Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 52-23 (2019) 37–42

Next Generation Sequencing Analysis of Wastewater Treatment Plant Process via Support Vector Regression

M.A. Prawira Negara * E. Cornelissen * A.K. Geurkink *
G.J.W. Euverink * B. Jayawardhana *

* *Engineering and Technology Institute Groningen, Faculty of Science and Engineering, University of Groningen, The Netherlands (e-mail:* {m.a.prawira.negara@rug.nl; emilecornelissen@gmail.com; b.geurkink@gmail.com; g.j.w.euverink@rug.nl; b.jayawardhana@rug.nl} *).*

Abstract: In this paper, we analyze next generation sequencing (NGS) data of wastewater treatment plant (WWTP) in the North Water facility for revealing the role of 1236 different genera of microorganisms in the aeration basin to the measured process data. Both the time-series data of NGS and process parameters are pre-processed and analyzed using support vector regression technique and is compared with the deep neural network approach. Local sensitivity analysis is performed on the resulting models. Both machine learning analyses show the importance of a subset of genera to the WWTP process and can be used to enrich the well-studied activated sludge model (ASM).

Keywords: Wastewater Treatment Plant, correlation analysis, process data, NGS data.

1. INTRODUCTION

It has been recognized that wastewater forms a significant part of waste from human activities and it requires treatment before it is environmentally safe to be discharged to the natural water resources (see, for example, Henze and Comeau (2008)). When it is discharged untreated, it pollutes water resources and can lead to disastrous ecological, environmental, as well as, economical impacts. Wastewater comes from two major sources: human sewage systems and process waste from manufacturing industries. The biological treatment of wastewater was firstly introduced in the early twentieth century and has become the basis of wastewater treatment worldwide. It involves the deployment of confining naturally occurring bacteria at very high concentrations in tanks. These bacteria, together with some protozoa and archaea, are collectively referred to as activated sludge. The concept of biological treatment is fairly simple. The microorganisms process small organic compounds, ammonia, and phosphate by consuming them for their growth; thereby the wastewater is cleansed. After the clarification step (where the activated sludge is sedimented), the treated wastewater is discharged to natural water resources, such as river or sea.

While the concept is simple, the control of the treatment process is very complex because of a large number of variables that affects it (Davies (2005)). These include changes in the composition of the bacterial flora in the treatment tanks and changes in the sewage flowing into the plant. The input can show variations in flow rate, chemical composition, pH, and temperature which in turns influence the population dynamics and metabolic process of the activated sludge. Some of the plants that treat industrial wastewater may have to cope with recalcitrant chemicals that are difficult to degrade by the activated sludge. Such industrial wastewater contains toxic chemicals and has extreme properties such as low or high pH or high salt concentration that can inhibit the functioning of the activated sludge.

There are a number of established models for describing the process dynamics in WWTP such as the well-studied Activated Sludge Models from the International Water Association (see Henze et al. (2000)). These ASM models are lumped dynamical models where the influence of the activated sludge is lumped to the kinetic laws/rate of few WWTP process variables. The unknown parameters are a fusion of metabolic activities of various microorganisms and consequently they have a large degree of uncertainties. This prevents the direct use of the models for optimization and for the model-based control design of WWTP. For improving the reliability and applicability of these models, it is recognized that the knowledge and real-time information of the microorganism population are indispensable for enriching the ASM models, as concluded in Muszynski et al. (2015); Liu et al. (2016); Bassin et al. (2017). In particular, the high-throughput data of microorganisms that are obtained from the Next Generation Sequencing (NGS) or other omics tools can be used to adjust/to adapt the uncertain parameters and decrease the uncertainties of the models. The large degree of heterogeneity of microorganisms, as well as, the time-varying and highly non-linear characteristics of the WWTP process dynamics, pose some challenges in linking the real-time data on microorganisms and other relevant information to the overall process dynamics.

