Bulletin of the Geological Society of Greece

Vol. 40, 2007

An innovative data mining procedure, using clean algorithm and factor analysis, for irregularly sampled temporal environmental data sets

Fakiris E. — Laboratory of Marine Geology and Physical Oceanography, Department of Geology, University of Patras

Papatheodorou G. — Laboratory of Marine Geology and Physical Oceanography, Department of Geology, University of Patras

Panagiotopoulos P. — Laboratory of Marine Geology and Physical Oceanography, Department of Geology, University of Patras

https://doi.org/10.12681/bgsg.17225

Δελτίο της Ελληνικής Γεωλογικής Εταιρίας τομ. XXXX, 2007
Πρακτικά 11ου Διεθνούς Συνεδρίου, Αθήνα, Μάιος 2007

Bulletin of the Geological Society of Greece vol. XXXX, 2007
Proceedings of the 11th International Congress, Athens, May, 2007

AN INNOVATIVE DATA MINING PROCEDURE, USING CLEAN ALGORITHM AND FACTOR ANALYSIS, FOR IRREGULARLY SAMPLED TEMPORAL ENVIRONMENTAL DATA SETS

Fakiris E.[1], Papatheodorou G.[1], and Panagiotopoulos P.[1]

[1] *Laboratory of Marine Geology and Physical Oceanography, Department of Geology, University of Patras*

Abstract

Environmental data are often irregularly collected in the time domain due to various reasons which affect the field sampling schedule. As a result, data sets with uneven time step and time periods with no measurements are frequently built. Many problems occur in such data sets when processed owing to that neither statistical nor spectral analysis methods can easily be applied to them without any specific pre-treatment. In our study it is demonstrated a unified methodological scheme especially designed to deal with incomplete and unevenly sampled temporal data sets. This method consists of the CLEAN algorithm and the Factor analysis. The proposed methodology is successfully applied to data sets that belong to two sampling sites of the Greek river Strimonas.

Key words: *Missing data, Fourier transform, CLEAN algorithm, Factor analysis, Environmental data*

Περίληψη

Οι περιβαλλοντικές βάσεις δεδομένων συχνά αντιμετωπίζουν τα προβλήματα της άτακτης δειγματοληψίας στον χρόνο και της έλλειψης μετρήσεων για κάποιες περιόδους. Το γεγονός αυτό εμποδίζει τη χρήση των κλασικών μεθόδων ανάλυσης χρονοσειρών, οι οποίες απαιτούν σταθερό χρονικό βήμα ενώ ταυτόχρονα τα χρονικά κενά εισάγουν δυσκολίες στην χρήση των περισσοτέρων μεθόδων πολυδιάστατης στατιστικής ανάλυσης. Η παρούσα εργασία προτείνει ένα πλήρες μεθοδολογικό σχήμα ανάλυσης χρονικών περιβαλλοντικών δεδομένων με δειγματοληπτική ανομοιογένεια, στο οποίο γίνεται χρήση του αλγορίθμου CLEAN και της Παραγοντικής ανάλυσης (Factor Analysis). Ο αλγόριθμος CLEAN έχει την ικανότητα να αναπλάθει τις αρχικές χρονοσειρές της βάσης δεδομένων χρησιμοποιώντας φασματική ανάλυση και να δημιουργεί καινούργιες με σταθερό χρονικό βήμα και έλλειψη κενών. Λαμβάνει χώρα δηλαδή τόσο συμπλήρωση των κενών της βάσης, όσο και «εξυγίανση» της δειγματοληψίας της. Η παραγοντική ανάλυση ομαδοποιεί τις μεταβλητές, ανάλογα με τον περιβαλλοντικό μηχανισμό από τον οποίο κάθε μια ελέγχεται και επιπλέον αποκαλύπτει τη χαρακτηριστική χρονική διακύμανση της κάθε ομάδας. Το συγκεκριμένο μεθοδολογικό σχήμα εφαρμόστηκε με πλήρη επιτυχία σε μια βάση υδροχημικών δεδομένων μεγάλης χρονικής περιόδου (1980-94) στον ποταμό Στρυμόνα.

Λέξεις κλειδιά: *Αλγόριθμος CLEAN, μετασχηματισμός Fourier, παραγοντική ανάλυση, περιβαλλοντικά δεδομένα.*

