Prediction algorithms to forecast air pollution in Delhi India on a decade

Prediction algorithms to forecast air pollution in Delhi India on a decade

Muhamad Rifki Taufik[1], Eka Rosanti[2], Tofan Agung Eka Prasetya[3], Tri Wijayanti Septiarini[4]

[1]University of Darussalam Gontor
[2]Occupational Safety and Health Department, Faculty of Health, University of Darussalam Gontor
[3] Occupational Safety and Health Department, Faculty of Vocational Studies, University of Airlangga
[4]Islamic Economics, Faculty of Economics and Management, University of Darussalam Gontor

mrifkitaufik@unida.gontor.ac.id

Abstract. According to the WHO Global Ambient Air Quality Database in the past two years, there are more than 4300 cities and settlements in 108 countries where have nearly doubled, especially in Delhi, India. Preventing unwanted events is a mandatory crucial step by forecasting air pollution identifying air quality levels and recognizing the associated health impacts. Aim of this paper to forecast air pollution using four prediction models i.e. Naïve Bayesian, Auto Regressive Integrated Moving Average (ARIMA), Exponential Smoothing, and TBATS. The data were obtained from the official website of the Indian government where this research analyzed time-series data from 2005-2015 consisted of PM10, SO_2, and NO_2 with time variables day, month, and year. The time series set was managed to be monthly in ten years. Moreover, the series was split into a training set and testing set with a ratio 75:25. The training set was utilized to build prediction models and the testing set would evaluate forecasting results. Forecasting results showed all models gave acceptable prediction and according to the error, the ARIMA and exponential smoothing models were the potential prediction model for air pollution data.

1. Introduction

Air pollution is defined by the Engineers Joint Council as "the presence in the outdoor atmosphere of one or more contaminants, such as dust, fumes, gas, mist, odor, smoke, or vapor, in quantities, of characteristics, and of duration such as to be injurious to human, plant or animal life or to property, or which unreasonably interferes with the comfortable enjoyment of life and property. According to World Health Organization (WHO) Global Ambient Air Quality Database in 2018, in the past two years, there are more than 4300 cities and settlements in 108 countries where have nearly doubled, with more and more locations identifying air pollution levels and recognizing the associated health impacts. In the comparative risk assessment [1], performed as part of the Global Burden of Disease (GBD) 2010 Project, air pollution ranked as a leading contributor to the burden of disease in South Asia.

According to [2], the air pollution in India has been considered as public concern and policy attention. India's current demographic condition has a very high population of young people who can be targeted by cardiopulmonary effects of pollutant exposure which can be observed in the next few decades as a cumulative effect [3]. The rapid growth in motor vehicle activity in India and other rapidly industrializing low-income countries is contributing to high levels of urban air pollution,

among other adverse socioeconomic, environmental, health, and welfare impacts. In the comparative risk assessment [1], performed as part of the Global Burden of Disease (GBD) 2010 Project, air pollution ranked as a leading contributor to the burden of disease in South Asia. Estimates of the burden in India show approximately 1.04 million premature deaths and 31.4 million disability-adjusted life years (DALYs) to be attributable to household air pollution (HAP) resulting from solid cooking fuels, and 627,000 premature deaths and nearly 17.8 million DALYs to be attributable to ambient air pollution (AAP) in the form of fine particulate matter ≤ 2.5 μm in aerodynamic diameter (PM2.5) [4]. According to [2], the high air pollution levels occur in Delhi and other major Indian cities, because of the concentration of motor vehicular and other energy, consuming activities in these cities and the high pollution intensity of these activities. Moreover, Delhi is a landlocked city with low wind speed and dry air so it can inhibit the dispersion of pollutants into the air [5]. Air pollution is categorized as invisible killer and more still needs to be done to further reduce the levels of air pollution [6]. It is necessary to study forecasting air pollution by using statistical method as a tool to identify.

2. Methodology

2.1. Data

This study analysed data obtained from Indian government website that provided online source data for air pollution. This data measured SO_2, NO_2, and PM10 in past ten years that provided data in monthly, annually, stations, and area. Data had been collected from the field instruments in micrograms (one-millionth of a gram) per cubic meter. This study only focus on time-series monthly average data in past a decade from 2005-2015. Time-series set split into two sets, training set and testing set with 75:25 ratio where according to [7], this ratio was suitable for time series prediction. Training set developed prediction model while testing set measured accuracy of prediction.

