Vol.3.Issue.4.2015 (Oct-Dec.)

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

*A Peer Reviewed International Research Journal*

http://www.bomsr.com

ISSN:2348-0580

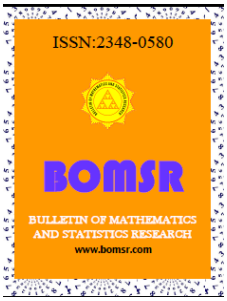
**RESEARCH ARTICLE**

# THE APPLICATION OF ARIMA MODEL ON 2014 AIR QUALITY INDEX IN YANQING COUNTY, BEIJING, CHINA

**JIE ZHU[1], RUOLING ZHANG[1], BINBIN FU[1], RENHAO JIN[1]**

[1]School of Information, Beijing Wuzi University, Beijing, China

Email : Renhao.jin@outlook.com (Renhao Jin)




**ABSTRACT**

In order to study the changes of air quality index (AQI) in Yanqing County, Beijing, Chinaand predict the trend of AQI value, this paperconstructed a time-series analysis. A non-stationary trend is found, andan ARIMA (1, 1, 2) model is found to sufficiently model the data.A short trend of AQI value is then predicted using the established model.

**Keywords:** Air Quality Index (AQI); prediction; ARIMA model.



## 1. INTRODUCTION

Beijingis the capital of China and one of the most populous cities in the world. Its population in 2013 was 21.15 million. The city proper is the 3rd largest in the world. The metropolis, located in northern China, is governed as a direct-controlled municipality under the national government, with 14 urban and suburban districts and two rural counties. It is home to the headquarters of most of China's largest state-owned companies and many large multinational companies, and is a major hub for the national highway, expressway, railway, and high-speed rail networks. As China economic is boosting over 20 years, Beijing is always an attraction in the world. However, in recent 2-3 years Beijing is air pollution problem is often in the headlines of many news articles.China government has noticed this problem and done a lot of measures to control the air pollution in Beijing. In this paper, the air quality index (AQI) is used as a comprehensive figure to measure the air quality.As the AQI increases, an increasingly large percentage of the population is likely to experience increasingly severe adverse health effects [1]. Different countries have their own air quality indices, corresponding to different national air quality standards.This paper only concerns the AQI defined by China government [2].The reasonable analysis and forecast of AQI can help the government make and check their air control police and let the hospitals to prepare their daily patient service.

China's Ministry of Environmental Protection (MEP) is responsible for measuring the level of air pollution in China. The AQI level is based on the level of 6 atmospheric pollutants, namely sulfur dioxide (SO2), nitrogen dioxide (NO2), suspended particulates smaller than 10 μm in aerodynamic diameter (PM10), suspended particulates smaller than 2.5 μm in aerodynamic diameter (PM2.5),

carbon monoxide (CO), and ozone (O3) measured at the monitoring stations in China [2]. Table 1 displays the AQI value and its corresponding level and health implications. As shown in Table 1, when AQI value is less than 100, the air is no effect for daily life, but when AQI is larger than 200, it can may case heavy adverse health effects.

In this paper, the study area is in Yanqing County of Beijing, which is situated in northeast Beijing and has an area of 1,993.75 square kilometers and a population of 317,000. It is an ecological conservation and development area of the capital, and well-endowed with natural resources and a picturesque landscape. The Yanqing County is famous tourism place is Beijing with over 30 unique tourist attractions including Badaling Great wall and Longqing Gorge. The 2022 winter Beijing Olympics will be held in this county as it has great Ice and Snow landscape. It is chosen to be study area as its tourism industry is highly determined by its air quality. As shown in Table 1, health implication of AQI is mainly related to outdoor activities. There is one air quality monitor to examine the air pollution and it publish the AQI value every day. The data is extracted from their everyday report from Jan. 1st 2014 to Dec. 29th 2014.

A lot of methods have been used to analysis and forecast of time series data, such as autoregressive model, autoregressive moving average model, autoregressive conditional heteroscedasticitymodel, autoregressive integrated moving average model (ARIMA) and so on [3]. In the balance of predict and explanation, ARIMA is a wildly used model. In this paper, the AQI data in YanqingCounty of 2014 is firstly examined and then ARIMA model is used to fit the data [4]. All computations are done by using SAS software (SAS® 9.4, SAS Institute Inc., Cary, N.C.) [5].

