Bogdan DZIADAK

Warsaw University of Technology

Short and Long-term Data Prediction for Water Quality Estimation

***Abstract**. In this paper two types of prediction algorithm are presented. Presented procedures are related to the distributed measurement system for clear water monitoring. First one relates to the situation when water parameters are rapidly change during ecological disaster. Second type of prediction is connected to long-term fluctuation of water quality due to natural factors.*

***Streszczenie.** W artykule zaprezentowane są dwa algorytm predykcji wyników pomiarów parametrów świadczących o jakości wody surowej. Podane prognoz zostaną wykorzystywane do oceny jakości wody – prognoza długoterminowa, jak również będą wspomagać system alarmowania o wystąpieniu katastrof ekologicznych – prognoza krótkoterminowa. Oba przedstawione predyktory bazują na sieci neuronowej typu MLP. (**Krótko- i długoterminowa predykcja danych do oceny jakości wody surowej**)*

Keywords: measurement systems, environmental protection, data prediction.
Słowa kluczowe: systemy pomiarowe, ochrona środowiska, predykcja.

Introduction

High water quality is important for developing ecosystem and natural environment, for industry and of course for people which each day use water to live. Quality of water changes during year according to the:

- Season – spring fertilization cause additional water contamination by ammonia, phosphates, and nitrates;
- Weather – chance of rain and flooding can rapidly change water resource in ecosystem;
- Temperature – high temperature accelerates the growth of algae and bacteria (especially cyanosis).

Modeling water quality is important because give the information haw to protect or treating water resource in ecosystem.

In this paper two types of prediction algorithm are presented. First one relates to the situation when water parameters are rapidly change during ecological disaster. Second type of prediction is connected to the long-term fluctuation of water quality due to natural factors. Presented procedures are related to the distributed measurement system for clear water monitoring which was build at Warsaw University of Technology, and installed on the Dobczyckie Lake near the Krakow (Poland). The author is co-designer of this system[1] – [3].

Metrological condition

According to Polish law for full estimate the quality of clear water, periodical monitoring of 134 parameters is needed. In cooperation with water supply company, author selected eight major parameters which permanently measurement will most benefit. Parameters and their limit values in one liter of water are listed in table 1. All parameters are used to modeling water quality, first four parameters determined water treatment technology. Chlorophyll inform about cyan-bacteria concentration. Last three parameters inform about chemical pollution.

Table 1: Basic parameters for estimate the quality of water [4]

	Parameter	Unit	limit water indicators
1.	Temperature	^{o}C	25
2.	Acidity	pH	6.5 – 8.5
3.	Conductivity	µS/cm	1000
4.	Suspension	mg/l	25
5.	Chlorophyl	um/l	20
6.	Ammonia	mg/l	0.5
7.	Chlorides	mg/l	250
8.	Hydrocarbons	mg/l	0.05

The system has two main task:

- Instant information about ecological disaster;
- Long term monitoring water parameters.

During disaster, pollution can penetrate via water intake to treatment reservoir, were biological filters (based on bacteria and algae) are installed. This penetration can entirely destroy the filters. Renovation the filters is difficult expensive and time-consuming. Therefore quickly information about danger level of contamination or danger trend of contamination is desirable. The predictor can be installed directly on measurement station or on the server. The most important thing is that the information immediately go to the users.

Short and long-term prediction

Short-term prediction is alarming part of build monitoring system. In this case the short time of measurement is more important than precision.

To create prediction algorithm be:

1. Define model of the time series which will be predicted
2. Construction prediction algorithm
3. Estimate quality of forecast, for this process author used to type of factors:

- Mean Absolute Error defined as(1):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - \hat{p}_i| \tag{1}$$

where: p_i is value of the measurement in time i
$\hat{p}_i$ is forecast of measurement
n is number of measurement

- Mean Absolute Percentage Error defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - \hat{p}_i|}{p_i} * 100\% \tag{2}$$

Based on characteristic used probes, and typical measurement process, is possible to identified the model of time series to predict. Dynamic characteristic of DHP probe for liquid hydrocarbon concentration measurement is presented on the Figure 1. The probe gives the stable value after 12 to 15 minutes. Characteristic is typical exponential curve, based on this shape is possible describe time series model via equation (3):

$$y(t) = a\left(1 - e^{-t/\tau}\right) + v \tag{3}$$

$$\lim_{t\to\infty} y(t) = \lim_{t\to\infty}\left(a\left(1 - e^{-t/\tau}\right)\right) = a \tag{4}$$

where: a is steady value, τ is time constant and the v is value represented signal noise. Based on this mode, calculation the forecast is reduced to calculate limit function $y(t)$ in infinity, assuming that value v is negligible.

