DATA MINING FOR ENVIRONMENTAL ANALYSIS AND DIAGNOSTIC: A CASE STUDY OF UPWELLING ECOSYSTEM OF ARRAIAL DO CABO*

*Gilberto Carvalho Pereira[1**]; Ricardo Coutinho[2***] and Nelson Francisco Favila Ebecken[1****]*

[1]Universidade Federal do Rio de Janeiro - Programa de Engenharia Civil
Centro de Tecnologia
(Cidade Universitária, Ilha do Fundão, Bloco B, Caixa Postal 68506 Rio de Janeiro, RJ, Brasil)

[2]Instituto de Estudos Paulo Moreira – Departamento de Oceanografia
(Rua Kioto, 235, Praia dos Anjos, 28430-000, Arraial do Cabo, RJ, Brasil)

**gcp@coc.ufrj.br; ***rcoutinhosa@yahoo.com; ****nelson@ntt.ufrj.br

ABSTRACT

The Brazilian coastal zone presents a large extension and a variety of environments. Nevertheless, little is known about biological diversity and ecosystem dynamics. Environmental changes always occur; however, it is important to distinguish natural from anthropic variability. Under these scenarios, the aim of this work is to present a Data Mining methodology able to access the quality and health levels of the environmental conditions through the biological integrity concept. A ten-year time series of physical, chemical and biological parameters from an influenced upwelling area of Arraial do Cabo-RJ were used to generate a classification model based on association rules. The model recognizes seven different classes of water based on biological diversity and a new trophic index (PLIX). Artificial neural networks were evolved and optimized by genetic algorithms to forecast these indices, enabling environmental diagnostic to be made taking into account control mechanisms of topology, stability and complex behavioral properties of food web.

RESUMO

A zona costeira brasileira apresenta grande extensão e variedade de ambientes. Contudo, pouco se sabe sobre sua diversidade biológica e o funcionamento dos ecossistemas. Como mudanças ambientais são constantes, é muito importante distinguir entre variabilidade natural e antrópica. Nesse cenário, o objetivo deste trabalho é apresentar a metodologia para o desenvolvimento de um Sistema Inteligente de Gerenciamento Integrado do Ecossistema Costeiro (SIGIEC) capaz de acessar o nível de qualidade e saúde ambiental através do conceito de Integridade Biológica. Foram usadas séries temporais de dez anos de parâmetros físicos, químicos e biológicos para extrair conhecimento e gerar modelos de regras de associação para classificar sete diferentes tipos de condições ambientais, analisadas através da diversidade biológica e um novo índice trófico (PLIX). Redes neurais artificiais foram otimizadas por algoritmos genéticos para fazer predições desses índices e apresenta-se um diagnóstico ambiental baseado na análise dos mecanismos de controle da topologia, estabilidade e propriedades do comportamento complexo de redes alimentares .

Descriptors: Data Mining, Intelligent Systems, Environmental Diagnostic, Ecosystem Management, Biological Integrity.

Descritores**:** Mineração de dados, Sistemas inteligentes, Diagnóstico ambiental, Gerenciamento de ecossistemas, Integridade biológica.

INTRODUCTION

The Brazilian coastal zone presents a large variety of ecosystems but little is known about its biodiversity instead of this, are subjected to discharge of contaminants via sewage, industrial effluents, dredged material, accidental chemical and oil drilling spills, storm, urban and agricultural runoff and atmospheric deposition from land-based activities like worldwide. Multiple agents may differentially cause acute and chronic impact (NELSON, 2003). According to COSTANZA et al. (1997) biodiversity is extremely valuable to humankind, accounting for over 60% of the economic value of the biosphere. Despite their immense value, marine ecosystems are deteriorating rapidly due to human activities, especially physical alteration of habitat, overexploitation, exotic species introduction, global climatic change, and pollution (HIKSON, et al., 2001). The most threatened systems are estuaries, mangrove, coral reefs and coastal rocks, all under strong

(*) Paper presented at the 1st Brazilian Congress of Marine Biology, on 15-19 May. Rio de Janeiro, 2006.

anthropic pressure. However, very little is known about the connectivity degree between these different ecosystems and their associated communities (COWEN et al., 2000). In Brazil, the determination of environmental quality patterns is still based in the concept of maximum admissible concentration level of pollutant according National Environmental Council (CONAMA). However, nowadays many international environmental agencies (e.g. United State Environmental Protection Agency, European Environmental Agency, Japan Environment Management Agency and Industry, etc.) have supported the use of biological criteria as indicators due to the concept of tolerance limits of biota be more close to the reality detecting the outcomes of many different disturb in ecosystem. In this way the ecological communities, plays a complex net of trophic interactions comprising many types of organism and should be used as a baseline indicator of ecological status through its biological integrity. Biological integrity here must be understood as an ideal condition when the community is minimally impaired by human activities. In order to determine the degree to which this community approaches biological integrity, it is necessary to measure attributes of their structure and function and be able to distinguish between natural and anthropogenic impacts. So, a deep inspection on data related to benthos, plankton, necton and environmental variables is necessary. As any other data set, this one can to come out with hidden patterns and relations that are needed to be extracted to increase our knowledge over the ecosystem structure and function (PEREIRA, 2005). Thus, the aim of this work is:

- Investigate the long-term ecological patterns among some environmental variables and meroplankton larval supply of epibenthic invertebrate fauna;
- Assess the system trophic status;
- Propose a methodology to develop an intelligent system to operate on the coastal zone integrated management as decision support system based on ecosystem health and biological integrity concepts.

