## Introduction

Statistical correlation between different time series provides some insight about the degree of natural correlation of data (e.g. due to their location). Moreover, the statistical correlation techniques are also applied when there are missing values in a time series that have to be filled in or when a data sample is extended backward or forward in time.

The correlation (or regression) techniques are executed separately, before the infilling or extension procedure; in this two-step procedure, in the first step the appropriate coefficients have been estimated, while in the second they are used to infill or extend the time series. Hydrognomon offers amongst others a multi-regression optimization method in order to estimate the optimal coefficient set per season, based on the criteria set by the user. Two more issues to be addressed is the seasonality of the variables and the maintenance of the statistical characteristics of the initial sample – that is before infilling or extension.

## Estimation techniques for the regression coefficients

In the section the estimation techniques for the regression coefficients which are implemented in Hydrognomon are featured. These methodologies cover both the need to estimate the degree of correlation between the time series and the needs for the expansion and infilling of the series.

### Mean

The mean of the time series can be used to fill in missing values ​​when no other possible connection can be found that could help to this cause. However, the application of this method leads to underestimation of the standard deviation of the sample filled in, which can be overcome by adding a random term to the values estimated (see section 7.3.1). The application of the mean technique can be represented in terms of regression by the coefficient *b* set to zero and the constant term *a* equal to the sample mean:

 7.1

This procedure can be applied mainly in time series with monthly time step, taking into account seasonality, i.e. counting twelve different values ​​for the constant term *a*. (see section 7.2.6). The methodology for the estimation of the mean can be found in the section 9.2.

### Simple – multiple regression

The simple linear regression is the well known method of line fitting between of single points at the Cartesian plane. In order to achieve that the principle of the least squares is used; i.e. the estimated line minimizes the sum of the squared distances between the line and the points. According to this estimation, a constant term *a* and a slope *b* is determined, which are the terms of the equation:

*y* = *a* + *b* *x* 7.2

where *x* is the independent variable and y the dependent. In the case that this correlation model is used for data infilling, then variable *x* will come from the time series of reference, while the variable *y* will refer to the time series with the missing values. The parameters of the simple linear correlations are given by the following relationships:

 7.3

and:

 7.4

where *n* is the number of common (simultaneous) measurements between the dependent variable *y* and the independent variable of *x*.

In the case of several independent and just a single dependent variable, an optimal solution is estimated for the parameters *a*0...*a*m of the equation:

*Υ* = *a*0 + *a*1*X*1 + ... + *amXm* 7.5

where *Y* is the dependent variable and *X*1...*Xm*, are the *m* independent variables. The solution is achieved through multiple linear regression. In order to carry out the regression all the included time series, both independent and dependent, should share a common measurement time period, during which there are no missing values in all time series. Therefore, from now on the term *time series* will be used for the subset of the time series of the common period. Similarly, any reference to the dispersion or the mean of the time series, will correspond only to the dispersion or the mean *of the values of the common period shared*.

The parameters *ai* are estimated by the following relationships (Christofides, 1994, p 14):

**** 7.6

where  the mean of time series *x*, s2(*x*) the variance of time series *x*, s(*xy*) the covariance of time series *x* and *y* and *d(ij)* the elements of the matrix:

**** 7.7

The unbiased value of the covariance of any two variables *x* and *y* is :

 7.8

Similarly to the mean, the relationships for the estimation of the variance can be also found in the section 9.2.

The above methodology is used to implement both single and multiple linear regression in Hydrognomon, as the former can be regarded as a simple subcase of the later; thus in both cases the determination coefficient is calculated from the generalized formula:

 7.9

where *Wi* is the error between the dependant variable and the regression model value, *yi* the dependent variable, *n* the dependent variable sample size and  the mean of the dependent variable. The correlation coefficient *r* is the square root of the determination coefficient *δ*.

