An Empirical Study of Sea Water Quality Prediction

Evaggelos V. Hatzikos[a,1] Grigorios Tsoumakas[b,*]
George Tzanis[b] Nick Bassiliades[b] Ioannis Vlahavas[b]

[a] *Department of Automation,*
Technological Educational Institute of Thessaloniki
P.O. BOX 141, 57400 Thessaloniki, Greece

[b] *Department of Informatics,*
Aristotle University of Thessaloniki,
54124 Thessaloniki, Greece

Abstract

This paper studies the problem of predicting future values for a number of water quality variables, based on measurements from under-water sensors. It performs both exploratory and automatic analysis of the collected data with a variety of linear and nonlinear modeling methods. The paper investigates issues, such as the ability to predict future values for a varying number of days ahead and the effect of including values from a varying number of past days. Experimental results provide interesting insights on the predictability of the target variables and the performance of the different learning algorithms.

Key words: Water Quality, Time Series, Prediction, Regression, Sensor Network

[*] Corresponding author
Email addresses: hatzikos@teithe.gr (Evaggelos V. Hatzikos), greg@csd.auth.gr (Grigorios Tsoumakas), gtzanis@csd.auth.gr (George Tzanis), nbassili@csd.auth.gr (Nick Bassiliades), vlahavas@csd.auth.gr (Ioannis Vlahavas).

[1] This work was supported by the ARCHIMEDES programme which is jointly funded by the Greek Ministry of Education (EPEAEK) and the European Union.

Preprint submitted to Elsevier Science *7 March 2007*

1 Introduction

The ability to predict one or more days ahead the quality of water in an ecosystem is a very important issue. Given such a predictive model, authorities will be able to foresee an increase of the pollution levels in the sea water and therefore instruct all the necessary precaution measures. Water quality prediction could also offer significant added value to several commercial applications, such as irrigation and piscicultures.

This paper is concerned with the prediction of future values for a number of water quality variables, based on data collected by an under-water measurement system. A set of sensors is used to measure the sea-water temperature, pH, conductivity, salinity, amount of dissolved oxygen and turbidity. The measured data are transmitted to a central monitoring station for analysis.

The paper investigates several aspects of the above problem using both exploratory and automatic analysis approaches. Initially, it studies the correlations and interactions between the different variables in search for evidence of an underlying mechanism governing the data. It then compares several linear and nonlinear modeling algorithms against the random walk model, which serves as a benchmark model in time series analysis tasks. In addition, the paper studies the ability to predict future values for a varying number of days ahead and the effect of including values from a varying number of past days.

The rest of the paper is organized as follows. In the next section we present background information on water quality variables, while in Section 3 we discuss the related work in the area of water quality prediction. Section 4 describes the data collection and pre-processing phases. In Section 5 our experimental setup is presented, including the design space, the algorithms and the evaluation method we have used. Section 6 contains the results of our experiments and discussion about these results. Finally, in Section 7 we present our conclusions and some directions for future research.

2 Background

There is a number of variables that indicate the quality of water. Some of the basic variables are *water temperature*, *pH*, *specific conductance*, *turbidity*, *dissolved oxygen*, *salinity*, *hardness*, and *suspended sediment.*

The temperature of water plays an important role in both environmental and industrial processes. Firstly, it affects the ability of living organisms to resist certain pollutants. Some organisms cannot survive when the water temperature takes a value beyond a specific range. The ability of water to hold oxygen is also affected by water temperature. Finally, low-temperature water is used for cooling purposes in power plants.

pH is a measure of the relative amount of free hydrogen and hydroxyl ions in the water. Water that has more free hydrogen ions is acidic, whereas water that has more free hydroxyl ions is basic. The values of pH range from 0 to 14 (this is a logarithmic scale), with 7 indicating neutral. Values less than 7 indicate acidity, whereas values greater than 7 indicate a base. The presence of chemicals in the water, affects its pH, which in turn can harm the animals and plants that live there. For example, an even mildly acidulous seawater environment can harm shell cultivation [2] . This renders pH an important water quality indicator.

Specific conductance is a measure of the ability of water to conduct an electrical current. It is highly dependent on the amount of dissolved solids (such as salt) in the water. Pure water, such as distilled water, has very low specific conductance, while sea water has high specific conductance. Specific conductance is an important water quality measure because it gives a good indication of the amount of dissolved material in the water.

Turbidity is the amount of particulate matter that is suspended in water. Turbidity measures the scattering effect that suspended solids have on light: the higher the intensity of scattered light, the higher the turbidity. Materials that cause water to be turbid include clay, silt, finely divided organic and inorganic matter, soluble colored organic compounds, plankton, microscopic organisms and others.

Each molecule of water contains an atom of oxygen. Yet, only a small amount of these oxygen atoms, up to about ten oxygen molecules per million of water molecules, is actually dissolved in the water. This dissolved oxygen is breathed by fish and zooplankton and is necessary for their survival. Rapidly moving water, such as in a mountain streams or large rivers, tends to contain a lot of dissolved oxygen, while stagnant water contains little. Bacteria in water can consume oxygen as organic matter decays. Thus, excess organic material in lakes and rivers can cause an oxygen-deficient situation to occur. Aquatic life can have a hard time in stagnant water that has a lot of rotting, organic material in it, especially in the summer, when dissolved-oxygen levels are at a seasonal low.

