ARIMA models for the analysis of the Precipitation Evolution

ALINA BĂRBULESCU and ELENA PELICAN
Department of Mathematics and Computers Science
Ovidius University
124, Mamaia Blv., Constanța, 900527
ROMANIA
alinadumitriu@yahoo.com http://alina.ilinc.ro/ID_262/index.html
epelican@univ-ovidius.ro http://csam.univ-ovidius.ro/~epelican

Abstract: In this article the temporal characteristics of precipitation evolution are investigated. The data base is formed by annual mean precipitation collected in the period 1965-2003, at 10 meteorological stations, in Dobrudja, a region of Romania. The dependence analysis shows that nine of these series don't have the long-range dependence property and are stationary. Their distributions have been studied and it has been found that four of ten series are Guassian noises. For the rest, AR or FARIMA models have been proposed and parameterized. It also has been shown that predictions based on these models are very close to the data available for the years 2004, 2005.

Key-Words: - ACF, correlation, long range dependence, precipitations, series, stationarity

1 Introduction

Weather modification is a topic of substantial worldwide interest for all countries. Building models and testing their validity is a step in understanding and predicting the weather evolution, which is a challenge for all scientists. Only a small number of studies is devoted to precipitation evolution in different regions of Romania, including the Black Sea coast [1, 2]. As a consequence, this article comes to complete the analyses made for a region of Romania.

Dobrudja is situated in the South – East of Romania, between the Black Sea and the lower Danube River.

Dobrudja's structure (excluding the Danube Delta) is that of a plateau with hilly aspect, with an average altitude between 100 and 180 m. Its climate is temperate - continental.

Table 1. The coordinates of meteorological stations

Station	Latitude	Longitude	Elevation (m)	1965-2003 average precipitation
Tulcea	+45:11	+28:49	4.36	452.94
Jurilovca	+44:46	+28:53	37.65	376.24
Corugea	+44:44	+28:20	219.20	417.59
Harsova	+44:41	+27:57	37.51	394.31
Cernavoda	+44:21	+28:03	87.17	475.07
Medgidia	+44:15	+28:16	69.54	438.75
Constanta	+44:13	+28:38	12.80	410.77
Adamclisi	+44:08	+28:00	158.00	474.82
Mangalia	+43:49	+28:35	6.00	438.76
Sulina	+45:09	+29:39	2.08	261.63

Researches showed that the frequency of droughty years is 89 %, the longest rainless period in this area being registered in the South of Dobrudja and the Black Sea coast [8].

The studied data represent mean annual precipitation, collected for 39 years at 10 meteorological stations, starting to 1965. The coordinates of these stations and the averages of precipitation over this period are given in Table 1 [8] and the data are represented in Fig.1.

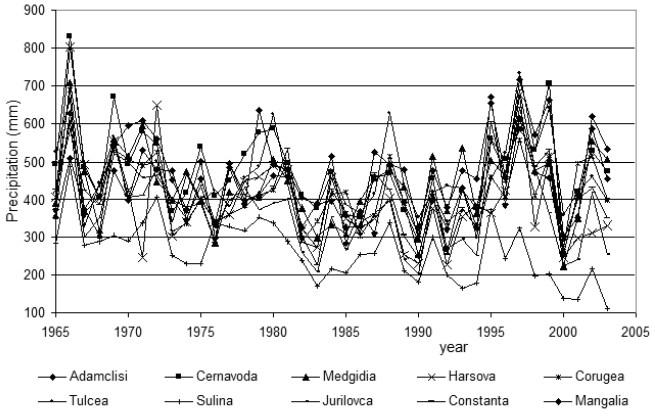

Fig.1. The mean annual precipitation at the studied stations in the period 1965 - 2003

2 Methodology

In order to obtain some models for the precipitation time series, the following steps were done:

- The analysis of long range dependence and the determination of Hurst coefficient;
- The analysis of break presence;
- Testing the hypothesis that the series are Gaussian

noises;

- The determination of a model for each time series if it doesn't form a Gaussian noise;

- The comparison of the models obtained for different stations.

To discuss our results some notions concerning the time series analysis are necessary.

A discrete process (a time series) in time is a sequence of random variables $(X_t)_{t\in \mathrm{N}^*}$ (shortly denoted by (X_t)).