1. Introduction

Environmental data often suffer from uneven sampling ratio, due to a variety of reasons that modify researchers' field planning. Even though a steady sampling ratio intention usually exists, rarely can it be followed with precision. In addition, it is probable that the in situ sampling can not be effectuated simultaneously for a relatively big number of variables, resulting to time step inequalities between them. The above mentioned facts pose great difficulties when attempting to mine the collected data, as long as neither time series analysis nor many statistical methods can be applied to the later ones. Although methods of time series analysis have been fully developed, only a few examples exist regarding the application of these techniques to environmental data. In most cases, this is a result of the requirements of the spectral analysis techniques. These techniques are based on the fast Fourier transformation (FFT) and their major drawback is the requirement of evenly spaced time series. Thus, the environmental data must undergo a pre-processing process before study of temporal variation is undertaken. The simplest forms of pre-processing are the linear or polynomial interpolation of the dataset and the splines. More specifically for hydrochemical data, simple interpolation techniques have been used in order to fill missing values. The main disadvantage of those interpolation procedures is that they disregard the general periodicities of the time series, as they apply locally mathematical formulas, introducing artefacts into the original dataset. A number of methods have been proposed for solving the treatment of incomplete and unevenly spaced data problem without dominantly affecting the results. Among them the spectral approach using the CLEAN algorithm is the most effective one for reconstructing time series with large or occasional gaps and irregularities in their sampling ratio. CLEAN algorithm developed by Roberts *et al.* (1987) is able to recover effectively most of the lost information even for a significantly smaller number of data points (Vio *et al.* 1992). It has been successfully applied to the analysis of astronomical and geophysical data (Dreher *et al.* 1986, Duvall *et al.* 1984, Negi *et al.* 1990, Tiwari and Rao 2000). Negi *et al.* (1996) applied the CLEAN algorithm to time series of secular variation of dolomite abundance in deep marine sediments in order to study the various quasi-periodic earth processes, including mass extinction phenomena. Baisch and Bokelmann (1999) used the CLEAN algorithm to investigate temporal changes of elastic propagation velocities and more recently Helsop and Dekker (2002) used it in conjunction with Monte Carlo simulation to study palaeoclimatic data.

In this paper we present, for the first time as far as we know, an application of CLEAN algorithm to hydrochemical data sets in order to convert them into time series with steady time step and study the temporal variation of the governing processes affecting their general form. After the transformation of the data sets using the CLEAN algorithm, multivariate statistical techniques can be successfully applied in order to group the variables with comparable mechanisms controlling their temporal variations. It is essential to mention that these techniques are ineffective or even impossible to apply when not dealing with compact databases (blank free) and this makes suitable pre-processing of information a high priority issue. Among the multivariate statistical techniques the Factor Analysis was chosen, as the most effective one. Thus, we demonstrate a methodological scheme consisting mainly of CLEAN algorithm and Factor analysis and show that it can be classified as a very important data mining tool which gives great insight into temporal variation of hydrochemical processes. The whole procedure is unified and automated using the MATLAB programming software. A variety of algorithms and scripts have been deployed to execute all the mathematical, statistical and visualizing operations necessary for the proposed methodology. In order to test and validate the method, hydrochemical datasets from two sampling sites of Strimonas River, northern Greece, were analysed. The data were collected under the inspection of the Greek ministry of agriculture.

2. Materials and Methods

2.1. Study area

River Strimonas has been monitored intensively for 14 years (1980 to 1994). Over 300 samples were collected for each one of the two sites, Mirtinos and Sidirokastro, at approximately one-month intervals. All the samples were analyzed for nineteen different physical and chemical water parameters, common for both sites, such as conductivity (EC) , pH, chloride (Cl), sulphate (SO4), acidic carbonic (HCO3), total of anions and kations (TAK), sodium (Na), magnesium (Mg), calcium (Ca), SAR, degree of alkalinity of sodium (Alk), total hardness (TH), dissolved oxygen (DO), rate of saturation (SAT), nitrite (NO2), nitric (NO3) ammonium (NH4) total phosphorus (TP), temperature of water (Tw). These two datasets were collected with the same procedures and protocol and were analysed using the standardised methods for water quality analysis.

2.2. Methodology overview

The proposed methodological scheme consists of three main parts: 1) Quality testing and preparation of the data. This part includes testing of the sampling conditions and detrending of the variables with significant trend, 2) application of the CLEAN algorithm in order to cleanse the time series by making their time step regular and the sampling points common for all variables, 3) application of Factor Analysis to the "CLEANed" data sets. This step includes a) extraction of the Factor loadings in order to group the variables and to investigate the major processes that control the data structure, b) calculation of the factor scores in order to study the temporal variations of the factors and 4) visualization and interpretation of the results.

2.3. Quality testing and pre-treatment of the data

2.3.1. Sampling conditions

Prior to the application of the analysis procedures a more detailed study of the sampling conditions is needed. As to visualize the ordinance of the sampling points within the time domain and judge whether they are intentional or not, the indicator function is used (Stefanakos and Athanasoulis 2001). This function is defined by:

Equation 1 – Indicator function

$$u(r) = \begin{cases} 1, & \text{if variable value at r is obtained} \\ 0, & \text{if variable value at r is missing} \end{cases}$$

and describes the existing value pattern of the measured data. Plotting of the former function offers a clear view of the sampling ratio intensions in real time and furthermore the percentage of the missing values can be estimated after averaging, using a steady time window, equal to the approximate sampling frequency. The reasons why the indicator function is a basic part of the proposed method will be analytically discussed in paragraph 2.4.2.

Moreover, in order to study the correlation between missing values and seasons, an existing-value diagram is also constructed. The existing-value diagram represents the existing-value seasonal averaging, plotted against time. The existing-value function is defined as:

Equation 2– Existing-value function

$$\bar{u}(r^a) = \sum_{j=1}^{j} u\left(j, r^a\right)$$

where $u(j,r^a)$: is the indicator function reindexed using the Buys-Ballot double index
j : seasonal index (3 months)
r^a : ranges within a calendar year.

