Dynamic soft sensors for detecting factors affecting turbidity in drinking water

Petri Juntunen, Mika Liukkonen, Markku J. Lehtola and Yrjö Hiltunen

ABSTRACT

Effective monitoring of water quality is critical for water safety. In particular, online monitoring based on modeling is useful in several applications such as process assessment, hazardous event detection or common fault diagnostics in the water processes. Soft sensors have lately established themselves as a good alternative for different tasks of process control such as the acquisition of critical process variables and process monitoring. In this paper, we introduce a dynamic method for predicting turbidity in drinking water. The goals of the work were to construct a dynamic real-time data-driven model to predict the turbidity in treated water and to find the most significant variables affecting turbidity. Both linear and non-linear regression methods are used in modeling. Our results show that the static linear or non-linear model ($r = 0.40$ and $r = 0.52$, respectively) is not able to follow the changes in turbidity, whereas the dynamic method can produce a reasonable estimate for turbidity ($r = 0.75$ for the dynamic linear and $r = 0.86$ for the dynamic non-linear model). In conclusion, the data analysis procedure seems to provide an efficient means of modeling the water treatment process online and of defining the most affecting variables.

Key words | dynamic modeling, multi-layer perceptrons, multi-linear regression, turbidity, variable selection, water treatment

Petri Juntunen (corresponding author)
Mika Liukkonen
Markku J. Lehtola
Yrjö Hiltunen
Department of Environmental Science,
University of Eastern Finland,
P.O. Box 1627,
70211 Kuopio,
Finland
E-mail: *petri.juntunen@uef.fi*

INTRODUCTION

Water quality is becoming an ever more important issue, as water of low quality causes many significant problems. In particular, a wide range of microbial and chemical constituents of drinking water can cause detrimental health effects, either in acute or chronic form (WHO 2006). On the other hand, water of bad quality can also be harmful from an economical perspective, as resources have to be directed to clear up the water supply system every time a problem occurs.

For these reasons, there is a growing pressure to improve water treatment and water quality management by online monitoring to ensure safe drinking water at a reasonable cost. In water treatment, as in many domains, process monitoring and control relies heavily on accurate and reliable sensor information. Whereas many process parameters can be measured continuously using relatively simple and cheap physical sensors, the determination of certain quantities of interest requires costly laboratory analyses which cannot be performed online (Valentin & Denoeux 2001).

Turbidity is a common water quality parameter, the purpose of which is to measure impurities in water. In a physical sense, turbidity is the reduction of clarity in water due to the presence of suspended or colloidal particles and is commonly used as an indicator for the general condition of drinking water (WHO 2006). Furthermore, turbidity has been used for many decades as an indicator of the efficiency of drinking water coagulation and filtration processes. For this reason, turbidity is an important operational parameter. High turbidity values refer to poor disinfection ability and possible fouling problems in the distribution network, for example, so turbidity should be minimized (Letterman 1999). However, turbidity is a sensible measurement and is affected by a large number of variables and phenomena, making turbidity challenging for modeling purposes.

doi: 10.2166/hydro.2012.052

The real-time monitoring network for water quality has been operational for the protection of water resources and detection of water quality. It is useful for management plans of water utility and local authorities to realize the changing characteristics of water quality. Real-time monitoring, characterized by a rapid response time, full compatibility with automation, sufficient sensitivity, high rate of sampling and minimal requirements for skill and training (Mays 2004), can provide data for multiple purposes such as environmental hazard assessment, water resources management and water hazard warning.

These challenges have created a need for new methods such as *soft sensors*, which could be used to support or compensate the direct measurements of water treatment, for example. A soft sensor consists of a combination of measurement signals and a model that can be used to form an estimate that matches a direct measurement. Soft sensors can therefore provide a tool for supporting or replacing the potentially difficult and expensive measurements. Soft sensors offer a valuable tool for the industry and have several advantages (Fortuna *et al.* 2007; Kadlec *et al.* 2009; Ma *et al.* 2009; Liukkonen *et al.* 2012):

- a low-cost alternative to expensive hardware sensors;
- can operate in parallel with hardware sensors, offering a back-up or decision support system (fault detection and diagnostics);
- can be implemented on existing hardware (e.g. to control the disinfection doses); and
- allow online estimation of data (e.g. to estimate risks of water safety, or acquisition of critical process variables).

In soft sensor development we are confronting the same challenges as in the modeling of water treatment processes, however. Water treatment processes are considered physically and chemically heterogeneous, as many features of raw water and water processes affect its performance (van Benschoten & Edzwald 1990a, b; Huang & Shiu 1996). For this reason, the parameters of raw water and water processes are generally regarded as complex and their mutual interactions non-linear (Baxter *et al.* 1999). Furthermore, successful applications of traditional mechanistic models are limited to idealized, artificial systems (Thomas *et al.* 1999). In real processes, the correlation between simulated and experimental data has been poor, but expensive *in situ* testing is needed (Thomas *et al.* 1999; Maier & Dandy 2000; Baxter *et al.* 2001; Maier *et al.* 2004).

