Water Science & Technology

Water Science & Technology Vol 85 No 9, 2565 doi: 10.2166/wst.2022.138

Evaluation of genetic models for COD and TSS estimation in wastewater through its spectrophotometric response

Daniel Carreres-Prieto [a,*], Javier Ybarra-Moreno[b], Juan T. García [a] and Fernando Cerdán-Cartagena[c]

[a] Department of Mining and Civil Engineering, Universidad Politécnica de Cartagena, Cartagena 30202, Spain
[b] Edam Ltda, Santiago de Chile, Chile
[c] Department of Information and Communications Technologies, Universidad Politécnica de Cartagena, Cartagena 30202, Spain
*Corresponding author. E-mail: daniel.carreres@upct.es

DC, 0000-0002-1852-1053; JTG, 0000-0002-7204-688X

ABSTRACT

In an urban wastewater treatment plant (WWTP), early knowledge of the pollutant load levels throughout the plant is key to optimize its processes and achieve better purification levels. Molecular spectrophotometry has begun to gain prominence in this wastewater characterization process, as it is a simple, fast, inexpensive and non-invasive technique. In this research work, different mathematical models based on genetic algorithms have been developed for the estimation of chemical oxygen demand (COD) and total suspended solids (TSS) from the spectral response of the samples, measured in the 380–700 nm range by means of a light-emitting diode (LED) spectrophotometer developed by the researchers. A field campaign was carried out in Mapocho-Trebal WWTP (Chile), where 550 samples were obtained in three different parts of the plant: at the inlet (raw wastewater), at the outlet (secondary treated wastewater) and at the outlet of the primary clarifier. A total of 18 estimation models have been calculated by mean of HeuristicLab software, which have presented a high accuracy, with a Pearson's coefficient between 80 and 90% in most cases. In order to achieve the most accurate models possible to characterize each part of the plant, specific models have also been developed, as well as combined models that are valid for all types of wastewater.

Key words: COD, genetic algorithm, LED spectrophotometer, TSS, wastewater pollutant characterization

HIGHLIGHTS

- Use of genetic algorithms and spectrophotometer for the generation of highly accurate COD and TSS estimation models from more than 550 samples from the visible spectrum (380–700 nm).
- Effect of visible spectrum wavelengths on the characterization of COD and TSS.
- The first approach towards the implementation of real-time continuous pollution monitoring by means of cost-effective LED-VIS spectrophotometer.

INTRODUCTION

The emergence of new and efficient automatic sensors at an affordable cost is enabling a large amount of online, continuous and real-time data to be provided in order to control the evolution of the pollutants load in influent and effluent of Wastewater Treatment Plants (WWTP). Deriving conclusions from this monitoring can lead to improved control strategies for such plants. Beraud *et al.* (2009) proposed to use genetic algorithms to deal with abundant new monitoring data in the search for optimized control rules. Do *et al.* (2021) also used genetic based control algorithm to optimize the oxygen dosage in the biological treatment to reduce the operating costs. Monitoring data requires the use of longer data series, to ensure their representativeness, taking into account the variability in load and flows as well as the non-linearity associated with the treatment processes; this justifies the use of evolutionary research techniques. Rauch & Harremoës (1999) proposed the application of nonlinear model predictive control by means of genetic algorithms based on wastewater quality parameters.

The UV spectrophotometric procedure has been considered as a simple, fast and reliable technique for the survey of wastewater global parameters such as chemical oxygen demand (COD), total organic carbon (TOC) or total suspended solids (TSS) (El Khorassani *et al.* 1999). UV/VIS spectrometry in the range of 200–400 nm in effluent secondary treated wastewater was studied by Carré *et al.* (2017), who showed that partial least square regression models (PLSR) from WWTP sampling were sufficiently robust and sensitive to be used in a reclaimed wastewater production process for routine COD and TSS monitoring. In the case of the COD, the most relevant wavelength was 374 nm, due to the significant influence of supra-colloidal and settleable matter. In this way, the correlation among pollutants and spectroscopy surrogates can be achieved through a combination between the molecular absorption spectroscopy in the ultraviolet spectral region (Thomas & Burgess 2007) and light scattering in the visible and short wave near infrared (VIS/SW-NIR) (Bogomolov *et al.* 2012; Bogomolov & Melenteva 2013) produced and balanced depending on the particle size and distribution in the matrix of wastewater. Jeong *et al.* (2007) used UV absorbance values of 300 nm as the most suitable wavelength, providing high enough sensitivity to adjust the measured values using a neural network, resulting in a similar level of accuracy to that achieved by other authors, like Rieger *et al.* (2004).

The variability of the sample compounds in both the solutes and the particles, and the photometric interferences among analytes that influence the absorbance obtained from the samples (Sooväli *et al.* 2006), justifies the need to use mathematical models to achieve accurate correlations among the absorbance and transmittance of the samples and the pollutants' concentration measurements. Most are based on linear regression, partial least square, evolutionary techniques such as support vector machines (Qin *et al.* 2012; Torres *et al.* 2013; Lepot *et al.* 2016; Pacheco *et al.* 2020) and genetic algorithms (Carreres-Prieto *et al.* 2020), achieving squared correlation coefficients of the calibrated function over 0.8 in most cases (Lepot *et al.* 2016).

In the market you can find equipment that makes use of spectrophotometry to characterize the pollutant load of water automatically, such as that developed by the company s::can GmbH, which is capable of characterizing more than 17 parameters (COD and TSS among them) from the spectral response measured between 190 and 750 nm as well as the use of additional probes. Also noteworthy is the ISA system of GO Systemelektronik GmbH, which makes use of variable wavelength spectrophotometry to carry out the characterization of multiple pollutant parameters, such as nutrients or suspended solids, among others, by means of UV-VIS xenon lamp, performing the analysis between 200 and 710 nm.

The use of light-emitting diodes (LED) as the emitting source in spectrophotometric measurements has been shown to be accurate and precise (Mohammad *et al.* 2015; Carreres-Prieto *et al.* 2019; Prairie *et al.* 2020). This enabled the possibility of economically-priced LED spectrophotometers, which are small in size, requiring very low power supply that can be extended along the sewerage network, and can be installed in the sewerage system, highlighting the equipment developed by Han *et al.* (2021), which uses LED technology to carry out its analysis between 235 and 275 nm to characterize the concentration of nitrates in natural water and wastewater effluents.

The manuscript presents several models for estimating the pollutant load, COD and TSS from spectrophotometric information measured experimentally through a device designed and constructed by the authors. Wastewater from three different parts of the Mapocho-Trebal WWTP will be used jointly and separately in combined and specific models. The models were generated by means of HeuristicLab software (Wagner *et al.* 2014) using the following genetic algorithm techniques: Classical Genetic Algorithm (CGA) (Rajasekaran & Pai 2003), The Age-Layered Population Structure (ALPS), (Hornby 2006) and Offspring Selection (OS) (Affenzeller & Wagner 2005). Tests comparing the measured and the predicted data will also be conducted to determine the capacity of the cost-effective device and the technique to provide the levels of COD and TSS in the wastewater at different points of the plant in a simple and fast way, from its spectrophotometric analysis. An early knowledge of the wastewater quality at different points of the WWTP gives valuable information to optimize each stage of the treatment processes.

MATERIALS AND METHODS

Experimental campaign

In order to study the suitability of molecular spectroscopy to achieve a system that allows a rapid characterization of wastewater at different stages of a WWTP, an experimental campaign has been carried out in the Mapocho-Trebal WWTP (Chile) between March 21, 2021 and August 3, 2021. To carry out the study, none of the samples have been filtered, in order to obtain estimation models that can be valid in real-time analysis. Table 1 shows the main characteristics of the WWTP where the

Table 1 | Characteristics of the Mapocho-Trebal WWTP

Parameters		Value
Population	Served(inhabitants)	2,807,000
	Equivalent inhabitants	3,674,880
Capacity (m^3/y)	Current (m^3/day)	760,320
COD	Influent (mg/L)	696
	Primary outlet (mg/L)	427
	Effluent (mg/L)	64
	Performance (%)	91
TSS	Influent (mg/L)	209
	Primary outlet (mg/L)	147
	Effluent (mg/L)	23
	Performance (%)	89

study was carried out. The plant generates a primary sludge by simple decantation of the wastewater, and secondary during the activated sludge biological treatment process. The anaerobic digesters have thermal hydrolysis and co-digestion to maximize biogas production and minimize sludge volumes.

The COD and TSS concentration are determined in the laboratory according to ISO 6060:1989 dichromate method with UV-VIS spectroscopy and SM 2540 F settleable solids, respectively. Moreover, spectrophotometric analysis in the 380–700 nm range is also studied in the samples. The points where the study was carried out were the:

- Inlet of the plant (influent raw wastewater).
- Primary clarifier outlet.
- Outlet of the plant (effluent secondary treated wastewater).

In view of the parts of the plant where samples were collected, a considerable variability in terms of COD and TSS concentrations was expected. This would lead to the achievement of more general models. The flow chart in Figure 1 shows how this sampling campaign was carried out until the mathematical estimation models shown in this manuscript were reached.

