Identification of time series model of heat demand using Mathematica environment

BRONISLAV CHRAMCOV
Tomas Bata University in Zlin
Faculty of Applied Informatics
nam. T.G. Masaryka 5555, 760 01 Zlin,
CZECH REPUBLIC
chramcov@fai.utb.cz http://web.fai.utb.cz

Abstract: - The paper presents possibility of model design of time series of heat demand course. The course of heat demand and heat consumption can be demonstrated by means of heat demand diagrams. The most important one is the Daily Diagram of Heat Supply (DDHS) which demonstrates the course of requisite heat output during the day. These diagrams are of essential importance for technical and economic considerations. Therefore forecast of the diagrams course is significant for short-term and long-term planning of heat production. The aim of paper is to give some background about analysis of time series of heat demand and identification of forecast model using Time series package which is integrated with the Wolfram Mathematica. The paper illustrates using this package for forecast model identification of heat demand in specific locality. This analysis is utilized for building up a prediction model of the DDHS

Key-Words: - Prediction, District Heating Control, Box-Jenkins, Control algorithms, Time series analysis, modelling

1 Introduction

Analysis of data ordered by the time the data were collected (usually spaced at equal intervals), called a time series. Common examples of a time series are daily temperature measurements, monthly sales, daily heat consumption and yearly population figures. The goals of time series analysis are to describe the process generating the data, and to forecast future values.

Forecasting can be an important part of a process control system. By monitoring key process variables and using them to predict the future behavior of the process, it may be possible to determine the optimal time and extent of control action. We can find applications of this prediction also in the control of the Centralized Heat Supply System (CHSS), especially for the control of hot water piping heat output. Knowledge of heat demand is the base for input data for the operation preparation of CHSS. The term "heat demand" means an instantaneous heat output demanded or instantaneous heat output consumed by consumers. The term "heat demand" relates to the term "heat consumption". It expresses heat energy which is supplied to the customer in a specific time interval (generally a day or a year). The course of heat demand and heat consumption can be demonstrated by means of heat demand diagrams. The most important one is the Daily Diagram of Heat Demand (DDHD) which demonstrates the course of requisite heat output during the day (See Fig. 1). These heat demand diagrams are of essential importance for technical and economic considerations. Therefore forecast of these diagrams course is significant for short-term and long-term planning of heat production. It is possible to judge the question of peak sources and particularly the question of optimal load distribution between the cooperative production sources and production units inside these sources according to the time course of heat demand [1]. The forecast of DDHD is used in this case.

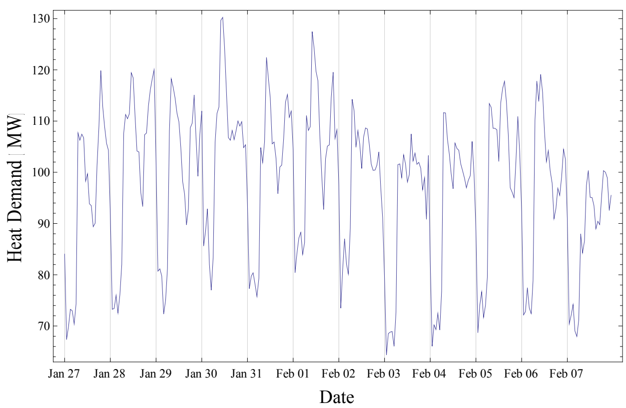

Fig. 1: DDHD for the concrete locality

In the other work [2], a model for operational optimization of the CHSS in the metropolitan area is presented by incorporating forecast for demand from customers. In the model, production and demand of heat in the region of Suseo near Seoul, Korea, are taken into account as well as forecast for demand using the artificial neural network.

In this paper we propose the forecast model of DDHD based on the Box-Jenkins [3] approach. This method works with a fixed number of values which are updated for each sampling period. This methodology is based on the correlation analysis of time series and works with stochastic models which enable to give a true picture of trend components and also that of periodic components. As this method achieves very good results in practice, it was chosen for the calculation of DDHD forecast.

Identification of time series model parameters is the most important and the most difficult phase in the time series analysis. This paper is dealing with the identification of a model of concrete time series of heat demand. We have particularly focused on preparing data for modeling as well as on estimating the model parameters and diagnostic checking.

