Study of Nitrogen Dioxide Concentration in New York based on Seasonal ARIMA Model

Ran Li

13/12/2020

Introduction

Air Pollution has become a global problem with the rapid development of economics these days. More and more problems in society are raised due to the severe air pollution such as chronic diseases and respiratory diseases in some regions, so it is worth studying the air pollutent time series in New York City, which is one of the busiest and crowdest city in the world. An approproiate model may help with forecasting air pollution and also help with policy making. Here we are going to build seasonal ARIMA model to the series and make prediction about air pollution.

Data

Data we used for analysis in this article is open data scraped from United States Environmental Protection Agency website.[1]. The whole dataset contains daily record for air pollution from January 1st 2000 to April 2016 (192 months and almost 5500 days in total) across the U.S in major cities including the New York City. In this article, we will focus on the study of monthly average of nitrogen dioxide(NO_2) concentration in NYC, so data taken from other regions are deleted to get a smaller dataset. Also, in order to make our analysis easier (but this will cause some error), data detected are in Bronx (one of counties in New York).Overall, by taking monthly average during the data cleaning procedure, we have a 192-month(frequency = 12) time series. Similarly, we can obtain monthly average for other major pollutents as well, but we are going to focus on NO_2 here.

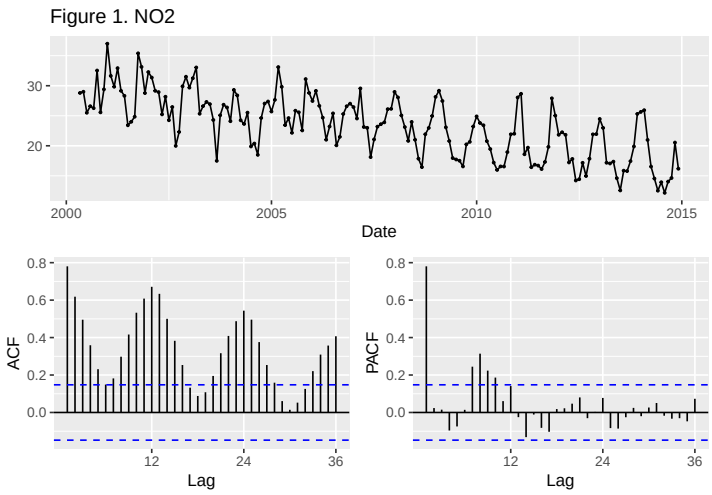
In order to test and compare models, we splits our series into training set containing data in first 176 months (January 2000 to December 2014) and testing set (January 2015 to April 2016), where the training set is used to build the Seasonal ARIMA model and the testing set for comparing model accuracy.Details will be discussed in result section.

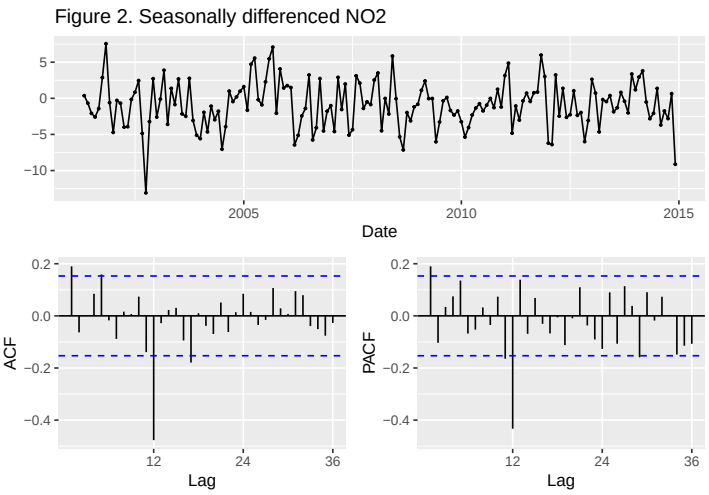
Results

Model Identification

We build the model based on data in first 176 months (training set). The plot of series is shown in Figure 1. It can be seen that there is similar "U-shaped" pattern yearly. Spectral analysis can be done to show this periodicity. See Figure 6 for the Periodogram of this series. It is clear that a narrow peak occurs at 12, giving us the predominant frequency to be 1/12. It has periodogram 63.51932 with 95% confidence interval [17.21914,2508.879], indicating that our data cycles every 12 months(1 year). Other peridominant frequencies have periodogram far less than this one, more information is shown in Table 4. Therefore, the s parameter in $SARIMA(p,d,q)(P,D,Q)_s$ is 12. Furthermore, this periodicity and the decreasing trend in the plot of series imply that the concentration of NO_2 is nonstationary.One can also tell the nonstationarity from the slowly decayed Autocorrelation Function(ACF), motivating us to perform the seasonal differencing by 1 (12 months, $D = 1$).

