MODELLING OF RESIDUAL ALUMINUM IN WATER TREATMENT PROCESS

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

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

corresponding author: petri.juntunen@uef.fi (Petri Juntunen)

Abstract

Aluminium compounds are widely used as coagulation chemicals in water treatment processes. However, the concentration of residual aluminium should be minimized in potable water because it causes acceptability problems for consumers. In this paper, residual aluminium of a water treatment process was modelled using linear and nonlinear methods. The aims were to construct process models and to determine whether the concentration of residual aluminium could be predicted reliably by using these methods. A variable selection procedure with both methods was used to find the most important factors affecting the concentration of residual aluminium. The results showed that the most effective variables were water temperature, Al/$KMnO_4$-ratio and silicate concentration. In conclusion, the results were promising, and the methods used showed potential considering the wider use of the data processing procedure in water treatment processes.

Keywords: Water treatment, residual aluminium, modelling, variable selection.

Presenting Author's biography

Petri Juntunen (M. Sc. in Env. Eng.) is a post graduate student in University of Eastern Finland in Process Informatics Group. His main time job is in Kuopion Vesi (waterworks of Kuopio municipality) as a design engineer. Petri Juntunen's main interest of his studies is intelligent methods of water treatment.

1 General

Aluminium compounds are widely used in water treatment processes because of their good ability to coagulate and flocculate both inorganic and organic compounds. The coagulation phenomena in physical and chemical level is described quite precisely and variables concerning the phenomena are known relatively well [1, 2]. However, coagulate compounds have some disadvantages including residual aluminium in distributed water. The residual aluminium causes turbidity in water networks resulting acceptability problems for consumers [3]. In addition, metals such as aluminium (Al) have also been implicated in some possible health effects [4].

Archived process history has the potential to be used for diagnostics and improvement of industrial processes, especially when used in combination with specific laboratory measurements. At present, artificial neural networks (ANN) are considered an advantageous way to analyze data in a variety of industrial processes, such as polystyrene production, wastewater treatment, energy production, and even soldering of electronics [5]–[14]. Many of these applications have also demonstrated that ANNs can provide an efficient automated method for modelling industrial process data. Especially the multilayer perceptrons (MLP) have proved their efficacy in the modelling of a wide spectrum of industrial processes [6,10,11,14].

In this paper, residual aluminium of a water treatment process was modelled using multiple linear regression (MLR) and neural networks (MLP).

2 Process and data

The experimental data was collected from the water treatment plant of Suomen Sokeri in Kirkkonummi, Finland. The plant is a chemical treatment plant and uses mainly surface water from Lake Humaljärvi or mixture of Lake Humaljärvi and Pikkala reservoir.

The process is a typical chemical process with chemical flotation and powdered granulated activated carbon (PAC) filtration. PAX-14, which is a aluminium based coagulation chemical, produced by Kemira Kemwater, was used as a coagulation chemical. The dose of the chemical varied between 30- 80 g/m^3. The dose was set as a function of raw water $KMnO_4$ so that ratio would be between 0.8-2 kgAl/kg$KMnO_4$. The final decision of the dose is controlled by process personnel.

Before flocculation, the pH value is adjusted with calcium hydroxide to 6.1, which is experimentally found as an optimum pH value. After filtration pH was adjusted to 8.2 to be suitable for distribution. Disinfection was done with UV-radiation and adding sodium hypochlorite.

The data used was produced from the process and laboratory measurements of 275 day period. The original process data period was 5 minutes, which was averaged to 1 day data to be comparable to the laboratory data. Before modelling, the outliers of the data were filtered out and the missing data points were added by linear interpolation. The laboratory and process variables are shown in Tab. 1 and in Tab. 2, respectively.