Salinity is the saltiness or dissolved salt content of a body of water. The salt content of most natural lakes, rivers, and streams is so small that these waters are termed fresh or even sweet water. The actual amount of salt in fresh

[2] Region of Central Macedonia, Directorate of Environment, Environmental Legislation, May 1999.

water is, by definition, less than 0.05%. The water is regarded as brackish, or defined as saline if it contains 3 to 5% salt. The ocean is naturally saline and contains approximately 3.5% salt. Some inland salt lakes or seas are even saltier. The Dead Sea, for example, has a surface water salt content of around 15%. Excessive salinity can be dangerous for shell cultivation, an economic activity that is very important for some regions.

The amount of dissolved calcium and magnesium in water determines its hardness. In areas with relatively hard water, someone may notice that it is difficult to get a lather up when washing his/her hands or clothes. Industries operating in such areas have to spend money in order to soften the water and avoid the damaging of equipment. Hard water can even shorten the life of fabrics and clothes.

Suspended sediment is the amount of soil moving along within a water stream. It is highly dependent on the speed of the water flow, as fast-flowing water can pick up and suspend more soil than calm water. If land is disturbed along a stream and no protection measures are taken, then excess sediment can harm the water quality of a stream.

3 Related Work

Reckhow (1999) studied Bayesian probability network models for guiding decision making for water quality of Neuse River in North Carolina. The author focuses both on the accuracy of the model and the correct characterization of the processes, although these two features are usually in conflict with each other.

Blockeel et al. (1999) studied two problems. The first one concerned the simultaneous prediction of multiple physico-chemical properties of river water from its current biological properties using a single decision tree. This approach is opposed to learning a different tree for each different property and is called predictive clustering. The second problem concerned the prediction of past physico-chemical properties of the water from its current biological properties. The Inductive Logic Programming system TILDE Blockeel and De Raedt (1998) was used for dealing with the above problems.

Dzeroski et al. (2000) addressed the problem of inferring chemical parameters of river water quality from biological ones, an important task for enabling selective chemical monitoring of river water quality. They used regression trees with biological and chemical data for predicting water quality of Slovenian rivers.

Lehmann and Rode (2001) investigated the changes in metabolism and water quality in the Elbe river at Magdeburg in Germany since the German reunification in 1990. They used weekly data samples collected between the years 1984 and 1996. They used univariate time series models such as autoregressive component models and ARIMA models that revealed the improvement of water quality due to the reduction of waste water emissions since 1990. These models were used to determine the long-term and seasonal behaviour of important water quality parameters.

Romero and Shan (2005) developed a neural network based software tool for prediction of the canal water discharge temperature at a coal-fired power plant. The variables considered in this system involve plant operating parameters and local weather conditions, including tide information. The system helps for the optimization of load generation among power plant generation units according to an environmentally regulated canal water discharge temperature limit of 95 Fahrenheit degrees.

Chau (2005) presented the application of a split-step particle swarm optimization (PSO) model for training perceptrons in order to predict real-time algal bloom dynamics in Tolo Harbour of Hong Kong. Experiments with different lead times and input variables have been conducted and the results have shown that the split-step PSO-based perceptron outperforms other commonly used optimization techniques in algal bloom prediction, in terms of convergence and accuracy.

The case-based reasoning system, presented in (Fdez-Riverola and Corchado, 2003, 2004), copes with water pollution. It specializes in forecasting the red tide phenomenon in a complex and dynamic environment in an unsupervised way. Red tides are the name for the sea water discolorations caused by dense concentrations of microscopic sea plants, known as phytoplankton. The system is an autonomous Case-Based Reasoning (CBR) hybrid system that embeds various artificial intelligence tools, such as case-based reasoning, neural networks and fuzzy logic in order to achieve real time forecasting. It predicts the occurrence of red tides caused by the pseudo-nitzschia spp diatom dinoflagellate near the North West coast of the Iberian Peninsula. Its goal is to predict the pseudo-nitzschia spp concentration (cells/liter) one week in advance, based on the recorded measurements over the past two weeks. The developed prototype is able to produce a forecast with an acceptable degree of accuracy. The results obtained may be extrapolated to provide forecasts further ahead using the same technique, and it is believed that successful results may be obtained. However, the further ahead the forecast is made, the less accurate it may be.

Hatzikos et al. (2005) utilized neural networks with active neurons as the modeling tool for the prediction of sea water quality. The proposed approach was concerned with predicting whether the value of each variable will move upwards or downwards in the following day. Experiments were focused on four quality indicators, namely water temperature, pH, amount of dissolved oxygen and turbidity.

4 Data Collection, Pre-Processing and Exploratory Analysis

This section describes the system that collected the data used in our study, the pre-processing approach that we followed and initial exploratory data analysis.

4.1 The Andromeda analyzer

The data used in this study have been produced by the Andromeda Analyzer (Hatzikos, 1998; Hatzikos, 2002). The system is installed in Thermaikos Gulf of Thessaloniki, Greece and consists of three local measurement stations and one central data collection station.

The local measurement stations (see Figure 1) are situated in the sea and serve the purpose of data collection. Each of them consists of the following parts:

- A buoy.
- A number of sensors.
- A reprogrammable logic circuit.
- Strong radio modems.
- A tower of 6 meters height for the placement of an aerial.
- Solar collectors interconnected for more power.
- Rechargeable batteries.

