International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Prediction of Air Quality Index Using Time Series Modelling: A Review Study

Vanshika Saini[1], Richa[2]

[1]Student, M.Sc. Statistics, School of Statistics, Maa Shakumbhari University, Saharanpur, Uttar Pradesh, India.
[2]Assistant Professor (Guest Faculty), School of Statistics, Maa Shakumbhari University, Saharanpur, Uttar Pradesh, India.

Abstract
Air pollution is a serious environmental issue that affects public health and climate change. Accurate forecasting of the Air Quality Index (AQI) is essential to mitigate its adverse effects and implement effective pollution control measures. This review paper evaluates various time series modelling approaches used for AQI forecasting, including traditional statistical models such as ARIMA and SARIMA, hybrid models that combine statistical and machine learning techniques and deep learning-based approaches such as LSTM and fuzzy time series models. The findings suggest that while ARIMA and SARIMA are effective for short-term forecasting, hybrid models and deep learning techniques provide better accuracy by capturing complex temporal patterns. However, challenges such as data quality issues, computational cost, and regional variations affect the reliability of these models. Future research should focus on developing efficient hybrid approaches to integrate real-time data sources, enhance model interpretability, and improve AQI forecast accuracy. This study provides insights into the strengths and limitations of different forecasting techniques, providing a basis for future advancements in air quality forecasting.

Keywords: Air, Air Pollution, Air Quality Index, Time Series, Analysis, Prediction, Forecasting, ARMA model, ARIMA model.

1. Introduction

Air, the invisible mixture of gases surrounding the Earth, is essential for sustaining life. It contains important components such as oxygen, nitrogen and carbon dioxide, which are crucial for survival [18]. However, the quality of the air we breathe is constantly deteriorating due to pollution, posing serious risks to public health, ecosystems and climate stability. Air pollution from both natural and man-made sources such as vehicle emissions, industrial activities and forest fires introduce harmful substances into the atmosphere. Air pollutants including particulate matter, carbon monoxide, nitrogen dioxide, ozone and sulfur dioxide [19] can cause respiratory and cardiovascular diseases, reduced agricultural productivity and long-term environmental degradation. The Air Quality Index (AQI) has emerged as an essential metric for monitoring and assessing air pollution levels. By converting the concentrations of various pollutants into a single numerical value, the AQI simplifies the communication of air quality information to the public. It helps identify pollution hotspots, assess health risks, and inform policy decisions aimed at reducing pollution. The index is particularly valuable for predicting short-term health

effects, such as respiratory distress and asthma exacerbations, and guiding interventions during high pollution conditions. Focusing on AQI is crucial to tackle the challenges posed by deteriorating air quality. It provides a comprehensive understanding of the impact of pollution on health and the environment, facilitating targeted actions to reduce these impacts. Furthermore, studying AQI trends helps identify high-risk areas and timing, helping governments and communities to implement measures such as emission controls, traffic restrictions, and public health advisories in a timely manner. By prioritizing AQI monitoring and prediction, we can create healthier living environments, improve quality of life, and ensure a sustainable future for generations to come. AQI prediction has gained significant attention in recent years due to its potential to inform proactive measures such as issuing health advisories, regulating industrial emissions, and implementing traffic controls. Time series modelling techniques have become critical in AQI forecasting, as they analyse historical data to identify trends, seasonality, and patterns. From traditional statistical models such as ARIMA to advanced methods such as machine learning and deep learning, researchers continue to explore innovative approaches to improve prediction accuracy.

This paper focuses on reviewing existing research in the field of AQI prediction using time series modelling. It discusses the development of forecasting techniques, evaluates their performance in different fields, and highlights their implications for public health and policy-making. The review also suggests future directions to enhance AQI forecasting methods.

2. Research Methodology

Research Design:

The methodology of this study involved reviewing a range of published research on AQI forecasting, particularly using time series modelling techniques, with literature covering the period 2006 to 2024. This review focuses specifically on empirical studies that have applied time series methods to forecast AQI, providing a detailed investigation of the effectiveness and evolution of these techniques. For the purpose of this research, AQI forecasting is defined as forecasting air quality based on historical data through various time series models. The aim of the study is to evaluate the performance of these models and analyse trends in AQI forecast accuracy across different regions.

