AQI PREDICATION USING TEMPORAL DATA MINING

[*1]Shashi Bhushan, [2]Dr. Sanjay Kumar Tiwari

*[*1]Research Scholar, Magadh University Bodhgaya, Bihar*

[2]Associate Professor, P. G. Department of Mathematics, Magadh University, Bodhgaya.

*[*1]Gmail: shashiibhushannn9@gmail.com*

ABSTRACT

The Air Quality Index (AQI) is an air quality standards indicator based on air pollutants that have negative impacts on human health and the environment. Because of several human activities, air pollution is growing very quickly, and it is the introduction of chemicals, particulate matter or biological materials into the atmosphere that cause human suffering and also harms the natural environment. Indeed, air pollution in metropolitan and industrial cities is one of the major environmental problems. So predicting pollution and avoiding these issues is very crucial. One of the most exciting and difficult functions is the forecast of air pollution using data mining. Many systems are designed to help data storage, inventory management and convenient statistics generation. India's air quality indicator is a standard measure used to indicate pollution (so2, no2, rspm, spm, etc.) from time to time. The main purpose of the current study is to predict the temporal AQI used by the previous day AQI and climate change is used to predict and visualize the temporary data mine using a gradient break and an unreasonable forecasting process. In Navigation Forecast, we divide the database into 85% data and 15% data based on data testing and training to determine seasonal variations and styles. Balance problems are often exploited by problems and forecasting uses an unreasonable prediction process and gradient downtime. Air quality forecasts based on historical data of previous years and predictions for less than a year as a reputable gradient using a recurring problem.

KEYWORDS-AQI, Dataset, Predication, Naïve Forecasting, Gradient Descent

1.INTRODUCTION

Temporal data mining is a rapidly changing learning area where a variety of disciplines meet, including statistics, temporal pattern recognition, temporal data details, efficiency and insight, advanced computing performance, and computing. Temporal data mining is a one-dimensional information system, the process of obtaining temporal data that calculates the formation of spatial data (spatial patterns or models) and any algorithm that calculates spatial patterns if the

local data is a spatial information algorithm. Temporal data mining often works from data, and the most popular strategies are those that focus on a large amount of data related to time, using the data collected for as long as possible to reach reliable conclusions. [1] Air pollution is a major global problem that involves the accumulation of pollutants such as ozone, carbon dioxide, nitrogen oxides, and particles such as PM 2.5 (Particle Matter). Among all these pollutants, PM 2.5 is very dangerous to human health because it can be inhaled directly into the lungs and spilled into the bloodstream, causing high levels of lung cancer. These air pollutants cause extreme effects on a human being's environment, such as human disease, sickness or death, risk to other living organisms, such as food harvesting, or harm to the natural environment. Therefore, air pollution must be monitored. Many decision support systems are designed to monitor data, but they are mainly restricted.Predicting air pollution using data mining is one of the most interesting and difficult tasks and we provide the predictive methods used to measure air pollution next day, next month, to prevent problems.[2]

The Air Quality Index (AQI) is a specific number used by government agencies, so this number helps indicate the air quality in a particular area. The AQI scheme converts the weighted parameters associated with air pollution into a single number or set of numbers. The AQI is used to manage national and regional air quality in major cities around the world [3]. The main objective of this study was to predict the provisional AQI using the previous AQI and climate change by predicting and presenting provisional data. Mines use gradient downtime and unpredictable forecasts. We have compiled data from the Indian government database that covers pollution issues across India. We have developed a system to predict air quality indicators in all accessible data areas. Maintain Indian air quality in any region. By guessing the air quality gauge, we can change the high pollution levels caused by pollution in India. A model based on pollution monitoring information from pollution sources, a climate-based model and a data-driven model. .. statistics, etc., advanced techniques used to predict air quality. Monitoring of pollution source data is difficult to obtain and the accuracy of this information is not high. This document is particularly applicable to the predicted air quality model of an urban area based on meteorological data and historical pollution data. Because air quality forecasts are highly dependent on local climates, the same source of pollution caused by air pollution varies with different weather conditions. The level of pollution on this day also affects air quality, which is why air quality forecasts have strong inaccurate indicators. [4]

