Employing data-driven models in the optimization of chemical usage in water treatment plants

Kandris K.[1,*], Romas E.[1], Tzimas A.[1], Gavalakis E.[1]

[1] Emvis Consultant Engineers SA, Paparrigopoulou 21, Ag. Paraskevi 153 43, Greece

*corresponding author e-mail: kkandris@emvis.gr

Abstract
One of the most challenging tasks in potable water production is the cost-efficient and consistent operation of water treatment plants (WTPs) that treat raw water of variable quality and quantity. To increase process stability and optimize the usage of resources, two data-driven models simulated coagulation in two WTPs. The data-driven models were successfully trained on monitoring data collected from the two WTPs (mean errors of effluent turbidity were below 0.5 NTU in both case studies) and were subsequently employed in the optimization of two historical periods of the WTPs. During this model-based backtesting of the WTPs, multiple operating scenarios were investigated on a daily time step in search of chemical doses that deliver a quality threshold for treated water at the minimum usage of chemicals. Results from the application of this model-based approach for WTP optimization indicated that a reduction of chemical costs equal to 6 % and 8 % would be probable for the two case studies respectively, without hampering the efficiency of raw water treatment. This work underscores that the large quantity of passive data that are amassed daily during the operation of WTPs can be turned into actionable intelligence that supports decision-making and enhances adaptive planning for water utility operators.
Keywords: Water treatment optimization, Data-driven modelling, Water treatment plant

1. Introduction

Optimization of coagulant doses is fundamental in water treatment plant (WTP) operation, as insufficient doses result in undesirable treated water quality, while high doses result in high treatment costs and potentially in water quality problems related to increased levels of residual aluminum.

Typically, jar tests are employed in the optimization of coagulant doses. Nevertheless, jar testing is time-consuming and, thereby, does not facilitate prompt responses to changes in raw water quality (Bertone et al., 2015). Modeling overcomes this drawback of jar testing.

In this work, data-driven algorithms were employed as inverse process models of the coagulation treatment process in order to specify the optimum coagulant dose. The inverse process model considered all available process inputs, the desired value of the process output (treated water turbidity) and searched for the optimal process control parameter, i.e. the coagulant dose, subject to operational constraints and standards of effluent water quality.

An indispensable constituent of model-based optimization is a model of coagulation that reproduces the observed efficiency of turbidity removal under varying raw water characteristics, distinctive coagulant types and different coagulant doses. Two candidate data-driven algorithms were trained to capture the complex and nonlinear relationships between physical, chemical and operational parameters of coagulation: random forests (RFs) and Gaussian Process regression (GPR) models. The best-performing algorithm was subsequently employed in a hindcasting rationale to pinpoint potential cost-saving treatment options in two WTPs located in Simbirizzi (Italy) and Aposelemis (Crete).

2. Development of data-driven models of coagulation

The development, training and validation of the data-driven models of coagulation was realized in four steps: (1) data collection, (2) data manipulation, (3) training and cross-validation, and (4) model selection.

Data collection (step 1) integrated available data into a single data set that were readily employed in model training. For the Simbirizzi WTP, daily measurements of flow rates, turbidity, pH and chemical dosing were available over a 24-month long period. For the Aposelemis WTP, flow rates, temperature, turbidity, pH, and chemical dosing were collected by a SCADA system in 5-minute intervals for a 14-month period.

Data manipulation (step 2) consisted of (a) filling information gaps, and (b) “cleaning” noisy data. Piecewise cubic spline interpolations with a window of one time-step filled the missing information in the training data set. Hampel filters were applied for noise removal using a one time-step window and a threshold of three standard deviations.

Model training and cross-validation (step 3) were jointly performed in a 10-fold cross-validation scheme. Two candidate models were trained and validated on the existing data, an RF and a GPR model. RFs produce non-linear functions from the mean response of ensembles of weak decision trees that are trained on random sub-samples of the training dataset (Kehoe et al., 2015). GPR models are nonparametric kernel-based probabilistic models that can be equivalent to any order of polynomials and, thus, they are appropriate for highly-nonlinear

functions with multiple extremes (Rasmussen, 2003). For both candidate models, a grid search method was used for their optimal formulation: (a) for RFs the size of the ensemble, the number of observations per leaf and the number of predictors at each node were investigated, (b) for GPR models we searched for the kernel function, the basis function, the noise variance and the hyperparameters that produce the best fit to the observed data.

Ultimately, model selection (step 4) was based on two performance metrics, the coefficient of determination (R^2) and the mean absolute errors (MAEs). The best-performing model was employed as an inverse process model to search for the minimum coagulant doses that produce the target effluent turbidity, which was equal to the mean observed effluent turbidity of the two WTPs, i.e. 3.0 NTU for Simbirizzi and 1.3 NTU for Aposelemis.

3. Results and Discussion

Model training and validation indicated that GPR models and RFs were equally competent in predicting effluent turbidity in the two WTPs, but GPR models were marginally better (data not shown). In addition, GPR models provide a robust assessment of prediction uncertainty. Thus, they were subsequently used in process optimization. For the Simbirizzi WTP, R^2 for the RF model was 0.52, while the respective value for the GPR model was 0.54. The latter accomplished a MAE equal to 0.45 NTU. For the Aposelemis WTP, 72% of the variability of the observed effluent turbidity was captured by the GPR model (R^2 = 0.72); the RF model accomplished an R^2 of 0.70. The MAEs were 0.23 NTU and 0.25 NTU, respectively.

