Contents lists available at ScienceDirect

Environmental Toxicology and Pharmacology

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

Environmentally-friendly technology for rapid identification and quantification of emerging pollutants from wastewater using infrared spectroscopy

C. Quintelas[a,*], A. Melo[a], M. Costa[a], D.P. Mesquita[a], E.C. Ferreira[a], A.L. Amaral[a,b]

[a] *CEB - Centre of Biological Engineering, University of Minho, 4710-057, Braga, Portugal*
[b] *Instituto Politécnico de Coimbra, ISEC, DEQB, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal*

ARTICLE INFO

Keywords:
FTIR spectroscopy
Emerging pollutants
kNN
PLS
OLS

ABSTRACT

The monitoring of emerging pollutants in wastewaters is nowadays an issue of special concern, with the classical quantification methods being time and reagent consuming. In this sense, a FTIR transmission spectroscopy based chemometric methodology was developed for the determination of eight of these pollutants. A total of 456 samples were, therefore, obtained, from an activated sludge wastewater treatment process spiked with the studied pollutants, and analysed in the range of 200 cm^{-1} to 14,000 cm^{-1}. Then, a k-nearest neighbour (kNN) analysis aiming at identifying each sample pollutant was employed. Next, partial least squares (PLS) and ordinary least squares (OLS) modelling approaches were employed in order to obtain suitable prediction models. This procedure resulted in good prediction abilities regarding the estimation of atrazine, desloratadine, paracetamol, β-estradiol, ibuprofen, carbamazepine, sulfamethoxazole and ethynylestradiol concentrations in wastewaters. These promising results suggest this technology as a fast, eco-friendly and reagent free alternative methodology for the quantification of emerging pollutants in wastewaters.

1. Introduction

The monitoring of emerging pollutants in wastewaters is a subject increasingly raising concern in the past few years. These compounds are challenging in terms of quantification due to the low concentrations, complex matrices and wide range of compounds with broad physical-chemical properties (Fedorova et al., 2014). The typical traditional methods applied to the quantification and monitoring of this class of compounds include solid-phase extraction coupled to liquid chromatography/tandem mass spectrometry (Li et al., 2018; Khan et al., 2012), high-performance thin-layer chromatography (HPTLC) (Shewiyo et al., 2012), and electrogenerated chemiluminescence biosensing methods (Zhang et al., 2019). The disadvantages of these methods are well-known and include being labour intensive, time and reagent consuming with expensive in both equipment and reagents. In this context, simple, reliable and rapid methods are needed to enable fast, reagent free, sensitive, and selective determination of emerging pollutants.

Recently, several fast, accurate, eco-friendly and reagent free methodologies based on infrared spectroscopy combined with chemometric analysis have gained visibility with special focus in the pharmaceutical industry (Noor et al., 2018). This technology is considered a powerful non-invasive and non-destructive analytical technique that allows measuring several parameters at once, being also sensitive to both chemical and physical attributes (Puchert et al., 2011). Furthermore, these authors also highlight among its main advantages being a green methodology, as it allows performing the analysis without the addition of chemicals (reagents free technology). Despite all these advantages, the application of this technique to the quantification of emerging pollutants in aqueous matrices is still a field under exploration.

The presence of emerging pollutants, as pharmaceuticals, personal care products, pesticides and others in aqueous systems occurs all over the world. Fram and Belitz (2011) investigated the occurrence of pharmaceutical compounds in groundwater used for public drinking-water supply in California. These authors analysed 1231 samples of groundwater and found pharmaceuticals compounds in 2.3 % of these samples. The pharmaceuticals found in higher concentration were paracetamol, caffeine, carbamazepine, codeine, p-xanthine, sulfamethoxazole, and trimethoprim. Moreover, pesticides and volatile organic compounds were also present in significant amounts in these samples. Zhou et al. (2016) analysed twelve selected pharmaceuticals including antibiotics, analgesics, antiepileptics and lipid regulators in

[*] Corresponding author.
E-mail address: cquintelas@deb.uminho.pt (C. Quintelas).

https://doi.org/10.1016/j.etap.2020.103458
Received 10 April 2020; Received in revised form 14 July 2020; Accepted 20 July 2020
Available online 22 July 2020
1382-6689/ © 2020 Elsevier B.V. All rights reserved.

water samples collected from 18 sampling sections along the three main urban rivers in Yangpu District of Shanghai, China, and found a number of emerging pollutants with relevant concentrations such as ibuprofen, carbamazepine, salicylic acid, and azithromycin. The occurrence of estrogens is also reported by several authors. Wang et al. (2015) analysed the occurrence and evaluated the ecological risk of five estrogens, estrone, 17β-estradiol (E_2), estriol, ethynylestradiol (EE_2), and bisphenol A in water, sediments and biota in Northern Taihu Lake (China). The authors found that estrone, E_2 and bisphenol A were widely distributed in water, while estriol and EE_2 were less frequently detected. All the target estrogens were widely found in sediments and biota.

Kosonen and Kronberg (2009) analysed the occurrence of antihistamines in sewage waters and in recipient rivers and found that three compounds, cetirizine, acrivastine and fexofenadine were detected in both influent and effluent wastewater samples at ng/L concentrations, while loratadine, desloratadine and ebastine could not be detected in the samples. As expected, the results showed that the level of antihistamines in wastewater is at the highest in Spring due to the outbreak of allergic reactions caused by high plant pollen amounts in the air. More recently, Kristofco and Brooks (2017) performed a global antihistamines survey in the environment, with special focus on the occurrence and hazards in aquatic systems. These authors analysed more than one hundred literature papers, mainly from Asia-Pacific, European and North American geographic regions, and found the occurrence of 24 antihistamines, including desloratadine, in water, sediment and tissue. It is important to highlight that monitoring data from Africa and South America was largely lacking, inferring that a larger number of antihistamines occurrence is to be expected worldwide.

Herbicides are one of the emerging pollutants of concern. Despite the fact that atrazine was banned in several countries, due to findings of atrazine concentrations in ground and drinking waters exceeding legislation values, monitoring of atrazine concentrations in the groundwater since then provides information about the resilience of this compound in groundwater (Vonberg et al., 2014). The monitoring data obtained by these authors shows that even 20 years after the atrazine ban, groundwater concentrations remain on a level close to the threshold value of 0.1 μg L^{-1} without any considerable decrease.

Given the amount of information confirming the presence of emerging pollutants in the environment, namely personal care products, antibiotics, antihistamines, sulphonamides, anticonvulsants, analgesic, antipyretic, pesticides and herbicides, among others, it is increasingly important to develop fast, easy and green methods to quantify its presence in aqueous systems. In this context, this study focus on the determination of seven pharmaceutical compound concentrations, an antipyretic (paracetamol – PRC), an antihistamine (desloratadine – DSL), an anti-inflammatory (ibuprofen – IBU), two estrogens (β-estradiol – E_2, ethynylestradiol – EE_2), an anticonvulsant (carbamazepine – CRB), an antibiotic (sulfamethoxazole – SMX) and a herbicide (atrazine – ATR), using Fourier-transform infrared (FTIR) spectroscopy, in aqueous solutions. FTIR spectroscopy was already tested for the determination of organic pollutants in wastewater: in 2004, Michel et al. (2004) developed a prototype mid-infrared sensor system for the determination of organic pollutants, as trichloroethylene, tetrachloroethylene and dichlorobenzene; more recently, Gowen et al. (2012) analysed the state of the art for the application of vibrational spectroscopy for analysis of water for human use and in aquatic ecosystems and found that despite all the promising results proved by the literature more works are need to performed before these techniques can be implemented as water quality monitoring tools.

The quantification of pharmaceuticals in wastewaters by FTIR, combined with chemometric analysis, represents a very promising fast, eco-friendly and reagent free alternative to the traditional methods for screening and estimation of emerging pollutants. To that effect, the present report presents a new chemometric approach, based on partial least squares (PLS) and ordinary least squares (OLS).

2. Materials and methods

2.1. Sample preparation

In batch biodegradation experiments, a 1.0 L glass beaker, containing 0.3 L of activated sludge suspension (3 g L^{-1}), was spiked with different initial concentrations of emerging pollutants, namely desloratadine (DSL), paracetamol (PRC), ibuprofen (IBU), β-estradiol (E_2), ethynylestradiol (EE_2), carbamazepine (CRB), sulfamethoxazole (SMX) and atrazine (ATR), within the range of mg L^{-1}. Each solution was prepared separately. The initial concentration of each assay is present, as Supplementary Material, in Table S1. The experiments were performed at room temperature and the agitation was kept constant at 150 rpm. A synthetic medium was fed to the system in the beginning of each experiment accordingly with Quintelas et al. (2019). Aqueous samples (1 mL) were taken at designated time intervals (within a 48 h interval) and analysed by ultra-high-performance liquid chromatography (UHPLC).

2.2. UHPLC analysis

The chromatographic analysis was performed using a *Shimadzu Corporation* apparatus (Tokyo, Japan) consisting of a UHPLC equipment (*Nexera*) with one multi-channel pump (*LC*-30AD), an autosampler (*SIL-30AC*), an oven (*CTO-20AC*), a diode array detector (*M-20A*) and a system controller (*CBM-20A*) with built-in software (*LabSolutions*).

For the PRC quantification, a *Kinetex2.6 u EVO C18* column (150×4.6 mm i.d.) supplied by *Phenomenex, Inc*. (CA, USA) was used. The mobile phase was 0.1 % phosphoric acid in water (pump A) and 0.1 % phosphoric acid in acetonitrile (pump B). Starting mobile phase composition was 95 % A, decreased to 5% A in 9 min, increased again to 95 % (9.01 min) and remaining in this percentage for 3 min. The flow rate was 1.8 mL min^{-1}. The samples were monitored by a diode array detector from 190 to 400 nm, and chromatograms were extracted at 248 nm. Column oven was set at 50 °C and the injection volume was 5 μL.

For the ATR quantification, a *Kinetex5 u EVO C18* column (150×4.6 mm i.d.) supplied by *Phenomenex, Inc*. (CA, USA) was used. The mobile phase was water (pump A) and acetonitrile (pump B). An isocratic method was employed with 15 % A and 85 % B. The flow rate was 1.0 mL min^{-1}. The samples were monitored by a diode array detector from 190 to 400 nm, and chromatograms were extracted at 220 nm. Column oven was set at 25 °C and the injection volume was 30 μL.

The same column was used for the quantification of DSL, with a mobile phase of potassium dihydrogen phosphate (0.05 M; pH 3) (pump A), acetonitrile (pump B) and methanol (pump C). An isocratic method was employed with 45 % A, 48 % B and 7% C. The flow rate was 0.8 mL min^{-1}. The samples were monitored by a diode array detector from 190 to 400 nm, and chromatograms were extracted at 247 nm. Column oven was set at 25 °C and the injection volume was 12 μL.

The quantification of IBU, CRB, E_2, EE_2 and SMX concentrations was performed accordingly to Quintelas et al. (2019). The standard errors for the UHPLC measurements were 0.023 mg L^{-1}, 0.078 mg L^{-1}, 0.048 mg L^{-1}, 0.029 mg L^{-1}, 0.016 mg L^{-1}, 0.122 mg L^{-1}, 0.043 mg L^{-1}, and 0.077 mg L^{-1}, respectively for ATR, DSL, PRC, IBU, SXF, CARB, EE_2 and E_2 and the values of R^2 for the model (calibration) curves were around 1 for all compounds.

2.3. Infrared scanning

The Fourier transform infrared (FTIR) spectra was recorded on a FTIR/FT-NIR spectrometer (*FTLA 2000, ABB, Thermo Electron Corporation*) equipped with an indium-gallium-arsenide (InGaAs) detector, from 14,000 to 200 cm^{-1}, in transmittance mode using a flow cell with a 0.7 mm pathlength. For each sample, 64 scans were made with a spectral resolution of 8 cm^{-1} and then averaged. Samples were temperature equilibrated at 23 °C (during approximately 3 min) in the instrument before scanning. The integration time was adjusted until the peaks at 8333–9091 cm^{-1} for FTIR were close to 60,000 intensity units. Grams / AI software (*Thermo Electron Corporation*) was used for spectrometer configuration, control, and data acquisition. Distilled water was used as background. A typical obtained FTIR spectrum (raw and pre-processed with SNV, MSC, 1D and 2D) is presented as Supplementary material (Figure S1).

2.4. Chemometric analyses

The ATR, DSL, IBU, CRB, SMX, E_2, EE_2 and PRC concentrations, monitored throughout the different experiments time length, were used as the Y dataset in the employed chemometric analyses, whilst the X dataset consisted of the collected FTIR spectra (ranging from 14,000 to 200 cm^{-1}). The following chemometric techniques were employed to the dataset sequentially: i) k-nearest neighbour (kNN) analysis to allocate each sample within its corresponding pollutant from the IR raw dataset; ii) preprocessing of the FTIR raw dataset by means of standard normal variate (SNV), multiplicative scatter correction (MSC), 1st derivative (1D) and 2nd derivative (2D); and iii) partial least squares (PLS) and ordinary least squares (OLS) modelling to obtain the predictive models regarding each studied pollutant.

2.4.1. k-nearest neighbour (kNN)

A k-nearest neighbour (kNN) analysis was next performed to the entire X dataset (FTIR wavelengths values) in order to validate the samples allocation within the corresponding studied pollutant. For that purpose. one third of the preallocated samples was chosen as the (allocation step) validation dataset, and the remaining two thirds as the (allocation step) training samples. A total of 1000 random validation and training (calibration) samples combinations were screened for robustness purposes. This methodology assumes that all samples correspond to points in an n-dimensional space, with its nearest neighbours defined in terms of the corresponding distance metric. Each sample point of the (allocation step) validation dataset was next compared with its three nearest neighbours in the training dataset, with the distance metric being the Euclidean distance. Furthermore, the majority rule was used to decide how to classify each validation sample point. Further details regarding the kNN technique can be found in Cover and Hart (1967) and Mitchell (1997).

2.4.2. Preprocessing of the IR raw dataset

Four different preprocessing methods were employed to the FTIR raw dataset; standard normal variate (SNV), multiplicative scatter correction (MSC), 1st derivative (1D) and 2nd derivative (2D). The SNV was performed on each collected sample spectrum by mean centring (removal of the spectrum average value), followed by scaling (division by the spectrum standard deviation). Further details regarding the SNV technique can be found in Barnes et al. (1989). The MSC was performed by shifting and rotating each sample spectrum to fit, as close as possible, to the data average spectra. This was achieved by an ordinary least squares first-degree polynomial, with the correction depending on the entire dataset average spectra. Further details regarding the MSC technique can be found in Martens and Naes (1989). Also 1st (1D) and 2nd (2D) derivatives were employed to the FTIR raw dataset. Each of these preprocessing steps was then fed to the OLS methodology.

2.4.3. Ordinary (OLS) and partial (PLS) least squares regression

The OLS analysis is a linear least squares method for estimating unknown parameters (Y data), from a set of explanatory variables (X dataset), in a linear regression model. In this sense, the OLS calculates the explanatory variables coefficients by minimizing the sum of the residuals (differences between observed and predicted Y values) squares in a given dataset. On the other hand, the PLS analysis constructs latent variables (LVs) from the original X dataset in new (and orthogonal) spaces, maximizing the captured predictive power of the X-space with regard to the Y-space. Further details regarding the OLS and PLS techniques can be found in Wold (1966) and Einax et al. (1997).

With the purpose of predicting the individual pollutants concentrations (Y data) from the FTIR wavelength dataset (X dataset), the PLS analysis based methodology firstly employed standard normal variate (SNV) and cross-validation (CV) tools to remove undesirable X data matrix variations and test its predictive significance. Two different methodologies were next employed: a) using the raw dataset [M1]; b) using an iterative method by the arrangement of the wavelength values according to weight similarity [M2]. The second methodology consisted of the following sequential steps: i) determination of each wavelength weights for the entire wavelength range in an initial PLS analysis; ii) arrangement of the wavelength values according to weight similarity; and iii) final PLS analysis with the averaged wavelength values. Both methodologies are described further in Quintelas et al. (2019).

In addition, an OLS methodology was also employed to create a linear model fit of the individual pollutant concentrations (Y data) from the raw dataset OLS [raw] and the four different pre-processed FTIR, namely SNV – OLS [SNV], MSC – OLS [MSC], 1D – OLS [1D] and 2D – OLS [2D] data (X datasets). A forward selection method was employed for the choice of the selected wavelengths, selecting the best variable first, next finding the second best, and so on until the obtained model ceased to improve or reached the maximum number of components allowed. Care was taken to accept solely results with probability values (p-values) for all coefficients (including the intercept) below 0.05, that is, statistically significant for a level of significance (α) of 0.05.

A total of 5000 PLS and OLS possible random validation and training (calibration) samples combinations, for each employed methodology and dataset, were screened to select the most unbiased training and validation datasets (regarding the calibration step). During this iterative procedure, the samples were randomly divided into two groups, the training set with two thirds of the samples, and the (external) validation set with the remaining one third of the samples in each iteration. The internal model validation was performed by cross-validation with the training dataset.

The conducted study was conducted in a twofold manner in order to i) strictly obey to the ASTM E1655 (2012) standard regarding the training (calibration) dataset or ii) obey to the ASTM E1655 standard considering the global (training and validation) dataset rather than solely the training (calibration) dataset. This standard recommends that the number of model components (k) should be no larger than one sixth of the number of calibration samples (n) [$n > 6\ (k + 1)$]. In accordance, whereas the first procedure resulted in a set of model parameters (wavelengths or PLS components) up to one sixth of the training samples alone (designated as procedure 1 – [P1]), the later resulted in a set of parameters up to one sixth of the global (training and validation) samples (designated as procedure 2 – [P2]), for each pollutant. The root mean square error (RMSE) value was used to define the best set of parameters following the above conditions.

All the above analyses were performed in Matlab 7.11 (The Mathworks, Inc. Natick, USA).

3. Results and discussion

3.1. Analytical data

The concentrations minimum, maximum, range and standard deviation (STD) values, number of samples in the training and validation sets, and number of model components in P1 and P2 procedures, for desloratadine (DSL), paracetamol (PRC), ibuprofen (IBU), β-estradiol (E_2), ethynylestradiol (EE_2), carbamazepine (CRB), sulfamethoxazole (SMX) and atrazine (ATR), in the experiments are presented in Table 1. A total of 456 samples were initially collected, 60 samples for each pollutant, with the exception of IBU with 36 samples. From this initial

Table 1
Minimum, maximum, range, standard deviation (STD), number of samples in the training and validation sets and number of model components in P1 and P2 procedures, for each studied pollutant.

	Min. (mg L^{-1})	Max. (mg L^{-1})	Range (mg L^{-1})	STD (mg L^{-1})	training set	validation set	P1 comp.	P2 comp.
ATR	1.17	19.30	18.13	3.41	40	20	7	10
DSL	0.68	5.51	4.83	1.24	30	14	5	7
PRC	0.29	9.01	8.72	2.24	38	19	6	10
E_2	0.09	2.75	2.67	0.79	39	19	7	10
EE_2	0.44	6.93	6.50	1.86	40	20	7	10
CRB	0.56	8.12	7.56	2.10	40	20	7	10
IBU	0.00	1.99	1.98	0.45	24	12	4	6
SMX	0.31	7.60	7.29	1.58	40	20	7	10

dataset, the samples presenting a concentration bellow the analytical instrumentation sensitivity level were discarded, resulting in a final set of 435 samples. The samples were then divided into two groups, the training set, with two thirds of the samples, and the validation set, with the remaining one third of the samples.

3.2. kNN results

In order to allow the concentration prediction by individually tailored PLS and OLS analyses, first the collected FTIR data was subjected to a k-nearest neighbour (kNN) analysis with the aim of identifying each dataset sample within the corresponding studied pollutant. For that purpose, the entire FTIR wavelength dataset was used for a total of 1000 random validation and training (calibration) sets, representing the samples allocation step. This resulted in a global identification percentage, for the allocation step validation dataset, of 98.6 % with solely 1.4 % of the samples misclassified, a figure is presented as Supplementary material (Figure S2).

Therefore, and given the high identification ability for each studied pollutant obtained by the kNN analysis, regarding the allocation step validation dataset, it could be concluded that, upon this step, individual PLS and OLS analyses could be performed to model each pollutant concentration. It should be emphasized, however, that the employed emerging pollutants identification methodology was only tested for the case where solely one of the studied pollutants was present in a sample. Further extension to samples presenting two or more of the studied emerging pollutants should yet be studied.

3.3. PLS and OLS results

For each method (OLS [raw], OLS [SNV], OLS [MSC], OLS [1D], OLS [2D], PLS [M1] and PLS [M2]) and studied pollutant, the regression equation, coefficient of determination (R^2), number of FTIR wavelengths (λ), number of PLS components, root mean square error (RMSE, in percentage of the studied range) and residual predictive deviation (RPD) values were determined. Table 2 presents the obtained results for a maximum number of model parameters (wavelengths or PLS components) up to one sixth of the training samples alone [P1], whereas Table 3 presents the obtained results for a maximum of up to one sixth of the global (training and validation) dataset [P2], for each pollutant.

The regression equation and R^2 values are presented for the global (training + validation) dataset, whereas the RMSE and RPD are presented both for the global dataset and for the validation dataset. An RPD parameter, i.e. the ratio between the population standard deviation (SD) and the prediction standard error of cross validation (SECV), larger than 3 is recommended for screening purposes (Fearn, 2002).

The first methodology employed for modelling the pharmaceuticals concentrations was the PLS analysis using both the raw dataset [M1] and the iterative method [M2]. The analysis of the prediction ability, revealed that the pollutants that presented RPD values above 3, were the ATR, CRB and IBU for the [M1] methodology and the ATR, DSL, CRB and IBU for the [M2] methodology, both for the global (training + validation) and validation datasets. Comparing these two methodologies, the best results (higher R^2 and RPD, and lower RMSE values) were obtained by the [M2] methodology for the ATR (R^2 of 0.989), DSL (R^2 of 0.912) and CRB (R^2 of 0.962), and by the [M1] methodology for the IBU (R^2 of 0.988) prediction. Care should be taken, however, when analysing the IBU results given that the [M2] methodology failed to go beyond the third PLS component, whereas the [M1] results were obtained by the use of four PLS components.

Under the P1 procedure, the PLS based methodologies allowed for an adequate prediction of half of the studied pharmaceuticals, with RMSE % values ranging from under 2% (ATR) to just below 7.5 % (DSL), and from just above 2% (ATR) to slightly under 8.5 % (DSL), for the global and validation datasets respectively. Furthermore, the [M2] methodology resulted, in most cases, in best prediction abilities than the [M1] methodology, proving its value, regarding the classical PLS [M1], facing a reduced set of components.

Regarding the OLS methodology, apart from the use of the raw dataset, OLS [raw], four different X datasets preprocessing methodologies were employed, namely the standard normal variate, OLS [SNV], multiplicative scatter correction, OLS [MSC], 1st derivative, OLS [1D] and 2nd derivative, OLS [2D]. The analysis of the prediction ability, revealed that the pharmaceuticals that presented RPD values above 3, were the ATR, PRC and CRB for the [raw], the ATR and DSL for the [SNV], the ATR, PRC and CRB for the [MSC], the ATR, PRC, E_2 and CRB for the [1D] and the ATR, PRC, E_2 and CRB for the [2D] methodologies, both for the global (training + validation) and validation datasets (with the exception of E_2 for [1D] regarding the validation dataset). Comparing these five methodologies, the best results (higher R^2 and RPD, and lower RMSE values) were obtained by the [SNV] methodology for the DSL (R^2 of 0.898), by the [1D] methodology for the ATR (R^2 of 0.991) and by the [2D] methodology for the PRC (R^2 of 0.976), E_2 (R^2 of 0.933) and CRB (R^2 of 0.968) predictions.

Under the P1 procedure, the OLS based methodologies allowed for an adequate prediction of five of the studied pharmaceuticals, with RMSE % values ranging from under 2% (ATR) to just under 8% (DSL), and from 1.5 % (ATR) to slightly above 8.5 % (E_2), for the global and validation datasets respectively. Furthermore, the [2D] methodology emerged as the one presenting the best prediction abilities wthin the OLS based methodologies facing a reduced set of components.

Taking into account both the PLS and OLS based methodologies, the IBU and DSL were best predicted by the PLS, and the ATR, PRC, E_2 and CRB by the OLS, under the P1 procedure. However, two of the studied pharmaceuticals, namely EE2 and SMX failed to be adequatelly predicted by this procedure, not surpassing an R^2 of 0.649 and 0.787, and an RPD (for the global dataset) of 1.77 and 2.88, respectively.

With the increase of the PLS and OLS model components, a general increase in the prediction ability occurred, as could be expected. However, it was observed that the adequateness of the different methodologies varied quite differently among each other.

Table 2
Regression equation, R^2, FTIR wavelengths (λ), PLS components, RMSE and RPD values for each studied pollutant and P1 procedure (glb – global and val – validation).

		Regression glb	R^2 glb	# IR λ	# PLS comp	RMSE glb %	RMSE val %	RPD glb	RPD val
ATR	PLS [M1]	y = 0.976 x	0.943		7	4.42	4.29	4.25	4.38
	PLS [M2]	y = 0.994 x	0.989		7	1.97	2.18	9.56	8.60
	OLS [raw]	y = 0.993 x	0.984	7		2.35	1.50	8.00	**12.56**
	OLS [SNV]	y = 0.963 x	0.917	7		5.29	6.05	3.55	3.11
	OLS [MSC]	y = 0.990 x	0.984	7		2.36	1.78	7.95	10.55
	OLS [1D]	y = 0.993 x	0.991	7		1.73	1.52	**10.87**	12.38
	OLS [2D]	y = 0.989 x	0.989	7		1.94	1.94	9.69	9.66
DSL	PLS [M1]	y = 0.846 x	0.730		5	12.83	16.12	2.00	1.59
	PLS [M2]	y = 0.953 x	0.912		5	7.39	8.32	**3.47**	3.08
	OLS [raw]	y = 0.866 x	0.727	5		12.17	14.66	2.11	1.75
	OLS [SNV]	y = 0.951 x	0.898	5		7.95	8.28	3.23	**3.10**
	OLS [MSC]	y = 0.883 x	0.762	5		11.99	12.97	2.14	1.98
	OLS [1D]	y = 0.884 x	0.852	5		9.77	14.26	2.63	1.80
	OLS [2D]	y = 0.901 x	0.841	5		9.98	12.82	2.57	2.00
PRC	PLS [M1]	y = 0.930 x	0.864		6	9.42	10.21	2.73	2.52
	PLS [M2]	y = 0.847 x	0.546		6	15.54	16.28	1.66	1.58
	OLS [raw]	y = 0.966 x	0.953	6		5.51	5.75	4.67	4.47
	OLS [SNV]	y = 0.867 x	0.695	6		13.12	12.30	1.96	2.09
	OLS [MSC]	y = 0.961 x	0.954	6		5.58	5.32	4.61	4.83
	OLS [1D]	y = 0.970 x	0.966	6		4.78	3.49	5.38	7.37
	OLS [2D]	y = 0.981 x	0.976	6		4.03	3.12	**6.38**	**8.23**
E_2	PLS [M1]	y = 0.905 x	0.858		7	10.97	13.80	2.69	2.13
	PLS [M2]	y = 0.918 x	0.762		7	12.73	17.24	2.31	1.71
	OLS [raw]	y = 0.910 x	0.799	7		12.23	13.83	2.41	2.13
	OLS [SNV]	y = 0.922 x	0.866	7		10.63	10.17	2.77	2.90
	OLS [MSC]	y = 0.903 x	0.822	7		12.40	14.60	2.38	2.02
	OLS [1D]	y = 0.942 x	0.918	7		8.40	10.09	3.51	2.92
	OLS [2D]	y = 0.955 x	0.933	7		7.53	8.62	**3.91**	**3.42**
EE_2	PLS [M1]	y = 0.848 x	0.549		7	16.63	19.78	1.72	1.45
	PLS [M2]	y = 0.855 x	0.649		7	16.18	21.98	**1.77**	1.30
	OLS [raw]	y = 0.814 x	0.488	7		18.58	21.63	1.54	1.32
	OLS [SNV]	y = 0.808 x	0.535	7		18.54	28.22	1.55	1.02
	OLS [MSC]	y = 0.817 x	0.478	7		18.67	23.00	1.53	1.25
	OLS [1D]	y = 0.850 x	0.611	7		16.39	19.07	1.75	**1.50**
	OLS [2D]	y = 0.817 x	0.547	7		17.67	24.74	1.62	1.16
CRB	PLS [M1]	y = 0.964 x	0.960		7	5.70	7.11	4.87	3.90
	PLS [M2]	y = 0.984 x	0.962		7	5.30	4.92	5.24	**5.64**
	OLS [raw]	y = 0.965 x	0.947	7		6.39	7.37	4.34	3.76
	OLS [SNV]	y = 0.910 x	0.774	7		12.14	16.59	2.29	1.67
	OLS [MSC]	y = 0.961 x	0.947	7		6.46	7.73	4.30	3.59
	OLS [1D]	y = 0.965 x	0.960	7		5.66	6.98	4.90	3.97
	OLS [2D]	y = 0.971 x	0.968	7		4.98	4.98	**5.57**	5.58
IBU	PLS [M1]	y = 0.9599 x	0.988		4	2.62	2.72	**8.72**	**8.40**
	PLS [M2]	y = 0.886 x	0.927		3	6.32	6.32	3.61	3.61
	OLS [raw]	y = 0.678 x	0.689	4		12.69	14.63	1.80	1.56
	OLS [SNV]	y = 0.654 x	0.680	4		12.86	15.18	1.77	1.50
	OLS [MSC]	y = 0.678 x	0.670	4		12.70	14.60	1.80	1.56
	OLS [1D]	y = 0.811 x	0.760	4		10.65	5.53	2.14	4.13
	OLS [2D]	y = 0.625 x	0.653	4		13.47	16.13	1.70	1.42
SMX	PLS [M1]	y = 0.921 x	0.782		7	9.31	9.70	2.32	2.23
	PLS [M2]	y = 0.933 x	0.876		7	7.51	8.63	**2.88**	**2.51**
	OLS [raw]	y = 0.907 x	0.815	7		9.00	11.52	2.41	1.88
	OLS [SNV]	y = 0.915 x	0.698	7		10.80	13.43	2.00	1.61
	OLS [MSC]	y = 0.902 x	0.810	7		9.14	11.38	2.37	1.90
	OLS [1D]	y = 0.933 x	0.854	7		8.04	10.12	2.69	2.14
	OLS [2D]	y = 0.934 x	0.854	7		8.02	9.44	2.70	2.29

* In all cases the obtained model *p-value* was below 0.001.

The analysis of the PLS prediction ability, revealed that all the pharmaceuticals presented RPD values above 3 for the [M1] methodology (except for the validation dataset for the DSL) whereas the EE_2 failed to achieve that goal regarding the [M2] methodology, both for the global (training + validation) and validation datasets. Comparing these two methodologies, the best results (higher R^2 and RPD, and lower RMSE values) were obtained by the [M2] methodology for the ATR (R^2 of 0.996), DSL (R^2 of 0.989) and CRB (R^2 of 0.962), and by the [M1] methodology for the PRC (R^2 of 0.989), E_2 (R^2 of 0.965), EE_2 (R^2 of 0.917), IBU (R^2 of 0.996) and SMX (R^2 of 0.978) predictions. Again, care should be taken when analysing the IBU results given that the [M2] methodology failed to go beyond the third PLS component, whereas the [M1] results were obtained by the use of six PLS components.

Under the P2 procedure, the PLS based methodologies allowed for an adequate prediction of all of the studied pharmaceuticals, with RMSE % values ranging from slightly above 1% (ATR) to 5.5 % (E_2), and from just above 0.5 % (IBU) to slightly under 9.5 % (EE_2), for the global and validation datasets respectively. Furthermore, and contrary to the P1 procedure, the [M1] methodology resulted, in most cases, in best prediction abilities than the [M2] methodology.

With respect to the OLS methodology prediction ability, the pharmaceuticals that presented RPD values above 3 were the ATR, PRC, CRB and SMX for the [raw], the ATR and DSL for the [SNV], the ATR, PRC,

Table 3
Regression equation, R^2, FTIR wavelengths (λ), PLS components, RMSE and RPD values for each studied pollutant and P2 procedure (glb – global and val – validation).

		Regression glb	R^2 glb	# IR λ	# PLS comp	RMSE glb %	RMSE val %	RPD glb	RPD val
ATR	PLS [M1]	y = 0.992 x	0.985		10	2.30	2.67	8.17	7.04
	PLS [M2]	y = 0.994 x	0.996		10	1.14	1.30	**16.49**	14.42
	OLS [raw]	y = 0.996 x	0.990	10		1.87	1.61	10.02	11.67
	OLS [SNV]	y = 0.987 x	0.947	10		4.24	4.53	4.44	4.15
	OLS [MSC]	y = 0.995 x	0.990	10		1.86	1.44	10.08	13.08
	OLS [1D]	y = 0.998 x	0.995	10		1.33	1.31	14.13	14.29
	OLS [2D]	y = 0.994 x	0.994	10		1.40	1.17	13.40	**16.10**
DSL	PLS [M1]	y = 0.919 x	0.900		7	8.11	11.35	3.16	2.26
	PLS [M2]	y = 0.987 x	0.989		7	2.67	1.97	**9.62**	**13.03**
	OLS [raw]	y = 0.906 x	0.853	7		9.57	11.34	2.68	2.26
	OLS [SNV]	y = 0.961 x	0.927	7		6.67	6.86	3.85	3.74
	OLS [MSC]	y = 0.902 x	0.849	7		9.60	9.94	2.67	2.58
	OLS [1D]	y = 0.952 x	0.937	7		6.45	8.12	3.98	3.16
	OLS [2D]	y = 0.951 x	0.926	7		6.89	8.19	3.73	3.13
PRC	PLS [M1]	y = 0.995 x	0.989		10	2.59	2.92	**9.95**	**8.81**
	PLS [M2]	y = 0.977 x	0.963		10	5.02	6.02	5.13	4.27
	OLS [raw]	y = 1.000 x	0.985	10		3.18	3.58	8.09	7.19
	OLS [SNV]	y = 0.913 x	0.855	10		10.18	12.28	2.53	2.09
	OLS [MSC]	y = 0.984 x	0.980	10		3.67	3.83	7.01	6.71
	OLS [1D]	y = 0.984 x	0.984	10		3.24	3.49	7.94	7.37
	OLS [2D]	y = 0.987 x	0.989	10		2.70	3.13	9.51	8.22
E_2	PLS [M1]	y = 0.976 x	0.965		10	5.45	5.50	**5.41**	5.36
	PLS [M2]	y = 0.950 x	0.939		10	7.27	8.48	4.05	3.48
	OLS [raw]	y = 0.939 x	0.877	10		10.02	11.20	2.94	2.63
	OLS [SNV]	y = 0.942 x	0.913	10		8.47	8.83	3.48	3.34
	OLS [MSC]	y = 0.944 x	0.880	10		9.89	9.38	2.98	3.14
	OLS [1D]	y = 0.959 x	0.960	10		6.00	6.50	4.91	4.53
	OLS [2D]	y = 0.972 x	0.965	10		5.47	4.86	5.38	**6.07**
EE_2	PLS [M1]	y = 0.933 x	0.917		10	8.32	9.30	**3.44**	**3.08**
	PLS [M2]	y = 0.947 x	0.809		10	11.69	12.79	2.45	2.24
	OLS [raw]	y = 0.873 x	0.719	10		12.80	18.01	2.24	1.59
	OLS [SNV]	y = 0.893 x	0.662	10		15.34	18.12	1.87	1.58
	OLS [MSC]	y = 0.891 x	0.708	10		14.11	19.21	2.03	1.49
	OLS [1D]	y = 0.910 x	0.797	10		12.10	13.72	2.37	2.09
	OLS [2D]	y = 0.907 x	0.813	10		12.08	14.42	2.37	1.99
CRB	PLS [M1]	y = 0.969 x	0.962		10	5.54	7.26	5.01	3.82
	PLS [M2]	y = 1.001 x	0.991		10	2.61	3.57	**10.63**	**7.76**
	OLS [raw]	y = 0.994 x	0.960	10		5.44	6.11	5.10	4.54
	OLS [SNV]	y = 0.947 x	0.853	10		10.11	13.35	2.74	2.08
	OLS [MSC]	y = 0.983 x	0.962	10		5.37	7.42	5.16	3.74
	OLS [1D]	y = 0.975 x	0.971	10		4.69	5.74	5.91	4.84
	OLS [2D]	y = 0.979 x	0.981	10		3.87	4.39	7.16	6.32
IBU	PLS [M1]	y = 0.988 x	0.996		10	1.43	0.55	**15.99**	**41.42**
	PLS [M2]	y = 0.886 x	0.927		3	6.32	6.32	3.61	3.61
	OLS [raw]	y = 0.836 x	0.861	6		8.52	9.57	2.68	2.39
	OLS [SNV]	y = 0.788 x	0.825	6		9.45	10.96	2.42	2.08
	OLS [MSC]	y = 0.838 x	0.862	6		8.43	6.20	2.71	3.68
	OLS [1D]	y = 0.880 x	0.904	6		7.10	5.91	3.21	3.86
	OLS [2D]	y = 0.806 x	0.836	6		9.24	9.09	2.47	2.51
SMX	PLS [M1]	y = 0.984 x	0.978		10	3.18	2.84	**6.80**	**7.61**
	PLS [M2]	y = 0.959 x	0.953		10	4.80	7.16	4.51	3.02
	OLS [raw]	y = 0.942 x	0.912	10		6.78	6.92	3.19	3.13
	OLS [SNV]	y = 0.900 x	0.794	10		9.46	13.33	2.29	1.62
	OLS [MSC]	y = 0.951 x	0.904	10		6.52	6.50	3.32	3.33
	OLS [1D]	y = 0.954 x	0.916	10		6.23	7.73	3.47	2.80
	OLS [2D]	y = 0.955 x	0.918	10		6.06	6.39	3.57	3.39

* In all cases the obtained model *p-value* was below 0.001.

CRB and SMX for the [MSC], all but the EE_2 (and the SMX validation dataset) for the [1D] and all but the EE_2 and IBU for the [2D] methodologies, both for the global (training + validation) and validation datasets. Comparing these five methodologies, the best results (higher R^2 and RPD, and lower RMSE values) were obtained by the [1D] methodology for the ATR (R^2 of 0.995), DSL (R^2 of 0.937) and IBU (R^2 of 0.904) and by the [2D] methodology for the PRC (R^2 of 0.989), E_2 (R^2 of 0.965), CRB (R^2 of 0.981) and SMX (R^2 of 0.918) predictions.

Under the P2 procedure, the OLS based methodologies allowed for an adequate prediction of of all of the studied pharmaceuticals with the exception of EE_2, with RMSE % values ranging from just under 1.5 % (ATR) to slightly above 7% (IBU), and from slighly above 1% (ATR) to just under 7% (DSL), for the global and validation datasets respectively. Furthermore, both the [1D] and the [2D] methodologies emerged as the ones presenting the best prediction abilities wthin the OLS based methodologies, somewhat in line with the P1 procedure.

Taking into account both the PLS and OLS based methodologies, all but the PRC (and the ATR and E_2 regarding the validation dataset), were best predicted by the PLS, under the P2 procedure. This result is significantly different from the one obtained under the P1 procedure (with the majority of the pharmaceuticals being best predicted by the OLS methodology). It seems, thus, that the increase in the allowable number of components leads to a greater performance increase for the PLS based methodologies. On the other hand, a more restrictive number

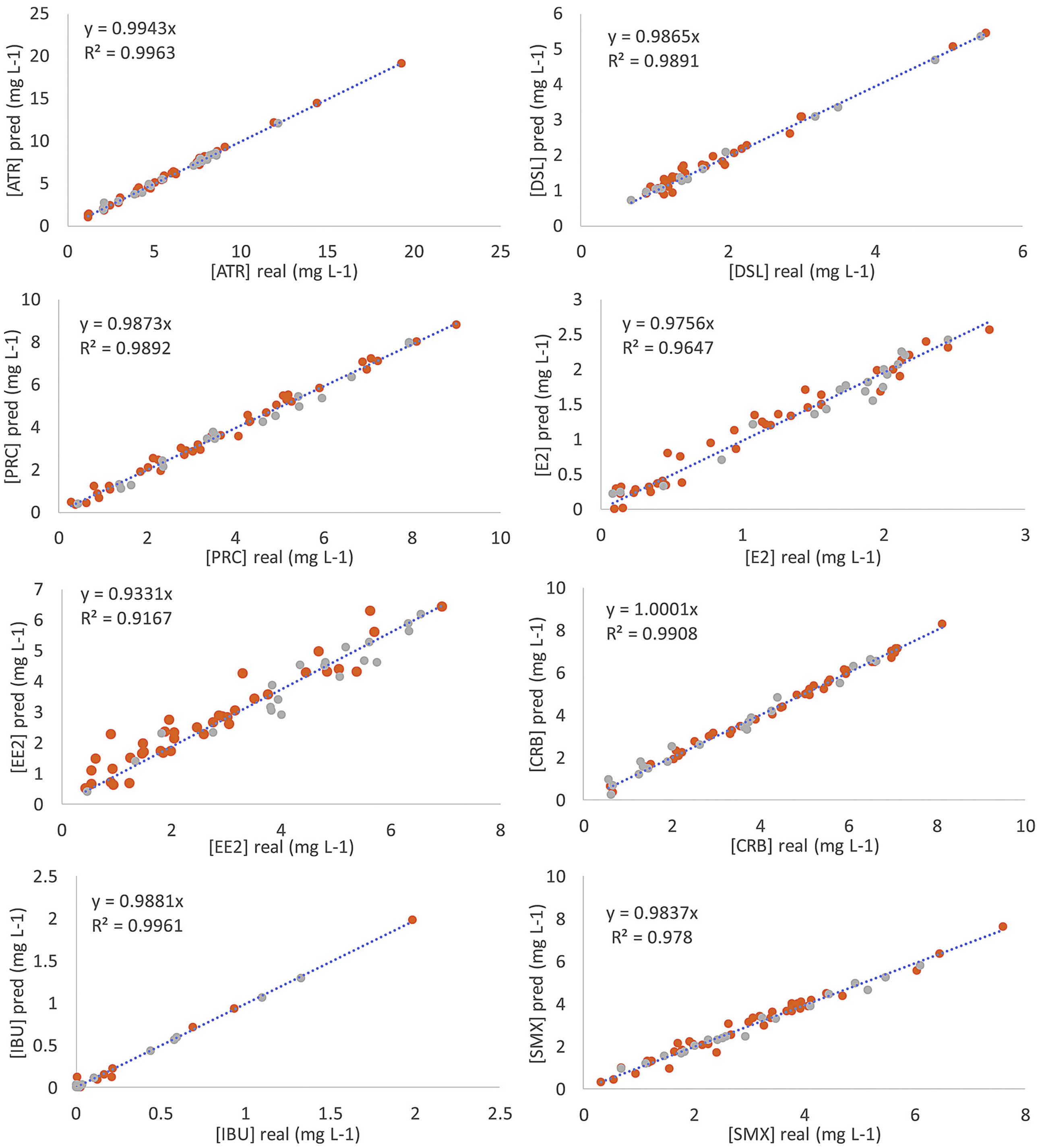

Fig. 1. Best model results for the ATR, DSL, PRC, E_2, EE_2, CRB, IBU and SMX predictions. The orange circles represent the training data and the gray circles represent the validation data.

of components seems to favor the use of OLS based methodlogies.

With an increased number of components, concerns may rise on possible overfitting of the prediction model to the training dataset. However, analysing the RMSE values for both the global and the validation dataset regarding the best models, no significant differences were found, with an average global RMSE to validation RMSE ratio of 1.059 for the P1 procedure and of 1.061 for the P2 procedure (resulting in a 6% difference). Given that the increased number of components in the P2 procedure led to better prediction abilities (average R^2 of 0.978, global RMSE % of 3.72 % and validation RMSE % of 3.94 % for the P2 and of 0.912, 6.50 % and 7.49 % for the P1, regarding the entire pharmaceuticals dataset), and seemed to be unaffected by overfitting problems, the P2 methodlogy results were chosen as the best prediction models.

Hence, these model results for the ATR, DSL, PRC, E_2, EE_2, CRB, IBU and SMX predictions, encompassing both the training (orange circles) and validation (grey circles) data are presented in Fig. 1.

Regarding the PLS models, each individual PLS component (or latent variable) results from a linear combination of the whole set of wavelengths. Therefore, and although the weight of each wavelength for a given PLS latent variable is different, no individual wavelength can be attributed to it.

In contrast, the employed wavelengths by the best OLS models varied from pharmaceutical to pharmaceutical, comprising, in ensemble, an IR region spanning from the 7200 to 200 cm^{-1}. The region that was most used by the prediction models was the 2000 to 200 cm^{-1} (39.7 % of the wavelengths), followed by the 4000 to 2000 cm^{-1} and 6000 to 4000 cm^{-1} (23.3 % and 24.7 % of the used wavelengths, respectively). These regions encompass the absorption frequency of the studied pharmaceuticals main chemical bonds (C−C, C═C, C–H, CH_2, CH_3, C−O, C═O, C−N, C═N, C−Cl, C–S, O−H, N−H, S═O and aromatic rings), as well as their combinations and first overtones.

On the other hand, the studied pharmaceutical compounds present IR characteristic wavelength bands in the regions of 500 to 1750 cm^{-1} and of 2750 to 3500 cm^{-1}, when analyzed as pure substances. Indeed, 57.5 % of the used wavelengths in the best models fell below 3500 cm^{-1}. It should be stressed, however, that these pollutants were not studied as pure substances but diluted in a quite complex matrix (wastewater). In this sense, a diversity of other compounds present IR characteristic bands that partially overlap some of the studied pharmaceuticals bands. In accordance, and in order to the models to assess the studied pharmaceutical weight for its corresponding bands and the weight of other (interference) compounds, a larger set of wavelengths (comprising information of the interference compounds) must be employed. Indeed, the importance of these added wavelengths in the best models can be proven by the prediction ability improvement from the P1 procedure to the P2 procedure.

3.4. Remarks and perspective

The application of FTIR techniques combined with chemometrics is not new and has been used by the pharmaceutical industry for different purposes as to thoroughly understand the feeding and mixing steps in the continuous manufacturing of solid oral dosage forms (Vargas et al., 2018), continuous blend potency determination in the feed frame of a tablet press (De Leersnyder et al., 2018), in-line and real-time monitoring of pharmaceutical hot melt extrusion (Vo et al., 2018) and as an analytical technique (Eldin and Shalaby, 2011; Said et al., 2011). However, all these works focus in the evaluation and determination in solid matrices and only a few papers described the quantification of pharmaceuticals in other matrices as syrup, suspensions, creams and ointments (Ziémons et al., 2010; Silva et al., 2012; Schlegel et al., 2017). Also, the analysis of herbicide residues applied in the soil, and in particular the presence of imazapyr, has already been quantified using near-infrared spectroscopy (NIRS) coupled to a chemometrics approach (Soto-Barajas et al., 2012), but, again, no data has been found for the quantification of herbicides in aqueous matrices by NIR.

Nowadays, FTIR techniques are becoming widespread in monitoring wastewater treatment processes. Several parameters can be measured with this technique as alcohols and volatile organic acids (Nespeca et al., 2017), biogas production (Stockl and Lichti, 2018), biochemical oxygen demand, chemical oxygen demand (COD), turbidity, total organic carbon (TOC) and volatile fatty acids (VFA), among others (Mesquita et al., 2017). On the other hand, the quantification of pharmaceuticals, and other emerging pollutants, in wastewaters by FTIR and chemometrics is just starting to have the attention of the scientific community (Quintelas et al., 2019). Indeed, more work is still needed to explore this technique and to improve the developed models. The occurrence of emerging or newly identified contaminants, as pharmaceutical compounds and herbicides, in the water resources is of continued concern for the health and safety of the community. In this context, the approach described in the present report proposes an eco-friendly and economically viable technology able to solve these problems.

4. Conclusions

Based on FTIR transmission spectroscopy spectra, a chemometric approach was developed for the quantification of emerging pollutants in wastewaters. In accordance, two partial least squares (PLS) and five ordinary least squares (OLS) modelling approaches were employed. The prior use of a k-nearest neighbour (kNN) analysis, aiming at validating the samples allocation within the corresponding studied pollutant, in order to allow the concentration prediction by individually tailored PLS and OLS analyses, allowed for the correct identification of 97.2 % of the samples with solely 2.8 % of the samples misclassified.

The procedure limiting the number of OLS and PLS model components to an upmost of one sixth of the number of calibration samples, led to adequate prediction abilities for the quantification of atrazine, desloratadine, paracetamol, β-estradiol, ibuprofen and carbamazepine. Under this procedure, the OLS based methodologies outperformed the PLS based methodologies for four of the above pharmaceuticals. The second procedure limiting the number of OLS and PLS model components to an upmost of one sixth of the number of global (calibration + validation) samples, led to adequate prediction abilities for the entire set of studied pharmaceuticals, including the ethynylestradiol and sulfamethoxazole. Under this procedure, all but the PRC were best predicted by the PLS. This later procedure allowed for better prediction abilities whuilst remaninig unaffected by overfitting problems.

In conclusion, this methodology can be considered as promising towards a future replacement of UHPLC and GC analysis for the routine quantification of emerging pollutants in wastewaters, representing presently a fast, eco-friendly and reagent free technique for pharmaceuticals screening and estimation in wastewaters. Indeed, the proposed method is a simpler, reagents free and faster technique presenting, at the same time, a high prediction ability for the quantification of atrazine, desloratadine, paracetamol, β-estradiol, ibuprofen, carbamazepine and sulfamethoxazole and slightly lower for the ethynylestradiol. It should be underlined, however, that the employed emerging pollutants identification methodology was only tested for the case where solely one of the studied pollutants was present in a sample. Further extension to samples presenting two or more of the studied emerging pollutants should yet be studied.

CRediT authorship contribution statement

C. Quintelas: Conceptualization, Investigation, Writing - original draft. **A. Melo:** Investigation. **M. Costa:** Investigation. **D.P. Mesquita:** Formal analysis. **E.C. Ferreira:** Supervision, Funding acquisition, Writing - review & editing. **A.L. Amaral:** Formal analysis, Writing - original draft.

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 thank the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UIDB/BIO/04469/2020 unit and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.etap.2020.103458.

References

ASTM E1655-17, 2017. Standard Practices for Infrared Multivariate Quantitative Analysis. ASTM International, West Conshohocken, PA. www.astm.org.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27.

De Leersnyder, F., Peeters, E., Djalabi, H., Vanhoorne, V., Van Snick, B., Hong, K., Hammond, S., Liu, A.Y., Ziemons, E., Vervaet, C., De Beer, T., 2018. Development and validation of an in-line NIR spectroscopic method for continuous blend potency determination in the feed frame of a tablet press. J. Pharm. Biomed. Anal. 151, 274–283.

Einax, J.W., Zwanziger, H.W., Geiss, S., 1997. Chemometrics in Environmental Analysis. Wiley-VCH, Weinheim, Germany.

Eldin, A.B., Shalaby, A.A., 2011. Comparison of FT-NIR transmission and HPLC to assay montelukast in its pharmaceutical tablets. Am. J. Analyt. Chem. 2, 885–891.

Fearn, T., 2002. Assessing calibrations: SEP, RPD, RER and R2. NIR News 13, 12–14.

Fedorova, G., Golovko, O., Randak, T., Grabic, R., 2014. Storage effect on the analysis of pharmaceuticals and personal care products in wastewater. Chemosphere 111, 55–60.

Fram, M.S., Belitz, K., 2011. Occurrence and concentrations of pharmaceutical compounds in groundwater used for public drinking-water supply in California. Sci. Total Environ. 409, 3409–3417.

Gowen, A.A., Tsenkova, R., Bruen, M., O'donnell, C., 2012. Vibrational spectroscopy for analysis of water for human use and in aquatic ecosystems. Crit. Rev. Environ. Sci. Technol. 42, 2546–2573.

Khan, G.A., Lindberg, R., Grabic, R., Fick, J., 2012. The development and application of a system for simultaneously determining anti-infectives and nasal decongestants using on-line solid-phase extraction and liquid chromatography–tandem mass spectrometry. J. Pharm. Biomed. Anal. 66, 24–32.

Kosonen, J., Kronberg, L., 2009. The occurrence of antihistamines in sewage waters and in recipient rivers. Environ. Sci. Pollut. Res. - Int. 16, 555–564.

Kristofco, L.A., Brooks, B.W., 2017. Global scanning of antihistamines in the environment: analysis of occurrence and hazards in aquatic systems. Sci. Total Environ. 592, 477–487.

Li, X.S., Li, S., Kellermann, G., 2018. Simultaneous determination of three estrogens in human saliva without derivatization or liquid-liquid extraction for routine testing via miniaturized solid phase extraction with LC-MS/MS detection. Talanta 178, 464–472.

Martens, H., Naes, T., 1989. Multivariate Calibration. Wiley, Chichester.

Mesquita, D.P., Quintelas, C., Amaral, A.L., Ferreira, E.C., 2017. Monitoring biological wastewater treatment processes: recent advances in spectroscopy applications. Rev. Environ. Sci. Biotechnol. 16, 395–424.

Michel, K., Bureau, B., Boussard-Plédel, C., Jouan, T., Adam, J.L., Staubmann, K., Baumann, T., 2004. Monitoring of pollutant in waste water by infrared spectroscopy using chalcogenide glass optical fibers. Sens. Actuators B Chem. 101, 252–259.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill, Maidenhead, U.K.

Nespeca, M.G., Rodrigues, C.V., Santana, K.O., Maintinguer, S.I., Oliveira, J.E., 2017. Determination of alcohols and volatile organic acids in anaerobic bioreactors for H2 production by near infrared spectroscopy. Int. J. Hydrogen Energy 42, 20480–20493.

Noor, P., Khanmohammadi, M., Roozbehani, B., Yaripour, F., Garmarudi, A.B., 2018. Determination of reaction parameters in methanol to gasoline (MTG) process using infrared spectroscopy and chemometrics. J. Clean. Prod. 196, 1273–1281.

Puchert, T., Holzhauer, C.V., Menezes, J.C., Lochmann, D., Reich, G., 2011. A new PAT/QbD approach for the determination of blend homogeneity: combination of on-line NIRS analysis with PC Scores Distance Analysis (PC-SDA). Eur. J. Pharm. Biopharm. 78, 173–182.

Quintelas, C., Mesquita, D.P., Ferreira, E.C., Amaral, A.L., 2019. Quantification of pharmaceutical compounds in wastewater samples by near Infrared Spectroscopy (NIR). Talanta 194, 507–513.

Said, M.M., Gibbons, S., Moffat, A.C., Zloh, M., 2011. Near-infrared spectroscopy (NIRS) and chemometric analysis of Malaysian and UK paracetamol tablets: a spectral database study. Int. J. Pharm. 415, 102–109.

Schlegel, L.B., Schubert-Zsilavecz, M., Abdel-Tawab, M., 2017. Quantification of active ingredients in semi-solid pharmaceutical formulations by near infrared spectroscopy. J. Pharm. Biomed. Anal. 142, 178–189.

Shewiyo, D.H., Kaale, E., Risha, P.G., Dejaegher, B., Smeyers-Verbeke, J., Heyden, Y.V., 2012. HPTLC methods to assay active ingredients in pharmaceutical formulations: a review of the method development and validation steps. J. Pharm. Biomed. Anal. 66, 11–23.

Silva, M.A.M., Ferreira, M.H., Braga, J.W.B., Sena, M.M., 2012. Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy. Talanta 89, 342–351.

Soto-Barajas, M., Gonzalez-Martin, I., Hernandez-Hierro, J.M., Prado, B., Hidalgo, C., Etchevers, J., 2012. NIR spectroscopy to identify and quantify imazapyr in soil. Anal. Methods 4, 2764–2771.

Stockl, A., Lichti, F., 2018. Near-infrared spectroscopy (NIRS) for a real time monitoring of the biogas process. Bioresour. Technol. 247, 1249–1252.

Vargas, J.M., Nielsen, S., Cárdenas, V., Gonzalez, A., Aymat, E.Y., Almodovar, E., Classe, G., Colón, Y., Sanchez, E., Romañach, R.J., 2018. Process analytical technology in continuous manufacturing of a commercial pharmaceutical product. Int. J. Pharm. 538, 167–178.

Vonberg, D., Vanderborght, J., Cremer, N., Putz, T., Herbst, M., Vereecken, H., 2014. 20 years of long-term atrazine monitoring in a shallow aquifer in western Germany. Water Res. 50, 294–306.

Wang, Y., Wang, Q., Hu, L., Lu, G., Li, Y., 2015. Occurrence of estrogens in water, sediment and biota and their ecological risk in Northern Taihu Lake in China. Environ. Geochem. Health 37, 147–156.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, P.R. (Ed.), Multivariate Analysis. Academic Press, New York, pp. 391–420.

Zhang, Y., Zhang, R., Yang, X., Qi, H., Zhang, C., 2019. Recent advances in electrogenerated chemiluminescence biosensing methods for pharmaceuticals. J. Pharm. Anal. 9, 9–19.

Zhou, H., Ying, T., Wang, X., Liu, J., 2016. Occurrence and preliminarily environmental risk assessment of selected pharmaceuticals in the urban rivers, China. Sci. Rep. 6, 1–10 34928.

Ziémons, E., Mantanus, J., Lebrun, P., Rozet, E., Evrard, B., Hubert, P., 2010. Acetaminophen determination in low-dose pharmaceutical syrup by NIR spectroscopy. J. Pharm. Biomed. Anal. 53, 510–516.