Chemosphere 293 (2022) 133610

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Implementing Early Warning Systems in WWTP. An investigation with cost-effective LED-VIS spectroscopy-based genetic algorithms

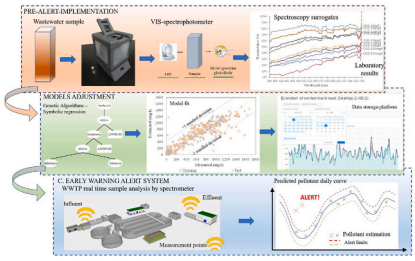
Daniel Carreres-Prieto [a,**], Juan T. García [a,*], Fernando Cerdán-Cartagena [b], Juan Suardiaz-Muro [c], Carlos Lardín [d]

[a] Department of Mining and Civil Engineering, Universidad Politécnica de Cartagena, 30202, Cartagena, Spain
[b] Department of Information and Communications Technologies, Universidad Politécnica de Cartagena, 30202, Cartagena, Spain
[c] Department of Electronic Technology, Universidad Politécnica de Cartagena, 30202, Cartagena, Spain
[d] Entidad de Saneamiento y Depuración de Aguas Residuales de la Región de Murcia (ESAMUR), c/Madre Paula Gil Cano, s/n, E-30009, Murcia, Spain

HIGHLIGHTS

- Spectroscopy in the VIS range 380–700 nm shows accuracy for estimating physico-chemical parameters in a field campaign in total of 43 WWTPs.
- Genetic Algorithms techniques have proven to be a tool to establish the correlation between transmittance and absorbance with water quality.
- First steps towards continuous and real-time monitoring of pollutants in WWTPs by means of a cost-effective LED-VIS device.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Y Yeomin Yoon

Keywords:
LED spectrophotometer
Wastewater pollutant characterization
Symbolic regression
Genetic algorithm

ABSTRACT

Measuring how the pollution load evolves in real time along sewer networks is key for proper management of water resources and protecting the environment. The technique of molecular spectroscopy for water characterization has increasingly widespread use, as it is a non-invasive technique that leads to the correlation of the physical-chemical conditions of wastewater with spectroscopic surrogates by a series of mathematical estimation models. In the present research work, different symbolic regression models obtained with evolutive genetic algorithms are evaluated for the estimation of chemical oxygen demand (COD); five-day biochemical oxygen demand (BOD_5); total suspended solids (TSS); total phosphorus (TP); and total nitrogen (TN), from the spectral response of samples measured between 380 and 700 nm and without the use of chemicals or pre-treatment. Around 650 wastewater samples were used in the campaign, from 43 different wastewater treatment plants (WWTP) in which both, raw/influent and treated/effluent, were examined through 18 models composed of Classical Genetic Algorithm (CGA), the Age-Layered Population Structure (ALPS), and Offspring Selection (OS) by mean of HeuristicLab software, to make a comparison among them and to determine which models and wavelengths are most suitable for the correlation. Models are proposed considering both raw and treated samples together (15) and only with tertiary treated wastewater reclaimed for agriculture irrigation effluent (3). The

* Corresponding author.
** Corresponding author.
E-mail addresses: daniel.carreres@upct.es (D. Carreres-Prieto), juan.gbermejo@upct.es (J.T. García), fernando.cerdan@upct.es (F. Cerdán-Cartagena), juan.suardiaz@upct.es (J. Suardiaz-Muro), carlos.lardin@esamur.com (C. Lardín).

https://doi.org/10.1016/j.chemosphere.2022.133610
Received 5 September 2021; Received in revised form 8 January 2022; Accepted 11 January 2022
Available online 17 January 2022

Pearson correlation coefficients were in the range of 67–91% for the test data in the case of the combined models. The results conform the first steps for a real-time monitoring of WWTP.

1. Introduction

Real-Time monitoring and Control (RTC) and Early Warning Systems (EWS) in Wastewater Treatment Plants (WWTP) provide valuable information on the evolution of pollutants transported by the influent during dry as well as extreme weather events, thereby enabling rapid adjustments in the plant's operating parameters. Moreover, the information serves to improve effluent quality, promote water reuse, and to prevent failures in biological processes caused by possible toxicity (Korshin et al., 2018; Jurga et al., 2017).

Spectroscopic surrogates, such as absorbance and transmittance, have a proven ability to monitor water quality in the influent and effluent of WWTP. Molecular absorption spectroscopy in the ultraviolet and visible (UV/VIS) spectral regions (Skoog et al., 2017; Thomas and Burgess, 2017) and light scattering in the visible and short wave near infrared (VIS/SW-NIR –(Near Infrared, 800–2500 nm)) (Bogomolov et al., 2012; Bogomolov and Melenteva, 2013) are techniques that allow the quantitative correlation of inorganic, organic, and biological species such as biochemical oxygen demand in five days (BOD_5); chemical oxygen demand (COD); total suspended solids (TSS); total nitrogen (TN); nitrate-nitrogen (NO3–N); and total phosphorus (TP), among others (Melendez-Pastor et al., 2013). Molecular absorption of photons in organic compounds is mainly observed in the UV/VIS range by Beer's law, which describes the absorption behaviour of media in terms of absorbance. Despite this, it is also usual that the complexity of the wastewater matrix, where the solute-solute interactions can alter the ability of the analytic species to absorb a given wavelength of radiation, is a limitation in the UV range of the spectrum. On the other hand, the scatter in the visible (VIS) spectrum, which is principally describing, in an indirect way, the particle density and their size distribution, is capable of investigating complex dispersive media.

To quantify the pollutant species in real samples of wastewater, where the existence of multiple compounds produces interferences, mathematical functions are usually adjusted to establish a correlation with the measured absorbance spectra. Multiple Linear Regression (MLR) and Partial least squares regression (PLSR) are mainly used to find linear correlations between absorbance and indicators of interest (Torres and Bertrand-Krajewski, 2008; Carré et al., 2017). Carreres-Prieto et al. (2020) first used symbolic regression in spectroscopic-based correlations which searched the mathematical formulas that best predicted the output through genetic algorithms.

Cost-effective spectrophotometer devices that can be readily deployed along the sewer network can contribute to achieve a real-time control monitoring platform to help in WWTP processes. The use of light-emitting diodes (LED) as light emitting sources, the simplification of optical lenses and of the charge-coupled device (CCD), while maintaining quantitative and robust predictive capabilities, are a field of innovation in recent years (Van Den Broeke et al., 2006; Carreres-Prieto et al., 2019).

The present work is based on an extensive experimental campaign of about 650 influent and effluent samples at 43 different WWTP located in the region of Murcia in south-eastern Spain. The physic-chemical conditions of wastewater samples have been used for correlation with spectroscopic surrogates by means of symbolic regression fitted from genetic algorithms. The equipment used is a cost-effective spectrophotometer with a visible spectrum range (380–700 nm), so the physical properties of light scattering have been used as the main input. The fitting equations obtained include: *i*) both influent and effluent values, 15 models and *ii*) effluent values, 3 models. This work represents a study based on GA-defined models capable of providing an accurate response for the development and implementation of a massive and real-time control platform to improve operations in WWTP.

2. Materials and methods

2.1. Experimental campaign

To achieve representative models that estimate the pollutant load of wastewater, it is recommended to have samples from a large number of different treatment plants (43 WWTP). This variability in water properties allows for more robust and generalist models, suitable for a wide range of water types and which can therefore become more useful models. Table A1 (Appendix A) shows the different WWTP used to carry out this study, indicating the number of inhabitants they support, their designs and operational flows, the average levels of COD, BOD_5, and TSS registered at the inlet and outlet of the WWTP, among other aspects. The WWTP cover a large part of the Region of Murcia (Spain). Most of them include a tertiary treatment to provide reclaimed water for agricultural irrigation. Given that the waters have different characteristics depending on the population they originate from, a sampling of all these waters could make it possible to achieve estimation models valid for all types of water. Most samples were taken between 11 January and June 22, 2021, with the exception of the samples from the Cabezo Beaza WWTP, which were taken during the period 2019 to April 2020.

