

COD and NH₄-N Estimation in the Inflow of Wastewater Treatment Plants using Machine Learning Techniques

Peter Kern, Christian Wolf, Daniel Gaida, Michael Bongards and Seán McLoone

Abstract— The in-line measurement of COD and NH₄-N in the WWTP inflow is crucial for the timely monitoring of biological wastewater treatment processes and for the development of advanced control strategies for optimized WWTP operation. As a direct measurement of COD and NH₄-N requires expensive and high maintenance in-line probes or analyzers, an approach estimating COD and NH₄-N based on standard and spectroscopic in-line inflow measurement systems using Machine Learning Techniques is presented in this paper. The results show that COD estimation using Random Forest Regression with a normalized MSE of 0.3, which is sufficiently accurate for practical applications, can be achieved using only standard in-line measurements. In the case of NH₄-N, a good estimation using Partial Least Squares Regression with a normalized MSE of 0.16 is only possible based on a combination of standard and spectroscopic in-line measurements. Furthermore, the comparison of regression and classification methods shows that both methods perform equally well in most cases.

I. INTRODUCTION

Chemical Oxygen Demand (COD), ammonium-nitrogen (NH₄-N) and phosphate (PO₄-P) are the substances in the inflow of a Wastewater Treatment Plant (WWTP) that are commonly used to operate and control the biological purification processes [1, 2]. While NH₄-N and PO₄-P are chemical substances, COD is a sum parameter which represents the amount of organic compounds in wastewater. On the basis that nearly all organic compounds can be fully oxidized, COD indicates the mass of oxygen consumed per litre of wastewater and is used as a substitute variable for carbon [3].

The in-line measurement of NH₄-N and COD concentrations in the WWTP inflow provides several advantages: (1) It is possible to optimize WWTP control strategies knowing in advance the amount of NH₄-N and COD flowing into the bioreactors. E.g. predictive control can be used to increase the oxygen (O₂) concentration when a load peak is detected in the inflow; (2) It allows the evaluation of mass balances and plant efficiency comparing inflow and effluent concentrations; (3) A calibration of WWTP simulation models such as the Activated Sludge Model (ASM) [4], requires a detailed measurement of these process variables in the WWTP inflow.

Based on recommendations of the DWA (the German Water Association) most plants in Germany are not equipped with probes for NH₄-N and COD. The main reasons are high maintenance and high cost of in-line measurement systems.

Peter Kern is with the Department of Electronic Engineering at the National University of Ireland Maynooth, Co. Kildare, IRELAND (e-mail: peter.kern.2009@nuim.ie).

Christian Wolf, Daniel Gaida and Michael Bongards are with the Gummertsbach Environmental Computing Center at Cologne University of Applied Sciences (e-mail: christian.wolf@fh-koeln.de, daniel.gaida@fh-koeln.de, michael.bongards@fh-koeln.de).

Depending on the operating principle of measurement systems, prices for NH₄-N probes lie between €3,500 for ion-selective probes and €20,000 for chemical analyzers. Prices for spectrometric COD probes vary between €10,000 and €25,000. Another reason is the harsh environment in the inflow. Raw-Inflow-Water can be characterized by the following attributes:

- strong variations in concentration
- strong temperature variations
- high fat / grease content
- high flow rates
- variations in pH-value

This results in higher maintenance costs due to the operation in bioreactors and involves regular cleaning and calibration in short intervals [5]. Thus, an alternative measurement solution which is sufficiently accurate and low-maintenance is needed and described in this paper. In this alternative approach NH₄-N and COD concentrations in the inflow are estimated based on process variables which are commonly measured at WWTPs or that are financially feasible compared to NH₄-N and COD, such as turbidity. The estimation of the target variables NH₄-N and COD is done by different regression and classification methods.

The development of so-called soft-sensors to estimate key variables in wastewater is a well-known approach. In the past soft-sensors were developed for COD [6, 8, 9], Biological Oxygen Demand (BOD₅) [8], Totally Suspended Solids (TSS) [6], NH₄-N [9] and nitrate (NO₃-N) [10]. Nevertheless, they rely on expensive or high maintenance surrogate measurement probes such as near-infrared, UV/vis and synchronous fluorescence spectroscopic in-line probes, which were specifically designed for COD or NH₄-N measurement, or on complex chemical analyzers. A review of previous research on soft-sensors for WWTP is given in Haimi et al. [7].