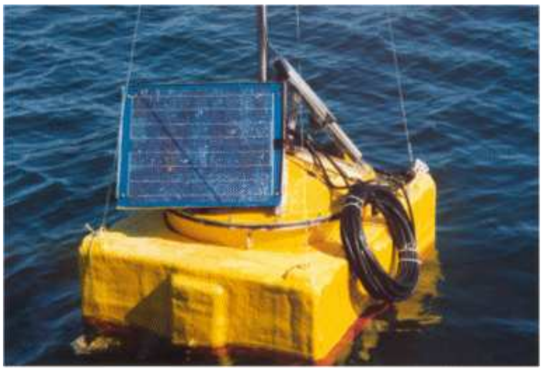


Fig. 1. One of the three local measurement stations of the Andromeda system.

The solar collectors and the batteries provide the electrical power needed by the sensors and electronics. The sensors measure water temperature, pH, conductivity, salinity, amount of dissolved oxygen and turbidity in sea-water at fixed time points. The reprogrammable logic circuit monitors the function of the local measurement station and stores the measurements in its memory. Moreover, it controls the communication via the wireless network and sends the measurements to the central data collection station.

The central data collection station monitors the communication with the local measurement stations and collects data from all of them. Data are stored in a database for the purpose of future processing and analysis. It consists of a Pentium computer operating in SCADA environment. The computer plays the role of *master* and controls the communication with the local measurement stations using the *hand-shake* protocol. The total number of measurements that are collected is between 8 and 24 daily. The frequency of measurements can be increased in case of emergency. This communication policy reduces the consumption of energy by the local stations, since they operate only when they have to send data to the central station.

Furthermore, the central station hosts an intelligent alerting system (Hatzikos et al., 2007) that monitors sensor data and reasons about the current level of water suitability for various aquatic uses, such as swimming and piscicultures. The aim of this intelligent alerting system is to help the authorities in the ”decision-making” process in the battle against the pollution of the aquatic environment, which is very vital for the public health and the economy of Northern Greece. The expert system determines, using fuzzy logic, when certain environmental parameters exceed certain ”pollution” limits, which are specified either by the authorities or by environmental scientists, and flags out appropriate alerts.

4.2 Data Preprocessing

The data that are studied in this paper were collected from April 14, 2003 until June 11, 2003 at an hourly basis with a sampling interval of 9 seconds. Given that the variation of the measurements from one hour to the next is typically very small, we decided to work on the coarser time scale of 24 hours, by averaging the measurements over days.

Two problems introduced in the data by the collection process are the following: a) there is a number of missing values due to temporary inefficiency of the sensors as well as problems in the transmission of the data, and b) the occurrence of special events near the local measurement stations, such as the crossing of a boat, have led to the recording of some outliers.

Fortunately, both of these temporary problems are automatically solved through the daily averaging process. During a day, the missing values are typically from

0 to 3, so the rest of the measurements can reliably give a mean estimate for the day. In addition, averaging ameliorates the effect of outliers. Specifically we calculate the median of all daily measurements, which trims away extreme values.

Based on the above remarks, the communication policy of the data collection system could be altered, in order to save energy if such a system was deployed in an ecosystem with limited sunlight. Instead of transmitting the data every hour, the local stations could transmit the average of their hourly measurements every k hours. For higher energy efficiency, the sensors themselves could operate every k hours and send their unique measurement. However, such a policy is less resilient to transmission failures and outliers. A different, adaptive, policy would let the local stations transmit their hourly measurements, only when the difference of at least one of the measurements with the previously transmitted corresponding measurement exceeds a predefined threshold. This would allow to save energy when hour to hour differences are negligible. The central station, assumes that the measurement values are the same if no values are transmitted.

4.3 Exploratory Analysis

We perform an initial exploratory analysis in order to have a first look at the data and assess the ability to make n-day ahead predictions. Figure 2 shows a plot of the values of the 6 variables over time, while Table 1 shows the correlation coefficient for each pair of variables.

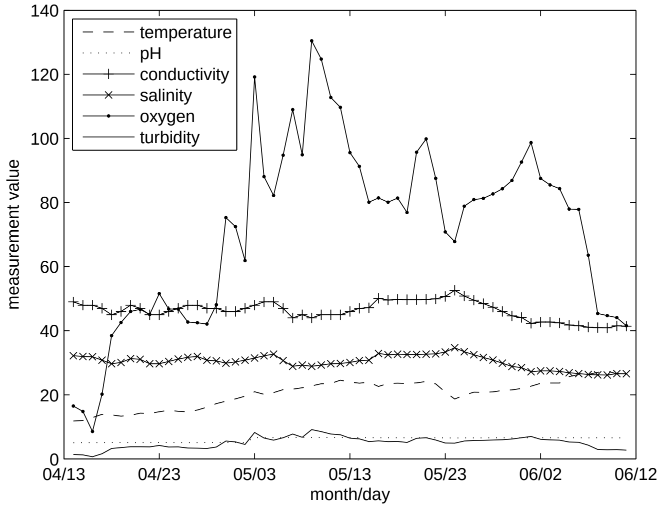


Fig. 2. Mean daily values of the 6 measuerements for the time period April 14 to June 11

	temperature	pH	conductivity	salinity	oxygen	turbidity
temperature	1.0000	0.8784	-0.3902	-0.4425	0.6191	0.5328
pH	0.8784	1.0000	-0.1287	-0.1803	0.7471	0.6796
conductivity	-0.3902	-0.1287	1.0000	0.9941	-0.0213	-0.0023
salinity	-0.4425	-0.1803	0.9941	1.0000	-0.0456	-0.0212
oxygen	0.6191	0.7471	-0.0213	-0.0456	1.0000	0.9921
turbidity	0.5328	0.6796	-0.0023	-0.0212	0.9921	1.0000