In order to influence the performance of the WWTP processes through the early operation of the plant parameters, samples were taken at the following points of the WWTP: influent wastewater at the inlet of the WWTP, raw water and treated water, at the outlet of the tertiary treatment. Samples were collected homogeneously during 24 h by means of an accumulated sample of 500 mL/h. The start time of the sampling period varied depending on the wastewater treatment plant, being between 9:00 and 14:30, and the samples were analyzed almost immediately after collection.

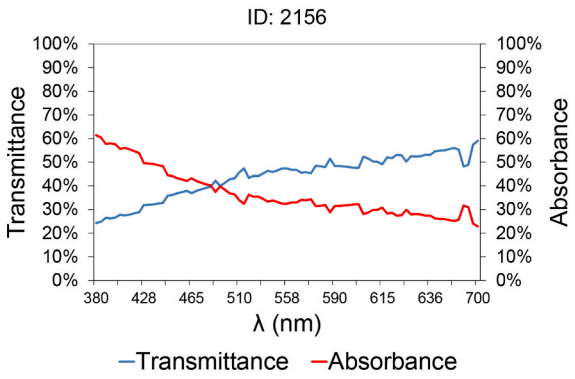
None of the samples were pre-treated by any filtering process in order to replicate the conditions of the future automated sampling of the continuous sensors. The tests performed were in accordance with Standard Methods (SM) and International Organization for Standardization (ISO). Mainly these were the dichromate method with UV–VIS spectroscopy (ISO 6060:1989) for COD; respirometric method (SM 5210 D) for BOD_5; settleable solids (SM 2540 F) in case of TSS; persulfate digestion with UV–VIS spectroscopy (SM 4500–NC) for TN and ascorbic acid method complemented with UV–VIS spectroscopy (SM 4500-P B) for TP.

In order to generate the statistical models for estimating the pollutant load from the spectrophotometric response of the samples, it is necessary to obtain, for any given sample, two sets of data: one corresponding to the spectral response measured in the range 380–700 nm (our input), and another containing the results of the analysis carried out in the laboratory (our target or output), to find the mathematical expressions that allow us to correlate both sets, i.e. to estimate one from the other.

2.2. Characteristics of water samples analyzed

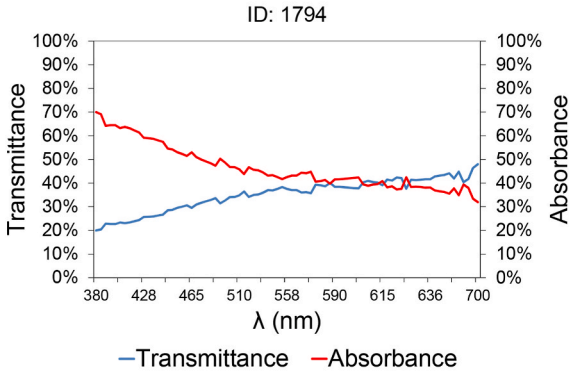
The waters present at the inlet and outlet of the WWTP have different properties, which are reflected in their own spectral response.

Fig. 1 shows the spectral response of four samples taken at random, where the first two (Fig. 1A and B) correspond to raw water samples, i.e., taken at the entrance of the WWTP. The samples in Fig. 1 C-D correspond to samples of treated water. In order to clarify the properties of each of the waters, each sample is accompanied by the analytical parameters measured in laboratories.



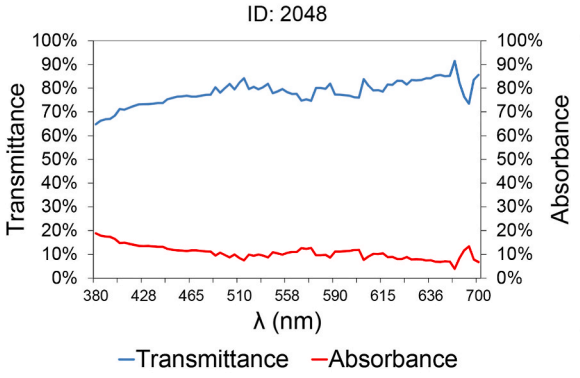
Polluting parameters	Value
COD	796 mg/l
BOD5	540 mg/l
TSS	264 mg/l
Phosphorus (P)	11.8 mg/l
Total Nitrogen (TN)	147 mg/l
NO_3-N	0 mg/l
PH	8.42
Conductivity	2120 μS/cm

A



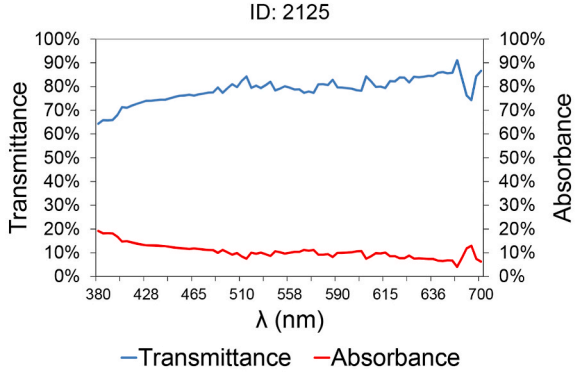
Polluting parameters	Value
COD	982 mg/l
BOD5	700 mg/l
TSS	244 mg/l
Phosphorus (P)	3.09 mg/l
Total Nitrogen (TN)	28.9 mg/l
NO_3-N	0 mg/l
PH	7.67
Conductivity	4040 μS/cm

B



Polluting parameters	Value
COD	16 mg/l
BOD5	3.7 mg/l
TSS	5.7 mg/l
Phosphorus (P)	1.73 mg/l
Total Nitrogen (TN)	2.58 mg/l
NO_3-N	0.889 mg/l
PH	7.85
Conductivity	3020 μS/cm

C



Polluting parameters	Value
COD	15 mg/l
BOD5	3 mg/l
TSS	3.4 mg/l
Phosphorus (P)	1.2 mg/l
Total Nitrogen (TN)	1.69 mg/l
NO_3-N	0.492 mg/l
PH	8.31
Conductivity	3890 μS/cm

D

Fig. 1. Spectral response of four samples taken at random, accompanied by their laboratory characterization.

As can be seen, the raw water samples present an ascending transmittance curve, with absorption values being low in the violet-blue region of the visible spectrum, and then gradually increasing as we approach the red region of the spectrum.

On the other hand, treated water, having a low concentration of pollutant load, especially suspended solids and organic matter, provides a more horizontal spectral response (transmittance), with no significant variations along the visible spectrum.

In order to visualize how these parameters evolve over time, Fig. 2 shows, for the different pollutant parameters analyzed in this research work, the daily averaged concentration in a specific WWTP over the course of a single month, showing both the values corresponding to raw water (blue graph) and treated water (orange line).

2.3. Spectrophotometric device

The equipment shown in Fig. 3A was used to carry out the study. This consisted of a low-cost spectrophotometer based on LED technology capable of performing multispectral analysis between 380 and 700 nm, with an accuracy comparable to commercial equipment based on incandescent lamps. The calibration of the equipment was performed and discussed in previous research carried out by Carreres-Prieto et al. (2019). As can be seen in Fig. 3B, it consists of a disk where 5 mm diameter LEDs are aligned with the sample under study, which is inserted through its upper part by means of a standard spectrophotometric test tube with 2.7 mL of volume. This procedure has enabled us to achieve a working range between 380 and 700 nm using just 33 LEDs that cover different areas of the visible spectrum.