**Table 1.** AQI and Health Implications by China's Ministry of Environmental Protection.

| AQI | Air Pollution Level | Health Implications |
|---|---|---|
| 0–50 | Excellent | No health implications |
| 51–100 | Good | Few hypersensitive individuals should reduce outdoor exercise |
| 101–150 | Lightly Polluted | Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise |
| 151–200 | Moderately Polluted | |
| 201–300 | Heavily Polluted | Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities |
| 300+ | Severely Polluted | Healthy people will experience reduced endurance in activities. There may be strong irritations and symptoms and may trigger other illnesses. Elders and the sick should remain indoors and avoid exercise. Healthy individuals should avoid outdoor activities |

## 2. Modeling

### 2.1. Description of the data

Atiming diagram is firstly plot using all the AQI data of 2014 in Yanqing County, Beijing. As shown in Figure 1, the AQI values range from 26 to 415 with the annual mean value 115. AQI values peak at spring and winter season, and for the other period of 2014 the AQI seems stationary. It is reasonable to have large AQI values in spring and winter months, as the temperature is relative low

in Beijing at that time, ranging from -10℃ to 5℃ and it often leads tofog and haze weather in low temperature. The number of days for every AQI Pollution level in Yanqing County, Beijing in 2014 are shown in Table 2, and 54.27% of days in 2014 are in Good or Excellent Air level. However, 12.95% days of 2014 in Yanqing are in Heavily Polluted or Severely Polluted. So, in general the air condition in Yanqing County is acceptable and suitable for the tourism industry.

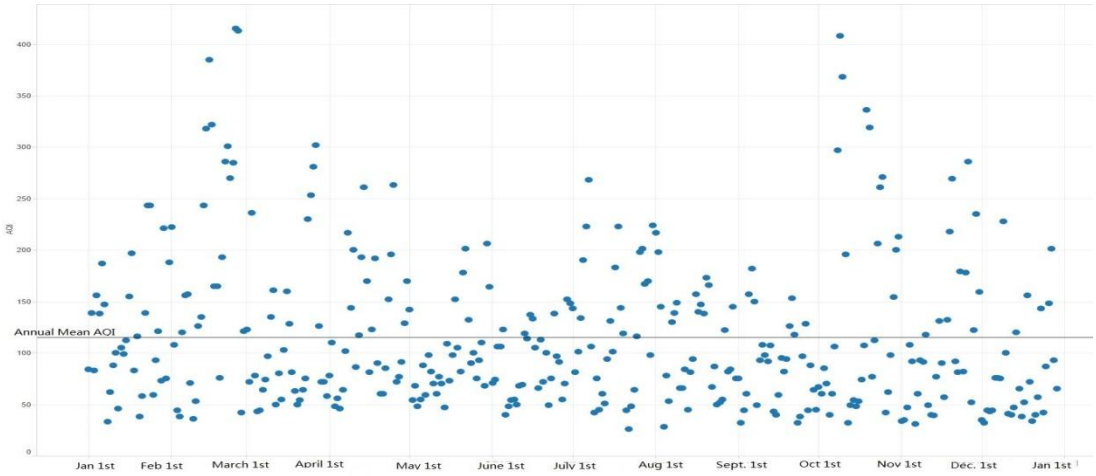



**Figure 1.**Thetiming diagram plot of AQI value of 2014 in Yanqing County, Beijing.

**Table 2.** The number of days for every AQI Pollution level in Yanqing County, Beijing in 2014.

| Air Pollution Level | # Days | Percentage | Cumulative Percentage |
|---|---|---|---|
| Excellent | 56 | 15.43 | 15.43 |
| Good | 141 | 38.84 | 54.27 |
| Lightly Polluted | 78 | 21.49 | 75.76 |
| Moderately Polluted | 41 | 11.29 | 87.05 |
| Heavily Polluted | 36 | 9.92 | 96.97 |
| Severely Polluted | 11 | 3.03 | 100.00 |

As shown in Figure 1, the two peaks on the two sides of the plot break the hypotheses of weaker stationary. A weaker form of stationarity commonly employed in time series is known as second-order stationarity, which only requires that 1st moment and auto-covariance do not vary with respect to time [6]. So, for a continuous-time random process X(t), it has the following properties: the mean function E{X(t)} must be constant [6,7]. It is unreasonable to assume the AQI values in the figure 1 have aconstant mean value from the Figure 1, thus a non-stationary model should be used to model the data [8].