Figure 2 is the corresponding seasonally differenced data, along with its ACF and PACF. The series seem to have more stable variance and expectation from the plot. ACF and PACF also show indication of SARIMA model. So next we are going to estimate the order by observing the ACF and PACF plot.





Based on ACF, for seasonal data, we can see there is obvious long spike at lag = 12, implying that the seasonal part should contain MA(1) component. Similarly, a long spike at lag = 1 indicates an MA(1) component for the nonseasonal part. From the PACF, we can see long spikes for lag = 1, 12 and lag = 24. So we propose first that seasoal part has AR(2) and nonseasonal part has AR(1). Combining together, we manually select $SARIMA(1,0,1)(2,1,1)_{12}$ model.

By slightly changing the order of model a little bit, we can obtain models that are very close to the one we selected manually, which are shown in Table 1 as below. Noticeably, among all these models, $SARIMA(1,0,0)(2,1,1)_{12}$ is the one selected automatically by using function `auto.arima()` in R. This model only differs in the order of moving average in nonseasonal part compared our manually selected one.

Model Testing and Selection

To compare models, we employee AIC, BIC creteria, prediction accuracy. Values of different creterion are shown in table one. The lower AIC and BIC, the better the model fits the dataset. Accuracy of model is identified by the Root Mean Square Error(RMSE). This is computed by predicting the monthly NO_2 concentration from January 2015 to April 2016 and comparing with actual values in the testing set. Lower RMSE implies better accuracy in prediction [9].

It is suprised to find that using these three creterion, neither our manually chosen model nor the automatically chosen SARIMA model performs the best (See Table 2). Instead, $SARIMA(2,0,1)(2,1,1)_{12}$ has the lowest AIC, BIC and RMSE. Further, this model also passes the Ljung-Box test for residuals (result in figure 3). The Q-statsitic is never significant for lags shown, together with the shape of QQ-plot support the normality and independence of white noise assumption for the residuals, indicating that this model takes enough information from the dataset. Therefore $SARIMA(2,0,1)(2,1,1)_{12}$ is selected and will be employed in forecasting. The model can be written as:

$$(1-\phi_1 B-\phi_2 B^2)(1-\psi_1 B^{12}-\psi_2 B^{24})(1-B^{12})x_t = (1+\theta_1 B)(1+\Theta_1 B^{12})w_t$$

where x_t is the monthly average of NO_2 concentration seires, w_t is mormal white noise error and B represents the backward operator.Other letters are parameters that will be estimated using maximum likelihood function. Results of estimation are shown in the following table 2.

We notice that even though our model passes the Ljung Box test for residuals, the p-values for the parameter estimate are not all significant, many estimates have large p-values indeed. But it is hard to find a SARIMA model that passes every test in this case. Limitation and weakness will be discussed further in the next section.

Forecasting

4 month forecasts of monthly average of NO_2 concentration (From May 2016 to April 2016) are made using $SARIMA(2,0,1)(2,1,1)_{12}$, plotted in Figure 4, where the slight blue regeion is the 80% confidenct interval and the dark blue band represents the 95% confidence interval. It seemes that the forecasts follow the pattern of the previous values

and there is a general decreasing trend in the data. Detailed information of forecasts for next four months can be found in table 3.

