International Journal of Innovation in Science and Mathematics
Volume 3, Issue 6, ISSN (Online): 2347–9051

Comparison of ARIMA model and Exponential Smoothing Model on 2014 Air Quality Index in Miyun County, Beijing, China

Renhao Jin*, Tao Liu, and Sha Wang
School of Information, Beijing Wuzi University, Beijing, China
*Correspondent Author's Email: Renhao.jin@outlook.com

***Abstract* – In order to study the changes of air quality index (AQI) in Miyun County, Beijing, China and predict the trend of AQI value, this paper constructed a time-series analysis-A non-stationary trend is found, and the ARIMA (3, 1, 3) model and Holt exponential smoothing model are found to sufficiently model the data. In comparison of these two model fittings, the Holt modelling result are better than ARIMA modelling's in terms of trend capturing and result MSE, and in this data it is better to apply the Holt model to predict the future AQI values.**

***Keywords* – Air Quality Index (AQI), Prediction, ARIMA Model, Exponential Smoothing Model, Holt Model.**

I. INTRODUCTION

Beijing is 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 Miyun County of Beijing, which is situated in northeast Beijing and has an area of 2,227 square kilometers and a population of half million. The Miyun County is famous tourism place is Beijing with SimaTai Great wall and large Miyun reservoir supplying water for the whole Beijing. 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 heteroscedasticity model, autoregressive integrated moving average model (ARIMA), Holt Exponential Smoothing Model and so on [3]. In these models, ARIMA and Holt methods are two wildly used models [4, 5]. In this paper, the performance of these two models are compared on the AQI data in Miyun County of 2014. All computations are done by using SAS software (SAS® 9.4, SAS Institute Inc., Cary, N.C.) [6, 7].

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

AQI	Air PollutionLevel	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
151–200	ModeratelyPolluted	reduce outdoor exercise
201–300	Heavily Polluted	Healthy people will be noticeably affected. People with breathing or heart problems

AQI	Air PollutionLevel	Health Implications
		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

II. MODELING

2.1 Description of the data

A timing diagram is firstly plot using all the AQI data of 2014 in Miyun County, Beijing. As shown in Figure 1, the AQI values range from 31 to 427 with the annual mean value 118. 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 to fog and haze weather in low temperature. The number of days for every AQI Pollution level in Miyun County, Beijing in 2014 are shown in Table 2, and 52.89% of days in 2014 are in Good or Excellent Air level. However, 12.39% days of 2014 in Miyun are in Heavily Polluted or Severely Polluted. So, in general the air condition in Miyun County is acceptable and suitable for the tourism industry.

