Input variable selection using machine learning and global sensitivity methods for the control of sludge bulking in a wastewater treatment plant

Nadja Hvala[a,*], Juš Kocijan[a,b]

[a] Jožef Stefan Institute, Department of Systems and Control, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

[b] University of Nova Gorica, Vipavska cesta 13, SI-5000 Nova Gorica, Slovenia

Abstract

Sludge bulking is a common and undesired phenomenon in wastewater treatment plants that negatively affects biomass settling characteristics, deteriorates treatment efficiency and causes severe operational problems. First-principles models for this phenomenon are not yet available. Therefore, data-driven models have been developed to predict sludge bulking. In this paper, the bulking phenomenon is studied from the control point of view and operating variables that can be used to control sludge bulking are identified. Identification is performed by designing a data-driven model using available process data as well as clustering and various classification methods. A global sensitivity analysis is applied to select the operating variables with the highest impact on sludge bulking. Application of the proposed approach to full-scale data has shown that increasing aeration intensity and limiting nitrogen sources are the most promising control actions for bulking control.

Keywords:

Wastewater treatment; Sludge bulking control; Global Sensitivity Analysis; Variable selection; Classification

* Corresponding author.
E-mail address: nadja.hvala@ijs.si (N. Hvala)

1. Introduction

Nowadays, wastewater treatment before discharge into the environment is an essential step to protect nature and human health. Most often, wastewater is treated in biological wastewater treatment plants (WWTPs), where a mixture of bacteria degrades the contaminating water components in an activated sludge process. The coexistence of different types of bacteria and the proper balance between them are among the most important operational challenges that, if they fail, can lead to operational problems and degradation of treatment performance.

One of the most serious operational problems in biological WWTPs is sludge bulking, which causes problems in solid-liquid separation. Bulking occurs when the suspended solids in the activated sludge process, i.e. biomass flocs, do not separate from the treated water by gravity settling in the settling tank. That is a common problem in modern biological nutrient removal (BNR) plants where long sludge retention times (SRTs) are used. Long SRTs favour the growth of filamentous bacteria. An appropriate balance between floc forming and filamentous bacteria improves sludge settling, but an excess of filamentous bacteria can lead to poor settling, foaming on the reactor surface, and dewatering problems in the sludge treatment process.

Bulking can be related to the characteristics of the wastewater and/or the operating conditions of the plant. However, the various causes have not been fully explored, and there are no first-principles or generally applicable models linking sludge bulking to process conditions. Also, recommended operational adjustments to limit the occurrence of bulking are inconsistent and often contradictory, e.g., increasing or decreasing various operational parameters such as sludge age, return sludge flow, waste sludge flow, oxygen concentration, etc. Due to the knowledge gaps and lack of formal theoretical descriptions of sludge bulking, data-driven models have been developed.

Data-driven models relate sludge bulking and process conditions based on empirical relationships derived from process data. Most commonly, models are designed as bulking *prediction* models aiming at forecasting the occurrence of bulking in advance (Liu et al., 2020). Bulking prediction models are

designed as time-series models (Liu et al., 2016a), multivariate models derived from other process data (Lou and Zhao, 2012, Bagheri et al., 2015, Deepnarain et al., 2019, Szeląg et al., 2020) or a combination of both (Liu et al., 2016b). Various methods have been proposed already, e.g. Artificial Neural Networks (ANN) (Capodaglio et al., 1991, Bagheri et al., 2015, Han et al., 2016), Principal Component Regression (PCR) (Lou and Zhao, 2012), Gaussian Processes Regression (GPR) (Liu et al, 2016b), Principal Component Analysis (PCA) and decision trees (Deepnarain et al., 2019). In most cases, the focus is on the quality of model prediction, i.e. whether the model can predict the occurrence of sludge bulking with high accuracy.

The issues of sludge bulking *diagnosis* have also been considered (Cheng et al., 2019, Han et al., 2021). On-line diagnosis consists of two steps. First, a data-driven model is used to detect the occurrence of bulking, and in the second step, the causes are identified. Liu et al. (2020) propose two further steps, i.e. remaining useful life prediction and maintenance strategy. In the preventive maintenance stage, the operating parameters should be adjusted to compensate for the incipient fault and keep sludge bulking below the control limit.

As presented above, on-line monitoring and diagnosis of sludge bulking involve data-driven prediction and identification of the causes of sludge bulking based on the temporal dynamics of process variables. However, prior knowledge of the key process variables associated with the conditions for sludge bulking is required. In addition, once the cause variable is identified, subsequent knowledge of the most promising *control* actions and adjustments of process operating parameters is also required (Nittami et al., 2021). Since this knowledge is very specific to each case and difficult to obtain directly from plant operations, it is expected to be obtained through data-driven model development. The model should include both potential causes and control variables that have been discovered to be related to sludge bulking conditions. Such a model will also provide information on the regions of bulking and non-bulking conditions in the space of plant operating parameters, which is important for control purposes.

For this purpose, the design of a data-driven model is considered in this paper. Modelling is intended for knowledge discovery, i.e. finding variables that are highly related to sludge bulking conditions. Therefore, the *selection of model input variables* is considered as one of the most important outcomes of modelling.

Different methods can be used for input variable selection. They can be divided into *model-based* and *model-free* (filter) methods. Model-based methods are further divided into wrapper methods and embedded methods (Guyon and Elisseeff, 2003). *Filters* are used to select a subset of variables as a pre-processing step, regardless of the modelling approach. Examples of filters are statistical analysis methods based on Pearson correlation coefficient, coefficient of determination R^2, F-test, or other similar criteria. *Wrappers* use a selected model to evaluate subsets of variables according to their predictive power. Some well-known wrapper methods in classical statistical approaches for variable selection in regression are forward selection and backward elimination (Andersen and Bro, 2010). In these cases, regressors in the selected model are systematically added or removed one by one until cross-validation results confirm the minimal set of regressors that provides the best model accuracy. A common feature of wrapper methods is that they are computationally intensive and can become intractable when the number of input variables is large. *Embedded* methods perform variable selection in the process of model training and are usually specific to a selected modelling approach. In this case, the task of variable selection is delegated to the model learning phase. An example of an embedded method is ANN training based on a pruning strategy where the irrelevant and/or redundant weights of a network are gradually removed.

When developing data-driven sludge bulking models, pre-existing knowledge is usually used to constrain the initial set of candidate variables. Variable selection is then based on statistical tests of candidate variables and/or a model-based search for the most appropriate combinations of input variables. Methods used include correlation analysis (Lou and Zhao, 2012), the Chi-squared test (Deepnarain et al., 2019), the variable importance in projection (VIP) method (Liu et al., 2016b,

Chmielowski et al., 2019), PCA and forward selection (Bagheri et al., 2015), the Fischer-Snedecor test followed by a search for different combinations of independent variables (Szeląg et al., 2020). In many of these cases, linear methods are used for pre-processing the input variables, e.g., correlation analysis and PCA, which may not discover the significant input variables in the case of nonlinear relations (Šindelář and Babuška, 2004). On the other hand, when using wrapper methods, even if the combinatorial problem of input variable selection is not extreme, the choice of input variables can be difficult when the differences in the performance of models with different sets of variables are small. This problem occurred in the development of binary classification models (Szeląg et al., 2020).

This paper proposes a model-based approach for the selection of input variables of sludge bulking models. The procedure follows the general scheme of wrapper methods (May et al., 2011) with the addition of variable ranking. Variable ranking allows the identification of the most influential model variables and is performed in our case by applying Global Sensitivity Analysis (GSA). GSA is a set of statistical techniques used to investigate the extent to which variation in model output can be attributed to variation in model inputs. Many GSA methods have been proposed in the literature, usually calculating a set of sensitivity indices for the various factors of the model. These indices can be used to estimate the impact of individual variables or groups of variables on model output. In this paper, we use the Variance-Based Sensitivity Analysis (VBSA or Sobol's method) (Sobol', 2001), which is one of the most popular methods in many disciplines (Wei et al., 2015, Makrygiorgos et al., 2020). Its advantages are that it provides global sensitivity over the entire input space, as opposed to local sensitivity at a particular model solution, and that it can be used for nonlinear and nonadditive systems. It is applied, among other things, to understand the dominant controls of a system (model) (Pianosi et al., 2015), which is the subject of this work. Performing a sensitivity analysis within the operating region of the process variables allows us to evaluate the impact of potential control variables on the process performance and thus estimate their ability to control sludge bulking. The approach is useful in cases where many combinations of process variables result in a similar performance, making it difficult to reduce input variables based on model performance alone. A similar machine learning framework for

identifying relationships between operational variables and effluent parameters in WWTPs was proposed by Wang et al. (2021). In their case, permutation importance (PI) was used as a measure of variable importance.

