*6th International Conference on Applied Informatics*
*Eger, Hungary, January 27–31, 2004.*

# Application Time Series Models on Medical Research

**Mária Fazekas**

Department of Economic- and Agroinformatics,
University of Debrecen
e-mail: kiss@thor.agr.unideb.hu

**Abstract**

In this paper we demonstrate applying time series models on medical research. The Hungarian mortality rates were analysed by autoregressive integrated moving average models and seasonal time series models examined the data of acute childhood lymphoid leukaemia.

The mortality data may be analysed by time series methods such as autoregressive integrated moving average (ARIMA) modelling. This method is demonstrated by two examples: analysis of the mortality rates of cerebrovascular diseases and analysis of the mortality rates of cancer of cervix. Mathematical expressions are given for the results of analysis. The relationships between time series of mortality rates were studied with ARIMA models. Calculations of confidence intervals for autoregressive parameters by tree methods: standard normal distribution as estimation and estimation of the White's theory and the continuous time case estimation. Analysing the confidence intervals of the first order autoregressive parameters we may conclude that the confidence intervals were much smaller than other estimations by applying the continuous time estimation model.

We present a new approach to analysing the occurrence of acute childhood lymphoid leukaemia. We decompose time series into components. The periodicity of acute childhood lymphoid leukaemia in Hungary was examined using seasonal decomposition time series method. The cyclic trend of the dates of diagnosis revealed that a higher percent of the peaks fell within the winter months than in the other seasons. This proves the seasonal occurrence of the childhood leukaemia in Hungary.

**Key Words and Phrases:** time series analysis, autoregressive integrated moving average models, mortality rates, seasonal decomposition time series method, acute childhood lymphoid leukaemia

# 1. Introduction

Time series analysis is a well-known method for many years. Box and Jenkins provided a method for constructing time series models in practice [1], [2]. Their method often referred to as the Box-Jenkins approach and the autoregressive integrated moving average models (ARIMA). This method has been applied in the beginning such fields as industry and economics later in medical research as well as [3], [4], [5], [6].

The method of seasonal time series analysis can be used in various fields of the medicine. With such time series one can detect the periodic trend of the occurrence of a certain disease [7], [8], [9]. Among other diseases, the seasonal periodicity of the childhood lymphoid leukaemia was also analysed using statistical methods [10], [11]. The pathogenesis of the childhood lymphoid leukaemia is still uncertain, but certain environmental effects may provoke the manifestation of latent genes during viral infections, epidemics or pregnancy.

The date of the diagnosis of patients were statistically analysed to determine the role, which the accumulating viral infections and other environmental effects may play during the conception and fatal period on the manifestation of the disease. Because the available data is rather limited and controversial, it seemed logical to make an in-depth analysis of the date of diagnosis of the acute lymphoid leukaemia in Hungarian children.

# 2. Methods

## 2.1. Autoregressive moving average models

The mortality data often change in the form of 'time series'. Data of frequencies of mortality rates are usually collected in fixed intervals for several age groups and sexes of the population. Let the value of the mortality rates $z_t$, $z_{t-1}$, $z_{t-2}$, ... in the years t, t-1, t-2, .... For simplicity we assume that the mean value of $z_t$ is zero, otherwise the $z_t$ may be considered as deviations from their mean. Denote $a_t$, $a_{t-1}$, $a_{t-2}$, ... a sequence of identically distributed uncorrelated random variables with mean 0 and variance $\sigma_a^2$. The $a_t$ are called white noise.

The autoregressive moving average model of order p, q (ARMA(p, q)) can be represent with the following expression [1], [12]: $z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$. Where $\phi_1, \phi_2, \ldots, \phi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ are parameters, p means the p order of autoregressive process and q denotes the q order of moving average process.

There are special cases of the ARMA(p, q) models: the autoregressive model of order p (AR(p) model) and the moving average model of order q (MA(q) model). The AR(p) [1], [12]: $z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t$. The MA(q) [1], [12]: $z_t = a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$. The special case of AR(p); when p=1; $z_t = \phi_1 z_{t-1} + a_t$. $z_t$ is linearly dependent on the previous observation $z_{t-1}$ and the random shock $a_t$.

The special case of MA(q); when q=1; $z_t$=$a_t + \theta_1 a_{t-1}$. In this case $z_t$ is linear expression of the present and previous random shock.

The time series that has a constant mean, variance, and covariance structure, which depends only on the difference between two time points, is called stationary. Many time series are not stationary. It has been found that the series of first differences is often stationary. Let $w_t$ the series of first differences, $z_t$ the original time series, than $w_t$=$z_t$-$z_{t-1} = \nabla z_t$. The Box-Jenkins modelling may be used for stationary time series [1], [12].