Spectrophotometric device

The study focused on the characterization of wastewater samples from their spectrophotometric response, as it is a quick and simple analysis that does not involve sample alteration. Previous work has developed equipment (Figure 2(a)) which, by means of LED technology, is able to perform the spectral analysis of the samples, within 380–700 nm, with an accuracy similar to that of a commercial equipment based on lamps. The samples are introduced into the cost-effective equipment developed by the authors from the top, using a standard 2.5 mL spectrophotometric test tube. A total of 81 wavelengths are investigated in terms of absorbance and transmittance using 33 limited-bandwidth LEDs. Other detailed characteristics of this equipment can be found in Carreres-Prieto *et al.* (2019). In Figure 2(b), a picture of the equipment being used in the laboratory is shown.

Genetic algorithms

To recognize more complex patterns than those found through linear regressions, genetic algorithms were used to calculate the estimation models, more specifically symbolic regression expressions (Koza 1992). The evolutionary nature of these algorithms makes it possible to achieve expressions that are more accurate. The models are synthesized into simple mathematical expressions that can be implemented for the low-cost equipment used, making use of a branch of genetic algorithms called symbolic regression. For the implementation of these models, the software HeuristicLab (Wagner *et al.* 2014) has been used.

Genetic algorithms base their operation on supervised learning (Niculescu-Mizil & Caruana 2005; Nasteski 2017), where the model is trained on data, i.e., it learns to extract patterns from the input data, and then validates the model with data that it has never seen before (test data). For that reason, the data have been separated into two sets: training and testing, with a 66–34% ratio, following the recommendation of Osowski (1996), who recommend a 70–30% ratio. However, the percentage of data for training has been slightly reduced in order to have more data for model validation.

In this manuscript, three types of estimation methods have been implemented: CGA, ALPS (Hornby 2006), and OS (Affenzeller & Wagner 2005).

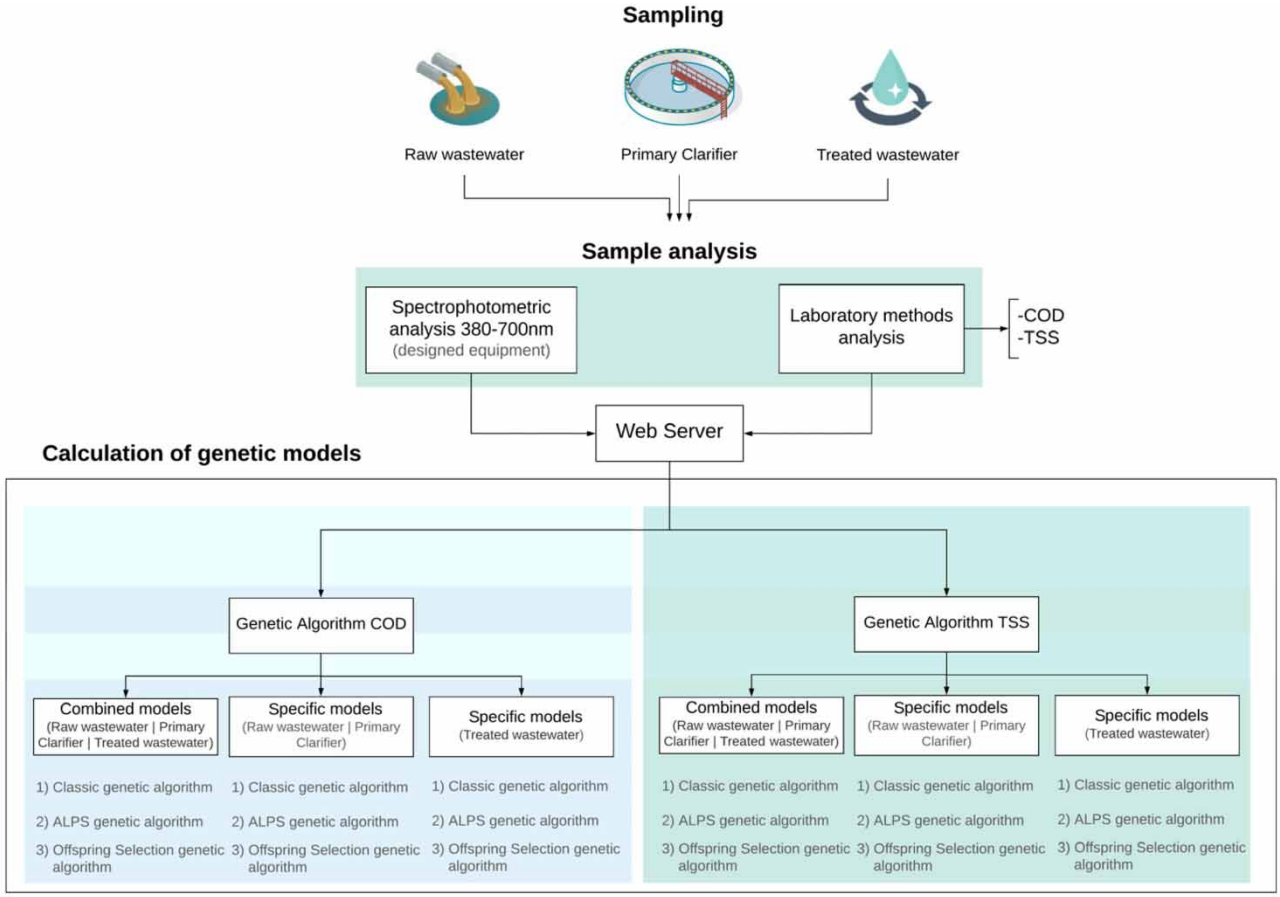


Figure 1 | Diagram of the campaign to obtain the mathematical models for COD and TSS estimation.

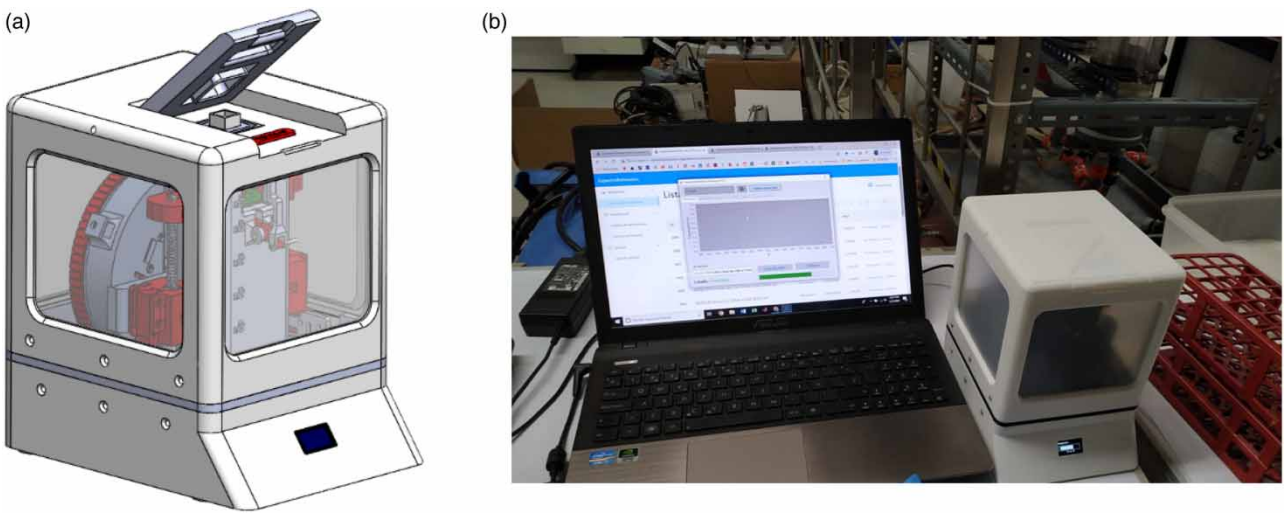


Figure 2 | (a) View of the equipment developed to be able to carry out the spectrophotometric analysis of wastewater samples. (b) View of the equipment being used in the laboratory.

RESULTS AND DISCUSSION

The models generated to estimate the pollutant load from the spectral response of the samples depend on the properties and type of wastewater under study. As the samples come from three different parts of the plant: *(i)* from the inlet (influent raw

water); *(ii)* from the outlet of the primary clarifier; and *(iii)* from the outlet of the secondary clarifier (secondary treated water), the following categorization of models is shown below for each of the parameters under study (COD and TSS):

- Combined model (C): valid models to estimate the pollutant load of raw wastewater, secondary treated water and primary clarifier outlet.
- Specific model for raw wastewater and primary clarifier effluent (R).
- Specific model for secondary treated wastewater (T).

All models presented were generated after removing outliers from each dataset using Box and Whisker plots. The original dataset is composed of 550 samples obtained in the three different parts of the WWTP. In the other side, a total of 162 variables, corresponding to absorbance and transmittance values measured at 81 different wavelengths, were used as the other inputs for the calculation of the models.

In order to show the properties of each type of wastewater analyzed, Figure 3 shows the spectral response recorded by the equipment for each type of wastewater analyzed and supported by the different models. The lower grey line shows a raw wastewater sample with a COD of 668 mg/L and TSS of 296 mg/L; the intermediate red plot shows the output of the primary clarifier (COD of 415 mg/L and TSS of 136 mg/L), and the blue chart shows effluent (secondary treated) wastewater (COD of 63 mg/L and TSS of 28 mg/L). As can be seen, the lower the pollutant load (COD and/or TSS), the higher the spectral response (transmittance).

