USE OF MACHINE LEARNING FOR DETERMINING PHYTOPLANKTON DYNAMIC ON STATION RV001 IN FRONT OF ROVINJ (NORTHERN ADRIATIC)

G. Volf[1*] – B. Kompare[2] – N. Ožanić[3]

[1]Department of Hydraulic Engineering, Faculty of Civil Engineering, University of Rijeka, R. Matejčić 3, 51000 Rijeka
[2]Department of Sanitary Engineering, Faculty of civil and Geodetic Engineering, University of Ljubljana, Jamova 2, 1000 Ljubljana
[3]Department of Hydraulic Engineering, Faculty of Civil Engineering, University of Rijeka, R. Matejčić 3, 51000 Rijeka

ARTICLE INFO

***Article history*:**
Received: 18.10.2013.
Received in revised form: 20.12.2013.
Accepted: 24.12.2013.

Keywords:
Machine learning
Phytoplankton
Northern Adriatic
Rovinj

Abstract:

The paper describes the use of machine learning (ML) and discusses various approaches in modeling phytoplankton based on data from station RV001 in front of Rovinj which well represents the main processes in the open northern Adriatic (NA). Station RV001 is an example of oligotrophic seawater in NA. In order to contribute to the understanding of phytoplankton dynamics at the observation station, based on data covering physical, biological and chemical parameters, ML techniques were used. The final result is a construction of models in the form of regression and model trees, respectively; there were models constructed to be used to explain the dynamics of phytoplankton concentrations at the mentioned station as a result of independent environmental variables. Models in an affordable way combine and show knowledge collected by measurements during 35 year period, which have greatly contributed to a better understanding of ecosystem functioning.

1 Introduction

Ecological systems rarely require only a simple statistical analysis, especially because collected data have unusual distribution, non-linearity, missing values, etc. Also, problems are very large databases which are very difficult to be handled. ML techniques are not always the solution to all the problems associated with data in ecology, but on the other hand, they can offer a significant set of tools that might be useful in solving the problems. ML presents classical statistical problems such as classification, regression, decision making, etc. What makes ML significant are its tools, techniques and strategies characterized by the use of various algorithms and computational resources used to handle large data sets, large number of variables and complex data structures. Today ML is widespread not only in various research fields but also in ecology.

The Adriatic Sea is subdivided into three regional pools (Northern, Central and Southern), which differ in bathymetry, physiography and biogeochemical characteristics. NA, Fig. 1, is the

[*] Corresponding author. Tel.:+385 51 265932
E-mail address: goranvolf@yahoo.com

shallowest, while his northwestern part is one of the most productive areas in the Adriatic, as well in the whole Mediterranean Sea [1-2]. Numerous rivers and streams discharge nutrient rich freshwaters into the NA shallow waters [3]. Semi-enclosed circulation of sea water body, characterized by cyclonic an anticyclonic atmospheric eddies prevails during spring and summer, significantly reducing the water exchange rate with the remainder of the Adriatic Sea [4-5].

To understand ecosystem functioning, it is of crucial importance to understand main biogeochemical and hydrological characteristics and processes affecting this ecosystem.

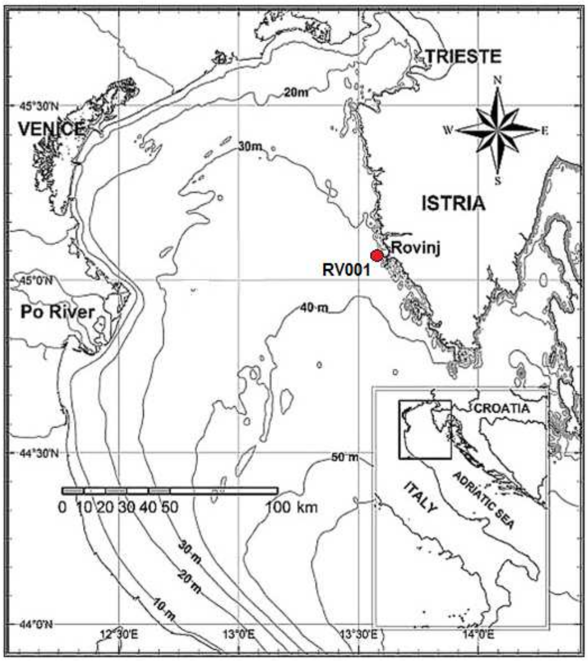

Figure 1. Northern Adriatic site with the displayed measuring stations RV001.

