


<b>EREM 74/1</b> Journal of Environmental Research, Engineering and Management Vol. 74 / No. 1 / 2018 pp. 7-20 DOI 10.5755/j01.erem.74.1.20083 © Kaunas University of Technology	<b>Predictive Analysis of Microbial Water Quality Using          Machine-Learning Algorithms</b>	
	Received 2018/02	Accepted after revision 2018/02
	 <a href="http://dx.doi.org/10.5755/j01.erem.74.1.20083">http://dx.doi.org/10.5755/j01.erem.74.1.20083</a>	

# Predictive Analysis of Microbial Water Quality Using Machine-Learning Algorithms

**Hadi Mohammed, Andreas Longva, Razak Seidu**

Water and Environmental Engineering, Faculty of Engineering and Natural Sciences, Norwegian University of Science and Technology (NTNU) in Ålesund, Larsgårdsvegen 2, 6009 Ålesund, Norway

**Corresponding author:** hadi.mohammedntnu.no

Water and Environmental Engineering, Faculty of Engineering and Natural Sciences, Norwegian University of Science and Technology (NTNU) in Ålesund, Larsgårdsvegen 2, 6009 Ålesund, Norway

Given the increasing recognition of machine learning tools for use in water quality monitoring, enhancing their applicability in full-scale plants require investigation of their capabilities and limitations in key aspects of the water supply chain. This study comprehensively evaluates the performances of three artificial neural network (ANN) training algorithms and three solvers for regression support vector machine (SVM) with different kernel functions in the estimation of the counts of faecal indicator bacteria from measured records of physico-chemical water quality parameters. In addition, input data were subjected to different normalization methods to determine their effects on the performances of both ANN and SVM models. The feedforward and the cascade forward algorithms yielded the lowest mean square error (MSE) values among the various ANN model configurations. No distinct disparity was found in the performances of the various solvers of regression SVM in the estimations. For the regression SVM kernel functions, the radial basis function (RBF) and the Gaussian kernel functions resulted in the lowest MSE values. Both the ANN and regression SVM have comparable abilities in predicting the levels of the faecal indicator organisms in raw water. However, the ANN models were more efficient in estimating intense variations in the levels of the indicator organisms in raw water.

**Keywords:** machine learning, feed forward, cascade forward, layer-recurrent, regression SVM, coliform bacteria.

## Introduction

The wide range of applications of machine learning algorithms and computational intelligence as decision

support tools has made them indispensable in the water supply industry today. Due to their robustness

and high accuracy in learning from imperfect data, the techniques significantly aid the automation of key processes in knowledge engineering, and have proven to be reliable replacements for some rather time-consuming regular activities in the industry. Unlike traditional multivariate regression methods that model linear relations among variables under the assumption of independence among variables, the machine learning approach applies different statistical, probabilistic and optimization techniques in finding and 'learning' regularities and patterns in records of system operational data, establishing complex nonlinear relationships between noisy and interdependent variables. This makes it possible to make inferences and decisions that are difficult to make using conventional statistical methodologies (Mitsell, 1997), resulting in improved efficiencies of the system. Machine learning techniques mainly include neural networks, instance-based or case-based learning, rule-based learning, analytic learning, ensemble learning and genetic algorithms (Langley and Herbert, 1995).

Improving the capacity of treatment plants to effectively manage water quality requires alternative methods of estimating the levels of indicator bacteria in raw water to augment conventional laboratory analysis methods. In recent years, researchers have applied various data-driven techniques to explain the influences of various physico-chemical water quality parameters on concentrations of faecal indicator organisms (FIOs) and other water quality parameters in raw water. These mostly include regression methods (Black et al., 2007; Juntunen et al., 2012) and artificial intelligence methods such as artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS) and support vector machine (SVM) (Singh et al., 2009; Kim et al., 2012; Heddam, 2014; Mohammed et al., 2017). For instance, a recent study carried out in Germany applied multiple regression method to explain the levels and variations of faecal indicator bacteria in river water in Germany. According to the authors, up to 70% of the variations in the levels of faecal indicator bacteria in the raw water samples are associated with variations in variables such as pH, rainfall and solar radiation. Other studies compared the performance of regression and ANN, ANFIS and SVM methods in the prediction of water quality parameters

(Abyaneh, 2014; Chandramouli et al., 2007; Zhang et al., 2015) and reported higher accuracy of these methods relative to conventional regression methods.

Recently, the application of ANN in water quality indices prediction has gained popularity due to its ability to approximate complex non-linear relationships between physico-chemical parameters and microbial organisms in water with high accuracy. However, with the exception of its 'black box' nature, optimizing the various parameters of the network in a way that will prevent overtraining may be challenging. In addition, the technique may be prone to overfitting (Tu, 1996). The performance of an ANN model can be affected by the selected model architecture, structure and the training algorithm used, since they mainly define how the inputs are transformed into outputs (Wu et al., 2014). Moreover, during the network training process, achieving optimal performance usually requires using different numbers of hidden layer neurons on a trial and error basis, to achieve optimal performance. Accordingly, establishing a particular set of protocols, including the data pre-processing method, training algorithm and hidden layer neurons for application in typical prediction problems in the water supply system is vital for repeating and generalizing the development of ideal ANN models for use in the water supply system.

Although SVM has been successfully applied in solving regression and time series problems, the method has not been widely applied in the prediction of faecal indicator organisms in raw water as compared with ANN. It has been reported that the difficulty associated with the selection of SVM model parameters makes its applicability limited (Lv et al., 2014). The lack of an optimal method for adaptation of regression SVM parameters has been reported in a recent review (Sapankevych and Sankar, 2009). Moreover, according to this study, although a vast majority of applications use the Gaussian kernel function, which has been widely accepted to be more efficient, there is no formal proof of the function's optimality. In addition, the choice of kernel functions in different time series prediction applications are arbitrary. Thus, to improve the applicability of this highly efficient method in the water quality management, it is vital to evaluate the response of different kernel functions in predictive models.

Accordingly, the main objective of this study is to investigate the suitability of different combinations

of data normalization methods, ANN training algorithms, number of hidden layer neurons in ANN structure as well as different regression SVM training algorithms and kernel functions in the prediction of coliform bacteria from water quality parameters. The study is based on measured records of physico-chemical parameters from the Oset drinking water treatment plant in Oslo, Norway. The models are expected to provide a fast and reliable approach to complement existing monitoring exercises in water treatment facilities in Norway and beyond.