Table 1
Correlation coefficients between the 6 variables

We notice that, as expected, there are correlations between the different variables. We know that changes in water temperature and clarity affects the amount of oxygen in water. In addition, water of low clarity could contain organisms, which can affect the acidity of water. Finally, the amount of salt in the water directly influences its ability to conduct electricity. The strongest correlations are that between salinity and conductivity, and between turbidity and oxygen.

The correlation of pairs of variables shows us the relation of the measurements at the same time points t and hence it is not particularly useful for assessing the relation of past values of the variables with current values. Such information can be obtained by examining the autocorrelation and cross-correlation function for all variables and pairs of variables respectively. However, strong correlation can help us in designing a power-efficient sensor platform. For example, since salinity and conductivity are so strongly correlated, only one of the sensors can operate, while the value of the other is calculated based on a linear function of the operating sensor's measurement.

Figure 3 presents plots of the autocorrelation function for the 6 variables. The bar graph depicts the autocorrelation coefficient (y-axis) over the lag number (x-axis), while the two horizontal lines correspond to the upper and lower confidence limits. The plots demonstrate that past values of each variable can assist in the prediction of future values.

Figure 4 presents plots of the autocorrelation function for temperature, pH, conductivity and dissolved oxygen. For simplicity of presentation, we do not show plots of the autocorrelation function for pairs of variables including salinity and turbidity. As seen in Table 1 these two variables are highly correlated with conductivity and dissolved oxygen respectively, so the results of the cross-correlation function are very similar.

The stem graph depicts the cross-correlation coefficient (y-axis) over the lag number (x-axis), while the two horizontal lines correspond to the upper and

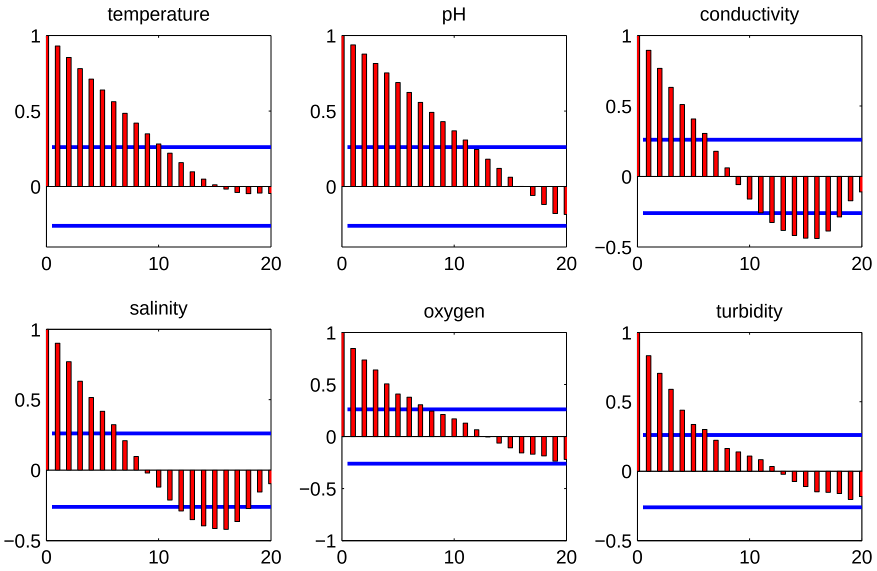


Fig. 3. Autocorrelation function for the 6 variables

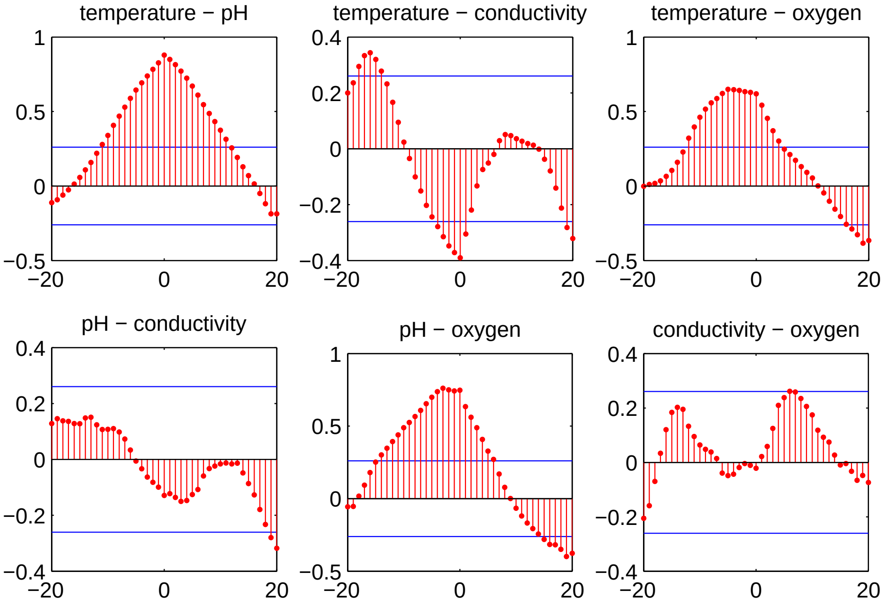


Fig. 4. Cross-correlation function for representative pairs of variables

lower confidence limits. The plots demonstrate that previous values of variables can assist in the prediction of future values for other variables, apart from the pairs (pH, conductivity), (pH, salinity), (conductivity, oxygen), (salinity, oxygen), (conductivity, turbidity), (salinity, turbidity). As expected, temperature is the most influential variable, affecting the future values of all the rest of the variables.

