



# Data-driven versus conventional N<sub>2</sub>O EF quantification methods in wastewater; how can we quantify reliable annual EFs?

V. Vasilaki<sup>a</sup>, S. Danishvar<sup>b</sup>, A. Mousavi<sup>b</sup>, E. Katsou<sup>a,\*</sup>

<sup>a</sup> Department of Civil & Environmental Engineering, Brunel University London, Uxbridge UB8 3PH, UK

<sup>b</sup> Systems Engineering Research Group (SERG), Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, UK

## ARTICLE INFO

### Article history:

Received 20 January 2020

Revised 25 June 2020

Accepted 26 June 2020

Available online 28 June 2020

### Keywords:

N<sub>2</sub>O emissions

Long-term monitoring campaign

Changepoint detection

Support vector Machine classification model

## ABSTRACT

A long-term N<sub>2</sub>O dataset from a full-scale biological process was analysed for knowledge discovery. Non-parametric, multivariate timeseries changepoint detection techniques were applied to operational variables (i.e. NH<sub>4</sub>-N loads) in the system. The majority of changepoints, could be linked with the observed changes of the N<sub>2</sub>O emissions profile. The results showed that even three-day sampling campaigns between changepoints have a high probability (>80%) to result to an emission factor (EF) quantification with ~10% error. The analysis revealed that support vector machine (SVM) classification models can be trained to detect operational behaviour of the system and the expected range of N<sub>2</sub>O emission loads. The proposed approach can be applied when long-term online sampling is not feasible (due to budget or equipment limitations) to identify N<sub>2</sub>O emissions “hotspot” periods and guide towards the identification of operational periods requiring extensive investigation of N<sub>2</sub>O pathways generation.

© 2020 The Authors. Published by Elsevier Ltd.

This is an open access article under the CC BY license. (<http://creativecommons.org/licenses/by/4.0/>)

## 1. Introduction

Nitrous oxide (N<sub>2</sub>O) emitted during biological nutrients removal, can significantly contribute to the total carbon footprint of Wastewater Treatment Plants (WWTPs). The recent roadmap to carbon neutrality in urban water published by Water and Wastewater Utilities for Climate Mitigation (WaCClIM) project and the International Water Association (IWA) (Ballard et al., 2018), states that direct N<sub>2</sub>O should be considered for the carbon footprint assessment and reporting.

N<sub>2</sub>O fluxes in wastewater processes are characterised by significant spatial and temporal variability due to the different interacting biological processes that consume or produce N<sub>2</sub>O and the variation of operational and environmental conditions (Daelman et al., 2015; Gruber et al., 2019). A recent analysis of N<sub>2</sub>O emission factors (EF) for over 70 full-scale wastewater treatment processes revealed that the sampling frequency and sampling techniques applied in N<sub>2</sub>O monitoring campaigns, can significantly affect the quantified EFs (Vasilaki et al., 2019). For instance, most of the monitoring campaigns lasting less than one month have reported EFs less than 0.3 % of the N-load. On the other hand, studies lasting over a year result in a median EF equal to 1.7 %

of the N-load. The IPCC guidelines for the estimation of N<sub>2</sub>O in WWTPs were updated in 2019; the suggesting an EF of 1.6 % for of the total N-load (IPCC, 2019). However, uncertainties remain; the use of measured emissions data is suggested for the estimation of country-specific EF in large WWTPs (IPCC, 2019). The development of process-based reliable N<sub>2</sub>O EFs requires long-term monitoring campaigns of over 1-year (Gruber et al., 2019; Vasilaki et al., 2019).

Long-term N<sub>2</sub>O sampling (continuous or via grab-samples) is still rarely performed in WWTPs. High cost and complexities of long-term online monitoring are the main limiting factor. There is still lack of a holistic low-cost, practical approach for the quantification of N<sub>2</sub>O EFs. Therefore, new approaches are required for the quantification of EFs, minimizing sampling rate and advising on the duration and frequency of sampling campaigns.

A large amount of raw, heterogeneous operational data is available from WWTP operations (Olsson et al., 2014). Several studies have demonstrated that utilisation of historical data (i.e. DO, mixed liquor suspended solids (MLSS), NH<sub>4</sub><sup>+</sup> concentrations) from WWTPs can feed statistical methods and predict the profile of target process variables or key performance indicators that cannot be monitored online; an overview can be found in the study of Haimi et al. (2013). Additionally, data-driven techniques have been extensively used to capture the non-linearities and complex structures of wastewater treatment processes towards their optimisation, monitoring and better control

\* Corresponding author.

E-mail address: [evina.katsou@brunel.ac.uk](mailto:evina.katsou@brunel.ac.uk) (E. Katsou).

(Haimi et al., 2013; Corominas et al., 2018; Newhart et al., 2019). Vasilaki et al. (2018) showed that variables monitored online can be utilised to provide insights on the long-term behaviour and abrupt changes of  $N_2O$  dynamics with the application of clustering and dimensionality reduction techniques. However, advanced information extraction methods have rarely been used to analyse data from  $N_2O$  monitoring campaigns. Recently, Sun et al. (2017) developed a back-propagation artificial neural network (ANN) to simulate  $N_2O$  emissions in an anaerobic-oxic (A/O) process. The authors demonstrated the feasibility and simplicity of predicting  $N_2O$  emissions with the use of data-driven models.

Univariate and multivariate changepoint detection techniques have been widely used to detect changes in underlying distribution of sequences and regime shifts in several applications including investigation of distributional changes in financial markets (Allen et al., 2018) and climate change investigation studies (Kotta et al., 2018). Li et al. (2015), recently, applied a non-parametric multivariate changepoint detection algorithm (edivisive; (James and Matteson, 2013)) to detect changes in water quality variables in a shallow lake (total nitrogen, total phosphorus and Chlorophyll) and linked the changepoints (CPs) with changes in the dynamics and patterns of sediment release.

In this work, the most prevalent sampling approaches in wastewater industry are presented and compared. A data-driven sampling approach and two conventional monitoring approaches have been compared and assessed. The result of the comparison proposes the most accurate and efficient amongst the three methods. Specifically, an approach that uses CPs to analyze the behaviour of online monitored variables linked with  $N_2O$  generation (i.e. DO,  $NH_4^+$ ), is proposed, to detect i) periods (between the CP intervals) with steady  $N_2O$  emissions profile and ii) changes in the temporal range and dynamics of  $N_2O$  emissions. Multivariate changepoint detection was applied to identify structural changes in the variables monitored online (i.e.  $NH_4^+$ , DO, flow-rate,  $NO_3^-$ ) in a full-scale Carrousel reactor. Subsequently, the CPs were linked with changes in the  $N_2O$  emissions behaviour and range during a 15-month monitoring campaign (Daelman et al., 2015). A classification model was developed to predict the range of  $N_2O$  emission loads (i.e. low, medium, high) based on the CP intervals. This approach can support operators to minimise GHG sampling requirements, without compromising long-term EF estimates. The accurate quantification of annual  $N_2O$  EF requires samples collection between all CP intervals and a few sampling days can be sufficient to estimate a representative EF for different CP intervals. The classification model can support the estimation of the  $N_2O$  emission range for new incoming data in the WWTP.

## 2. Materials and methods

### 2.1. Process description and the source of data

$N_2O$  measurements and the extensive data-set of the operational variables from the studies of Daelman et al. (2015) and Vasilaki et al. (2018) were used in the analysis. The dataset belongs to one of the longest  $N_2O$  monitoring campaigns undertaken in the wastewater sector (15 months). A plug-flow reactor linked with two subsequent parallel Carrousel reactors was monitored. A full description of the WWTP can be found in the study of Daelman et al. (2015).

