International Journal of Innovations & Research Analysis (IJIRA)
ISSN :2583-0295, Impact Factor: 5.449, Volume 03, No. 04(II), October- December, 2023, pp 149-155

A COMPREHENSIVE STUDY OF TIME SERIES MINING TECHNIQUES FOR ANALYSIS OF AIR QUALITY DATA

Dr. Leena Bhatia*
Apoorva Verma**
Dr. Nitish Pathak***

ABSTRACT

Air is one of the most fundamental elements required by all living beings to survive on the planet. Our planet Earth is surrounded by a thin layer of air that extends up to several kilometers above the surface of the earth and that forms our atmosphere. Air is a mixture of various components like nitrogen (N2) 78.08%, oxygen (O2) 21%, carbon dioxide (CO2) 0.04%. Rest of the components like water vapor, argon, dust, and smoke, are present in very small amounts. These are the components that play a major role in determining the quality of the air that we have around us. When the proportions of these components get disturbed, this leads to air pollution. Air pollution has become a major environmental problem affecting the health of people worldwide. In recent years, due to the availability of air pollution data, time series mining techniques has emerged as a promising approach predicting air pollution levels by identifying underlying trends, patterns and also forecasting the various components present in air. Many researchers are continuously working to analyze the air quality. Several methods, like analytical, statistical, and ensemble methods, have been adopted for getting accurate results for the analysis and forecasting of the air quality. In this paper, we have studied the works of several scholars, the models proposed by them, an analysis made by them, and the predictions they have made for forecasting air quality. This paper presents a concise literature summary of the works and focuses on their research gaps which may be helpful for further studies and analysis of air quality.

Keywords: *Air Quality, Boosting Algorithms, Forecasting Models, Predictors, Time Series data.*

Introduction

Due to an increase in anthropogenic activities likes rapid growth in population, urbanization, and the development of modern industries, the quality of the air has dynamically weakened. According to the 2018 W.H.O. report, air pollution is a significant environmental and public health issue, causing an estimated seven million premature deaths annually worldwide. Developing countries like India are mostly affected by air pollution. SPM, RSPM, SO_2, NOx, and other organic and inorganic pollutants, including traces of metal pollutants, are produced and increased by vehicular emissions from urban areas [12]. Air

* Associate Professor, Department of Computer Applications, S.S. Jain Subodh PG College, Jaipur, Rajasthan, India.

** Research Scholar, Rajasthan Technical University Kota, Assistant Professor Department of MCA, Xavier Institute of Management and Informatics, St. Xavier's College, Jaipur, Rajasthan, India.

*** Associate Professor, Department of Information Technology, Bhagwan Parshuram Institute of Technology, New Delhi, India.

pollution is a gradual process that causes several diseases among humans like ischemic heart disease, stroke, chronic obstructive pulmonary disease (COPD), lung cancer, and acute lower respiratory infections in children. Lead, ozone, particulate matter, nitrogen dioxide, carbon monoxide, and Sulphur dioxide are the major air pollutants. Trees and crops can also suffer damage from air pollution. So, reducing air pollution can help both the environment and society.

To monitor the air quality, the government has set up several air quality monitoring stations all over India. The Central Pollution Control Board (CPCB) has executed a National Air Quality Monitoring Program (NAMP) for monitoring ambient air quality. There are 804 operating stations covering 344 cities/towns in 28 states and 6 union territories of the country [10]. The complexity and dynamic nature of air pollution data make it challenging to analyze and forecast air pollution trends accurately. Time series mining techniques have emerged as a promising approach to address this challenge.

Literature Review

The primary aim of this literature review includes the study of time series data mining with respect to air quality forecasting, the study of air and various factors affecting air quality, the study of various algorithms applied to time series data mining, and finding the gap between various studies that are still being studied or explored. This review has been conducted in a systematic manner by exploring highly indexed databases like Google Scholar, Web of Science journals, journals on the UGC Care List, and a few other reputed journals published from year 2015 till the current year. IEEE Xplore, Science Direct, SSRN, and Springer are the major databases in which articles are found. The key search terms included 'time series mining',' algorithms', 'air quality', 'forecasting', etc. with different combinations. The research papers included in this review are those that are relevant, have good citations, and are published in Q-indexed journals.

