

## Automatic seasonal auto regressive moving average models and unit root test detection\*

Siana Halim<sup>†</sup>, Indriati N Bisono

Industrial Engineering Department Petra Christian University, Surabaya, Indonesia

(Received December 12 2007, Accepted March 1 2008)

**Abstract.** It is well known that in reality, sequential data more likely exhibit a non-stationary time series or a seasonal non-stationary time series than the stationary one. Therefore, a hypothesis is needed for testing those properties in the time series. Various tests are available in the literature. However in this study, we applied unit root test of Dickey Fuller, Augmented Dickey Fuller and Seasonal Dickey Fuller. We also designed a forecasting program using R 2.6.1. This program executes raw data and gives information of the best time series model in the sense of minimum AIC (Akaike Information Criterion). Using this program, a user who doesn't have a grounded background in time series analysis will be able to forecast short-period future values time series data accurately. The analysis of data consists of Box-Cox transformations, unit root test, removing unit root and seasonal components, finding the best time series model for the data, parameter estimation, models diagnostic checking, and forecasting of the future value time series.

**Keywords:** auto regressive integrated moving average, box-cox transformation, unit root test

### 1 Introduction

Forecasting of future demands is one of the important steps that productions planning control engineers need to do. There are many methods of forecasting, from simple models, such as Holt Winter, Moving Average, to the complicated ones, such as the Box-Jenkins methods. In the market many ready used statistical packages for forecasting, e.g., E-views, ITSM of Brockwell and Davis [6], are available. However, to be able to use one of those packages, a user should have a well grounded background in time series analysis. This prerequisite is rarely fulfilled by most of engineers. Therefore, we developed an automatic forecasting subprogram which executes raw data and gives information of the best proposed time series model in the sense of minimum AIC (Akaike Information Criterion).

In this paper, we present an implementation of an algorithm for automatic forecasting on R. version 2.6.1. R is a free software environment for statistical computing and graphics [13]. The R is available from the Comprehensive R Archive Network (CRAN) at <http://CRAN.R-project.org/>. We also illustrate the use of the subprogram through some examples.

### 2 Models

In modelling the time series data as seasonal autoregressive moving average (SARIMA) models automatically, we need several steps. They are: Box-Cox transformations [4], unit root test, removing unit root and seasonal components, finding the best time series model for the data, parameter estimation, models diagnostic checking, and forecasting of the future value time series.

\* We are very grateful to Melissa and Cyn Thia for helping us in R programming.

<sup>†</sup> Corresponding author. Tel: +62-31-2983425; E-mail address: halim@petra.ac.id, mlindri@petra.ac.id.

## 2.1 The Box - Cox Transformation

The SARIMA models are good for modeling homoscedastic time series, i.e., the variances of the series are constant. Supposing, they are not constant, we should model the series in the heteroscedastic models such as Auto Regressive Conditional Heteroscedastic (ARCH) models [9] or Generalized ARCH (GARCH) models [2] and their derivatives. However, we also can transform the series using Box-Cox transformation. The Box-Cox transformation is a classical technique to stabilize variances. The transformation function is

$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & 0 < \lambda < 1.5 \\ \log(y), & \lambda = 0 \end{cases} \quad (1)$$

where  $y$  is the original data and  $\lambda$  is the chosen parameter. For this version, we occupied the Box-Cox transformation before we modeled the series into SARIMA.

## 2.2 The ARIMA ( $p, d, q$ ) ( $P, D, Q$ )<sup>s</sup>

The Autoregressive Integrated Moving Average, ARIMA( $p, d, q$ )( $P, D, Q$ )<sup>s</sup> is a general model for a non seasonal and non-stationary time series [3]. The model can be written as

$$\phi(B)\Phi^S(1-B)^d(1-B^S)^D y_t = \theta(B)\Theta(B)^S \varepsilon_t \quad (2)$$

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) \quad (3)$$

$$\Phi B^S = (1 - \Phi_1 B^S - \Phi_2 B^{2S} - \dots - \Phi_P B^{PS}) \quad (4)$$

$$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) \quad (5)$$

$$\Theta(B)^S = (1 - \Theta_1(B)^S - \Theta_2(B)^{2S} - \dots - \Theta_Q(B)^{QS}) \quad (6)$$

where,

$p, d, q$ : order of nonseasonal AR, differencing and MA respectively.

$P, D, Q$ : order of seasonal AR, differencing and MA respectively.