STUDIED AREA

The case studied is the extractive marine reserve of Arraial do Cabo (23°S, 42°W), northeast of Rio de Janeiro state (Fig. 1), Brazil. At this place, there is sometimes, upwelling events, where inorganic nutrients are supplied to euphotic zone by the exchange of water between nutrient-depleted surface water (Brazil current) and nutrient-rich deeper water coming from the up flow of South Atlantic Central Water (SACW), resulting in direct impact on the quantity and composition of species shifting the trophic structure well described in Valantin, (1988). This place is one of the most attractive sea and landscape for tourist and recreational activities like diving, sailing and fishing significantly contributing to the local economy, but there is an urban disorder increasing at the city. Many people pull out shells from the bottom of a complex of hypersaline coastal lagoon to a calcareous industry which takes sea water for system refreshment. Others, less fortunate, lives removing organism, like mussels, from intertidal zone and sells to the regional market beyond the fact, this place has became a operational support base of oil drilling companies because the small harbor and its proximity of petrol Campus basin. It can still be seem, a little marine farm in a small cove. Moore et al. (1997), states that coastal zone is an environment where conflicting interests meet; developmental, industrial and conservational. Management is a question of reconciling these differing viewpoints.

DATA AND METHODOLOGY

Available data were routinely collected and concerns to a weekly harvested medium-term time-series of physical, chemical and biological gradients coming from November of 1994 to December of 2004, as shown in Table 1, with their mean and standard deviation values. Physical and chemical variables demonstrate the variability of hydrological characteristic of environment as a function of interchangeable periods of upwelling and subsidence events defining quality patterns of different water masses. The biological variables are composed by chlorophyll (mg/l.) measurements as estimation of microalgal biomass but probably it also contains all free living autotrophic bacteria of water column both influenced qualitative and quantitatively by nutrient entrances that in the other hand, supply it self as feeding material for meroplankton larvae which are expressed in numbers of organisms per cubic meter of water. LD variable is a very young type of larva which does not allow a right identification but certainly it will be a mollusk. The functioning of a complex system, like ecosystems, not always is perceived by experts due to their dynamical behavior, huge number of elements interconnected and emergent properties difficult to interpret. Notwithstanding, recent technological advances allow the application of efficient Data Mining algorithms for knowledge discovery. Data Mining is a multidisciplinary research field that includes different areas such as statistical methods, machine learning, data base, expert systems, data visualization techniques and high performance computing in a high connected manner (FAYYAD et al., 1996) whose process can be seen in Figure 2.

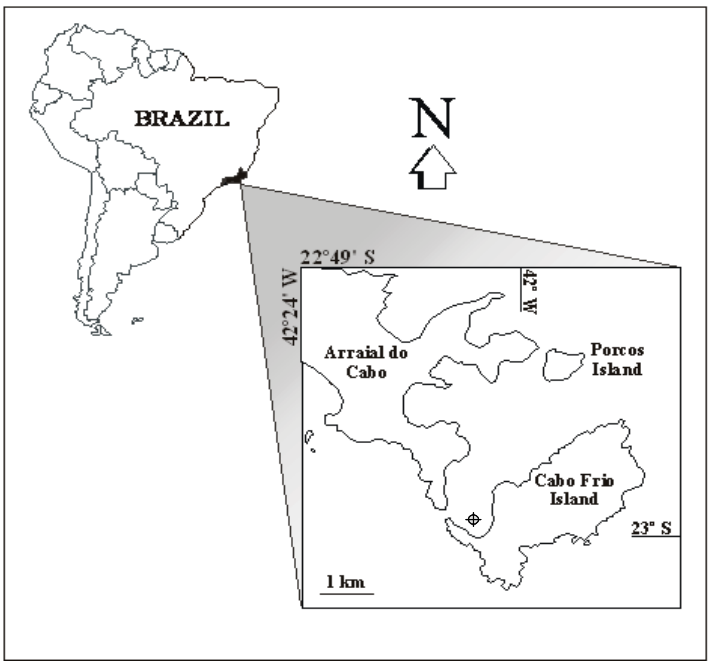

Fig. 1. The studied area.

Table 1. Summary of available data.

Variables	Mean	S. Dev.
Temperature (ºC)	22.59	1.88
Salinity (g/L)	35.48	0.62
Oxygen (ml/L)	5.3	0.5
Phosphate (PO4) (mg/l)	0.27	0.21
Nitrite (NO2) (mg/l)	0.07	0.09
Nitrate (NO3) (mg/l)	0.62	0.94
Chlorophyill-a (mg/m^3)	0.95	0.84
Cirripedia (Org/m^3)	155.12	308.71
Mytilidae (Org/m^3)	99.01	424.93
Polychaeta (Org/m$^{3)}$	18.06	82.54
Decapoda (Org/m^3)	17.71	42.67
LD (Org/m^3)	72.19	195.95
Ostreidae (Org/m^3)	28.64	82.38
Cypris (Org/m^3)	49.41	450.45
Ascidiaceae (Org/m^3)	17.08	73.32
Bryozoa (Org/m^3)	2.45	8.58
Bivalvia (Org/m^3)	9.92	28.8

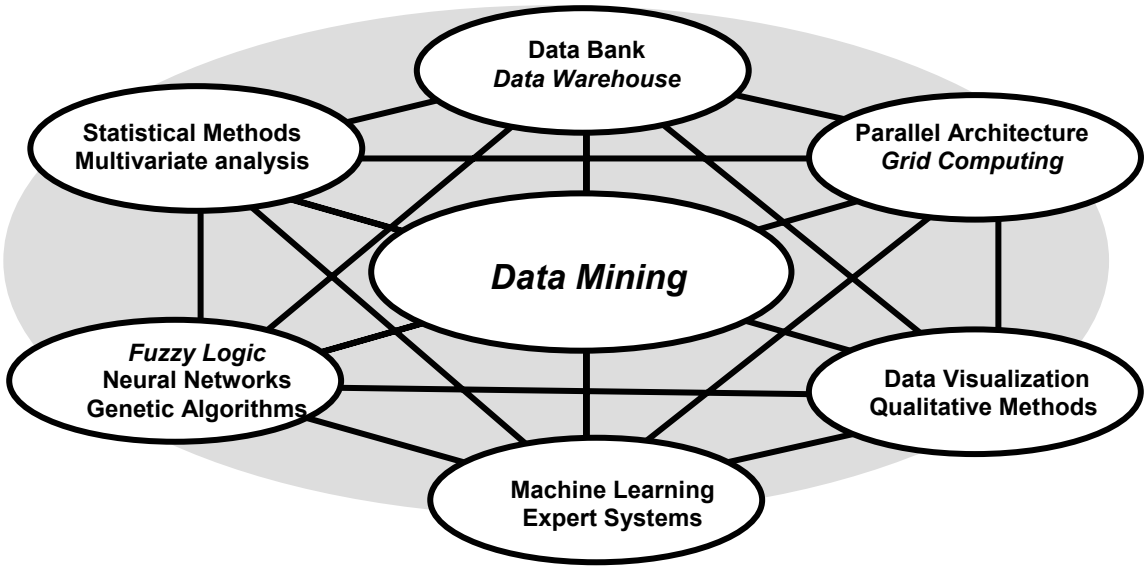

Fig. 2. Data mining process.