From the literature survey, we found that the major focus of various researchers while choosing the algorithm was the ARIMA model, which stands for Autoregressive Integrated Moving Average and is a statistical model used to analyze and forecast the concentration of the various pollutants present in the air [8],[9]. According to a few studies, it has been found that the ARIMA model alone is not enough to give satisfactory results, so, many researchers have combined the ARIMA model with some other techniques like empirical mode decomposition(EMD), and principal component analysis (PCA) for forecasting the results [1]. Aladağ E. (2021) has used wavelet transformation and seasonal adjustment along with the ARIMA model for forecasting the results [2].

The other variation of ARIMA is the seasonal ARIMA model, also known as SARIMA, which has also been used along with the factor analysis technique for predicting the concentration of PM2.5 in Lahore, Pakistan. [5] Liu. T et.al. have used another variation of the ARIMA model, i.e., hybrid ARIMAX to develop a numerical model for forecasting air quality in areas of Hong Kong, China [11]. Apart from the ARIMA model various other statistical tools and techniques were also used for research, like Correlation Analysis [3]. Data Mining Techniques are also used for the analysis of time series data. Taneja, S., Sharma et.al. have used Linear regression and Multilayer Perceptron for trend analysis in time series data of the air in the major polluted areas of Delhi[13]. Machine Learning Techniques in the current scenario turn out to be most promising in the case of forecasting and trend analysis. Thus, various methods of Machine Learning were used by the researchers for forecasting the air pollution, and trend analysis. Kumar K., & Pande, B. P. (2022) used Support Vector Machine (SVM), XG Boost, Gaussian Naïve Bayes, Random Forest(RF), and K-Nearest Neighbor Algorithm(KNN) for prediction of air pollution in 23 different cities in India. Other types of methods which can be used for more accurate results are deep learning techniques like Simple Recurrent Neural Networks (RNN), Long Short Term Memory(LSTM), Bi-Directional LSTM(Bi LSTM), Gated Recurrent Units(GRUs) and Variation Auto Encoder(VAE) which help in better variable capturing in time series data, thus giving more accurate and quick results[14].

Given below is a summarized table (Table 1) of the literature review, which gives the crux of the papers studied, the techniques used by them, their evaluation criteria, the datasets they have used, and the major findings and drawbacks that can be drawn from their work.

Table 1

Reference	Algorithm used	Evaluation Criteria	Dataset Used	Major Findings	Drawbacks
[8]	ARIMA	MAE,MSE, RMSE	Data from 5 monitoring stations in Hyderabad	1. Forecasted concentration of air pollutants with relatively low prediction errors.	1. Authors considered only limited pollutants SO2, NO2, PM2.5 and PM10