Selection Criteria and Sources of Data:

For this review, studies were selected based on their focus on AQI forecasting using time series modelling techniques. Only papers published between 2006 and 2024 were considered, ensuring the inclusion of the most recent research in this area. The selected studies specifically used time series methods such as ARIMA and SARIMA to forecast AQI. Relevant research articles were sourced from academic databases, ensuring the inclusion of diverse and high-quality studies on AQI forecasting using time series modelling techniques. This approach helped in compiling a comprehensive review of time series-based AQI forecasting methods.

3. Review of Literature

Air Quality Index (AQI) forecasting has been widely studied using various time series modelling approaches. These methodologies range from traditional statistical models like ARIMA and SARIMA to more advanced machine learning and deep learning models such as LSTM and hybrid approaches integrating multiple techniques. This review synthesizes findings from various research studies, analysing the effectiveness of different time series forecasting techniques applied to AQI prediction. Sidhu et al.

(2024) investigated predictive modelling of AQI across various cities and states in India, with a focus on the impact of stubble burning in Punjab. Their study utilized data from 22 monitoring stations in Delhi, Haryana, and Punjab, sourced from the Central Pollution Control Board (CPCB). They pre-processed the data by handling missing values and outliers before applying multiple models, including SARIMAX, LSTM, Random Forest, SVM, CatBoost, and XGBoost. The evaluation based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) revealed that the Random Forest model outperformed others in AQI forecasting, particularly in pollution-affected areas. Pant et al. (2023) focused on AQI forecasting in Dehradun, a highly polluted region in India, using time series modelling. Their approach combined machine learning with the Akaike Information Criterion (AIC) for optimal model selection. They employed a Seasonal Auto-Regressive Moving Average (ARMA) model, which was found to be effective in capturing seasonal variations in AQI. Their findings emphasized the importance of time series modelling in predicting pollution levels and informing policymakers about health risks associated with air pollution, particularly during winter. Atoui et al. (2022) conducted a study in Zahleh, Lebanon, evaluating AQI levels using Naïve models, Exponential Smoothing, TBATS, and SARIMA. Their findings indicated that SARIMA was the most effective model, achieving high accuracy in AQI prediction. This study highlighted the importance of selecting appropriate models for specific environmental conditions, as some techniques performed better in capturing seasonal and temporal variations. Mani and Viswanadhapalli (2022) examined AQI forecasting in Chennai, India, using a combination of ARIMA and Multi-Linear Regression (MLR). Their study integrated NO_2, Ozone (O_3), $PM_{2.5}$, and SO_2 sensor data from CPCB to train their model. Their ARIMA model achieved an accuracy of over 80% for short-term predictions, whereas the hybrid ARIMA-MLR approach showed improvements in overall forecast accuracy. Alyousifi et al. (2020) introduced a Fuzzy Time Series Markov Chain (FTSMC) model for AQI forecasting in Klang, Malaysia. Addressing limitations of traditional fuzzy time series models, they proposed an optimal partitioning method to enhance prediction accuracy. Their results, validated through RMSE and MAPE metrics, indicated that the FTSMC model outperformed other statistical models, demonstrating its effectiveness in improving AQI forecasting and air quality management. Koo et al. (2020) compared several forecasting models, including ANN, ARIMA, TBATS, and fuzzy time series models, for API prediction in Kuala Lumpur. Their research, based on six years of daily air pollution data, found that Singh's fuzzy time series model was the most accurate, with an RMSE of 1.4704 and a MAPE of 4.364%. The study emphasized the computational efficiency and accuracy of fuzzy time series models compared to traditional approaches. Sethi and Mittal (2020) investigated AQI prediction in Gurugram, India, using univariate and multivariate time series models. Their study compared ARIMA and Vector Auto Regression (VAR) models, concluding that ARIMA outperformed VAR in predictive accuracy. This study reaffirmed ARIMA's effectiveness in univariate AQI forecasting while highlighting the limitations of VAR in modelling air pollution dynamics. Bhalgat et al. (2019) employed machine learning techniques to predict SO_2 concentrations in urban and industrial areas. Their research integrated meteorological, traffic, and industrial parameters to enhance model accuracy. By incorporating time series models into their approach, they demonstrated the importance of data-driven forecasting in managing environmental pollution and mitigating health risks. Naveen and Anu (2017) conducted a study in Thiruvananthapuram District, Kerala, India, using ARIMA and SARIMA models for AQI prediction. They found that ARIMA provided more satisfactory results than SARIMA, particularly in short-term forecasting. Their study underscored the importance of model optimization in improving prediction accuracy. Passamani and Masotti (2016) analyzed air pollution trends in an Alpine Italian