Moreover, in water treatment there are observable cycles 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. Annual 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 (Bratby 2006; Juntunen *et al.* 2010a). In addition to cyclic behavior, episodic events such as rapid, previously unseen changes of lime feed or pH may cause sudden changes in turbidity (Juntunen *et al.* 2010b). Moreover, the phenomena existing in the process are state dependent, meaning that they work differently under different process conditions (Juntunen *et al.* 2011). For these reasons, process dynamics is a typical problem in data-based systems, and successful modeling often requires an ability to adapt to changing conditions. Particularly in water treatment, the behavior of the process can depend heavily on the state of the process, for which adaptivity is generally required from applications.

Multivariate analysis methods such as factor analysis and principal component analysis (PCA) have been widely used in analyzing hydrological systems (Giraudel & Lek 2001; Lee *et al.* 2006). However, the limitations of conventional multivariate statistical methods arising from the challenges mentioned earlier are also known (Giraudel & Lek 2001). Artificial neural network (ANN) techniques are a powerful tool for multivariate and non-linear analysis, and offer an alternative to traditional statistical methods for optimal monitoring and determination of dynamic systems (Hong *et al.* 1998). ANNs are successfully exploited in modeling hydraulic and hydrologic phenomena within various tasks such as river basin management (Solomatine & Ostfeld 2008), rainfall–runoff prediction (Hettiarachchi *et al.* 2005) and simulation of water networks (Martinez *et al.* 2008; Tabesh *et al.* 2009). There are also water treatment applications in which multivariate analysis methods are utilized in assessing the performance of water treatment. For example, Bieroza *et al.* (2012) have analyzed decomposed fluorescence data using several multivariate methods, and Ricardo *et al.* (2011) have used a Projection to Latent Structures (PLS) method to model a single unit process in the pilot scale.

An important aspect to bear in mind is that, regardless of the technique applied, the performance of a data-driven method depends on the quality of the data used. A model based on insufficient data or data of bad quality is not necessarily able to predict the output reliably and precisely. Since neural networks are empirical models, problems can also occur if the data samples do not reflect the entire range of practicable machine operation or do not even include the optimal process settings. In other words, if the advantages of a data-driven method are to be fully exploited, careful and systematic acquisition of process data must be implemented first.

In this paper, we present a novel approach for predicting turbidity of treated water. The approach is based on a dynamic soft sensor, which utilizes multivariate regression in computation. Because process data typically consist of a large number of variables, a group of them is selected adaptively before a dynamic predictive model is created; this is then used in estimating the degree of turbidity in the future. The results of the dynamic model are compared with the traditional (static) multiple linear regression (MLR) and multi-layer perceptron (MLP) models.

MATERIALS AND METHODS

Static versus dynamic models

In a static data-driven model, the whole dataset (except the validation dataset) is used in training the model; in a dynamic model, only the latest part of the dataset is utilized by moving the window forward after each modeling round. In ideal conditions, static models may be adequate for modeling an industrial process. Process dynamics is a typical problem in data-based systems however, and it has been shown that, in some applications, adaptive models perform better (Sbarbaro *et al.* 2008; Ma *et al.* 2009; Liukkonen *et al.* 2012).

Historically, the first soft sensors relied on offline modeling based on recorded historical data. In such a case, the collected historical recordings are used to build a static data-driven model. However, there are some challenges for static soft sensors in representing industrial processes. First of all, the historical data should contain all possible future states and conditions of the process, such as the states in which the process can be operated or the states related to environmental changes (Kadlec *et al.* 2011). In addition, these diverse process states may involve totally different behavior (Juntunen *et al.* 2011), which leads to difficulties in constructing a generic and robust model which could represent all different conditions.

Furthermore, most of the processes are characterized by time-dependent behavior and thus require a strategy for online adaptation (Kadlec *et al.* 2011), which is very difficult to estimate during the model design phase. In particular, this is a challenge for processes such as water treatment which are sensitive to environmental effects. For these reasons, the performance of static models typically deteriorates during their online operation and therefore adaptive soft sensing techniques are needed (Kadlec *et al.* 2009).

MLR

In MLR (Cohen 1968), the purpose is to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data samples. In principle, the MLR model with observations and variables is:

$$Y = e + a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n \quad (1)$$

where Y is the value of the response variable; X is the value of the predictor (explanatory) variable; $a_{0,\ldots,n}$ is the unknown coefficient to be estimated; and e signifies the uncontrolled factors and experimental errors of the model. The fitting works 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 also called least-squares fitting.

MLP

MLPs are well-known feed-forward neural networks (Meireles *et al.* 2003; Haykin 2009) 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. A MLP network is trained with data samples, which leads 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 dataset.

MLP neural networks must be trained for each problem separately. A popular MLP training technique is the back-propagation algorithm (Werbos 1974), in which the output values are compared with the correct answer from the original data in order to calculate a 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 (Haykin 2009):

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.

The ANN consisted of the process parameters as inputs, one hidden layer with seven 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 (Mackay 1992) was exploited in training, and the mean squared error (*mse*) 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. Seventy percent of the data were selected for training, 15% for independent validation and 15% for testing. In variable selection, the training and validation data were selected using cross-validation; in the soft sensor method validation of the results was carried out in a continuous series at the end of the training period (Mathworks 1998).