Currently, there is a wide range of free and commercial products, which offer an extremely wide spectrum of possibilities for the time series analysis and subsequent forecast of these time series. Our workplace is equipped with a Mathematica environment, which is used for education and academic research. Mathematica environment is the product of the Wolfram Research company [4], and is one of the world's most powerful global computation system. Time Series is a package of Wolfram Mathematica [5]. It is a fully integrated environment for time-dependent data analysis. Time Series performs univariate and multivariate analysis and enables you to explore both stationary and nonstationary models. It is possible to fit data and obtain estimates of the model's parameters and check its validity using residuals and tests such as the portmanteau.

2 Preparing real data for modeling

In order to fit a time series model to data, we often need to first transform the data to render them "wellbehaved". By this we mean that the transformed data can be modeled by a zero-mean, stationary type of process. We can usually decide if a particular time series is stationary by looking at its time plot. Intuitively, a time series "looks" stationary if the time plot of the series appears "similar" at different points along the time axis. Any nonconstant mean or variability should be removed before modeling. The way of transforming the raw data of heat demand into a form suitable for modeling are presented in this section. Time series package enables to use many transformations. These transformations include linear filtering, simple exponential smoothing, differencing, moving average, the Box-Cox transformation and others. Only differencing is considered for time series analysis of heat demand.

2.1 Differencing the time series of heat demand

The graph [see Fig.1] shows that the values of heat demand signalling a possible nonconstant mean. Therefore it is necessary to use a special class of nonstationary ARMA processes called the autoregressive integrated moving average (ARIMA) process. Equation (1) defines this process with order *p, d, q* or simply ARIMA(*p,d,q*).

$$(1-B)^d \phi(B) z_t = \theta(B)\varepsilon_t \qquad (1)$$

Non-negative integer d is degree of differencing the time series, p represents the order of autoregressive process and q is order of moving average process. $\phi(B)$ and $\theta(B)$ are polynomials of degrees p and q. ARIMA(*p,d,q*) series can be transformed into an ARMA(*p,q*) series by differencing it d times. Using the definition of backward shift operator B it is possible to define differencing the time series z_t for $d=1$ in the form (2).

$$\nabla z_t = (1-B) = z_t - z_{t-1} \qquad (2)$$

Determination of a degree of differencing d is the main problem of ARIMA model building. Differencing is an effective way to render the series stationary. In Time series package it is possible to use the function `ListDifference[data,d]` to difference the data d times

In practice, it seldom appears necessary to difference more than twice. That means that stationary time series are produced by means of the first or second differencing. A number of possibilities for determination of difference degree exist. It is possible to use a plot of the time series, for visual inspection of its stationarity. In case of doubts, the plot of the first or second differencing of time series is drawn. Then we review stationarity of these series. Investigation of sample autocorrelation function (ACF) of time series is a more objective method. If the values of ACF have a gentle linear decline (not rapid geometric decline), an

ISBN: 978-1-61804-004-6

autoregressive zero is approaching 1 and it is necessary to difference. The work [6] prefers to use the behaviour of the variances of successive differenced series as a criterion for taking a decision on the difference degree required. The difference degree d is given in accordance with the minimum values of variance $\sigma_z^2, \sigma_{\nabla z}^2, \sigma_{\nabla^2 z}^2 \ldots$.

Sometimes there can be seasonal components in a time series. These series exhibit periodic behaviour with a period s. For these time series a multiplicative seasonal ARMA model of seasonal period s and of seasonal orders P and Q and regular orders p and q is defined in the form (3).

$$(1-B)^d (1-B^s)^D \phi(B) \Phi(B^s) z_t = \theta(B) \Theta(B^s) \varepsilon_t \qquad (3)$$

$\Phi(B)$ and $\Theta(B)$ are polynomials of degrees $P*s$ and $Q*s$. Model in the form (3) is referred to as SARIMA $(p,d,q)\times(P,D,Q)_s$.