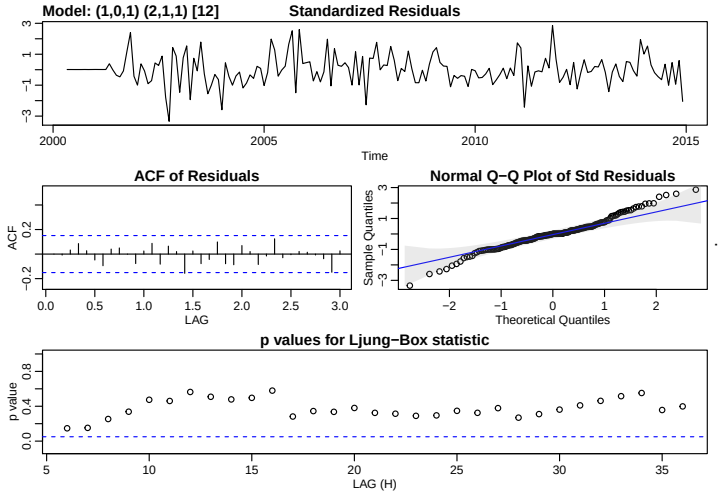


Figure3. Diagnostics of Residuals

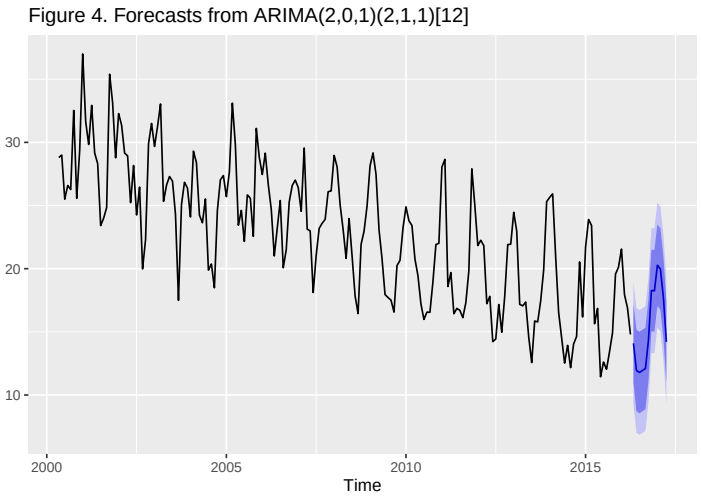


Table 1: Model Comparison

	AIC	BIC	RMSE
(1, 0, 1)(2, 1, 1)12	795.88	814.48	1.61
(1, 0, 0)(2, 1, 1)12	817.77	833.27	1.58
(2, 0, 1)(2, 1, 1)12	790.63	812.33	1.49
(1, 0, 0)(2, 1, 2)12	819.16	837.76	1.53
(1, 0, 1)(1, 1, 1)12	794.11	809.60	1.61
(1, 0, 1)(1, 1, 2)12	796.02	814.62	1.60

Table 2: Estimate of Coefficients for (2, 0, 1)(2, 1, 1)12

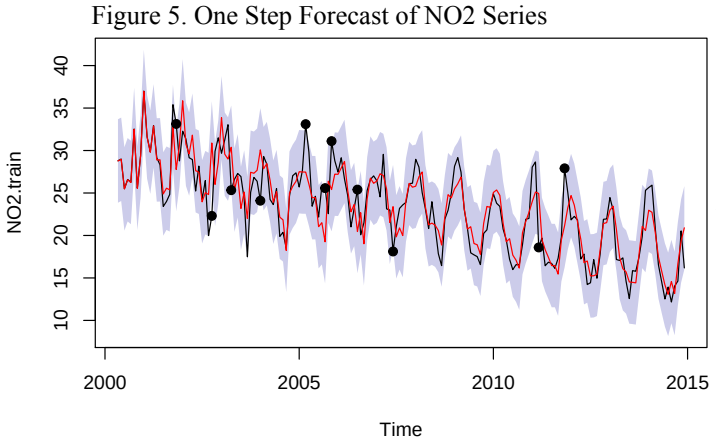
	Estimate	SE	p.value
ar1	-0.24	0.26	0.35
ma1	0.48	0.23	0.04
sar1	-0.03	0.13	0.84
sar2	0.03	0.12	0.79
sma1	-0.80	0.12	0.00
constant	-0.07	0.01	0.00