The approach is presented for a full-scale WWTP where a severe problem of a sludge bulking phenomenon is encountered throughout the year. Microscopic analysis revealed that the filamentous bacterium *Microthrix parvicella* was present in the biological reactors of the WWTP. Its presence can be in theory associated with certain operating conditions. These conditions and associated process variables were considered as potential model regressors in the data-driven model design. The model was designed using various classification methods in the Matlab classification toolbox. As a pre-processing step for classification, the model output, i.e., measured sludge settleability was clustered into bulking and non-bulking states.

The original contributions are as follows:

- The procedure for selecting input variables based on global sensitivity analysis as a variable importance measure.
- The application of various machine learning methods to design data-driven models of sludge bulking.
- The demonstration of the proposed method on a full-scale WWTP case study.

This paper is organised as follows. In the next section, we first present the WWTP case study, followed by the description of the proposed procedure for input variable selection as well as the clustering and classification methods. In Section 3, we demonstrate the proposed method on a full-scale WWTP and discuss the results. The paper ends with the conclusions describing the main results and perspectives for future work.

2. Materials and methods

2.1 WWTP case study

The case study under consideration is a WWTP for 95000 PE (population equivalent) treating municipal and industrial wastewater. The plant was upgraded in 2016 for complete nitrogen and phosphorus removal with biological treatment and chemical precipitation, respectively. The treatment facilities consist of mechanical treatment (screens, grit and grease chamber, primary clarifier), a biological stage with suspended biomass activated sludge process (nine mixed and/or aerated reactors) and sludge treatment (thickening, anaerobic digestion, dewatering). The reactors of the biological stage with active biomass are operated as a three-cascade system, each cascade consisting of three tanks, i.e. denitrification reactor (DN), combined reactor operated as denitrification or nitrification reactor (DN/N) and nitrification reactor (N). Sludge settling takes place in secondary clarifiers, where solid particles are separated from treated wastewater by gravity settling. The plant operation is monitored by laboratory analyses and on-line sensors as shown in Fig. 1. All data, together with other measured signals, are stored in a supervisory control and data acquisition (SCADA) system and are available for data-driven analysis and modelling of sludge bulking.

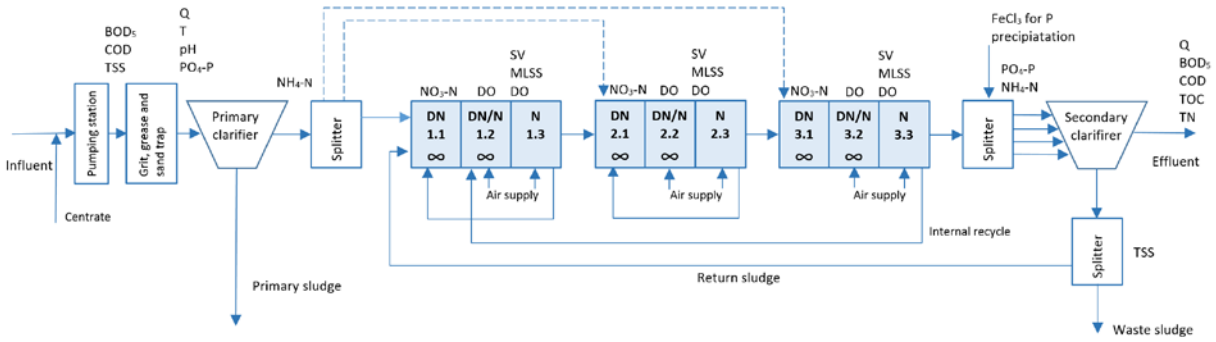


Fig. 1. Schematic layout of WWTP with indicated laboratory measurements and on-line sensors. Laboratory measurements include 5-day biological oxygen demand (BOD_5), chemical oxygen demand (COD), total suspended solids (TSS), mixed liquor suspended solids ($MLSS$) and settled sludge volume (SV). On-line measurements and sensors include flow (Q), temperature (T), pH, orthophosphate (PO_4-P), ammonia-nitrogen

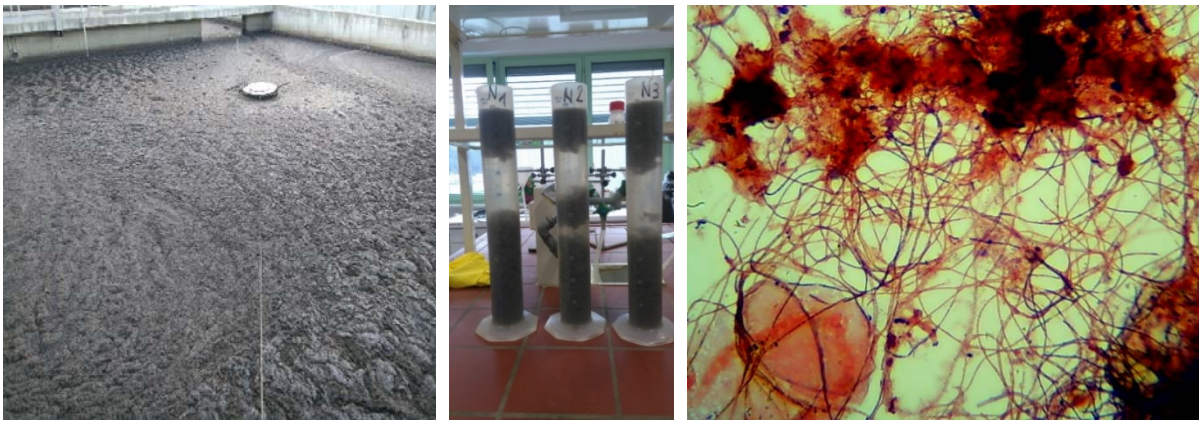
(NH_4-N), nitrate-nitrogen (NO_3-N), total nitrogen (TN), total organic carbon (TOC), total suspended solids (TSS) for return sludge and dissolved oxygen concentration (DO).

Sludge settleability is measured using a standard laboratory test in which 1 litre of the mixed activated sludge sample from the biological reactors is settled for 30 minutes (Jin et al., 2003). The settled sludge volume (SV) after 30 minutes and the measured mixed liquor suspended solids concentration ($MLSS$) in the reactor are then used to calculate the sludge volume index (SVI) as follows

$$SVI = \frac{SV\ (\mathrm{mL/L})}{MLSS\ (\mathrm{g/L})} \tag{1}$$

The generally accepted threshold value for SVI is 150 mL/g. Sludge settling is considered as appropriate if SVI < 150 mL/g, while for SVI > 150 mL/g poor settling and sludge bulking occurs.

In the considered case study, the problem of sludge bulking is severe and occurs throughout the year. It is indicated by poor settling of solids, lifting of the sludge in the settling test (Fig. 2) and by very high SVI values, which rarely fall below 150 mL/g. Occasional microscopic images of sludge samples show that the presence of filamentous bacteria is significant, with *Microthrix parvicella* recognised as the predominant type (Fig. 2). Despite bulking, effluent quality is most of the time below the legislation limits (daily average effluent values of TN < 15 mg/L, NH_4-N < 10 mg/L, TP < 2 mg/L), but some very restrictive measures have to be taken in plant operation, e.g. limited influent flow to the biological stage to reduce the risk of solids washout from the secondary clarifiers during the first flush and rain events, low $MLSS$ concentrations in the biological reactors to prevent sludge accumulation and foaming.



a b c

Fig. 2. Severe bulking conditions presented with a) foam on the reactor surface, b) poor settling and sludge uplift in the 30-minute laboratory test, and c) the presence of filamentous bacteria *Microthrix parvicella* in the microscopic image of the sludge sample.

