ARIMA model selection in Matlab

Radek Hřebík[1], Jana Sekničková[2]

Abstract. This paper aims to discuss and suggest an approach to analyzing and modelling of economic time series. Econometric theory deals with the problem of right models. As time series analysis methodology is selected the Box-Jenkins methodology representing the stochastic approach. Aim of this paper is to propose an interactive application to enable user not only automated selection of time series model but also to inform the user about everything important that has been done in process of automated selection. This paper deals with Matlab as one of not very typical tool for construction of time series models. Matlab was selected to show and emphasize the power of such tool commonly used at technical universities. The tool gives to user a wide range of possibilities to assess the model. We suggest an application which enables automated model selection and works in two modes, for basic and advanced user. Aim of two user groups is not to discriminate the beginners and not to bore the advanced users. The automated model selection is planned to be divided into two phases. The first phase includes model identification and the second is based on model verification.

Keywords: ARIMA, Box-Jenkins, model selection, Matlab.

JEL Classification: C44
AMS Classification: 90C15

1 Introduction

The time series analysis and modelling represent important process and are very often demanded in various areas of life. The reason is simple because in most cases it is needed to predict future values of time series. The answer to why predict future values is quite clear. Almost everybody wants to know something about the future progress, about the future opportunities. In some fields it may be also the main content to predict future values. As example illustrating this claim can serve the compilation of public budgets. Budget for next period is almost always based on the prediction of main indicators. Econometric theory gives answer to this and the existence of econometric models and their usefulness is nowadays almost doubtless. But as it always is, nothing is fully ideal. There is high probability that the right model exists, but how to choose them? Number of models that are available today is huge. Surely exist the ways how to select the model, but they are not automated and require experience of users. Moreover, the model selection starts already with the selection of an analysis approach.

Current theory offers two main approaches to time series analysis and model. The selection is between two main approaches – the deterministic and stochastic approach. The deterministic approach assumes that all the facts affecting the time series can be explained exactly, for example with some existing models or so on. In case of stochastic approach there is calculated with a random component. As time series analysis methodology used by authors of this paper is selected the Box-Jenkins methodology representing the stochastic approach.

In 1970 Box and Jenkins made autoregressive integrated moving average (ARIMA) models very popular by proposing a model building methodology with three steps – model identification, estimation of parameters and diagnostic checking (Box and Jenkins [1] and Box, Jenkins and MacGrego [2]) and using obtained model for forecasting. Unfortunately, building an ARIMA model is often a difficult task for users because it requires good statistical practice, knowledge in the field of application and very specialized user-friendly software for time series modelling. There exist a lot of freeware or shareware econometric tools helping users to analyze and model time series. The main aim of such tools is to create econometric model specified by user and inform user about the basic parameters of created model. In many cases the user has to make the decision which model to select and what more, the user is responsible for model verification. Because of using this methodology the model selection is focused on suggestion of some approach to select the right model representing the concrete (ARIMA) process.

[1] Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Software Engineering in Economics, Břehová 7, Praha 1, 115 19, Czech Republic, Radek.Hrebik@seznam.cz.
[2] University of Economics, Prague, Faculty of Informatics and Statistics, Department of Econometrics, W. Churchill Sq. 4, 130 67, Praha, Czech Republic, Jana.Seknickova@vse.cz.

Another problem relates with the fact that the number of time series to be analysed is often large and so Box-Jenkins methodology requires both experience and time for successful modelling. There are reasons why we discuss the problem of automated time series analysis. But we are not the first. Also Hoglund and Ostermark in [4] presented study about automatic ARIMA modelling. Unfortunately, authors worked with computer program ANSECH (Mélard [9]) that is not so popular as the others. Tashman and Leach [11] published the article about automatic forecast software but in this work was presented automatic forecasting and not entire modelling process with analysis.

Although the model selection problem is nothing new, it has still no exact solution how to select the right model without human interaction. This paper aims not only to discuss and suggest some solution in this field but aims also to serve as an introductory paper to other possible developers of such application. Application for automated selection discussed in this paper is based on using not typical programming environment to time series analysis. The reason for the aim of this paper is to start a discussion and to inspire the reader to their contribution to solving the automated selection problem. As in most cases it is not possible to select only one model, the offer of more suitable models is also discussed.