Tab. 1 Laboratory data for modelling

Variable	unit
pH	
KMnO4	mg/l O_2
Hardness	mmol/l
Color	mgPt
Conductivity	mS/m
Silicate	mg/l
Turbidity	NTU

Tab. 2 Process data for modelling

Variable	unit
Intake of surface water	m^3/h
Intake of ground water	m^3/h
Total intake of water	m^3/h
Aluminium feed	l/h
Flow_to_filter_1	m^3/h
Flow_to_filter_2	m^3/h
Flow_to_filter_3	m^3/h
Flow_to_filter_4	m^3/h
Filter wash water	m^3/h
Portion of Lake Humaljärvi/ Pikkala resevoir water intake	%
Level of surface water	m
Pressure of filter 1	bar
Pressure of filter 2	bar
Pressure of filter 3	bar
KMnO4	mg/l
Flocculation pH	
Temperature of water	C^o
Aluminium Dose	g/m^3
Aluminium dose/raw water $KMnO_4$	

3 Methods

3.1 Multiple linear regression

In multiple linear regression [15], 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 given in (1):

$$Y = e + a_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n \quad (1)$$

where Y is the value for the response variable, X is the value of the predictor (explanatory) variable, $a_{0...n}$ equal the unknown coefficients 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.

3.2 Multilayer perceptrons

Multilayer perceptrons (MLP) are well-known feed-forward neural networks [5], which consist 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. At the first phase, the input layer distributes the inputs to the first hidden layer. Next, the hidden neurons summarize the inputs based on predefined weights, which either weaken or strengthen the effect of each input. The weights are determined by learning from examples (i.e. data samples), which is called supervised learning. Eventually, the inputs are processed by a transfer function, and the result is transferred as a linear combination to the next layer, which is generally the output layer. The performance of the model is then evaluated with an independent validation data set.

Matlab version 7.6 [18] software was used in data processing. The neural network used in the analysis consisted of the process parameters as inputs, one hidden layer with 5 hidden neurons and the output neuron representing the predicted amount of residual aluminium. Hyperbolic tangent sigmoid (*tansig*) transfer function was used for the hidden layer, and the linear (*purelin*) transfer function for the output layer. The resilient back-propagation (*trainrp*) algorithm [19] was exploited in training, and mean squared error (*mse*) as the error function in training.

3.3 Variable selection

In recent years, variable selection has become a relevant part of data analysis [16]. The purpose of variable selection is to select a subset of p variables from the set of P variables without considerably degrading or with possibly improving the performance of the model. While the so called exhaustive subset selection methods involve the evaluation of a very large number of subsets, the number of different subsets to be evaluated can be reduced significantly by using suboptimal search procedures [17], such as the sequential forward selection method, which was used for the selection of variables in this case. In the approach, the variables are progressively included to larger and larger subsets so that the prediction performance for residual aluminium is maximized. To select p variables from the set of P variables: (1) Search for the variable that gives the best value for 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 are selected.

4 Results

The results of variable selection for the residual aluminium are presented in Tab. 3 for the linear regression and in Tab. 4 for the MLP method. The evolution curves of variable selection for residual aluminium are presented in Fig. 3. The figure can be interpreted so that the model performance improves by adding variables to it, in the order presented in Tabs. 3 and 4, until the improvement gradually stops, achieving the best model at four variables by using the nonlinear MLP method and at six by using the linear MLR method. The validation of the models was made using the 5-fold cross validation method. The results for these five validation sets are shown in Fig. 2 (MLR) and Fig. 3 (MLP).

Tab. 3 Results for the MLR variable selection

Variables	**Correlation coefficient**
Temperature of water	0,722
Aluminium dose/raw water $KMnO_4$	0,764
Silicate	0,782
Turbidity	0,787
Portion of Humaljärvi/ Pikkala water intake	0,789
$KMnO_4$	0,791

Tab. 4 Results for the MLP variable selection

Variables	**Correlation coefficient**
Temperature of water	0,717
Silicate	0,795
Aluminium dose/raw water KMnO4	0,812
Turbidity	0,814
$KMnO_4$	0,807
Portion of Humaljärvi/ Pikkala water intake	0,814

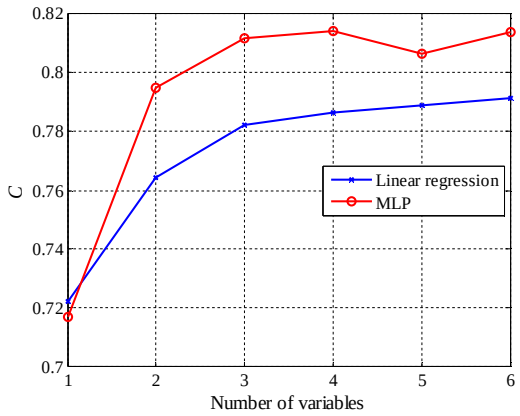

Fig. 1 Evolution curves of MLR and MLP variable selection for residual aluminium.

