Use of fuzzy method to estimate river nutrient loads from scarce observation

K. Buzás
Department of Sanitary and Environmental Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary

Abstract. Evaluation of data time series in order to get information about water systems is one of the routinely needed tasks. The results are always associated with uncertainties, of which one arises from data scarcity. Traditional methods, such as regression analyses etc. become rapidly useless with decreasing number of data available. A method based on fuzzy set theory was applied to get more reliable information about the system from scarce databases. Monitored daily flow and water quality data of the medium size Zala River in Hungary were considered as elements of fuzzy sets. Fuzzy rules were generated from data pairs (flow, suspended solids concentration, water temperature and phosphorus load as inputs and output, respectively) from which combined rule bases were set up. These rule bases can be considered as a tool of mapping from the input space to the output space using defuzzification procedure. The method is trainable: it can learn from observations. It is demonstrated that the method is capable to generate daily phosphorus loads and annual balance with acceptable accuracy when it is trained only by weekly, biweekly or monthly data pairs. In comparison to other approaches the tool is well suited to utilize better the information content of scarce observations. Furthermore, monitoring costs can be considerably decreased without substantial information loss since sampling of expensive and labour intensive parameters can be reduced.

Keywords: Lack of data; water quality; uncertainty; fuzzy method; nutrient load

Introduction

Uncertainties about complex physical and biochemical processes arise from lack of knowledge as well as lack of data. Various types of uncertainties are discussed extensively in the literature (Beck, 1987; Ganoulis, 1994), where a difference is made between objective and subjective (or secondary) ones. There are several reasons why man-induced uncertainties arise: (i) data uncertainties that may be due to sampling or measurement method, and/or analysis errors; (ii) modelling uncertainties caused by applying the wrong model and/or parameter estimate errors; and finally, (iii) operational uncertainties related to errors of construction, operation, and maintenance of engineering systems. In contrast with uncertainties related to natural randomness, the degree of these uncertainties can be decreased if more information is obtained, and/or the model applied is improved, etc.

Evaluation of data time series in order to get information about water systems is one of the routine tasks. In the case of eutrophication the estimation of nutrient load is crucial since water quality management requires a reliable knowledge on nutrient loads and balances. It is typical that we have hydrological data (for example flow and water temperature) sampled with high frequency (sometimes daily or even continuously measured), while water quality data are scarce. The frequency of river samples taken either occasionally or routinely is mainly up to weekly, at best. Questions always arising are as follows: Is water quality sampling frequency satisfactory? What are the errors of calculated nutrient loads from scarce observation? How can we refine the estimates?

Traditionally, methods based on probability calculus and supported by a substantially and thoroughly developed theoretical basis have been used for handling uncertainties. The application of these methods and the reliability of results are limited by the amount of data available. Depending on the method and number of data, reliability may decrease so much
that the calculation itself may become senseless. Sometimes it is very costly to increase the number of data (requiring sampling and laboratory analysis). At other instances such as using historical data time series it is not even possible to do so. Therefore in the latter case the information content of scarce data cannot be extracted dependably.

Routine sampling, however, has two limitations in every case: (i) samples do not cover all characteristic situations (e.g. flood events are omitted); and (ii) there is a limited amount of data available for statistical evaluation. The weight of the first problem can be considerably decreased by sampling characteristic events (low flows, floods, periods of snow melting, etc.) instead of routine sampling. The other limitation cannot be eliminated without additional sampling, and extra expenses. The number of data is usually increased by using long time series (sometimes of several years). This, however, may bring another uncertainty factor in the calculation: changes might occur in the watershed in the meantime, which decrease the homogeneity of the time series. Such changes include, for example, the alterations in agricultural soil cultivation practices and/or in fertilizer application, the varying share of phosphate-free detergents in the household consumption, or upgrading the treatment technology (nutrient removal) at wastewater treatment plants, etc.

When data are scarce, the fuzzy set theory may be used to handle and quantify imprecision. Fuzziness occurs when the boundary of a piece of information is not clear-cut. A fuzzy set $A$ can be characterized by a membership function, which assigns to each object of a domain its grade of membership in $A$ (Zadeh, 1965). The more an element or object can be said to belong to a fuzzy set $A$, the closer to 1 its grade of membership. The focus of fuzzy sets theory is thus placed upon its nonstatistical characteristics in nature that refers to the absence of sharp boundaries in the information. This feature of fuzzy set theory makes it able to represent mathematically a class of linguistic problems, among others decision problems. The application of fuzzy sets has a long tradition in the water management, and there are several examples for its use in linear regression, control and forecasting in uncertain environments (Hesmathy and Kandel, 1985; Koo and Shin, 1986; Baffaut and Chameu, 1990; Bardossy et al., 1990; Duckstein and Bogardi, 1991; Ganoulis, 1994), as well. However, numerical water quality data processing is rather rare (Müller, 1994). Hereinafter a method based on fuzzy rules, which are generated from numerical data (observations) is used to get more information for estimate annual phosphorus load from scarce water quality data.

