International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT2063197

A Survey on Air Quality Prediction Using Traditional Statistics Method

S. Karthikeyani, S. Rathi
Department of Computer Science and Engineering, Government College of Technology,Coimbatore, Tamilnadu, India

ABSTRACT

Air pollution is the release of pollutants into the atmospheric air which are harmful to human health and the planet as a whole. Car emissions, dust, pollen, chemicals from factories and mold spores may be suspended as a particle. In this survey, the analyzes are made revolving on air quality prediction using the traditional statistics method. The prediction using air pollutants are PM2.5, PM10, NO2, NOx, NO, SO2, CO, O3 and meteorological parameters such as Absolute Temparathure(AT) and Relative Humidity(RH). In this comparison experiments, common predicted algorithms are Naive Method, Auto-Regressive Integrated Moving Average(ARIMA), Exponentially Weighted Moving Average(EWMA), Linear Regression(LR), LSTM model, Prophet Model are analyzed.

Keywords : Air pollution, RMSE value, gases

I. INTRODUCTION

Air pollution has been determined that developing country's human health is getting majorly affected, majorly due to wherever there is no infrastructure to observe or keep its management. It's been proved that there's a correlation between atmospheric pollutants and diseases related to lungs and respiratory illness. The World Health Organization (WHO) has developed the tips to limit the bound of gases like Ozone(O3), Nitrogen dioxide(NO2) and Sulphur dioxide (SO2).

The main objective of this paper is to find out the best prediction and forecasting methods of air pollutants PM2.5, PM10, NO2, NOx, NO, NH3, SO2, CO, O3. The traditional statistical method has been implemented which include, Naive technique, Auto-Regressive Integrated Moving Average (ARIMA), Exponentially Weighted Moving Average (EWMA), Linear Regression, LSTM model, Prophet model. This paper deals with the following sections :

Section 1: It gives the introduction and related traditional statistics models.

Section 2: It deals with related works with various forecasting models.

Section 3: It explains the result analysis of the pollution in the context of the predictive statistical model in the context of various forecasting models.

II. Related Work

A. Dataset Description

The dataset collected from the Central Pollution Control Board which consists of an hourly concentration of air pollutants such as PM2.5, PM10, NO2, NOx, NO, NH3, SO2, CO, O3 and meteorological parameters such as AT and RH.

From Date and Time: 1/1/2020 and 00:00
To Date and Time: 31/1/2020 and 12:00

Station Name: SIDCO kurichi, Coimbatore- TNPCB
City name: Coimbatore

S.NO	ATTRIBUTES	DESCRIPTION
1	PM2.5	True hourly average concentration of PM2.5 in microgram/meter cube
2	PM10	True hourly average concentration of PM10 in microgram/meter cube
3	NO2	True hourly average concentration of NO2 in microgram/meter cube
4	NOx	True hourly average concentration of NOx in microgram/meter cube
5	NH3	True hourly average concentration of NH3 in microgram/meter cube
6	SO2	True hourly average concentration of SO2 in microgram/meter cube
7	CO	True hourly average concentration of CO in microgram/meter cube
8	O3	True hourly average concentration of O3 in microgram/meter cube

B. Data Preprocessing

For preprocessing the time series, we make sure that there is no None(NULL) values in the dataset; if there is, we can replace them with either 0 or average or preceding. By taking the place of values is like a choice over dropping so that the flow of the time series is provided. However, in our data the some of the last values appear to be NULL so that dropping will not affect the flow.

C. Naive Technique

A naive technique is assumed that the predicted value at a time 't' to be the actual value of the variable at a time 't-1' or rolling mean of series are used to weigh, however that statistical models and machine learning models will perform and emphasize their need. It is one among the option of time series information.

The statistics for all 733 observations across equivalent spaced timeline which is useful to understand the data. The naive method is setting the predictive value at present equal to actual value at previous time and calculate the root mean square (RMSE) for quantifying the performance of this method.

The RMSE value of PM2.5 -7.8084, PM10- 14.8405, NO2 – 11.7679, NOx – 11.9860, NO – 4.2396, NH3- 11.7260, SO2- 3.3803, CO- 0.1373, O3-28.2841.

D. Exponential Smoothing Model

The exponential smoothing or holt winter method applies three times- level smoothing lt, trend smoothing bt, and seasonal smoothing St, with α, β and γ are smoothing parameters with m. Here, we have trained the model once with the training set and then we keep on making predictions.

A lot of realistic approach is to re-train the model after one or longer steps. As we have a tendancy to

get the prediction for time 't 1' from training data 'til time 't', the next prediction for time 't 2' is created using the training data 'til time 't 1' as the actual value at 't 1' are identfied then. The methodology of making predictions for one or more future steps and then re-training the model is called rolling forecast or walk-forward validation. Although the training of statistical models is not time-consuming, walk-forward validation is the most preferred solution to get the most accurate results.

