Proceedings of the International Conference on Industrial Engineering and Operations Management

Manila, Philippines, March 7-9, 2023

A Heuristic Algorithm for Determining the Order of ARIMA Models

Shima Soltanzadeh, Melika Modarres Vahid, and Majid Khedmati[1]
Department of Industrial Engineering
Sharif University of Technology
Tehran, Iran
shima.soltanzadeh@ie.sharif.edu, melika.modarres@ie.sharif.edu,
Khedmati@sharif.edu

Abstract

Autoregressive integrated moving average (ARIMA) models have been proven successful in application and simple in comprehension and consequently, they have been widely applied to different fields in forecasting. The order of an ARIMA model is determined subjectively based on the judgment of the experts where, the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for a given time series are used to determine the potential orders of the model. In this paper, a new heuristic algorithm is proposed for determining the order of ARIMA models. The proposed method determines the order of the ARIMA models, objectively, based on the Mean Squared Error (MSE), Akaike Information Criterion (AIC), and Schwarz Bayesian Information Criterion (BIC). In this regard, the order of the models is determined objectively and as a result, the forecasting results would be more accurate. The performance of the proposed method is evaluated based on a real-world dataset of global temperature anomaly where, the results show that the proposed method performs accurately and efficiently in determining the order of ARIMA models.

Keywords

Box-Jenkins models, Order determination of ARIMA model, Time series forecasting, Heuristic algorithm

1. Introduction

Forecasting plays a crucial role in many fields including finance, business, environmental sciences, medicine and so on. Many different methods such as time series analysis, regression analysis and artificial intelligence can be implemented in forecasting. The chosen method mainly depends on the purpose of the forecasting as well as the accuracy of each method. Time series data are among the most valuable data in the forecasting field. Time series data refers to the chronological sequence of observations on the variable of interest taken at equally spaced points in time. In recent years, different models have been developed to analyze the time series data. These models have different purposes and accuracies including autoregressive integrated moving average (ARIMA) models, regression models, exponential smoothing models, and Holt-Winter models (Gooijer & Hyndman 2006). ARIMA models are one of the most popular and widely used statistical methods for forecasting time-series data. The ARIMA models characterize the existing trend in a sequence and extrapolate the trend to anticipate the future (Montgomery et al. 2015).

In spite of the wide application of ARIMA models, difficulty of constructing an adequate model based on the information provided by finite number of observations remains. Difficulty arises in different stages of constructing the model. It should be noted that building an ARIMA model consist of three stages including model identification, parameter estimation, and diagnostic checking and selection between competing models. There are several means for determining the potential order of ARIMA models. In this process, the order of ARIMA models is determined commonly based on the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. In other words, the researcher investigates the ACF and PACF plots and suggests some potential orders and candidate models. It should be noted that the subjectivity in the first step of model building process can result in more iterations of other steps. In addition, this problem would be intensified when the number of observations is less than required number of observations.

[1] Corresponding Author.

In most of the research efforts performed in the literature, ACF and PACF plots are used to identify the potential models. However, there are many pitfalls including the subjectivity of this approach. It is quite clear that the most difficult part of ARIMA model building is the initial model selection stage. The selection of models can be time-consuming if a large number of time series are to be analyzed. One way of looking at the issue of model selection, which does result in a definite answer, is to use automatic criteria to estimate the order of models (Newbold, 1983). Liu (1989) proposed a filtering method for identifying seasonal ARIMA models when traditional methods, such as autocorrelation function (ACF) and partial autocorrelation function (PACF) methods, do not provide a clear-cut model. Ozaki (1977) used minimum Akaike's Information Criterion estimation (MAICE) to address the difficulty in ARIMA model determination where, MAICE procedure selects a model whose structure and associated parameters produce a minimum AIC.

Hoglund and Ostermark (1991) introduced the cartesian ARIMA (CARIMA) search algorithm. They developed an automatic procedure for modeling time series in the spirit of ARIMA methodology. Hyndman and Khandakar (2008) described the implementation of two automatic univariate forecasting methods in the forecast package for R. Rahkmawti et al. (2019) used TSClust approach to forecast inflation and evaluated accuracy in ARIMA model identification based on model selection criteria. They compared AIC, BIC, AICc (AIC corrected), RMSE and MAPE, and concluded that BIC is the best model selection criteria because it leads to the highest average accuracy. Awe et al. (2020) proposed an alternative algorithm based on the principles of Cartesian products of sets in mathematics for ARIMA model selection. In this paper, a heuristic approach is proposed for identifying the orders of ARIMA models where the subjectivity of the decision-making process by the researcher is eliminated, and since it is an automated process, the order identification is faster than traditional approaches.