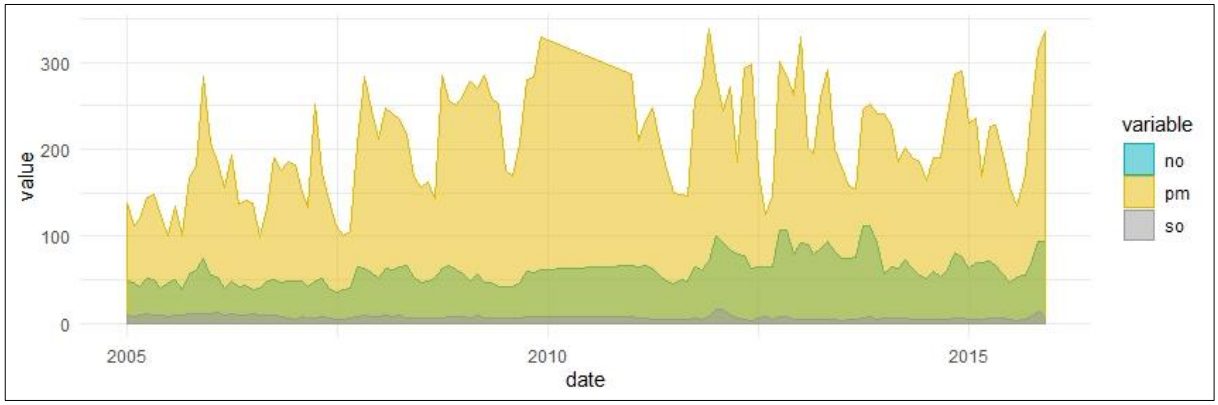


Figure 1. The time series plot for PM10, NO_2, and SO_2 since 2005-2015.

2.2. Statistical methods

2.2.1. ARIMA (Autoregressive Integrated Moving Average)

ARIMA modelling also called Box-Jenkins modelling is an approach to modelling ARIMA processes mathematical models used for forecasting. The approach uses previous time series data plus an error to forecast future values. More specifically, it combines a general autoregressive model AR(p) and general moving average model MA(q). AR(p) uses previous values of the dependent variable to make predictions and MA(q) uses the series mean and previous errors to make predictions. The approach was first proposed by [8], who detailed ARIMA's estimation and prediction procedures [9]. ARIMA models work on the assumption of stationarity which the data requires a constant variance and mean, otherwise the data need to be transformed before using ARIMA.

2.2.2. Naïve Bayesian

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem. Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, and recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real life cases, the predictors are dependent, this hinders the performance of the classifier.

2.2.3. Exponential Smoothing

According to [10], a data set consists of two components such as signal and noise. Smoothing can be described as a way to separate signal (data set) and noise as smooth as possible in order to obtain signal estimation. A simple exponential weight smoother can be used to achieve smoother separation by applying discount factor θ. It can be expressed as:

$$\tilde{y}_T = (1-\theta)y_T + \theta\tilde{y}_{T-1}, \quad (2)$$

where
$\tilde{y}_T$ is defined as the fitted value of y_T,
θ is defined as the discount factor,
y_T is defined as the observation value.
The simple exponential smoothing also can be represented in other equation by defined $\alpha = 1-\theta$,

$$\tilde{y}_T = \alpha\, y_T + (1-\alpha)\tilde{y}_{T-1}, \quad (3)$$

where α represents the weight place on the end observation and $(1-\alpha)$ represents the weight place on the smoothed value of the previous observation $(0 \leq \alpha \leq 1)$.

2.2.4. TBATS

An alternative approach developed by [11] uses a combination of Fourier terms with an exponential smoothing state space model and a Box-Cox transformation, in a completely automated manner. As with any automated modelling framework, there may be cases where it gives poor results, but it can be a useful approach in some circumstances."TBATS" is an acronym denoting its salient features:

T for trigonometric regressors to model multiple-seasonalities
B for Box-Cox transformations
A for ARMA errors
T for trend
S for seasonality

A TBATS model differs from dynamic harmonic regression in that the seasonality is allowed to change slowly over time in a TBATS model, while harmonic regression terms force the seasonal patterns to repeat periodically without changing. One drawback of TBATS models, however, is that they can be slow to estimate, especially with long time series.

3. Result and discussions

3.1. Trend and stationary data

A decade air pollutions trend was depicted on Figure 1. The widest range was PM10 where the lowest value was 15 and the highest point reached 892 ppm where within 10 years had increased year by year. This record was almost four times more than [12] reporting PM10 was 208+-137 ppm based on GIS. Moreover, NO was in range 6.3 ppm to 224 ppm and minimum value for SO_2 was 2 ppm and maximum on 49 ppm. The pollution levels were approximately two to three times those monitored in the summer months [13]. Those time-series seem had high variance where the data with lower variance was expected to be more applicable to forecast. Before developing prediction model, the data was differencing to create likely a stationary data. Stationary process is the main thing in traditional time series particularly in forecasting model such as ARIMA, moving average, exponential smoothing, etc. Stationary time series are kind of time series which have constant statistical properties such as means, variance, autocorrelation, etc [14]. Since stationary time series have constant means, variance or autocorrelation, it can predict easily.

The PM10 standard is generally used to measure air quality. The PM10 standard includes particles with a diameter of 10 ppm or less (0.0004 inches or one-seventh the width of a human hair). These small particles are likely to be responsible for adverse health effects because of their ability to reach the lower regions of the respiratory tract [6].Particulate matter has a serious impact compared to other particulates due to the high level of suspension to the atmosphere so that besides harming humans as well as plants, animals, buildings, etc [15]. PM10 indicated increase year-by-year and trend fluctuated on a decade. Figure 2 showed PM10 the time series plot before-after differencing and the Auto Correlation Function (ACF) plot before-after differencing. Based on PM10 ACF, differencing decreased lags from seven to two which come out from the horizontal line. The lower ACF implied the time-series was more stable.