The RMSE value of PM2.5- 7.92, PM10- 15.01, NO2- 11.81, NOx- 12.04, NO- 4.30, NH3- 11.90, SO2- 3.41, CO- 0.13, O3-25.66.

E. Auto Regressive Integrated Moving Average (ARIMA)

For a stationary time series, autoregression models show the value of a variable at time 't' as a linear function of values 'p' time steps preceding it. Mathematically it can be written as –

$$y_t = C + \phi 1 y_{t-1} + \phi 2 y_{t-2} + \ldots + \phi_p y_{t-p}$$

Where,

'p' is the auto-regressive trend parameter and it is set to 5.

ϵt is white noise, and

Y_{t-1}, Y_{t-2}...Y_{t-p} indicates the value of variable at previous time periods.

The value of p can be scaled by using various methods. A way to find the apt worth of 'p' is plotted by the auto-correlation plot.

For a stationary statistic, a moving average model show the worth of a variable at time 't' as a linear function performs a residual errors from 'q' time steps preceding it. The residual error is calculated by contrast the value at the time 't' to moving average of the values preceding.

Mathematically it may be written as –

$$y_t = C + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

Where

'q' is that the moving-average trend parameter and it is set to 2.

ϵt is white noise, and

$\epsilon t-1, \epsilon t-2 \ldots \epsilon t-q \epsilon t-1, \epsilon t-2 \ldots \epsilon t-q$ measures the error terms at previous time periods

Value of 'q' can be labelled in many ways. A way of finding the apt value of 'q' is plotted by the partial auto-correlation plot.

For non-stationary statistic, we have a tendancy to set 'd' parameter as 1. Also, the value of the auto-regressive trend parameter 'p' and therefore the moving average trend parameter 'q', is calculated on the stationary statistic time series i.e by plotting ACP and PACP after differencing the statistic time series.

The RMSE value for ARIMA model is: PM2.5- 5.92,PM10-11.05, NO2- 10.54, NOx-10.96, NO-2.58, NH3-10.96, SO2- 2.55, CO-0.11, O3- 23.22.

F. Linear Regression

In statistics, linear regression is linear approach to modelling the relationship between a dependant variable and one or more independent variables. The description of linear equation that merge a selected set of input values(x), the solution to which the predicted output for the set of input values(y).

Here,

$$Y = B0 + B1 * x1$$

Where,

B0 is the bias coefficient,

B1 is the coefficient for the height column.

Before fitting to suit a linear model to observed data, a modeler ought to initial confirm whether there's a relationship between the variables. This does not necessarily imply the one variable causes the other, but there is some significant association between two

variables. A scatterplot will be a useful tool in regulating the strength of the link between two variables. If there is no association between the projected informatory and dependent variables (i.e., the scatterplot doesn't indicate any increasing or decreasing trends), then fitting a regression towards the mean model to the data won't offer a helpful model. A valuable numerical measure of association between two variables is that the parametric statistic, which is a value between -1 to 1 indicating the strength of the association for the observed data of two variables.

The RMSE value is: PM2.5- 6.55, PM10-11.43, NO2-13.26, NOx- 11.54, NO-5.08, NH3- 11.96, SO2- 3.99, CO-4.08, O3-32.55.

G. Long Short-Term Memory network

The Long Short-Term Memory network, or LSTM network, is a repetitive neural network that is trained Exploitation Backpropagation Through Time and overcomes the vanishing gradient problem. LSTMs are sensitive to the size of the input data, specifically once the sigmoid functions are used. Rather than neurons, LSTM networks have memory blocks that are connected through layers[9]. It is often an honest follow to resize the data to the range of 0 to -1, also called as normalizing.

- The LSTM network expects the input data (X) to be supplied with a selected array structure within the kind of: [samples, time steps, features].
- At present, our data is within the form of: [samples, features] and we are framing the matter collectively time step for every sample. We are able to transform the prepared train and test input data into the expected structure using numpy.reshape().
- The network encompasses a visible layer with one input, a hidden layer with four LSTM blocks or neurons, and an output layer that produces a one value prediction.
- The network is trained for a hundred epochs and a batch size of 1 is used.

The RMSE value: PM2.5- 6.77, PM10- 14.98, NO2-18.26, NOx-15.34, NO-7.08, NH3-12.96, SO2-5.99, CO-3.08, O3-24.55.

H. PROPHET METHOD

Time series model consists of three components: trend, seasonality, and holidays. They are combined in the following equation:

$$y(t) = g(t)+s(t)+h(t)+\mathcal{E}_t$$

where,

- g(t): piecewise linear growth curve for modelling non-periodic changes in time series
- s(t): periodic changes (e.g. weekly/yearly seasonality)
- h(t): effects of holidays with improper schedules
- εt: error term accounts for any uncommon changes not accommodated by the model.