2.4. Genetic algorithms

Genetic algorithms (GA) (Augusto and Barbosa, 2000; Harik et al., 1999) have been used as a technique for the study and generation of models, where the structure and its parameters are not defined, known as symbolic regression (Koza, 1992). The evolutionary nature of GA allows them to learn and to extract patterns from the input data and they are capable of overcoming certain limitations of multivariate linear regression techniques (Schmidt and Finan, 2018), such as the need for

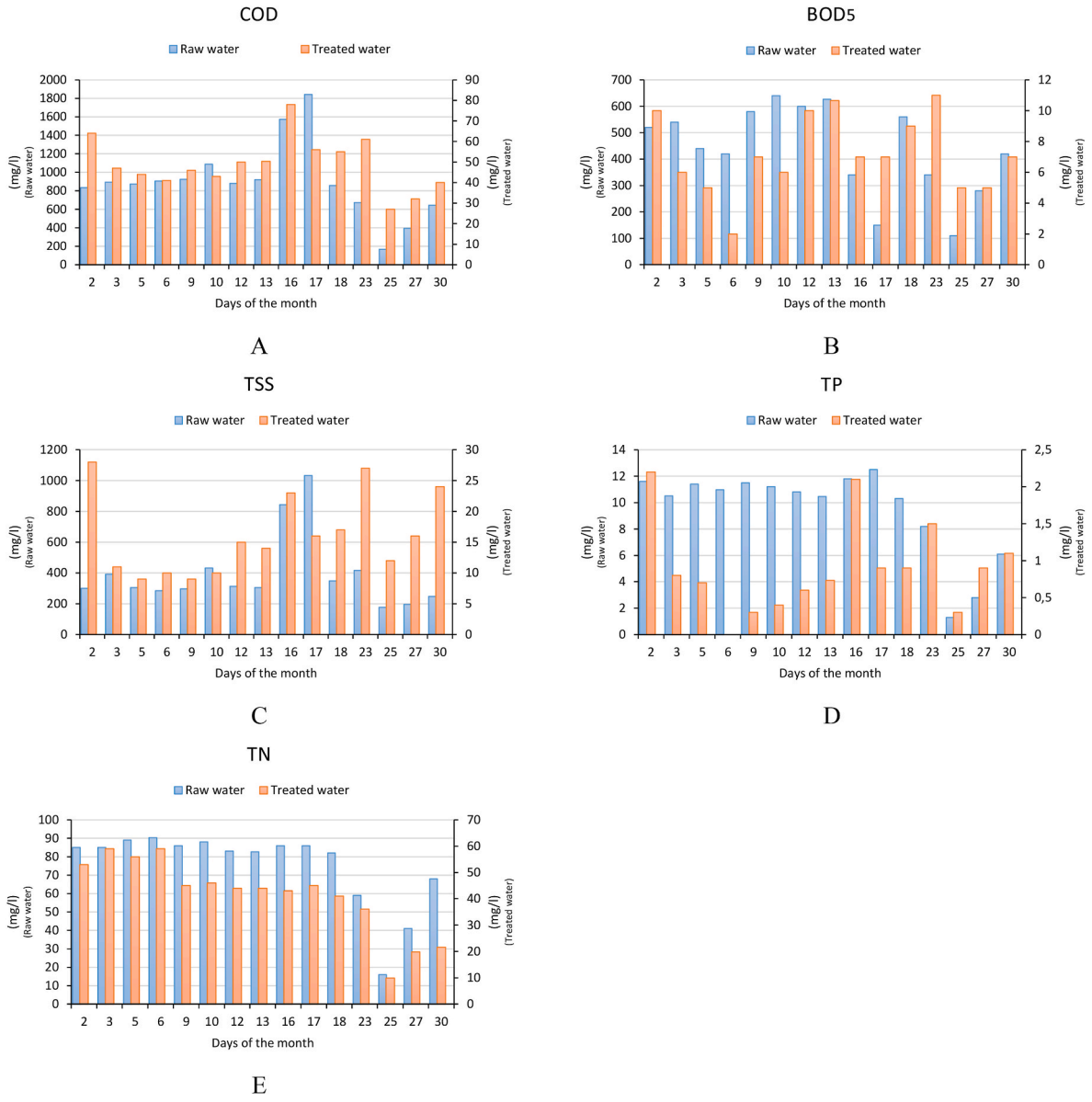


Fig. 2. Evolution over one month of the following pollutant parameters, both at the WWTP inlet (raw water) and outlet (treated water): **(A)** COD, **(B)** BOD_5, **(C)** TSS, **(D)** P, **(E)** TN.

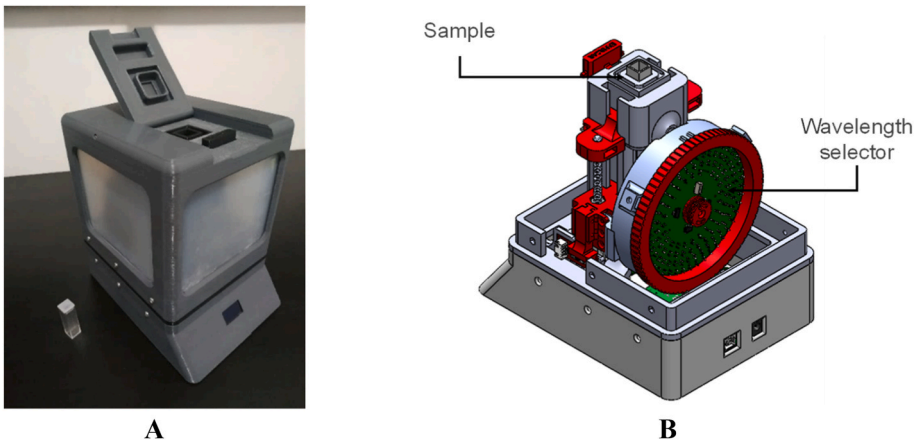


Fig. 3. View of the equipment developed to carry out the spectrophotometric analysis in the different WWTP. (**A**) External view of the equipment (**B**) Internal structure of the device.

the residuals to follow a normal distribution. As a result, their use is beginning to spread within the field of wastewater quality analysis (Cho et al., 2004; Huang et al., 2015; Holenda et al., 2007).

Within genetic algorithms, three different computational techniques have been implemented:

- **Classical Genetic Algorithm (CGA)** (Rajasekaran and Pai, 2003)

Part of a population starts from randomly generated mathematical functions (individuals) that are crossed with each other to give rise to a new generation that has characteristics of the parents plus a random component. These individuals are then evaluated according to how well they can estimate the response variable (e.g., COD) and those that exceed a certain threshold will spawn the next generation and so on, up to a finite number of generations.

- **Age-Layered Population Structure (ALPS)** (Hornby, 2006)

Developed by the researcher Gregory S. Hornby, from the NASA Ames Research Center, this technique allows the population of individuals to be segregated into different layers according to their age while introducing new randomly generated individuals in the youngest layers, which improve the results.

- **Offspring Selection (OS)** (Affenzeller and Wagner, 2005)

This technique solves one of the problems of AGCs, where with the passage of generations, individuals could become worse than their parents. In this case, new parents are randomly selected and a new offspring is generated and tested again to see whether or not it is better than its parents. This ensures that the next generation will be at least as good as the previous one.

The working range between 380 and 700 nm using just 33 LEDs is divided into 81 bands, where surrogates, absorbance and transmittance, are measured at each sample. So, a total of 162 variables were used as inputs for the calculation of the models. For the implementation of these models, the software HeuristicLab (Wagner et al., 2014), has been used. All models have been generated after removing outliers from each dataset using Box and Whisker plots.

Following the criteria of Osowski (1996), which recommends a 70-30% ratio, we have used a 66-34% ratio, in order to have more data to validate the models, by mean HeuristicLab interface.

2.5. Data storage platform

The results of the spectrophotometric analysis of the equipment are sent automatically to a web server where they will be available to be analyzed and compared with the information provided by the physical-chemical conditions analyzed in the samples. All this information is downloaded in a.CSV file containing both the spectral response of the samples analyzed with the equipment in Fig. 3 as well as the contaminant parameters characterized in the laboratory. This file is used for the generation of the estimation models.

3. Results and discussion

3.1. Mathematical models proposed from genetic algorithms

A series of mathematical models have been developed from the spectral response measured by the device to estimate the following contaminant parameters: COD, BOD_5, TSS, TP, and TN. All the models presented in this research work include both, raw wastewater samples (taken at the inlet of the WWTP) as well as treated water samples (taken at the outlet of the WWTP), without the need to subject the samples to chemicals or pre-treatments.

In order to clarify the presentation of the different models, Table 1 shows a summary of each one of the studied models, indicating technique, number of generations, Pearson coefficient achieved, mutation rate, number of maximum generations, as well as other distinctive aspects of each technique.

A total of 18 models are studied, in the case of considering all the samples – influent and effluent – together, and three further models considering just treated wastewater. Some of the algorithms calculated based on different techniques and configurations are available in the Supplementary Information section (SI). The following nomenclature has been used: *T* for Transmittance and *A* for Absorbance followed by the wavelength at which they have been measured. For instance, T_{385} corresponds to the Transmittance measured at 385 nm.