What we aim is to propose an interactive application to enable user not only automated selection of time series model but also to inform the user about everything important that has been done. The application aim is also to enable user to suggest his model that was not suggested by our application. The model parameters have to be estimated. In terms of parameter estimation would be used only the implemented methods in existing software (Matlab). It is not the aim of this paper to focus on methods of parameter estimations. To verify the model will be also used methods and functions implemented in used software.

There already exists a lot papers talking about the theory of ARIMA processes, such as Emmenegger [3] or Makridakis and Hibon [8], and about the right models for something like typical or ideal situations but reality seems quite different. In theory is described the ideal process but sometimes it is quite difficult to apply it to the real data. Papers presenting some approaches to these examples from real life which do not fit the model as described in theory are rare. Almost always there is some appeal on analytic decision based on his or her experience. Situation when the data do not fit the model exactly is very common and without human interaction is very hard to select right model to be estimated. The reason of writing paper trying to automate the ARIMA model selection is simple because the question of automated selection is not closed yet.

2 ARIMA and Matlab

In this part of the paper we focused on the selecting criteria for the ARIMA model. As already said the problem of automated model selection is nothing new, but our asset we see in such discussion connected with Matlab and in range of planned use of our application. This paper does not aim to give some genial and cure-all approach. The aim is also to inveigle the reader to his own work and self-evident to show and discuss some example of such approach.

This paper deals with one of not very typical tool for construction of time series models. For model selection, estimation and verification is used high-level computing language Matlab. The reason for using Matlab is implemented suitable econometric tool for parameter estimation helping us in our decisions how to select the right fitting ARIMA model. Matlab is often used for time series modelling, e.g. in Kugiumtzis and Tsimpiris [7] or in Peng and Aston [10].

The reason why we have selected Matlab is to show and emphasize the power of such tool commonly used at technical universities where is often available for students. Such tool gives the user a wide range of possibilities to construct and assess the model. The strength of tool as Matlab and aim of this paper shows one of the reasons why not to use a typical econometric software. Using typical econometric software is nothing bad and in practise it is also seen as a best way. But if the user wants to work more with the constructed models and wants to try some own improvement in the common econometric software is not permitted to do some changes. Here we see the biggest advantage of tool, such as Matlab, to our purposes. The implementation of used functions to evaluate basic characteristics of time series is available to user. So not only to program own functions in Matlab but in many cases is also possible to improve the existing functions. Because there are many of provided functions looking the same as user created functions.

Because the authors are affiliated with two universities it is also at this point possible to discuss and compare Matlab with other software used in special courses focused on time series modelling. In case of Czech Technical University there are two special courses on econometric and applied econometric. In both courses are students working with software to analyse and model time series. As starting program is in this courses used Gretl, but because the students already have experience with Matlab from other courses, they very often use the Matlab instead of Gretl. The example of using Matlab is also included in the course syllabus. When we mention Univer-

sity of Economics in Prague, there are more courses focused on econometric field. But in fact in the courses which are similar in content we have to conclude that the number of students using Matlab is not as high as in case of Czech Technical University in Prague but is increasing in recent years. Using Matlab gives more opportunity to study the problem [5].

Selecting Matlab to analyse time series using the Box-Jenkins methodology is a very pleasant way. It was already mentioned the use of Matlab implemented functions to estimate the model parameters. The reason why we are able to use functions to estimating parameters is the econometrics toolbox. This toolbox provides functions for modeling economic data. User is able not only to select model to be estimated, but also simulate and forecast. Unused in our case of use to automated model selection, but important to mention is, that toolbox provides for example also Monte Carlo methods for simulating systems of linear and nonlinear stochastic differential equations. Of course the toolbox offer a variety of diagnostics for model selection, including hypothesis, unit root (in Matlab: `h = adftest(Y,Name,Value)`), and stationarity tests (in Matlab e.g.: `h = kpsstest(y,Name,Value,...)`).

