Assessing and Forecasting of Epidemiological Data using Time Series Analysis

THEODOR D. POPESCU, ADRIANA ALEXANDRU, MARILENA IANCULESCU
National Institute for Research and Development in Informatics
8-10 Averescu Avenue, 011455 Bucharest, ROMANIA
(Theodor.Popescu, Adriana.Alexandru, Marilena.Ianculescu)@ici.ro

Abstract: The paper gives an overview of time series modeling and forecasting, using multiplicative $SARIMA$ models, with application in assessing and forecasting of epidemiological data. After general view of the main models and the methodological issues used in Box-Jenkins approach, the paper presents a case study having as subject the modeling and forecasting of a time series representing the measles infections, in Great Britain, 1971-1994, quarterly recorded, and an example of intervention analysis, using as exogenous data the measles infections, and as endogenous variable the number of vaccinated persons, in the same time period. The intervention analysis proved to be a useful approach to model interrupted time series, when the time series is affected by the effect of population vaccination.

Key–Words: Time series analysis, modeling, forecasting, intervention analysis, Box-Jenkins approach, epidemiological data, case study.

1 Introduction

Although the incidence of epidemic diseases has reached historic lows in many parts of the world, these diseases still causes substantial morbidity globally. So, it is of great interest to explore the degree to which new sources of data, combined with existing public health data, can be used to evaluate the landscape of immunity and the role of vaccination in the eradication of epidemic diseases. In this context, different public health surveillance systems have been developed to facilitate the detection of abnormal behavior of infectious diseases and other adverse health events. To achieve this goal, different approaches have been used for assessing and forecasting of infectious disease incidence. Time series analysis enjoys of great interest in this field. It makes use of statistical models able to forecast the epidemiological behavior of the historical surveillance data. Different methods have been reported in the literature, such as: exponential smoothing, generalized regression, decomposition methods, and multilevel time series models, among others.

Seasonal autoregressive integrated moving average $(SARIMA)$ models have been extensively used for epidemic time series forecasting including the hemoragic fever renal syndrome, [1], [2], dengue fever,[3], [4], and tuberculosis, [5]. Also, the problem of modeling and forecasting of measles infection is present in many papers, [6], [7], [8] [9], among others, using different approaches.

The paper gives a general view on the time series models, regression and intervention models, with application in modeling and forecasting of epidemiological surveillance data. The approach is used in modeling and forecasting of measles infections in Great Britain, 1971-1994, for a multiplicative $SARIMA$ model and an interrupted time series ITS model, taking into account the number of vaccinated persons.

2 Time Series Models

The statistical approaches adopted in time series modeling and forecasting usually rely on multiplicative $SARIMA$ (Seasonal Auto Regressive Integrated Moving Average) model. A such model has the following form for the time series z_t, [10]:

$$\phi(B)\Phi(B^s) \bigtriangledown^d \bigtriangledown_s^D z_t = \theta(B)\Theta(B^s)a_t \qquad (1)$$

where a_t a white noise and

$$\begin{aligned}
\phi(B) &= 1 + \phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p \\
\theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \\
\Phi(B^s) &= 1 + \Phi_s B^s + \Phi_{2s} B^{2s} + \ldots + \Phi_{Ps} B^{Ps} \\
\Theta(B^s) &= 1 + \Theta_s B^s + \Theta_{2s} B^{2s} + \ldots + \Theta_{Qs} B^{Qs}
\end{aligned}$$

with B the time delay operator, $Bz_t = z_{t-1}$, $\bigtriangledown z_t = (1 - B) = z_t - z_{t-1}$, nonseasonal differentiating operator, and $\bigtriangledown_s z_t = (1 - B^s) = z_t - z_{t-s}$, seasonal

differentiating operator: d is the nonseasonal differentiating order, D is the seasonal differentiating order and s is the seasonal period of the series.

The model $SARIMA(p,d,q)(P,D,Q)_s$ is presented in Fig.1, where (p,d,q) and (P,D,Q) denote nonseasonal orders and seasonal orders of the model, respectively.

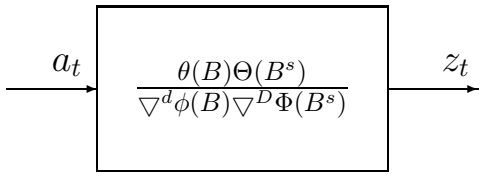

Figure 1: Model $SARIMA(p,d,q)(P,D,Q)_s$

The multiplicative form of the model simplifies the stationarity and invertibility conditions checking; these conditions can be separately checked, for seasonal and nonseasonal coefficients of the model. Starting from the general model form of the model $SARIMA(p,d,q)(P,D,Q)_s$ it can be obtain related models.