Many studies have been undertaken, resulting in a substantial amount of knowledge about the NA ecosystem and its productivity. Just a few decades ago, the NA was eutrophic for most of the time during the year, but environmental protection measures put in force since that time are now giving noticeable results. The latest study performed on long term data carried strong evidence that the still common perception of the NA as a very eutrophic basin is no longer appropriate, at least for its northern part and in recent years [2].

Phytoplankton plays a central role in the health and productivity of marine ecosystems, in addition to being considered as a sensitive indicator of speed and severity of global climate change [6]. Most commonly, data analyses have been performed with only classical and just recently with advanced statistical approaches such as principal component analysis (PCA) [7-9]. Although these techniques provide very useful insights into the data, they are sometimes limited in terms of interpretability due to their black-box nature. On the other hand, a branch of ML methods and tools have been proven to produce descriptive, e.g. transparent-box models, which generally allow for much easier interpretation [10-14].

The advantages of ML in the case of regression and model trees to build an understandable and interpretable description and prediction models of phytoplankton dynamics in the NA at station RV001 will be presented here. Also, knowledge collected by measurements during 35 year period will be shown using the models, contributing thus to a better understanding of ecosystem functioning.

2 Data description

The data set comprises physical, chemical, and biological parameters. Data were collected at station RV001 which is 1 Nm off Rovinj on the western Istrian coast by the Center for Marine Research (CMR) in Rovinj, Croatia, Fig. 1. The water column was sampled with 5 l Niskin samplers at 0.3 m, 5 m, 10 m, and 20 m, and at 2 m above the bottom from 1979 to 2007 with almost a monthly frequency. An analysis of pH was performed aboard the research vessel immediately after sample collection by Radiometer pH meters. Temperature was measured with reversing thermometers and salinity with Beckman RS 7c or Yeo-Kal MKII high precision salinometers in the ashore laboratory. The samples for total phytoplankton counts (micro and nano fractions) were preserved with lugol solution and counted according to Utermöhl [15] using Carl Zeiss inverted microscopes.

The Po River flow data measured daily at Pontelagoscuro, Fig. 1, from January 1966 to December 2007 were obtained from the Agenzia Regionale Prevenzione e Ambiente dell'Emilia Romagna, Servizio Idrometeorologico, Parma. The data used for building the phytoplankton models are shown in Table 1.

Table 1. Data used for phytoplankton models.

Symbol	Symbol interpretation	Unit
Month	Month of sampling	
Year	Year of sampling	
Flow	Po river flow	m^3/s
Temp	Sea temperature	°C
Sal	Salinity	
Dene	Density	kg/m^3
pH	pH	
Phyto	Total phytoplankton	cell/L
Phyto_pred	Total phytoplankton shifted for one month	cell/L

The data were pre-processed with regard to modelling and research goals. For the phytoplankton models, the entire span of the historic data was used. At each station, the measured parameters for the top 11 m of the water column, e.g. above the thermocline were averaged and taken as one layer (related to eutrophication).

3 Modeling with machine learning tools

The main task of ML is to learn a *concept* from given *examples*. The entire procedure consists of a concept, examples (measurements), *learning algorithm* and *learning scheme* or model, Fig. 2. Each example consists of attribute and class values. The attributes are descriptors of the class, e.g. independent variables, while the class represents the dependent variable. The learning algorithm then, from the examples and some background knowledge, generates the learning scheme (model), which is a presentation of what has been learned, e.g. the class values are presented in terms of the attribute values. Based on the class value type (e.g. numeric, nominal, discrete, continuous…), the learning scheme (model) can be a decision tree, regression tree, classification rules, decision tables, and so on. Different ML algorithms allow for different levels of expert knowledge introduction in the learning procedure.

Typically, the quality of data-driven models is dependent on the examples (data) quality and quantity. To learn a concept successfully, a sufficient number of representative examples is needed.

Regression trees are hierarchical structures composed of nodes and branches where the internal nodes contain tests on the input attributes. Each branch of an internal test corresponds to an outcome of the test, and the predictions for the values of the target attribute (class) are stored in the leaves, which are the terminal nodes in the tree. If we have a single value for the class prediction, we deal with simple regression trees, while if linear equation is used for prediction in the leaf we refer to (regression) model trees.