After Creating an instance of a prophet object we can able to work a model of historical information, by calling the fit method on the prophet object and spending it in data frame. Since we have tendency to use daily periodicity data in this dataset, we are going to leave frequency at its default and set the periods argument to thirty one days indicating that we would like to forecast thirty one days into the future[10]. At this point, Prophet will create a new dataframe assigned to the forecast variable that contains the forecasted values for future dates, also the uncertainty intervals and components for forecast.

The RMSE value is: PM2.5- 6.45, PM10- 12.66, NO2-15.24, NOx-.12.78, NO-3.55, NH3-13.54, SO2-5.43, CO-1.23, O3-32.55.

III. CONCLUSION

From this survey, among all the traditional statistics method ARIMA model can get a good predictor

effect. None of the above model can include all these factors, but predict model can still help government and other authorities to take advanced measures to enhance the quality of air condition. The ARIMA model is better than any model in terms of trend capturing and results of RMSE. This data is better to apply the ARIMA model to predict the future AQI values.

Gas	Naive	EWMA	ARIMA	LR	PR
PM2.5	7.80	7.92	5.92	6.77	6.45
PM10	14.84	15.01	11.05	14.98	12.6
NO2	11.76	11.81	10.54	18.26	15.24
NOx	11.98	12.04	10.96	15.34	12.78
NO	4.23	4.30	2.58	7.08	3.55
NH3	11.72	11.90	10.96	12.96	13.54
SO2	3.38	3.41	2.55	5.99	5.43
CO	0.13	0.13	0.11	3.08	1.23
Ozone	28.28	25.66	23.22	24.55	32.55

IV. REFERENCES

[1]. D'Amato G. et al.,(2010). Urban Air Pollution and Climate Change as Environmental Risk Factors of Respiratory Allergy. J. InvestigAllergolClinImmunol, (20(2), 95-102)

[2]. Zhu, J.; Zhang, R.; Fu, B.; Jin, R.(2015) Comparison of ARIMA model and exponential smoothing model on 2014 air quality index in Yanqing county, Beijing, China. Appl. Comput. Math 4, 456–461. https://doi.org/10.11648/j.acm.20150406.19

[3]. Armstrong, J.S., F. Collopy, 1992. Error measures for generalizing about forecasting methods: empirical comparisons. International Journal of Forecasting 8, 69-80.

[4]. Kadilar, G.Ö.; Kadilar, C(2017) Assessing air quality in Aksaray with time series analysis. In Proceedings of the AIP Conference Proceedings, Antalya, Turkey, 18–21 April 2017; AIP Publishing: Melville, NY, USA, Volume 1833, p. 020112.

[5]. Fahrmeir L, Kneib T, Lang S(2009) Regression – Modelle, Methoden und Anwendungen. 2nd edition. Berlin, Heidelberg: Springer.

[6]. Rybarczyk, Y.; Zalakeviciute(2016) R. Machine learning approach to forecasting urban pollution. In Proceedings of the 2016 IEEE Ecuador Technical Chapters Meeting (ETCM), Guayaquil, Ecuador, 12–14 October.; IEEE: Piscataway, NJ, USA; pp. 1–6.

[7]. Giovanni Raimondo,; Alfonso Montuori,; Walter Moniaci,; Eros Pasero1,; and EsbenAlmkvist An Application of Machine Learning Methods to PM10 Level Medium-Term Prediction.

[8]. Johnston, F.R.; Boyland, J.E(1999).; Meadows, M.; Shale, E. Some properties of a simple moving average when applied to forecasting a time series. J. Oper. Res. Soc. 50, 1267–1271. https://doi.org/10.1057/palgrave.jors.2600823

[9]. Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi(2019) Antisymmetric RNN: A dynamical system view on recurrent neural networks. In International Conference on Learning Representations, Feb 2019. https://arxiv.org/abs/1902.09689

[10]. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. & Riddell, A., 'Stan(2017): A probabilistic programming language', Journal of Statistical Software 76(1).

[11]. Tsay, R.S., and Tiao, G.C. (1984), "Consistent Estimates of Auto-regressive Parameters and Extended Sample Auto-correlation Function for Stationary and Non-stationary ARMA models," Journal of American Statistical Association, 79, 84-96.

Cite this article as : S. Karthikeyani, S. Rathi, "A Survey On Air Quality Prediction Using Traditional Statistics Method", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.942-946, May-June-2020. Available at
doi : https://doi.org/10.32628/CSEIT2063197
Journal URL : http://ijsrcseit.com/CSEIT2063197