In some situations, it is known that some external events, or inputs, can affect the variables for which the practitioner intends to forecast the future time series values. The dynamic models, used in this case, are special kind of $SARIMA$ model and are called intervention models or interrupted time series (ITS) models, [11]. As examples of practical interventions can be mentioned: the effect of medication on the health of the patient, vaccination campaigns, the effect of the exchange of the laws in legislation to prevent the morbidity and mortality, etc.

A such intervention model can be represented like a transfer function (TF) model (see Fig. 2), where z_t is the value of the endogenous variable at time t, $\mathbf{u}_t = [u_{1t},\ldots,u_{rt}]^T$ is the vector of exogenous variables, and a_t is a white noise error.

$$\begin{aligned}\Omega_i(B) &= \omega_{i0}+\omega_{i1}B+\omega_{i2}B^2+\cdots+\omega_{in_i}B^{n_i}\\ &\quad i=1,2,\ldots,r\\ \Delta_i(B) &= 1+\delta_{i1}B+\theta_{i2}B^2+\cdots+\delta_{in\delta_i}B^{n\delta_i}\\ &\quad i=1,2,\ldots,r\end{aligned}$$

$\phi(B)$, $\theta(B)$,$\Phi(B^s)$ and $\Theta(B^s)$ have been described above.

The models are identified by the mean of the autocorrelation (ACF) and the partial autocorrelation functions $(PACF)$.

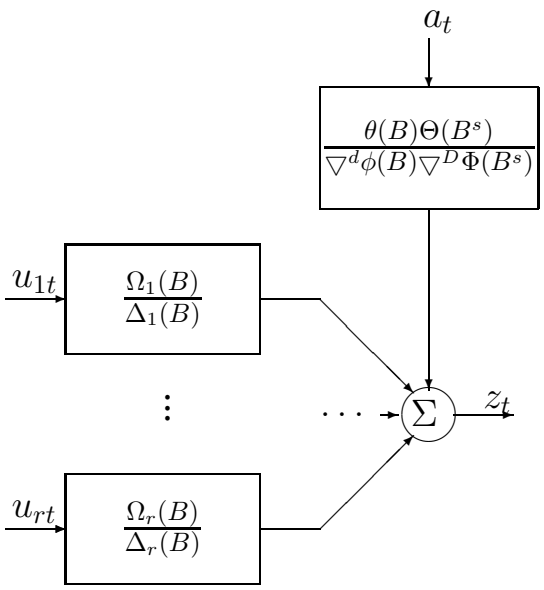

Figure 2: Transfer function (TF) model

3 Methodological Aspects

The time series model construction usually includes the following stages, [10]:

- Identification (specification) of the time series model using some data analysis tools (different graphical representations, autocorrelation functions (ACF) and partial autocorrelation functions $(PACF)$) in order to determine the types of transformations to obtain stationarity and to estimate the degree of differentiation needed to induce stationarity in data, as well as the polynomial degrees of autoregressive and moving average operators in the model.
- Model parameter estimation of the time series implies the use of efficient methods (such as maximum likelihood, among others) for parameter estimation, standard errors and their correlations, dispersion of residuals, etc.
- Model evaluation (validation) aims to establish the model suitability, or to make some simplifications in structure and parameter estimates, using residual analysis.

The process often add a preliminary stage of data preparation and a final stage of model application, or forecasting, [12].

Visual analysis of series data allows a first image on the series' non-stationarity and on the presence of a seasonal pattern in the data. The final decision on the inclusion of seasonal elements in the time series

model will be taken after the autocorrelation function (ACF) and partial autocorrelation function ($PACF$) analysis, as well as after the estimation results analysis; the visual analysis of the data can provide useful additional information.

Significant changes in the mean value of the series data require non seasonal differentiation of the first order, while the varying of the rate for average value imposes the nonseasonal differentiation of the second order of the series. Strong seasonal variations usually require, not more than the seasonal differentiation of the first order of the series'data. Autocorrelation function of the series offers information on the nonseasonal and seasonal degrees to be used to obtain the stationarity of the data, as well as on the model structure, [13], [14]. Also, in the validation-diagnosis stage, the attention will be focused on the coefficients of seasonal autocorrelations, using t statistic test.