## Materials and Methods

### Study site

Fig. 1 shows the location of the study area. Maridalen Lake, where the Oset water treatment plant sources its raw water from is located in the northern outskirts of Oslo, is the largest lake within the Oslo municipality. The lake has a surface area of 3.83 km<sup>2</sup> and

an elevation of 149 m and it is surrounded by forest catchment area of 252 km<sup>2</sup>. With two main primary inflows from the northern part (Skajærsjølva and Dausjølva), the lake has an average annual flow of 184 million cubic meters of water, and drains mainly through the Akerselva River to the south. To prevent contamination due to human activities around the lake, the surrounding municipalities have imposed restrictions on mostly recreational activities to some distance away from the lake and adjoining streams (Oslo municipal water and waste department, 2012).

### Data set

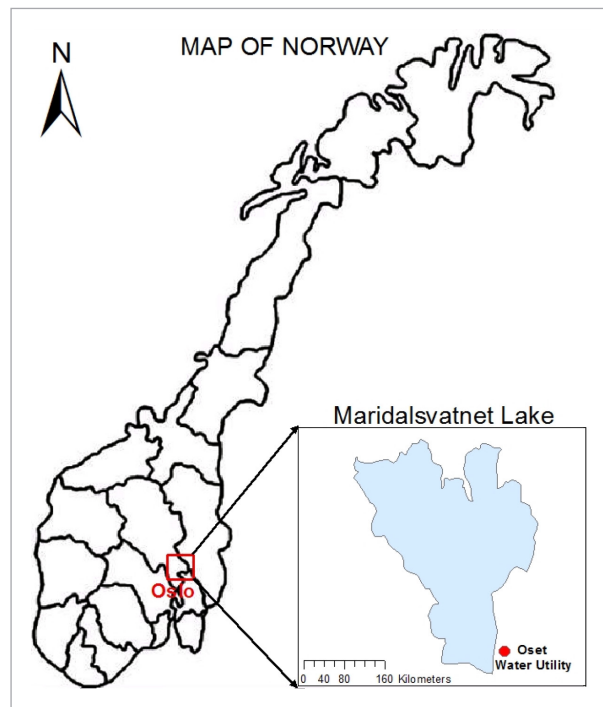
The study is based on observed counts of coliform bacteria and measured water quality parameters including pH, temperature (°C), conductivity ( $\mu$ ), turbidity (NTU), colour (mg Pt/L) and alkalinity (mmol/L) at the raw water intake point of the Oset drinking water treatment plant in Oslo, Norway. The data consist of 208 weekly records taken from January 2012 to December 2015. With a capacity of 390,000 m<sup>3</sup>/day, the Oset treatment plant, known to be the largest municipal water treatment plant in Scandinavia, provides safe drinking water to about 90% of the inhabitants of Oslo and depends on raw water drawn at a depth of 32 m from Maridalen Lake. The microbial data and the physico-chemical parameters are data taken as part of the routine monitoring exercise at the water treatment plant.

### Data normalization

Prior to the training of ANN and SVM, it is useful to carry out data normalization. The main purpose is to adjust all data variables to a common scale, to avoid bias. This facilitates the learning process particularly for the network. Moreover, by using non-linear transfer functions at the output nodes of the network, it is necessary to transform the desired outputs to match the actual range of the network. Experiences in some studies show that considerable improvement in the efficiency of standard ANN and other artificial intelligence models can be achieved when the input data are normalized before training (Jayalakshmi and Santhakumaran 2011). This study compares two different data normalization methods; minimum-maximum and z-score as described in the following expressions:

**Fig. 1**

Map of study area showing location of Lake Maridalsvatnet in Oslo, Norway



\_ min-max normalization:

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

\_ z-score normalization:

$$x' = \frac{x_i - \mu}{\sigma} \quad (2)$$

where  $x'$  is the normalized data point  $x_i$ ,  $x_{min}$  and  $x_{max}$  are the minimum and maximum values of each data variable, and  $\mu$  and  $\sigma$  are their respective means and standard deviations.

### ANN models

ANN is an information processing technology system that stimulates the human brain nervous system (Negnevitsky, 2005). It is an efficient tool for modelling complex relationships and processes. The successful application of ANN modelling in water quality studies cannot be overemphasized. Various training algorithms can be used in training the network.

### Feed-forward ANN model

It is the most widely used form of supervised ANN in which signals are allowed to travel in only the forward direction, from input nodes to output nodes. It uses multiple layers of neurons with non-linear transfer functions to learn complex relationships between input and output vectors. Feedforward ANN is generally governed by the expression:

$$y_j = F_j(\sum_{i=1}^m w_{j,i} \cdot y_i + b_i) \quad (4)$$

where  $y_j$  is the output,  $F_j$  is the transfer function of the  $j^{\text{th}}$  neuron in a layer,  $w_{j,i}$  is the weight that connects the output  $y_i$  of the  $i^{\text{th}}$  neuron from one layer to the input of the  $j^{\text{th}}$  neuron in the next layer,  $b_i$  is the bias weight on the  $j^{\text{th}}$  neuron of each layer. It is mostly solved using the error back-propagation algorithm. The error function is expressed as:

$$E_p = \frac{1}{2} \sum_{j=1}^m (t_{pj} - y_{pj})^2 \quad (5)$$

where  $t_{pj}$  represents the desired target value  $p$  and  $y_{pj}$  is the  $j^{\text{th}}$  output of the final layer.

### Cascade-forward ANN model

The technique is similar to the feed-forward in the use of error backpropagation method for updating weights. However, here, an additional weight connection from the input nodes to all subsequent layers is included. The speed at which the network learns the relationship existing between inputs and targets might be improved by the additional connections (Al-allaf and AbdAlKader, 2011).

### Layer-recurrent ANN model

Layer-recurrent networks (RNN) are a class of ANN which allow connections between units in a loop. The main distinguishing feature from a feedforward network is the presence of additional connections between units, forming a feedback loop that enables preservation of sequential information. That is, for each time step  $t$ , the decision reached is affected by the decision from the preceding time step  $t-1$ . Accordingly, for an input sequence  $x = (x_1, \dots, x_T)$ , the network outputs ( $y_j$ ) for a node,  $j$  depends on connection weights and the current input signal as well as preceding states of the network as:

$$y_j = Ax'(t) \quad (6)$$

where  $A$  is the weight matrix of the output layer neurons that is connected to the hidden neurons. The output,  $x'(t)$  of a given input vector  $x(t)$  at time,  $t$  is given by:

$$x'(t) = f(w_h x'(t-1) + w_{h0} x(t-1)) \quad (7)$$

$f()$  is a function that characterizes the hidden nodes.

