The Pearson's correlation - a measure for the linear relationships between time series?

Werner R.[1], Valev D. [1], Danov D.[2]

[1] *Solar-Terrestrial Influences Institute, Bulgarian Academy of sciences, Stara Zagora, Bulgaria.*
[2] *Solar-Terrestrial Influences Institute, Bulgarian Academy of sciences, Sofia, Bulgaria*

In many studies the correlation coefficient is used to characterize the relationship between different time series. To bring out long-time relations, the time series are usually filtered by a moving average procedure. However, the time series are autocorrelated and by the average procedure the autocorrelation is increased in comparison with the correlation of the non-averaged series. This paper demonstrates the influence of the moving average on the correlation coefficient and the relation to the causality is discussed. It is concluded that in the case when as a consequence of the high autocorrelation the linear model is not adequate, it is under question whether there is any sense to specify the linear correlation coefficient.

Introduction

Many research papers on space physics, solar-earth interaction, helio physics, climate sciences and geophysics employ statistical methods as a result of the lack of concrete physical models to examine the relation between two variables. Therefore, the focus is on finding relations between different metric scaled variables using a correlation as a measure for the strength of the relationship of the variables and/or a regression analysis is applied to determine the direction of the relationship. However, in several papers the mathematical requirements connected with the application of the correlation and the regression analysis are not controlled. Some studies have shown linear relationships between the averaged quantities without giving any indication for the variability of the averaged quantities or at least for the number of data points that have been included into the averages [1]. What is more, in many cases relationships between different time series are studied, which are typically autocorrelated.

In the first place, this paper reminds the conditions and the properties of the Pearson's correlation and the linear regression, after which we demonstrate the influence of the averaging on the correlation on a sample of climate time series.

Correlation and linear regression

Basically, as a measure for the correlation, the Pearson product-moment correlation coefficient[1] is used and is usually called a correlation coefficient. The two variables between the linear relations should be determined and should be randomly distributed. The Pearson product-moment correlation coefficient exists for any bivariate probability distribution for which the population covariance is defined and the marginal population variances are defined and are not zero. The t-test and the F-test can be applied to test the hypothesis whether it is a linear relationship between the variables X and Y or not, or the hypothesis about a hypothetical value of the correlation coefficient. Both tests are based on the assumption that the variables are normally distributed and both tests are sensitive to the deviations from the normal distribution. The normality of the distributions of the samples X and Y can be examined, for example, by the help of the Chi-Squared, Kolmogorov-Smirnov, Shapiro-Wilkins and the Lilliefors Goodness-of-Fit Tests (see [2] and the citations herein).

The value of the correlation coefficient is in the interval from -1 up to +1, where for values smaller than zero the negative sign stands for anti-correlation and the values larger than zero denote a positive correlation. A correlation of zero does not mean that there isn't a relationship between the variables. It only means that a linear correlation does not exist, however, the relation can be quadratic or it's of other functional relations. For a very strong correlation, even for the absolute value of one, the relation is not necessarily of causality.

Since the nature of the correlation coefficient makes it a measure for the linear correlation between two variables, it is closely connected with the linear regression analysis, where it not only gives the answer of the question whether the relationship is linear or not, but it also gives a linear relation between the two variables, the linear regression equation or, shortly, the regression. In the multivariate case of the mostly used type of regressions, considered here, the dependent variable Y (also called a regressant, or a predicted variable) should be continuous and with a normal distribution, but the distribution of the explanatory variables X (also called predictors, regressors or an independent variable) is not necessarily random. The predictors should be linearly independent and are assumed as error-free. In the climate science, the linear coefficients are also called forcing parameters. The error term $\varepsilon = Y - \hat{Y}$, where Y is the vector of the observations and $\hat{Y}$ - the vector of the values, estimated by the linear model, has the distribution $N(0,\sigma)$, which means: $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. The errors have to be uncorrelated, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ and have to be independent from the regressor X and the estimations, i.e. $\hat{Y}$ $Cov(X, \varepsilon) = Cov(\hat{Y}, \varepsilon) = 0$. The assumptions for normal distribution of the regressor and for the errors allow to estimate the confidence intervals of the regression coefficients and of the regressor prediction intervals as well as to prove some hypotheses, based on the Student, Fisher or χ^2 test. The error variance of all observations must be constant, which means that they are homoscedastic, and can be tested by the Levene test, Barlett test, Cochran test or by the F_{max} test. The correlation coefficient

