NEURAL NETWORK TO ANALYZE WASTEWATER TREATMENT PLANT WITH CEPT

Signe Moe[1,5,*], Bård Myhre[2], Anne Marthine Rustad[1], Herman Helness[3], Frank Batey[4]
[1]Dept. of Mathematics and Cybernetics, SINTEF Digital, Oslo, Norway
[2]Dept. of Connectivity Technologies and Platforms, SINTEF Digital, Oslo, Norway
[3]Dept. of Infrastructure, SINTEF Community, Trondheim, Norway
[4]Engineering Services Dept., Trondheim Municipality, Trondheim, Norway
[5]Dept. Engineering Cybernetics, NTNU, Trondheim, Norway

Abstract Overview

Wastewater treatment is a complex process that is difficult to model and optimize. Operational data from the wastewater treatment plant Høvringen (HØRA) in Trondheim, Norway (a chemically enhanced primary treatment (CEPT) plant with coagulation, flocculation and particle separation by settling) has been used to train and test an artificial neural network (ANN) to predict the turbidity of the treated water given relevant input. Suspended solids (SS) content causes cloudiness of a fluid and can therefore be used to determine treatment efficiency through turbidity measurements. The main finding is that the ANN is able to learn how process input parameters influence the quality of the treated wastewater, and that combining machine learning with a technical understanding of the system provides valuable insight into operating wastewater treatment plants. Chemical dosage is currently determined by the inflow rate to the plant. However, it is known that flow pattern, reject flows and influent wastewater quality are important. The importance of these key factors were confirmed in the training process, which indicates that chemical dosage could be determined as a function of 1) SS concentration at the plant inlet, 2) the average hydraulic retention time (V/Q), 3) averaging the inflow parameters over a corresponding number of hours, and 4) whether or not the reject water pumps are on or off. Use of ANNs to design more advanced process control systems has the potential to increase chemical dosage precision and robustness of chemically enhanced primary treatment towards influent variations. This will be important for the degree of water purification and may lower the consumptions of chemical and thereby operational costs.

Keywords

Machine Learning, Artificial Neural Networks, Wastewater Treatment

Introduction

The modelling of water treatment processes is challenging because of its complexity, nonlinearity, and numerous contributory variables. Furthermore, it is of importance since poor water treatment has negative impacts on society and the environment. The overall scope of this work was to investigate whether machine learning (ML) can contribute to optimized control and operation of wastewater plants. Due to their ability to learn complex functions from data ML techniques are highly relevant for process industries such as water supply and treatment. For instance, the influences of various parameters on concentrations of faecal indicator organisms and other water quality parameters have been investigated using regression methods (Black et al., 2007; Juntunen et al., 2012) and ANNs (Singh et al., 2009). Wastewater treatment plans are required to ensure a certain treatment efficiency. For CEPT

[*] To whom all correspondence should be addressed

plants, this key performance indicator is determined by comparing SS concentration at the in- and outlet of the plant or process step. In this project, operational data was used to train and evaluate the performance of an ANN predicting the turbidity of treated wastewater given relevant input data. The aim of the study was thereby to investigate if ANNs have potential for application in process control.

Treatment Process at HØRA

The HØRA plant is equipped with several sensors logging measurements every two minutes. The dataset used for training spans the period 04.01.2018-01.21.2019 and consists of roughly 214 000 points.

Wastewater from a sewage and drainage system covering 95 km^2 and roughly 150 000 inhabitants is pretreated by grit removal and screens before the CEPT process. The turbidity of the untreated water $turb_{in}$ [g/m^3] is measured before the screen. If the capacity of the plant is exceeded, the excess water is discharged to the Trondheim fjord through three overflows, $Q_{overflow,1\text{-}3}$ [m^3/h]. In the plant, the water flow $Q_{A\text{-}D}$ [m^3/h] is treated in four parallel lines where chemicals (polymers) $Pol_{A\text{-}D}$ [L/h] are added to enhance coagulation and flocculation. The wastewater then proceeds through a three chambered flocculation step before entering the sedimentation step where particles are removed. The turbidity of the treated water $turb_{out}$ [g/m^3] is measured and the water discharged. The solids proceed to sludge treatment, where additional water is removed, stored in tanks with level h_{dec} and h_{flow} [m] and fed back to the plant inlet by pumps $p_{dc,1\text{-}2}$, $p_{fw1\text{-}2}$, $p_{dw,1\text{-}2}$.

Training and evaluation

Of the total dataset, 3% was separated and used as test data. The test data was kept in chronological order around a third into the total set, thereby spanning roughly nine days in July 2018. Similarly, 3% was used as validation data. These data points were randomly sampled from the entire dataset (excluding test data) and used by the model design phase to test different parameters and model options. After a grid search to determine hyperparameters, an ANN architecture of 100, 100 and 10 neurons and an L1 regularization constant of 0.03 was chosen. Training was performed using the Adam optimizer, 32 epochs and a batch size of 128 and a mean square error loss.

Initially, the measurements described in the above section were modelled directly. However, this resulted in a very poor prediction model subject to noisy and inaccurate estimates. A hypothesis was constructed based on knowledge of the system: The measured inputs at a specific time are not related to $turb_{out}$ at the same time – there is a certain residence time in the flocculation and sedimentation basins. Hence, a spearman correlation analysis was conducted to find the connection between $turb_{out}$ now and inputs measured n hours ago. Due to the discrete nature of the pump data these were omitted from the analysis. The results are shown in Figure 1 and the following conclusions were drawn. 1) The input which correlates most with $turb_{out}$ is $turb_{in}$. Thus, the most important factor for how clean the treated water is, is how dirty it is when entering the plant. 2) There is no clear correlation or pattern related to overflow measurements. This is natural due to the fact that overflow rarely happens. 3) Input measured between 0-8 hours ago has the highest correlation with the current turbidity out. This suggests that water is mixed well in the flocculation and sedimentation tanks: some water flows through quickly while some has a higher residence time of several hours. 4) There is a periodic oscillation in the correlation analysis of 24 hours. This is related to the relatively stable domestic water use, which peaks in morning and afternoons and is easily detected in inflow measurements.

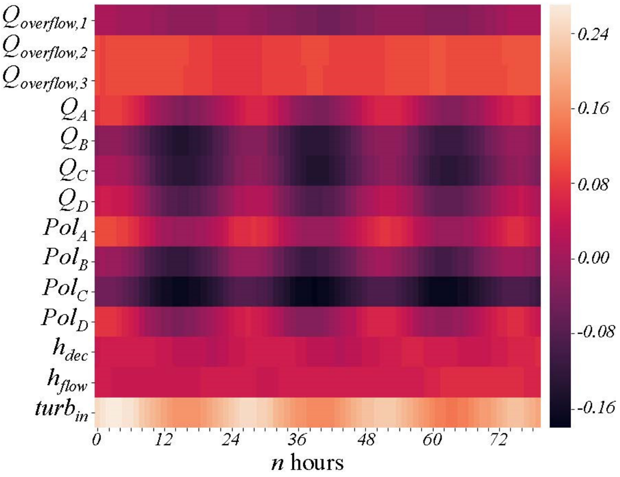

Figure 1. Correlation (scale to the right) between the plant inputs (y-axis) at a specific time and the turbidity out n *hours (x-axis) later?*

The input to the ANN was therefore defined as the average over the last n hours rather than the current measurements. The final train, validation and test loss were assessed for values of n up to 24 hours. The train and validation loss were comparable for all values of n, indicating no overfitting. Furthermore, they were quite constant and low for $n \geq 8$ hours, indicating that including a larger "memory" in the data does not result in a better prediction model. This also corresponds well with the correlation analysis. However, for test data, the final loss varied and there was no value for n which resulted in a lower or more consistent loss than others. Since the test data is chronologically picked from within the data set, it can be concluded that the inputs in this period do not closely resemble the inputs the network was trained on. Hence, the model performance on these data is quite random since it is attempting extrapolation, a known weakness of neural networks. In fact, the available dataset does not span an entire year, which would be necessary to capture the natural seasonal changes.

The results of $n = 8$ hours are shown in Figure 2 for the entire dataset. The prediction model has good performance and is able to predict $turb_{out}$ fairy accurately. The performance is equally good on validation as training data. In Figure 2, the model also performs well on test data.

Feature importance analysis

Based on system knowledge, Q_{A-D}, Pol_{A-D} and $turb_{in}$ were considered essential information for estimating $turb_{out}$. However, it was desirable to investigate the effect of the remaining inputs. For instance, there is feedback from the sludge treatment where water with a high concentration of dirt is extracted and fed back into the water treatment. To gain more insight into the individual importance of these inputs, new models with the same architecture were trained as described in the previous section. However, the features related to overflow and sludge treatment have systematically been removed to investigate which (if any) are essential for training a good prediction model for $turb_{out}$. The results are shown in Figure 3.

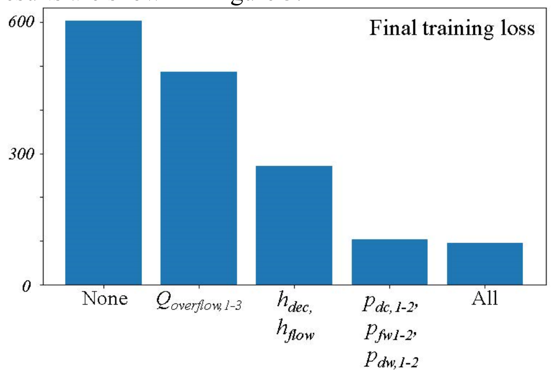

Figure 3. Feature importance analysis. Final training loss for ANN where none, overflow, level, pump status and all of the above are included as model inputs.

Clearly, by omitting all of them the resulting model is much less accurate. Similarly, the train loss decreases when including overflow or level measurements as inputs but is still significantly higher than the baseline of including all available information. However, completely omitting overflow and level and only including pump status results in a train loss of the same magnitude as including all. Thus, it can be confirmed that the feedback from the sludge treatment greatly affects $turb_{out}$.

Conclusion

This extended abstract has presented the training and evaluation process of using an ANN to estimate the outlet turbidity of HØRA to investigate if an ANN could be used predict outlet wastewater quality and thereby have potential for application in process control. The following conclusions have been drawn. 1) The variables identified by the ANN and which could predict the outlet turbidity were those expected from system knowledge. 2) There is a certain residence time in the tanks which must be accounted for in the prediction model to get accurate results. Correlation analysis indicate 0-8 hours. 3) The feedback from sludge treatment, run by pumps, affects the outlet turbidity to a large extent. In addition, the inlet turbidity is the input parameter with the highest correlation with outlet turbidity.

Polymer is currently added based on the wastewater inflow to the plant. Given the above findings, we recommend investigating the effects of basing it on 1) the average inflow parameters the last ~8 hours, and 2) whether or not the reject water pumps are on or off, and/or 3) measured inlet turbidity in.

These measures may result in a higher degree of water purification and/or a lowered consumption of polymer.

Acknowledgments

This work was supported by the Regional Research Fund of Mid-Norway (RFFMIDT), project 285262.

References

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. Applied and env. microbiology, 73(12), pp. 3965-3974.

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 Comp. Intelligence and Soft Computing, vol. 2012.

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.

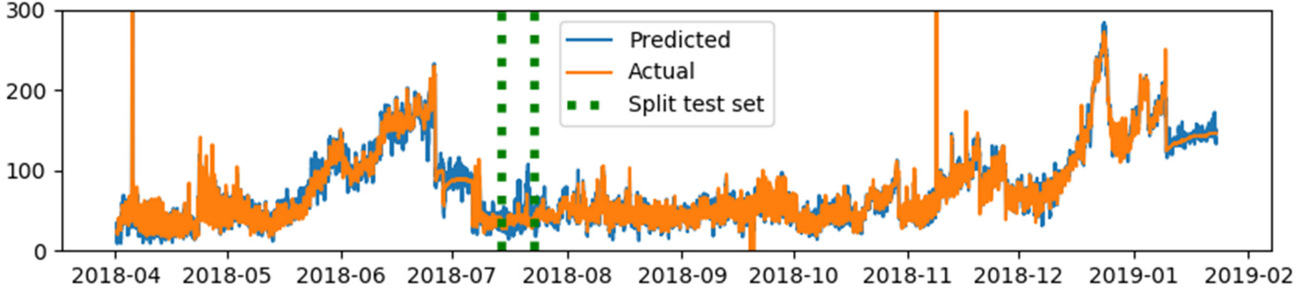

Figure 2. Predicted and actual $turb_{out}$ for entire dataset. The model performs equally well on train and validation data. Test data is picked chronologically from the middle of the data set. Here the performance varies as the training data does not span the same input space.