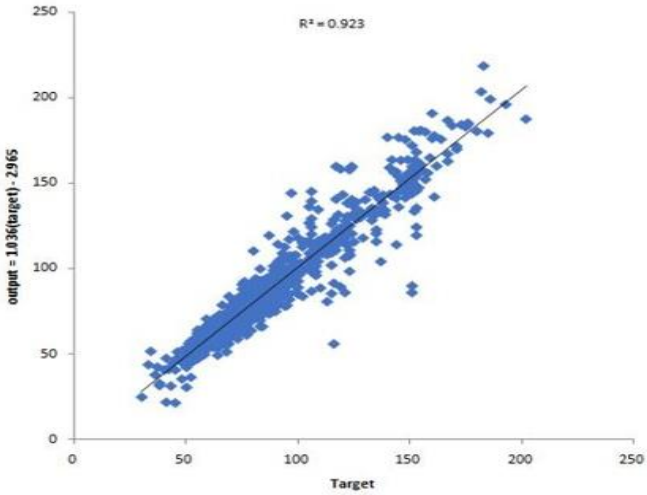

Fig. 17. Scattered Plot Wavelet-ARIMA Karachi.

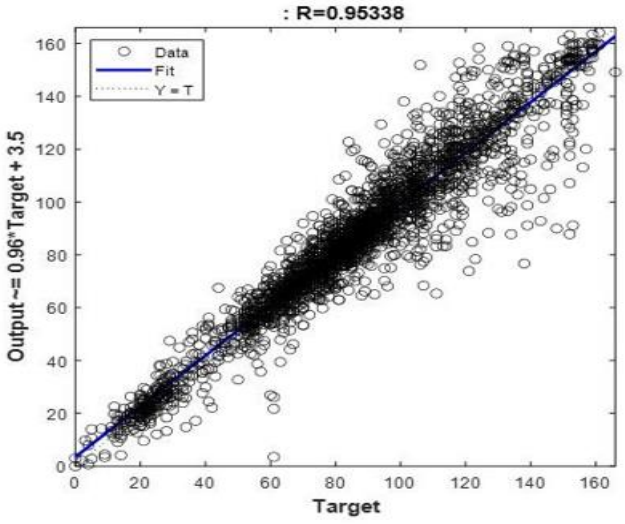

Fig. 15. Scattered Plot ARIMA Lahore.

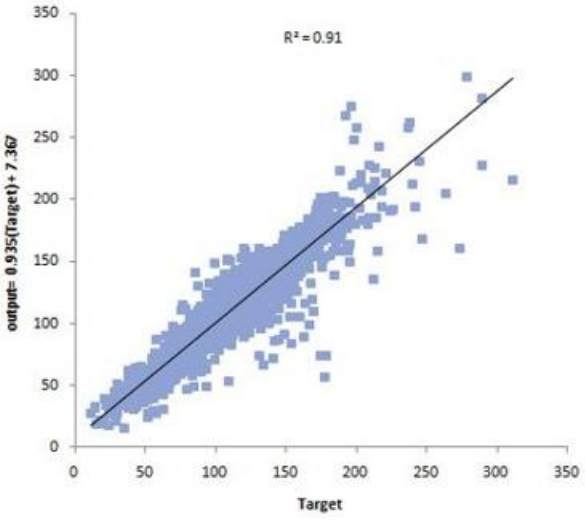

Fig. 18. Scattered Plot Wavelet-ARIMA Lahore.

ISSN No:-2456-2165

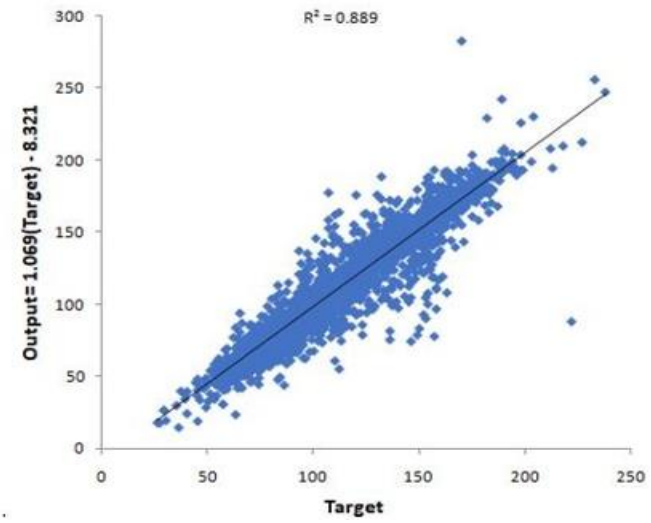

Fig. 19. Scattered Plot Wavelet-ARIMA Peshawar.

