

Circadian Patterns Recognition in Ecosystems by Wavelet Filtering and Fuzzy Clustering

Stefano Marsili-Libelli and Simone Arrigucci

Dept. of Systems and Computer Engineering, University of Florence
Via S. Marta, 3 - 50139 Florence, ITALY
Email:marsili@ingfi1.ing.unifi.it

Abstract: This paper presents a method for extracting representative patterns from a set of data representing circadian cycles. The analysis is based on a combination of wavelet filtering and fuzzy clustering. The data are first processed with a discrete wavelet decomposition in order to filter out the noise and isolate the relevant circadian cycle. It is shown that the second level decomposition yields the best cycle approximation, filtering out measurement noise and other artefacts and preserving the main cycle features. From the filtered data the following discriminating features are extracted: minimum and maximum daily values, and the slope of the line passing through these extreme points. These features are then processed with a fuzzy clustering algorithm, in order to isolate significant behaviours. The Fuzzy Maximum Likelihood Estimates (FMLE) method was used for its variable metric, being able to conform the cluster shape and volume to the data. This combined algorithm is applied to the physico-chemical data from the Orbetello lagoon with the aim of detecting ecologically meaningful behaviours. The results show that relevant daily patterns are indeed isolated and in particular combination of variables leading to the dystrophic crisis are correctly interpreted. The relevance of the selected patterns is confirmed by their distribution over the calendar day, which corresponds to a clear seasonal patterns.

Keywords: Wavelet filtering; Fuzzy clustering; knowledge-based systems; artificial intelligence;

1. INTRODUCTION

Ecosystems are subject to fluctuations with a wide range of periods, from the short-term random variations to daily cycles and seasonal changes. Of all these, the diurnal cycle is the most important, being related to the day-night sequence which drives most ecological process. Daily patterns are a major feature of ecosystems and their seasonal changes may reveal important information regarding its functioning and, if properly interpreted, can represent a valuable tool in ecosystems forecasting and management.

A previous study of the circadian cycles (Marsili-Libelli, 2004) used a fuzzy method for the automatic pattern recognition, but its weak point was the heuristic choice of the model patterns. To complement the previous algorithms, this paper proposes a new method to construct a consistent, objective knowledge-base of significant patterns. Circadian cycles are daily fluctuations in biological activities.

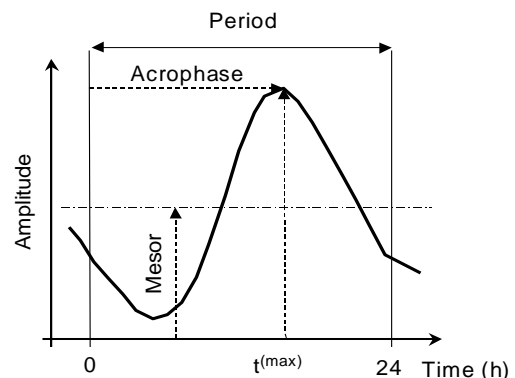


Fig. 1. Basic features of the circadian cycle.

They are the consequence of the daily variations of the solar radiation and follow the general pattern of Figure 1. The important parameters of a circadian cycle are the maximum value (acrophase), its timing and the mean daily value (mesor). The basic idea is to extract the underlying circadian pattern from noisy data using wavelet filtering and partition these approximations using a fuzzy clustering method. The result of this combined

procedure is a number of "typical" patterns, which can be used as the required knowledge-base.

The reasons for applying this two-step procedure are now examined:

1. Data are normally affected by noise in many ways. In order to extract the information related to the circadian pattern, a wavelet decomposition is performed. This filtering technique uses windows of variable size and is capable of performing a joint time-frequency analysis revealing aspects of data that other signal analysis techniques miss, especially if the data record is short. It can also de-noise the data without appreciable degradation, separating components of different scales: low-frequency, which are usually ecologically meaningful, and high-frequency disturbances.
2. Once the basic patterns are isolated, significant features are extracted and grouped into a number of meaningful behaviours through fuzzy clustering. In particular the Fuzzy Maximum Likelihood Estimates (FMLE) algorithm is used for its ability to produce clusters of varying volume and shape (Babuska, 1998). This attractive feature is a consequence of being based, like the well-known Gustafson and Kessel method, on an adaptive distance norm derived from the fuzzy covariance matrix.