Author realized prediction algorithm based on function model were based on minimum sum of square criterion the coefficients of $y(t)$ and forecast of measurement are calculate (4).

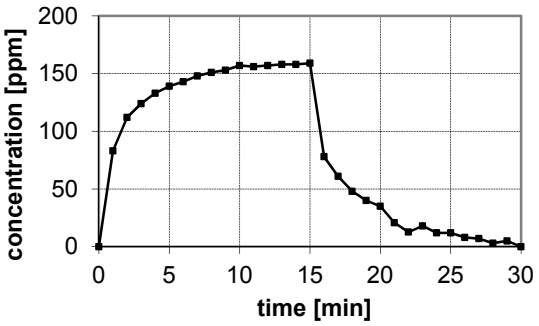

Fig. 1. Dynamic characteristic of the DHP probe

After estimation quality of the prognosis proved that error of forecast ($MAPE > 5\%$) is too high to implemented this algorithm in system.

As the next, prediction algorithm based on the Neural Network was created. The NN are very often used in many areas [5], also in prediction [6].

Author used the Multi Layer Perceptron (MLP) net. This predictor work in configuration presented on the figure 2. Predictor used specified number of fallowing measurement which are the same as input of net and give the regression of the measurement.

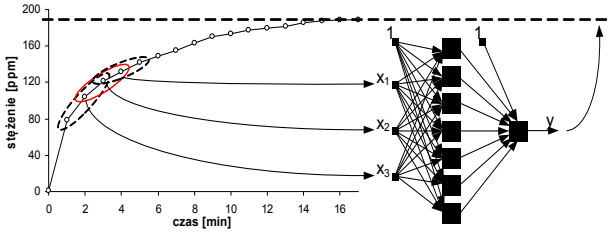

Fig. 2. Predictor based on MLP neural network

Author tested predictors which has from 3 to 6 input one hidden layer which has from three to fifteen neurons. All neurons can adopt one from four activation function (linear, exponential, hyperbolic-tangent and sigmoid). For neurons which has continuous and differentiable activation function the minimization of target function can be realized via iteration using gradient algorithm. As optimization algorithm for NN learning the Quasi-Newton method Broyden – Flacher – Goldfrab – Shanon (BFGS) was used.

Learning NN to prediction concentration liquid hydrocarbons in water was in supervisor mode. The learning series included 32000 pairs input record and output value. In this series was real measurement collected by authors and simulation measurement generated with different noise value, according to function (3) . The series was divided for four subsets. Each subsets contain cycles hydrocarbon measurements form 3 to 102 ppm with resolution 1ppm, and time interval one minute. The value of noise v was random and has limit:

- In first subset $v \in (-0.5,\ 0.5)ppm,\ \tau = 3min$
- In second subset $v \in (-1,\ 1)ppm,\ \tau = 3min$
- In third subset $v \in (-3,\ 3)ppm,\ \tau = 3min$
- In fourth subset $v \in (-0.5,\ 0.5)ppm,\ \tau \in (2.5,\ 3.5)min$

Series contain situation when concentration contamination rise, down, and be stable. Algorithm can predict rise and decries of contamination. Prediction in wider range of measurement is not necessary because in natural environment concentration liquid hydrocarbons highest than 100ppm is not observed, even during ecological disaster.

Author created networks which has 3, 4, 5, and 6, input for each mode 40 networks differing in number of hidden neurons was tested. In table 2 the best result was presented.

All nets were tested using new measurement series included 24000 pairs input and output value. The best predictors gives good prognosis after one, maximum 2 minute after then contamination start to rise. So the algorithm shortens time of measurement 6 – 7 times. In this case is only then minutes, but we should known that the intake output is 2500 litters per second (100 t. m^3/day).

Table 2: Parameters of the best predictors

	Structure	Testing Error	Activation function	
			Hidden	Output
1	MLP 3-14-1	0.000132	Thah.	Sigmo.
2	MLP 4-12-1	0.000086	Thah.	Sigmo.
3	MLP 5-15-1	0.000082	Thah.	Exp.
4	MLP 6-9-1	0.000072	Thah.	Sigmo.