The process (X_t) is said to be weakly stationary (or stationary) if it has a finite mean and the covariance depends only on the lag between two points in the series.

A stationary process (X_t) is called a white noise if X_t, X_{t+h} are uncorrelated for every $h \neq 0$, identically distributed, with the null expectance, and the same variance.

A discrete process is said to be autoregressive of *p* order, and is denoted by AR(*p*), if:

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + Z_t, \forall\, t \in \mathrm{N}^*, \varphi_p \neq 0,$$

where (Z_t) is a white noise.

A discrete process is said to be a fractionally integrated ARMA process, or more precisely an FARIMA(*p*, *d*, *q*) process, with $0 < |d| < 0.5$ if:

$$(1-B)^d \phi(B) X_t = \theta(B) Z_t,$$

where $\phi(z)$ and $\theta(z)$ are polynomials of degrees *p* and *q*, respectively, satisfying $\phi(z) \neq 0$ and $\theta(z) \neq 0$ for all *z* such that $|z| \leq 1$, *B* is the backward shift operator, and (Z_t) is a white noise with the mean zero and variance σ^2. The operator $(1-B)^d$ is defined by the binomial expansion

$$(1-B)^d = \sum_{j=0}^{\infty} \pi_j B^j,$$

where $n_0 = 1$ and

$$\pi_j = \prod_{1<k\leq j} \frac{k-1-d}{k},\ j = 1,2,\ldots \text{ [4].}$$

A time series (X_t) has the long range dependence property if it has correlations that persist over all time scales.

Using different methods [5] the analysis of long-range dependence property of studied time series was performed and the self-similarity parameter *H,* which measures the intensity of long range dependence in a time series, was determined. If the process is persistent, *H* is greater than 0.5.

A break in a time series is a change of probability low at a certain moment [6]. The break tests permit to detect a change in a time series mean. The methods used for this purpose were: the Pettitt test [7] and the segmentation procedure of Hubert [8, 9], since they work even if the series are not normally distributed. In the situations of contradictory results of these tests, Buishard test and change point analysis were also performed.

The Pettitt test is a non-parametric one, the null hypothesis to be tested being:

H_0: There is no break in the series (X_t).

Hubert's segmentation procedure is able to detect the multiple breaks in time series. The method also gives the instants of the breaks.

In addition, the following tools were used:

- Kolmogorov – Smirnov, and Jarque - Bera tests, and Q - Q plots – to test the normality;

- the autocorrelation functions and Box-Ljung statistics – to test the correlation [10];

- the Levene test - to test the homoscedasticity hypothesis, i.e.:

H_0: the series has the same variance.

3 Results

For all the series Hurst's coefficients were determined. Only the value corresponding to Sulina station (0.6) was bigger than 0.5. Thus this is the only series that has the LRD property.

The results of the break tests are presented in Table 2, where "yes" signifies that the hypothesis of breaks' absence is accepted (at the confidence level of 95%, in the case of Pettitt test) and "no", followed by a year represents the point where the break is present.

Table 2. Results of break tests

Station	Pettitt	Hubert
Adamclisi	yes	yes
Cernavoda	yes	yes
Corugea	yes	no, 1972
Constanta	yes	yes
Harsova	yes	no, 1969
Jurilovca	yes	yes
Mangalia	yes	yes
Medgidia	yes	yes
Tulcea	yes	yes
Sulina	no, 1981	no, 1981

In 70% of cases, there is no break in the time series. Since in the case of Corugea and Cernavoda stations only one of the tests rejected the hypothesis of break absence and there is only a few data before the break points determined by the segmentation procedure, the models will not take into account two sub-periods.

The results and the models for all the series are presented in the following.

3.1 Adamclisi series

The existence of correlation in the data series was studied using the autocorrelation function (Fig.2).

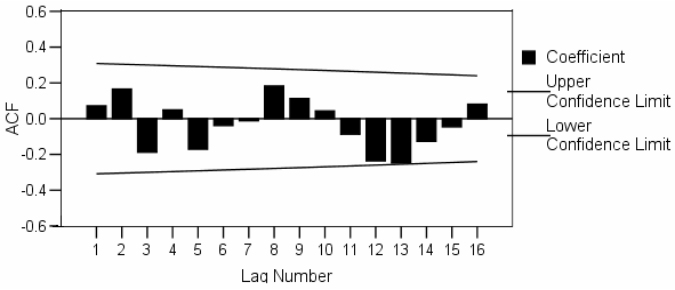

Fig.2. ACF of Adamclisi series

Since all the absolutes values of ACF are between the confidence limits, at the confidence level of 95%, the autocorrelation hypothesis was rejected.