Unlike the feedforward network, layer recurrent network uses an extension of backpropagation known as the backpropagation through time (BPTT) algorithm, which unfolds the network in time through the creation of various copies of the recurrent units so it can be treated like a feedforward network with associated weights. Thus, the algorithm is updated in descript time steps. The errors at the hidden nodes are propagated backward as:

$$E_{pj}(t-1) = \sum_h^m E_{ph}(t) u_{hj} f'(s_{pj}(t-1)) \quad (8)$$

where  $u_{hj}$  is the weight matrix mapping the previous hidden layer ( $s_{pj}(t-1)$ ) to the current one.  $h$  and  $j$  are respective indices for hidden nodes at time steps  $t$  and  $t-1$ . A standard three-layer feedforward network with the backpropagation algorithm was used to predict the levels of coliform bacteria (CFU/100 mL). The six inputs were water pH, temperature ( $^{\circ}\text{C}$ ), conductivity ( $\mu$ ), turbidity (NTU), colour (mgPt/L) and alkalinity (mmol/L).

### Regression SVM models

Support vector machine, developed by Vapnik and his collaborators (Vapnik, 1995), is a learning technique that is originally meant for binary classification. The principal idea of the method is to obtain a hyper-plane that separates different classes of data points. To enable a linear separation of data, SVMs employ kernel functions to map data from the input space into a higher dimensional space, creating two parallel hyper-planes to separate the data. Therefore, in classification, the geometric margin between the two hyper-planes is maximized, while minimizing the classification error (Cristianine and Taylor, 2000; Singh et al., 2011).

In regression SVM as applied in this study, the aim is to find the optimal hyper-plane with the minimum distance from all data points (Lin et al., 2008; Pan et al., 2008). That is, instead of a yes or no output in a typical classification problem, regression SVM is trained to output a numerical value. That is, the hyperplane with the least distance to all data points is optimized. Thus, for a given vector of water quality variables (inputs), and observed faecal indicator organisms (target), a linear function, that yields an estimate of the target such that the deviation between the estimate and the target is less than the insensitive loss function ( $\epsilon$ ) for all training data is obtained (Wu et al., 2010). Thus,

$$E_{pj}(t-1) = \sum_h^m E_{ph}(t) u_{hj} f'(s_{pj}(t-1)) \quad (8)$$

$$f(x) = wx + b, \quad x \in R^n, b \in R \quad (9)$$

$$|y_i - f(x_i)| < \epsilon \quad (10)$$

where  $w \in R^n$  is a weight vector and  $b$  is the bias. Addition of an error term to the linear estimation above leads to the following expression:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i + \xi_i^* \quad (11)$$

where  $((y_i - w^T x_i - b) \leq \epsilon + \xi_i, w \cdot x + b - y_i \leq \epsilon + \xi_i^*, \xi_i, \xi_i^* \geq 0)$  for all  $i = 1, \dots, m$ .  $\xi_i$  and  $\xi_i^*$  are slack variables that are used in the optimization of the separation planes. Thus, all estimation errors that are smaller than  $\epsilon$  do not enter the objective function.  $C$  is a positive parameter known as the penalty parameter, and is responsible for adjusting the level of errors in the estimation. With the introduction of Lagrangian multipliers, the following optimization problem can be obtained:

$$w(\alpha, \alpha^*) = -\epsilon \sum_{i=1}^m (\alpha_i^* + \alpha_i) + \sum_{i=1}^m (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{i,j=1}^m (\alpha_i^* - \alpha_i) (\alpha_i^* - \alpha_i) k(x_i, x_j) \quad (12)$$

Here,  $0 \leq \alpha_i, \alpha_i^* \leq C$ , for  $i = 1, \dots, m$ , and  $\sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0$ , where  $k(x_i, x)$  is the kernel function and  $\alpha_i$  and  $\alpha_i^*$  are Lagrangian multipliers used in the optimization process. Finally, the regression estimate is obtained as:

$$f(x) = \sum_{i=1}^m (\alpha_i^* - \alpha_i) K(x_i, x) + b \quad (13)$$

where  $b = y_i - \sum_{i=1}^m (\alpha_i^* - \alpha_i) k(x_i, x)$  ( $C \geq \alpha_i, \alpha_i^* \geq 0, i = 1, \dots, m$ ).

The use of the structural risk minimization principle in the formulation of SVMs gives the method a greater ability for generalization. With the exception of the various parameters used in SVM for both classification and regression problems, the type of the kernel function used is a key determinant of the performance of the SVM method (Pan et al., 2008). In this study, different kernel functions are applied to determine a more appropriate one for estimating the concentrations of faecal indicator organisms in raw water from the measured physico-chemical parameters. We applied linear, polynomial and Gaussian kernel functions, defined as follows:

– linear kernel:

$$k(x, x^*) = (x^T \cdot x^* + C) \quad (14)$$

– polynomial kernel:

$$k(x, x^*) = (x^T \cdot x^* + C)^d \quad (15)$$

– Gaussian kernel (RBF):

$$k(x, x^*) = \exp\left(-\frac{\|x-x^*\|^2}{2\sigma^2}\right) \quad (16)$$

### Model calibration and evaluation

Fig. 2 shows the procedure for the calibration and evaluation of machine learning models as applied in this study. Across the various configurations of both ANN and regression SVM models, the input dataset was divided into a training set (70%) and testing (30%). After the algorithm learns from the data, several performance measures are used to evaluate the quality of the solution to classification, regression and clustering problems typically modelled in machine learning. The performance indices may include mean square error (MSE), root mean square error (RMSE) and confusion

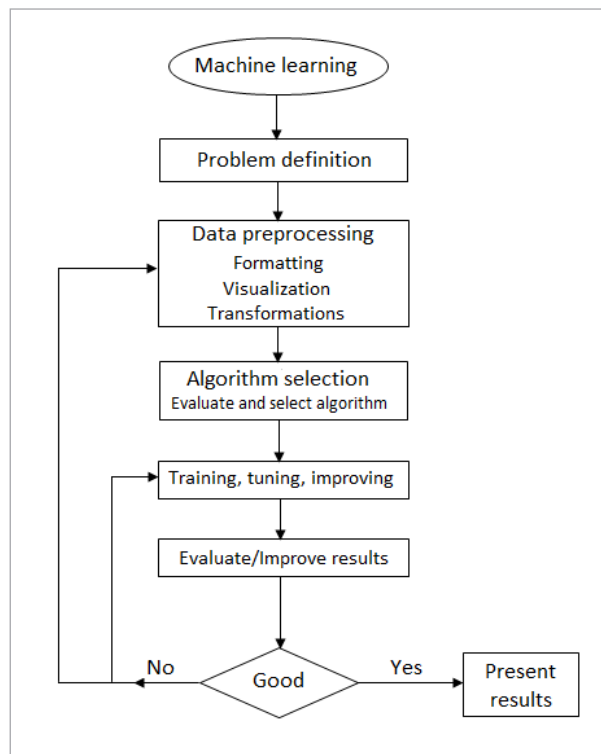
matrix indices, such as true positive, true negative, false positive, false negative, etc. In this study, visual inspection of plots that compared the outputs from the various combinations of input data normalization methods, and numbers of hidden layers were first used to assess the performances of the various neural network-training algorithms. For the regression SVM models, the outputs from the different kernel functions were compared graphically. Finally, the overall performances of the models were compared using the mean square error values of the model predictions.