The paper is organized as follows. Section 2 describes the ARIMA methodology in details. The proposed approach is described in Section 3. Section 4 provides a numerical example on global temperature anomaly and discusses the result of the proposed method which is used to forecast the given data set. Finally, section 5 presents conclusion and recommendations for future research.

2. ARIMA Model Building

2.1. Autoregressive integrated moving average models

ARIMA models first were introduced by Statisticians George Box and Gwilym Jenkins in 1970's. Hence, they are also known as Box-Jenkins models. ARIMA models assume that the current observation is a linear function of past values and errors. The acronym ARIMA describes the key aspects of the model. In brief, they are:

- Autoregressive (AR). A model that relies on the dependent relationship between a given observation and a given number of lagged observations.
- Integrated (I). The process of moving raw observations through differencing or transformation in order to make the time series stationary.
- Moving average (MA). A model which exploits the relationship between an observation and the residual error from a moving average model that is applied to lagged observations.

All of these components are explicitly mentioned in the model as parameters. These classes of models are denoted as ARIMA (p, d, q), where p is referred to the number of autoregressive terms, d is referred to the number of differencing needed for stationarity, and q is referred to the number of moving average terms. In terms of Y the general forecasting equation is:

$$y_t = \delta + \sum_{i=1}^{p} \varphi_i \, y_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \quad (1)$$

where $t = 1, 2, 3, \dots, n$ denotes the time values, n is the total number of observations in the time series, y_t denotes the value of the time series variable y at time t, φ_i denotes the autoregressive parameters, θ_j denotes the moving average parameters, and ε_t represents the error term at time t.

ARIMA methodology consists of three iterative steps including model identification, parameter estimation, and selection between competing models (Mudelsee 2019).

Before identifying the order of the model, one must reproduce the times series to achieve a stationary time series. Stationarity is a crucial condition in ARIMA modeling, and is satisfied when statistical characteristics such as mean and autocorrelation of observations are almost constant over time. To reproduce a stationary time series different method such as differencing, transformation and a wide range of smoothing methods can be applied (Montgomery et al., 2015). The objective of these operations is to produce a set of stationary residuals.

2.1.1. Model Identification

To determine an ARIMA model, three items should be checked: 1) A time series plot of the data to look for possible trend, seasonality, constant variance or non-constant variance, 2) ACF plot to determine the order of MA model, and 3) PACF plot to determine the order of AR model. ACF describes the dependence structure of a stationary time series. This function determines the correlation between y_t and y_{t-k}. On the other hand, PACF determines the correlation between y_t and y_{t-k} after adjusting for the correlations between the two observations. It is common for time series observations to be correlated with observations at previous time points.

This step of the model identification is the trickiest part because the order will not be easy to be identified. In particular, the order identification of a mixed ARIMA model by observing the behavior of autocorrelations and partial autocorrelations are difficult and impractical. The ACF of an AR(1) process, for example, follows an exponential curve, whereas the ACF of a MA(1) process shows a single peak at $k = 1$. Furthermore, the potential candidates are highly subjective and accordingly, different experts may recognize different candidates. Furthermore, there will be several competing models to be taken into account. Basically, the expert has to guess and continue to next steps.

2.1.2. Parameter Estimation

After determining the potential models, parameters for each model should be estimated. There are several methods for estimating the parameters including least square method, maximum likelihood (ML), conditional least square (CLS), and back casting method (BC) which is also called back forecasting. However, many available software packages implement one or more of the introduced approximation methods (Newbold et al. 1994).

2.1.3. Selecting Between Competing Models

As stated above, in most of the cases, more than one model is identified as potential models for the same data set. The final model is determined by evaluating and comparing the performance of models based on some selection criteria. This step will probably require some type of cross-validation procedure. In this regard, the performance of each model is evaluated from two perspectives, i) how model fits the historical data and ii) how it is successful in forecasting future observations. After comparing the performance of models based on training data, the forecasting accuracy should be evaluated based on test data.

The Akaike's information criterion (AIC) (Akaike 1974), Schwarz Bayesian information criterion (BIC) (Schwarz 1978), and mean square error (MSE) are recommended for evaluating the goodness of fit for potential candidates (Jayawardena, 2020). These criteria are defined as:

$$\mathrm{AIC} = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{2p}{T} \quad (2)$$

$$\mathrm{BIC} = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{p\ \ln(T)}{T} \quad (3)$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}[e_t(1)]^2 \quad (4)$$

where $e_t(1) = y_t - \hat{y}_t(t-1)$ is called one-step-ahead forecast error. The AIC and BIC penalize the sum of squared residuals for including additional parameters in the model. MSE measures the variability in forecast errors. A model with small values of AIC, BIC, and MSE is considered as a good model (Montgomery et al., 2015). It should be also noted that among different models which represent the data equally well, one chooses the most parsimonious one. In other words, due to principle of parsimony, the model which contains the least number of parameters is chosen.

