MACHINE LEARNING AS A DECISION SUPPORT TOOL FOR WASTEWATER TREATMENT PLANT OPERATION

THIBAULT MERCIER[1], ABEL DEMBELE[1], THIERRY DENOEUX[2] & PASCAL BLANC[1]
[1]Suez Smart Solutions, France
[2]Université de Technologie de Compiègne, CNRS, Heudiasyc, France

ABSTRACT

Wastewater treatment is a significant environmental challenge. It is also an economic challenge for all operators, who face more and more demanding national and supranational regulations. Optimizing wastewater treatment processes requires physical, biological and chemical models with various degrees of complexity. From an operational perspective, programmable logic controllers are generally used. Those controllers follow strategies implemented by technicians with various degrees of expertise. This may lead to over- or under-aeration, which can be very costly. Commonly used strategies are mostly based on business rules and expert guidelines, which do not necessarily consider specific operating conditions. In this study, focused on the aeration process, a machine learning approach is applied to predict the daily operating time of aerators. Two types of models, according to the data considered, have been evaluated. The first model considers only the operation data as explanatory variables (pollutant concentrations and inflow), while the second model includes exogenous weather data (temperature, hygrometry, rainfall depth). The best model reaches a mean error less than 1%.

Keywords: machine learning, wastewater treatment, aeration process, model verification.

1 INTRODUCTION

This study is part of a project aiming to optimize wastewater treatment processes. It is based on machine learning techniques. This article covers the aeration process, which consists in bringing oxygen to micro-organisms (bacteria) while brewing water to reach homogenisation. The whole process is based on heterotrophic bacteria, which consume organic carbon for their anabolism and catabolism. This biological procedure, called nitrification, ends up producing nitrate (NO_3). A second phase in anaerobic environment reduces nitrate into dinitrogen (N_2). To optimize the process, it is recommended to go through several aerobic and anaerobic phases [1]. Aerators bring the oxygen to the pool; they are activated and deactivated several times in a day to alternate between phases. The control of this process is, in some cases, performed by technicians responsible for pump management, often without any specific guidelines. This practice leads to over-aeration or under-aeration problems, depending on the technician experience. This article aims to propose a machine learning algorithm to predict the daily aeration time. The goals are to harmonize the aerator operation, to minimize unnecessary fluctuations and to understand, through machine learning, how daily operational decisions are made. Expected benefits include, firstly, to avoid extreme aeration time and, secondly, to reduce energy consumption while meeting operational quality standards. To achieve this goal, statistic and machine learning-based approaches have been applied using both operation and weather data. The following sections describe the experimental site, the data, the modelling approaches and the preliminary results of this study.

2 DATA AND EXPERIMENTAL SITE

The data used in this investigation concern an experimental site located in the south east of France. With less than 50,000 population equivalent, it can be considered as a small or medium-size plant. For confidential reasons, the name and the exact location will not be

doi:10.2495/WRM190101

disclosed. Collected data are composed of a three-year historical record of operation and weather data. The list of parameters and data quantity are given in Table 1. Fig. 1 presents the observed aeration time over the period, expressed as a fraction of the observed mean time. The behaviour seems to follow some seasonal trends, but the behaviour looks otherwise quite erratic. This observation is the source of one of the objectives: to smooth the aeration process operation.

Table 1: List of variables and data quantity.

	Variable	Number of values
Operational data	Inflow (m^3/day)	1,096
	BOD_5 (kg/day	72
	COD (kg/day)	156
	Nitrogen (kg/day)	70
	TSS (total suspended solid) (kg/day)	156
	Phosphorus (kg/day)	72
Weather data	Rainfall depth (mm/day)	1,096
	Temperature (°C)	1,096
	Sunshine duration (h/day)	1,096

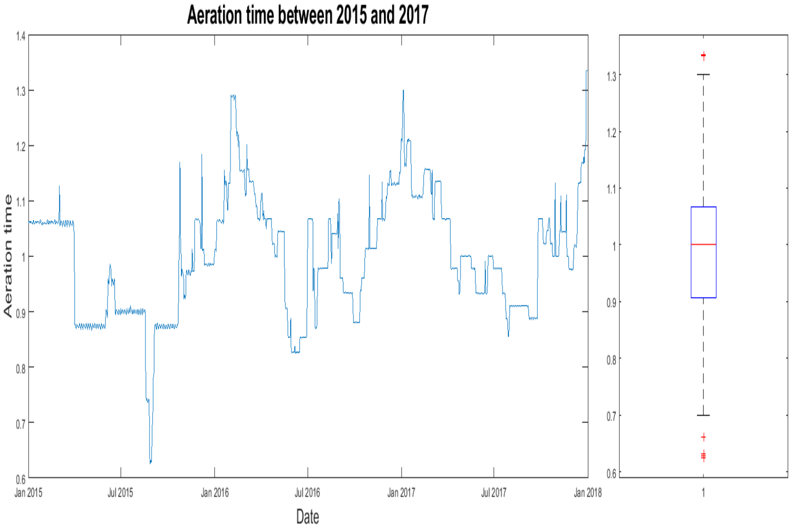

Figure 1: Observed aeration time between 1 January 2016 and 31 December 2017 (left) and data distribution (right).

Before training models, a pre-processing step was carried out. *Aquadvanced wastewater plant* is a software tool developed by Suez Smart Solution to help wastewater plant manager in their daily operations. Aquadvanced extracts data from a database supplied by several sensors or users. Some of these data were erroneous and some corrections were needed. For example, two values were set to 0 for the nitrogen inflow. Automatic outlier detection and data imputation are among our future objectives. At that time, we were not able to reconstruct these data and they were simply ignored.

3 MODELLING AND RESULTS

3.1 Modelling

Many regression models have been trained. A not exhaustive list is presented hereafter:

- Constant: the mean value of the available observations is taken as prediction value.
- Full regression: Linear regression using all the data of available variables.
- Restrained regression: Linear regression using only variables with significant p-values.
- Quadratic restrained regression: Quadratic regression using variables from the restrained regression.
- Principal component analysis (PCA) regression: Linear regression using the principal components (95% of explained variance).
- Multi-model regression: Linear regressions over separated clusters identified by using the k-means algorithm.
- Boosted trees: Aggregation of weak learners (regression trees).

These models are based on operational data. They are referred to as *operational models*.

Weather data have been used differently. An algorithm that computes a mean weighted by the similarity between observations has been implemented. The observations may be represented in a p-dimensional space (where p is the number of variables). In this space, we can compute a measure of distance between the learning set and the explicative variables. With this distance, we are then able to compute similarity. Hereafter, these models will be referred as *meteorological models*. These models could model seasonal trends but not very local behaviour. To increase robustness and solve this problem a temporal similarity was introduced. A variable representing the elapsed time between observation is added to the model to compute this similarity. The resulting new models will be identified as *historical models*.

3.2 Results and discussion

The performances of the operational models are presented in Table 2. The normalized mean errors are between 9.56% and 7.75%. These errors have been computed by cross validation [2]. For the *operational models* 5 folds were used, while, for *meteorological models* and *historical models*, a leave-one-out cross validation was applied.

Table 2: Performances of *operating models*.

Model	Observed error
Constant	9.00%
Full regression	9.14%
Restrained regression	8.48%
Quadratic restrained regression	8.09%
PCA regression	9.38%
Multi model regression	9.56%
Boosted trees	7.75%

The most efficient *operating model* was the boosted regression trees algorithm with a normalized mean error of 7.75%. The values for some functional variables were not available daily. Thus, it led us not to include them systematically, in the most complex models. This had probably an impact on performances of the functional models. The k-means algorithm identified two classes based on the inflow. This result, which was predictable as the inflow has a large variance, do not lead to good performances. A possible reason is that the inflow might be, for a given wastewater treatment plant (WTP), nearly seen as a constant. Consequently, the variation does not have a significant effect on aeration time.

Including weather variables, we could use our similarity algorithm with different subsets of variables. The most efficient model used less than five variables. Among these variables, we find temperature, which is known to have a significant influence on plant operation [3], [4]. These *meteorological models* reached a cross-validation error of 6.69%. Fig. 2 presents the results over the year 2017. In this paper, the aeration time is expressed as a fraction of the mean time.

Finally, addition of a time component variable allows us to further increase the model's performances to a mean error of 0.96% (Fig. 3). The maximal error of 17% was obtained for a day where the observed aeration time was 40% under the mean. This anomaly could not be imputed to a failure of the aerators, nor to a high variation in other variables at our disposal. We have for sure no explanation about this very particular observation. Thus, in nominal operation, the performances of this model can be considered as satisfactory.