Firstly it is necessary to determine a degree of seasonal differencing - D. In seasonal models, necessity of differencing more than once occurs very seldom. That means D=0 or D=1. The first seasonal differencing with period s is defined in the form (4).

$$\nabla_s^D z_t = (1-B^s) z_t = z_t - z_{t-s} \qquad (4)$$

It is possible to decide on the degree of seasonal differencing on the basis of investigation of sample ACF. If the values of ACF at lags $k*s$ achieve the local maximum, it is necessary to make the first seasonal differencing (D=1) in the form $\nabla_s z_t$. In Time series package it is possible to difference the data d times with period 1 and D times with the seasonal period s and obtain data in the form (5) using function `ListDifference[data,{d,D},s]`.

$$\nabla^d \nabla_s^D z_t = (1-B)^d (1-B^s)^D z_t \qquad (5)$$

An example of the determination of the difference degree for our time series of heat demand is shown in this part of paper. The course of time series of heat demand [see Fig.1] exhibits an evident non-stationarity and also seasonality. It is necessary to difference. We use the course of sample ACF and values of estimated variance of differenced series for determination of degree of regular differencing and degree of seasonal differencing. The clear periodic structure in the time plot of the heat demand is reflected in the correlation plot [see the Fig.2]. The pronounced peaks at lags that are multiples of 24 indicate that the series has a seasonal period $s=24$. That represents a seasonal period of 24 hours by a sampling period of 1 hour. We also observe that the ACF decreases rather slowly to zero. This suggests that the series may be nonstationary and both seasonal and regular differencing may be required.

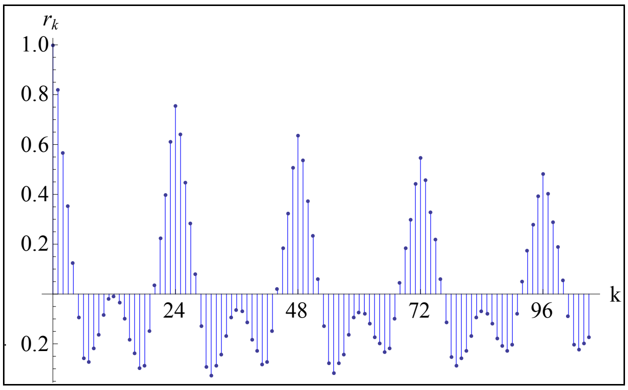

Fig. 2: The course of sample ACF of time series of heat demand

The course of first regular differenced time series is shown in Fig. 3. The differenced series looks stationary now. This fact is confirmed by the sample ACF of differenced data (see Fig.4).

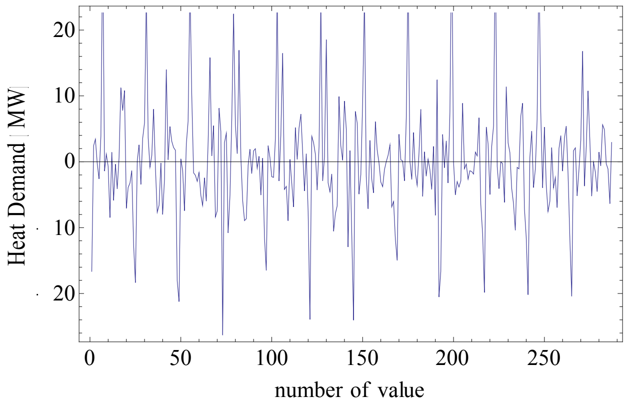

Fig. 3: The course of first regular differenced time series of heat demand

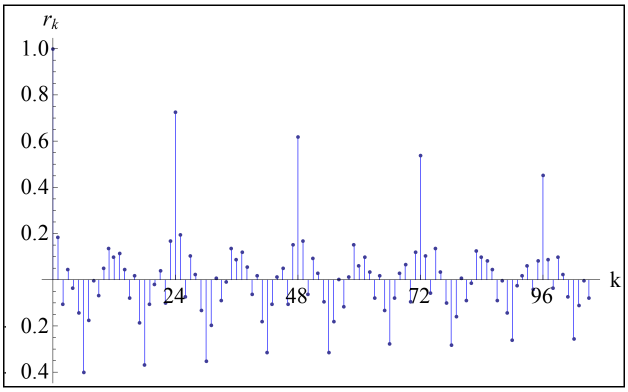

Fig. 4: The course of sample ACF of regular differenced time series of heat demand

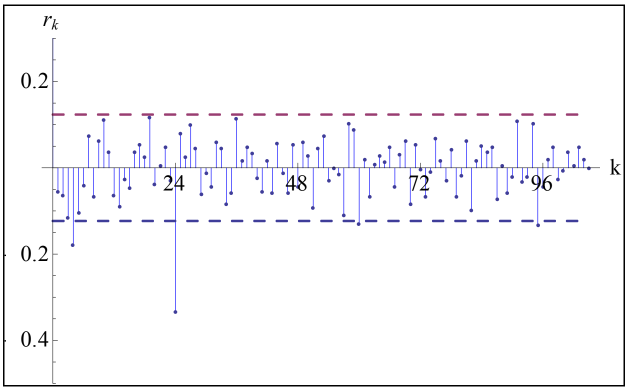

Fig. 5: The sample ACF of data after regular and seasonal differencing with period 24