In other to access the data structure, a matrix of 17 variables and 511 objects, and see how it is ordinate, two statistical approaches was made. Firstly, a Principal Component Analysis (PCA) was performed. Based on a correlation matrix, this technique establishes a set of orthogonal factors providing information about ecological similarities of samples. Table 2 depicts the results of this analysis. It is worth to say, this arrangement (all factors) accounts for only 66,27 % of total variance of data set, in such a way this result does not allow to discard any variable. The second statistical approach and attempting to portray a more clear structure and composition of ecosystem functional units, a clustering analysis were performed. There is, in literature, many available algorithms for this task, the choice depends on the application, type of data and the subject of what are reached. For clustering variables, the well-known Ward method used with Spearman coefficient was applied. The Figure 3 presents a dendrogram produced by this approach. It shows seven groups and is close to the seven PCA components. Almost every cluster has the same variables and two big clusters corresponding to the macrostructure of ecosystem with environmental variables in one side and biological ones at the other.

In the other hand, to access of matrix objects it was applied the K-means clustering method (HAN, 2001) sing a Euclidean distance as a measure of similarity, depicted at Figure 4. Traditionally, the classification of water masses is made through temperature/salinity diagrams like Table 3 from the Brazilian Navy hydrograph department. These intervals were implemented and applied to the available data showing fifteen examples that does not belong to any of these intervals (data not shown) suggesting a new class of water. Coincidentally, The Figure 4 presents seven clusters at the random level two of similarity.

The next step, all variables was discretized into five intervals (low, mean-low, average, mean-high and high) according interviews with plankton experts who set the cut points of variables. Over the discretized matrix was applied a modified Apriori algorithm (AGRAWAL, 1994), in order to mine associations rules (if-them type). It is able to relate environmental and biological interval of variables and presents a base line of patterns of occurrence that can be used as an initialization tool for visual inspections of food web trophic interactions. Examples of associations rule are:

If TEMP3 and SAL4 and ASC5	them CHLO1	[8.20%, 100%]	(1)
If TEMP 3 and SAL2 and ASC1	them OXI1	[0.19%, 100%]	(2)
If SAL2 and OXI3 and TEMP4	them ASC2	[0.19%, 100%]	(3)
If OXI3 and TEMP5 and CIR2	them ASC4	[0.19%, 100%]	(4)
If TEMP3 and OXI3 and ASC2	them SAL2	[0.19%, 100%]	(5)

Table 2. Loading variables factors.

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Temperature	-0,01526	0,438974	0,042074	-0,03489	-0,61646	0,222757	0,147892
Salinity	-0,0413	0,010708	-0,03088	-0,03397	0,074022	**-0,77989**	0,121073
Oxygen	0,05243	0,058959	-0,00121	-0,01063	**0,717425**	-0,07693	-0,0089
Phosphate	0,033359	-0,69605	-0,12586	0,029123	0,068003	0,20494	0,135587
Nitrite	-0,02423	**-0,74020**	0,067472	-0,02516	0,056542	-0,06867	-0,11146
Nitrate	0,010184	-0,67343	0,070083	-0,04332	-0,06991	-0,1587	-0,02735
Chlorophyll-a	-0,05441	0,000942	-0,00419	-0,02502	**0,745648**	0,124475	0,063845
Cirripedia	0,044237	-0,02226	**0,877982**	0,087916	-0,05515	0,003237	0,048265
Mytillidae	**0,910663**	0,052887	0,0785	0,017311	0,021604	0,013369	-0,06695
Decapoda	0,25986	0,044231	0,137367	**0,792501**	-0,03531	-0,08866	0,287693
Polychaeta	0,056358	-0,00506	**0,824314**	0,040364	0,038222	0,024274	0,14484
LD	0,306453	-0,07542	0,299894	-0,05914	-0,14375	0,120434	0,555725
Ostreidae	**0,822054**	0,037852	0,180147	-0,00789	0,01522	-0,05596	0,086852
Cipris	-0,06507	0,010937	0,025042	**0,893926**	0,003344	0,013917	-0,13204
Ascidiaceae	-0,01127	0,091166	0,063857	0,085879	0,072653	-0,11072	**0,874755**
Bryozoa	0,650738	-0,14575	-0,20175	0,17108	-0,02188	0,196483	0,186756
Bivalvia	0,037816	0,098235	0,003916	-0,08131	0,0245	**0,779906**	0,076564
Var. Expl.	2,110354	1,733892	1,671236	1,490104	1,502387	1,436741	1,322542
Prp.Total	0,124138	0,101994	0,098308	0,087653	0,088376	0,084514	0,077797

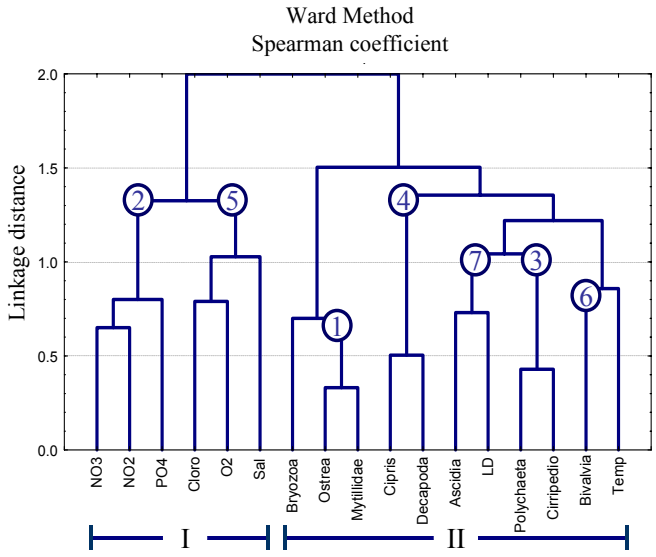

Fig. 3. Ward dendrogram with Spearman coefficient.

