2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science

Air Quality Index Forecasting using Auto-regression Models

2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS) 978-1-7281-4862-5/20/$31.00 ©2020 IEEE 10.1109/SCEECS48394.2020.216

Nimisha Tomar
Department of CSE
Maulana Azad National Institute of Technology, Bhopal
nimisha.tomar28@gmail.com

Durga Patel
Department of CSE
Maulana Azad National Institute of Technology, Bhopal
durga7654patel@gmail.com

Akshat Jain
Department of CSE
Maulana Azad National Institute of Technology, Bhopal
akshatjain1011@gmail.com

***Abstract*— As the impact of air pollutants is increasing drastically on living beings in recent years, it has become very necessary to address the issue of air pollution control by scientists and environmentalists. To ensure the same, it is important to forecast air quality in terms of parameters that cause air pollution directly or indirectly, which generally affects the living population. US Environmental Protection Agency (USEPA) has suggested a method to estimate air quality index in a region which constitutes different concentration of criteria air pollutants such as RSPM, SO2, NO2, and SPM. The motive of our research is to model a predicting model for forecasting daily AQI, that can be put to use for local and regional air quality management.**

Keywords— AQI, AR, ARIMA.

I. INTRODUCTION

With recent awareness of humans regarding the various issues of pollution, many steps are taken to reduce or eradicate them for a sustainable future. These issues may be related to water pollution, soil pollution, or air pollution. One of the main problems is air pollution as it affects human health directly. Acid Rains, Respiratory disorders like asthma, tuberculosis, and various cancers are consequences of Air pollution. The meteorological factors like speed and direction of the wind, humidity, temperature, rainfall patterns, solar radiation and artificial factors like gases forming industrial waste, pollutant gases from vehicles; and used fuel from the radioactive activity and hydrogen bomb testing, etc. play significant roles in causing and alleviating air pollution. First of all to control the air pollution a method is required to forecast the measure of air quality or air quality index. So, this model aims to predict the accuracy by which the air quality index can be predicted.

Air pollution is generally caused due to various industries, power plants, vehicles and also natural disasters and at the current rate the air will not be breathable after 2045, so it is the right time to take strict action against it and having proper measures to prevent this from happening.
So, accurate air quality forecasting can alert the authorities as well as the common population making it an important aim for society.

II. RELATED WORK

Till date, many researchers have proposed several methods to solve this problem of air quality prediction. These methods are discussed below in this section.

Kalapanidas et al. [1] discussed the NEMO prototype that was built to support the short-term forecast of Nitrogen dioxide. They classified pollution level into four levels (a) low, (b) med, (c) high, and (d) alarm by using a lazy learning approach, the case-based reasoning system.

Athanasia et al. [2] describes a novel classifier, namely σ-FLNMAP, (σ-fuzzy lattice) neuro computing classifier to predict and classified O3 concentrations into three levels based on Wind velocity, Temperature, Relative humidity, etc. and pollutants such as SO2, NO, NO2, O3 (ozone level) and so on.

[3] Here, meteorological calculations and air quality index values were given as data to feed-forward back-propagation neural networks and predicted daily concentration levels of pollutants for the next three days. They concluded non-geographic plain model gives considerably higher error than the distance-based geographic model.

Corani [4] examined the performances of feed-forward neural networks with pruned neural networks and lazy

978-1-7281-4862-5/20/$31.00 ©2020 IEEE

learning and predicted hourly O3 and PM10 concentrations based on data from the previous day.

Jiang et al. [5] examined various models on the air quality forecasting assignment such as physical and chemical models, regression models, and the conclusions determine that statistical models are more reliable than the chemical physical and classical models.

In [6], the author examined multiple statistical models on PM2.5 data, and their conclusions indicated that models based on linear regression can be proved to be performing more reliable than the other models.

Many works have been introduced to implement machine learning and deep learning models to forecast the air properties. Some of the algorithms are presented here:

[7] Li et al. evaluated the quality of air using Spatio-temporal interpolation approaches.

[8] Used the affinity graph method to administer with the air condition of a geological area and recognizing the preferable areas to set adviser services.

[9] Developed a cryptic semi Markov model to forecast PM2.5 concentration.

[10] Used the methods based on fundamentals of regression and neural networks to forecast the intensity of air contaminants in the air.

III. DATASET

Dataset/Source: Kaggle

Structured/Unstructured data: Structured Data in CSV format.