Stopping criteria of model generation adopted is based on both maximizing the fitness as well as controlling the size of the model, to avoid bloat and parsimony and overfitting of the data used. In this sense, the adopted values of depth and length of the symbolic regression models proposed are up to 5 and to 20, respectively, to avoid parsimony and overfitting. The depth is the distance from its root node to the furthest leaf node accounting for the number of divisions of the model. The length is the number of elements in the regression which is equal to the total number of nodes.

It should be noted that the tests performed for the different types of genetic algorithms have provided very similar estimation values. However, some models with a certain configuration (genetic algorithm technique, type of mathematical operations used, use of conditionals, etc.) have shown, for a specific pollutant parameter, better fits than

Table 1
Summary of models calculated in the research study.

Model (Eq.)	Param.	Type of water*	GA	Operat.**	Pearon's coefficient		Mut. rate	Max. gen.
					Test	Training		
1	COD	C	CGA	Arth	81%	91%	15%	50
S1	COD	C	OS	Arth	89%	90%	20%	50
S2	COD	C	ALPS	Arth	88%	89%	25%	500
4	BOD_5	C	OS	Arth	81%	85%	20%	25
S3	BOD_5	C	CGA	Arth	82%	83%	15%	50
S4	BOD_5	C	ALPS	Arth	79%	85%	25%	500
5	TSS	C	OS	Arth	86%	83%	15%	100
S5	TSS	C	ALPS	Arth and Log/Exp	80%	88%	25%	500
S6	TSS	C	CGA	Arth	84%	84%	15%	50
S7	TN	C	CGA	Arth	76%	73%	15%	50
S8	TN	C	OS	Arth	69%	75%	20%	25
S9	TN	C	ALPS	Arth	72%	75%	25%	500
S10	TP	C	ALPS	Arth	75%	70%	25%	500
S11	TP	C	CGA	Arth	72%	71%	15%	50
S12	TP	C	OS	Arth	67%	74%	20%	25
2	COD	T	CGA	Arth	53%	52%	15%	100
S13	BOD_5	T	CGA	Arth	19%	38%	15%	50
S14	TSS	T	OS	Arth and Po	43%	40%	20%	25

*C = combined raw and treated; T = treated.
**Arth = arithmetic; Po = power.

others (Table 1). Besides, these don't always select the exact same wavelengths. Therefore, in order to clarify the presentation of the results, in the main manuscript, only a single model is shown for each parameter, corresponding to the one whose configuration (technique, mathematical operations, mutation rate, etc.) has provided the best result. The rest of the models being calculated can be found in the Supplementary Information, in order to show which models, that can have selected different wavelengths, allow similar results to be obtained, a key aspect in the development of simpler and simpler equipment.

3.1.1. Chemical oxygen demand (COD)

For the calculation of models for the estimation of the chemical oxygen demand, a total of 636 data were used, after elimination of outliers. Of the total of samples, 420 were used for training, while the remainder were used for model evaluation. Equation (1) shows the model calculated using the CGA, which, of the algorithms calculated, is the one with the best relationship between the number of variables and the fit, with a Pearson coefficient of 81% for testing and 91% for training. This model was calculated for a maximum of 50 generations with a mutation rate of 15%.

$$COD_{(mg/l)}=(c_0{*}T_{380}+c_1{*}A_{425}{*}c_2{*}A_{405}{*}c_3{*}T_{485}){*}c_4+c_5$$
$$c_0=0.18988; c_1=0.43341; c_2=3.4384; c_3=1.7748; c_4=3917.6; c_5=-619.49 \quad (1)$$

The rest of the searched models, ALPS and OS, are presented in the supplementary information. Regardless of the genetic algorithm technique used, all the models calculated (Equations (S1-S2)) have a very similar.

The fit of the model is seen in Fig. 4, where it is observed that both the estimated values during training (orange squares) and during the test phase (grey triangles), are close to the values measured in the laboratory, organizing themselves around the fit line, with an interval of ±1 the standard deviation. This fit is also observed in the other calculated models (Figs. S1–S2).

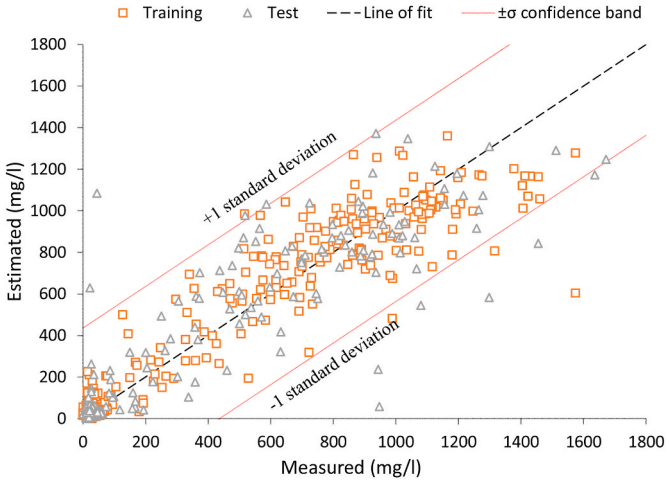


Fig. 4. Correlation plot between laboratory measured COD values and those estimated using the model of Equation (1), based on the CGA, for both training (squares) and test (test) data.

A model considering only treated water has also been fitted using the CGA technique. In comparison to previous models, this allows a better fit to be achieved - in terms of the treated wastewater – in order to have a real-time monitoring system for predicting values in the effluent from tertiary treatment for agricultural irrigation. A total of 287 data were used, after elimination of outliers. Of the total of samples, 189 were used for training, whilst the remainder were used for model evaluation. Equation (2) shows the model calculated, with a Pearson's coefficient of 53% for testing and 52% for training. This model was calculated for a maximum of 100 generations, with a mutation rate of 15%. The fit of the model is seen in Fig. 5.

$$COD_{(mg/l)}=\frac{c_0{*}T_{395}{*}c_1{*}A_{430}}{c_2{*}T_{656}{*}c_3{*}A_{660}}{*}\frac{c_4{*}T_{395}{*}c_5{*}A_{405}}{c_6{*}T_{656}{*}c_7{*}A_{697}}{*}c_8+c_9 \quad (2)$$
$$c_0=0.44446;\ c_1=2.579;\ c_2=1.6263;\ c_3=0.67182;\ c_4=0.44446;\ c_5=2.6203; c_6=1.6263;\ c_7=1.6582;\ c_8=37.119; c_9=-6.3138$$

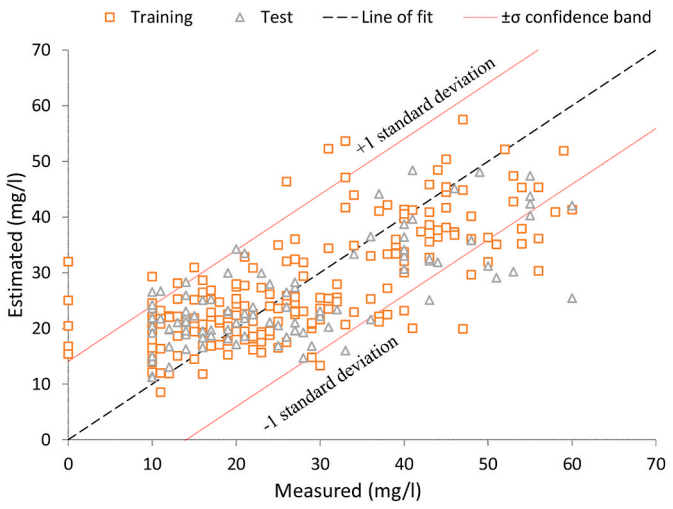


Fig. 5. Correlation plot between laboratory measured COD values of the treated effluent and those estimated using the model of Equation (2), based on the CGA, for both training (squares) and test (test) data.

In the Supplementary Information, can be found the models and adjustment corresponding to the treated wastewater in terms of BOD and TSS, Equation (S13-S14) and Figures (S16-S17).

It should be noted that, although the treated water only model gives slightly better fit to the measured values, if we make a comparison with the model of Equation (1), which is valid for both treated and raw water, we observe that its Pearson's correlation coefficient is much higher (89% for tests and 90% for training) than that obtained for a specific model for treated water (Equation (2)). This is due to the variability of the data.