The paper may be regarded as a sequel to a previous one (Marsili-Libelli, 2004) where the knowledge-base was constructed with a mixed statistical-fuzzy technique. The method is demonstrated with reference to daily variation of dissolved oxygen (DO) data from a eutrophic lagoon, described in the next section.

2. PATTERN DETECTION IN THE ORBETELLO LAGOON DATA

The Orbetello lagoon, schematically shown in Figure 2, is located along Italy's west coast. It consists of two shallow coastal reservoirs with a combined surface of approximately 27 km², an average depth of 0.8 m. Two water-quality monitoring stations, indicated by the two circles in Figure 2, transmit hourly physico-chemical data to the Orbetello Lagoon Managerial Office headquarters. These data include Dissolved Oxygen (DO), Oxido-Reduction Potential (ORP), pH and temperature.

The submersed vegetation is composed of macroalgae (*Chaetomorpha linum*, *Cladophora vagabunda*, *Gracilaria verrucosa*, *Ulva rigida*) and macrophytes (*Ruppia maritima*). Given the large availability of nutrients and the limited water renewal, when the macroalgae decompose after an

excessive growth, an oxygen imbalance may occur, causing anoxia.

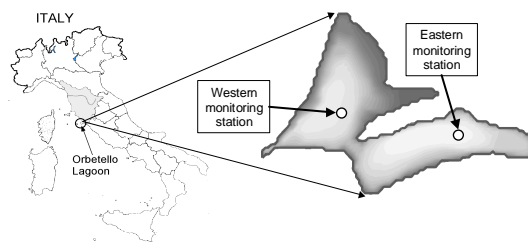


Fig. 2. The Orbetello lagoon with the location of the two monitoring stations.

This kind of dystrophic crisis is well studied (Christian *et al.*, 1998). During the normal growth phase, macroalgae represent a sink for dissolved inorganic nitrogen and oxidising processes prevail: day-time dissolved oxygen (DO) has a well defined afternoon peak, often well above the saturation level. When the growth phase ends, the fast anoxic decomposition enriches the sediment with reduced organic nitrogen (Christian *et al.*, 1998). These reducing conditions can be detected by low, almost constant DO and low daytime oxidation-reduction potential (ORP) in the water.

Thus the daily cycle of these variables contains important season-dependent information, which this method attempts to extract from noisy data using the two-step procedure outlined in Section 1. The algorithm is organised in a hierarchical structure, where DO is the primary variable from which the clustering features are extracted. The secondary variables ORP, pH and temperature are processed at a later stage, depending on the DO results and are used to reveal finer details in a cluster structure.

3. WAVELET PREPROCESSING

Frequency-based techniques have been used extensively for data filtering, but their drawback is that the time-information is completely lost. This is a serious shortcoming if the signal is non-stationary, because trends, bursts and specific one-time events may be missed. This is due to the use of infinite functions, usually sinusoids, as the basis functions. The question then arises why not choose a basis function that has a finite duration, instead of choosing an infinite-duration one? Wavelets (Strang and Nguyen, 1996; Torrence and Compo, 1998) are finite-duration signals which can be used to replace sinusoids as basis functions for filtering. In this way time and frequency analysis can be combined through a variable windowing technique and local analysis of short-duration events can be performed without losing the power of frequency analysis.

