Łukasz MAKOWSKI

Warsaw University of Technology

Bayesian method to evaluate uncertainty of data fusion used to estimate cyanobacteria levels in Dobczyckie reservoir

Abstract. *Distributed measurement systems are more common in protection of strategic natural resources such as clean water. Bayesian methods can be applied to algorithms of such systems to improve quality of immediate situation assessment and hazard notification. In this paper a practical example of such approach is discussed.*

Streszczenie. *Rozproszone systemy pomiarowe zdobywają popularność w ochronie strategicznych zasobów naturalnych takich jak woda pitna. Metody bayesowskie mogą zostać użyte w algorytmach takich systemów, w celu poprawy jakości natychmiastowej oceny sytuacji i powiadamiania o zagrożeniach. W tym artykule omówiony jest praktyczny przykład realizacji takiego podejścia.* ***(Zastosowanie metod bayesowskich do estymacji liczebności sinic w Zbiorniku Dobczyckim)***

Keywords: distributed measurement systems, data estimation algorithms, Bayesian methods
Słowa kluczowe: rozproszone systemy pomiarowe, algorytmy estymacji danych, metody bayesowskie

Introduction

Blue-green algae also known as cyanobacteria might be hazardous to ecosystem and living creatures including humans. Unfortunately cyanobacteria are common specie, residing also in aquatic environments including clean water reservoirs.

Fig. 1. Dam on Dobczyckie reservoir and its position on map of Poland

The Dobczyckie reservoir, which picture is shown on figure 1, is placed in south of Poland, near Krakow. It is a clean water source for more than 500 thousand people. Estimation of cyanobacteria levels in this reservoir are investigated and some results of this research are presented in this paper.

Some limitations of current measurement system

It is futile to count numerous cyanobacteria cells even in small sample of water. However, value of chlorophyll-a is measurable parameter, consequent to their amount in water [1]. This parameter actually can be defined with use of complex laboratory methods or very expensive probes. On the other hand biology of cyanobacteria is researched for a long time and it is known that their quantity is related to level of total phosphorus (TP) and total nitrogen (TN) dissolved in water [2]. Measurements of these compound groups are easier *in situ* than counting cyanobacteria or measuring chlorophyll-a, and may lead to earlier estimation of their number.

Data fusion with Levenberg-Marquardt Algorithm (LMA) was successfully used to estimate level of cyanobacteria in Dobczyckie reservoir [3]. However, explanation of difference between computed and real values, which was considerably high at some points, was considered to be insufficient. Therefore there is a need for approach with more probabilistic explanation of processes that occur in environment and to provide better information about confidence level over calculated estimations.

Assumptions for probabilistic approach

Two benefits of Bayesian approach are crucial: a dynamic linear model can represent time-varying parameters and it makes feasible to simultaneously model different aspects of time series such as level trend and variance.Cyanobacteria level that is considered to be unknown variable and other, measurable water parameters are continuous functions which may be observed at times $t = 1 \ldots T$.

Modelling starts from simple autoregression of order 1, which has the form:

$$y_t = \mu + \rho_1 y_{t-1} + u_t \qquad t = 2, \ldots T \tag{1}$$

Predicted, discrete cyanobacteria level is estimated as y_t and y_{t-1} is one-step lagged observation. Autocorrelation between successive values is given by ρ. Signal error is denoted as u_t. As the environment limits values of y_t, signal is centred and the level of outcome μ is equal to zero, which simplifies equation.

Having operator B that denotes earlier samples of signal y_t, general equation for order p autoregression might be written as follows:

$$y_t - \rho_1 B^1 y_t - \rho_2 B^2 y_t \ldots - \rho_p B^p y_t = u_t \tag{2}$$

which can be further simplified to:

$$\rho(B) y_t = u_t \tag{3}$$

Except of TP and TN also dissolved oxygen (DO), acidity (pH) and turbidity (FTU) have high correlation with cyanobacteria quantity level. Surprisingly water temperature and season has no impact on cyanobacteria blooms. Finally, current number of cyanobacteria may also cap or boost further colony growth, depending on amount of nutrients dissolved in water. Therefore problem evolves to search for a multivariate model.