Four plots in Figure 3 pictured NO_2 time series plot and ACF plot before and after differencing. Within a decade, NO_2 had big movement where increased slightly on 2012. NO_2, on the other hand, can in particular make children susceptible torespiratory diseases especially in winter [16].The primary cause of the increase in NO2 concentration is meteorological and photochemical conditions. In calm condition, the pollutant concentration is high in the air because the pollutant dispersion is slower [17]. The variance for NO_2 was large enough where the ACF plot for NO_2 before differencing had fourteen (14) lags which come out from horizontal line. Then, after differencing ACF remained two lags which come out from horizontal line. Effort to create stationary data pointed out NO_2 trend more stable than before differencing.

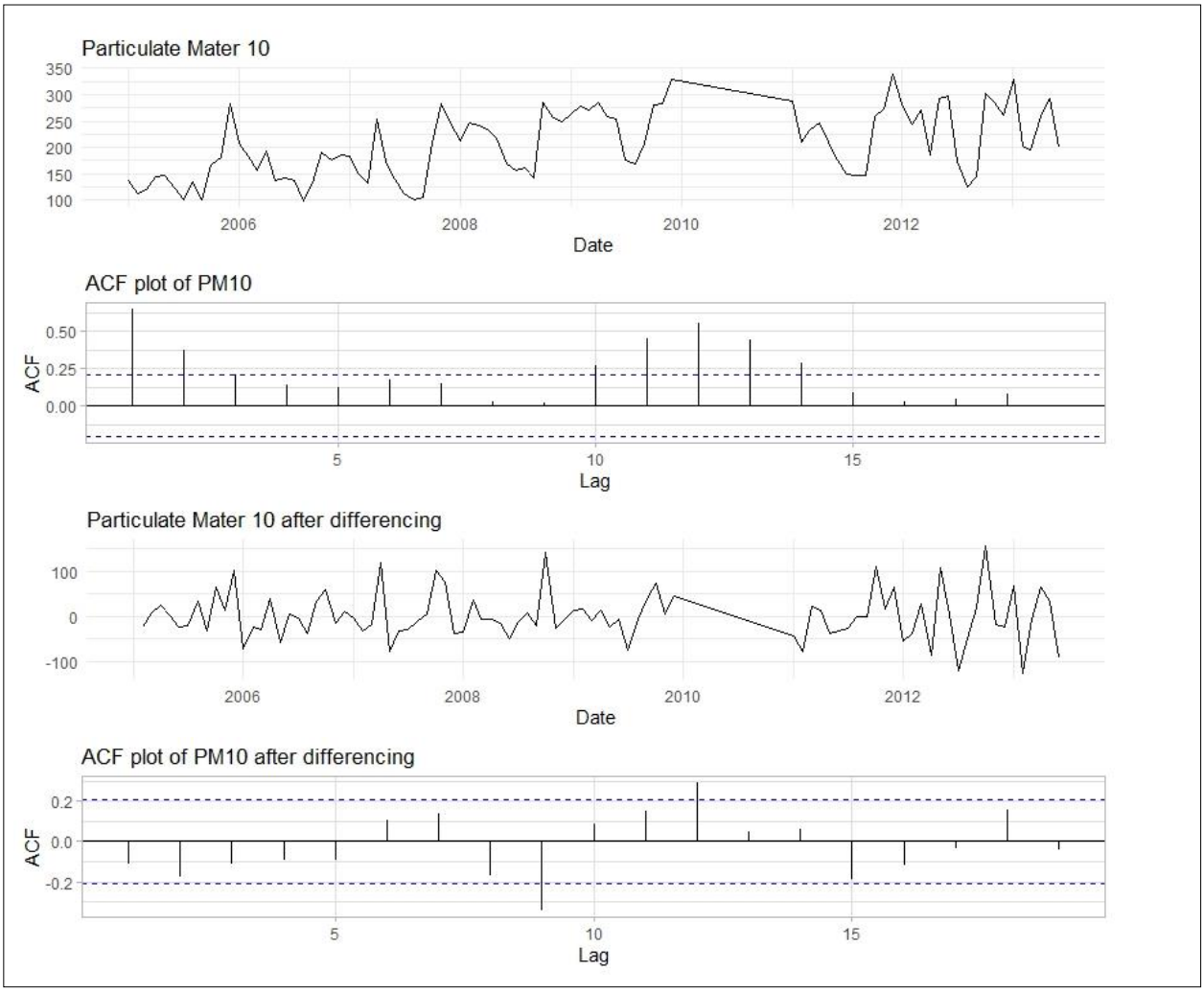


Figure 2. The time series plot for PM10 and ACF before and after differencing.