In order to prove the data normality, the Q-Q plot of precipitation sample was analysed. Since the points that represent the observed values are distributed around the straight line representing the expected normal values (Fig.3), the hypothesis of the time series normality was accepted.

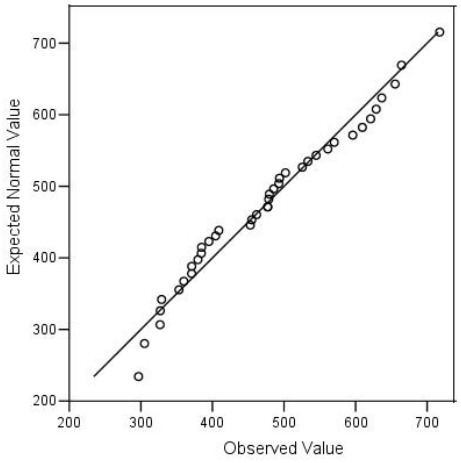

Fig.3. The Q-Q plot of precipitation series at Adamclisi station

To test the series for homoscedasticity, let us denote by:

- N – the selection volume (in our case 39);
- k – the number of groups in which the sample is divided (in our case $k = 3$);
- n_i - the number of data in each group (in our case, $n_i = 13$, $i = 1, 2, 3$);
- s_i^2 - the selection variance of the group i, $i = \overline{1,k}$;
- s_p^2 - the pooled estimate for the variance:

$$s_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)s_i^2;$$

$$X^2 = \frac{(N-k)s_p^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2}{1 + \frac{1}{3(i-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right)}.$$

If $X^2 < \chi^2_{1-\alpha}(k-1)$, at the significance level α, then the hypothesis H_0 is accepted.

In our case,

$$\alpha = 0.05,\ \chi^2_{1-\alpha}(i-1) = \chi^2_{0.95}(2)$$

is the quintile of χ^2 distribution, and $i - 1 = 2$ is the degree of freedom.

Since

$$X^2 = 1.2211 < 5.991 = \chi^2_{0.95}(2),$$

the homoscedasticity hypothesis is accepted.

The change point analysis was also performed. The CUSUM chart (Fig. 4) comes to confirm that there is no change in the series evolution.

Fig. 4. CUSUM chart for Adamclisi series

Concluding, at a reasonable confidence level (95%), the data are independent, identically and normally distributed. So, the process is a Gaussian noise.

This result was expected, since the estimation H = 0.488 was obtained for this station.

3.2 Cernavodă and Medgidia series

It was proved that Cernavodă and Medgidia series are also Gaussian noises.

It was expected, since the values of H were respectively 0.507, 0.492.

3.3 Corugea series

Since the segmentation procedure of Hubert indicates a break in 1972, the cumulated sum chart has also been studied (Fig.5).

The CUSUM chart shows that in 1972 there is a change in the mean of data series. After performing the Buishard and Pettitt tests we accept the null hypothesis:

H_0: There is no break in the time series

even at a confidence level of 90%. (Fig.6)