Like for many other lakes, the estimation of annual phosphorus load is an essential problem of the management of the Lake Balaton, the largest shallow lake in Central Europe. The watershed of the lake is drained by the River Zala, which is the largest river characterized by 7 m$^3$/s mean flow and by a number of small watercourses. Samples have been taken uniquely daily at the mouth of Zala for about twenty years. In contrast, there are a few water quality data for the creeks, which makes difficult and uncertain the load estimation. The uncertainty of results remains unknown, according to practical experience, however, it may even be as high as some hundred percent. Reliability can be considerably increased by more sophisticated methods such as taking flow as well as suspended solids concentration into account and using regression analyses (Clement, 2001).

Fuzzy methodology
The method developed by Wang and Mendel (1992) has been used to supplement scarce water quality data subsequently and to compute annual P loads. This method consists of five phases: (i) the input and output spaces are divided into fuzzy regions; (ii) fuzzy rules are produced from related data pairs; (iii) the degree of each rule is specified; (iv) a combined fuzzy rule base is developed; and (v) the rule base is used for mapping from the input space to the output space by a defuzzification procedure.
A given set of input/output data pairs is provided by the sampling program: \((x_1^{(1)}, x_2^{(1)}; y^{(1)}), (x_1^{(2)}, x_2^{(2)}; y^{(2)} \ldots )\). The first step to produce fuzzy rules from the data pairs is to define fuzzy regions in the input and output space, respectively. The possible domain interval of both, the inputs and the outputs are divided into any odd number of partial ranges (regions) in such a way that they overlap each other. The length of region may differ in the case of each variable. One fuzzy membership function is assigned to each region. The shape of membership functions may also differ. A triangular one has been applied in our case.

Generating fuzzy rules from numerical data, first the membership value of \(x_1^{(i)}, x_2^{(i)}, \text{ and } y^{(i)}\) is specified for each region. The data pairs \(x_1^{(i)}, x_2^{(i)}, \text{ and } y^{(i)}\) are assigned to the region where their membership value \((m(x_1), m(x_2), m(y))\) is the highest. Denote \(R_{x_1}^i, R_{x_2}^i\text{ and } R_y^i\) the regions. The fuzzy rule is set as: \(\text{IF } x_1 \in R_{x_1}^i \text{ AND } x_2 \in R_{x_2}^j, \text{ THEN } y \in R_y^k\).

Rules can be derived from all data pairs; therefore it is highly possible that many of them will be in conflict with some others (those where the input regions are identical but the output regions are different). In order to eliminate this conflict, a degree is assigned to each rule and only the rule of the highest degree is accepted. The degree of a rule will be the product of the membership values of inputs and output. Expert knowledge may also be integrated into the objective data analysis at this point. It is a general phenomenon to have unrealistic, “wild” data pairs in time series. They result from errors made in one of the steps of the data generation procedure. The expert when analysing time series can evaluate previously the reliability of such data, assigning a fuzzy membership value \((m_e)\). For reliable data we assign high membership values (e.g. close to 1), while the “bad” ones will be provided with lower values. Therefore not only data but data pairs as well are considered as part of a fuzzy set. This way the above degree of a fuzzy rule can be modified by multiplication with \((m_e)\).

In the case of two inputs and one output a combined rule base set up from the fuzzy rules forms a two dimensional matrix. The number of fuzzy regions applied for the inputs defines the size of the matrix. Each box contains the number of the output region included in the rule with the highest degree. The rule base can be considered as the fuzzy mapping of primary information to be obtained from measurement data. If a fuzzy rule base completed on the basis of measurement data available, the output related any inputs could be obtained by a defuzzification procedure. In our case the centroid defuzzification formula were used.

This method has many useful features for solving practical problems. As the fuzzification/defuzzification procedure consists of processing data measured, it is able to “learn” from the “examples”, i.e. from new sampling data. This adaptivity has high importance for mapping systems that change in the course of time. The method can be viewed as a trainable and model-free fuzzy system. Latter means that it is not an application requirement to have any, often highly non-linear (or even not existing) mathematical model of the system. These conditions frequently occur with environmental systems. The fuzziness is introduced into the system by linguistic (expert) fuzzy rules and/or fuzziness of data. If the fuzzy rule base can be completed in a way that it does not contain any empty boxes, it will be suitable for mapping any non-linear and continuous functions with desired degree of accuracy (Wang and Mendel, 1992). The method is simple and it is built up step by step. It allows a much quicker learning process than the neural network method applied for similar purposes, however, its price is more memorial capacity required.