Dataset description:

The data consist of measurements of different pollutant and meteorological quantities taken at hourly basis, in which each row consists of single hour measurements. We have created time chunks of 11 days. Training data has 8 days of each time slice available. Following has been provided in the training data:

- Row ID
- time_slice ID
- slice_position (starts at 1 for each slice of data, increments every hour)
- most_common_month (most common month in each data slice) weekday (day of the week, as a string)
- hour (local time)
- solar_radiation_64
- direction_of_wind (angle of the wind given in angle, e.g. a wind from the west is "45")
- speed_of_wind
- Speed_of_wind_2038 ("2038" is site no)
- Maximum_ambient_Temperature_(site no)
- Minimum_ambient_Temperature_(site no)
- Barometer. Pressure_(site no)
- Maximum.Barometer.Pressure_(site no)
- Minimum.Barometer.Pressure_(site no)

The variables indicated with the "_(site_no)" are available for various sites, and, "_(target_no)" will vary across different targets.[15]

IV. METHODOLOGY

For serially correlated data, Time series models are proved to be better as compared with other models. Therefore ARMA model is used. Here AR is auto-regression and MA is moving average. But they are not applicable on non-stationary series. So the biggest challenge is to stationarize and predict the time series if it is not stationary with help of stochastic models. Methods like Detrending, Differencing etc. are some ways of converting a non stationary series to a stationary series.

Methods to stationarize series: We can check the stationarity of the time series as soon as we know about the different patterns, cycles, trends, and seasonality of the given series. One of the most used test is the Dickey – Fuller for checking stationarity. Following are some techniques mostly employed to make a time series stationary:

1. **Detrending** : It means to eliminate the trend element from the time series. For instance, if the equation of time series is:

f(c) = (mean + trend * c) + error

2. **Differencing**: It is the generally adopted technique to eliminate non-stationarity. It means to minimize the differences of the terms and not the actual term. For instance,

f(c) – f(c-1) = ARMA (p , q)

This differencing is called as the Integration component in AR(I)MA. Following are the parameters:
p: AR
d: I
q: MA

3. **Seasonality**: This can simply be combined in the ARIMA model directly.

Models to be used for forecasting are-

(1) AR (auto-regressive) Model-

It simply uses the values from last time and then put them in regression equations to predict the next value. It is easily able to give good forecasts on a variety of conditions. As it uses data from previous steps so it is called auto-regressive.

yt = m + nt*xt

Here,

Y = prediction

M and N = coefficients

X = input value

M and N are calculated while training to give the best results.

Here the value of the next time step can be predicted with help of the previous steps :

Y = x(t+1) = m + n1*x(t) + n2*x(t-1) + n3*x(t-2)

(2) ARIMA Models-

Auto-Regressive Integrated Moving Average is the most customary way to predict the future vales by making them stationary by logging or deflating. A value is stationary or stable if its statistical characteristics are always fixed. Its variation around its mean has a constant amplitude and is consistent i.e.; its power spectrum remains constant. The ARIMA has a linear equation where the predictor consists of intervals of the subordinate variable or prediction fallacies.

Y (forecasted value) = generally constant (in some cases is sum of values of y or errors)

The model is pure autoregressive ("self regressed") if it only consists of lagged values of y. This is a special case of a regression model. The ARIMA model is not a linear regression model if some of the predictors are lag of the errors because the last period error cannot be termed as independent variable as errors are calculated from time to time. While using lagged errors as predictors the problem arises that prediction is no longer the linear function of the coefficients even though they are linear function of past data. Therefore lagged errors are estimated by non-linear optimization techniques in ARIMA. Here autoregressive terms are lags of the series, moving average terms are forecast errors and time series is termed as integrated as they are versions of stationary series.

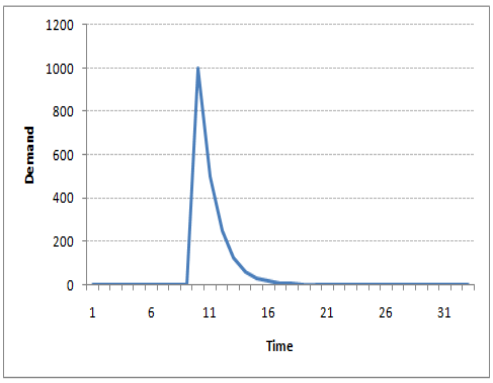

Fig 1. General graph for an auto-regressive model[14]

ARIMA (p,q,d) model-

P - number of auto-regressive terms

d - number of non-seasonal differences required for stationarity

Q - number of lagged forecast errors in forecasting equation

Now,

d=0 => yt = Y(t)

d=1 => yt = Y(t) - Y(t-1)

d=2 => yt = (Y(t) - Y(t-1)) - (Y(t-1) - Y(t-2))

d=3 => yt = (Y(t) - Y(t-1)) - (Y(t-1) - Y(t-2)) - (Y(t-2) - Y(t-3))