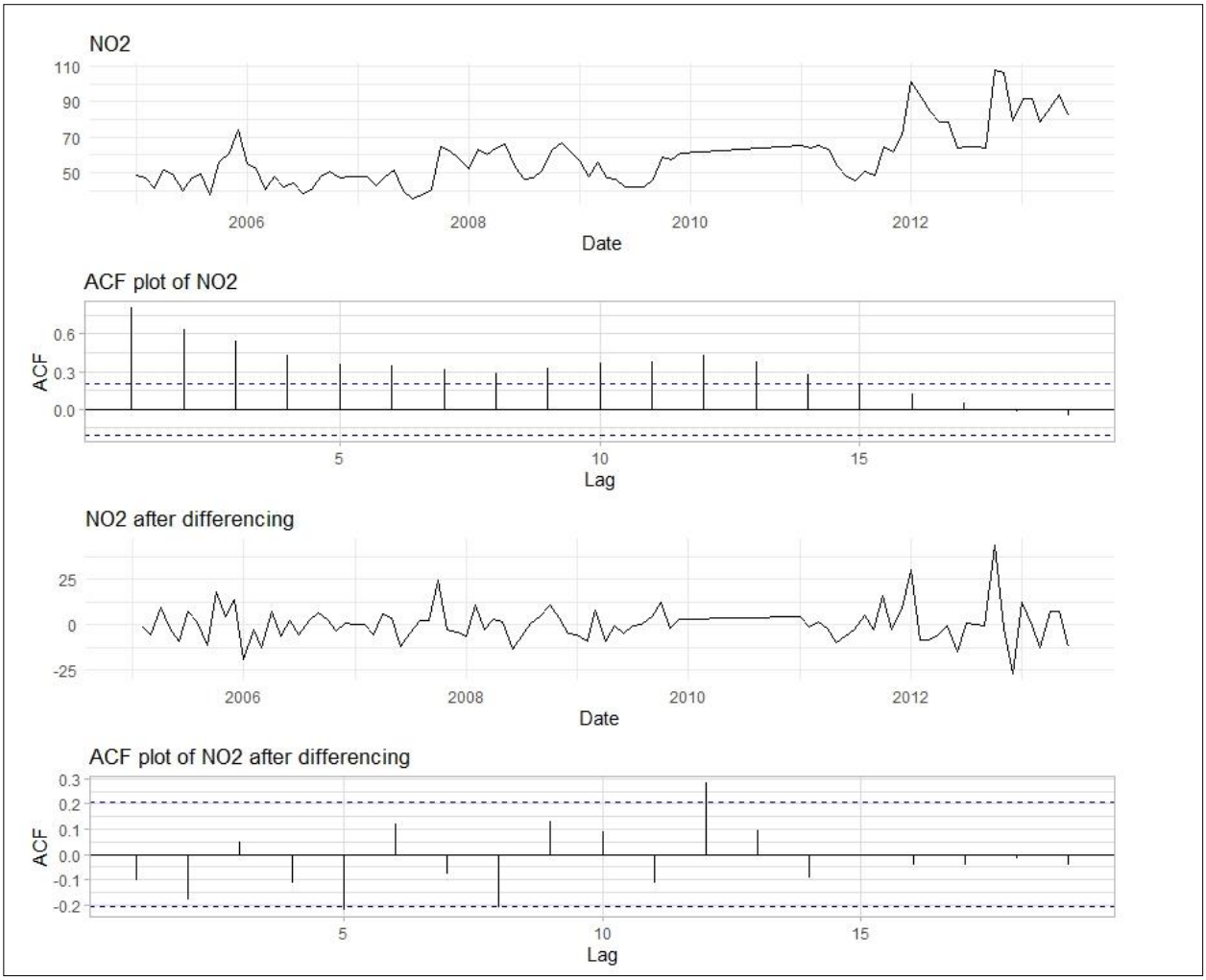


Figure 3. The time series plot for $NO_2$and ACF before-after differencing.

Furthermore, Figure 4 marked sulfur dioxide or SO_2 decreasing from 2005 until increasing on 2012. Oxides of sulfur can oxidize and form sulfuric acid, thereby leading to the damage of lungs and various lung disorders such as wheezing and shortness of breath [16]. The trend was stable enough which had only two lags, however the data would be differencing instead and created no lags on ACF plot. This result showed in Figure 4. The differencing result showed that there is no lag which come out from the horizontal line. It means that the differencing process is adequate successful.

