

# Environmental data inference through fuzzy clustering

S. Marsili-Libelli<sup>a</sup> and E. Giusti<sup>a</sup>

<sup>a</sup> *Department of Systems and Computers, University of Florence, Via S. Marta 3, 50139 Firenze, Italy (marsili/giusti@dsi.unifi.it)*

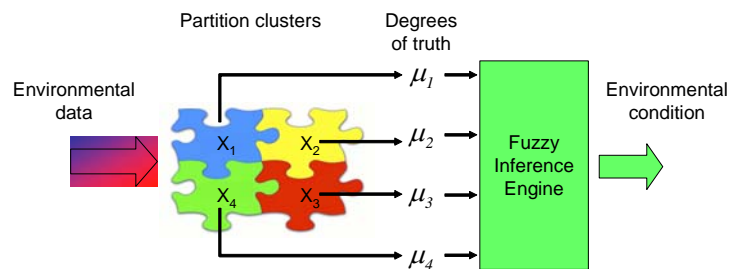
**Abstract:** Retrieving the relevant information from noisy environmental data is a difficult task, which requires advanced filtering techniques. This paper presents a method for extracting representative prototypes from a set of data containing daily and seasonal fluctuations. It is based on a combination of wavelet filtering and fuzzy clustering. Discriminating features are extracted with a Fuzzy Maximum Likelihood Estimates (FMLE) clustering algorithm, which was selected for its variable metric enabling the adaptation of the cluster shape and volume to the data. The results show that the discriminating power of this algorithm is considerable, as demonstrated by the application to two differing domains: the discrimination of dissolved oxygen circadian cycles in the Orbetello lagoon and the daily and seasonal fluctuations in photosynthesis in the Arno river. In both cases the isolated patterns have a clear ecological meaning and reveal the relevant ecosystem variations on differing time-scales.

**Keywords:** Fuzzy modelling, pattern recognition, fuzzy clustering, Water quality, photosynthesis, dissolved oxygen.

## 1. INTRODUCTION

Extracting the information from environmental data may become a very complex process for a number of reasons: the noise affecting the data is non-Gaussian and difficult to model, the information is contained in the data rate of change rather than in the data itself, and complex patterns may be hidden in the data. This paper describes an information retrieval procedure developed with fuzzy tools, in order to extract information from environmental noisy data.

An efficient fuzzy inferential system may be composed of a large number of IF-THEN predicates, so that practical application may easily reach unmanageable proportions. To avoid the “curse of dimensionality” data are grouped into fuzzy clusters depending on the extent to which they share some common features. In this way the antecedents are replaced by partition clusters, whose degrees of truth are used to activate the fuzzy inference engine inferring the current environmental condition



**Figure 1.** General structure of the Sugeno fuzzy inference algorithm with clustered antecedents.

The generic  $i$ -th fuzzy rule of the Sugeno fuzzy inference algorithm of Figure 1 has the form

$$R_i : \text{IF } x \in X_1 \text{ AND } \dots x \in X_n \text{ THEN } y = y_i. \quad (1)$$

and the resulting environmental condition is obtained by weighted average defuzzification

$$y = \frac{\sum_{i=1}^n \mu_i y_i}{\sum_{i=1}^n \mu_i} \quad (2)$$

where  $n$  is the number of rules and  $\mu_i$  is the degree of activation of each cluster.

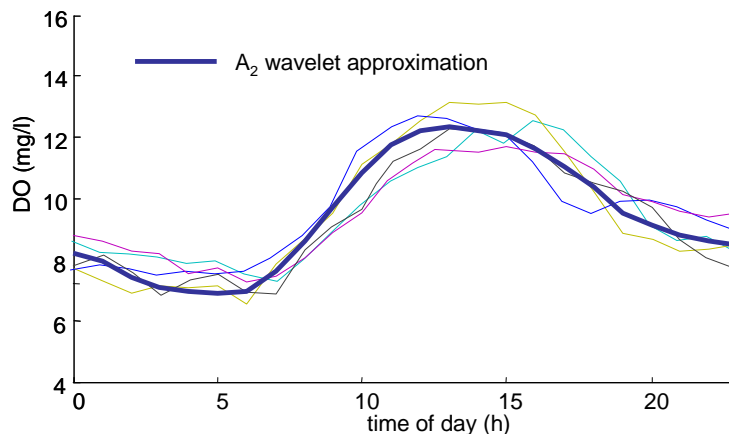
### 1.1. The pattern recognition algorithm

The algorithm described in this paper is aimed at extracting meaningful behavioural patterns from a set of data representing circadian cycles. It is based on a combination of wavelet filtering and fuzzy clustering.

