International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 14 Issue 2, February-2025

Study on Air Pollution using Multivariate Time Series Analysis

Nidhi Bhardwaj

ABSTRACT

Air pollution vastness has always been a theme of apprehension due to intense development and urbanization excess a period of time. The escalating scale of pollutants in ambient air since last years has deteriorated the air quality in National Capital region at an alarming rate. This bring us to our study on air quality in NCR which includes analysis, prediction and preventive measures of future air quality has been carried out using multivariate analysis. In this paper the multivariate analysis using the approaches of data mining for forecasting.

Keywords: Multivariate, Air Pollution, Data Mining, Forecasting.

INTRODUCTION

Air is the mixture of gases such as nitrogen, oxygen less amount of some other gases that surrounds the earth and help to people, animals and plants for breathe. So now a day's air is very essential for us. India is second highest populated and third most polluted country in the world. According to World health organization study says that 13 of the 20 most polluted cities in the world are in India. It causes tuberculosis, cancer, lungs, skin and eye infection. Vehicle exhaust fumes, fossil fuel-based power plants, emissions from industrial facilities, agricultural and construction activities are all sources of air pollution.Air pollutants have largely classified in two group outdoor and indoor pollutants. Outdoor pollutants contain remains of fossil fuel, metallic particles in the atmosphere from industrial emissions, automobile exhaust, carbon particles, toxic gases i.e. nitrogen, Sulphur dioxide, carbon monoxide and ozone, tobacco and smoke. Whereas, indoor pollutants includes toxic gases produced from construction materials and from kitchen fuels found in the air.It is important to analyze air pollution and forecast future preventive measures for sustainable environment. Researches in air pollution are making suitable decision making for action against air pollution.

RELATED WORK

S. Nidhi [1] this paperforecasts the amount of air pollution in Delhi using data analysis methods. J. K. Sethi and M. Mittal [2] analyzed the data of Gurugram using multivariate time series analysis models namely :Vector Autoregression(VAR) and Autoregression Integrated Moving Average(ARIMA). M. Dhirendra and G. Parmila [3] provides the Artificial neural network (ANN) - A multilayer perception model is being used to anticipate air pollution in Agra.Nath, P., Saha, P., Middya [4] analyses the historical pollution data of Kolkata with the help of time series analysis techniques, which provide forecast for the next two years using different statistical and deep learning method.V, Naveen & N, Anu [5] the seasonal Auto Regressive Integrated Moving Average (SARIMA) and Auto Regressive Integrated Moving Average (ARIMA) methods were used to analysesthe datasets from the Kerala State Pollution Control Board (KSPCB) for air quality forecasting.Nadeem, Imran [6] forecast the concentration of Sulphur Dioxide, Respirable Suspended Particulate Matter(RSPM), Nitrogen Dioxide pollutants using ARMA and ARIMA models.R. Bhardwaj and D. Pruthi [7] analyses the data of Odd-Even Scheme(15 April 2016-30 April 2016), using pre and post Odd Even Scheme in the Dawarka area adjacent to Indira Gandhi International Airport, Delhi that deals with the estimation of Hurst exponent, regression coefficient, Fractal Dimension and Predictability Index of Air Pollutants. N. Djebbri and M. Rouainia [8] analyses and prediction the data of industrial zone(GL1K) complex in Skikadaduring the period from 2013 to Jan 2014 using the method Nonlinear Auto Regressive Model(NARX) method based on Artificial Neural Networkconsidering two contaminants, and the goal is to anticipate NOX and CO pollutant concentrations.S. Taneja, N. Sharma, K. Oberoi and Y. Navoria [9] analyze the existing trend and prediction about future in Delhi Air Pollution using some data mining techniques such as multilayer perceptron and linear regression in which considering some air pollutants.A. Yadav and D. Toshniwal [10] a study has been carried out by using the air quality data from Mumbai, New Delhi, Chennai and Bengaluru. Purposed a modified ARIMA model that is suitable to numeric data streams indicated by SDA has been suggested (Streaming Data ARIMA).P. Singh, T. L. Narasimhan [11] purpose a solution based on Long Short-Term Memory (LSTM) networks that are known to perform

well on sequential prediction problems.S. Mahajan, L. Chen and T. Tsai [12] the hourly forecast of PM2.5 was performed and compared to the Holt-Winters and ARIMA models.K. Nandini and G. Fathima [13] analyze the data of Bangalore, with the help of machine learning techniques predict the air pollution. In this study K-mean algorithm is used for clustering and compared the result of accuracy and error rate of multinomial logistic regression and decision tree algorithm are used to analyzed the data on R programming Language. Multinomial Logistic Regression give high accuracy in comparison to Decision Tree. Liu, Huixiang and Li, Qing [14]purpose of this study to analyze the data of the city Beijing and predicting the concentration of NOx in Italian city.