Variable selection

The enormously increased amount of information available in recent years has caused the selection of variables, or reduction of model inputs, to become a relevant part of data analysis (Blum & Langley 1997; Jain *et al.* 2000; Guyon & Elisseeff 2003; Liu & Motoda 2008). The objective of this selection is to improve the prediction performance of the model, to provide faster processing of the data and to provide a better understanding of the process (Guyon & Elisseeff 2003). For instance, when exploiting ANNs in computation, 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 (Whitney 1971). One of these is the *sequential forward selection* method, which was used for the selection of variables in this case.

In the sequential forward selection method, 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.

Dynamic soft sensor

We have developed a dynamic approach to predict the turbidity of treated water, which can be utilized in an online application. Because process data typically consist of a

large number of variables, a group is adaptively selected before a dynamic predictive model is created. The method uses N previous values of turbidity with corresponding process data and p adaptively selected variables to create a p-dimensional regression model that can be used to predict the turbidity in the future. Multivariate regression is used in creating the model. The stages of the algorithm are presented in Figure 1.

Two different values of N were used: one for the variable selection and another for the final soft sensor. The values of N were selected using a trial-and-error approach, in which prior knowledge of the process conditions such as day and week cycles was utilized to get the approximate initial values for the search. In this study, $N = 168$ in variable selection and $N = 24$ in the prediction model (soft sensor). In other words, the group of variables used by the soft sensor is updated once a week and the soft sensor model is updated daily. The value of the parameter p, i.e. the number of variables, was set to 5 because it seemed to yield the best results in this case.

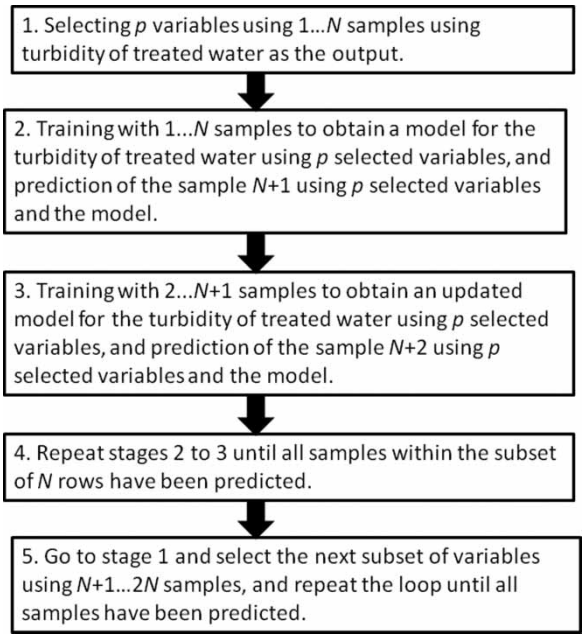

Figure 1 | The flowchart of the dynamic data-driven soft sensor.

RESULTS

Case study: Itkonniemi water treatment plant

Itkonniemi plant is one of the water treatment plants of Kuopio Waterworks (see Figure 2). In the plant water is first bank filtrated. After filtration, water is purified with physical and chemical oxidization processes followed by chemical coagulation in a chemical purification process. After coagulation, the coagulated floc is separated by sedimentation or flotation followed by sand filtration. In the final stage, water is disinfected by chlorination. The separation process is divided in three parallel working sections (lines 500, 600 and 700). The purification process can also use water from Lake Kallavesi as a raw water source, which has a large effect on the process. Waters produced in Itkonniemi and another plant (Jänneniemi) are mixed in Itkonniemi waterworks and are disinfected by chlorination.

The data include process and laboratory measurements from a period of 1,017 days with a resolution of 1 hour. The total number of variables is 64 (see Table 1).

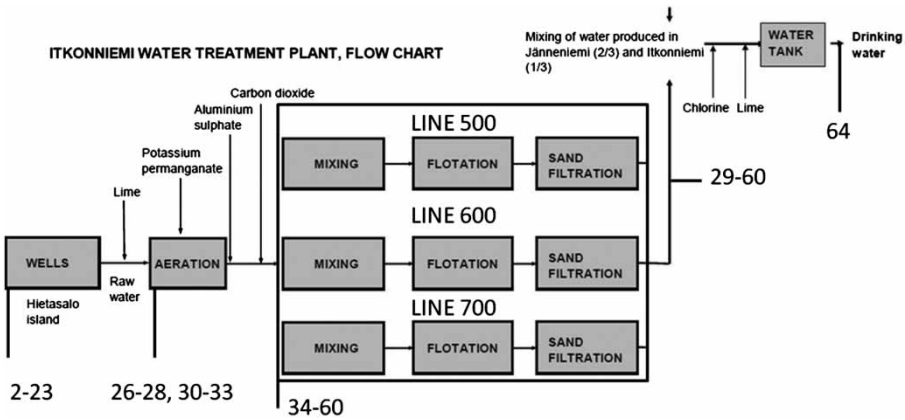

Figure 2 | Flow chart of the drinking water purification process in Itkonniemi. The numbers refer to the measuring points for the variables (see Table 1).

