Utilizing variable selection methods in modelling potable water quality
Jani Tomperi and Kauko Leiviskä

ABSTRACT

Predictive models are potential tools for assessing changes in a water treatment process and water quality, and to be used as assistance in process control to ensure the production and distribution of high quality drinking water to consumers with lower operation costs. In this work, mathematical variable selection methods were utilized to find the optimal subsets of variables to develop predictive models for the two quality parameters of drinking water. The found variable subsets were evaluated using three modelling methods and comparisons with the expert knowledge-based models were carried out. The study pointed out the importance of expert knowledge and showed that predicting the quality of treated water is possible but finding an optimal subset of input parameters for a model that predicts the precise value of a quality parameter is challenging.

Key words | aluminium, artificial neural network, cross-validation, genetic algorithm, multiple linear regression, turbidity

INTRODUCTION

The accurate monitoring and control of a water treatment process are vital in producing and distributing high quality drinking water with reasonable expenses to consumers. The efficient functionality of a treatment process and the quality of drinking water are sensitive to disturbances which may occur suddenly especially when surface water is used as the raw water source. For example, human actions can pollute the environment or floods and erosion caused by heavy rains or fast melting of snow affect the quality of the raw water source and make water treatment more challenging. Modern water treatment plants (WTPs) include several on-line sensors and manually taken samples are analysed extensively around the process, but the quality of drinking water cannot be measured until water is leaving the distribution system. On this account, a predictive model based on measurements at the early stage of the water treatment process is a valuable tool for assessing the upcoming quality of distributed water. Predictive information used as assistance in process control ensures the distribution of high quality water to consumers with lower operating costs and avoiding health-related risks. However, modelling the operation of a WTP or part of it is very challenging due to the nonlinear and complex nature of the treatment process and the partly unknown factors affecting process functionality (WHO 2017; Price & Heberling 2018).

Several indicators can assess the efficiency of a water treatment process but one way is to evaluate the drinking water quality parameters of turbidity and residual chemical coagulation. Chemical coagulation with aluminium or ferric salts is the most commonly used in water treatment processes to reduce organic matter, colour and turbidity of surface raw water. A dysfunctional treatment process or overdosed coagulation chemical may lead to a declined quality of treated water which causes in addition to general annoyance possible minor and short-lived health effects on consumers and economic losses to a WTP (WHO 2017). The residual aluminium level in drinking water has been
found to be affected by, *inter alia*, raw water temperature, raw water organic matter, raw water and coagulation pH, raw water turbidity, coagulation chemical, and the silicate concentration (*Juntunen et al.* 2012; *Kimura et al.* 2013; *Tomperi et al.* 2013). Residual aluminium can be minimized for example by optimizing pH, avoiding an excessive dosing of aluminium, a good mixing of coagulants, optimum paddle speed in the flocculation process and efficient floc filtration (WHO 2017).

Turbidity in water is caused by suspended particles which reduce the clarity of water. Suspended particles in treated water are usually from the raw water source due to an inadequate water treatment process or from the re-suspension of sediments. High turbidity is not by itself harmful or a threat to health but the particles that cause turbidity can protect microorganisms from the effects of disinfection and stimulate bacterial growth which causes health effects for consumers. Among others, the coagulation chemical dose, the turbidity of raw water, the temperature of water, raw water KMnO₄ (potassium permanganate) value and water pH have been found to affect the turbidity of drinking water (*Juntunen et al.* 2012). Residual aluminium and turbidity in drinking water have a strong mutual correlation (*Tomperi et al.* 2011).

In a water treatment process, changes can occur frequently and very often micro-scale interactions are poorly understood, which makes it impossible to develop a useful mechanistic model. Data-driven models can capture relationships using the desired input–output mapping and physical processes do not have to be known explicitly as when using mechanistic models. Among others, artificial neural network (ANN) and multiple linear regression (MLR) methods are capable of modelling drinking water quality (*Baxter et al.* 2001; *Shetty et al.* 2003; *Juntunen et al.* 2012; *Rak* 2013; *Cuesta Cordoba et al.* 2014).

