Reliability of Principal Component Analysis and Pearson Correlation Coefficient, for Application in Artificial Neural Network Model Development, for Water Treatment Plants

Jayaweera CD, Aziz N. *
School of Chemical Engineering, Engineering Campus, Universiti Sains Malaysia, Seri Ampangan,14300 Nibong Tebal, Seberang Perai Selatan, Penang, Malaysia
chamanthidj@gmail.com, chnaziz@usm.my*

Abstract. The Pearson Correlation Coefficient (PCC) and Principal Component Analysis (PCA) are methodologies commonly used for linear variable selection. PCC has been extensively used for variable selection, due to its simplicity and as it assists in recognizing the degree of correlation between input and output variables. Meanwhile, PCA has been used for recognizing variables that have high variances influencing the output variable. However, the use of linear forms of variables selection methodologies in non-linear modelling such as artificial neural networks (ANN) is questionable. In this work, the acceptability of PCC and PCA in variable selection for ANN modelling of the coagulation process in water treatment, is analysed. ANN models, aiming to predict coagulant dosage, treated water (TW) turbidity, TW pH and residual Aluminium, were developed. In order to compare the validity of inputs selected via PCC and PCA, an exhaustive search strategy of variable selection was carried out. The results showed that using the variables selected using PCA did not contribute in improving ANN model development. Meanwhile, variables selected by PCC were successfully used for all ANNs developed, except for TW pH prediction. The results also demonstrated that PCC and PCA are incapable of capturing collective effects of variables, on the output parameter.

1. Introduction
Coagulation process is primarily undertaken in water treatment plants for removal of turbidity. Alum is one of the most commonly used coagulants. The required dosage of coagulant is conventionally determined by jar tests, which consume time, chemicals and thus, is costly and not feasible to be used effectively, to respond to rapid changes of turbidity. Therefore, research has been carried out, to develop Artificial Neural Network (ANN) models, using past data of coagulant dosages and other factors such as pH, turbidity and alkalinity to approximate required dosages of coagulant with adequate accuracy.

The majority of studies that had developed ANN models for coagulation have adopted Multi-Layer Perceptron (MLP) architecture, as it had yielded in satisfying results. Theoretically, the Levenberg-Marquardt algorithm is considered a more superior learning algorithm for MLP models, than the back propagation algorithm, due to its ability to escape local minima.

Data preprocessing is one of fundamental aspects employed to ensure the validity of data used. [1] studied the effect of preprocessing and concluded that models trained with preprocessed data had a rapid learning rate and a smaller terminal error.

The Pearson Correlation coefficient (PCC) is a frequently used tool in the selection of input parameters in modeling coagulation to predict the optimal coagulant dosage [2,3]. PCC is used to determine the most related variables to the output parameter and eliminate inter-related input
parameters. [2,3] had used a combined approach of using PCC and an exhaustive search strategy to arrive at the optimal model predicting the coagulant dosage. [2] have recognized that historical alum dosage (Al(t-1)) as an inherent factor in developing their model, thus this factor is adopted in this work. PCC has been used for the determination of input parameters for the prediction of the optimal coagulant dosage, but it is yet to be used for models predicting treated water quality parameters. The use of predictive models for monitoring treated water qualities is advantageous over the use of online measuring instruments, as it spares the cost spent on infrastructure and maintenance.

Principal Component Analysis (PCA) is another method used for selection of inputs as frequently as PCC [4,5]). The variables that account for the highest variance in the data could be recognized using PCA, which often mandatorily contribute to increase accuracy of a prediction model. Thus, PCA is used in this research to identify the variables with the highest variance.

Although PCC and PCA are often used in input parameter determination successfully, they are linear methods, and therefore the use of them in non-linear models is questionable as pointed out by [6]. Consequently, the objective of this work is to investigate the acceptability of using PCC and PCA for input parameter selection in the development of ANN models for coagulant dosage, Treated Water (TW) turbidity, TW pH and residual Aluminium prediction.