In the estimation stage, the use of the initial estimates of the model parameters of the value of 0.1 leads to good results in most cases; better initial estimates for model parameters can be obtained based on the autocorrelation and partial autocorrelation functions. The criteria Akaike Information Criterion (AIC), Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC), [15], Adjusted Root Mean Square Error (ARMSE) and Absolute Mean Percent Error (AMPE), [13], offer information on the parameter estimation quality.

Forecasting is what the whole procedure is designed to accomplish. Once the model has been selected, estimated and checked, it is usually a straight forward task to compute forecasts. The forecasting problem can be solved, in the most direct way, using the multiplicative $SARIMA$ model of the form (1). The description of the model by an infinitely weighted sum of current values and the earlier noise is proving useful, in particular, to estimate the variance of forecasting values, as well as to determine their confidence intervals. Standards and practices for time series forecasting are given in [16].

4 Case Study

The case studies making the object of this section has as subject the modeling and forecasting of a time series representing the measles infections, in Great Britain in the period 1971-1994, quarterly recorded, and an example of intervention analysis, using as the exogenous data the number of measles infections, and as endogenous variable the number of vaccinated persons, in the same time period, using a transfer function model.

4.1 Modeling and Forecasting of Measles Infections, in Great Britain 1971-1994

The time series representing the measles infections, in Great Britain in the period 1971-1994, quarterly recorded, is presented in Fig. 3.

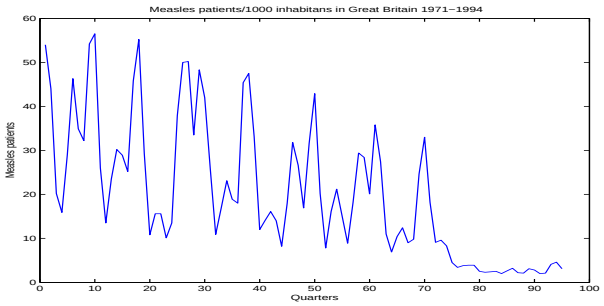

Figure 3: Number of measles infections/1000 inhabitants, Great Britain, 1971-1994.

We present in Fig. 4 the autocorrelation (ACF) and partial autocorrelation $(PACF)$ functions of the original data, and the Ljung Box Q (LBQ) test.

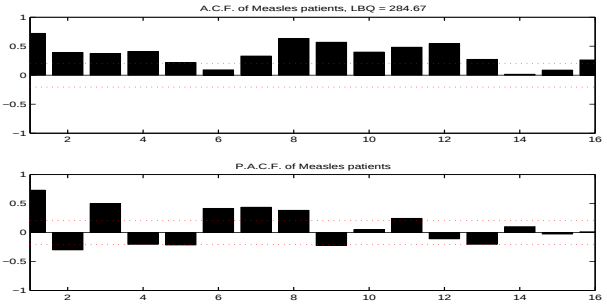

Figure 4: ACF and PACF functions of measles infections/1000 inhabitants, Great Britain, 1971-1994.

It can be noted, from the data analysis, the nonstationary and seasonal character of the series. Because the data are quarterly recorded, it can be supposed the presence in the data series of a seasonal component of period $s = 4$ (yearly); it is also confirmed by the autocorrelation function (ACF). So, the original time series has been seasonal differentiated with period $s = 4$, and it is presented in Fig. 5.

The ACF and $PACF$ of differentiated series, and Ljung Box Q test, are given in Fig. 6.

Starting from these functions, the following $SARIMA$ model structure resulted:

$$(1 + \Phi_4 B^4 + \Phi_8 B^8)(1 - B^4) z_t = \\ = (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta_4 B^4) a_t \qquad (2)$$

and $v[a_t] = \sigma^2$.

The model parameter estimation has been performed using the Broyden-Fletcher-Goldfarb-Shanno

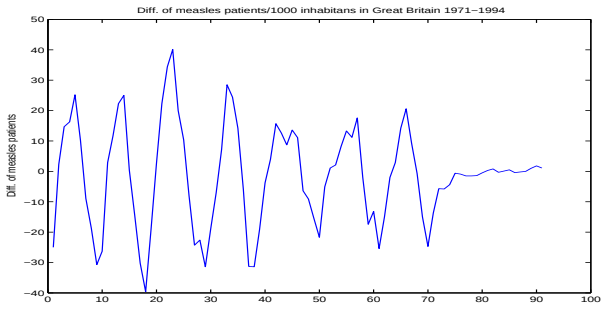

Figure 5: Differentiated series with s = 4.

