Time series air quality forecasting with R Language and R Studio

I Setiawan[1]

[1] Department of Computer and Informatics Engineering, Politeknik Negeri Bandung, Jalan Gegerkalong Hilir, Bandung, Indonesia

E-mail: irwan@jtk.polban.ac.id

Abstract. The purpose of this study is to demonstrate how to make air quality forecasting to predict the Nitrogen Dioxide quality index in the future. In this paper, we demonstrate exploratory data analysis and compare the performance of the Autoregressive Integrated Moving Average and Exponential Smoothing Model. We used R Language and R Studio to integrate all the datasets, exploratory data analysis, data preparation, performing Autoregressive Integrated Moving Average and Exponential Smoothing methods, model evaluation, and visualization. This study used data from the automatic remote air quality-monitoring station located in an urban area in Madrid, Spain. The dataset in the period from 1 January 2001 to 31 December 2017. The dataset recorded six pollutants such as Nitrogen Dioxide, Particulate Matter 10 micrometres, Sulphur Dioxide, Carbon Monoxide, Ozone and Particulate Matter 2.5 micrometres. In this study, we focus only on Nitrogen Dioxide pollutants. From our model, we saw that exponential smoothing has better accuracy compared to the Autoregressive Integrated Moving Average. We also exposed that Nitrogen Dioxide pollutant shows unhealthy for sensitive group's level in November to March and has the lowest level in June and July.

1. Introduction

Air pollution is a remarkable problematic in big cities, where healthiness concerns and transportation constraints are constantly growing. Some pollutants cause immense disturbance to the environment. To defend human healthiness and the atmosphere, the World Health Organization (WHO) has distributed recommendations. To keep populations from harmful air, many nations have air quality forecasting programs to estimate the concentrations of pollutants such as O_3, NO_2, $PM_{2.5}$, and PM_{10} [1].

Some information is recycled to deliver early air quality warnings that allow decision-makers and people to take preventive measures such as taking public transportation or temporarily stopping primary emission sources to reduce air pollution and limit their exposure to an unhealthy level of air pollution. Accurate air quality forecasting can offer great social and financial benefits by facilitating advanced planning for individuals, families, or organizations to reduce pollutant emissions.

Driven by a major improvement and the unique challenges of air quality estimates in the past two decades, this study aims to show exploratory data analysis and compare the performance of the Autoregressive Integrated Moving Average and Exponential Smoothing model.

2. Literature survey

Nieto, Lasheras, Gonzalo, and Juez research to assess the utility of VARMA, ARIMA, MLP, and SVM in forecasting future PM_{10} concentrations. They used seven years of air quality monitoring data taken from air quality monitoring stations, which located in the metropolitan area of Avilés. The mean concentration of pollutants (SO_2, NO, and NO_2) and PM_{10} are used to forecast the average concentration of PM_{10} on monthly basis. The result of their research is that the ARIMA model performs better when forecasting one month, while SVM gives the best performance to forecast from one to nine months ahead [2]. They also doing similar research for Oviedo, northern Spain. The SVM model performs better when forecasting seven months ahead [3].

Zhang *et al*. use the ARIMA model to analyze the trend and forecasting of $PM_{2.5}$ in Fuzhao, China. Two years of time series data of meteorological parameters and pollutant concentrations assessed. Their model shows that $PM_{2.5}$ concentrations experienced seasonal fluctuations higher in cold periods and lower in two warm periods [4]. Cadenas, Rivera, Amezcua, and Heard developed ARIMA and NARX model to predict wind speed in La Mata, Oaxaca, Mexico. Their simulation shows that the multivariate NARX model gives more accurate results compared to the univariate ARIMA model [5]. ARIMA model also used by Shukur and Lee to forecast daily wind speed through hybrid KF-ANN [6].