The variability among the data studied influences the estimation of Pearson's coefficient, which is determined by Equation (3) (being reference values, x and estimated values, y).

$$Pearson's\ coefficient = \frac{\sigma_{xy}}{\sigma_x * \sigma_x} = \frac{\sum_i^n (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_i^n ((x_i - \bar{x}))^2} * \sqrt{\sum_i^n ((y_i - \bar{y}))^2}} \quad (3)$$

In the combined model (Equation (1)), we have used both raw water and treated water values for the COD calculation, i.e., we have mixed very high values with very low values. This makes the covariance (numerator Equation (3)) larger than if we only have treated water data (Equation (2)), since all the data are similar to each other. Moreover, in a dataset where the pollutant load values are low, as is the case for treated water, any small deviation of the estimate with respect to the reference values is accentuated. This is why the Pearson coefficient of Equation (2) is lower than that of Equation (1).

3.1.2. Biochemical oxygen demand (BOD_5)

A total of 637 data after outliers were used to calculate BOD_5, with a 66%–34% ratio between training and test data. Equation (4) shows the model obtained using the "Offspring Selection", which only makes use of simple arithmetic operations and presents a Pearson coefficient of 81% and 85% for the test and training data, respectively.

Table 2
Comparison of weights of the wavelengths used in the models: Equations (1), (4) and (5) and Equations (S7 and S10).

Spectrum	λ*	COD Eq (1)	BOD_5 Eq (4)	TSS Eq (5)	TN Eq (S7)	TP Eq (S10)
Violet	380	4% (T)				7% (T)
380–427 nm	400		25% (A)			
	405	39% (A)				
	420		25% (A)	5% (T)		10% (T)
			25% (T)	32% (A)		
	425	39% (A)			18% (T)	
Blue	430			16% (A)		
427–476 nm	465					15% (T)
	470				24% (T)	
Cyan	485	17% (T)				
476–497 nm						
Green	500				19% (T)	
497–570 nm	510					8% (T)
	515			44% (T)		
	520					12% (T)
	522					11% (T)
	532					18% (T)
	560					18% (T)
Yellow	574				20% (T)	
570–581 nm						
Red	627		5 (A)		11% (T)	
618–780 nm	650		19 (A)			
	656			3% (A)		
	660				9% (T)	

T: Transmittance, A: Absorbance.

The model was calculated after a maximum of 25 generations and a mutation rate of 20%. Fig. S3 shows the correlation between the BOD_5 values estimated by the model of Equation (4) and those obtained in the WWTP control laboratory, since almost all the values are concentrated with respect to the adjustment line with an interval of ±1standard deviation. The rest of the models (CGA and ALPS), are shown in Equations (S3-S4) within Supplementary Information.

3.1.3. Total suspended solids (TSS)

The determination of the models for the estimation of TSS was carried out on 603 data, of which 396 were used for training and the rest for validation, maintaining the 66%–34% ratio of the previous cases. The model of Equation (5) makes use of simple arithmetic operations calculated by the "Offspring Selection", which has a Pearson coefficient of 86% and 83% for test and training, respectively. If we look at the model from Equation (S5), for instance, we observe that it has a slightly better fit (al less in training data), than the model in Equation (5). This improvement in the fit is due to the use of exponential functions, but their use makes the model much more sensitive to small variations in the input values (transmittance and absorbance).

$$BOD_{5(mg/l)} = \frac{(c_0 * A_{650} - c_1 * A_{400}) - (c_2 + c_3 * A_{627})}{(c_4 * T_{420} - (c_5 - c_6 * A_{420}))} * c_7 + c_8 \quad (4)$$

$c_0 = 628.94$; $c_1 = -578.55$; $c_2 = 153.96$; $c_3 = 665.43$; $c_4 = 1099.7$; $c_5 = 635.49$; $c_6 = 909.53$; $c_7 = 476.58$; $c_8 = 139.74$

$$TSS_{(mg/l)} = \left(\frac{(c_0{*}A_{420} + (c_1{*}T_{515} - c_2{*}T_{420}))}{(c_3{*}A_{430} + (c_4{*}T_{515} - c_5{*}A_{656}))}\right){*}c_6 + c_7 \quad (5)$$

$c_0 = 2.1402;\ c_1 = 1.5927;\ c_2 = -0.36435;\ c_3 = 1.0293;\ c_4 = 1.6349;\ c_5 = -0.35326;\ c_6 = 1224.6;\ c_7 = -1497.8$

Fig. S6 shows the correlation between the TSS values estimated by the model of Equation (5) and those obtained in the WWTP control laboratory. In addition, the Pearson coefficients have remained practically constant in all models, which is observed in the high similarity of the correlation curves in Fig. S6 and Figs. S7–S8, where practically all of the estimated TSS values with respect to those measured in the laboratory are within the interval of ±1 standard deviation.

3.1.4. Total nitrogen (TN) and total phosphorus (TP) models

GA models searching for TN are presented in the SI information (Figs. S7–S9). TP is presented in Figs. S10–S12. The Pearson coefficient range achieved was 69–73% and 67–75% for the test data in the cases of TN and TP, respectively. The standard deviation interval was ±33.23 mg/L for nitrogen and ±4.29 mg/L for phosphorus. Although the results are less accurate than for the three previous pollutants searched, they continue to offer valuable information of wastewater characteristics. More details are presented in the SI information.

3.1.5. Comparison between parameters for combined models

Table 2 shows a comparison between the different models shown in Equations (1), (4) and (5), Equation (S7) for TN and Equation (S10) for TP, showing the different wavelengths (regressors) used and the weight that each of them adopted in their respective models until 100% is reached. In the Supplementary Information, in addition to the weight of each of the variables, the relative variable frequency for each regressor is shown for the optimal model researched.

As can be seen, all the models for all the pollutant parameters under study have certain wavelengths that are key to their characterization.

In the case of COD and BOD_5, a greater weight of wavelengths belonging to the violet region of the spectrum is observed (Youquan et al., 2010; Thomas et al., 1996), representing around 83% in the case of COD (adding the weight of 380 and 425 nm) and around 75% in the case of BOD5 (400–420 nm). The different models presented for COD and BOD5, make use of similar wavelength groups, although in the case of Equation (S1) (COD, OS), it is observed that the near-infrared region (Stephens and Walker, 2002), plays an important role, as it is the case of 625 nm (31%), something similar to what is observed in the model of Equation (4) for BOD5, where 650 nm represents a weight around 20%. All techniques (Classical Algorithm, ALPS, and Offspring Selection) resulted in a selection of similar spectral regions (Tables S1–S2 and Tables S3–S4), while in the BOD_5 model, from Equation (S4), there is a more homogeneous distribution of wavelengths.

The case of the TSS is slightly different: the wavelengths belonging to the green region of the spectrum (515 nm) had the greatest weight in the model of Equation (5), with a weight of 44%. This is in contrast to the other models calculated in Equation (S6) where a greater weight of the wavelengths close to infrared was observed, more specific 645 nm, an aspect that is also detailed in the literature (Jha and Garg, 2010). In fact, all the models presented in the Supplementary Information section (Equations (S5-S6)) make use of similar spectral regions, although the influence of the blue region of the spectrum was also observed, such as 420 nm in Equation (5), 410 nm in Equation (S5), or, 395 and 490 nm in Equation (S6), with a very similar distribution of weights.

This denotes one of the characteristics of genetic algorithms, the fact that there are infinite solutions where, although it is true that in some of them the best fit is obtained with the use of wavelengths close to the IR, this does not exclude that from other wavelengths (from the green region in this case); it is also possible to estimate the TSS with an accuracy comparable to that if the 600 nm wavelengths had been used. This feature allows to explore the development of simpler and more economical equipment.

The different models calculated for the estimation of Total Nitrogen from the spectral response Equations (S7-S9), which are presented in the supplementary information section, presented a greater predisposition to make use of shorter wavelengths (Shi et al., 2013; Reeves and Van Kessel, 2000), especially in the case Equations (S7 and S9). In fact, their impact, also understood as weight, increased as we approached the infrared region of the spectrum. In all the models, the wavelengths closest to infrared had weights between 12% and 56% Table (S7–S9).