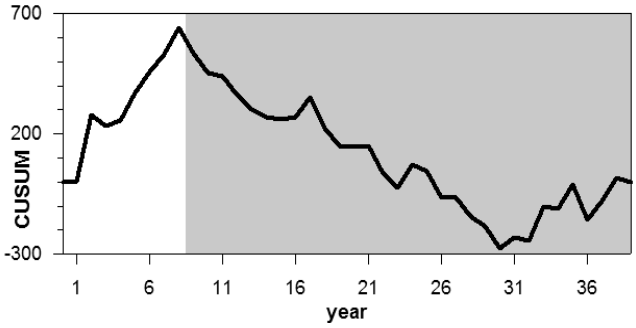

Fig. 5. CUSUM chart for Corugea series

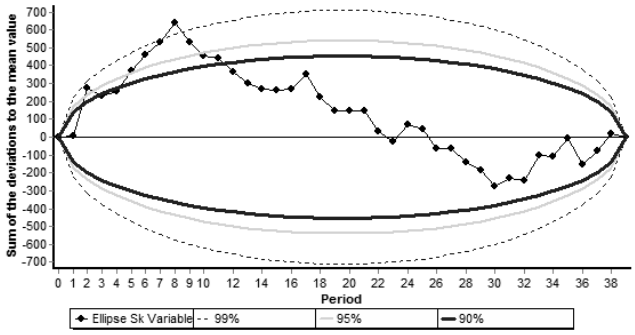

Fig.6. Bois'ellipse associated with Buishard test

The value of X^2 statistics is 2.0178 < 5.991, so the series has the homoscedasticity property.

The value of Jarque – Bera statistic (3.825) and the corresponding *p*-value (0.14771) lead us to accept the normality hypothesis.

The values of Ljung – Box and McLeod - Li statistics (8.4159, respectively 4.5781) and the corresponding *p* - values (0.9887, respectively 0.9998) come to confirm the hypothesis that the series is independent and identically distributed.

Since Hurst's coefficient is 0.491, we can conclude that the series forms a Gaussian white noise.

3.4 Corugea series

The result of the Kolmogorov – Smirnov test leads us to accept the hypothesis that the data are normally distributed and correlated (Fig. 7).

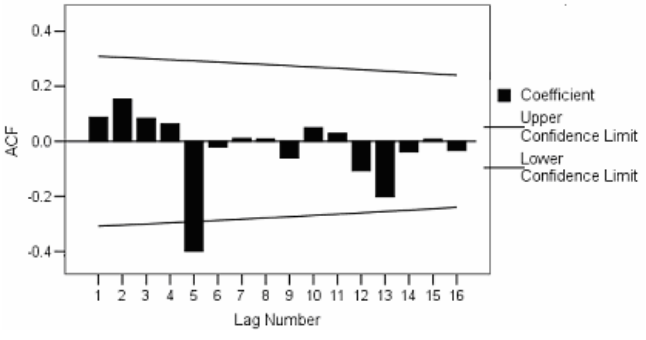

Fig.7. ACF of Constanta series

For the series transformed by taking logarithms, and denoted by (Y_t), the best model was:

$$Y_t = 0.9783Y_{t-1} + Z_t,$$

where (Z_t) is a white noise with the variance 0.0456.

3.5. Harsova series

Applying the Jarque - Bera test, we reject the hypothesis that the data are normally distributed.

The hypothesis that the series is not correlated is also rejected at the confidence level of 95%.

Studying the CUSUM chart associated we reject the hypothesis that there is a break in the series.

Taking the data logarithms, the resulted series, (Y_t), is also correlated, but normally distributed. The following AR(1) model was determined:

$$Y_t = 0.9786Y_{t-1} + Z_t,\ t \geq 2,$$

where (Z_t) is the residual, which is a Gaussian white noise with the variance 0.202.

3.6 Jurilovca series

After the application of Kolmogorov - Smirnov test and the study of ACF we accept the hypotheses that the data are normally distributed and correlated. Applying a logarithmic transformation, we fit an AR(1) model (Fig.8) for the transformed series, (Y_t). It has the equation:

$$Y_t = 5.885 + 0.316Y_{t-1} + Z_t,\ t \geq 2,$$

where (Z_t) is the residual, which is a white noise with the variance 0.083.

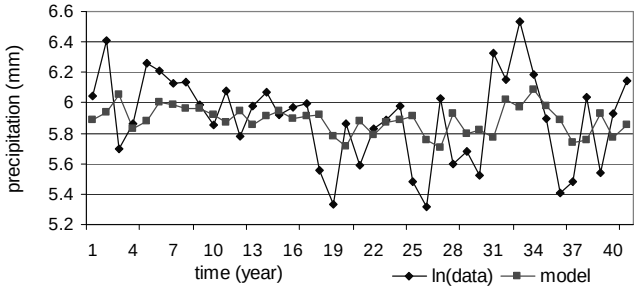

Fig.8. The model for Jurilovca station

3.7 Mangalia series

After testing the normality hypothesis and analysing the ACF function at the confidence level of 95%, we accept the hypotheses that the data are normally distributed and correlated.