Table 1 | The data variables used in modeling and their calculated lags in hours

Variable no.	Variable	Units	Lag	Description
1	Hour of day		0	Hours from 1 to 24
2–12	Well 1–11, surface	m	−5	The surface level in ground water wells 1–11
13–22	Flow from well 1–9	$m^3\ h^{-1}$	−5	Total water flow from wells 1–9
23	Lake water used	$m^3\ h^{-1}$	−3	Total flow from the lake
24	Level of the lake	m	−3	The surface level of the lake
25	Flow to aeration	$m^3\ h^{-1}$	−3	Total water flow to aeration
26	Solvent water	$m^3\ h^{-1}$	−3	The flow of solvent water
27	Pre lime feed	$m^3\ h^{-1}$	−3	The rate of pre lime feed
28	Start lime feed	$m^3\ h^{-1}$	−3	The rate of start lime feed
29	Post lime feed	$m^3\ h^{-1}$	0	The rate of post lime feed
30	CO_2 feed	kg h^{-1}	−3	The rate of CO_2 dosage
31	Aeration temperature	°C	−3	Temperature of water in aeration
32	Aeration, pH		−3	pH of water in aeration
33	pH after lime feed		−3	pH of water after lime feed
34	Line 500, flow	$m^3\ h^{-1}$	0	Total water flow to line 500
35	Line 500, level in sedimentation	m	0	Water level in the sedimentation phase of line 500
36–39	Line 500, 1–4. Filtration pressure	m	0	Filtration pressures of filters 1–4 on line 500
40–42	Line 600, 1–3. Flow	$m^3\ h^{-1}$	0	Total water flow to line 600
43–44	Line 600, 1–3. Air flow	$m^3\ h^{-1}$	0	Air flow to aeration on line 600
45–50	Line 600, 1–6. Filtration pressure, m	m	0	Filtration pressures of filters 1–6 on line 600
51	Line 700, 1 & 2 flow	$m^3\ h^{-1}$	0	
52	Line 700, 3 & 4 flow	$m^3\ h^{-1}$	0	
53–56	Line 700, 1–4. Air flow	$m^3\ h^{-1}$	0	Air flow to aeration on line 700
57–60	Line 700, 1–4 Filtration pressure	m	0	Filtration pressures of filters 1–4 on line 700
61	Filter wash flow	$m^3\ h^{-1}$	0	Total flow to filter wash
62	Workday/weekend	1/0	0	Mon–Fri = 1, Sat–Sun = 0
63	Temperature of air	°C	0	Outside temperature
64	Flow to consumption	$m^3\ h^{-1}$	0	Total flow of treated water to consumption

Twenty-two of the variables are hydrological and quality parameters of raw water; the following 38 variables are online measurement data from the process and the rest of the variables are different time variables.

Process lags, or *delays*, can be determined using a cross-correlation method in which the correlations between variables are calculated in a time window. The correlation coefficients between two variables are calculated in each time step and the maximum absolute value of these coefficients represents the process lag between these two variables. The lags between individual variables constitute a matrix, which can be used in determining the final lags with respect to each variable. The final process lags are determined by the lags indicated by the most considerable correlations. In this case, the lags were defined twofold by first calculating them based on the physical dimensions of unit operations and then comparing the results to those from cross-correlation calculations. The embedded lags varied over 0–5 hours.

Applications

At the first stage, the turbidity of treated water was modeled using the measurement data and the static linear and non-linear methods. At the second stage, the turbidity was modeled using the dynamic linear and non-linear methods. The five input variables were selected for the models by the forward selection procedure.

The main results (correlations and root mean square errors) are shown in Table 2. Using the dynamic linear models, the average correlation between the observed values

Table 2 | The correlations and RMSE values between the observed values of turbidity and those estimated by the different models

	Linear		Non-linear	
	Correlation	RMSE	Correlation	RMSE
Static	0.40	0.34	0.52	0.26
Dynamic	0.76	0. 30	0.86	0.18

of turbidity and those estimated by the dynamic model was 0.75. By using the static model, we achieved a correlation of 0.40. Using the dynamic non-linear models, the average correlation between the observed values of turbidity and those estimated by the dynamic model was 0.86, whereas using a static model we achieved a correlation of 0.52.

A sample of the measured values and those estimated by different methods are shown in Figure 3. In the modeling and variable selection phase, the process data were used for producing dynamic predictive models for turbidity. The variables selected for the static model and those selected most frequently for the dynamic models are shown in Table 3.

In Figures 4 and 5, the features of the dynamic models are further described on a weekly basis. In these figures, the occurrences of the most frequently selected variables and the correlation between the model and the measured turbidity are shown. The value of correlation is the overall correlation between the measured and estimated values of turbidity. In other words, the value of the variable is 1 if it has been selected to the respective model and 0 if it has not been selected. Furthermore, the weekly average value of turbidity is shown.