The dependence structure of a stationary time series $z_t$ is described by the autocorrelation function: $\rho_k$ = correlation($z_t$;$z_{t+k}$); k is called the time lag. This function determines the correlation between $z_t$ and $z_{t+k}$.

To identify an ARIMA model Box and Jenkins have suggested an iterative procedure [1]:

- for provisional model may be chosen by looking at the autocorrelation function and partial autocorrelation function
- parameters of the model are estimated
- the fitted model is checked
- if the model does not fit the data adequately one goes back to the start and chooses an improved model.

Among different models, which represent the data equally well, one chooses the simplest one, the model with fewest parameters [1], [12].

The relation between two time series $z_t$ and $y_t$ can be give by the cross correlation function ($\rho_{zy}(k)$); $\rho_{zy}(k)$ = correlation($z_t; y_{t+k}$); where k= $0, \pm 1, \pm 2, \ldots$. The cross correlation function determines the correlation between the time series as a function of the time lag k [1].

## 2.2. Estimations for confidence intervals

For estimation the parameter of first order autoregressive model two methods are well known: apply the standard normal distribution as estimation and the White method [13]. These methods cannot be applied in non-stationary case. Little known for estimation of the parameter of first order autoregressive parameter is the application of estimation for continuous time case processes [13], [14]. This method can be applied in each case properly.

## 2.3. Seasonal time series

The time series usually consist of three components: the trend, the periodicity and the random effects. The trend is a long-term movement representing the main direction of changes. The periodicity marks cyclic fluctuations within the time series. The irregularity of the peaks and drops form a more-or-less constant

pattern around the trend line. Due this stability the length and the amplitude of the seasonal changes is constant or changes very slowly. If the periodic fluctuation pattern is stable, it is called a constant periodic fluctuation. When the pattern changes slowly and regularly over the time, we speak of a changing periodicity. The third component of the time series is the random error causing irregular, unpredictable, non-systematic fluctuations in the data independent from the trend line.

An important part of the time series analysis is the identification and isolation of the time series components. One might ask how these components come together and how can we define the connection between the time series and its components with a mathematical formula. The relationship between the components of a time series can be described either with an additive or a multiplicative model.

Let $y_{i,j}$ (i=1,. . . , n; j=1,. . . , m) marks the observed value of the time series. The index $i$ stands for the time interval (i.e. a year), the $j$ stands for a particular period in the time interval (i.e. a month of the year). By breaking down the time series based on the time intervals and the periods we get a matrix-like table. In the rows of the matrix are the values from the various periods of the same time interval; while in the columns are the values from the same periods over various time intervals.

$y_{1,1}; y_{1,2}; \ldots; y_{1,m};$
$y_{2,1}; y_{2,2}; \ldots; y_{2,m};$
$y_{3;1}; y_{3,2}; \ldots; y_{3,m};$
. . .
$y_{n,1}; y_{n,2}; \ldots; y_{n,m.}$

Let $d_{i,j}$ (i=1,2,. . . , n; j=1,2,. . . , m) mark the trend of the time series, $s_{i,j}$ (i=1,2,. . . , n; j=1,2,. . . , m), the periodic fluctuation and $\varepsilon_{i,j}$ (i=1,2,. . . , n; j=1,2,. . . , m), the random error. Using these denotations the additive seasonal model can be defined as $y_{i,j}=d_{i,j}+s_{i,j} + \varepsilon_{i,j}$, (i=1,2,. . . , n; j=1,2,. . . , m), the multiplicative model as $y_{i,j}=d_{i,j}*s_{i,j} * \varepsilon_{i,j}$; (i=1,2,. . . , n; j=1,2,. . . , m).

The trend of a time series can easily be computed with moving averages or analytic trend calculation. Moving averaging generates the trend as the dynamic average of the time series. Analytic trend calculation approximates the long-term movement in the time series with a simple curve (linear, parabolic or exponential curve) and estimates its parameters.

The indices of the periodic fluctuation are called seasonal differences (in the additive model) or seasonal ratios (in the multiplicative model). These indices represent the absolute difference from the average of the time interval using the additive model or the percentile difference using the multiplicative model. Seasonal adjustment is done by subtracting the $j$ seasonal difference from the $j$ data value of each $i$ season (additive model) or by dividing the $j$ data value of each $i$ season by the $j$ seasonal ratio (multiplicative model). The seasonally adjusted data reflect only the effect of the trend and the random error.