These features make this method highly applicable for process control. However, they would not be sufficient for processing environmental (water quality) data if the problem to be solved arises from data scarcity. In this case we have not enough data pairs to complete the fuzzy rule base as the possible input space is very extensive and there is only a limited number of data pairs to be used for training. There is, however, a further and very useful feature of the method: we can fill the empty boxes up without any further data. This process
is based on the primary information (limited given rules) obtained from actual data and stored in the filled boxes. If characteristic but not measured values are set for the input regions of empty boxes, the missing output regions of the fuzzy rule base can be specified by applying steps (i) to (iv). Starting from the empty boxes that are the nearest neighbours of the full ones filled up during the training period, the completion process expands to the whole entire input area, which mostly needs iteration.

Implementation
Let us consider the river to be a complex, non-linear system that provides measurable answers to the inputs. There may be various input parameters (for example, precipitation in the watershed, waste water discharges, etc.). In order to get the information content of time series of water quality data, let us take the flow (Q) and suspended solids concentration (SS) or water temperature (T), alternatively as input values. Daily phosphorus load (TP) is selected to be the output of the system.

Application for the Zala River is provided in two phases. First, the mapping system is trained using 365 data pairs of a set of years (1992-1998), then daily TP values are generated. Secondly, we illustrate how the mapping system can be trained using incomplete database. Then the daily phosphorus loads are generated, and compared with sampled (365) data.

Figure 1 shows the time series of daily measured and generated phosphorous loads in the year of 1994. This year was rainy enough to demonstrate the effects of runoff and erosion on phosphorus load, besides point sources emissions. The value range of each input has been divided into 25 fuzzy regions, that of the output into 27 fuzzy regions. As Figure 1 shows, mapping provides adequate results even near the peak values. However, considering the phosphorus load of the Lake Balaton it is more important that the generated annual load is only 0.97% higher than that of estimated from daily data. Having done the calculations for the other years, it has turned out that the annual load estimate always shows less than 3% difference between the generated and the measured one. This has been taken as a proof for the fact that the method is capable of mapping the behaviour of the system in question. The practical importance of this capability, however, lies in the fact whether it is also functional with scarce data basis.

To illustrate how the method is capable of utilizing the information content of scarce data, three options of input configurations were evaluated. Since the correlation between the flow and the phosphorus load of Zala River has been found strong (Somlyódy and van Straten, 1986), it was obvious that daily TP generation should be based on the daily flow

![Figure 1](https://iwaponline.com/wst/article-pdf/43/7/257/429828/257.pdf)

**Figure 1** Sampled and generated daily phosphorus loads of the River Zala in 1994 (the fuzzy system is trained using daily data)
time series, in the first run. In the next step, besides the daily flows, the values of daily water temperatures were taken into consideration as input for training and mapping. The latter have been used to account better for snow melting and heavy storm events. The phosphorus load is also correlated with the suspended solids concentration on an eroded watershed as the watershed of River Zala. Accordingly, it was expected that the generation could be further improved by involving SS as input parameter besides flow.

It was also expected that the rate of data scarcity could considerably influence the estimation. For the training process of fuzzy method deficient time series (simulating weekly, biweekly and monthly routine sampling frequencies) were developed from the complete (daily) database. For example, the routine weekly measurements were produced by selecting the measurement data of the first, then the second, etc. day of each week. Then we generated the daily TP and compared them with the sampled data. The evaluation criterion was the accuracy of the generated annual phosphorus load, accepting the measured one as “precise” load. Besides the accuracy of annual load, the computing capability considering the dynamics of river was also measured by the correlation between the daily measured and generated time series. The process was executed for seven years separately, of which the year of 1994 was chosen to demonstrate the results.

As can be seen in Table 1 the average accuracy of annual load is certainly acceptable, even if the sampling frequency is low. The increase of average error is more or less proportional to the rate of decrease of sampling frequency, while the range of error is widened. Considering the similar ranges of minimum errors, we can conclude that the success of load generation rather depends on the relevancy of data used for training than that of number. Sampling a few but characteristic flood events and low flow periods instead of frequent routine samples is far efficient for training the fuzzy method both in methodological and economic aspects. This feature of method resulted in other conclusion: there is no gains and consequently reason to use long time series for training. When the mapping was trained with seven years data, the annual loads were less correct than those were generated only from sampling of the given year. This follows from the formation procedure of fuzzy rules. Additional information getting from water temperature or suspended solids concentration data slightly improves the accuracy. However, the lower the sampling frequency the less is the gains. It should be underlined that the method is also capable to reproduce the dynamics of the River Zala, inasmuch as the measured and generated daily loads are well correlated in every case. This is illustrated in Figures 2 (a–c).