2. Procedure
One year’s data corresponding to year 2005 from the Segama water treatment plant, Sabah, was used for the development of ANN models to predict alum dosage, TW turbidity, TW pH and residual Aluminium concentration. The input variables were selected from raw water quality parameters that include RW Color, RW Turbidity, RW Alkalinity, RW Total Dissolved Solids (TDS) and RW pH. Input variables were selected based on raw water quality parameters for simplicity and robustness. Raw water and treated water quality ranges are given in table 1.

| Table 1 - Water quality ranges |
|--------------------------------|
| **Raw water**                  | **Treated water**              |
|                                 | pH    | Turb (NTU) | Col (HU) | TDS | Alkalinity (mg/l) | pH    | Turb (NTU) | Col (HU) | TDS | Alkalinity (mg/l) | Al residual (mg/l) | Coagulant dosage |
| Min                             | 6.6   | 11         | 36       | 60  | 50                | 0.19  | 0          | 90      | 20  | 0.01                | 20                 |
| Max                             | 7.9   | 3405       | 656      | 160 | 170               | 7.5   | 5.24       | 6       | 170 | 0.2                 | 170                |

Matlab 2017 was used to develop the feed-forward ANN models. The number of neurons in the hidden layer was decided based on the approach used by [7], where a maximum of 2I+1 hidden neurons is used, I being the number of input variables. The transfer function used for the input and output layers were tansg and purelin, respectively. Training data constituted of 60% of the data, 20% of data was used for validation, and the remaining 20% of the data was used for testing. Data division was carried out randomly using the Matlab’s default option. The networks were trained using the Levenberg-Marquardt algorithm. The best model was based on the lowest test mean squared error and highest regression coefficient (R).

As indicated in the flowchart shown in Fig. 1, input parameter selection for developing ANN models was carried out using an exhaustive search strategy of the best combinations of RW variables. The selected input parameters were compared to input parameters selected using PCC and PCA. All values were normalized to carry out the principal component analysis, by dividing each variable by its maximum value. The exhaustive search strategy is a technique used for input selection where the optimum model is selected from all possible combinations of input parameters. It should be noted that it was intended to develop a model with minimal number of variables. In this work, four variables were found sufficient to achieve the acceptable model performance.
3. Results and discussion

As shown in Table 2, the cumulative variance of the first two components is 89% and therefore they sufficiently represent the variability of the entire variable space. Thus, the first two principal components are chosen to represent the data set.

Table 2 - Principal components

| Principal Component | Eigen vectors | Percentage total Variance | Cumulative variance |
|---------------------|---------------|----------------------------|---------------------|
| 1                   | 0.038         | 67                         | 0.67                |
| 2                   | 0.013         | 22                         | 0.89                |
| 3                   | 0.004         | 7                          | 0.96                |
| 4                   | 0.001         | 2                          | 0.98                |
| 5                   | 0.001         | 2                          | 1                   |

Table 3 shows the coefficients of each variable obtained for each principal component, from the analysis carried out. According to table 3, RW TDS and RW Alkalinity are the most significant variables, since they have high values for coefficients in both principal components. Consequently, both of them are deemed to be chosen input parameters as per the PCA.

Table 3 - Coefficients of variables of significant principal components

| Variables     | PC1     | PC2     |
|---------------|---------|---------|
| RW Color      | -0.071  | -0.080  |
| RW Turbidity  | -0.105  | -0.183  |
| RW Alkalinity | 0.728   | -0.682  |
| RW TDS        | 0.668   | 0.703   |
| RW pH         | 0.080   | 0.025   |

An exhaustive search of the best combination of input variables was carried out by testing the performance of models with different combinations of the available input raw water quality parameters. The results of this search are given in table 5. It could be noted that the variables identified by the PCA are not among the best combinations in any of the models shown in table 5. Therefore, it
is evident that PCA is not an efficient method for input parameter selection for the neural network models developed.

Pearson correlation coefficients of each raw water quality with respect to the output variable of each model are shown in Table 4.

| Output          | Alum dosage | RW Alkalinity | RW Color | RW pH  | RW TDS | RW turbidity | Residual Al |
|-----------------|-------------|---------------|----------|--------|--------|--------------|-------------|
| TW Turbidity    | -0.86       | -0.13         | 0.48     | 0.18   | -0.33  | 0.99         | Insignificant |
| TW pH           | 0.165       | 0.3587        | -0.1056  | 0.917  | 0.4154 | -0.1773      | Insignificant |
| Residual Al     | 0.1054      | -0.02         | 0.1097   | 0.062  | 0.012  | 0.1307       | 1           |
| Alum dosage     | 1           | -0.180        | 0.481    | -0.292 | -0.407 | 0.8536       | 0.107224    |

, RW parameters that have high correlation with respect to the output parameter are taken as input parameters, when using PCC. For example, the Pearson correlation coefficients relating Alum dosage with RW turbidity and RW color are relatively higher. Therefore RW turbidity and RW color are considered as input parameters to the model predicting Alum dosage. This result obtained by PCC agrees with the combination shown in Table 5. According to Table 4, input parameters for the model predicting TW turbidity are, RW turbidity, Alum dosage and RW color, which also agree with the combination shown in Table 5. Similarly, input parameters for models predicting residual Al based on Table 4 agree with the combination shown in Table 5.