Kadilar and Kadilar assess air quality in Aksaray, Turkey, using seasonal ARIMA. Their focus is on the SO_2 parameter. They conclude that the SARIMA model provides reliable and satisfactory predictions for air quality assessment and justification [7]. Katimon, Shahid and Mohsenipour using ARIMA to model water quality and hydrological variables. Their study found that their model gives 95 percent confidence bound which indicates the suitability of the ARIMA model in forecast water quality and hydrological variables[8]. Zhu *et al*. proposed two hybrid models to forecast AQI data. One year of data used and are collected from Xingtai, China. They saw that their hybrid models give higher forecasting precision value compared to ARIMA, SVR, EMD-GRNN, Wavelet-GRNN, and Wavelet-SVR [9]. Sharma, Mitra, Sharma, and Roy proposed to use recurrent neural network and long-short-term-memory to estimate AQI. Their simulation result shows that the proposed method outperforms state-of-the-art technique AQI estimation in terms of root mean square error and Min/Max aggregation of AQI values [10]. Liu, Lau, Sandbrink, and Fung mention a mixed forecast strategy to ARIMAX for normal values of $PM_{2.5}$ and O_3 and numerical models for outputs above 75 percent of historical observations [11].

3. Methodology

In this study, we use R Language and R Studio to integrate all the data, exploratory data analysis, data preparation, model evaluation, and visualization. The forecasting step is shown in Figure 1.

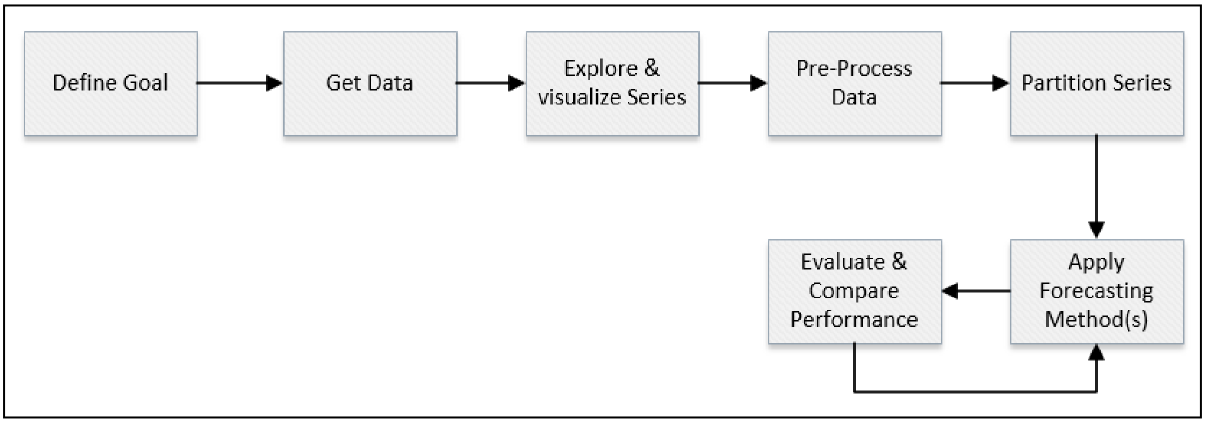
Define Goal
Get Data
Explore & visualize Series
Pre-Process Data
Partition Series
Apply Forecasting Method(s)
Evaluate & Compare Performance

Figure 1. Forecasting step.

4. Results and discussion

The dataset taken from Madrid City Council is about Madrid's air quality datasets. The dataset recorded air quality data from 24 air quality control stations between the years 2001 and 2017 (17 years). Each observation based on an hourly basis and measured in µg/m3 (micrograms per cubic meter of air). These automatic remote stations uncover three types of locations. First is Urban location, representative of the exposure of the urban population in general, second is Traffic, located in such a way that its pollution level is mainly influenced by emissions from a street or road, and last is Suburban, located on the borders of the city, in the area where the highest levels of ozone are found. In this study, we only focus to study air quality data from one of the urban location air stations. The name of the station is Farolillo, which located in Calle Faralillo.