The analysis and the development of the methodological approach were based on data obtained from the Carrousel reactor 1. The data matrix used in the analysis, the location of the sensors in the system and the details of the operational control are provided in the study of Vasilaki et al. (2018). The system includes the following probes: DO (DO1, DO2, DO3) in the beginning, middle and end of the Carrousel reactor, ammonium nitrogen ( $NH_4-N$ )

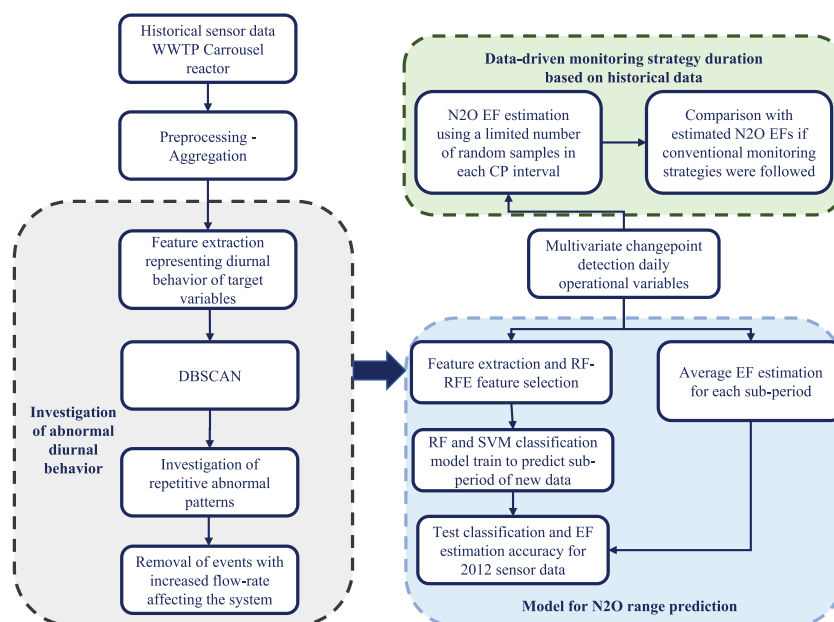
and nitrate nitrogen ( $NO_3-N$ ) in the effluent of the Carrousel reactor,  $NH_4-N$  from the middle of the second oxic zone in the plug flow reactor, temperature and influent flow-rate. The flow-chart of the secondary treatment is provided in the supplementary information.

The behaviour of  $N_2O$  emissions at Carrousel reactor 1 showed a high level of volatility during the 15-month monitoring campaign and characterised by significant diurnal and seasonal variations (see supplementary material). The daily emission loads ranged from  $< 0.004$  kg  $N_2O$  / day to  $>150$  kg  $N_2O$  / day. Daelman et al. (2013) simulated different sampling strategies, based on data collected from the same plug-flow – Carrousel reactor linking EFs with different sampling strategies. The authors simulated the sampling strategies using a long-term dataset. They reported that short-term campaigns (grab sample, 24 h and 7-day sampling), cannot accurately estimate annual EFs, while there is a high probability to underestimate actual emissions. The relative error of the estimated annual  $N_2O$  emissions ranged between -22 % and 35 % (95 % of the cases) by simulating a 50-days  $N_2O$  sampling campaign (random 24h periods on working days were selected). The authors found that long-term offline/online sampling capturing seasonality and temperature effects is needed for reliable EF assessment. Reliable estimation of  $N_2O$  emissions, can provide guidance on  $N_2O$  mitigation measures and support WWTPs towards their carbon neutrality goals. However, there is high cost and resources related to long-term, of  $N_2O$  online monitoring. Therefore, minimizing the sampling requirements can help water utilities to integrate  $N_2O$  monitoring in practice.

Vasilaki et al. (2018) applied a changepoint detection technique on the  $N_2O$  emissions hourly time-series in Carrousel reactor 1, combined with hierarchical  $k$ -means clustering to investigate the  $N_2O$  emission patterns and identify links with the variables monitored online. The study concluded that i) the dependencies between  $N_2O$  and other operational variables (i.e.  $NH_4^+$ ,  $NO_3^-$ , DO) varied in different sub-periods, ii) the system disturbances are mainly linked with events of elevated influent flow-rates, iii) specific ranges of operating variables have historically resulted in low or high ranges of  $N_2O$  in the different sub-periods. These findings have been used to develop the methodological framework of Section 2.2.

### 2.2. Methodological framework and data analytics

Fig. 1 summarises the methodology applied in the current study. Pre-processed data obtained from the work of Vasilaki et al. (2018) were used in the analysis. The previous examination showed that disturbances (i.e. precipitation events) significantly affect the  $NH_4-N$  effluent concentrations. Thus, the first step of the analysis was to isolate and categorize abnormal diurnal pattern of the influent flow-rate that affected system performance. Subsequently, sensor data were used to segment the behaviour of the system into sub-periods with different behaviour and operational variables ranges (i.e.  $NH_4-N$ , DO). The aim was to investigate whether changes in the  $N_2O$  emissions coincide with the changes in the range of operational variables. For this purpose, multivariate changepoint detection techniques were applied to categorize one-year historical sensor data into sub-periods exhibiting different pattern. The sequential segmentation of the operational variables enabled the quantification of  $N_2O$  EF over 1 year using a small number of random samples between segments. Average estimated  $N_2O$  emissions were then compared with the respective EFs from conventional monitoring techniques (equivalent sampling duration), following the methodology applied in the study of Daelman et al. (2013). Finally, features were extracted representing the diurnal pattern of the operational variables and classification models were trained to predict the range of  $N_2O$  emissions.



**Fig. 1.** Methodology used for the data-driven estimation of duration of monitoring campaign and sampling frequency. The analysis is based on historical data and the development of SVM and RF classifiers to predict  $N_2O$  emissions range.

The analysis was based on the  $N_2O$  emission ranges between the changepoint segments.

### 2.2.1. Identification and isolation of influent-flow-rate increase

A variation in the behaviour of the online monitored variables was observed during the campaign. Previous analysis (Vasilaki et al., 2018) showed that abrupt and rapid increases in the influent flow-rate were linked with precipitation events and often resulted into peaks in ammonium concentration in the effluent of the Carrousel reactor. The following steps were performed in order to detect and isolate diurnal influent flow-rate patterns that affected the system's performance (Fig. 1): i) features were extracted representing the diurnal behaviour of influent flow-rate and ammonium concentration (i.e. daily mean), ii) the selected features were transformed into a lower dimension space using principal component analysis (PCA) and iii) density-based spatial clustering of applications with Noise (DBSCAN) was applied to detect days that did not exhibit the expected dynamics and range of the target variables.

PCA (Jolliffe, 2002) was applied to reduce feature dimensionality by eliminating a proportion of variance in data. Subsequently, DBSCAN (Ester et al., 1996) was applied to the features and clusters with regions of high and low density. DBSCAN has been applied in various studies to identify outliers considering multivariate sensor data (i.e. precipitation, humidity) (Saeedi Emadi and Mazinani, 2018). DBSCAN was applied to the first three PCs extracted from the selected feature vector (explaining ~90% of the total variance) to isolate data at a distance greater than a pre-defined distance. DBSCAN was executed in the R package (Hahsler et al., 2017); details are provided in Vasilaki et al. (2020).