The course of sample ACF for data after regular and seasonal differencing with period 24 also confirms seasonal period $s=24$ [see Fig. 5]. On the basis of the executed analysis, it is necessary to make the first differencing and also the first seasonal differencing with period 24 of time series of the heat demand in the form (6).

$$\nabla\nabla_{24} z_t = (1-B)(1-B^{24})\ z_t = \\ = z_t - z_{t-1} - z_{t-24} - z_{t-25} \qquad (6)$$

For the comparison, it is possible to calculate the variance of time series and differenced series according to [6]. The results are presented in the Tab. 1. These results confirm transforming the time series of heat demand by differencing in the form (6). These differenced data are prepared for the next modeling and forecasting.

Table 1: The values of variance of differenced series

Raw data of heat demand	$\hat{\sigma}_z^2 = 221.649$	
Regular differencing	$\hat{\sigma}_{\nabla z}^2 = 80.5422$	$\hat{\sigma}_{\nabla^2 z}^2 = 131.093$
Seasonal differencing	$\hat{\sigma}_{\nabla\nabla_{24} z}^2 = 35.4418$	

3 Selecting the orders of model

After differencing the time series, we have to identify the order of autoregressive process AR*(p)* and order of moving average process MA*(q)* and seasonal orders *P* and *Q*. There are various methods and criteria for selecting the orders of an ARMA or an SARIMA model. The sample ACF and the sample partial correlation function (PACF) can provide powerful tools to this. The traditional method consists in comparing the observed patterns of the sample autocorrelation and partial autocorrelation functions with the theoretical autocorrelation and partial autocorrelation function patterns. These theoretical patterns are shown in Tab 2.

Table 2: Behaviour of theoretical autocorrelation and partial autocorrelation function

Model	ACF	PACF
AR(p)	Tails off	Cuts off after *p*
MA(q)	Cuts off after *q*	Tails off
ARMA(p,q)	Tails off	Tails off

The expression *Tails off* in Table 1 means that the function decreases in an exponential, sinusoidal or geometric fashion, approximately, with a relatively large number of nonzero values. Conversely, *Cuts off* implies that the function truncates abruptly with only a very few nonzero values. In the case of SARIMA model, the *Cuts off* in the sample correlation or partial correlation function can suggest possible values of $q + s*Q$ or $p + s*P$. From this it is possible to select the orders of regular and seasonal parts. The standard errors of the ACF and PACF samples are useful in identifying nonzero values. As a general rule, we would assume an autocorrelation or partial autocorrelation coefficient to be zero if the absolute value of its estimate is less than twice its standard error.

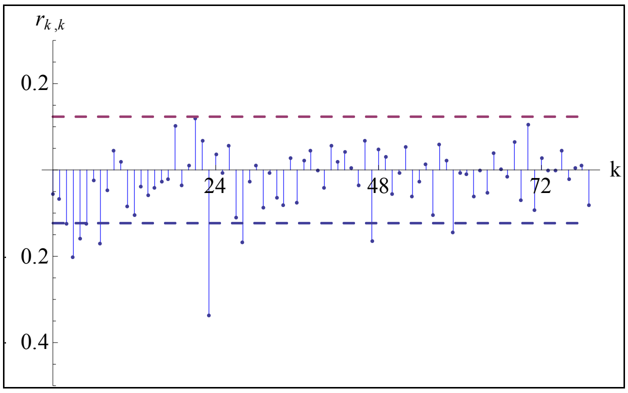

Fig. 6: The sample PACF of data after regular and seasonal differencing with period 24

The sample ACF and PACF of our differenced series of heat demand are shown in the Fig. 5 and Fig. 6. The single prominent dip in the ACF function at lag 24 shows that the seasonal component probably only persists in the MA part and $Q=1$. This is also consistent with the behavior of the PACF that has dips at lags that are multiples of the seasonal period 24. The correlation plot also suggests that the order of the regular AR part is small or zero. Because sample ACF cuts off after 4 lag, order of the regular MA part can achieve the value of 4. Based on these observations, it is possible to consider the models with $P=0$, $Q=1$ and $p \le 1$, $q \le 4$.