### Homogenous line

In many cases it is desirable the restriction *a*0 = 0 to be set. In this case the rest coefficients are estimated by the same formula as above:

 for *i* = 1, ..., *m* 7.10

except that s(*x*(*j*)*y*) is not the covariance but the sum of products of values ​​of the time series (Christofides, 1994, p 16). This holds true both for the above equation and the equation 7.7. If there is a single independent variable, the slope of the line passing through the origin is given by the relationship (Koutsoyiannis, 1997):

 7.11

Again, the determination coefficient *δ* is estimated by the generalized formulation 7.9.

### Organic correlation

The organic correlation is mainly applied to the extension of samples when the maintenance of the initial statistical characteristics of the original (pre-infill) sample is desired. The correlation coefficients are estimated by keeping the same mean and variance as the original sample; in this case the requirement of minimizing the mean square error is abandoned. This method for estimating the coefficients is known as organic correlation or Maintenance Of Variance Extension – MOVE.1 (Koutsoyiannis, 1997, Christofides, 1994).

The application of the organic correlation is achieved through an independent variable *x*. In this case, the parameters *a* and *b* of the equation:

*y* = *a* + *b* *x* 7.12

are estimated by maintaining the statistical characteristics of the original sample. The coefficients are given by (Koutsoyiannis, 1997, p 202):

 7.13

 7.14

where *sx*, *sy* the sample values of the standard deviation of the variables *x*, *y* and *rxy* the sample correlation coefficient:

 7.15

where *sxy* the sample covariance of the variables *x* and *y*. The determination coefficient is estibated by relationship 7.22 (the outcome is exactly the same if the generalized relationship 7.9 is used).

### Autocorrelation

Hydrological time series often exhibit autocorrelation, i.e. there is a relationship between each value and the values which precede it (Christofides 1994, Christofides, 1998). Namely:

*yi* = *a*0 + *a*1*yi*-1 + *a*2*yi*-2 + ... + *akyi*-*k* + *ε* 7.16

where *ε* a random term of a given distribution. This relationship, wherein each value is associated with *k* predecessors, is called *AR(k) model*, where *k* is called the *order* of the model. The determination of the coefficients *αi* is exactly the same as the determination in linear regression; in this case the independent times series are the displacements of 1, 2, ..., *k* positions of the time series (time series is correlated with itself *k* times, with a lag of 1, 2, ... *k*). Thus, if the time series exhibits seasonality, there will be 12 independent correlations (e.g, in monthly time series: correlation of all Januarys with all Decembers, all Februarys with all Januarys, etc.) and the model is called PAR(*k*).

Autocorrelation is sometimes combined with cross-correlation, as follows:

*yi* = *a*0 + *a*1*yi*-1 + *a*2*yi*-2 + ... + *akyi*-*k* + *ak*+1*x*i(1) + *ak*+2*x*i(2) + ... + *ak*+*mxi*(*m*) 7.17

Here the independent variables are *k*+*m*; these are the *k* displacements of time series *y* plus the other *m* time series.

This is implemented in the Hydrognomon environment by introducing *m* time series as independent variables of the cross-correlation and *k* time series as independent variables of the autocorrelation (the later are time series of the dependent variable diplaced by 1, 2, ... *k* time units). This method can be used in missing data infilling of monthly river discharge values, because they exhibit strong statistical autocorrelation ​​ (see also Efstratiadis, et al, 2000).

### Seasonality

Several of the hydrometeorological variables exhibit seasonality; a periodic pattern in their values linked to the annual alternation of the seasons. Seasonality is simulated by splitting the time series in twelve subsets, each one corresponding to a calendar month. Thus, the ​​regression parameters are estimated separately for each independent monthly subset of values; the methodology presented below, will be applied twelve times, as twelve samples are formed by the splitting of the annual time series.

After their estimation, the regression parameters are represented as a matrix consisted of twelve rows (one for each month) and a number of columns according to the number of independent time series. If there are missing values to be completed, then the appropriate parameters which correspond to the missing value month are used.