The following sections present the genetic models that adjust the pollutants COD and TSS from spectrophotometry surrogates, i.e. the transmittance and absorbance. The models are defined combining the data from the three different parts of the WWTP under study or in specific models where raw and primary clarifier wastewater are combined and the secondary treated wastewater is adjusted separately.

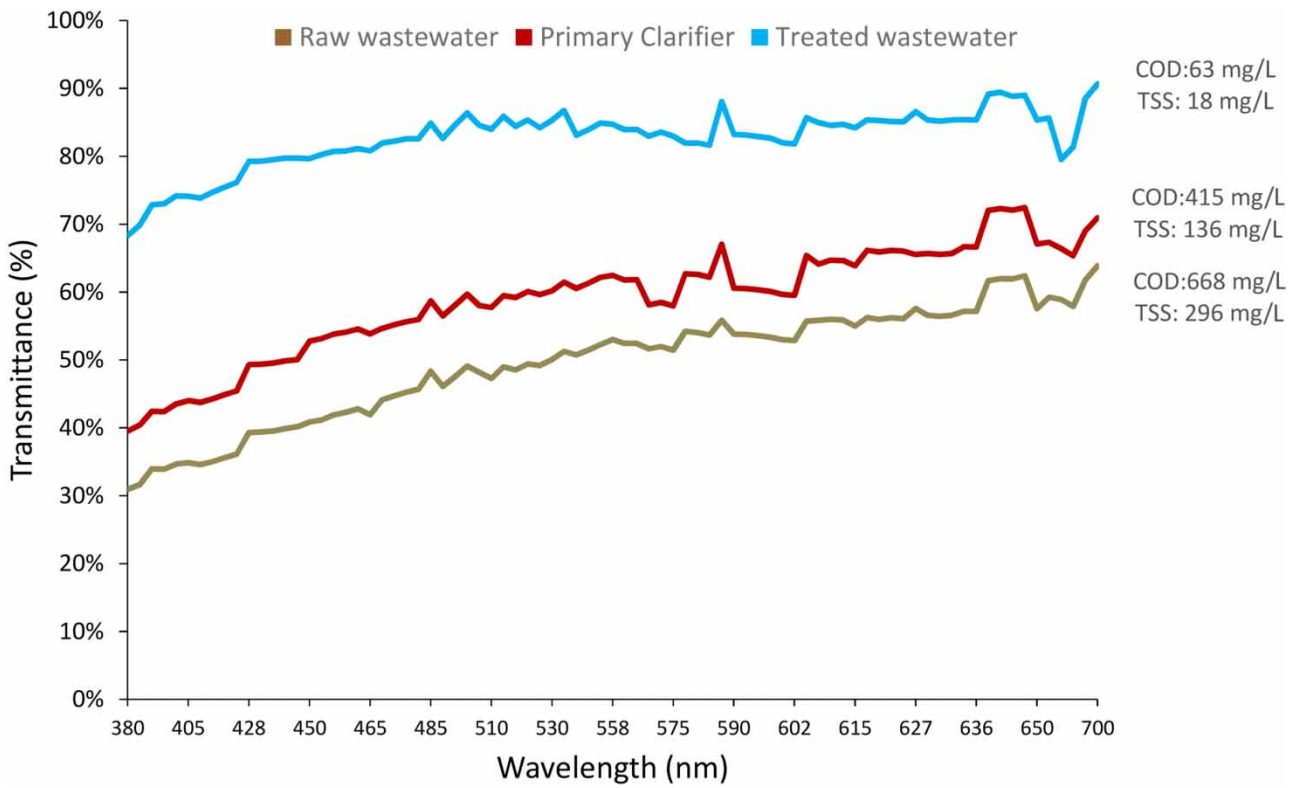


Figure 3 | Comparative graph of the spectral response and contaminant load of water samples taken at the Mapocho-Trebal WWTP, corresponding to: raw water (grey), primary clarifier (red) and secondary treated water (blue).

Chemical oxygen demand (COD) estimation models

For COD estimation, a total of nine different mathematical models have been developed using different genetic algorithm techniques. Table 2 summarizes principal outcomes of each model, indicating aspects such as Pearson's Coefficient (Lee Rodgers & Nicewander 1988; Benesty *et al.* 2009) and the maximum number of generations, or the mutation rate, among others. For clarity, we designate the values of transmittance and absorbance measured at a certain wavelength x as T_x and A_x. Equation (1) shows the model calculated by CGA. The rest of the equations are presented in the Supplementary Information (SI) document with the reference included in Table 2. These models have been calculated on a total of 545 samples after eliminating outliers, which were taken at the WWTP inlet, the outlet of the primary clarifier and the outlet of the secondary clarifier, where 66% of the data was used to train the model (358 samples) and the remaining 34% (187 samples) to validate the model.

$$COD_{(mg/l)} = \left(c_0 * \frac{c_1 * A_{622}}{c_2 * T_{415}} + c_3 * T_{550} * \frac{c_4 * A_{400}}{c_5 * T_{465}}\right) * c_6 + c_7 \quad (1)$$

$c_0 = -0.7921$; $c_1 = -0.59153$; $c_2 = -0.81183$; $c_3 = 1.8822$; $c_4 = -0.57704 c_5 = -0.7921$; $c_6 = 1294.9$; $c_7 = -131.08$

Figure 4 shows a comparison on 15 random samples, between the COD values measured in the laboratory and values obtained by the different genetic algorithm models calculated in the cases of: *(i)* raw and treated wastewater, combined model, Figure 4(a) of models 1, S1 and S2; *(ii)* raw wastewater, Figure 4(b), for models S3–S5; *(iii)* treated water, Figure 4(c) for the models S6–S8.

Figure 4 shows that the models present considerable variation in estimation levels. In general terms, those adjusted with the OS GA technique, S2, S5 and S8, are those with the best estimate, i.e. closer to the reference value (laboratory measured, blue column).

This is clearly seen in Table 2, where the Pearson's coefficient is relatively high in all data, especially in the case of the combined models (C), as well as in the raw water and primary clarifier effluent (R). However, in the case of treated water (T), the Pearson's coefficient adopts lower values, but this is due to the low variability of the treated water data, which means that any small deviation of the estimates from the reference data is accentuated. In spite of this, the latter (T) achieve estimated values closer to reference values than those reached by the C and R models in the case of considering treated wastewater.

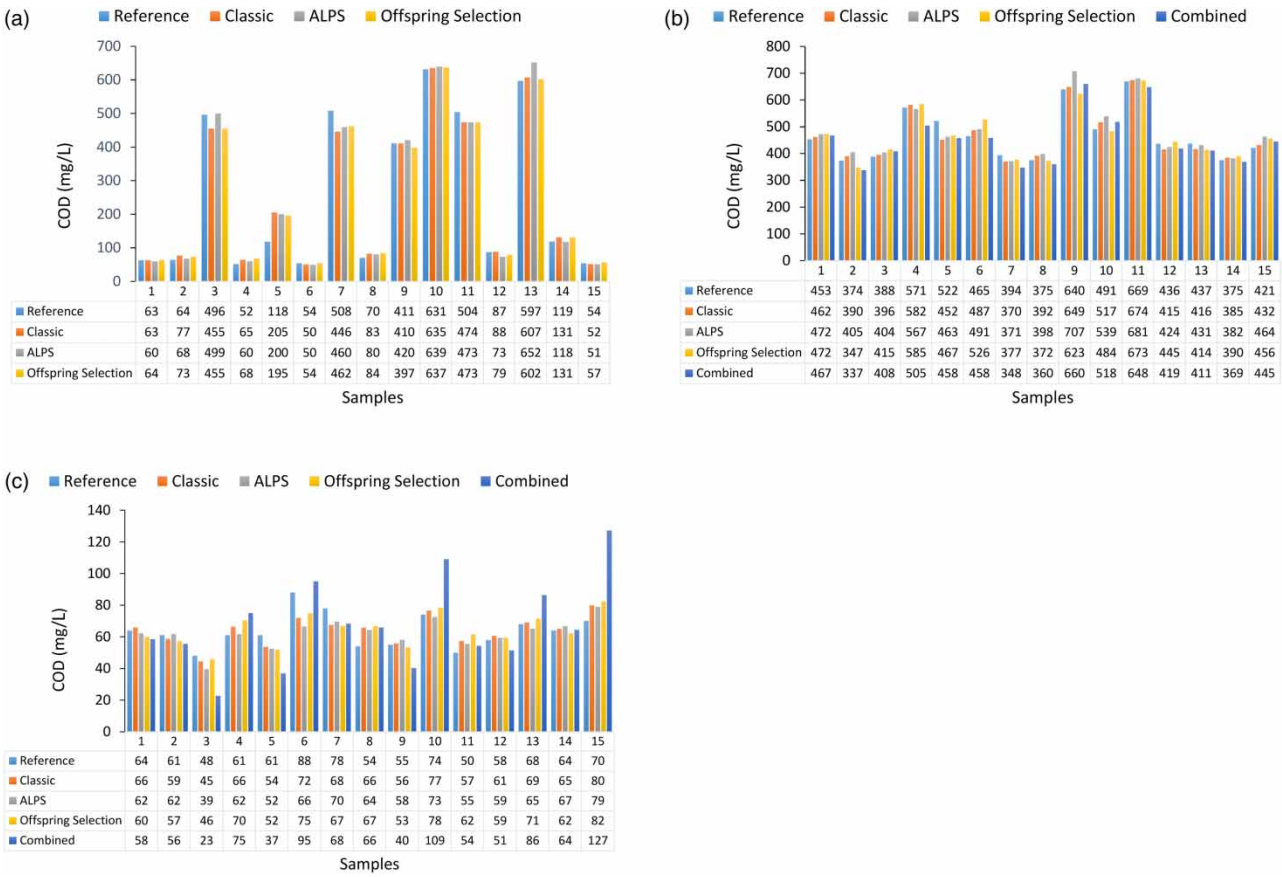
The true extent of the models' estimation capacity can be seen in Figure 4, which shows a comparison of the different models for 15 randomly drawn samples.