In this work, the optimal subsets of variables for developing predictive drinking water quality models (residual aluminium and turbidity) are sought using automatic variable selection methods based on mathematical analysis of the data. ANNs, cross-validation (CV) and MLR methods are used for developing models to evaluate the results of variable selection. The developed models are compared with each other and with the models developed based on expert knowledge.

**MATERIALS AND METHODS**

**Water treatment plant**

The data used in this work were collected from a WTP which uses the surface water of a nearby lake as raw water. In the treatment process, before adding a coagulation chemical the pH of raw water is adjusted to the optimal value with calcium hydroxide. An aluminium-based coagulation chemical is dosed as a function of raw water KMnO₄ value. Potassium permanganate is an oxidizing agent used for measuring the chemical oxygen demand (COD), which indicates the amount of organic compounds in water. After the filtration, water pH is adjusted to the optimal level for distribution. UV-radiation and sodium hypochlorite are used for disinfection.

**Data**

To ensure sufficient results in modelling, the data should be fully representative of the full spectrum of all possible conditions. In cases like water treatment where varying environmental factors (temperature, rainfall, etc.) affect the quality of raw water and the operating process, the source dataset must encompass in minimum one full year of measured data. The dataset used in this study covered a period of 16 consecutive months including on-line process measurements and daily-performed laboratory measurements of raw water and treated water. On-line measurements were daily-averaged and combined with the corresponding laboratory data to reduce the amount of data and ease the computational load after it was discovered that the hour-averaged data from the period were computationally too heavy to achieve results in a reasonable processing time. However, daily-average values can be considered adequate in a slow water treatment process where delay between inflow and outflow is several hours. All the data processing, variable selection and modelling were performed using Matlab (The MathWorks Inc.) and self-developed scripts.

At first, the dataset was inspected and incorrect (unrealistic) values were manually deleted and replaced with linear interpolation. Before variable selection and modelling, the dataset was scaled between \([-2, +2]\) using the nonlinear
scaling method based on generalized moments, norms and skewness. The method (presented in more detail in Juuso (2013)) removes the nonlinearity of the variables while scaling.

**Variable selection**

Variable selection is one of the most important steps in model development to find the relevant variables from a large dataset. Only significant input variables should be selected into a model because some variables may be strongly correlated with each other, noisy or have no significant relationship with an output variable and thus will not be informative. Non-essential input variables increase computational complexity, make the model training more difficult and the prediction results worse. An over-fitted model that has an excellent performance in training but is unusable with new upcoming data may occur if the model contains too many variables which are fitted also to random noise included in the data. In this work, in addition to selecting a subset of input variables manually based on expert knowledge, five mathematical variable selection methods described below were utilized for finding the optimal subsets for the water quality parameters.

In correlation-based selection, variables are selected by the absolute value of their correlation coefficient against the output variable. At first, variable pairs that had mutual correlation |0.83| or larger were sought and the variable with a lower correlation coefficient was removed from the set. Then variables were arranged in downward order by their correlation coefficient for selection for the model development (MathWorks 2017).

Forward selection adds one variable at a time from the original dataset to a model. Initially, every variable from the variable set is evaluated and the best variable is added. At the following steps, variables that are not already included in the model are evaluated and variables are added one at a time to the model based on their performance. Adding is continued until the performance of the model does not improve. Stepwise selection is a modified forward selection method, which adds the best variable to, or deletes the worst variable from, a variable subset at each round. Adding and deleting is based on variables’ statistical significance in regression (Hall 1999; Guyon & Elisseeff 2003; MathWorks 2017).

A genetic algorithm (GA) is a biological evolutionary-based optimization method for various problems. The new populations of chromosomes are generated using genetic operators, reproduction (selection and crossover) and mutation, to improve the population for solving an optimization problem. For feature selection, a subset is represented as a binary string (chromosomes) of the length of the total number of variables. The value of each position in the string represents the presence or absence of a particular variable. A variable is selected if the bit is 1 and not selected if the bit is 0. Each variable is evaluated to determine its fitness, or its ability to survive and move into the next generation. The new variables are created by the iterating of crossover and mutation processes. The results of GA variable selection are dependent on the tuning parameter values, which are optimized manually one by one (Siedlecki & Sklansky 1989; Davis 1991).