The choice of input parameters shown in Table 5 could be backed by the analytical studies done on coagulation in water treatment [8,9]. Although NOM is not measured using color, NOM is usually responsible for color. NOM affects the turbidity removal as it interferes with the coagulation process by forming complexes with Alum, thus explaining the influence of RW color in models predicting TW turbidity and Alum dosage. However this cannot be confirmed as corresponding data on Total Organic Carbon or UV-254 are unavailable.

Although pH is considered an important parameter affecting turbidity removal, residual aluminium concentration and alum dosage [8-11], the models developed do not demonstrate such a correlation. The lack of influence of pH on the models developed in this work, could be explained by how pH affects the coagulation mechanism of a coagulant. Charge neutralization is the dominant mechanism for alum at pH less than 6.5, at which the form of the prevalent monomer species highly depend on the pH, and therefore majorly affect turbidity removal and alum demand. The dominant mechanism switches to sweep flocculation at pH greater than 6.5 due to hydrolysisation of monomer species [12], during which the effect of pH on alum demand and turbidity removal is lesser. According to Table 1, the pH range of the water for the duration considered was 6.5 to 7.5. Therefore, sweep flocculation can persist throughout the range of data considered. The decisive value of pH affecting the residual Aluminium content is 6 [15], as Al exists in soluble forms at pH less than 6. At higher pH, the solubility of Al declines and it is easier to reduce the residual Al concentration.

Table 6 demonstrates the performance of the TW pH model with different combinations of input variables. It could be noted in Table 6 that the mean squared error of models developed using RW TDS and RW Alkalinity is higher than the MSE of models developed with RW turbidity and alum.
dosage. However the results of the PCC suggest that RW Alkalinity and RW TDS demonstrate a significantly higher correlation to TW pH than the others, thus revealing a limitation of PCC.

The performance of the optimum models in table 5 is shown in table 7. The correlation coefficient of the turbidity model is poorer compared to other models, as the fluctuation of RW turbidity and TW turbidity is high and therefore it is difficult to establish a solid correlation.

In order to investigate the suitability of PCC for input parameter selection, this study was extended using an exhaustive search of higher number of input variables, including treated water parameters in the choice for input parameters for the prediction of alum dosage. The results obtained are demonstrated in table 8.

It could be noted that under the combination of four variables, RW color is not considered as an input parameter, and it has been replaced by RW Alkalinity and TW turbidity. Based on Table 4, TW turbidity has a high correlation with alum dosage, but RW alkalinity shows low correlation with alum dosage. Therefore PCC was incapable of capturing the collective effect of the two variables on the alum dosage. However, it has been reported that alkalinity has a direct influence on the alum dosage [18]. The requirement for alum increases with decreasing alkalinity, due the corresponding decrease in the concentration of nucleating agents that stimulates coagulation. Therefore, a major
drawback of the PCC is its inability to capture collective effects of input parameters on the output parameter.

However, several studies had been successfully carried out using PCC for input selection [2, 3]. Table 9 demonstrates that the performance of the model developed using PCC (model 1) and the model from table 8 with four variables (model 2) only differs slightly. Therefore PCC is a reliable tool for developing models with reasonable performance.

Table 9 - Test performance of models predicting alum dosage

|               | Correlation coefficient | Mean squared error (mg/l) |
|---------------|-------------------------|---------------------------|
| Model 1       | 0.9686                  | 12.78                     |
| Model 2       | 0.9706                  | 11.69                     |

The regression plots of the two models shown in figure 2, confirms the results in table 8, as the plot of model 1 has a few more scattered points than model 2.

4. Conclusion
Input parameters for models predicting coagulant dosage, TW turbidity, TW pH and residual Aluminium were determined using an exhaustive search of input variables, where the best combination of inputs were selected by testing the performance of ANN models. Input variables selected by principal component analysis and Pearson correlation coefficient, were compared with the best combination of variables selected from the exhaustive search. Variables selected by the PCA did not agree with the best combination of variables in any model. However the inputs selected by PCC agreed with models predicting coagulant dosage, TW turbidity and residual Aluminium. It was found that PCC was not able in selecting optimum parameters for the model predicting pH. It was also shown that PCC was not able to capture the collective influences of variables on the output parameter. However, all models developed in this study are capable of predicting the optimum coagulant dosage and treated water qualities with reliable accuracy. It should be noted that these models are valid for plants maintaining pH at neutral ranges, with moderately alkaline water, the source of which is not exposed to many industrial processes.