The information from predictor will allow for quickly closure the intake, thus to protect the biological filters. Illustration the measurement process and forecasting is presented on the figure 3.

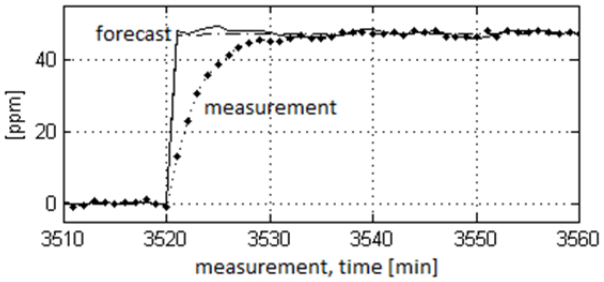

Fig 3. Measurement process and forecasting (MLP 6-9-1)

In the table 3 is presented set the best predictors. Errors MAE and MAPE was calculate according to formula (1) and (2). The number of errors higher than 10ppm represented errors in all testing series.

Table 2: Errors of the best predictors

	Structure	MAE [ppm]	MAPE [%]	Errors >10ppm
1	MLP 3-14-1	0.99	4.14	87
2	MLP 4-12-1	0.82	3.56	47
3	MLP 5-15-1	0.78	3.58	57
4	MLP 6-9-1	0.75	3.29	46

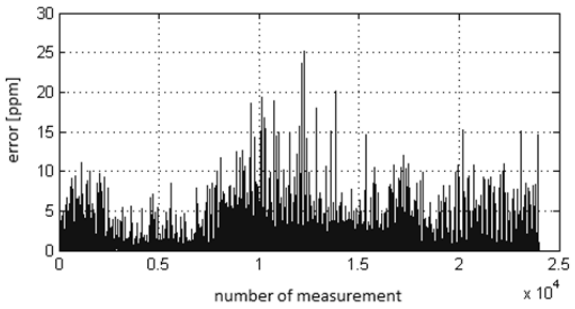

Fig.4. Absolute error of prediction for MLP 6-9-1

Analyzing errors the best result gives the NN which has six input and 9 neurons in hidden layer. Value of MAPE lower than 4% mean that the algorithm work properly. Number errors higher than 10ppm is the smallest for 4

input-net but the statistical error are better in 6 input-net. Absolute errors for 6 input NN for all series is presented on figure 4, and histogram of this error is presented on the figure 5. On histogram we can see that the more than 18.6 thousand forecasts to 24000 has absolute error smaller than 1ppm.

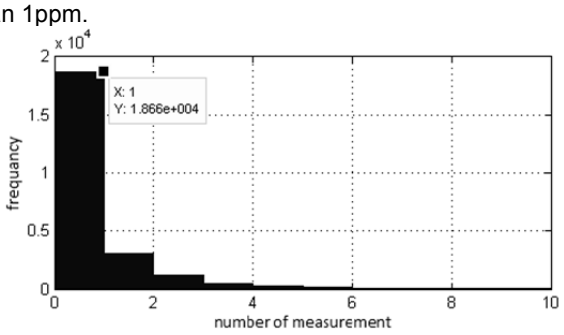

Fig. 5 Histogram of absolute error for MLP 6-9-1 predictor

Long term prediction for water quality modelling

One of the most monitoring parameters is chlorophyll a because is directly indicator of cyan-bacteria. It is possible to measure chlorophyll concentration using dedicated probes but this probe are very expensive. Considering the fact, author with colleges create algorithm to predict chlorophyll, based on the level of another parameters, which are easier to measure. At the beginning the algorithm based on data fusion and Bayes modelling function [7], [8].

Author create long-term predictor, for chlorophyll level estimation, base on the Neural Network MLP. In this case as a learning data the real measurement, from two years monitoring of Dobczyckie Lake was used. The time interval between fallowing measurement is one week. In base was 100 fallowing measurement of 14 parameter. Author tested three type of Neural Network:

- 14 input – input record of this net contain 14 parameters (date, temperature, turbidity, pH, O2, O dissolved, silica, phosphates, phosphor, N, B.O.D, Ch.O.D., suspension, conductivity);
- 6 input – input record contain 6 parameters (date, temperature, turbidity, pH, suspension, conductivity);
- 4 input – input record contain 4 parameters (temperature, turbidity, pH, conductivity).

For each type of NN was tested 20 nets. To the learning process author used only 70 pairs output and input records. This small amount of learning data caused that the errors are high. The bests nets are listed in table 3. Two last predictors bas on parameters which are measurement directly by measurement station. This parameters will be measure in 1 hour interval, so in nearly time the weights of net can be recalculate.