Towards the development of the aforementioned enriched/adaptive ASM models, we study in this paper the

2405-8963 Copyright © 2019. The Authors. Published by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2019.11.006

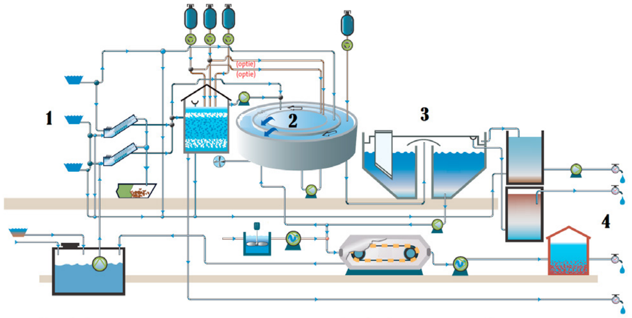

Fig. 1. Schematic diagram of the North Water's Saline Wastewater Treatment Plant, where 1) influent 2) aeration tank 3) sedimentation tank 4) effluent. The picture is courtesy of North Water.

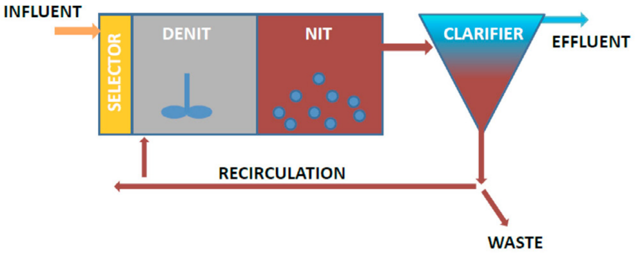

Fig. 2. Block diagram of the aeration tank and clarifier.

relation of NGS data (which are obtained from an industrial WWTP of the North Water facility located in Oosterhorn, Delfzijl, Netherlands) to the measured process variables. The analysis is based on a machine learning technique, the so-called Support Vector Regression, which can handle well high-degree of uncertainties and unstructured relation between rich set (high-degree of heterogeneity) of input and output data (with a small number of time points). Such preliminary analysis can extremely be useful to identify a subset of the microbial population that can explain most of the variations observed in the process dynamics. More importantly, a targeted low-cost and fast measurement to this population subset (such as, using the qPCR (quantitative polymerase chain reaction) technology) allows us to provide an additional feedback control loop to the process control of WWTP.

A generic schematic overview of the WWTP of North Water is shown in Fig. 1. The influent of the WWTP is industrial salt wastewater and the activated sludge processes take place in the aeration tank which is indicated by 2 in Fig. 1. In the sedimentation tank (which is given by 3 in Fig. 1) the activated sludge is separated from the water and partly fed back to the aeration tank. We refer to Fig. 2 for the simplified block diagram of the system. The treated water is then discharged into the Wadden Sea.

Next Generation Sequencing (NGS) of the 16S rRNA gene from the bacteria in sludge samples is used to analyze the microbial communities. The analysis of the NGS data shows the presence and abundance of the microorganisms in the sludge based on the millions of DNA sequence data. According to Xu (2014), the DNA sequences are compared to 16S rDNA sequences in public databases to identify the microorganisms in the sludge. Based on this comparison, an overview of the microorganisms that occur in the sludge sample is created. The frequency of the occurrence of the same sequence in the NGS dataset is related to the abundance of a specific microorganism in the sludge sample.

The rest of the paper is organized as follows. In Section 2, we review the Support Vector Regression technique that we use to explain the relation between the microorganism data and the measured process variables. In Section 3, we present the data pre-processing that we have used to both dataset. We present our SVR analysis in Section 4 where we compare it with the results that are obtained using the deep neural network. Finally, in Section 5, we present the conclusions.

2. SUPPORT VECTOR REGRESSION

Support Vector Machines (SVM) is a supervised learning method where data can either be classified according to a pre-determined set of kernel functions or be fitted to a given set of basis functions, see also, Cortes and Vapnik (1995) and Drucker et al. (1996). The latter use of SVM for regression analysis is also referred to as the Support Vector Regression (SVR) method.

In the following, we will briefly describe the SVR method where nonlinear basis functions are used. Suppose that $x_i \in \mathbb{R}^n$ with $i = 1, 2, \ldots, m$ are a collection of input vector sampled from the vector space of $\mathbb{R}^n$ and m is the total number of samples. Let $y_i \in \mathbb{R}$ be the corresponding samples of scalar output data. In SVR, the nonlinear regression between x_i and y_i assumes the following regression model

$$y_i \approx w^T \phi(x_i) + b, \tag{1}$$

where $\phi : \mathbb{R}^n \to \mathbb{R}^q$ is the $q-$dimensional kernel functions and $w \in \mathbb{R}^q$ is the parameter vector to be identified.