In Figure 4(a), regarding the combined models, we observe that all of the calculated values of the samples present a high precision in the estimations, very close to the reference values (laboratory). With regard to sample 9, for instance, whose COD value measured in the laboratory is 411 mg/L, the CGA was the one that presented the closest fit, with 410 mg/L, followed by the OS, with 397 mg/L; however, there is no significant discrepancy between the models.

Table 2 | Summary of COD estimation models

				Pearson's coefficient				
Model	**Param.**	**Type of water***	**GA**	**Training**	**Test**	**Mut. rate**	**Max. gen.**	**Optimal gen.**
1	COD	C	CGA	88.38%	91.42%	15%	50	50
S1	COD	C	APLS	89.82%	88.48%	25%	500	491
S2	COD	C	OS	89.61%	87.62%	20%	10	10
S3	COD	R	CGA	69.85%	64.30%	15%	50	50
S4	COD	R	APLS	66.11%	72.06%	25%	500	476
S5	COD	R	OS	68.24%	61.35%	20%	18	18
S6	COD	T	CGA	41.96%	28.86%	35%	80	71
S7	COD	T	APLS	41.91%	27.79%	25%	500	476
S8	COD	T	OS	44.21%	39.16%	20%	20	20

*C, combined raw and treated; R, Raw water; T, Treated water.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Reference	63	64	496	52	118	54	508	70	411	631	504	87	597	119	54
Classic	63	77	455	65	205	50	446	83	410	635	474	88	607	131	52
ALPS	60	68	499	60	200	50	460	80	420	639	473	73	652	118	51
Offspring Selection	64	73	455	68	195	54	462	84	397	637	473	79	602	131	57

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Reference	453	374	388	571	522	465	394	375	640	491	669	436	437	375	421
Classic	462	390	396	582	452	487	370	392	649	517	674	415	416	385	432
ALPS	472	405	404	567	463	491	371	398	707	539	681	424	431	382	464
Offspring Selection	472	347	415	585	467	526	377	372	623	484	673	445	414	390	456
Combined	467	337	408	505	458	458	348	360	660	518	648	419	411	369	445

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Reference	64	61	48	61	61	88	78	54	55	74	50	58	68	64	70
Classic	66	59	45	66	54	72	68	66	56	77	57	61	69	65	80
ALPS	62	62	39	62	52	66	70	64	58	73	55	59	65	67	79
Offspring Selection	60	57	46	70	52	75	67	67	53	78	62	59	71	62	82
Combined	58	56	23	75	37	95	68	66	40	109	54	51	86	64	127

Figure 4 | Comparison of 15 randomly selected samples between the COD value measured at the WWTP laboratory (Reference, blue column) and the estimated value calculated by: (a) Combined models 1, S1 and S2; (b) Specific raw wastewater models S3–S5; and (c) Specific treated wastewater models S6–S8.

On the other hand, at low COD values, all the combined models show a similar level of fit, although it is observed that, in general terms, the OS is the most adequate model in all cases, since it provides an accurate estimate while making use of few wavelengths and a reduced number of training generations.

Although a combined model (valid for waters with high and low COD levels) also provides very accurate estimates, tests have shown that specific models for each type of water provide a better fit. This is clearly seen in Figure 4(b) and 4(c) for raw water-primary clarifier (R) and treated water (T), respectively. If we look at sample 4 in Figure 4(b), for a reference value of 571 mg/L, the CGA shows an error of 2% (582 mg/L), similar to the OS (2.3%), with the ALPS showing the best result, with an error of 0.75%. The combined model presented in Equation (1) showed an error of almost 12% (505 mg/L). This is more evident in the treated water samples (Figure 4(c)), where the combined models show a greater deviation, as seen in sample 10, where, while the specific models (Equations (S6–S8)), show a deviation of less than 6%, the combined model (Equation (1)) has shown a deviation of 47%, a fact that is more clearly observed in other samples, such as 4, 6, 13 or 15.

The accuracy of all models can be seen clearly in the scatter diagram in Figure 5 for combined models, where there is a high correlation between the reference values (laboratory) and the estimated values (where the training data of the model is represented by an orange square, and the test data by a grey triangle), with practically all the points being within the $\pm 1\%$ standard deviation interval.

A comparison for all samples and estimation models combined is included in Table S1 of the Supplementary Information.

As observed in Equation (1) and the rest of the equations included in SI, each model makes use of a certain number of wavelengths, although the developed equipment performs a multispectral analysis between 380 and 700 nm that considers 162 values divided between absorbance and transmittance for each sample. Table 3 shows the weight of each of the selected wavelengths in the models performed.

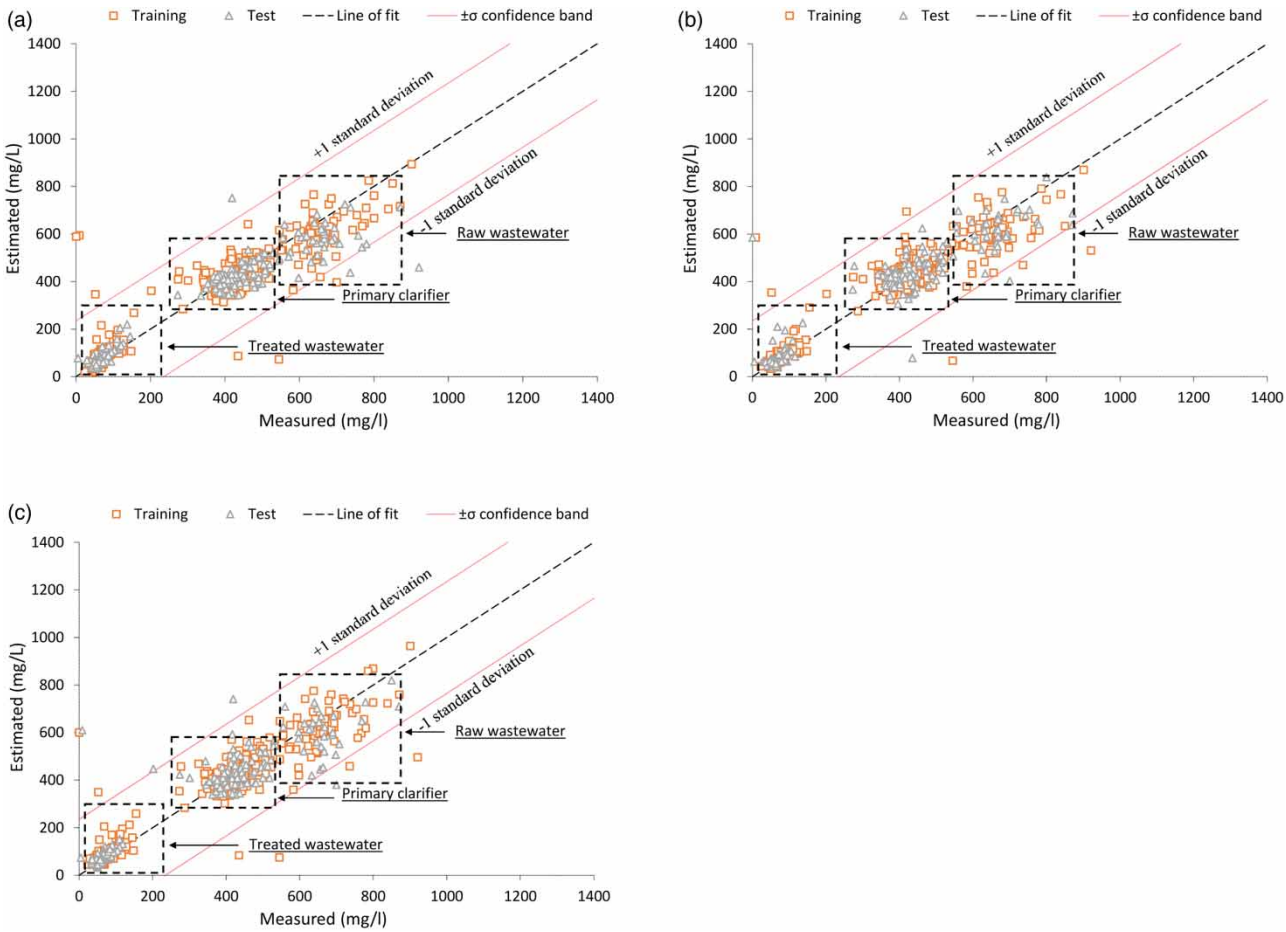


Figure 5 | Scatter plot between laboratory measured COD values (Measured) and those estimated by: (a) Model 1, Classic Genetic algorithm, (b) Model S1, ALPS algorithm, and (c) model S2, Offspring Selection algorithm.