2.1 Application user modes

Our suggested application to enable automated model selection is working in two modes. We decided to this division because we want to offer an application with possibly the widest range of use. One application user mode is set for basic user and second for advanced user. The aim is to extend the number of possible users and especially not to discriminate or unnecessarily discourage the beginners. In the case of advanced users we do not want to bore them. Our aim is to enable advanced user the intervention into process of building the right time series model.

The main difference between the two types of users is that the basic user is not informed about all the results of partial processes. Because the interactive user is welcome, such user in advanced mode is able to select whether he or she wants to step in at some checkpoints or not. So our suggested application offers menu to advanced user and there it gives the possibility to select the partial processes.

In case of stepping in there will be as output the selected model with the detailed characteristics for the selection. Because our aim is to get some relevant feedback about successfulness of our application in real use the user will not be able to refuse suggested models. The evaluation of suggested models is very important to the future assessment of successfulness. Although the user is not able to refuse suggested model but because the interactive application is he or she able to suggest some own ARIMA model which lacks in the list of suggested models. When user is not interested in this selection application will continue and give all the steps as output at the end.

Because the application is enabling user to input own suggestion of ARIMA model it is very suitable to uncover the behaviour of time series which led to the selection. The identification part includes test of stationarity of time-series (unit root test should be implemented), and also calculation of values of autocorrelation and partial autocorrelation function. The selection phases will be discussed in the next, at this place is only necessary to mention this.

So after one or more models have been selected to be estimated (in Matlab: `fit = estimate(model,Y)`) follows the model verification. It is not needed to show the basic user all models but only models that satisfy our criteria. The reason for our approach to not to show some models is, that models not passing the verification are for basic user useless to be discussed in next. Of course this approach goes with the problem of that no model has reached criteria. This case is in our application taken with special care and there is created some report informing about the state of art. The user of application will be informed about the result that no model was found, but will be able to see what models where suggested and why the verification failed. In case of advanced needs the problem of no verified model no special attention because the user is always informed completely. In this phase it is necessary to check constant standard deviation of random element (ARCH test in Matlab: `[h,pValue,stat,cValue] = archtest(res,Name,Value)`), autocorrelation of random element (Ljung–Box Q test in Matlab: `[h,pValue,stat,cValue] = lbqtest(res,Name,Value)`), normal distribution of random element (Kolmogorov-Smirnov test in Matlab: `h = kstest(x,CDF,alpha,type)`), significance of parameters (Matlab results include *t*-values, *t*-test in Matlab: `h = ttest(...,alpha)`) and more models passing verification tests (Akaike information criterion – AIC, Bayesian information criterion – BIC in Matlab: `[AIC,BIC] = aicbic(LLF,NumParams,NumObs)`).

The aim of both user categories is to collect data that give us the possibility to evaluate the successfulness of application. To completely accomplish the aim of developed application for automated selection is needed to select how to quantify the results of estimated models. The reason of such approach to our software is that we want to have some feedback from advanced users. We are conscious that every case of real data time series

analysis is unique process and so the collecting of data we see as the best way how to improve our application. The data separated into two categories of basic and advanced users enable us also to view the difference between the using the application.

One of possible the possible conclusions can be also that the two groups of user are unnecessary. But we think that such conclusion is not possible. We see the potential benefit of two user groups in the evaluation of successful estimated models. Because we are aware that basic users can be using time series not suitable for Box-Jenkins methodology.

2.2 Main selection phases

The automated model selection in our application we suggest to divide into two phases. As already said according to the theory is the analysis process in case of Box-Jenkins methodology composed from three phases – model specification, parameters estimation and model verification. The reason why we mention only two phases instead of the three known from theory is that we focus on our contribution to phases. This is the reason why the estimation is not presented as separate phase.

The first phase includes model identification. The identification is based on automated approach which is presented in this paper. The second phase is based on model verification and will be also performed by our application in automated way.