#### 1.1.1. Wavelet denoising

The data are first processed with a discrete wavelet decomposition in order to isolate the relevant circadian cycle, and it is based on a combination of wavelet filtering and fuzzy clustering. The wavelet filtering represents a useful data pre-processing tool to separate noise from the relevant deterministic information allowing a more discriminating features extraction. This is later accomplished by applying a fuzzy clustering algorithm to the denoised data. The Fuzzy Maximum Likelihood Estimates (FMLE) clustering method was preferred for its variable metric, which adapts the cluster shape and volume to the data. The results show that the discriminating power of this combined technique is much higher than that of each technique used separately.

Wavelet denoising [Strang and Nguyen, 1996; Percival and Walden, 2000] is performed to alleviate the noise problem and provide a stable numerical derivative, if required. As thoroughly described in [Marsili-Libelli and Arrigucci, 2004] wavelet denoising is obtained by a Multi Level Digital Wavelet Decomposition (MLDWD) where the scale factor is halved at each stage. The decomposition is based on powers of two of the circadian period  $T_s$ , providing a signal decomposition into an *approximation*, grouping the low-frequency components of the signal, and a *detail*, retaining the high-frequency components. From extensive data experimentation, it was concluded that the Level 2 approximation is a good denoised reconstruction of the true circadian cycle, as thoroughly discussed in [Marsili-Libelli and Arrigucci, 2004; Marsili-Libelli, 2004]. The Meyers [Meyers et al., 1993] was selected for its properties of symmetry, orthogonality and finite support.



**Figure 2.** Meyers second approximation  $A_2$  (thick line) of several daily patterns (thin lines) for the daily DO variations in the Orbetello lagoon.

### 1.1.2. Fuzzy clustering

Once the basic features are denoised, pattern recognition is applied to group significant behaviours a number of meaningful fuzzy clusters. Of the many techniques available [Bezdek, 1981; Babuska, 1998; Abonyi, 2003] the Fuzzy Maximum Likelihood Estimates (FMLE) algorithm is used here for its ability to produce clusters of varying volume and shape, thus adapting to the dimension of the data. This flexibility is a consequence of basing the partition on an adaptive distance norm derived from the fuzzy covariance matrix, as shown in Eq. (4).

The FMLE fuzzy clustering algorithm arranges the data  $\mathbf{x}_k$  into  $c$  clusters through the constrained minimisation of the partition functional

$$J(\mu_{i,k}, m) = \sum_{i=1}^c \sum_{k=1}^N (\mu_{i,k})^m d_{ik}^2, \quad (3)$$

where  $m$  is the fuzzy exponent and the distance  $d_{ik}$  depends on the norm-inducing matrix, which in the FMLE case is the fuzzy covariance matrix  $\Sigma$ , defined as

$$\Sigma_i = \frac{\sum_{k=1}^N \mu_{ik} (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T}{\sum_{k=1}^N \mu_{ik}} \quad i = 1, \dots, c. \quad (4)$$

The distance in Eq. (3) then becomes

$$d_{ik_f} = \frac{(\det \Sigma_i)^{\frac{1}{2}}}{\left(\frac{1}{N} \sum_{k=1}^N \mu_{ik}\right)} \exp\left(\frac{1}{2} (\mathbf{x}_k - \mathbf{v}_i)^T \Sigma_i^{-1} (\mathbf{x}_k - \mathbf{v}_i)\right) \quad (5)$$

and the partition membership functions are obtained as

$$\mu_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik_j}}{d_{ij_z}}\right)^{\frac{2}{m-1}}} \quad i = 1, \dots, c \quad k = 1, \dots, N \quad (6)$$

The resulting membership  $\mu_{ik}$  in Eq. (6) represents the degree of membership of the  $k$ -th data point  $\mathbf{x}_k$  to the  $i$ -th cluster with center  $\mathbf{v}_i$ .

