MODELLING AND FORECASTING FOR NONSTATIONARY PROCESS MANAGEMENT

Toir Makhamatkhujaev [1] *and* ***Shokhjakhon Abdufattokhov*** [2]
t.makhamatkhujaev@polito.uz; sh.abdufattohov@polito.uz;
Turin Polytechnic University in Tashkent, Uzbekistan

Abstract. Predicting the evolution of complex systems is noted as one of the ten grand challenges of modern science. Time series data from complex systems capture the dynamic behaviours and causalities of the underlying processes and provide a tractable means to predict and monitor system state evolution. However, most of the time series observed in our life are nonstationary, often exhibiting trends which can be seen in several forms. Such trends are often removed by differencing the data an appropriate number of times, in which case the series is known as an integrated process. Box and Jenkins recommended this approach, and it is widely used in many research areas. In the paper, a sequence of main steps of the Box-Jenkins model is highlighted and demonstrated in a case study of the Real Gross Domestic Product of the US example. Several simple cases of the ARMA model are introduced and analyzed, followed by building and selecting an appropriate model to explain the evolution of an observed time series.

Keywords: Non-stationary processes, linear model, complex systems, random walk.

Introduction

Time series modeling is a dynamic research area which has attracted attentions of researchers community over last few decades. The main aim of time series modeling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the past. Due to the indispensable importance of time series forecasting in numerous practical fields such as business, economics, finance, science and engineering, etc. [1, 2, 3] proper care should be taken to fit an adequate model to the underlying time series. It is obvious that a successful time series forecasting depends on an appropriate model fitting. A significant efforts have been done by researchers over many years for the development of efficient models to improve the forecasting accuracy. As a result, various important time series forecasting models have been evolved in literature. One of the most popular and frequently used stochastic time series

models is the Autoregressive Integrated Moving Average (ARIMA) [2, 4] model. The basic assumption made to implement this model is that the considered time series is linear and follows a particular known statistical distribution, such as the normal distribution. ARIMA model has subclasses of other models, such as the Autoregressive (AR) [4], Moving Average (MA) [4] and Autoregressive Moving Average (ARMA) [4] models. The popularity of the ARIMA model is mainly due to its flexibility to represent several varieties of time series with simplicity as well as the associated Box-Jenkins methodology [2, 4] for optimal model building process. Below, we give brief details of ARMA Box-Jenkins methodology step by step and demonstrate the methodology in case study of Real Gross Domestic Product of US example.

Methodology

An *ARMA(p, q)* model is a combination of *AR(p)* and *MA(q)* models and is suitable for univariate time series modeling. In an *AR(p)* model the future value of a variable is assumed to be a linear combination of p past observations and a random error together with a constant term. Mathematically the *AR(p)* model can be expressed as [4]:

$$y_t = a_0 + \sum_{n=1}^{p} a_n y_{t-n} + \epsilon_t \quad = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + \cdots + a_p y_{t-p} + \epsilon_t \qquad (1)$$

Here y_t and ϵ_t are respectively the actual value and random error (or random shock) at time period t, a_n (n = 1,2, ... , p) are model parameters and a_0 is a constant. The integer constant p is known as the order of the model. Sometimes the constant term is omitted for simplicity. Usually for estimating parameters of an AR process using the given time series, the Yule-Walker equations [4] are used.

Just as an *AR(p)* model regress against past values of the series, an *MA(q)* model uses past errors as the explanatory variables. The *MA(q)* model is given by [4]:

$$y_t = \lambda + \sum_{n=1}^{q} b_n \epsilon_{t-n} + \epsilon_t \quad = \lambda + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + b_3 \epsilon_{t-3} + \cdots + b_q \epsilon_{t-q} + \epsilon_t \qquad (2)$$

Here b_n (n = 1,2, ... , q) are model parameters and λ is a mean of the process. The integer constant q is known as the order of the model. The random shocks are assumed to be a white noise [4] process, i.e. a sequence of independent and identically distributed (i.i.d) random variables with zero mean and a constant variance σ^2. Generally, the random shocks are assumed to follow the typical normal distribution. Thus, conceptually a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations. Autoregressive *(AR)* and moving average *(MA)* models can be effectively combined together to form a general and useful class of time series models, known as the *ARMA* models. Mathematically an *ARMA(p, q)* model is represented as [4]:

$$y_t = \mathrm{c} + \sum_{n=1}^{p} a_n y_{t-n} + \sum_{n=1}^{q} b_n \epsilon_{t-n} + \epsilon_t \text{ with } \mathrm{c} = \lambda + a_0 \qquad (3)$$

Usually ARMA models are manipulated using the lag operator [4] notation. The lag or backshift operator is defined as $Ly_t = y_{t-1}$. Polynomials of lag operator or lag polynomials are used to represent ARMA models as follows:

AR(P) model: $\epsilon_t = a(L)y_t$

MA(Q) model: $y_t = b(L)\epsilon_t$

ARMA(p, q) model: $a(L)y_t = b(L)\epsilon_t$

Here $a(L) = 1 - \sum_{n=1}^{p} a_n L^n$ and $b(L) = 1 + \sum_{n=1}^{q} b_n L^n$

The next issue to our concern is how to select an appropriate model that can produce accurate forecast based on a description of historical pattern in the data and how to determine the optimal model orders. Statisticians George Box and Gwilym Jenkins developed a practical approach to build ARIMA model, which best fit to a given time series and also satisfy the parsimony principle. Their concept has fundamental importance on the area of time series analysis and forecasting [2].

The Box-Jenkins methodology does not assume any particular pattern in the historical data of the series to be forecasted. Rather, it uses a three step iterative approach of model identification, parameter estimation and diagnostic checking to determine the best parsimonious model from a general class of ARIMA models [2]. This three-step process is repeated several times until a satisfactory model is finally selected. Then this model can be used for forecasting future values of the time series. The Box-Jenkins forecast method is schematically shown in Figure 1.

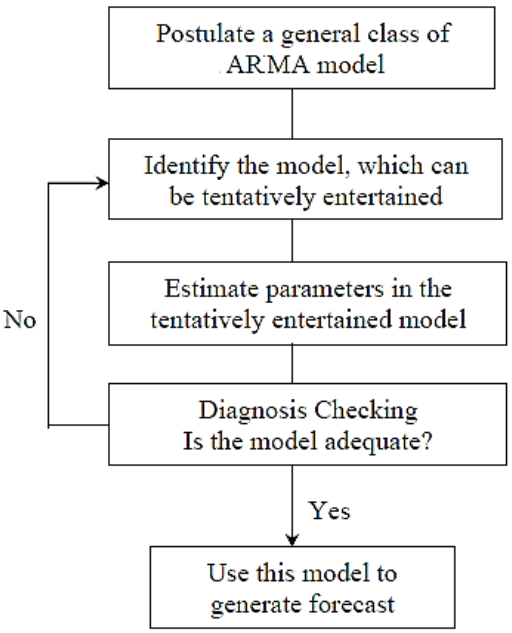


Figure 1. The Box-Jenkins methodology for optimal model selection

A crucial step in an appropriate model selection is the determination of optimal model parameters. One criterion is that the sample ACF and PACF, calculated from the

training data should match with the corresponding theoretical or actual values [4, 5, 6]. Other widely used measures for model identification are Akaike Information Criterion (AIC) [5, 6] and Bayesian Information Criterion (BIC) [5, 6] which are defined below:

$$AIC(p) = n\,log(\sigma_e{}^2/n) + 2p \qquad (4)$$
$$BIC(p) = n\,log(\sigma_e{}^2/n) + p + p\,log(n) \qquad (5)$$

Here n is the number of effective observations, used to fit the model, p is the number of parameters in the model and $\sigma_e{}^2$ is the sum of sample squared residuals. The optimal model order is chosen by the number of model parameters, which minimizes either AIC or BIC. Other similar criteria have also been proposed in literature for optimal model identification.

Case Study

As a case study, we consider the analysis of quarterly U.S. RGDP(Real gross domestic product) from 01.01.1947 to 01.07.2017, with n = 283 observations,

which is not seasonally adjusted. Real gross domestic product is the inflation

adjusted value of the goods and services produced by labor and property located in the United States [7]. Data is taken from [8] is divided into two time frames. From 01.01.1947 to 01.01.2014 will be used as a training data set and the remaining which is from 01.04.2014 to 01.07. 2017 will be used as a test set to validate the forecasting accuracy.

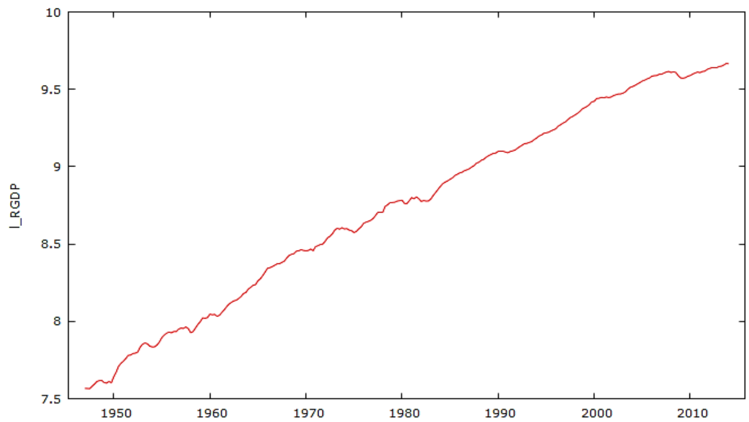


Figure 2. Log of time series of quarterly U.S. RGDP from 1947 to 2014

From Figure 2, we can see that sample time series plot is non-stationary, upward trend. Therefore, we used differencing to make it stationary in order to go further. The first differencing removes the trend from our data and we are able to notice that the variability in the first half of the data is larger than in the second half of the data. Thus, we ended up with the second differencing. From Figure 3, it appears to be covariance stationary with mean zero. We can therefore reject the null hypothesis of non-

stationarity and claim that the series with log second differencing is a realization of stationary process.