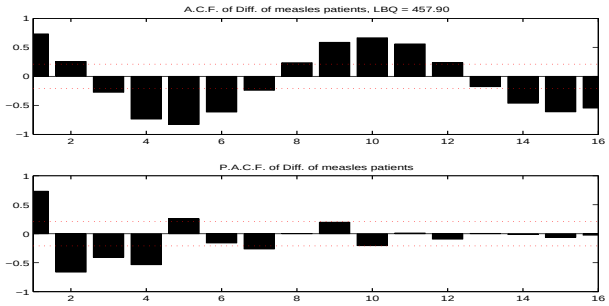

Figure 6: ACF and $PACF$ of differentiated series with s = 4.

(BFGS) optimization algorithm, [17]. The results are presented in Table 1, with the objective function = 315.7083, nr. of iterations = 24 and information criteria: AIC = 6.7728 and SBC = 6.9341; the correlation matrix of model parameter estimates denotes the model quality.

Table 1: Results for $SARIMA$ model parameter estimation

Param.	**Estim.**	**Appr. Std. Dev.**	**t-test**
Φ_4	-0.461	0.104	-4.435
Φ_8	-0.509	0.100	-5.084
θ_1	1.043	0.106	9.824
θ_2	0.507	0.088	5.755
Θ_4	-0.492	0.096	-5.110
$v[a_t]$	40.653	5.986	6.790

The model residuals are presented in Fig. 7, and residual ACF, $PACF$, Ljung Box Q test, are given in Fig. 8.

The estimation results confirm the model quality, according with the Box-Jenkins methodology used in time series analysis, [13].

The forecasting, for the resulted model, has been performed, started from the 92 quarter for a horizon time of 4 quarters and the 95% confidence limits, to compare the original data with the forecasting results.

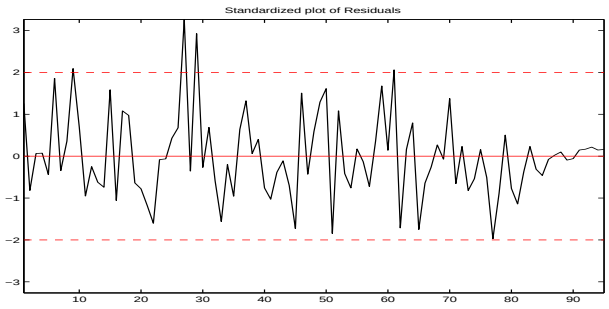

Figure 7: Model residuals

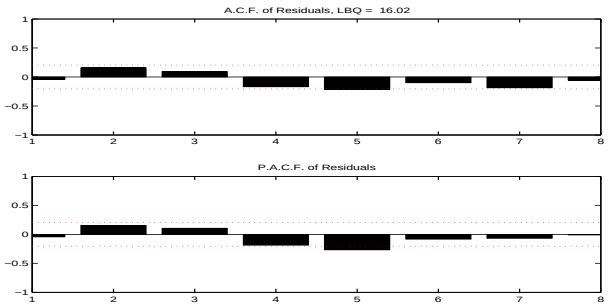

Figure 8: ACF and PACF of model residuals.

It can be noted that the forecasting results follow the evolution trend of the original time series, and are in the confidence limits 95%. The forecasting results and confidence limits are given in Fig. 9.

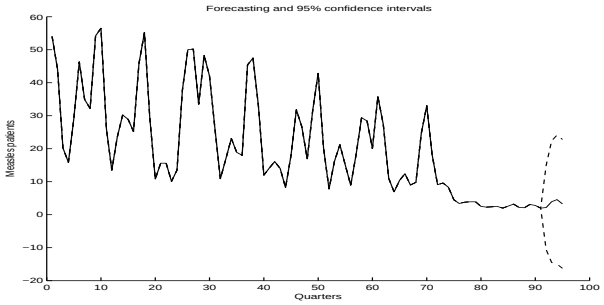

Figure 9: Forecasting results and confidence limits 95% for 4 quarters.