Using the regression model (1) and for a given q-dimensional kernel function ϕ, the SVR method looks for w and b that solve a given nonlinear programming problem. In the kernel-based regression model as in (1), we assume a linear relation between the kernel output $\phi(x_i)$ to the output scalar y_i which simplifies the regression problem as we can extend tools from linear regression techniques in a straightforward manner. The kernel choice as well as the particular selection of adjustable kernel parameters have an important influence on the performance of the kernel. For some data sets, the kernel parameters in ϕ need to be optimized and can be done along side the regression step, see Eitrich and Lang (2006). Usually, this is done by performing a grid search, where all possible combinations of parameters are investigated according to Hsu et al. (2010).

A loss/error function is typically introduced to quantify the regression performance. It gives a distance measure between the approximated output $w^T \phi(x_i) + b$ to the measured output y_i. In a standard regression problem, this loss function is given by a quadratic function. However, in SVR, an ϵ-insensitive loss function is used instead. Roughly speaking, this function is equal to zero when the error is within a range of ϵ and is equal to the 1-norm of the error minus ϵ otherwise. In addition to this, in order to guarantee the feasibility of the optimization problem, two slack variables are introduced in our analysis, see for instance, Scholkopf and Smola (2002).

Thus, the corresponding nonlinear regression problem of (1) is given by the following minimization problem where both the ϵ-insensitive term and slack variables $\xi, \hat{\xi}$ are used.

$$\underset{w,\xi,\hat{\xi}}{\text{minimize}} \left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) \right] \tag{2}$$

subject to

$$\begin{cases} y_i - w^T\phi(x_i) - b \leq \epsilon + \xi_i & i = 1, \ldots, n \\ w^T\phi(x_i) + b - y_i \leq \epsilon + \hat{\xi}_i & i = 1, \ldots, n \\ \xi_i, \hat{\xi}_i \geq 0 & i = 1, \ldots, n. \end{cases} \tag{3}$$

As a dual problem to the above minimization problem, we can reformulate it by introducing Lagrange multipliers α and α^* as follows

$$\underset{\alpha,\alpha^*}{\text{minimize}} \left(\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\phi(x_i)^T\phi(x_j) + \epsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \right) \tag{4}$$

subject to

$$\begin{cases} 0 \leq \alpha_i^* \leq C \\ 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0. \end{cases} \tag{5}$$

In this case, the identified weight/parameter vector w is given by

$$w = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\phi(x_i). \tag{6}$$

Consequently, when a new data x is given, we can predict its output by computing

$$y = w^T\phi(x) + b \tag{7}$$

where b is computed using the Karush-Kuhn-Tucker (KKT) complementarity conditions. The KKT complementary condition we used are

$$\begin{cases} \alpha_i(\epsilon + \xi_i - y_i + \phi(x_i)) = 0 \\ \alpha_i^*(\epsilon + \xi_i^* + y_i - \phi(x_i)) = 0 \\ \xi_i(C - \alpha_i) = 0 \\ \xi_i^*(C - \alpha_i^*) = 0. \end{cases} \tag{8}$$

Hence b can be computed as follows

$$\begin{aligned} b &= y_i - w^T\phi(x_i) - \epsilon \quad \text{for} \quad 0 \leq \alpha_i \leq C \\ b &= y_i - w^T\phi(x_i) + \epsilon \quad \text{for} \quad 0 \leq \alpha_i^* \leq C \end{aligned} \tag{9}$$

3. DATA PROCESSING

3.1 NGS Data

The NGS dataset, taken from the WWTP in Oosterhorn, was made available for this paper by BioClear. The NGS analysis of 32 active sludge samples was performed for 2.5 years, starting in week 42, 2014 and ending in week 6, 2017.

The dataset was arranged in multiple sheets of data, each sheet representing one hierarchical rank of taxonomy. The available taxonomical ranks in the dataset are (in hierarchical order): Class, Order, Family, and Genus. For this study, the genus rank was chosen to develop and build the models. This rank shows the highest level of detail of the four available ranks. In total, 1236 different genera were identified in the 32 samples. The dataset was imported in Python V2.7 and stored into a data-frame. However, not all genera were identified in each of the 32 samples. This resulted in an empty cell for that particular genus in the sample. These cells are filled with a value well below the detection limit of that sample. Lastly, the values were raised to the power of 10.

