Advancements in machine learning for estimating parameters of wastewater treatment plants

Natalya Kolyeva[1*], *Alexander* Rastyagaev[1], *Lyudmila* Kortenko[1], *Sergey* Rozhkov[1], *Mariia* Sbitneva[1], and *Aleksandr* Kuznetsov[1]

[1]Ural State University of Economics, Ulitsa 8 Marta, 62/45, 620144, Yekaterinburg, Russian Federation

Abstract. The aim of the study is to develop and validate machine learning methods for calculating the parameters of aeration tanks of wastewater treatment plants at the stage of technical and commercial proposal. Research methods included: generalization of known scientific and technical results, theoretical studies were conducted using the theory of fluid motion in the boundary layer, the theory of kinetics of enzymatic reactions of organic pollutants in wastewater, machine learning methods and statistical decision theory. Experimental studies were conducted on a laboratory setup to study the kinetics of wastewater sedimentation. As a result of the study, a model of the XGBoost algorithm was developed, which successfully coped with the task of optimization of calculations, providing high accuracy, and this, in turn, opens up new opportunities for improving the efficiency of design of wastewater treatment plants.

1 Introduction

In recent years, machine learning (ML) methods have become actively applied in various fields, including ecology and engineering design. Machine learning is attractive due to its ability to analyze large amounts of data, identify complex dependencies and adapt the operation of individual pieces of equipment and systems as a whole to new conditions, which makes it a promising tool for optimizing calculations of aeration tank process parameters.

The main document regulating the calculation of technological parameters of wastewater treatment facilities in the Russian Federation is the Code of Rules. The updated version [1-3], provides designers with an opportunity to use alternative methods of calculation of biological treatment facilities, including mathematical models. This opens new horizons for engineers, but at the same time poses them a difficult task – the choice of methodology in order to minimize risks and ensure the required quality of wastewater treatment. This issue becomes especially relevant in the design of aeration tanks.

In 2018-2019, papers were published in [4-6], which made a significant contribution to the development of computational methods. In recent years, the controversy around this

* Corresponding author: nkoleva@mail.ru

topic has only intensified: experts analyze existing approaches, discuss international experience and propose their own solutions. In this context, special attention should be paid to the studies [7], which provide an in-depth analysis of aeration basin calculation methods, including ATV-DVWK-A131E, Danilovich-Epov methodology and ASM2d.

The authors conclude that the most correct methods are those based on enzyme kinetics formulas. However, the choice of a specific approach is left to the technologist, who should take into account the specifics of the project and the requirements for the quality of treated water.

2 Research results

The design of aerobic wastewater treatment plants, particularly aeration tanks, is a complex and multi-stage process that requires consideration of many factors such as wastewater composition, activated sludge load, oxygen concentration, temperature conditions and other parameters. Traditional calculation methods based on empirical formulas and regulations have a number of limitations:

– labor-intensive calculations: classical methods require considerable time to perform manual calculations and verify the results;

– low adaptability: existing approaches do not always take into account changing operating conditions, which can lead to design errors;

– ограниченная точность: эмпирические формулы часто не учитывают нелинейные зависимости между параметрами, что снижает точность расчетов.

In this regard, there is a need to develop a new approach that will automate the process of calculating the technological parameters of aeration tanks, thereby increasing the accuracy and adaptability of calculations, as well as reducing time costs at the stage of preparing a technical and commercial proposal.

The aim of the study is to develop a model for calculating the technological parameters of aeration tanks using machine learning methods, which provides high accuracy and speed of calculations.

Objectives of the study:

– to analyze the processes of the enterprise to identify bottlenecks;

– analyze classical methods of aeration basin calculation and identify their limitations;

– investigate the applicability of machine learning methods to solve the problem of calculating aeration tank parameters;

– develop and train a machine learning algorithm;

– evaluate the accuracy of the model and compare its results with classical methods.

The main requirements for the solution are:

– high accuracy of calculations (error not more than 5%);

– possibility to take into account non-linear dependencies between parameters;

– minimization of time spent on calculations;

– integration of the model into existing design systems;

Expected results:

– a machine learning model capable of calculating aeration tank parameters with high accuracy;

– automation of the calculation process.

Thus, the set research task is aimed at solving the actual problem of designing aerobic treatment facilities by introducing modern machine learning algorithms into existing enterprise processes, which will improve the efficiency and accuracy of calculations.

Let us consider the main parameters affecting the volume of the aeration tank and divide them into input (attributes) and target (labels).

1. Input parameters:

1.1 Wastewater characteristics (these data characterize the properties of the incoming wastewater):

– wastewater volume (Q), m³/day;
– biochemical oxygen demand - BOD5, mg/l;
– suspended solids concentration (SS), mg/l;
– nitrogen (N) concentration, mg/l;
– phosphorus (P) concentration, mg/l;
– wastewater temperature (T) °C.

1.2 Treatment plant characteristics (these data describe the operating conditions:

– operating activated sludge concentration in the aeration basin (a), g/l;
– activated sludge load (Ns), g BOD5/g sludge*day;
– operating dissolved oxygen concentration (OC), g/L.