They used two techniques Support Vector Regression and Random Forest Regression for AQI and NOx concentration prediction. Random Forest give better performance in predicting the NOx concentration and SVR shows better performance in prediction of AQI.C R, Aditya & Deshmukh, Chandana [15] in this paper implementation of two techniques such as logistic regression and autoregression with the help of logistic regression detect whether a data sample is polluted or not. Further with the implementation of autoregression technique predict the future values of PM 2.5 based on the previous data.Dyuthi Sanjeev [16] purpose of this study to present how the machine learning algorithm can be used to analyze and predict the quality of air. In this paper, the implementation of three machine learning techniques such as Random forest, Support vector machine and Artificial neural network are discussed.PoojaBhalgat, SejalPitale, Sachin Bhoite[17] analyze the air pollution data of Maharashtra state and the data is gather from Kaggle website. In this paper AR model and ARIMA model is used for predicting the value of SO2 and the future work of this paper is also consider the PM2.5,NO2 and other component and also find the Air Quality Index.R. O. Sinnott and Z. Guan [18] In this paper, air pollution data, especially particulate matter of less than 2.5 micrometer's (PM 2.5), was gathered from a variety of web-based resources, and the resulting data was analysed with various machine learning models, including linear regression, artificial neural network, and long short term memory recurrent neural network, and the LSTM model was found to be the most accurate.Maleki, Heidar &Sorooshian [19]

The ideal of this paper is to analyses the data of Ahavaz megacity the applied algorithm involved nine factors in the input stage(five meteorological parameters, pollutant attention 3 and 6 h in advance, time, and date), 30 neurons in the retired phase, and eventually one affair in last position. When comparing performance between using 5 and 10 of data for confirmation and testing, the further dependable results were from using 5 of data for these two stages. For all six criteria adulterants examined(O3, NO2, PM10, PM 2.5, SO2, and CO) across four spots, the correlation measure(R) and root-mean square error(RMSE) values when comparing prognostications and measures were 0.87 and 59.9, independently. When comparing modeled and measured AQI and AQHI, R2 was significant for three spots through AQHI, while AQI was significant only at one point. This study demonstrates that ANN has connection to metropolises similar as Ahvaz to read air quality with the purpose of precluding health goods.

Kostandina Veljanovska1 & Angel Dimoski2[20] in this paper compare four machine learning algorithm such as K-nearest neighbor, neural network, support vector machine and decision tree using the database of air pollution contains data from measurement stations in the Republic of Macedonia's capital city for each day of 2017. XiaosongZhao , Rui Zhang [21] in this paper Attempt to predict Air Quality Classification (AQC) on 3 different industrial cities in the United States using a deep learning method.

Deep learning's Recurrent Neural Network (RNN) is used to construct a major prediction model. RNN can process and memorizing sequential data, such as daily air quality data over a given time period. The experimental results demonstrate the performance of three models: SVM, Random Forest, and RNN. When compared to two machine learning approaches, the proposed RNN model produces the best results. T. M. Amado and J. C. Dela Cruz[22] in this paper proposed a methodology of implementing five models of machine learning such as support vector machine, naïve Bayesian classifier, k nearest neighbor, random forest and neural network. S. Sur, R. Ghosal and R. Mondal [23] in this paper, uses a variety of methods and algorithms to identify air pollution hot - spots and predict levels of pollution in a specific area of Delhi.

Time series AQI data is collected in Delhi by CPCB sensors. SVM is used to classify hotspots, and LSTM and PROPHET are used to analyse time series data samples containing pollutants such as PM2.5, PM10, CO, and NO. M. Huang, T. Zhang, J. Wang and L. Zhu[24] the forecasting model, which integrates data mining algorithms and the BP neural network algorithm, is focused on air pollution monitoring data collected from Shijiazhuang ambient air quality monitoring. To begin, this model employs data mining technology to identify the factors that influence air quality. Second, it uses the data from these factors to train the neural network. Finally, the predicting model's evaluation test is performed. The findings indicate that: Because of its higher forecasting accuracy.