For most of the models presented in Table 2, the violet region of the spectrum (380–427 nm) represents about 30–50% of the weight of the models. In other models, such as Equation (S1 or S4), it has been the blue region (427–476 nm) or cyan (476,497 nm) in the case of Equations (S4 and S7), which has presented these weights. Violet, blue and cyan (380–497) represent around 40% in most models presented. These results are as expected, as low wavelengths present more sensitivity to light absorbance when organic compounds are present, since organic matter has a greater sensitivity to short wavelengths (ultraviolet), which increases with shorter wavelength. Studies such as Rieger *et al.* (2004), Carré *et al.* (2017) or Chen *et al.* (2014) have shown that organic matter presents a peak at 200–400 nm. Depending on the matter that passes through the light, the jump between orbitals will be greater or smaller (Kalsi 2007). As organic matter tends to have larger jumps, it needs more energy to jump from one orbital to another, which translates into higher absorbance values.

Carré *et al.* (2017) noted that lower wavelengths between 200 and 400 nm present higher absorbance values (or lower transmittance values) in organic compounds. These wavelengths between 200 and 400 nm showed high sensitivity but also higher variability, which is not the preferred characteristic for statistical models. Therefore, their study established that the wavelength of 374 nm is the most representative for the characterization of COD (presenting more stable transmittance values).

Chen *et al.* (2014) introduced the variable pathlength – as the distance that light travels through the sample – in the absorbance measurements for the prediction of COD by Partial Least Square regression models. In their research, they concluded that as the pathlength increases, wavelengths between 394 and 406 nm are proposed for correlation models, confirming that the visible region was mostly selected from long pathlength measurements, whereas relatively short pathlength spectra contributed more in the UV region. Also, the visible regions of the spectrum were chosen to cover the absorption of colloidal fraction, which is important when samples are not filtered, as in this research.

Table 3 | Comparison of wavelengths used and their contribution weights in the calculated COD models

		C			R			T		
Spectrum	**λ***	**CGA 1**	**ALPS S1**	**OS S2**	**CGA S3**	**ALPS S4**	**OS S5**	**CGA S6**	**ALPS S7**	**OS S8**
Violet 380–427 nm	390							10% (T)		
	395									
	400	47% (A)				23% (A)	27% (A)	21% (A)		41% (T)
	410			28% (A)						
	415	1% (T)				12% (A)				
	425			26% (A)						
Blue 427–476 nm	428			30% (A)						
	440		4% (A)		40% (T)		28% (A)			
	445					2% (A)				
	455						27% (T)			
	460		38% (A)							
	465	43% (T)								
Cyan 476–497 nm	485									
	490				42% (T)	24% (A)				
	495								52% (T)	
Green 497–570 nm	500		41% (A)						48% (A)	
	505							51% (A)		40% (A)
	515		5% (A)							
	550	7% (T)				15% (A)				
	555		3% (T)		8% (A)	20% (T)				
Orange 581–618 nm	585									
	586						3% (T)			
	591			3% (T)				11% (T)		
	609		7% (T)	13 (A)						
	615		2% (T)							
Red 618–700 nm	622	2% (A)								
	624						12% (A)			
	625				9% (A)					
	627							7% (T)		
	630									12% (T)
	632						3% (T)			
	636					5% (T)				
	639									7% (A)

*Wavelengths (nm).

Based on this, in the present manuscript, the tests carried out have shown that the wavelengths in the violet region of the visible spectrum, closer to the UV region, present enough sensitivity for the prediction of the COD and are the ones that have the highest significance in the COD characterization models, achieving high levels of fitness.

This greater weighting of the wavelengths closer to the UV region is also observed in the specific models. In the case of S3–S5, for raw and primary clarifier effluent wastewater, we observe that there is a great preponderance of wavelengths closer to violet, such as 400 nm, which represent between 22 and 27% in each of the models. In the case of treated models, the wavelengths closer to the violet region of the spectrum (including the blue region) are those that have a greater weight in the estimation of COD, regardless of the type of wastewater being analysed.

Also, the supracolloidal and sedimentable matter contained in the samples causes an important effect in spectroscopy surrogates that react to all the wavelengths. These have the characteristic that the longer the wavelength, the lower the variability in the measurements (Van den Broeke *et al.* 2006). For this reason, the algorithms tend also to select the near-IR region to model its behaviour, since it has greater spectral stability. In averaged terms, the red and near-IR (618–700) represent a similar average in the most models calculated, with a weight between 2 and 12% in that region, as can be seen able to see in Table 3, although this average is observed in the orange spectrum region in models such as Equation (S1 or S6) as well.

Total suspended solids (TSS) estimation models

Table 4 shows a summary of the different models calculated for TSS, indicating their mutation rate, Pearson coefficient, or type of genetic algorithm used, among others. In this section, both specific models for raw and treated water are shown, as well as models that are valid for both. The combined models were calculated by mean of 497 samples out of a total of 550,

Table 4 | Summary of TSS estimation models

				Pearson's Coefficient				
Model	**Param.**	**Type of water***	**GA**	**Training**	**Test**	**Mut. rate**	**Max. gen.**	**Optimal gen.**
S9	TSS	C	CGA	87.01%	84.80%	15%	50	46
S10	TSS	C	APLS	89.17%	79.75%	25%	500	455
S11	TSS	C	OS	85.73%	86.06%	20%	15	15
S12	TSS	R	CGA	69.33%	59.99%	15%	50	50
S13	TSS	R	APLS	69.09%	59.99%	25%	500	495
S14	TSS	R	OS	67.37%	66.49%	20%	15	15
S15	TSS	T	CGA	53.59%	41.05%	15%	50	45
S16	TSS	T	APLS	59.44%	39.59%	25%	500	478
S17	TSS	T	OS	55.43%	44.14%	20%	20	20

*C, combined raw and treated; R, Raw water; T, Treated water.

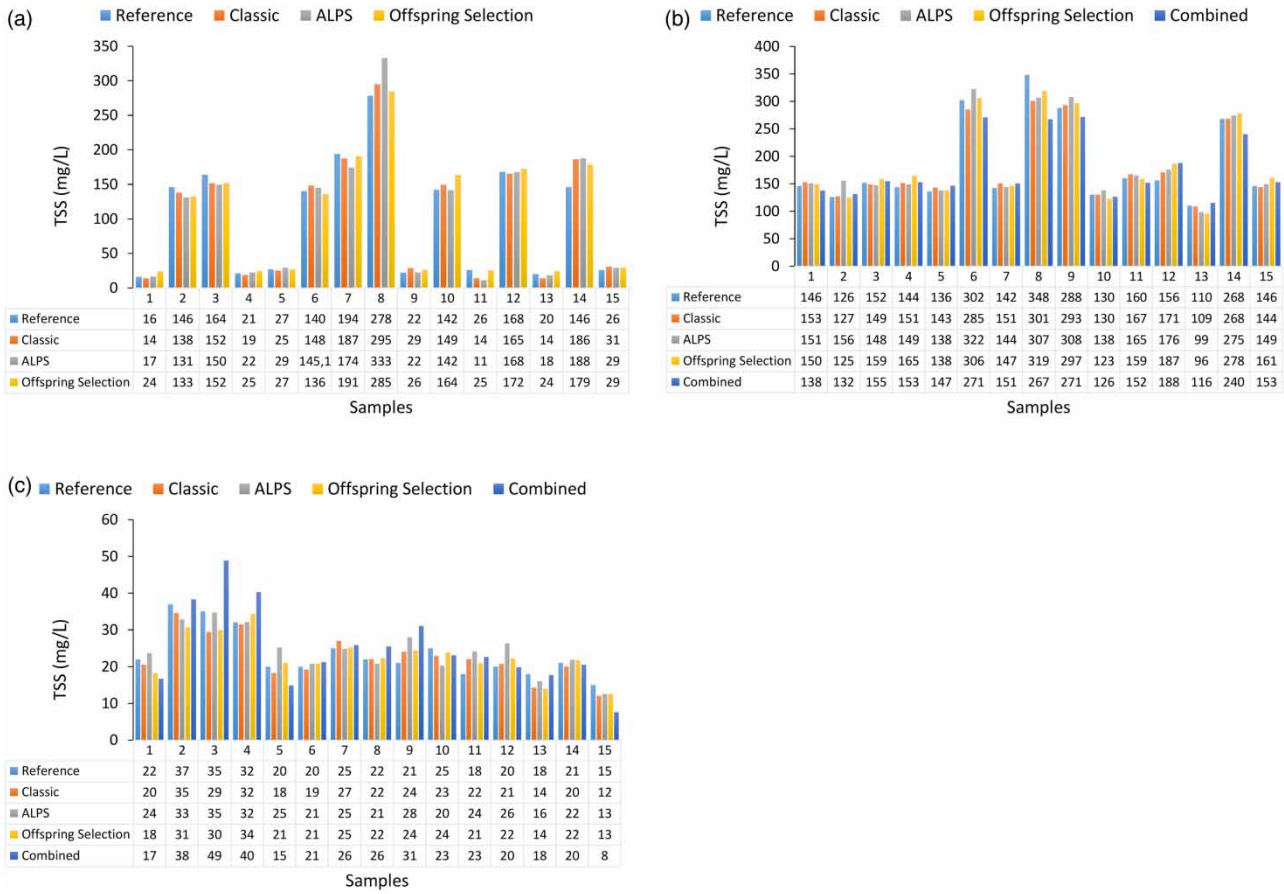


Figure 6 | Comparison of 15 randomly selected samples between the TSS value measured at the WWTP laboratory (Reference, blue column) and the estimated value calculated by: (a) Combined models S9–S11, (b) Specific raw wastewater models S12–S14 and (c) Specific treated wastewater models S15–S17.

after eliminating outliers, while the specific models for raw and treated water used a total of 326 and 196 samples, respectively, after eliminating outliers.