In term of y it can be seen as

$\hat{Y} = \mu + \phi(1)*y(t-1) + \ldots + \phi(p)*y(t-p) - \theta(1)*e(t-1) - \ldots - \theta(q)*e(t-q)$

Generally θ is negative in the equation which is moving average parameter as the convention given by Box and Jenkins. It may be plus sign also. It is generally regarded to know your software which you are using to know which sign is to be used while reading output. For getting the best suited model for Y start by standardizing the series by order of differencing and removing the gross feature of seasonality. By fitting a random trend at this point sometimes different series might be constant; however stationarized series may have some auto correlated errors implying that some AR as well as MA terms also required in forecasting equation.

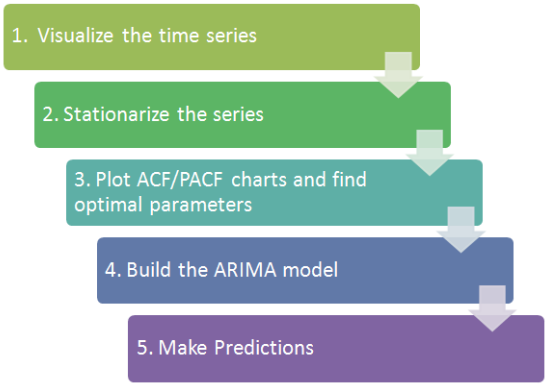

Fig 2 . Steps for building ARIMA time series model for air quality index forecasting [14].

V. RESULTS AND CONCLUSION

The disturbing problem of air pollution has caused severe changes to the earth. Hence to regulate the pollutant level, AR and ARIMA models can be expanded to make the forecast of AQI.

Auto-regression implemented on time series dataset to predict the Air Quality Index value seven days before the current date, produced the MSE to be 27.00. MSE can be minimized by lowering the difference between the present and the date on which the value of the AQI is to be predicted.

The management of AQI is swiftly becoming one of the most critical tasks. People must know what the level of pollution in their surroundings is and take appropriate action towards battling with it. The results show that auto-regression models can be conveniently and effectively employed to recognize the quality of air and forecast the level of Air Quality Index value of the near future. The suggested system will help ordinary people as well as those in the meteorological department to identify and foretell pollution levels and take the required action by that.

VI. REFERECES

[1] Kalapanidas, E.; Avouris, N. Short-term air quality prediction using a case-based classifier. Environ. Model.Softw. 2001, 16, 263–272.

[2] Athanasiadis, I.N.; Kaburlasos, V.G.; Mitkas, P.A.; Petridis, V. Applying machine learning techniques on air quality data for real-time decision support. In Proceedings of the First international NAISO Symposium on Information Technologies in Environmental Engineering (ITEE'2003), Gdansk, Poland, 24–27 June 2003.

[3] Kurt, A.; Oktay, A.B. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. Appl.2010, 37, 7986–7992.

[4] Corani, G. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 2005, 185, 513–529.

[5] Jiang,D.;Zhang,Y. Progress in developing anANN model for air pollution index forecast. Atmos. Environ. 2004, 38, 7055–7064.

[6] Ni, X.Y.; Huang, H.; Du, W.P. Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. Atmos. Environ. 2017, 150, 146–161.

[7] L. Li, X. Zhang, J. Holt, J. Tian, and R. Piltner, "Spatiotemporal interpolation methods for air pollution exposure," in Symposium on Abstraction, Reformulation, and Approximation, 2011.

[8] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '15, 2015, pp. 437– 446.

[9] M. Dong, D. Yang, Y. Kuang, D. He, S. Erdal, and D. Kenski, "PM 2.5 concentration prediction using hidden semi-markov model-based times series data mining," Expert Syst. Appl., vol. 36, no. 5, pp. 9046–9055, Jul. 2009.

[10] S. Thomas and R. B. Jacko, "Model for forecasting expressway pm2.5 concentration – application of regression and neural network models." Journal of the Air & Waste Management Association, vol. 57, no. 4, pp. 480–488, 2007.

[11] Pooja Bhalgat, Sejal Pitale and Sachin Bhoite. "Air Quality Prediction using Machine Learning Algorithm". International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656

[12]https://machinelearningmastery.com/autoregressionmodels-time-series-forecasting-python

[13]https://machinelearningmastery.com/arima-fortime-series-forecasting-with-python/

[14]https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

[15]https://www.kaggle.com/c/air-pollution-prediction