			city, Jan2017-Dec2019	2. Concentration of air pollutants found to be higher in winter.	whereas Ozone and CO which are the major pollutants were left out. 2. There is no comparison of ARIMA model's predictive performance with other time series models, which could provide insights into the relative effectiveness of the model for air pollution prediction.
[1]	Empirical Mode Decomposition (EMD)+ Principal Component Analysis (PCA)+ ARIMA models	Forecasting and Accuracy	Monthly crude oil prices from January 2000 to December 2016.	1. A decomposition ensemble model with the reconstruction of intrinsic mode functions (IMFs) based on ARIMA models showed better forecasting accuracy compared to other models such as ARIMA, EMD-ARIMA, and PCA-ARIMA. 2. Crude oil prices are influenced by a combination of long-term trends, seasonal patterns, and short-term fluctuations. The decomposition ensemble model effectively captured these different components, resulting in more accurate forecasts. 3. Authors identified key economic indicators such as gross domestic product (GDP), inflation rate, and exchange rate as significant predictors of crude oil prices.	1. Study only considered monthly data, which may not capture short-term fluctuations and changes in market conditions. 2. Analysis only focused on crude oil prices and did not consider other factors that may affect crude oil demand and supply, such as geopolitical events, technological advances, and environmental policies. 3. The paper does not provide a detailed explanation of the PCA-ARIMA model used in the study, which may make it difficult for readers to replicate the analysis. 4. The study did not perform out-of-sample testing to evaluate the model's predictive accuracy on new data, which could provide additional insights into the model's effectiveness.
[2]	ARIMA model+ wavelet transformation + Seasonal adjustment.	RMSE, MAE,MAPE	Hourly data from January 2014 to December 2014 in Guangzhou, China.	1. The Proposed model showed better forecasting as compared to traditional ARIMA model 2. Wavelet Transformation helps in the effective decomposition of time series data, which make it easier to identify underlying patterns and trends 3. Authors have also used seasonal adjustment technique	1. Authors have used data only from single location. 2. The major focus was on a single pollutant i.e. Particulate matter and rests of the pollutants were ignored. 3. Other meteorological factors like weather conditions, traffic density, and industrial emissions were not considered.
[3]	Statistical Techniques(correlation analysis+ principal component analysis+ cluster analysis) Visualization tools (Heat Maps, contour plots+ time series plots)	Correlation	Hourly data from three monitoring stations from São Carlos-SP (Brazil) 2014 to 2015	1. There was a clear association between high PM10 levels and meteorological factors such as wind speed and temperature. 2. The authors developed a new visualization tool, called the "PM10 Cube," which allows for the exploration and visualization of PM10 data across multiple dimensions.	Research considers only one air pollutant PM10 and do not consider rest and area of research is limited to only a single place with only hourly data of a single year, hence it is difficult to draw the generalized results. The study does not provide information on the sources of PM10, which would be helpful in developing effective mitigation strategies.

[4]	Random Forest(RF)+Recurrent Neural Network(RNN)+Long Short Term Memory(LSTM)+ Convolutional Neural network(CNN)+Gated Recurrent Unit(GRN)	Generates new time series data	Used various global datasets	1. The authors propose a method that generates new time series data points by transforming and combining existing data points, which is then used to train forecasting models. 2. The authors show that the proposed method outperforms other state-of-the-art data augmentation methods in terms of forecasting accuracy.	The study does not provide a detailed explanation of the data augmentation method and the specific transformations used to generate new data points. The study only evaluates the proposed method on a limited number of global forecasting models and may not be generalizable to other models.
[5]	SARIMA+Factor analysis	Analyze and predict the concentration of PM2.5	Air Quality dataset of Lahore city, Pakistan	The major factors affecting concentration of PM2.5 are weather conditions, traffic volume and industrial emission.	The study did not examine the effectiveness of different pollution control measures and policies.
[6]	Interpolation+Principal Component Analysis(PCA)	Analysis of level of PM2.5 in Jaipur City	Dataset from major ground-level monitoring station of Jaipur city.	The Authors found the concentration of pollutants varied significantly across different parts of the city.	Study is limited to the scope of analysis. Authors did not consider the ozone and carbon monoxide pollutants as they also have the adverse impacts on health. the study did not provide any policy recommendations or strategies for mitigating air pollution in Jaipur
[7]	Statistical Analysis- Correlation and t-tests	Correlation analysis and T-testing	Ground-level monitoring stations and satellite remote sensing to analyze the concentration of PM2.5 in the atmosphere in the IGP region.	The study found that the COVID-19 pandemic had a significant impact on air pollution levels in the IGP region, with a significant decrease in PM2.5 concentrations observed during the lockdown period. However, the PM2.5 concentrations increased significantly after the lifting of lockdown restrictions, indicating that the measures taken during the lockdown period were not sustainable in the long term.	The Author did not consider other major pollutants which are prevailing in IGP region like SO2, NOx and Ozone.
[9]	ARIMA	Forecasting concentrations of air pollutants for next 5 years	Data of three pollutants – Sulphur Dioxide, Nitrogen Dioxide, PM10 from 4 monitoring stations of Nanded city, Maharashtra	1. Findings are useful for policy makers and researchers in developing strategies to mitigate air pollution in Nanded city and similar urban areas.	Focused only on forecasting air pollution concentrations and did not examine the health impacts of exposure to air pollutants. Meteorological factors were not considered.
[11]	Hybrid ARIMAX+NumericalModelling	RMSE	Dataset of Hong Kong	1. The numerical model developed in the study was effective in forecasting the concentration of air pollutants, including particulate matter and nitrogen dioxide, in the Hong Kong region. 2. The paper highlights the importance of using high-resolution data to improve the accuracy of air quality forecasting. 3. The study showed that meteorological	The study only focused on air quality forecasting in the Hong Kong region, and the results may not be generalizable to other regions.