province using a dynamic multiple time series model. Their approach identified long-term improvements in air quality, demonstrating the effectiveness of autoregressive stochastic factor models in analysing pollution trends. Kadiyala and Kumar (2014) explored ARMAX/ARIMAX models for indoor air quality prediction in public transportation settings. Their study emphasized the need for accurate AQI forecasting to guide vehicle manufacturers in designing ventilation systems for safe air exchange rates. Wu and Kuo (2012) utilized vector time series models, coupled with ARCH and GARCH methodologies, to analyse AQI trends in Taiwan. Their findings indicated that ozone (O_3) levels were influenced by lagged values of PM_{10} and NO_2, while SO_2 levels directly impacted CO concentrations. Their research highlighted the dynamic interdependencies between different air pollutants and their impact on AQI. Kumar and Goyal (2011) developed a daily AQI forecasting model for Delhi, using ARIMA and Principal Component Regression (PCR). Their study demonstrated that combining statistical techniques improved predictive performance and helped in developing more reliable forecasting models. Hoi et al. (2009) proposed a time-varying statistical model (TVAREX) based on the Kalman filter for PM_{10} forecasting in coastal cities. Comparing TVAREX with an ANN model, they found that TVAREX was more efficient in capturing pollution episodes, emphasizing its potential in real-time applications. Chelani and Devotta (2006) addressed the challenges of air quality time series forecasting by developing a hybrid approach combining ARIMA with nonlinear dynamic models. Their research demonstrated that hybrid models significantly improved forecasting accuracy by capturing both linear and nonlinear dependencies in air pollution data. The reviewed studies highlight several key trends in AQI forecasting. Traditional statistical models like ARIMA and SARIMA remain effective for short-term predictions, whereas hybrid approaches integrating statistical and machine learning models improve long-term accuracy. Deep learning techniques such as LSTM and fuzzy time series models provide promising results, especially for handling nonlinear dependencies. Challenges such as missing data, seasonal variations, and computational complexity continue to impact AQI forecasting reliability. Future research should explore hybrid methodologies, address data inconsistencies, and apply novel deep learning techniques to enhance air quality prediction across diverse environmental conditions.

Table 1: Summary Table of Reviewed Studies

Authors & Year	Models Used	Dataset Source	Performance Metrics	Key Findings
Sidhu et al. (2024)	SARIMAX, LSTM, RF, XGBoost	CPCB (India)	RMSE, R^2	RF performed best for AQI prediction
Pant et al. (2023)	ARMA	Dehradun AQI data	AIC, RMSE	Seasonal ARMA model effective
Atoui et al. (2022)	SARIMA, TBATS, Exponential Smoothing	Lebanon AQI data	RMSE, MAE	SARIMA most accurate
Mani & Viswanadhapalli (2022)	ARIMA, MLR	CPCB (India)	RMSE, MAE	Hybrid ARIMA-MLR improved accuracy