$y_t$ : data at period  $t$ .

$\varepsilon_t$ : the independent, identical, normally distributed error (random shock) at period  $t$ .

$B$ : backward shift, where  $B^m y_t = y_{t-m}$ .

$S$ : seasonal order.

## 2.3 The Unit Root Test

A time series is considered to be stationary if the roots of the characteristic equation (e.g. equation (3)) are all greater than unity in absolute value. For the case of a simple AR(1); the characteristic equation is  $1 - \phi B = 0$  that gives a root of  $B = 1/\phi$ . If the root is greater than unity in absolute value, then  $y_t$  is stationary. Thus, the AR(1) model is stationary when  $|\phi| < 1$ . If the root equal to one, we call such root as a unit root. Several tests are available for testing the existence of unit root. Among them, we used the Dickey-Fuller and the Augmented Dickey - Fuller tests.

### 2.3.1 Dickey - Fuller Test

Dickey and Fuller [7] suggested three regression equations to detect the existency of unit roots, i.e.,

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t \quad (7)$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + \varepsilon_t \quad (8)$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + a_1 t + \varepsilon_t \quad (9)$$

The hypotheses of Dickey Fuller (DF) test are

$H_0: \gamma = 1$  (time series is non stationary)

$H_1: \gamma < 1$  (time series is stationary)

The equation (7) is the simplest form of the DF test; in this test we do not encounter the existence of drift and trend on the series. When the underlying data is given by equation (7), but it is not known whether  $y_0$  in the series equals zero, then it is better to allow a constant  $a_0$  to enter the regression model when testing for a unit root, i.e., equation (8). If we let a time trend  $t$  enters, the regression model in equation (9) is employed to test for a unit root. Perron [12] has put forward the sequential testing procedure. If we fail to reject the null using the most general specification, testing continues on down to more restricted specifications. The testing stops as soon as we are able to reject the null hypothesis of a unit root existence.

### 2.3.2 Augmented Dickey-Fuller Test

If a simple AR(1) DF model is used when in fact  $y_t$  follows an AR( $p$ ) process, then the error term will be autocorrelated to compensate for the misspecification of the dynamic structure of  $y_t$ . Thus, assuming  $y_t$  follows an  $p^{th}$  order AR process, we employ the generalized Augmented Dickey Fuller test i.e.,

$$\begin{aligned} y_t &= \Psi_1 y_{t-1} + \Psi_2 y_{t-2} + \dots + \Psi_p y_{t-p} + \varepsilon_t \\ \Delta y_t &= \Psi^* y_{t-1} + \Psi_1 y_{t-1} + \Psi_2 y_{t-2} + \dots + \Psi_p y_{t-p} + \varepsilon_t \end{aligned} \quad (10)$$

where  $\Psi^* = (\Psi_1 + \Psi_2 + \dots + \Psi_p) - 1$  and  $\varepsilon_t \sim \text{IID}(0, \sigma^2)$

The hypotheses of Augmented Dickey Fuller test are

$H_0: \Psi^* = 0$  (there is a unit root)

$H_1: \Psi^* < 0$  (there is no unit root)

Harris and Sollis [10] gave more explanation about the unit root test. The procedure of a unit root test can be seen in Fig. 1.

## 2.4 Diagnostic Check

The diagnostic check is a procedure that used to check the residuals. The residuals should fulfill the models assumptions. They should be independent and Normally distributed. Suppose these assumptions were not fulfilled then we should choose another model for the series. We used the Ljung-Box statistic for testing the independency of the residuals. We also need to do the statistical inferences of the parameters and the goodness of fit of an estimated statistical models.

### 2.4.1 Akaike Information Criterion (AIC)

AIC, proposed in Akaike [1], is a measure of the goodness of fit of an estimated statistical models. The AIC is formulated as

$$AIC = 2k + n + \ln \left( \frac{RSS}{n} \right) \quad (11)$$

where

$k$ : number of parameters or regressors

$n$ : number of data

$RSS$ : residual sum of square.

The preferred model is the one with the lowest AIC value. The AIC methodology attempts to find a model that best explains the data with minimum free parameters.

### 2.4.2 Preliminary Parameter Estimation

Not all parameters in the models are significant. The ratios

$$\left| \frac{\text{Parameter}}{1.96 \times \text{Std error}} \right| > 1 \quad (12)$$

may suggest trying a model in which some of the parameters are set to zero [8]. Then, we need to re-estimate the model after each parameter is set to zero.

