Jul 13th, 8:30 AM - 8:50 AM

ASSESSING INPUT-OUTPUT RELATIONS IN ENVIRONMENTAL DATA BY MEANS OF FUZZY CLUSTERING AND BAYESIAN INFERENCE

L. Bigozzi
University of Florence, lisa.bigozzi@stud.unifi.it

E. El Basri
University of Florence, emanuele.elbasri@stud.unifi.it

F. Ianniciello
University of Florence, francesca.ianniciello@stud.unifi.it

S. Marsili-Libelli
University of Florence, stefano.marsililibelli@unifi.it

I. Simonetti
University of Florence, irene.simonetti@stud.unifi.it

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

ASSESSING INPUT-OUTPUT RELATIONS IN ENVIRONMENTAL DATA BY MEANS OF FUZZY CLUSTERING AND BAYESIAN INFERENCE

L. Bigozzi[a], E. El Basri[b], F. Ianniciello[a], S. Marsili-Libelli[c], I Simonetti[b]

[a] *Graduate student, Department of Civil and Environmental Engineering, School of Engineering, University of Florence, Italy (lisa.bigozzi/francesca.ianniciello@stud.unifi.it)*
[b] *PhD Student, Department of Civil and Environmental Engineering, School of Engineering, University of Florence, Italy (emanuele.elbasri/irene.simonetti@stud.unifi.it)*
[c] *Department of Information Engineering, School of Engineering, University of Florence, Italy (stefano.marsililibelli@unifi.it)*

Abstract: When dealing with complex environmental datasets, it is also difficult to establish the strength of the input-output relation among variables. Correlation analysis may yield a preliminary indication, but is limited to the linear case. Mutual Information (MI) is a more powerful method which can establish input-output dependence regardless of the nature of their interaction. However, to avoid the heavy computational demand of MI, a simple method is presented based on fuzzy clustering and Bayes' rule. After a preliminary conditioning phase, the data are grouped by fuzzy clustering and approximated with the value of the most relevant centroid. Then the prior and likelihood probabilities are computed by frequentist methods by counting the occurrences of each sample with respect to the precomputed clusters. In this way the MI can be quickly computed, to yield the relative importance of the informative content of each input.

Keywords: Data mining; Uncertainty reduction; Environmental data processing; Fuzzy systems; Mutual Information; Bayesian inference.

1 INTRODUCTION

The causal relationships among environmental variables are not always apparent, so that it is often difficult to establish which variable is influencing which and to what extent. Ecosystem managers often ask which environmental variable has an impact on an observed phenomenon. In other words they are asking for a quantitative estimate of mutual influence among variables.

The simplest measurement of variable interdependence is the correlation analysis. but it is limited to a linear relation. On the other hand, nonlinear relationships are often encountered in ecosystems and cannot be properly captured by the linear correlation methods. Conversely, the Mutual Information (MI) criterion (Fraser and Swinney, 1986) uses the joint and marginal probabilities to quantify their dependence, hence more insight can be gleaned into the variables interactions by using this criterion, even in the case of a nonlinear relationship.

Given two random variables x and y, the Mutual Information (MI) measures their general interdependence better than the cross-correlation, which is limited to a linear dependence, and yields the reduction in uncertainty of *y* due to knowledge of the variable *x.* In the case of *N* discrete samples $\left(x_i, y_i \mid i = 1,...,N\right)$ the Mutual Information criterion is defined as

$$MI = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{xy}\left(x_i, y_j\right) ln\left(\frac{p_{xy}\left(x_i, y_j\right)}{p_x\left(x_i\right)\cdot p_y\left(y_j\right)}\right) \tag{1}$$

where $p_{xy}(x_i, y_j)$ is the joint probability density function (pdf) between variables x_i and y_j, while $p_x(x_i)$ and $p_y(y_j)$ are their marginal pdfs. The MI criterion of eq. (1) defines the extent to which the prior knowledge of *x* reduces the uncertainty on *y*. If they are independent, their MI is zero, because $p_{xy}(x,y) = p_x(x) \cdot p_y(y)$ hence the ratio is equal to one and its logarithm is zero, but if there is a dependence its extent will be expressed by the value of MI, regardless of the linear or nonlinear relation between the two variables.