### 2.2.2. Changepoint detection

Following a similar approach to the respective one applied in the study of Li et al. (2015), the aim was to investigate whether distributional changes and level shifts of variables conventionally monitored in wastewater systems can be used to detect changes in the range and formation of  $N_2O$  emissions. Identifying changes in the online data collected from wastewater treatment processes is not straight forward; the time-series consist of a combination

of seasonal, gradual and abrupt changes. For this purpose, the e-divisive algorithm was used from the R package *ecp* (James and Matteson, 2013). E-divisive changepoint detection algorithm is a non-parametric method that slices the time-series by detecting changes in the characteristic functions of the underlying distributions (that define a probability distribution) between segments. The method assumes that the  $\alpha$  absolute moment (for  $\alpha \in (0, 2]$ ) exists and that observations are independent. E-divisive is an iterative procedure where in each iteration one single changepoint that divides the time-series into two segments that maximise the difference between the characteristic functions of the segments is detected. Subsequently, the statistical significance of the changepoint is evaluated based on a permutation test (James and Matteson, 2013). The procedure is repeated until the statistically significant changepoints (CPs) of variables have been identified. In the implementation, 21 days (3 weeks) were selected as the minimum distance between CPs, in order to account for seasonal variability. The confidence was defined as equal to 95% in order to control the false-positive rate of CPs. The identified CPs indicate distributional changes of the operational variables in the system. CP intervals are defined as the periods between CPs.

### 2.2.3. $N_2O$ emission factor estimation based on changepoint detection

Different  $N_2O$  sampling scenarios were tested following the methodology described in Daelman et al. (2013), to evaluate whether changepoint detection applied to historical data can reduce the required number of samples for the determination of  $N_2O$  EFs. For this purpose, the daily emission load was calculated ( $kgN_2O/day$ ), for the first year of the monitoring campaign in the Northern Carrousel reactor ( $N_2O$  emissions dataset - *DatN<sub>2O</sub>*).

Three different scenarios were considered; i) random 3-day monitoring between the CP intervals (total samples equal to 36 days/year) (sampling strategy 1 -ST1), ii) monitoring  $N_2O$  emissions for 3 random days each month for 1 year to account for seasonal variability of  $N_2O$  emissions (36 days/year) (sampling strategy 2 -ST2), iii) random monitoring for 36 days/year (sampling strategy 3 -ST3).

In all sampling strategies, it is assumed that emissions were monitored continuously for 24 h (starting from 00:01 a.m. of the chosen day). Emissions averaged over the 24 h periods, represent

the daily average  $N_2O$  emitted (kg  $N_2O-N$  / d). In ST1, Dat $N_2O$  dataset was used to extract randomly 3 days from each CP interval (total 36 days). In ST2, 3 days were extracted randomly from each month (total 36 days), whereas in ST3 36 days were randomly selected over the whole Dat $N_2O$  dataset (total 36 days). Subsequently, the average  $N_2O$  emissions over these 36 days in all scenarios were estimated and were considered to represent the annual EF estimates. This procedure was repeated 10,000 times. Therefore, for each scenario, 10000 average annual  $N_2O$  emission loads were simulated, and a frequency histogram with the potential annual  $N_2O$  emission estimates was developed and compared with the observed average  $N_2O$  emissions.

#### 2.2.4. Feature extraction

The data included in the analysis are characterised by 24-hour cyclical trend, therefore 24 hours were selected as the time interval. A 24-element vector was developed for each variable monitored in the system (representing hourly average). Subsequently, features were extracted, and a feature vector was developed, representing the behaviour of the system; in total >100 features were extracted that can be grouped into three main categories (see supporting information). The first category consists of first-order statistical features including measures of central tendency (i.e. mean), measures of variability (i.e. standard deviation), measures of shape (kurtosis, skewness) and basic statistical functions such as daily maximum, minimum and interquartile range (IQR). First-order features are calculated using the real values of the time-series and provide information on the diurnal behaviour of the variables. The second category consists of second-order features calculated based on the differences between neighbouring values (Nanopoulos et al., 2001). The third group of the features was developed based on specific diurnal sub-events. It captures the pattern of operational variables under specific conditions. The intensity and presence of these events and patterns can provide information on the temporal pattern of the system. In the system, aerator 1 operates under on/off pattern (when ammonium is higher than 1.2 mg/l), while aerators 2 and 3 operate always and peak when ammonium is higher than 0.6 and 0.9 mg/l, respectively. Therefore, one subset of this group of features aimed at capturing the behaviour of DO and nitrate concentrations when ammonium concentrations in the Carrousel reactor effluent were higher than 1.2 mg/L and lower than 0.6 mg/L. Additionally, the concentration of ammonium and nitrate in the plug-flow reactor and the flow-rates provide an indication of the loads entering the Carrousel reactor. Subsequently, the second subset of features belonging to this group, aimed at capturing the diurnal duration of low (<10 mg/L) or high (>18 mg/L) ammonium concentration in the plug-flow reactor and the respective pattern of nitrate concentration. Calculated second-order features of values were also used to calculate a subset of features belonging to group three. For instance, there are periods with a strong relationship between  $N_2O$  concentration and nitrate concentration in the Carrousel effluent. Therefore, the diurnal duration and strength of increasing/decreasing nitrate concentrations were calculated and linked with and the response of other variables in the system. A detailed list of the tested features is provided in the supplementary information.

Finally, Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) (Torres et al., 2011) was applied, to deconstruct the Temperature and  $NH_4-N$  concentration in the plug-flow reactor, into Intrinsic Mode Functions (IMFs) representing different oscillatory components. These variables were selected to investigate seasonal and cyclical components. CEEMDAN is an extension to Empirical Mode Decomposition (EMD) (Huang et al., 1998) and to the ensemble empirical mode decomposition (EEMD) (Wu and Huang, 2009). Results from CEEMDAN decomposition are provided in the supplementary information.

In this study, CEEMDAN was used to extract the trend of the variables monitored in the system. Trend and variability of the variables are defined based on the study of Wu et al. (2007) following the methodology applied in Antico et al. (2014). Specifically, the variability consists of the modes with oscillatory characteristics less than 3 months; the trend is extracted by subtracting all the variability modes with oscillatory periods less than 90 days from the original variable. The frequency matching the peak of the raw periodogram is used to define the oscillatory period. The behaviour of wastewater treatment plant processes is affected by environmental factors (i.e. temperature); the exact seasonal trend is strongly linked with the local environmental conditions and on many occasions cannot be extrinsically derived. CEEMDAN is an adaptive approach based on information extracted from the raw data (Antico et al. 2014) and it can be useful to extract trends when analyzing wastewater treatment processes data. In the algorithm, the noise level was set to 0.02, the realisations to 1000 and the maximum sifting iterations to 1000. Subsequently, the aggregated daily mean of ammonium concentration and temperature were added to the set of features.

#### 2.4.5. Feature selection

The objective of the feature selection is to isolate feature-subsets, that can distinguish days belonging to different ranges of  $N_2O$  emissions. Feature selection has been widely applied in environmental modelling, i.e. for groundwater quality monitoring (Rodriguez-Galiano et al., 2018) and in renewable energy prediction problems (Salcedo-Sanz et al., 2018). In the wastewater sector, Tomperi et al. (2017) recently used five different features (i.e. based on forward selection, stepwise regression and genetic algorithms) together with a multivariable linear regression, to optimise the prediction of quality wastewater parameters from process measurements and high-resolution optical monitoring.

In this study, a recursive feature elimination (RFE) approach (Guyon et al., 2002) applied to the feature vector for the selection of features. It implements backward elimination of features, wrapped with a standard random forest classification algorithm is Random forest classification (Breiman, 2001) is a nonparametric machine learning method where multitude of decision trees are constructed from a random subset of the features and trained in a bootstrap sample of the training set (consisting of around 2/3 of the data producing uncorrelated predictions); the final class prediction consists of the repeated outputs of these trees. It is one of the most powerful methods for feature selection and classification. The algorithm was implemented in caret package (Kuhn, 2008) in R software; details are provided in the supplementary information.

The selected features were used to build and compare an SVM classification model using the same training and test sets (see supplementary information). This procedure was repeated 50 times for the different training and test dataset splits. The performance of the classification models was evaluated based on the accuracy, kappa (see supplementary material).