Problem Definition

Major source of air pollution in India Country and around the world are industrial emissions and automobile exhaust . Air pollution leads to health hazardous for the human being. So researchers are required to do research in air pollution and provide forecasting and preventive measures using multivariate time series data.

RESEARCH METHODOLOGY

In national capital region major sources of air pollution are automobile drain and industrial emission. Effects of air pollution mainly health hazardous, so duty of researches to analyze the air pollution data for better quality air index in national capital region of India.

Data Set
The data has been taken from secondary source which is central pollution control board (CPCB). The dataset contains multiple attributes which help us to obtain effective analysis of data sets.

PROPOSED METHODOLOGY

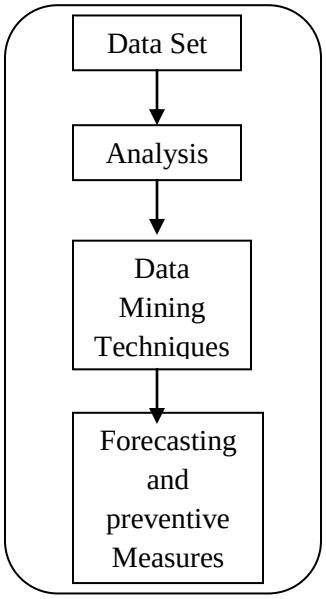

Figure 1: Proposed Approach for Multivariate time series analysis

Figure 1 shows that the approach for the proposed methodology which starts from collection of data set, analysis which includes cleaning of data. Data mining techniques are used for forecasting and analysis for preventive measures.

Data Mining Techniques
There are numerous application domains of data mining with time series data, with multivariate time series analysis for forecasting being one of the most important. Computational methods, such as the Genetic Algorithm or the Neural Network. Some papers in the literature use hybrid approaches, which means they combine more than one technique. The method that is commonly used for dealing with time-based data is known as "time series data," and the models that we develop for it are known as "time series modeling." Time series data is generally based on time (minutes, hours, days, months, and years) and is used for forecasting and predicting future events based on historical events. Time series modeling is also categorized in two ways Univariate time series and Multivariate time series.

Classification: Data mining mainly follows three approaches are supervised learning, unsupervised learning and reinforcement learning. So classification comes under the category of supervised learning. This method is used to obtain critical and relevant information about data. This data mining technique aids in the classification of data into various categories. Data mining techniques are classified using the following criteria such as data source mined, on the bases of rules. There are several clustering techniques such as Support Vector Machine, Decision Tree, Baysian Classifier Model, Neural Network etc.

Regression: Regression analysis is a data mining procedure that identifies and analyses the relationship between variables as a result of the existence of another factor. It is used to define the likelihood of a given variable.A data mining function that predicts a quantitative or continuous value.

Both classification and regression procedures provide the same purpose (predictive analysis), but classification divides the data into distinct groupings. The regression method is used to determine the connection between dependent and independent variables. It converts a data point into a real-valued prediction variable.

Regression also indicates changes in one variable as a result of changes in another one. One essential aspect of regression is that it describes the link between variables in greater depth than correlation - it indicates the strength of the association between variables. Regression problems are frequently addressed as classification problems with quantitative class labels. There are various regression techniques such as Linear Regression, Lasso Linear Regression, Ridge Regression etc.

Association :This data mining approach aids in the discovery of a relationship between two or more things. It discovers a hidden pattern inside the data collection. Association rules are if-then statements that help to demonstrate the likelihood of interactions among data items in huge data sets in various types of databases. Association rule mining seems to have a wide range ofapplications, although it is most typically used to aid in sales correlations in data or medical data sets.There are several association techniques such as Apriority Algorithm, Dynamic Item Set Counting etc.

Clustering :Is an unsupervised learning strategy for determining the probability of grouping unlabeled data items. Clustering determines which data items are more important to each other and which differ from data components in other groupings. Clustering may be divided into two types: conventional clustering and conceptual clustering. There are several clustering algorithm such as K-Mean clustering, Hierarchical Model etc.

Sequential pattern analysis :The sequential pattern is a data mining approach that is used to uncover sequential patterns by examining sequential data. It consists of locating interesting subsequences in a group of sequences, where the value of a sequence may be quantified using several parameters like as length, occurrence frequency, and so on. In other sense, this data mining approach aids in the discovery or recognition of comparable patterns in transaction data over time.