Given a wavelet function $\psi(a,b,t)$ where a and b represent the scaling and the shifting factors, the continuous wavelet transform (CWT) is defined as the integral of the signal $s(t)$ multiplied by the scaled wavelet $\psi(a,b,t)$

$$C(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(t) \psi\left(\frac{t-b}{a}\right) dt \quad (1)$$

Given the finite time duration of $\psi(a,b,t)$ the integration limits in eq. (1) can in practice be limited. In CWT a and b vary continuously. If the wavelet is stretched ($a \gg 1$) it contains mostly low frequencies and gives a global view, whereas compressed wavelets ($a \ll 1$) contain mostly high frequencies and give details view of a small portion of the signal $s(t)$. Thus by varying a the analysis can be concentrated on global or specific features of the signal. However, computing Eq. (1) for any combination of (a,b) would be greatly time-consuming with little insight into the process generating the signal. Restricting the CTW to dyadic scaling, i.e. $\left(\frac{T_s}{2}, a, \frac{T_s}{4}, a, \frac{T_s}{8}, \dots, a, \frac{T_s}{2^n} \right)$ the

Discrete Wavelet Transform (DWT) is obtained, based on powers of two of the base window T_s , which in our case represents the circadian period of 24 h. DTW yields a signal hierarchical decomposition into *Approximations*, grouping the high-scale, low-frequency components of the signal, and *Details*, retaining the low-scale, high-frequency parts.

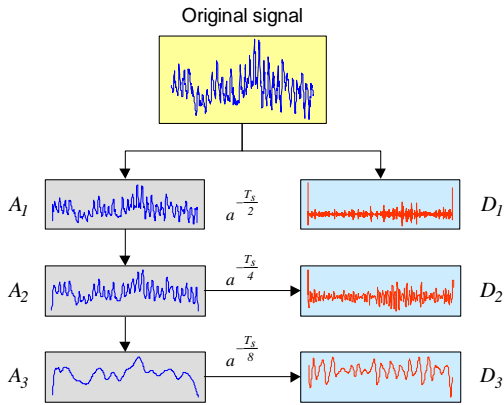


Fig. 3. Discrete-time wavelet multilevel decomposition tree.

This process can be iterated in the scale $a = \frac{T_s}{2^k}$, as shown in Figure 3, halving the window T_s at each step in order to get increasingly low-frequency, long-time approximations and higher-frequency details. In performing this process of *Multilevel Decomposition* (MD) it should be considered that at each step the available data are halved and the

time scale is doubles with respect to the basic sampling time T_s , as shown in Figure 4.

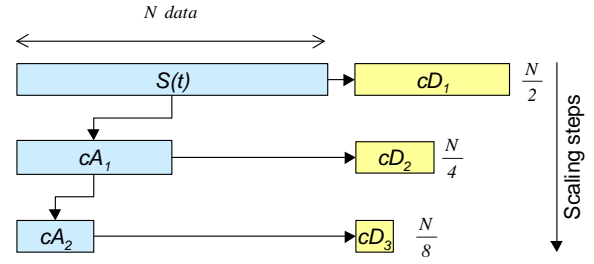


Fig. 4. Data halving at each scaling step of MD.

Therefore the iteration should be stopped before the data become so scarce that edge effect become important and the time-scale of the transformation exceeds the range of interest.

3.1. Settings for wavelet analysis of circadian patterns

For the analysis of circadian patterns a Meyers wavelet (Meyers et al., 1993) was selected for its properties of symmetry, orthogonality and finite support. The ratio between time window and sampling interval is important. Since the object of the study is the circadian pattern, it seems obvious to select a $T_s = 24 h$ window which, given a sampling interval $\Delta = 1 h$, includes 24 values. With these conditions, the multilevel decomposition cannot be iterated beyond the second level or the edge effects would seriously deteriorate the results. Moreover, restricting the analysis to Level 2 has a clear physical meaning:

- **Level 1** implies 2 h differences, hence the approximation A_1 represents the de-noised signal, filtering out the measurement noise, assuming that a random 1 sample fluctuation is the combined effect of the DO probe error and short-term environmental changes (wind gusts, passing clouds, etc.). Likewise, the detail D_1 represents the noise contribution due to these short-lived effects;
- **Level 2** is performed at 4 h intervals, producing a signal free of climatic artefacts. Thus approximation A_2 represents the true circadian cycle, rich in ecological information (algae photosynthesis and respiration, decomposition processes, etc.). Likewise, detail D_2 accounts for the environmental variability on a 4 h horizon.