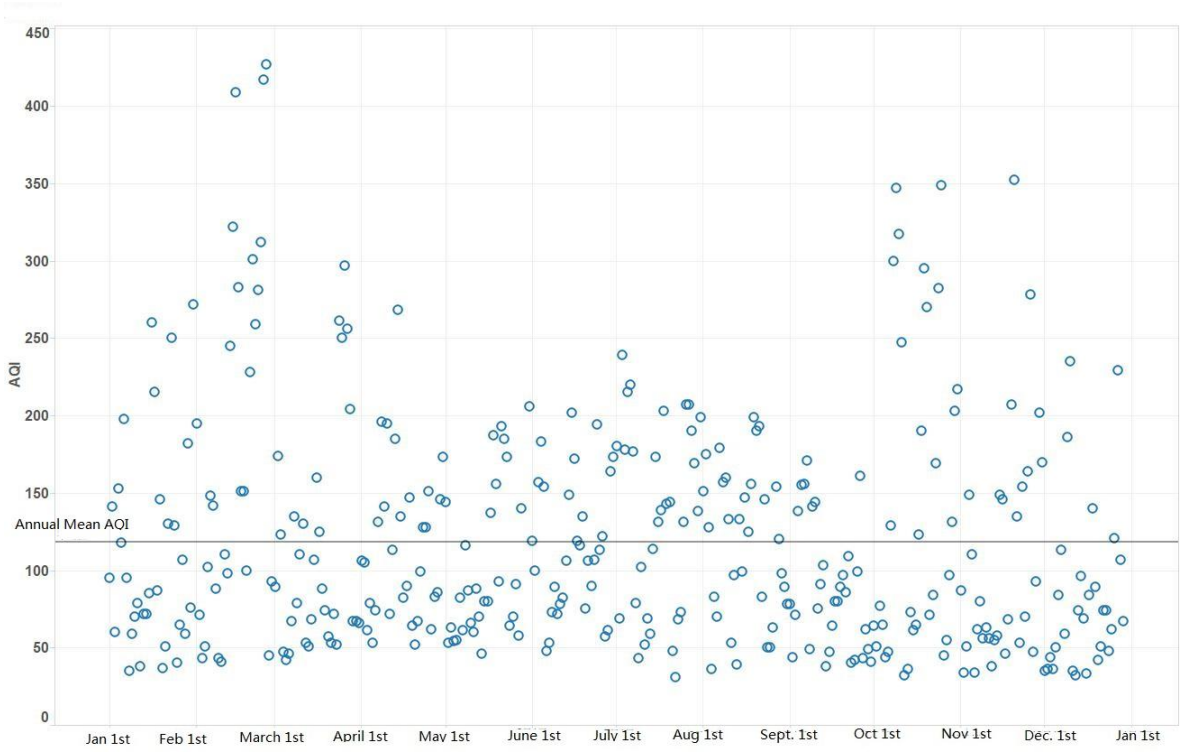

Fig.1. The timing diagram plot of AQI value of 2014 in Miyun County, Beijing.

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

Air Pollution Level	# Days	Percentage	Cumulative Percentage
Excellent	48	13.22	13.22
Good	144	39.67	52.89
Lightly Polluted	74	20.39	73.28
Moderately Polluted	52	14.33	87.60
Heavily Polluted	35	9.64	97.25
Severely Polluted	10	2.75	100.00

As shown in Figure 1, the two peaks on the two sides of the plot break the hypotheses of weaker stationary [8]. It also has no linear trend in diagram and very difficult to match the trend in the Figure to any curve model such as polynomial models and exponential models, thus a non-stationary model could be fitted to the data [8, 9]. Because the data is only in one year, i.e., from Jan. 1st 2014 to Dec. 29th 2014, it cannot be fitted with seasonal effects in the

model. In non-stationary models, Holt exponential smoothing model and ARIMA models are two prefer models which are generally used [10].

2.2 ARIMA and Holt modelling

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 denoted ARIMA (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 [11]. 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, 12].

Holt exponential smoothing method is the most popular double exponential smoothing method, proposed by Holt (1957) with extending simple exponential smoothing to allow forecasting of data with a trend. Holt method is to concentrate on the series of increments X_t-X_(t-1), and then estimate the slop parameter to a linear trend by exponential smoothing of these differences. The Holt method can be expressed as following formulas,

$$\tilde{X}_t = \alpha X_t + (1-\alpha)(\tilde{X}_{t-1} + b_{t-1})$$

And

$$b_t = \gamma(\tilde{X}_t - \tilde{X}_{t-1}) + (1-\gamma)b_{t-1}$$

The formula for prediction is

$$\tilde{X}_{T+i} = \tilde{X}_T + i\,b_T$$

In this formulation, two weighting parameters (α and γ) are used for the two updating equations. In the model fitting of the Holt exponential smoothing method, the initial parameters of $\tilde{X}_0$ and b_0 need to be determined. In this study, the initial value of the smoothing series $\tilde{X}_0$ is set to be X_1, i.e., $\tilde{X}_0 = X_1$. The initial value of the trend series b_0 can be defined by many ways, and a simple method is defined that for an arbitrarily n,

$$b_0 = (X_{n+1} - X_1)/n.$$

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 for white 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 3, while as shown in Figure 4, the partial autocorrelation is also no more than 3. This means that it is enough to choose the model in the set of {p≤3 and q≤3}. From the discussion above, it concludes that the ARIMA (p, 1, q) is suitable to AQI data of Miyun 2014, but the parameters p and q need to be determined [13, 14].