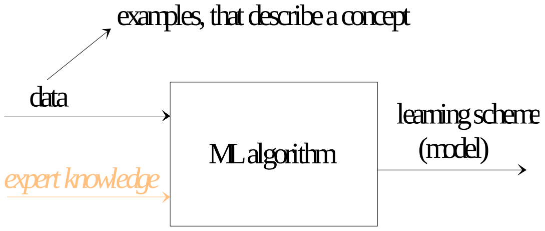

Figure 2. Machine learning procedure.

4 Results and discussion

4.1 Description of the experiments

For the experiments, the ML algorithm M5P for model and regression trees integrated in the WEKA modelling software was used. For the first model, total phytoplankton was set as a target (dependant) variable, whereas date (year and month of measurement), the Po River flow, temperature, salinity, density and pH (see Table 1) were used as independent variables (descriptors) to make a phytoplankton model. For the second model where we predict phytoplankton concentrations one month ahead, total phytoplankton shifted for one month e.g. one measurement (Phyto_Pred) was set as a target (dependant) variable, whereas the Po River flow, month, temperature, salinity, density and pH were independent variables (descriptors). The above parameters were mainly used because they provide the best representation of the parts of the ecosystem on top of which the target variable relays. The Po River flow rates were used as a rough measure of the eutrophication pressure acting on the investigated ecosystem combined with nutrient concentrations in the sea as a measure of the eutrophication degree.

While comparing data to previous research [14], some parameters for building models were omitted because mainly they did not have big influence on building models in previous research (like nutrients and their ratios). Here the goal is to get simple and yet efficient models only from measured data.

4.2 Phytoplankton models

Using the data from station RV001, Table 1, Fig. 1, two types of phytoplankton models were constructed (see Fig. 4 and Fig. 5). The main processes in the open NA waters are well represented by Station RV001 mainly because the local ones are not pronounced at all [16]. The accuracy of the models is expressed by the correlation coefficient (R) between modeled and measured concentrations of phytoplankton obtained by testing data using 10-folds cross-validation for the test option. The correlation coefficient for the first model built by regression trees, Fig. 3, amounts to 0.7, and for the second model built by model trees to 0.82, Fig. 4.

Taking into account the complexity of tackled domain and the lack of adequate data, it is evident that the correlation coefficients are quite high and obtained results very satisfactory.

The goal of the first model built by regression trees is to explain how the phytoplankton concentration was changing at station RV001 and to identify the most influential factors of this dynamics. The model was constructed using 5 of 7 variables (factors) that describe the concentration of phytoplankton in the observed period (1979-2007). These variables are year, month, salinity, the Po River flow and temperature. The model consists of 14 leaves in which average values of phytoplankton concentrations are situated. These values point to certain changes in the dynamics of phytoplankton concentration.