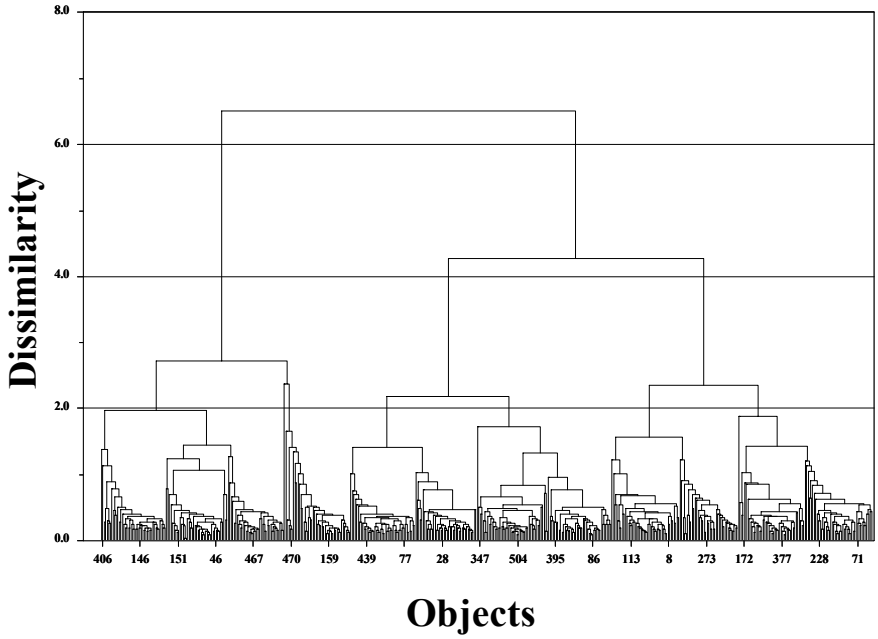

Fig. 4. K-means dendrogram of objects.

Table 3.Temperature and salinity characteristics of Cabo Frio water masses.

TYPE OF WATERS	T°C	S %
ACAS	T<18	S<36
ACAS/COASTAL WATER	18<T<20	35,4<S<36
COASTAL WATER	T>20	S<35,4
ACAS/TROPICAL WATER	18<T<20	S>36
COSTAL WATER/TROPICAL WATER	T>20	35,4<S<36
TROPICAL WATER	T>20	S>36

(from Navy Hydrographyc Department)

Each variable names was abbreviated with its first three or four letters like Temp for temperature; Sal for salinity; Chlo to chlorophyll; Pol to Polychaeta; Asc to Ascidiaceae and so on. All abbreviations are followed by a number that correspond to the five intervals of discretization. Thus, interval 1 for instance in environmental attributes means low concentration of a given nutrient or salt, in the case of biological variable (larvae), it means absence of the organism, while number 5, is the highest values respectively. The exception is chlorophyll where interval 1 means very low concentration once it is also a chemical measurement. The percentages between clasps mean firstly the support value which is the occurrence of the rule along the data set and the second, its confidence level. Some rules can appear very frequently in the data set, while others are rare. The subject is to find interesting rules which is unknown or defy the expert knowledge. This approach can still come out the sensibility or tolerance level of studied organisms to environmental variability.

Another possibility, is to set the clustering results (k-means classes in this case) at the consequent part of rules to become a target to be reached and use this special set of rules as a classifier rule model of the different water masses (LIU, 1998). The result is a classifier build on associations. This approach can still give an insight on the population composition and preferential interrelations among species. The following are examples of classification rules:

If TEMP3 and BRY1 and POL1 and CHLO1 them Class 1 [3.76%, 100%] (6)

If DEC3 and BRY5 and POL4 and CHLO2 them Class 2 [0.57%, 100%] (7)

If LD2 and MIT2 and CIP1 and CHLO3 them Class 3 [12.26%, 85%] (8)

If NTO2 and NTI3 them Class 4 [1.37%, 100%] (9)

If BIV5 and NTI1 and MIT4 them Class 5 [12.52%, 81%] (10)

If BRY1 and CHLO5 them Class 6 [2.52%, 100%] (11)

If TEMP3 and CIR5 and DEC3 them Class 7 [0.96%, 100%] (12)

It is clear that the absence of bryozoans and polychaeta (rule 6) indicates the class number 1 while high incidences of bryozoan going along with mean-high values of polychaeta occurrence (rule 7) classify the class number 2. In addition, the class 4 (rule 9) can be separated only by the nitrite and nitrate concentrations. It can be verified that class 5 is characterized by the huge concentrations of bivalves larvae followed of mean-high counts of mytilidae. Chlorophyll is the main attribute to identify class 6 and in the same way, high numbers of cirripedia and average values of decapoda is used to establish class 7.

Once this type of model can access into identification and classification of different conditions of ecosystem, the next question is how to measure and express its condition, species richness and trophic status. The answer for these tasks is the development and usage of some indices which embody easiness to measure in operational conditions, comprehensibility to be understandable in its equation, represent objects and process of ecosystem behavior and predictable. The real subject of management work is to adopt conservation strategies that give guarantee of a minimum threshold of quality of life support so, a measure of biodiversity as richness seems necessary. The literature is full of these indices but this work used the derived information theory Margalef´s (1951) index as follow:

$$D = (S - 1) / \log_2 N \quad (13)$$

where:

S is the number of species in the sample and N is the total number of individuals.