Forecasting includes generating a number, number set, or scenario that matches a future event. Short and long term planning is very important. By definition, prediction is based on prior knowledge and not on automatic predictions based on instinct, feelings, or imagination. For example, descriptive definitions, a process sometimes used in prediction, often mean that their purpose is to explain or "predict." As the weather forecast approaches, the weather accuracy improves. For example, tomorrow's forecast will be more accurate than next month's forecast; The forecast for next month will be more accurate than the forecast for next year. Projections for the coming year will be more accurate than those for the next decade.

William J. Stevenson lists a number of characteristics that are common to a good forecast.1.Accurate-In order to make a comparison with alternative forecasts, some degree of precisionshould be determined and indicated.2.Reliable-If the user is to create some degree of confidence, the prediction technique should continuously provide a good prediction.3.Timely-It takes a certain quantity of moment to react to the prediction, so that the prediction horizon has to allow the time required to make adjustments.4. Cost Effective: The cost of making a forecast should not exceed the benefits derived from the forecast. An air pollution index can be defined as a system that converts the values (weight) of individual parameters related to air pollution, such as concentration, visibility, humidity, etc., into a single number or set of numbers. As shown in Table 1

Table 1: Description For various range of AQI

Standards	Description
0-50	The AQI is normal The air quality is better and therefore no cause for fear.
51-100	The AQI is moderately high The air quality is moderate and should be conserved.
101-150	The AQI is high The air quality is unhealthy for sensitive peoples.
150-200	The AQI is high The air quality is unhealthy.
201-300	The AQI is high The air quality is unhealthy attention is required.
301-500	The AQI is very high The air quality is hazardous immediate attention is required.

2.METHODOLOGY-

The design of the plan is emphasized in a building diagram that describes the structure and behavior of the plan. Construction begins with a menu that follows the information management in the database. When the processing is complete, statistical graphs are generated annually and AQI is calculated as shown in Figure 1.

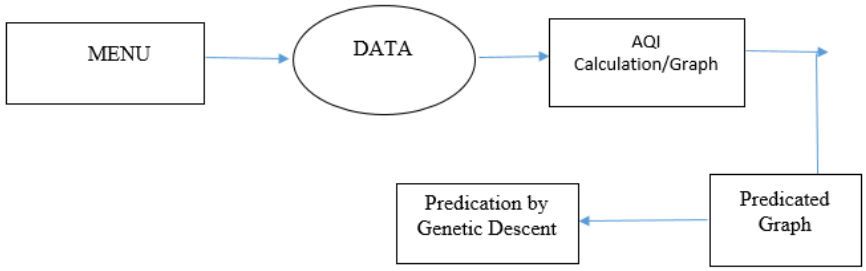

Figure.1 System Architecture

3.BLOCK DIAGRAM

1. **Data Collection-**We collected online data fromair quality monitoring sites from 1990 to 2014.The air pollutant data in this study included the concentrations of O3, PM2.5 and SO2.We chose the meteorological factors that would influence the levels of air pollutants, including air temperature, relative humidity, wind speed and direction, wind rainfall, accumulation of precipitation, visibility, dew point, wind direction, pressure and weather conditions.

2. **Performance Evaluation-**The statistical criteria such as mean absolute error (MAE), mean absolute percent error (MAPE), correlation coefficient (R), and root mean square error (RMSE) were selected to assess the efficiency of each regression model. The following equations are provided.

$$MAE = \frac{\sum_{k=1}^{n} | t_k - y_k |}{n}$$

Mean absolute error (MAE)

$$MAPE = \frac{\sum_{k=1}^{n} \left| \frac{t_k - y_k}{t_k} \right|}{n} \times 100\%$$

Mean absolute percent error (MAPE)

$$R = \frac{\sum_{k=1}^{n} (t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (t_k - \bar{t})^2 \sum_{k=1}^{n} (y_k - \bar{y})^2}}$$

Correlation Coefficient (R)

$$RMSE = \sqrt{\frac{1}{n} \sum_{k=1}^{n} | t_k - y_k |^2}$$

Root Mean Square Error (RMSE)

If n is the number of data points, k y is the expected value, k t is the observed value, t is the average of the observed data, and y is the average of the observed data. The quality of test information is evaluated because it represents the accuracy of each regression model.