#### 2.4.6. Support vector machine

SVMs, initially developed by Cortes and Vapnik (1995) are a range of supervised non-parametric machine learning algorithms with applications in several sectors, including wastewater (Corominas et al., 2018). In SVM algorithms kernel functions can be used to map the observation into a high-dimensional (possibly infinite) feature space; the 'maximum margin hyperplane' is then selected in this space. This is the separating hyperplane that has the farthest minimum distance to observations belonging to different classes. The SVM algorithm is developed in R using kernlab (Karatzoglou et al., 2004), and the radial basis function (RBF) kernel as described in Vasilaki et al. (2020). Repeated 10-fold cross validation (3 repetitions) was applied to select the optimal cost

and gamma ( $\gamma$ ) parameters over a two-step grid-search with the caret package (Kuhn, 2008) following the methodology proposed in Hsu et al. (2003). The cost defines the penalty of instances misclassified or instances exceeding the maximal margin whereas  $\gamma$  determines the amplitude of the kernel. One versus one approach was applied to classify into more than two classes. The dataset was randomly divided into test and train, with 75% of the available data used for training the SVM model and 25% used for testing. Additionally, data from the last period of the monitoring campaign (~30 days), not used during training, were also tested, to evaluate the performance of the model.

### 3. Results and discussion

#### 3.1. Detection of abnormal events

The DBSCAN features (supplementary material); were selected in order to represent the diurnal behaviour of the target variables. The *eps* parameter was determined based on the “knee” of k-nearest neighbours of the data plotted in increasing order (see supplementary information). Based on this procedure *MinPts* and *eps* were set equal to 6 and 0.4 respectively.

The pattern of the influent flow-rate and effluent  $\text{NH}_4\text{-N}$  concentration, in the days isolated by DBSCAN are shown in the supplementary material. In total 155 days were isolated and are mainly characterised by events with elevated influent flow-rate or/and peaks of the  $\text{NH}_4\text{-N}$  concentration in the Carrousel efflu-

ent. Subsequent inspection showed that these events varied in intensity and duration; therefore, they were categorized into three major groups. Group 1 consists of system disturbances with duration equal or less than 24 h. Days belonging to group 1 are characterised by elevated influent flow-rate (days with precipitation) and peaks in the effluent  $\text{NH}_4\text{-N}$  concentration during the same day and thus, low removal efficiency of  $\text{NH}_4\text{-N}$ . In group 1, the system resumes to normal  $\text{NH}_4\text{-N}$  removal efficiency after 24h. Group 2 consists of system disturbances lasting more than 24h. Multiple days with precipitation, at close temporal proximity, that affect the system performance for several days were assigned to this group. Finally, elevated influent flow-rate events lasting less than 24h, affected significantly the behaviour of  $\text{NH}_4\text{-N}$  concentration in the Carrousel effluent for several days. These occasions were assigned to group 3. In total, 54 different events were detected (155 days) that belong to one of the three groups.

Fig. 2 shows examples of the events belonging in groups 1, 2 and 3 and the pattern of  $\text{N}_2\text{O}$  emissions. Blue lines represent the events detected by DBSCAN and red lines represent the normal operational conditions. Overall, ~30% of the events belong to group 1. The average daily influent flow-rate is ~4000  $\text{m}^3/\text{h}$ ; therefore, days in group 1 have moderate increase of the influent flow-rate (flow-rate peaks <7000  $\text{m}^3/\text{h}$  and  $\text{NH}_4\text{-N}$  concentration in the Carrousel peaks < 6  $\text{mg}/\text{L}$ ). Fig. 2 (a-d) shows an event with the highest influent flow-rate peak;  $\text{N}_2\text{O}$  emissions are not significantly affected. Overall, the behaviour of  $\text{N}_2\text{O}$  emissions for 1-day events at temperatures between 12–16°C is not significantly affected compared

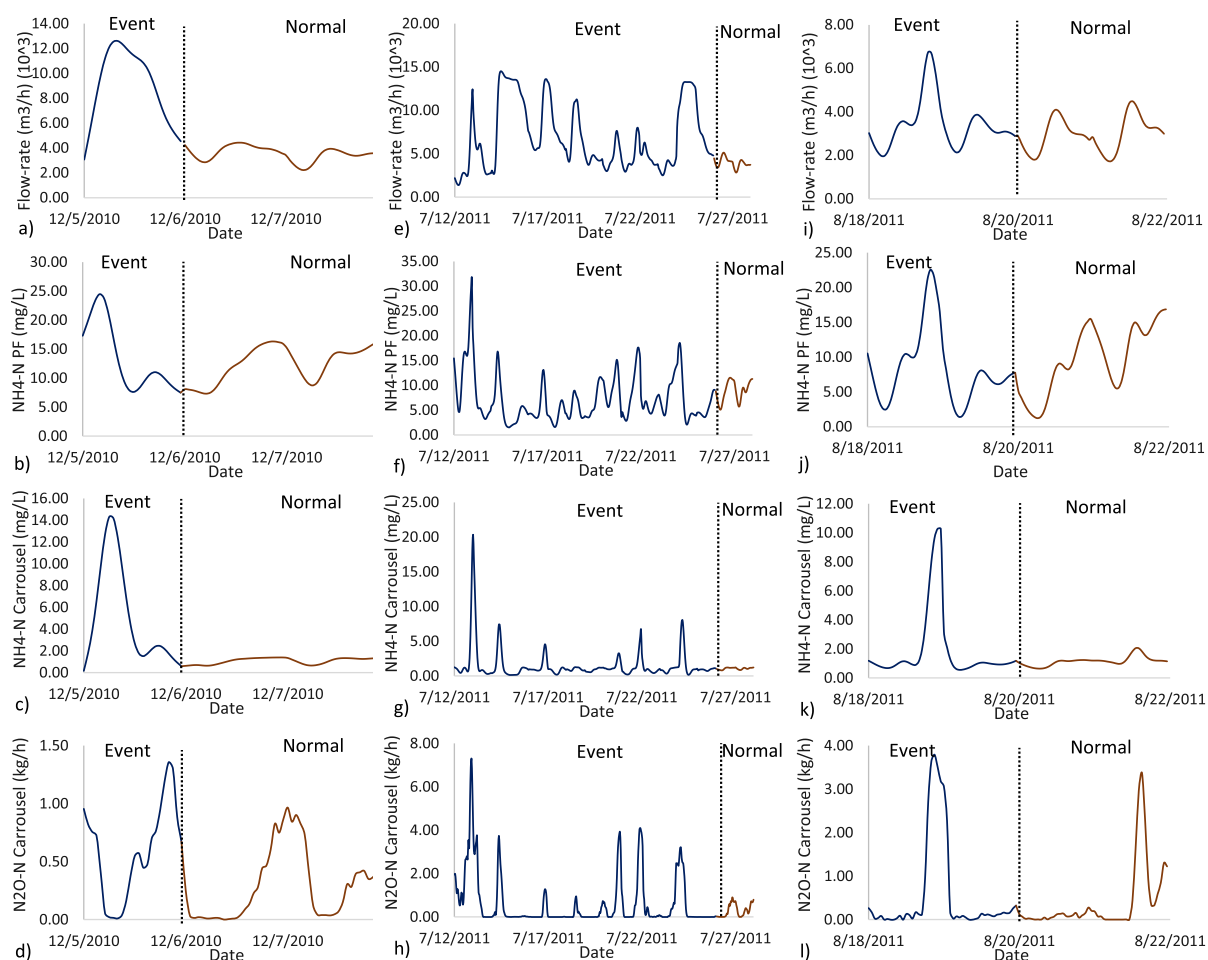


Fig. 2. Flow-rate,  $\text{NH}_4\text{-N}$  concentrations in the plug-flow and Carrousel effluent and  $\text{N}_2\text{O}$  emissions' profiles for days detected by DBSCAN belonging to events group 1 (a-d), group 2 (e-h) and group 3 (i-l).

to the average behaviour of emissions the day prior to the event (Fig. 2 (d)). However, significant peaks, of N<sub>2</sub>O emissions, coinciding with group 1 events, are observed at higher temperatures. Days belonging to group 2 are characterised by influent flow-rate peaks above 8000 m<sup>3</sup>/h, whereas NH<sub>4</sub>-N concentration peaks are higher than 9 mg/L in the effluent of the Carrousel reactor (Fig. 2 (e-h)). Again, the pattern of N<sub>2</sub>O emissions varies. However, emissions tend to drop after the peak of the influent flow-rate (NH<sub>4</sub>-N concentrations in the plug-flow <8 mg/L). Finally, most N<sub>2</sub>O emissions peaks between June and September 2019 belong to group 3 (Fig. 2 (e-h)).