In the NGS dataset, samples are taken on average every four weeks, resulting in 32 total samples. Unfortunately, these samples are not spread evenly over the total period, but differ in interval time. Since the process data has a higher, weekly frequency, the NGS dataset is re-sampled into a weekly frequency as well. This was done using interpolation. The main two criteria for an interpolation technique are that it naturally approximates the missing data points and that it cannot be lower than zero. Values lower than zero are not acceptable since the occurrence of genera cannot be negative. One of the interpolation techniques that satisfies these two criteria is Piecewise Cubic Hermite Interpolating Polynomial (PCHIP). Therefore, the PCHIP interpolation technique was chosen and applied to the dataset.

A dataset with 1236 columns or features is extremely high for any machine learning model to deal with. Therefore, to reduce the number of features, Principal Component Analysis (PCA) is used for a dimensionality reduction of the dataset. The retained variance per component reduces quickly after the first components and with 28 components, the cumulative retained variance exceeds 0.99. Thus, with that number of components, 99% of the variance from the original dataset is retained. Consequently, a number of 28 principal components is chosen.

3.2 Process Data

The process data from the same WWTP consists of 108 time samples, where the first sample was taken in week 40 of 2014 and the last sample was taken in week 10 of 2017. The process data consists of the chemical composition of the wastewater influent and effluent. In the available dataset of the process parameters, there are some missing values for certain samples. The same PCHIP technique as for the NGS data is used to fill in these values. However, the data points are not re-sampled as done with the NGS dataset, since the frequency of the data is already weekly. The volatility of the process parameters is significantly higher than that of the NGS data. A low-pass filter was applied to overcome this difference by removing the higher frequencies in the data, leaving out a smoother graph. To smooth the process data we used the moving average filter (MA).

3.3 SVR Model

Prior to training the regression model (1) using the WWTP data, we firstly divide the collected data into three subsets: a training, a validation, and a test set. The training set is used to fit the model to the data, and the validation set is used to tune the parameters of the model,

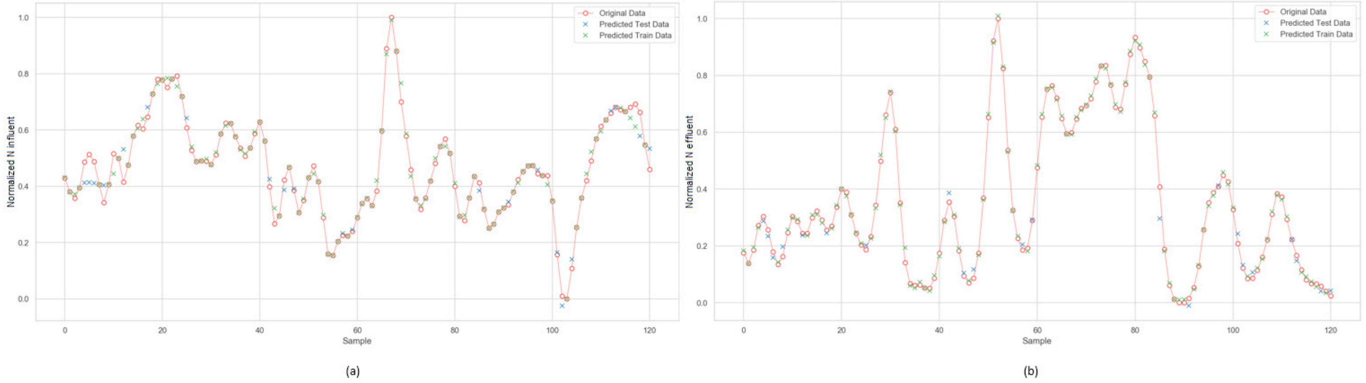

Fig. 3. Comparison between SVR model and the measured data, (a) Nitrogen Influent (b) Nitrogen Effluent

whereas the test set will be used after the model is created to evaluate the performance of the model. Consequently, the model will not see the test set until the evaluation. In this study, 20% of the dataset is used for testing the developed models. The remaining 80% of the dataset is then further divided into a training set and validation set. When we refer to the regression model (1), the scalar output data y_i refers to each of the measured process parameters in the WWTP process data and x_i corresponds to the sample of input vector data obtained from NGS, e.g., the vector of genera density/population.