The MI criterion has been largely applied in feature selection problems and often integrated into expert systems (Wang et al., 2004; Hoque et al., 2014; Bennasar et al., 2015) and it was also used in selecting inputs to neural networks in hydrological data-driven models (Battiti, 1994; Sharma, 2000; Bowden et al., 2005). The main difficulty in applying eq. (1) is the practical estimation of the joint pdf $p_{xy}(x_i, y_j)$. While histograms were originally used (Fraser and Swinney, 1986), more recent applications were based on kernel density estimation (Sharma, 2000; Bowden et al., 2005).

In this paper we follow a different approach and compute the conditional pdf $p_{xy}(x_i, y_j)$ as the posterior probabilities obtained by applying the Bayes' rule after the data have been grouped by fuzzy clustering. The whole algorithm is illustrated in Figure 1. It is assumed a priori which variable is the consequent $y \in \mathbb{R}^{1\times N}$ and which variables are the antecedents $(x_1, x_2, ..., x_p) = \boldsymbol{X} \in \mathbb{R}^{p\times N}$. After data regularization and denoising through spline smoothing, each antecedent variable is clustered into *c* classes, then its values are discretized by substituting each value with centroid of the cluster providing the maximal degree of membership. The likelihoods and marginal probabilities of the variables are then computed as the relative frequencies of each class. These quantities are then used to represent the conditional probabilities in eq. (1) required to compute the Mutual Information, where the joint probabilities are computed according to the Bayes' rule.

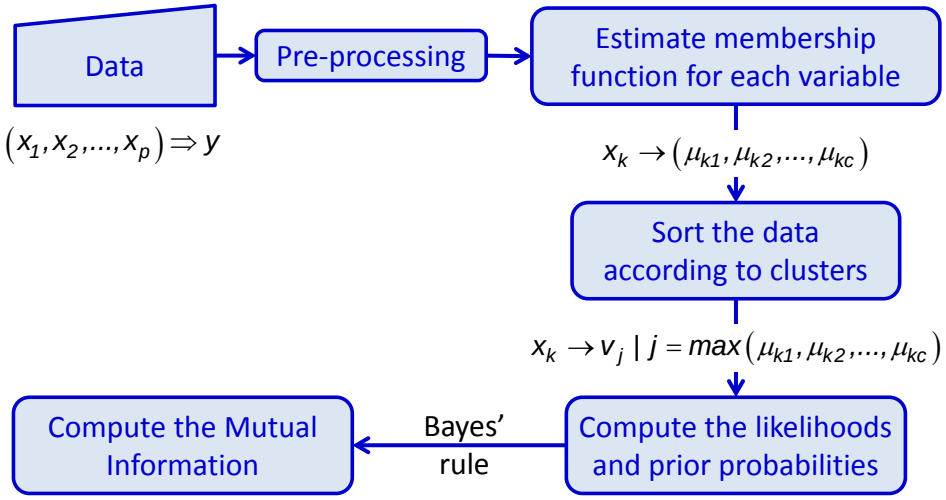

Figure 1. Algorithm flow diagram for computing the Fuzzy Mutual Information from a set of environmental data.

2 MUTUAL INFORMATION ALGORITHM

2.1 Trial dataset

To demonstrate the algorithm, a dataset describing the water quality in the Orbetello lagoon, in the Tyrrhenian Sea, central Italy, is considered. This shallow lagoon, whose dynamics has been thoroughly described by deterministic models (Giusti and Marsili-Libelli, 2005;Giusti and Marsili-Libelli, 2006; Giusti and Marsili-Libelli, 2009; Giusti et al., 2010), is a very fragile ecosystem, where oxygen depletion has caused severe dystrophic crises (Christian et al., 1996; Azzoni et al., 2001; Giordani et al., 2009). Though the inner workings of the nutrients/oxygen balance are fully explained by the previous literature, managers and stakeholders who are not familiar with mathematical models often ask for a simple data-driven approach to establish causal relations among environmental variables. The present analysis is aimed at determining which environmental variable among pH, oxido-reduction

potential (ORP), water temperature (T), and Salinity (Sal) has the highest influence on the dissolved oxygen (DO) concentration.