The remainder of the paper is organized as follows. Section II gives an overview of the materials and methods used, describes the measurement campaign conducted, the data sets used for COD and NH₄-N estimation and the data pre-processing undertaken. A brief description of the applied regression and classification methods is also given. Then Section III presents results for the application of these machine learning techniques to the estimation of COD and NH₄-N inflow concentrations. Finally, conclusions are presented in Section IV.

Seán McLoone is with Queen's University Belfast, UK as Director of the Energy, Power and Intelligent Control Research Cluster (e-mail: s.mcloone@qub.ac.uk).

We thank Endress+Hauser Conducta for the provision of the online-measurement systems that were used in the measurement campaign and the staff of the Rospe WWTP for their support.

II. MATERIALS AND METHODS

A. Measurement Campaign

In order to properly measure wastewater composition in the WWTP inflow, a set of in-line probes were installed at the Rospe WWTP of the Aggerverband¹. Overall, the measurement campaign for this paper was conducted in 2012 over a period of two months. The Rospe WWTP is municipal, connected to approximately 20,000 population equivalents (PE) and treats the wastewater of the German city of Gummersbach. As there is currently no relevant industrial discharger connected, the wastewater can be considered to be mostly municipal. Furthermore, the Rospe WWTP is continuously operated with upstream de-nitrification. Before the water reaches the primary treatment, it is passed through a 6 mm screen and a grid chamber. The in-line probes already installed at the plant as well as the newly installed probes are situated between the screen and grid chamber. Therefore, it is still raw water with only large compounds that could potentially mechanically damage the probes removed. Standard inflow instrumentation at the Rospe WWTP consists of a magnetic flow meter (MID), a pH-probe, a conductivity (Cond.) sensor and a temperature sensor.

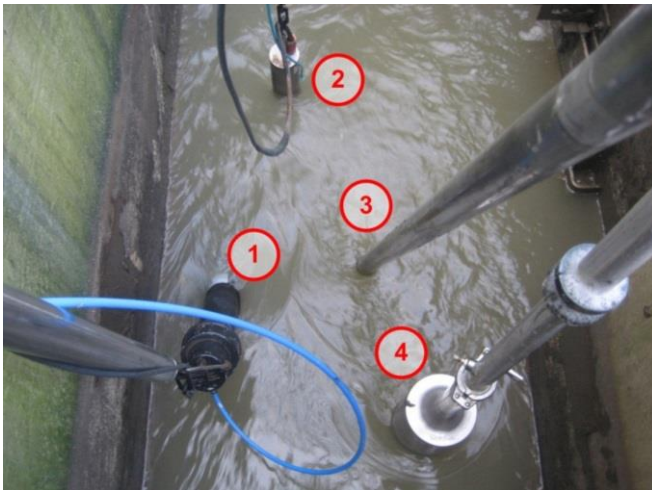


Figure 1. Online-measurement probes in the inflow of the Rospe WWTP

Figure 1 depicts the additionally installed in-line probes for WWTP inflow characterization:

1. Endress + Hauser - ISEmax CAS40D (ion-selective $\text{NH}_4\text{-N}$ measurement probe)
2. Trios ProPS-WW (spectrometric COD measurement probe 190 nm -360 nm)
3. Endress + Hauser - Turbimax CUS51D (turbidity measurement probe 860 nm)
4. Endress + Hauser - STIP-scan (spectrometric COD measurement probe 200 nm-680 nm)

The Turbimax CUS51D uses reflected light at 860 nm to determine the turbidity in formazin nephelometric units (FNU). Costs for this probe are approximately €2,500. The Stip-scan

COD probe provides not only a measurement value for COD but also the spectral absorption coefficient at 254 nm (SAC254) and 433 nm (SAC433) as well as a turbidity measure (Absorptiometric Turbidity Units (ATU)). The ISEmax CAS40D is only used to determine $\text{NH}_4\text{-N}$ and the Trios ProPS-WW was used as a backup for the Stip-scan as it can measure COD, TOC and nitrate ($\text{NH}_3\text{-N}$).