2. Target data is the value to be predicted:

–aeration tank volume (V), m³.

In order to increase the representativeness of the sample and improve the quality of model training, dairy wastewater will be chosen as the object of study for a number of reasons [8]:

– dairy wastewater is water contaminated with dissolved organics, nitrogen and phosphorus compounds, which determines the need for biological treatment methods for this type of wastewater;

– EnviroChemie has more than twenty years of experience in dairy wastewater treatment, with a project database of more than one hundred such facilities;

– when forming the sample, we assume that the wastewater has passed all the necessary preliminary treatment stages and is ready to be fed to the biological stage.

3 Conclusion

In order to prepare a dataset with characteristics of dairy wastewater with volumes from 400 to 4000 m³/day, we will create a synthetic dataset based on typical characteristics of dairy wastewater and aeration tank parameters. This will allow us to perform modeling and further processing [9].

Based on our earlier assumptions, we include the following parameters in the dataset:

– wastewater volume (Q), m³/d (range from 300 to 4000);

– BOD5 initial concentration (C_BOD5in), mg/l (range from 800 to 2000, depends on the technological processes at the plant);

– concentration of C_BOD5out after treatment, mg/l (fixed value regulated by the requirements of the RF legislation – 3);

– suspended solids concentration (SS), mg/l (range from 30 to 120);
– nitrogen concentration (N), mg/l (range from 60 to 90);
– phosphorus (P) concentration, mg/L (range 15 to 30);
– wastewater temperature (T), °C (range 18 to 22)
– working activated sludge concentration in the aeration basin (a), g/l (fixed value set by the engineer as part of the calculations and depending on the activated sludge separation technology, for classical technology it is taken as 3 g/l);
– activated sludge (Ns) load, g BOD5/g sludge*day. (fixed value, which should not exceed 0.2 for dairies);
– aeration tank volume (V), m³ (target variable to be calculated).

Algorithm for generating the dataset:

− wastewater volume, BOD5 concentration, suspended solids concentration, total nitrogen concentration, total phosphorus concentration, wastewater temperature, aeration basin activated sludge working concentration and activated sludge load are random values generated in the data ranges of real facilities;

−aeration tank volume is calculated by a simplified formula as the ratio of the product of daily flow rate and the difference of BOD5 concentrations at the aeration tank inlet and outlet to the product of activated sludge concentration and activated sludge load:

$$V = Q \cdot (C\_BOD5in - C\_BOD5out)/Ns \cdot a$$

Figure 1 shows a snippet of Python source code for generating a dataset based on the data and algorithm described above.

```
n_samples = 1000

np.random.seed(42)
data = {
    "Q": np.random.uniform(300, 4000, n_samples),
    "C_BOD5in": np.random.uniform(800, 2000, n_samples),
    "C_BOD5out": 3,
    "SS": np.random.uniform(30, 120, n_samples),
    "N": np.random.uniform(60, 90, n_samples),
    "P": np.random.uniform(15, 30, n_samples),
    "T": np.random.uniform(18, 22, n_samples),
    "a": 3,  #
    "Ns": 0.2,
}

df = pd.DataFrame(data)

df["Volume"] = (df["Q"] * (df["C_BOD5in"] - df["C_BOD5out"])) / (df["Ns"] * df["a"] * 1000)

df.head()
```

Fig. 1. Code for generating a synthetic dataset

When generating the dataset, we used a special parameter “seed”, which sets the initial value for the random number generator. The use of this parameter is necessary for reproducibility of results when creating random values, otherwise, the function “random” from the library “NumPy” would give us different values each time it is accessed, which would make it difficult to compare different models, because each of them would be trained on data different from the previous one.

At the next step, a dictionary of features in the specified ranges of values was generated. After that, this dictionary was converted into a special data structure – “DataFrame”, which is analogous to an Excel or SQL table [10]. Each column in our “DataFrame” represents a separate attribute (e.g., runoff volume, BOD5 concentration, etc.), and each row represents a data sample (e.g., data for one object). Thus, a dataset of 1000 rows was generated, the first 5 rows of which are shown in Figure 2.

	Q	C_BOD5in	C_BOD5out	SS	N	P	T	a	Ns	Volume
0	1685.798440	1022.159515	3	53.553512	80.181090	23.579938	19.574542	3	0.2	2863.495866
1	3817.642934	1450.281137	3	52.228092	83.900442	27.081485	19.893743	3	0.2	9208.671009
2	3008.377585	1847.535003	3	111.562912	67.514037	26.402414	21.418190	3	0.2	9248.429596
3	2515.036392	1678.669864	3	52.459158	78.746223	17.308499	19.360018	3	0.2	7023.951146
4	877.268970	1767.873377	3	54.475475	77.152379	17.238742	21.478599	3	0.2	2580.447749

Fig. 2. Fragment of the generated dataset

Let's set up a machine learning model to predict the aeration tank volume based on the generated dataset.

Since the target variable in our problem is a continuous variable that can take any value within a certain range, this property makes the problem ideally suited for regression analysis. Unlike classification problems where the target variable is categorical (e.g., “small”, “medium”, “large” volume), regression allows us to predict an exact numerical value, which is critical for engineering calculations. Also, regression models allow not only to predict the volume value, but also to establish a quantitative relationship between the input parameters and the target variable. This opens up additional opportunities to optimize the calculation process, as engineers can accurately estimate how changes in input parameters will affect the aeration tank volume.