3. **Naive Forecasting**- How to estimate how the actual expiration time is used as a weather forecast, without changing it or trying to find the causal factors. Used to compare only predictions for better (advanced) methods. Naïve Forecast offers a comparison benchmark separate from the final prediction showing whether or not the original final prediction is improved. Naïve Forecast is like baseline prediction based on facts and information, but for many organizations it is still unexplored. It is a method that is implemented in forecasting at the primitive level and promotes the fundamental comparison standard. The Organizations analyze whether the naïve forecast is inferior or superior from the final forecast generated by the organization.

4.TECHNIQUES USED

1. **LINEAR REGRESSION-**

Linear regression is a numerical manner in which interaction between two continuous variables is reviewed and revised.Linear regression is a linear approach to modeling the connection between a scalar dependent y variable and one or more explanatory X-denoted variables.Simple linear regression is called the case of a descriptive inconsistent

value. The method is called multiple linear regression for more than one variable.The connections are modeled in linear regression using linear predictor features whose information estimate unknown model parameters. These models are referred to as linear models. Specifies the lowest square waning line for the set of n information points,

y = ax + b

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\operatorname{Cov}[x, y]}{\operatorname{Var}[x]} = r_{xy}\frac{s_y}{s_x},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x},$$

2.GRADIENT DESCENT ALGORTHIM-

Gradient reduction is an optimization algorithm used to determine the parameters (coefficients) of an activity (f) that reduces labor costs (costs). Gradient reduction is best used when parameters cannot be calculated by analysis (for example, using straight algebra) and should be obtained using the optimization algorithm. Gradient reduction is an algorithm for reducing activity. Gradient downtime can reduce performance on very large data sets. Because the repetition of the gradient descent algorithm requires prediction of each situation in the training database, it can take some time if you have several million cases. for large amounts of data, you can use a different gradient drop called a stochastic gradient drop.

In linear regression, gradient descent is not only used; it is a more general algorithm. Now we'll learn how to use gradient descent algorithm to minimize some arbitrary feature f and then apply it to a cost function to determine its minimum. We will start off by some initial guesses for the values of θ0 and θ1 and then keep on changing the values according to the formula:

$$\theta j := \theta j - \alpha \partial \partial \theta j f(\theta 0, \theta 1) \text{ for } j = 0,1$$

Here, the learning rate is called α, and it determines how large a step is needed when updating the parameters. The learning pace is a positive number at all times.We want to update both j=0

and j=1 concurrently, i.e. calculate the right-hand side of the above equation, and then update the parameter values to the freshly calculated ones. This process will be repeated until convergence is attained.In this paper gradient descent are used for compassions between actual value and predicated value.

5.RESULTS AND DISCUSION

To predict the air quality index for a specific region, we need the polluting focus of all available gases on the cpcb.nic.in website, which contains all the information that pollutes cities each year. AQI (Air Quality Index) equations will be applied in order to calculate AQI using linear regression algorithm for a given year. Several datasets are inserted into the folder and the infinite information is set to null values. As shown in figure.2