This methodology should not be used for time series with annual time step and thus the user is prevented from applying it. In the case of time series with time step smaller than monthly, monthly parameters will be calculated based on the dates of the time series values. The estimation of parameters in smaller time scales than the monthly (e.g. daily) are beyond any meaning because in such scales autocorrelation becomes even stronger and the intermittent behavior of some short term-periodic phenomena is introduced (mainly dominated by the alternation of day and night).

### Scatter diagram

The scatter diagram gives an overview of the regression model fit to the actual data. The creation of the chart is a simple procedure for the case of linear regression of two variables (either of simple regression, zero constant term or organic). Each variable is depicted in one of the two axes x, y, regression is a line described by the simple formulation *y* = *a* + *bx*, while the values ​​of the time series depicted as a pair of conjugate points (*xi*, yi). An example can be seen in Figure 7.1 for the rainfall cross-correlation of the Aliartos and Ag. Triadia stations during February. In the case of seasonality, we shall have 12 independent samples and therefore an equal number of scatter plots.

Multiple regression results should be depicted in *n*+1 dimensions, depending on the regression degree (*n*). If *n* = 1 (simple regression) the scatter diagram can be illustrated in 2 dimensions; however if *n* = 2 (two independent variables) the scatter diagram has to be illustrated in three-dimension space, and hence the more the independent variables the more the scatter plot dimensions. This display problem is anticipated by a simple transformation in order to produce a two-dimensional scatter plot; nevertheless in this way a certain amount of information is lost.



Figure 7.1: Scatter plot of simple linear regression of monthly rainfall for Aliartos and Ag. Triadia stations during February (*δ*=0.83).



Figure 7.2: Scatter plot of multiple linear regression of monthly rainfall during April. The dependent variable is Aliartos station while the dependent are Amfissa, Atalanti and Ag. Triadia stations (*δ*=0.90).

In the case of multiple linear regression, the regression relationship is formulated as *y*=*a*0+*a*1\**x*1+*a*2\**x*2+… . If *x*1=*x*2=…=*xm*, where *xm* is the mean rainfall of the independent variables, then the above formulation is depicted as a straight line, and the points *x*1*i*, *x*2*i*, *x*3*i*, …, *xni*, *yi­* of the *n*+1-dimensioned which correspond to the measurements at *n* +1 stations are transformed into points of 2D space as follows:

 7.18

An example is demonstrated in Figure 7.2. The dependent variable (values ​​of rainfall at Station Aliartos) is positioned on the *y* axis, while the mean rainfall of the stations of Amfissa, Atalanti and the And Ag. Triada is positioned on the *x* axis. The sample taken corresponds to the rainfall for April.

Note that in the case of multiple linear regression the scatter diagram, unlike the case of simple linear regression, gives only o portion of the total information part and is used mainly to provide qualitative measure of the good-of-fitness degree. Therefore, it should be used critically in conjunction with the value of the determination coefficient and the existence or not of near zero values of linear coefficients for any individual variables.

### Optimization

The problem of regression optimization between a dependent station *y* and various independent *xi* is often encountered. This problem becomes even more complicated in the case of seasonality, because the determination of optimal solution has to be achieved for each month separately. The optimization measure is the determination coefficient. It must be underlined that for each independent variable introduced, there is a corresponding increase in the determination coefficient and thus the optimization problem would be meaningless if there were not some restrictions set. Such restrictions are:

A maximum time series size *k*. If *n* is the total size of the time series then *k*≤*n*. *k* should be taken as 1 (simple linear regression) or 2-3 in order to have a better oversight in the correlated stations.

* A unique time series *i* is mandatorily used in a given month set.
* A unique time series *i* is not mandatorily used in a given month set.

Negative parameters are allowed or not.