### 2.2. ARIMA model

In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). They are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied to reduce the non-stationarity. ARIMA models are generally denotedARIMA (p, d, q)where parameters p, d, and q are non-negative integers, p is the order of the Autoregressive model, d is the degree of differencing, and q is the order of the Moving-average model [9]. ARIMA models form an important part of the Box-Jenkins approach to time-series modelling.When two out of the three terms are zeros, the model may be referred to the non-

zero parameter, dropping "AR", "I" or "MA" from the acronym describing the model. For example, ARIMA (1,0,0) is AR(1), ARIMA(0,1,0) is I(1), and ARIMA(0,0,1) is MA(1). [3, 4, 10]

The basic idea of ARIMA model is to view the data sequence as formed by a Stochastic Process on time. Once the model has been identified, the model can be used to predict the future value from the past and present value of the time series. To identify the parameters in ARIMA (p, d, q),firstly thescatter plots of time series, self-correlation function and partial auto correlation function plot are used to test its variance, trend and seasonal variation, stability of sequence recognition. For general applications, the time series data are not stationary series. The next step is to do some data manipulation on the non-stationary sequence. If the data series is non-stationary, and there is a certain growth or decline, the data difference is need to be proceed. The parameter d is the order of difference to transform the original non-stationary time series data to stationary time series data. After the data processing, the correlation function value or partial correlation function values should be not significantly different from zero. According to the identification rules on time series, the corresponding model can be established. If a partial correlation function of a stationary sequence is truncated, and self-correlation function is tailed, it can be concluded the sequences for AR model; if partial correlation function of a stationary sequence is tailed, and the self- correlation function is truncated, it can be determined that the MA model can be fitted for the sequence. If the partial correlation function of a stationary sequence and the auto-correlation function are trailed, then the ARMA model is suitable for the sequence [9, 11].

Following ARIMA model procedure, a First order differencing is computed for the data, and then a timing diagram of the differencing data is computed and shown in Figure 2. The differencing data shows a stationary pattern, although several outliers exist, thus it is suitable to let parameter d=1. Auto-correlogram (Figure 3) is also done on the differencing data, which displays a short-term autocorrelation and confirms the stationary of the differencing data. To make an accurate inference of the data, autocorrelation check forwhite noise is also done on the differencing data. As shown in Table 3, the white noise hypotheses is rejected on lag 6, 12, 18 and 24 with very small p-values. All these results shows that an ARMA model can be fitted to the first order differencing data. From the Figure 3 of the Auto-correlogram, it is safe to determine that q is no more than 2, while as shown in Figure 4, the partial autocorrelation is no more than 4. This means that it is enough to choose the model in the set of $\{p \leq 4 \text{ and } q \leq 2\}$. From the discussion above, it concludes that the ARIMA (p, 1, q) is suitable to AQI data of Yanqing 2014, but the parameters p and q need to be determined [11, 12].

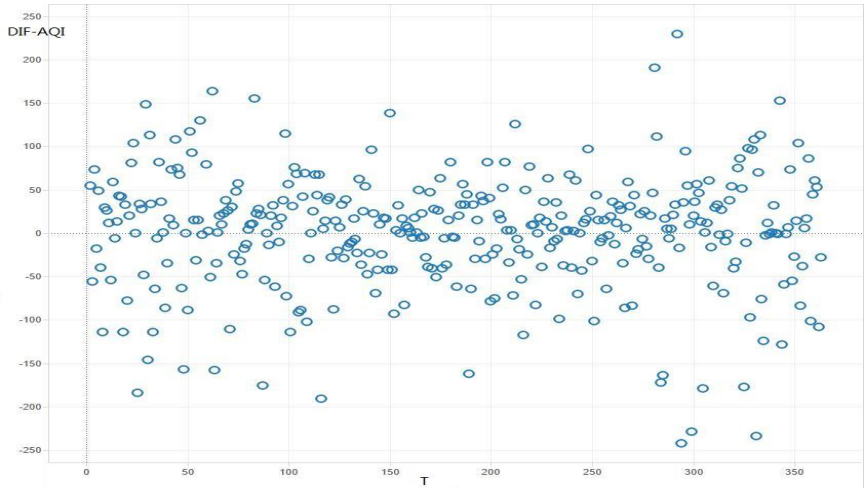



**Figure 2.** Thetiming diagram plot of the First order differencing data on AQI data of 2014 in Yanqing County, Beijing.