A successive projections algorithm (SPA) is a forward selection method in multivariate calibration. SPA uses simple operations in a vector space to minimize the collinearity between selected variables. It starts with one variable and adds a new variable at each iteration until a specific number of variables is reached. SPA selects variables whose information content is minimally redundant. SPA is described in detail in Araújo et al. (2001). In this work, SPA was used as preselection method before the GA.

**Modelling**

**Artificial neural network**

An ANN typically consists of at least three layers: an input layer, one or more hidden layers and an output layer (Figure 1). External inputs (x) of the network are received by neurons in the input layer. Each input is multiplied by interconnection weights (w) and sent forward to every neuron in the hidden layer where they are summed and processed by a nonlinear transfer function (f). The output of the network is given by the neurons on the output layer. The number of hidden layers, the number of neurons, the learning rate and initial weights, for instance, influence the network training and prediction accuracy. Finding the optimal number of hidden layers, nodes and parameter.
values is often very laborious because they are usually found by trial and error (Dayhoff 1990).

In this work, the neural network consisted of measured variables selected by the variable selection methods as inputs, the predicted value of water quality as output and one hidden layer (five neurons). Resilient back-propagation was used as the training function and the mean squared error (MSE) as the performance function. The aim of the training process is to minimize the output error, the difference between the predicted output and measured output, by adjusting the interconnection weights. The calculated error is back-propagated to the neural network through each layer and the weights are adjusted to decrease the error. Hyperbolic tangent sigmoid was used as the transfer function for the hidden layer, and the linear transfer function for the output layer (Dayhoff 1990; Baxter et al. 2001).

**Multiple linear regression**

MLR can be used to describe a quantitative relationship between independent variables and a dependent variable as a linear system, to predict future scores on the dependent variable or to test specific hypotheses based on a scientific theory or prior research. MLR may not be useful with nonlinear features and can only ascertain relationships, but not be sure about the underlying causal mechanism. With sparse data MLR can outperform the nonlinear models (Araújo et al. 2007; Hastie et al. 2009). The MLR equation is a weighted linear combination of the independent variables:

\[ Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e \]  

where \( b_0 \) is a constant value, \( b_1...b_n \) are regression coefficients, \( X_1...X_n \) are independent variables and \( e \) is the error.

**K-fold CV**

For efficient training and validation, both subsets of the data have to be long and representative enough of all possible conditions. Splitting the dataset into two subsets may cause a significant loss of data in training the model. K-fold CV is a resampling method and one way to predict the fit of a model for a validation set without an explicit validation set. The whole data are used for training and validating the model (in this work, a linear regression model) by randomly partitioning the dataset into \( k \) subsets of equal size and using \( k-1 \) parts of the data for training and one part for validation and repeating this \( k \) times until each of the subsets is used once as the validation data. Final estimation is produced by combining these \( k \) results of the folds (Rao et al. 2008; Arlot & Celisse 2010).
RESULTS AND DISCUSSION

The results of the variable selection for the residual aluminium and turbidity are presented in Tables 1 and 2, respectively. The listed variables are sorted by their importance (the order of selection) and limited to the first ten selected variables. Using SPA with GA did not significantly reduce the amount of selected variables nor improve the modelling results, contrary to Sorsa et al. (2015) where SPA applied before GA greatly improved the reliability of the genetic search.