To identify the degree of differencing is needed to study the autocorrelation function (ACF). Here is the possible automated approach very significant. It is talked about the unit roots and existence of them can be simply identified from ACF values. To decide about the unit roots existence is needed to have the rule saying when the ACF values are closed to one, the decision is based on statistic approach and the differencing is done when there is more than 90% probability that the first values are closed to one.

In case of differencing there is made back control if the differenced time-series really fulfil the stationary. If the results are in this phase not significant there is implemented also the possibility of second differencing. In case of failure of differencing the application tries to make logarithmic transformation as one of the other options to make time series stationary.

After suggestion how to make time series stationary there is another very import decision to select the autoregressive (AR) and moving average (MA) part of ARMA model. In this case are very important autocorrelation function (ACF) and partial autocorrelation function (PACF) values. The theory gives clear description of ACF and PACF in case of typical AR, MA or mixed ARMA processes. To decide about the autoregressive or moving average part of process is needed to research the values on its statistical significance and compare them with typical models.

Matlab provides implemented functions to evaluate ACF (in Matlab: `autocorr(Series,nLags,M,nSTDs)`) and PACF values (in Matlab: `parcorr(Series,nLags,M,nSTDs)`) there is no need to define special functions to get theses values. The decision how to interpret the values of ACF and PACF is the theme of our research and makes the engine of our proposed application. Our suggestion is based on fitting real time series. The end of this phase is the selection of model prepared to be evaluated and tested. Why do this and not so only to evaluate all common used models, evaluate them and then compare? The reason is simple, because the evaluation of model is much more demanding than the evaluation of the right approach to model selection. With the term right approach we mean the approach which would not be more demanding then evaluation of different models.

There is nothing wrong on selecting more models to be evaluated and at this place is our program coming into second phase of use. In this phase will the models be constructed and consecutively verified. As it is seems this phase connects the estimation, fully done by Matlab, and the verification phase. The assessment of models is very important and the question is what the right criteria?

Verification has three basic steps. Autocorrelation of random element, normal distribution and statistical significant of estimated parameters. The aim is to decide what model has passed the validation criteria. Ideal case of every three criteria fulfilled is rather rare. As default we work with 10% probability that the test has not confirmed our expectation. The user in basic mode is able to change the default value but only to all the verification tests as whole. Advanced user is selecting the probability of failing to each test separately. As the tests are used commonly used and recommended tests. As in case of parameters estimation also the verification tests are made using implemented in Matlab.

The suggested application deals with ARCH test to ensure about constant standard deviation, to exclude the autocorrelation is used Ljung–Box Q test and research of normal distribution procures the Kolmogorov-Smirnov test. The question of significant parameters is solved by Matlab when estimating model parameters because the estimation result includes *t*-values which are interpreted by our application at this point.

If after verification tests there are more models passing verification then is nothing better than apply the criteria specially designed to this situation. The mainly used are Akaike information criterion (AIC) and Bayesian information criterion (BIC). Both these criteria, providing model quality measurement, are implemented in Matlab so there is no need to create own functions for the application proposes.

3 Conclusion

In this paper was discussed the automated time series analysis. Paper is focused on Box-Jenkins methodology representing the stochastic approach, and proposes an approach helping user to select the right time-series model to estimate the future values. The aim of the paper was to discuss and suggest an application enabling user the automated selection of ARIMA model. The application uses Matlab so the phase of estimation parameters is fully done using Matlab implemented functions. The authors aim is automated approach to specification and verification of the model.

Application is designed for two groups of users. The first group are the basic users. To them provides the application very easy way to get acquainted with the time series analysis during the fully automated model selection. In case of advanced users is application more interactive. Advanced users are able not only to suggest own ARIMA models to be estimated and verified but also to set parameters for verification. To the possibilities of advanced users also belongs the detailed overview of specification process. As the authors are aware that proposed application does not mean the end of their work they are ready to improve the application in future.