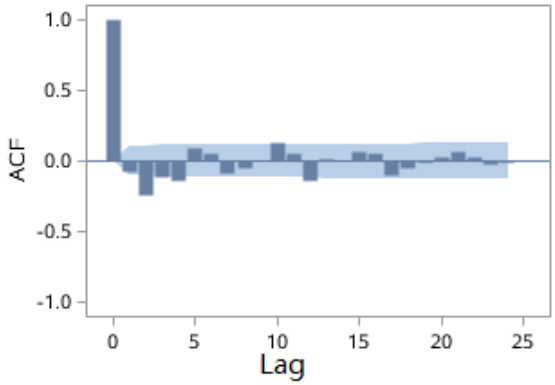



**Figure 3.** The Autocorrelogram on the first order differencing data of original AQI data.

**Table 3.**Autocorrelation check for white noise on the differencing data at lag 6, 12, 18 and 24.

| Autocorrelation Check for White Noise | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Lag | Chi-Square | DF | P-Value | Autocorrelations | | | | | |
| 6 | 40.02 | 6 | <.0001 | -0.079 | -0.239 | -0.116 | -0.144 | 0.094 | 0.050 |
| 12 | 59.80 | 12 | <.0001 | -0.090 | -0.052 | 0.004 | 0.135 | 0.053 | -0.145 |
| 18 | 67.64 | 18 | <.0001 | 0.009 | 0.006 | 0.064 | 0.052 | -0.108 | -0.045 |
| 24 | 70.47 | 24 | <.0001 | -0.011 | 0.032 | 0.069 | 0.025 | -0.026 | -0.012 |

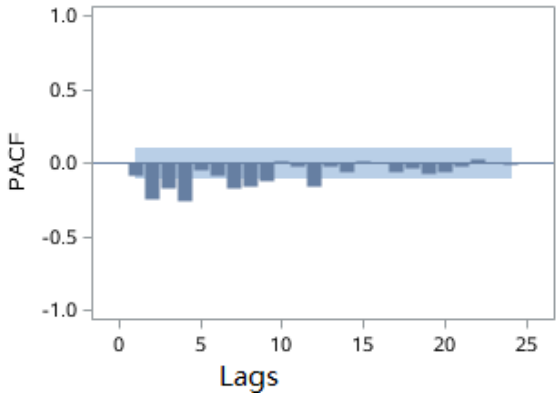



Figure 4. The partial Autocorrelogram on the first order differencing data of original AQI data.

At the significant level of 0.05, all the ARIMA (p, 1, q) models with $\{p \leq 4 \text{ and } q \leq 2\}$ are compared. The model with all parameters are significantly different from 0 and least AIC value is selected as the best model [13]. The model ARIMA (1, 1, 2) is chosen and the estimated parameters are shown in Table 4. The constant term are eliminated as its p-value is 0.92.

**Table 4.**The parameters estimated in the ARIMA(1,1,2).

| Parameter | Estimate | STD | Pr>t |
|---|---|---|---|
| MA1,1 | 0.66926 | 0.08618 | <.0001 |
| MA1,2 | 0.30687 | 0.08234 | 0.0002 |
| AR1,1 | 0.35640 | 0.08517 | <.0001 |

The final model is

$$(1 - 0.3564\text{B})(AQI_t - AQI_{t-1}) = (1 - 0.66926B - 0.30687B^2)\varepsilon_t$$

Using the built model, five steps predictions and their confidential intervals are calculated and shown in Figure 5.