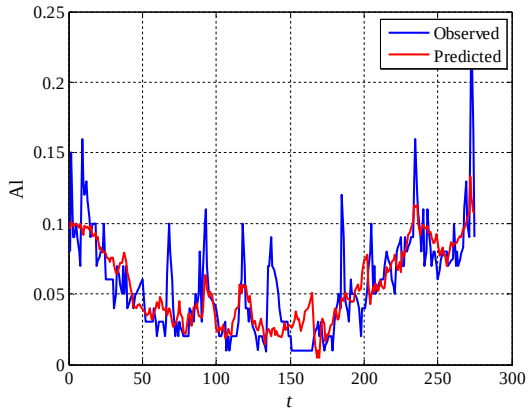

Fig. 2 Validation results of multiple linear regression.

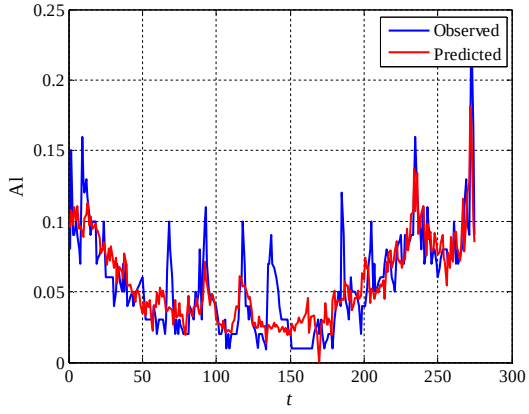

Fig. 3 Validation results of MLP.

5 Discussion

In this paper, the concentration of residual aluminium was modelled using multiple linear regression and neural networks. As the results in the Tab. 3 and Tab. 4 show, the models of both methods were fairly good. Especially, the prediction accuracy for the baseline of residual aluminium was excellent, although reasons for some sharp concentration peaks could not be found (see Fig. 2 and 3). The four most important variables (water temperature, Al/$KMnO_4$-ratio, silicate concentration and turbidity) were the same using both methods, which indicates good reliability of the results.

The nonlinear MLP method possibly seems to overcome the MLR method in two ways (see Fig. 1). First, the prediction ability of the MLP model is slightly better. This indicates that the problem would have some nonlinearity, although, the differences of the models are minor. Second, the best model can be achieved with a lesser number of variables by the MLP method, which would be beneficial in industrial applications. However, the use of a linear method could be reasonable in cases where nonlinearity does not prevail; in such cases the faster linear methods would be more efficient.

Some restrictions for the performance of the method used may follow from the fact that the variable selection was implemented by adding variables to the model one by one. This means that all the possible combinatorial effects are not evaluated using the most sophisticated methods. So it is possible that there are two or more variables whose mutual interaction may have a considerable effect on the concentration, although their singular effect on the model is insignificant. Nonetheless, developing data analysis methods for process industry involves always more or less compromises. Modelling the influence of all variable combinations would certainly be possible, but in reality the computing time for that would be long, which would evidently reduce the usability of the method in any real-world process applications.

The results show that the used approaches have several benefits. Apparently, the methods are flexible, have a high computing power, and are able to find nonlinear connections. Trained with real process data, the models are also able to adapt to exceptional situations in the process. Additionally, the methods are suitable for cases where physical processes are not well known or are highly complex. Generally speaking, although the created models assimilated good prediction abilities, the models can later be improved by adding more data samples or some process variables that were not used here/present this time. In practice, a linear model with two or three best variables could help to develop a controller to minimize the residual aluminum in the process. Thus, the data analysis procedure used seems to provide an efficient means of modelling the real water treatment process.

6 Conclusion

In conclusion, the results were promising considering the wider use of the data-driven variable selection and modelling in water treatment processes, and the potential of the method used was clearly indicated.

7 References

[1] J. E. Van Banshoten et al. Chemical aspects of coagulation using aluminum salts-I. Hydrolytic reactions of alum and polyaluminum chloride. *Water research*, 24, 12:1519-1526, 1990.