TABLE I shows all process variables measured during the campaign. They were measured with a sample time of three minutes over a period of two months. After reviewing an excerpt of 14,380 samples and eliminating corrupt data, a subset of 9,843 samples from March 2012 was chosen. This leads to an investigated period of approximately three weeks. Looking at the min and max values in TABLE I it is obvious that measured concentrations are relatively low compared to a typical WWTP inflow, which might be due to infiltration water.

TABLE I. INPUT PROCESS VARIABLES

Variable	Probe	Min Value	Max Value	Mean	Std. (σ)
COD [mg/l]	Stip-scan	3.1	462.0	169.2	85.6
$\text{NH}_4\text{-N}$ [mg/l]	CAS40D	0.1	27.0	9.63	4.1
FNU_{860} [m^{-1}]	CUS51D	0.0	322.9	53.4	31.0
ATU [m^{-1}]	Stip-scan	0.0	34.4	8.0	4.47
SAC_{254} [m^{-1}]	Stip-scan	0.47	88.4	32.4	14.9
SAC_{433} [m^{-1}]	Stip-scan	0.02	10.8	3.19	1.98
Cond. [$\mu\text{S}/\text{cm}$]	WTW	0.0	1095	412	105.0
Temperature [$^{\circ}\text{C}$]	WTW	5.3	9.6	7.6	0.81
Flow rate [l/s]	MID	0.0	357.6	134.3	53.7
pH-Value	WTW	6.9	8.5	7.4	0.22

B. Data Preparation

The raw data sets are preprocessed using global outlier elimination followed by local outlier elimination to allow for better regression and classification results. Global outliers are samples that have a deviation of the median which is greater than a multiple of the median absolute deviation (MAD) of the complete dataset. For local outlier detection, the method described by Menold et al. [11] is used. In the last preparation step, the data is normalized and divided into a training and validation dataset (70 %, 19 days) and a chronologically separated test dataset (30 %, 8 days). From the training data two sets are randomly selected and used as training and validation data for parameter optimization of the applied Machine Learning methods. The test dataset contain a period of missing data as can be seen in Figure 2. In addition, the values in the test data are larger than observed in the training data, hence achieving accurate predictions on the test data is challenging.

For each training and test dataset, all observations of variables that are used as predictors are stored in a matrix $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$, where n is the number of predictors and each \mathbf{u}_i consists of m distinct observations of a variable,

¹ Aggerverband - A local water association managing the water-related tasks of the Agger river basin

while the observations of the estimated variable are saved in vector \mathbf{x} .

C. Data Sets

Generation of Input Data Sets

For the practical implementation and application of the estimation methods at a WWTP, it is important to investigate which in-line measurements in the WWTP inflow are necessary to achieve acceptable results for COD and $\text{NH}_4\text{-N}$ estimation. Therefore, three different input data sets \mathbf{U} were generated.

- $\mathbf{U}_1 = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ contains all $n = 9$ input variables except the respective target variable \mathbf{x} .
- $\mathbf{U}_2 = (\mathbf{u}_Q, \mathbf{u}_{pH}, \mathbf{u}_{cond}, \mathbf{u}_T)$ contains only the standard inflow in-line measurements (Flow rate, pH-value, conductivity and temperature).
- $\mathbf{U}_3 = (\mathbf{u}_{FNU}, \mathbf{u}_{pH}, \mathbf{u}_Q)$ contains the variables FNU, pH-value and flow rate which were determined to have significant influence on predictions based on a backward elimination sensitivity analysis.

While several other combinations of input variables were tested, these three datasets are the most interesting combinations for the practical application. The first dataset is a quasi-reference for the best result that can be achieved using all input variables. The second one shows what is possible without investment into additional in-line probes and the third one represents a pragmatic trade-off between investment costs and prediction results (only an additional turbidity probe has to be installed).

Generation of Data Classes

For the application of classification methods, the data was divided into five classes representing very low, low, medium, high and extremely high COD and $\text{NH}_4\text{-N}$ concentrations respectively. This classification has two main benefits for practical implementation: (a) direct visualization of WWTP inflow states for the operator in a “traffic light” fashion through class labels and (b) precise estimation of extreme inflow events in a separate class. Furthermore, the number of classes as well as their spans were also determined based on typical concentrations of the substances in the inflow and to minimize model complexity by capturing inflow dynamics with a minimal number of classes.