Lets denote variables having impact on cyanobacteria as a vector $\mathbf{Y}_t = (y_{1t}, y_{2t}, \ldots y_{Kt})'$, and provide coefficients matrices $\mathbf{\Phi}$ of size $K \times K$, and signal errors $\mathbf{U}_t \sim \mathbf{N}_K(0, \mathbf{V})$ where $\mathbf{V}$ is a covariance matrix. Then we can rewrite above equations with matrix approach of multivariate autoregression:

$$\mathbf{Y}_t - \mathbf{\Phi}_{p1} \mathbf{Y}_{t-1} - \mathbf{\Phi}_{p2} \mathbf{Y}_{t-2} \ldots - \mathbf{\Phi}_{pp} \mathbf{Y}_{t-p} = \mathbf{U}_t \tag{4}$$

Applying presented approach to normal distribution we can calculate conditional likelihood between cyanobacteria

level and measurable water parameters with equation:

(5)
$$f(\mathbf{Y}_2|\mathbf{Y}_1,\rho,\frac{1}{\sigma^2}) \propto \frac{1}{\sigma^{(n-p)}} exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\left[\rho\left(B\right)\mathbf{Y}_t\right]^2\right)$$

Coefficients computing procedure

To calculate sum from equation 5, which is a matrix version of equation 3, matrices of coefficients $\mathbf{\Phi}$ that fit lagged vectors $\mathbf{Y}_{t-n}$ with error vector $\mathbf{U}_t$ are necessary. On the basis of known time series given by matrix $\mathbf{Y_t}$, and assumed error vector $\mathbf{U}_t$, these coefficients can be found using curve fitting method such as Levenberg-Marquardt Algorithm (LMA).

Variance of $\mathbf{U}_t$ and its precision are constant across whole time series. In final result an average value of signal becomes its basis level with variance as uncertainty. Ignoring this fact one can get quite reliable solution but without confidence level provided by uncertainty distribution.

To obtain curve fitting all K^2 entries from each matrix $\mathbf{\Phi}$ must be calculated where K is number of water parameters under consideration. With $K = 5$ and base of $p = 4$ recent measurements it makes 100 parameters to compute.

Lets denote curve as $\mathbf{C} = \mathbf{Y}_t$, which corresponds with equation 4. Entries of $\mathbf{C}$ can be computed independently using just a single row K of every p-lagged coefficients matrix $\mathbf{\Phi}_{pp}$ and corresponding observation vectors $\mathbf{Y}_{t-p}$:

(6)
$$\begin{aligned} C_K = U_K + &\Phi_{p1,K,1}Y_{t-1,1} &+\ldots+ \Phi_{p1,K,K}Y_{t-1,K} &+\\ +&\Phi_{p2,K,1}Y_{t-2,1} &+\ldots+ \Phi_{p2,K,K}Y_{t-2,K} &+\\ +& &\ldots & +\\ +&\Phi_{pp,K,1}Y_{t-p,1} &+\ldots+ \Phi_{pp,K,K}Y_{t-p,K} & \end{aligned}$$

Above formulae might be simplified to:

(7)
$$C_K = U_K + \sum_{p}\sum_{n=1}^{K}\Phi_{pp,K,n}Y_{t-p,n}$$

For every parameter K, LMA might be used independently to contribute part of $\mathbf{\Phi}_{pp}$ entries. Repeating this procedure will provide the whole matrix. In practical approach of programmer it is a series of loops over multidimensional arrays. Application to solve this problem was written in Python language with help of NumPy and SciPy scientific packages.

Equation for estimation of chlorophyll-a level

Chlorophyll level, which is main indicator for quantity of cyanobacteria in water, is to be modelled on the basis of other values that are easier to measure *in situ*. However, a proper modelling equation must be constructed before coefficients calculation procedure can be initiated. This article is based on already tested, previously investigated methodology [3]. However, equation was completely overhauled and data vector $\mathbf{Y}$ used with new model is defined as:

(8)
$$\mathbf{Y} = \begin{bmatrix} chlorophyll-a \\ 0,5^{TP\times\log(1+COD)} \\ \left(pH\times\frac{TN}{TP}\right)^2 \\ pH \\ \sqrt[3]{FTU} \end{bmatrix}$$

One may ask, from where this and not other equation vector $\mathbf{Y}$ originated. Simple answer is that there is little to none suggestion how parameters should look like. Researchers agree that cyanobacteria level is correlated with TN:TP ratio but it is also known that this simple relation is not enough to describe complex cellular life of cyanobacteria and thus statistics of environmental ecology [4, 5, 6]. Therefore, author used his knowledge gathered from literature, intuition, experience with the problem, and lots of tests with different versions of the model equation. We may assume that further progress in this modelling is highly possible. However, at this point the vector $\mathbf{Y}$ presented above was chosen as best known result.