In the example below, the optimal regression parameter sets are estimated for each month limited by some restrictions. The dependent variable is the time series of rainfall in Aliartos station, while the independents include time series from stations 1: Ag. Triada, 2: Amfissa, 3: Ano Bralos, 4: Atalanti, 5: Athan. Diakos, 6: Davlia. In the following window (Figure 7.3), the user can set the appropriate restrictions.

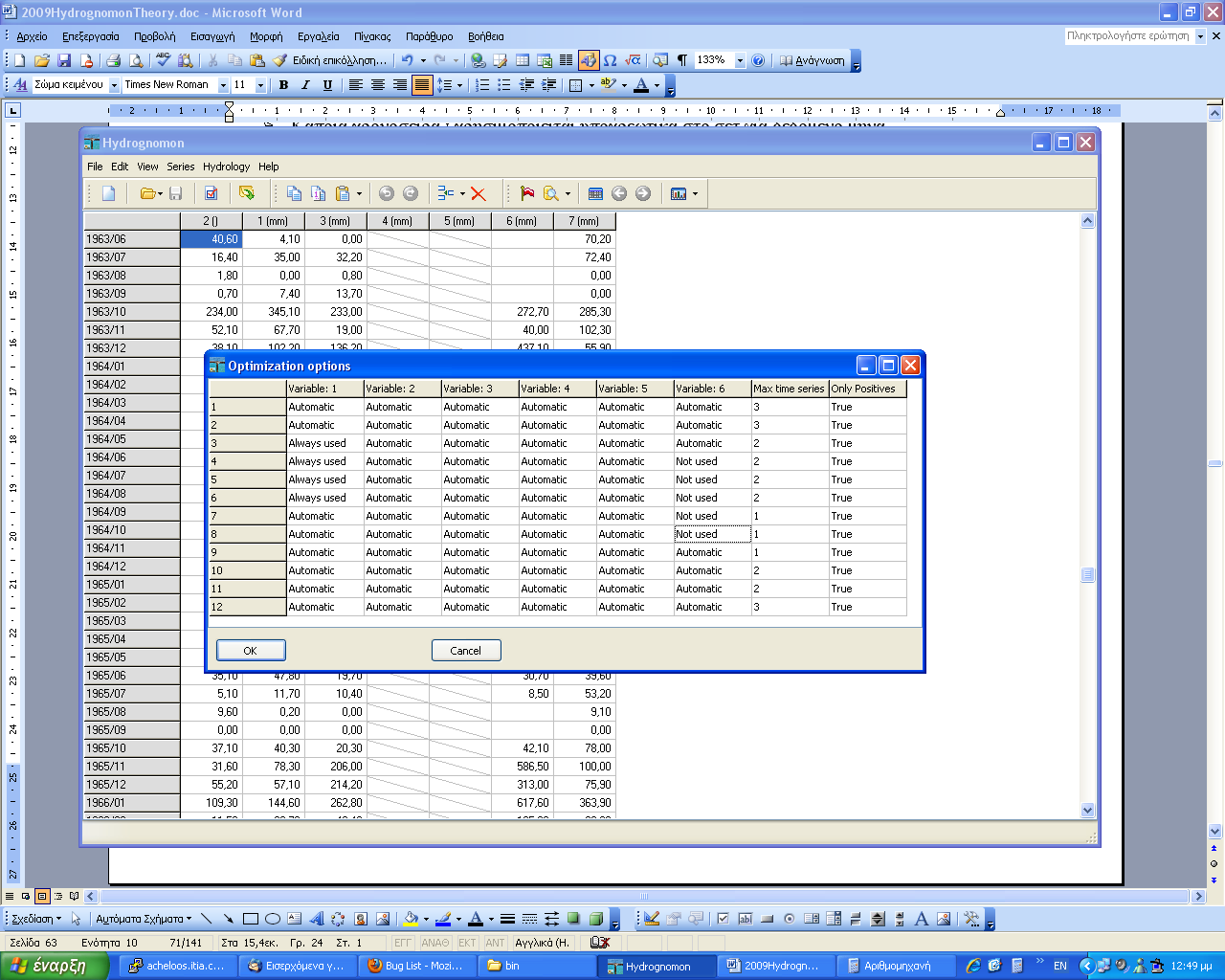


Figure 7.3: Optimization restrictions window.