TABLE II gives an overview of the generated classes. For both data sets five classes were chosen, where the fifth class contains the highest COD and $\text{NH}_4\text{-N}$ concentrations representing extreme inflow events.

TABLE II. DATA CLASSES

Class	COD [mg/l]	Number of Data Points	$\text{NH}_4\text{-N}$ [mg/l]	Number of Data Points
1	$0 \leq \text{COD} < 100$	2501	$0 \leq \text{NH}_4\text{-N} < 5$	1306
2	$100 \leq \text{COD} < 150$	1359	$5 \leq \text{NH}_4\text{-N} < 8$	2302
3	$150 \leq \text{COD} < 200$	2064	$8 \leq \text{NH}_4\text{-N} < 15$	5109
4	$200 \leq \text{COD} < 300$	3393	$15 \leq \text{NH}_4\text{-N} < 20$	1002
5	$300 \leq \text{COD} < \infty$	526	$20 \leq \text{NH}_4\text{-N} < \infty$	124

D. Regression Methods

Multivariate Linear Regression (MLR)

In MLR the optimal parameter vector \mathbf{b}^* of the linear predictor $\hat{\mathbf{x}} = \mathbf{U} \cdot \mathbf{b}^*$ is determined by solving the classical least squares problem:

$$\mathbf{b}^* = \arg \min_{\mathbf{b}} (\mathbf{x} - \mathbf{U} \cdot \mathbf{b})^T \cdot (\mathbf{x} - \mathbf{U} \cdot \mathbf{b}). \quad (1)$$

Partial Least Squares (PLS)

In PLS the original input matrix \mathbf{U} and the estimated variable \mathbf{x} are both projected into a new space in which the covariance between the projected \mathbf{U} and \mathbf{x} (so called \mathbf{U} scores and \mathbf{x} scores) is maximal. Thus, the \mathbf{U} scores may contain less predictors than \mathbf{U} ; Furthermore, all predictors in the \mathbf{U} scores are orthogonal to the preceding predictors in the \mathbf{U} scores. Here MATLAB's *plsregress* command, which implements the SIMPLS algorithm [12], is used for PLS.

Multilayer Perceptron (MLP)

MLPs belong to the family of artificial neural networks and have several desirable properties like universal function approximation, good generalization and the availability of robust efficient training algorithms [13,14]. In the regression problems considered here a three layer feed forward MLP with 20 hidden neurons is used for modelling. Training is performed using the Levenberg-Marquardt training algorithm which is part of MATLAB's Neural Network toolbox.

Support Vector Regression (SVR)

The most commonly used form of Support Vector Regression (SVR) is called ϵ -SVR and was introduced by [15]. The parameter ϵ defines the upper bound of the prediction error of the SVR model. In its simplest form a SVR is a linear regression model but usually a SVR is nonlinear by introducing so-called kernel functions. The kernel function maps the predictors into a high-dimensional feature space so that also highly nonlinear regression models can be learned. For the SVR used for COD and $\text{NH}_4\text{-N}$ estimation a radial basis function (RBF) kernel is used as it is perfectly suited for a nonlinear relation between \mathbf{U} and \mathbf{x} . Another advantage of the RBF kernel is its ability to capture linear relations as the linear kernel is a special case of the RBF kernel as proven by [16].

E. Classification Methods

Linear Discriminant Analysis (LDA)

LDA finds a linear transformation \mathbf{A} for predictors \mathbf{U} so that the linearly transformed predictors \mathbf{U}^* , so-called features, can be better linearly separated than \mathbf{U} . The matrix \mathbf{A} is found by maximizing the well-known Fisher discriminant criterion [17].

Random Forests (RF)

RF works with an ensemble of decision trees which map the predictors \mathbf{U} on the respective classes. Each tree is trained with a random subset of the training data resulting in slightly different trees. The classification problem is solved by using the majority vote of all decision trees in the random forest [18].