2.3.2. Detrending process

Before continuing with the data mining procedure i.e. application of the CLEAN algorithm and the Factor analysis technique, it is necessary to detrend the time-series properly. Detrending process provides a suitable dataset for factorial analysis. The application of factor analysis to the detrended data correlates time-series with similar periodicities rejecting the correlation between time-series on the basis of their similar trends. It is also generally accepted that detrending improves spectral analysis results. There is a variety of methods to perform detrending processes. Notwithstanding, the dominant trend of the data can be removed just by fitting simple linear regression (Raike *et al.* 2003). We choose the linear regression for our detrending process in view of its simplicity and suitability concerning our purpose. A main issue of that particular procedure is to detect variables with significant trends. Judgeless detrending of all our temporal data, according to their linear regression, would be baseless without knowing whether their trends are reliable or not. In our case, a trend is defined as the presence of a non parametric rank correlation between a variable and the relevant time. As a correlation coefficient we use Kendall's tau-b, which examines only whether the temporal change is positive or negative, and disregards the magnitude of the change. Ranging from -1 to +1 it is a measure of the consistency of a monotonic relationship (Mitikka and Ekholm 2002). A value of exactly -1 or +1 is obtained only if there is a consistent decrease or increase throughout the time series. A confidence level of $P<0.01$ is used. Thus, there is a risk of 1 % that the test indicated a trend when actually there was no trend. The Kendall's Tau-b statistical test is especially suitable for environmental data because (i) it is not particularly sensitive to the missing values or outliers, and (ii) requires no assumption of normality (Helsel and Hirsch 1992, Raide *et al.* 2003).

2.4. Application of the CLEAN algorithm

As we mentioned in paragraph 2.2, we use the CLEAN algorithm in an effort to modify our data so that all variables have regular time step and their sampling points are common. After this kind of cleansing and compacting of the data set, it is ready to be treated as a set of time series. These time series can be compared to each other straight forward, without the need of any assumption, e.g. to consider a value representative for the period it belongs or average per month etc.

2.4.1. Short description of the CLEAN algorithm

CLEAN algorithm is an effective tool for spectral analysis especially appropriate for unequally spaced time series which was introduced by Robert *et al.* (1987). This technique is based on a complex one-dimensional version of the CLEAN deconvolution algorithm widely used in image reconstruction. The main advantages of the algorithm are (i) it removes artifacts related to missing data; (ii) it provides clean stable peaks (Tiwari and Rao 2000) and (iii) does not require a formal statistical test (Negi *et al.* 1996). Furthermore, Vio *et al.* (1992) showed that CLEAN algorithm is able to recover effectively most of the lost information even for significantly fewer number of data points. In CLEAN method, a raw (dirty) frequency spectrum is calculated using DFT, which contains real peaks and sidelobes. This dirty spectrum is then iteratively cleaned. The largest spectral peak is found and is subtracted with its side lobes from the original dirty spectrum. In the next iteration, the now largest peak is detected in the residual dirty spectrum and compensated for. The iterations are repeated until a defined noise level or number of iterations is reached. After the CLEANing of the dirty frequency spectrum, all side-lobes are removed. The final CLEAN spectrum is constructed from the accumulated clean spectral components, which are produced by the iterations. Inverse discrete fourier transform is then applied to reconstruct the time-series, using a predefined time step interval. Usually the time step interval for the IDFT is defined as the minimum time difference between the samples, but in our case we will use a -Δt- equal to that defined by the harmonic analysis of the indicator functions, as it will be explained in paragraph 2.4.3. The detailed computational procedure governing CLEAN algorithm is given in Robert *et al.* (1987). A briefly description of the equations incorporating CLEAN algorithm has been presented and discussed by Negi *et al.* (1990, 1996).

2.4.2. Set up and convergence of the CLEAN algorithm

A compatible program was built in MATLAB 7, in order to execute the CLEAN algorithm to all the variables of a data set. Three inputs are needed to run the CLEAN algorithm. These are: 1) a matrix of the raw data (with blanks), 2) the time step interval that the time series will have after applying Inverse Discrete Fourier Transform (iDFT) to the "CLEANed" spectrum and 3) a number indicating how many iterations will be done before the spectrum is considered "CLEAN" enough (see paragraph 2.4.1.) .

2.4.3. Determining the output time step interval

After the DFT spectrum of a particular variable has been CLEANed, inverse Discrete Fourier Transform (IDFT) procedure converts the original data to time series with steady time step. The reconstruction (inverse transform) is done for a predefined time step output, as far as the input data had not a particular sampling ratio but in the best case an approximate one (see indicator function's spectrum of fig. 1.d.) . In our method, we propose determination of the time step according to the indicator function described in paragraph 2.3.1. in a way that will be discussed below. Fast Fourier Transform (FFT) analysis is applied to the indicator function that belongs to the best variable's time series (i.e. the variable with the least gaps and irregularities) and thus the prevailing sampling frequency intention according to the field planning is defined. According to the Nyquyst criterion, the output time series should have at least double the frequency that the raw ones have. This frequency is to be used as the time step interval output of the IDFT, when applied to the CLEANed spectrum.