In this example, one can see that there are at maximum three time series for December to February, at maximum two time series for the months of March to June and October to November and at maximum one time series for the months of July to September. The variable 1 (Ag. Triada) is mandatorily used from March to June, while the variable 6 (Davlia), is not necessarily used for the months from April to August. Where the setting is ‘Automatic’ a test is performed to find the optimal solution. The negative coefficient values ​​are not allowed (‘Only Positives’ column). The result of the optimization is illustrated in the Table (7.1):

Table 7.1: Multiple regression results by optimization technique.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *n* | *δ* | *r* | *δ*cr | *a*0 | *a*1 | *a*2 | *a*3 | *a*4 | *a*5 | *a*6 |
| Oct | 25 | 0.854 | 0.924 | 0.160 | 1.859 | 0.327 |  |  | 0.342 |  |  |
| Nov | 25 | 0.855 | 0.925 | 0.160 | -0.679 |  |  |  | 0.684 |  | 0.233 |
| Dec | 25 | 0.805 | 0.897 | 0.160 | -5.208 | 0.279 | 0.070 |  | 0.517 |  |  |
| Jan | 24 | 0.846 | 0.920 | 0.167 | 7.671 | 0.232 |  | 0.109 | 0.337 |  |  |
| Feb | 24 | 0.860 | 0.927 | 0.167 | -19.925 | 0.563 |  | 0.032 | 0.189 |  |  |
| Mar | 24 | 0.696 | 0.834 | 0.167 | 6.027 | 0.262 | 0.277 |  |  |  |  |
| Apr | 25 | 0.887 | 0.942 | 0.160 | -10.533 | 0.576 | 0.241 |  |  |  |  |
| May | 25 | 0.693 | 0.832 | 0.160 | 4.261 | 0.160 |  |  | 0.732 |  |  |
| Jun | 22 | 0.929 | 0.964 | 0.182 | -0.927 | 0.384 |  |  | 0.504 |  |  |
| Jul | 17 | 0.631 | 0.794 | 0.235 | 1.972 |  |  |  | 0.668 |  |  |
| Aug | 24 | 0.823 | 0.907 | 0.167 | 0.419 | 0.727 |  |  |  |  |  |
| Sep | 22 | 0.710 | 0.843 | 0.182 | 2.251 | 0.566 |  |  |  |  |  |

All the limitations posed are met, while the variable 5 was not utilized (Athan. Diakos station).

## Data infilling – extension

Often time series contain missing values, that is e.g. because of organ failure or mistypes in data processing. These gaps may hamper certain data processing techniques, so it is desirable to fill them in. The infilling, even if it's rough estimate with large error margins, can be very helpful, especially when one wants to extract a time series of smaller resolution. For example, if a monthly time series of rainfall is missing the month of January, then it is not possible to estimate the annual rainfall because there available data covers only 11 months. In order to calculate the annual sum the application of a suitable infilling technique for January rainfall is needed. Even though the estimation error could be large, its influence to the annual value will be small. It must be noted though, that data infilling should be used only when there were occasional data gaps, e.g. up to four (4) months in a year or ten (10) days in a month for daily time series.

However, when there are extensive time series from neighboring stations it is possible to fill longer missing data gaps or even estimate values ​​beyond both ends of the existing time series (before and after the initial sample). This process is called *data extension* of the time series and is essentially the same process with data infilling - estimation of values that ​​do not exist - but has a different purpose and therefore is based on a different technique. Data infilling is applied only to a few values, in order to efficiently aggregate a time series to larger scales and thus the accuracy of the estimate is not very meaningful. On the other hand, an extended time series may lead to a better estimate of the statistical parameters or its properties are closer to the properties of the natural process which represents. Therefore, there should be no deterioration of its statistical characteristics, which can be achieved by the application of the organic correlation technique (it preserves the statistical properties of the sample), always in conjunction with the introduction of a random term (see section 7.3.1). The infilling and extension theory is described by Koutsoyiannis (1997, p. 227-236).