Support Vector Machines (SVM)

While SVR uses the idea of support vectors for regression, SVM uses the same principle for classification problems. For the classification problem under consideration a C-Support Vector Classification is used with soft margin optimization and a RBF kernel [19] using the SVM implementation LIBSVM [20]. In order to determine optimal parameters for the RBF kernel and the margin, a grid search was conducted for each data set.

F. Comparison of Regression and Classification

Standard for comparison

Due to the fact that regression and classification results are not directly comparable a reasonable reference standard has to be defined. Two aspects have to be considered when defining such a standard: (1) The mass of substance which flows into the plant during a certain period of time; (2) The variations of the inflow concentrations. For this analysis this standard is based on a two hour composite sample of the measured process variables, which is a typical time span used by the Aggverband for laboratory analysis and which has proven to be suitable for the calibration of simulation models. Equation (3) describes the procedure for COD, $t_j = t_0, t_0 + n_m \cdot \Delta t, \dots$:

$$COD_{CS}(t_j) = \frac{1}{n_m} \sum_{i=0}^{n_m-1} COD(t_j + i \cdot \Delta t), \quad (2)$$

where n_m is the number of measurements for the composite sample COD_{CS} and Δt the sample time of the in-line probes.

Back Transformation from Classes to Concentrations

For the comparison of regression and classification results, the predicted classes have to be transformed back to concentrations given in mg/l . This is done using the mean value of all measurements in a particular class.

$$\bar{C} = \left(\sum_{i=1}^N x_i \right) N^{-1} \quad (3)$$

Here \bar{C} is the class mean value, N the number of data points in the particular class and x_i the measured sample i in class C . The results could be described as virtual composite samples.

G. Error Measures

Normalized Mean Squared Error (NMSE)

To enable the comparison between COD and NH_4 -N results the NMSE is used as a standard performance measure. The NMSE is defined as

$$NMSE = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (\hat{x}_i - x_i)^2}{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2} \quad (4)$$

where x_i is the true concentration and \hat{x}_i is the predicted concentration of the i th sample, respectively.

Normalized Misclassification Rate (NMCR)

The NMCR [21] is an error measure which gives an equal weighting to each class independent of the number of data

points per class. The NMCR is calculated from the confusion matrix \mathbf{K} , whose rows sum to 100.

$$NMCR = 100 - \frac{1}{C} \text{sum}(\text{diag}(\mathbf{K})) \quad (5)$$

III. RESULTS

Before making a direct comparison and performance evaluation of regression and classification results against real measured COD and NH_4 -N samples, regression and classification results are presented and evaluated separately in the following two subsections.

A. Regression Results

Looking at TABLE III, it becomes obvious, that the estimation for dataset $U_{1,COD}$ using all input variables achieves very good results. This was expected, as the dataset contains SAC_{254} , which represents the absorption at the wavelength of 254 nm, where carbon has the highest absorption within the UV/vis band. This is why the variable SAC_{254} is often used as a replacement for direct COD measurement. Datasets $U_{2,COD}$ and $U_{3,COD}$ without the SAC_{254} achieve considerably worse results. Furthermore, the additional turbidity probe whose measurements are included in dataset $U_{3,COD}$ only provides a small improvement. For COD estimation, the best performance was obtained by RF_{reg} and SVR. The NH_4 -N datasets show similar results. While the best results are achieved for dataset U_{1,NH_4} , the results for datasets U_{2,NH_4} and U_{3,NH_4} are considerably worse.

TABLE III. REGRESSION RESULTS FOR TEST DATA – NMSE [$\times 100$ %]

Dataset	MLR	RF_{reg}	MLP	PLS	SVR
$U_{1,COD}$	0.02	0.02	0.05	0.02	0.02
$U_{2,COD}$	0.28	0.30	0.90	0.28	0.36
$U_{3,COD}$	0.30	0.16	0.21	0.30	0.21
U_{1,NH_4}	0.17	0.47	0.72	0.16	0.65
U_{2,NH_4}	0.49	0.86	1.26	0.49	0.83
U_{3,NH_4}	0.56	0.70	1.13	0.56	0.90

Interesting is the fact that for NH_4 -N the linear methods MLR, and PLS show significantly better performance than the more complex non-linear methods.