Acknowledgement
The financial support from Universiti Sains Malaysia through postgraduate conference fund to the first author and the cooperation of Sabah Water Supply Department and LDWS for supplying the Segama Water Treatment Plant data are greatly acknowledged.

References
[1]. Joo, D. (2000). The effects of data preprocessing in the determination of coagulant dosing rate. Water Research, 34(13), pp.3295-3302.
[2]. Wu, G. and Lo, S. (2010). Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial neural network. Expert Systems with Applications, 37(7), pp.4974-4983.

[3]. Wu, G. and Lo, S. (2008). Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. Engineering Applications of Artificial Intelligence, 21(8), pp.1189-1195.

[4]. Lamrini, B., Benhammou, A., Le Lann, M. and Karama, A. (2005). A neural software sensor for online prediction of coagulant dosage in a drinking water treatment plant. Transactions of the Institute of Measurement and Control, 27(3), pp.195-213.

[5]. Hernandez H. and Le Lann M. (2006), Development of a Neural Sensor for On-Line Prediction of Coagulant Dosage in a Potable Water Treatment Plant in the Way of Its Diagnosis, Advances in Artificial Intelligence - IBERAMIA-SBIA , pp. 249-257.

[6]. R. May, G. Dandy and H. Maier, "Review of Input Variable Selection Methods for Artificial Neural Networks", 2011.

[7]. H. Maier (2004), Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters, Environmental Modelling & Software, 19(5), pp. 485-49.

[8]. A. Matilainen, M. Vepsäläinen and M. Sillanpää (2010), Natural organic matter removal by coagulation during drinking water treatment: A review, Advances in Colloid and Interface Science, 159(2), pp. 189-197.

[9]. R. Jiao, R. Fabris, C. Chow, M. Drikas, J. van Leeuwen and D. Wang (2016), Roles of coagulant species and mechanisms on floc characteristics and filterability, Chemosphere, 150, pp. 211-218.

[10]. Z. Yang, B. Gao and Q. Yue (2010), Coagulation performance and residual aluminum speciation of Al2(SO4)3 and polyaluminum chloride (PAC) in Yellow River water treatment, Chemical Engineering Journal, 165(1), pp. 122-132.

[11]. A. Robenson, S. Shukor and N. Aziz (2009), Development of Process Inverse Neural Network Model to Determine the Required Alum Dosage at Segama Water Treatment Plant Sabah, Malaysia.

[12]. C. Kim and M. Parnichkun, MLP, ANFIS, and GRNN based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant, Journal of Water Supply: Research and Technology - Aqua, vol. 66, no. 1, pp. 49-61, 2016.

[13]. E. Towler, B. Rajagopalan and R. Summers (2009), Using Parametric and Nonparametric Methods to Model Total Organic Carbon, Alkalinity, and pH after Conventional Surface Water Treatment, Environmental Engineering Science, 26(8), pp. 1299-1308.

[14]. F. N. Valentin (2009), An Hybrid Neural Network Based System for Optimization of Coagulant Dosing in a Water Treatment Plant, Citeserx.ist.psu.edu.

[15]. K. Griffiths and R. Andrews (2011), The application of artificial neural networks for the optimization of coagulant dosage, Water Science & Technology: Water Supply, 11(5), p. 605.

[16]. M. Kennedy, A. Gandomi and C. Miller (2015), Coagulation modeling using artificial neural networks to predict both turbidity and DOM-PARAFAC component removal, Journal of Environmental Chemical Engineering, 3(4), pp. 2829-2838.

[17]. S. Dharman, V. Chandramouli and S. Lingireddy (2012), Predicting Total Organic Carbon Removal Efficiency and Coagulation Dosage Using Artificial Neural Networks, Environmental Engineering Science, 29(8), pp. 743-750.

[18]. J. Gregor, C. Nokes and E. Fenton (1997), Optimising natural organic matter removal from low turbidity waters by controlled pH adjustment of aluminium coagulation, Water Research, 31(12), pp. 2949-2958.