DISCUSSION

In this study we have constructed soft sensors for predicting the turbidity of treated water. The sensors have static/dynamic and linear/non-linear features combined with variable selection. Our results show that the static models are not able to follow the changes in turbidity ($r = 0.4$ and $r = 0.5$ for the linear and non-linear models, respectively), whereas the adaptive model can produce a reasonable estimate for it ($r = 0.75$ and $r = 0.86$ for the linear and non-linear models, respectively). This dynamic behavior is probably due to the cyclic

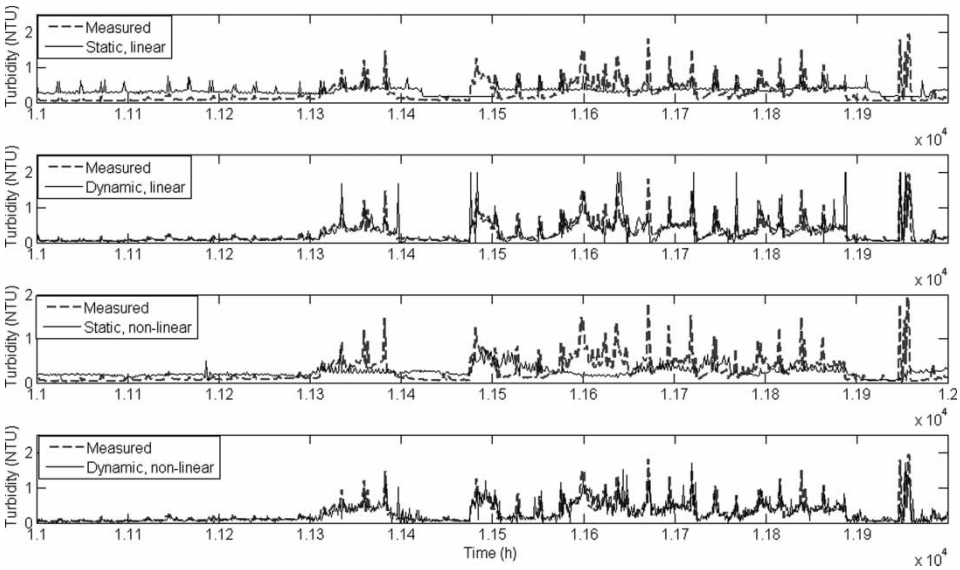

Figure 3 | The results of the linear and non-linear models. A sample period (c. 1 month) of the modeled and measured turbidity.

Table 3 | The variables selected by each approach. For dynamic models, the variables selected most frequently are presented

Selection round	Static model, linear	Dynamic model, linear	Static model, non-linear	Dynamic model, non-linear
1.	Solvent water feed	Filter wash flow	Solvent water feed	Filter wash flow
2.	Filter wash flow	Line 600 1, filtration pressure	pH in aeration	Hour of day
3.	Line 700, filtration pressure	Line 500, level in sedimentation	Line 700, filter pressure 2	Line 600, filter pressure 2
4.	Well 4, surface	Line 600 6, filtration pressure	Flow to the process	Line 600, filter pressure 1
5.	Post lime feed	pH after lime feed	Flow to consumption	Line 600, filter pressure 4

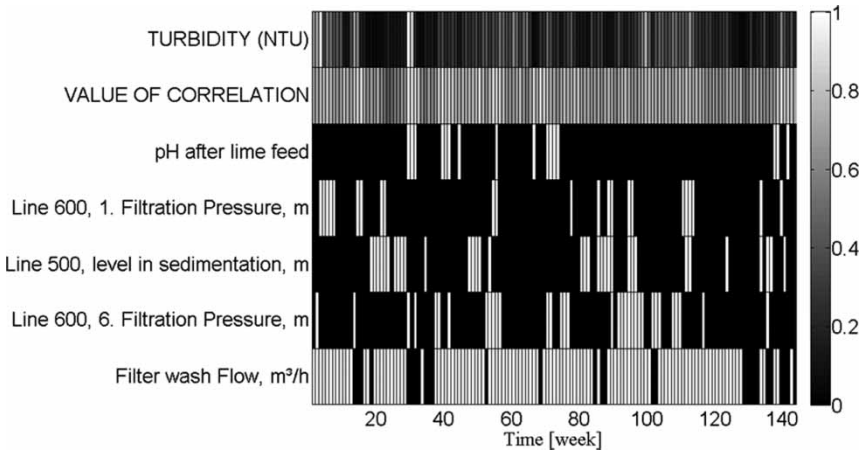

Figure 4 | The presence of the most frequently selected variables in the dynamic linear models, average correlations and values of turbidity. The value of the variable is 1 if it has been selected to the respective model, and 0 if it has not been selected.