The analysis system's operational behaviour can help operators to identify events that affect performance and apply mitigation strategies (i.e. regulate the anaerobic supernatant stream to the mainstream line to reduce the system loads).

### 3.2. Change-point detection

Change-point detection was applied to identify changes in the profiles of the operational variables that affect the performance of the Carrousel reactor. Daily averages of i) NH<sub>4</sub>-N load (kg/h) in the plug-flow reactor (as an indication of the influent ammonium in the Carrousel reactor), ii) NH<sub>4</sub>-N load (kg/h) in the effluent of the Carrousel reactor, iii) DO1 and DO2 concentrations were used. Daily averages of these variables were considered in order to avoid the diurnal cyclic characteristics of the variables. The volumetric flow-rate of the plug-flow and Carrousel reactors were calculated according to Daelman et al. (2015).

The minimum transition interval was equal to 21 days (~3 weeks) on the assumption that biological processes can be affected by seasonality. The multivariate change-point detection analysis identified 12 statistically significant CPs (with significance level <<0.05) for the first year of the monitoring campaign. Fig. 3 shows the identified CPs and the respective profile of N<sub>2</sub>O emissions for each period interval between CPs. On many occasions, CPs coincide with the changes of the N<sub>2</sub>O emissions profile during the monitoring period. For instance, the highest drop in the average N<sub>2</sub>O emissions between adjoining periods (CPs 6 and 7) coincides with a drop in the ammonium load and an increase in the average nitrate-nitrogen load in the plug-flow reactor. Similarly, the drop of N<sub>2</sub>O emissions between CPs 2 and 3, coincides with an increase in the DO1 concentration in the Carrousel reactor.

### 3.3. Accuracy of the monitoring strategy based on system CPs

As shown in Section 3.2, the pattern and range of N<sub>2</sub>O emissions changes between the detected CP intervals. Quantification of reliable N<sub>2</sub>O EFs in wastewater treatment processes is still not straightforward; monitoring campaign duration and strategy significantly affect the reliability of the results. Seasonal effects have also significant impact on N<sub>2</sub>O emissions (Vasilaki et al., 2019).

The aim of this analysis is to simulate a knowledge-based N<sub>2</sub>O sampling campaign between CPs and evaluate EFs following a similar approach to the study of Daelman et al. (2013). Additionally, the knowledge-based sampling campaign is compared with two alternative monitoring strategies: i) random 24-h monitoring and ii) random 24-h monitoring for specific days at each month capturing the seasonal variability.

Fig. 4 (a) shows the relative frequency histogram of the estimated annual N<sub>2</sub>O load (kg/day) when 24h sampling for 36 random days is applied, during the first year of the monitoring campaign ( $n=10000$  repetitions). The red vertical line represents the measured average annual N<sub>2</sub>O load (equal to 39- kg/day). In total 43% of the simulations resulted to an EF ranging between 35 and 43 kg/day (less than 10% error from the actual annual N<sub>2</sub>O load

**Table 1**

Classes considered in the classification based on the CP intervals.

Class	N <sub>2</sub> O emissions (kg/d)	Change-point intervals	N <sub>2</sub> O emissions level
1	8.3 (±8.3)	1, 2, 3	Low
2	68.3 (±23.2)	4, 5	Medium
3	141.7 (±31.6)	6	High
4	58.9 (±25.0)	7, 8, 9	Medium
5	7.2 (±7.9)	10, 11, 12	Low

quantified by the monitoring campaign). Additionally, the probability to underestimate the N<sub>2</sub>O load by more than 10% is equal to ~30%. Fig. 4 (b), shows the histogram of the estimated annual N<sub>2</sub>O load (kg/day) when 24h sampling for three random days between the CPs (12 CP intervals) was assumed ( $n=10000$  repetitions). Overall, the likelihood to estimate an average N<sub>2</sub>O load between 35 and 43 kg N<sub>2</sub>O/day was equal to ~80%, with >99% of the simulated N<sub>2</sub>O estimates ranging between 32 and 46 kg/day. In this case, the probability to underestimate the emissions by more than 10% was approximately 5%. Finally, when random sampling for 3 days per month was tested (Fig. 4 (c)), the probability to estimate an N<sub>2</sub>O load ranging between 35 and 43 kg/day, was equal to ~70%, whereas the probability to underestimate the emissions by more than 10%, was approximately 25%.

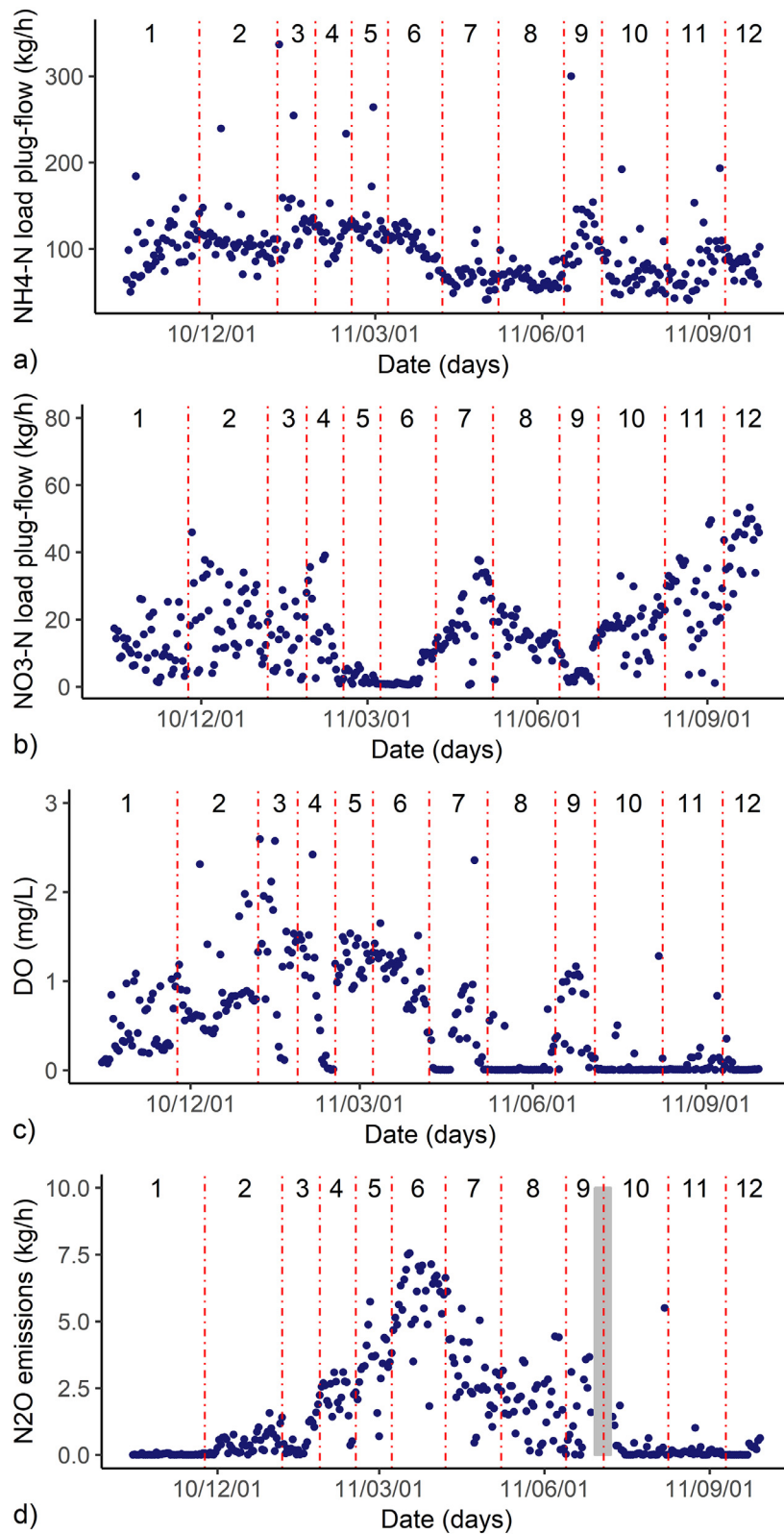
The behaviour of the operational variables needs to be considered together with seasonal effects when sampling campaigns are planned. In the investigated system, limited sampling days between the CPs could give a realistic quantification of the actual EF during a whole year. The proposed approach can be applied to identify N<sub>2</sub>O emissions "hotspot" periods and guide towards the identification of the operational periods that require intensive investigation of N<sub>2</sub>O pathways and mitigation measures.