5 Experimental Setup for Automatic Analysis

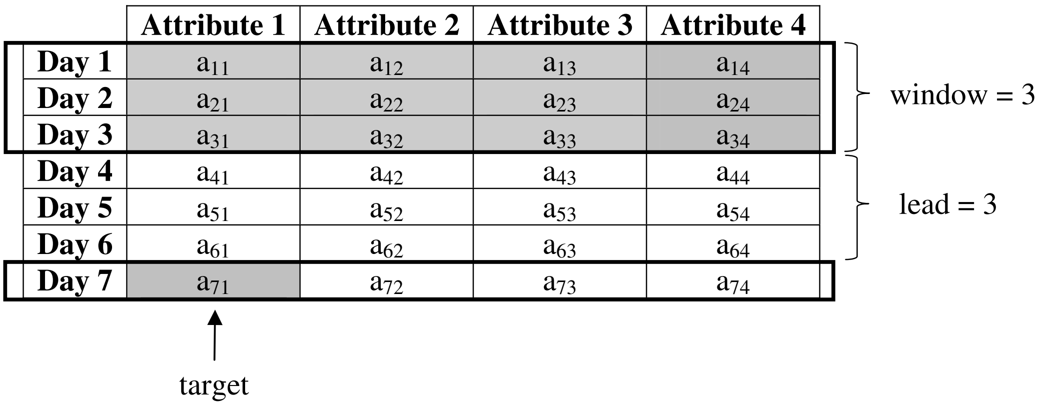
This section describes the experimental setup of the automatic data analysis using machine learning algorithms. It presents the various parameters investigated in the experiments (design space), the different learning algorithms and the evaluation process.

5.1 Design Space

A first parameter in our design space was the target attribute, which takes 6 different values, as we are interested in predicting the future values of all 6 variables monitored by the Andromeda analyzer. The input attributes correspond to values of previous days for all variables, including the target one.

One of the parameters that were studied in our experiments was the number of the preceding days that will be used for generating the prediction model. This is referred to as *window* or *time lag*. Another parameter was the *time lead*, that is the number of the intermediate days between the last day used for generating the attributes and the day we are going to predict the target variable. In the rest of the paper we will use the terms *window* and *lead* for the above parameters respectively.

Figure 5 depicts how a training example is generated from the original data given specific values for the lead and window parameters and a target attribute a_1. Figure 6 displays an example of how datasets are derived from the original dataset and given specific values for the lead and window parameters and a target attribute a_j.



	Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**
Day 1	a_{11}	a_{12}	a_{13}	a_{14}
Day 2	a_{21}	a_{22}	a_{23}	a_{24}
Day 3	a_{31}	a_{32}	a_{33}	a_{34}
Day 4	a_{41}	a_{42}	a_{43}	a_{44}
Day 5	a_{51}	a_{52}	a_{53}	a_{54}
Day 6	a_{61}	a_{62}	a_{63}	a_{64}
Day 7	a_{71}	a_{72}	a_{73}	a_{74}

Training Example: a_{11} a_{12} a_{13} a_{14} a_{21} a_{22} a_{23} a_{24} a_{31} a_{32} a_{33} a_{34} $\mathbf{a_{71}}$

Fig. 5. How a training example is generated from the original data.