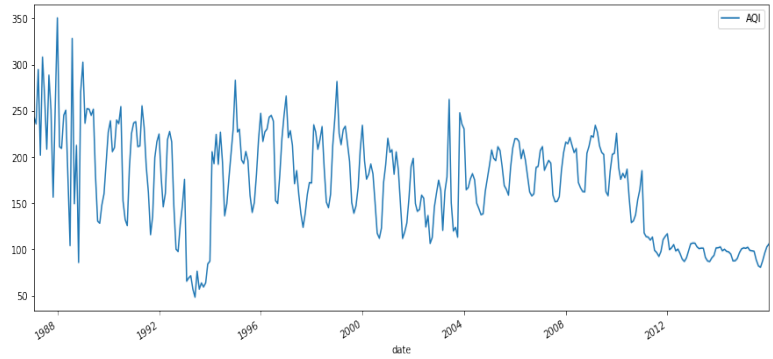

Figure:2 Time Series Visualization

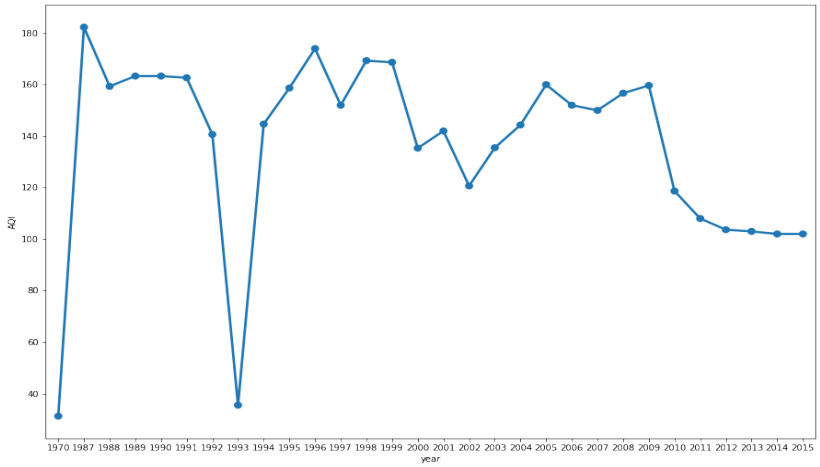

Figure.3 Visualization output year wise

In this graph AQI is the average value of AQI of each year across India.A specific data point's air quality index is the aggregate of the highest indexed pollutant in that region. As shown in figure 3. That pollutants max-sub index is taken as that specific location's air quality index.

Air Quality Index Predication-Using Naïve Forecast strategy, we split the information set into two components of the first 85 percent and stored 15 percent of information in test and train datasets to recognize the enormous seasonal differences and trends.Naïve Forecast is like baseline prediction based on facts and information, but for many organizations it is still unexplored. In this paper we define the naïve forecast strategy result according to the year wise. In which we show the difference between actual and predication values used in between AQI and year (1988-2016).We calculated our data points ' moving average and plotted the moving average. As shown in figure 4..

