Development of a Virtual Sensor for COD Measurement in a Wastewater Treatment Plant

Antonio Díaz-Longueira, Míriam Timiraos, Álvaro Michelena, Óscar Fontenla-Romero, and Jose Luis Calvo-Rolle

Ciencia e Técnica Cibernética (CTC), Department of Industrial Engineering, Universidade da Coruña, 15071 A Coruña, Spain
Fundación Instituto Tecnológico de Galicia, Department of Water Technologies, National Technological Center, Cantón Grande 9, Planta 3, C.P. 15003, A Coruña, Spain.
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: a.diazl@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.46*

Abstract: The objective of the work is to develop a system that allows predicting, from a global perspective, the behavior of the process in a wastewater treatment plant. To do this, the chemical oxygen demand, a variable present in water, is estimated indirectly, avoiding difficult and complex measurements. This estimation is carried out in real time through the relationship between easily measured variables. This modeling will be done through the use of machine learning techniques. Different regression techniques are applied and compared. The dataset contains variables such as pH, conductivity, suspended solids and etc. In this way, a non-physical indirect sensor is implemented. Thresholds are established for the detection of deviations in the sensor parameters.

1 Introduction

Water is a scarce and irreplaceable good, with great importance in the field of health and production in our country and the world. The constant advance of climate change and the unstoppable growth of the world's population are affecting its availability, making it an increasingly scarce resource. In view of this situation, and as a possible measure in this situation, the possibility of reusing wastewater (Salgot and Folch, 2018) appears. This reuse is subject to eliminating harmful agents that may be present. With the aim of reducing wastewater pollution to values valid for reuse, wastewater treatment plants (WWTP) appear, capable of reducing the polluting load (EDAR, n.d.).

To ensure and make sure of the correct operation of these facilities, it is essential to know the state of the water, by means of certain markers that make it possible to establish the type and degree of contamination of the water, both in the inflow and outflow of the WWTP. These markers include physicochemical variables that are costly and/or technically complicated to measure. Instead of using physical sensors to measure these markers, it is possible to try to estimate, based on other variables with a simpler measurement, their value. An indicator of the degree of water contamination is the chemical oxygen demand, which provides an idea of the presence of both organic and inorganic agents (Clesceri et al., 1999).

This estimation is carried out in real-time based on the existing relationship with these third variables. In this way, a non-physical indirect sensor is implemented, which makes use of machine learning techniques to model this relationship with the rest of the variables and predict their value. For this reason, regression techniques will be used. With the indirect sensor already implemented, it is possible to establish thresholds for the detection of deviations in the parameters and create an alarm system. These thresholds may be established by different methods.

2 Materials and methods

This section describes the machine learning techniques and algorithms, the metrics and procedures used for the evaluation of the models, the graphs where the results will be plotted and the data set used in the analysis.

2.1 Machine learning techniques and algorithms

To try to carry out a study that is as heterogeneous and varied as possible, different supervised learning techniques and algorithms were evaluated to analyze and compare their performance in the sought estimates. The following techniques were used:

- Recursive Least Squares (RLS).
- K-Nearest Neighbors (KNN).
- Decision Tree (DT).
- Support Vector Regression (SVR).
- MultiLayer Perceptron (MLP).

2.2 Model evaluation

For the evaluation and subsequent selection of the models, different metrics were used to determine their performance. The metrics used were:

- Mean Absolute Error (MeanAE).
- Mean Squared Error (MSE).
- Root Mean Squared Error (RMSE).
- Symmetric Mean Absolute Percentage Error (SMAPE).
- Coefficient of determination (R^2).

2.3 Dataset

The data set used is a real set, taken from measurements from 3 wastewater treatment plants. These measurements were carried out over 3 months: June, July and August, with one measurement per day. In addition to the physicochemical variables related to water, there is also a variable that indicates the daily volume of water processed by the WWTP, two that establish the day and month of the measurement and a last one indicating the WWTP from which the data were obtained. These last four variables were not used in the development of the work.