Note that, the adequacy of the model is checked by checking the residuals. In fact, diagnostic testing is conducted to assess whether there are significant autocorrelations among the residuals. If the appropriate model has been chosen, there will be zero autocorrelation in the residuals (Pierce and Box 1970). If the model does not fit the data adequately, one must go back to previous steps and choose a better model.

3. Proposed Method

As discussed in the previous section, in the classical approach, the competing models are selected based on ACF and PACF plots that is quite subjective. In this process, overlooking an optimal model due to human error is common. In other words, it is possible that the best model being overlooked because of poor judgment. Furthermore, when there are many time series to be analyzed, this can cause a difficulty and exhaustion. Accordingly, a new heuristic algorithm is proposed in this paper based on selection criteria addressed in section

2. The proposed algorithm represents an objective perspective on the matter. Instead of observational judgment to determine the potential models, the model is identified based on some selection criteria. In conventional ARIMA modeling, the potential models are chosen and then selection criteria are calculated for each of the candidates in order to find the best possible model. However, in the proposed method, the selection criteria are calculated for all combinations of (p, q). The proposed algorithm consists of three simple steps. In the first step, an upper bound on p and q must be determined. This upper bound should be chosen in such a way that all the potential ARIMA models are considered. In the second step, the AIC and BIC are calculated for all possible combinations of (p, q) whitin the upper bound. Considering the superiority of small values of AIC and BIC, the smallest calculated amounts for AIC and BIC is called (p_1, q_1) and (p_2, q_2), respectively, and accordingly, ARIMA(p_1, d, q_1) and ARIMA (p_2, d, q_2) are considered the two final models from which the final model is chosen. Finally, MSE is calculated for both candidates and the model with the smallest value of MSE is chosen and the related (p, q) is determined as the final order of the ARIMA model. After fitting the chosen model to the given data, residuls must be checked to see whether or not there are specific patterns. In other words, residulas should be checked if they behave like white noise. It should be noted that, the proposed heuristic order selection algorithm is designed to analyze and forecast nonseasonal data sets. The proposed algorithm does not search for the seasonal order of seasonal ARIMA (SARIMA) models where this subject is suggested for future research. The proposed heuristic algorithm is represented in Algorithm 1.

Algorithm1. The proposed heuristic algorithm

The proposed algorithm to determine the orders P and Q:

Step 1: Specify an upper bound on orders **p** and **q**
Step 2: Based on three criteria (AIC, BIC, and MSE):
For **p**: **1 to P** Do
    For **q**: **1 to Q** Do
        Calculate BIC;
        Determine **p** and **q** based on BIC;
        $(\mathbf{p_1}, \mathbf{q_1}) \leftarrow (\mathbf{p}, \mathbf{q})$;
        Calculate AIC;
        Determine **p** and **q** based on AIC;
        $(\mathbf{p_2}, \mathbf{q_2}) \leftarrow (\mathbf{p}, \mathbf{q})$;
    End For
End For
Step 3: Choose between $(\mathbf{p_1}, \mathbf{q_1})$ and $(\mathbf{p_2}, \mathbf{q_2})$ based on MSE;
$(\mathbf{P}, \mathbf{Q}) \leftarrow (\mathbf{p}, \mathbf{q})$.

4. A Case Study

In this section, the performance of the proposed algorithm is evaluated based on a case study. In the past decade, global warming has gained a lot of attention. The researchers and world leaders are both concerned with the effects of this universal phenomenon. As an important key factor in climate impact, estimates of air temperatures have been investigated in many fields including agricultural, ecological, environmental, and industrial fields (Cifuentes et al., 2020; Ye et al., 2013; Mudelsee 2019).

The medians of temperature anomaly from 1850 to 2018 have been used in this paper where the related observations are shown in Figure 1. As discussed previously, the first step is to apply differencing on data to obtain stationary time series. The observations after first-order differencing are shown in Figure 2 where the results show stationarity of the time series.