After taking logarithms, an AR(1) model (Fig.9) was fitted for the transformed series, (Y_t). Its equation is:

$$Y_t = 0.09698Y_{t-1} + Z_t,\ t \geq 2,$$

where (Z_t) is the residual, which is Gaussian,

independent and identically distributed, with the variance 0.1079.

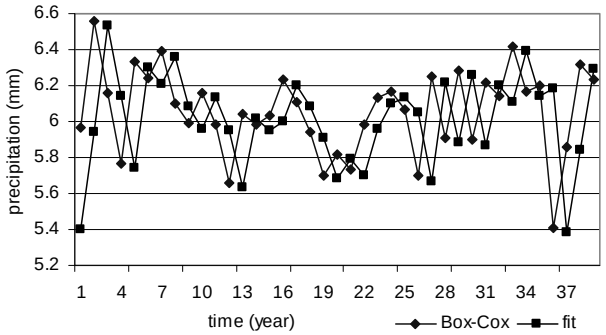

Fig.9. The model for Mangalia series after a Box – Cox transformation

3.8 Tulcea series

Testing the normality and correlation hypotheses, we reject the hypothesis that the series is normally distributed and we accept the hypothesis that it is not correlated. The same tests made after taking the square root of data, allow us to accept the hypotheses of normality (Table 3) and independence for the series $X_t' = \sqrt{X_t}$.

Table 3. Resuls of Kolmogorov-Smirnov test

Kolmogorov-Smirnov			Shapiro-Wilk		
statistic	df	Sig.	statistic	df	Sig.
0.093	39	0.200	0.974	39	0.506

The portmanteau test leads us to accept the hypothesis that series (X_t') is a Gaussian noise.

3. 9 Sulina series

Sulina station is situated at 8 km offshore, so the precipitation variation is not similar to that registered inside Dobrudja.

The series is normally distributed, homoscedastic, but dependent and has the long range property. The results of all the break tests, including Lee and Heghinian, Buishard and CPA, lead us to accept the hypothesis that there is a break in 1981. As a consequence we tried to determine a model for the entire series and two models for the subsequent periods: 1965-1981 and 1982-2003.

In Fig.10, Pettitt's test U – variable evolution is represented.

For entire series Sulina, two types of models were proposed:

a. After considering a Box – Cox transformation ($\lambda = 0$) and the mean subtraction:

$$X_t = 0.4021 \cdot X_{t-1} + Z_t,$$

where (Z_t) is a white noise. (Fig.11)

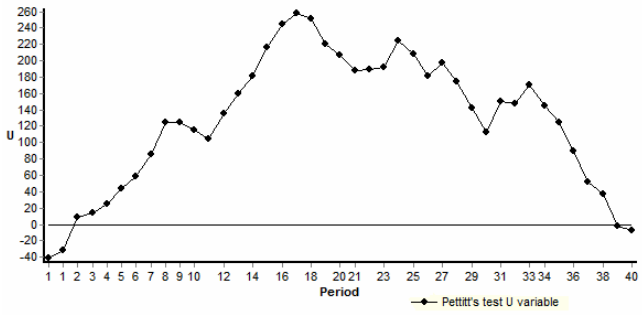

Fig.10. Pettitt's test U variable

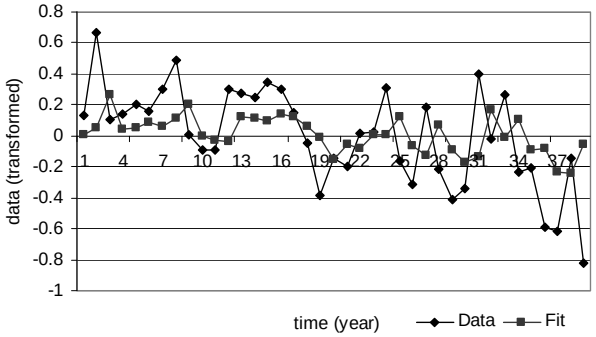

Fig.11. Model for Sulina series after a Box – Cox transformation

b. After the mean subtraction:

$$(1-B)^{0.28} X_t = Z_t,$$

where (Z_t) is a white noise, with the variance 4618.11.