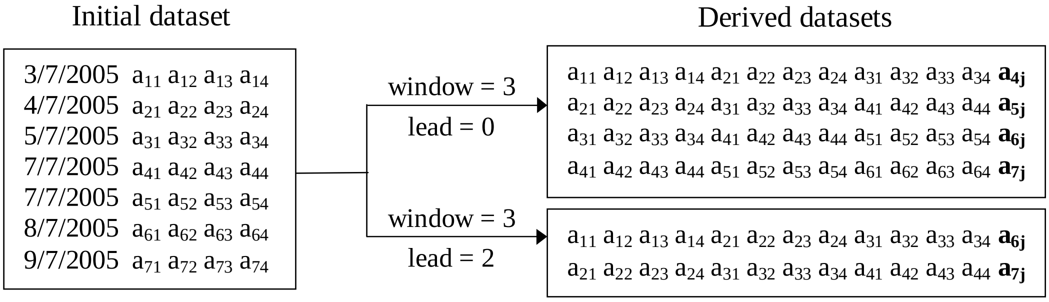

Initial dataset
Derived datasets
3/7/2005 a_{11} a_{12} a_{13} a_{14}
4/7/2005 a_{21} a_{22} a_{23} a_{24}
5/7/2005 a_{31} a_{32} a_{33} a_{34}
7/7/2005 a_{41} a_{42} a_{43} a_{44}
7/7/2005 a_{51} a_{52} a_{53} a_{54}
8/7/2005 a_{61} a_{62} a_{63} a_{64}
9/7/2005 a_{71} a_{72} a_{73} a_{74}
window = 3
lead = 0
window = 3
lead = 2
a_{11} a_{12} a_{13} a_{14} a_{21} a_{22} a_{23} a_{24} a_{31} a_{32} a_{33} a_{34} $\mathbf{a_{4j}}$
a_{21} a_{22} a_{23} a_{24} a_{31} a_{32} a_{33} a_{34} a_{41} a_{42} a_{43} a_{44} $\mathbf{a_{5j}}$
a_{31} a_{32} a_{33} a_{34} a_{41} a_{42} a_{43} a_{44} a_{51} a_{52} a_{53} a_{54} $\mathbf{a_{6j}}$
a_{41} a_{42} a_{43} a_{44} a_{51} a_{52} a_{53} a_{54} a_{61} a_{62} a_{63} a_{64} $\mathbf{a_{7j}}$
a_{11} a_{12} a_{13} a_{14} a_{21} a_{22} a_{23} a_{24} a_{31} a_{32} a_{33} a_{34} $\mathbf{a_{6j}}$
a_{21} a_{22} a_{23} a_{24} a_{31} a_{32} a_{33} a_{34} a_{41} a_{42} a_{43} a_{44} $\mathbf{a_{7j}}$

Fig. 6. How training datasets are generated from the original data.

We have experimented with 10 different values of window length that was ranging between 1 and 10. Finally, we have experimented with the lead parameter ranging from 0 to 5. A total number of 360 datasets (6 targets x 10 windows x 6 leads) have been generated from the original dataset in order to study all the parameters mentioned above.

5.2 Algorithms

For the conduction of our experiments we used the Weka library of machine learning algorithms Witten and Frank (2005). The following algorithms have been used in our experimental setup:

- SimpleLinearRegression (SLR). This class of Weka library implements an algorithm for learning a simple linear regression model. The algorithm chooses the attribute that results in the lowest squared error and can only deal with numeric attributes.
- SMO. This class of Weka library implements the sequential minimal optimization algorithm of Smola and Scholkopf (1998) for training a support vector regression model. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default.
- IBk. This is a k-nearest neighbours classifier. The algorithm normalizes attributes by default and can do distance weighting. We have used this algorithm with two different numbers of nearest neighbors ($k \in 1, 3$).
- M5P. This is a class that implements routines for generating M5 regression trees. This algorithm uses the M5 pruning criterion.

Apart from the above learning algoriths, we also included the random walk model for the purpose of comparison with a simple baseline method. The random walk model simply states that the future value of a variable will be equal to its current value supporting in that way the unpredictability of the modeling object.

5.3 Evaluation

In order to evaluate the results of our experiments we have used the 10-fold cross validation method. In particular, the performance of a classifier on a given dataset D is evaluated as follows. The dataset is split into 10 subsets D_i, $i = 1..10$ of approximately equal size. Each of these datasets D_i is used for testing the performance of an algorithm that has been trained on the union of the rest subsets $\bigcup D_j, j \neq i$. The error of the classifier is calculated by the mean of the 10 errors for all subsets.

Normalized root mean squared error (NRMSE) is used as the performance evaluation metric in the following discussion. The NRMSE metric is equal to the RMSE divided by the mean value of the target variable. This allows comparisons across the different target variables.

6 Results and Discussion

This section discusses the results of the experiments, independently for each design variable, as well as for pairs of variables.

6.1 General independent results for each design variable

Figure 7 shows plots of the average predictive performance for the different values of the four design variables. In plot (a) we notice that all algorithms exhibit better performance than the Random Walk baseline method. The lazy learning algorithm kNN, especially for $k = 3$, achieves the best results among the different learning algorithms. One interesting result in plot (b) is that dissolved oxygen and turbidity are harder to predict than the rest of the variables. Another interesting result is shown in plot (c), where the predictive performance decreases when we incorporate the values of variables for the past 2 to 3 days, while it starts to increase again from 4 to 10 days. In plot (d), we notice that, as one would expect, the performance decreases when we try to predict the target variable for more days ahead.

6.2 Pairwise results of design variables

Figure 8(a) shows the average NRMSE of the different algorithms with respect to the different values of the lead variable. We can group the algorithms into three behavioral clusters. Algorithms IB1, IB3 and MLP aren't strongly