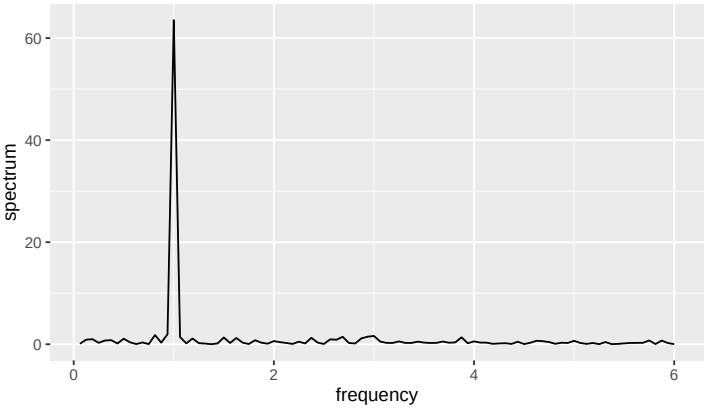
Table 3: Forecast of Nitrogen Dioxide Concentration from May 2016 to August 2016

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
May 2016	14.09	10.94	17.24	9.28	18.90
Jun 2016	11.95	8.72	15.17	7.02	16.87
Jul 2016	11.78	8.56	15.01	6.85	16.72
Aug 2016	11.94	8.71	15.17	7.00	16.88

Table 4: Predominant frequency and Spectrum

Frequency	Spectrum	Low95	High95
0.083	63.519	17.219	2508.878
0.078	1.960	0.531	77.404
0.068	1.799	0.488	71.057
0.250	1.626	0.441	64.208





Discussion

Qualitatively, from our model, we can tell a clear decreasing trend and periodicity in the monthly nitrogen dioxide concentration. This "U-shaped periodicity" is also observed in other countries as well, for example Shenzhen, China.[8]

Spring and winter tends to have higher immision of NO_2 due to temperature and wind speed, it is shown that other major air pollutents including PM_{10}, O_3 and SO_2 also behave in similar periodic pattern.[7] Also, from the decreasing trend, we can see that New York air pollution is getting better in the latest decade, maybe due to effective environmental policies and increased public awareness.

From Figure 5, we can see that the predicted values are pretty close to the actual values where most of them are within the 95% confidence interval with the 1.485278% mean absolute percentage error. This model overall performs a nice fitness to our existing data and thus provide a method for making prediction of air pollution. Indeed, there has already been many sucessful satatistical practices in building ARIMA models for air pollution data in many countries around the world[4],[5]. The effectiveness of prediction of ARIMA model has been appreciated by the many researches in different areas. Similar approach here may also apply to series of other major pollutents or the prediction of the Air Quality Index(AQI).

Weakness and Next Steps

(1) Many of our estimate of parameters based on maximum likelihood are not significant. Even though it is not always easy to find a model that goes through all the model testing process and behaves nice enough in prediction for real world series, we are trying to select optimal ones among those.
(2) Even though ARIMA model performs nicely in short term, we can see that prediction becomes less accurate as time moves on. Increasing the accuracy for longer-range forecasts may be a step to consider next.
(3) In this model, we only consider the regression with its past values, but true air pollution may have correlation with other variables as well. Models that with other information considered may perform a better fittness and prediction.[8]

References

[1] https://aqs.epa.gov/aqsweb/airdata/download_files.html
[2]Box, G., Jenkins, G., (1976). Time Series Analysis: Forecasting and Control. Holden-Day, Boca Raton.
[3] Zhang, L., Lin, J., & Qiu, R. (2018). Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. *Ecological Indicators*.
[4] Lee, M. H., & Abd.Rahman, N. H. (2012). Seasonal ARIMA for Forecasting Air Pollution Index: A Case Study. *American Journal of Applied Sciences*, 570-578.
[5]Guarnaccia, C. (n.d.). ARIMA Models Application to Air Pollution Data in Monterrey, Mexico. *Mathematical Methods and Computational Techniques in Science and Engineering II*.
[6]Ye, Z. (2019). Air Pollutants Prediction in Shenzhen Based on ARIMA and Prophet Method. *E3S Web of Conferences*.
[7]Czarnecka, M., Nidzgorska-Lencewicz, J. (2011). Impact of weather conditions on winter and summer air quality. *Int. Agrophys.*, 25(1), 7-12.
[8]Mou, J., Zhao, X., Fan, J., & Yan, Z. (2017). Time Series Prediction of AQI in Shenzhen Based on ARIMA Mode. *Journal of Environmental Hygiene*, 7(2), 102-107.
[9]Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.