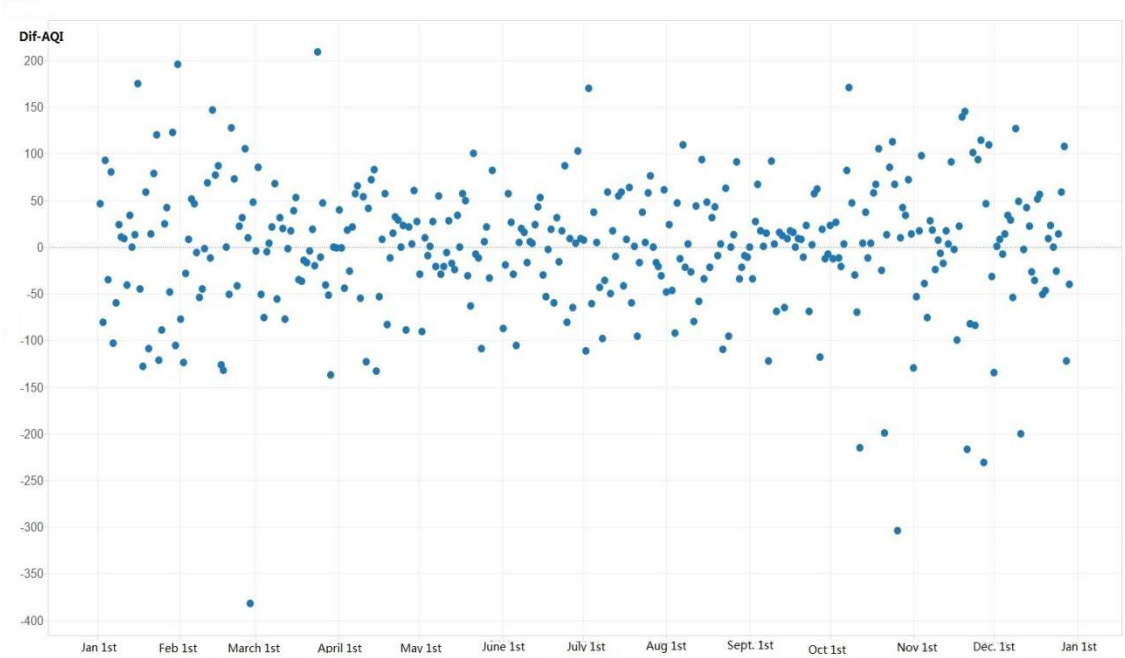

Fig.2. Thetiming diagram plot of the First order differencing data on AQI data of 2014 in Miyun County, Beijing.

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	33.36	6	<.0001	-0.104	-0.209	-0.172	-0.071	0.067	0.026
12	46.70	12	<.0001	-0.039	-0.057	-0.015	0.116	0.052	-0.132
18	55.12	18	<.0001	0.007	-0.002	0.108	-0.024	-0.000	-0.112
24	58.58	24	0.0001	-0.022	0.030	0.084	0.019	-0.019	-0.034

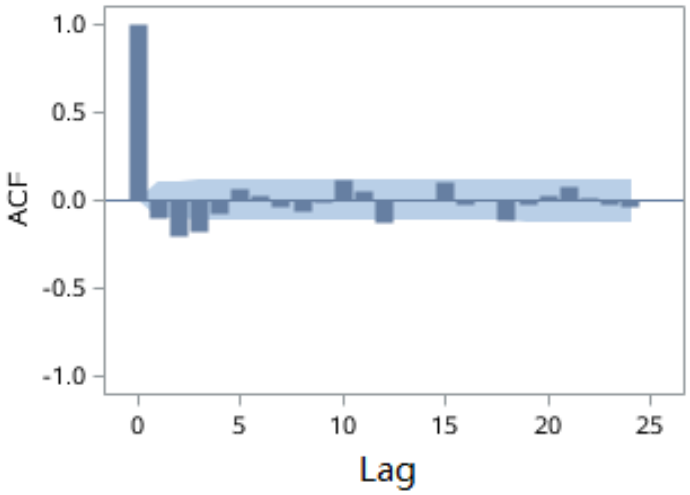

Fig.3. The Autocorrelogram on the first order differencing data of original AQI data

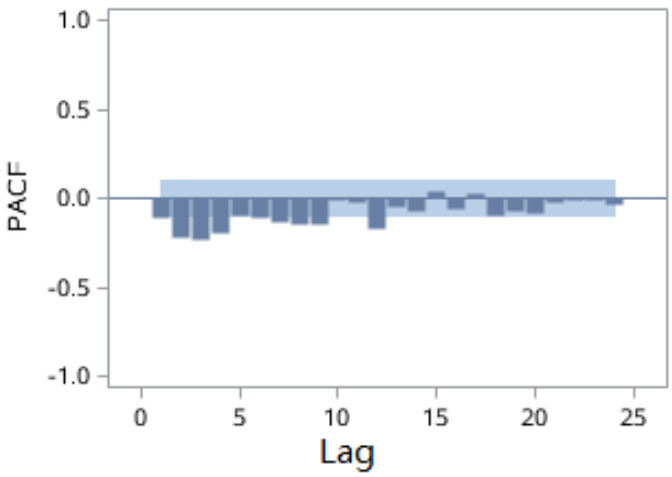

Fig.4. The partial Auto correlogram on the first order differencing data of original AQI data.