### Model sensitivity analysis

Evaluating the sensitivities of each input parameter is necessary to determine their respective influences on the predictive abilities of the models. Moreover, in predicting the concentrations of the faecal indicator organisms in raw water, it is essential to determine which physical or chemical parameters of raw water affect the variations in the indicator organisms. While each parameter may directly or indirectly influence the occurrence of the indicator organisms in raw water, identifying the most important surrogates provides vital information for the management of the microbial quality of raw water. Therefore, the relative importance of the various input parameters were evaluated through a stepwise omission of each input parameter in the models. At each stage, the MSE value of the model predictions was calculated and compared with the corresponding value for the model with all inputs included. In this case, the omission of more important inputs could significantly raise the MSE in the model. Due to the different ANN and SVM configurations used in this study, resulting in different MSE values, a model configuration with the least MSE was selected each from the ANN and SVM models during the sensitivity analysis.

**Fig. 2**

Workflow of machine learning model calibration and evaluation



## Results and discussions

### Raw data

Results of the initial statistical analysis of the raw data set are shown in Table 1. Considerable skewness is noted in the data set, with only pH, conductivity and alkalinity being approximately symmetrical. In addition, some water quality parameters including



**Table 1**

Descriptive statistics of measured raw water parameters

	Mean					Min.	Max.	SD	Var.	Skew.
	Winter	Sprig	Summer	Autumn	Overall					
1	2	3	4	5	6	7	8	9	10	11
pH	6.60	6.55	6.47	6.41	6.50	6.27	7.00	0.11	0.13	<b>-0.08</b>
Temperature	7.06	5.21	7.15	8.89	7.07	2.80	13.70	3.23	10.4	1.72
Conductivity	2.53	2.62	2.64	2.61	2.59	2.08	3.02	0.14	0.02	<b>-0.88</b>
Turbidity	0.51	0.39	0.41	0.47	0.45	0.20	2.29	0.21	0.05	2.87
Color	27.43	27.05	25.73	24.45	26.31	19.00	57.00	6.84	51.5	2.19
Alkalinity	0.08	0.44	0.08	0.08	0.17	0.07	0.13	0.01	0	<b>0.67</b>
Coliform	5.17	0.27	0.54	24.07	7.49	0	300	22.41	502	8.36
E. coli	0.52	0.14	0.11	3.45	1.12	0	300	15.90	252	13.8

Min. = minimum, Max. = maximum, SD = standard deviation, Var. = variance, Skew. = skewness

temperature and colour showed high variabilities, respectively ranging from 0 to 13°C and 19 to 57 mg Pt/L. As shown in Table 1, higher counts of the faecal indicator organism in raw water were observed in autumn-winter turn over periods of each year, with the highest counts of coliform bacteria (300 CFU/100 mL) occurring at this period of 2014. In addition, greater proportion of the 208 observed data points for the faecal indicator organisms were zeros (approximately 56% and 84% for coliform bacteria and *E. coli*, respectively). Among the four main seasons in Norway, the mean concentrations of the faecal indicator organisms in raw water are higher in winter and autumn. For instance, compared with the four-year mean concentration of approximately 8 CFU/100 mL, the mean winter season concentration of coliform bacteria in raw water was approximately 5 CFU/100 mL. For the autumn seasons, the mean concentration is 24 CFU/100 mL.

### ANN model response

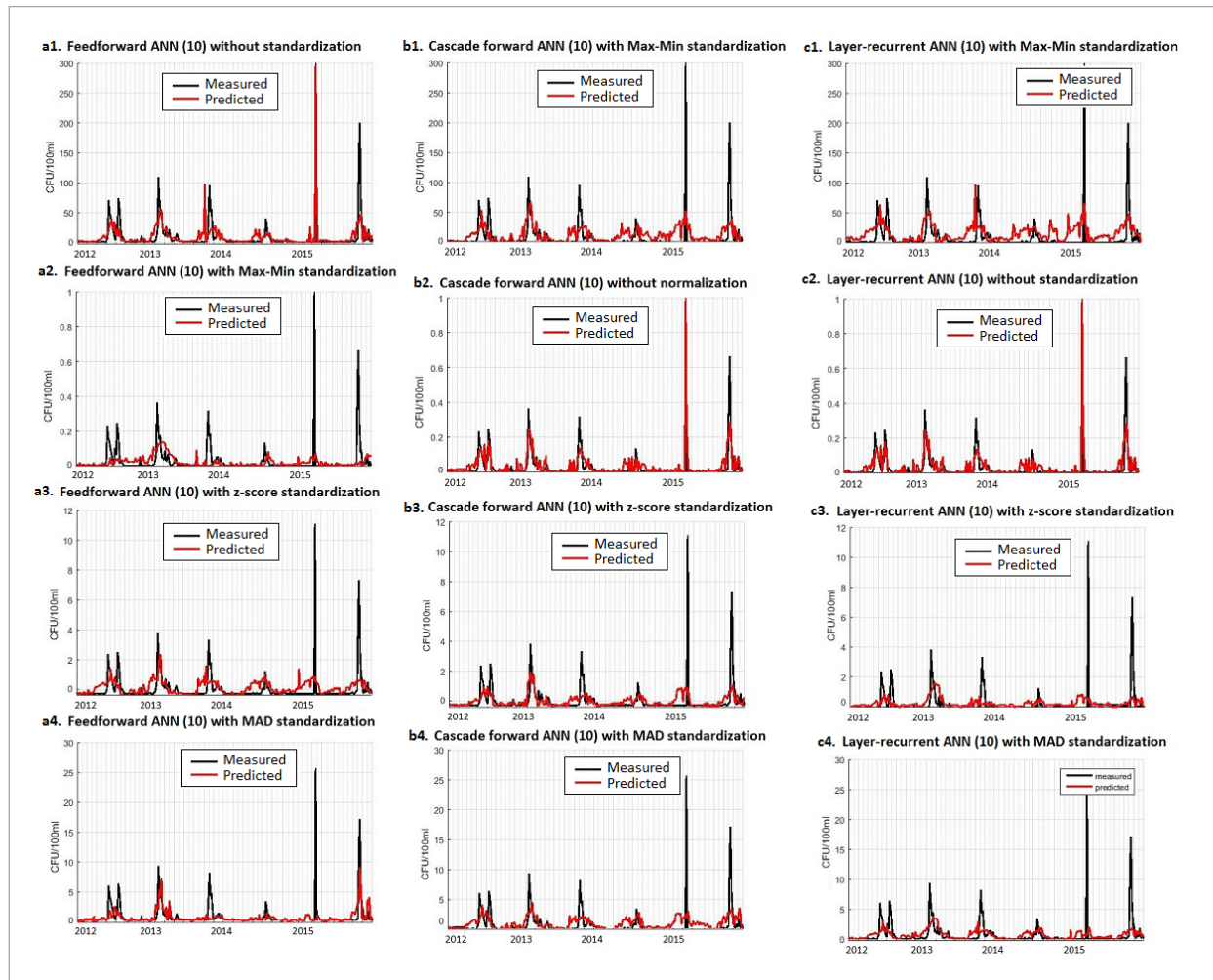
Fig. 3 shows outputs of the performances of the feed-forward ANN model. Only the results from the feed-forward ANN models for 10 hidden layers are shown in this figure. The effect of the various raw data standardization methods on the performances of the ANN models are distinct from the plots. It can be observed