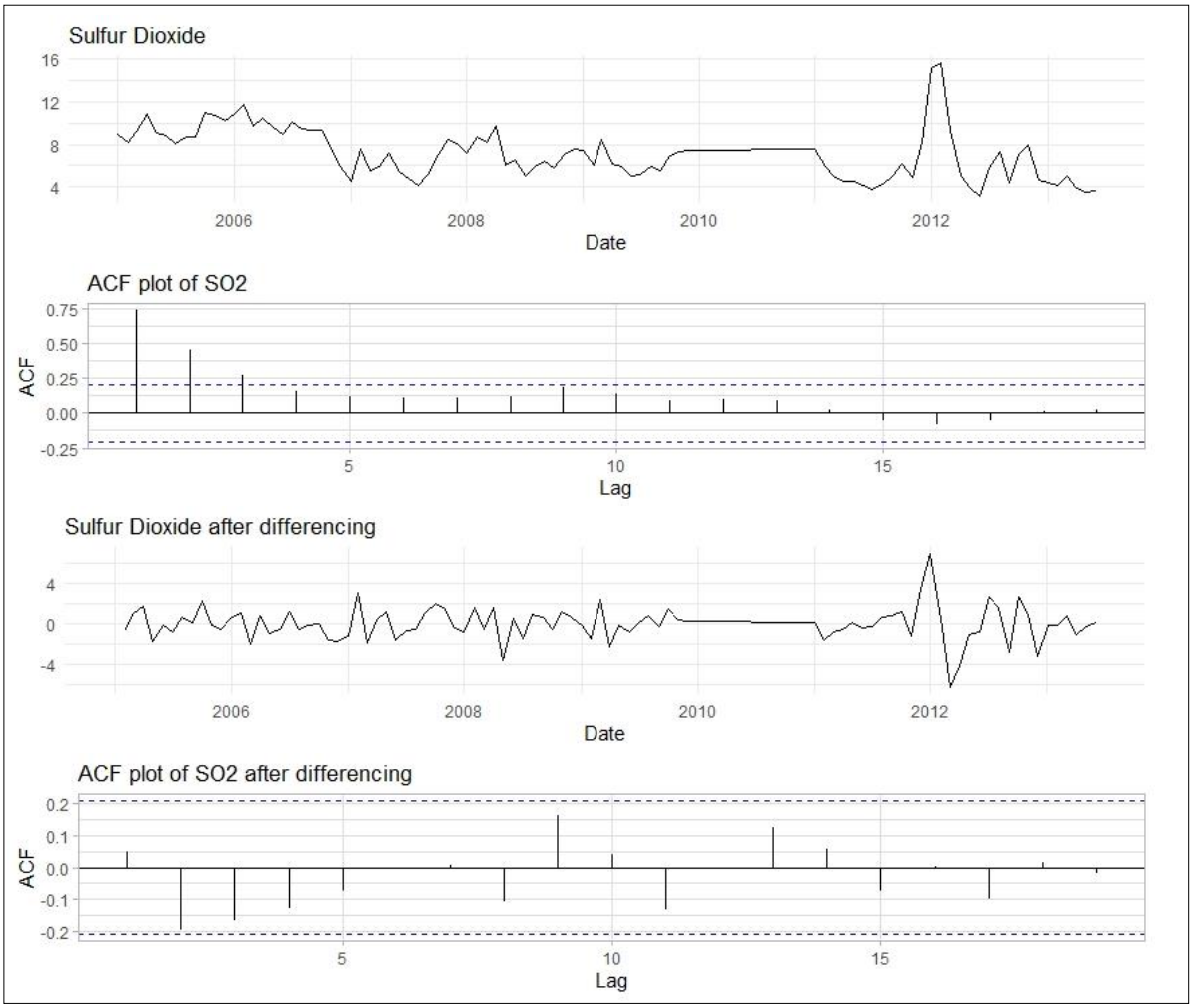


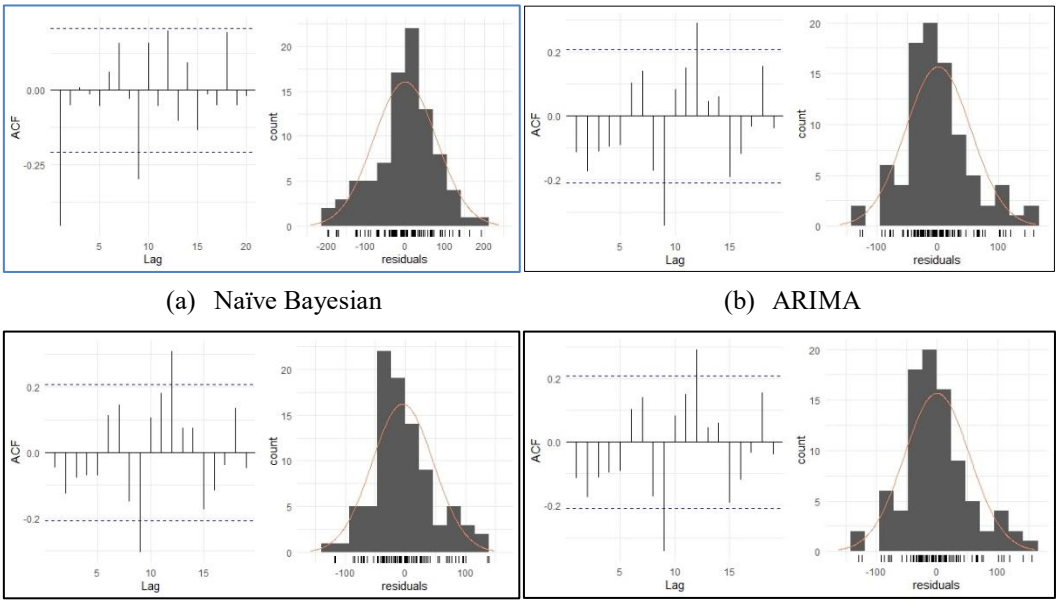
Figure 4. The time series plot for SO_2 and ACF before-after differencing.