The data set used for the development of the work is composed of 8 physical-chemical variables of water. These variables are pH, conductivity (Cond), biochemical oxygen demand (DBO), chemical oxygen demand (DQO), nitrates (N), phosphates (F), suspended solids and settleable solids (V60). Figure 1 shows the correlation matrix between the different variables. The case of DQO stands out.

Figure 1: Correlation matrix

The measurements do not follow a specific periodicity; it is possible that no measurements were taken on a specific day, or that some of the physical-chemical variables were not measured. For the development of the work, since it is not known which variables will be used for a prediction, only those records in which all the variables are present will be used.

This implies that, of the 276 expected records, the data set will be smaller.

3 Experiments

This section details the experiments performed to design the best indirect sensor. For this purpose, and looking for the best model to implement, the performance of different regression techniques will be evaluated. They will be configured through their hyperparameters until the best prediction is reached.

The three variables with the highest correlation with COD were used. According to Figure 1, they are DBO, nitrates and phosphates.

The comparison between experiments will be performed based on the metrics obtained in 3-kfold cross-validation, selected due to the small size of the dataset. To know the performance of a model and to be able to compare it with that of the others, first of all, the metrics observed are the coefficient of determination and the SMAPE. Since the SMAPE does not handle over- and under-forecasting in the same way, as soon as these metrics do not allow us to select a clear winner, we will proceed to focus on the other metrics.

The following are the hyperparameters modified in each of the techniques, and the values tested.

3.1 Recursive Least Squares

The models generated were tested by forcing the coefficients to be positive (*positive* with possible values of True or False) and to calculate or not the independent term (*intercept* with possible values of True or False).

The configurations of this technique will receive their name following the following construction: RLS + intercept value + positive value. In this way, the configuration with *positive* set to True and *intercept* set to False will be named RLSFalseTrue.

3.2 K-Nearest Neighbors

It was obtained, experimentally, that the best results are obtained when using an odd number of neighbors between 3 and 9. Two functions were evaluated for assigning weights to each neighbor: *uniform* y *distance*.

The configurations of this technique will receive their name following the following construction: KNN + neighbor + weight function. In this way, the configuration with 5 neighbors and *uniform* weight distribution will be named KNN5uniform.

3.3 Decision Tree

Two methods for determining the best split at each node (criterion) were tested: *absolute_error* y *squared_error*. The maximum depth of the diagram was also modified, from 1 to 7. This range was obtained experimentally.

The configurations of this technique will receive their name following the following construction: DT + depth + criterion. In this way, the configuration with a maximum depth of 3 and squared_error as criterion will be named DT3squared_error.

3.4 Support Vector Regression

Experimentally, it was proven that the best results appeared for values of the regularization coefficient of 10 and 0.1. Also experimentally, it was found that the best *epsilon* values were 1. Finally, three kernels were tested: *linear*, *sigmoidal* and *tansig*.

The configurations of this technique will receive their name following the following construction: SVR + regularization coefficient + kernel. In this way, the configuration with a regularization coefficient of 10 and a linear kernel will be named SVR10linear.

3.5 MultiLayer Perceptron

Experimentally, it was proven that the best results were obtained when working with an intermediate layer of 8 to 9 neurons, so these were the tested values. Also, three activation functions were analyzed for the input and intermediate layers: *linear*, *sigmoidal* and *tangent-sigmoidal*.

The configurations of this technique will receive their name following the following construction: MLP + hidden neurons + activation function. In this way, the configuration with 10 hidden neurons and a linear function activation will be named MLP10linear.

4 Results

This section collects the results of the different experiments. A table is included with each ML method, with the configurations tested. The best result of each technique is highlighted in bold.

Table 1 shows the mean value of metrics obtained by the configurations in the Recursive Least Squares method.