Inverse process modeling in the Simbirizzi WTP indicated that coagulant costs could have been 8% less for the historical 24-month period examined herein without a decline in coagulation performance (see Figure 1). In absolute terms, this reduction is equal to 57,300 kg poly-aluminum chloride (PACl) and corresponds to 1,590 kg of CO_2 emissions. Model application proposed coagulant doses like those used in the WTP for most of the hindcasting period. Yet, for 37 days a cost reduction higher or equal to 500 kg/d could have been achieved. Cost reduction was achieved in two ways. On one hand, model application detected coagulant doses that achieve better turbidity removal combined with reduced cost. On the other hand, model application suggested doses that offered an acceptable performance decline combined with an inversely proportional cost reduction.

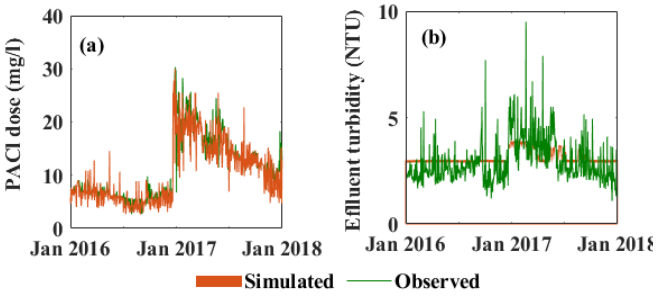

Figure 1. Observed versus simulated (a) coagulant dose and, (b) effluent turbidity for the coagulation stage of Simbirizzi WTP for a target effluent turbidity of 3.0 NTU. Simulated turbidities are depicted as intervals of 95 % confidence.

In Aposelemis WTP, model application revealed that a chemical cost reduction of 6% could have been achieved for the 14-month period considered. Potential reduction of chemical dosing corresponds to nearly 300 kg of CO_2 emissions. The difference between model-based and observed doses is attributed to the potential reduction of alum doses (Figure 2c and 2d). Inverse process modeling proposed 16% lower alum doses, for the days that alum was supplied as a coagulant. The potential benefit achieved from alum optimization was mitigated when PACl was used as coagulant: modeling proposed nearly 5% higher PACl dosing compared to the actually added doses (Figure 2a and 2b). Nonetheless, increased dosing offered higher turbidity removal.

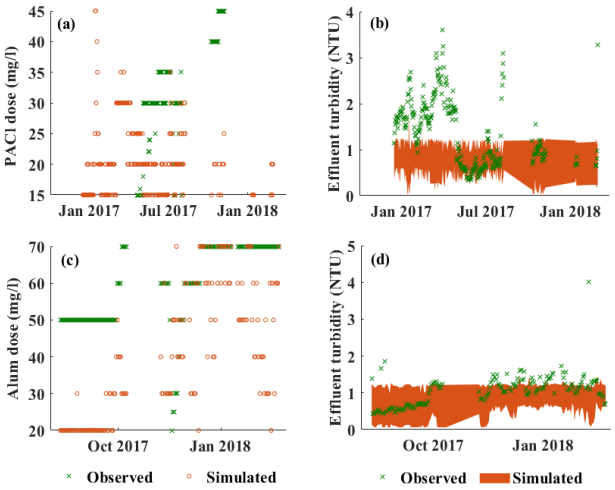

Figure 2. Observed versus simulated coagulant doses (a, c) and effluent turbidity (b, d) for the coagulation stage of Aposelemis WTP for a target effluent turbidity of 1.3 NTU. Simulated turbidities are depicted as an interval of 90 % confidence.

4. Conclusions

Data-driven models, i.e. GPR models and RFs, predicted satisfactorily effluent turbidity following coagulation under diverse coagulant doses, types and raw water characteristics in two WTPs. Mean turbidity errors were lower than 0.5 NTU, constituting, thus, model-based predictions adequate.

The GPR-based models of coagulation were employed in an inverse mode for the optimization of the respective process of Simbirizzi and Aposelemis WTPs. Model-based optimization revealed that chemical usage can be reduced without hampering the efficiency of turbidity removal by the coagulation process, even in WTPs that already function in an optimization rationale.

This work underscored that the large quantity of passive data that are amassed daily during the operation of WTPs can be turned into actionable intelligence that supports decision-making and enhances adaptive planning for water utility operators. The modeling approach followed herein can be a part of a tool that improves chemical dosing and safeguards against spikes of critical water quality parameters. If this tool is integrated with forecasted raw water quality, it can enhance preparedness

of the water utility operator against possible changes in the water quality influx.

Acknowledgements

This work was funded as part of the SPACE-O project which has received funding from EU H2020 Research & Innovation Programme under GA No. 730005.

The authors kindly thank G. Zedda and S. Gyparakis for the provision of operational data from the Simbirizzi and Aposelemis water treatment plants, respectively.

References

Bertone, E., Stewart, R. A., Zhang, H., & O'Halloran, K. (2016). Hybrid water treatment cost prediction model for raw water intake optimization. *Environmental modelling & software*, **75**, 230-242.

Kehoe, M. J., Chun, K. P., & Baulch, H. M. (2015). Who smells? Forecasting taste and odor in a drinking water reservoir. *Environmental science & technology*, **49**(18), 10984-10992.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. *In Summer School on Machine Learning* (pp. 63-71). Springer, Berlin, Heidelberg.