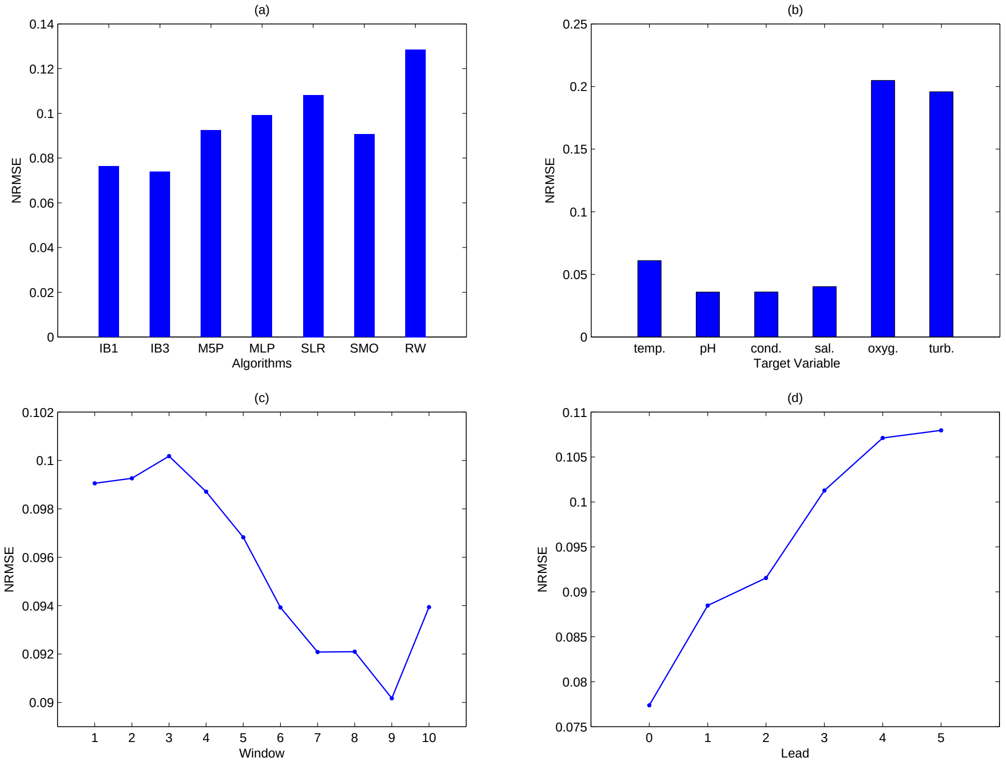


Fig. 7. Average NRMSE of the 4 design variables

affected by the increase of the lead value. In fact, we notice a decrease of the error for 3 days ahead prediction (lead=2). Algorithms RW and SLR are on the other extreme, as their error increases linearly with increasing values of lead. Notice that SLR is actually the best performing algorithm for next day prediction (lead=0). The error of SMO and M5P has an increasing trend with increasing values of lead too, but with a smaller rate compared to RW and SLR.

Figure 8(b) shows the average NRMSE of the different algorithms with respect to the different values of the window variable. We notice that the error of most of the algorithms (IB1, IB3, SLR, SMO, RW) decreases with the increase of window size. A larger window contains more information, as it includes the values of variables for a longer time period of the past. However, we also notice that the error of M5P and MLP increases with window size. A larger window size has as a consequence a large dimensionality of the input data. High dimensionality combined with few training sample might lead to overfitting, as it is probably the case for the two algorithms.

Figure 8(c) shows the average NRMSE of predicting the different target variables with respect to the different values of the lead variable. We notice that for all target variables, making predictions more days ahead is a more difficult task. Figure 8(d) shows the average NRMSE of predicting the different target variables with respect to the different values of the window variable.