The choice of the output time step is particularly essential because a time step output smaller than half the sampling intension would be pointless and would induce artefacts and noise to the original time series. On the other hand, time step output bigger than sampling intension would produce coarse time series and would cause loss of information.

2.4.4. Determining the number of iterations

An easy way to define the number of the iterations that will provide the best results in our analysis is to repetitively execute CLEAN algorithm to a specified variable by progressively increasing the iterations number and then plot the iterations number against the misfit of the CLEANed data to the raw data. Misfit values are calculated according to the formula:

Equation 3 - Misfit function

$$M(\%) = 100 \cdot (1 - \sigma_{rc})$$

Where: σ_{rc} is the correlation coefficient between the raw and the CLEANed data of a variable. As far as the matrices of the CLEANed and raw data are not of the same length, a new matrix has to be created with the same sampling scheme as the raw data by suitably interpolating to the CLEANed one. Thus the σ_{rc} refers to the correlation coefficient between the raw samples of the variable and the matching ones extracted by interpolating to the CLEANed variable. After a number of iterations the above equation reaches to convergence. This means that the error induced to the data because of their reconstruction has been minimized. Hence, the iterations number that corresponds to the convergence point can be safely used as the appropriate one for the CLEAN algorithm.

Applying the above method to the best and to the worst sampled variable of the dataset on the basis of their indicator functions, will give a clue about the reliability of the method in view of the particular data set. We choose to determine the iterations number according to the worst variable so that no possibility that the algorithm has not reached to convergence for a variable exists.

2.5. Application of Factor Analysis to the CLEANed data

Having applied the CLEAN algorithm to the data sets, a compact dataset has been created, the dependent variable of which is the time step and the independent ones are the hydrochemical measurements. Grouping of the variables on the basis of their temporal variations is a very useful task towards the data mining process.

2.5.1. Short description of the Factor Analysis method

Factor analysis is a generic term that describes a variety of mathematical procedures applicable to the analysis of data matrices. The most important feature of factor techniques is their ability to reduce a large number of variables down to a smaller number of factors (data reduction technique).

Six main stages or steps in the application of R-mode factor analysis can be recognized: (i) a data matrix (n x m) as basic input is required (where n: observations, m: variables), (ii) the correlation coefficients matrix among the variables (m x m) are computed, (iii) the -m- eigenvalues and eigenvectors are extracted from the correlation matrix, (iv) the selection of the number of factors using certain criteria, (v) the rotation of factor axes in order to achieve the "simple structure" of factor loadings matrix, and (vi) the matrix of factor scores is computed (Papatheodorou *et al.* 2006).

Table 1 - Centralized presentation of main quantitative parameters. $^{(+,-)}$ indicates variables with significant (sig. level < 0.01) increasing or decreasing trends

	Mirkinos			Sidirokastro		
	Kendal's Sig. level	Missing values	misfit	Kendal's Sig. level	Missing values	misfit
T_w	0.17	14.50%	0.05%	0.15	12.28%	0.10%
E.C	$3\ 10^{-5\,(+)}$	0.00%	0.42%	$14\ 10^{-5\,(+)}$	0.00%	0.26%
Ph	0.48	7.25%	0.19%	0.06	4.68%	0.22%
Cl	0.8	2.17%	0.21%	0.02	3.68%	0.30%
SO_4	$10^{-6\,(+)}$	15.22%	0.40%	$2\ 10^{-6\,(+)}$	6.43%	0.31%
HCO_3	0.89	15.22%	0.30%	$9\ 10^{-3\,(+)}$	6.43%	0.21%
TAK	$17\ 10^{-6\,(+)}$	15.22%	0.36%	$5\ 10^{-6\,(+)}$	6.43%	0.22%
Na	$10^{-6\,(+)}$	15.22%	0.11%	$5\ 10^{-5\,(+)}$	6.43%	0.12%
Mg	0.09	15.22%	0.38%	$7\ 10^{-4\,(+)}$	6.43%	0.27%
Ca	0.95	15.22%	0.22%	$0.001^{(+)}$	6.43%	0.28%
SAR	$10^{-6\,(+)}$	15.22%	0.08%	$0.01^{(+)}$	6.43%	0.13%
Alk	$10^{-6\,(+)}$	15.22%	0.12%	0.03	7.02%	0.13%
T.H	0.31	15.22%	0.37%	$6\ 10^{-5\,(+)}$	6.43%	0.26%
D.O	$10^{-6\,(+)}$	13.77%	0.25%	$10^{-6\,(+)}$	11.70%	0.15%
SAT	$10^{-6\,(+)}$	11.59%	0.42%	$10^{-6\,(+)}$	9.36%	0.27%
NO_3	0.41	13.77%	0.11%	0.08	5.85%	0.20%
NO_2	$10^{-6\,(-)}$	13.77%	0.29%	$10^{-6\,(-)}$	5.85%	0.28%
NH_4	$31\ 10^{-5\,(-)}$	14.49%	0.22%	$5\ 10^{-6\,(-)}$	7.02%	0.64%
T.P	10^{-6} *	13.77%	0.15%	0.92	5.85%	0.25%