[2] J. E. Van Banshoten et al. Chemical aspects of coagulation using aluminum salts-II.Coagulation of fulvic acid using alum and polyaluminum chloride. *Water research*, 24, 12:1527-1535, 1990.

[3] World Health Organization. Guidelines for drinking-water quality [electronic resource]: incorporating first addendum. Vol. 1, Recommendations. – 3rd ed. 2006.

[4] A. Campbell. The role of aluminium and copper on neuroinflammation and Alzheimer's decease. *Journal of Alzheimer's decease*, 10:162-172, 2006.

[5] S. Haykin, Neural Networks and Learning Machines, 3rd ed., Pearson Education Inc., Upper Saddle River, New Jersey, 2009.

[6] M. Heikkinen, V. Nurminen, T. Hiltunen, and Y. Hiltunen A Modeling and Optimization Tool for the Expandable Polystyrene Batch Process. *Chemical Product and Process Modeling*, 3 (article 3), 2008.

[7] M. Heikkinen, T. Hiltunen, M. Liukkonen, A. Kettunen, R. Kuivalainen, and Y. Hiltunen. A Modelling and Optimization System for Fluidized Bed Power Plants. *Expert Systems with Applications*, 32:10274-10279, 2009.

[8] M. Heikkinen, H. Poutiainen, M. Liukkonen, T. Heikkinen, and Y. Hiltunen. SOM-based Subtraction Analysis to Process Data of an Activated Sludge Treatment Plant. In I. Troch and F. Breitenecker, editors, *Proc. MATHMOD 09 VIENNA* [CD-ROM]. Argesim Publishing House, Vienna, 2009, pp. 1021-1026.

[9] M. Liukkonen, E. Havia, H. Leinonen, and Y. Hiltunen. Application of Self-Organizing Maps in Analysis of Wave Soldering Process. *Expert Systems with Applications*, 36:4604-4609, 2009.

[10] M. Liukkonen, T. Hiltunen, E. Havia, H. Leinonen, and Y. Hiltunen. Modeling of Soldering Quality by Using Artificial Neural Networks. *IEEE Trans. Electronics Packaging Manufacturing*, 32:89-96, 2009.

[11] M. Liukkonen, M. Heikkinen, T. Hiltunen, E. Hälikkä, R. Kuivalainen, and Y. Hiltunen. Modeling of Process States by Using Artificial Neural Networks in a Fluidized Bed Energy Plant. In I. Troch and F. Breitenecker, editors, *Proc. MATHMOD 09 VIENNA* [CD-ROM], Argesim Publishing House, Vienna, 2009, pp. 397-402.

[12] M. Liukkonen, M. Heikkinen, E. Hälikkä, R. Kuivalainen, and Y. Hiltunen. Emission Analysis of a Fluidized Bed Boiler by Using Self-Organizing Maps. In M. Kolehmainen, P. Toivanen, and B. Beliczynski, editors, *Lecture Notes in Computer Science 5495*, Springer-Verlag Berlin Heidelberg, 2009, pp. 119-129.

[13] M. Liukkonen, E. Havia, H. Leinonen, and Y. Hiltunen. Quality-oriented Optimization of Wave Soldering Process by Using Self-Organizing Maps. *Applied Soft Computing* (in press), 2010.

[14] M.R.G. Meireles, P.E.M. Almeida, and M.G. Simões. A Comprehensive Review for Industrial Applicability of Artificial Neural Networks. *IEEE Trans. Industrial Electronics*, 50:585-601, 2003.

[15] J. Cohen. Multiple Regression as a General Data-Analytic System. *Psychological Bulletin*, 70:426-443, 1968.

[16] H. Liu, H. Motoda, editors. Computational Methods of Feature Selection. Chapman & Hall, United States of America, 2008.

[17] A.W. Whitney. A Direct Method of Nonparametric Measurement Selection. *IEEE Transactions on Computers*, C-20:1100-1103, 1971.

[18] Mathworks. *Using Matlab version 5*. The Mathworks Inc., Natick, 1998.

[19] Riedmiller M. and Braun H., 1993. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In: Ruspini H. (ed.), *Proc. IEEE International Conference on Neural Networks (ICNN)*, San Francisco, pp. 586–591.