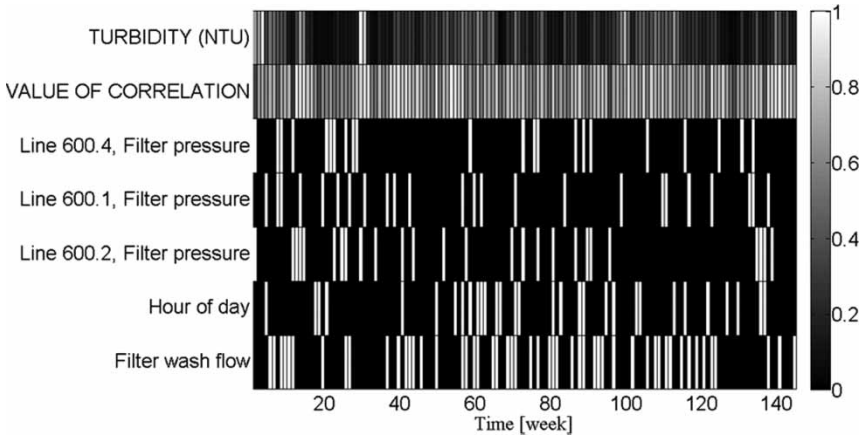

Figure 5 | The presence of the most frequently selected variables in the dynamic non-linear model, average correlations and values of turbidity. The value of the variable is 1 if it has been selected to the respective model, and 0 if it has not been selected.

and/or state-specific behavior of the process, which can also be seen in several process variables.

The accuracies of both dynamic models are acceptable. The prediction accuracy for the baseline of turbidity is especially good, although the reasons for the sharp concentration peaks cannot be found. However, turbidity is sensitive to disturbances such as air bubbles and solid particles, the presence of which cannot be estimated. The most important variables were the same using both methods, indicating good reliability of the results. There seem to be some periods in which the models do not work so well however, which reduces the overall goodness of the estimate produced by the dynamic models. In these periods, the correlation between the measured and the estimated values can be as low as 0.4; the correlation is generally 0.7–0.9. One possible reason for this phenomenon is the limited accuracy of the turbidity meter when turbidity is <0.1 NTU. The precise information on the high values is more valuable however, because they indicate the quality issues in the process. In conclusion, this study indicates that the models produced for processes which have cyclic and/or episodic behavior should have dynamic elements in the models.

The results show that the non-linear models yield better results than the linear models, which is an expected result because some phenomena of the water treatment processes usually have non-linear elements (Maier *et al.* 2004). For this

reason, experiments with non-linear regression methods are likely to increase the performance of the dynamic models. However, as a faster and more robust method, a linear model can for example be used for data pre-handling purposes.

The soft sensor approach has two clear benefits: first, the selected variables inform the process personnel of the most affecting variables in the very moment, which can help to increase the knowledge of the process and the problems, for example. This knowledge can be exploited in offline (decision making, optimization) or online (quality monitoring, operating the process, adjusting the disinfection level, etc.) processes.

Secondly, the soft sensor can predict the future turbidity of the process which can be used in decision making for operating the process. For example, in Figure 4 we can see a 3-week episode around week 30 with an exceptionally high level of turbidity. In the same period we can see that 'pH after lime feed' variable has a strong tendency to be selected. Further analysis showed that the pH level was exceptionally high during that period, which implicates an un-optimal pH control during that time.

These kinds of features provide several potential applications concerning process diagnostics, online water quality monitoring and optimizing the operation of the process. Furthermore, the soft sensor could be used as a tool for online water quality monitoring. The benefits of the applications would be, for example, to obtain improved knowledge of short-term changes online, feed-forward control the post-processes of water treatment such as disinfection or optimize the sampling of water.

Because the turbidity of the treated water indicates the general quality of the water, soft sensors can give valuable information of the water quality in the near future. This is important knowledge when we are constructing a real-time water-quality monitoring system at a water treatment plant. It is possible to estimate the current (or the near future) quality of water, risk level of the process or water safety in a general level, saving time for pro-active operations or for additional sampling, for example. Furthermore, as process diagnostic tools, the soft sensors can support the existing measurements. The output of the soft sensors could be compared with the direct measurements, which could indicate potential faults in the sensors or abnormal situations in the process.

One of the potential functions of the soft sensors is to *replace* a real sensor with a soft sensor. For our dynamic approach, this kind of application is challenging because the method needs calibration to some extent (either sample-based calibration or sensor-based calibration periods). The challenge is to minimize the number and duration of those calibration periods and to gather maximum information for the model at same time, questions of sensitivity, identifiability and uncertainty analysis methods (Denis-Vidala & Joly-Blanchard 2004). Research into mechanistic models for water quality modeling (Freni *et al.* 2011) and data-based models such as sensitivity analysis methods for the MLPs (Lee & Hsiung 2009; Yeh & Cheng 2010) has been conducted. It should be emphasized that, in this case, the main purpose was to create an online application which would be robust and fast and which would provide satisfactory accuracy for the estimation of turbidity. These are interesting issues for further research.

An important aspect to bear in mind is that, regardless of the technique applied, the performance of a data-driven method depends on the quality of the data used. A model based on insufficient or poor-quality data is not necessarily able to predict the output reliably and precisely. Since neural networks are empirical models, problems can also occur if the data samples do not reflect the entire range of practicable machine operation or do not even include the optimal process settings. In other words, if the advantages of a data-driven method are desired to be fully exploited, careful and systematic acquisition of process data must be implemented first.

In summary, the results presented in this paper show that the developed adaptive soft sensor is an efficient way of monitoring the performance of a water treatment process online and provides a useful tool for process diagnostics. It is also possible to utilize the soft sensor as an indicator for diagnosing the process more thoroughly in problematic situations, in the form of an early warning system for example. In addition, the approach can provide more in-depth knowledge for the process experts, who can use it for diagnosing the process and making decisions concerning the dosing of chemicals for instance.