2.5.2. Calculation and interpretation of the factor scores

After the number of the factors is decided and the factor loadings are determined, factor scores were calculated by applying matrix multiplication between the mxn CLEANed data set (D) and the nxf matrix of the factor loadings (L),

Equation 4 - Factor scores definition function

$$S(m\times f)=D(m\times n)\cdot L(n\times f)$$

where -m- is the number of time steps, -n- is the number of the variables and -f- is the number of factors. In order to emphasize the variables with high loadings and ignore the variables with comparably low loadings, the CLEANed data matrix is multiplied by the 3rd power of the factor loadings. Use of the third power ensures that the sign of the loadings will not be altered. This is done bearing in mind that remarkably high positive or low negative loadings in a particular factor indicate variables that are controlled by the same environmental mechanism which however affects them quite the opposite. Multiplication of a standardized variable with its negative factor loading causes inversion of its values and thus the corresponding time structure will be emphasized when added to the variables with high loadings, of the same factor. The formula for calculating the factor scores finally used is the following:

Equation 5 - Factor scores used function

$$S(m\times f)=std\left[\,std\left[(D(m\times n)\right]\times\left[L(n\times f)\right]^{3}\,\right]$$

where -std- indicates the standardization process. As far as we know, no scientific works using the factor scores of temporal data in such a manner for temporal data has ever been done.

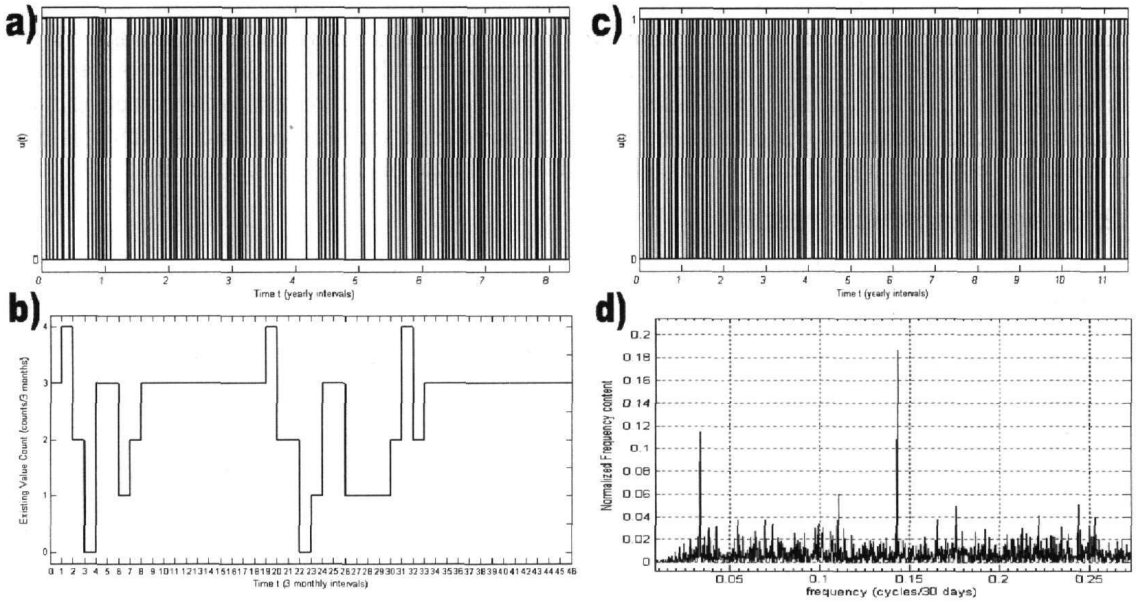

Figure 1 – Sampling structures for Mirkinos site. a) and c) indicator functions diagram of Ca and E.C. respetively, b) existing-value diagram of Ca and d) FFT spectrum of the indicator function for E.C.

3. Results

3.1. Sampling inspection and pre-treatment of the data

Mirkinos and Sidirokastro sampling sites were examined for their sampling scheme according to the indicator and existing-values functions, described in paragraph 2.3.1. Figure 1.a and b show characteristic indicator and existing-value diagrams that were created for Ca of Mirkinos site. Black vertical lines of the indicator function diagram correspond to successfully sampled values while spaces refer to periods of no sampling. At the existing value diagram the number of samples is defined for each season (3 months period) and is plotted against the corresponding time window.

It is obvious that the sampling scheme is quite irregular through the field survey period and there are time gaps without measurements that lasted even 3 to 6 months. These diagrams offer great insight to the data and define the variables to be pre-treated. Figure 1.c shows the indicator function of E.C. (Mirkinos site) that is the variable with the best sampling scheme for the particular site. Even in this case the sampling rate is not even but varies slightly.