In this study, we found many missing values in each year. Missing values in a time series create "holes" in the series. ARIMA models and smoothing method cannot be directly applied to time series with missing values, because the relationship between consecutive periods is modeled directly [12]. To handle this, we used the mean value of each year to replace the missing values.

Table 1 shows the minimum, mean, 1st quartile, median, 3rd quartile, maximum and standard deviation of four pollutants. As we can see, there is a huge difference between the median and maximum values. This indicates that many outliers (extreme values) may present in the data. To deal with this, we used maximum value instead of the mean value. The reason is there can be peak hours in a day where the values of pollutants can go higher and drop down suddenly. If we used the average value, it could be not relevant.

Table 1. Summary of the data set.

	O_3	PM_{10}	NO_2	SO_2
Minimum	0.00000	0.00000	0.00000	0.000000
Mean	43.52594	26.21618	43.77296	7.942037
1st Quartile	11.00000	11.00000	19.42000	3.930000
Median	40.30000	20.00000	36.00000	6.500000
3rd Quartile	67.00000	33.00000	60.26000	10.460000
Maximum	210.00000	402.00000	407.20001	144.600006
Stdev	33.89758	23.38312	31.18935	6.356575

Daily maximum and monthly average emission of NO_2 illustrates in Figure 2. The first line chart shows changes in the number of daily maximum emission of NO_2, and the second chart shows the monthly average emissions of NO_2 between 2001 and 2017. There was a fall in the number of monthly average emissions. Most of the monthly average emission was below 150, but there was a significant number of monthly average emissions of NO_2 in 2002, reaching almost 300.

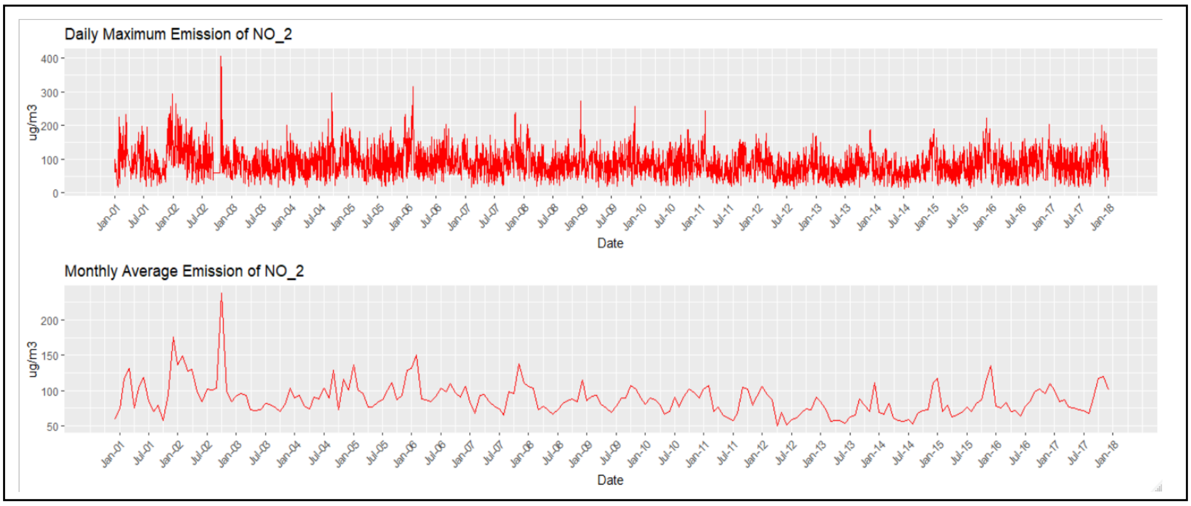

Figure 2. Daily maximum and monthy average emission of NO_2.

The value of NO_2, as seen in Figure 3, it reaches the maximum in February and October with values 320 and 410 respectively and has the lowest in March. The daily maximum value of NO_2 stood at 300 at the beginning of the day (Figure 4). Over the subsequent hours, the maximum value decreased and reached the lowest at 4 A.M., followed by a period of volatility. During the night, the maximum value increased dramatically and reach it's maximum at 11 P.M. The trend was increased.