To be able to the future improvement have the authors included the evaluation of successfulness. The application is collecting data on the results of the use. There are collected only data connected with the results of automated selections. The term collect sounds at first sight not very good, but the aim is not to select any data connected with user or their analysed data, the data to collect means only the data about successfulness of our application. To the goal of selecting data is to quantify the successfulness of our application in real use and consecutively to improve our application based on the results of considerable amount of real data. The assessment will be made separately for basic and advanced users. We find the assessment of collected data as the best way to future improvement of our application. The problem also may be users are using wrong data to be fitted on ARIMA model.

The phase of specification includes the decision about stationary of time-series. There is primary used the 90% probability and attention is paid also to standard deviation. In case of autoregressive and moving average part of model is the decision based on ACF and PACF values. The values for the investigated time series are compared with the values for typical models known from theory.

Model verification is default set to 90% probability of accomplishment criteria. Users in basic mode are able to change the probability only as a one number applied to all the verification tests. Advanced users may determine own probabilities for each verification test. Because it is possible that more models will satisfy the verification test, we have in also mentioned the criteria helping to select one of more suitable models which are also implemented in our application.

Because of the experience with special courses focused on econometric at universities see the authors also the following possibility of use. The proposed application can serve for students at these special courses. Because the application is ready to evaluate the results, there is a great chance to get data from closely aimed and specialized groups of users. So in ideal case can be data collected also at other selected universities. The specialized group of user ensures the relevance of obtained data and the chance to possible improvement of our application is rapidly increasing. The reporting of errors can serve not only to quantifying the application successfulness but also, when used as tool for students, to assess what mistakes students are making.

Acknowledgements

This work was supported by grant no. SGS11/166/OHK4/3T/14 at Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering and also by grant no. IGA F4/1/2012 of Inter grant agency at University of Economics, Prague, Faculty of Informatics and Statistics.

References

[1] Box, G. E. P. and Jenkins, G. M. (1968) 'Some Recent Advances in Forecasting and Control I.', *The Royal Statistical Society Series C-Applied Statistics,* Vol. 17, Issue 2, pp. 91.

[2] Box, G. E. P., Jenkins, G. M. and MacGrego, J. F. (1974) 'Some Recent Advances in Forecasting and Control 2.', *Journal of the Royal Statistical Society Series C-Applied Statistics,* Vol. 23, Issue 2, pp. 158-179.

[3] Emmenegger, J.F. (1996) 'Time and frequency analysis of economic time series', *Zeitschrift fur angewandte Mathematik und Mechanik*, Vol. 76, Issue 3, pp. 417-418.

[4] Hoglund, R. and Ostermark, R. (1991) 'Automatic ARIMA Modeling by the Cartesian Search Algorithm', *Journal of Forecasting*, Vol. 10, Issue 5, pp. 465-476.
[5] Hřebík, R., Sekničková, J.: 'Preference for econometric software in tuition', *Proceedings of the 9th International Conference on Efficiency and Responsibility in Education (ERIE 2012)*, Prague, pp. 230-238.
[6] Hřebík, R., Sekničková, J.: 'Automated selection of appropriate time-series model', *Proceedings of Quantitative Methods in Economy 2012 Conference* , May 30 – June 1, Bratislava (Slovak Republic), 2012
[7] Kugiumtzis, D. and Tsimpiris, A. (2010) 'Measures of Analysis of Time Series (MATS): A MATLAB Toolkit for Computation of Multiple Measures on Time Series Data Bases', *Journal of Statistical Software,* Vol. 33, Issue 5, pp. 1-30.
[8] Makridakis, S. and Hibon, M. (1997) 'ARMA models and the Box Jenkins methodology', *Journal of Forecasting*, Vol. 16, pp. 147–163.
[9] Mélard, G. (1982), 'Software for time series analysis', *Proceedings in Computational Statistics*, Vienna, pp. 336–341.
[10] Peng, J. and Aston, J. A. D. (2011) 'The State Space Models Toolbox for MATLAB', *Journal of Statistical Software*, Vol. 41, Issue 6, pp. 1-26.
[11] Tashman, L. J. and Leach, M. L. (1991) 'Automatic forecast software: a survey and evaluation', *International Journal of Forecasting*, Vol. 7, pp. 209–230.