# 3. Results

## 3.1. Analysing the mortality rates

The SPSS program-package was used for analysing. ARIMA models were identified for some mortality rates. The results are demonstrated two cases from Hungarian mortality rates.

The Figure 1 illustrates the mortality rates of cancer of cervix for age class 0-64 years and over age 65. The autocorrelation functions decay for both data series [Figure 2].

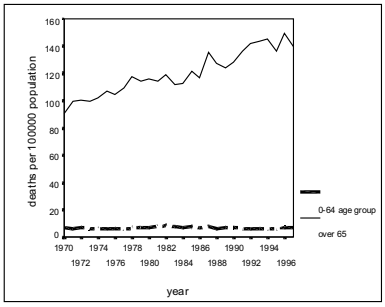


Figure 1: Mortality rates of cancer of cervix age-class 0-64 and over age 65

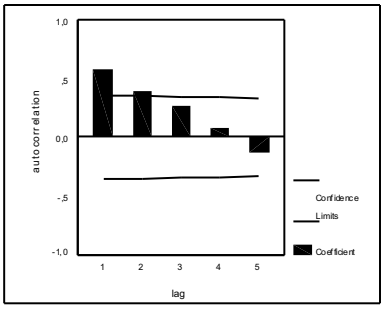


Figure 2: The autocorrelation function for mortality rates for age class 0-64

The partial autocorrelation functions have a significance value at k=1 lag. The first order autoregressive model can be acceptable on the basis of autocorrelation and partial autocorrelation functions. So the stochastic equation for age class 0-64 years: $z_t$=0,576$z_{t-1}+\varepsilon_t$. The model for over age 65 is the following: $z_t$=0,703$z_{t-1}+\epsilon_t$. When the fitted model is adequate then the autocorrelation of residuals have $\chi^2$ distribution with (K-p-q) degree of freedom [4]. On the basis of test the selected models were adequate because $\chi^2_{0-64} = 1,956$; $\chi^2_{\text{over }65} = 1,651$; $\chi^2_{0,05;5} = 11,07$.

The Figure 3 demonstrates the cross correlation function before fitting model. The cross correlation function for the residuals can be seen in the Figure 4 after fitting model. From behaviour of residuals we may be conclude that between examined time series isn't "synchronisation" [4].

The change in the mortality rates of cerebrovascular diseases for over age 65 between female and male are well illustrates in the Figure 5. The stochastic equation for the mortality rates of female: $z_t$=0,809$z_{t-1} + \varepsilon_t$; for data of male: $z_t = 0,792z_{t-1} + \varepsilon_t$. On the basis of the $\chi^2$ test the selected models were adequate; because $\chi^2_{female} = 3,886$; $\chi^2_{male} = 1,7461$; $\chi^2_{0,05} = 11,07$ [4].

The cross correlation function for residuals demonstrates in the Figure 6. It has significance value at k=0 lag on 95% significance level. It may be concluded that there is "synchronisation" between time series. In that years when the mortality rates for female increased the mortality rates for male increased as well.

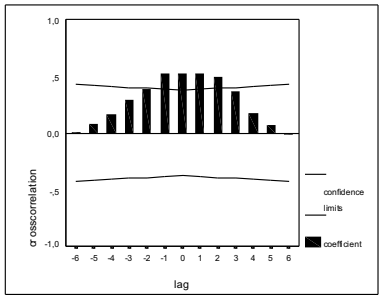


Figure 3: The cross correlation function for mortality rates of cancer of cervix between age class 0-64 and over age 65

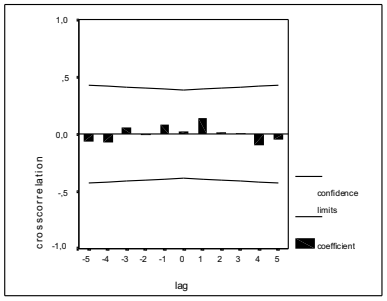


Figure 4: The cross correlation function of residuals for mortality rates of cancer of cervix between age class 0-64 and over age 65

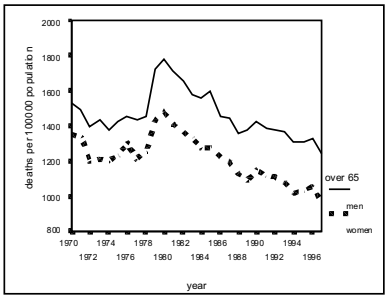


Figure 5: Mortality rates of cerebrovascular diseases over age 65 between female and male