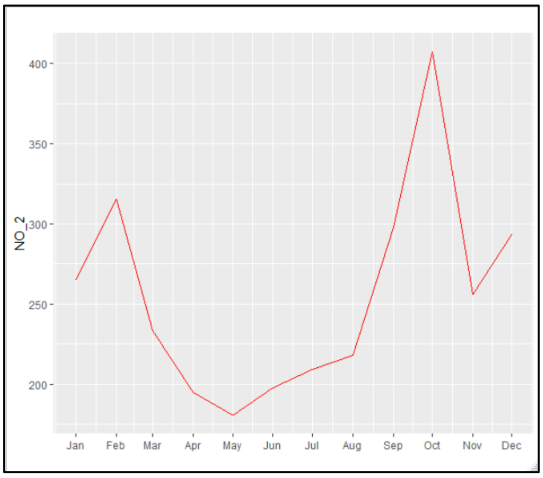

Figure 3. Maximum value for each month.

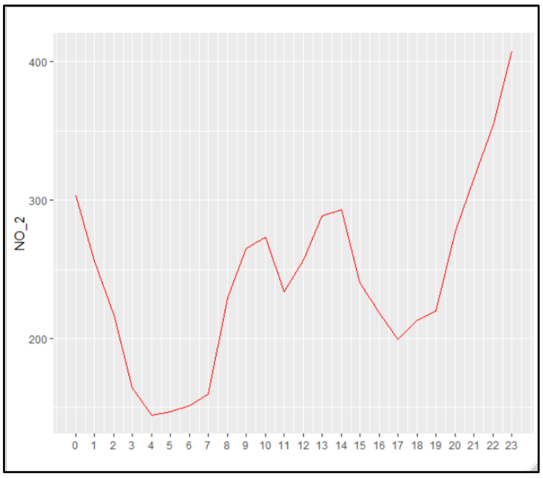

Figure 4. Maximum value for each hour.

Figure 5 shows an additive decomposition of Farolillo NO_2 pollutant data. The two components shown separately in the two panels of the Figure can be added together to reconstruct the data shown in the top panel. The seasonal component was stable over time. The trend-cycle has captured the sudden fall in the data in early 2004.

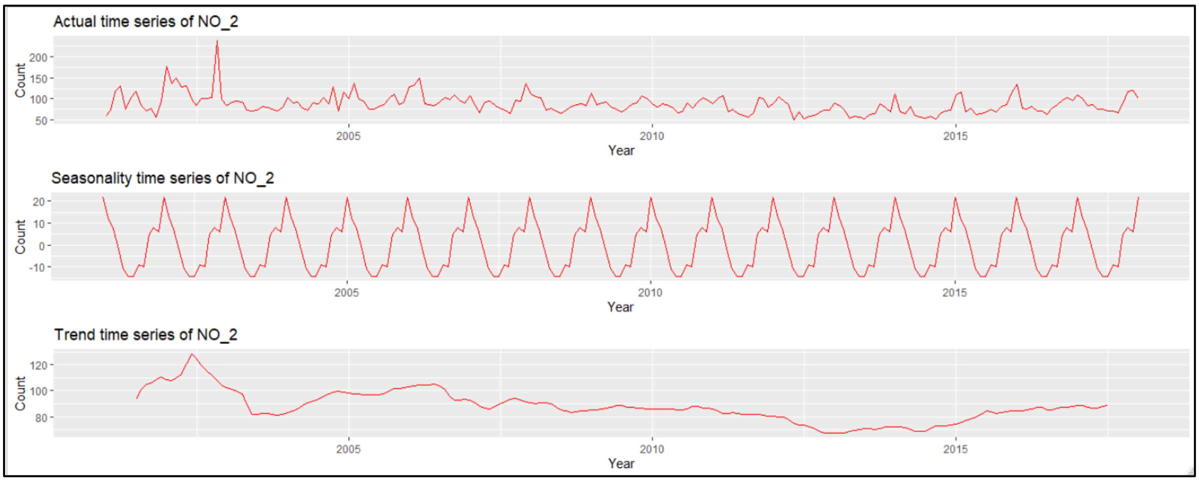

Figure 5. Decomposition of NO_2 time series.