The second one is of FARIMA type and better fit the data than the previous one, so it was used to forecast the mean annual precipitation evolution (Fig. 12).

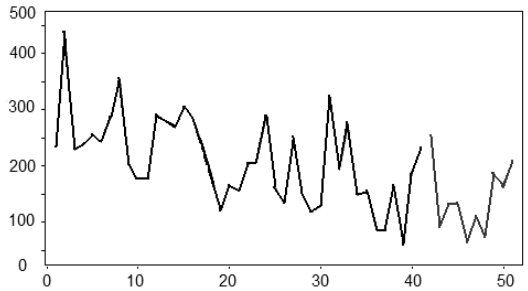

Fig.12. Forecast of Sulina annual series

The subseries Sulina_1 (1965 - 1981) is Gaussian, independent and identically distributed, with the variance 3795.05.

For the subseries Sulina_2 (1982 – 2003), after the mean subtraction, the best model determined was a Gaussian noise, with the variance 4296.75.

4 Conclusions

The analyses made on the data from 10 meteorological stations for the period 1965 – 2003 show that the series of annual precipitations didn't register high variability in that period.

Four series are Gaussians noises, nine are normally distributed and one has the long range property.

Between the models, four were of AR type and one of fractionally integrated ARIMA type. The best of them were selected in concordance with the Akaike' selection criterion, after the analysis of different ARIMA models.

The models were used to predict the mean annual precipitations for 2004 and 2005 and the results were compared with the measurements. They are satisfactory and sometimes better than those obtained using other types of methods [11].

References:

[1]V. S. Barabanov, V. V. Efimov, and M. V. Shokurov, Modeling of the Specific Features of Climate in the Black - Sea Region, *Physical Oceanography*, vol. 12, no. 4, 2002, pp. 10 -23

[2]V. V. Efimov, M. V. Shokurov, and V. S. Barabanov, Statistical Modeling of Monthly Anomalies of Atmospheric Precipitations for the Region of the Ukraine and Black Sea, *Physical Oceanography*, vol.12, no.1, 2002, pp. 191 - 199

[3] C. Maftei and A. Barbulescu, Statistical analysis of climate evolution in Dobrudja region, *Lecture Notes in Engineering and Computer sciences*, WCE 2008, vol.II, IAENG, pp. 1082-1087

[4] P. Brokwell, R.A. Davis, *Introduction to time series and forecasting*, Springer, 2002, ch. 5 and ch.10

[5] M. Taqqu and V. Teverovsky, On Estimating the Intensity of Long-Range Dependence in Finite and Infinite Variance Time Series, in *A practical Guide to Heavy Tails: Statistical Techniques and Applications*, R.J. Adler, R.E. Feldman and M.S. Taqqu, editors, Birkhauser, Boston, 1998, pp. 177 – 217

[6] H. Lubes, J.M. Masson, E. Servat, J.-E. Paturel, B. Kouame, and J.F. Boyer, Caractérisation de fluctuations dans une série chronologique par applications de tests statistiques, *Rapport* n^o 3, *programme* ICCARE, 1994, 21 pp.

[7] A. N. Pettitt, A non - parametric approach to the change-point problem, *Applied Statistics* vol. 28, no.2, 1979, pp. 126 – 135

[8] P. Hubert, J.P. Carbonnel and A. Chaouche, Segmentation des séries hydrométéorologiques. Application à des séries de précipitations et de débits de l'Afrique de l'Ouest, *Journal of Hydrology*, 110, 1989, pp. 349 – 367

[9] P. Hubert and J. P. Carbonnel, Segmentation des séries annuelles de débits de grands fleuves africains, *Bulletin de liaison du* CIEH 92, 1993, pp. 3 - 10

[10] D.J. Seskin, *Handbook of parametric and nonparametric statistical procedures*, Chapmann and Hall/CRC, Boca Raton, 2007, ch. 12

[11] A. Barbulescu, E. Pelican, On the Sulina Precipitation Data Analysis Using the ARMA models and a Neural Network Technique, *Proceedings of the 10th WSEAS International Conference on Mathematical and Computational Methods in Science and Engineering*, MACMESE'08, Part II, WSEAS Press, 2008, pp. 508-511

Acknowledgements. The research was supported by Grant ID_262/2007