				parameters, such as temperature and wind speed, have a significant impact on air pollutant concentrations and should be considered in air quality forecasting models.	
[14]	Simple Recurrent Neural Network(RNN)+Long Short Term Memory(LSTM)+ Bi Directional LSTM(Bi LSTM)+Gated Recurrent Units(GRUs)+Variational Auto Encoder(VAE)	RMSE+MAE +MAPE+RMSLE	Covid19 Data set of 6 countries Italy,Spain,France,USA,China, and Autralia	1. VAE captured all variability in data and provided more accurate forecasting in comparison to other algorithms.	Lack of transparency in data pre-processing Scope of the study is limited as the comparison is done only among a few famous deep learning algorithms. Lack of statistical analysis
[13]	Linear Regression+Multi Layer Perceptron	Correlation +Trend analysis	Dataset of Delhi from 4 places	1. The paper identifies key factors that contribute to air pollution – traffic density, industrial emission and weather conditions	No specific information about the data studied. No future work is clearly mentioned. The paper does not mention limitation of data mining techniques they have used.
[10]	Support Vector Machine(SVM)+XGBoost+Gaussian Naïve Bayes+Random Forest(RF)+K-Nearest Neighbor Algorithm(KNN) For Visualization – Box plot	Correlation, MAE,RMSE	six years of air pollution data from 23 Indian cities for air quality analysis and prediction	1. Gaussian Naïve Bayes performed with highest accuracy. 2. Support Vector machine exhibit lowest accuracy. 3. XG Boost performed best among other models giving highest linearity between predicted values and actual data.	The work can be extended by deploying deep learning techniques on the data set.

Predicators used for air Quality Prediction

As per the survey done above we have discussed the forecasting models available for air quality prediction, now we need to define the various indexes available for air quality prediction.

Mean Absolute Error (MAE)

MAE is the ideal valuation metric for regression models. A model's mean absolute error can be defined as the average or mean of the absolute values of the distinct prediction errors in a test set. The mean absolute error of a model with respect to a test set is the average of the absolute values of the individual prediction errors. A prediction error can be described as a difference between a true value and a predicted value.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (1)$$

Where

MAE = mean absolute error

Y_i = prediction

X_i = true value

n = total number of data points

Mean Squared Error (MSE)

The Mean Square Deviation (MSD) is another name for it. Also called the risk function, this function measures the average squared difference between estimated and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (2)$$

Where,

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

Root Mean Square Error (RMSE) – It is the most commonly used deviation in climatology, forecasting, and regression analysis. It calculates the concentration of data around the line of best fit. Euclidean Distance measures how far predictions differ from the actual measured values.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{X}_i)^2}{N}} \quad (3)$$

Where,

RMSD	=	root-mean-square deviation
i	=	variable i
N	=	number of non-missing data points
X_i	=	actual observations time series
$\hat{X}_i$	=	estimated time series.