Some statistical techniques
Simple moving average (SMA) :One of the most basic forecasting techniques. Essentially, this approach computes the average. It entails accumulating the values from the previous 'n' periods and then dividing the total by 'n'. As a result, the moving average value is used to anticipate the future period.

Exponential smoothing
As the observation gets older, exponential smoothing assigns it a diminishing weight. When the observation was made more recently, the exponential weight increased. Exponential smoothing is most commonly employed for short-term forecasting, and it is not appropriate for long-term forecasting.

Simple exponential smoothing
For short-term forecasting, you can apply a basic exponential smoothing method on time series data that can be characterized using an incremental approach with consistent level and no seasonality.

Holt's exponential smoothing
In this scenario, utilize holt's exponential smoothing on a time series data that can be explained using an additive model with an increasing or decreasing trend but no seasonality.

Winter's three parameter linear and seasonal exponential smoothing
In the event of time series data that can be characterized using an additive model with growing or decreasing trend and seasonality, the holt-Winter exponential smoothing approach can be used to make short-term forecasting.

Autoregressive integrated moving average(ARIMA)
An autoregressive model is a statistical model that uses time series data to better understand a dataset or to forecast future trends based on existing data.
AR - determines a link between the number of lag observations included in the model and the number of lag observations themselves (p).

The degree of differencing is given by I – Integrated, which calculates the difference of raw observations utilised to make the time stationary (d).

MA - Moving average is a model that uses the relationship between an observation and a residual error from a moving average model (q).

Performance Evaluation Metrics

Performance evaluation metrics is mainly used for error calculating or find out the accuracy of the model. It is mainly used in forecasting model. In performance evaluation metrics error find out on the basis of actual value and forecast value. Some performance evaluation metrics are Mean Squared Error(MSE), Root Mean Squared Error(RMSE), Mean Absolute Error(MAE) and Mean Absolute Percentage Error(MAPE).

In these metrics A_i denotes the actual values of a variable, F_i denotes the predicted values of a variable, and N *are* the number of observations available for analysis.

Mean Squared Error(MSE) : refers to average or mean of the square of the difference between actual values and forecasting or estimated values in statistics.

$$\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)$$

Root Mean Squared Error(RMSE) : refers to square root of the mean or average of the square of all errors is the root mean squared error (RMSE). The RMSE error standard is extensively used by experimenter for exploration purpose. RMSE is a good measure of delicacy, but only to compare vaticination crimes of different models or model configurations for a particular variable and not between variables, as it's scale-dependent.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2}$$

Mean Absolute Error(MAE) : refers to the average magnitude of the difference between actual values and forecasting or estimated values.

$$\frac{1}{N}\sum_{i=1}^{N}|F_i - A_i|$$

Mean Absolute Percentage Error(MAPE) : refers to average or mean of all anticipated absolute percentage errors. Error is defined as the difference between the actual value or observed value and the and forecasting or projected value. To compute MAPE, percentage errors are added together without respect to sign. Because the errorsis expressed in percentages, this metric is simple to comprehend. The problem of positive and negative mistakes cancelling each other is also avoided when absolute percentage errors are employed. As a result, MAPE has management appeal and is a frequent forecasting metric.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|F_i - A_i|}{A_i}$$

CONCLUSION

We conclude that this paper gives techniques for creating awareness and decrease pollution by adopting proper preventive measures. The analyses of multivariate time series forecasting in NCR region provides better understating of air pollutants and their future effects.

REFERENCES

[1]. S. Nidhi, “Forecasting air pollution load in Delhi using data analysis tools”, Proceedings of the Elsevier International Conference on Computational Intelligence and Data Science , pp. 1077-1085, 2018.

[2]. J. K. Sethi and M. Mittal, "Analysis of Air Quality using Univariate and Multivariate Time Series Models," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 823-827, doi: 10.1109/Confluence47617.2020.9058303.

[3]. Mishra, Dhirendra. (2014). Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra. Atmospheric Pollution Research. 6. 10.5094/APR.2015.012.

[4]. Nath, P., Saha, P., Middya, A.I. *et al.* Long-term time-series pollution forecast using statistical and deep learning methods. *Neural Comput&Applic* (2021). https://doi.org/10.1007/s00521-021-05901-2.

[5]. V, Naveen & N, Anu. (2017). Time Series Analysis to Forecast Air Quality Indices in Thiruvananthapuram District, Kerala, India. International Journal of Engineering Research and Applications. 07. 66-84. 10.9790/9622-0706036684.