2.2 Prior knowledge of bulking conditions

The filamentous bacterium *Microthrix parvicella* is commonly found in activated sludge WWTPs and is most often responsible for solid-liquid separation problems. The microbiological review paper (Nielsen et al., 2009) indicates that *M. parvicella* is well adapted to the hydrolysis, uptake and growth on lipids and greases. Substrates are taken up under both aerobic and anaerobic conditions and stored as lipid reserves, but filaments grow only under conditions with nitrate or oxygen as an e-acceptor. The growth characteristics of *M. parvicella* presented in Rossetti et al. (2005) show that several factors influence its growth and can be used in control strategies to prevent bulking:

i. *M. parvicella* requires fairly *high sludge ages* for its survival in WWTPs (solids retention times SRTs > 10 d), which is in agreement with its presence in nutrient removal plants with longer SRTs for nitrification. Besides, it may proliferate in scums and physical barriers where the retention time is higher than the nominal sludge age.
ii. *M. parvicella* grows at appreciable rates at *low temperatures* (T < 12-15 °C), which implies a significant competitive advantage over other bacteria during the cold season.

iii. It appears to be sensitive to high oxygen tensions, suggesting a microaerophilic preference. Therefore, it proliferates in conventional plants with spatial or temporal *low DO concentrations* or in nutrient removal plants with anoxic-aerobic zones (Comas et al., 2008, Guo et al., 2012, Zhang et al., 2017).

iv. Another factor promoting the growth of *M. parvicella* is also the *availability of nitrogen compounds*. It has been hypothesized that incomplete denitrification and formation of nitric oxide may cause toxic effects on floc-forming organisms preventing them to utilize slowly biodegradable substrate and thus outcompete filamentous bacteria. Besides, it has been suggested that incompletely nitrifying BNR plants have available ammonia under aerobic conditions that can be used as a preferential nitrogen source for *M. parvicella*. In a set of laboratory experimental investigations, it has been shown that in the alternating anoxic-aerobic conditions the AA bulking may occur (Casey et al., 1993), which is associated with a group of low F/M (food to mass ratio) bacteria, including *M. parvicella*. AA bulking was observed to be related to the presence of nitrate (NO_3-N) or nitrite (NO_2-N) in the anoxic zone or on the transition between anoxic to the aerobic zone. The process conditions inducing AA bulking are as follows: (i) either nitrate or nitrite or both are present in the anoxic zone immediately preceding the aerobic zone (NO_3-N > 5 mg/L and/or NO_2-N > 1 mg/L) (Lakay et al., 1999, Musvoto et al., 1999), (ii) residual ammonia is present under aerobic conditions (NH_4-N > 1 mg/L) (Tsai et al., 2003), (iii) an aerobic mass fraction between 30-40 % of the total if nitrite is present in excess (Musvoto et al., 1999). High concentrations of NO_3-N in the secondary clarifier influent were also identified as critical conditions for the development of rising sludge in the activated sludge systems (Comas et al., 2008).

Besides the literature review on *M. parvicella* growth conditions, the plant personnel observe that some operating conditions may favour the increased filamentous growth. For example, the plant is hydraulically underloaded, therefore bulking is more severe at low input flows. Besides, the mixing of the reactors is not sufficient, and in severe bulking conditions, the biomass retains in the poorly mixed parts of the reactors. Also, the occurrence of bulking seems to increase at higher input ammonia loads.

Based on the theoretical knowledge and practical observations above, the following potential reasons for sludge bulking were identified:

- low temperatures,
- low dissolved oxygen concentrations,
- low aeration intensity contributing to insufficient mixing of reactors,
- availability of excess ammonia and nitrate.

The related process variables were considered in the data-driven analysis as candidate input variables for the bulking control model design.

2.3 Procedure for input variable selection

The problem of input variable selection is to choose a (small) subset from the available variables that gives the optimal form of the model. The optimal input variable set contains the fewest input variables required to describe the behaviour of the output variable, with a minimum degree of redundancy and with no uninformative variables. Such a set of variables results in a more accurate, efficient, cost-effective and easily interpretable model. The procedure for variable selection applied in this paper is presented in Fig. 3. The first three steps are based on a general conceptual approach of wrapper algorithms presented in May et al. (2011) with the addition of variable importance ranking based on global sensitivity analysis (GSA) and adjusted for the sludge bulking model design. The procedure consists of the following steps:

Step 1: Selection of candidate input variables based on a priori knowledge of the process.

Step 2: Training the model with all possible input variables using machine learning.

Step 3: Selection of the model with the best cross-validation performance.

Step 4: Variable importance ranking for the selected model using global sensitivity analysis.

Step 5: Reduced model training, i.e., performing *Step 2, Step 3 and Step 4* with the reduced set of input variables.

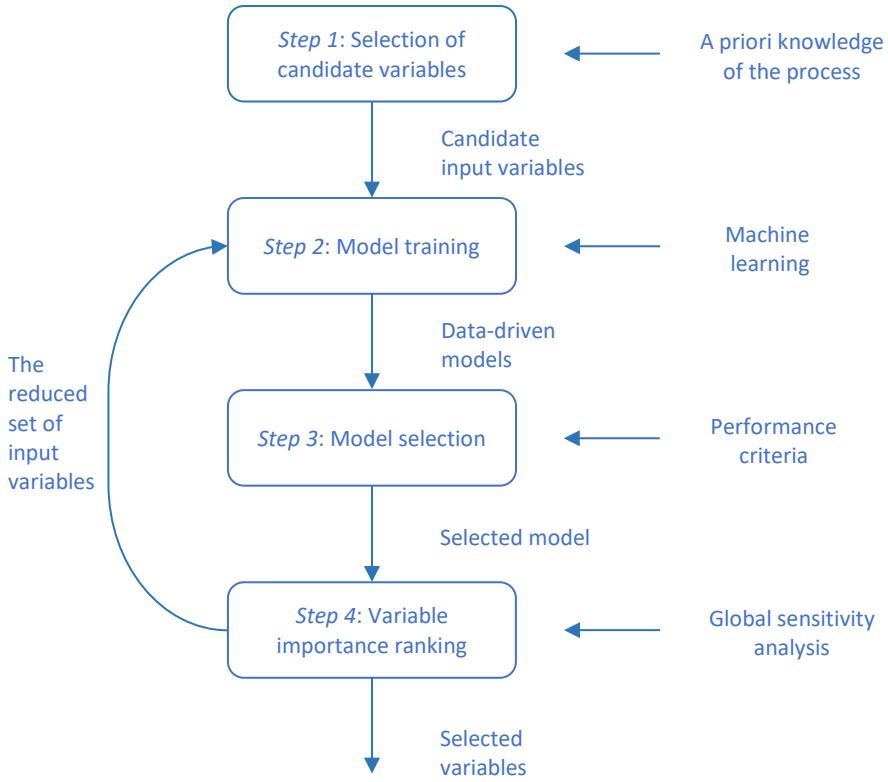


Fig. 3. Procedure for input variable selection based on global sensitivity analysis.

By performing *Steps 1-3*, this procedure first generates a complete model with all possible input variables and then in *Step 4* removes those variables that have little influence on the model output. This is essentially a backward elimination procedure, but the inclusion of global sensitivity analysis for variable importance ranking avoids the need to repeatedly generate new models to test each combination of variables. The sensitivity analysis indicates whether a particular process variable significantly affects the process output in the tested area.

2.4 Global sensitivity analysis

Global Sensitivity Analysis (GSA) and the calculation of Sobol's indices was performed by Matlab Toolbox SAFE (Pianosi et al., 2015). For the model under investigation, described by a function $Y = f(\mathbf{X})$, where $\mathbf{X} = (X_1, \dots, X_n)$ represents a n-dimensional set of input variables, the first-order

sensitivity indices S_i and total-order sensitivity indices S_{Ti} are calculated by observing variations in Y when varying X_i (Wei et al., 2015)

$$S_i = \frac{\mathrm{Var}_{X_i}(E_{\mathbf{X}_{\sim i}}(Y|X_i))}{\mathrm{Var}(Y)} \tag{2}$$

$$S_{Ti} = 1 - \frac{\mathrm{Var}_{\mathbf{X}_{\sim i}}(E_{X_i}(Y|\mathbf{X}_{\sim i})}{\mathrm{Var}(Y)} \tag{3}$$

where $\mathrm{Var}(Y)$ is the total variance of model output, Var_{X_i} is the variance when varying X_i, $\mathrm{Var}_{\mathbf{X}_{\sim i}}$ is the variance when varying all variables except X_i. S_i represents the "main effect", i.e. the contribution to the output variance by varying X_i alone. S_{Ti} represents the "total effect", i.e. all contributions to the output variance from X_i, including the variance caused by all its interaction with other variables.

The sensitivity indices have the property $0 \leq S_i \leq S_{Ti} \leq 1$. In variable importance ranking, more influential model variables have higher S_i and S_{Ti} values. For X_i to be non-influential, $S_i = 0$ is a necessary but not sufficient condition, while $S_{Ti} = 0$ is a necessary and sufficient condition (Saltelli et al., 2004). Therefore, S_i is used for selecting important variables while S_{Ti} is more suitable for screening non-influential variables.

Numerical calculation of sensitivity indices was performed by Monte Carlo simulation, which involves generating a sequence of randomly distributed points in the space of input variables. For the calculation of sensitivity indices, a sequence of 10000 $\boldsymbol{X}$ data points was generated based on uniform distribution and Latin Hypercube Sampling in the defined space of input variables. The range of each X_i variable was determined based on full-scale measured data as lower adjacent (LA) and upper adjacent (UA) of X_i. These values present the most extreme observations in the n-point data-set that are within the lower (LL) and the upper (UL) limits. These limits are defined from the sample quartiles ($\hat{q}_{0.25}$, $\hat{q}_{0.75}$) and the interquartile range ($\widehat{IQR}$):

$$\widehat{IQR} = \hat{q}_{0.75} - \hat{q}_{0.25} \tag{4}$$

$$LL = \hat{q}_{0.25} - 1.5\,\widehat{IQR} \quad (5)$$

$$UL = \hat{q}_{0.75} + 1.5\,\widehat{IQR} \quad (6)$$

2.5 Clustering

The output classes are determined using the *Jenks natural breaks classification method* (Jenks, 1967, Amirruddin et al., 2020). The rationale for using this method is that we do clustering of one-dimensional data. *k-means clustering*, as an example of other clustering methods, is the generalization for multivariate data. The Jenks natural breaks classification method iteratively repeats three steps. In the first step, the sum of squared deviations from the class means is calculated. In the second step, the sum of squared deviations from the average mean of classes is calculated. In the third step, a piece of data is moved from a class with a larger variance to a class with a lower variance. Clustering was performed using Matlab code (MS, 2020).