At the significant level of 0.05, the parameters in all the ARIMA (p, 1, q) models with {p≤3 and q≤3} are all significantly different from 0, but the ARIMA (3, 1, 3) has the minimum AIC value, thus the final model is ARIMA (3, 1, 3). The constant term are eliminated as its p-value is 0.34. The final ARIMA model is

$$(1 - 0.95746B + 1.09473B^2 - 0.44404B^3)(AQI_t - AQI_{t-1}) = (1 - 1.30923B + 1.15103B^2 - 0.78914B^2)\,\varepsilon_t$$

Comparing with ARIMA modelling, Holt modelling is relative simple. The least squared method is often used to estimate parameters of Holt method, and also chosen in this study. The two parameters of Holt modelling are estimated as α=0.11 and β=0.25. To compare the performance of these two models, the fitted Holt exponential smoothing model, ARIMA model and timing diagram plot of AQI value of 2014 in Miyun County, Beijing are shown in Figure 5. It can be seen from the plot that Holt model fitting result well capture the trend in the data as the green line match the blue line almost everywhere. The mean squared error (MSE) of these two model fitting results are also calculated, and the MSEs of Holt model fitting and ARIMA model fitting are 3842.41 and 6012.47 respectively. So, the Holt modelling result are better than ARIMA modelling's in terms of trend capturing and result MSE, and in this data it is better to apply the Holt model to predict the future AQI values. It is noted that Holt smoothing model only can be used to predict future short steps, just like all the other time series methods.

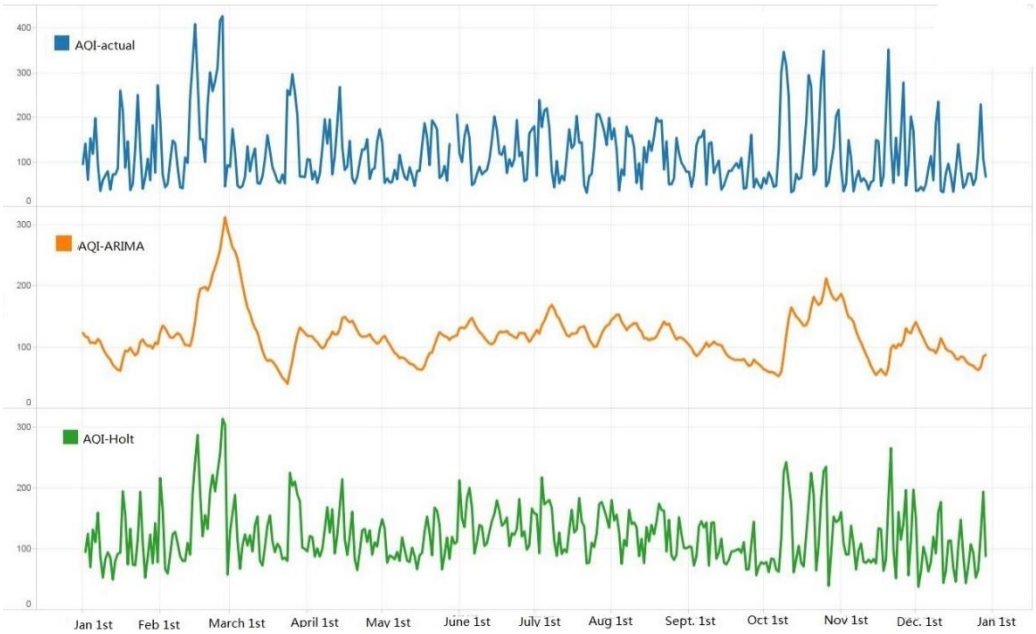

Fig.5. The fitted ARIMA model, Holt exponential smoothing model and timing diagram plot of AQI value of 2014 in Miyun County, Beijing.

III. Conclusion

This paper does a study on 2014 the air quality index (AQI) in Miyun 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. In the ARIMA model fitting, comparing with several models, ARIMA (3, 1, 3) is chosen as the final model. In the Holt exponential smoothing model fitting, least squared method is used to model the data. In comparison of these two model fittings, the Holt modelling result are better than ARIMA modelling's in terms of trend capturing and result MSE, and in this data it is better to apply the Holt model to predict the future AQI 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

Acknowledgment

This paper is funded by the project of National Natural Science Fund, Logistics distribution of artificial order picking random process model analysis and research (Project number: 71371033); and funded by intelligent logistics system Beijing Key Laboratory (No.BZ0211); and funded by scientific-research bases---Science & Technology Innovation Platform---Modern logistics information and control technology research (Project number: PXM2015_014214_000001); University Cultivation Fund Project of 2014-Research on Congestion Model and algorithm 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] Anders Milhoj (2013). Practical Time Series Analysis Using SAS. NC: SAS Institute Inc, Cary.

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

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

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

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

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

[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.

[12] 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.

[13] Engle Robert F. Dynamic Conditional Correlation: A Simple Class of Multivariate GARCH Models [J]. Journal of Business and Economic Statistics, 2002, 20 (3):341-347.

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

International Journal of Innovation in Science and Mathematics
Volume 3, Issue 6, ISSN (Online): 2347–9051

AUTHOR'S PROFILE

Renhao Jin
Lecturer in Statistics, School of Information, Beijing wuzi University, Beijing, China. Current research areas:Spatial and spatial-temporal statistics, Data Mining and statistical modelling.
Email: Renhao.jin@outlook.com