B. Classification Results

The classification results (TABLE IV) confirm the regression results. While the results for dataset $U_{1,COD}$ are very good with a NMCR of only 5.33 %, results for datasets $U_{2,COD}$ and $U_{3,COD}$ cannot compete with NMCRs of up to 70 %. All results for NH_4 -N seem to be as bad as the results for dataset $U_{2,COD}$ and $U_{3,COD}$, with the best result achieved by RF_{class} for dataset U_{1,NH_4} with a NMCR of 50.25 %. What is not captured by the NMCR is the “strength” of the misclassification. Keeping in mind that the classes are on a cardinal scale, it makes a big difference for the usability of classification results whether the algorithm confuses adjacent classes or non-adjacent classes. After back transformation (II.F) followed by

calculation of the NMSE, this limitation of the NMCR is compensated as can be seen in the following subsection C.

TABLE IV. MEDIAN (20 REPETITIONS) CLASSIFICATION RESULTS FOR TEST DATA - NMCR [%]

Dataset	RF _{class}	LDA	SVM
U _{1,COD}	5.33	9.33	15.69
U _{2,COD}	52.54	45.40	70.54
U _{3,COD}	35.38	36.22	48.65
U _{1,NH₄}	50.25	51.10	53.38
U _{2,NH₄}	59.87	63.04	61.30
U _{3,NH₄}	59.11	58.81	71.60

C. 2-h Mean Comparison of Regression and Classification Results

For the final comparison of classification and regression results, classification results were transformed back to concentrations using equation (4). In a second step all results (regression and classification) were transformed according to the described 2h reference standard using equation (3). The same was done to the real measured reference values of COD and NH₄-N. From this point on the results are considered to be virtual 2h composite samples. The final step for the comparison is the application of the NMSE (5) to the different datasets and the reference data respectively.

TABLE V gives an overview of the performance of all applied regression and classification methods. The results show that for most cases regression and classification methods achieve similar NMSE values. This effect is obvious for dataset U_{1,COD}, U_{2,COD} and U_{3,COD} in particular. While dataset U_{1,COD} achieves good results no matter which method is used, it is the least interesting dataset for practical utilization, due to the fact that in-line SAC₂₅₄ probes are in the same price range as in-line COD probes. Significantly more interesting are the results for datasets U_{2,COD} and U_{3,COD}. For dataset U_{2,COD} the best performing classification and overall method is LDA which achieves a NMSE of 0.18, while for dataset U_{3,COD} the best classification method with a NMSE of 0.12 is RF_{Class}. The overall best result for U_{3,COD} is achieved by RF_{reg} with a NMSEs of 0.11. This shows that the turbidity probe improves the overall estimation performance over a period of 2 hours, which could not be seen in TABLE IV. Looking at the NH₄-N results only U_{1,NH₄} is of interest for practical utilization. U_{2,NH₄} and U_{3,NH₄} show significant deviations and were not able to

follow the dynamics. In particular the height of the peaks could not be predicted.

Figure 2 shows the comparison results for U_{3,COD}. The middle part is not considered, due to a data gap. It can be seen that the estimated data not only follows the dynamics of the real measured data but also matches the peaks. Nevertheless, RF has problems in the second half of the validation period, where the COD levels rise. This can be attributed to the fact that values as high as these were not present in the training data.

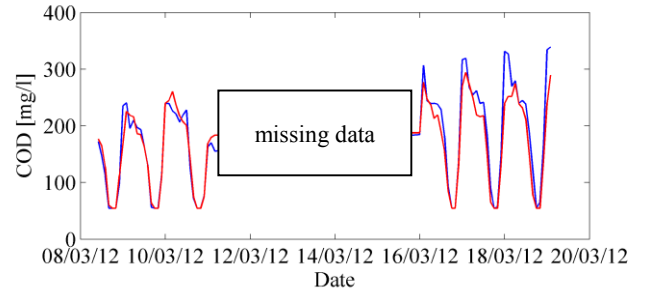


Figure 2. RF Class. 2h mean U_{3,COD} (blue: measured, red: predicted)

Figure 3 shows the results for SVR for U_{2,COD}. While the results are still sufficient in the first half, it is obvious that SVR has problems capturing the peak concentrations. Nevertheless, SVR achieves surprisingly good results in the second half of the validation period, although COD levels are much higher and values this high are not present in the training data.