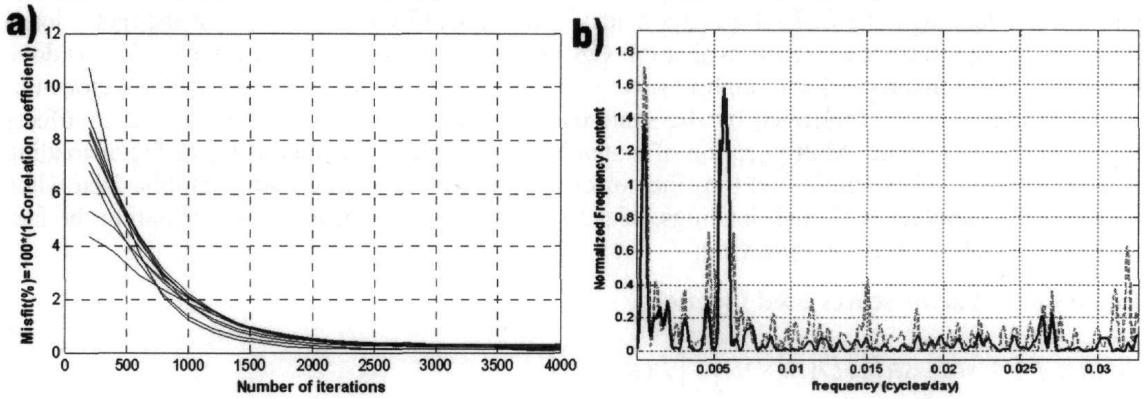

Figure 2 – Set up and convergence of the CLEAN algorithm: a) Centralized misfit diagram for ten variables of Sidirokastro site and b) "CLEANed" (solid black line) versus "dirty" spectrum (dashed gray line) of variable E.C., Mirkinos site

Both datasets of Strimonas River were examined for their trends as described in paragraph 0.. Nine variables with significant trends were located in Mirkinos and thirteen in Sidirokastro sampling site (see Table 1, Kendal's Sig. level field). These variables were detrended by subtracting their fitted linear trend. Detrending process is not essential however it ensures that Factor Analysis will not correlate variables with similar trends but it will concentrate on periodic similarities.

3.2. Spectral analysis results

Two main steps are essential for the application of the CLEAN algorithm (see paragraph 0). These are a) determination of the output time step of the inverse discrete Fourier transform applied to the CLEANed spectrum and b) choice of the iterations needed for the misfit (see Equation 3) to reach convergence. FFT spectrum was created (see. Figure **2**1d) for the indicator function of the most completely sampled variable (i.e. E.C., Mirkinos site), in order to determine the time series' time step. Figure 2d suggests a month's (frequency 0.033 corresponds to 30 days period) main sampling intension, although smaller periodicities are also apparent in the indicator function's spectrum (e.g. 7 days). Thus, a Fifteen days' time step output is chosen to comply with the Nyquyst criterion. To determine the iterations number 10 randomly selected variables (from Mirkinos site) were used to create a misfit plot as shown in Figure 2a. An iterations number equal to 3.000 is considered completely suitable to all the variables while it provides almost perfect match between the raw and the CLEANed data.

To further validate the method results, diagrams plotting the initial data versus the processed ones were created. Figure 3 demonstrates the relation between the CLEANed and the raw data for two characteristic variables of Mirkinos site (NO_3 and Ca) that have a sufficient number of periods characterized by missed samples. These time windows are marked in the figures with transparent grey regions and they include the values predicted by the CLEAN algorithm for the particular periods. Since environmental data are controlled by periodic phenomena it is secure to consider interpolated data close enough to reality. As far as all raw data points are identified with the predicted ones, it is certain that the interpolated values reflect the truth.