Thus it can be concluded that a Level 2 Approximation is sufficient to extract the required information from the daily signal. The problem of edge effects is now considered. Unfortunately 24 is not a power of 2, hence the dyadic scaling does not provide optimal data scaling. One possible answer

to this problem could be the processing of three consecutive days and consider the decomposition (A_2, D_2) of the central day only. However, Figure 5 shows that the advantage of this method is minimal *vis à vis* the increased computational burden and the increased fraction of excluded days because of missing data (triplets would be needed instead of single days).

A quantitative way to appreciate the degradation due to edge effects in processing each single day can be obtained by computing the percentage of energy corresponding to the approximation E_a and the sum of the percentages of details energy E_d . The variation of E_a in processing single days or triplets is less than 0.1 %, whereas the E_d variation may be as high as 50%. This confirms the choice of A_2 as a stable decomposition, truly representative of the circadian cycle.

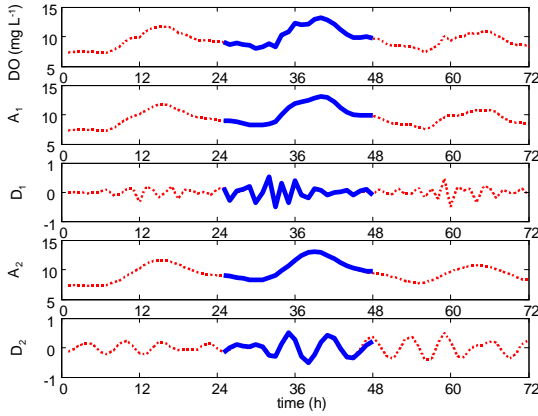


Fig. 5. Comparison between Level 2 MD of a triplet (dotted line) and the central day alone (thick line).

3.2. Features extraction from filtered data

The characteristics of wavelet filtering for this application can be summarised as follows: a second level Discrete Wavelet Transform (DWT) is performed on single days composed of 24 hourly samples and the second approximation (A_2) is considered to be a good denoised reconstruction of the underlying circadian cycle. From this filtered signal the following features are extracted for clustering: minimum and maximum DO values (DO_{min} , DO_{max}) and the slope α of the line connecting these values, as shown in Figure 6. Therefore the data points \mathbf{x} are triplets of the following kind

$$\mathbf{x}_k = [DO_{min}^k \quad DO_{max}^k \quad \alpha^k] / k = 1, 2, \dots, N.$$

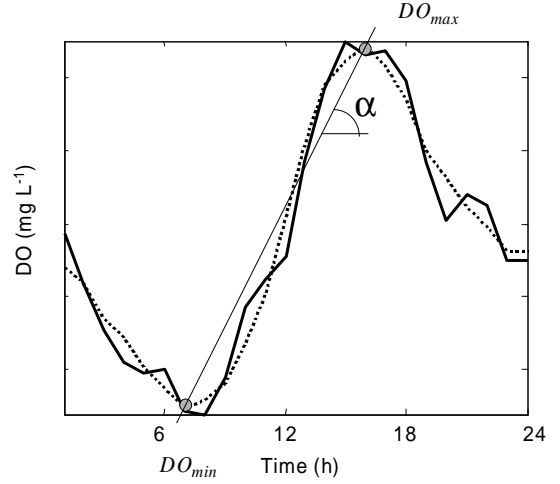


Fig. 6. Features of a DO circadian cycle.