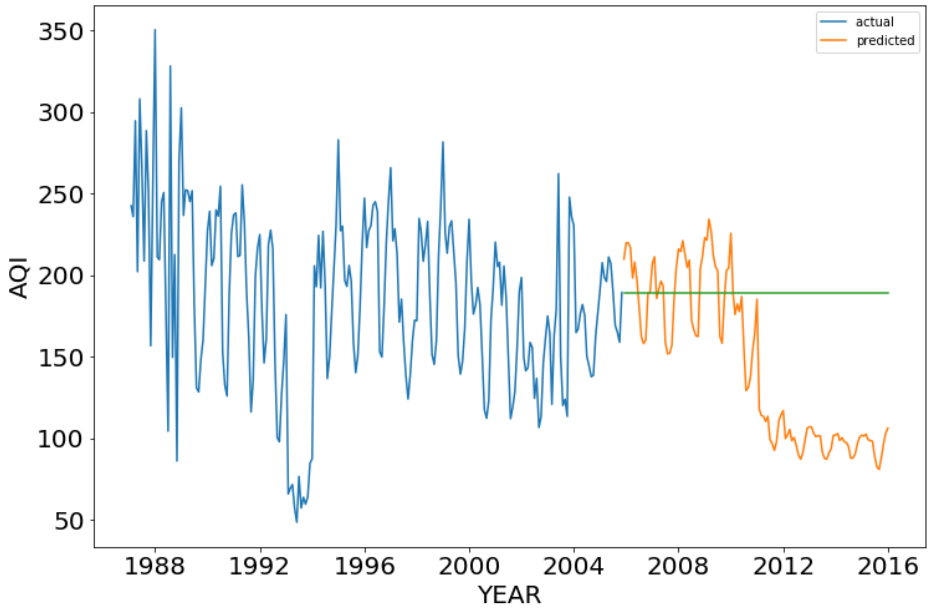

Figure.4 Naïve Forecast

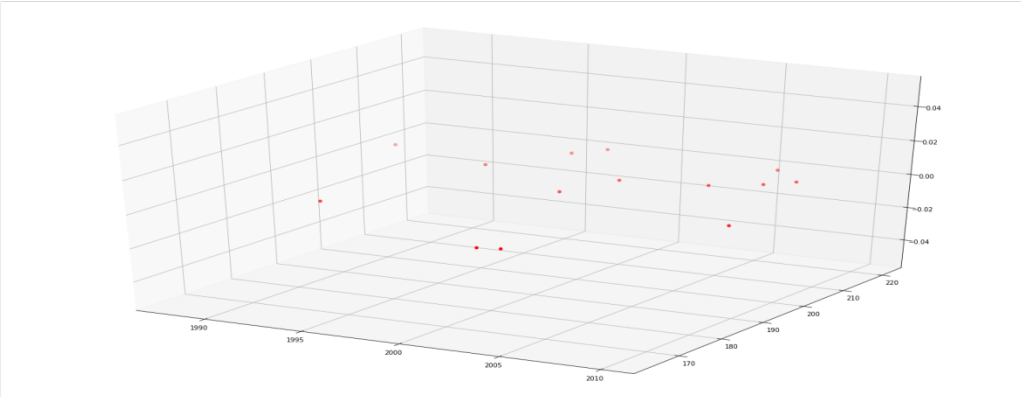

Figure.5 Temporal Data Mining Output

It predicated the temporal data mining in 3-D axis form in which they are show result in year wise.As shown in figure 5.

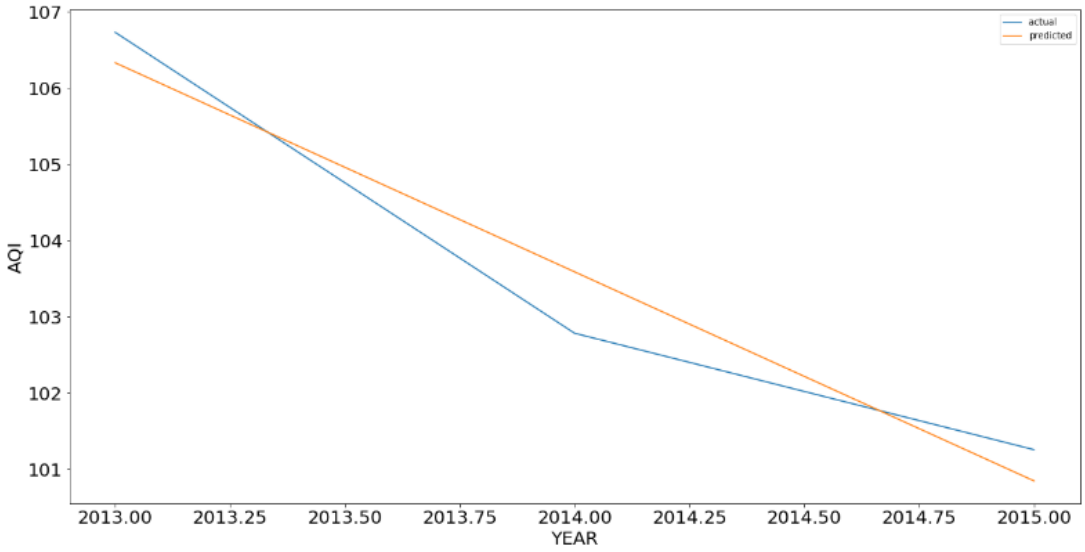

Figure.6 Actual Vs. Predicated in Gradient Descent (2013-2015)

Gradient Descent used iterations and compare actual and prediction values year wise. Different outputs of results are used year wise. As shown in figure 6. The main problem affected by people is air contamination, since air includes countless substances that can be manufactured or performed regularly.