Figure 6 shows the observed values compared to the rolling forecast prediction using ARIMA. As we can see, the forecasts did not align with the true values very well. Figure 7 shows the forecasting of NO_2 pollutants for the year 2018 and 2019 using ARIMA. The blue line shows the forecasted values and the confidence intervals (grey area) shows that it can go anywhere.

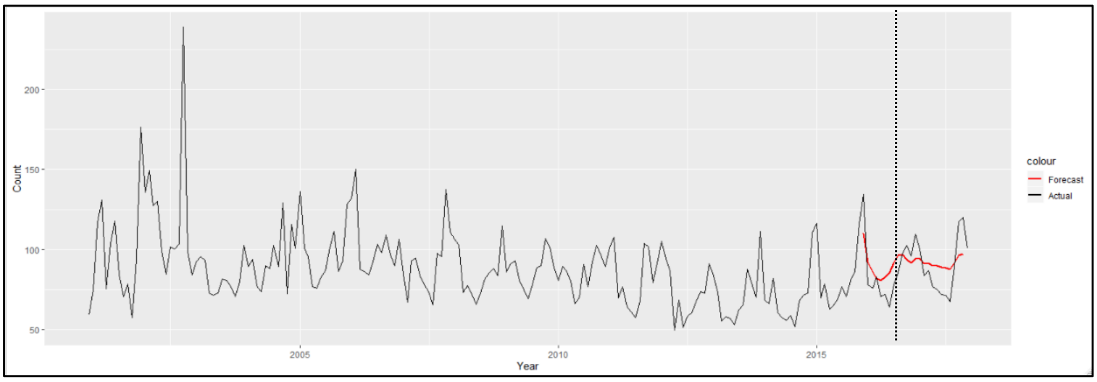

Figure 6. ARIMA training vs testing Plot.

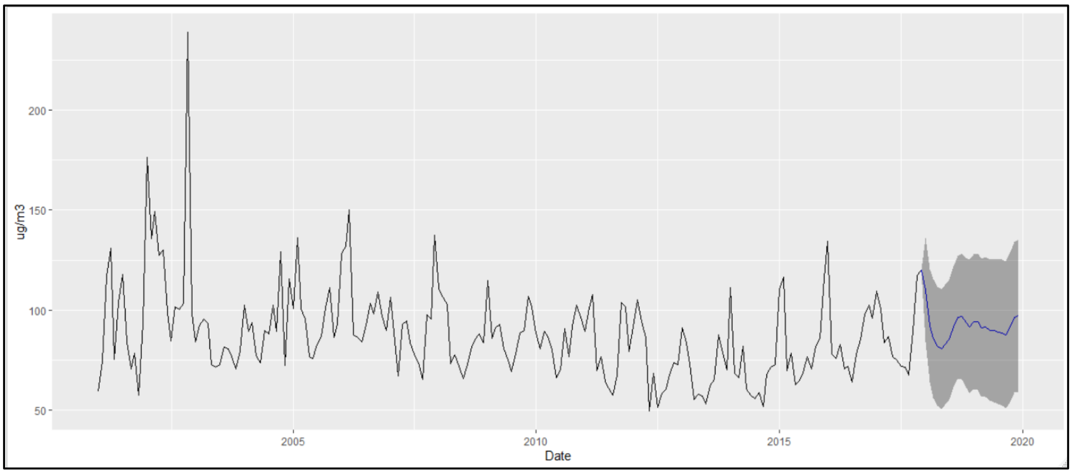

Figure 7. ARIMA forecast of NO_2 for 2018 – 2019.

Figure 8 shows the observed values compared to the rolling forecast prediction using ETS. As we can see, the forecasts align with true values very well. Figure 9 shows the forecasting of NO_2 pollutants for the year 2018 and 2019 using ETS. The blue line shows the forecasted values.