2.6 Classification

The classification was performed with the Matlab (Version R2020a) Classification Learner application, which is part of the Statistics and Machine Learning toolbox. It allows choosing between different algorithms for the training and validation of classification models and their comparison. The groups of methods are *Decision Trees* (Fine, Medium and Coarse), *Discriminant Analysis* (Linear and Quadratic), *Logistic Regression, Support Vector Machines* (SVM) (Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian and Coarse Gaussian), *k-nearest neighbour classifiers* (Fine, Medium, Coarse, Cosine, Cubic and Weighted KNN), *Ensemble classifiers* (Boosted Trees, Bagged Trees, Subspace Discriminant, Subspace KNN and RUSBoosted Trees). A total of 23 classification algorithms were tested for the model design. Among them, Linear Discriminant and Linear SVM belong to linear methods, all other methods are nonlinear.

2.7 Training and validation

The process data for classification was divided into two sets. The first set was used to train the model, while the second set was used to test the obtained statistical classifier. In this way, the consistency of the results could be assessed. Before classification, the data was normalized using the Matlab function *normc*, which normalizes each input variable independently to length 1.

In the classification model training, the first data set was used (training data). The models based on different classification algorithms were evaluated and compared using *k*-fold cross-validation. The model with the best performance in *k*-fold cross-validation was tested for its accuracy Acc in predicting the measured output:

$$Acc = \frac{\text{Total correct predictions}}{\text{Total predictions}} \tag{7}$$

The overall performance of the classifier was evaluated by different criteria, i.e. accuracy in *k*-fold cross-validation in model training, accuracy in training data, accuracy in test data, true positive rate (TPR) per class. For good model performance, TPR should be close to 100% and similar for all classes (Chmielowski et al., 2019).

3. Results and Discussion

For the design of the sludge bulking model, on-line and laboratory measurements from the considered full-scale WWTP were collected in the period from January 2019 to July 2020. Continuously measured on-line signals were sampled at a one-day interval as daily average values.

3.1 Clustering the model output

The most common process variable to observe sludge settleability is the sludge volume index SVI (1). It can be easily clustered into classes of good and poor settleability based on the established threshold value of 150 mg/L as described in Section 2.1. However, in the presented WWTP the SVI exceeds 150 mg/L most of the time. Besides, the sedimentation rate is often very poor, meaning that the sludge does not settle at all and the volume of the settled sludge *SV* is equal or close to the limit value of 1000 mL/L. In such conditions, the use of *SVI* is not appropriate because of its highly nonlinear relation to *SV* and *MLSS* (1). For these reasons, the SV is used as an observed output variable. SV is measured in each of the three nitrification reactors (N1.3, N2.3 and N3.3 in Fig. 1). The average value of these three measurements is used for clustering.

The data of measured SV in the observed period include 324 data points. The clustering algorithm has identified the following two classes, i.e. 'NB' class representing non-bulking conditions and good sludge settleability with SV within 346 - 810 mL/L, and 'B' class representing bulking conditions and poor sludge settleability with SV within 810 - 1000 mL/L. Hence, the measured data were classified as follows

$$y_k = \begin{cases} \text{'B'}, & \text{if } (SV_k > 810 \text{ mL/L}) \\ \text{'NB'}, & \text{otherwise} \end{cases}, k = 1, \dots, n. \quad (8)$$

The results of clustering are shown in Fig. 4 and Fig. 5. From Fig. 4 it can be seen that the variation of the settled sludge volume can be associated with seasonal variations and temperature changes. The

sludge bulking is most severe from January to June when the sludge does not settle at all or rises to the top during the 30-minute settling test. The conditions improve in summer at higher wastewater temperatures and continue in autumn due to better-conditioned sludge during the summer. Occasional short-term changes from bulking to non-bulking conditions or vice versa indicate that other input variables also affect sludge bulking and are expected to be discovered by the application of the classification approach. Fig. 5 shows that after the division of data in two classes based on the variance of deviations to the classes mean, around one-third of data is in class 'NB' (good settleability) and two-third of data is in class 'B' (poor settleability).

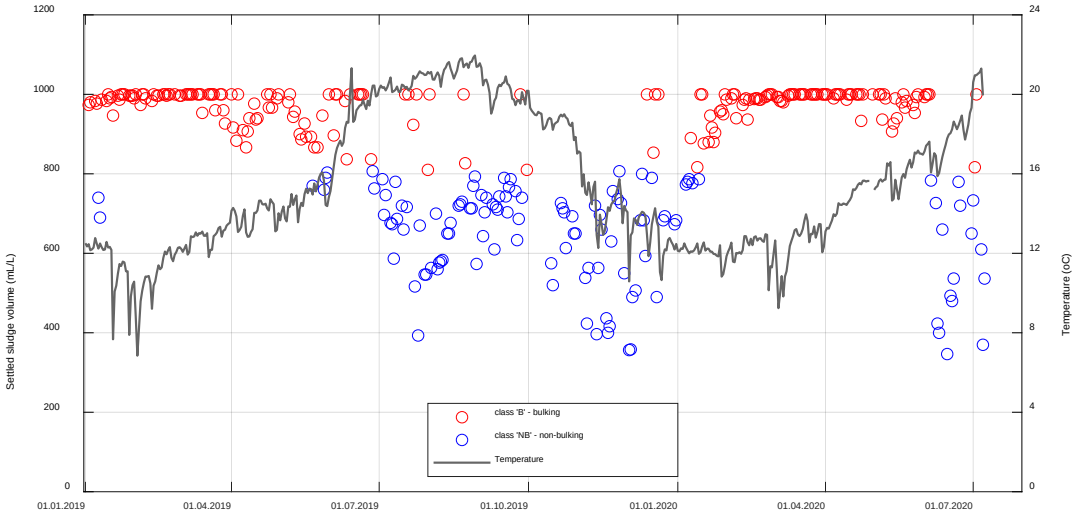


Fig. 4. Classification of settled sludge volume (SV) into 'B' and 'NB' classes using Jenks natural breaks classification method.

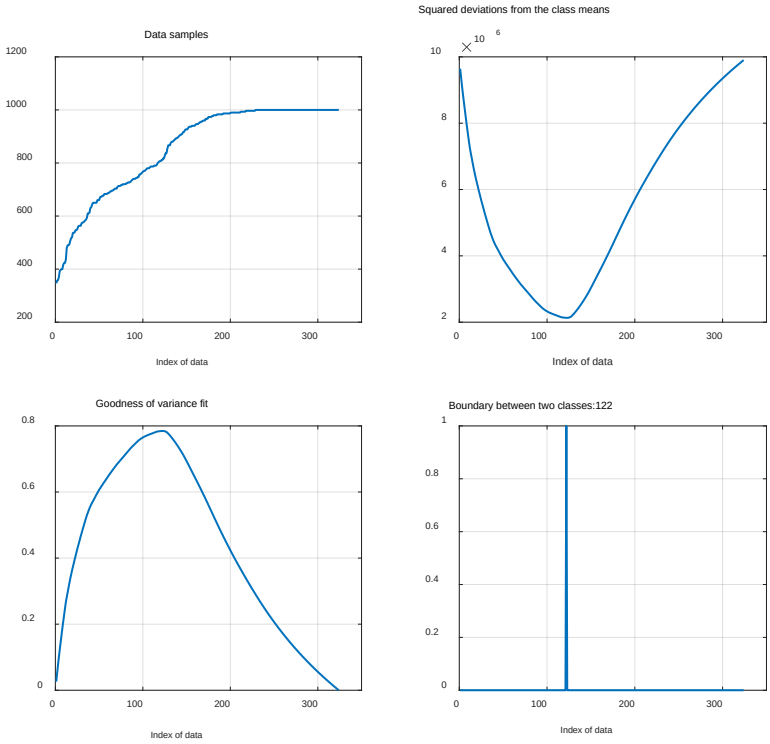


Fig. 5. The final result of using Jenks natural breaks classification method. Sorted data by its value is in the upper left figure, squared deviations from the class mean is given in the upper right figure, goodness of variance fit is shown bottom left and bottom right is the boundary between both classes.