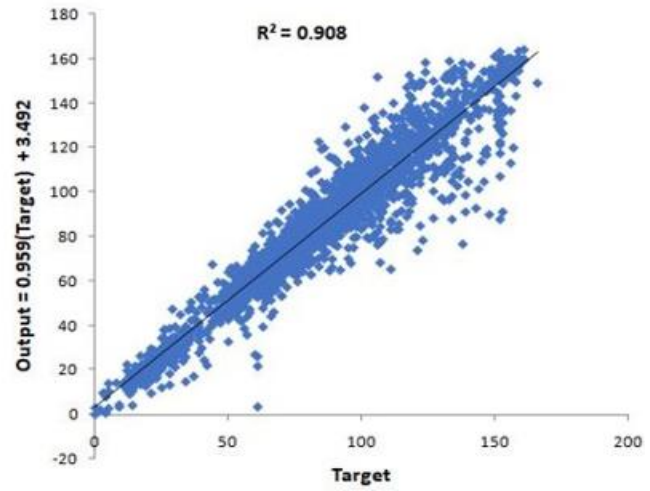

Fig. 20. Scattered Plot Wavelet-ARIMA Islamabad.

IX. CONCLUSION

In this study two models were developed that are ARIMA and Wavelet-ARIMA and short term forecasting is carried. PM 2.5 concentration of four cities of Pakistan by using both the models.The Accuracy of the models is checked using MAE, RMSE andR2 method and it is found out that the accuracy of ARIMA model is to be more than the Wavelet-ARIMA model. The results showed that the ARIMA model is superior to the Wavelet-ARIMA model in forecasting results in Pakistan as while choosing the order of wavelet decomposition we have been very concerned we used both the ways by using the formula as well as by checking theAIC/BIC value for other levels and then selecting the best order. When we compared the results the Least AIC/BIC was coming at 4th order decomposed signal, and the ARIMA model is coming out to be better modelfor our time series which is contrary to our previous knowledge by Yong Chenga (2018). The larger the amount of original data, the moreaccurate will be the prediction. The PM 2.5 concentrations mainly depends on two factors the spatial and temporal scale. To capture the results in better way i.e. the changes in the PM 2.5 concentration,we have taken the data hourly values and that is the main pointthat even without decomposition of data the ARIMA model directly implemented on the data is giving us more good results as it iscapturing all hidden information from the data. This can be improvedfurther if the data is taken on lower scale.

REFERENCES

[1] Zhou, J., Xing, Z., Deng, J., Du, K., 2016. Characterizing and sourcing ambient pm 2.5, over key emission regions in China i: water-soluble ions and carbonaceous fractions. Atmos. Environ. 135, 20–30.
[2] Yan, J., Lai, C.H., Lung, S.C., Chen, C., Wang, W.C., Huang, P.I., et al., 2017. Industrial PM 2.5 cause pulmonary adverse effect through rhoa/rock pathway. Sci. Total Environ. 599–600, 1658–1666.
[3] Tai, A.P.K., Mickley, L.J., Jacob, D.J., 2010. Correlations between fine particulate matter (PM 2.5) and meteorological variables in the United States: implications for the sensitivity of PM 2.5 to climate change. Atmos. Environ. 44 (32), 3976–3984.
[4] Weimiao, Q., Jing, C., Juncai, H., Xinghong, C., Xiaomin, W., 2017. Research on a nonlinear forecast method based on wrf-cmaq model. Environ. Sci. Technol.
[5] Wang, Y., Wang, C., Shi, C., Xiao, B., 2018. Short-term cloud coverage prediction using the ARIMA time series model. Remote Sensing Letters 9 (3), 275–284.
[6] Chenga, Hong Zhanga, Zhenhai Liua, Longfei Chena, Ping Wang 2019 Hybrid algorithm for short-term forecasting of PM 2.5 in China Yong Atmospheric Environment 200 (2019) 264–279
[7] Xue, D., Liu, Q., 2014. Prediction of surface so2 concentration in shanghai using artificial neural network. Appl. Mech. Mater. 522–524, 44–47.
[8] Shakeel , Arshad, Saeed, Ahmed, Khan HMT, Noreen, Ali and Munir Application of GIS in Visualization and Assessment of Ambient Air Quality for SO2 and NOx in Sheikhupura City, Pakistan
[9] Suling Zhu, Xiuyuan Lian, Haixia Liu, Jianming Hu, Yuanyuan Wang, Jinxing Che 2017 Daily air quality index forecasting with hybrid models: A case in China
[10] Yu-Fei Xing, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Lian 2016 The impact of PM 2.5 on the human respiratory system
[11] Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter, External Review Draft
[12] World economic forum
[13] Yu-Fei Xing, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Lian 2016 The impact of PM 2.5 on the human respiratory system