We can conclude that the mapping method is capable to generate the missing data of daily loads with a fairly high degree of accuracy using only limited information (scarce data). As far as the annual load concerned, accuracy of the calculation is definitely higher than that of traditional methods (Clement and Buzás, 1999).

Table 1 Error of generated annual phosphorus load in the year of 1994

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Sampling frequency</th>
<th>Weekly</th>
<th>Biweekly</th>
<th>Monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Load error, %</td>
<td>Max.</td>
<td>Min.</td>
<td>Average</td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td>9.3</td>
<td>0.14</td>
<td>4.6</td>
</tr>
<tr>
<td>Q·T</td>
<td></td>
<td>12.8</td>
<td>0.1</td>
<td>3.6</td>
</tr>
<tr>
<td>Q·SS</td>
<td></td>
<td>3.6</td>
<td>1.83</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Q = flow; T = water temperature; SS = suspended solids concentration; r² = correlation coefficient between measured and generated daily TP loads

Downloaded from https://iwaponline.com/wst/article-pdf/43/7/257/429828/257.pdf on 02 November 2019
Table 2 Error of generated annual mean phosphorus concentration

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Weekly</th>
<th>Biweekly</th>
<th>Monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average error of annual mean TP concentration generated, %</td>
<td>$r^2$</td>
<td>Average error of annual mean TP concentration generated, %</td>
</tr>
<tr>
<td>Q - T</td>
<td>3.1</td>
<td>0.70</td>
<td>5.5</td>
</tr>
<tr>
<td>Q - SS</td>
<td>4.9</td>
<td>0.74</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Q - flow; T - water temperature; SS - suspended solids concentration; $r^2$ - correlation coefficient between measured and generated daily TP concentrations

Figure 2 Sampled and generated daily total phosphorus loads of the River Zala in 1994. The fuzzy system is trained using: weekly (a), biweekly (b) and monthly (c) flow and water temperature data.
It was supposed that the good results are partly due to the high correlation ($r^2=0.84$ in 1994) between the flow and the phosphorus load. Therefore the method was also tested for the generation of daily total phosphorus concentrations, which are practically not correlated either with flow or with the flow-water temperature or flow-suspended solids concentration data pairs ($r^2=-0.2, 0.5$ and $0.58$, respectively). Comparing the results with the load generation it can be concluded that the method is also capable to compute the annual mean phosphorus concentration with an acceptable accuracy (Table 2) on the base of scarce, uncorrelated hydrological data. However, less effectiveness is observable regarding the dynamics. Although the dynamics of the daily phosphorus concentration profile almost has been lost, as can be seen in Figures 3–5, the seasonal changes are generated well.

It should be underlined that the accuracy of generation depends also on the way we formulate the domain intervals of fuzzy regions. While in generating TP loads equally spaced triangular membership functions were used, the altering spaced ones proved to be better for TP concentrations. The high degree of freedom in the first phase of model set up indicates that the performance of method could be further improved.

![Figure 3](image-url) Sampled and generated daily phosphorus concentration of the River Zala in 1994. The fuzzy system is trained using weekly: (a) flow and water temperature, and (b) flow and suspended solids concentration data.

![Figure 4](image-url) Sampled and generated daily phosphorus concentration of the River Zala in 1994. The fuzzy system is trained using biweekly: (a) flow and water temperature, and (b) flow and suspended solids concentration data.

![Figure 5](image-url) Sampled and generated daily phosphorus concentration of the River Zala in 1994. The fuzzy system is trained using monthly: (a) flow and water temperature, and (b) flow and suspended solids concentration data.
Conclusions
From the study the following conclusions can be drawn.
i. The fuzzy-based, trainable, and model-free procedure developed by Wang and Mendel can be successfully implemented for mapping dynamics of TP loads of River Zala strongly influenced by non-point source impacts.
ii. If the data used for training are sufficiently representative, the method is capable to generate missing data with an acceptable accuracy.
iii. Even in the case of a scarce data base containing less representative data, the fuzzy mapping system generates the annual mass balance with higher accuracy than the traditional methods using the same set of data. It can be considered as a tool better suitable to extract the information content of scarce historical data, than traditionally applied methods (e.g. regression analyses).
iv. Monitoring costs can be decreased without substantial information loss since fewer parameters with reduced frequency are satisfactory to be monitored. The method can be also applied for decreasing uncertainty of load estimates if the sample size is kept.
v. Although the method is quite robust, easy to apply, and performs surprisingly well, it does not substitute the mathematical models of the system, (if any) but it rather supplements them. Although it represents a universal approximator, the effective application needs detailed knowledge about the substantial processes of the river system.

References
With application to Lake Balaton. Springer-Verlag, Berlin, p 386.