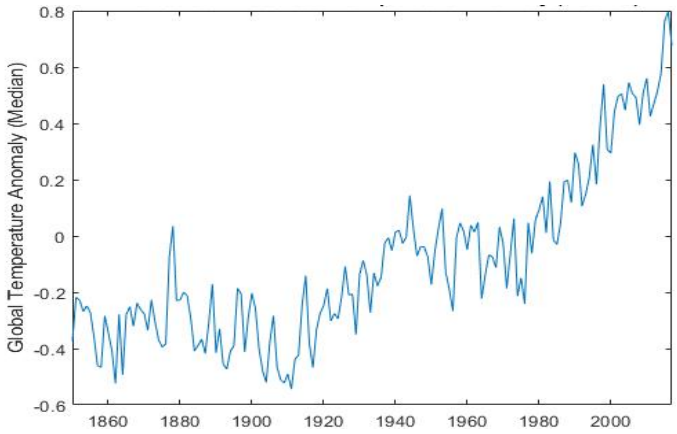

Figure 1. The global temperature anomaly time series

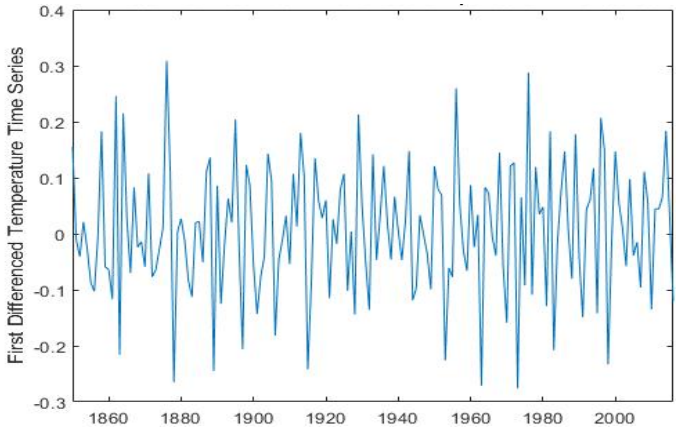

Figure 2. The global temperature anomaly differenced time series

Then, the proposed algorithm is applied on the stationary time series where the final results are shown in Table 1. The results are the same as the results of applying the classic approach using ACF and PACF plots. After comparing the competing models, the optimal model would be ARIMA (3,1,2). Also, the criteria for the optimal model are shown in Table 2.

Table 1. Results of the proposed algorithm

Order	**p**	**q**
	3	2

Table 2. The criteria for optimal model

Criteria for optimal model		
AIC	**BIC**	**MSE**
-281.478	-265.847	0.0120

Finally, the hybrid model proposed by Asadi et al. (2012) is used to estimate the parameters of the model. An in-depth discussion of this method can be found in Asadi et al. (2012). The results are shown in Table 3 and the related model is shown in Equation (5).

Table 3. Estimated parameters

	Estimated parameter
AR (1)	0.366163
AR (2)	-0.040087
AR (3)	-0.012429
MA (1)	0.227594
MA (2)	0.680866
Constant	0.002758

$$y_t = 0.366163y_{t-1} - 0.040087y_{t-2} - 0.012429y_{t-3} - 0.227594\varepsilon_{t-1} - 0.688666\varepsilon_{t-2} + 0.002758 \quad (5)$$

Following the classical approach for order determination of ARIMA models, with regard to the ACF and PACF plots which are shown in Figure 3, different models are nominated. The parameters of the potential models are estimated by Minitab software. Final model is chosen based on the selection criteria. It can be easily shown that, following the classic approach, the same results are obtained.