Subsets for the residual aluminium were very divergent, even though stepwise and forward selection resulted in similar subsets due to the similarity of the methods. Some variables like raw water temperature and KMnO4 are found to be important and selected for every subset. The variable subsets for turbidity are even larger and more divergent. Variables like raw water KMnO4 and temperature are

### Table 1 | Variable selection results for the residual aluminium models

<table>
<thead>
<tr>
<th>Correlation selection</th>
<th>Forward selection</th>
<th>Genetic algorithm</th>
<th>SPA + GA</th>
<th>Stepwise selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW temperature&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW temperature&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NaOCl&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NaOCl&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW temperature&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Coagulant chemical/KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Ca(OH)&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Filter 1 pressure&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
</tr>
<tr>
<td>DW turbidity</td>
<td>RW turbidity&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 4 pressure&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW turbidity&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>NaOCl&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW temperature&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NaOCl&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>RW pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Ca(OH)&lt;sub&gt;2&lt;/sub&gt;</td>
<td>RW turbidity&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 2 flow&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Ca(OH)&lt;sub&gt;2&lt;/sub&gt;</td>
</tr>
<tr>
<td>Mixing pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 4 pressure</td>
<td>Filter 4 flow</td>
<td>Ca(OH)&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Filter 4 pressure</td>
</tr>
<tr>
<td>Ca(OH)&lt;sub&gt;2&lt;/sub&gt; flow</td>
<td>RW bacteria</td>
<td>Return flow</td>
<td>RW turbidity</td>
<td>RW bacteria</td>
</tr>
<tr>
<td>RW bacteria</td>
<td>Return flow</td>
<td>Filter 1 pressure</td>
<td>Return flow</td>
<td>Return flow</td>
</tr>
<tr>
<td>RW colour</td>
<td>Filter 2 flow</td>
<td>Filter 3 pressure</td>
<td>RW temperature</td>
<td>RW pH</td>
</tr>
<tr>
<td>Coagulant</td>
<td>RW pH</td>
<td>Filter 4 pressure</td>
<td>RW bacteria</td>
<td>Filter 2 flow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+6</td>
<td></td>
<td>+1</td>
</tr>
</tbody>
</table>

<sup>a</sup>Selected for the model.

RW, raw water; DW, drinking water; Ca(OH)<sub>2</sub>, calcium hydroxide; NaOCl, sodium hypochlorite.

### Table 2 | Variable selection results for the turbidity models

<table>
<thead>
<tr>
<th>Correlation selection</th>
<th>Forward selection</th>
<th>Genetic algorithm</th>
<th>SPA + GA</th>
<th>Stepwise selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>LW conductivity&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Coagulation chemical&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW pH&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Alum/KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Mixing pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Alum/KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>RW flow&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
</tr>
<tr>
<td>DW aluminium</td>
<td>Alum/KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Coagulation chemical&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW coliform&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW temperature&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Coagulation chemical&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 4 pressure&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW conductivity&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 4 pressure&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 4 pressure&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mixing pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>LW conductivity&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Return flow&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Alum/KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
</tr>
<tr>
<td>RW temperature&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Coagulation chemical</td>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Filter 2 flow&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 2 pressure&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>DW pH</td>
<td>RW pH</td>
<td>Mixing pH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Filter 1 flow</td>
<td>Mixing pH</td>
</tr>
<tr>
<td>RW colour&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RW UV intensity</td>
<td>RW pH</td>
<td>Mixing pH</td>
<td>LW conductivity</td>
</tr>
<tr>
<td>RW pH</td>
<td>Filter 2 pressure</td>
<td>RW UV intensity</td>
<td>Return flow</td>
<td>Ca(OH)&lt;sub&gt;2&lt;/sub&gt;</td>
</tr>
<tr>
<td>RW bacteria</td>
<td>RW conductivity</td>
<td>RW KMnO4&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Filter 3 flow</td>
<td>Coagulation chemical&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1</td>
<td></td>
<td>+1</td>
</tr>
</tbody>
</table>

<sup>a</sup>Selected for the model.

RW, raw water; DW, drinking water; Ca(OH)<sub>2</sub>, calcium hydroxide; NaOCl, sodium hypochlorite; LW, limewater.
found to be important also for turbidity. Based on the results, finding one optimal subset of input variables is difficult, but certain variables are important in developing predictive models for drinking water quality parameters. For practical predictive model development the amount of input variables should be decent to avoid over-fitting and to keep the computational load low. Thus, none of the found subsets are practical as is and manual selection based on expert knowledge is needed to reduce the number of variables before model development.