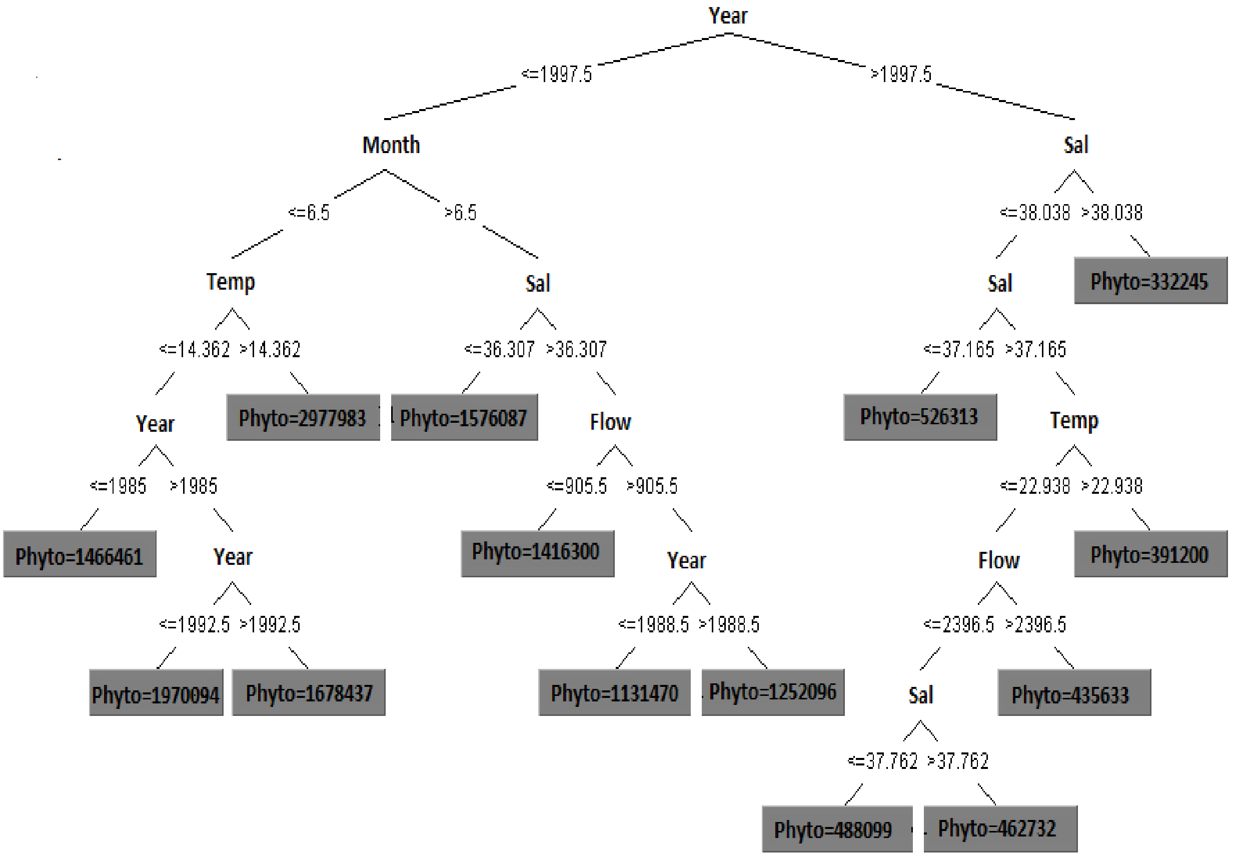

Figure 3. Phytoplankton model obtained using regression trees.

The built model confirms some of the conclusions from previous studies of the phytoplankton dynamics in the NA, and gives an easy-to-read structured knowledge representation. The model indicates that during 1998 an important change in the phytoplankton dynamics occurred at the observation station [14]. The phytoplankton concentrations before 1998 were significantly higher than after that year. After 1998 salinity is the main signal indicating changes of the impact of freshwater inputs to the area, but also of the inflow of more saline waters from the central Adriatic. A reduction of riverine nutrients input and extended saline waters intrusion contributed to lower phytoplankton concentrations after 1998, most often throughout the investigated area of the NA. It can be seen that month (July) is the main indicator of changes in the phytoplankton concentration in the period before 1998, which suggests two other important indicators, salinity and temperature.

Before 1998 the model indicates three significant changes: in 1985, 1989 and in 1993. The changes in 1993 coincide with unusually high freshwater discharges in the NA in autumn [17]. In October 1993, the Po River flow rates were markedly higher than any monthly averages for all months of the year since 1917 when the measurements started (CMR, internal database).

The model also points out to changes in the phytoplankton concentration in 1989 that are related to the Po River flow. Changes in 1989 coincide with mucilage event when winds blowing from the sea to the land dominated and the coast was contaminated by gelatinous material for weeks. Also, at the end of the 1980s and early 1990s changes in the entire copepod community were observed in the Gulf of Trieste, which can be associated with the change of Northern Ionian Gyre (NIG) circulation in 1987 [18]. The same authors related other changes in the abundance of some species from the late 1990s to the early 2000s to the reversal of the NIG circulation in 1997. Changes in 1985 are most likely related to the reduction of polyphosphate contents in detergents, with a consequent marked decrease of phosphorus compound in river waters [19].

When comparing this model and model from previous research [14] it can be seen that the correlation coefficient here is higher for smaller tree although smaller set of data for building the model were used. However, it must be noticed that in [14] the model was chosen for the whole NA and tested at all observation stations. Pointing to relevant changes which happened in observed ecosystem, models are also quite similar. Both models recorded changes in 1998, 1993 and in middle 1980s. The difference is in changes of phytoplankton in 1989 and in 2000, which are quite specific [14]. In this model temperature also appears as an indicator of trophic changes in the ecosystem.

The second model for predicting phytoplankton concentrations one month ahead is presented in Fig. 4. The model is built using model trees which instead of a single target variable in their leaves contain an equation for the description of that variable (see Table 2). Although the model is small and simple, it has an acceptable correlation coefficient of 0.82 obtained by testing data using 10-folds cross-validation for test option. The model was created to demonstrate the use of model trees in predicting the phytoplankton concentrations. Fig. 5 presents a good prediction of the peak values of phytoplankton concentrations.

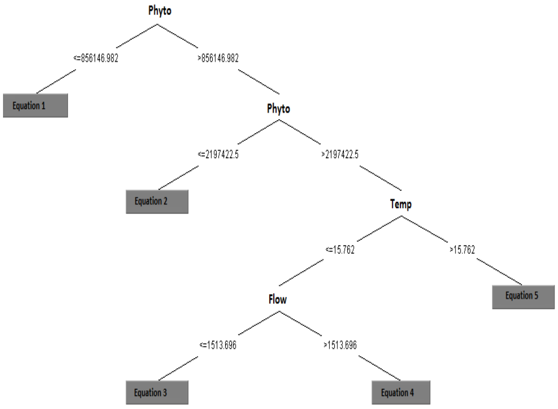

Figure 4. Phytoplankton model obtained using model trees.

To test model capabilities, other test options were also verified where the lowest, but still a good correlation coefficient of 0.57 was obtained to supply test set option. For this test option, the model was trained on data from 1979 to 2005, and then tested on data from 2006 and 2007. Looking back to previous research [14], it can be seen that the use of interpolated data gives better models and higher correlation coefficients. Theoretically, more data must result in better models. But it must be mentioned that the goal here was to get a simple and yet efficient model for predictions generated only from measured data.

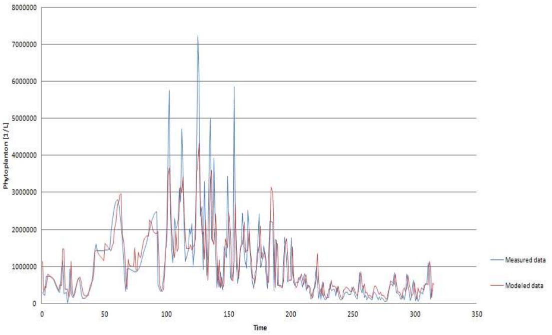

Figure 5. Comparison of measured and modeled values of phytoplankton.

Table 2. Equations for model presented in Fig. 4.

No.	Equation
1.	Phyto_pred= -18245.461*Month-14431.3336*Temp+45291.7626*SAL-57854.5192*Dene+0.7582*Phyto+405925.6985
2.	Phyto_pred= -42548.0029*Month-149115.9989*Temp+584937.4679*SAL-662894.1897*Dene+0.4208*Phyto-132054.7933
3.	Phyto_pred= 14639.1135*Month+223.2072*Flow-109523.9196*Temp+246004.9391*SAL-271526.5835*Dene+0.2132*Phyto+2049152.0359
4.	Phyto_pred= 14639.1135*Month+57.7314*Flow-109523.9196*Temp+246004.9391*SAL-271526.5835*Dene+0.2132*Phyto+2738296.8914
5.	Phyto_pred= -14639.1135*Month-104897.5165*Temp+246004.9391*SAL-271526.5835*Dene+0.2132*Phyto+1759245.6482

5 Conclusions

Regression trees are useful for describing a given ecosystem, while model trees can be used for predictions. The model derived by using regression trees successfully identifies some of the already known changes in the observed ecosystem and gives an easy-to-read structured knowledge representation. As indicated previously, the salinity and temperature appear to be the most important indicators of trophic changes in the observed ecosystem.

The second model for predicting phytoplankton concentrations although being simple achieved a satisfactory correlation coefficient and identified peak values successfully. This kind of model can be a highly useful water management tool as a self standing model predicting the phytoplankton concentrations or integrated in more complex watershed models which include nutrient generation watershed activities. Such an integrated model can be used for controlling the nutrient loadings from the watershed. Model trees in theory give higher correlation coefficients than regression trees, mainly because target variable in model trees is described by the equation, while in regression tree we have only the mean value of target variable. The use of ML methods, in this example, regression and model trees are very useful and give a different view on the data obtained from other analyses.

Acknowledgements

The authors thank all who contributed to the data collection, analysis and interpretation, with the special thanks to the crew of RV “Vila Velebita” for help during sampling. We are grateful to Dr. R. Precali and Dr. N. Atanasova for valuable comments during the review of the manuscript. This work is a part of the scientific projects: “Mechanism of long-term changes in the northern Adriatic ecosystem” (098-0982705-2731), „Hydrology of sensitive water resources in karst“ (114-0982709-2549) and Project “Jadran” funded by the Ministry of Science, Education and Sports of the Republic of Croatia. Authors also greatly appreciate the comments of two anonymous reviewers who helped us to substantially improve the paper.