On the other hand, Total Phosphorus presented a greater weight of wavelengths between 400 and 550 nm, as seen in Equations (S10-S12); this can be observed in Table (S10–S12).

In the case of suspended solids the light scatter and shading represents the major influence in the measurements along the whole studied spectrum in a quite homogeneous way (Van Den Broeke et al., 2006). Higher wavelengths, like near infra-red, presenting lower variability and deviation in the response, are among those preferred and chosen by the algorithms.

In the case of inorganic substances, whose sensitivity in the range of the spectrum studied in present work is almost constant (Carré et al., 2017; Sarraguça et al., 2009), is shown that the wavelengths closer to infrared and to the violet-blue region of the spectrum are commonly selected by the algorithms for these species. According to this, Sarraguça et al. (2009) verified that a NIR based model offered relative standard deviation comparable with those obtained using the UV–visible techniques.

Carré et al. (2017) conclude that lower wavelengths – 200 to 400 nm – show higher values of absorbance for the organic compounds, where the algorithm finally selected the 374 nm wavelength as the most representative for COD characterization, also considering the effect of supracolloidal and sedimentable matter. Jeong et al. (2007) established that the 300 nm wavelength provided more significant spectral values for COD wastewater characterization. In fact, this sensitivity increases as the wavelength decreases, but also experiences a greater variability in transmittance values, therefore, mathematical adjusted models prefer to use wavelengths that are close to the violet region of the visible spectrum, since in that region, the spectral response is more stable. This relationship of light adsorption is determined by the atomic composition of matter. When light passes through a compound, the energy of the light is used to make an electron in the outermost shells move from a bonding orbital to a non-bonding (empty) orbital. Depending on the matter through which the light passes, the jump between orbitals will be greater or smaller (Kalsi, 2007). Organic matter tends to have large jumps, which means that a lot of energy is needed to jump from one orbital to another, which results in higher absorbance and lower

Table 3
Comparison of error rates of the models of Equations (1), (4) and (5), Equation (S8), and Equation (S10).

Parameter	PBias (%)	$\overline{R^2}$
COD	0.066%	0.876
BOD_5	1.913%	0.837
TSS	−0.014%	0.842
TN	2.186%	0.740
P	0.947%	0.719

transmittance.

After analysing the different characterization models calculated, a comparison was made between the models based on the Percent Bias, *PBias,* which measures the average tendency of estimated values to be larger or smaller than reference one. (Equation (6) (Gupta et al., 1999; Moriasi et al., 2007), and adjusted R-squared ($\overline{R^2}$), (Equation (7)). Where n is the number of samples; and $X_{reference}$ and $X_{estimated}$ are the values of the polluting parameters (COD, BOD_5, TSS, P, TN and NO_3–N) obtained by the analytical methods used by the wastewater treatment plant and by the calculation models, respectively, and $\overline{X_{reference}}$ is the average of reference values.

$$PBias(\%) = \frac{\sum_i^n \left(X_{reference_i} - X_{estimated_i}\right)}{\sum_i^n X_{reference_i}} * 100, \quad (6)$$

$$\overline{R^2} = 1 - \frac{\frac{\sum_i^n \left(X_{reference_i} - X_{estimated_i}\right)^2}{n}}{\frac{\sum_i^n \left(X_{reference_i} - \overline{X_{reference}}\right)^2}{n}} \quad (7)$$

In order to clarify the presentation of the results, only those models belonging to Equations (1), (4) and (5), and Equation (S7) for TN, and Equation (S10), for TP, are shown in Table 3, since they had the best accuracy-number of variables relationship.

As can be seen in Table 3, for all models, the level of fit is high, above 83% in most cases, with the estimate of nutrients being somewhat lower, between 71 and 74%. Furthermore, as we can see, the PBias values are almost all positive, which indicates that the models, on average, provide slightly higher values than the reference values, although, in all cases, the deviation is less than 2.2%.

3.1.6. *Tertiary-treated reclaimed wastewater effluent specific models*

The combined samples models, including both raw and treated, previously presented are valid for the estimation of the pollutant load of wastewater. Even so, when the intention is to predict the treated wastewater physic-chemical parameters, a better fit is achieved when a specific model using just treated measures is found, representing a PBias between −0,23 and −1.178, that is, around 0.89–1.42% lower than the error recorded with the combined model, that is, the specific models for treated water provide values closer to the reference values than the combined models. These are presented above in Equation (2) and Fig. 5 for treated COD and Equations (S13, S14) and Figs. S15 and S16 for the treated BOD_5, and TSS models, respectively. In the case of the BOD_5 model, a low correlation coefficient of 27% for the test data of that model is explained by the low variability of the measured values, where most of the samples have very low values that complicate the adjustment. Even though this specific model for treated reclaimed wastewater allows to predict this parameter with an accuracy in the range of ±1 standard deviation, it adopted a low value of 2.85 mg/L when considering only treated wastewater. Tables S13 and S14 present the wavelengths used by the models.

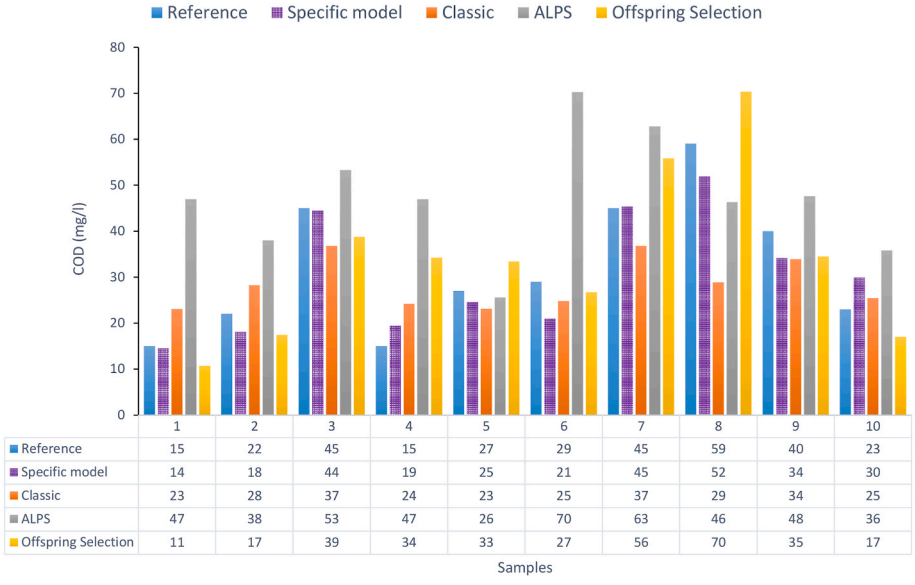
In order to illustrate more clearly how a specific model provides better results at low contaminant load levels, i.e. in treated water samples, Fig. 6 shows a comparison between the values obtained by the combined models and a specific model for treated water, for COD on 10 samples taken at random. As can be seen, although all the models provide very similar values (except ALPS in some cases), the specific model of Equation (2) is the one that presents an estimate closest to the reference value.

4. Conclusions

In the present research work, 18 different models (15 with raw and treated wastewater and 3 with only treated wastewater) based on symbolic regression have estimated the physical-chemical characteristics of the wastewater (COD, BOD_5, TSS, TP, and TN) from spectrophotometric analysis of samples between 380 and 700 nm, without the use of chemicals or pre-treatment of samples.

A key aspect when it comes to generating estimation models is the fact that these can fit all possible types of water, including influent and effluent to WWTP. For this reason, the study has focused on 43 different WWTP, an extensive sampling campaign of around 650 samples and the development of mathematical expressions that can be valid for estimating the pollutant load in both raw water samples (input to the wastewater treatment plant) as well as treated water (output from the plant). When using samples with such different properties (raw and treated water) such variability in the data makes it difficult to apply techniques such as multivariate linear regression analysis, as that type of analysis requires the data to follow a normal distribution, which is not possible when combining different types of water. For this reason, the use of genetic algorithms (GA) represents a very suitable option for the analysis of the pollutant load of wastewater, given that their evolutionary nature allows said restriction to be bypassed.