In the same way, to give an insight into trophic conditions Wollenweider et al. (1998) developed the Trix index integrating Chlorophyll (ChA), Oxygen as absolute percent deviation from saturation (aD%O), mineral nitrogen (minN) and phosphorous (P) presented as:

$$TRIX = k/n * \Sigma (M_i - L_i) / (U_i - L_i) \quad (14)$$

where:

k = 10 (scale the results between 0 and 10), n = 4 (number of variables are integrated, M_i = measured value of variable i, U_i = upper limit of variable i, L_i = lower limit of value i.

Although Trix index were originally developed to lakes, Wollenweider (1998) extended it to estuarine and coastal areas. Trix is defined by a linear combination of the logarithms of its four state variables and makes the link between physical-chemical state of environment and the first trophic level, primary producers. However, our data presents many meroplankton larval variables. The idea here is to maintain the same expression, however, with a different set of attributes. It was used Chlorophyll as estimation of primary production, nitrogen/phosphate ratio as the primary production can be nutrient limited according Pereira et al. (2002), oxygen concentration and the sum of total larva in the sample. This new set of integrated variables is now called PLIX, a planktonic index, which can express two trophic levels and be easily extended to total zooplankton. Like Trix, the new index Plix suffered a log transformation and the data exceeding the mean ± 2.5 STD was desconsidered. In practice, only few data are eliminated and this procedure allow to obtain normalization and variance stabilization. Margalef, (1951) states that this kind of transformations of rough data really proves appropriate for parameters referred to phytoplankton populations and to the environmental factors strongly influenced by organisms. The maximum Plix observed unit along the time series was 5.92 in relation to larval increase which demonstrate the herbivore pressure in the system. The minimum and average values of Plix were 1.82 and 3.42 respectively in lower concentrations of larva.

An artificial neural network (ANN) model was used to map these indices and simulate predictions. Recently, this technique can be verified in many ecological applications (e.g. prediction of environmental factors, Maier, 1993; algal production, Scardi, 1996; modeling algal blooms Recknagel, 1997, Maier et al., 1998; classification of habitat characteristics, Lek, 1996, Ozesmi, 1999, Bradshaw, 2000; stream fishes assemblages, Oakes et al., 2005; hierarchic models of community, Olden et al., 2006 among others).

This approach was also tested to classify water masses like rules based model. A typical ANN consists of interconnected processing elements that are arranged in layers; an input layer, one or more hidden layers and an output layer. Let us first have a look at this smallest unit of the network, a neuron i with its couplings to outer neurons. A neuron j in this state S_j sends a signal with strength J_{ij} S_j to neuron i , where J_{ij} denotes the coupling between neuron i and neuron j. Summing all incoming signals at neuron i leads to the so-called membrane potential m_i of neuron i . This potential determines the new state of neuron i, in the simplest case the new state is justed defined as S_i^{new}.

The information to be processed enters the network through an input layer of neurons which influences an output layer in a way that is given by the couplings in-between the neurons. In this system, the weighted synapses are in general unidirectional. The unidirectional or feedfoward networks of this type are usually called perceptrons (BLOCK, 1962). These weighted inputs ($w_{ji}x_i$) are summed and a threshold value (θ_j) is added, producing a single activation level for processing element (Ij):

$$I_j = \Sigma\, w_{ji}\, x_i + \theta_j \qquad (15)$$

This activation level constitutes the argument of a transfer function (f(.)), such as a linear, sigmoid or hyperbolic tangent function, which produces the node output:

$$y_j = \mathrm{f}\,(I_j) \qquad (16)$$

This output is passed to the weighted input connections of many other processing elements.

The most successful method to train this network is known as error back-propagation (RUMELHART, et al., 1986). The desired relationship is learned by repeatedly presenting examples of desired input/output relationship to the network and adjusting the model coefficients (i.e. the connection weights) in order to get the best possible match between the historical values and those predicted by the model. The training process involves the following basics steps:

(1) The connection weights are assigned small, arbitrary values.
(2) A training sample is presented to the network, producing a network output.
(3) The global error function is calculated:

$$E = \tfrac{1}{2} \sum (o_\mathrm{d} - o_\mathrm{p})^2 \qquad (17)$$

Where E is the global error function; o_d is the desired output; o_p is the output predicted by the network.

(4) The connection weights (w) are adjusted using the gradient descent rule:

$$\Delta w(\mathrm{t}) = -\,\eta \partial \mathrm{E}/\partial w + \mu \Delta w\,(\mathrm{t\text{-}1}) \qquad (18)$$

Where η is the learning rate; μ is the momentum value.

The weights may be update after the presentation of each sample or after a number of training samples have been presented to the network. The number of training samples presented to the network between weight updates is called the epoch size (ε).

Steps 2 to 4 are repeated until certain stopping criteria are met. For example, training may be stopped when a fixed number of training samples have been presented to the network, when the global error function is sufficiently small or when there is no further improvement in the forecasts obtained using an independent data set.

The problem in building a neural network is not so much to define the local learning rule, but to find out how to arrange the neurons in the network and how to choose their couplings in order to obtain a desired learning behavior. Usually, the success or failure of artificial neural networks models are related to their architectures and topology. The concept in the present approach is the application of evolutionary methods to the structure of neural networks and internal parameters. The structure of the neural network will be determined by the genetic algorithm (GOLDBERG, 1989) and no global learning rule has to be specified for a given problem.

These algorithms are a computational abstraction of biological evolution that can be used to solve optimization problems. These models consist on three basic elements: a fitness measure which governs an individual's ability to influence future generations, a selection and reproduction process which produces offspring for next generation and genetic operators which determine the genetic makeup of the offspring. The individuals also called chromosomes represent possible solutions on the problem. Chromosomes are chains of bits or binary code vectors which takes the genetic information, it means, the number of layers of neural architecture, number of neurons of such layer, its connectivity, weights of synaptic values parameters or the best input configuration for a given output. The power of genetic algorithm (GA) drives largely from the concept of "implicit parallelism", the simultaneous allocation of trials to many regions of the search space. For any selection algorithm, the allocation of trials to individuals induces a corresponding allocation of hyperplanes or substrings represented by individuals. The search is not directionless but makes use of the probabilistic generation of control parameters to direct the search. The main control parameters of a GA are: the population size, the selection mechanism, the crossover rate, the mutation rate and the number of generations allowed for the evolution of required structure. In its simplest form a GA is a cycle following the steps accordingly.