3.2. Analytics result

3.2.1. Residual

All air pollution gases, PM10, NO_2 and SO_2, were applied into four prediction models which are ARIMA, Naïve Bayesian, exponential smoothing, and TBATS. Those models had been developed using training set where pattern and trend of stationary data indicated prediction value. Goodness of fit models each gas were measured and depicted in following graphs. Each model will be analyse the residuals in order to know the fitted performance of each data set.

Figure 5 showed residual based on four prediction models to forecast PM10. There were two lines exceeded the blue line of ACF plots on all of models. Since successive residuals do not seem to be correlated, and the forecast errors seem to be normally distributed with mean zero and constant variance, the constructed four prediction models provide a predictive model. All residual indicated the models gave acceptable result where the residuals gave normally distributed.

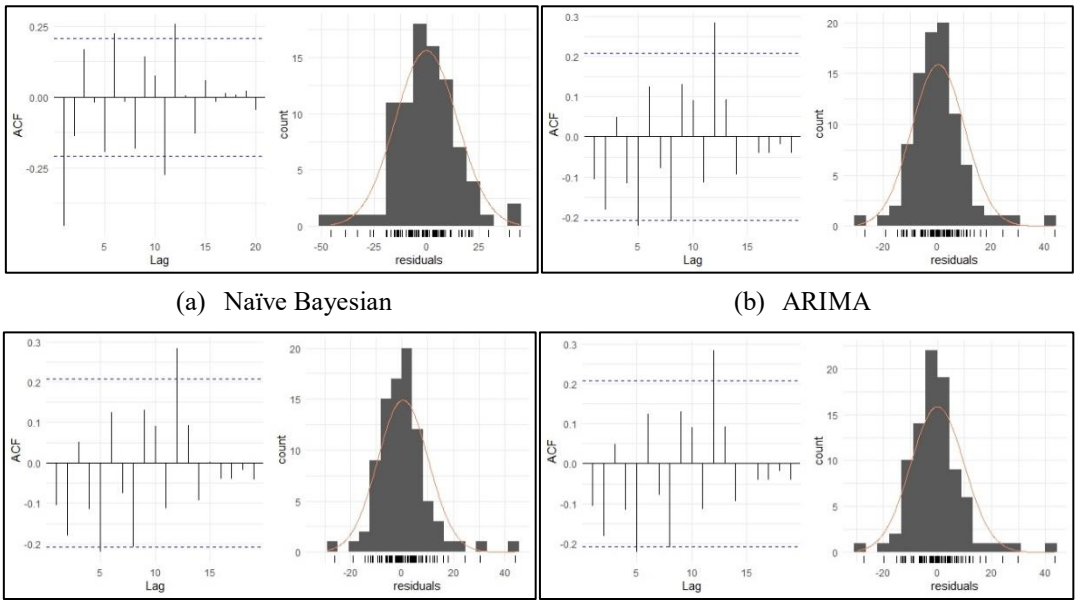


(a) Naïve Bayesian (b) ARIMA

(c) TBATS (d) ETS

Figure 5. The ACF plot and histogram of models residual for PM10.

Fitted value and predicted value of nitrogen dioxide measured the accuracy and indicated model goodness of fit. Figure 6 depicted residual result in predicting NO_2 within ten years. The residual checking consisting of the ACF plot and histogram for NO_2. Since successive residuals do not seem to be correlated, and the forecast errors seem to be normally distributed with mean zero and constant variance, the constructed four prediction models provide a predictive model. All residual indicated the models gave acceptable result where the residuals gave normally distributed.