In the present research work, three different techniques based on genetic algorithms have been used by mean HeuristicLab software, to



	1	2	3	4	5	6	7	8	9	10
Reference	15	22	45	15	27	29	45	59	40	23
Specific model	14	18	44	19	25	21	45	52	34	30
Classic	23	28	37	24	23	25	37	29	34	25
ALPS	47	38	53	47	26	70	63	46	48	36
Offspring Selection	11	17	39	34	33	27	56	70	35	17

Fig. 6. Comparison of estimation between combined models and specific model for COD calculation.

generate estimation models for each of the pollutant parameters considered: Classical Genetic Algorithm (CGA), Age-Layered Population Structure (ALPS) and Offspring Selection. All the models, considering both influent and effluent, showed a level of correlation in terms of the R^2 correlation coefficient, above 83% for COD, BOD_5 and TSS, and over 71% on average for TN and P, when considering the 34% of the data used to test the models.

In the event that greater accuracy is needed, for instance to monitor tertiary-treated reclaimed wastewater, as is found in an important proportion of the 43 WWTP in the study, GA is also feasible to be used just with the treated wastewater, where more accurate models with lower intervals of standard deviations are achieved. Three of these models are presented for the COD, BOD_5, and TSS parameters. A lower correlation coefficient for the Pearson correlation coefficient of the test data is due to the data characteristics that do not detract from the accuracy and validity of the fitted model.

Each GA technique presents differences in the selection of the wavelengths when adjusting each of the different pollutant parameters. In the case of COD and BOD_5, although the violet wavelengths (380–425 nm) are of great importance in the characterization of the samples, other wavelengths have shown to be relevant for the adjustment and have also been selected by the algorithms. This explains, for example, why Equation (S4) (BOD_5), make use of near-red wavelengths whilst others do not. The study of COD usually focuses on analysis in the UV region of the spectrum, but this does not prevent useful information from being extracted from other wavelengths, such as near-infrared wavelengths (Equation (S1)). The proposed models, adjusted from VIS-spectrum spectroscopy surrogates, are mainly based in the scatter of the water matrix to the VIS wavelengths. In the case TN, around 40% of the weight of the adjustment comes from transmittance in the range of red/yellow light (570–780 nm) and green (497–570 nm), although violet region plays a role as well. In the case of TP, the wavelength distributions are very similar to the previous case (TN), where the green region seems to be the most significant one, however, in the model of Equation (S12), near-infrared region is also important, but to a lesser extent.

In this research work, a great variety of models with different configurations (number of variables, exponential/logarithmic operations, different generation criteria, among others) have been calculated in order to show the effect that each of these configurations has on the estimation of each pollutant parameter.

Likewise, one aspect that stands out in the use of GA as a modelling technique is that it is possible to describe a parameter from different parts of the spectrum. This is clearly observed in the calculation of COD, where it is observed that, as indicated by the literature and the research carried out in this manuscript, the wavelengths closer to the UV region of the spectrum have a greater significance in its calculation, which is why their weight in the models is very significant. However, and as shown in the model of Equation (S3) of the Supplementary Information, it is possible to obtain models that make exclusive use of other regions of the spectrum, as is the case of the red/near infrared region, achieving levels of adjustment very close to those obtained with the use of the violet region of the visible spectrum as was also concluded by Sarraguça et al. (2009), due to different variables like are the suspended solids and the complexity of the wastewater matrix, own an influence in this process of characterisation.

This is important, since it allows the development of simpler systems to characterize water quality from other parts of the spectrum, the generation of which is associated with a lower energy cost.

This research work shows an estimation of the pollutant load of wastewater and treated water that can be used to build a real-time pre-alert system that helps to visualize the evolution and shifts of the concentrations of pollutants. Spectrophotometry is presented as a tool for estimating water quality with acceptable levels of accuracy, without the need to use chemicals or pre-treatments.

The present work not only allows for an optimization of the water treatment processes by speeding up the analysis processes and enabling a faster response, but also for better protection of the environment through an early estimation of unauthorized discharges into the environment. Moreover, by synthesizing the estimation models into simple mathematical expressions, it opens the door to the development of low-cost, low-computational-capacity equipment which, based on spectrophotometric analysis, is capable of carrying out continuous monitoring of the sewage network in real time.

Author contribution

Daniel Carreres-Prieto: Investigation, Data curation, Formal analysis, Methodology, Software, Supervision, Validation, Writing - original draft, Writing - reviewed manuscript, **Juan T. García:** Investigation, Formal analysis, Methodology, Supervision, Validation, Writing - original draft, Writing - reviewed manuscript, **Fernando Cerdán-Cartagena:** Formal analysis, Funding acquisition, Methodology, Software, Supervision, Validation, Writing - reviewed manuscript, **Juan Suardiaz-Muro:** Formal analysis, Funding acquisition, Methodology, Supervision, Validation, Writing - reviewed manuscript, **Carlos Lardín:** Formal analysis, Field campaign, Funding acquisition, Methodology, Writing - reviewed manuscript.

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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to thank the help and availability received from the company Munuera Laboratories during the field campaign.

Appendix A

Table A1
Wastewater treatment plants used during the study.