6.CONCULSIONS

It is quickly becoming one of the most significant duties to regulate air pollutant concentrations. It is essential for individuals to understand what the amount of pollution is in their environment

and to take a step towards combating it. The result indicates that the unsupervised learning methods (logistic regression and gradient descent) can be used efficiently to identify air qualityindex and predict future of RSPM levels.With this model, we can predict the AQI and inform the respected region of the state as well as it is a progressive learning model which can trace back to the specific place that needs attention, given that the time series data of each region requires attention

The implemented model will help ordinary citizens as well as those in the meteorological department in detecting and predicting amounts of pollution and following the required action appropriately.This will also assist individuals to set up a information source for small towns that are generally left out compared to big towns.

Funding:

There is no funding from any Research or Funding Agency

Conflict of Interest:

The authors declare that we have no conflict of interest.

Authorship change form

I have confirm the all authors

REFERNCES

1. Lin, W., Orgun, M.A. and Williams, G.J., 2002, December. An Overview Of Temporal Data Mining. In *AusDM* (pp. 83-90).
2. Azid, A., Juahir, H., Toriman, M.E., Kamarudin, M.K.A., Saudi, A.S.M., Hasnam, C.N.C., Aziz, N.A.A., Azaman, F., Latif, M.T., Zainuddin, S.F.M. and Osman, M.R., 2014. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, & Soil Pollution*, *225*(8), p.2063.
3. Ma, Y., Richards, M., Ghanem, M., Guo, Y. and Hassard, J., 2008. Air pollution monitoring and mining based on sensor grid in London. *Sensors*, *8*(6), pp.3601-3623.
4. Huang, M., Zhang, T., Wang, J. and Zhu, L., 2015, September. A new air quality forecasting model using data mining and artificial neural network. In *2015 6th IEEE*

International Conference on Software Engineering and Service Science (ICSESS) (pp. 259-262). IEEE.

5. Dixian Zhu, Changjie Cai, Tianbao Yang and Xun Zhou: A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big data and cognitive computing.
6. Carbajal-Hernández, José Juan "Assessment and prediction of air quality using fuzzy logic and autoregressive models." Atmospheric Environment 60 (2012): 37-50.
7. Kumar, Anikender and P. Goyal, " Forcasting of daily air quality index in Delhi", Science of th Total Environment 409, no. 24(2011): 55175523.
8. Wang Z et al , " A nested air quality prediction modelling system for urban and regional scales : Application for high high-ozone episode in Taiwan " Water, Air and Soil Pollution130.1-4(2001):391-396.
9. Russo Ana Frank Raischel and Pedro G.Lind, "Air quality prediction using optimalneural networks with stochastic variables", Atmospheric Environment 79(2013): 822-830.
10. Sivacoumar R, et al, "Air pollution modelling for an industrial complex and model performance evaluation ", Environmental Pollution 111.3 (2001) 471-477.
11. Singh Kunwar P., Shikha Gupta and Premanjali Rai, "Identifying pollution sources and prediction urban air quality using ensemble learning methods", Atmospheric environment80 (2013): 426-437.
12. Peng, H., 2015. *Air quality prediction by machine learning methods* (Doctoral dissertation, University of British Columbia).
13. M. Caselli & L. Trizio & G. de Gennaro & P. Ielpo. "A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model." Water Air Soil Pollut (2009) 201:365–377.
14. S.Bordignon, C. Gaetan and F. Lisi, "Nonlinear models for ground-level ozone forecasting." Statistical Methods and Applications, 11, 227-246, (2002).
15. Edwin Diday. Symbolic data analysis: a mathematical framework and tool for data mining. In Advances in Data Science and Classification, pages 409–416. Springer, 1998.