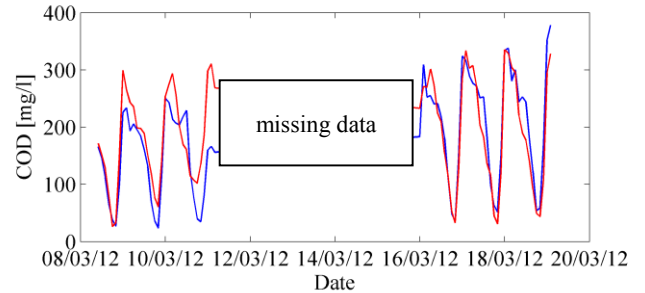


Figure 3. SVR 2h mean U_{2,COD} (blue: measured, red: predicted)

The best result for NH₄-N presented in Figure 4 is achieved for data set U_{1,NH₄}. Although NH₄-N is considered harder to estimate than COD, estimation of NH₄ concentrations yields good results. The dynamics are captured well and even peak concentrations are estimated well for the second half of the test data. Estimation for the first half of the test data is slightly worse with peak concentrations underestimated and an offset in predictions relative to the measured values. However, overall these results are of high value for WWTP operation and advanced control.

TABLE V. NMSE COMPARISON RESULTS FOR TEST DATA: VIRTUAL 2H-COMPOSITE SAMPLES [X100 %]

Dataset	RF _{class}	LDA _{class}	SVM _{class}	MLR _{reg}	RF _{reg}	MLP _{reg}	PLS _{reg}	SVR _{reg}
U _{1,COD}	0.00	0.01	0.02	0.00	0.01	0.02	0.00	0.00
U _{2,COD}	0.24	0.18	1.07	0.23	0.26	0.89	0.23	0.30
U _{3,COD}	0.12	0.14	0.39	0.20	0.11	0.16	0.19	0.14
U _{1,NH₄}	0.51	0.28	0.48	0.13	0.48	0.70	0.13	0.64
U _{2,NH₄}	0.70	0.54	0.64	0.47	0.70	1.20	0.47	0.82
U _{3,NH₄}	0.92	0.52	0.77	0.51	0.72	1.15	0.51	0.88

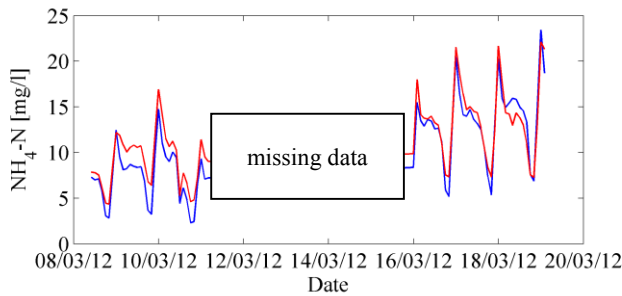


Figure 4. PLS 2h mean U_{1,NH₄} (blue: measured, red: predicted)

In general, it can be seen that for U_{2,COD} and U_{3,COD} the non-linear methods perform best, while for U_{2,NH₄} and U_{3,NH₄} the linear methods achieve the best results. While regression and classification performance is similar for easily predictable datasets like U_{1,COD}, classification methods achieve better results for datasets that are harder to estimate.

IV. CONCLUSION

The results show that it is possible to estimate COD in the WWTP inflow based on standard measurements with sufficient accuracy for use in optimization and control strategies. Furthermore, it is possible to achieve better prediction results through installation of an additional in-line turbidity probe. The prediction of NH₄-N with standard equipment is not feasible due to low estimation accuracy whereas it is possible to achieve sufficiently accurate results, using all input variables. For practical application this means that it is recommended to estimate COD based on standard equipment or after installation of an additional in-line turbidity probe, while NH₄-N estimation is a valid option for operators who already have an in-line COD probe in the WWTP inflow.

In addition, this analysis shows that it is possible to achieve sufficient estimation results using classification instead of regression methods. This not only facilitates the transparent presentation of inflow concentrations to the operator but also allows for specifically adapted training to focus on correct estimation of interesting individual classes.