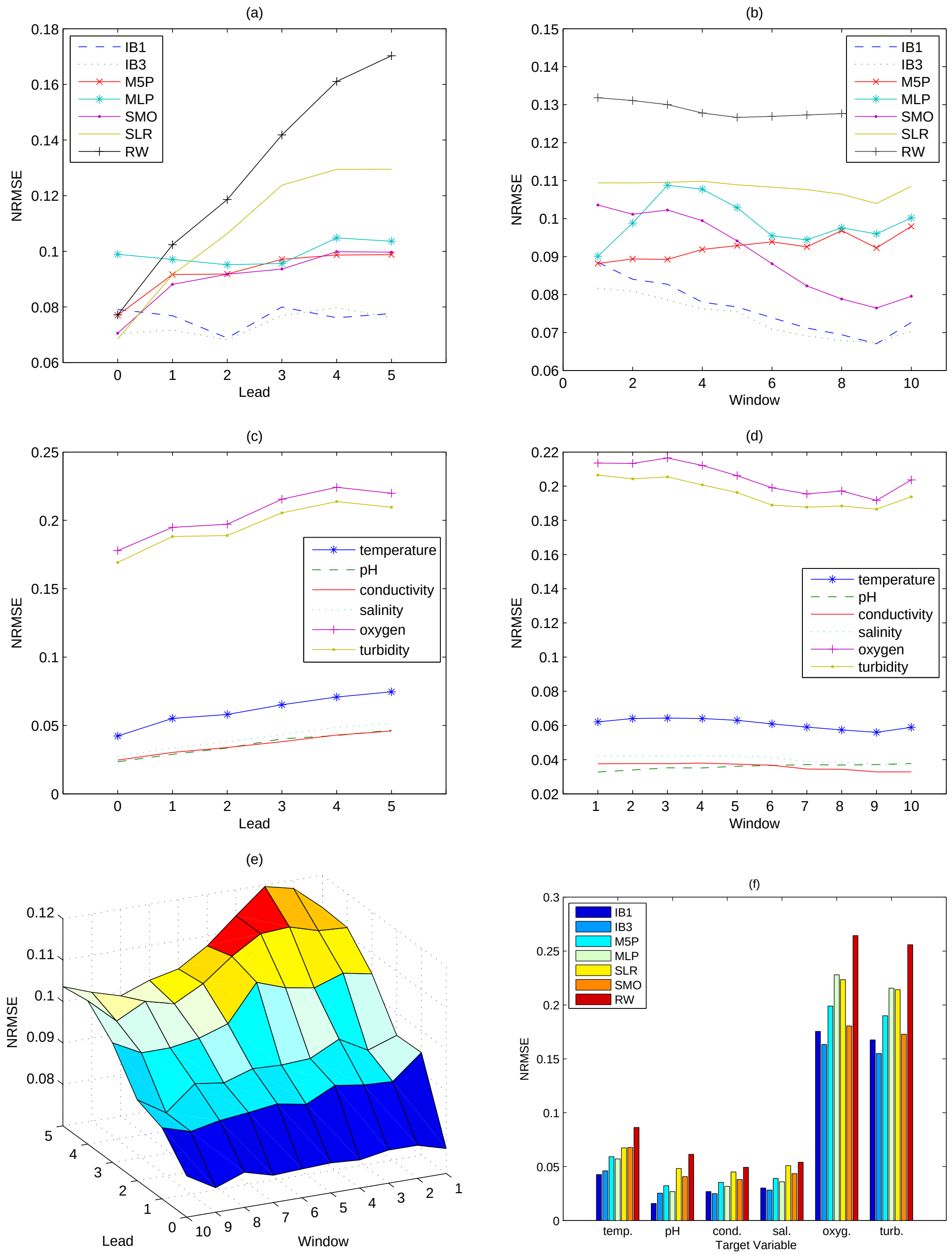


Fig. 8. Average NRMSE for the 6 pairs of design variables

We notice in general that the larger the window size the better the predictive performance. This holds especially for difficult to predict variables, such as dissolved oxygen and turbidity. In contrast, we notice that for the pH variable, which is already easy to predict, the error increases with the window size. This also shows that pH might have a shorter temporal dependence on the input variables and that including more past values is just adding noise to the modelling process.

Figure 8(e) shows the NRMSE error surface for the different values of window and lead. It is interesting to notice, that short-time predictions (small lead) are accurate even with a small window size, while if we want to make predictions further ahead, a bigger window size is required.

Finally, Figure 8(f) shows the NRMSE error of the different algorithms for the different target variables. We notice that the performance of algorithms may vary from one target to the next. For example IB1 is better than IB3 for predicting temperature and pH, while IB3 is better than IB1 for the rest target variables. In general however, the relative performance of the different algorithms exhibits a more or less uniform behavior across all target variables.

7 Conclusions and Future Work

This paper has studied the problem of water quality prediction, based on measurements from sensors deployed in the sea. It performed both exploratory and automatic analysis of the collected data with a variety of methods. The results showed that machine learning algorithms can help make accurate predictions several days ahead and are better than the Naive prediction that the value will be similar to today. Among the different learning algorithms, the nearest neighbor classifier achieved the best overall performance. In addition, we noticed that the furthest ahead the prediction, the largest the window of past values we have to incorporate in the model.

In the future, we plan to integrate the water quality prediction algorithms we presented in this paper within the intelligent alerting system of Hatzikos et al. (2007), so that the alerting system will be able to issue early warnings based on predicted hydrological parameters values.

Furthermore, we intend to investigate various energy-preservation policies and the trade-of between prediction accuracy and data quality, which will allow us to deploy the water quality monitoring system in aquasystems with limited sunlight.

References

Blockeel, H., De Raedt, L., 1998. Top-down induction of first order logical decision trees. Artificial Intellgence 101 (1–2), 285–297.

Blockeel, H., Dzeroski, S., Grbovic, J., 1999. Simultaneous prediction of multiple chemical parameters of river water quality with tilde. In: Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery. Vol. 1704 of LNAI. Springer-Verlag.

Chau, K., 2005. A split-step pso algorithm in prediction of water quality pollution. In: Proceedings of the 2nd International Symposium on Neural Networks.

Dzeroski, S., Demsar, D., Grbovic, J., 2000. Predicting chemical parameters of river water quality from bioindicator data. Applied Intelligence 13 (7–17).

Fdez-Riverola, F., Corchado, J., 2003. Cbr based system for forecasting red tides. Knowledge-Based Systems 16 (321–328).

Fdez-Riverola, F., Corchado, J., 2004. Fsfrt: Forecasting system for red tides. Applied Intelligence 21 (251–264).

Hatzikos, E., Anastasakis, L., Bassiliades, N., Vlahavas, I., 2005. Simultaneous prediction of multiple chemical parameters of river water quality with tilde. In: Proceedings of the 2nd International Scientific Conference on Computer Science. IEEE Computer Society, Bulgarian Section.

Hatzikos, E., Bassiliades, N., Asmanis, L., Vlahavas, I., 2007. Monitoring water quality through a telematic sensor network and a fuzzy expert system. Expert Systems 24 (4), (to appear).

Lehmann, A., Rode, M., 2001. Long-term behaviour and cross-correlation water quality analysis of the river elbe, germany. Water Research 35 (9), 2153–2160.

Reckhow, K., 1999. Water quality prediction and probability network models. Canadian Journal of Fisheries and Aquatic Sciences 56, 1150–1158.

Romero, C., Shan, J., 2005. Development of an artificial neural network-based software for prediction of power plant canal water discharge temperature. Expert Systems with Applications 29, 831–838.

Smola, A., Scholkopf, B., 1998. A tutorial on support vector regression. Tech. rep., NeuroCOLT2 Technical Report NC2-TR-1998-030.

Witten, I. H., Frank, E., 2005. Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann.