Hindawi Publishing Corporation
Applied Computational Intelligence and Soft Computing
Volume 2012, Article ID 846321, 9 pages
doi:10.1155/2012/846321

Research Article

Modelling of Water Quality: An Application to a Water Treatment Process

Petri Juntunen,[1] Mika Liukkonen,[1] Marja Pelo,[2] Markku J. Lehtola,[1] and Yrjö Hiltunen[1]

[1] *Department of Environmental Science, University of Eastern Finland, P.O. Box 1627, 70211 Kuopio, Finland*
[2] *Finnsugar Ltd., Sokeritehtaantie 20, 02460 Kantvik, Finland*

Correspondence should be addressed to Petri Juntunen, petri.juntunen@uef.fi

Received 10 October 2011; Revised 19 December 2011; Accepted 25 December 2011

Academic Editor: Cheng-Jian Lin

The modelling of water treatment processes is challenging because of its complexity, nonlinearity, and numerous contributory variables, but it is of particular importance since water of low quality causes health-related and economic problems which have a considerable impact on people's daily lives. Linear and nonlinear modelling methods are used here to model residual aluminium and turbidity in treated water, using both laboratory and process data as input variables. The approach includes variable selection to find the most important factors affecting the quality parameters. Correlations of ~0.7–0.9 between the modelled and real values for the target parameters were ultimately achieved. This data analysis procedure seems to provide an efficient means of modelling the water treatment process and defining its most essential variables.

1. Introduction

Water quality is becoming an ever more important issue, as water of low quality causes many significant problems. In particular, there is a wide range of microbial and chemical constituents of drinking water that can cause either acute or chronic detrimental health effects, and the detection of these constituents in treated water is often time-consuming, complex, and expensive [1]. On the other hand, water of bad quality can also be harmful from an economic perspective, as resources have to be directed towards improving the water supply system every time a problem occurs. For these reasons, there is growing pressure to improve water treatment and water quality management in order to ensure safe drinking water at reasonable costs. Systematic assessments of raw water, treatment processes, and operational monitoring issues are needed to meet these challenges.

There are many parameters which can be used to measure the quality of water, of which turbidity is a common one, the purpose being to measure impurities in the water. In a physical sense, turbidity is a reduction in the clarity of water due to the presence of suspended or colloidal particles, and it is commonly used as an indicator of the general condition of drinking water [1]. Furthermore, turbidity has been used for many decades as an indicator of the efficiency of drinking water coagulation and filtration processes, so that it is an important operational parameter for this reason, too. High turbidity values refer to poor disinfection and possibly to fouling problems in the distribution network, so that turbidity should be minimized [2]. However, turbidity is a quite sensible and faulty measurement, and many variables and phenomena are influencing it. This makes turbidity challenging for modeling purposes [1, 2].

Another important quality parameter for treated water is residual aluminium, especially when aluminium flocculants are used in the treatment process [2]. Residual aluminium causes turbidity in water networks, resulting in acceptability problems for consumers [1]. Usually the phenomenon can be seen when residual aluminium exceeds 0.1–0.2 mg/L, which are the usual guideline levels for residual aluminium [1]. In addition, metals such as aluminium have been implicated in the pathogenesis of Alzheimer's disease [3]. Some epidemiological studies show that there can be a correlation between neural disorders and Al concentrations of 0.1 mg/L in the drinking water [3].

Many chemical and physical features of raw water affect the water treatment process. Many organic and inorganic compounds in suspended, colloid, or solved form influence

the flocculation process. Organic compounds, which are usually measured by a $KMnO_4$ test, play an essential role in the process. Furthermore, many inorganic compounds such as the silicate or the pH of raw water also affect the process. As an example of physical parameters, the water temperature has a remarkable influence on the flocculation in water treatment processes [4–6]. Naturally process conditions also have a great effect. The dose of the flocculation chemical is naturally the key parameter, as is the adjusted pH value. Further, hydraulic variables such as flow to the process or filters affect the performance [2].

Moreover, in water treatment there are observable cycles or episodic events present which cause the process to behave dynamically. The variation in water consumption is one of these, causing changes not only within a day but also within a week and even within a year. Year cycles can be distinguished even more clearly if surface water is treated, because the water temperature is observed to have some effects on the process [7]. In addition, the phenomena existing in the process are usually state dependent, meaning that a certain phenomenon in the process may work differently in different process conditions [8].

As a general tool which can be of assistance in improving water treatment, the modelling of water processes has confronted many challenges. Since the treatment processes involved are physically and chemically heterogeneous [4–6], the water and process parameters are generally complex and their mutual interactions nonlinear [9]. Furthermore, successful applications of traditional mechanistic models are limited to idealized, artificial systems [10], so that the correlation between simulated and experimental data from real processes has been poor and expensive *in situ* testing has been needed [9–11].

A process-oriented approach performs optimizations based on real process data. In practice, the variables used in modelling are derived from archived laboratory or process data resources. At present, data-based multivariate methods such as multiple linear regression (MLR) and artificial neural networks such as multilayer perceptrons (MLPs) are considered advantageous for analysing process data. Many applications have demonstrated that they provide an efficient automated method for modelling industrial process data [12–14]. Data-based modelling has also been used in connection with wastewater treatment [15], water resources [16, 17], water distribution systems [18], and water treatment. The most general applications to water treatment processes involve the prediction of quality parameters such as turbidity, colour [9, 19], or the optimal dosing of flocculation chemicals [9, 11].

MLPs, above all, have proved their efficiency in the modelling of water treatment processes. The method has many advantages: the MLP technique is robust and allows the development of multivariate, nonlinear models without any physical or chemical knowledge of the process [11], which means that it offers a computationally powerful alternative for complex problems in which nonlinearity is present. The drawback with MLP models, however, is that they have more complex mathematical formulae than more explicit models such as those based on linear regression. MLP also requires substantially more knowledge from the user than do simpler statistical methods. It is therefore reasonable to use this technique only in applications where linear methods have failed.

Traditionally, multivariate analysis methods such as factor analysis and principal component analysis (PCA) have been widely used in analyzing hydrological system. However, the limitations of conventional multivariate statistical methods arising from the challenges mentioned earlier are known [20].

In summary, understanding the complex relationships and phenomena prevailing in large systems is a challenging task. The quality of water in a treatment process, for example, may be affected by several factors which either are not known thoroughly or which have not been verified on an experimental basis. For these reasons, data-based modelling methods, such as MLP, would be preferable for modelling of water treatment processes.

In this paper we employ a multivariate linear regression method (MLR) and a nonlinear modelling method (MLP) to model turbidity and residual aluminium in a water treatment process using both process and laboratory data as model inputs. Because process data typically consist of a large number of variables, we use variable selection as a diagnosis tool to find the most important input variables affecting the outputs. We compare the results of the MLR and MLP models to explore their applicability to real-life modelling purposes.

2. Process and Data

The experimental data were collected from the water treatment plant of Suomen Sokeri in Kirkkonummi, Finland. The plant uses mainly surface water from Lake Humaljärvi or a mixture of this with water from the Pikkala reservoir. The process is a typical chemical process with a coagulation and flocculation unit, flotation, and powdered activated carbon (PAC) filtration (see Figure 1). PAX-14, an aluminium-based coagulation chemical produced by Kemira Kemwater, is used in doses varying between 30 and 80 g/m^3. The dose is set as a function of the raw water $KMnO_4$ content, so that the $Al/KMnO_4$ ratio will be between 0.8 and 2 $kgAl/kgKMnO_4$. For most of the time, the ratio is near 1.3. The final decision regarding the dose is in the hands of the process personnel. The process is shown in Figure 1.

The pH value is adjusted to 6.1–6.3 with calcium hydroxide before flocculation, this having been found experimentally to be the optimum pH for removing organic compounds and turbidity. After filtration it is readjusted to 8.2 to be suitable for distribution. Finally, the water is disinfected with UV radiation and by adding sodium hypochlorite.

The data were obtained from process and laboratory measurements over a period of 373 days. The original process data period was 5 minutes, which was averaged to daily data in order to be comparable to the laboratory data. Before modelling, the outliers in the data were filtered out manually and the missing data points filled in by linear interpolation.

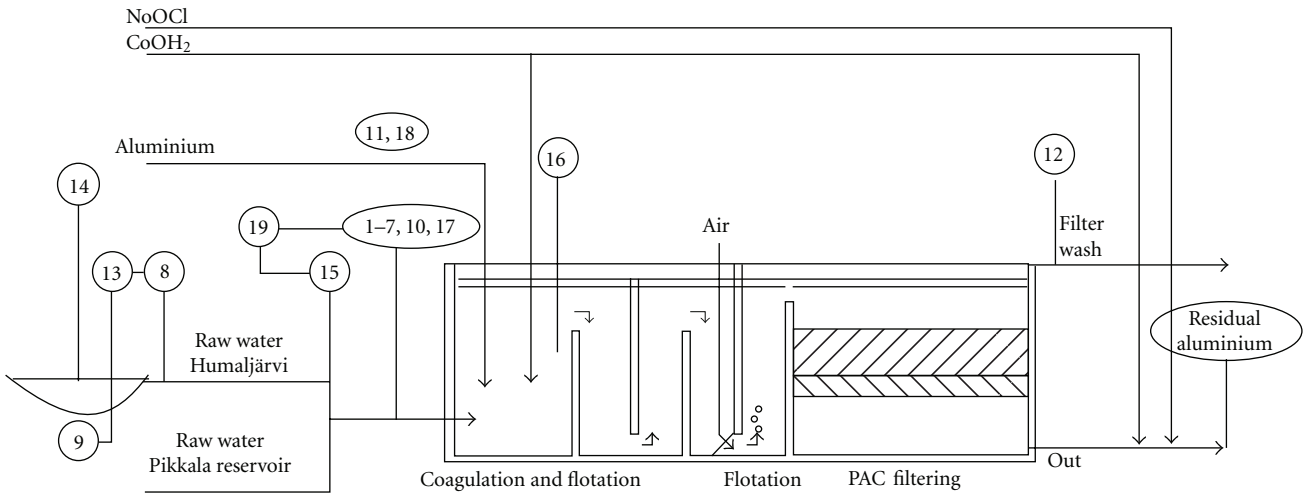

Figure 1: The water treatment process used by Suomen Sokeri. The numbers refer to the measuring points for the variables (see Tables 1 and 2).

The raw water and process variables are shown in Tables 1 and 2, Figure respectively.

Some of the process variables were left out before modelling, so that only those variables were chosen which could potentially have an effect on the output when manipulated by the controllers, for example. Consequently, all the variables measured after the output measuring point were omitted, as also were those variables which cannot be converted to on-line measurements. In addition, variables containing data of bad quality (e.g., too many missing data points) were ruled out manually. The measuring points for the process variables are shown in Figure 1.

3. Methods

3.1. Multiple Linear Regression (MLR). MLR [7] can be used to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data samples. An MLR model with N observations and P variables is defined by

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_P x_{iP} + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, N, \tag{1}$$

where y denotes the value of the response variable, x is the value of the predictor (explanatory) variable, b_0 is a constant, $b_1 \cdots b_P$ equal the unknown coefficients to be estimated, and ε comprises the uncontrolled factors and experimental errors in the model. The fitting is performed by minimizing the sum of the squares of the vertical deviations from each data point to the line that fits best for the observed data, which is known as least squares fitting.

3.2. Multilayer Perceptrons (MLPs). MLP networks are well-known feed-forward neural networks [12, 14] consisting of processing elements, called neurons, and connections. The neurons are arranged in three or more layers: an input layer, one or more hidden layers, and an output layer. An MLP network is trained with data samples, leading to a supervised learning procedure. The network input signals are processed forward through successive layers of neurons on a layer-by-layer basis. In the first phase the input layer distributes the inputs to the first hidden layer. Next, the hidden neurons summarize the inputs based on predefined weights, which either weaken or strengthen the effect of each input. The weights are determined by learning from examples (i.e., data samples), which is called supervised learning. Eventually, the inputs are processed by a transfer function, and the result is transferred as a linear combination to the next layer, which is generally the output layer. The performance of the model is then evaluated with an independent validation data set.

MLP neural networks must be trained for each problem separately. A popular MLP training technique is the back-propagation algorithm [21], in which the output values are compared with the proper answer from the original data in order to calculate the value for a predefined error function. Eventually the iterative training procedure defines a set of weights which minimize the error between the actual and expected outputs for all input patterns. In summary, the back-propagation training proceeds in two phases [12].

(1) Forward Phase. The network weights are fixed and the input is forwarded through the network until it reaches the output.

(2) Backward Phase. The output of the network is compared with the desired response to obtain an error signal, which is propagated backwards in the network. In the meantime, the network weights are adjusted successively to minimize the error.

3.3. Selection of Variables. The enormously increased amount of information available in recent years has caused the selection of variables or reduction of model inputs,

Table 1: Raw water variables used for modelling and their correlations with residual Al and turbidity.

	Variable	Unit	Correlation with residual Al	Correlation with turbidity
1	pH of raw water		0.21	0.21
2	$KMnO_4$ of raw water	mg/L O_2	0.18	0.33
3	Hardness of raw water	mmol/L	0.27	0.22
4	Colour of raw water	mg Pt	0.19	0.27
5	Conductivity of raw water	mS/m	0.34	0.38
6	Silicates on raw water	mg/L	0.35	0.28
7	Turbidity of raw water	NTU	0.07	0.06

Table 2: Process variables used for modelling and their correlations with residual Al and turbidity.

	Variable	Unit	Correlation with residual Al	Correlation with turbidity
8	Intake from Lake Humaljärvi	m^3/h	−0.16	−0.27
9	Intake from the pikkala reservoir	m^3/h	0.25	0.39
10	Total intake of water	m^3/h	−0.033	−0.07
11	Aluminium feed	L/h	0.11	0.23
12	Filter wash water	m^3/h	−0.36	−0.12
13	Proportion of Lake Humaljärvi/Pikkala reservoir water intake	%	−0.52	−0.38
14	Surface level of Lake Humaljärvi	m	0.28	0.14
15	$KMnO_4$ of raw water	mg/L	0.065	0.44
16	Flocculation pH			
17	Water temperature	°C	−0.72	−0.42
18	Aluminium dose	g/m^3	0.25	0.46
19	Aluminium dose/raw water $KMnO_4$		0.052	−0.08
20	Flow to filter 1	m^3/h	0,078	0.03
21	Flow to filter 2	m^3/h	0,22	0.19
22	Flow to filter 3	m^3/h	−0.097	−0.11
23	Flow to filter 4	m^3/h	0.31	−0.32

to become a relevant part of data analysis [22–25]. The objective of this selection procedure can be to improve the prediction performance of the model, to provide faster processing of the data or to provide a better understanding of the process [24]. When exploiting artificial neural networks for computation purposes, for instance, reducing the number of model inputs may shorten the computing times significantly. With respect to certain tasks such as process diagnostics, however, it is also useful to discover the main factors affecting the physical phenomena.

In practice, the aim is to select a subset p from the set of P variables without appreciably degrading the performance of the model and possibly improving it. Although exhaustive subset selection methods involve the evaluation of a very large number of subsets, the number to be evaluated can be reduced significantly by using suboptimal search procedures [26]. One of these is the *sequential forward selection* method, which was used for the selection of variables in this case.

In sequential forward selection, the variables are included in progressively larger subsets so that the prediction performance of the model is maximized. To select p variables from the set P,

(1) search for the variable that gives the best value for the selected criterion;

(2) search for the variable that gives the best value *with* the variable(s) selected in stage 1;

(3) repeat stage 2 until p variables have been selected;

3.4. Application of Methods. At the first stage variables were selected using multiple linear regression and a sequential forward search. The data were divided into two subsets: a training subset comprising 2/3 of the total number of samples, to be used for training the model, and a validation data set consisting of the remaining 1/3 of the samples, to be used as an independent means of testing the model. The first eight variables that improved the performance of the model most were finally chosen, because in practice the models did not seem to improve beyond this point.

Next, variables were selected using an MLP network with a back-propagation algorithm and a sequential forward search. The data were divided into three subsets: a training subset, comprising 2/3 of the total number of samples, to be used for training the network, of which a test subset containing 20% of the training data was reserved for back-propagation error calculations and a validation data set, consisting of the remaining 1/3 of the samples, to be used as an independent means of testing the model.

The artificial neural network consisted of the process parameters as inputs, one hidden layer with 5 neurons and the output neuron describing the predicted variable. The parameters of the neural network and the training algorithm were determined experimentally. The radial basis (*radbas*) transfer function was used for the hidden layer and the linear (*purelin*) transfer function for the output layer. The Bayesian regularization back-propagation (*trainbr*) algorithm [27] was exploited in training, and the sum squared error (*sse*) as the error function in training. Matlab (version 7.11) software with the Neural Network Toolbox (version 7.0) was used for the data processing.

4. Results

4.1. Modelling of Turbidity. The variables selected for water turbidity using MLR and MLP are presented in Tables 3 and 4, respectively. Evolution curves for the selecting of variables using the MLR and MLP techniques are shown in Figure 2. The results for predicting the validation data using

Table 3: Variables selected for turbidity using multiple linear regression.

Round	Variables	Correlation coefficient (C)
1	Aluminium dose	0.55
2	Intake from the pikkala reservoir	0.62
3	Turbidity of raw water	0.66
4	Proportion of Lake Humaljärvi/Pikkala reservoir water intake	0.70
5	Colour of raw water	0.71
6	$KMnO_4$ of raw water	0.71
7	Flocculation pH	0.71
8	Water temperature	0.71
9	Aluminium dose/raw water $KMnO_4$	0.71
10	Silicates in raw water	0.71

Table 4: Variables selected for turbidity using multilayer perceptrons.

Round	Variables	Correlation coefficient (C)
1	Aluminium dose	0.56
2	Intake from the pikkala reservoir	0.66
3	Turbidity of raw water	0.70
4	Proportion of Lake Humaljärvi/Pikkala reservoir water intake	0.75
5	Water temperature	0.75
6	$KMnO_4$ of raw water	0.76
7	Flocculation pH	0.77
8	Intake from Lake Humaljärvi	0.78
9	$KMnO_4$ of raw water	0.76
10	pH of raw water	0.77

MLR and MLP with 8 variables are given in Figures 3 and 4, respectively.

In addition, a two-sample F-test was conducted between the outputs of the MLR and MLP models with 8 variables. The test showed that the null hypothesis cannot be rejected with a P value of 0.5323 using the 0.95 confidence level; that is, there is no significant difference between the linear and nonlinear models.

4.2. Modelling of Residual Aluminium. The variables selected for residual aluminium in the water using MLR and MLP are presented in Tables 5 and 6, respectively. Evolution curves for the selecting of variables using linear regression and the MLP technique are shown in Figure 5. The results for predicting the validation data using MLR and MLP with 8 variables are given in Figures 6 and 7, respectively.

In addition, a two-sample F-test was conducted between the outputs of the MLR and MLP models with the 8 variables. The test showed that the null hypothesis can be rejected with a 0.95 confidence and a P value of 0.0485 that is, there is a statistically significant difference between the linear and nonlinear models.

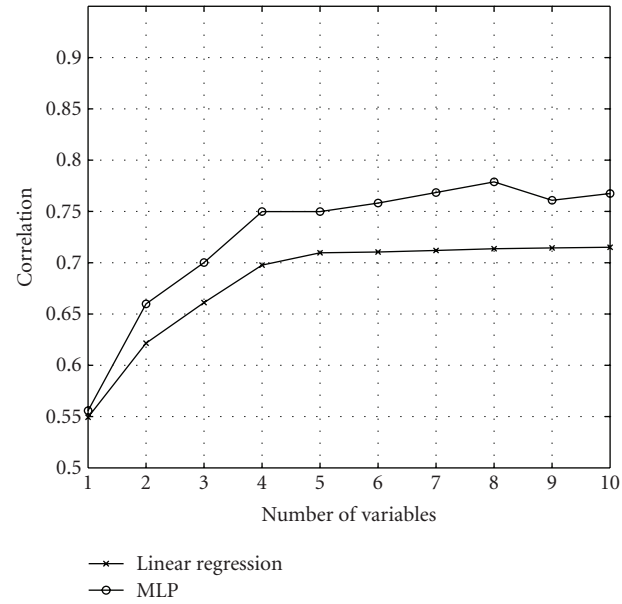

Figure 2: Evolution curves for turbidity. The goodness of the model improves at first as variables are added, but then the improvement gradually stops.

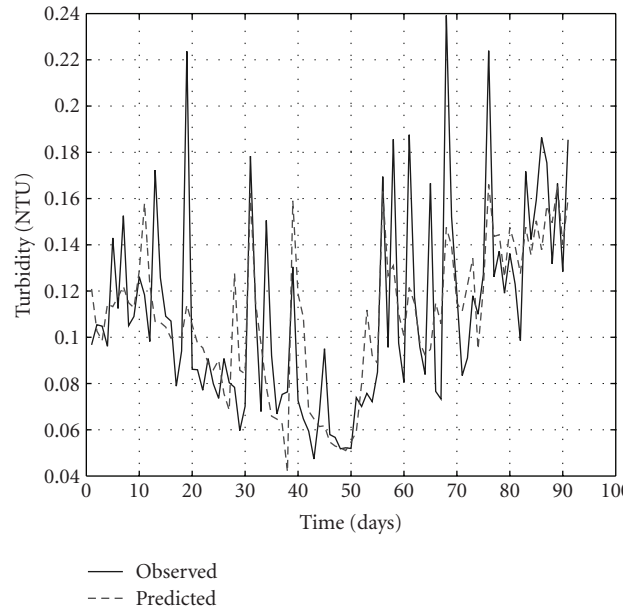

Figure 3: Observed turbidity and the values predicted by the MLR model when using the 8 best variables.

5. Discussion

The quality of drinking water is an important matter, because water of low quality may cause health-related and economic problems which have a considerable impact on people's daily lives. Monitoring and controlling water quality is a challenging task; however, as the quality of water in a treatment process may be affected by numerous factors

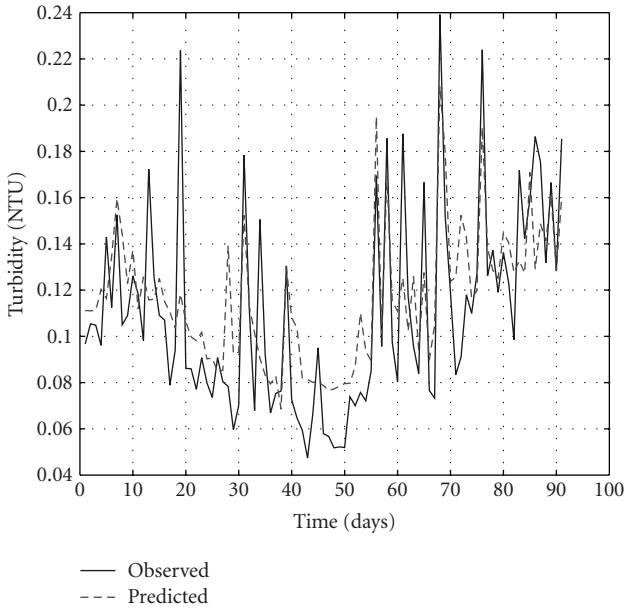

Figure 4: Observed turbidity and the values predicted by the MLP model when using the 8 best variables.

Table 5: Variables selected for residual aluminium using multiple linear regression (MLR).

Round	Variables	Correlation coefficient (C)
1	Water temperature	0.72
2	Aluminium dose/raw water $KMnO_4$	0.79
3	Silicates in raw water	0.81
4	pH of raw water	0.81
5	Hardness of raw water	0.82
6	Flocculation pH	0.82
7	Conductivity of raw water	0.82
8	Total intake of water	0.82
9	Surface level of Lake Humaljärvi	0.82
10	Colour of raw water	0.82

Table 6: Variables selected for residual aluminium using multilayer perceptrons (MLPs).

Round	Variables	Correlation coefficient (C)
1	Water temperature	0.76
2	Aluminium dose/raw water $KMnO_4$	0.86
3	Silicates in raw water	0.88
4	Surface level of Lake Humaljärvi	0.88
5	Hardness of raw water	0.88
6	Turbidity of raw water	0.89
7	Aluminium dose	0.90
8	pH of raw water	0.91
9	$KMnO_4$ of raw water	0.89
10	Aluminium feed	0.91

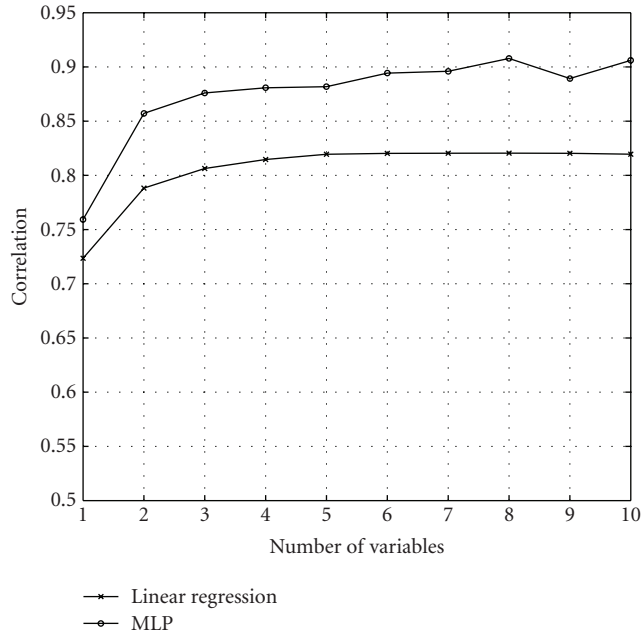

Figure 5: Evolution curves for residual aluminium. The goodness of the model improves at first as variables are added, and then the improvement gradually stops.

which are either not thoroughly known or which have not been verified on an experimental basis. The modelling of water quality has, therefore, become more important in recent years.

Both linear and nonlinear modelling methods were used in this paper to model turbidity and residual aluminium in a water treatment process. The general conclusion is that in both cases the goodness of the nonlinear model was slightly better than that of the linear one, which would indicate that both problems have some nonlinear features. On the other hand, the improvement in the goodness of the model is not great, which seems to suggest that simpler computational methods may be applicable to these problems.

As for turbidity, the results (see Figure 3) show that the linear model is able to predict the generic trend, whereas the majority of the peaks are modelled better by MLP (see Figure 4). It seems, however, that the reasons for some of the sharp peaks remain obscure regardless of the method used. In particular, the F-test test did not show any significant difference between the linear and nonlinear model. Overall, it is reasonable to use the linear method if the objective is only to reveal generic trends and the nonlinear one if the objective is to predict the extreme values as accurately as possible. In addition, MLR is more suitable for applications which require explicit models or fast calculation, for example, in adaptive soft sensors.

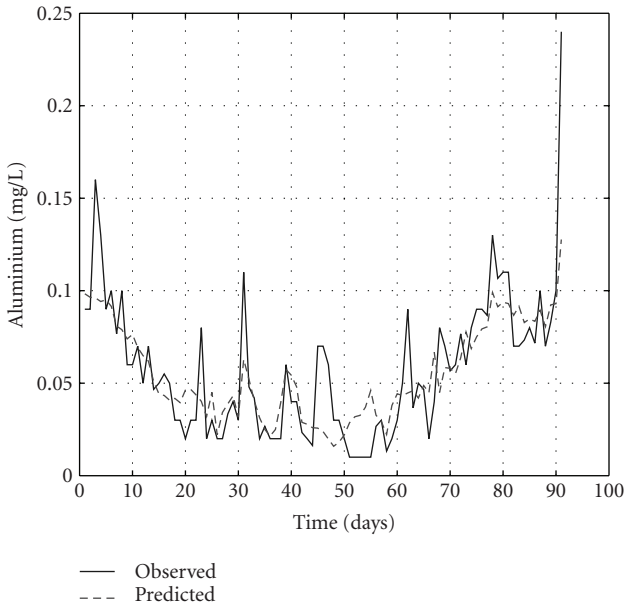

Figure 6: Observed residual aluminium and the values predicted by the MLR model when using the 8 best variables.

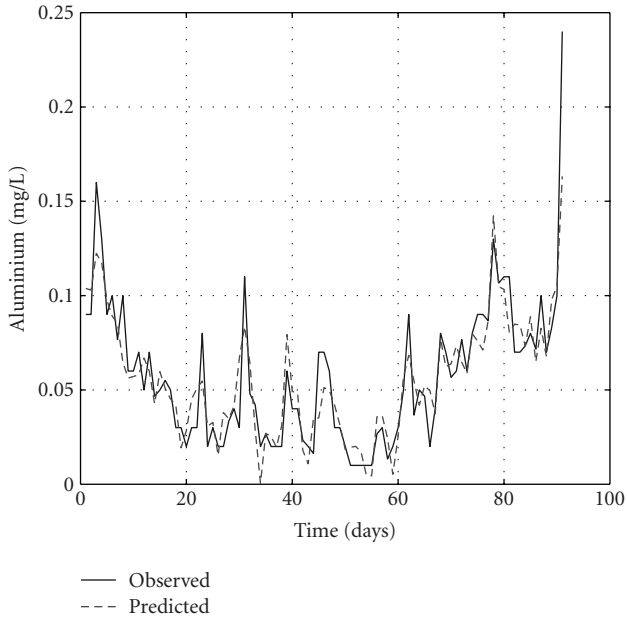

Figure 7: Observed residual aluminium and the values predicted by the MLP model when using the 8 best variables.

Slightly better models can be achieved for residual aluminium (see Table 5). In fact, the fit of the nonlinear model for aluminium ($C \approx 0.9$) is very good. Beside producing more accurate estimates, MLP also seems to be superior to the linear method because it is able to predict both the generic trend and the concentration peaks, whereas the linear method cannot find the reasons for the peaks, as can be seen in Figures 6 and 7. In addition, the F-test showed a statistically significant difference between the models. MLP may, therefore, be regarded as the preferable method for modelling residual aluminium.

