Journal of Environmental Management 261 (2020) 110219

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: http://www.elsevier.com/locate/jenvman

Research article

Towards an online mitigation strategy for N_2O emissions through principal components analysis and clustering techniques

Giacomo Bellandi [a,*], Stefan Weijers [b], Riccardo Gori [c], Ingmar Nopens [d]

[a] Politecnico di Milano, DICA – Department of Civil and Environmental Engineering, Piazza Leonardo da Vinci, 32, 20133, Milan, Italy
[b] Waterschap de Dommel, Postbus 10.001, Boxtel, NL-5280, DA, the Netherlands
[c] Department of Civil and Environmental Engineering, University of Florence, Via di S. Marta 3, 50139, Florence, Italy
[d] BIOMATH, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000, Gent, Belgium

ARTICLE INFO

Keywords:
Greenhouse gas
Nitrous oxide
Wastewater treatment
PCA
Carbon footprint
Control

ABSTRACT

Emission of N_2O represents an increasing concern in wastewater treatment, in particular for its large contribution to the plant's carbon footprint (CFP). In view of the potential introduction of more stringent regulations regarding wastewater treatment plants' CFP, there is a growing need for advanced monitoring with online implementation of mitigation strategies for N_2O emissions. Mechanistic kinetic modelling in full-scale applications, are often represented by a very detailed representation of the biological mechanisms resulting in an elevated uncertainty on the many parameters used while limited by a poor representation of hydrodynamics. This is particularly true for current N_2O kinetic models. In this paper, a possible full-scale implementation of a data mining approach linking plant-specific dynamics to N_2O production is proposed. A data mining approach was tested on full-scale data along with different clustering techniques to identify process criticalities. The algorithm was designed to provide an applicable solution for full-scale plants' control logics aimed at online N_2O emission mitigation. Results show the ability of the algorithm to isolate specific N_2O emission pathways, and highlight possible solutions towards emission control.

1. Introduction

Wastewater treatment processes (WWTPs) can be considered to contribute to global warming in different ways and one of the most effective can be the emission of N_2O (*inter alia:* Law et al., 2012b; Kampschreur et al., 2009). At a global level, N_2O is a greenhouse (298 times more potent than CO_2) and ozone depleting gas of major concern (IPCC, 2013; Ravishankara et al., 2009) and the emissions from the wastewater treatment sector account for about 3% of global anthropogenic emissions (IPCC, 2014). Efforts were concentrated in understanding the specific bio-chemical processes responsible for N_2O production (Schreiber et al., 2012) and the WWTP design and operational factors impacting its emission (Daelman et al., 2013; Kampschreur et al., 2009; Monteith et al., 2005).

Measurements on full-scale WWTPs showed that N_2O emissions can represent more than 78% of a WWTP's carbon footprint (CFP) (Daelman et al., 2015). In addition to this, literature studies show the emission of up to 7% of the influent nitrogen load in the form N_2O (Kampschreur et al., 2008). However, the fraction of influent N that is emitted as N_2O shows important variations among plants (Kampschreur et al., 2008; Mampaey et al., 2013).

Considerable efforts have been put into modelling the ammonium oxidizing bacteria (AOB) pathways known to be responsible for N_2O production (i.e. AOB denitrification and incomplete NH_2OH oxidation) either with a single-pathway solution (Law et al., 2012a; Mampaey et al., 2013) or considering both AOB pathways (Ni et al., 2014). However, given the heterogeneity of WWTP process conditions, the potential variability of N_2O emissions, and the diversity of available models, consensus on model selection, dominant pathways and their implementation is yet to be reached.

At present, most advanced WWTPs have the availability of a large amount of data from sensors scattered over the plant, which is largely underexploited. Modern small WWTPs generate up to 500 signals, whereas larger ones typically register over 30 k signals (Olsson et al., 2014). These data are, in some sense, lost in most of the cases, as they are stored in databases and not transformed into actionable knowledge for

* Corresponding author.
E-mail addresses: Giacomo.Bellandi@polimi.it (G. Bellandi), SWeijers@dommel.nl (S. Weijers), Riccardo.Gori@dicea.unifi.it (R. Gori), Ingmar.Nopens@ugent.be (I. Nopens).

https://doi.org/10.1016/j.jenvman.2020.110219
Received 29 August 2019; Received in revised form 22 January 2020; Accepted 27 January 2020
Available online 2 March 2020

system optimization. As a result, the investments made for these sensors is only marginally payed back. Resources are thus dissipated on installing and maintaining on-line sensors without making proper use of potentially hidden information. Sub-optimal operation of WWTPs is still the norm rather than the exception (Villez et al., 2016).