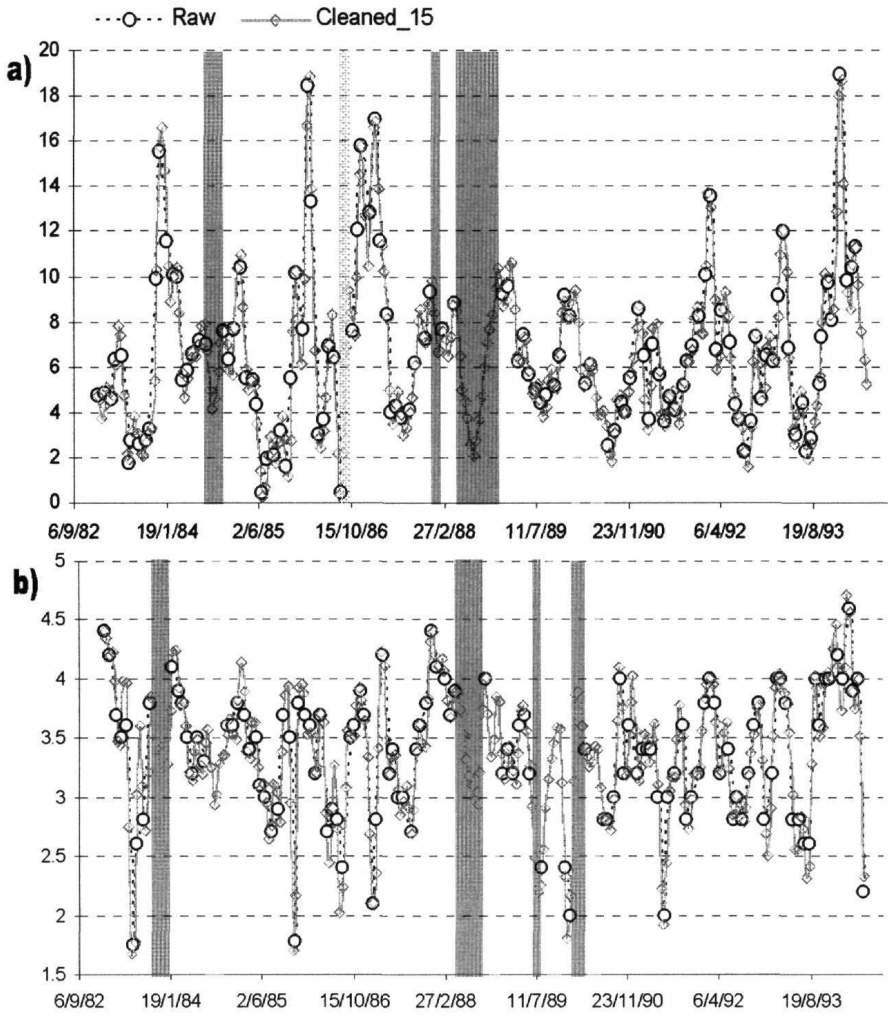

Figure 3 – Initial time series plotted versus the CLEANed ones (15 days time step output). a) NO_3 and b) Ca, Mirkinos site

3.3. Factor analysis results

Factor analysis was applied to both datasets of Strimonas River and the prevailing relations between the variables have been investigated. For both Mirkinos and Sidirokastro sites a six factors model was decided to be used as the most appropriate. For brevity reasons we will examine analytically only the two factors with the greater total variance explained. For both sites in the first factor variables E.C., HCO_3, T.A.K. and T.H. share significantly high positive loadings while the second factor is characterized by the variables: Na, S.A.R. and Alk (Alcalinity of sodium). The former represents more than 20 % of the total variance of the data and the latter represents more than 15 % of the total variance. The first factor corresponds to the mechanisms that control the salinity and total hardness of the river while the second factor corresponds to human pollution caused by agricultural activities and urban wastes.

The scores of the two major factors for each site were calculated according to Equation 5. A comparison between the first two factors of each site and the variables that they represent is illustrated in Fig. 4. It is obvious that factor scores can reveal the temporal expressions of the variables that they represent with remarkable accuracy. This is a very useful tool when data mining is to take place for a data set that consists of a great number of variables. The variables can be easily grouped according to the mechanisms that they are controlled from and then a characteristic temporal variation can be visualized for each group.