that the model generally captures significant variations in the counts of coliform bacteria in raw water. In terms of estimating periods of intense variation in the count of coliform bacteria in raw water, the feed-forward ANN model without data standardization outperforms the others. However, the close similarity in the performances of the models from the three standardization methods indicate that there may be biasing in the one without standardization (Fig. 3 a1). This shows that the relative contributions of the input parameter in the model are not equally distributed. It can also be noted that out of the various normalization methods, only the min-max method resulted in output values that fall within the range of 0 and 1.

These are within the range of outputs produced by typical activation functions used in training neural network. Interestingly, the training of the network proceeded faster than the other normalization methods. Although the output of the model with min-max standardization (Fig. 3 a1) shows the least performance in estimating intense variations in the level of coliform bacteria, relative to the other standardization methods, it results in the lowest mean square prediction error. The effects of input data standardization was equally noticeable from the performances of the other ANN training algorithms used in this study. However, unlike the feedforward network, the

Fig. 3

Performances of feedforward, cascade forward and layer-recurrent neural networks with 10 hidden layer neurons in predicting coliform bacteria under different data standardization methods



performances of the other training algorithms improved when the input data was normalized using the min-max method. As shown in Fig. 3, considerable variations in the counts of coliform bacteria in raw water were estimated after the normalization. When the input data set was normalized using the z-score and MAD methods, for the same number of neurons in the hidden layer, each of the three ANN training algorithms showed adequate performances. However, the feedforward network was more efficient in estimating the variabilities in the count of coliform bacteria than the other two. The z-score normalization, however, yielded some negative counts of the

microorganism, which obviously resulted in negative outputs in some cases.

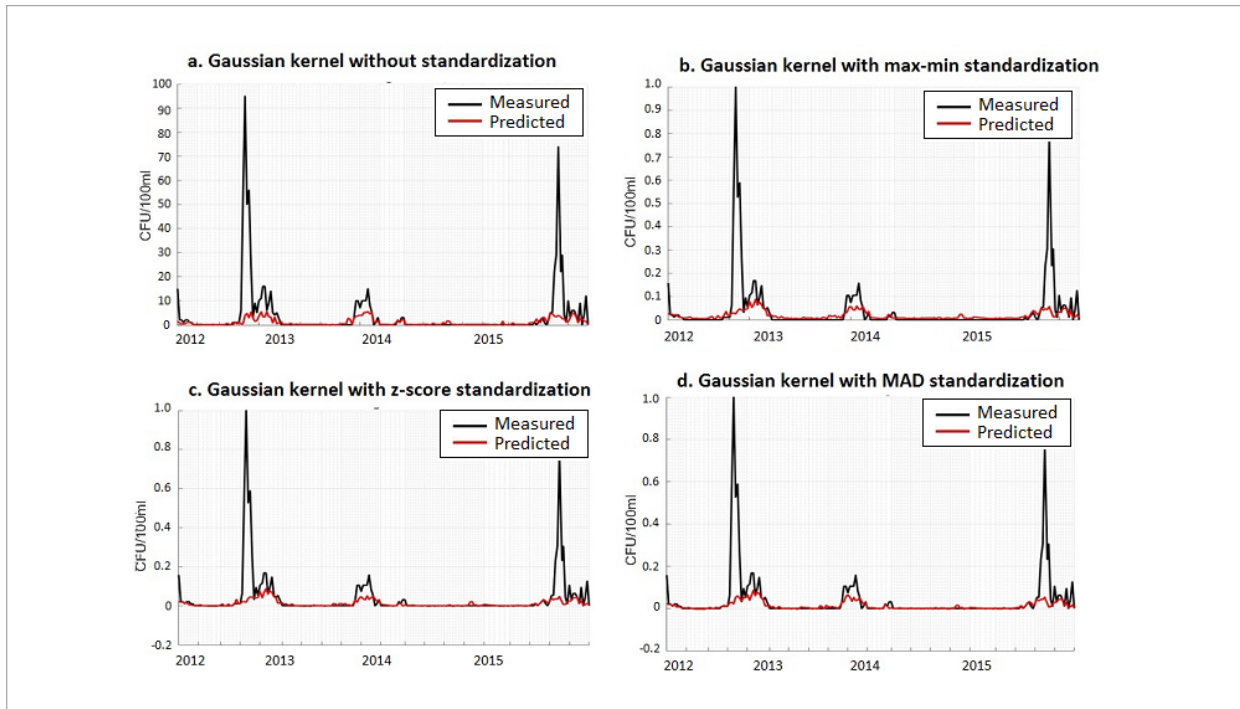
### SVM model response

Fig. 4 shows the response of the regression SVM model with the Gaussian kernel function using input data with different normalization methods. The regression SVM model generally failed to account for periods of intense variations in the counts of bacteria. However, the model clearly estimates zero counts of the indicator organism with high accuracy. Typical records of indicator organisms observed in raw water contain a large number of zeros, thus the ability of a model to



**Fig. 4**

Performances of regression SVM model in the prediction of coliform bacteria under different kernel functions



estimate zero counts as well as higher counts from real time measurements of water quality variables makes it useful in the water supply system.

Amongst the various kernel functions, the Gaussian and the RBF functions yielded estimates that were much closer to the counts of the indicator organism observed in raw water. When the polynomial kernel function was applied, the system failed completely in learning from the data set. This resulted in much larger estimates, especially when no input data normalization was applied. Finally, different SVM training algorithms used in this study yielded similar results, with comparable mean square error of predictions.