In any case, for the infilling/extension to be accomplished, there must be one or more time series from the neighboring stations. These time series shall be called *independent time series*, while the one to be filled in/extended are the *dependent time series*. For each time moment which is going to be filled in/extended there must be values in every other independent time series (if even one value is missing no infilling/extension will be made at that time moment).

The time period which is consists of all the missing values is called *infilling period* or *period 2*, while the time period in which all time series, independent and dependent, appear to have values in common time moments is called *common period* or *period 1.* The data size of common and infilling period is denoted as *N*1 and *N*2, respectively.

All the techniques that were demonstrated in the previous section (simple or multiple linear regression, organic correlation, autocorrelation, etc) can be used to implement extension/infilling, with or without consideration of seasonality. Each value that comes from extension/infilling, should be flagged with the *infilling* flag, in order to be identifiable.

A 5% importance factor should be set or else the regression would be meaningless (Koutsoyiannis, 1997, p. 233) and the correlation coefficient should be at least:

 7.19

The above inequality can also be written in terms of determination coefficient as:

 7.20

It is obvious that the former relationships have meaning only for *N*1>4. In the case of the organic correlation an upper barrier is also introduced:

 7.21

As the determination coefficient of the organic correlation is:

*δ* = 2|*rxy*|-1 7.22

then the above restriction can be formulated as:

 7.23

It should be noted that the relationship (7.22) is the one demonstrated in the outcome of organic correlation in the Hydrognomon.

### Random term introduction

A problem encountered in the infilling of the time series using simple or multiple linear regression is the underestimation of the variance of the sample filled in. One way around this problem is the introduction of a random term *ε* to the infilling values ​​in order to keep the variance at its initial pre-infilling value (Christofides, 1994). In this technique the random term used is the one which was first proposed by Matalas and Jacobs (1964):

 7.24

where *z* the values taken from a normal distribution random number generator (see section 9.7.1), *R* the (multiple) linear correlation coefficient and *a* is given by the relationship:

 7.25

where *m* is the number of reference time series (independent variables), *Ν*1 the number of values measured at the same time moments in both the dependent and the independent variables and *Ν*2 the number of missing values. The introducing random term is meaningless if the size of missing values is small; in this case *z* may not fit adequately to the normal distribution. If the initialization of the random number generator is at a specific value each time, then the same results are reproduced. However, the user may wish to initialize the generator at a random value.

The technique of the random term introduction can also be used in the place of the organic correlation; it may retain more efficiently the original correlation coefficient. The main disadvantage of this methodology is the non-repeatability of the results due to the randomization factor.

A computational approach of the above process is the estimation of the statistical characteristics of the measured error *ε* (see Koutsoyiannis, 1997, p. 238); an approach also used in the Hydrognomon. For the general case of multiple linear regression the measured errors are:

 7.26

where *y* is the dependent variable and *xj* the independent variables. If we estimate the statistical characteristics of *εi*, i.e. the standard deviation *σw* and mean *μw*, then the random error terms can be expressed by the normal distribution random number generator:

*ε* = *N*(*μw*, *σw*) 7.27

The value *μw* is (or should be) very small and can be regarded as zero.

### Truncate to zero

For physical variables in which negative values ​​are meaningless, such as rainfall, it is possible to truncate to zero all the negative values ​​that may show up after the infilling. Such are the cases in which the intercept of the linear regression is less than zero. The truncate is based on the relationship:

 7.28

This technique has very small impact to the standard deviation of the final time series and relatively small to the mean.