### 3.4. Feature selection and classification

A classification algorithm was constructed to predict low, medium or high N<sub>2</sub>O emissions based on the operational behaviour of the system. The categorization of the different classes was based i) on the CP analysis, ii) on the seasonal effects. Therefore, two periods (from the CP intervals) characterised by similar N<sub>2</sub>O ranges but not sequential, were assigned in two different classes. Table 1 shows the average N<sub>2</sub>O emissions in each class and the change-point intervals for each class (the change-point intervals are shown in Fig. 3). However, five only classes were considered in the feature selection and construction of the classification algorithms.

Feature selection is a significant step of several high-dimensional classification applications. However, many studies have shown that selected features depend on the training sample, and thus, a feature selection algorithm can be unstable (He and Yu, 2010; Kalousis et al., 2007). Therefore, in many cases feature selection stability needs to be considered together with model prediction accuracy in the evaluation of the classification/regression performance (Pes et al., 2017; Saeys et al., 2008).

The most common features for feature subset sizes equal to 6 and 10 for the resampling perturbations are shown in the supplementary information. Overall, 4 features coincided in all subsets with feature size equal to 6 and 8 features coincided in almost all subsets with feature size equal to 10. The selected features are divided between feature group 1 (i.e. first-order statistical features - maximum DO2, minimum influent flow-rate), feature group 3 (that capture the pattern of operational variables under specific conditions) and feature group 4 (i.e. trend extracted by CEEMDAN). For instance, the features describing the pattern of DO2 concentration for NH<sub>4</sub>-N higher than 1.2 mg/L in the Carrousel effluent were selected in all feature subsets. The trends of NH<sub>4</sub>-N concentration

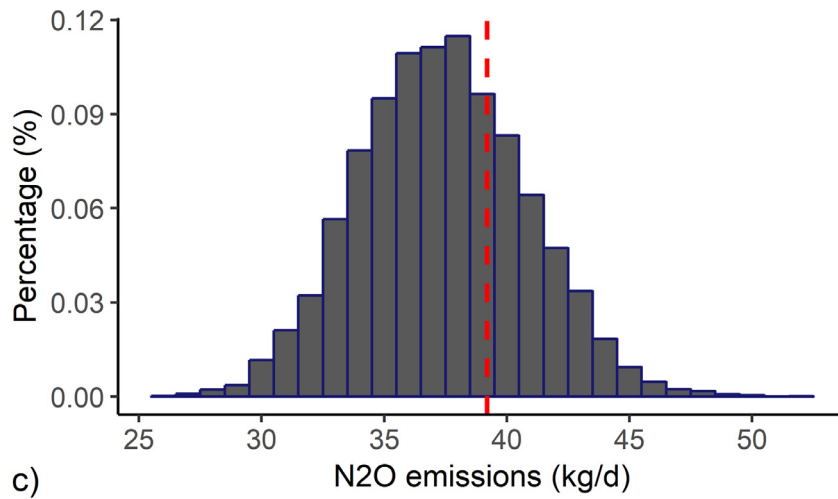
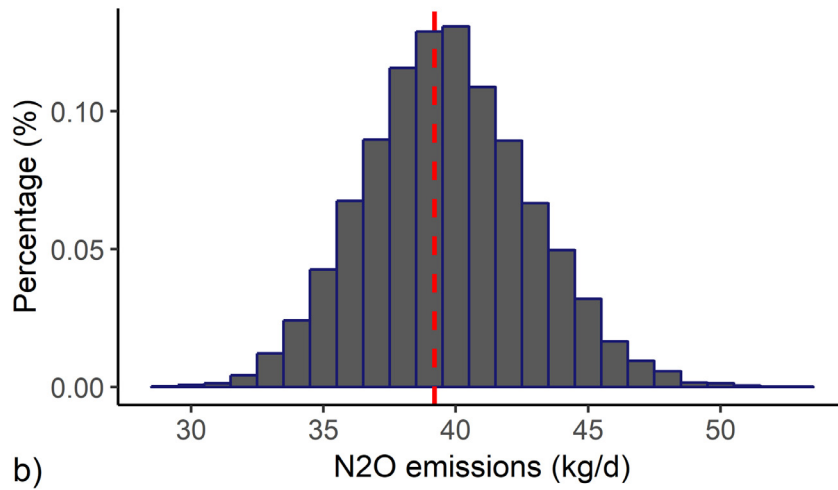
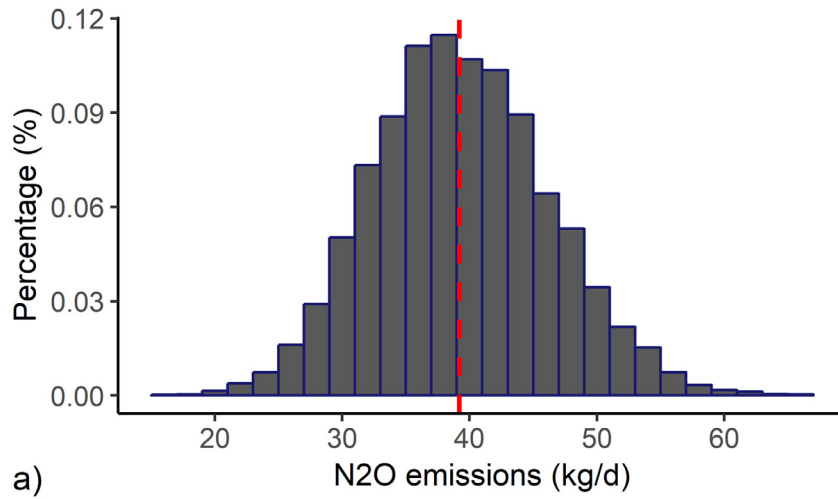


**Fig. 3.** CPs intervals for a)  $\text{NH}_4\text{-N}$  load (kg/h), b)  $\text{NO}_3\text{-N}$  load (kg/h), c) DO1 and DO2 average concentration, d)  $\text{N}_2\text{O-N}$  emissions behaviour between CP intervals.

in the plug-flow reactor and temperature extracted by CEEMDAN, were also included in all feature subsets.

The results of the SVM and RF classification models from the different resampling perturbations for both the train and test data sets are shown in Table 2. Feature subsets that minimise the clas-

sification error were selected in each resampling. The results show similar results both for the RF and SVM classifiers, whereas classification accuracy in the test dataset is high even for small feature subsets (> 97%). In 58% of the resampling perturbations, the size of the best feature subset was equal to 6. In total, 8 variables



**Fig. 4.** Relative frequency histogram of the estimated average annual  $N_2O$  load for (a) a simulated monitoring campaign with duration equal to 36 random days, (b) a simulated monitoring campaign with 3 random days between CP intervals and (c) a simulated monitoring campaign with 3 random days between different months for 1 year.