Table 3: Parameters of the best predictors

	Structure	MAE [ug/l]	Test correlation
1	MLP 14-5-1	2.34	0.800
2	MLP 6-3-1	2.48	0.865
3	MLP 4-5-1	3.09	0.710

Analyzing data from table 3 the best result has 14 input net. However to implementation in the system the MLP 6 - 3-1 net was chosen. This net has the simplest structure, and need only parameters which will be measure by measurement station floating on the leak. The error is compare to the 14-imput net, and has the best correlation between prognosis and measurement result. Using this net author predicted level of chlorophyll during 30 weeks. The prognosis have one week time horizon. The result of this research are presented on the figure 6.

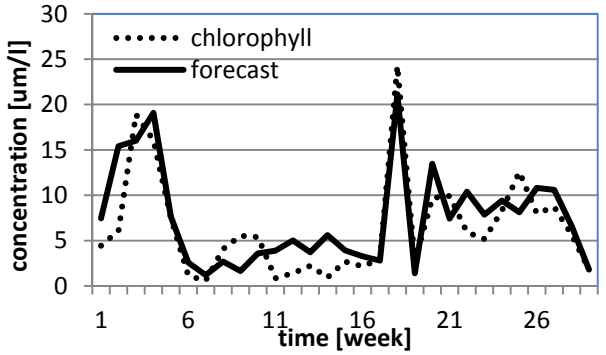

Fig. 6. Level of chlorophyll in lake and its forecast

Conclusion

Short and long-term prediction concentration component of raw water is very helpful in alarming situation and for modelling water quality. Both presented algorithm based on MLP predictor. Short-term forecast based on typical one step predictor. This solution can shorten time of measurement six times, what is very desirable during ecological disaster. Long-term predictor merge data fusion and one step predictor. Long term prognosis are used to modelling changes of biological parameters, as cyan-bacteria.

ACKNOWLEDGMENTS

Work of Bogdan Dziadak has been supported by the European Union in the framework of European Social Fund through the Warsaw University of Technology Development Programme.

References

[1]. Dziadak B., Michalski A., "Quality engineering tools used to designe and optimization of Mobile Measurement Station structure", *IEEE Instrumentatation and Measurement Magazine*, 13, (2010), 01, 33 - 38

[2]. Dziadak B., Michalski A., "Evaluation of the Hardware for a Mobile Measurement Station", *IEEE Industrial Electronics Transactions on*, 58, (2011), 7, 2627 – 2635.

[3]. Dziadak B., Kalicki A., Staroszczyk Z., Makowski Ł., Michalski A., "Framwork architecture of a large scale distributed measurement system for enviromental protection", Inovative Technological Solutions for Sustainable Development (2010), 235 – 263.

[4]. Bochnia T., Kaszowski J.; Opracowanie założeń dot. monitoringu jakości wód powierzchniowych w trybie *on-line*. Raport MPWiK, (in Polish), (2011).

[5]. Kołodziej M., Majkowski A., Rak R.," Wykorzystanie maszyny wektorów wspierających (SVM) do klasyfikacji sygnału EEG na użytek interfejsu mózg-komputer", Pomiary Automatyka Kontrola (in Polish), (2011), 12, 1546 – 1548.

[6]. Siwek K., Osowski S., Sowoński M., "Evolving the Ensemble of Predictors Model for Forecasting the Daily Average PM10", International Journal of Environment and Pollution, 46 (2011), ¾, 199 – 215.

[7]. Makowski Ł., "Bayesian method to evaluate uncertainty of data fusion used to estimate cyanobacteria levels in Dobczyckie reservoir", *Przegląd Elektrotechniczny*, (2012), 04a, 126 – 128.

[8]. Dziadak B., Kalicki A., Staroszczyk Z., Makowski Ł., Michalski A., „Wykorzystanie fuzji danych do estymacji liczebności sinic w jeziorze Dobczyckim. *Przegląd Elektrotechniczny*, (in Polish), (2011) 9a, 87 – 90.

***Authors**: dr inż. Bogdan Dziadak,* bogdan.dziadak@ee.pw.edu.pl *Politechnika Warszawska, Instytut Elektrotechniki Teoretycznej I Systemów Informacyjno-Pomiarowych, ul. Koszykowa 75, 00-662 Warszawa.*