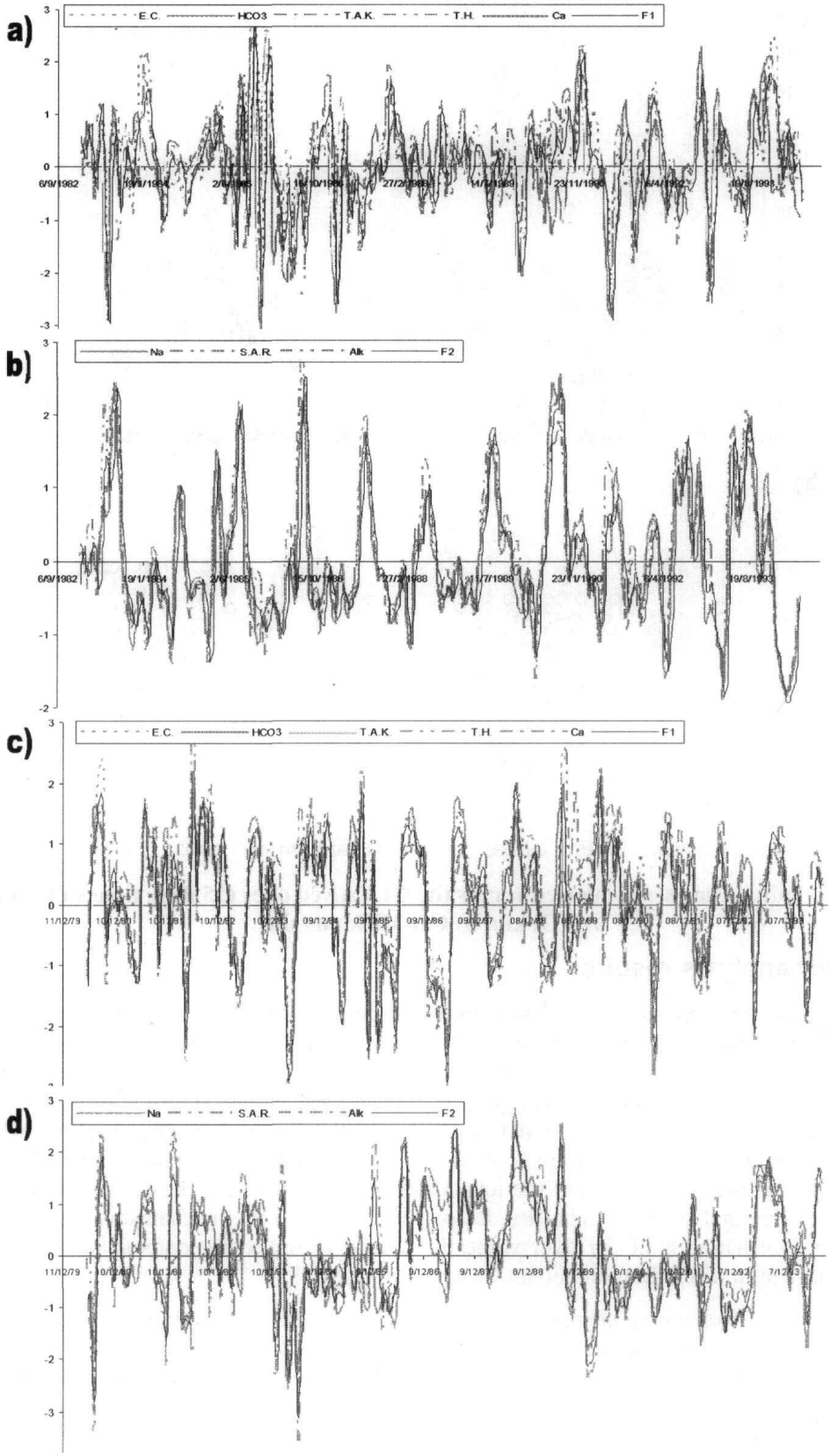

Figure 4 – Factor scores plotted versus the variables that they represent. a) and b) first and second factor of Mirkinos site, c) and d) first and second factor of Sidirokastro site

4. Acknowledgements

The leading author would like to thank the National Scholarships Foundation (NSF) of Greece for the financial support.

5. References

Baisch, S., and Bokelman, G.H.R., 1999. Spectral analysis with incomplete time series: an example from seismology, *Computers & Geosciences,* 25, 739-750.

Davis, J.C., 1986. *Statistics and data analysis in geology*, Wiley, New York, 647pp.

Drehar, J.W., Robert, D.H., and Lehar, J., 1986. Very large array observations of rapid non-periodic variations in OJ287, *Nature*, 320, 239-242.

Duvall, T.L., Jr., and Harvey, J.W. 1984. Rotational frequency splitting of solar oscillations, *Nature*, 310, 19-22.

Heslop, D., and Dekkers, M.J., 2002. Spectral analysis of unevenly spaced climatic time series using CLEAN: signal recovery and derivation of significance levels using a Monte Carlo simulation, *Physics of the Earth and Planetary Interiors* ,130, 103-116.

Mitikka, S., and Ekholm, P., 2003. Lakes in the Finnish Eurowaternet: Status and trends, *The Science of the Total Environment*, 310 37–45

Negi, J.G., Tiwari, R.K., and Rao, K.N.N., 1996. Clean periodicity in secular variations of dolomite abundance in deep marine sediments, *Marine Geology*, 133, 113-121.

Oliver, M.A., Webster, R., Edwards, K.J., and Whittington, G., 1997. Multivariate autocorrelation and spectral analyses of a pollen profile from Scotland and evidence for periodicity, *Rewiew of Palaeobotany and Palynology*, 96, 121-144.

Papatheodorou, G., et al., 2006. A long-term study of temporal hydrochemical data in a shallow lake using multivariate statistical techniques, *Ecological Modelling,* 193, 759-776.

Raike, A., Pietilainen, O.-P., Rekolainen, S., Kauppila, P., Pitkanen, H., Niemi, J., Raateland, A., and Vuorenmaa, J., 2002. Trends of phosphorus, nitrogen and chlorophyll *a* concentrations in Finnish rivers and lakes in 1975 –2000, *The Science of the Total Environment*, 310 (2003) 47 –59.

Ritzi, R.W., Wright, S.L., Mann, B., and Chen, M., 1993. Analysis of Temporal Variability in Hydrogeochemical Data Used for Multivariate Analysis, *Ground Water*, 31(2), 221-229.

Robert, D.H., Lehar, J., and Drever, J.W. 1987. Time series analysis with clean derivation of spectra, *Astron. J.,* 93, 968-989.

Schulz, M., and Stattegger, K., 1997. Spectrum: spectral analysis of unevenly spaced paleoclimatic time series, *Computers & Geosciences*, 23(9), 929-945.

Spangenberg, A., and Bredemeier, M., 1999. Application of spectral analysis to meteorological and soil solution chemistry data, *Chemosphere*, 39(10), 1651-1665.

Stefanakos, Ch. N., and Athanasoulis, G.A., 2001. A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data, *Applied ocean research*, 23, 207-220

Suk, H., and Lee, K.K., 1999. Characterization of a Ground Water Hydrochemical System through Multivariate Analysis: Clustering into Ground Water Zones, *Ground Water*, 37(3), 358-366.

Tiwari, R.K., and Rao, K.N.N., (1999) Solar and tidal reverberarions of deglaciation records from the tropical western Pacific: a clean spectral approach, *Geofizika*, 16-17, 33-41.

Vio, R., Christiannis, Lossi O., and Provenzale, A., (1992) Time series analysis in astronomy: An application to quasar variability studies, *Astron. J.*, 391, 518-530.