3.1 Information criteria for order selection

The order of model is usually difficult to determine on the basis of the ACF and PACF. This method of identification requires a lot of experience in building up models. From this point of view it is more suitable to use the objective methods for the tests of the model order. A number of procedures and methods exist for testing the model order [7]. These methods are based on the comparison of the residuals of various models by means of special statistics so-called information criteria. The Time series package in Wolfram Mathematica use the two commonly functions. Formula (7) is called Akaike's information criterion (AIC) and the second function (8) is called Bayesian information criterion (BIC).

$$AIC(p,q) = \ln\hat{\sigma}^2 + 2(p+q)/n \quad (7)$$

$$BIC(p,q) = \ln\hat{\sigma}^2 + (p+q)\ln n/n \quad (8)$$

Here $\hat{\sigma}^2$ is an estimate of the residual variance and n is a number of residuals. To get the AIC or BIC value of an estimated model in Time series package it is possible simply to use the functions `AIC[model,n]` or `BIC[model,n]`. Since the calculation of these values requires estimated residual variance, the use of these functions will be demonstrated later.

4 Estimation of model parameters

After selecting a model (model identification), parameter estimation of the selected model has to be looked for. The Time series package includes different commonly used methods of estimating the parameters of the ARMA types of models. Each method has its own advantages and limitations. Apart from the theoretical properties of the estimators (e.g., consistency, efficiency, etc.), practical issues like the speed of computation and the size of the data must also be taken into account in choosing an appropriate method for a given problem. Often, we may want to use one method in conjunction with others to obtain the best result. The maximum likelihood method and the conditional maximum likelihood method are used for estimating the parameters of our selected model.

The maximum likelihood method of estimating model parameters is often favored because it has the advantage among others that its estimators are more efficient (i.e., have smaller variance) and many large-sample properties are known under rather general conditions. The function `MLEstimate[data,model,{φ1,{φ11,φ12}},…]` fits selected *model* to *data* using the maximum likelihood method. The parameters to be estimated are given in symbolic form as the arguments to *model*, and two initial numerical values for each parameter are required. The exact maximum likelihood estimate can be very time consuming especially for large n or large number of parameters. Therefore, an approximate likelihood function is used in order to speed up the calculation. The likelihood function so obtained is called the conditional likelihood function. The Time series package in Mathematica environment use the function `ConditionalMLEstimate[data,model]` to fit *model* to *data* using the conditional maximum likelihood estimate.

Estimation of the parameters of the models of differenced series of heat demand is presented here. On the base of conclusions in section 4 we consider nine models of time series of heat demand in the form SARIMA$(p,1,q)\times(0,1,1)_{24}$. That means SARIMA$(p,0,q)\times(0,0,1)_{24}$ model for differenced series of heat demand. We use the result from the Hannan-Rissanen estimate (function `HannanRissanenEstimate`) as our initial values of AR(p) and MA(q) processes. In addition to that this result was considered by selection of models. The initial value of Θ_1 is determined from our sample partial correlation function as $-r_{48,48}/r_{24,24} \cong -0.45$.

After definition of models the conditional maximum likelihood estimate of considered models is obtained. Then we use the exact maximum likelihood method to get better estimates for the parameters of models. For comparison of selected model AIC and BIC information criterion are used. The results of estimation are presented in the Tab. 3. Adequacy of these models may be examined by means of Portmanteau test.

Table 3: Evaluation of selected models for differenced series of heat demand

Type of model SARIMA$(p,d,q)\times(P,D,Q)_s$	Information criterions		Portmanteau statistic Q_{20}	Quantile of Chi-Square distribution $\chi^2_{1-\alpha}(K\text{-}p\text{-}q\text{-}Q)$
	AIC	BIC		
SARIMA$(0,0,1)\times(0,0,1)_{24}$	3.2716	3.2988	43.01	28.8693
SARIMA$(1,0,0)\times(0,0,1)_{24}$	3.2717	3.2988	43.02	28.8693
SARIMA$(1,0,1)\times(0,0,1)_{24}$	3.1791	3.2199	35.77	27.5871
SARIMA$(0,0,2)\times(0,0,1)_{24}$	3.2595	3.3003	57.35	27.5871
SARIMA$(0,0,3)\times(0,0,1)_{24}$	3.1917	3.2460	36.21	26.2962
SARIMA$(0,0,4)\times(0,0,1)_{24}$	3.1308	3.1987	20.88	24.9958
SARIMA$(1,0,4)\times(0,0,1)_{24}$	3.1320	3.2135	19.33	23.6848
SARIMA$(0,0,5)\times(0,0,1)_{24}$	3.1280	3.2095	19.59	23.6848
SARIMA$(0,0,6)\times(0,0,1)_{24}$	3.1297	3.2248	21.47	22.3620

4.1 Diagnostic checking - Portmanteau test

After fitting, a model is usually examined to see if it is indeed an appropriate model. There are various ways of checking if a model is satisfactory. The commonly used approach to diagnostic checking is to examine the residuals. If the model is appropriate, then the residual sample autocorrelation function should not differ significantly from zero for all lags greater than one. We may obtain an indication of whether the first *K* residual autocorrelation considered together indicate adequacy of the model. This indication may be obtained by means of Portmanteau test. The Portmanteau test is based on the statistic in the form (9), which has an asymptotic chi-square distribution with *K-p-q* degrees of freedom.

$$Q_K = n(n+2)\cdot\sum_{k=1}^{K}\frac{r_k^2(\hat{\varepsilon})}{n-k} \tag{9}$$

Where *n* is a number of residuals, $r_k^2(\hat{\varepsilon})$ is value of sample ACF of residual at lag *k*.
If the model is inadequate, the calculated value of Q_K will be too large. Thus we should reject the hypothesis of model adequacy at level α if Q_K exceeds an appropriately small upper tail point of the chi-square distribution with *K-p-q* degrees of freedom (10).

$$Q_K > \chi^2_{1-\alpha}(K-p-q) \tag{10}$$

The Mathematica (Time series package) function `PortmanteauStatistic[residual,K]` gives the value of Q_K. The values of Portmanteau statistic using the residuals given the considered models and observed data are displayed in the Tab. 3. These values are compared with 5 percent value chi-square variable with *K-p-q-Q* degrees of freedom. (we consider $K=20$ and $\alpha=0.95$). Based on these results we would conclude that the first 5 models are not satisfactory whereas there is no strong evidence to reject the next 4 models.

5 Conclusion

This paper presents possibility of model design of time series of heat demand in Mathematica environment – Time series package. The results of this paper confirm supposition that this package is applicability to analysis of time series of heat demand. Some models were proposed and the adequacy of these models was tested by means of Portmanteau statistic. The proposed model is possible to use for prediction of heat demand in the concrete locality. This prediction of heat demand plays an important role in power system operation and planning. It is necessary for the control in the Centralized Heat Supply System (CHSS), especially for the qualitative-quantitative control method of hot-water piping heat output – the Balátě System.

Acknowledgement:
This work was supported in part by the Ministry of Education of the Czech Republic under grant No. MSM7088352102 and National Research Programme II No. 2C06007.

References:

[1] BALÁTĚ, Jaroslav. *Design of Automated Control System of Centralized Heat Supply*. Brno, 1982. 152 s. Thesis of DrSc. TU Brno, Faculty of Mechanical Engineering.

[2] PARK, Tae Chang, et al. Optimization of district heating systems based on the demand forecast in the capital region. *Korean journal of chemical engineering*. 2009, 26, 6, s. 1484-1496. ISSN 0256-1115.

[3] BOX, George E; JENKINS, Gwilym M. *Time series analysis : forecasting and control.* Rev. ed. San Francisco : Holden-Day, 1976. 575 s. ISBN 0-8162-1104-3..

[4] *Wolfram Research: Mathematica, technical and scientific software* [online]. c2011 [cit. 2011-03-17]. Available from WWW: <http://www.wolfram.com/>.

[5] *Time Series: A Fully Integrated Environment for Time-Dependent Data Analysis* [online]. c2011 [cit. 2011-03-17]. Wolfram Research. Available from WWW: <http://www.wolfram.com/products/applications/timeseries/>.

[6] ANDERSON, O. D. *Time Series Analysis and Forecasting – The Box-Jenkins approach*. London and Boston: Butterworths,1976. 180 p., ISBN: 0-408-70926-X.

[7] CROMWELL, J.B., LABYS W.S., TERRAZA M. *Univariate Tests for Time Series Models*. Thousand Oaks-California: SAGE Publication, 1994. ISBN 0-8039-4991-X.