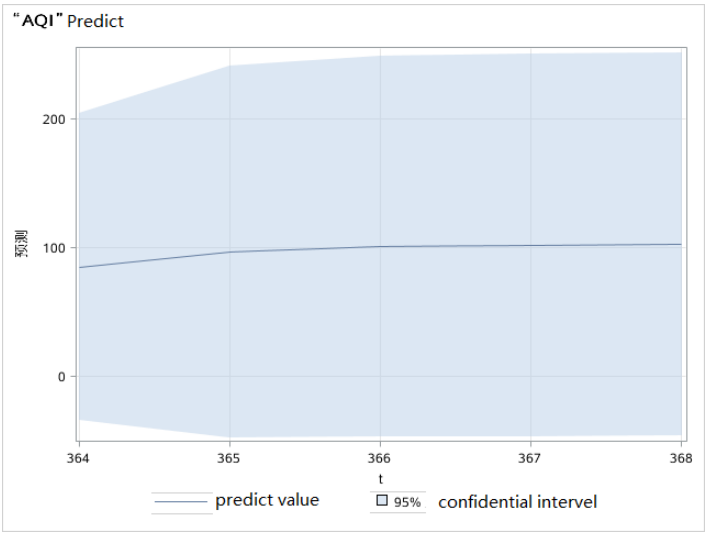



Figure 5. The next 5 steps prediction of AQI value based on the final model. The predicted value are shown with its 95% confidence limits.

## 3. Conclusions

This paper does a study on 2014 the air quality index (AQI) in Yanqing County, Beijing, China. In the process of model building, the original AQI data is found to be non-stationary, but the first order differencing data of original AQI data is stationary. By comparing with several models, ARIMA (1, 1, 2) is chosen as the final model and it succeeds in predicting five steps trends of AQI and their 95% confidential intervals. Considering the fluctuations of AQI, this model can be applied to predict future values. The fluctuations of AQI value are non-rational, and it is influenced by many factors. No model can include all these factors, but this predict model can still help government and other authorities to take advanced measures to the coming air condition.

**Acknowledgements**

This paper is funded by the project of National Natural Science Fund, Logistics distribution of artificial orderpicking random process model analysis and research (Projectnumber: 71371033); and funded by intelligent logisticssystem Beijing Key Laboratory (No.BZ0211); and funded by scientific-research bases---Science & TechnologyInnovation Platform---Modern logistics information andcontrol technology research (Project number:PXM2015_014214_000001); University Cultivation FundProject of 2014-Research on Congestion Model andalgorithm of picking system in distribution center(0541502703).

**REFERENCES**

[1]. Garcia, Javier; Colosio, Joëlle (2002). Air-quality indices: elaboration, uses and international comparisons. Presses des MINES.

[2]. "People's Republic of China Ministry of Environmental Protection Standard: Technical Regulation on Ambient Air Quality Index". Access:http://kjs.mep.gov.cn/hjbhbz/bzwb/dqhjbh/jcgfffbz/201203/W020120410332725219541.pdf

[3]. Box, G.E.P., Jenkins, G.M., and Reinsel, G.C.(1994), Time Series Analysis: Forecasting and Control, 3rd edition, Prentice Hall: Englewood Cliffs, New Jersey.

[4]. Box, G.E.P., and Pierce, D. (1970), "Distribution of Residual Autocorrelations in Auto-regressive-Intergrated Moving Average Time Series Models," Journal of the American statistical Association, 65, 1509-1526.

[5]. SAS Institute Inc, (2014). SAS/STAT® 9.4 User's Guide: The ARIMA Procedure (Book Excerpt). NC: SAS Institute Inc, Cary.

[6]. Bollerslev T. Generalized autoregressive conditionalheteroskedasticity [J]. Journal of Econometrics, 1986, 31 (3):309-317.

[7]. Engle R.F. Autoregressive conditional heteroskedasticity withestimates of the variance of United Kingdom inflation [J].Econometric, 1982, 50 (4): 989-1004.

[8]. Engle R.F., Kroner F.K. Multivariate SimultaneousGeneralized ARCH [J].Econometric Theory, 1995, 11:135-149.

[9]. 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.

[10]. Cox, D. R., & Wermuth, N. (1991). A simple approximation for bivariate and trivariate normal integrals. International Statistical Review/Revue Internationale de Statistique, 59(2), 263-269.

[11]. Engle Robert F. Dynamic Conditional Correlation: A SimpleClass of Multivariate GARCH Models [J]. Journal of Businessand Economic Statistics, 2002, 20 (3):341-347.

[12]. Engle R.F., Lilien D.M., Robins R.P. Estimating time-varyingrisk Premia in the term structure: The ARCH-M model [J].Econometrica, 1987, 55: 395-406.

[13]. Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in B.N. Petrov and F.Csaki, ed. 2nd International Symposium on Information Theory, 267-281. Akademia Kiado: Budapest.