1. Construct randomly an initial population of chromosomes,
2. Calculate and evaluate each chromosome fitness,
3. If the result is satisfactory, stop, if not,
4. Select the best chromosomes for reproduction based on its fitness (higher),
5. Create new offspring by application of crossover and mutation operators,
6. Form a population for next generation,
7. If process has converged, return the best chromosome as a solution, otherwise go to step two.

RESULTS AND DISCUSSION

The Figure 5 demonstrates a confusion matrix of the errors per classes of classification rules model. The algorithm uses all data set within random partitions to train, test and validation. It shows an accuracy of 87,63% and a mean error rate of 12,37%. Numbers in diagonal are right classifications. In the other hand, Table 4 presents the results of the best three neural networks models for water masses classification (K-means classes). The genetic algorithm not only randomly shifts data set into train, test and validation but also optimizes the weight matrix of model and makes a feature selection. Because of the high dependence of this type of model to data, a 10 fold cross validation approach was performed. Many types of architecture were tested but it always results in Multilayer Perceptrons with four layers demonstrating complex patterns.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	← Classified as
172	0	19	0	1	0	1	(1): 1
5	8	5	0	2	0	0	(2): 2
19	2	156	0	1	0	0	(3): 3
1	0	1	12	0	0	0	(4): 4
2	0	1	0	33	0	0	(5): 5
0	0	0	0	0	15	0	(6): 6
1	0	0	0	0	0	29	(7): 7

Fig. 5. Error matrix of classification of seven water classes by rule based model.

Table 4. Errors of three best neural models.

Type of Architecture	Train	Test	Validation
MLP- 2/4/7/1	0.1025	0.1141	0.1066
MLP- 5/9/7/1	0.0555	0.0037	0.0070
MLP- 17/17/7/1	0.0192	0.0221	0.0553

Inputs vary from two to seventeen and all three models shows the second hidden layer with seven neurons going into a single output neuron, the class. Although one could say all models present good performance, the winner network is in the middle which have smallest validation (0,0070) error and smallest architecture (5/9/7/1) that means less computational cost. The five and most important selected input variables chosen by this model to reach the present results were Salinity, Mytilidae, Nitrite, Bivalve and LD. Some could interpret this set of variables as the salinity been the great important physical stressor that drives the larval recruitment at the studied site while nitrite suggest high bacterial activity. The three other variables are all mollusks suggesting the sensitivity of this group to the environmental stressors. It is clear that neural network models are much more precise than rule-based models but do not show any information or explicit knowledge about the structure of community as the rules model did. For this reason, this type of model was used to classify and predict the used indices. Table 5 presents the performance of Multilayer Perceptron neural network comparing the error (Root Mean Square Error - RMSE) of models applied to the classification task. Plix index presents excellent previsibility in classifying environment. It is probably due to its highest linearity than others. In the other hand, someone could use the model to assess how would be the future trophic status if a given event happens. For this intention, Table 6 shows an error (RMSE) comparison, and again Plix demonstrates better performance over Trix index. As visualization example, the interannual (10 years), monthly averaged, variation of Plix values and the output of prediction of neural network model can be viewed in Figure 6.

Table 5. Neural network indices error of classification.

Index	Train	Test	Validation	Correlation R^2
TRIX	0.6272	0.3457	0.5362	0.6818
PLIX	0.0027	0.0027	0.003	**0.9999**
MARGALEF	0.8369	0.1905	0.2651	0.9355

Table 6. Neural network prediction error of indices.

Index	Train	Test	Validation	Correlation R^2
TRIX	0.7114	0.3445	0.6001	0.3275
PLIX	11.711	0.8348	0.8167	**0.7407**
MARGALEF	0.8889	0.4177	0.7295	0.6239

Based on association rules model, Figure 7 presents, for example, three different population structures related to three different water mass conditions. In Figure 7a the coastal water is worm, present good oxygen level but is pour in chlorophyll and low nutrient despite this, demonstrate the biggest larvae concentration and the absence of cypris. Figure 7b, tropical waters of Brazil current, not so worm, got a bit more chlorophyll probably by little entrance of phosphorous but less larval concentration and the absence of bryozoans larvae, while Figure 7c represents the cold deep water from ACAS, very oxygenated, big entrance of nitrogen forms, low chlorophyll concentration and very few larvae. In spite of Figure 7.a and b are quite different (similar biodiversity and different trophic status), there is in Figure 7 no aparent food for larvae suggesting they are probably feeding on another energy resousce like organic matter. This situation can occour under special conditions of wind, absence of strong currents and low tide leaving the waters from a small river and mangrove in north come into the coastal area. In Figure 7b the climatic conditions have change and chlorophyll-a gave some response to the little entrance

of nitrogen but the phytoplankton composition should determine the structure of zooplankton community, an example of bottom-up control of food web. Figure 7c show a critical condition of the environment in which the upwelling event happens. Could deep water reach surface carrying high amounts of nitrate, some nitrite and phosphorous. It is a question of time to the chlorophyll response and bloom formation. Turbulence and mixing process can explain high values of oxigen.

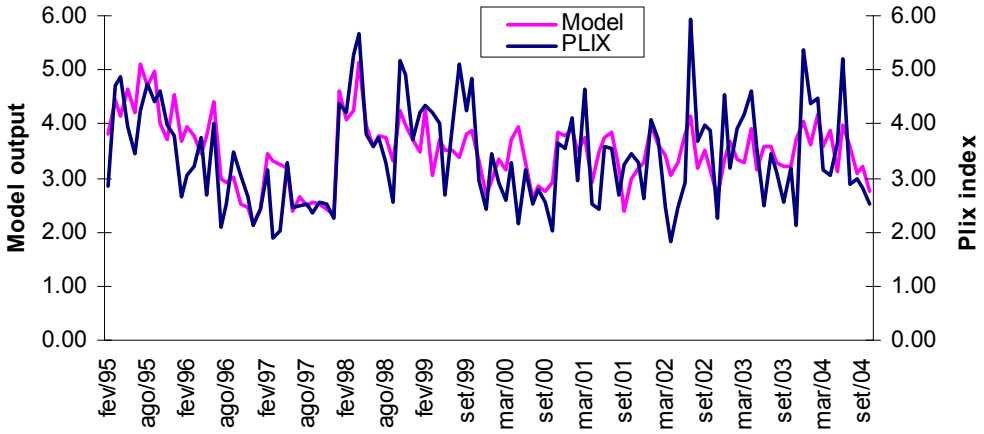

Fig. 6. Average interannual prediction of monthly values of Plix index.

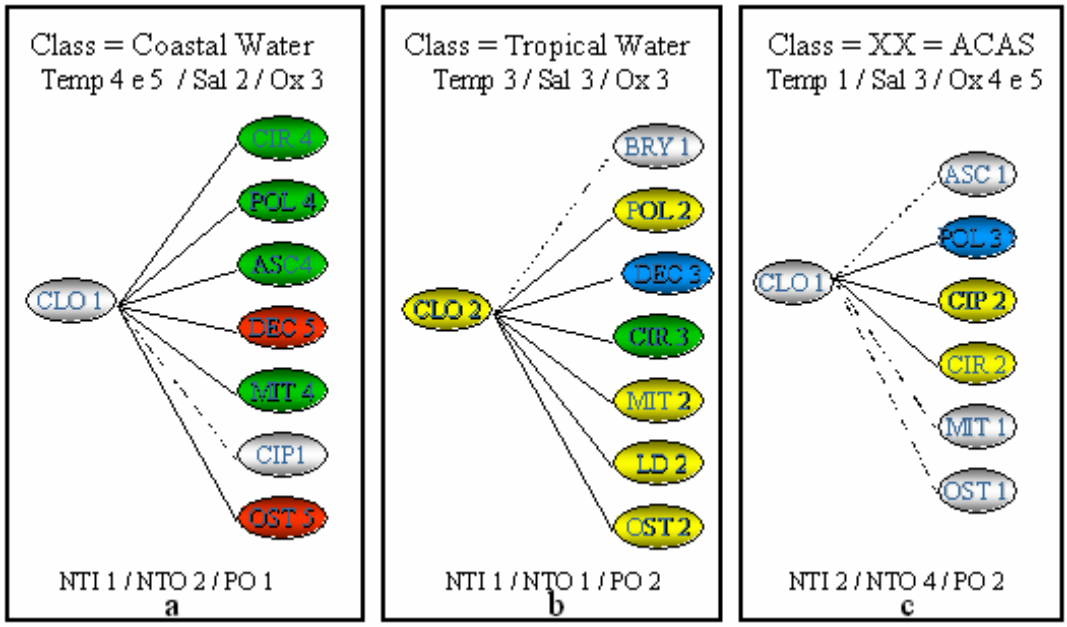

Fig. 7. Population structure of three different water conditions: in a, the coastal water mass; b, tropical water of Brazil Current; and c, upwelled water from ACAS.

CONCLUSIONS

This work tried to present a methodological development to make an ontology over the ecosystem and its biological integrity coming these concepts into management. We demonstrate that knowledge systems based on rules can be used in environmental programs and management strategies. It is clear from a management and conservation point of view, the choice of an indicator can not follow the easyest and more abundant groups, some times, rare groups are more sensibles. This is the case demonstrated by the set of rules 1-5 that under conditions of normal oxigen (rule 5) ascidians larvae are present within mean-low concentration, and in rule number 2 is absent. This kind of larva depicts a high sensitivity to physical parameters increasing its concentation according temperature and salinity The Apriori algorithm and the generated association rules, demonstrate specific assemblages of meroplankton larvae related to environmental conditions and can be used as inicialization method for ecological network, where the biomass of different population can stablish the neurons threshold. The structural analysis of the architecture arrangement will provide insigths over the ecosystem functioning. The knowledge regarding nutrient-phytoplankton-zooplankton relationship is not only important for understanding the dynamics of the system, but also in the development of models coastal zone management. In order to best evaluate the performance of proposed models a ten fold cross validation approach was adopted. Many research to build classifiers have focus on minimum erro but, in real world aplications, different kind of errors can have different cost. The precision estimation of models is important to evaluate its behavior when operating within unknown data. In this way, artificial neural networks provide a powerfull tool for classification and prediction of environmental conditions, they can learn patterns of interactions among independent variables whithout any a priori assumptions. The genetic algorithm evolved neural networks with recurrent archtectures showing two hidden layers demonstrating high complexity of patterns. Traditionally, the identification of organism in natural samples is a time consuming task resulting in high cost and usually does not represent the true of nature because of low sampling frequency. As future works we intend to use an “in situ” flow cytomety to rapidly enumeration and identification of organism to model the microbial food web and advance in algorithms for knowledge extraction of weighted matrix of the networks, once the structural complexity does not allow the direct analysis of the archtectures, at this moment.

ACKNOWLEDGEMENTS

The authors are grateful to Almirante Paulo Moreira Institute the of for data availability and to the Brazilian Research Agency (Capes) for the financial support.

REFERENCES

AGRAWAL, R.; SRIKANT, R.. **Fast algorithms for mining associations rules**. *VLDB-94,* 1994.

BLOCK, H.D. The perceptron: A model for brain fuctioning. **Revs. Mod. Phys**., v. 34, p. 123-135, 1962.

BRADSHAW, C.J.A.; PURVIS, M.; RAYCOV, R.; ZHOU, Q.; DAVIS, L.S. Predicting patterns in spatial ecology using neural networks: modelling colonization by New Zealand fur seal. In: DENZER, R., SWAYNE, D.A, PURVIS; M.; SCHIMAK, G. (Ed.). **Environmental Software Systems**. Environmental Information and Decision Support, n. 167. Dordrecht: Kluwer Academic Publishers, 2000. p. 57-65.