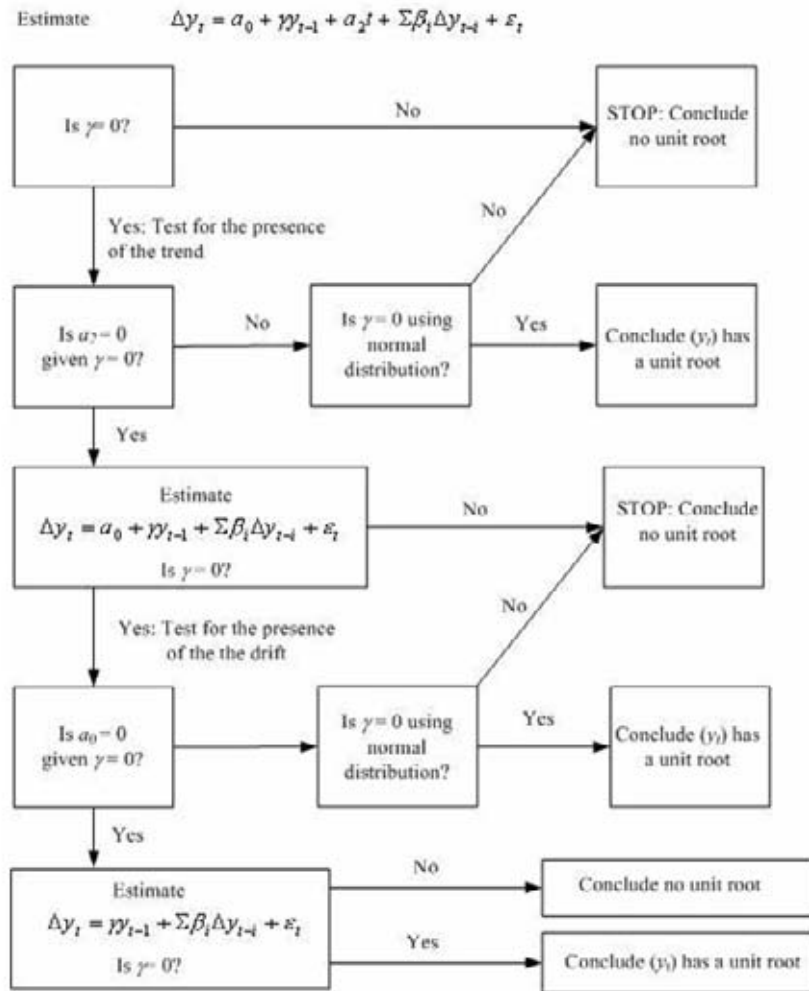


Fig. 1. Procedure of a unit root test [? ].

### 2.4.3 Ljung - Box Statistics

Ljung-Box [11] statistic tests whether a group of autocorrelations of a time series are less than zero. The test statistic is calculated as

$$Q = T(T + 2) \sum_{k=1}^s \frac{r_k^2}{T - k} \tag{13}$$

where

$T$ : number of observations

$s$ : length of coefficients to test autocorrelation

$r_k$ : autocorrelation coefficient (for lag  $k$ )