References

[1] Sournia, A.: *La production primaire planctonique en Méditerranée*, Bull. Ėtude Commun. Medit., 5 (1973), 128.

[2] Mozetič, P., Solidoro, C., Cossarini, G., Socal, G., Precali, R., Francé, J., Bianchi, F., De Vittor, C., Smodlaka, N., Fonda Umani, S.: *Recent Trends Towards Oligotrophication of the Northern Adriatic: Evidence from Chlorophyll a Time Series.* Estuaries and Coasts, 33 (2009), 362-375.

[3] Raicich, F.: *On the fresh water balance of the Adriatic Sea*, Journal of Marine Systems, 9 (1996), 305-319.

[4] Supić, N., Orlić, M., Degobbis, D.: *The Istrian countercurrent and its year to year variability*. Estuarine, Coastal and Shelf Science, 51 (2000), 385-397.

[5] Grilli, F., Marini, M., Degobbis, D., Ferrari, C. R., Fornasiero, P., Russo, A., Gismondi, M., Djakovac, T., Precali, R., Simonetti, R.: *Circulation and horizontal fluxes in the*

northern Adriatic Sea in the period June 1999-July 2002. Part II: Nutrients transport. Science of the Total Environment, 353 (2005), 115-125.

[6] Widdicombe, C.E., Eloire, D., Harbour, D., Harris, R.P., Somerfield, P.J.: *Longterm phytoplankton community dynamics in the western English Channel*, Journal of Plankton Research, 32(2010), 643-655.

[7] Bernardi Aubry, F., Berton, A., Bastianini, M., Socal, G., Acri, F.: *Phytoplankton succession in a coastal area of the NW Adriatic, over a 10-year sampling period (1990-1999)*. Continental Shelf Research, 24 (2004), 97-115.

[8] Tedesco, L., Socal, G., Bianchi, F., Acri, F., Veneri, D., Vichi, M.: *NW Adriatic Sea biogeochemical variability in the last 20 years (1986-2005)*, Biogeosciences, 4 (2007), 673-687.

[9] Marić, D., Kraus, R., Godrijan, J., Supić, N., Djakovac, T., Precali, R.: *Phytoplankton response to climatic and antropogenic influences in the nort-eastern Adriatic during last four decades*, Estuarine, Coastal and Shelf Science, 115 (2012), 98-112.

[10] Kompare, B.: *The Use of Artificial Intelligence in Ecological Modelling*, Ph.D Thesis, FGG, Ljubljana; Royal Danish School of Pharmacy, Copenhagen, 1995.

[11] Kompare, B., Todorovski, L., Ďzerovski, S.: *Modelling and prediction of phytoplankton growth with equation discovery: case study-Lake Glumsø, Denmark*. Verh. Int. Verein. Limnol., 27 (2001), 3626-3631.

[12] Atanasova, N., Gal, G., Kompare, B.: *Modelling Dinoflagellate Dynamics in Lake Kinneret*. Verh.-Int. Ver. Theor. Angew. Limnol., 31 (2008), 100-104.

[13] Džeroski, S.: *Machine learning applications in habitat suitability modelling*. Artificial Intelligence Methods in the Environmental Sciences, 2 (2009), 397-411.

[14] Volf, G., Atanasova, N., Kompare, B., Precali, R., Ožanić, N.: *Description and prediction models of phytoplankton in thenorthern Adriatic*. Ecological modelling, 222 (2011), 2501-2511.

[15] Utermöhl, H.: *Zur Verfollkommnung der quantitative Phytoplankton-Methodik*. Mitt. int. Ver. theor. angev. Limnol., 17 (1958), 47-71.

[16] Precali, R., and Djakovac, T.: *Towards oligotrophication of the northern Adriatic: A reality?* ASLO Aquatic Sciences Meeting, Nica, 25-30 Jan 2009, 211 (2009).

[17] Supić, N., Đakovac, T., Krajcar, V., Kuzmić, M., Precali, R.: *Effects of excessive Po River discharges in the northern Adriatic*. Fresenius Envir. Bull., 15, 3(2006), 193-199.

[18] Conversi, A., Peluso, T., Fonda Umani. S.: *The Gulf of Trieste: A changing ecosystem,* Journal of Geophysical Research, 114 (2009).

[19] Pagnotta, R., Caggiati, G., Piazza, D., Ferrari, F.: *Il controllo della qualita` delle acque superficiali del bacino padano*. Inquinamento, 4 (1995), 8-14.