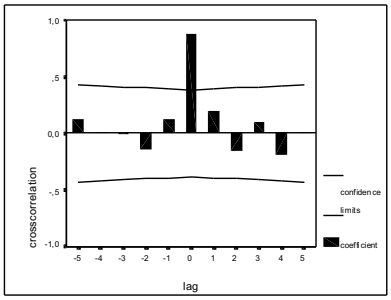


Figure 6: The cross correlation function of residuals for mortality rates of cerebrovascular diseases between groups

The confidence intervals were carried out by three mentioned methods. For the calculations of the confidence limits we used the tables of the known exact distribution of the maximum-likelihood estimator of the damping parameter of an autoregressive process [13], [14]. The confidence intervals for different significance levels for the first order autoregressive parameter of stochastic equation for male can be seen in the following table.

| $\phi \approx 0{,}792$(MALE) | p=0,1 | p=0,05 | p= 0,01 |
|---|---|---|---|
| Normal distribution | (0,5956;0,9884) | (0,5523;1,0317) | (0,4736;1,1104) |
| White method | (0,5993;0,9847) | (0,5596;1,0244) | (0,4791;1,1105) |
| Continuous time process | (0,6949;0,9424) | (0,6649;0,9723) | (0,6116;0,9978) |

### 3.2. Analysing the periodicity of acute childhood lymphoid leukaemia

The databank of the Hungarian Paediatric Oncology Workgroup contains the data of all the patients with lymphoid leukaemia diagnosed between 1988 and 2000. In this time interval a total of 814 children were registered (of which 467 were boys). The patients were 0-18 years old, with a mean age of 6,4 years and a median of 5,4 years.

The components of the time series can be identified and isolated using statistical program packages. The analysis of the seasonal periodicity of the acute childhood lymphoid leukaemia was done with the SPSS 9.0 statistical program package.

The analysis of the periodicity of acute childhood lymphoid leukaemia was performed on the basis of the date of the diagnosis (year + month) of the disease. We analysed three data series. The first data series contained the number of all the patients diagnosed monthly, the second contained the number of those patients younger than the value of the median, the third series contained the number those older than the value of the median.

The seasonal trend of all patients revealed 9 peaks (peak=the value of the cyclic trend greater than 6), see Figure 7. 6 of these peaks fell within the winter months (November-February), 1 in the autumn period (September-October), 1 in the summer months (June-August) and 1 in the spring months (March-May).

The seasonal trend of the younger age group showed 7 peaks (peak=cyclic trend greater than 3) in the winter, 1 in the spring and 1 in the summer months.

The seasonal trend of the older age group showed 7 peaks (peak=cyclic trend greater than 3) in the winter, 1 in the spring, 1 in the autumn and 4 in the summer months.

## 4. Discussions

The Box-Jenkins models may be useful for analysing epidemiological time series. The method described the relationships between time series of mortality rates. It reveals strong synchronised behaviour of cerebrovascular diseases between the sexes. For time series of mortality data for cancer of cervix for age class 0-64 years and age class over 65 no such synchronisation is found between subgroups.

From the analysis of the first order autoregressive parameters it may be seen that by applying the normal distribution as estimation and White method the confidence intervals are near equal. For the upper estimations of confidence limits we can get larger than one applying these methods. Applying the continuous time process for the estimation of the confidence intervals they are much smaller and it can be used in each case [13].

Analysis of the seasonality of childhood lymphoid leukaemia in Hungary was performed both on the total number of patients and on the data series divided at the median. This way the characteristics can be observed more easily.

A certain periodicity was found in the dates of the diagnosis in patients with

leukaemia. Although there was some difference in the patterns of the cyclic trend peaks of the three time series, the majority of the peaks fell within the winter months in all three-time series. This was more significant in the group of all the patients and in the younger age group. The results of the analyses proved the seasonal occurrence of the childhood lymphoid leukaemia. Some studies reported similar seasonality [15], while other studies denied any kind such periodicity [16]. Our results prove the seasonal occurrence of the childhood lymphoid leukaemia in Hungary. Due to the controversial nature of the available international data, further studies should be carried out.