	WWTP	Province	Population		Capacity (m3/a)		SST			COD			BOD_5		
			Served	Equivalent	Design (m3/a)	Current (m3/a)	In (mg/l)	Out (mg/l)	Perf (%)	In (mg/l)	Out (mg/l)	Perf (%)	In (mg/l)	Out (mg/l)	Perf (%)
1	**Abanilla**	Murcia	3.626	15.711	547.500	779.051	294	4	98.6	739	18	97.6	442	4	99.1
2	**Abarán**	Murcia	13.371	12.626	1.642.500	726.065	257	5	98.1	596	28	95.3	381	3	99.2
3	**Albudeite**	Murcia	1.296	1.043	365.000	45.738	205	8	96.1	756	31	95.9	499	4	99.2
4	**Alcantarilla**	Murcia	41.447	62.342	4.745.000	2.588.649	301	6	98.0	835	33	96.0	527	4	99.2
5	**Alguazas**	Murcia	9.102	37.629	5.475.000	1.076.650	371	4	98.9	1.208	22	98.2	765	3	99.6
6	**Archena**	Murcia	24.413	54.425	2.737.500	1.792.326	459	6	98.7	1.104	27	97.6	665	3	99.5
7	**Baños y Mendigo**	Murcia	218	344	173.375	21.521	352	10	97.2	591	36	93.9	350	3	98.9
8	**Barinas**	Murcia	756	1.982	197.100	73.884	390	4	99.0	906	21	97.7	588	4	99.3
9	**Barqueros**	Murcia	1.030	1.872	109.500	60.376	443	17	96.2	1.245	58	95.3	679	6	99.1
10	**Beniel Nueva**	Murcia	11.900	25.818	1.825.000	1.245.618	659	4	99.4	944	26	97.2	454	3	99.3
11	**Blanca**	Murcia	5.184	5.636	730.000	356.464	271	4	98.5	559	19	96.6	346	3	99.1
12	**Cabezo Beaza**	Murcia	176.223	173.924	12.775.000	9.031.284	470	18	96.2	924	52	94.4	422	12	97.2
13	**Cabezo de la Plata**	Murcia	104	358	44.165	44.165	248	11	95.6	1.024	30	97.1	702	3	99.6
14	**Calasparra**	Murcia	9.505	26.938	2.190.000	659.778	408	3	99.3	1.426	23	98.4	894	3	99.7
15	**Campos del Río**	Murcia	1.998	1.635	547.500	88.252	212	5	97.6	649	21	96.8	406	3	99.3
16	**Cañada de la leña**	Murcia	93	28	21.900	6.166	68	19	72.1	172	56	67.4	99	6	93.9
17	**Cañares/ Bronchos**	Murcia	442	195	1.350.500	54.371	805	2	99,7	3.253	2.751	37.5	156	3	98,0
18	**Casas Nuevas**	Murcia	152	220	73.000	8.170	847	6	99.3	1.190	28	97.6	590	4	99.3
19	**Ceutí Nueva**	Murcia	11.774	36.311	2.920.000	1.052.685	448	11	97.5	1.274	33	97.4	755	3	99.6
20	**Cieza**	Murcia	33.797	63.567	3.650.000	2.485.914	362	5	98.6	872	23	97.4	560	3	99.5
21	**Corvera**	Murcia	2.443	2.464	109.500	133.534	284	2	99.3	682	24	96.5	404	3	99.3
22	**El Cantón**	Murcia	66	506	18.250	18.250	324	18	94.4	1.058	34	96.8	608	5	99.2
23	**El Raal**	Murcia	15.940	23.706	2.737.500	3.950.557	151	7	95.4	240	21	91.3	131	4	96.9
24	**El Valle**	Murcia	194	464	511.000	58.089	378	5	98.7	343	17	95.0	175	3	98.3
25	**Fortuna**	Murcia	7.557	11.544	912.500	423.126	445	9	98.0	975	34	96.5	598	4	99.3
26	**Fuente Librilla**	Murcia	579	1.418	146.000	44.776	266	19	92.9	1.122	38	96.6	694	4	99.4
27	**Hacienda Riquelme**	Murcia	224	658	574.875	64.366	145	5	96.6	313	24	92.3	159	3	98.1
28	**Jumilla Nueva**	Murcia	24.588	70.595	4.380.000	1.739.564	825	3	99.6	1.761	24	98.6	889	3	99.7
29	**La Murta**	Murcia	91	545	44.165	15.378	353	4	98.9	1.271	28	97.8	776	3	99.6
30	**Lorqui**	Murcia	6.622	26.108	1.825.000	1.221.497	376	4	98.9	835	19	97.7	468	3	99.4
31	**Macisvenda**	Murcia	504	557	41.975	26.219	242	5	97.9	730	30	95.9	465	3	99.4
32	**Molina Norte**	Murcia	68.296	218.823	9.125.000	6.093.740	490	6	98.8	1.456	36	97.5	786	3	99.6
33	**Mosa Trajectum**	Murcia	144	285	642.400	42.568	177	3	98.3	270	15	94.4	147	3	98.0
34	**Mula Nueva**	Murcia	15.496	17.210	2.190.000	672.031	335	2	99.4	892	18	98.0	561	3	99.5
35	**Murcia Este**	Murcia	375.775	553.451	36.500.000	36.952.999	277	9	96.8	577	32	94.5	328	5	98.5
36	**Pliego**	Murcia	3.631	4.490	547.500	162.769	583	3	99.5	1.150	23	98.0	604	3	99.5
37	**Pol. Ind. Fortuna**	Murcia	0	584	65.700	22.945	797	30	96.2	855	68	92.0	557	13	97.7
38	**Santomera Norte**	Murcia	14.956	16.139	2.190.000	1.137.404	242	5	97.9	526	33	93.7	311	4	98.7
39	**Sucina Nueva**	Murcia	1.924	3.634	1.825.000	173.650	212	3	98.6	681	22	96.8	458	3	99.3
40	**Torres de Cotillas N.**	Murcia	19.996	53.597	4.380.000	1.602.051	641	9	98.6	1.281	21	98.4	733	3	99.6
41	**El Trampolín**	Murcia	149	158	73.000	13.930	329	37	88.8	396	33	91.7	248	4	98.4
42	**Yecla**	Murcia	31.876	43.586	2.920.000	1.648.354	490	7	98.6	1.015	20	98.0	579	3	99.5
43	**Yecla Raspay**	Murcia	97	109	18.250	8.385	147	5	96.6	477	18	96.2	286	4	98.6

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2022.133610.

References

Affenzeller, M., Wagner, S., 2005. Offspring selection: a new self-adaptive selection scheme for genetic algorithms. In: Adaptive and Natural Computing Algorithms. Springer, Vienna, pp. 218–221.

Augusto, D.A., Barbosa, H.J., 2000. Symbolic regression via genetic programming. In: Proceedings, vol. 1. IEEE, pp. 173–178. Sixth Brazilian Symposium on Neural Networks.

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. Chemometr. Intell. Lab. Syst. 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 Chem. 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 Sci. Technol. 76 (3), 633–641.

Carreres-Prieto, D., García, J.T., Cerdán-Cartagena, F., Suardíaz, J., 2019. Spectroscopy transmittance by LED calibration. Sensors 19, 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.

Cho, J.H., Sung, K.S., Ha, S.R., 2004. A river water quality management model for optimising regional wastewater treatment using a genetic algorithm. J. Environ. Manag. 73 (3), 229–242.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1999. Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. J. Hydrol. Eng. 4 (2), 135–143.

Harik, G.R., Lobo, F.G., Goldberg, D.E., 1999. The compact genetic algorithm. IEEE Trans. Evol. Comput. 3 (4), 287–297.

Holenda, B., Domokos, E., Rédey, A., Fazakas, J., 2007. Aeration optimization of a wastewater treatment plant using genetic algorithm. Optim. Control Appl. Methods 28 (3), 191–208.

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.

Huang, M., Ma, Y., Wan, J., Chen, X., 2015. A sensor-software based on a genetic algorithm-based neural fuzzy system for modeling and simulating a wastewater treatment process. Appl. Soft Comput. 27, 1–10.

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. Environ. Monit. Assess. 133 (1), 15–24.

Jha, S.N., Garg, R., 2010. Non-destructive prediction of quality of intact apple using near infrared spectroscopy. J. Food Sci. Technol. 47 (2), 207–213.

Jurga, A., Gemza, N., Janiak, K., 2017. A concept development of an early warning system for toxic sewage detection. In: E3S Web of Conferences, vol. 17. EDP Sciences, 00036.

Kalsi, P.S., 2007. Spectroscopy of Organic Compounds, 6th ed. New age international, New Delhi.

Korshin, G.V., Sgroi, M., Ratnaweera, H., 2018. Spectroscopic surrogates for real time monitoring of water quality in wastewater treatment and water reuse. Curr. Opinion Environ. Sci. Health 2, 12–19.

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

Melendez-Pastor, I., Almendro-Candel, M.B., Navarro-Pedreño, J., Gómez, I., Lillo, M.G., Hernández, E.I., 2013. Monitoring urban wastewaters' characteristics by visible and short wave near-infrared spectroscopy. Water 5 (4), 2026–2036.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50 (3), 885–900.

Osowski, S., 1996. Sieci Neuronowe W Ujęciu Algorytmicznym. Wydawnictwa Naukowo-Techniczne, Warsaw, Poland.

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

Reeves III, J.B., Van Kessel, J.S., 2000. Near-infrared spectroscopic determination of carbon, total nitrogen, and ammonium-N in dairy manures. J. Dairy Sci. 83 (8), 1829–1836.

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. Anal. Bioanal. Chem. 395 (4), 1159–1166.

Schmidt, A.F., Finan, C., 2018. Linear regression and the normality assumption. J. Clin. Epidemiol. 98, 146–151.

Shi, T., Cui, L., Wang, J., Fei, T., Chen, Y., Wu, G., 2013. Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. Plant Soil 366 (1), 363–375.

Skoog, D.A., Holler, F.J., Crouch, S.R., 2017. Principles of Instrumental Analysis. Cengage learning.

Stephens, A.B., Walker, P.N., 2002. Near-infrared spectroscopy as a tool for real-time determination of BOD5 for single-source samples. Trans. ASAE 45 (2), 451.

Thomas, O., Burgess, C. (Eds.), 2017. UV-visible Spectrophotometry of Water and Wastewater. Elsevier.

Thomas, O., Theraulaz, F., Agnel, C., Suryani, S., 1996. Advanced UV examination of wastewater. Environ. Technol. 17 (3), 251–261.

Torres, A., Bertrand-Krajewski, J.L., 2008. Partial Least Squares local calibration of a UV–visible spectrometer used for in situ measurements of COD and TSS concentrations in urban drainage systems. Water Sci. Technol. 57 (4), 581–588.

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

Wagner, S., Kronberger, G., Beham, A., Kommenda, M., Scheibenpflug, A., Pitzer, E., Affenzeller, M., 2014. Architecture and design of the HeuristicLab optimization environment. In: Advanced Methods and Applications in Computational Intelligence. Springer, Heidelberg, pp. 197–261.

Youquan, Z., Huimin, W., Ziyu, L., Yuchun, L., Shifu, F., 2010. Novel method for on-line water COD determination using UV spectrum technology. Chin. J. Sci. Instrum. 9 (1).