To create the model, part of data already collected *in situ* was used as reference, and part is used in verification process. Both data sets were about one year length with measurements that had been made at least once a week.

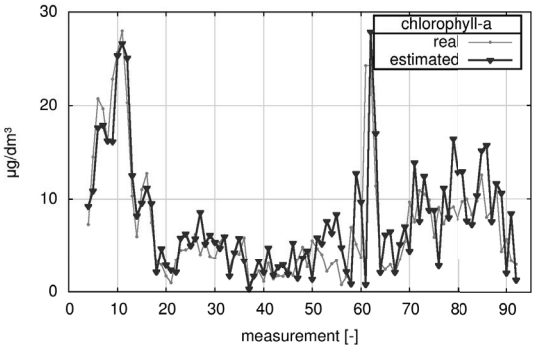

Fig. 2. Estimated chlorophyll-a level compared with real values

Having computed coefficients it is now possible to use them in further calculations, and to built procedure of estimation cyanobacteria level in Dobczyckie reservoir. Result of this step is not-lagged plot of chlorophyll-a level estimated on the basis of five other, measured parameters (TN, TP, COD, pH, FTU). Plot and corresponding real data are shown on figure 2 where samples up to 40 were used in fitting procedure, and forthcoming did not affect coefficients. We may notice that both plots are highly correlated.

Probabilistic approach to model

Probabilistic approach is based on equation 5, where we need to calculate standard deviation σ and series of core values.

Assuming that estimation was correctly regressed to real signal, standard deviation can be calculated on priors and used for whole series of results. To calculate σ of sum given by equation 4, assuming that its Y_{t-p} are independent variables, we may use central limit theorem (CLT). Denoting consecutive entries of $\Phi_{pp,K,K}$ as b_i and their related data vectors $Y_{t-p,K}$ as X_i, the theorem calculation can be approximated with equation:

(9)
$$\sigma\left(Y_t\right) = \sqrt{\sum_{i=1}^{n} b_i^2\sigma^2\left(X_i\right)}$$

Computation is dominated by one of parameters (b1), that corresponds to previous value of chlorophyll-a. That makes sense as forthcoming value should be close to previous state, up to a point. Values of water parameters related to other coefficients must change significantly to affect cyanobacteria estimation.

To validate this approach we calculate standard deviations:

- of real chlorophyll-a level $\sigma_r = 7.41$,
- of values estimated over train series $\sigma_t = 6.74$,

– of values estimated over verification series $\sigma_v = 5.30$,
– calculated with use of CLT $\sigma_{CLT} = 9.52$.

Standard deviation calculated with CLT is close enough to σ_r. As it is calculated over variables from vector given by equation 8 it might be assumed that it represents uncertainty hidden in all series of single vector entries. This imposed uncertainty is small enough that σ_{CLT} does not differ too much from standard deviations of other time series. Therefore its value proves right decision the use of vector 8 and parameters calculated with LMA. Moreover, it is the worst case from values presented above thus a safe choice for next computations.

The core is a series of sums over $\mathbf{\Phi}_{pp}\mathbf{Y}_{t-p}$ which are based on priors. Sum of squared cores is used in exponentiation. Main goal at this stage of research is to estimate chlorophyll-a level based on other, measurable parameters. Quality of estimation must be improved before prediction of future values series will be possible. Therefore in equation 5 sum is over only one element $Y_{n=p+1}$, which is based on previous states of model given by equation 6. Therefore conditional likelihood of next value Y, based on $p = 4$ priors equals to:

$$f(\mathbf{Y}_2|\mathbf{Y}_1, \rho, \frac{1}{\sigma^2}) \propto \frac{1}{\sigma} exp\left(-\frac{1}{2\sigma^2}\left[\rho\left(B\right)\mathbf{Y}_t\right]^2\right) \tag{10}$$

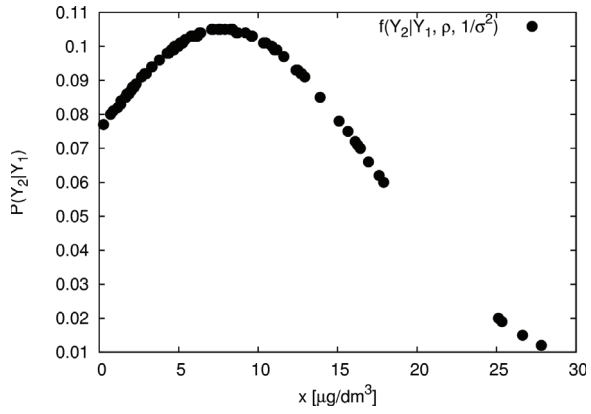

Fig. 3. Distribution of likelihood of estimation conditioned on priors

Therefore conditional likelihood that next value Y_2 can be estimated on priors Y_1 based on theory ρ, which has assumed variance σ^2, is given by normal distribution. Further from average estimated value is, the probability of its correctness lowers what is shown on figure 3.