2.2 Data pre-processing

The environmental data may significantly differ in their accuracy and continuity. Missing data or outliers are often encountered. Though the algorithm to be described later is fairly robust, it is always a good practice to regularize the data by removing unwanted noise, replacing outliers and filling in missing data. Smoothing by cubic splines can provide the necessary regularization without eroding the information in the data. It can also replace outliers with their smoothed approximation and provide reasonable estimates for sparse missing data. Cubic spline approximation balances approximation vs. smoothing via a smoothing parameter α whereby more emphasis can be put on the approximation $(\alpha \to 1)$ rather than on smoothness $(\alpha \to 0)$. The composite objective function to be optimized is

$$\alpha \sum_{k=1}^{N} \underbrace{\left(y(k)-y_s(k)\right)^2}_{approximation} + (1-\alpha) \underbrace{\int \lambda(t) \left| \frac{d^2 y_s(k)}{dt^2} \right|^2 dt}_{smoothing} , \tag{2}$$

where y represent the data and y_s their smoothed approximation. The selection of α can be guided by comparing the frequency spectra of the original and smoothed time-series (Marsili-Libelli, 2016). The original and smoothed trial datasets used in this study are shown in Figure 2.

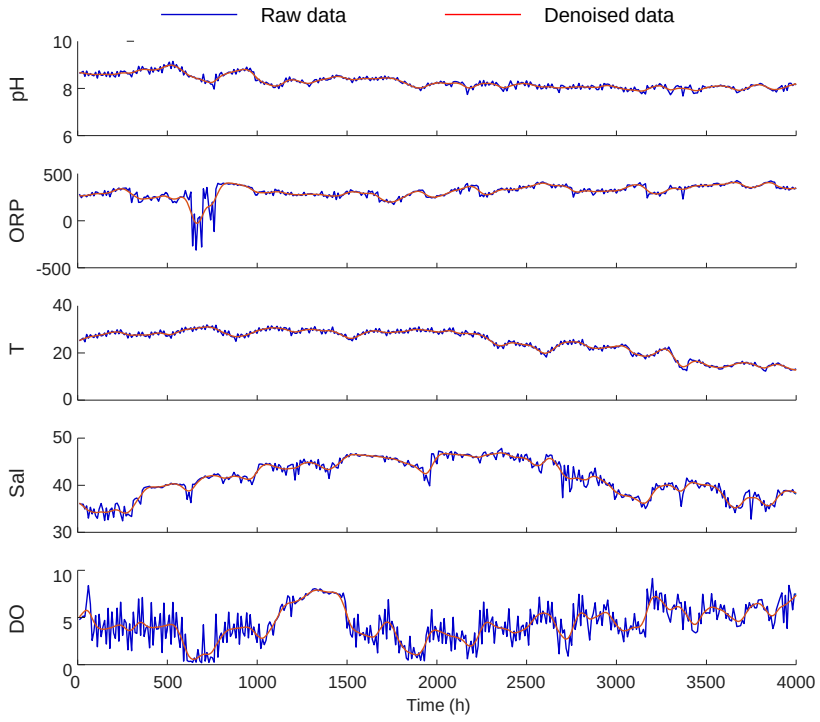

Figure 2. The trial data set used to demonstrate the FMI algorithm, before and after denoising. It consists of over 4000 hourly samples of the water quality parameters drawn from the monitoring station in the Orbetello lagoon between June and October 2003.

2.3 Classification by fuzzy clustering

For the subsequent computation of the likelihoods in the Bayes' rule, it is crucial that each antecedent is processed independently from all the others, therefore the values of each antecedent variable are separately clustered into *c* groups using the Fuzzy C-means (FCM) algorithm (Bezdek, 1981; Bezdek and Pal, 1995). Bezdek's idea was to minimize a fuzzy partition functional subject to the constraint of total membership

$$\min_{\mu,v} \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m d_{ik}^2 \quad with \quad d_{ik}^2 = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T (\boldsymbol{x}_k - \boldsymbol{v}_i) \; s.t. \sum_{i=1}^{c} \mu_{ik} = 1 \; . \tag{3}$$