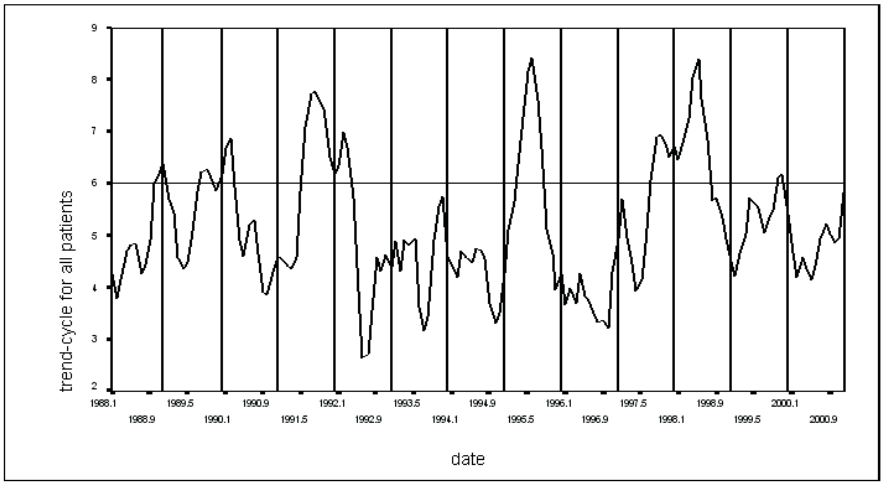


Figure 7: Seasonal trend of all patients of acute lymphoid leukaemia diagnosed monthly in the observed period.

# References

[1] Box, G.E.R., Jenkins, G.M.: *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco 1976.

[2] Jenkins, D.M., Watts, D.G.: *Spectral Analysis and its Applications*, Holden-Day, San Francisco 1968.

[3] Allard, R.: Use of Time Series Analysis in Infectious Disease Surveillance. *Bull. World Health Organ*, Vol. 76, (1998), 327-333.

[4] Helfenstein, U.: Detecting Hidden Relationships between Time Series of Mortality Rates. *Methods Inf. Med.*, Vol. 29, (1990), 57-60.

[5] Helfenstein, U., Ackermann-Liebrich, U., Braun-Fahrlander, C., Uhrs Wanner, H.: The Environmental Accident at 'Schweizerhalle' and Respiratory Diseases in Children: A Time Series Analysis. *Statistics in Medicine*, Vol. 10, (1991), 1481-1492.

[6] Rios , M., Garcia, J.M., Cubedo, M., Perez, D.: Time Series in the Epidemiology of Typhoid Fever in Spain. *Med. Clin.*, Vol. 106, Num. 18 (1996), 686-9.

[7] Fleming, D.M., Cross, K.W., Sunderland, R., Ross, A.M.: Comparison of the Seasonal Pattern of Asthma Identified in General Practitioner Episodes, Hospital Admissions and Deaths. *Thorax*, Vol. 8, (2000), 662-665.

[8] Saynajakangas, P., Keistinen, T., Tuuponen, T.: Seasonal Fluctuations in Hospitalisation for Pneumonia in Finland. *Int J Circumpolar Health*, Vol. 60, Num. 1 (2001), 34-40.

[9] Lani, L., Rios, M., Sanchez, J.: Meningococcal Disease in Spain: Seasonal Nature and Resent Changers. *Gac Sanit*, Vol. 15, Num. 4 (2001), 336-340.

[10] Cohen, P.: The Influence on Survival of Onset of Childhood Acute Leukaemia (ALL). *Chronobiol Int*, Vol. 4, Num. 2 (1987), 291-297.

[11] Harris, R.E., Harrel, F.E., Patil, K.D., Al-Rashid, R.: The Seasonal Risk of Paediatric/Childhood Acute Lymphocyte Leukaemia in the United States. *J Chronic Dis*, Vol. 40, Num. 10 (1987), 915-923.

[12] Csaki, P.: ARMA Processes. In: Tusnady, G., Ziermann, M. (eds): *Time Series Analysis*. Technical Publishing House, Budapest, 1986. 49-84.

[13] Arato, M., Benczur, A.: Exact Distribution of the Maximum Likelihood Estimation for Gaussian-Markovian Processes. In: Tusnady, G., Ziermann, M. (eds): *Time Series Analysis*. Technical Publishing House, Budapest, 1986. 85-117.

[14] Arato, M.: *Linear Stochastic Systems with Constant Coefficients: A Statistical Approach*. Springer, Berlin, 1982.

[15] Vienna, N.J., Polan, A.K.: Childhood Lymphatic Leukaemia Prenatal Seasonality and Possible Association with Congenital Varicella. *Am J Epidemiol*, Vol. 103, (1976), 321-332.

[16] Sorenson, H.T., Pedersen, L., Olse, J.H., et al. Seasonal Variation in Month of Birth and Diagnosis of Early Childhood Acute Lymphoblastic Leukaemia. *J. A. M. A.*, Vol. 285, 168-169.

## Postal addresses

**Mária Fazekas**
*Department of Economic-*
*and Agroinformatics*
*University of Debrecen*
*138, Boszormenyi Street, Debrecen,*
*Hungary*