As we can see, the Pearson coefficients between the training and test data are very similar in almost all the models presented.

A comparison between the different models is shown in Figure 6, on ten randomly selected samples, for the estimation of TSS: Combined model, in Figure 6(a), models S9–S11; raw wastewater and primary clarifier effluent (R), in Figure 6(b), models S12–S14; and secondary treated wastewater, Figure 6(c), S15–S17. As shown in Figure 6(a), the models presented in Equations (S9–S11), in all cases, provide estimates of the TSS concentration very close to the laboratory (reference) measurement. Although the level of fit is similar, it has been observed that the OS has represented the best estimates in most samples. If we pay attention to sample 8, related to raw water, it presents a reference value of 278 mg/L, where the CGA has provided a value of 295 mg/L (deviation of 6%), 285 mg/L for OS (deviation of just under 2.5%), while the ALPS has provided a value of 333 mg/L, i.e. with a deviation of almost 20%. This is observed in most cases, where ALPS usually shows a higher deviation compared to the other techniques, although these are not very significant in most cases.

Although the settings presented by the combined models are very close to the reference ones, a specific model for a type of water (raw or treated) will always provide better results. This is observed both in Figure 6(b) (raw water and primary clarifier) and especially in Figure 6(c) for treated water, where in this latter type of water (treated), it can be seen more clearly how the combined model (Equation (S9)) presents deviations of up to 50% with respect to the reference values for most cases.

In the case of sample 15 in Figure 6(c), we observe that for a reference value of 15 mg/L, the estimates are very close, being 12 mg/L (Equation (S15)), 13 mg/L (Equation (S16)) and 13 mg/L (Equation (S17)).

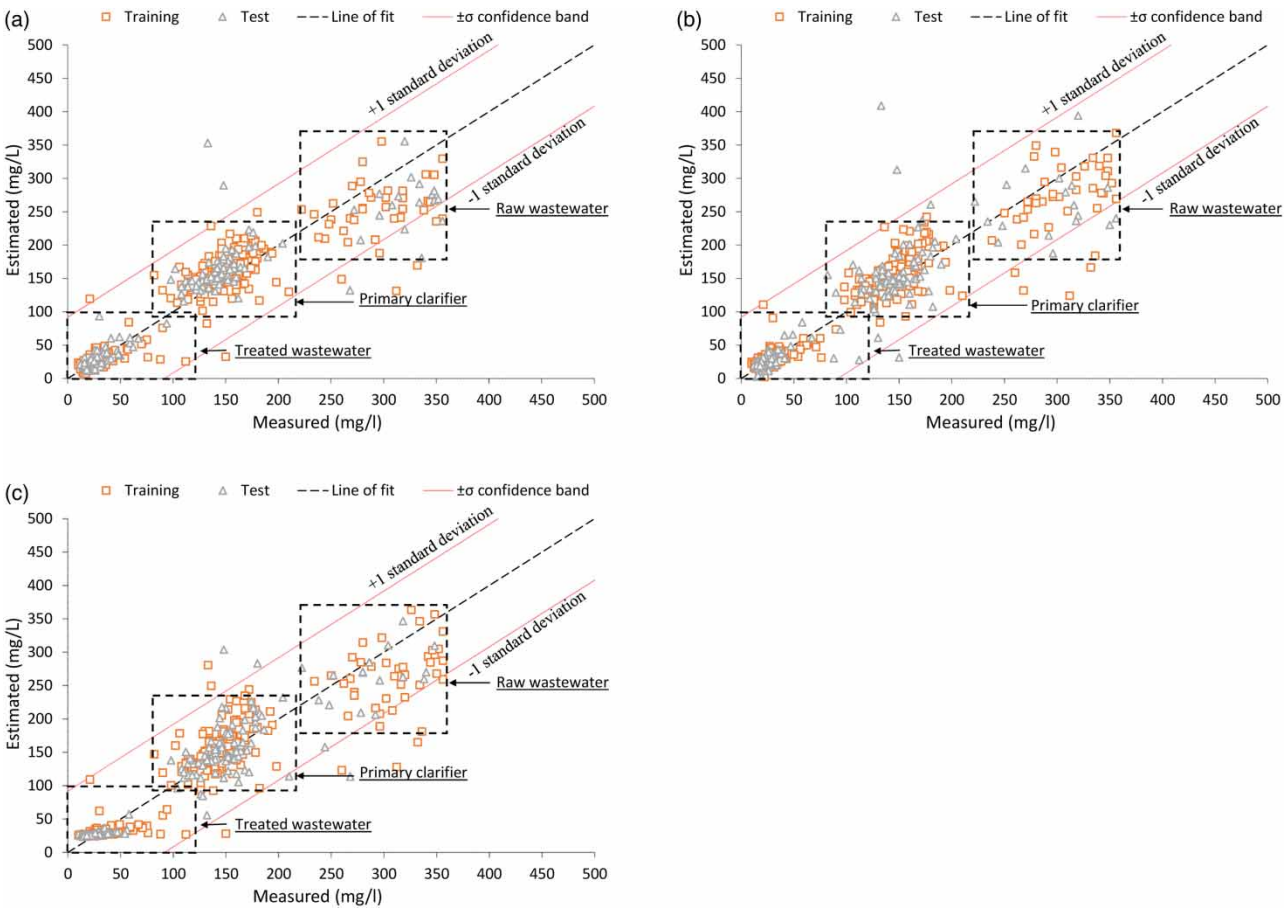


Figure 7 | Scatter plot between laboratory measured TSS values (Measured) and those estimated by: (a) S9, Classic Genetic algorithm, (b) S10, ALPS algorithm, and (c) S11, Offspring Selection algorithm.

The accuracy of the combined models can be seen in the scatter plot in Figure 7. As can be seen, in all cases, the points are on the diagonal. Although the number of points outside the interval ±1% of the standard deviation is very similar in all the models, Figure 7(b) for S10 and Figure 7(c) for S11 show a higher concentration of points around the main diagonal, indicating that there is a greater correlation between what has been measured in the laboratory (Measured) and what has been estimated with the model (Estimated). In the case of the classical algorithm (Figure 7(a)), a lower concentration of points on the diagonal is observed, although not very significant.

A comparison for all samples and estimation models combined for TSS is included in Table S2 of the Supplementary Information.

Figure 8 presents the scatter plots of the models calculated specifically for the secondary treated wastewater in the cases of S15 (Figure 8(a)), S16 (Figure 8(b)) and S17 (Figure 8(c)). As was previously discussed, although the Pearson's coefficients obtained for these models are lower (Table 4), the accuracy of the results is higher than that obtained with the combined and raw and first clarifier models, as shown with Figure 6.

In all the models calculated, it is observed that the wavelength groups used for the estimation of the TSS are very similar. Table 5 shows a comparison of the different wavelengths selected by the evolutionary models.

Similar to COD, in all models (Equations (S9–S17)), a great influence of the violet-blue region of the spectrum (380–476 nm) is observed, with weights between 70 and 80% in most models, while in other regions such as Red (618–780 nm), its influence is around 10% in most cases, although the whole spectrum is observed in the determination of the TSS, always providing a lower weight than the violet-blue region.

