PROCEEDINGS OF SCIENCE

The Analysis and Forecasting of Time Sequence Based on Seasonal Autoregressive Integrated Moving Average Model

Zhimin Deng [1]
College of Automation Science and Technology, South China University of Technology, Guangzhou, 510641, China
E-mail: 401200861@qq.com

Xiaofeng Yang
Guangzhou Institute of standardization, Guangzhou, 510051, China
E-mail: 540208138@qq.com

This paper focus on the modeling of quantities of commodity barcodes registrated in Guangzhou. With analysis of the sample time sequence, difference method is applied to turn the sample sequence into stationary one. By studying the autocorrelation function(ACF) and partial autocorrelation function(PACF) figures, seasonal autoregressive integrated moving average (SARIMA) models are put forward. After simulation, SARIMA$(0,1,1)(0,1,1)^{12}$ model is proved to be of better accuracy, and the result demonstrates the goodness of fit. By utilizing this model, the paper forecast the quantities of commodity barcodes registration in the next six months. The result indicates the quantities of commodity barcodes registration will develop in the future, with year-on-year growth of about 10%.

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1]This work is supported in part by Guangdong Province and Ministry Cooperation Projects of Industry-University-Research (2013B090600062), and Science and Technology Planning Project of Guangdong Province (2014B040401001)

PoS(ISCC2015)002

1. Introduction

When drawing up projects for the development in the future, it is always necessary to forecast the future sequences based on previous and recent datas. In this field, SARIMA models are usually applied. In recent years, lots of scholars have applied SARIMA models to analyse time sequences. For instance, SARIMA model was used to forecast the short-term traffic flow[1]. The air pollution in small urban area was analysed and forecasted based on SARIMA model[2]. And the water demand in Iran was estimated based on SARIMA model[3]. In theory, SARIMA model is suitable for the forecasting of time sequences with nonstationarity and seasonality[4-5]. In this paper, we establish a SARIMA model to forecast the quantities of commodity barcodes registration in the future.

2. SARIMA Model

SARIMA model is an expanded form of ARMA model. ARMA (Autoregressive and Moving Average) model is a method to forecast time sequences. It is composed of two parts: AR(Autoregressive) model and MA (Moving Average) model. Given time sequence $\{y_t\}$, if the forecated variable could be expressed as a linear combination of previous observations, it can be called AR(p) model as showed in formula (2.1), where c is the constant, ε_t is the residual error of the model and obey a normal distribution which has zero mean and constant variance, and p is the number of lags in time sequence. Differenced with AR(p) model, in MA(q) model the forecasted variable is a function of the current and previous errors, as presented in formula (2.2).

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t \tag{2.1}$$

$$y_t = c + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{2.2}$$

If operator B is introduced, supposing $B^j x_t = x_{t-j}$, then formular (2.1) and (2.2) could be briefly illustrated as the following formulas:

$$\varphi(B) y_t = c + \varepsilon_t \tag{2.3}$$

$$y_t = c + \theta(B) \varepsilon_t \tag{2.4}$$

where $\varphi(B) = 1 - \varphi_1 B + \cdots + \varphi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$.

Combining AR(p) model and MA(q) model, we will get ARMA(p,q) model, which can be expressed as formula (2.5) or (2.6):

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \tag{2.5}$$

$$\varphi(B) y_t = c + \theta(B) \varepsilon_t \tag{2.6}$$

In particular, it requires the time sequence to be stationary when establishing ARMA(p,q) model for forecasting. But as a matter of fact, it is hard to make the time sequence strictly stationary. In order to turn nonstationarity into stationarity, difference is usually applied. The ARMA(p,q) model established for time sequence with d order difference is called ARIMA(p,d,q) model, which could be represented as following:

$$\varphi(B)(1-B)^d y_t = c + \theta(B)\varepsilon_t \tag{2.7}$$

For time sequence with seasonality, ARIMA(p,d,q) model should be made seasonal difference to form SARIMA(p,d,q)(P,D,Q)^S model as represented as following:

$$\varphi(B)\Phi(B^S)(1-B)^d(1-B^S)y_t = c + \theta(B)\Theta(B^S)\varepsilon_t \tag{2.8}$$

where $\Phi(B^S) = 1 - \Phi_1 B^S - \cdots - \Phi_P B^{PS}$ and $\Theta(B^S) = 1 - \Theta_1 B^S - \cdots - \Theta_Q B^{QS}$.