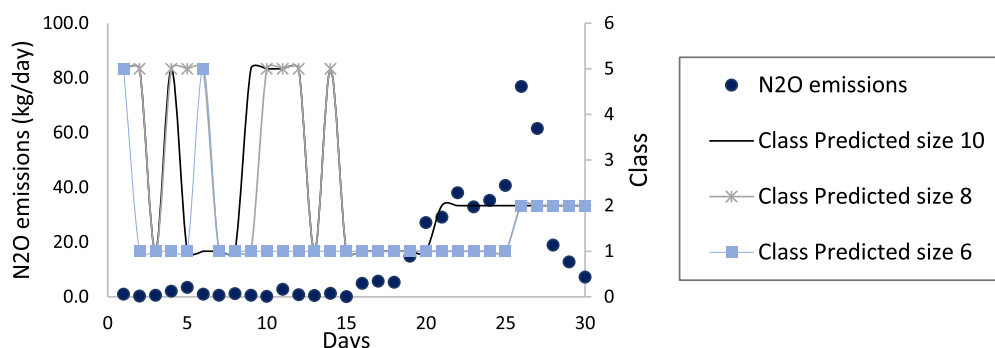


Fig. 5. Daily N<sub>2</sub>O emissions and predicted classes (SVM classification) for the last period of the monitoring campaign (data not used during changepoint detection).

**Table 2**  
Evaluation of SVM and RF classifiers for different feature subset sizes.

Feature subset size	Model	Dataset	Accuracy (%)	Kappa (%)
6	RF	Train	97.4 (± 1)	96.6 (± 1)
		Test	97.4 (± 2)	96.7 (± 4)
	SVM	Train	98.9 (± 1)	98.6 (± 1)
		Test	95.1 (± 3)	93.7 (± 4)
8	RF	Train	97.3 (± 2)	96.6 (± 2)
		Test	98.2 (± 2)	97.7 (± 2)
	SVM	Train	98.7 (± 1)	98.4 (± 2)
		Test	95.9 (± 2)	94.8 (± 2)
10	RF	Train	97.4 (± 3)	96.6 (± 3)
		Test	98.3 (± 1)	97.7 (± 2)
	SVM	Train	99.0 (± 1)	98.8 (± 2)
		Test	96.3 (± 3)	95.2 (± 4)

were selected for 14% of the resampling perturbations, 10 for 20% of the resampling perturbations whereas all the features were selected for 4% of the resamples.

Subsequently, data from the last period of the monitoring campaign were tested to assess the predictive capabilities of the models in previously unseen operational periods. In total ~30 days were tested (precipitation events were not considered). Fig. 5 displays the predicted classes (SVM classification) for different best feature subset sizes (majority rule for the different resampling perturbations) and the daily N<sub>2</sub>O emissions. In the first 20 days all models have predicted low emission-risk classes. Class 1 represents operational conditions from the beginning of the monitoring campaign, whereas class 5 represents conditions from the end of the monitoring campaign (after ~1 year). The investigated period is ~ 1 year after the start of the monitoring campaign and the observed N<sub>2</sub>O emissions are again low. Therefore, the alternation of the predicted classes between class 1 and 5 can be expected, mainly due to the impact of seasonal effects on the influent concentration and on N<sub>2</sub>O generation. Finally, the model with best feature subset size equal to 10 was the only one able to detect the change in the N<sub>2</sub>O range after the first 20 days.

Additional data are required to investigate the generalisation capabilities of the SVM classifier. The methodological approach is able to predict the range of N<sub>2</sub>O emissions, as long as the system operates within the predefined and investigated range.

Under the investigated conditions, the accuracy of the classifier with feature subset size equal to 10, was satisfactory, even when data, from the second year of the campaign were tested (these data were not used during training, 30 days). Therefore, the SVM classifier can be used (with caution) to detect periods with operational behaviour that has been historically linked with elevated emissions.

The development of mitigation measures in the predicted high-risk N<sub>2</sub>O emission periods, can be supported with the integra-

tion of mechanistic models or practical, simplified theoretical approaches. The latter facilitates the identification of potential triggering mechanisms linked with the period-specific operational conditions. For example, a simplified N<sub>2</sub>O risk-based model was developed by Porro et al. (2014) considering thresholds of ASM state variables linked with the generation of N<sub>2</sub>O emissions (i.e. DO, nitrite, COD:N) based on the treatment step (nitrification, nitrification, transition zones).

Additional long-term monitoring campaigns, in continuous wastewater systems are required, to validate the proposed strategy and standardise the selection of operational variables that should be considered during changepoint detection and classification. Finally, the development of a detection approach needs also to be integrated in the procedure, that will detect new, unobserved operational states and provide feedback to the algorithms on recalibration requirements.

#### 4. Conclusions

This study shows that information hidden in conventional variables monitored in wastewater can be mined to reduce N<sub>2</sub>O sampling frequency without compromising the quantification of annual N<sub>2</sub>O EFs and ultimately predict the risk of elevated emissions. The application of changepoint detection in the process operational variables provided insights on structural changepoints of the N<sub>2</sub>O emissions profile. Limited 24-h N<sub>2</sub>O samples between the CP intervals are sufficient to estimate the average N<sub>2</sub>O EF for the whole year, while conventional strategies resulted in lower accuracy of the N<sub>2</sub>O EF. An SVM classification model was constructed to predict operational periods linked with specific N<sub>2</sub>O emission ranges. The results indicate that analysis of historical data and investigation of seasonal effects can be of paramount importance in the planning of monitoring campaigns. The proposed approach can be applied when long-term online sampling is not technically and economically feasible. The proposed solution is capable of pinpointing the N<sub>2</sub>O emissions “hotspot” periods and guiding towards the identification of operational periods that require extensive investigation of N<sub>2</sub>O pathways and mitigation measures.

#### Author statement

All authors have contributed for the development of the current work.

#### 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

This paper is supported by the **Horizon 2020** research and innovation programme, SMART-Plant under grant agreement No **690323**.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.compchemeng.2020.106997](https://doi.org/10.1016/j.compchemeng.2020.106997).