4. FUZZY CLUSTERING FOR CIRCADIAN PATTERN RECOGNITION

These three features are then processed using the Fuzzy Maximum Likelihood Estimates (FMLE) fuzzy clustering algorithm (Babuska, 1998). The fuzzy clustering algorithm arranges the data $\{\mathbf{x}_k / k = 1, \dots, N\}$ into c clusters through the minimisation of the partition functional

$$J(\mu_{i,k}, m) = \sum_{i=1}^c \sum_{k=1}^N (\mu_{i,k})^m d_{ik_A}^2 \quad (2)$$

where the distance $d_{ik_A}^2$ depends on the norm-inducing matrix \mathbf{A} , which in the FMLE case is a function of the fuzzy covariance matrix \mathbf{R} .

$$\mathbf{R} = \frac{\sum_{k=1}^N \mu_{ik} (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T}{\sum_{k=1}^N \mu_{ik}} \quad (3)$$

$$d_{ik_A}^2 = \frac{(\det \Sigma_i)^{\frac{1}{2}}}{\left(\frac{1}{N} \sum_{k=1}^N \mu_{ik}\right)} \exp\left(\frac{1}{2} (\mathbf{x}_k - \mathbf{v}_i)^T \mathbf{R}^{-1} (\mathbf{x}_k - \mathbf{v}_i)\right) \quad (4)$$

The resulting membership $\mu_{i,k}$ represents the degree of membership of the k -th data point \mathbf{x}_k to the i -th cluster with center $\mathbf{v}_i / i = 1, \dots, c$. The effectiveness of the partition can be evaluated in information-theoretic terms using the normalised partition entropy H_n , defined as (Babuska, 1998).

$$H_n = -\frac{1}{1 - \frac{c}{n}} \sum_{k=1}^n \sum_{i=1}^c \mu_{i,k} \log(\mu_{i,k}) \quad (5)$$

This quantity was used to decide the appropriate number of clusters c in the partition, as the one which minimises the H_n value for $c > 2$. Processing the daily DO data in the years 2001, 2002 and

2003 resulted in a number of clusters with a clear ecological meaning. The number of cluster for each year, determined with the partition entropy of eq. (5), varied from $c = 3$ for 2001 to $c = 6$ in 2002 and $c = 4$ for 2003.

As an example, the clusters obtained for 2001 are shown in Figure 7.

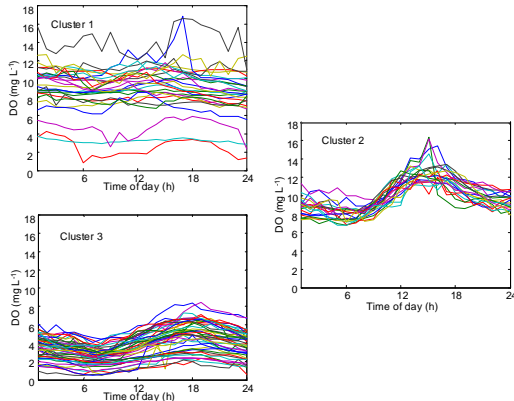


Fig. 7. DO clusters obtained from the 2001 data.

For the year 2001 three basic behaviours were isolated, with cluster 2 representing the typical spring cycle with a strong afternoon peak, due to intense photosynthesis, and cluster 3 typical of the summer situation with low oxygen due to high temperatures and declining algal activity. Cluster 1 groups atypical behaviours, possibly induced by adverse weather conditions, preventing the development of the normal circadian cycle. The distribution of these patterns over the year, together with that of unclustered patterns is shown in Figure 8.

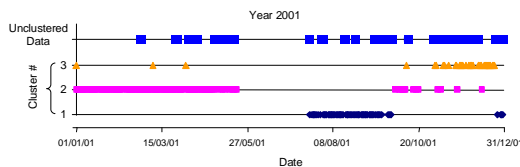


Fig. 8. Distribution of clustered and unclustered patterns over 2001.

The same analysis for 2002 produced six clusters whose distribution over the year is shown in Figure 9. This year presented a wider variability, which resulted in a larger number of clusters. As to their characteristics, cluster 3 was similar to cluster 2 of 2001 (high algal activity) and cluster 4 was similar to cluster 3 of 2001 (summer low activity).