COWEN, R. K.; K. LWIZA, M. M.; SPONAUGLE, S.; PARIS, C.; OLSON. B. D. B. Connectivity of marine populations: open or close? **Science,** v. 287, p. 857-859, 2000.

COSTANZA, R.; D’ARGE, R. DE GROOT; FARBER, S.;GRASSO,M.; HANNON, B.; LIMBURG, K.; NAEEM, K.; O’NEILL,R.V.; PARUELO,J.; RASKIN, R.G.; SUTTON, P.; VAN DER BELT, M. The value of the world’s ecosystem services and natural capital. **Nature**, v. 387, p. 253-260, 1997.

CHRISTENSEN, J. Auditing conservation in an age of accountability. **Conserv. Practice**, v.4, p. 12-19, 2003.

EUROPEAN ENVIRONMENTAL AGENCY – EEA. ELUNIS – **European Nature Information System**. Electronic Source: http://eunis.eea.eu.int/index.jsp, Last web site update: 04.12.2003.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; PADHRAIC, S., UTHURUSAMY, R. (Ed.). **Advances in knoledge discovery and Data Mining**, Cambridge; London: MIT Press, 1996. 611 p.

GOLDBERG, D.E. **Genetic algorithm in search, optimization, and machine learning**. New York: Addison-Wesley, 1989.

HAN, J.; KAMBER, H. Data Mining – Concepts and Thechniques, Morgan Kaufmann Publishers, Academic Press, San diego, CA, USA, 550 p., 2001.

HIXON, M.A.; BOERSMA, P.; HUNTER, M. L.; MICHELI, FIORENZA, NORSE, E.; POSSINGHAM, H.P.; SNELGROVE, P.V.R. Oceans at Risk: Research priorities in marine conservation biology. In: SOULÉ, M.E.; G.H. ORIANS, G.H. (Ed.). **Conservation biology**: research priorities for the next decade. Washington, D.C.: Island Press; 2001. p. 125-154.

LEK, S.; DELACOSTE, M.; BARAN, P.; DINOPOULUS, I.; LAUGA, J.; AULAGNIER, S. Aplication of neural networks to nonlinear relationship in ecology, **Ecol. Model**., v. 90**,** p. 39-52, 1996

LIU, B.; HSU, W.; MA, Y. **Building an accurate classifier using association rules.** KDD-98. New York, p. 27-31, 1998.

MAIER, H. R.; DANDY, G. C. **The application of artificial neural networks to the prediction of salinity**. Adelaide: Department of Civil and Environmental Engineering, The University of Adelaide, 1993. 464 p. (Research Report No. R101).

MAIER, H. R.; GRAEME, C. D.; MICHAEL, D. B. Use of artificial neural networks for modelling cyanobacteria *Anabena* spp.in the river Murray, South Australia, *Ecol. Model*., v. 105, p. 257-272, 1998.

MARGALEF, R. Diversidad de especies en las comunidades naturales. **Publ. Inst. Biol. Apl.**, Barcelona, v. 9, p. 15-27, 1951.

MOORE, T.; MORRIS, K.; BLACKWELL, G.; GIBSON, S. Extraction of beach landforms from dems using a Coastal Management Expert System. Annual Conference of GeoComputation, 2., 1997, New Zealand. **Proc.,** 1997.

NELSON, S. M.; ROLINE, R. A. Effects of multiple stressors on hyporheic invertebrates in lotic system. **Ecol. Indicat**., v. 3, n. 2, p. 65-79, 2003.

OAKES, R. M.; GIDO, K. B.; FALKE, J. A.; BROCK, B. L. Modelling of stream fishes in the Great Plains, USA. **Ecology Freshwat. Fish** , v. 14, p. 361-374, 2005.

OLDEN, J. D.; POFF, N. L.; B. P. BLEDSOE, B. P. Incorporating ecological knowledge into ecoinformatics: an example of modeling hierarchically structured aquatic communities with neural network. **Ecol. Informatics** , v. 1, p. 33-42, 2006.

ÖZESMI, S. L.; ÖZESMI, U. An artificial neural network approach to spatial habitat modelling with interspecific interaction. **Ecol. Model**., v. 116, n. 1, p. 15-31, 1999.

PEREIRA, G.C. **Data Mining for environmental analisys and diagnostic**. Ph.D. dissertation, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. (In portuguese). 2005.

PEREIRA,G. C.; COUTINHO, R.; EBECKEN, N. F. F. 2002. **Biological response neural network prediction in coastal upwelling field**. Oil and Hydrocarbon Spill, 3., 2002. **Proc.**, 2002. p. 301-310.

RECKNAGEL, F.; FRENCH, M.; HARKONEN, P.; YABUNAKA, K. Artificial neural network approach for modelling and prediction of algal blooms. **Ecol. Model**., 96, pp. 11-28, 1997.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. "Learning representations by back-propagation error". **Nature**, v. 323, p. 533-536, 1986.

SCARDI, M. Artificial neural network as empirical models for estimating phytoplankton production. **Mar. Ecol. Prog. Ser**., v. 139, p. 289-299, 1996.

USEPA - U.S. Environmental Protection Agency. **Guidelines for ecological risk assesment.** Risk Assessment Forum. Washington, DC: Office of Research and Development. U. S. Environmental Protection Agency (EPA/630/R-95/002F), 1998.

VALANTIN, J. L. *A dinâmica do plâncton na ressurgência de Cabo Frio-RJ*. **Inst. Pesqui. Mar.**, Rio de Janeiro. Coletânea de trabalhos, In: F.P. Brandini (editor) Memórias de III EBP Curitiba, 1988.

VOLLENWEIDER, R. A.; GIOVANARDI, F. G.; MONTANARI, RI-NALDI. Characterization of the trophic conditions of marine coastal waters, with special reference to the NW adriatic Sea: proposal for a trophic scale, turbidity and generalized water quality index. **Environmentrics,** v. 9, p. 329-357, 1998.

(Manuscript received 09 June 2006; revised 16 April 2007; accepted 04 July 2007)