### Comparison of model outputs

Since a large number of model response plots were generated using different combinations of input data normalization methods, network training algorithms, etc., only selected few plots were included in this paper. To assess the overall performances of each of the different configurations of the models, the mean square errors (MSE) of the estimates were calculated

and used as the main model performance index. Tables 2 and 3 show the results of the performances of the ANN model configurations while the results of the regression SVM models are shown in Tables 4 and 5. Without data normalization, the best performances of the feedforward (MSE = 84 CFU/100 mL) and cascade forward (MSE = 92 CFU/100 mL) ANN models were achieved when the network contained 20 neurons in the hidden layer. The least MSE for the layer-recurrent network (MSE = 97 CFU/100 mL) in this case was achieved with the 10 hidden layer neurons instead. Results of the models with data normalization prior to model training were transformed back into their original forms to enable fair comparison with the results of the models with the raw data set (without normalization). For both ANN and SVM models for coliform bacteria, the normalizations clearly improved the performances of the models. For instances, for the same ANN feedforward model architecture for prediction of coliform bacteria, the min-max data normalization resulted in MSE value of 2.02 CFU/100 mL, compared with MSE value of 84 CFU/100 mL achieved in the

**Table 2**

MSE of ANN models for coliform bacteria with various training algorithms, hidden layer sets, and normalizations

ANN	Hidden layers	MSE (CFU/100 mL)		
		Raw data	Min-Max	z-score
1	2	3	4	5
Feedforward	10	137.66	3.04	83.92
	20	84.57	2.02	93.31
	50	132.13	3.25	90.11
Cascade-forward	10	106.42	2.98	88.07
	20	92.81	2.65	139.47
	50	98.91	4.68	73.76
Layer-recurrent	10	97.58	4.04	70.36
	20	122.50	2.43	74.15
	50	106.70	3.85	87.05

**Table 3**

MSE of ANN models for *E. coli* with various training algorithms, hidden layer sets, and normalizations

ANN	Hidden layers	MSE (CFU/100 mL)		
		Raw data	Min-Max	z-score
1	2	3	4	5
Feedforward	10	1.22	3.45	97.38
	20	0.08	2.48	96.01
	50	0.07	4.52	97.31
Cascade-forward	10	0.09	4.37	106.11
	20	0.09	4.46	83.18
	50	0.10	3.65	88.29
Layer-recurrent	10	0.14	3.51	81.41
	20	0.10	2.62	74.11
	50	0.09	3.03	70.19

model without any normalization. Similar significant improvements were produced in the other configurations of both ANN and SVM models. Interestingly, however, the data normalizations failed to improve the performances of the models for *E. coli*, as the

least MSE rather increased from 0.07 CFU/100 mL (for the model without data normalization) to 2.48 CFU/100 mL (with min-max normalization) and 70.19 CFU/100 mL (with z-score normalization) in the ANN models. The corresponding least MSEs in the SVM models similarly increased from 0.14 CFU/100 mL to 3.75 CFU/100 mL and 102.81 CFU/100 mL. This may be due to the significant disparities in the characteristics of the actual observation data for the two faecal indicator organisms (56% and 84% of data points for coliform bacteria and *E. coli* respectively were zeros). It is difficult to train a model when the output data set is full of zeros. However, this is the typical nature of the occurrence of these indicator organisms in raw water, thus, making the establishment of a reliable for their prediction necessary in drinking water supply.

Further, for each normalization method, the MSE values for the three ANN training algorithms are approximately similar. The least MSE (2.02 CFU/100 mL) of the coliform bacteria model under the min-max data normalization was close to values achieved for both

**Table 4**

MSE of regression SVM models for coliform bacteria with different training algorithms, kernel functions, and normalizations

SVM	Solver	MSE (CFU/100 mL)		
		Raw data	Min-Max	z-score
1	2	3	4	5
Linear	SMO	140.09	4.51	108.58
	ISDA	140.33	4.55	102.93
	L1Qp	140.51	4.46	103.36
Gaussian	SMO	131.54	4.21	95.56
	ISDA	132.94	4.31	96.64
	L1Qp	131.56	4.08	96.54
RBF	SMO	131.57	4.24	95.69
	ISDA	132.14	4.05	96.30
	L1Qp	130.31	4.28	95.86
Polynomial	SMO	6011.53	5.39	103.89
	ISDA	7399.48	8.93	104.61
	L1Qp	3330.11	6.39	152.32

**Table 5**

MSE of regression SVM models for *E. coli* with different training algorithms, kernel functions, and normalizations

SVM	Solver	MSE (CFU/100 mL)		
		Raw data	Min-Max	z-score
1	2	3	4	5
Linear	SMO	0.14	6.24	113.08
	ISDA	0.14	6.21	113.06
	L1Qp	0.14	6.25	113.14
Gaussian	SMO	0.14	4.01	102.98
	ISDA	0.14	3.75	103.14
	L1Qp	0.14	4.21	102.99
RBF	SMO	0.15	4.09	102.81
	ISDA	0.14	3.91	104.08
	L1Qp	0.14	4.31	103.91
Polynomial	SMO	256.84	6.16	104.09
	ISDA	1399.45	6.95	106.36
	L1Qp	175.47	6.63	105.48

the cascade forward (2.65 CFU/100 mL) and layer-recurrent (2.43 CFU/100 mL) architectures with 20 sets of hidden layer neurons. The overall precision of the regression SVM models were slightly lower than the ANN models (due to larger MSE values). For instance, while a least MSE of 2.02 CFU/100 mL was achieved with the feedforward ANN with min-max normalization, the corresponding value in the regression SVM model was approximately twice this MSE (4.05 FCU/100 mL) achieved with the RBF algorithm. Similarly, with the z-score normalization method, slightly higher MSE values were achieved in the regression SVM models compared with the ANN models, as shown in Tables 2 to 5.

### Results of model sensitivity analysis

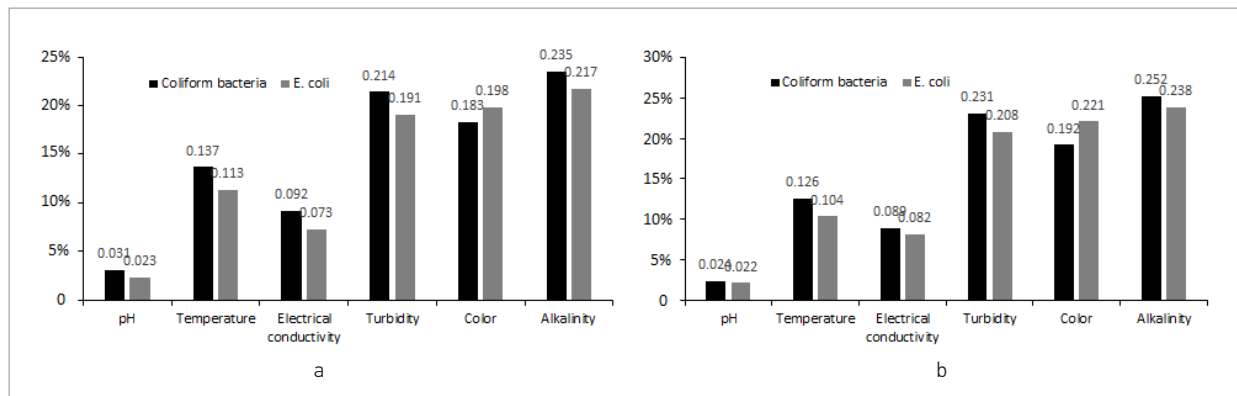
The sensitivities of the various physical and chemical water quality parameters to the performances of the ANN and regression SVM models in the prediction of faecal indicator bacteria in raw water are shown in Fig. 5. Raw water turbidity, colour and alkalinity had