[6]. Nadeem, Imran & Ilyas, Ashiq &Uduman, P.S.. (2020). Forecasting Ambient Air Quality Of Chennai City In India. GEOGRAPHY ENVIRONMENT SUSTAINABILITY. 13. 12-21. 10.24057/2071-9388-2019-97.

[7]. R. Bhardwaj and D. Pruthi, "Time series and predictability analysis of air pollutants in Delhi," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016, pp. 553-560, doi: 10.1109/NGCT.2016.7877476.

[8]. N. Djebbri and M. Rouainia, "Artificial neural networks based air pollution monitoring in industrial sites," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-5, doi: 10.1109/ICEngTechnol.2017.8308151.

[9]. S. Taneja, N. Sharma, K. Oberoi and Y. Navoria, "Predicting trends in air pollution in Delhi using data mining," 2016 1st India International Conference on Information Processing (IICIP), 2016, pp. 1-6, doi: 10.1109/IICIP.2016.7975379.

[10]. A. Yadav and D. Toshniwal, "Extracting Patterns and Variations in Air Quality of Four Tier I Cities in India," 2017 IEEE World Congress on Services (SERVICES), 2017, pp. 17-20, doi: 10.1109/SERVICES.2017.12.

[11]. P. Singh, T. L. Narasimhan and C. S. Lakshminarayanan, "DeepAir: Air Quality Prediction using Deep Neural Network," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 869-873, doi: 10.1109/TENCON.2019.8929470.

[12]. S. Mahajan, L. Chen and T. Tsai, "An empirical study of PM2.5 forecasting using neural network," 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud &Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1-7, doi: 10.1109/UIC-ATC.2017.8397443.

[13]. K. Nandini and G. Fathima, "Urban Air Quality Analysis and Prediction Using Machine Learning," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, pp. 98-102, doi: 10.1109/ICATIECE45860.2019.9063845.

[14]. Liu, Huixiang& Li, Qing & Yu, Dongbing& Gu, Yu. (2019). Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. Applied Sciences. 9. 4069. 10.3390/app9194069.

[15]. C R, Aditya & Deshmukh, Chandana & K, Nayana & Gandhi, Praveen &astu, Vidyav. (2018). Detection and Prediction of Air Pollution using Machine Learning Models. International Journal of Engineering Trends and Technology. 59. 204-207. 10.14445/22315381/IJETT-V59P238.

[16]. Dyuthi Sanjeev, 2021, Implementation of Machine Learning Algorithms for Analysis and Prediction of Air Quality, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 03 (March 2021).

[17]. PoojaBhalgat, SejalPitale, Sachin Bhoite, “Air Quality Prediction using Machine Learning Algorithms”, International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656.

[18]. R. O. Sinnott and Z. Guan, "Prediction of Air Pollution through Machine Learning Approaches on the Cloud," 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), 2018, pp. 51-60, doi: 10.1109/BDCAT.2018.00015.

[19]. Maleki, Heidar &Sorooshian, Armin & Goudarzi, Gholamreza &Baboli, Zeynab & Tahmasebi Birgani, Yaser & Rahmati, Mojtaba. (2019). Air pollution prediction by using an artificial neural network model. Clean Technologies and Environmental Policy. 21. 10.1007/s10098-019-01709-w.

[20]. Kostandina Veljanovska1 & Angel Dimoski2, Air Quality Index Prediction Using Simple Machine Learning Algorithms,2018, International Journal of Emerging Trends& Technology in Computer Science (IJETTCS).

[21]. Xiaosong Zhao , Rui Zhang, Jheng-Long Wu, Pei-Chann Chang and Yuan Ze University, A Deep Recurrent Neural Network for Air Quality Classification, 2018, Journal ofInformation Hiding and Multimedia Signal Processing.

[22]. T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," TENCON 2018 - 2018 IEEE Region 10 Conference, 2018, pp. 0668-0672, doi: 10.1109/TENCON.2018.8650518.

[23]. S. Sur, R. Ghosal and R. Mondal, "Air Pollution Hotspot Identification and Pollution Level Prediction in the City of Delhi," 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), 2020, pp. 290-294, doi: 10.1109/ICCE50343.2020.9290698.

[24]. M. Huang, T. Zhang, J. Wang and L. Zhu, "A new air quality forecasting model using data mining and artificial neural network," 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2015, pp. 259-262, doi: 10.1109/ICSESS.2015.7339050.