The choice of the best number of clusters ( $c$ ) was performed by applying three different cluster validity measures:

- *Fuzzy hypervolume:*

$$V_h = \sum_{i=1}^c [\det(\Sigma_i)]^{\frac{1}{2}}, \quad (7)$$

where  $\Sigma_i$  are the cluster covariance matrices. The first minimum of this measure yields the best cluster partition;

- *Average Partition Density:*

$$D_A = \frac{1}{c} \sum_{i=1}^c \frac{S_i}{[\det(\Sigma_i)]^{\frac{1}{2}}}, \quad (8)$$

where  $S_i$  is the sum of the membership degrees of the data that lie within a hyper-ellipsoid with radius equal to the standard deviations of the cluster features. The first maximum of this measure yields the best cluster partition;

- *Average Cluster Flatness :*

$$t_A = \frac{1}{c} \sum_{i=1}^c \frac{\lambda_{in}}{\lambda_{i1}}. \quad (9)$$

It is based on the ratio between the smallest and largest eigenvalue of the fuzzy covariance matrices  $\Sigma_i$ . The first minimum of this measure yields the best cluster partition.

## 2. APPLICATIONS OF THE ALGORITHM

The combined wavelet denoising and fuzzy clustering algorithm is used as a pattern recognition procedure and applied to the detection of variations in two water quality environmental time-series: the patterns in the daily dissolved oxygen in the Orbetello lagoon and the seasonal fluctuations in the river water quality parameters measured in the Arno river, both in Tuscany, central Italy, as shown in Figure 3.

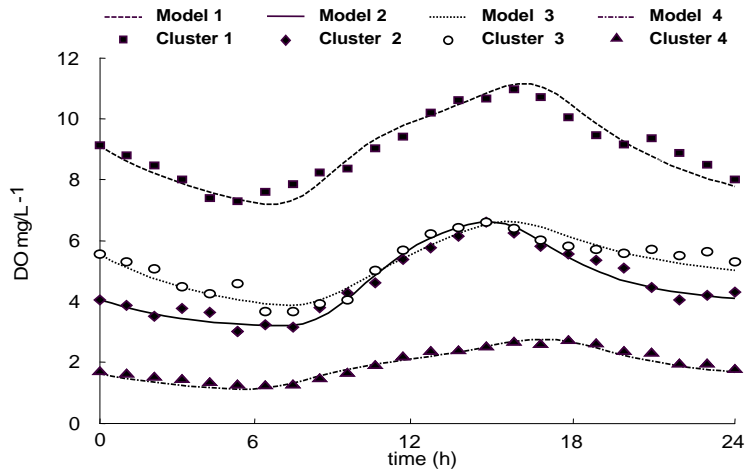
### 2.1. Daily dissolved oxygen patterns in the Orbetello lagoon

This problem has already been treated in the past [Marsili-Libelli and Arrigucci, 2004; Marsili-Libelli, 2004], thus only the new results will be presented here. Starting with the problem of circadian pattern recognition for eutrophication early warning [Marsili-Libelli, 2004] the interest in dissolved oxygen dynamics has shifted to the completion of a complex ecological model of the lagoon [Giusti and Marsili-Libelli, 2006], encompassing all the lagoon dynamics, except the dissolved oxygen. In adding this last feature to the model, it was recognized that the operating conditions of the lagoon varied so widely over the years that no single model would be able to accommodate such large variations [Marsili-Libelli and Giusti, 2007], hence the need to develop a bank of differing models and use the most appropriate of them for the current conditions, which are detected by applying the procedure outlined in the previous section. Using the same clustering parameters for discriminating the diurnal cycle (minimum and maximum Dissolved Oxygen concentration and the slope of the line connecting these two extremal points) as in [Marsili-Libelli and Arrigucci, 2004; Marsili-Libelli, 2004] the new pattern recognition results are now used to select the most appropriate dissolved oxygen dynamical model, developed separately, for a long-term simulation [Marsili-Libelli and Giusti, 2007]. In fact it was found that no single model (i.e. parameter set) yields an acceptable agreement between data and the DO model, thus the problem arises of choosing the best combination of parameters in order to adapt the model to the long-term DO variations caused by environmental changes induced by the seasonality and varying physico-chemical conditions. The pattern recognition algorithm was used for model parameters patching to obtain the best response on a long time horizon. After obtaining four meaningful clusters, this number being given by the partition indicators previously mentioned in Eqs. (7 – 9), four model parameter sets were obtained by calibrating the model with respect to the prototypical pattern of each cluster, obtaining the model-data agreement on Figure 4. The actual parameter set to be used in the model for each day is obtained by fuzzy implication of the four basic prototypes, as shown in Figure 5. The degree of truth of each pattern is obtained by fuzzy comparison between the current DO pattern and the prototypes. This is obtained by using again Eq. (3), but this time in a one-pass way (i.e. dropping the k index) and not iteratively because the prototypes are now known quantities [Marsili-Libelli and Müller, 1996]



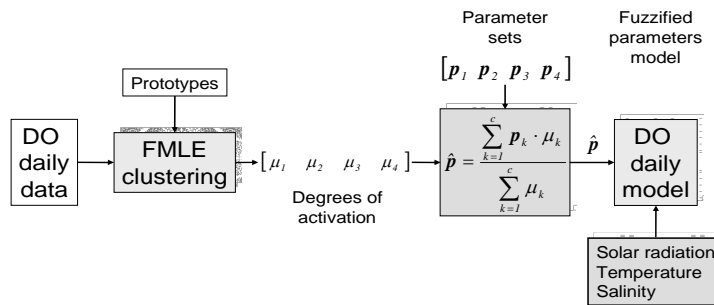
**Figure 3.** Location of the two case studies described in the paper.

$$\mu_i = \frac{1}{\sum_{j=1}^c \left( \frac{d_{i_x}}{d_{j_x}} \right)^{\frac{2}{m-1}}}, \quad i = 1, \dots, c \quad (10)$$

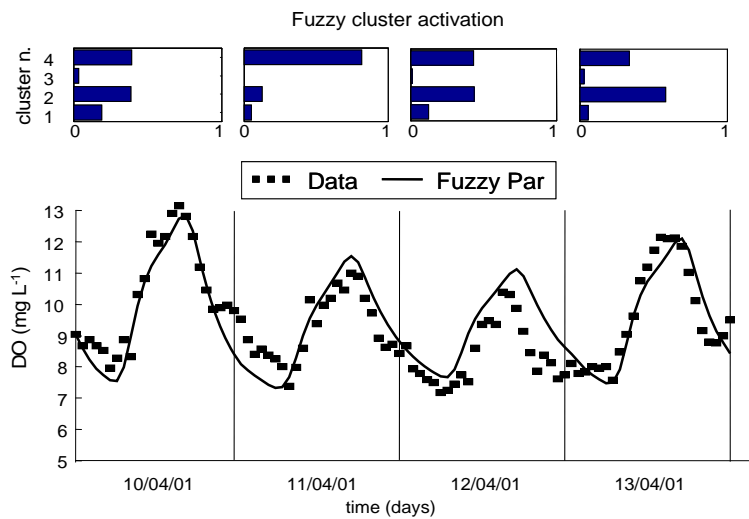


**Figure 4.** Calibrated response of the four basic DO models, each compared with the corresponding prototypical data sets, representing the four prototypical daily DO patterns in the lagoon.

where the  $d_{\bullet, \Sigma}$  are the  $\Sigma$ -norm distances of the current DO pattern parameters ( $DO_{\min}$ ,  $DO_{\max}$ ,  $\alpha$ ) from the cluster prototypes, previously computed with Eq. (6), where the norm is induced by the fuzzy covariance matrix  $\Sigma_i$  defined by Eq. (4).



**Figure 5.** Model patching with fuzzy parameters. The model parameter set results from the fuzzy combination of the four parameter sets through the degrees of activation of each pattern.

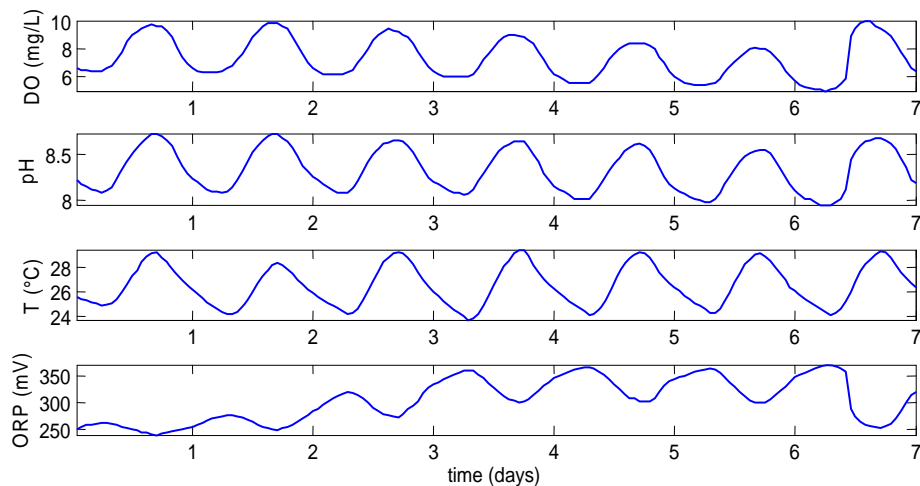


**Figure 6.** Performance of the patched DO model of Figure 5. The upper boxes represent the degree of activation of the four prototypes.

Figure 6 shows the performance of the patched model of Figure 5. Thus, in this application the pattern recognition algorithm is used to select the best combination of parameter values to be used in a dynamical model, in order to adapt to the varying environmental conditions in a complex ecosystem.

## 2.2. Daily and seasonal analysis of water quality data in the Arno river

As a second application of the pattern recognition algorithm of sect. 1.1, the detection of significant daily and seasonal trends in river quality was considered. This problem arises from the need to validate the data produced by the many automated monitoring stations deployed along the main rivers of Tuscany and their interpretation in terms of environmental changes. The basic water quality parameters (Dissolved oxygen, pH, Oxido-Reduction Potential and temperature) continuously measured by automated monitoring stations along the river. Significant variations of these parameters occur with daily and seasonal frequency. An example of these patterns is shown in Figure 7, where all the four main quality indicators follow a very clear diurnal cycle superimposed to a seasonal drift. The DO daily cycle is due to photosynthesis and is synchronous with the temperature variations.



**Figure 7.** Typical patterns in the main four quality variables recorded in the Arno river in the week 17 – 23 July 2003.

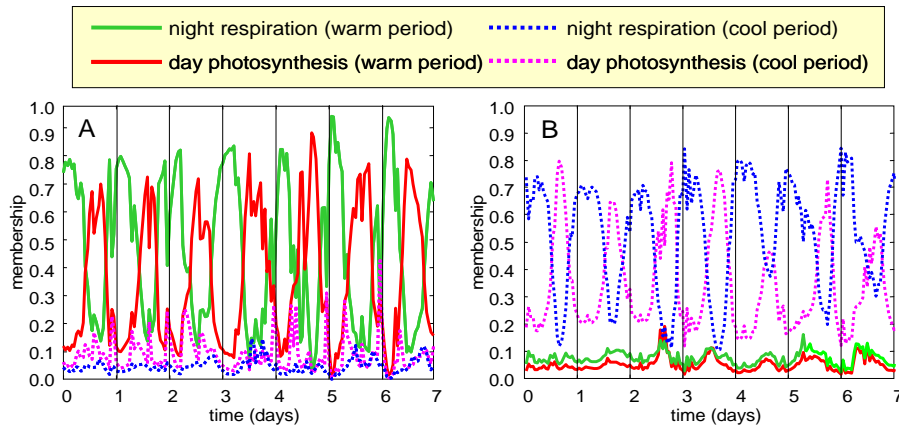
During the night DO decreases because the photosynthetic contribution is absent, whereas the degradation processes (night respiration) are still active. Photosynthesis is active during daytime and induces a  $\text{CO}_2$  depletion resulting in an increase in daytime pH [Chapra, 1997]. The Oxido-Reduction-Potential follows a similar path, but shows more emphasis on a long-time variation due to alkalinity changes.

The purpose of this study is to detect anomalies in the natural cycle and provide a pre-emptive discriminating power for the detection of potentially dangerous eutrophication levels. Hourly samples covering the time span of 2003 – 2007 were analysed and four clusters were selected to represent the information contained in the data.

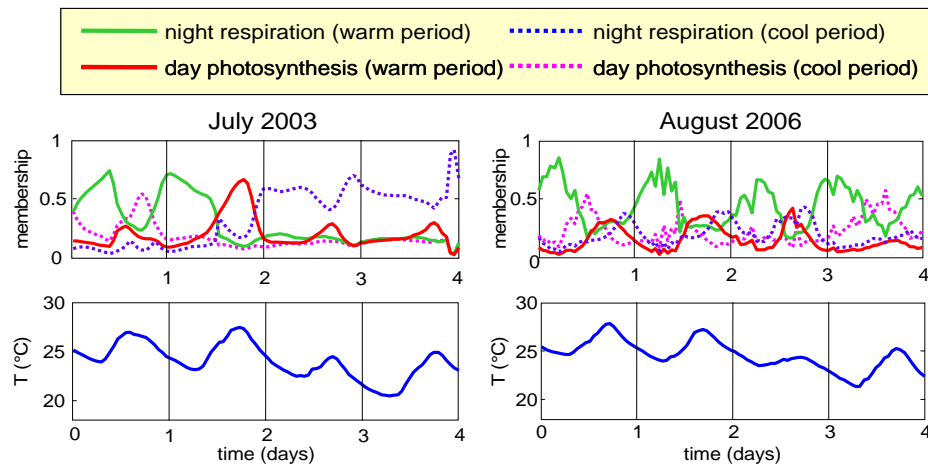
Analysing the daily patterns, all the validity measures of Eq. (6 – 8) indicated an optimal partition of four clusters, two of which can be related to the daily photosynthesis in warm and temperate periods and two to the night respiration in the same thermal conditions, as shown in the example of Figure 8, where the differing level of photosynthesis are clearly discernible, with an exchange between the warm and cool clusters.

In addition to isolating steady state steady patterns, like the ones in Figure 8, the fuzzy clustering algorithm can detect transition induced by a change in environmental variable, such as temperature. Figure 9 shows two changes from a warm to a comparatively cooler period, shown by a decreasing temperature. However, the change in membership reveals the differing patterns of the two transitions. In the July 2003 change analysis shows a marked shift from warm to cool period behaviour with limited day photosynthesis

due to cloudy weather, whereas in the August 2006 sample the shift is much less sharp, with a sustained warm night respiration, due to the mature algal population and massive daytime radiation.

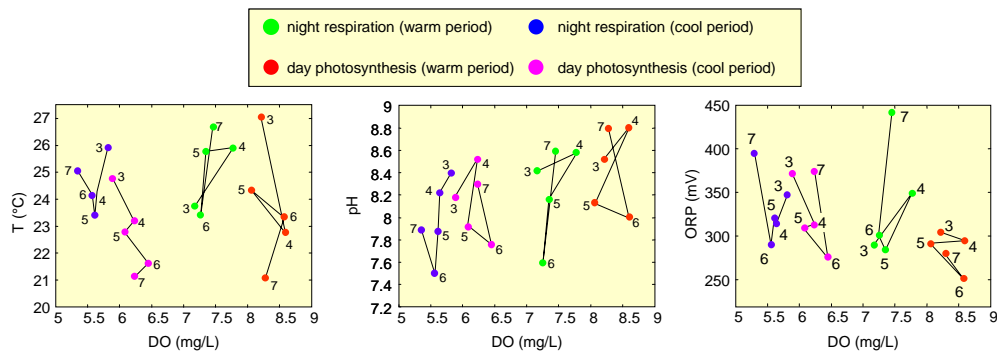


**Figure 8.** Steady-state alternating photosynthesis/respiration daily patterns during a warm (A) and a temperate (B) week. The vertical lines indicate the midnight of each day.



**Figure 9.** Transition from a warm period to a cool one. In both cases the temperature (shown in the bottom graphs) is the most evident indicator, whereas the memberships (top graphs) show differing transition modes. The vertical lines indicate midnight of each day.

Further to short-term analysis, more information can be extracted from the composite time-series covering the whole 2003 – 2007 period. Centroid displacements could be used to detect climatic anomalies reflected in the water quality parameters.



**Figure 10.** Centroid displacement showing the anomaly of 2004, obtained by analysing the composite time series. The numbers stand for the year after 2000.

As an example, Figure 10 shows the centroid variations in the various parameter combinations, clearly showing a continuing trend over the five years which have been considered and can possibly be related to climate changes.

Inspecting the centroid displacements over the years, shown in Figure 10, it appears that the most discriminating variable is Dissolved Oxygen (DO), with the four clusters maintaining the same relative position in all the figures, coherent with the daily trends shown in Figure 8. Clearly, the warm behaviour (day and night) correspond to the highest DO values, whereas the DO values of the cool behaviour is consistently lower. The middle plot (DO vs. pH) also reveals that an increased photosynthesis increases the pH as a consequence of CO<sub>2</sub> depletion, whereas the opposite is true for ORP.

### 3. CONCLUSION

A fuzzy pattern recognition algorithm has been presented, composed of a denoising wavelet filter and a fuzzy clustering procedure based on the Fuzzy Maximum Likelihood Estimates (FMLE) method. This composite procedure provided robust partitions and could be used in a number of environmental applications. Two of them have been presented here: the detection of daily patterns in the dissolved oxygen data from the Orbetello lagoon, and the analysis of water quality time-series from the automated measuring stations in the Arno river. In the first case, the algorithm was used to mechanize the necessary parameter-patching procedure to feed a long-term DO dynamical model for the lagoon. In the second example, the algorithm was used on two different time-scales: on a daily basis it provided a discrimination between daytime photosynthesis and night time respiration in differing climatic conditions, whereas on a yearly horizon, it could discriminate significant environmental anomalies.

As to the generality of the algorithm, in principle the two-stage procedure (wavelet denoising followed by fuzzy clustering) can be applied to any environmental problem, provided that proper indicators are preliminarily selected. The required level of wavelet decomposition should then be checked with a frequency analysis. Lastly, the effectiveness of the fuzzy partition should be checked with the indicators Eqs. (7 – 9). If these waypoints are observed, the algorithm can be applied to a wide class of environmental problems.

### 4. REFERENCES

- Abonyi J., *Fuzzy model identification for control*. Birkhäuser, pp. 273, Boston, 2003.
- Babuska, R., *Fuzzy modeling for control*. Kluwer Publ. Co., Amsterdam, 260 pp., 1998.
- Bezdek J.C., 1981. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, pp. 256, 1981.
- Chapra, S., *Surface Water-quality Modeling*. McGraw Hill Int. Ed., New York, pp. 844, 1997.
- Marsili-Libelli, S. and A. Müller, Adaptive fuzzy pattern recognition in the anaerobic digestion process, *Pattern Recognition Lett.* 17, 651 – 659, 1996.
- Marsili-Libelli S. and S. Arrigucci, Circadian patterns recognition in ecosystems by wavelet filtering and fuzzy clustering. In Pahl C., Schmidt S. and T. Jakeman (eds.) *iEMSs 2004 International Congress*, Osnabrück, Germany, 2004.
- Marsili-Libelli S., Fuzzy pattern recognition of circadian cycles in ecosystems. *Ecol. Model.*, **174**, 67 - 84, 2004.
- Giusti, E. and S. Marsili-Libelli, An integrated model for the Orbetello Lagoon ecosystem. *Ecol. Model.* 196, 379 – 394, 2006.
- Marsili-Libelli, S. and E. Giusti, Spatio-temporal modelling of the dissolved oxygen dynamics in the Orbetello lagoon by pattern recognition and fuzzy patching. Submitted to *Ecological Modelling*, 2007.
- 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.
- Percival, D.B., and A.T. Walden, *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, pp. 594, 2000.
- Strang, G. and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, 1996.