ISSN:2790-1688 Volume-14-(2025)

Comparative Analysis of ARMA and ARIMA Models for Air Quality Prediction

Zhenning Yan

Ningbo Yinzhou Lanqing School, Ningbo, 315000, China

Abstract. Accurate prediction of urban air pollution is crucial for environmental management and public health. In this study, two classical time series models were comparatively analyzed to simulate and predict the concentrations of PM2.5, NO2, and O3 in Shanghai from 2019 to 2021. Based on the Interquartile Range method, we performed data preprocessing. The AutoRegressive Moving Average and the AutoRegressive Integrated Moving Average models were applied to predict the air pollutants. We used various evaluation metrics to compare the model performance. The conclusion shows that ARIMA is more suitable for air pollution applications as it captures the nonlinear trend of the data through differential methods.

Keywords: Air pollution forecasting; ARMA; ARIMA; Time series modeling.

1. Introduction

Over the past decades, urban air pollution has become a pressing global problem with far-reaching implications for public health, environmental sustainability, and climate dynamics. To measure air quality, a number of existing indicators, such as fine particulate matter (PM2.5), nitrogen dioxide (NO2), and ozone (O3) are the main pollutants whose monitoring data can help us to have a more quantitative understanding of their dynamics [5,19,20]. Whichever of these indicators is mentioned above, will exhibit complex behavior at different times. Specifically, they embody significant seasonality, stochastic volatility, long-range dependence, and non-stationary structural changes. These characteristics arise from multiple interacting factors, including anthropogenic emissions, meteorological variations, chemical transformations, and atmospheric transport processes. With increasing urbanization and evolving environmental regulations around the globe, there is a need to more accurately detect and predict these dynamics. In order to balance the interpretability of the detection process and the accuracy of the prediction results, rigorous mathematical modeling is a suitable approach.

In the process of modeling time series data of air quality mathematically, three types of classical modeling systems have been established by previous generations. The first category is the classical statistical model, which has a complete mathematical analysis principle and a clear and transparent model structure and is trusted by scholars. For example, the AutoRegressive Moving Average model (ARMA) [3], the AutoRegressive Integrated Moving Average model (ARIMA) [2], and the Seasonal AutoRegressive Integrated Moving Average model (SARIMA) [4]. This type of model is widely used in trend prediction and long and short-term fluctuation modeling of air quality. For example, Gao used the ARIMA model to predict air pollutants in several cities in Hunan Province [1]. After finite difference processing, the model results can be evaluated with an accuracy of more than 0.8. However, the model also has drawbacks, such as insufficient ability to recognize periodic pollution indicators. Therefore, some scholars have improved the original model and proposed a seasonal ARIMA model with a stronger response to cyclical fluctuations, i.e., the SARIMA model. Zhang and Wang analyzed the AQI indicators in Beijing in 2016, which effectively captured the cyclical fluctuations in the time series data [6].

Due to the fact that the actual data on air pollution is often multidimensional and has a nonlinear structure, the traditional model has a performance bottleneck for this. Therefore, scholars purposefully propose a second class of hybrid models, which are usually a combination of two underlying models. This type of model often makes up for the shortcomings of one model by another model, i.e., applying hybrid strategies to improve the robustness and fitting accuracy of the model.

For example, AutoRegressive Integrated Moving Average – Artificial Neural Network model [7,8,9], if the time series data has a complex nonlinear structure and changes drastically, or the data is contaminated. Then the residuals after analyzing through the ARIMA model are likely to still have a nonlinear relationship. The ANN neural network model can well extract residual information and improve the data fitting accuracy. For example, Wang used the ARIMA model to fit the PM2.5 time series data in the Guangzhou area after 2021 [10]. The residuals are put into the neural network to learn the complex nonlinear part of it. After comparison, the statistical assessment indexes such as MAE and RMSE of this method are lower than a single data analysis model, and it achieves a satisfactory fitting performance.