4.2 Modeling and Forecasting of Measles Infections, in Great Britain Function of Vaccinations 1971-1974

In this case an intervention model, a transfer function (TF) model, has been used, with the exogenous variable the number of measles infections, z_t, and with endogenous variable the percent of vaccinated persons, u_t, in the time period making the object of the analysis. The percent of measles vaccinations, Great Britain, 1971-1994 is presented in Fig. 10.

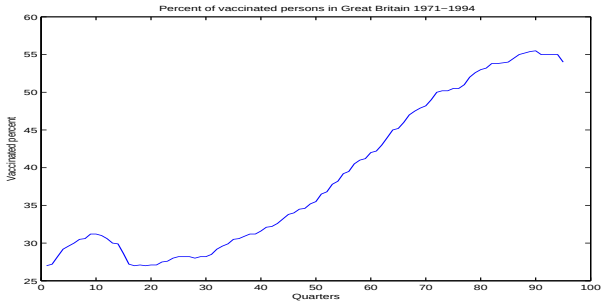

Figure 10: Percent of measles vaccinations, Great Britain 1971-1994.

After preliminary analysis of the data, and different model structures, resulted the following structure of the transfer function model, representing the intervention model:

$$(1-B^4)z_t = \frac{\omega_1}{1+\delta_1 B}u_t + \\ + \frac{(1+\theta_1 B+\theta_2 B^2)(1+\Theta_4^4)}{1+\Phi_4 B^4+\Phi_8 B^8}a_t \qquad (3)$$

with $v[a_t]=\sigma^2$ and $s=4$, due to the nostationarity of the data. For the model parameters and variance, σ^2, have been used as initial values 0.1. Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm, [17], was used for parameter estimation, resulting the following values for model parameters (see Table 2):

Table 2: Results for TF model parameter estimation

Param.	Estim.	Appr. Std. Dev.	t-test
Φ_4	-0.580	0.090	-6.397
Φ_8	-0.349	0.086	-4.065
θ_1	1.055	0.088	11.893
θ_2	0.529	0.079	6.683
Θ_4	-1.000	0.037	-26.382
ω_1	-0.289	0.118	-2.443
δ_1	0.873	0.063	13.721
$v[a_t]$	25.682	4.137	6.208

for an objective function = 290.7013, nr. of iterations = 50 and information criteria: AIC = 6.2884, and SBC = 6.5035; the correlation matrix of TF model parameter estimates denotes the model quality.

The model residuals are presented in Fig. 11, and the residual ACF and $PACF$, with Ljung Box Q test, are given in Fig. 12.

The results confirm the model quality, according with the Box-Jenkins methodology used, [13]. The forecasting results, for the transfer model resulted,

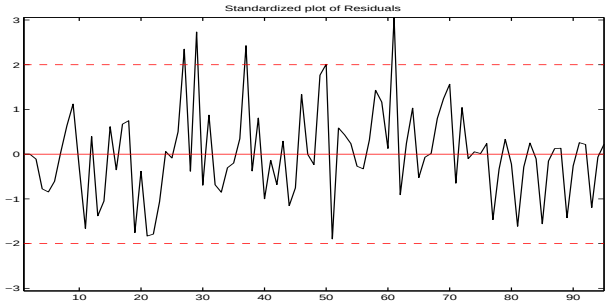

Figure 11: Transfer function model residuals.

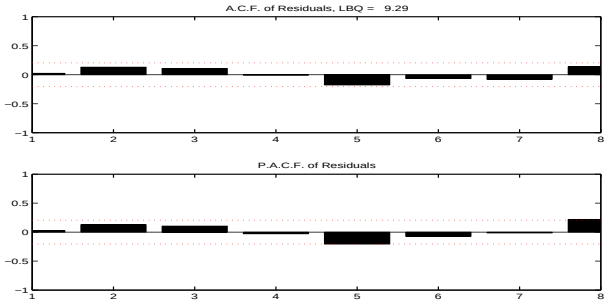

Figure 12: ACF and $PACF$ transfer function residuals.

started from the 92 quarter for a horizon time of 4 quarters and the 95% confidence limits are given in Fig. 13; the values used, as percent of vaccinations for the forecasting measles infections, in forecasting, represent the values recorded for the last 4 quarters of the original series. It can be noted that the forecasting results follow the evolution trend of the time series of measles infections, and are in the confidence limits 95%.