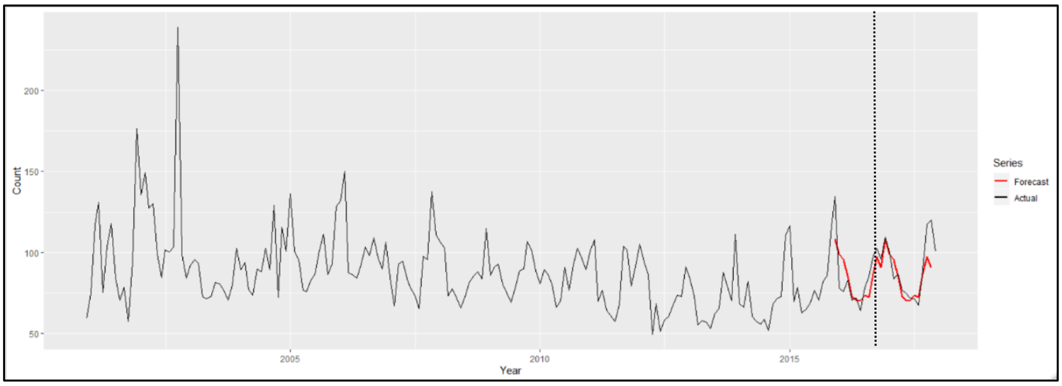

Figure 8. ETS training vs testing plot.

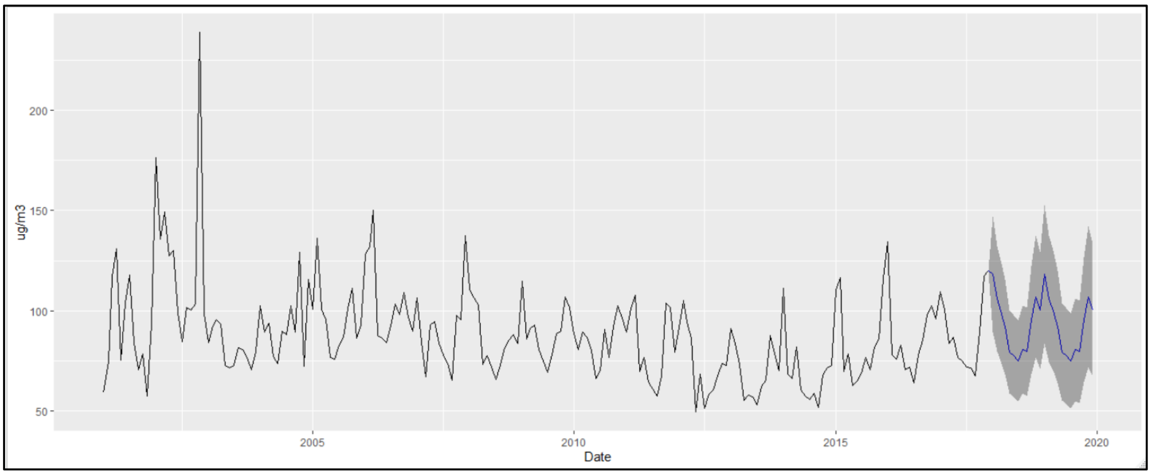

Figure 9. ETS forecast of NO_2 for 2018-2019.

We could see in Figure 10, there are expected hikes in January, February, March, November, and December, and five times in a year, the pollutants reach an unhealthy level. People sensitive to pollutants have the most chance of getting affected between November to March. From our model, we observed that ETS stayed strong with accuracy and it was able to overcome the ARIMA model.