[1] It is also named Bravais-Pearson product-moment correlation coefficient.

$$\rho = Cov(X,Y)/\sqrt{Var(X)Var(Y)} \qquad (1)$$

is related to the slope b of the linear model by :

$$b = \rho\sqrt{Var(Y)/Var(X)}\,. \qquad (2)$$

It is easy to calculate the variance of the error term using the definition of the correlation coefficient and to obtain the formulae for the estimation of the correlation coefficient r

$$r = \sqrt{\frac{Var(\hat{Y})}{Var(Y)}}\,. \qquad (3)$$

With the last equation the coefficient of determination r^2 is defined as the ratio of the explained variance to the total variance, and describes which part of the total variance can be explained by the used regression. The last two equations demonstrate the close connection between the correlation and the bivariate linear regression. In the case of multivariate linear regression, the adjust coefficient of determination must be applied.

Similarly to the correlation, a strong linear relationship does not implicate causal connections. High correlation can be spurious or nonsense. We speak about spurious or about illusory correlation or regression when a third variable influences the two correlated variables.

Application of the regression method to time series

Up to the present the understanding that the correlation is not influenced by changes over the time was implicit. This means that the data are in an equilibrium state. This is especially important when the regression is used for the purpose of prognoses. The regression method is also applicable for analysis of time series. Here time series are series of a sequence of data points, very often equidistantly spaced in time. They are usually additively or multiplicatively separate in a trend, a cyclical and/or a seasonal component and an irregularly (noise) component. The trend, the cyclical as well as the seasonal component are sometimes combined with a smooth component. The trend of a time series can be analyzed by a linear regression as a simple polynomial function of a not very high degree p. Another method to determine the trend is based on moving averages, on wavelet decomposition of the series or on the application of specific filters in both, the time and the frequency space. By moving average the variance of the averaged time series $\overline{y}_t$ in relation to the original series y_t is decreased and in the case of central moving average it is determined by the weighting coefficients w_i

$$Var(\overline{y}_t) = Var(y_t)\sum_{i=1}^{2q+1} w_i^2\,, \qquad (4)$$

where 2q+1 is the odd number of averaged consecutive observations.

A linear trend in a time series can also be separated by the determination of the first differences and of a polynomial of a higher degree and, consequently, by differences of higher orders.

The cyclic and seasonal components can be determined by a harmonical analysis in the time space or by spectral methods in the frequency domain. In many science disciplines the relationships between different time series often focus the interest in order to find connections between the variables. In the case of only one regressor or predictor, the dynamical linear regression equation can be written as

$$Y_t = \alpha + \beta X_t + \varepsilon_t\,. \qquad (5)$$

The Y_t and X_t are now pairs of Y and X measured at a moment *t*. Equation (5) is the same as the equation of a simple linear regression. Naturally, the process is now not in equilibrium. Moreover, the regressor Y and the predictor X are typically autocorrelated. The autocorrelation function

$$\rho_{t,\tau} = \frac{\mathrm{cov}(y_t, y_{t-\tau})}{\sigma^2(y_t)} \qquad (6)$$

is defined similarly to equation (1), only the index *i* for the sample pair $[y_i,x_i]$ is replaced by the time index t and t-τ, where τ is the index number of the time series y_t shifted to itself. Sometimes the unnormalized term $cov(y_t,y_{t-\tau})$ is called autocovariance. The autocorrelation of Y and/or X is reflected on the autocorrelation of the error term and the assumption (5) is not valid. Now we will discuss the consequences of the autocorrelation, following the description in [3]:

1. If the regression parameters are estimated by the ordinary least square method, the estimations of the parameters are unbiased (i.e. they are not changed due to the errors autocorrelation), however, the estimations are not efficient. This means that the estimated confidence intervals are influenced as a result of the autocorrelations.
2. If the error autocorrelation and the predictor autocorrelation are of the same order, the variance of the errors is biased. The sign of the differences between the estimated and the real value depends on the signs of the autocorrelations. Normally, the positive sign prevails. Then the estimation of the error variances is that they are too small and the determination coefficient and the F-value are overestimated. The bias decreases with increasing the time series length.
3. The calculation of the variance of the slope $Var(\hat{\beta})$, which is easy to be made from the difference of the real value β and its estimation $\hat{\beta}$, in the simple bivariate case gives $Var(\hat{\beta}) = \sigma_\varepsilon^2/\sum_{i=1}^{n}(x_i - \overline{x})$. Thus, the biased variance of the error term produces also a biased variance of the slope. In the case of an autoregression of the first order for the bivariate linear regression Hibbs [4] gives the approximation

$$Var(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \overline{x})}\left(\frac{1+\gamma\lambda}{1-\gamma\lambda}\right)\,, \qquad (7)$$

where γ and λ are the regression coefficients of the autocorrelation series of the residual and the predictor series, respectively. The term in the brackets on the left side of the equation can be interpreted as a correction term describing the

deviation of the variance, not taking into account the autocorrelation.