In this work, we will use the following radial basis functions.

$$\phi_i(x) = \exp(-\gamma\|x - a_i\|^2), \tag{10}$$

where ϕ_i is the i-th element of ϕ, and $a_i \in \mathbb{R}^n, \gamma \in \mathbb{R}$, $i = 1, \ldots, q$ are the kernel parameters that need to be optimized as briefly discussed before. We use the training data x_i as the centre of each basis function ϕ_i, i.e., $a_i = x_i$. This means also that the dimension of ϕ and w as in (1) is n.

Grid search methodology with 5-fold cross-validation on the training set is applied to retrieve the optimal values for the model parameters C (regularization cost parameter), ϵ (determining the ϵ-insensitive loss function) and γ (kernel parameter) for each of the process parameters. The search was done by using exponential sequences for C, ϵ and γ in the ranges of

$$\begin{aligned} C &= [2^1, 2^3, 2^5, ..., 2^{15}] \\ \gamma &= [2^{-15}, 2^{-13}, 2^{-11}, ..., 2^3] \\ \epsilon &= [2^{-11}, 2^{-9}, 2^{-7}, ..., 2^{-1}] \end{aligned} \tag{11}$$

These three ranges resulted in a total of 420 different combinations. In combination with the 5-fold cross validation, this resulted in 2100 runs of the model per process parameter. After fitting the model with the optimized parameters, the test set is used as input for the model. Since the model has not known this dataset yet, it is a validation of the model. The test output is compared with the measured data using the performance criteria R^2. Fig. 3 shows the comparison between SVR model and the measured data.

3.4 Sensitivity Analysis

SVR is a black-box model. Thus, the intrinsic relations between the inputs of microbial communities and the predicted outputs of process parameters are not known. A sensitivity analysis (SA) was performed on all selected trained models, to show the relative importance of each genus to the prediction of the process parameters.

One Factor At a Time (OFAT) technique is used for this SA. OFAT is a local sensitivity analysis technique, only measuring the influence of a genus on a process parameter for a certain change of that genus. Thus, only a local area of the influence of each genus is explored. Algorithm 1 shows how SA works.

Algorithm 1 Algorithm for One Factor At a Time

```
for all selected process parameters $j$ do
    load model
    load all input variables $x$
    $y_j$ = model.predict($x$)
    for all genera $i$ do
        $x^* = x$
        $x^*_i = 1.1x_i$
        $\Delta x_i = x^*_i - x_i$
        $y^*_j$ = model.predict($x^*$)
        $\Delta y_j = y^*_j - y_j$
        $S_{y_j \leftarrow x_i} = \frac{\frac{\Delta y_j}{y_j}}{\frac{\Delta x_i}{x_i}}$
    end for
end for
```

4. RESULTS

4.1 Model Performance

The measured process parameters where measured output is compared with predicted output of both the test set and the training set. In general, all models scored very well according to the performance criteria, since all models had an R^2 score higher than 0.85 (except for INF_ SO_4 with an R^2 score below 0.85). Therefore, all process parameters are selected for the sensitivity analysis.

To further prove that the performance of the model created by SVR can be used, we then compared it with a Deep Neural Network (DNN) model. The comparison of the R^2 score between SVR and DNN can be seen in Table 1. We can see that the SVR produced better models than DNN (70% of the models are better).

4.2 Result of Sensivity Analysis

The sensitivity analysis with OFAT showed that *Thalassospira* has a positive correlation with K and TOD in the effluent and Cl, COD, EC, and Na in the influent (Fig 4). *Thalassospira* species are usually isolated from marine environments, as reported by Hutz et al. (2011). Since the influent of this WWTP contains more salt than usual these correlations with the electric conductivity (EC), NaCl, and KCl) can be explained. Furthermore, this analysis also showed that the anaerobic genera *Desulfocapsa* has a negative SA with the SO_4 in the influent and effluent. As this species ranked in the top 10, this means that a high number of this species in the activated sludge is correlated with the concentration of SO_4 in the effluent and influent. However, the concentration of SO_4 does not necessarily imply the presence of *Desulfocapsa*. The sensitivity ranking of the top 10 genera for each process parameter in the SVR model is shown in Fig 4. Both SVR and DNN show similar results for the sensitivity analysis.