The results of variable selection indicate that most of the phenomena behind residual aluminium could be explained with two of the best correlating variables (temperature and Al dose/$KMnO_4$ ratio), whereas aluminium dose, intake from the Pikkala reservoir, and turbidity of raw water were the best variables for explaining the turbidity. In the sense of water chemistry, the most important process parameters are usually Al dose and pH [2]. According to our results, the Al dose (or Al/$KMnO_4$ ratio) is an important variable, because it was selected in the second round of variable selection. This implies that there could be potential for optimizing the dosing of Al, for example, by making a more sophisticated controller for dosing. In contrast, the pH value was not among the most important variables, which implies that the pH value would be already optimized in the process.

Furthermore, it is worth remembering that the selected variables are not necessarily the same as those which have the best correlations with turbidity or residual aluminium (see Tables 1 and 2). This is because the variable selected in each round is always the one that adds most information to the model in the particular round. In other words, the procedure for selecting variables takes multivariate interactions into account, whereas calculating the simple correlations does not.

According to the literature [9, 11, 17], water hydrology and water treatment processes have a nonlinear nature, so nonlinear methods should be used in the modelling. However, the results show no significant difference in the turbidity case, and only a small difference in the residual aluminium case between the linear and nonlinear models. One possible explanation would be phenomena such as seasonality or episodic events affecting the process. In this case such phenomena would be water temperature (strong seasonal dependence) and intake from Pikkala (episodic event). This is supported by the fact that it has been shown earlier that seasonality or episodic events also have a strong influence on the water treatment processes [7, 8]. Sometimes nonlinear models may give better results than linear ones, but a large part of the nonlinearity may arise from the seasonality and/or episodic events and, on the other hand, from the multivariate interactions between the variables connected with these phenomena. In this case, we could find the variables explaining these phenomena using a linear multivariate method, and the resulting model could capture most of this behaviour, so that especially in the case of turbidity the difference between the two methods used was not significant. Thus, nonlinear models are not always needed, although they can somewhat improve the goodness of models.

The results show that the approach used here has several benefits. The approach itself is evidently flexible regardless of the computational method, and it has a high computing power. Trained with real process data, the method is also able to adapt to exceptional situations in the process. In addition, the approach is suitable for cases where the physical processes are not well known or are highly complex. Generally speaking, although the resulting models assimilated good

prediction abilities, they could have been improved later by adding more data samples or some process variables that were not used on this occasion.

Some limitations in the performance of the approach may follow from the fact that the variable selection was implemented by adding variables to the model one by one. This means that all the possible combinatorial effects were not evaluated as they would be in more sophisticated approaches. It is, therefore, possible that there may be two or more variables whose mutual interaction may have a considerable effect on the concentration, although their individual effects on the model may be insignificant. On the other hand, the forward selection method makes variable selection robust with regard to interdependences between the variables. It would certainly be possible to model the influence of all combinations of variables, but in reality the computing time required for that would be long, especially in processes involving a large number of variables, and this would obviously reduce the usability of the method in any real-world process applications. The correlations observed here are nevertheless of the same order as those observed by [9], who did not use variable selection in their approach.

Data-based modelling has been used in various water treatment applications in recent years, for example, for predicting raw water quality parameters, optimal flocculation dosages, or the quality parameters of treated water [11, 17]. This study shows that data-based modelling combined with variable selection provides an efficient tool for analysing specific problems affecting water treatment processes. In addition, it shows that online data and process data can be combined in the same data set for these purposes. Furthermore, there is no need to select the variables to be included in the model manually, because they can be selected during modelling by the procedure of variable selection. In addition, this study shows that nonlinear models are not always necessary if simpler and more explicit linear multivariate models perform well enough.

The method has many potential applications. First of all, regression models can be used as real-time prediction models for estimating water quality parameters. In addition, predictive models can be used for proactive management of the process and for forecasting or evaluating water quality and the risks related to it. Process diagnostics is another potential field of application. The selecting of variables, for example, provides valuable information on the factors affecting water quality. Moreover, it is possible to construct a data-based soft sensor for water quality parameters which could be used further for control or fault detection purposes.

When used off-line, the model can help to evaluate quality and/or risk aspects of water safety in different scenarios, including the accumulation of aluminium in consumers. Using virtual process interfaces, it is possible to obtain a better understanding of these things and to conduct scenario analyses based on process histories. The variable selection technique allows the most efficient variables to be selected for the piloting phase, which will reduce the number of pilot tests. Laboratory tests may still be needed after modelling, but the number of variables can be limited to the most effective ones. Overall, the approach as introduced here provides a simple and economical tool for analysing and optimizing water treatment processes and a fruitful way of investigating interactions between the variables affecting these processes.

6. Conclusions

As drinking water quality guidelines continue to become more stringent, modelling methods which utilize process histories will offer valuable tools for process modelling and control in water treatment plants and provide an alternative to conventional methodologies. Moreover, these modelling techniques allow such utilities to increase their process knowledge and, therefore, facilitate process control. The results are promising as far as the wider use of the data-driven selection of variables and modelling in water treatment processes is concerned, and the approach used here undoubtedly has considerable potential.