As previously commented, the wavelengths used for each parameter are consistent with what is established in the literature. In the case of inorganic matter, as part of those that we can find within the TSS (in addition to organic matter), it is observed

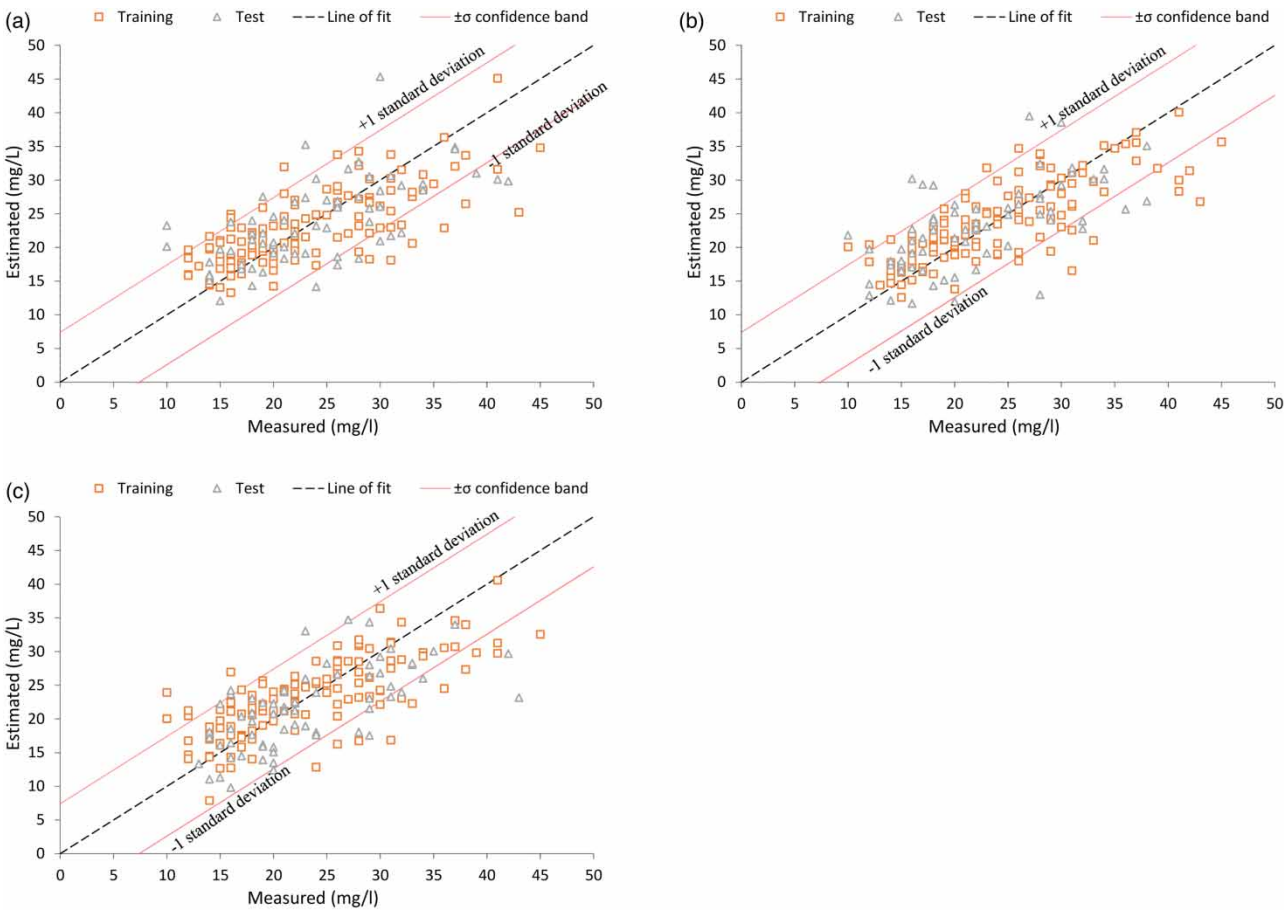


Figure 8 | Scatter plot between laboratory measured TSS values (Measured) and those estimated by: (a) S15, Classic Genetic algorithm, (b) S16, ALPS algorithm, and (c) S17, Offspring Selection algorithm.

Table 5 | Comparison of wavelengths used and their contribution weights in the calculated TSS models

		C			R			T		
Spectrum	**λ***	**CGA S9**	**ALPS S10**	**OS S11**	**CGA S12**	**ALPS S13**	**OS S14**	**CGA S15**	**ALPS S16**	**OS S17**
Violet 380–427 nm	395								5% (T)	6% (T)
	400	42% (T)	20% (T)						48% (A)	
	410								13% (T)	
Blue 427–476 nm	435			30% (A)						
	440					23% (A) 5% (T)	34% (T)			
	450	41% (T)				35% (T)				
	455		29% (T)							
	460			21% (A)			33% (T)			
	461			30% (A)	43% (T)					
Cyan 476–497 nm	490		31% (T)					22% (A)		
	495				16% (A)					
Green 497–570 nm	500						11% (T)			
	505							23% (A)		
	510								8% (T)	
	521					8% (T)				
	522				13% (A)					
	530							25% (A)		62% (T)
	550		8% (T)							
	557	5% (T)								
	558			6% (A)						
	565								2% (T)	
Orange 581–618 nm	583									24% (T)
	585					0.2% (A)		9% (A)		
	586					4% (T)				
	607			7% (A)						
	609					13% (T)	15% (T)		5% (A)	
	610							6% (A)		
	615				6% (T)					8% (A)
Red 618–780 nm	622	4% (T)	3% (T)							
	624			6% (A)						
	625				6% (T)					
	630				9% (T)					
	631		3% (T)		9% (T)	12% (T)				
	632	4% (T)								
	635						7% (T)			
	636	4% (T)								
	642							6% (A)		
	655								14% (A)	
	656								4% (T)	
	660		7% (T)							

*Wavelengths (nm).

that it is sensitive to the entire visible spectrum. The work of Sarraguça *et al.* (2009) shows that wavelengths closer to the infrared and violet-blue region of the spectrum are usually a common denominator in these species and verified that an NIR-based model provided a relative standard deviation comparable to those obtained using UV-visible techniques.

CONCLUSIONS

In the present research work, a total of 18 mathematical models based on genetic algorithms have been developed to estimate the concentration of COD and TSS in wastewater samples taken at three locations of the Mapocho-Trebal WWTP (Chile), specifically: At the inlet of the plant (raw wastewater), at the outlet of the primary clarifier, and at the outlet of the plant (secondary treated wastewater).

Of these 18 models, 9 have been used to calculate COD and the other half to calculate TSS from the spectral response of the samples analysed within the 380–700 nm range, by means of equipment developed for this purpose, without any pre-treatment processes. For each parameter, models in three categories have been developed: Combined models (valid for any type of wastewater in the plant), specific models for raw wastewater and primary decanter, and specific models for secondary treated wastewater, in order to make a comparison of which is more suitable for the estimation process, as well as whether the discrepancies between them were significant or not.

Within each category, three different models were developed using three GA techniques: CGA, ALPS and OS. To obtain these models, a sampling campaign was carried out during 6 months in 2021, in a WWTP located in the city of Chile: Mapocho-Trebal, which serves a population of 3,674,880 equivalent inhabitants, achieving some 550 samples analysed both in the laboratory and with the spectrophotometer developed.

Genetic algorithms have proven to be a very effective tool in the characterization of wastewaters from the spectral response, as they are non-linear models. This is relevant, given that an analysis by more conventional methods, such as multivariate linear regression, is greatly affected by the variability of the data (variance) and cannot recognize non-linear patterns. The tests carried out have shown that although it is possible to obtain linear models to estimate COD and TSS, they have two disadvantages with respect to the genetic algorithms developed in this research work that usually give rise to much lower fits, around 70% for these models, and they are found to be mainly useful for samples with a high contaminant load, not being valid either for secondary treated wastewater samples or for combined models (raw wastewater and secondary treated wastewater).

This is because the spectral response of a secondary treated wastewater sample is almost horizontal (within the visible spectrum), with small fluctuations, while one with a higher concentration of solids or organic matter (higher turbidity) presents a spectral response in the form of an upward slope, which makes it much easier to detect points on the graph (wavelengths, i.e. regressors) that can be decisive in obtaining a linear model.

However, a model that combines samples with high and low COD and TSS values also presents difficulties for estimation by linear models, because the variability of the data makes it difficult to obtain a good fit using this type of technique.

Based on these limitations, non-linear models obtained from genetic algorithms represent an extraordinary solution to the characterization problem, thanks to their evolutionary character, which learns through crossover and mutation of individuals (randomly generated mathematical functions) until reaching in one generation an individual (mathematical function) that is able to estimate a certain pollutant parameter with a higher accuracy.

All models have a high accuracy, which can be observed in the Pearson's coefficient. In the case of the combined models, the COD estimation models are around 90% for both training and test, values similar to those obtained for the TSS, with 80–85% for test and 85–90% for training. In comparisons between the combined and specific models for a given type of wastewater, it has been shown that a specific model, in general terms, provides a better fit than a model that is valid for very different wastewaters, but the differences do not seem to be very significant, with discrepancies being within levels that the authors consider acceptable for an early warning system for the characterization of the pollutant load.

Each of the calculated models has a different structure and variables (wavelengths). This is due to the fact that the starting data for each model are taken at random. However, in practically all the models, it is observed that the violet-blue region of the visible spectrum has a significant weight in the characterization of COD and TSS for most models calculated, with levels ranging from 80% in some models for COD, and around 30–70% for TSS, if we add the weights of all wavelengths involved in the violet region (380–476 nm) in Tables 3 and 5, although it has been observed in some cases that the wavelengths closer to the infrared region also have an important weight in most models, although on a lower average than the violet-blue region, representing little more than 10% in the best of cases, with a few exceptions.

The results and models presented in this work are intended to serve as a complement to the wastewater analysis and characterization systems within the WWTP, as well as to serve as a basis for the development of new monitoring and early warning systems for unauthorized discharges, which are simpler, faster and more economical, in order to achieve the real-time control so necessary for proper management of wastewater resources and environmental protection.