5. CONCLUSION

By analyzing the (local) sensitivity of each modeled process parameter to each input (each genus), an indication of the influence of the microbial structure on process performance was found. Some of these sensitivity scores can be explained by looking at the function of a specific organism in an environment, but most of them remain unknown. It is difficult to further assess the plausibility of this sensitivity analysis, due to two reasons. First, the functions of many genera are still unknown. Second, a strong sensitivity does not automatically mean that there is a causal relation between the two factors. In a next step, we will generate NGS datasets of mRNA of the samples and determine the metabolic pathways that are active in the genus. Using this data we can optimize the SVR and SA to predict the behavior of a WWTP in relation to the influent composition.

Table 1. Comparison between SVR and DNN

Parameter	SVR	DNN	Better result
BOD-5 effluent	0.8954	0.7212	SVR
Cl effluent	0.8938	0.8134	SVR
COD effluent	0.9259	0.8555	SVR
EC effluent	0.9057	0.8667	SVR
K effluent	0.9990	0.9970	SVR
N effluent	0.9221	0.8678	SVR
Na effluent	0.9994	0.9989	SVR
NH_4 effluent	0.9285	0.8347	SVR
Nkj effluent	0.9014	0.8246	SVR
NO_2 effluent	0.9697	0.9215	SVR
NO_3 effluent	0.9325	0.8350	SVR
PO_4 effluent	0.8806	0.9474	DNN
SO_4 effluent	0.9301	0.8068	SVR
TOD effluent	0.9989	0.9963	SVR
TSS effluent	0.8793	0.7709	SVR
BOD-5 influent	0.8671	0.8925	DNN
Cl influent	0.9640	0.9655	DNN
COD influent	0.8516	0.9068	DNN
EC influent	0.9651	0.9459	SVR
K influent	0.9994	0.9979	SVR
N influent	0.9171	0.8966	SVR
Na influent	0.9994	0.9984	SVR
NH_4 influent	0.9705	0.9725	DNN
Nkj influent	0.9535	0.9784	DNN
NO_2 influent	0.9614	0.9443	SVR
NO_3 influent	0.9370	0.9451	DNN
PO_4 influent	0.9221	0.9268	DNN
SO_4 influent	0.7739	0.4766	SVR
TOD influent	0.9987	0.9971	SVR
TSS influent	0.8984	0.9388	DNN

ACKNOWLEDGEMENTS

The data used in this paper is taken from North Water's Wastewater Treatment Plant in Delfzijl. North Water provided the analytical data of the chemical composition of the influent and effluent and the microbiological samples of the activated sludge compartment of the WWTP are processed by BioClear (2014-2017).

REFERENCES

Bassin, J., Rachid, C., Vilela, C., Cao, S., Peixoto, R., and Dezotti, M. (2017). Revealing the bacterial profile of an anoxic-aerobic moving-bed biofilm reactor system treating a chemical industry wastewater. *International Biodeterioration & Biodegradation*, 120, 152–160.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Davies, P. (2005). *The Biological Basis of Wastewater Treatment*. Strathkelvin Instruments Ltd, Motherwell.

Drucker, H., Burges, C., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 155–161.

Eitrich, T. and Lang, B. (2006). Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of Computational and Applied Mathematics*, 196, 425–436.

Henze, M. and Comeau, Y. (2008). *Biological Wastewater Treatment*. IWA Publishing, London.

Henze, M., Gujer, W., Mino, T., and van Loosdrecht, M. (2000). *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA Publishing, London.

Hsu, C., Chang, C., and Lin, C. (2010). A practical guide to support vector classification. `https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`. Online; accessed 12 Oktober 2018.

Hutz, A., Schubert, K., and Overmann, J. (2011). Thalassospira sp. isolated from the oligotrophic eastern mediterranean sea exhibits chemotaxis toward inorganic phosphate during starvation. *Applied and Environmental Microbiology*, 77, 4412–4421.

Liu, T., Liu, S., Zheng, M., Chen, Q., and Ni, J. (2016). Performance assessment of full-scale wastewater treatment plants based on seasonal variability of microbial communities via high-throughput sequencing. *PLoS ONE*, 11.