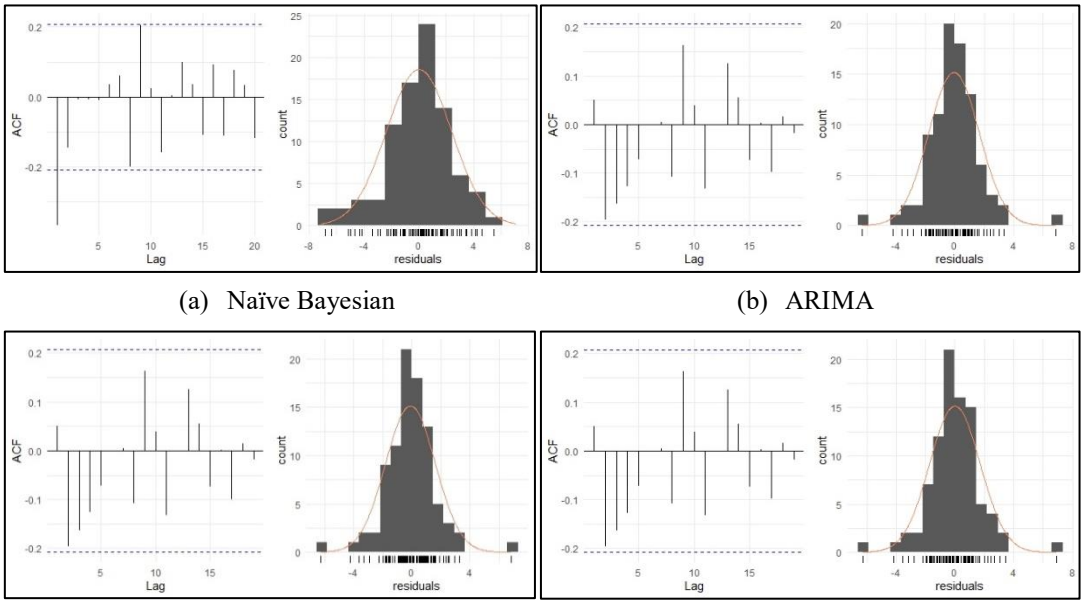
(a) Naïve Bayesian (b) ARIMA

(c) TBATS (d) ETS

Figure 6. The ACF plot and histogram of models residual for NO_2.

Previous models in PM10 and NO_2 had about two lags in residual result. However, SO_2 seem well fitted to prediction model since there was no lags on ACF plot except on Naïve Bayesian model. Since

successive forecast errors do not seem to be correlated, and the forecast errors seem to be normally distributed with mean zero and constant variance, the constructed four prediction models provide a predictive model.



(a) Naïve Bayesian (b) ARIMA

(c) TBATS (d) ETS

Figure 7. The ACF plot and histogram of models residual for SO_2.

3.2.2. RMSE

The model accuracy based on their prediction demonstrate on Table 1. The root mean square error (RMSE) is being as a tool to evaluate the performance of all models. According to the evaluation value on PM10, ARIMA and exponential smoothing have the same value in training error of RMSE. Yet, exponential smoothing has the smallest in testing error. Since, the testing error is smaller than training error, it means that exponential smoothing is out perform to predict PM10.

ARIMA prediction was the most accurate than other but not slightly different with TBATS and exponential smoothing. The testing error is smaller than the training error indicating that ARIMA is adequate predictive model to predict NO_2. A study also claimed ARIMA was out performed than exponential smoothing [7]. Lately, exponential smoothing prediction accuracy has the smallest error in training dataset which is 1.70. But, the testing error is higher than in training set which is 2.17.

Table 1. The models evaluation.

	RMSE		
	PM10	NO_2	SO_2
Naive bayes			
Training	79.26	14.65	2.36
Testing	103.08	17.79	**2.16**
ARIMA			
Training	**53.27**	**9.84**	1.71
Testing	38.26	13.13	2.17
TBATS			
Training	58.75	9.88	1.73
Testing	38.48	13.15	2.17
ETS			
Training	**53.27**	9.88	**1.70**
Testing	**38.18**	**13.12**	2.17

4. Conclusion

Purpose of this study is forecasting ambient of air pollutions was to anticipate unwanted event in near future. Four models attempted to predict three air pollution substances. This study concluded that all models gave acceptable performance where ARIMA and exponential smoothing were potentially predictive models among others. However, each dataset has the unique characteristic so the forecasting model cannot guarantee to perform potentially for all dataset.

References

[1] Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Adair-Rohani, H., and Andrews, K.G 2012*A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010*(The Lancet)vol 380(9859) p 2224–2260

[2] Lim S S, Vos T, Flaxman A D, Danaei G, Shibuya K, Adair-Rohani H and Andrews K G 2012 A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010 *The Lancet* vol 380(9859) pp 2224–2260

[3] Badami, M.G 2005 *Transport and urban air pollution in India*(Environmental Management)vol 36(2) p 195-204

[4] Badami M G 2005 Transport and urban air pollution in India *Environmental Management* vol 36(2) pp 195-204

[5] Gordon T, Balakrishnan K, Dey S, Rajagopalan S, Thornburg J, Thurston G, Agrawal A, Collman G, Guleria R, Limaye S, Salvi S, Kilaru V, Nadadur S 2018 *Air pollution health research priorities for India: Perspectives of the Indo-U.S. Communities of Researchers*(Environment International 119): 100-108.

[6] Gordon T, Balakrishnan K, Dey S, Rajagopalan S, Thornburg J, Thurston G, Agrawal A, Collman G, Guleria R, Limaye S, Salvi S, Kilaru V and Nadadur S 2018 Air pollution health research priorities for India: Perspectives of the Indo-U.S. Communities of Researchers *Environment International* 119 pp 100-108

[7] Balakhrisnan, K., Cohen, A. and Smith, K.R 2014*Addressing the burden of disease attributable to air pollution in India: the need to integrate across household and ambient air pollution exposures*(Environmental Health Perspectives)vol 122(1) p A6-A7

[8] Balakhrisnan K, Cohen A and Smith K R 2014 Addressing the burden of disease attributable to air pollution in India: the need to integrate across household and ambient air pollution exposures *Environmental Health Perspectives* vol 122(1) pp A6-A7