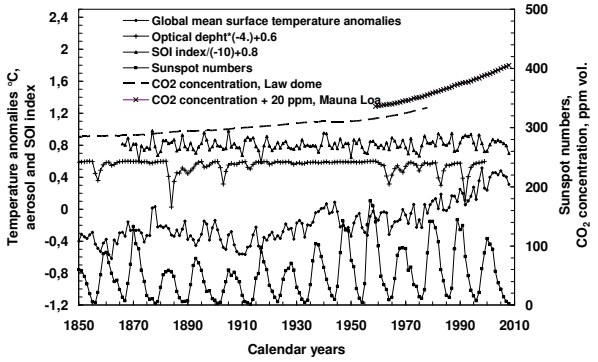

Fig.1. Time series of the global mean annual temperature anomalies and some of the important temperature impact factors.

The above discussion underlines the importance of the graphical inspection of the error term for the estimation of the stationarity of the error term and to see whether the error term contains a deterministic trend, (a remain of) a trade cycle and/or seasonal cycle and it answers the question for their autocorrelation. A very common test for autocorrelation of the error term is the Durbin-Watson test.

Many studies involve the application of the moving average as a low pass filter of the time series and the determination of correlation coefficients is applied directly without detrending the series. Sometimes even the time series used to study the correlation are filtered in a different manner. The preceding detrending of the series is very important because the existing trends produce an inflation of the series correlation. If two series are strongly linearly time-correlated, the correlation between these series will be very high. However, this high correlation is not conclusive in regards to the causality of the processes. Conclusions for their causality can be drawn only in the case when the variabilities (of the de-trended series) are correlated with one another. In other words, with the classical regression method applied to the time series, it is possible to study only the correlation of the short-term variations of the series.

Time series example – Data use

Now we will demonstrate the influence of the averaging interval length of the moving average on the correlation with and without detrending the time series by the help of an example from the climate sciences. A basic parameter in the climate science is the global annual mean surface temperature (T) depending on the time. Here we will use the data set HADCRUT3 taken from the Met Office Hadley Centre for Climate Change. A download of the data set is to be found at: http://hadobs.metoffice.com/hadcrut3/diagnostics/global/nh+sh/annual. In the climate sciences, the temperature series are often compared to the series which are related to physical measures driving the earth climate such as the solar sunspot number, the CO_2 mixing ratio, the aerosol index and the South Oscillation index (SOI) ((http://www.cru.uea.ac.uk/ftpdata/soi.dat), based on the difference between the Sea Level Pressure at Darwin and Tahiti (http://www.cru.uea.ac.uk/ftpdata/soi_dar.dat, http://www.cru.uea.ac.uk/ftpdata/soi_tah.dat), related to El

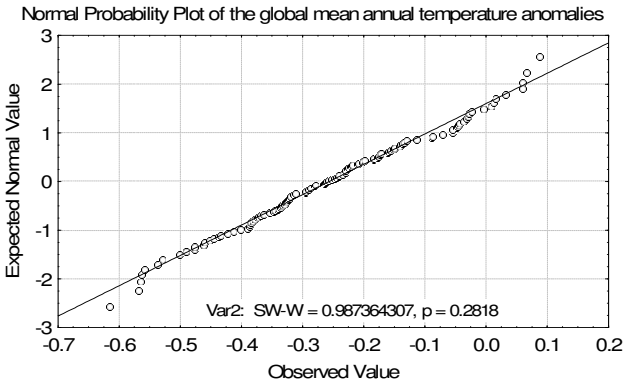

Fig. 2a. Normal probapility plot of the global mean annual temperature anomalies from 1865 up to 1993.

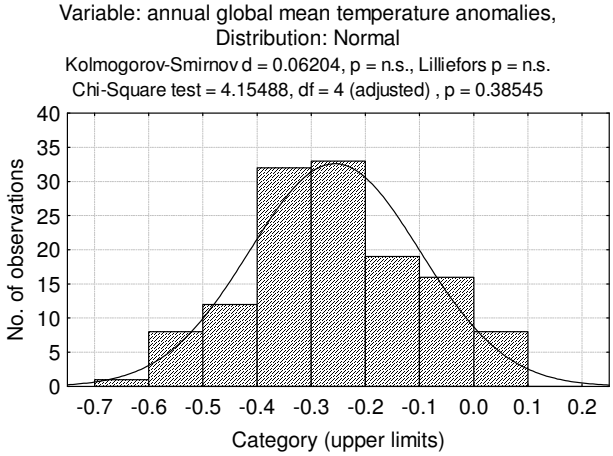

Fig.2b. Histogram of the same time series as in Fig. 2a and the fitted normal distribution. Results of the statistical tests of the normal distribution.

Niño events. The sunspot number is an index describing the variation of the solar irradiance, one of the measures related to the solar activity. The increasing CO_2 mixing ratio during the industrial period is one of the forcing factors of the global earth climate, by irradiation absorption of the earth in the infrared spectral range (greenhouse effect). A lot of aerosol particles are emitted as a result of the volcanic activities, which change the atmospheric absorption properties expressed by the optical depth. http://data.giss.nasa.gov/modelforce/strataer/tau_line.txt. The annual time series, the global temperature, the sunspot number, the CO_2 concentration ratio μ and the atmosphere optical depth tau and the SOI index are shown in Fig.1. The aerosol data and the CO_2 concentration determined from the ice core drilling of Law dome are taken from the Historical data related to the global climate change, compiled by Wm. Robert Johnston (updated 25 March 2008) and are to be found at:
http://www.johnstonsarchive.net/environment/co2table.html.
Since the annual Law dome CO_2 data span the time interval from 1850 up to 1978, additionally CO_2 data is used from the measurements at Mauna Loa, covering the time span from 1959 up to 2008:
ftp://ftp.cmdl.noaa.gov/ccg/co2/trends/co2_annmean_mlo.txt
The CO_2 time series in Fig. 3. is the logarithm of the ratio μ / μ_0 with μ_0=280 ppmv, where 280 ppm is used as a pre-

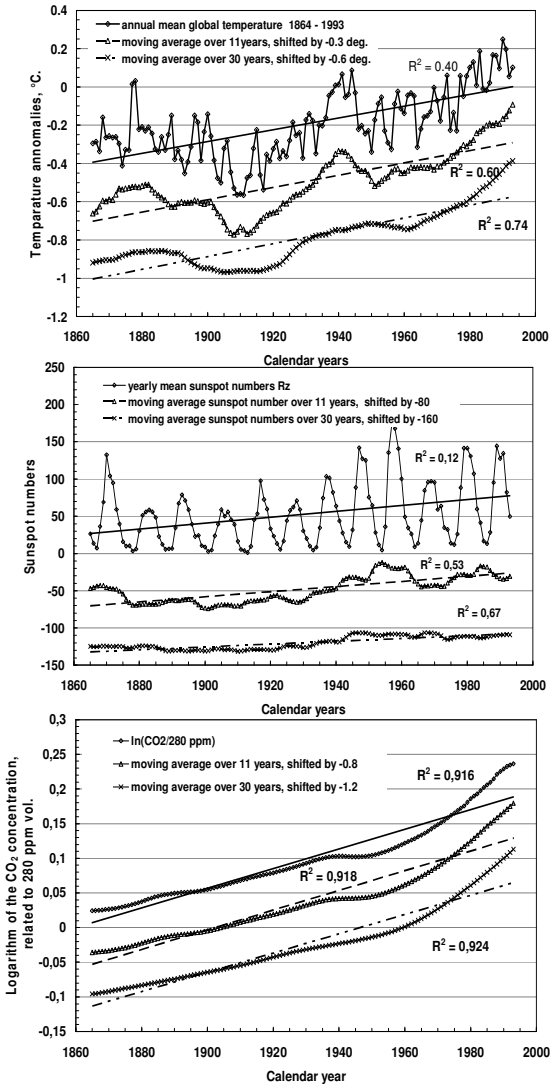

Fig.3. The original time series of the global mean annual temperature anomalies (on top), the sunspot numbers (in the middle), the logarithm of the CO_2 concentrations (bottom) and their 11 year and 31 year moving average as well as their linear trends, respectively.

industrial value. It is well known that CO_2 does not force the global temperature linearly. The CO_2 climate forcing term is described rather by the form of its logarithm which is a result of the saturation effect in the CO_2 absorption bands [5].

Analysis and Discussion

The aerosol indices vary sporadically. The higher atmosphere aerosol content leads to reduction of the atmosphere transparence for the solar radiation, which causes a temperature decrease (or, a compensation effect of the temperature increase) at a time scale of some years. This is one of the reasons why the global temperature over the examined time interval does not show a clear Gaussian distribution over the whole interval from 1850 up to 2008, and another reason is probably the rising of the global temperature by external forcing. If we constrain on the time interval from 1865 up to 1993 (as explained below), and if we do not take into account the strong warming period during the last two decades, the normality can not be rejected for the global temperature distribution. Fig. 2 shows the normal probability plot for the same distribution and the results of different statistic tests are presented, using the statistic program STATISTICA 6. We will determine the correlation coefficients for the bivariate relation of the global annual mean surface temperature and the sun spot numbers, on the one hand, and the CO_2 concentration, on the other hand. The sun spot and the CO_2 concentration are used here for their long time variations, which are studied in relation to the global, mean temperature in a multitude of scientific papers. Here at first the correlations are determined for the annual time series and then for the 11 year and also of the 31 year averaged time series as well as those of the linear detrended series. The averaging over 11 years is used to outline variations longer than the solar cycle variations [6,7,8]. The averaging over 30 (here for simplification 31) years is chosen to suppress variations shorter than the climatic epochs [9]. To compare the results using the same sample numbers in every series we constrain to the use of the time interval from 1865 up to 1993 after providing the averaging procedures, to exclude edge effects.

It is very important that the time series, included in the correlation and/or regression analysis are filtered in the same manner. It has been proven that the frequently cited finding of a strong correlation between the solar cycle length and the mean global temperature by Friis-Christensen, E., Lassen [10] is not correct and was produced by an implication of different filter techniques on the time series [11,12].

From the two time series for CO_2 we have composed one general time series, where the Law dome values after 1978 were predicted by the measured Mauna Loa CO_2 concentrations by the linear regression equation obtained by the values of the overlapping series part. Fig. 3 shows the original series of the global temperature anomalies, the sun spot numbers and the CO_2 concentrations, the moving averaged series and the corresponding linear fits. Close to the linear trend lines, the determination coefficients of the time series and their linear fit are given. It is clearly seen that the absolute deviations of the series from their fits decrease by stronger averaging and the determination coefficients consequently increase. As a result of the averaging, however, the autocorrelation of the series also rises. For example, the autocorrelation coefficient to the lag 1 of the annual temperature series is 0.737 and is reduced to 0.651 by linear detrending. However, the autocorrelation coefficient at lag 1 of the averaged time series of the global temperature exceeds the value of 0.96, it is also close to 1 and the autocorrelation function decreases very slowly, outlining the non-stationarity of the series (also in the case of the detrended temperature series). According to some authors, which „naively“ determine the correlation coefficients, we have determined the them and the determination coefficients for the bivariate multivariate series with using the sunspot numbers and the CO_2 forcing as predictors while the temperature is used as a predicted variable. The results for the annual time series, averaged with and without detrending, are summarized in Table 1. The determination coefficient of the relation between the temperature and the sunspot numbers for the annual time series is very low and the slopes of the relations are not significant even in the case when the autocorrelation effects are neglected. This is in agreement with the well

known fact that the direct impact of the 11 year solar cycle signal on the global climate is very weak. The global mean temperature variations at the time scale of some years are produced predominantly by the El Niño events expressed by the SOI index. The temperature variations are also influenced by the atmosphere aerosol and the sulfate concentration, by the concentrations of the green house gases other than CO_2, methane for example, and probably by other climatic factors. Table 1 shows that the correlation coefficients strongly increase if the time series are averaged over longer time intervals and decrease in the multivariate case for the detrended time series.

Table.1.
Correlation and determination coefficients

	Correlation coefficients and determination coefficients between the global mean temperature anomalies and					
	the sunspot numbers (SSN) (bivariate)		the CO_2 concentration (bivariate)		the SSN and CO_2 (multivariate)	
	non detrended series	lin. detrended series	non detrended series	lin. detrended series	non detrended series	lin. detrended series
annual series	0.135 0.018	0.056 0.003	0.558 0.311	0.227 0.051	0.559 0.312	0.229 0.053
11 year moving average	0.729 0.531	0.381 0.145	0.842 0.708	0.556 0.309	0.887 0.750	0.690 0.477
31 year moving average	0.855 0.732	0.475 0.226	0.927 0.859	0.704 0.496	0.947 0.897	0.887 0.787

Conclusions

Smaller correlation coefficients were obtained for the detrended time series than for the non-detrended ones. This leads to the conclusion that the high correlations are produced by the averaging procedures. This fact is confirmed by the analysis of the residuals of the time series and their linear model estimations. The residual series from the multivariate linear model T(SSN,CO_2) for the 11 year averaged series is shown in Fig. 4 as an example. The residuals are far from random and are strongly non-stationary. The residual series are highly auto correlated. For the shown example, the estimated correlation coefficient r_e = 0.967. The correlations for the predictors r_{SSN} = 0.985 and r_{CO2} = 0.967. The Durbin-Watson statistics gives the quantities d_{SSN}=0.06 and d_{CO2}=0.008. For the significance level of 0.95 the critical lower value $d_{129,2,0.95}$ is 1.71. Therefore, d is much smaller than d_L and the hypothesis that H_0: ρ=0, which means that the series are not autocorrelated, has to be rejected. These high autocorrelations generate very strong biased estimations of the slope variances. We have found for the correction terms in equation (14) values of the order of 20 up to 30 for the slope variance of the linear models for the averaged time series and the slopes either reach the limit of significance or are not significant. As a consequence, the linear models are not adequate and it is under question whether the specification of the correlation coefficients is justified. The models have to be changed to include predictors in order that the model is more adequate. If the residuals are random, their autocorrelation can be studied and the regression parameters and the confidence intervals can be estimated by the application of the Cochrane-Orcutt-method [13] or by the bootstrap method [13].

Acknowledgement

The authors want to acknowledge to the Ministery of Education and Science to support this work under the contract DVU01/0120.

REFERENCES

[1] EOS, Forum, Vol. 90, Number 32, 11 August 2009.
[2] J. Hartung, B. Elpelt, K.-H. Klösner, Statistik. Lehr- und Handbuch der angewandten Statistik, R. Oldenbourg Verlag München Wien, 2005.
[3] H. Thome, Zeitreihenanalyse. Eine Einführung für Sozialwissenschaftler und Historiker, R. Oldenbourg Verlag München Wien, 2005.
[4] D.A. Hibbs, Problems of statistical estimation and causal inference in time-series regression models. In H.L.Costner (ed.) Sociogical methodology, San Francisco et.al., pp. 252-308, 1974.
[5] G. Myhre, E.J Highwood, K.P. Shine, F. Stordal, New estimates of radiative forcing due to well mixed greenhouse gases, Geophysical Research Letters, Vol. 25, NO. 14, pp. 2715-2718, 1998.
[6] S.K. Solanki, M. Fligge, A reconstruction of total solar irradiance since 1700, GRL 26, 2465–2468, 1999.
[7] J.L. Lean, J. Beer, R.S. Bradley, Reconstruction of solar irradiance since 1610: Implications for climate change, Geophysical Research Letters, 22 (23), pp. 3195–3198, 1995.
[8] N.A. Krivova, S.K. Solanki, Solar variability and global warming: a statistical comparison since 1850, Advances in Space Research, 34, pp. 361-364, 2004.
[9] K.Georgieva, C. Bianchi, B.Kirov, Once about global warming, Mem.S.A.It., Vol. 76, pp. 969-972, 2005.
[10] E. Friis-Christensen, K. Lassen, Length of the Solar Cycle: An Indicator of Solar Activity Closely Associated with Climate, Science, 254, pp. 698–700, 1991.
[11] Laut, P., Gundermann, J., "Solar cycle lengths and climate: A reference revisited", J. Geophys. Res., 105, pp. 27 489–27 492, 2000.
[12] R. E. Benestad, A review of the solar cycle length estimates, Geophys. Res. Lett., 32, L15714, 2005.
[13] W.G. Cochran, G.H Orcutt, Application of least squares regression to relationships containing autocorrelated error terms, Journal of the American Statistical Association, vol. 44, pp. 32-61, 1949.
[14] Efron, B., R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, 1994.