The data from 2003 produced the most interesting structure: the basic DO clustering resulted in the four clusters of Figure 10. However, a finer structure can be extracted from cluster 2, observing that these DO patterns correspond to very different ORP cycles. Tracing back these

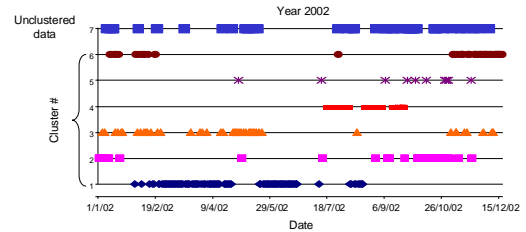


Fig. 9. Distribution of clustered and unclustered patterns over 2002.

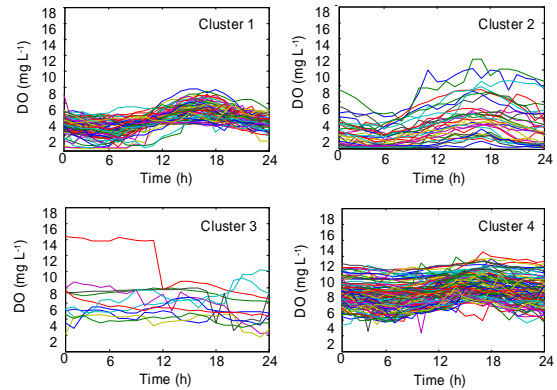


Fig. 10. DO clusters obtained from the 2003 data.

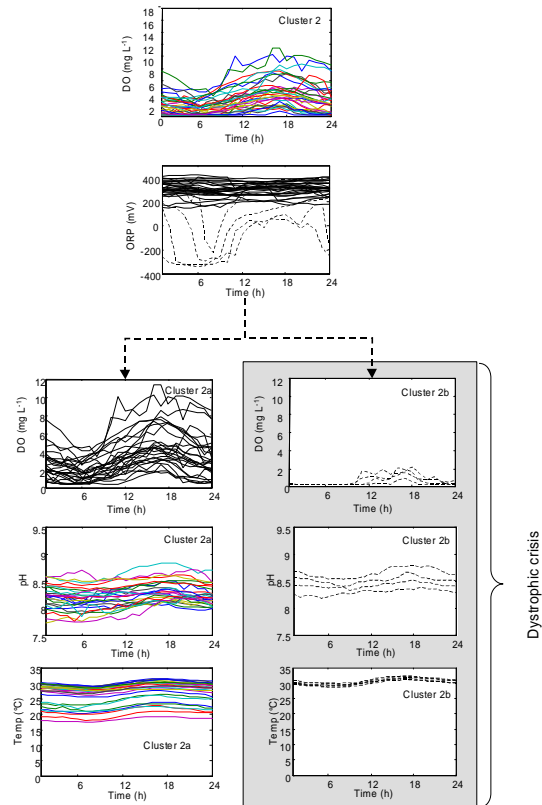


Fig. 11. Finding a finer structure in cluster 2.

patterns to the original DO cycles in cluster 2, this can be further decomposed into two subsets, one of which is clearly representative of the dystrophic crisis, as shown in Figure 11. This separation is

confirmed by the two other variables, pH and temperature, yielding the composite cluster structure for 2003 shown in Figure 12.

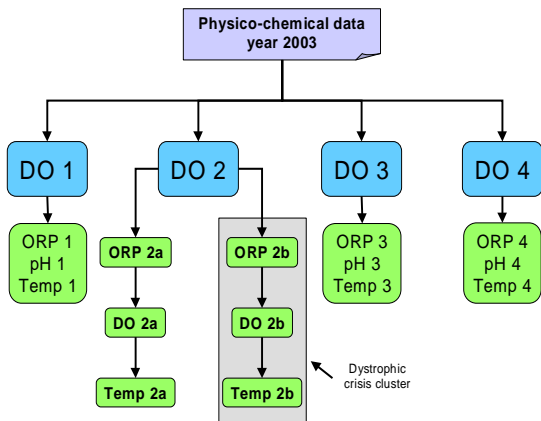


Fig. 12. Cluster structure for the year 2003.