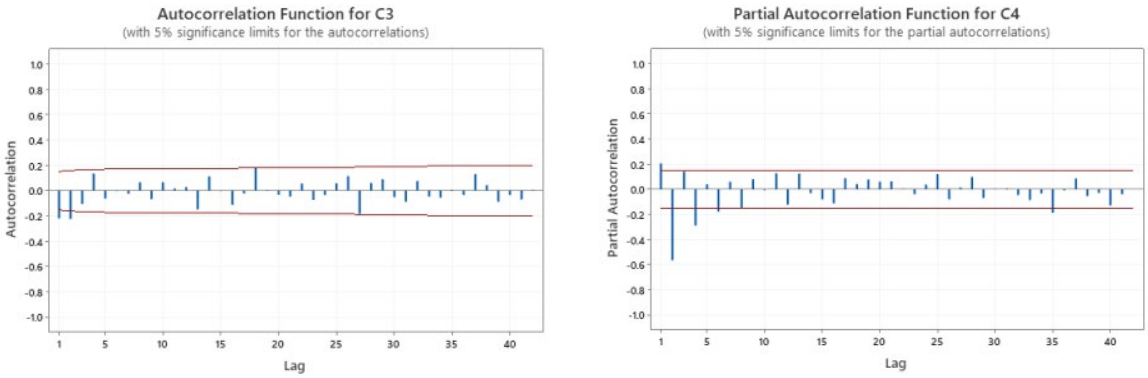

Figure 3. ACF and PACF plot for differenced data set

5. Conclusions

To determine the order of ARIMA model effectively, a large number of historical data is often required. Furthermore, order estimation based on ACF and PACF plots often results in finding many potential models. These competing models can be biased. In this paper, a new heuristic algorithm has been introduced to determine the order of ARIMA models. The proposed method determines the order of the ARIMA model based on statistical selection criteria MSE, AIC and BIC. Also, the proposed method creates more objective results. The proposed algorithm is applied on a case study related to temperature anomaly where the results are promising. It indicates that the proposed method is efficient for determining the order of the ARIMA model.

Future research should consider the potential effects of seasonality more carefully. For example, SARIMA models can be taken into account. In addition, parsimoniousness can be addressed in the proposed algorithm.

References

Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control,* 19(6), 716-723, 1974.

Asadi, S., Tavakoli, A., & Hejazi, S.R. A new hybrid for improvement of auto-regressive integrated moving average models applying particle swarm optimization. *Expert Systems with Applications,* 39(5), 5332-5337, 2012.

Awe, O., Okeyinka, A.E., & Fatokun, J.O. An Alternative Algorithm for ARIMA Model Selection. 2020 International Conference in Mathematics, *Computer Engineering and Computer Science (ICMCECS),* 1-4, 2020.

Box, G.E.P., & Pierce, D.A. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association,* 65(332), 1509-1526, 1970.

Box, G.E.P, Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. Air temperature forecasting using machine learning techniques: a review. *Energies*, 13(16), 4215, 2020.

De Gooijer, J.G., & Hyndman, R.J. 25 years of time series forecasting. *International journal of forecasting*, 22(3), 443-473, 2006.

Höglund, R., & Östermark, R. Automatic ARIMA Modelling by the Cartesian search Algorithm. *Journal of Forecasting*, 10(5), 465-476, 1991.

Hyndman, R.J., & Khandakar, Y. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3), 1-22, 2008.

Jayawardena, A.W. Time series analysis and forecasting. Environmental and Hydrological Systems Modelling, 139-210, 2020.

Liu, L.M. Identification of seasonal arima models using a filtering method. *Communications in Statistics - Theory and Methods*, 18(6), 2279-2288, 1989.

Montgomery, D.C., Jennings, C.L., & Kulahci, M. Introduction to time series analysis and forecasting. John Wiley & Sons, 2015.

Mudelsee, M. Trend analysis of climate time series: A review of methods. *Earth-Science Reviews*, 190, 310-322, 2019.

Newbold, P. ARIMA model building and the time series analysis approach to forecasting. *Journal of Forecasting*, 2(1), 23-35, 1983.

Newbold, P., Agiakloglou, C., & Miller, J. Adventures with ARIMA software. *International Journal of Forecasting*, 10(4), 573-581, 1994.

Ozaki, T. On the order determination of ARIMA models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(3), 290-301, 1977.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, 461-464, 1978.

Ye, L., Yang, G., Van Ranst, E., & Tang, H. Time-series modeling and prediction of global monthly absolute temperature for environmental decision making. *Advances in Atmospheric Sciences*, 30(2), 382-396, 2013.

Author biographies:

Shima Soltanzadeh is a Ph.D. candidate in the Department of Industrial Engineering at the Sharif University of Technology. She earned her B.S. and M.S. in Industrial and Systems Engineering from Amirkabir University of Technology. Her research interests are in stochastic optimization, machine learning applied to healthcare systems engineering, customer behavior, and supply chain management. Her e-mail address is shima.soltanzadeh@ie.sharif.edu.

Melika Modarres Vahid received her M.S. in Industrial Engineering (Engineering Management) from Sharif University of Technology. Her research work focuses on time series analysis and forecasting, interrupted time series analysis, applied statistics, strategic management, and system dynamics. Her e-mail address is melika.modarres@ie.sharif.edu.

Majid Khedmati is Associate Professor of Industrial Engineering at Sharif University of Technology. He received his BSc in Industrial Engineering from Iran University of Science and Technology in 2010, and MSc and PhD degrees both in Industrial Engineering from Sharif University of Technology in 2012 and 2015, respectively. His research interests are in the areas of data science, machine learning, quality engineering, and applied statistics. His research papers have been published in Quality and Reliability Engineering International, Computers and Industrial Engineering, Expert Systems with Application, and Annals of Operations Research, Communications in Statistics-Simulation and Computation. His e-mail address is Khedmati@sharif.edu.