REFERENCES

[1] Åmand, L., Olsson, G. and Carlsson, B., 2013. Aeration control – a review. *Water Science & Technology*, 67 (11), 2374.
 [2] Olsson, G., 2005. *Instrumentation, control and automation in wastewater systems*. 1st ed. London: IWA Publ; IWA.

[3] American Public Health Association, 1981. Standard methods for the examination of water and wastewater. 15th ed. Washington: American Public Health Association, presented at the IEEE Summer power Meeting, Dallas, TX, June 22–27, 1990, Paper 90 SM 690-0 PWRS.
 [4] M. Henze, International Water Association. Task Group on Mathematical Modelling for Design and Operation of Biological Wastewater Treatment, Activated sludge models ASM1, ASM2, ASM2d and ASM3. London: IWA Pub., 2000.
 [5] M. Graner, T. Hilmer, and M. Bongards, „Einsatz ionenselektiver Messgeräte für die Online-Messung von Stickstoffverbindungen auf Kläranlagen – ein Erfahrungsbericht“, 2005, Bd. 1890, S. 81–92.
 [6] R. S. Brito, H. M. Pinheiro, F. Ferreira, J. S. Matos, and N. D. Lourenço, “In situ UV-Vis spectroscopy to estimate COD and TSS in wastewater drainage systems,” *Urban Water J.*, pp. 1–13, Jun. 2013.
 [7] H. Haimi, M. Mulas, F. Corona, and R. Vahala, “Data-derived soft-sensors for biological wastewater treatment plants: An overview,” *Environ. Model. Softw.*, vol. 47, pp. 88–107, Sep. 2013.
 [8] Q. Yang, “Simultaneous Determination of Chemical Oxygen Demand (COD) and Biological Oxygen Demand (BOD5) in Wastewater by Near-Infrared Spectrometry,” *J. Water Resour. Prot.*, vol. 01, no. 04, pp. 286–289, 2009.
 [9] D. J. Dürrenmatt and W. Gujer, “Data-driven modeling approaches to support wastewater treatment plant operation,” *Environ. Model. Softw.*, vol. 30, pp. 47–56, Dec. 2011.
 [10] F. Corona, M. Mulas, H. Haimi, L. Sundell, M. Heinonen, and R. Vahala, “Monitoring nitrate concentrations in the denitrifying post-filtration unit of a municipal wastewater treatment plant,” *J. Process Control*, vol. 23, no. 2, pp. 158–170, Feb. 2013.
 [11] P. H. Menold, R. K. Pearson, and F. Allgower, „Online outlier detection and removal“, in Proceedings of the 7th Mediterranean Conference on Control and Automation (MED99), 1999, S. 1110–1133.
 [12] S. de Jong, “SIMPLS: An alternative approach to partial least squares regression,” *Chemom. Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, Mar. 1993.
 [13] S. S. Haykin, *Neural networks: A comprehensive foundation*, 2. Aufl. Upper Saddle River, N.J: Prentice Hall, 1999.
 [14] S. McLoone, M. D. Brown, G. Irwin, and A. Lightbody, „A hybrid linear/nonlinear training algorithm for feedforward neural networks“, *IEEE Transactions on Neural Networks*, Bd. 9, Nr. 4, S. 669–684, 1998.
 [15] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.
 [16] S. S. Keerthi und C.-J. Lin, „Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel“, *Neural Computation*, Bd. 15, Nr. 7, S. 1667–1689, 2003.
 [17] R. O. Duda, *Pattern classification*, 2nd ed. New York: Wiley, 2001.
 [18] A. Jaialtilal. randomforest-matlab - Random Forest (Regression, Classification and Clustering) implementation for MATLAB (and Standalone), 2010
 [19] C. Cortes und V. Vapnik, „Support-vector networks“, *Machine learning*, Bd. 20, Nr. 3, S. 273–297, 1995.
 [20] C. C. Chang und C. J. Lin, „LIBSVM: a library for support vector machines“, Taiwan, 2001.
 [21] C. Wolf, D. Gaida, A. Stuhlsatz, T. Ludwig, S. McLoone, and M. Bongards, “Predicting organic acid concentration from UV/vis spectrometry measurements - A comparison of machine learning techniques,” *Trans. Inst. Meas. Control*, vol. 35, no. 1, pp. 5–15, Sep. 2011.