the greatest influence on the prediction of both coliform bacteria and *E. coli*. For instance, by removing turbidity from the ANN models (Fig. 5 A), the MSE values increased from 2.02 CFU/100 mL to 2.47 CFU/100 mL and from 2.48 CFU/100 mL to 2.95 CFU/100 mL, respectively, for coliform bacteria and *E. coli* predictions. These correspond to 21% and 19% increases in the MSEs compared with the models with all the input variables. The corresponding percent increases in the MSEs resulting from the removal of colour and alkalinity were approximately 18% (for coliform bacteria), 19% (for *E. coli*) and 23% (for coliform bacteria), 22% (for *E. coli*), respectively. Similar increases in the various MSEs were obtained in the SVM models as shown in Fig. 5 (B). However, the magnitudes of the increases were considerably higher in the SVM models. In addition, the importance of water temperature on the prediction of the faecal indicator bacteria in both models is evident from the increases in the MSEs of up to 14% for coliform bacteria and 12% for *E. coli*. The water quality parameter that showed the least sensitivity in the predictions was the water pH, with MSE increases of approximately 3% (ANN models) and 2% (SVM models), and the changes were similar in both coliform bacteria and *E. coli* prediction models.

Modelling the occurrence of faecal indicators in drinking water sources enables early warning information to the water utility operators prior to the treatment of raw water, such that treatment processes can be optimized where necessary. Detection methods of faecal indicator organisms (FIBs) in raw water are still being improved, mainly to reduce the detection time. Moreover, conventional weekly sampling and analysis of FIBs in raw water at water utilities may not give the actual contamination levels, particularly during heavy rainfall (Tryland et al., 2011). For this reason, public health burdens associated with waterborne outbreaks occurring after peak rainfall events can be reduced if FIBs in raw water are estimated in 'real time' on a daily basis before treatment. ANN and regression SVM are highly efficient machine learning techniques that are capable of learning complex relationships among variables in a system. Enhancing the applicability of these data-driven techniques in drinking water supply systems require clarification of the strengths and weaknesses of various architectures that can be used

**Fig. 5**

Sensitivity of input parameters to the performances of the ANN models (A) and regression SVM models (B)



to reliably estimate the concentrations of FIBs in raw water. Results of this study distinctly indicate that while various algorithms of ANN and regression SVM and input data normalization methods adequately accommodate the variations in typical FIBs observed in raw water, a simple feedforward ANN model with 10 hidden layer neurons is enough for real time prediction of FIBs in raw water, with lower chances of overfitting. In addition, as shown in Figs. 3 and 4, applying normalizations on the input data sets does not necessarily improve the performances of both the ANN and regression SVM models. Compared to the z-score normalization, the use of the minimum-maximum method results in better model performances, particularly by the MSE measure. The main limitation of the models applied in this study is the absence of a validation stage, where all the models would be tested on different data sets. However, the objective at this stage of the study was more focused on establishing the simplest model configuration for use in this respect, as well as clarifying the effects of different solution algorithms, hidden layer sets, and normalization methods on the accuracies of the models.

## Conclusions

The model results showed that all the three ANN training algorithms adequately estimated the counts of the faecal indicator organisms in raw water with

comparable efficiencies. Similarly, no distinct disparity in model performances were observed when different training algorithms were applied to the regression SVM. In the ANN models, acceptable estimates were achieved with 10 and 20 sets of hidden layer neurons. Although networks were faster with 50 hidden layer neurons in all configurations used in this study, extreme variations in the model estimates were observed in this case. Further, although the two data normalization methods applied in this study considerably improved the performances of the ANN and SVM models for the coliform bacteria, they failed to improve the accuracies of *E. coli* prediction using both machine learning techniques. Moreover, the feedforward network was more efficient in estimating the counts of the faecal indicator organisms in raw water when no normalization was applied to the input data set. The effect of the normalization methods on the performances of the machine learning techniques in predicting faecal indicator organisms in raw water may depend on the data characteristics of the observed indicator bacteria used as model output. Amongst the various kernel functions applied to the regression SVM, the RBF and the Gaussian functions were more efficient, with the least MSE errors. Finally, results of this study suggest that compared with regression SVM, ANN is highly efficient in estimating necessary variations in FIBs in raw water from measured values of physico-chemical parameters of raw water with lower mean square prediction errors. Although regression SVM adequately estimates the variation in the indicator organism with MSE comparable with ANN, SVM fails to

estimate periods of intense variations in the level of the indicator organism in raw water, an information that is highly desired for optimizing treatment in water utility services.

### Acknowledgement

The authors wish to show gratitude to the management

of the Oset drinking water treatment plant in Oslo, Norway, for providing the water quality data used in this study.

### Funding

The Norwegian Research Council through the KLIMA-FORSK project provided funding for this study.