ACKNOWLEDGEMENTS

The authors wish to thank the help and availability received from the company and technical personnel from EDAM Ltda. during the field campaign.

FUNDING

The author Daniel Carreres Prieto wishes to thank the financial support received from the Seneca Foundation of the Región de Murcia (Spain) through the program devoted to training novel researchers in areas of specific interest for the industry and with a high capacity to transfer the results of the research generated, entitled: 'Subprograma Regional de Contratos de Formación de Personal Investigador en Universidades y OPIs' (Mod. B, Ref. 20320/FPI/17). The present research has been funded by the project *MONITOCOES: New intelligent monitoring system for microorganisms and emerging contaminants in sewage networks*. Reference: RTC2019-007115-5 by the Ministry of Science and Innovation – State Research Agency, within the RETOS COLABORACIÓN 2019 call, which supports cooperative projects between companies and research organizations, whose objective is to promote technological development, innovation and quality research.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Affenzeller, M. & Wagner, S. 2005 Offspring selection: a new self-adaptive selection scheme for genetic algorithms. In: Ribeiro, B., Albrecht, R.F., Dobnikar, A., Pearson, D.W., Steele, N.C. (eds) *Adaptive and Natural Computing Algorithms*. Springer, Vienna, doi:10.1007/3-211-27389-1_52.

Benesty, J., Chen, J., Huang, Y. & Cohen, I. 2009 Pearson correlation coefficient. In: *Noise Reduction in Speech Processing*. Springer Topics in Signal Processing, vol 2. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-00296-0_5.

Beraud, B., Lemoine, C. & Steyer, J. P. 2009 Multiobjective genetic algorithms for the optimisation of wastewater treatment processes. In: do Carmo Nicoletti, M., Jain, L.C. (eds) *Computational Intelligence Techniques for Bioprocess Modelling, Supervision and Control*. Studies in Computational Intelligence, vol 218. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-00296-0_5.

Bogomolov, A. & Melenteva, A. 2013 Scatter-based quantitative spectroscopic analysis of milk fat and total protein in the region 400–1100 nm in the presence of fat globule size variability. *Chemometrics and Intelligent Laboratory Systems* **126**, 129–139.

Bogomolov, A., Dietrich, S., Boldrini, B. & Kessler, R. W. 2012 Quantitative determination of fat and total protein in milk based on visible light scatter. *Food Chemistry* **134** (1), 412–418.

Carré, E., Pérot, J., Jauzein, V., Lin, L. & Lopez-Ferber, M. 2017 Estimation of water quality by UV/Vis spectrometry in the framework of treated wastewater reuse. *Water Science and Technology* **76** (3), 633–641.

Carreres-Prieto, D., García, J. T., Cerdán-Cartagena, F. & Suardiaz-Muro, J. 2019 Spectroscopy transmittance by LED calibration. *Sensors* **19** (13), 2951.

Carreres-Prieto, D., García, J. T., Cerdán-Cartagena, F. & Suardiaz-Muro, J. 2020 Wastewater quality estimation through spectrophotometry-based statistical models. *Sensors* **20** (19), 5631.

Chen, B., Wu, H. & Li, S. F. Y. 2014 Development of variable pathlength UV–vis spectroscopy combined with partial-least-squares regression for wastewater chemical oxygen demand (COD) monitoring. *Talanta* **120**, 325–330.

Do, H. T., Van Bach, N., Van Nguyen, L., Tran, H. T. & Nguyen, M. T. 2021 A design of higher-level control based genetic algorithms for wastewater treatment plants. *Engineering Science and Technology, an International Journal* **24** (4), 872–878.

El Khorassani, H., Trebuchon, P., Bitar, H. & Thomas, O. 1999 A simple UV spectrophotometry procedure for the survey of industrial sewage system. *Water Science and Technology* **39** (10–11), 77–82.

Han, Y. Z., Ji, W. X., Jiang, B. C., Tian, Y. C., Shen, S. Q., Zhou, D., Li, Y., Shuang, C.-D., Li, W.-T., Lu, H. & Li, A. M. 2021 Developing a miniaturized spectrophotometer using 235 and 275 nm UVC-LEDs for fast detection of nitrate in natural water and wastewater effluents. *ACS ES&T Water* **1** (12), 2548–2555.

Hornby, G. S. 2006 ALPS: the age-layered population structure for reducing the problem of premature convergence. In: *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*. pp. 815–822.

Jeong, H. S., Lee, S. H. & Shin, H. S. 2007 Feasibility of on-line measurement of sewage components using the UV absorbance and the neural network. *Environmental Monitoring and Assessment* **133** (1), 15–24.

Kalsi, P. S. 2007 *Spectroscopy of Organic Compounds*. New Age International, Ludhiana, India.

Koza, J. 1992 *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA.

Lee Rodgers, J. & Nicewander, W. A. 1988 Thirteen ways to look at the correlation coefficient. *The American Statistician* **42** (1), 59–66.

Lepot, M., Torres, A., Hofer, T., Caradot, N., Gruber, G., Aubin, J. B. & Bertrand-Krajewski, J. L. 2016 Calibration of UV/Vis spectrophotometers: a review and comparison of different methods to estimate TSS and total and dissolved COD concentrations in sewers, WWTPs and rivers. *Water Research* **101**, 519–534.

Mohammad, K. A., Zekry, A. & Abouelatta, M. 2015 LED based spectrophotometer can compete with conventional one. *International Journal of Engineering & Technology* **4** (2), 399.

Nasteski, V. 2017 An overview of the supervised machine learning methods. *Horizons. b* **4**, 51–62.

Niculescu-Mizil, A. & Caruana, R. 2005 Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning*. pp. 625–632.

Osowski, S. 1996 *Sieci Neuronowe W Uj¸eciu Algorytmicznym*. Wydawnictwa Naukowo-Techniczne, Warsaw, Poland.

Pacheco Fernández, M., Knutz, T. & Barjenbruch, M. 2020 Multi-parameter calibration of a UV/Vis spectrometer for online monitoring of sewer systems. *Water Science and Technology* **82** (5), 927–939.

Prairie, M. W., Frisbie, S. H., Rao, K. K., Saksri, A. H., Parbat, S. & Mitchell, E. J. 2020 An accurate, precise, and affordable light emitting diode spectrophotometer for drinking water and other testing with limited resources. *PloS One* **15** (1), e0226761.

Qin, X., Gao, F. & Chen, G. 2012 Wastewater quality monitoring system using sensor fusion and machine learning techniques. *Water Research* **46** (4), 1133–1144.

Rajasekaran, S. & Pai, G. V. 2003 *Neural Networks, Fuzzy Logic and Genetic Algorithm: Synthesis and Applications*. PHI Learning Pvt. Ltd, Coimbatore, India.

Rauch, W. & Harremoës, P. 1999 Genetic algorithms in real time control applied to minimize transient pollution from urban wastewater systems. *Water Research* **33** (5), 1265–1277.

Rieger, L., Langergraber, G., Thomann, M., Fleischmann, N. & Siegrist, H. 2004 Spectral in-situ analysis of NO_2, NO_3, COD, DOC and TSS in the effluent of a WWTP. *Water Science and Technology* **50** (11), 143–152.

Sarraguça, M. C., Paulo, A., Alves, M. M., Dias, A. M., Lopes, J. A. & Ferreira, E. C. 2009 Quantitative monitoring of an activated sludge reactor using on-line UV-visible and near-infrared spectroscopy. *Analytical and Bioanalytical Chemistry* **395** (4), 1159–1166.

Sooväli, L., Rõõm, E. I., Kütt, A., Kaljurand, I. & Leito, I. 2006 Uncertainty sources in UV-Vis spectrophotometric measurement. *Accreditation and Quality Assurance* **11** (5), 246–255.

Thomas, O. & Burgess, C. 2007 *UV-visible Spectrometry of Water and Wastewater*. Elsevier B.V, Amsterdam, The Netherlands.

Torres, A., Lepot, M. & Bertrand-Krajewski, J. L. 2013 Local calibration for a UV/Vis spectrometer: PLS vs. SVM. A case study in a WWTP. In: *7th International Conference on Sewer Processes & Networks*. pp. 1–8.

Van Den Broeke, J., Langergraber, G. & Weingartner, A. 2006 On-line and in-situ UV/vis spectroscopy for multi-parameter measurements: a brief review. *Spectroscopy Europe* **18** (4), 15–18.

Wagner, S., Kronberger, G., Beham, A., Kommenda, M., Scheibenpflug, A., Pitzer, E., Vonolfen, S., Kofler, M., Winkler, S., Dorfer, V. & Affenzeller, M. 2014 Architecture and design of the HeuristicLab optimization environment. In: Klempous, R., Nikodem, J., Jacak, W., Chaczko, Z. (eds) *Advanced Methods and Applications in Computational Intelligence*. Topics in Intelligent Engineering and Informatics, vol 6. Springer, Heidelberg. https://doi.org/10.1007/978-3-319-01436-4_10.

First received 19 January 2022; accepted in revised form 15 April 2022. Available online 23 April 2022