3.2 Classification model design

The process input and output data to be used for classification include around 250 data points from full-scale plant operation. The number varies slightly depending on the selected input variables. Three-quarters of the data were used for training and one quarter for testing. When training the classification model, the models were evaluated using 4-fold cross-validation. The number of folds, i.e. four, was determined to ensure that both training and test subsets contained a sufficient number of cases in 'B' and 'NB' classes. The following sub-sections present individual steps from Fig. 3 for input variable selection for the classification model.

3.2.1 Step 1: Selection of candidate variables

In Section 2.2, four potential causes of sludge bulking were identified. Each of them could be associated with some measured process variables (see Fig. 1), i.e. temperature (*T*); dissolved oxygen concentrations (*DO*) in six aerated reactors (DN/N and N tanks); aeration intensities (*AE*) in six aerated reactors, where aeration intensity is determined as the percentage of the time in a day that the reactor is aerated; nitrogen concentrations including input ammonia mass flow (Φ_m), determined from influent flow (*Q*) and influent ammonia concentration ($NH4_i$), nitrate concentrations (*NO3*) in DN reactors and ammonia concentration at the outlet of the biological reactors ($NH4_o$). These variables were identified as candidate input variables and can be classified into four main groups:

1) Disturbance variables (Q, Φ_m, T).
2) Dissolved oxygen concentrations in aerated reactors (DO_1, DO_2, DO_3, DO_4, DO_5, DO_6).
3) Aeration intensities in aerated reactors (AE_1, AE_2, AE_3, AE_4, AE_5, AE_6).
4) Nitrogen (N) concentrations ($NO3_1$, $NO3_2$, $NO3_3$, $NH4_o$).

The first group represents input disturbances, i.e. those variables that are related to input wastewater characteristics and environmental conditions and could not be intentionally varied. The other three groups include process operating variables that are potentially related to bulking conditions and could be manipulated in bulking control. A total of 19 variables are used as model candidate variables.

3.2.2 Steps 2 & 3: Training and model selection

First, classification models were designed with all candidate input variables. The obtained 4-fold cross-validation accuracy on training data ranged from 65 to 89 % for the different classification algorithms. The results of the best performing classification models are shown in Table 1. We can see that the models have high accuracy on both training and test data and were used for sensitivity analysis and variable selection in the next step.

Table 1. Classification models with all input variables and 4-fold cross-validation accuracy higher than 85%.

Model	Accuracy in 4-fold cross-validation (%)	Accuracy in training data (%)	Accuracy in test data (%)
Linear Discriminant	89.0	89.91	90.28
Linear SVM	87.6	89.91	90.28
Ensemble Subspace Discriminant	86.7	88.07	88.89
Medium Gaussian SVM	86.2	91.28	88.89
Logistic Regression	85.8	90.83	90.28

3.2.3 Step 4: Variable importance ranking

To find the most important variables related to sludge bulking, the global sensitivity analysis presented in Section 2.4 was used. As a first step, GSA was performed for the groups of variables to evaluate the contribution of the different groups (disturbances, DO, AE, N concentrations) to sludge bulking conditions. Group sensitivity was performed for the models in Table 1. The results are shown in Fig. 6 with a boxplot diagram. The red lines show the median and the red crosses show the mean values. We can see that disturbance variables have the highest impact on classification model performance. This could be expected since the temperature is one of the most influential parameters affecting the growth of *M. parvicella*, which can be directly seen in the data. The results also indicate high sensitivity to other groups of variables, especially to N concentrations and to a lesser extent to DO concentrations and aeration intensities. This is an important result for the control of sludge bulking since it indicates that operating variables contribute to sludge bulking. Therefore, appropriate adjustment of operating conditions can potentially prevent or mitigate sludge bulking.

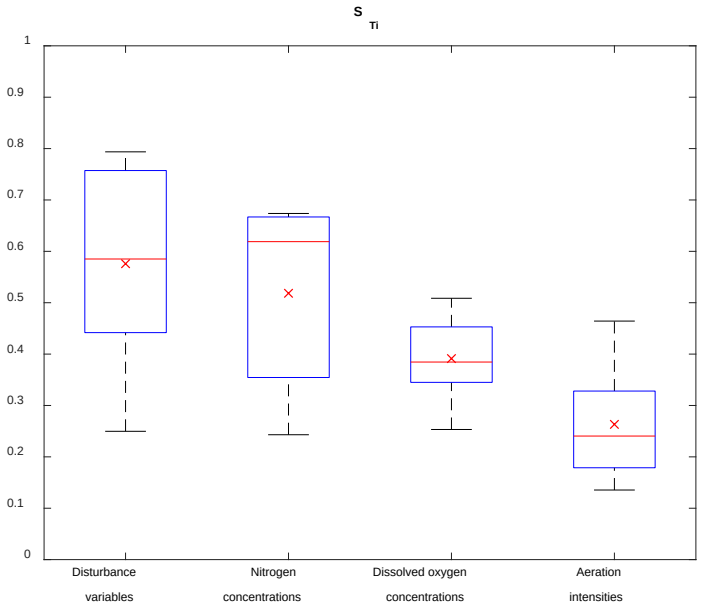


Fig. 6. The total-order sensitivity index for the groups of variables obtained with the best performing models with all input variables.

In the next step, a sensitivity analysis was also performed for individual variables. Fig. 7 shows the first-order and total-order sensitivity indices for the best performing classification models. The highest first-order values are obtained for temperature (T) and nitrate concentrations $NO3_1$ and $NO3_3$. The total-order indices additionally indicate the importance of some other variables. Higher mean and 75th percentile values of the total-order index are also obtained for dissolved oxygen concentrations DO_4 and DO_3, aeration intensity AE_2, mass flow Φ_m, influent flow Q and nitrate concentration $NO3_2$. All these variables have a mean total interaction effect greater than 0.1. This set of 9 variables was further considered for the reduced model training. Repeated GSA for the 9 selected variables revealed high total-sensitivity indices for T, $NO3_3$, $NO3_1$, medium for DO_4, Q, Φ_m, and low for AE_2, $NO3_2$, DO_3.

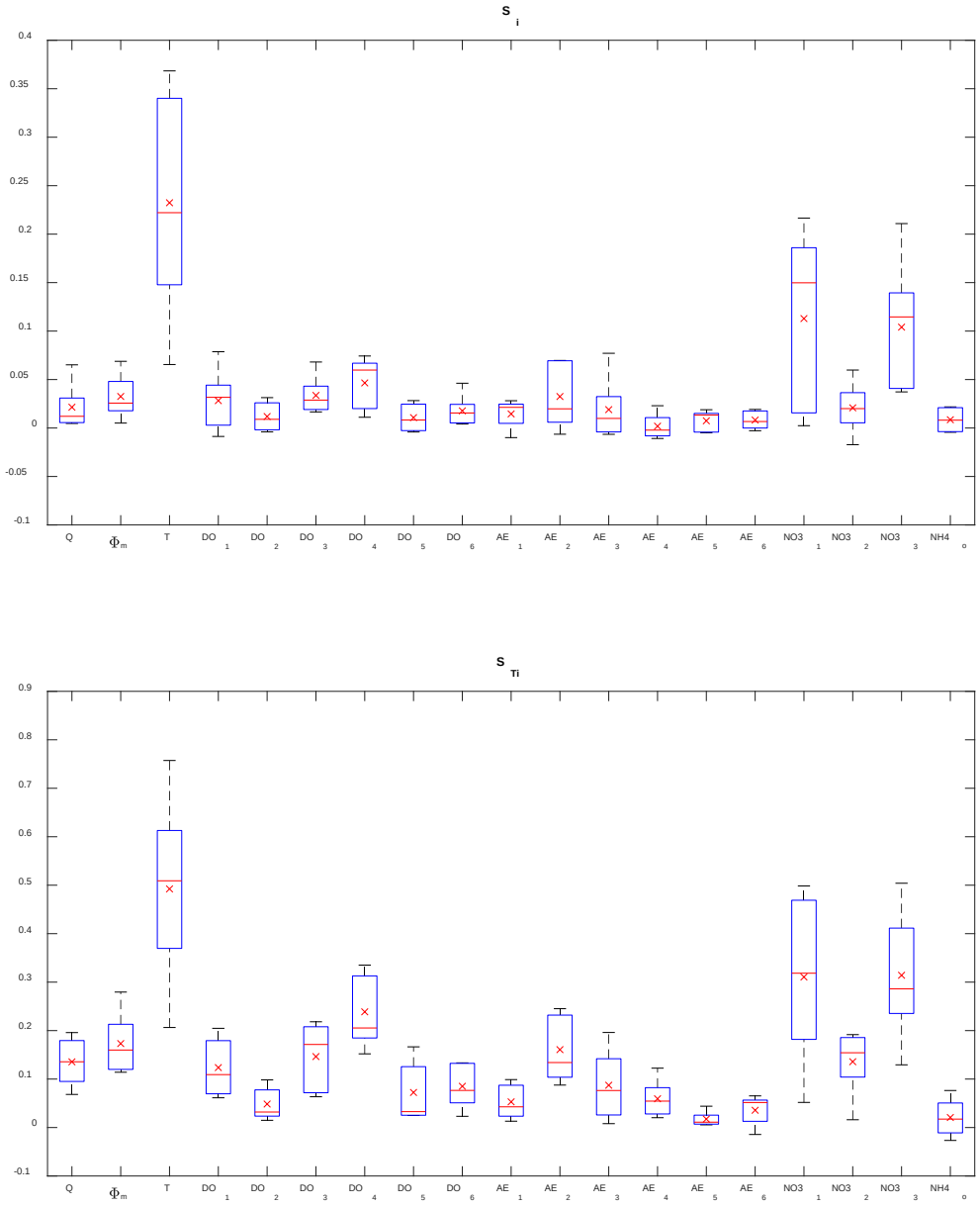


Fig. 7. The first-order and the total-order sensitivity indices of individual variables for the best performing models with all input variables.

