Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2023.29715

301 (2023) 106–114
July

Sequential algorithm of building the regression-classification model for total nitrogen simulation: application of machine learning

Krzysztof Barbusiński[a], Bartosz Szeląg[b,*], Anita Białek[c], Ewa Łazuka[d], Emilia Popławska[d], Joanna Szulżyk-Cieplak[d], Roman Babko[e], Grzegorz Łagód[d]

[a]*Silesian University of Technology, Konarskiego 18, 44-100 Gliwice, Poland, Tel.: +48 32237-11-94; email: krzysztof.barbusinski@polsl.pl (K. Barbusiński)*
[b]*Warsaw University of Life Sciences, Nowoursynowska 166, 02-787 Warsaw, Poland, Tel.: +48 22 59 310 00; email: bartoszszelag@op.pl (B. Szeląg)*
[c]*Kielce University of Technology, Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland, Tel.: +48 41342-47-35; email: anita_bialek@interia.eu (A. Białek)*
[d]*Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland, Tel.: +48 81538-43-22; emails: e.lazuka@pollub.pl (E. Łazuka), e.poplawska@pollub.pl (E. Popławska), j.szulzyk-cieplak@pollub.pl (J. Szulżyk-Cieplak), g.lagod@pollub.pl (G. Łagód)*
[e]*Schmalhausen Institute of Zoology NAS of Ukraine, 01030 Kyiv, Ukraine, Tel.: +38 044235-10-70; email: rbabko@ukr.net (R. Babko)*

Received 11 November 2022; Accepted 19 June 2023

ABSTRACT

Total nitrogen (TN) concentration is one of important indications of wastewater quality and also a parameter important for wastewater treatment plant performance evaluation. Since the variability of total nitrogen in the effluent from the wastewater treatment plant is the result of the processes taking place in the bioreactor, the processes can be described by mechanistic models, for example, activated sludge models. However, calibration of many parameters is required in such models, and can leads to problems in identifying their proper numerical values. The paper proposes a novel way to deal with this problem by presenting a methodology for building a model for simulating TN, based on sequential structure. In the applied approach, regression models for simulation of TN are first created using Extreme Gradient Boosting (XGBoost), and random forest (RF) methods. In the case of unsatisfactory predictive ability, a division of the dependent variable into a classifier form is made. In the next stage, classification models are created by RF and XGBoost methods and sensitivity analysis is performed by calculating Shapley indices. Two classification models were built that allow for the identification of TN_{eff} variability ranges. The new approach using two models instead of one is preferable because it allows control and optimization of the bioreactor operation.

Keywords: Total nitrogen simulation; Wastewater parameters; Operating and control of WWTPs; Machine learning; Extreme Gradient Boosting (XGBoost); Random forest (RF); Regression and classification models

* Corresponding author.

Presented at the 15th Scientific Conference on Micropollutants in the Human Environment, 14–16 September 2022, Częstochowa, Poland