It is also interesting to observe how the clustered and unclustered data are distributed over the year, as shown in Figure 13, with cluster 2 prevailing during the warm season, and in particular with cluster 2b, representative of the dystrophic crisis, placed exactly when the crisis actually occurred, as shown in Figure 14.

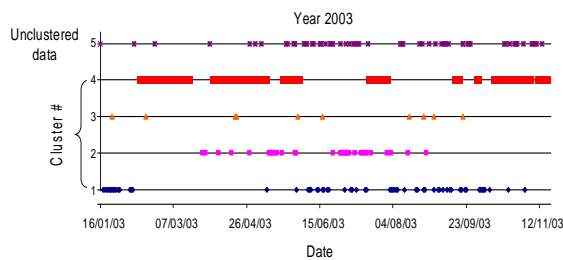


Fig. 13. Distribution of clustered and unclustered cycles over the calendar day in 2003.

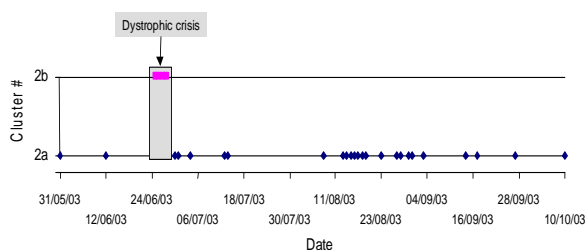


Fig. 14. Placement of cluster 2b during the dystrophic crisis.

5. CONCLUSION

This paper has presented a method for extracting representative prototypes of circadian cycles. It is based on a combination of wavelet filtering and

fuzzy clustering. The data were first processed with a discrete wavelet decomposition in order to isolate the relevant circadian cycle, which is well represented by the second level approximation A_2 . From the filtered cycles three relevant features were extracted: maximum and minimum values and the slope of their connecting line. The cycles were then classified on the basis of these features using the Fuzzy Maximum Likelihood Estimates (FMLE) clustering algorithm, which was preferred for its variable metric, being able to conform the cluster shapes to the data.

The application of the algorithm to the physico-chemical data from the Orbetello lagoon resulted in a consistent classification of relevant circadian cycles and in particular it isolated the patterns corresponding to the dystrophic crisis observed in extreme eutrophic conditions. Though the pattern structure varied over the three years of the study (2001 - 2003), the method allowed positive cycle identification, whose distribution over each year was consistent with the observed behaviours and seasonal variability.

6. ACKNOWLEDGEMENT

The Orbetello Lagoon Managerial Office is acknowledged for supporting this research under contract n. 96/988 of 24.07.2003.

7. REFERENCES

- Babuska, R., *Fuzzy modeling for control*. Kluwer Publ. Co., Amsterdam, 260 pp., 1998.
- Christian, R.R., Naldi, M. and Viaroli, P., 1998. Construction and analysis of static, structured models of nitrogen cycling in coastal ecosystems. in Koch A.L., Robinson J.A. Milliken G.A. (eds.), *Mathematical Modelling in Microbial Ecology*, Chapman & Hall, New York.
- Marsili-Libelli, S., Fuzzy pattern recognition of circadian cycles in ecosystems. *Ecol. Model.*, **174**, 67 - 84, 2004.
- Meyers, S.D., Kelly, B.G., and J.J. O'Brien. An introduction to wavelet analysis in oceanography and meteorology: with application to the dispersion of Yanai waves. *Mon. Weather Rev.* **121**, 2858–2866, 1993.
- Strang, G. and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, 1996.
- Torrence, C. and A. Compo, A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, **79**, 61 –78, 1998.