In eq. (3) the Euclidean distance between the x_k*-th* sample and the v_i*-th* centroid is weighted by the degree of membership μ_{ik}, with the fuzzy exponent $m \in [1,\infty)$ defining the fuzziness of the partition. The FCM algorithm provides the optimal partition $(\boldsymbol{U})$ and the centroids $(\boldsymbol{V})$ as

$$\boldsymbol{U}: \; \mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}} \quad \begin{matrix} k = 1,...,N \\ i = 1,...,c \end{matrix} \quad and \quad \boldsymbol{V}: v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m v_k}{\sum_{k=1}^{N} (\mu_{ik})^m} \quad i = 1,...,c \; . \tag{4}$$

In this way each sample is replaced by the degree of membership to the clusters defined by the fuzzy exponent *m* and by the centroids $(v_1, v_2, ..., v_c)$, namely

$$x_k \rightarrow (\mu_{1k}, \mu_{2k}, ..., \mu_{ck}, m) | k = 1,...,N \; . \tag{5}$$

The partition efficiency is checked through the fuzzy partition entropy H_n defined as (Bezdek, 1981)

$$H_n = -\frac{1}{1 - \frac{c}{N}} \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik} \cdot log(\mu_{ik}) / N \; . \tag{6}$$

2.4 Estimating membership function for each antecedent variable

Once a partition of *c* centroids $(v_1, v_2, ..., v_c)$ has been created, all the data of that particular variable can be classified with respect to that partition according to the membership function of eq. (4). This numerical approximation, however has some drawbacks, given by the fact that for data far from the extreme centroids, all the memberships tend to *1/c*, while for intermediate data there are membership "humps" due to the total membership constraint in eq. (3). For these reasons, the numerical classification obtained by eq. (4) is approximated by analytical functions to the numerical membership data. The boundary (leftmost and rightmost) mfs are approximated with sigmoid Z or S curves, while the middle mfs are approximated by asymmetrical Gaussian curves. These analytical functions are fitted to the numerical mfs by least-squares, as described in Marsili-Libelli, (2016).

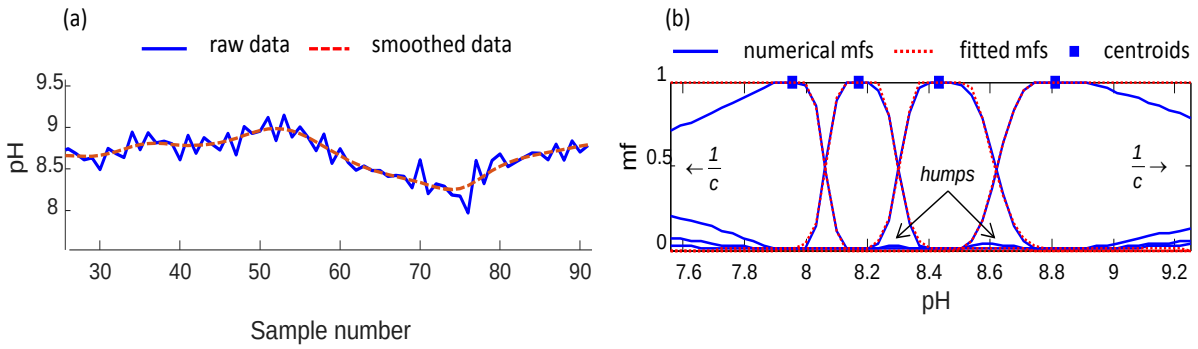

Figure 3. An example of data pre-processing: in (a) the pH data are denoised by cubic spline smoothing, while (b) compares the numerical clustering (with *c = 4* and *m = 1.75*) to their analytical mfs approximation, removing the humps and the inappropriate *1/c* asymptotic behaviour.

As to the number of clusters, it was observed that the relative partition entropy eq. (6) steadily decreased as the number of partitions increased, eventually stabilising around c = 8 for all the variables, as shown in Figure 4. Thus, a partition with eight clusters was adopted.

2.5 Variables discretization

Enumeration of all the possible values assumed by the variables would increase the problem dimensionality to an unmanageable level. For this reason the previous classification is used to reduce the set of the possible values of the variables. Each variable is replaced by the value of the centroid corresponding to its maximum degree of membership. Figure 5 compares the original data with their discretized counterpart obtained by the classification of the data.

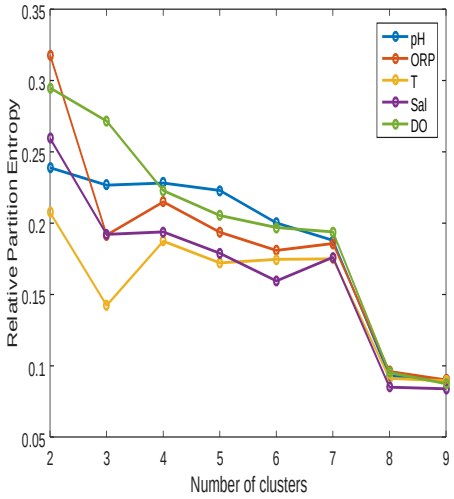

Figure 4. Partition entropy eq. (6) decrease with the increasing number of clusters.

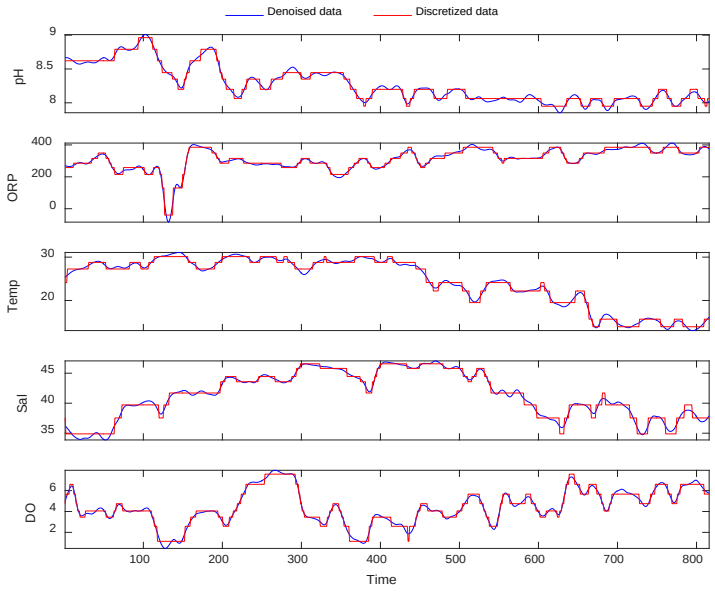

Figure 5. The eight-level discretized water quality variables obtained with an FCM partition are compared to the original denoised data.

2.6 Estimating the joint probabilities via Bayes' rule

Considering one antecedent at a time, so that the input/output implication becomes $x \Rightarrow y$, according to Bayes' rule (see e.g. Stone, 2013), the conditional probability of the output *y* subject to the occurrence of the antecedent *x* is given by

$$P(y \mid x) = \frac{P(x \mid y) \cdot P(y)}{P(x)}. \tag{7}$$

Since we have partitioned all the variables into *c* classes, their probabilities can be approximated by their relative frequencies, computed as the number of times that the value of each variable is approximated by the value of the centroid with maximal membership. Thus, eq. (7) can be rewritten by considering the classes of occurrence rather than the value of the variables, thus

$$P(y_j \mid x_i) = \frac{P(x_i \mid y_j) \cdot P(y_j)}{P(x_i)}, \tag{8}$$

where the priors, the likelihoods and the marginal probabilities are estimated respectively as

$$P(y_j) = \frac{\sum_{k=1}^{N} (y_k \subset v_j)}{\sum_{k=1}^{N} y_k}; \quad P(x_i \mid y_j) = \frac{\sum_{k=1}^{N} (x_k \subset v_i)}{\sum_{k=1}^{N} (y_k \subset v_j)}; \quad P(x_i) = \frac{\sum_{k=1}^{N} (x_k \subset v_i)}{\sum_{k=1}^{N} x_k}, \tag{9}$$

where in each of the three summation the notation $\sum_{k=1}^{N}\left(y_k \subset v_j\right)$ or $\sum_{k=1}^{N}\left(x_k \subset v_j\right)$ means counting the occurrences of either variable y_k or x_k being approximated by the v_j centroid with maximal membership. With these quantities the conditional occurrences defined by eq. (8) can be put in matrix form as follows

$$P(y \mid x) = \begin{bmatrix} P(x_1 \mid y_1) & \ldots & P(x_1 \mid y_c) \\ \ldots & \ldots & \ldots \\ P(x_c \mid y_1) & \ldots & P(x_c \mid y_c) \end{bmatrix} \in \mathbb{R}^{c \times c} \ , \tag{10}$$

from which the Fuzzy Mutual Information can be derived from eq. (1) using the marginal probabilities computed via the quantities defined by eqs. (9).

$$FMI = \frac{1}{c}\sum_{i=1}^{c}\sum_{j=1}^{c} P(x_i \mid y_j) ln\left(\frac{P(x_i \mid y_j)}{P(x_i)\cdot P(y_j)}\right) \ , \tag{11}$$

The contribution of each input is then normalized to the total FMI

$$fmi(j) = \frac{FMI(j)}{\sum_{j=1}^{c} FMI(j)} \ . \tag{12}$$

3 DISCUSSION

The trial dataset of Figure 2 was processed by the FMI algorithm just described and the results are shown in Figure 6. It can be seen that the oxido-reduction potential (ORP) has the highest influence on the dissolved oxygen, while the temperature (T) is the second most important parameter and salinity (Sal) coming last. This ranking is consistent with the chemistry of the lagoon (Christian et al., 1996; Azzoni et al., 2001; Giusti and Marsili-Libelli, 2006; Giusti and Marsili-Libelli, 2009), whereby the oxidized or reduced state of the sediment has a primary influence on dissolved oxygen by modulating the release of nutrients. Considering the various steps in Figure 1, the data pre-processing proved to be a crucial stage, because data irregularities could wrongly influence the subsequent clustering. The number of clusters in the partition should also be carefully considered, in terms of improved discrimination and best data separation.

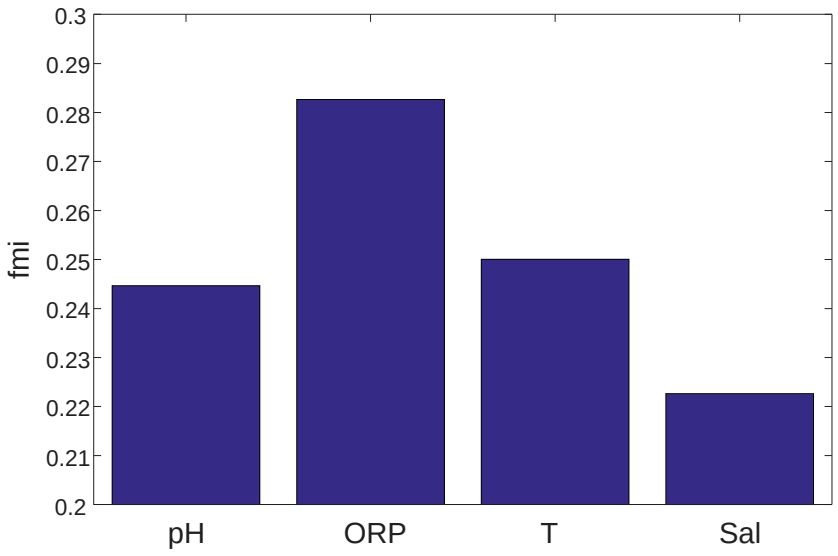

Figure 6. Relative fuzzy mutual information for the dataset of Figure 2.

4 CONCLUSION

This paper has presented an alternate method to compute the Mutual Information for datasets of environmental interest. One of the difficulties in the practical computation of the Mutual Information criterion (1) is the evaluation of the joint probabilities, for which the use of Gaussian kernel function has been suggested (Sharma, 2000), later extended to datasets which do not follow a Gaussian distribution (Li et al., 2015). In all cases the resulting algorithms are computationally demanding.

In this paper a simpler alternative approach was proposed, by estimating the joint probabilities via the Bayes' rule, preceded by data denoising and fuzzy clustering. By approximating the variables with the clusters, the probability of occurrence of each input variable was estimated by a frequentist approach, which counted the number of occurrences in each cluster. In this way the priors, the likelihoods and the posterior probabilities to be used in the Bayes' rule could be easily computed.

The method was tested on a water quality dataset drawn from the Orbetello lagoon and it produced a ranking among the variables which influence the dissolved oxygen (DO) level. The oxido-reduction potential (ORP) appears to be the first factor controlling the DO. This is in agreement with the observations, whereby this factor actually provides information about the oxidized or reduced state of the sediment and the consequent nutrient resuspension. It becomes the primary indicator of a dystrophic crisis, when ORP plummets to negative values in the order of minus several hundreds of millivolts. The second most important factor is the water temperature, almost equalled by pH which is an indicator of the photosynthetic oxygen production. In fact, when this process increases, the resulting CO_2 depletion results in a pH increase. Salinity is comparatively less important, as it influences the oxygen saturation level, but a lesser extent than temperature. So the ranking shown in Figure 6 is compatible with known mechanisms at the basis of the DO dynamics in shallow waters.

5 BIBLIOGRAPHY

Azzoni, R., Giordani, G. M., Bartoli, M., Welsh, D. T., and Viaroli, P. L., 2001. Iron, sulphur and phosphorus cycling in the rhizosphere sediments of a eutrophic Ruppia cirrhosa meadow (Valle Smarlacca, Italy). *Journal of Sea Research*, 45, 15–26.

Battiti, R., 1994. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks*, 5, 537–550.

Bennasar, M., Hicks, Y., and Setchi, R., 2015. Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, 42, 8520–8532.

Bezdek, J. . and Pal, N. R., 1995. On cluster validity for the Fuzzy c-Means model. *IEEE Trans. on Fuzzy Systems*, 3, 370–379.

Bezdek, J. C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, pp. 256.

Bowden, G. J., Maier, H. R., and Dandy, G. C., 2005. Input determination for neural network models in water resources applications. Part 2. Case study: Forecasting salinity in a river. *Journal of Hydrology*, 301, 93–107.

Christian, R. R., Forés, E., Comin, F., Viaroli, P. L., Naldi, M. C., and Ferrari, I., 1996. Nitrogen cycling networks of coastal ecosystems: influence of trophic status and primary producer form. *Ecological Modelling*, 87, 111–129.

Fraser, A. M. and Swinney, H. L., 1986. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33, 1134–1140.

Giordani, G., Zaldívar, J. M., and Viaroli, P., 2009. Simple tools for assessing water quality and trophic status in transitional water ecosystems. *Ecological Indicators*, 9, 982–991.

Giusti, E., Marsili-Libelli, S., Renzi, M., and Focardi, S., 2010. Assessment of spatial distribution of submerged vegetation in the Orbetello lagoon by means of a mathematical model. *Ecological Modelling*, 221, 1484–1493.

Giusti, E. and Marsili-Libelli, S., 2006. An integrated model for the Orbetello lagoon ecosystem. *Ecological Modelling*, 196, 379–394.

Giusti, E. and Marsili-Libelli, S., 2005. Modelling the interactions between nutrients and the submersed vegetation in the Orbetello Lagoon. *Ecological Modelling*, 184, 141–161.

Giusti, E. and Marsili-Libelli, S., 2009. Spatio-temporal dissolved oxygen dynamics in the Orbetello lagoon by fuzzy pattern recognition. *Ecological Modelling*, 220, 2415–2426.

Hoque, N., Bhattacharyya, D. K., and Kalita, J. K., 2014. MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, 41, 6371–6385.

Li, X., Maier, H. R., and Zecchin, A. C., 2015. Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models. *Environmental Modelling and Software*, 65, 15–29.

Marsili-Libelli, S., 2016. *Environmental Systems Analysis with MATLAB*, CRC Press, Boca Raton, FL., USA, pp. 546.

Sharma, A., 2000. Seasonal to interannual rainfall ensemble forecasts for improved water supply management: A strategy for system predictor identification. *J. of Hydrology*, 239, 232–239.

Stone, J. V., 2013. *Bayes' Rule, a tutorial introduction to Bayesian analysis*, Sebtel Press, Sheffield, UK, pp. 170.

Wang, X., Wang, Y., and Wang, L., 2004. Improving fuzzy c-means clustering based on feature-weight learning. *Pattern Recognition Letters*, 25, 1123–1132.