## References

- Abyaneh, H. Z. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, 12(1), <https://doi.org/10.1186/2052-336X-12-40>
- Al-allaf, O. N. and AbdAlKader, S. A. (2011). Nonlinear Autoregressive Neural Network for Estimation Soil Temperature: A Comparison of Different Optimization Neural Network Algorithms. In Special issue of ICIT 2011 conference, pp. 43 - 51.
- Black, L. E., Brion, G. M. and Freitas, S. J. (2007). Multivariate logistic regression for predicting total culturable virus presence at the intake of a potable-water treatment plant: Novel application of the atypical coliform/total coliform ratio. *Applied and environmental microbiology*, 73(12), pp. 3965-3974. <https://doi.org/10.1128/AEM.02780-06>
- Chandramouli, V., Brion, G., Neelakantan, T. R and Lingireddy, S. (2007). Backfilling missing microbial concentrations in a riverine database using artificial neural networks," *Water research*, 41(1), pp. 217-227. <https://doi.org/10.1016/j.watres.2006.08.022>
- Cristianine, N. and Taylor, J.S. (2000). *An Introduction to Support Vector Machine and other Kernel based Learning Methods*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511801389>
- Heddam, S. (2014). Modeling hourly dissolved oxygen concentration (DO) using two different adaptive neuro-fuzzy inference systems (ANFIS): a comparative study. *Environmental monitoring and assessment*, 186(1), pp. 597-619. <https://doi.org/10.1007/s10661-013-3402-1>
- Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering*, 3(1), pp. 89 - 93. <https://doi.org/10.7763/IJCTE.2011.V3.288>
- Juntunen, P., Liukkonen, M., Pelo, M., Lehtola, M. J. and Hiltunen, Y. (2012). Modelling of water quality: an application to a water treatment process. *Applied Computational Intelligence and Soft Computing*, Article No. 4. <https://doi.org/10.1155/2012/846321>
- Kim, K. B., Kim, J. H., Jeong, Y., Jeong, Y. S. and Chung, S. J. (2012). Prediction of Coastal Fecal Indicator Bacteria Concentrations Using Multivariate Data Analysis. *Journal of Environmental Science and Engineering. A*, 1(4A), pp. 440 - 447.
- Pat, A. and Simon H.A. (1995). Applications of machine learning and rule induction. *Communications of the ACM* 38.11, 54-64. <https://doi.org/10.1145/219717.219768>
- Lin, S. W., Lee, Z. J., Chen, S. C. and Tseng, T. Y. (2008). Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied soft computing*, 8(4), pp. 1505-1512. <https://doi.org/10.1016/j.asoc.2007.10.012>
- Lv, J., Zou, W., Wang, W., (2014). Water quality prediction using support vector machine with differential evolution optimization. *ICIC express letters. Part B, Applications: an international journal of research and surveys*, 5(3), pp. 763-768.
- Mitchell, T. (1997). *Machine Learning*. New York: McGraw Hill.
- Mohammed, H., Hameed, I. A. and Seidu, R. (2017). Adaptive neuro-fuzzy inference system for predicting norovirus in drinking water supply. *International Conference on Informatics, Health & Technology (ICIHT)*, Riyadh, 2017, pp. 1-6. <https://doi.org/10.1109/ICIHT.2017.7899134>
- Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Pearson Education.
- Oslo municipal water and waste department.. 2012. Retrieved online: [http://www.vannavlopsetaten.oslo.kommune.no/vannet\\_vart/drikkevann/vannkilder/restriksjoner/](http://www.vannavlopsetaten.oslo.kommune.no/vannet_vart/drikkevann/vannkilder/restriksjoner/) [Accessed 15/11/2016].
- Pan, Y., Jiang, J., Wang, R. and Cao, H. (2008). Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds. *Chemometrics and Intelligent Laboratory Systems*, 92(2), pp. 169-178. <https://doi.org/10.1016/j.chemolab.2008.03.002>
- Sapankevych, N. I. and Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*, 4(2), pp. 25 - 38. <https://doi.org/10.1109/MCI.2009.932254>



Singh, K. P., Basant, A., Malik, A. and Jain, G. (2009). Artificial neural network modeling of the river water quality—a case study. *Ecological Modelling*, 220(6), pp. 888-895. <https://doi.org/10.1016/j.ecolmodel.2009.01.004>

Singh, K. P., Basant, N. and Gupta, S. (2011). Support vector machines in water quality management," *Analytica chimica acta*, 703(2), pp. 152-162. <https://doi.org/10.1016/j.aca.2011.07.027>

Tryland, I., Robertson, L., Blankenberg, A. G. B., Lindholm, M., Rohrlack, T., & Liltved, H. 2011 Impact of rainfall on microbial contamination of surface water. *International Journal of Climate Change Strategies and Management*, 3(4), 361-373. <https://doi.org/10.1108/17568691111175650>

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11), pp. 1225-1231. [https://doi.org/10.1016/S0895-4356\(96\)00002-9](https://doi.org/10.1016/S0895-4356(96)00002-9)

Vapnik, V., and Cortes, C. (1995). Support vector networks. *Machine Learning*, vol. 20, pp. 273-297. <https://doi.org/10.1007/BF00994018>

Wu, C., Lv, X., Cao, X., Mo, Y. and Chen, C. (2010). Application of support vector regression to predict metallogenic favourability degree," *International Journal of Physical Sciences*, 5(16), pp. 2523-2527.

Wu, W., Dandy, G. C. and Maier, H. R. (2014). Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environmental Modelling & Software*, 54, pp. 108-127. <https://doi.org/10.1016/j.envsoft.2013.12.016>

Zhang, Z., Deng, Z. and Rusch, K. A. (2015). Modeling fecal coliform bacteria levels at Gulf Coast Beaches. *Water Quality, Exposure and Health*, 7(3), pp. 255-263. <https://doi.org/10.1007/s12403-014-0145-3>

## Mikrobinio vandens kokybės prognozė analizė naudojant algoritmus

**Hadi Mohammed, Andreas Longva, Razak Seidu**

Vandens ir aplinkos inžinerija, Inžinerijos ir gamtos mokslų fakultetas, Norvegijos mokslo ir technologijų universitetas (NTNU), Ålesund, Larsgårdsvegen 2, 6009 Ålesund, Norvegija

Atsižvelgiant į tai, kad vis dažniau naudojamasi algoritmais paremtos priemonės, skirtos naudoti vandens kokybės stebėjimui, didinant jų pritaikymą visame pasaulyje, reikia iširti jų pajėgumus ir apribojimus pagrindiniuose vandens kokybės nustatymo grandinės aspektuose. Šiame tyrime išsamiai įvertintos trijų dirbtinių neuroninių tinklų (angl. ANN) mokymo algoritmų ir trijų regresijos palaikymo vektorius mašinos (SVM) sprendimų, turinčių skirtingų branduolio funkcijų, rezultatai vertinant fekalinių indikatorinių bakterijų skaičių iš išmatuotų fizikocheminio vandens kokybės įrašų parametrai. Be to, įvesties duomenims buvo taikomi skirtingi normalizavimo metodai, siekiant nustatyti jų poveikį ANN ir SVM modelių veikimui. Persiuntimo ir kaskadinio tipo pirmųjų algoritmai duoda mažiausias vidutines kvadratinės klaidas (MSE) reikšmes tarp įvairių ANN modelių konfigūracijų. Skaičiavimuose nenustatyta jokie skirtingo SVM regresijos sprendimų rezultatai. Regresijos SVM branduolio funkcijoms, radialinės bazinės funkcijos (RBF) ir Gauso branduolio funkcijos sukūrė mažiausias MSE vertes. Tiek ANN, tiek regresijos SVM yra panašūs sugebėjimai prognozuoti fekalijų indikatorius organizmo koncentraciją užterštame vandenyje. Tačiau ANN modeliai buvo veiksmingesni vertinant intensyvius indikatorinių organizmų pakitimus užterštame vandenyje.

**Raktiniai žodžiai:** mechanizmo algoritmas, poslinkis į priekį, kaskados į priekį, sluoksnio pasikartojimas, regresijos SVM, koliforminių bakterijų skaičius.

Gauta:

2018 m. vasaris

Priimta spaudai:

2018 m. vasaris