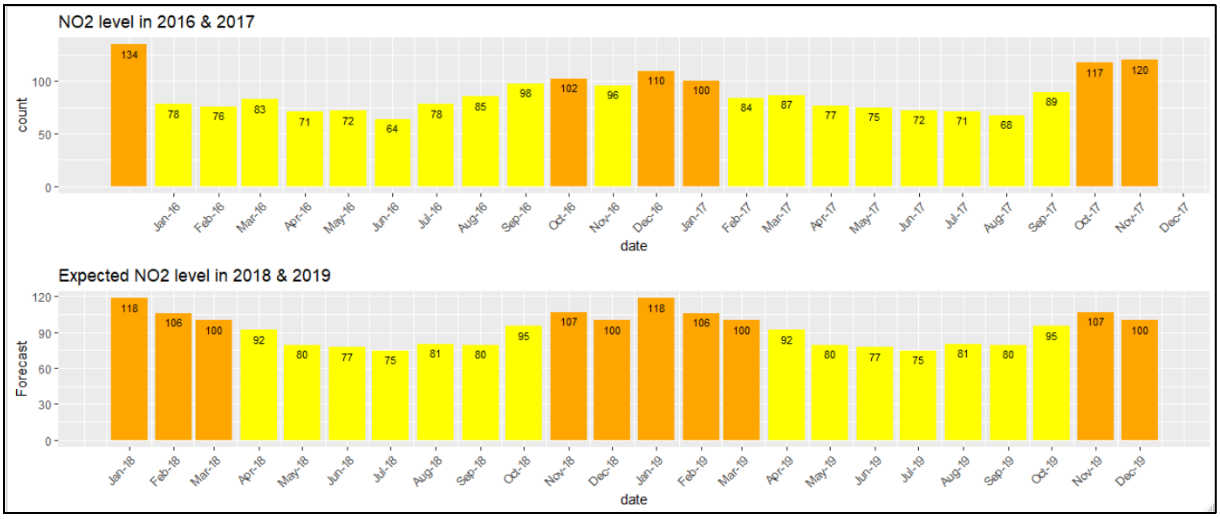

Figure 10. Expected NO_2 pollutant level in 2018 and 2019.

5. Conclusions

Autoregressive Integrated Moving Average and Exponential Smoothing model was used to demonstrate forecasting air quality using R Language and R Studio. Seventeen years recorded air quality data used. The results showed that the Exponential Smoothing model more effective than Autoregressive Integrated Moving Average. Air pollution forecasting is important to prevent contamination or maximally reduce the danger of pollution incidents, and it plays an important role in a warning and controlling air pollution. The pollution index series is non-stationary and chaotic, making it difficult to achieve an accurate estimate for the air pollution index.

6. References

[1] Zhang Y, Bocquet B, Mallet V, Seigneur C and Baklanov A 2012 *Atmospheric Environment* **60** 632-655

[2] Nieto P G, Lasheras F S, Garcia-Gonzalo E and Juez F D C 2018 *Stochastic Environmental Research and Risk Assessment* **32** 3287-3298

[3] Xinyu H and Rongrong L 2019 *Energies* **12**

[4] Zhang L, Lin J, Qiu R, Hu X, Zhang H, Chen Q, Tan H, Lin D and Wang J 2018 *Ecological Indicators* **95** 702-710

[5] Cadenas E, Rivera W, Amezcua R C and Heard C 2016 *Energies* **9**

[6] Shukur O B and Lee M H 2015 *Renewable Energy* **76** 637-647

[7] Kadilar G O and Kadilar C 2017 *AIP Conference Proceedings*

[8] Katimon A, Shahid S and Mohsenipour M 2018 *Sustainable Water Resources Management* **4** 991-998

[9] Zhu S, Lian X, Liu X, Hu J, Wang Y and Che J 2019 *Environmental Pollution* **9**

[10] Sharma A, Mitra A, Sharma S and Roy S 2018 *Artificial Neural Networks and Machine Learning-ICANN*

[11] Liu T, Lau A K, Sandbrink K and Fung J C2018 *Journal of Geophysical Research: Atmospheres* **123**

[12] Shmueli G and Lichtendahl K C JR 2016 *Practical Time Series Forecasting With R* (USA: Axelrod Schnall)