Conclusions

This paper methodically reviewed the recent methods of modeling and algorithms available for time series data, particularly in the field of air forecasting. In this paper time series data in context to air is explained and various air pollutants which affect the air and how they are not good for us are explained. All the popular works of different scholars have been reviewed and the important points have been studied. Along with the models, various indexes which were used in the studies by the researchers are also explained in detail Time series mining techniques are a valuable tool for analyzing air quality data and predicting future levels of pollution. Each technique has its advantages and limitations, and the choice of technique depends on the specific research question and data characteristics. Future research should explore the use of ensemble techniques that combine multiple time series mining techniques to improve the accuracy and robustness of air pollution predictions. Additionally, research should investigate the use of time series mining techniques for the analysis of complex air pollution data, such as spatiotemporal and multivariate data. Overall, time series mining techniques have great potential to improve our understanding of air pollution and inform effective policy decisions and interventions.

Author Contributions

In the paper I Apoorva Verma has been the main author of the paper. Collection of data, research papers from highly indexed papers and reviewing them has been done by me. My co-author Dr. Leena Bhatia who is also my research supervisor has majorly worked on conclusion part, reviewing the paper and making the required corrections..

Funding

There is no funding to be declared

Acknowledgments

Not Applicable

Conflicts of Interest

The authors declare that they do not have competing financial interests or personal relationships that may have influenced their work.

References

1. Aamir, M., Shabri, A., & Ishaq, M. (2018). Improving forecasting accuracy of crude oil prices using decomposition ensemble model with reconstruction of IMFs based on ARIMA model. Malaysian Journal of Fundamental and Applied Sciences, 14(4), 471-483.
2. Aladağ, E. (2021). Forecasting of particulate matter with a hybrid ARIMA model based on wavelet transformation and seasonal adjustment. Urban Climate, 39, 100930.
3. Alexandrina, E. C., Ortigossa, E. S., Lui, E. S., Gonçalves, J. A. S., Corrêa, N. A., Nonato, L. G., & Aguiar, M. L. (2019). Analysis and visualization of multidimensional time series: Particulate matter (PM10) from São Carlos-SP (Brazil). Atmospheric Pollution Research, 10(4), 1299-1311.
4. Bandara, K., Hewamalage, H., Liu, Y. H., Kang, Y., & Bergmeir, C. (2021). Improving the accuracy of global forecasting models using time series data augmentation. Pattern Recognition, 120, 108148.

5. Bhatti, U. A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H., ... & Yuan, L. (2021). Time series analysis and forecasting of air pollution particulate matter (PM 2.5): an SARIMA and factor analysis approach. IEEE Access, 9, 41019-41031.

6. Dadhich, A. P., Goyal, R., & Dadhich, P. N. (2018). Assessment of spatio-temporal variations in air quality of Jaipur city, Rajasthan, India. The Egyptian Journal of Remote Sensing and Space Science, 21(2), 173-181.

7. Das, M., Das, A., Ghosh, S., Sarkar, R., & Saha, S. (2021). Spatio-temporal concentration of atmospheric particulate matter (PM2. 5) during pandemic: a study on most polluted cities of indo-gangetic plain. Urban Climate, 35, 100758.

8. Gopu, P., Panda, R. R., & Nagwani, N. K. (2021). Time series analysis using ARIMA model for air pollution prediction in Hyderabad city of India. In Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 1 (pp. 47-56). Springer Singapore.

9. Kulkarni, G. E., Muley, A. A., Deshmukh, N. K., & Bhalchandra, P. U. (2018). Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra, India. Modeling Earth Systems and Environment, 4, 1435-1444.

10. Kumar, K., & Pande, B. P. (2022). Air pollution prediction with machine learning: a case study of Indian cities. International journal of environmental science and technology : IJEST, 1–16 Advance online publication.

11. Liu, T., Lau, A. K., Sandbrink, K., & Fung, J. C. (2018). Time series forecasting of air quality based on regional numerical modeling in Hong Kong. Journal of Geophysical Research: Atmospheres, 123(8), 4175-4196.

12. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. Frontiers in public health, 8, 14.

13. Taneja, S., Sharma, N., Oberoi, K., & Navoria, Y. (2016, August). Predicting trends in air pollution in Delhi using data mining. In 2016 1st India international conference on information processing (IICIP) (pp. 1-6). IEEE.

14. Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. Chaos, Solitons & Fractals, 140, 110121.