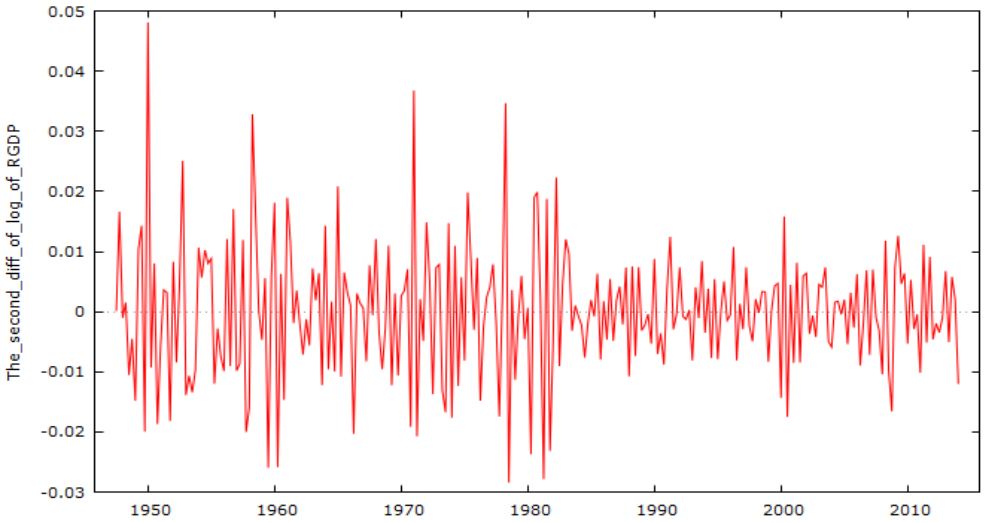


Figure 3. Log of time series of quarterly U.S. RGDP from 1947 to 2014

Table 1. AIC and BIC criterion

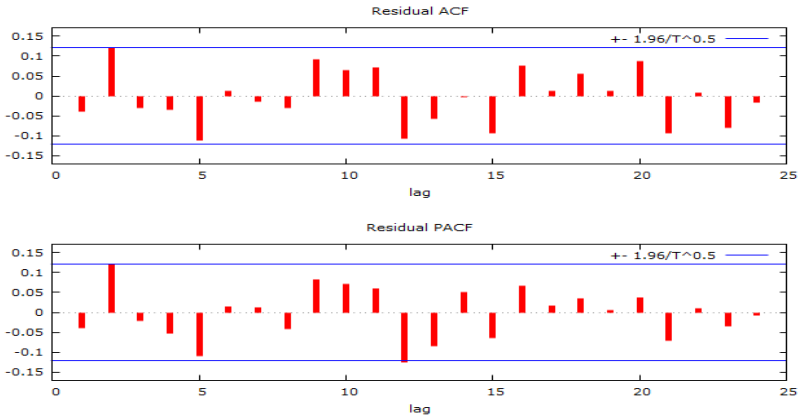
Orders p,q of ARMA model	(0,1)	(1,0)	(1,1)	(2,1)	(1,2)	(2,2)
AIC	-1713.56	-1691.68	-1734.64	-1724.70	-1734.53	-1739.82
BIC	-1702.80	-1684.52	-1720.31	-1706.81	-1716.61	-1718.34

From Table 1, analyzing AIC values it is concluded that the ARMA(2,2) model is most suitable for our time series. From BIC values, however, the ARMA(1,1) is judged to be better suited. The fact that AIC and BIC provided different indications about the best fitting models is not surprising because BIC penalizes larger models more than AIC. Thus, BIC tends to produce more parsimonious best fitting models than AIC. Table 2 shows error analysis of ARMA(1,1) and ARMA(2,2) models. It is seen that ARMA(1,1) model out performs the other models as it has the minimum error.

Table 2. Comparison of errors using the second difference of log of RGDP

	ARMA(1,1)	ARMA(2,2)
RMSE	0.0040678	0.0041388
MAE	0.0029783	0.003115

(a)



(b)

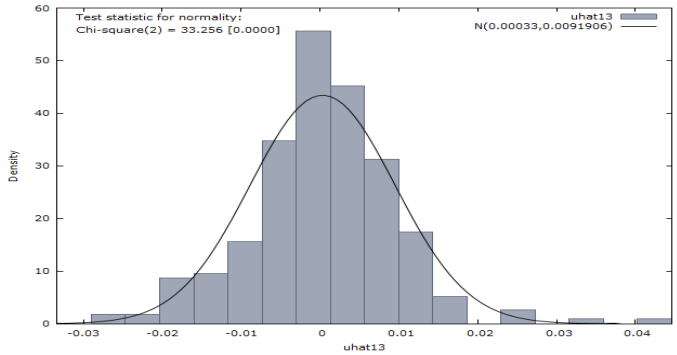


Figure 4. (a) - Residuals of ACF and PACF, *(b)* - Normality plot

From Figure 4, it can clearly be seen that residuals of our model are white noise sequence with mean zero and variance σ^2, also the normality plot illustrate that our residuals are normally distributed, which means our model can be useful. Thus, we accept this model.

After model identification, the next step is to find parameters to fit the data. We estimated the parameters of ARMA(1,1) model using the Maximum Likelihood Estimation for the whole training data set, that is, form 1947 to 2014 and estimated values are $a_1 = 0.382211$ and $b_1 = -0.960320$.

The ARMA(1,1) model has the estimated representation as :

$$y_t = 0.3821\, y_{t-1} - 0.9603\, \epsilon_{t-1} + \epsilon_t \quad \text{with } \epsilon_t \sim WN(0, 0.49) \qquad (6)$$

Using the model in equation (6), we forecast the whole data (training + test) and it is visualized in Figure 5. The output performance are RMSE = 0.012453 and MAE = 0.011904.

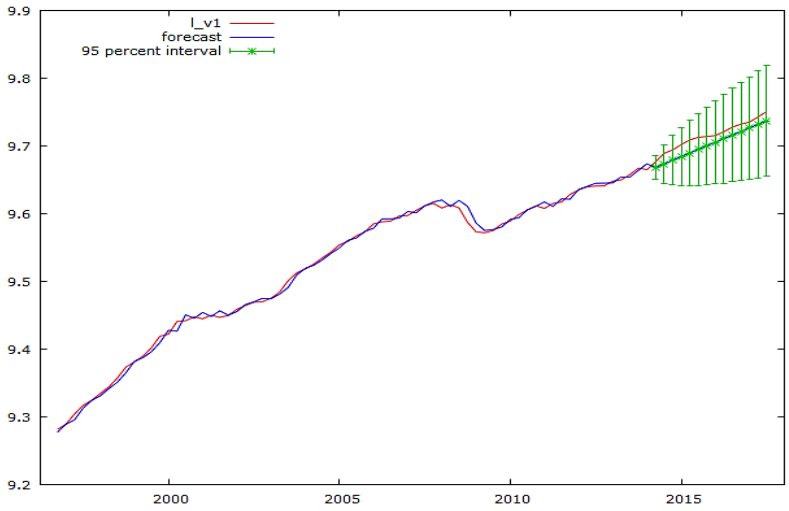


Figure 5. ARMA(1,1) model forecast for log of original RGDP.

Conclusion

In this paper, the Box-Jenkins method highlighted to model the nonstationary time series. We have considered a few important performance measures for evaluating the accuracy of forecasting models. It has been understood that for obtaining a reasonable knowledge about the overall forecasting error, more than one measure should be used in practice. The forecast accuracy was considered using the RMSE and MAE together with AIC and the BIC techniques. Moreover, our satisfactory understanding about the considered forecasting models and their successful implementation can be observed from residual and forecast diagrams. However in some cases, significant deviation can be seen among the original observations and our forecasted values. In such cases, we may choose a suitable data preprocessing, and this will be one of our mission in our future work to improve the forecast performances.

References:

1. G.P. Zhang, “A neural network ensemble method with jittered training data for time series forecasting”, Information Sciences 177 (2007), p. 5329–5346.

2. G.P. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model”, Neurocomputing-50 (2003), p. 159–175.

3. H. Tong, “Threshold Models in Non-Linear Time Series Analysis”, Springer-Verlag, New York, 1983.

4. K.W. Hipel, A.I. McLeod, “Time Series Modelling of Water Resour ces and Environmental Systems”, Amsterdam, Elsevier 1994.

5. J. Faraway, C. Chatfield, “Time series forecasting with neural networks: a comparative study using the airline data”, Applied Statistics 47 (1998), p. 231–250.

6. J.M. Kihoro, R.O. Otieno, C. Wafula, “Seasonal Time Series Forecasting: A Comparative Study of ARIMA and ANN Models”, African Journal of Science and Technology (AJST) Science and Engineering Series Vol. 5, No. 2, p. 41-49.

7. http://www.bea.gov/national/pdf/nipaguid.pdf

8