Table 2 shows the mean value of metrics obtained by the configurations in the K-Nearest Neighbors.

Table 1: Mean value of metrics by the RLS

Configuration	MeanAE	SMAPE	MSE	RMSE	MaxError	R2
RLSTrueTrue	**69,554**	**5,584**	**8020,624**	**89,009**	**245,365**	**0,781**
RLSTrueFalse	**69,554**	**5,584**	**8020,624**	**89,009**	**245,365**	**0,781**
RLSFalseTrue	74,006	6,096	8888,193	93,335	257,115	0,761
RLSFalseFalse	74,006	6,096	8888,193	93,335	257,115	0,761

Table 2: Mean value of metrics by the KNN

Configuration	MeanAE	SMAPE	MSE	RMSE	MaxError	R2
KNN3distance	63,754	5,013	9627,722	95,873	385,558	0,742
KNN3uniform	64,074	5,039	9605,909	96,251	387,222	0,742
KNN5distance	63,515	4,929	9114,417	93,655	379,033	0,756
KNN5uniform	64,063	4,986	8885,404	92,860	376,933	0,762
KNN7distance	63,089	4,918	8857,006	92,510	367,355	0,763
KNN7uniform	**63,406**	**4,951**	**8663,092**	**91,774**	**367,048**	**0,768**
KNN9distance	62,373	4,846	8872,268	92,466	369,004	0,762
KNN9uniform	63,954	4,965	8809,346	92,514	368,889	0,764

Table 3 shows the mean value of metrics obtained by the configurations in the Decision Tree method.

Table 3: Mean value of metrics by the DT

Configuration	MeanAE	SMAPE	MSE	RMSE	MaxError	R2
DTabsolute_error2	85,552	16,183	14478,422	120,312	444,000	0,586
DTabsolute_error5	84,142	17,261	14283,554	119,385	414,833	0,584
DTabsolute_error6	88,583	17,201	15310,806	123,573	399,000	0,553
DTabsolute_error7	86,358	17,266	15681,057	124,625	420,833	0,537
DTsquared_error2	84,031	15,663	13666,092	116,779	433,070	0,614
DTsquared_error5	**72,930**	**17,033**	**10608,092**	**102,972**	**401,584**	**0,697**
DTsquared_error6	72,035	16,922	11176,430	105,543	416,976	0,679
DTsquared_error7	77,131	17,102	12385,229	111,223	423,194	0,648

Table 4 shows the mean value of metrics obtained by the configurations in the Support Vector Regression method.

Table 5 shows the mean value of metrics obtained by the configurations in the MultiLayer Perceptron method.

Table 4: Mean value of metrics by the SVR

Configuration	MeanAE	SMAPE	MSE	RMSE	MaxError	R2
SVR10linear	67,894	16,182	7891,394	88,251	265,448	0,766
SVR10rbf	124,656	13,063	25681,210	159,469	441,298	0,267
SVR10sigmoid	162,113	12,829	44650,823	209,667	498,596	-0,264
SVR0.1linear	**67,588**	**16,127**	**7796,304**	**87,756**	**260,655**	**0,768**
SVR0.1rbf	154,143	12,707	40419,934	199,541	483,021	-0,145
SVR0.1sigmoid	154,680	12,711	40700,431	200,229	483,797	-0,153

Table 5: Mean value of metrics by the MLP

Configuration	MeanAE	SMAPE	MSE	RMSE	MaxError	R2
MLPlinear8	86,098	6,979	12166,413	108,842	281,640	0,628
MLPlinear9	**74,370**	**6,138**	**9073,254**	**94,628**	**282,360**	**0,734**
MLPlinear10	78,792	6,586	9683,534	97,898	276,400	0,721
MLPtansig8	559,587	75,912	349629,048	590,234	1017,871	-9,447
MLPtansig9	550,588	73,704	339717,577	581,715	1008,872	-9,148
MLPtansig10	537,830	70,625	325936,970	569,668	996,115	-8,734
MLPsigmoid8	604,963	88,111	402400,503	633,426	1063,248	-11,039
MLPsigmoid9	584,359	82,296	377848,954	613,769	1042,644	-10,300
MLPsigmoid10	597,566	86,008	393466,340	626,357	1055,850	-10,769

5 Conclusions and future works

The objective of the work was to develop an indirect, non-physical sensor that would allow estimating the COD value through third variables, thus facilitating the measurement of this pollution marker. The sensor developed presents acceptable R^2 values. This is due to the strong correlation that is present between the variables used with respect to chemical oxygen demand. The highest value obtained is 0.781, achieved by using the RLS technique when the independent term is calculated, since forcing the coefficients does not affect performance. The value of the error metrics is also good, since it makes a relative error of $5,584\%$.The rest of the sensors present similar results, but because they use more complex techniques, it was decided to use the one mentioned above. In Figure 2 we check the performance of the indirect sensor.

The sensor obtained is relatively reliable, with a prediction that is closer to the ideal and does not make large errors. Once the sensor has been designed, but not yet implemented, the introduction of thresholds for detecting parameter deviations and generating early warnings can be established based on statistical methods, such as standard deviation or percentage margins; through expert knowledge indicating security thresholds and other methods, such as machine learning techniques for anomaly detection.

With the aim of improving indirect sensors, and increasing the accuracy of the prediction, these sensors could be developed specifically for a certain WWTP, instead of trying to generalize. For this, it would be necessary to increase the size of the dataset with new measurements. The method could be modified when selecting the variables to be used, opting for other requirements than the correlation between them. It would also be interesting to study the possibility

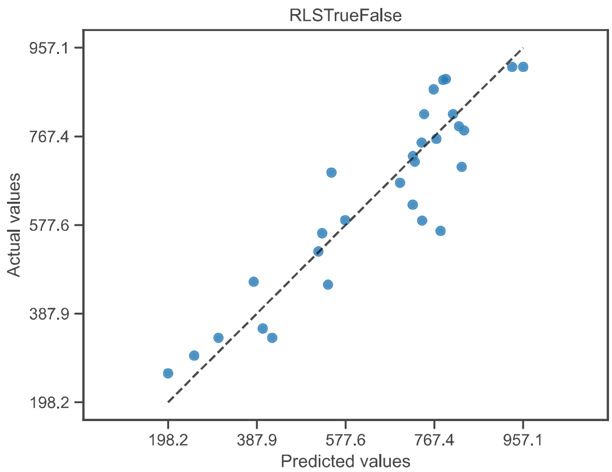

Figure 2: Predicted vs. actual values

that the physical-chemical variables could be part of a time series, for which it would be appropriate to apply other regression techniques such as Long-Short Term Memory (LSTM).

Acknowledgement

Míriam Timiraos's research was supported by the "Xunta de Galicia" (Regional Government of Galicia) through grants to industrial PhD (*http://gain.xunta.gal/*), under the "Doutoramento Industrial 2022" grant with reference: 04_IN606D_2022_ 2692965.

Álvaro Michelena's research was supported by the Spanish Ministry of Universities (*https://www.universidades.gob.es/*), under the "Formación de Profesorado Universitario" grant with reference: FPU21/00932.

CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

Bibliography

L. Clesceri, A. Greenberg, A. Eaton, and M. Franson. Standard methods for the examination of water and wastewater, ed. american public health association-american water works association-water environment federation, usa, 19 p., 1999.

EDAR. Estación depuradora de aguas residuales (edar), n.d. URL *https://www.miteco.gob.es/es/agua/temas/saneamiento-depuracion/sistemas/edar/*.

M. Salgot and M. Folch. Wastewater treatment and water reuse. *Current Opinion in Environmental Science & Health*, 2:64–74, 2018.