The hypotheses of Ljung-Box test are

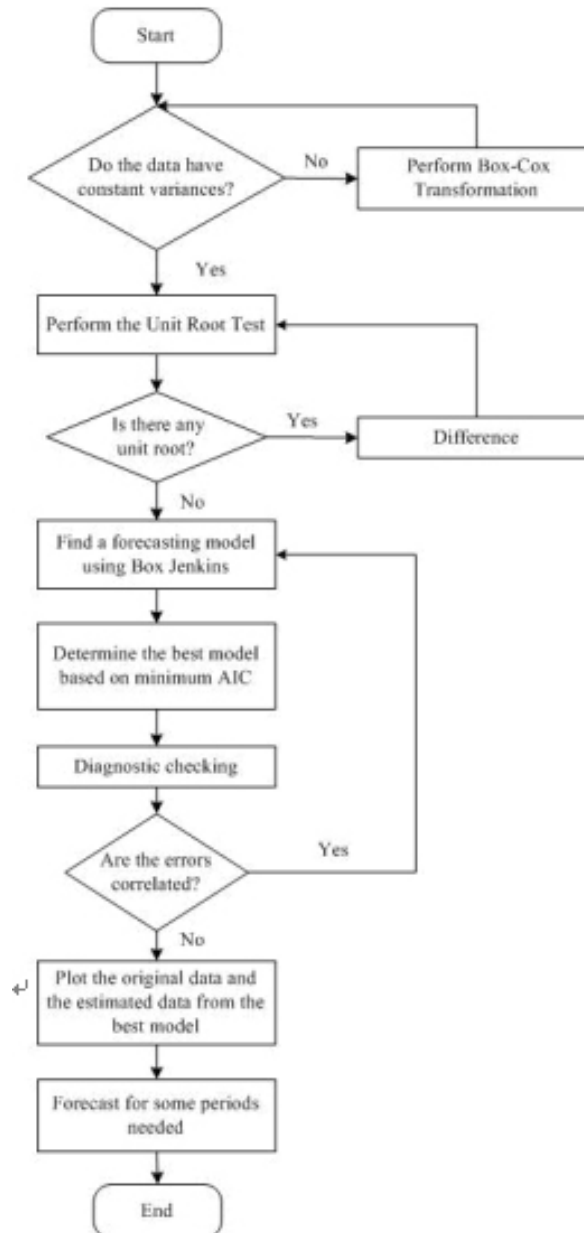
$H_0$ : residual is white noise

$H_1$ : residual is not white noise

If the sample value of  $Q$  exceeds the critical value of a chi-square distribution with  $s$  degrees of freedom then at least one value of  $r$  is statistically different from zero at the specified significance level.

### 3 The Automatic Forecasting Procedure

When data arrive at hand, they can not be modeled immediately. We need some pre-processing, namely stabilizing variances and ensuring data to be stationary, to fulfill the assumptions on the time series modeling.



**Fig. 2.** Flowchart for an automatic forecasting subprogram.

For this first version procedure, we restrict models only for homoscedastic data, i.e., data with constant variances. These are reasonable models for forecasting the market demands on industries. But, the models would not be appropriate for forecasting foreign exchanges, stocks exchanges and others financial commodities which are more volatile.

At the first step we used the Box-Cox Transformation, in equation (1), to stabilize residual variances. We chose a value of parameter  $\lambda$  ( $0 < \lambda < 1.5$ ) several times until the variance of data's plot look likes homogeneous. Now, we can assume that our data have constant variances, but not yet stationary. We precede the data processing with testing the existence of unit roots. The non stationary conditions can occurred in a

series due to the existences of drift, trend, and seasonality. The number of unit roots suggests the differences, i.e. the  $d$  for a non-stationary and non-seasonal model and the  $D$  for a seasonal model in (2), which should be constructed.

By differencing the series, we would have stationary data, so we can use the autocorrelation as well as partial autocorrelation function (PACF) as a tool for detecting the order of the series. We took the largest lag before the cut off in those functions, say lag  $q$  for ACF and lag  $p$  for PACF, modeled the series in  $AR(p)$  or  $MA(q)$ . Then we checked the AIC value and tested the significance of each parameter on the models. Remove the parameter which is the most not significant and then re-estimate the parameters on the model. We stopped modeling when all parameters are significant or when the difference between two consecutives AIC is less than two. We chose one possible model between those two models, i.e.,  $AR(p)$  or  $MA(q)$  and then we compared the best one with the  $ARMA(p, q)$  model. The best of this final comparison is our propose model.

Diagnostic check is our last step in model construction for checking the adequacy of the model to the assumptions. We used the Ljung-Box statistic (13). Once all of the assumptions are fulfilled, we give the plot of original data versus the proposed model. Finally we can use this proposed model to forecast the future data in some short periods. Since, it is well known in time series analysis that forecast values for a long period will converge into the mean of the series.

The flowchart for an automatic forecasting subprogram is given in Fig. 2. We developed this subprogram based on R.6.1. In R, all packages for parameters estimation, unit root tests, forecasting and graphics are available, so we do not need to write our code from scratch. What we have to do is to combine all packages that we need in forecasting, and write the R's code following our algorithm. The users only need to input the data which can be in Excel in the .csv format, executes the subprogram in R, and they will get the proposed model as well as the forecast data.

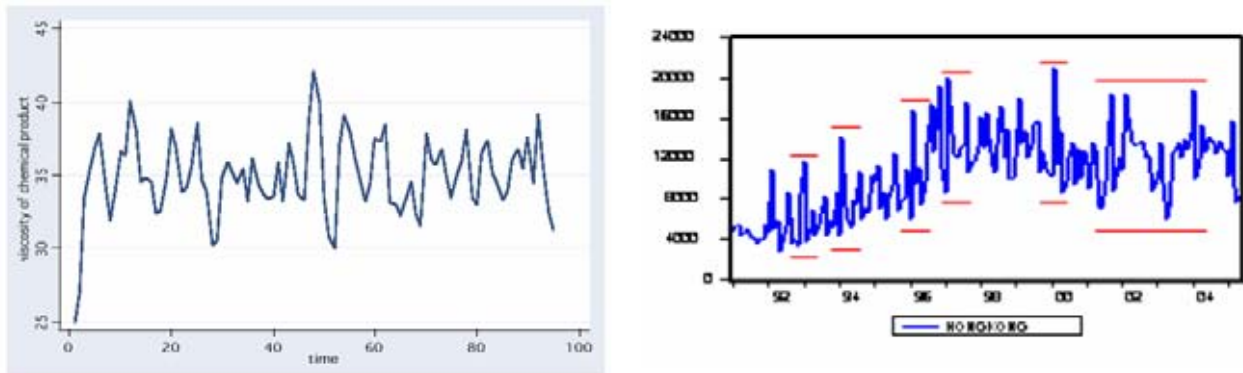


Fig. 3. Data with constant variances

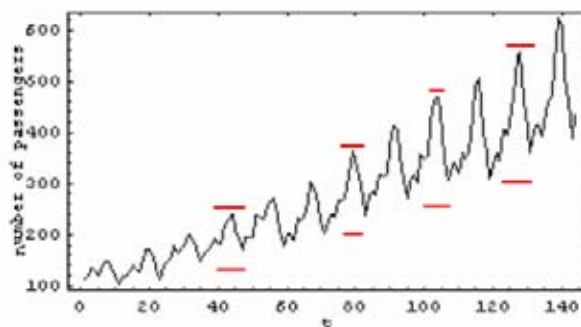


Fig. 4. Data with nonconstant variances

The performance of this procedure is as good as if an expert of time series analysis models a homoscedastic series and if he or she chose the AIC as his or her goodness of fit test. Another advantage of this procedure is that there will be no cost for buying and maintaining the software, since it is developed under R. However, there is a lack in this version, that is, it needs visually detection from users in the first step to determine the homogeneous of the data. Then, the users need nothing to do except waiting for the result. It took around four minutes to perform the result under a PC - Pentium (R) 4, 2.40 GHz, and 256MB of RAM.

#### 4 The Test Cases

As we stated above, that users should decide whether the series have constant variance or not. As this subprogram is dedicated for practitioners who are less familiar with statistics background, we plot the data and train the users to recognize the appearance of nonconstant variances in the data through some examples (see Fig. 3 and Fig. 4).

To show the performance of the proposed procedure, we used several test cases to test the performance of our subprogram. First, we used the viscosity of chemical XB 77-5 data ([3]: pp. 471). The output of the subprogram is

```
The data is stationary, doesn't consist of trend and drift.
The possible order of MA(q) are 0 1 3 4
There is no possibility of seasonal MA(Q)
The possible order of AR(p) are 1 2
There is no possibility of seasonal AR(P)

Based on the AIC, the proposed model is AR(2) with AR seasonal 0-order
with AIC = 422.0743

Call:
arima(x = x.trans, order = c(p.opt.d1,0), seasonal = list(order =
c(P.opt, D1, 0)))

Coefficient:
      ar1      ar2      intercept
      0.6821    -0.4333     0.0163
s.e      0.0980     0.1037     0.2935

sigma^2 estimated as 4.544:log likelihood =
- 207.04, AIC = 422.07
```

The output of this subprogram confirmed that the data is stationary and do not consist of trend and drift. As there are no seasonal and trend, the subprogram develop the possible MA and AR models, then find the best model based on the minimum AIC that is AR(2). This should be a reasonable model since the subprogram checked all possible model and choose the one with minimum AIC. The procedure also gives the forecast data as we needed, plots of the data and the forecast. The point forecasts and 95% prediction intervals are as follows, see Tab. 4:

Obs	Forecast	Lower Bound	Upper bound
96	33.55752	29.29437	37.82066
97	35.58574	30.42532	40.74616
98	35.98431	30.82209	41.14652

The second case is the demand of Super Tech Videocassette Tape data ([3]: pp. 504), showing the output of the subprogram if the data is non-stationary The output of the subprogram is:

The data are nonseasonal  
 The data are non-stationary and consist of drift.  
 The data can be stationaryized with difference 1

The possible order of MA(q) are 0 1 5 6  
 There is no possibility of seasonal MA(Q)  
 The possible order of AR(p) are 1 5 6  
 There is no possibility of seasonal AR(P) with Based on the AIC, the proposed model is AR(6) with AR seasonal 0- order with AIC = 798.6504

Call:  
 arima(x = x.trans, order = c(p.opt.dl,0), seasonal = list(order = c(P.opt, D1,0)))

Coefficient:

	ar1	ar2	ar3
	0.2756	-0.1008	0.0636
s.e	-0.0779	0.0794	0.0817
	ar4	ar5	ar6
	0.0431	-0.1930	-0.1828
s.e	0.0814	0.0822	0.0816

sigma^2 estimated as 7.862:log likelihood =  
 - 392.33.04, AIC = 798.6504

In this case, it is shown that the data are nonseasonal but they are non-stationary and consist of drift. There are three possibilities models, the best one in the sense of minimum AIC is AR(6). The plot of data as well as its data forecast is depicted in Fig. 5.

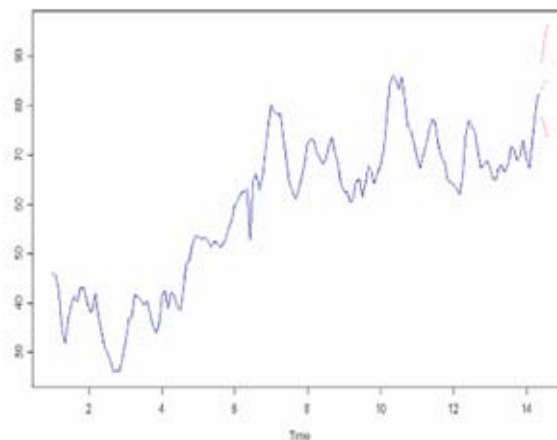


Fig. 5. Plot of the demand of super tech video cassette tape data and the forecast.

The point forecasts and 95% prediction intervals are as follows, see Tab. 4:

## 5 Conclusion and Remark

In this paper we presented the flowchart for an automatic seasonal non-stationary homogenous forecasting. This subprogram executes raw data and gives information of the best proposed time series model in the sense of minimum AIC (Akaike Information Criterion). We also presented two cases and its solution using

Obs	Forecast	Lower Bound	Upper bound
162	83.22874	77.62084	88.83665
163	84.79393	75.70423	93.88362
164	84.86081	73.37939	96.34223

this automatic subprogram. Another criterions, such as, Bayesian Information Criterion (BIC), maximum likelihood, can also be used as a measurement for choosing a best model. Later on, this subprogram should get improvement for handling heteroscedastic data.

## References

- [1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, **19**(6): 716–723.
- [2] T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 1986, **31**: 307–327.
- [3] L. Bowerman, R. O’Connell. *Forecasting And Time Series: An Applied Approach*, 3rd edn. Duxbury Press, California, 1993.
- [4] G. Box, D. Cox. An analysis of transformations. *Journal of Royal Statistical Society*, 1964, **26**: 211–246. Series B.
- [5] G. Box, G. Jenkins. *Time Series Analysis forecasting and control*. Holden-Day, California, 1976.
- [6] P. Brockwell, R. Davis. *Introduction to Time Series and Forecasting*, 2nd edn. Springer-Verlag, New York, 2002.
- [7] D. Dickey, W. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 1979, **74**: 427–431.
- [8] W. Enders. *Applied Econometrics Time Series*, 2nd edn. Wiley, United States of America, 2003.
- [9] R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 1982, **50**: 987 – 1008.
- [10] R. Harris, R. Sollis. *Applied Time Series Modelling and Forecasting*. John Wiley & Sons, England, 2003.
- [11] G. Ljung, G. Box. On a measure of lack of fit in time series models. *Biometrika*, 1978, **65**: 553–564.
- [12] P. Perron. Trends and random walks in macroeconomic time series: Further evidence from a new approach. *Journal of Economic Dynamics and Control*, 1988, **12**: 297–332.
- [13] R. D. C. Team. R: A language and environment for statistical computing. **in:** *R Foundation for Statistical Computing*, Vienna, Austria, 2007. ISBN 3-900051-07-0, URL <http://www.R-project.org/>, 2007.