As the development of the age, there are increasing numbers of machine learning models being used in the field of environmental statistics. In particular, deep learning neural network models have attracted the attention of scholars for their advantages in time series forecasting. Different from traditional statistical modeling approaches, deep neural networks can autonomously learn autocorrelations and complex nonlinear relationships in time series data [11,12]. The advantage of machine learning approaches is that they do not require any assumptions, such as ARMA models that require the data to be sufficiently smooth. Sharma and Singh in 2023 compare the performance of Long Short-Term Memory Networks (LSTM), RIMA, and Exponential Smoothing Models in forecasting of AQI time series data [13,14]. The LSTM model was measured to be robust when the data is volatile, such as when natural disasters, industrial pollution, etc. occur. The LSTM model is especially suitable for short-term high-frequency data, which will achieve better results. However, machine learning models always have the disadvantage of low interpretability, which is also the price of autonomous learning. They are more sensitive to small sample data, and the accuracy of their results is easily disturbed when insufficient air quality data can be collected. Most of the neural network models require a significant amount of training time, and their principles demand further mathematical explanations to be given.

Based on the characteristics of the above multiple time series forecasting models, the ARMA and ARIMA models, which have the clearest mathematical principles among them, are selected in this paper to analyze the air quality data of the Shanghai Municipality in China. The purpose of the study is to mathematically model and forecast the trend of urban air quality in Shanghai, China, and the characteristics of long- and short-term fluctuations. For example, the smoothness, seasonality, and development trend of air quality indicators such as PM2.5 and other air quality indicators. In addition, the difference in the performance of mathematical models ARMA and ARIMA in modeling different pollutants is also one of the research objectives of this paper. In order to achieve the above research objectives, this paper firstly describes and analyzes the ARMA and ARIMA models in Section 2, and then collects the air quality data of Shanghai in the past three years. After that, the air quality monitoring data of Shanghai for the last three years, such as daily and hourly time series data of PM2.5, NO2, and O3, are collected. The data are preprocessed, e.g., dealing with missing values, outliers, etc. The visualization results of their descriptive statistics are given in Section 3. After that, the paper analyzes the time series data for smoothness and fits it using the ARMA model before differencing. After obtaining the valid results, the time series data are further analyzed using the ARIMA model, which has included the differentiation step. Finally, we compare the results of the two models for different pollutants and different differencing steps and discuss the evaluation of the results. The results of this paper can provide data support for the environmental regulatory authorities in Shanghai, methods for pollution early warning and early intervention, and a basis for the government in policy formulation and management.

2. Models

In environmental statistics, ARMA and ARIMA are two trusted classical models [15-16]. They are different mainly because of the presence or absence of a different step. This depends on whether

the data analyzed are smooth or not. We first give the mathematical analysis of the two models in this section.

2.1 ARMA

Autoregressive Moving Average (ARMA) model, as a classical time series model, is often used to characterize the statistical properties of time series data and forecast future trends. ARMA is a combination of two sub-models that combine the features of Autoregressive (AR) [17] and Moving Average (MA) [18] models, respectively, as follows.

2.1.1 AR

The main role of autoregressive models is to predict future values based on past observations of a time series. For example, an AR model can predict the current or future values of an indicator at a certain time using past observations of PM2.5 in the air, i.e.,

$$X_t = c_1 + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ? \quad \alpha_r X_{t-r} + \grave{o}_t, \tag{1}$$

where X_t is the value of air quality, e.g. PM2.5, at moment t. $\alpha_1, ¼, \alpha_r$, are model parameters that represents the lagged effect of the data at the first r moments. In other words, it means that the firs r sets of observations still have an effect on future observations at a later time. This is manifested as $\alpha_1 X_{t-1}, ¼, \alpha_r X_{t-r}$. In the model c_1 is a constant term, usually a finite real number, i.e., c_1 Î □ . The error term $\grave{o}_t$ in the model is noteworthy in that it represents random fluctuations in the data that cannot be fully explained by the model. When the model is established accurately enough, the value of the error term $\grave{o}_t$ is usually sufficiently small.

AR modeling is a part of ARMA modeling when this paper examines the relationship between air quality and time. It can help to model and describe how the indicator values of air quality are affected by its historical state. For example, the level of oxygen O2 in the air may be affected by the concentration of oxygen in the past few days.

2.1.2 MA

The main modeling object of the moving average (MA) model is the forecast error of past series. Statistical models always have limited bias, but we always attempt to minimize the size of the error. This is where modeling the past prediction error can effectively reduce the future prediction error. The MA model is a typical example of using k previous prediction errors to predict the current observation, which is expressed mathematically as

$$X_t = c_2 + \sigma_1 \varepsilon_{t-1} + \sigma_2 \varepsilon_{t-2} + ? \quad \sigma_k \varepsilon_{t-k} + e_t, \tag{2}$$

where X_t represents the current or future value predicted by the model at moment t. $\sigma_1, ¼, \sigma_k$, are the model parameters that represent the effect of the prediction error at previous moments on the current predicted value. In other words, its lagged effect is specified as $\sigma_1 \varepsilon_{t-1}, ¼, \sigma_k \varepsilon_{t-k}$. c_2 denotes the mean value of the previous time series data, which is usually a constant. Unlike the previous model, e_t, as the residual of the data, represents the difference between the actual observations and the model predictions at each moment. However, it can be used equally well as a measure of how accurately the model fits the time series data.

The advantage of the MA model in air quality analysis and prediction is that it takes the error of each previous prediction into account. This has the advantage of constantly correcting the model itself. The effect of all previous prediction errors on the current prediction value, i.e., the manifestation of the lag effect, is taken into account during each iteration.

2.1.3 The combination of AR and MA

The ARMA model used in this paper, on the other hand, is a combination of the above AR and MA models. This has the advantage of taking into account both the past observations of the time

series data as well as the forecasting errors from the past forecasting process. The combined ARMA mathematical model is denoted as

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ? \quad \alpha_r X_{t-r} + \grave{o}_t + \sigma_1 \varepsilon_{t-1} + \sigma_2 \varepsilon_{t-2} + ? \quad \sigma_k \varepsilon_{t-k} + e_t + C, \quad (3)$$

where C is formed by combining c_1 and c_2, and the rest of the previously explained notation is not repeated here.

ARMA is applicable to air quality time series data for two reasons. Firstly, the AR term in ARMA is mainly oriented to deal with the inherent trends in air quality data. Whereas the MA term corrects the model using the error of each previous prediction. The ARMA model is due to its clear logic and computer reasoning. It can combine historical trends and short-term fluctuations to provide reliable prediction results for the data analyzed in this paper.

2.2 ARIMA

The ARIMA model is an improved ARMA. The difference between them is mainly in the fact that the ARIMA model has an additional difference step. This step is designed mainly for non-smooth time series data. After the differencing step, the time series data will become smoother compared to before. And the more differentiation steps, then the smoother the data will become. The format of the difference involved in the ARIMA model is as follows，

$$\Delta y_t = y_t - y_{t-1} \quad \text{or} \quad \Delta^d y_t = y_t - y_{t-1} - ? \ , \quad (4)$$

where y_t represents the observation at moment t, y_{t-1}represents the observation at the previous moment of t, and so on. d represents how many orders the step is in differential format. With more orders, the time series data will be smoother. The ARMA model after the differential format will be shown below. In other words, the differenced time series will obey an ARMA model.

$$\Delta^d y_t = c + \phi_1 \Delta^d y_{t-1} + ? \quad \phi_p \Delta^d y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ? \quad \theta_p \varepsilon_{t-q}. \quad (5)$$

There are a number of reasons that ARIMA is particularly suited to air quality data analysis, one in particular is obvious. One of the difference operations eliminates nonlinear trends in the original series. For example, if the data fit a higher-order curvilinear function, the effectiveness of the ARMA model will be diminished at that point. However, the step of differencing can make the data smooth. The ARMA model maintains excellent accuracy even when faced with data information with higher-order nonlinear relationships. In the subsequent experiments, this paper will show the role of ARMA and ARIMA in the field of data analysis. The data used is the air quality data of Shanghai city during 2019-2021.

3. Numerical Experiment

3.1 Experiment 1 Data and its pre-processing

In this paper, the experiment will use the daily air quality data of Shanghai during a three-year period (2019-2021). The three pollutants PM2.5, NO2, and O3 are the main subjects of the study. The data contains a total of 1,096 daily observations for each pollutant. First, the 5% missing values were treated by linear interpolation and the outliers were treated by the Interquartile Range (IQR) method. After completing the data cleaning, the distribution characteristics are shown in Figure 1.

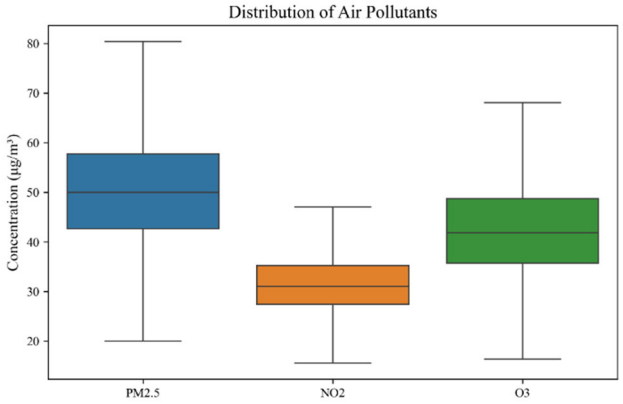

Fig.1 Air Pollutant Distribution

The box plots in Fig. 1 show the distribution characteristics of the three pollutants, PM2.5, No2, and O3. PM2.5 has the highest variability (IQR: 20-80), as can be seen from the lengths of its upper and lower lines. The possible reason for this is that Shanghai, as a first-tier city with daily progress in industrial development, emits dust that significantly increases the concentration of PM2.5 in the air. In addition, Shanghai has the highest number of cars in the country, which is one of the main reasons for this pollutant. The distribution of NO2 is the most symmetrical, and the cause of its generation is mainly from combustion. NO2 is the most symmetrical and is mainly generated by the combustion of fossil fuels, which is essential in Shanghai, such as industrial boilers and the use of natural gas for domestic use, etc. The distribution of O3 is positively skewed, with occasional high concentrations. The results of this analysis indicate that the ozone hole in Shanghai is small, and the main sources are nitrogen oxides (NO_x) and volatile organic compounds (VOCs) generated by photochemical reactions under sunshine.

Fig.1 Air Pollutant Distribution

The box plots in Fig. 1 show the distribution characteristics of the three pollutants, PM2.5, No2, and O3. PM2.5 has the highest variability (IQR: 20-80), as can be seen from the lengths of its upper and lower lines. The possible reason for this is that Shanghai, as a first-tier city with daily progress in industrial development, emits dust that significantly increases the concentration of PM2.5 in the air. In addition, Shanghai has the highest number of cars in the country, which is one of the main reasons for this pollutant. The distribution of NO2 is the most symmetrical, and the cause of its generation is mainly from combustion. NO2 is the most symmetrical and is mainly generated by the combustion of fossil fuels, which is essential in Shanghai, such as industrial boilers and the use of natural gas for domestic use, etc. The distribution of O3 is positively skewed, with occasional high concentrations. The results of this analysis indicate that the ozone hole over Shanghai is small, and the main sources are nitrogen oxides (NO_x) and volatile organic compounds (VOCs) generated by photochemical reactions under sunshine.

3.2 Experiment 2 Temporal Patterns

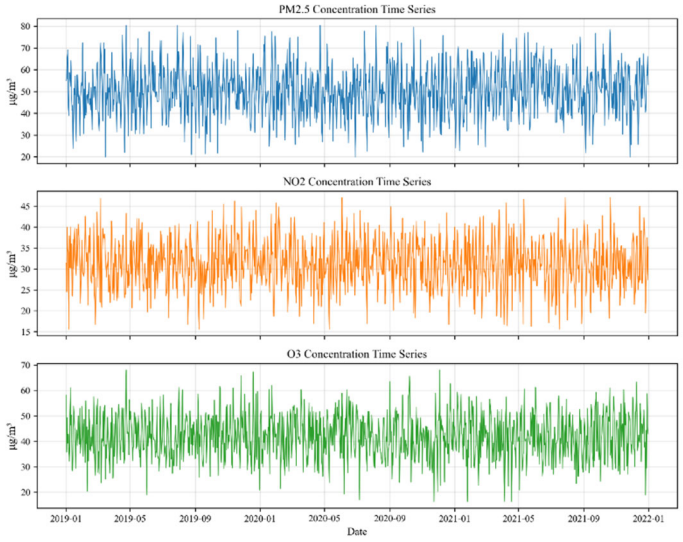

Fig. 2 Time Series of Pollutant Concentrations

The raw time series of pollutant concentrations are shown in Figure 2. PM2.5 and NO_2 show a certain pattern of winter peaks, which may be due to the increased consumption of fossil fuels. O_3 shows a clear summer peak, which is consistent with photochemical activity under high solar radiation. The observations in this figure indicate the need for models that can deal with seasonality and non-stationarity.

3.3 Experiment 3. Stationarity Analysis

Table 1 Augmented Dickey-Fuller Test Results

Pollutant	ADF Statistic	p-value	Stationary
PM2.5	-6.433	0.000	Ture
NO_2	-6.510	0.000	Ture
O_3	-5.970	0.000	Ture

In this experiment, we employ the Augmented Dickey-Fuller (ADF) test to assess the stationarity of each pollutant time series. Table 1 demonstrates that all three series are stationary at a 1% level of significance (p-value less than 0.01). Thus, the feasibility of the ARIMA model is tested. However, the ARIMA model can be further improved.

3.4 Experiment 4 Forecast Comparison

The present experiment was carried out in order to make a comparison of the prediction performance. For both ARMA and ARIMA models, Figure 3 compares the predicted and actual values of NO_2. Due to the fact that the difference step stabilizes the data, the ARIMA model is better able to capture the temporal pattern of pollutants, especially during periods of rapid fluctuations.

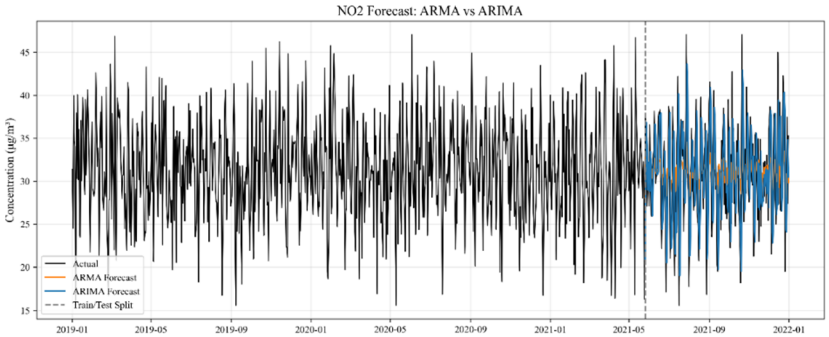

Fig. 3 Model Forecast Comparison for NO_2

As can be seen in Figure 3, the prediction error of the ARIMA model is significantly smaller than that of the ARMA model. Specifically, if the ARMA model (orange) is used for forecasting, the long-term trend of the data is not adequately captured. This leads to a significant deviation of the predicted values from the actual values. In contrast, if the ARIMA model (blue) is used for improvement, the model is able to better track changes in the actual data due to the correction of non-stationarity by the difference method.

Table 2 Model Evaluation Indicators

Pollutant	Model	AIC	BIC	RMSE	MAE	R^2
PM2.5	ARMA	4216.32	4231.45	8.924	7.213	0.612
PM2.5	ARIMA	4089.56	4109.73	6.217	5.032	0.782
NO_2	ARMA	3598.21	3613.34	4.356	3.521	0.543
NO_2	ARIMA	3487.92	3503.05	3.892	3.124	0.689
O_3	ARMA	3821.78	3836.91	7.843	6.324	0.487
O_3	ARIMA	3702.15	3717.28	5.962	4.812	0.724

In Table 2, we present a comparative assessment of the ARMA and ARIMA models for three pollutants- PM2.5, NO_2, and O_3-using several performance metrics, including AIC, BIC, RMSE, MAE, and R^2. Summarizing the results in the table, we can still conclude that the ARIMA model is superior to the ARMA model. The above results are reflected in the lower values of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), both of which are used to measure the complexity and goodness of fit of the model. Therefore, the lower the value of the above metrics, the better the balance of the model. The Root Mean Square Error (RMSE) is calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{6}$$

The Mean Absolute Error (MAE), defined as

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|. \tag{7}$$

The above two metrics serve to quantify the mean of the prediction error. Coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \tag{8}$$

which is used to measure the proportion of variance explained by the model.

4. Conclusion

By studying the time series of major air pollutants, this paper explores and analyzes their characteristics. Various applications of ARMA and ARIMA models are presented. Based on the preprocessing of Shanghai air quality data, we applied the two models to make predictions. The empirical results show that although both models can effectively capture the statistical dynamics of air pollutants, the ARIMA model shows superior performance in terms of accuracy and model evaluation criteria.

Specifically, ARIMA outperforms ARMA in terms of AIC, BIC, RMSE, MAE, and R^2 for air pollutants in Shanghai. Secondly, both theoretical and empirical results reflect the enhancement of model stability by the differential methods. Especially for pollutants with nonlinear or seasonal patterns, ARIMA is more robust when dealing with real-world air quality data, which are often noisy, non-stationary, and affected by exogenous factors.

This study has important implications for environmental detection and environmental policy planning. Accurate air quality forecasts can help relevant authorities to pre-empt future pollution

trends. Therefore, promoting public health and environmental sustainability. In future work, we plan to combine ARIMA with machine learning or meteorological inputs in a hybrid model. Such means of combining novel technologies will have great research potential in today's intelligent society.

References

[1] Gao, W., Xiao, T., Zou, L., Li, H., & Gu, S. (2024). Analysis and Prediction of Atmospheric Environmental Quality Based on the Autoregressive Integrated Moving Average Model (ARIMA Model) in Hunan Province, China. Sustainability, 16(19), 8471.

[2] Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. Time series analysis and its applications: with R examples, 75-163.

[3] Choi, B. (2012). ARMA model identification. Springer Science & Business Media.

[4] Dubey, A. K., Kumar, A., García-Díaz, V., Sharma, A. K., & Kanhaiya, K. (2021). Study and analysis of SARIMA and LSTM in forecasting time series data. Sustainable energy technologies and assessments, 47, 101474.

[5] De Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., ... & Hoek, G. (2018). Spatial PM2. 5, NO2, O3 and BC models for Western Europe–evaluation of spatiotemporal stability. Environment international, 120, 81-92.

[6] Zhang, H., Wang, S., Hao, J., Wang, X., Wang, S., Chai, F., & Li, M. (2016). Air pollution and control action in Beijing. Journal of Cleaner Production, 112, 1519-1527.

[7] Wang, L., Zou, H., Su, J., Li, L., & Chaudhry, S. (2013). An ARIMA-ANN hybrid model for time series forecasting. Systems research and behavioral science, 30(3), 244-259.

[8] Babu, C. N., & Reddy, B. E. (2014). A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data. Applied soft computing, 23, 27-38.

[9] Alsuwaylimi, A. A. (2023). Comparison of ARIMA, ANN and Hybrid ARIMA-ANN models for time series forecasting. Information sciences letters, 12(2), 1003-1016.

[10] Wang, J., Wang, R., & Li, Z. (2022). A combined forecasting system based on multi-objective optimization and feature extraction strategy for hourly PM2. 5 concentration. Applied soft computing, 114, 108034.

[11] Sen, R., Yu, H. F., & Dhillon, I. S. (2019). Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. Advances in neural information processing systems, 32.

[12] Bhanja, S., & Das, A. (2018). Impact of data normalization on deep neural network for time series forecasting. ArXiv:1812.05519.

[13] Sharma, S., & Sen, S. (2023). Real-time structural damage assessment using LSTM networks: regression and classification approaches. Neural computing and applications, 35(1), 557-572.

[14] Singh, K. N., Sharma, K., Avinash, G., Kumar, R. R., Ray, M., Ramasubramanian, V., ... & Lal, S. B. (2023). LSTM based stacked autoencoder approach for time series forecasting. Journal of Indian society of agricultural statistics, 77, 71-78.

[15] Kumar, U., & Jain, V. K. (2010). ARIMA forecasting of ambient air pollutants (O 3, NO, NO 2 and CO). Stochastic environmental research and risk assessment, 24, 751-760.

[16] Wang, J., Zhou, Q., & Zhang, X. (2018, December). Wind power forecasting based on time series ARMA model. In IOP conference series: Earth and environmental science, 199(2), 022015.

[17] Slini, T., Karatzas, K., & Moussiopoulos, N. (2002). Statistical analysis of environmental data as the basis of forecasting: an air quality application. Science of the total environment, 288(3), 227-237.

[18] Gardenier, T. K. (1982). Moving averages for environmental standards. Simulation, 39(2), 49-58.

[19] Orellano, P., Reynoso, J., Quaranta, N., Bardach, A., & Ciapponi, A. (2020). Short-term exposure to particulate matter (PM10 and PM2. 5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. Environment international, 142, 105876.

[20] Orellano, P., Reynoso, J., Quaranta, N., Bardach, A., & Ciapponi, A. (2020). Short-term exposure to particulate matter (PM10 and PM2. 5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. Environment international, 142, 105876.