3. Analysis of Sample Time Sequence

This research choose monthly quantities of commodity barcodes registrated in Guangzhou as the object of study. The datas come from Guangzhou Institute of Standardization. Among them, we select the datas during Jan. 2004 to Jun. 2014 as sample, named y_t, in which there are 120 observations(shown in Figure 1). And datas during Jul. 2014 to Jun. 2015 are used to test the accuracy of the model.

In order to test the stationarity, ACF figure(shown in Figure 2) and unit root test(shown in Table 1) are applied. Figure 2 shows that the autocorrelation coefficients of sequence y_t fluctuate and can't tend to 0, and as indicated in Table 1, the value of ADF is larger than the test critical value in 10% level. Both of these suggest the nonstationarity of sequence y_t.

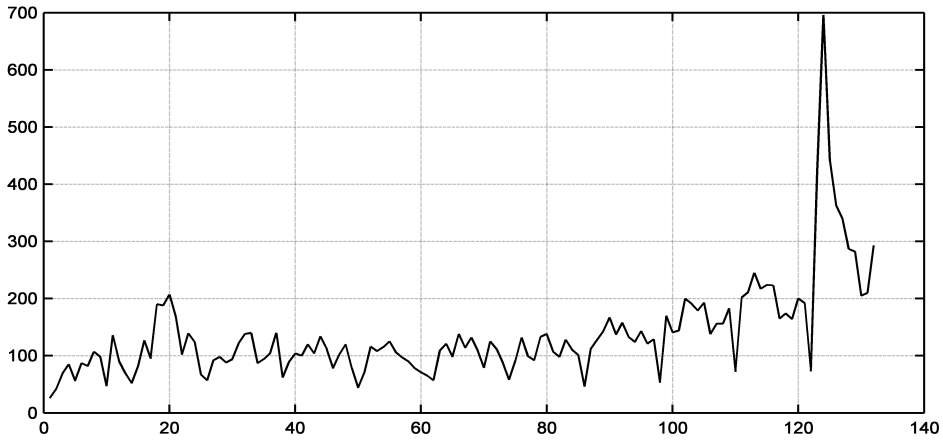

Figure 1: Sample Time Sequence

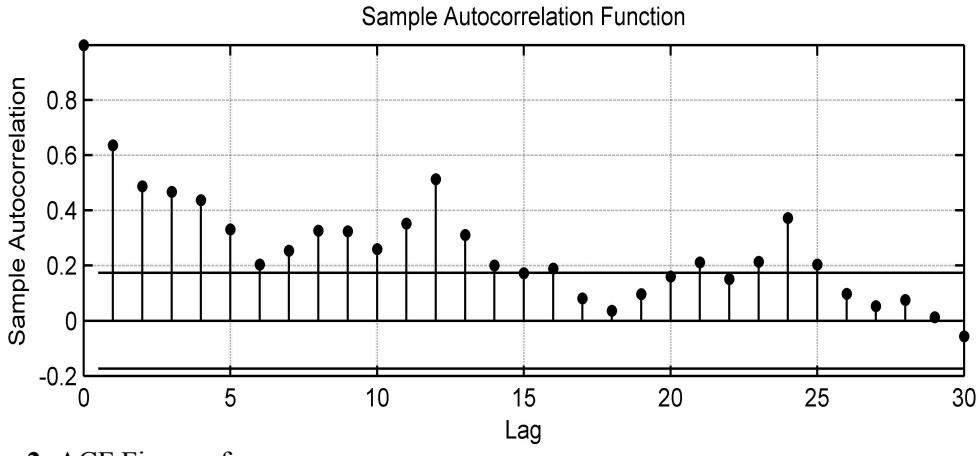

Figure 2: ACF Figure of y_t

	ADF test statistic	Test critical values		
		1% level	5% level	10% level
t-statistic	-1.568624	-2.583593	-1.943406	-1.615024

Table 1: ADF Test of y_t

4 Forecasting of Time Sequence based on SARIMA

4.1 Forecasting of Time Sequence based on ARIMA

According to the modeling strategy of Box and Jenkins[6], the time sequence must be stationary when establishing ARIMA(p,d,q) model. In order to transform the sequence y_t to be stationary, difference disposal is applied[7]. After the first-order differnce, the new sequence is named Δy_t, and Figure 3 shows its ACF and PACF. It seems the stationarity gets greatly improved.

Then we can establish ARIMA model for sequence Δy_t. After comparing the different combinations of (p,q), it is when p=3 and q=2 that the ARIMA model has the least MAPE(Mean Absolute Percent Error) value. Therefore, ARIMA(3,1,2) is the optimal model for sequence Δy_t, and the result of forecasting is as shown in Table 2. However, we can see that ARIMA model has poor accuracy in forecasting.

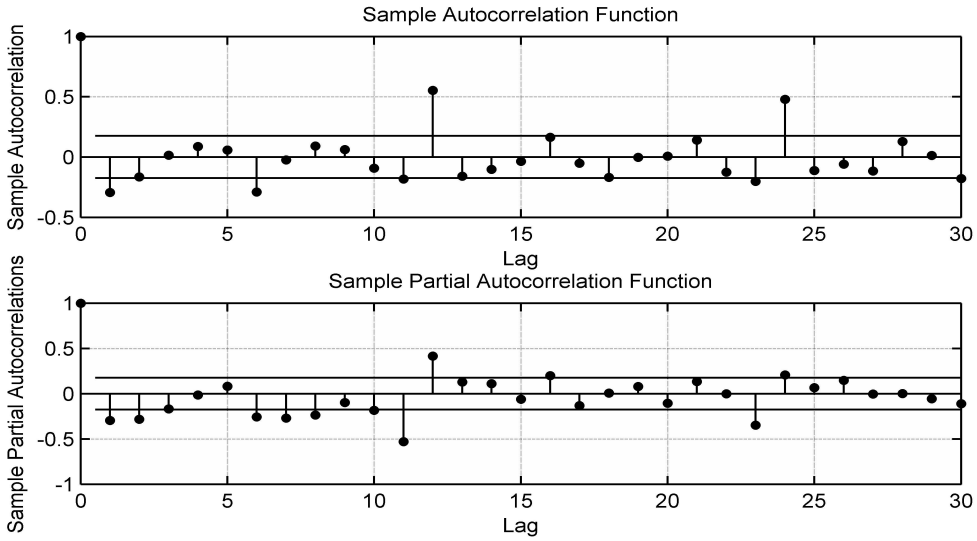

Figure 3: ACF and PACF of Δy_t

In 2014	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Actual values	340	287	282	205	210	293
Forecasting values	382	320	306	231	247	270
Relative errors	9.41%	11.4%	8.51%	12.68%	17.62%	-7.84%
In 2015	Jan.	Feb.	Mar.	Apr.	May.	Jun.
Actual values	207	143	301	399	346	389
Forecasting values	229	124	293	372	311	354
Relative errors	10.63%	-13.29%	-2.66%	-6.76%	-10.11%	-9.00%

Table 2: Result of Forecasting During Jul. 2014 to Jun. 2015 Based on ARIMA(3,1,2) Model

4.2 Forecasting of Time Sequence based on SARIMA

Through observing ACF and PACF figures of Δy_t, we can find when time lags are equal to the multiples of 12, the autocorrelation and partial autocorrelation coefficients are not significantly close to 0, which suggests that seasonal character may exist in Δy_t. In this case, we apply seasonal difference with time lags of 12 to Δy_t, and then get a new sequence named $\Delta\Delta_{12}y_t$. Figure 4 shows the ACF and PACF figures of sequence $\Delta\Delta_{12}y_t$, and Table 3 shows that the ADF value of $\Delta\Delta_{12}y_t$ is below the test critical values in 1% level, both of which determine that $\Delta\Delta_{12}y_t$ is a stationary sequence.

After analysis, there are six models for selection. For determining the optimal model, the MAPE of each model should be compared, and the result is showed in Table 4. Obviously, the optimal model should be SARIMA(0,1,2)(0,1,1)^{12} because its MAPE value is the smallest among all the models. By using maximum likehood estimation method[8], the result of fitting the model is

$$y_t = 2y_{t-1} + y_{t-2} - 4y_{t-13} + 2y_{t-14} - y_{t-24} + 2y_{t-25} - y_{t-26} + \varepsilon_t - 0.569\varepsilon_{t-1} - 0.858\varepsilon_{t-12} + 0.488\varepsilon_{t-13} \quad (4.1)$$

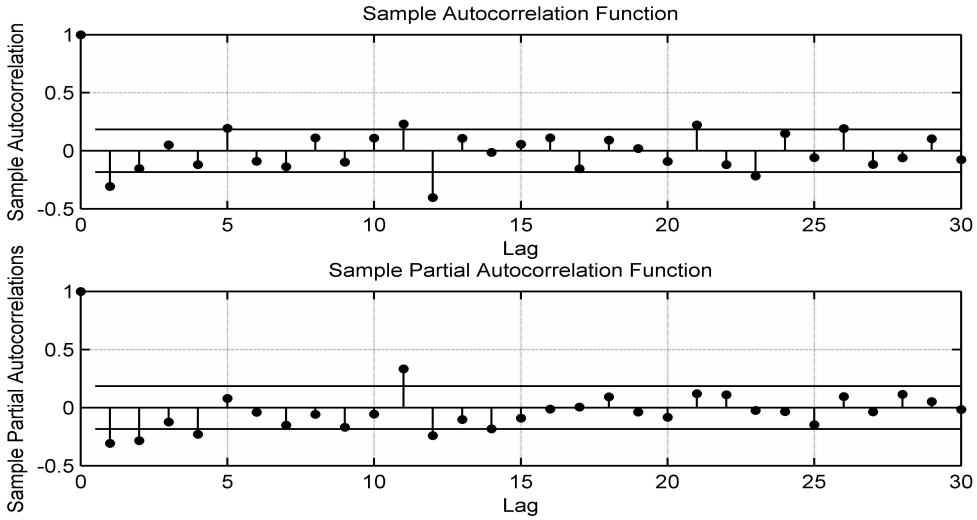

Figure 4: ACF and PACF figure of ΔΔ12 yt

	ADF test statistic	Test critical values		
		1% level	5% level	10% level
t-statistic	-4.213318	-2.585587	-1.943688	-1.61485

Table 3: ADF Test of $\Delta\Delta_{12}\, y_t$

Models	SARIMA(2,1,1)(1,1,1)^{12}	SARIMA(3,1,1)(1,1,1)^{12}	SARIMA(2,1,2)(1,1,1)^{12}
MAPE	9.46	9.98	10.32
2.			
Models	SARIMA(3,1,2)(1,1,1)^{12}	SARIMA(0,1,2)(0,1,1)^{12}	SARIMA(0,1,1)(0,1,1)^{12}
MAPE	11.55	6.45	6.85

Table 4: MAPE Values of Models

If the SARIMA(0,1,2)(0,1,1)^{12} model is utilized to forecast the monthly quantities of commodity barcode registrated in Guangzhou during Jul. 2014 to Jun. 2015, the result is as shown in Table 5.

In 2014	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Actual values	340	287	282	205	210	293
Forecasting values	328	299	272	216	203	288
Relative errors	-3.53%	2.79%	-3.54%	5.37%	-3.30%	-1.71%
In 2015	Jan.	Feb.	Mar.	Apr.	May.	Jun.
Actual values	207	143	301	399	346	389
Forecasting values	220	150	320	381	334	367
Relative errors	6.28%	4.90%	6.31%	-4.51%	-3.47%	-5.66%

Table 5: The Result of Forecasting and Relative Errors During Jul. 2014 to Jun. 2015

As the table above suggests, the relative errors are all below 10%. Compared with ARIMA model, SARIMA$(0,1,1)(0,1,1)^{12}$ model has better performance in forecasting for sequence y_t.

In addition, the relative errors during Jul. 2014 to Dec. 2014 are generally smaller than that during Jan. 2015 to Jun. 2015, which indicates SARIMA model is suitable for shor-term forecasting. With the extension of forecasting time, the accuracy will decline.

If the model is used to estimate the quantities of commodity barcodes registrated in Guangzhou during Jul. 2015 to Dec. 2015, the result is as shown in Table 6. Figure 5 indicates the quantities of commodity barcodes registration will develop in the future, with year-on-year growth of about 10%. However, compared with the growth from 2013 to 2014(shown in Table 7), the growth from 2014 to 2015 will decline obviously. This reveals that the economic growth in China is slowing down.

In 2015	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Forecasting values	383	350	333	263	294	334

Table 6: The Result of Forecasting During Jul. 2015 to Dec. 2015

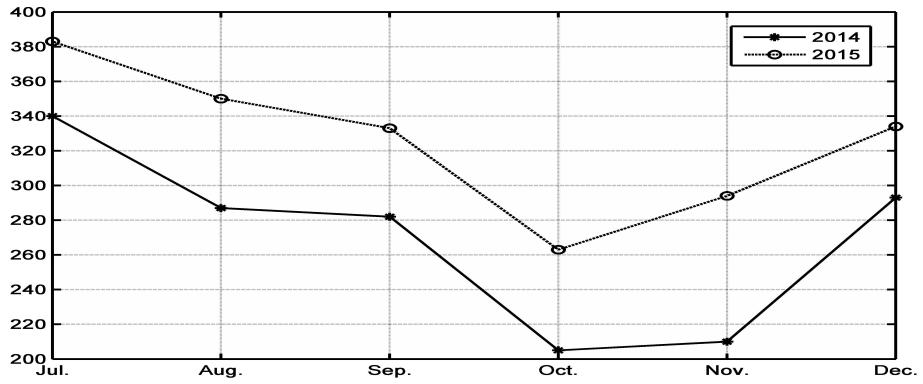

Figure 5: Comparison Between 2014 and 2015

	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
2013	224	223	165	174	164	200
2014	340	287	282	205	210	293
growth	53.57%	28.70%	70.90%	17.82%	28.05%	46.50%

Table 7: Growth From The Second Half of 2013 to The Second Half of 2014

5. Conclusion

Through analyzing the characteristic of the sample, the research establishes SARIMA$(0,1,1)(0,1,1)^{12}$ model for forecasting. The result suggests that the model is practical. Yet, there are some dificiencies on the model. Except the sample data, the other factors data are ignored, which may cauce errors to the forecasting result. Therefore, it needs to carry further research on modeling for enhancing the accuracy of forecasting.

References

[1] S.V. Kumar, L. Vanajakshi. *Short-term traffic flow prediction using seasonal ARIMA model with limited input data*[J]. European Transport Research Review, 7(3): 1-9(2015).

[2] S.G. Gocheva-llieva, A.V. Ivanov, D.S. Voynikova and D.T. Boyadzhiev. *Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach*[J]. Stochastic Environmental Research and Risk Assessment, 28(4): 1045-1060(2014).

[3] H.A. Mombeni, S. Rezaei, S. Nadarajah and M. Emami. *Estimation of Water Demand in Iran Based on SARIMA Models*[J]. Environmental Modeling & Assessment, 18(5): 59-565(2013).

[4] M. Braun, T. Bernard, O. Piller and F. Sedehizade. *24-Hours Demand Forecasting Based on SARIMA and Support Vector Machines*[J]. Procedia Engineering, 89(1): 926-933(2014).

[5] H.S. Kim, D.S. Kang, J.H. Kim. *The BDS statistic and residual test*[J]. Stochastic Environmental Research and Risk Assessment, 17(1): 104-115(2003).

[6] G.E.P. Box, G.M. Jenkins. *Time Series Analysis, Forecasting and Control*[M]. Holden-day, San Francisco, pp:109-130(1970).

[7] F. Barthelemy, M. Lubrano. *Unit roots tests and SARIMA models*[J]. Economics Letters, 50.2: 147-154(1996).

[8] J.H. Lee, D.W. Shin. *Maximum likelihood estimation for arma models in the presence of ARMA errors*[J]. Communications in Statistics - Theory and Methods, 26(5): 1057-1072(1997).