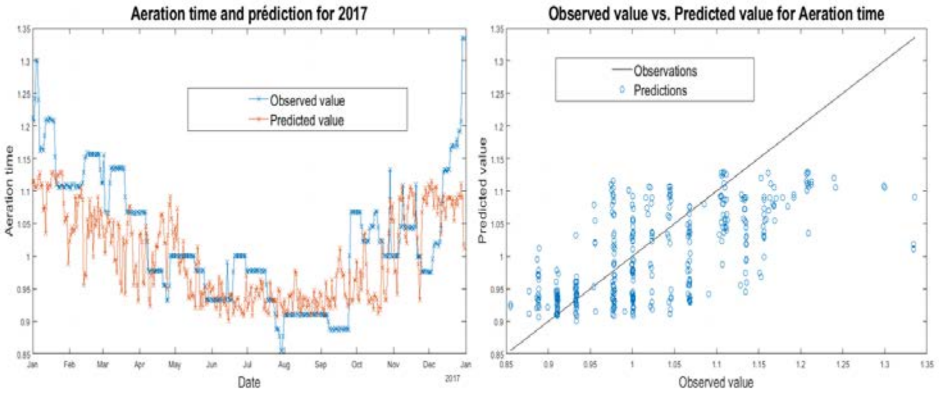

Figure 2: Result of a *meteorological model* for 2017.

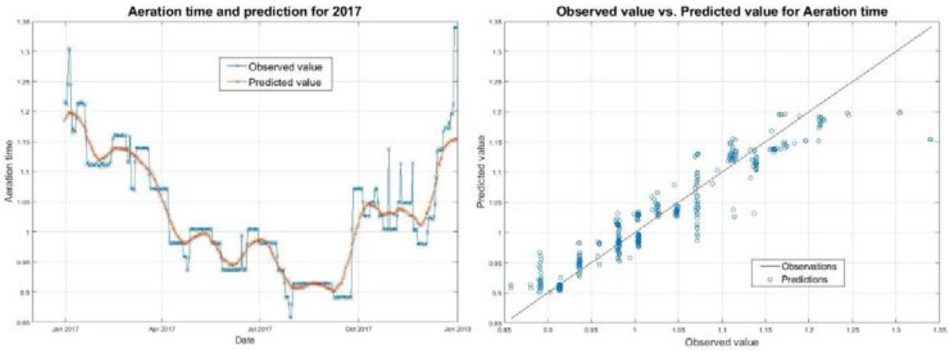

Figure 3: Results of a *historical model* for 2017.

4 CONCLUSION AND PERSPECTIVES

During this study, several statistical and machine learning approaches have been assessed. We could evaluate three different kinds of models (operational, meteorological and historical) based on various variables. The best model uses weather data, in addition to the operational data, to predict the daily aeration time. The achieved prediction error is less than 1%, which is acceptable for this application. Furthermore, we achieved two other objectives, which were: to smooth aeration daily time and to learn how choices are made by operators. Indeed, after asking them how they made operational decisions, they answered that they were based mainly on weather parameters. This methodology must be experimented on each wastewater treatment plant as each plant's behaviour might not be based on the same variable.

We may expect that the plant operation is (highly) dependent of non-numeric considerations such as distinction between working day or holiday. This assumption will lead us to add categorical data in future models. We will, then, must determine a way to compute similarities [5], [6] on categorical data, and how to weight this similarity with numerical similarity as computed in this study.

ACKNOWLEDGEMENTS

This study was founded by the Suez Smart Solutions. The data has been made available thanks to the Aquadvanced® Plant team.

REFERENCES

[1] Boeglin, J.C., Traitements biologiques des eaux résiduaires. *Techniques de l'ingénieur. Génie des procédés*, **4**, J3942, 1-J3942, 28, 1998.

[2] Arlot, S., A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, pp. 40–79, 2010.

[3] Grenier, M., Bonsteel, J., Lai, G. & Perry, L. Optimizing small/medium water treatment plants for turbidity removal. www.xcg.com/wp-content/uploads/2014/03/6.2.1-OWWA-2007-Conference-Paper-Optimizing-Sm_Med-WTPs-Turbidity-Removal.pdf. Accessed on: 20 Feb. 2019.

[4] Guo, H. et al., Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences*, **32**, pp. 90–101, 2015.

[5] Boriah, S., Chandola, V. & Kumar, V., Similarity measures for categorical data: A comparative evaluation. *Proceedings of the SIAM International Conference on Data Mining, SDM 2018,* Atlanta, Georgia, USA, 24–26 April, 2018.

[6] Jia, H., Cheung, Y.M. & Liu, J., A new distance metric for unsupervised learning of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, **27**, pp. 1065–1079, 2015.