3.2.4 Step 5: Reduced model training

Based on the reduced set of variables, the models with different number and combinations of variables were tested. The models with 5-7 variables give the best performance. The finally selected input variables of the reduced model are

- three disturbance variables (Q, Φ_m, T) and
- four control variables (DO_3, DO_4, $NO3_1$, $NO3_3$).

For the selected set of variables, the performance of different classification algorithms was tested. The results of the best models for the six groups of classification methods are shown in Table 2. The table shows the accuracy of models, the performance per class, and the percentage of data in the available full-scale data-set that change from 'B' to 'NB' state by manipulating each model input variable within the $[LA, UA]$ interval.

The model with the best performance is Linear SVM. Its performance accuracy is 89.6 %, 90.5 % and 89.0 % in 4-fold cross-validation, training data and test data, respectively. As we can see, the performance is similar to models with a complete set of variables in Table 1.

We can also see that the methods give a different performance, but they can be roughly grouped into two groups. The first three groups of methods (1 to 3 in Table 2) give similar performance, which can be characterized by higher accuracy, better performance per class but also higher sensitivity to control variables. The other three groups of methods (4 to 6 in Table 2) give lower accuracy, worse performance per class, in particular 'NB', but also lower sensitivity to control variables. The interesting result is that, for disturbance variables, the sensitivity of almost all models is more similar. Thus, although models 1 to 3 perform better and should be more trusted, their performance with respect to control variables should be treated with caution, as they may be affected by a small number of data in the data set or by model overfitting.

This problem was investigated by constructing a simplified classification model with only the two most influential variables (T and $NO3_3$) as model inputs. The aim was to observe the space of input variables visually in a two-dimensional graph. Plots of the classification results for different methods are shown in the supplementary material (Fig. S1). The results are presented for the best performing method in each of the six groups of classification algorithms in the case of the 2-input variable model. The results show that all models identify the regions of good and poor settleability at high and low temperatures,

respectively. At intermediate temperatures, data points of both 'B' and 'NB' classes are present but their number is relatively small. Therefore, the sensitivity to $NO3_3$ is identified differently by different models. For example, in the cases of Ensemble Subspace Discriminant and Coarse Tree methods, the model is almost insensitive to $NO3_3$, while in the cases of other methods close to a linear relationship in the models was identified. This difference in sensitivity to $NO3_3$ was obtained despite the similar performance of all models. Model accuracy in the cases presented was 78.4 to 80.2 % in the 4-fold cross-validation, 79.3 to 82.3 % for the training data and 81.8 to 83.1 % for the test data.

Table 2. Results of different classification methods for the reduced model with three disturbance variables (Q, Φ_m, T) and four control variables (DO_3, DO_4, $NO3_1$, $NO3_3$).

	Group of classification methods	The method with the best performance in a group	Accuracy in 4-fold cross-validation (%)	Accuracy in training data (%)	Accuracy in test data (%)	Performance per class (% of data)		The percentage of data points changed from 'B' to 'NB' state within the tested range of each variable[1]						
								Disturbance variables			Control Variables			
						True positive rate 'B'	True positive rate 'NB'	T	Φ_m	Q	$NO3_3$	DO_4	$NO3_1$	DO_3
1	Support Vector Machines (SVM)	Linear SVM	89.6	90.54	89.04	92	86	+++	+	+	++	++	+	+
2	Logistic Regression Classifier	Logistic regression	89.2	90.99	91.78	92	85	+++	+	+	+++	++	++	++
3	Discriminant Analysis	Linear Discriminant	88.7	91.44	90.41	90	86	+++	+	+	+++	++	++	+
4	Ensemble Classifiers	Ensemble Subspace Discriminant	87.4	87.84	93.15	92	78	+++			+	+		
5	Nearest Neighbour Classifier	Cubic KNN	86.0	86.49	86.30	92	76	+++	+	+		+		
6	Decision Trees	Coarse Tree	80.6	87.39	82.19	84	75	+++	+++	+++				

[1] Legend: +++ high (> 75%), ++ medium (> 35% & ≤ 75%), + low (> 0% & ≤ 35%), empty (0%)

3.3 Implication of modelling results on plant operation

The models obtained were used to evaluate the changes that need to be imposed on each variable to achieve non-bulking conditions. This knowledge is useful for a daily operation to pursue non-bulking plant operating conditions. The evaluation was performed by considering bulking data 'B' in the available full-scale data-set and observing the required change (increase, decrease) of each model input variable to change from 'B' to 'NB' state. An example of the analysis for six random data points and the best performing Linear SVM model is presented in the supplementary material (Fig. S2). The results show that the settleability is improved at higher Q, lower Φ_m, higher T, higher DO_3, lower DO_4, lower $NO3_1$, and higher $NO3_3$. It should be noted that these results are consistent for all data points for which the 'NB' state is reached within the $[LA, UA]$ interval and for different classification models.

For some variables, the suggested changes could be well associated with process knowledge and plant observations. The effect of higher temperature on better sludge settleability has already been discussed. The effects of Q and Φ_m are also consistent with plant observations. Sludge settleability deteriorates during long periods of dry weather conditions when Q is low and sludge accumulates at the bottom of the tanks due to low hydraulic load and poor mixing. Such conditions with longer SRTs are favourable for *M. parvicella* growth. An additional feature of plant operation is also occasional high influent ammonia concentration, and thus high Φ_m, caused by an additional load from the sludge treatment line during periods of sludge dewatering. These conditions represent N sources for increased filamentous growth.

Concerning the control variables, maintaining appropriate nitrate concentrations is suggested as the most powerful control action according to sensitivity analysis. The proposed changes to improve settleability are to decrease the nitrate concentration in the anoxic tank of the first cascade ($NO3_1$) and increase the concentration in the third cascade ($NO3_3$). This could indicate that a nitrate gradient should be established along the biological stage with lower nitrate concentrations in the initial tanks and higher in the final tanks. This would prevent the availability of N sources in the initial stage, where

the presence of *M. parvicella* is greatest. The proposed control adjustments concern also DO levels in the reactors of the second cascade. This could also be related to poor mixing conditions since biomass retention is greatest in this part of the plant and more aeration is required to improve mixing.

From the data-driven analysis, it could be concluded that besides temperature, the availability of nitrogen in the initial tanks of the biological stage, both in the form of input ammonia or nitrate, is identified to be related to sludge settleability. Low hydraulic load and poor mixing are also potential causes of poor sludge settleability, while the effects of low DO concentrations could not be confirmed.

3.4 Discussion

We derived the models of sludge settleability following an established SVI criterion and adapted it to the measurement of the settled sludge volume SV. However, clustering in two classes and performing binary classification have hindered the classification process. In particular, the changes in operating conditions that result in the change of SV within the class are not reflected in the classification data. Therefore, they could not be identified by the model. On the other hand, modelling SV as a continuous variable would cause other problems related to insufficient and unreliable data. Therefore, an appropriate number of classes should be determined in the future.

The presented work also shows that observing only the model quality on training and test data may not be a sufficient criterion for assessing the quality of the model for the intended model use. In particular, it was only after model simplification and visual inspection of the process data that it became apparent that the available process data may not be sufficient to identify the desired relationships in the data. The application of several classification methods was favourable in this case to encounter different predictions for different models. However, in the future, also other methods should be considered, for example, Bayesian learning approaches or methods that deal with the model uncertainty in prediction systematically if insufficient data for model identification (Schweidtmann et al., 2020). An example of such methods would be Gaussian Process models (Kocijan, 2016).

4. Conclusions

We have proposed a procedure that is important for the prevention of sludge bulking in WWTPs. The procedure identifies process operating variables that can be used in the control of sludge bulking. It relies on data-driven clustering and classification using global sensitivity analysis for input variables ranking. The proposed input variable selection method can be classified as a wrapper method with backward elimination of input variables. It allows keeping a large number of input variables but limits the number of tested candidate models. The procedure helps in ranking model input variables in cases where different combinations of variables give similar performance and reduction based solely on model performance is not possible.