Alyousifi et al. (2020)	FTSMC	Klang, Malaysia	RMSE, MAPE	FTSMC outperformed traditional models
Koo et al. (2020)	ANN, ARIMA, TBATS, Fuzzy Time Series	Kuala Lumpur API data	RMSE, MAPE	Fuzzy time series models were the most accurate
Sethi & Mittal (2020)	ARIMA, VAR	Gurugram AQI data	RMSE, MAE	ARIMA outperformed VAR in AQI forecasting
Bhalgat et al. (2019)	ML Techniques, Time Series Models	SO_2 data from urban/industrial areas	RMSE	Highlighted importance of meteorological and industrial parameters in AQI forecasting
Naveen & Anu (2017)	ARIMA, SARIMA	Thiruvananthapuram AQI data	RMSE	ARIMA provided better results than SARIMA for short-term forecasting
Passamani & Masotti (2016)	Dynamic Multiple Time Series Model	Alpine Italian province AQI data	RMSE	Showed long-term air quality improvements
Kadiyala & Kumar (2014)	ARMAX, ARIMAX	Public transport indoor AQI data	RMSE	Recommended forecasting techniques for improving air ventilation systems
Wu & Kuo (2012)	Vector Time Series Models, ARCH, GARCH	Taiwan AQI data	RMSE	Identified interdependencies between air pollutants
Kumar & Goyal (2011)	ARIMA,PCR	Delhi AQI data	RMSE	Hybrid ARIMA-PCR model improved AQI prediction
Hoi et al. (2009)	TVAREX, ANN	PM_{10} data from coastal cities	RMSE	TVAREX performed better for real-time forecasting
Chelani & Devotta (2006)	ARIMA, Nonlinear Dynamic	Delhi NO_2 data	RMSE	Hybrid approach captured both lin-

	Models			ear and nonlinear trends

4. Research Gap

Despite significant progress in AQI forecasting, several research gaps still remain unresolved. Many studies have focused on specific cities or regions, leading to a lack of generalization in forecasting models across different geographic locations. The dominance of traditional statistical models such as ARIMA and SARIMA highlights the need for more extensive exploration of hybrid approaches that integrate machine learning and deep learning techniques. While deep learning models such as LSTM have shown promise, their high computational cost and the need for large datasets limit their widespread application. Another important challenge is data quality, as many studies do not adequately address the impact of external factors such as missing values, sensor inaccuracies, or meteorological conditions. Furthermore, most research has been focused on short-term AQI forecasting, while limited efforts have been devoted to long-term trend analysis. Future studies should focus on improving model adaptability across regions, enhancing data pre-processing techniques, and exploring new methodologies that balance computational efficiency with forecast accuracy. Additionally, comparative evaluation of different hybrid models under different environmental conditions is needed to establish best practices for AQI prediction.

5. Result

Based on the reviewed studies, hybrid models integrating statistical and machine learning techniques, such as ARIMA combined with Random Forest, XGBoost or LSTM, have shown the best accuracy in AQI forecasting. These models effectively utilize both time-dependent trends and complex non-linear patterns, leading to better forecasting performance. Traditional models such as ARIMA and SARIMA are widely used, especially for short-term AQI forecasting, due to their simplicity and efficiency. However, they struggle to handle long-term forecasting and highly fluctuating pollution levels. Fuzzy time series models and deep learning-based approaches, including LSTM and GRU, have been highly effective in capturing dynamic trends but require extensive data preprocessing and computational resources. Studies also indicate that hybrid models provide the most consistent results across different geographic regions and seasonal variations, making them a promising option for future AQI forecasting efforts.

6. Conclusion

This review paper provides a comprehensive analysis of AQI forecasting using time series models. The findings show that while traditional models such as ARIMA and SARIMA are widely used due to their simplicity and efficiency, they are often limited in capturing long-term trends and complex non-linear relationships in air quality data. Hybrid models that integrate statistical methods with machine learning techniques such as Random Forest and XGBoost have demonstrated improved accuracy by taking advantage of both time-dependent patterns and complex pollutant interactions. Deep learning models such as LSTM and fuzzy time series approaches have shown promise in handling non-linearity, but require extensive data preprocessing and high computational resources. The literature highlights that no single model is universally superior, and model selection should be based on factors such as data availability, computational efficiency, and regional pollution characteristics. Future research should

focus on refining hybrid methods, incorporating real-time data sources, and addressing data quality challenges to enhance the reliability and applicability of AQI prediction models in different environmental contexts.

7. Futuristic Approach

Future research in AQI forecasting should focus on developing hybrid models that integrate statistical, machine learning, and deep learning techniques for improved accuracy and efficiency. Incorporating real-time data sources such as satellite imagery and IoT-based sensors can enhance the predictive power of these models. Additionally, explainable AI (XAI) techniques can be explored to improve the interpretability of complex forecasting models, ensuring that policymakers can make informed decisions based on model predictions. Another promising avenue is the application of transfer learning, where pre-trained models can be adapted for AQI forecasting across multiple regions without requiring large amounts of localized training data. Addressing challenges related to data quality, feature selection, and computational efficiency will be critical in advancing AQI forecasting methods.

References

1. Sidhu K.K., Balogun H., Oseni K.O., “Predictive Modelling of Air Quality Index (AQI) Across Diverse Cities and States of India using Machine Learning: Investigating the Influence of Punjab's Stubble Burning on AQI Variability”, 2024 Apr 11.
2. Pant A., Joshi R.C., Sharma S., Pant K., “Predictive modeling for forecasting air quality index (AQI) using time series analysis”, Avicenna Journal of Environmental Health Engineering, 2023 Jun 29, 10(1), 38-43.
3. Atoui A., Slim K., Andaloussi S.A., Moilleron R., Khraibani Z., “Time Series Analysis and Forecasting of the Air Quality Index of Atmospheric Air Pollutants in Zahleh, Lebanon”, Atmospheric and Climate Sciences, 2022 Aug 23, 12(4), 728-49.
4. Mani G., Viswanadhapalli J.K., “Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models”, Journal of Engineering Research, 2022, 10(2A), 179-94.
5. Alyousifi Y., Othman M., Sokkalingam R., Faye I., Silva P.C., “Predicting daily air pollution index based on fuzzy time series markov chain model”, Symmetry, 2020 Feb 17, 12(2), 293.
6. Central Pollution Control Board (CPCB), Government of India. (accessed 20th April, 2020) https://cpcb.nic.in/
7. Koo J.W., Wong S.W., Selvachandran G., Long H.V., Son L.H., “Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models”, Air Quality, Atmosphere & Health, 2020 Jan 13, 77-88.
8. Sethi J.K., Mittal M., “Analysis of air quality using univariate and multivariate time series models”, In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020 Jan 29, 823-827.
9. Bhalgat P., Pitale S., & Bhoite S., “Air quality prediction using machine learning algorithms”, International Journal of Computer Applications Technology and Research, 2019, 8(9), 367-370.
10. Naveen V., & Anu N., “Time series analysis to forecast air quality indices in Thiruvananthapuram District, Kerala, India”, International Journal of Engineering Research and Application, 2017, 7(6),

66-84.
11. Passamani G., & Masotti P., "Local atmospheric pollution evolution through time series analysis". Journal of Mathematics and Statistical Science, (2016). 2(12), 781-788.
12. Plank, C.E., "Handbook of Analysis of Air Quality In Urban Environments", Auris Reference Ltd., UK, 2015.
13. Kadiyala A., & Kumar A., "Multivariate time series models for prediction of air quality inside a public transportation bus using available software", Environmental Progress & Sustainable Energy, 2014, 33(2), 337-341.
14. Wu E. M. Y., & Kuo S. L., "Air quality time series based GARCH model analyses of air quality information for a total quantity control district", Aerosol and Air Quality Research, 2012, 12(3), 331-343.
15. Kumar A., & Goyal P., "Forecasting of daily air quality index in Delhi", Science of the Total Environment, 2011, 409(24), 5517-5523.
16. Hoi K. I., Yuen K. V., & Mok K. M., "Prediction of daily averaged PM10 concentrations by statistical time-varying model", Atmospheric Environment, 2009, 43(16), 2579-2581.
17. Chelani A. B., & Devotta S. Air quality forecasting using a hybrid autoregressive and nonlinear model. Atmospheric Environment, (2006). 40(10), 1774-1780.
18. https://education.nationalgeographic.org/resource/air/
19. EPA (2005). Six common air pollutants. U. S. environmental protection agency. https://www.epa.gov/criteria-air-pollutants