Muszynski, A., Tabernacka, A., and Miobedzka, A. (2015). Long-term dynamics of the microbial community in a full-scale wastewater treatment plant. *International Biodeterioration & Biodegradation*, 44–51.

Scholkopf, B. and Smola, A. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Massachusetts.

Xu, J. (2014). *Next-Generation Sequencing*. Caister Academic Press, Norfolk.

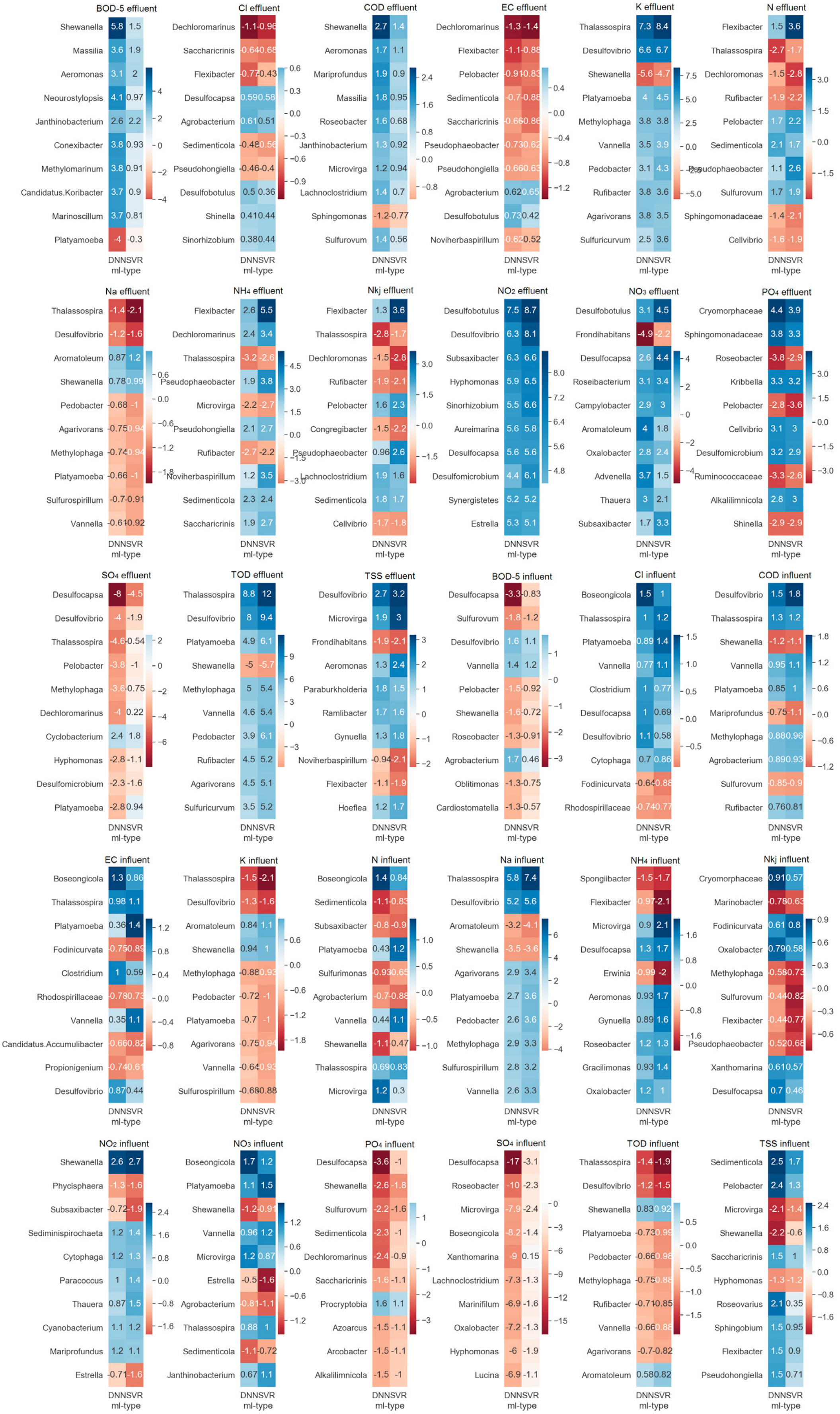

Fig. 4. Ranking of the SA of the top 10 genera using OFAT on the trained models SVR and DNN using each process parameter