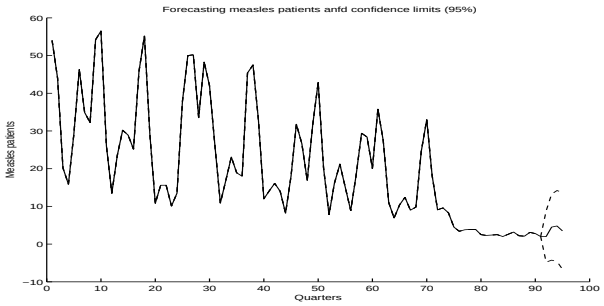

Figure 13: Forecasting results and confidence limits 95% for 4 quarters using transfer function model.

5 Conclusions

The time series modeling and forecasting of epidemiological surveillance data using seasonal multiplicative $SARIMA$ models and the attractive features of

the Box-Jenkins approach provide an adequate description to the data in this field. The $SARIMA$ processes are a very rich class of possible models and it is usually possible to find a process which provides an adequate description to the data. Also, the intervention analysis, or ITS model using, proved to be a useful approach to model interrupted time series, in this case, when such time series are affected by the effect of medication on the health of the patient, population vaccination policies, some law constraints, etc. The case study presented in the paper proved the efficiency of the approach.

The Box-Jenkins methodology, applicable to a wide variety of statistical modeling situations, provides a convenient framework which allows an analyst to think about the data, and to find an appropriate statistical model which can be used to help answer relevant questions about the data.

Acknowledgments

The authors thank the Ministry of Research and Innovation for its support under the 2019-2022 Core Program, Cod PN 301 200, Project RO-SmartAgeing.

References:

[1] Q. Li, N. N. Guo, Z. Y. Han, Y. B. Zhang, S. X. Qi, et al., Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic Fever with renal syndrome, *The American Journal of Tropical Medicine and Hygiene*, 87, 2012, pp. 364-370.

[2] Q. Liu, X. Liu, B. Jiang and W. Yang, Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model, *BMC Infectious Diseases*, 11, 2011.

[3] S. Wongkoon, M. Jaroensutasinee and K. Jaroensutasinee, Development of temporal modeling for prediction of dengue infection in Northeastern Thailand, *Asian Pacific Journal of Tropical Medicine*, 5, 2012, pp. 249-252.

[4] P. L. Luz, B. V. M. Mendes, C. T. Codeco, C. J. Struchiner and A. P. Galvani, Time series analysis of dengue incidence in Rio de Janeiro, Brazil, *American Journal of Tropical Medicine and Hygiene*, 79, 2008, pp. 933-939.

[5] M. Rios, J. M. Garcia, J. A. Sanchez and D. Perez, A statistical analysis of the seasonality in pulmonary tuberculosis, *European Journal of Epidemiology*, 16, 2000, pp. 483-488.

[6] S. Sharmin and I. Rayhan, Modelling of infectious diseases for providing signal of epidemics: A measles case study in Bangladesh, *J. Health. Popul. Nutr.*, 29, 2011, pp. 567-573.

[7] M. G. Roberts and M. I. Tobias, Predicting and preventing measles epidemics in New Zealand : application of a mathematical model, *Epidemiol. Infect.*, 124, 2000, pp. 279–287.

[8] O. N. TTAR et al., Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model, *Ecological Monographs*, 72, 2002, pp. 169-184.

[9] A. Sumi et al., Prediction analysis for measles epidemics, *Jpn. J. Appl. Phys.*, 42, 2003.

[10] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 2-nd Edition, Holden Day, San Francisco , 1976.

[11] G. E. P. Box and G. C. Tiao, Intervention analysis with applications to economic and environmental problems, *Journal of the American Statistical Association*, 70, 1975, pp. 70–79.

[12] S. Makridakis, S. C. Wheelwright and R. J. Hyndman, *Forecasting: Methods and Applications*, New York: John Wiley & Sons, 1998.

[13] A. Pankratz, *Forecasting with Univariate Box-Jenkins Models*, Wiley, New York, 1983.

[14] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer-Verlag, New York, 1996.

[15] S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer, 2008.

[16] J. Scott Armstrong, Standards and practices for forecasting, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, J. Scott Armstrong (ed.), MA: Kluver Academic Publishers, 2001, pp. 1–46.

[17] J. Casals, A. G. ,Hiernaux, M. Jerez,S. Sotoca and A. Trindade, *State-Space Methods for Time Series Analysis: Theory, Applications and Software*, Chapman and Hall/CRC, 2016.