[9] Banarjee BD, Sharma KA, Ghozh C, Bayan P 2018 *Health Effects of Air Pollution among Residents of Delhi: A SystematicReview*(International Journal of Health Sciences & Research) vol 8 no.(1):273-282.
[10] Banarjee B D, Sharma K A, Ghozh C, Bayan P 2018 Health Effects of Air Pollution among Residents of Delhi: A Systematic Review *International Journal of Health Sciences & Research* vol 8(1) pp 273-282
[11] Rizwan SA, Nongkynrih B, and Gupta SK 2013 *Air pollution in Delhi: its magnitude and effects on health*(Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine) vol 38(1) p 4.
[12] Rizwan S A, Nongkynrih B and Gupta S K 2013 Air pollution in Delhi: its magnitude and effects on health *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine* vol 38(1) p 4
[13] Septiarini TW and Musikasuwan S 2018*Investigating the performance of ANFIS model to predict the hourly temperature in Pattani, Thailand* (Journal of Physics: Conference Series IOP Publishing)Vol. 1097 No. 1, p. 012085
[14] Septiarini TW and Musikasuwan S 2018 Investigating the performance of ANFIS model to predict the hourly temperature in Pattani, Thailand *Journal of Physics: Conference Series IOP Publishing* Vol. 1097(1) p. 012085
[15] Box G 1989*An Unexpected Route to Time-Series-A Citation Classic Commentary on Time-Series Analysis-Forecasting and Control By Box, Gep, And Jenkins, Gm* (Current Contents/Physical Chemical & Earth Sciences)vol 30 no.(44):22-
[16] Box G 1989 An Unexpected Route to Time-Series-A Citation Classic Commentary on Time-Series Analysis-Forecasting and Control By Box, Gep, And Jenkins, Gm *Current Contents/Physical Chemical & Earth Sciences* vol 30 (44):22-
[17] Hyndman R, Koehler AB, Ord JK, and Snyder RD 2008*Forecasting with exponential smoothing: the state space approach*(Springer Science & Business Media)vol 19
[18] Hyndman R, Koehler A B, Ord J K and Snyder R D 2008 Forecasting with exponential smoothing: the state space approach *Springer Science & Business* Media vol 19
[19] Montgomery DC, Jennings CL, and Kulahci M 2015*Introduction to time series analysis and forecasting*(John Wiley & Sons) vol 21
[20] Montgomery D C, Jennings C L and Kulahci M 2015 Introduction to time series analysis and forecasting *John Wiley & Sons* vol 21
[21] De Livera AM, Hyndman RJ, and Snyder RD 2011*Forecasting time series with complex seasonal patterns using exponential smoothing*(Journal of the American Statistical Association)vol 106(496) no.1513-27
[22] De Livera A M, Hyndman R J and Snyder R D 2011 Forecasting time series with complex seasonal patterns using exponential smoothing *Journal of the American Statistical Association* vol 106(496) p 1513-27
[23] Guttikunda SK and Calori G 2013*A GIS based emissions inventory at 1 km× 1 km spatial resolution for air pollution analysis in Delhi, India*(Atmospheric Environment) vol 67 p 101-11
[24] Guttikunda S K and Calori G 2013 A GIS based emissions inventory at 1 km× 1 km spatial resolution for air pollution analysis in Delhi, India *Atmospheric Environment* vol 67 p 101-11
[25] Guttikunda SK and Gurjar BR 2012*Role of meteorology in seasonality of air pollution in megacity Delhi, India*(Environmental monitoring and assessment) vol 184 no 5 p 3199-211
[26] Guttikunda S K and Gurjar B R 2012 Role of meteorology in seasonality of air pollution in megacity Delhi, India *Environmental monitoring and assessment* vol 184(5) p 3199-211
[27] Zhang J and Smith KR 2003*Indoor air pollution: a global health concern*(British medical bulletin) vol 68 no.1:209-25.
[28] Zhang J and Smith K R 2003 Indoor air pollution: a global health concern *British medical bulletin* vol 68(1) p 209-25

[29] Pandey JS, Kumar R, and Devotta S 2005 *Health risks of NO2, SPM and SO2 in Delhi (India)*(Atmospheric Environment) vol 39 no.(36):6868-74.
[30] Pandey J S, Kumar R and Devotta S 2005 Health risks of NO2, SPM and SO2 in Delhi (India) *Atmospheric Environment* vol 39(36) p 6868-74
[31] Patel J and Soni HB 2017*Assessment Of Ambient Air Quality And Air Quality Index In Golden Corridor Of Gujarat, India: A Case Study Of Dahej Port*(International Journal of Environment) vol 6 no.(4):28-41.
[32] Patel J and Soni H B 2017Assessment Of Ambient Air Quality And Air Quality Index In Golden Corridor Of Gujarat, India: A Case Study Of Dahej Port *International Journal of Environment* vol 6(4) pp28-41
[33] Bassin, Mamta P 2010 *Analysis Of Ambient Air Quality Using Air Quality Index – A Case Study*(International Journal of Advanced Engineering Technology) vol 1 no.(2):106-114.
[34] Bassin, Mamta P 2010 Analysis Of Ambient Air Quality Using Air Quality Index – A Case Study *International Journal of Advanced Engineering Technology* vol 1(2) pp 106-114