CONCLUSIONS

Due to the complex and dynamic character of water treatment processes, it may be difficult to create static models for water

quality applications. In this paper, we have presented a novel dynamic computational approach for predicting the turbidity of treated water. Using a case study, we have demonstrated that dynamic models seem to provide a good platform for data-driven soft sensors in the water treatment process.

As our experiments suggest, the turbidity of treated water can be estimated using the dynamic soft sensors presented, shown by the goodness of the estimate ($r = 0.75$ for linear and $r = 0.86$ for non-linear dynamic model). Since turbidity is one of the key variables for monitoring water quality, the estimated turbidity can provide an indirect means of estimating water quality in water treatment. Especially high turbidity values can be estimated well, which is useful when considering an early warning system for increasing turbidity, for example. The non-linear model ($r = 0.86$) yields a more accurate estimate for turbidity than the linear model ($r = 0.75$), which suggests that the process involves non-linear elements. The dynamic model yields a more accurate estimate for turbidity than the static model, which can be seen in both linear and non-linear models.

In summary, the soft sensor could provide a basis for numerous applications such as online water quality monitoring, predictive modeling of turbidity/water quality, process diagnostics and acquisition of expert knowledge, early warning systems (e.g. indicator for possible problems in treatment) and fault detection.

ACKNOWLEDGEMENTS

The writing of this paper was supported by Maa- ja Vesitekniikan tuki Ry. The material on which it is based was produced in the POLARIS project financed by the Finnish Funding Agency for Technology and Innovation (Tekes), which the authors thank for its financial support.

REFERENCES

Baxter, C. W., Stanley, S. J. & Zhang, Q. 1999 Development of a fullscale artificial neural network model for the removal of natural organic matter by enhanced coagulation. *Journal of Water Supply: Research and Technology - Aqua* **48** (4), 129–136.

Baxter, C. W., Zhang, Q., Stanley, S. J., Shariff, R., Tupas, R.-R. T. & Stark, H. L. 2001 Drinking water quality and treatment: the use of artificial neural networks. *Canadian Journal of Civil Engineering* **28** (Suppl. 1), 26–35.

Bieroza, M., Baker, A. & Bridgeman, J. 2012 New data mining and calibration approaches to the assessment of water treatment efficiency. *Advances in Engineering Software* **44**, 126–135.

Blum, A. L. & Langley, P. 1997 Selection of relevant features and examples in machine learning. *Artificial Intelligence* **97**, 245–271.

Bratby, J. 2006 *Coagulation and Flocculation in Water and Wastewater Treatment*. IWA Publishing, Cornwall.

Cohen, J. 1968 Multiple regression as a general data-analytic system. *Psychological Bulletin* **70**, 426–443.

Denis-Vidala, L. & Joly-Blanchard, G. 2004 Equivalence and identifiability analysis of uncontrolled nonlinear dynamical systems. *Automatica* **40**, 287–292.

Fortuna, L., Graziani, S., Rizzo, A. & Xibilia, M. 2007 *Soft Sensors for Monitoring and Control of Industrial Processes*. Springer-Verlag, London.

Freni, G., Mannina, G. & Viviani, G. 2011 Assessment of the integrated urban water quality model complexity through identifiability analysis. *Water Research* **45**, 37–50.

Giraudel, J. L. & Lek, S. 2001 A comparison of self-organising map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling* **146**, 329–339.

Guyon, I. & Elisseeff, A. 2003 An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157–1182.

Haykin, S. 2009 *Neural Networks and Learning Machines*, 3rd edn. Pearson Education Inc., Upper Saddle River, New Jersey.

Hettiarachchi, P., Hall, M. J. & Minns, A. W. 2005 The extrapolation of artificial neural networks for the modelling of rainfall–runoff relationships. *Journal of Hydroinformatics* **7** (4), 291–296.

Hong, Y. S., Bhamidimarri, S. & Charleson, T. 1998 A genetic adapted neural network analysis of performance of the nutrient removal plant at Rotorua. In: *Proceedings of Institute of Professional Engineers New Zealand (IPENZ) Annual Conference*. Simulation and Control Section, Wellington, New Zealand, vol 2, p. 43.

Huang, C. & Shiu, H. 1996 Interactions between alum and organics in coagulation. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **113**, 155–163.

Jain, A. K., Duin, R. P. W. & Mao, J. 2000 Statistical pattern recognition: a review. *IEEE Transactions of Pattern Analysis and Machine Intelligence* **22**, 4–37.

Juntunen, P., Liukkonen, M., Pelo, M., Lehtola, M. & Hiltunen, Y. 2010a Modeling of residual aluminium. In: *Proceedings of the 7th EUROSIM Congress on Modelling and Simulation Vol. 2*. Czech Technical University, Prague.

Juntunen, P., Liukkonen, M., Pelo, M., Lehtola, M. & Hiltunen, Y. 2010b Modeling of turbidity in water treatment processes. In: *Proceedings of the 51st SIMS Congress on Modelling and Simulation*. Finnish Society of Automation, Helsinki.