Scatter plots are used for visually studying the relationship between two variables by plotting the independent and dependent variables as a collection of points. In Figure 2, the relationships of the process measurements to the residual aluminium (on the horizontal axes) are shown. As the residual aluminium and turbidity have mutual dependence, the relationships between the process variables and turbidity are not presented here. Inspection confirms that the residual aluminium in drinking water is dependent for example on raw water KMnO$_4$, turbidity and the temperature of raw water. Heavy rains and melting snow cause the level of the raw water source to rise and the value of KMnO$_4$ to increase. This also causes changes in the colour of the raw water as well as turbidity and residual aluminium of the treated drinking water to increase. Hence, these variables also have mutual dependences.

For the model development, variables marked with note (*) in Tables 1 and 2 were selected. In MLR and ANN modelling the original dataset was divided into training (12 months) and validation (4 months) subsets. In CV, the whole dataset was used in both training and validation. The performance of the models is evaluated and compared by using Root Mean Square Error (RMSE) and coefficient of determination ($R^2$). In Table 3, the performance values of the residual aluminium and turbidity models developed by the ANN, CV and MLR methods are shown. In addition, the results of the models using manually selected variables (coagulation chemical, raw water KMnO$_4$, raw water pH, raw water temperature and raw water turbidity) are presented. These variables were selected because they are found to be important variables affecting the performance of the treatment process and the quality of treated water. They can also be measured on-line and are located in the early stage of the treatment process which ensures the true prediction ability of the model.

Even though the water treatment process has some non-linearities, the ANN models did not perform well. In general, the performances of the ANN models were weaker than the performances of the MLR and CV models. This may be because the optimal parameter values were not found in ANN model training. The length of the training dataset was too short to include all the required
conditions to develop a robust model as the validation period included some phenomena that were not in the training set. In general, the best performance was achieved using five-fold CV with variables found using the forward selection method. The accuracy of the best model can be considered good for predicting the level of residual aluminium in drinking water, yet it is not possible to achieve the exact values in every situation. The models based on manually selected variables also performed well compared with other models. The best MLR residual aluminium model was developed with the manually selected variables. The variable subset found using SPA combined with GA was in practice useless for accurate model development. Predicting the water turbidity was even more challenging. The performances of the CV models were decent for predicting the turbidity levels and changes that occur. Even though the training accuracies of the ANN and MLR models were acceptable, their $R^2$ values in validation were poor.

**CONCLUSION**

In this work, mathematical variable selection methods were utilized for finding the optimal subset of input variables to develop predictive models for the water quality parameters, residual aluminium and turbidity in treated water, that indicate the functionality of the treatment process. Three modelling methods were used for developing models to evaluate the results of variable selection.

The study pointed out the importance of expert knowledge in model development. Finding the optimal subset of input variables for the models is very challenging. The number of selected variables was too high and manual elimination of variables for developing a practical model was needed. However, the selection methods showed the importance of certain variables (the temperature and KMnO$_4$ of raw water, for example), which tally with the visual inspection of variable dependences and expert knowledge from earlier studies.

The training dataset and the tuning parameters affected the accuracy of the ANN and MLR models. CV resulted in models which can be considered satisfactory for predicting the level of the quality variables and the changes that occur even though the exact values are not possible to achieve. Thus, the models can be used to assist in controlling the process for a better quality of water and to avoid health-related risks and economic losses.

**ACKNOWLEDGEMENTS**

The data used in this study were collected during the project ‘Development of a comprehensive water quality management system’. The writing of this paper was financially supported by KAUTE Foundation which is thereby greatly acknowledged. Aki Sorsa, D.Sc. (Tech.), is greatly acknowledged for providing solutions in variable selection issues.

**REFERENCES**


Hastie, T., Tibshirani, R. & Friedman, J. 2009 The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd edn. Springer-Verlag, New York, USA.


First received 16 March 2018; accepted in revised form 4 October 2018. Available online 17 October 2018.