The application on a full-scale WWTP indicates that bulking could be controlled by limiting N-sources at the beginning of the biological stage. However, the validity of the model is limited if insufficient data for model identification is available. Therefore, we propose to couple data-driven classification with experimental design and modelling techniques, e.g. Gaussian process models, to provide data with sufficient excitation and to systematically address model uncertainty.

Acknowledgements

This work was financially supported by the Slovenian Research Agency, program P2-0001, and Public Water Utility JP Komunala Kranj, d.o.o. The authors would like to thank the WWTP personnel, B. Bajželj, L. Janeš and M. Margetič, for their assistance with the collection of plant data and information on plant operation. B. Bajželj is the author of Fig. 2a and Fig. 2b, L. Janeš is the author of Fig. 2c.

References

Amirruddin, A. D., Muharam, F. M., Ismail, M. H., Ismail, M. F., Tan, N. P., Karam, D. S. (2020). Hyperspectral remote sensing for assessment of chlorophyll sufficiency levels in mature oil palm

(Elaeis guineensis) based on frond numbers: Analysis of decision tree and random forest. *Computers and Electronics in Agriculture*, **169**, 105221.

Andersen, C. M., Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, **24**, 728-737.

Bagheri, M., Mirbagheri, S. A., Bagheri, Z., Kamarkhani, A. M. (2015). Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. *Process Safety and Environmental Protection*, **95**, 12-25.

Capodaglio, A. G., Jones, H. V., Novotny, V., Feng, X. (1991). Sludge bulking analysis and forecasting: Application of system identification and artificial neural computing technologies. *Water Research*, **25** (10), 1217-1224.

Cheng, H., Wu, J., Huang, D. (2019). A novel fault identification and root-causality analysis of incipient faults with applications to wastewater treatment processes. *Chemometrics and Intelligent Laboratory Systems*, **188**, 24-36.

Chmielowski, K., Czekański, A., Leśniańska, A. (2019). Using Data Mining to Predict Sludge and Filamentous Microorganism Sedimentation. *Pol. J. Environ. Stud.*, **28** (5), 3105-3113.

Comas, J., Rodríguez-Roda, I., Gernaey, K.V., Rosen, C., Jeppsson, U., Poch, M. (2008). Risk assessment modelling of microbiology-related solids separation problems in activated sludge systems. *Environmental Modelling & Software*, **23**, 1250-1261.

Deepnarain, N., Nasr, M., Kumari, S., Stenström, T. A., Reddy, P., Pillay, K., Bux, F. (2019). Decision tree for identification and prediction of filamentous bulking at full-scale activated sludge wastewater treatment plant. *Process Safety and Environmental Protection*, **126**, 25–34.

Guo, J., Peng, Y., Wang, S., Yang, X., Wang, Z., Zhu, A. (2012). Stable limited filamentous bulking through keeping the competition between floc-formers and filaments in balance. *Bioresource Technology*, **103**, 7-15.

Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157-1182.

Han, H. G., Li, Y., Guo, Y. N., Qiao, J. F. (2016). A soft computing method to predict sludge volume index based on a recurrent self-organizing neural network. *Applied Soft Computing*, **38**, 477-486.

Han, H. G., Dong, L. X., Qiao, J. F. (2021). Data-knowledge-driven diagnosis method for sludge bulking of wastewater treatment process. *Journal of Process Control*, **98**, 106-115.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International yearbook of cartography*, **7**, 186-190.

Jin, B., Wilén, B. M., Lant, P. (2003). A comprehensive insight into floc characteristics and their impact on compressibility and settleability of activated sludge. *Chemical Engineering Journal*, **95** (1-3), 221-234.

Kocijan, J. (2016). *Modelling and Control of Dynamic Systems Using Gaussian Process Models*. Springer International Publishing, Cham.

Lakay, M.T., Hulsman, A., Ketley, D., Warburton, C., de Villiers, M., Casey, T.G., Wentzel, M.C., Ekama, G.A. (1999). Filamentous organism bulking in nutrient removal activated sludge systems. Paper 7: Exploratory experimental investigations. *Water SA*, **25**,383-396.

Liu, Y., Xiao, H., Pan, Y., Huang, D., Wang, Q. (2016a). Development of multi-step soft-sensors using a Gaussian process model with application for fault prognosis. *Chemometrics and Intelligent Laboratory Systems*, **157**, 85-95.

Liu, Y., Guo, J., Wang, Q., Huang, D. (2016b). Prediction of Filamentous Sludge Bulking using a State-based Gaussian Processes Regression Model. *Scientific Reports*, **6**, 31303.

Liu, Y., Yuan, L., Huang, S., Huang, D., Liu, B. (2020). Integrated design of monitoring, analysis and maintenance for filamentous sludge bulking in wastewater treatment. *Measurement*, **155**, 107548.

Lou, I., Zhao, Y. (2012). Sludge Bulking Prediction Using Principle Component Regression and Artificial Neural Network. *Mathematical Problems in Engineering*, 237693.

Makrygiorgos, G., Maggioni, G.M., Mesbah, A. (2020). Surrogate modeling for fast uncertainty quantification: Application to 2D population balance models. *Computers & Chemical Engineering*, **138**, 106814.

Mathworks (2020). Statistics and Machine Learning Toolbox™ User's Guide R2020a.

May, R., Dandy, G., Maier, H. (2011). Review of Input Variable Selection Methods for Artificial Neural Networks. In: Suzuki K, editor. Artificial Neural Networks - Methodological Advances and Biomedical Applications. *InTech*; 2011. doi:10.5772/16004.

MS (2020). Clustering via Jenks Natural Breaks (https://github.com/MSH19/Clustering-via-Jenks-Natural-Breaks-), GitHub. Retrieved July 13, 2020.

Musvoto, E.V., Lakay, M.T., Casey, T.G., Wentzel, M.C ., Ekama, G.A. (1999). Filamentous organism bulking in nutrient removal activated sludge systems. Paper 8: The effect of nitrate and nitrite. *Water SA*, **25**, 397-408.

Nielsen, P.H., Kragelund, C., Seviour, R.J., Nielsen, J.L. (2009). Identity and ecophysiology of filamentous bacteria in activated sludge. *FEMS Microbiology Reviews*, **33**, 969–998.

Nittami, T., Kasakura, R., Kobayashi, T., Suzuki, K., Koshiba, Y., Fukuda, J., Takeda, M., Tobino, T., Kurisu, F., Rice, D., Petrovski, S., Seviour, R. J. (2020). Exploring the operating factors controlling Kouleothrix (type 1851), the dominant filamentous bacterial population, in a full-scale A2O plant. *Scientific Reports*, **10**, 6809.

Pianosi, F., Sarrazin, F., Wagener, T. (2015). A Matlab toolbox for Global Sensitivity Analysis, *Environmental Modelling & Software*, **70**, 80-85.

Rossetti, S., Tomei, M.C., Nielsen, P.H., Tandoi, V. (2005). ‘‘Microthrix parvicella’’, a filamentous bacterium causing bulking and foaming in activated sludge systems: a review of current knowledge. *FEMS Microbiology Reviews*, **29**, 49–64.

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M. (2004). *Sensitivity analysis in practice. A guide to assessing scientific model*. John Wiley & Sons.

Schweidtmann, A.M., Weber, J.,M., Wende, C., Netze, L., Mitsos, A. (2020). Obey validity limits of data-driven models. Preprint.

Sobol’, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, **55**, 271-280.

Szeląg, B., Drewnowski, J., Łagód, G., Majerek, D., Dacewicz, E., Fatone, F. (2020). Soft Sensor Application in Identification of the Activated Sludge Bulking Considering the Technological and Economical Aspects of Smart Systems Functioning. *Sensors*, **20**, 1941.

Šindelář, R., Babuška, R. (2004). Input selection for nonlinear regression models. *IEEE Transactions on fuzzy systems*, **12**(5), 688-696.

Tsai, M.W., Wentzel, M.C., Ekama, G.A. (2003). The effect of residual ammonia concentration under aerobic conditions on the growth of Microthrix parvicella in biological nutrient removal plants. *Water Research*, **37**, 3009-3015.

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. *Science of the Total Environment*, **784**, 147138.

Wei, P., Lu, Z., Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, **142**, 399-432.

Zhang, X., Zheng, S., Xiao, X., Wang, L., Yin, Y. (2017). Simultaneous nitrification/denitrification and stable sludge/water separation achieved in a conventional activated sludge process with severe filamentous bulking. *Bioresource Technology*, **226**, 267-271.