## References

- Allen, D.E., McALEER, M., Powell, R.J., Singh, A.K., 2018. Non-parametric multiple change point analysis of the global financial crisis. *Ann. Finan. Econ.* 13, 1850008. <https://doi.org/10.1142/S2010495218500082>.
- Antico, A., Schlotthauer, G., Torres, M.E., 2014. Analysis of hydroclimatic variability and trends using a novel empirical mode decomposition: application to the Paraná River Basin [WWW Document]. *J. Geophys. Res.* <https://doi.org/10.1002/2013JD020420>.
- Ballard, S., Porro, J., Trommsdorff, C., 2018. *The Roadmap to a Low Carbon Urban Water Utility: An International Guide to the WaCCIM Approach*. IWA Publishing Place of publication not identified.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Corominas, L.L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., Poch, M., 2018. Transforming data into knowledge for improved wastewater treatment operation: a critical review of techniques. *Environmental modelling & software, special issue on environmental data science. Applications to air quality and water cycle* 106, 89–103. <https://doi.org/10.1016/j.envsoft.2017.11.023>.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20, 273–297. <https://doi.org/10.1007/BF00994018>.
- Daelman, M.R.J., van Loosdrecht, M.C., Volcke, E.I., De Baets, B., 2013. Influence of sampling strategies on the estimated nitrous oxide emission from wastewater treatment plants. *Water Res.* 47, 3120–3130.
- Daelman, M.R.J., van Voorthuizen, E.M., van Dongen, U.G.J.M., Volcke, E.I.P., van Loosdrecht, M.C.M., 2015. Seasonal and diurnal variability of N<sub>2</sub>O emissions from a full-scale municipal wastewater treatment plant. *Sci. Total Environ.* 536, 1–11. <https://doi.org/10.1016/j.scitotenv.2015.06.122>.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. in: *Kdd*. pp. 226–231.
- Gruber, W., Villez, K., Kipf, M., Wunderlin, P., Siegrist, H., Vogt, L., Joss, A., 2019. N<sub>2</sub>O emission in full-scale wastewater treatment: proposing a refined monitoring strategy. *Sci. Total Environ.* 699, 134157. <https://doi.org/10.1016/j.scitotenv.2019.134157>.
- Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. <https://doi.org/10.1023/A:1012487302797>.
- Hahsler, M., Piekenbrock, M., Arya, S., Mount, D., 2017. dbSCAN: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. R package version 1–0.
- Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: an overview. *Environ. Model. Softw.* 47, 88–107. <https://doi.org/10.1016/j.envsoft.2013.05.009>.
- He, Z., Yu, W., 2010. Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* 34, 215–225. <https://doi.org/10.1016/j.compbiolchem.2010.07.002>.
- Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. *A practical guide to support vector classification*. Taipei.
- Huang Norden, E., Zheng, S., Long Steven, R., Wu Manli, C., Shih Hsing, H., Quanan, Z., Nai-Chyuan, Y., Chao, T.C., Liu Henry, H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. A* 454, 903–995. <https://doi.org/10.1098/rspa.1998.0193>.
- IPCC, 2019. Chapter 5.6 wastewater treatment and discharge. In: Bartram, D., Short, M.D., Ebbe, Y., Farkas, J., Gueguen, C., Peters, G.M., et al. (Eds.), 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse gas Inventories. IPCC Guidelines for National Greenhouse gas Inventories, IPCC.
- Jolliffe, I.T. (Ed.), 2002. *Principal Components in Regression Analysis*, in: *Principal Component Analysis*, Springer Series in Statistics. Springer New York, New York, NY, pp. 167–198.
- James, N.A., Matteson, D.S., 2013. ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data. [arXiv:1309.3295 \[stat\]](https://arxiv.org/abs/1309.3295).
- Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst.* 12, 95–116. <https://doi.org/10.1007/s10115-006-0040-8>.
- Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab - an S4 package for Kernel Methods in R. *J. Stat. Softw.* 11, 1–20.
- Kotta, J., Herkül, K., Jaagus, J., Kaasik, A., Raudsepp, U., Alari, V., Arula, T., Haberman, J., Järvet, A., Kangur, K., Kont, A., Kull, A., Laanemets, J., Maljutenko, I., Männik, A., Nöges, P., Nöges, T., Ojaveer, H., Peterson, A., Reihan, A., Rõõm, R., Sepp, M., Suursaar, Ü., Tamm, O., Tamm, T., Tõnisson, H., 2018. Linking atmospheric, terrestrial and aquatic environments: regime shifts in the Estonian climate over the past 50 years. *PLOS ONE* 13, e0209568. <https://doi.org/10.1371/journal.pone.0209568>.
- Kuhn, M., 2008. Building predictive models in R using the caret package. *J. Stat. Softw.* 28. <https://doi.org/10.18637/jss.v028.i05>.
- Li, Y., Liu, Y., Zhao, L., Hastings, A., Guo, H., 2015. Exploring change of internal nutrients cycling in a shallow lake: a dynamic nutrient driven phytoplankton model. *Ecol. Model.* 313, 137–148. <https://doi.org/10.1016/j.ecolmodel.2015.06.025>.
- Nanopoulos, A., Alcock, R., Manolopoulos, Y., 2001. Feature-based classification of time-series data. *Int. J. Comput. Res.* 10, 49–61.
- Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y., 2019. Data-driven performance analyses of wastewater treatment plants: a review. *Water Res.* 157, 498–513. <https://doi.org/10.1016/j.watres.2019.03.030>.
- Olsson, G., Carlsson, B., Comas, J., Copp, J., Gernaey, K.V., Ingildsen, P., Jeppsson, U., Kim, C., Rieger, L., Rodriguez-Roda, I., 2014. Instrumentation, control and automation in wastewater—from London 1973 to Narbonne 2013. *Water Sci. Technol.* 69, 1373–1385.
- Pes, B., Dessì, N., Angioni, M., 2017. Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. *Inf. Fusion* 35, 132–147. <https://doi.org/10.1016/j.inffus.2016.10.001>.
- Porro, J., Kampschreur, M.J., Pijuan, M., Volcke, E., Daelman, M., Guo, L., Nopens, I., Vanrolleghem, P.A., Yuan, Z., Chandran, K., 2014. Measuring nitrous oxide emissions from biological wastewater treatment, art or science? IWA Specialist Conference on Global Challenges: Sustainable Wastewater Treatment and Resource Recovery. International Water Association (IWA).
- Rodriguez-Galiano, V.F., Luque-Espinar, J.A., Chica-Olmo, M., Mendes, M.P., 2018. Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. *Sci. Total Environ.* 624, 661–672. <https://doi.org/10.1016/j.scitotenv.2017.12.152>.
- Saeedi Emadi, H., Mazinani, S.M., 2018. A novel anomaly detection algorithm using DBSCAN and SVM in wireless sensor networks. *Wireless Pers. Commun.* 98, 2025–2035. <https://doi.org/10.1007/s11277-017-4961-1>.
- Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (Eds.), *Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pp. 313–325.
- Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., García-Herrera, R., 2018. Feature selection in machine learning prediction systems for renewable energy applications. *Renew. Sustain. Energy Rev.* 90, 728–741. <https://doi.org/10.1016/j.rser.2018.04.008>.
- Sun, S., Bao, Z., Li, R., Sun, D., Geng, H., Huang, X., Lin, J., Zhang, P., Ma, R., Fang, L., Zhang, X., Zhao, X., 2017. Reduction and prediction of N<sub>2</sub>O emission from an Anoxic/Oxic wastewater treatment plant upon DO control and model simulation. *Bioresour. Technol.* 244, 800–809. <https://doi.org/10.1016/j.biortech.2017.08.054>.
- Tomperi, J., Koivuranta, E., Kuokkanen, A., Leiviskä, K., 2017. Modelling effluent quality based on a real-time optical monitoring of the wastewater treatment process. *Environ. Technol.* 38, 1–13.
- Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P., 2011. A complete ensemble empirical mode decomposition with adaptive noise. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4144–4147.
- Vasilaki, V., Conca, V., Frison, N., Eusebi, A.L., Fatone, F., Katsou, E., 2020. A knowledge discovery framework to predict the N<sub>2</sub>O emissions in the wastewater sector. *Water Res.* 178, 115799. <https://doi.org/10.1016/j.watres.2020.115799>.
- Vasilaki, V., Massara, T.M., Stanchev, P., Fatone, F., Katsou, E., 2019. A decade of nitrous oxide (N<sub>2</sub>O) monitoring in full-scale wastewater treatment processes: a critical review. *Water Res.* 161, 392–412. <https://doi.org/10.1016/j.watres.2019.04.022>.
- Vasilaki, V., Volcke, E.I.P., Nandi, A.K., van Loosdrecht, M.C.M., Katsou, E., 2018. Relating N<sub>2</sub>O emissions during biological nitrogen removal with operating conditions using multivariate statistical techniques. *Water Res.* 140, 387–402. <https://doi.org/10.1016/j.watres.2018.04.052>.
- Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.* 1, 1–41. <https://doi.org/10.1142/S1793536909000047>.
- Wu, Z., Huang, N.E., Long, S.R., Peng, C.-K., 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. *PNAS* 104, 14889–14894. <https://doi.org/10.1073/pnas.0701020104>.