Acknowledgments

The writing of this paper was supported by Maa- Ja Vesitekniikan Tuki Ry. The material was produced in the POLARIS project financed by the Finnish Funding Agency for Technology and Innovation (Tekes). The authors gratefully acknowledge this financial support.

References

[1] World Health Organization, *Guidelines for drinking-water quality*, vol. 1, Recommendations, 3rd edition, 2006.

[2] R. D. Letterman, Ed., *Water Quality & Treatment, Handbook of Community Water Supplies*, AWWA, 1999.

[3] A. Campell, "The role of aluminium and copper on neuroinflammation and Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 10, pp. 165–172, 2006.

[4] J. E. Van Benschoten and J. K. Edzwald, "Chemical aspects of coagulation using aluminum salts - I. Hydrolytic reactions of alum and polyaluminum chloride," *Water Research*, vol. 24, no. 12, pp. 1519–1526, 1990.

[5] J. E. Van Benschoten and J. K. Edzwald, "Chemical aspects of coagulation using aluminum salts - II. Coagulation of fulvic acid using alum and polyaluminum chloride," *Water Research*, vol. 24, no. 12, pp. 1527–1535, 1990.

[6] C. Huang and H. Shiu, "Interactions between alum and organics in coagulation," *Colloids and Surfaces A*, vol. 113, no. 1-2, pp. 155–163, 1996.

[7] P. Juntunen, M. Liukkonen, M. Lehtola, and Y. Hiltunen, "Cluster analysis of a water treatment process by self-organizing maps," in *Proceedings of the 8th IWA Symposium on Systems Analysis and Integrated Assessment*, E. Ayesa and I. Rodríquez-Roda, Eds., pp. 553–558, WATERMATEX, 2011.

[8] P. Juntunen, M. Liukkonen, M. Lehtola, and Y. Hiltunen, "Dynamic modelling approach for detecting turbidity in drinking water," in *Proceedings of the 52nd International Conference of Scandinavian Simulation Society*, E. Dahlquist, Ed., 2011.

[9] C. W. Baxter, Q. Zhang, S. J. Stanley, R. Shariff, R. R. T. Tupas, and H. L. Stark, "Drinking water quality and treatment: the

use of artificial neural networks," *Canadian Journal of Civil Engineering*, vol. 28, supplement 1, pp. 26–35, 2001.

[10] D. N. Thomas, S. J. Judd, and N. Fawcett, "Flocculation modelling: a review," *Water Research*, vol. 33, no. 7, pp. 1579–1592, 1999.

[11] H. R. Maier, N. Morgan, and C. W. K. Chow, "Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters," *Environmental Modelling and Software*, vol. 19, no. 5, pp. 485–494, 2004.

[12] S. Haykin, *Neural Networks and Learning Machines*, Pearson Education, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[13] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven Soft Sensors in the process industry," *Computers and Chemical Engineering*, vol. 33, no. 4, pp. 795–814, 2009.

[14] M. R. G. Meireles, P. E. M. Almeida, and M. G. Simões, "A comprehensive review for industrial applicability of artificial neural networks," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 3, pp. 585–601, 2003.

[15] M. Heikkinen, H. Poutiainen, M. Liukkonen, T. Heikkinen, and Y. Hiltunen, "Self-organizing maps in the analysis of an industrial wastewater treatment process," *Mathematics and Computers in Simulation*, vol. 82, no. 3, pp. 450–459, 2011.

[16] A. M. Kalteh, P. Hjorth, and R. Berndtsson, "Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application," *Environmental Modelling and Software*, vol. 23, no. 7, pp. 835–845, 2008.

[17] H. R. Maier and G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications," *Environmental Modelling and Software*, vol. 15, no. 1, pp. 101–124, 2000.

[18] M. S. Gibbs, G. C. Dandy, and H. R. Maier, "Calibration and optimization of the pumping and disinfection of a real water supply system," *Journal of Water Resources Planning and Management*, vol. 136, no. 4, Article ID 023003QWR, pp. 493–501, 2010.

[19] C. W. Baxter, S. J. Stanley, and Q. Zhang, "Development of a full-scale artificial neural network model for the removal of natural organic matter by enhanced coagulation," *Journal of Water Supply: AQUA*, vol. 48, no. 4, pp. 129–136, 1999.

[20] J. L. Giraudel and S. Lek, "A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination," *Ecological Modelling*, vol. 146, no. 1–3, pp. 329–339, 2001.

[21] P. J. Werbos, *Beyond regression: new tools for prediction and analysis in the behavioral sciences*, Doctoral thesis, Harvard University, Cambridge, Mass, USA, 1974.

[22] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[23] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[25] H. Liu and H. Motoda, Eds., *Computational Methods of Feature Selection*, Chapman & Hall, Boca Raton, Fla, USA, 2008.

[26] A. W. Whitney, "Direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, vol. C-20, no. 9, pp. 1100–1103, 1971.

[27] D. J. C. MacKay, "A practical bayesian framework for backpropagation networks," *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.