Supplementary Material for

Input variable selection using machine learning and global sensitivity methods for a full-scale wastewater treatment plant

Nadja Hvala[a,*], Juš Kocijan[a,b]

[a] Jožef Stefan Institute, Department of Systems and Control, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

[b] University of Nova Gorica, Vipavska cesta 13, SI-5000 Nova Gorica, Slovenia

[*] Corresponding author.
E-mail address: nadja.hvala@ijs.si (N. Hvala)

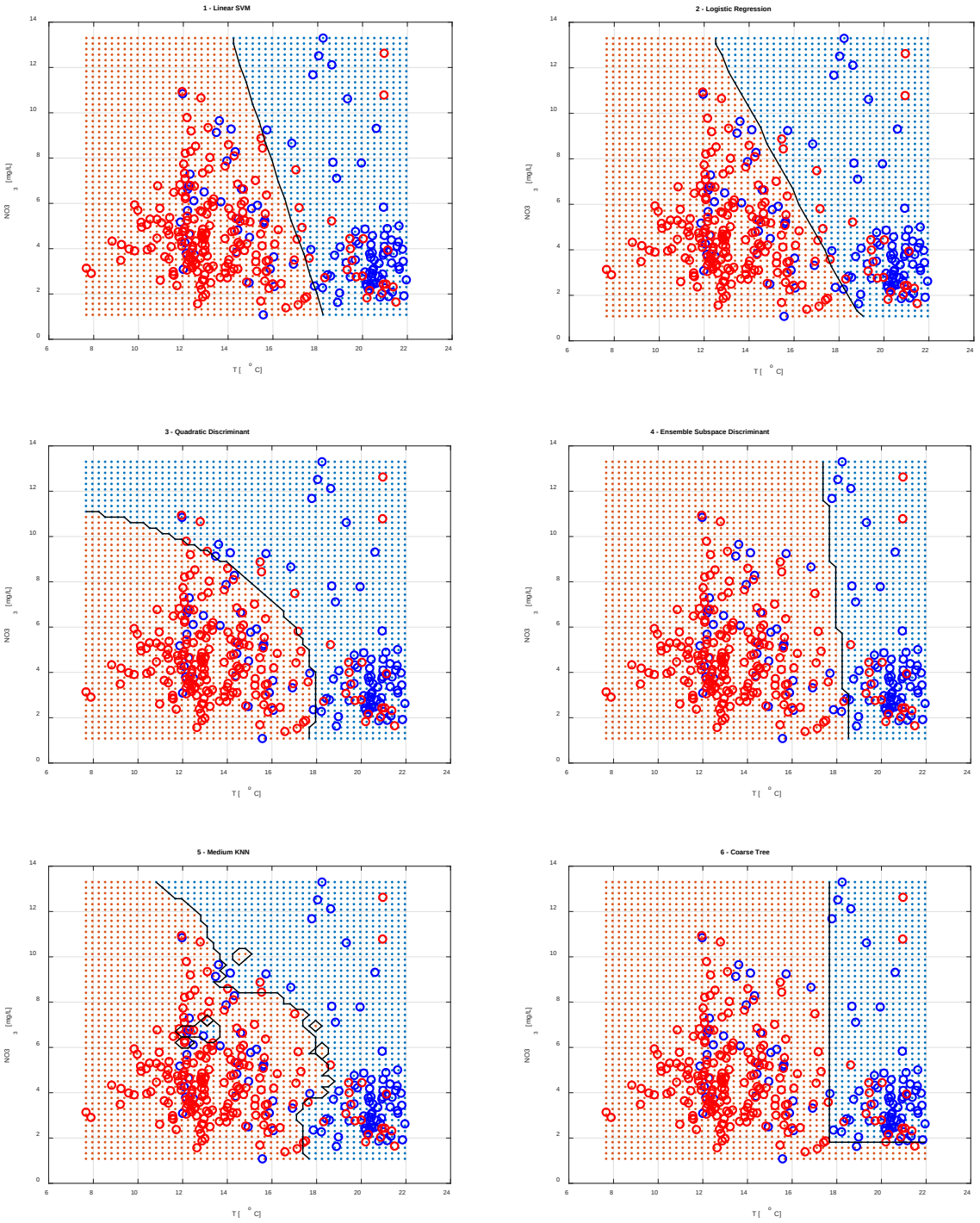


Fig. S1. Classification results for different classification methods in the case of a 2-input-variable model (*T*, $NO3_3$). Red dots indicate the regions of bulking conditions ('B') and blue dots indicate regions with non-bulking conditions ('NB') as determined by the classification model. The border between the regions is indicated by the black line. Circles present measured data points in 'B' (red circles) and 'NB' (blue circles) classes. Good model quality is indicated by the red circles on the red dotted area and the blue circles on the blue dotted area.

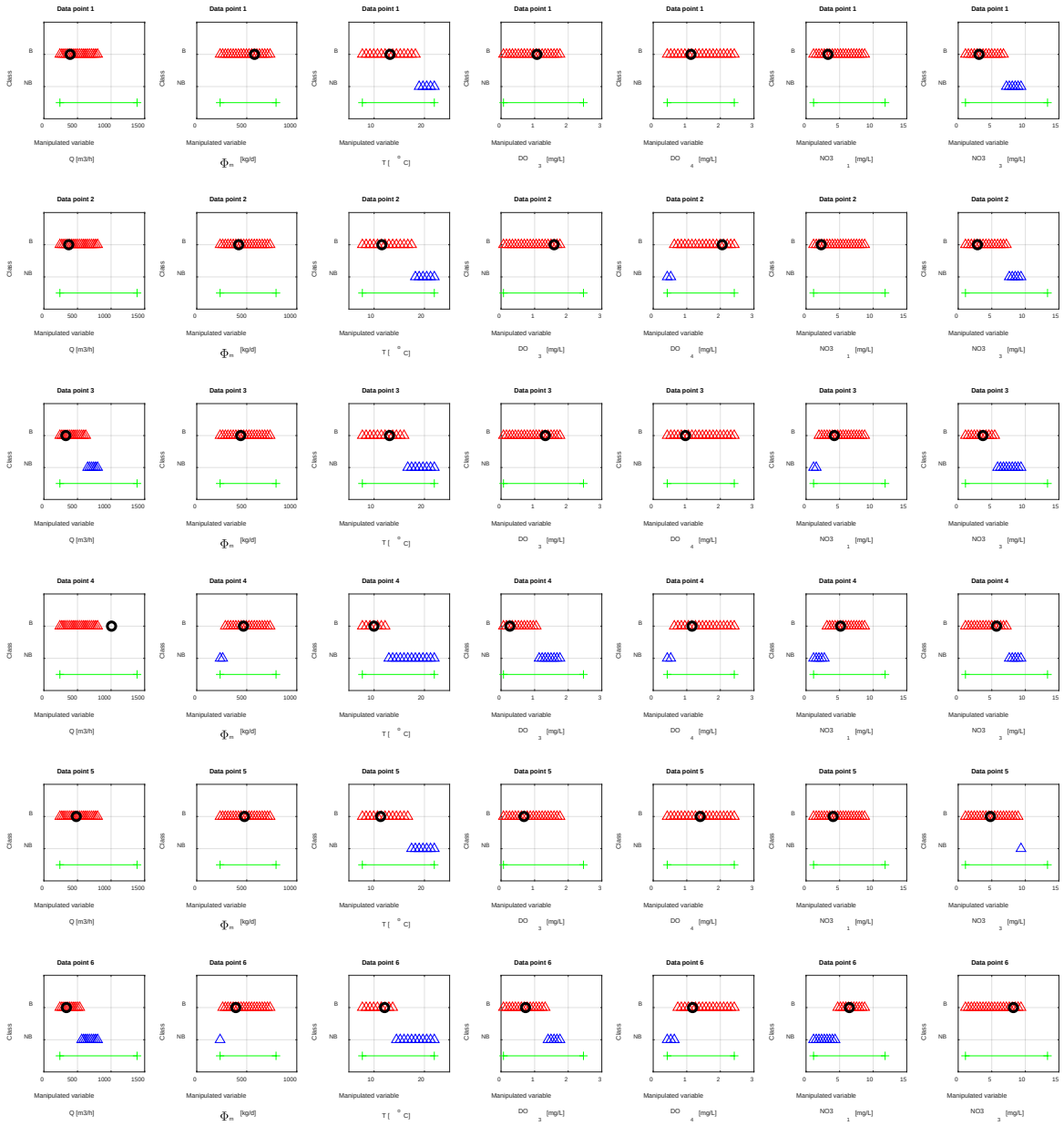


Fig. S2. An example of sensitivity test to model input variables shown for six random data points in the data set. Plots in a row show one data point and the variation of the settleability class as predicted by the model if a selected input variable is manipulated within the $[LA, UA]$ interval. The red triangles indicate the bulking class 'B', the blue triangles indicate the non-bulking class 'NB', the black circles present the measured value of the input variable and the measured output class, the green intervals show the region of measured values of each input variable.