We will use three models to solve the problem of calculating the aeration tank volume:

– linear regression (“Linear regression”), which was chosen as the base model to establish a linear relationship between the input parameters and the target variable. This model is easy to implement and interpret, making it ideal for initial analysis. However, linear regression does not account for nonlinear dependencies, which may limit its accuracy in complex problems.

– Random Forest is an ensemble model consisting of multiple decision trees. It allows taking into account nonlinear dependencies and interactions between features, which is especially important in problems with a large number of parameters. The random forest exhibits high accuracy and robustness to overfitting, making it a powerful tool for regression analysis.

– gradient bousting (“XGBoost”) is another ensemble model that consistently improves predictions by combining weak models. XGBoost demonstrates high accuracy and speed with large amounts of data. This model was chosen to account for complex nonlinear dependencies and interactions between traits.

At the next step, we will divide the data into a feature matrix (X) and a vector of the target variable (y), and divide the dataset we obtained into training and test samples in the ratio of 80% for training and 20% for test using the function “train_test_split” of the “scikit-learn” library. This is necessary to prevent overtraining of models, the test sample is an “independent” dataset, which is not used during training, which helps to make sure that the model does not just remember the training data, but has learned to generalize them.

Next, we train all three of our models on the training sample and obtain the values predicted by the model on the test data. Figure 16 shows a code fragment reflecting the processes of training the XGBoost model, its operation, and saving it to a file for further use.

```
xgb_model = XGBRegressor(n_estimators=100, max_depth=10, learning_rate=0.1, random_state=42)
xgb_model.fit(X_train, y_train)
joblib.dump(xgb_model, "/content/drive/MyDrive/Colab Notebooks/Aeration tank model/xgboost_aerotank_model.pkl")

#
y_pred_xgb = xgb_model.predict(X_test)
```

Fig. 3. Model setup and training

It should be noted that the model was trained on a synthetically generated data set, which, despite being close to real data, still requires correction, since synthetic data are limited in their representativeness and do not fully reflect the full range of possible aeration tank operating conditions. Consequently, in order to increase the reliability and adaptability of the information system to different conditions, it should be trained on real data collected from operating treatment plants. In addition, the structure of the information system, despite its high accuracy, can be complicated to take into account various technological parameters

of aeration tank operation and their non-linear dependencies, which will make it more universal and able to solve a wider range of engineering problems.

Thus, despite the limitations associated with the use of synthetic data and the simplified structure of the information system, it is already capable of performing accurate preliminary calculations. Once finalized, the algorithm can become the basis for automation of wastewater treatment plant design, and machine learning methods can become a new standard in the field of technological calculations of wastewater treatment plant parameters. This will not only ensure the speed and accuracy of calculations, but also allow engineers to focus on solving more complex problems

References

1. SP 32.13330.2018. Kanalizacija. Naruzhnye seti i sooruzhenija. SNiP 2.04.03-85. Utv. prikazom Ministerstva stroitel'stva i ZhKH RF (Minstroj Rossii) ot 25 dekabrja 2018 g. № 680/pr, vveden v dejstvie s 26 ijunja 2019 g. (M.: Standartinform, 2019.)
2. SP 529.1325800.2023 Opredelenie osnovnyh raschetnyh gidrologicheskih harakteristik. Utv. prikazom Ministerstva stroitel'stva i ZhKH RF № 654/ pr ot 11 sentjabrja 2023. (M. FAU «FCS». 2023.)
3. SP 131.13330.2020. Svod pravil. SNIP 23-01-99*, vveden 24.05.2021 / Stroitel'naja klimatologija. (M.: Minstroj Rossii, 2020)
4. T.S. Semenova, O.I. Sergienko, Best Available Technologies of Water Supply and Drainage **S2**. 118-122 (2024)
5. Y.A. Menshutin, Water Supply and Sanitary Engineering **8**. 36-40 (2024). DOI 10.35776/VST.2024.08.04
6. B.B. Orazbayev, L.T. Kurmangaziyeva, G.K. Shambilova, A.A. Muratbekova, Bulletin of the Karaganda University. Chemistry Series. **3(95).** 102-106 (2019). DOI 10.31489/2019Ch3/102-106.
7. V.N. Shvetsov, K.M. Morozova, S.V. Stepanov, Water Supply and Sanitary Engineering. **9.** 26-39 (2018)
8. N.A. Chernikov, N.V. Tvardovskaya, I.M. Okhremenko BRICS Transport. **2**, 2 (2023). DOI 10.46684/2023.2.2. EDN AGSEGJ.
9. I.V. Pavlova, I.N. Postnikova, I.V. Isakov, D.A. Presnyakova, Izvestiya Vuzov. Applied chemistry and biotechnology **1(12).** 90-96 (2015)
10. Y.A. Ermolin, Water Supply and Sanitary Engineering **12**. 40-44 (2023). DOI 10.35776/VST.2023.12.06.