Juntunen, P., Liukkonen, M., Pelo, M., Lehtola, M. & Hiltunen, Y. 2011 Cluster analysis of a water treatment process by self-organizing maps. *Watermatex 2011: Conference Proceedings*. IWA, London.

Kadlec, P., Gabrys, B. & Strandt, S. 2009 Data-driven soft sensors in the process industry. *Computers and Chemical Engineering* **33**, 795–814.

Kadlec, P., Grbic, R. & Gabrysa, C. 2011 Review of adaptation mechanisms for data-driven soft sensors. *Computers and Chemical Engineering* **35**, 1–24.

Lee, C.-J. & Hsiung, T.-K. 2009 Sensitivity analysis on a multilayer perceptron model for recognizing liquefaction cases. *Computers and Geotechnics* **36**, 1157–1163.

Lee, W. S., Kwon, Y. S., Yoo, J. C., Song, M. Y. & Chon, T. S. 2006 Multivariate analysis and self-organising mapping applied to analysis of nest-site selection in Black-tailed gulls. *Ecological Modelling* **193**, 602–614.

Letterman, R. D. (ed.) 1999 *Water Quality & Treatment, Handbook of Community Water Supplies*. AWWA, Denver, CO.

Liu, H. & Motoda, H. (eds) 2008 *Computational Methods of Feature Selection*. Chapman & Hall, USA.

Liukkonen, M., Hälikkä, E., Hiltunen, T. & Hiltunen, Y. 2012 Dynamic soft sensors for NO_x emissions in a circulating fluidized bed boiler. *Applied Energy* **97**, 483–490.

Ma, M.-D., Ko, J.-W., Wang, S.-J., Wu, M.-F., Jang, S.-S., Shieh, S.-S. & Wong, D. S.-H. 2009 Development of adaptive soft sensor based on statistical identification of key variables. *Control Engineering Practice* **17**, 1026–1034.

MacKay, D. J. C. 1992 A practical Bayesian framework for backpropagation networks. *Neural Computation* **4** (3), 448–472.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. *Environmental Modeling & Software* **15** (1), 101–124.

Maier, H. R., Morgan, N. & Chow, C. W. K. 2004 Use of artificial neural networks for predicting optimal alum doses and optimal water quality parameters. *Environmental Modeling and Software* **15** (5), 105–124.

Martinez, F., Hernandez, V., Alonso, J., Rao, Z. & Alvisi, S. 2008 Optimizing the operation of the Valencia water distribution network. *Journal of Hydroinformatics* **9** (1), 65–78.

Mathworks 1998 *Using Matlab Version 5*. The Mathworks Inc., Natick.

Mays, L. 2004 *Water Supply Systems Security*. The McGraw-Hill Companies, CA, USA.

Meireles, M. R. G., Almeida, P. E. M. & Simões, M. G. 2003 A comprehensive review for industrial applicability of artificial neural networks. *IEEE Transactions of Industrial Electronics* **50**, 585–601.

Ricardo, A., Oliveira, R., Velizarov, S., Reis, M. & Crespo, J. 2011 Multivariate statistical modelling of mass transfer in a membrane-supported biofilm reactor. *Process Biochemistry* **46**, 1981–1992.

Sbarbaro, D., Ascencio, P., Espinoza, P., Mujica, F. & Cortes, G. 2008 Adaptive soft-sensors for on-line particle size estimation in wet grinding circuits. *Control Engineering Practice* **16**, 171–178.

Solomatine, A. & Ostfeld, A. 2008 Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics* **10** (1), 3–22.

Tabesh, M., Soltani, J., Farmani, R. & Savic, D. 2009 Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. *Journal of Hydroinformatics* **11** (1), 1–17.

Thomas, D. N., Judd, S. J. & Fawcett, N. 1999 Flocculation modeling: a review. *Water Research* **33** (7), 1579–1579.

Valentin, V. & Denoeux, T. 2001 A neural based software sensor for coagulation control in a water treatment plant. *Intelligent Data Analysis* **5**, 23–39.

van Benschoten, J. E. & Edzwald, J. K. 1990a Chemical aspects of coagulation using aluminium salts-I. Hydrolytic reactions of alum and polyaluminium chloride. *Water Research* **24** (12), 1519–1526.

van Benschoten, J. E. & Edzwald, J. K. 1990b Chemical aspects of coagulation using aluminium salts-II. Coagulation of fulvic acid using alum and polyaluminium chloride. *Water Research* **24** (12), 1527–1535.

Werbos, P. J. 1974 Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Doctoral Thesis, Harvard University, Cambridge, MA.

Whitney, A. W. 1971 A direct method of nonparametric measurement selection. *IEEE Transactions on Computers* **C-20**, 1100–1103.

World Health Organization 2006 Guidelines for drinking-water quality. Available from: http://www.who.int/water_sanitation_health/dwq/gdwq3rev/en/. (Accessed October 2012).

Yeh, I.-C. & Cheng, W. L. 2010 First and second order sensitivity analysis of MLP. *Neurocomputing* **73**, 2225–2233.

First received 16 March 2012; accepted in revised form 17 August 2012. Available online 17 November 2012