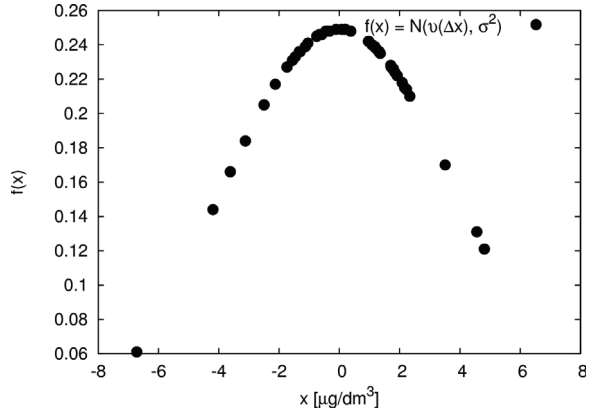

Fig. 4. Distribution of estimation absolute error

However, having real values it is possible to check actual uncertainty of presented estimation algorithm. On figure 4 it is shown a distribution of absolute error of estimated series. Assuming it actually is normal distribution then average absolute error is $\mu = 0.63$ and $\sigma = 4.01$. Therefore 68% of estimated values compared to real values are no further from them by more than $4.01 mg/dm^3$.

Conclusions

Main objective of presented research was to increase quality of chlorophyll-a estimation, which value is directly related to cyanobacteria presence. Fitting quality was improved with new enhanced model equation, which halved χ^2 distance between estimated and real values compared to previously published results.

Main limitation of current model is quality of modelling all parts of $\mathbf{Y}$ vector. Optimization was done concerning chlorophyll-a level, which is most important from practical point of view. But at the same time other entries of $\mathbf{Y}$ are modelled with significantly lower quality. Therefore it is still a problem to create a model with which better time series prediction could be produced. To sum up this issue it might be said that described model is an improved estimator but still not a predictor.

Despite some limitations the approach presented in this paper provides a useful tool to speed up assessment of dangerous situation, such as unusually high fertilization, and makes it easier to quickly counteract beginning of cyanobacteria bloom. Furthermore usage of oxidants during water treatment process might be managed more economically, which can lower costs for business and citizens, and provide them with water of better quality.

This work has been supported by the European Union in the framework of European Social Fund through the Warsaw University of Technology Development Programme.

REFERENCES

[1] Håkanson L., Bryhn A.C., Hytteborn J.K. On the issue of limiting nutrient and predictions of cyanobacteria in aquatic systems Science of the Total Environment, 379 (2007), pp. 89 - 108
[2] Smith V.H. Havens K., East T., James T., N:P ratios, light limitation, and cyanobacterial dominance in a subtropical lake impacted by non-point source nutrient pollution. Environmental Pollution, 122 (2003), pp. 379 - 390
[3] Dziadak B., Kalicki A., Staroszczyk Z., Makowski Ł., Michalski A., Wykorzystanie fuzji danych do estymacji liczebności sinic w jeziorze Dobczyckim. Metrologia Wspomagana Komputerowo - proceedings, 2011 (in Polish)
[4] Tandeau de Marsac N., Lee H.M., Hisbergues M., Castets A.M., Bédu S. Control of nitrogen and carbon metabolism in cyanobacteria Journal of Applied Phycology, 13 (2001), no.4, pp. 287 - 292
[5] Smith V.H. Havens K., East T., James T., N:P ratios, light limitation, and cyanobacterial dominance in a subtropical lake impacted by non-point source nutrient pollution, Environmental Pollution, 122 (2003) pp. 379 - 390
[6] Yandigeri M.S., Yadav A.K., Meena K.K., Pabbi S. Effect of mineral phosphates on growth and nitrogen fixation of diazotrophic cyanobacteria Anabaena variabilis and Westiellopsis prolifica. Antonie van Leeuwenhoek, 97 (2010), no. 3, pp. 297 - 306

Author: Ph.D. Łukasz Makowski, Institute of Theory of Electrical Engineering Measurement and Information Systems, Faculty of Electrical Engineering, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warszawa, Poland, email: lukasz.makowski@ee.pw.edu.pl