In the literature several applications of data mining tools to wastewater treatment for process understanding, monitoring (fault detection), and control of industrial processes such as wastewater treatment are reported (Gernaey et al., 2004). Clustering techniques have been applied to characterize industrial wastewaters (Dürrenmatt and Gujer, 2011), while Pareto efficiency algorithms have been proven to be effective in defining the optimal sensor placement (Villez et al., 2016). Several variants of Principal Component Analysis (PCA) have been proven to be effective in the control of different aspects of sequencing batch reactors (SBRs) (Villez et al., 2008). However, to the best of the authors' knowledge, there is neither research nor applications of a data mining technique for N_2O production monitoring in WWTPs yet.

In this work, we propose a practical application of PCA coupled with clustering techniques aimed at providing a realistic solution for online implementation of N_2O emission monitoring and control. Results show that WWTP's historical data can be used to train a monitoring tool describing full-scale N_2O production. The variables that are normally measured on a full-scale can be used for full-scale N_2O minimization.

2. Materials and methods

2.1. Full-scale data

A one-month long dataset from one of the biological reactors of the WWTP of Eindhoven (The Netherlands) was used to identify potential clues related to the emissions of N_2O from this treatment step. The dataset, with a frequency ranging from 1 to 15 min, was collected during an extensive field measurement campaign. Data from the supervisory control and data acquisition (SCADA) system at the WWTP of Eindhoven, and N_2O concentrations measured in the liquid, were used to unravel possible relations among variables that are normally measured in WWTPs and N_2O liquid concentrations. The variables monitored from the SCADA system were NH_4, NO_x, dissolved oxygen (DO) and airflow (Q_{air}), while concentrations of N_2O in the liquid phase were measured by means of two full-scale probes (Unisense Environment, Denmark) located at the beginning and the end of the summer aeration package (Fig. 1).

The sensors of NH_4, NO_x, DO and the N_2O sensor 2, were located reasonably close to one another, whereas N_2O sensor 1 was located about 70 m upstream, at the beginning of the aeration compartment. This ensured a high resolution of information at the end of the aeration compartment and, at the same time, a monitoring location for the N_2O concentration entering the aerated zone.

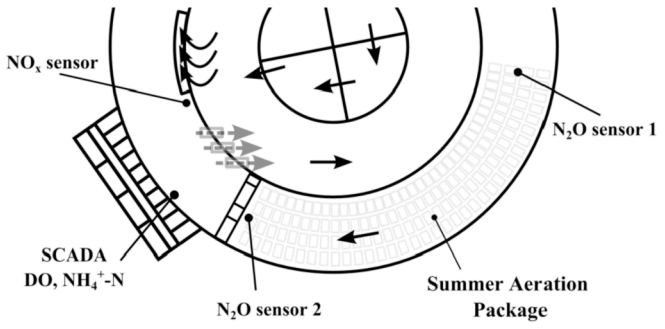


Fig. 1. Sensors location in the outer ring of the bioreactor of Eindhoven.

2.2. Data preparation

Pre-processing of the dataset is important in order to provide good quality and validated input data, free of potential biases for the following data mining steps. However, to ease its practical implementation in full-scale, data preparation needs to be kept minimal and robustly applicable to a large amount of data for a given WWTP. In this work, it was decided to use the information contained in the entire dataset to build a representative daily pattern for each variable.

After a first cleaning step with a moving average, all variables were grouped for each quarter of an hour contained in a day. Thus, extracting the ith percentile, allowed to build a distribution for each time step over the whole month. The best performing percentile for our purpose was observed to be the 70th, which returned a close representation of a typical daily pattern for every variable. The use of higher percentiles than the 70th, results in important data losses as the time series variability sensibly decreases. On the other hand, using smaller percentiles than the 70th, favoured the appearance of less frequent daily dynamics and emphasised noise.

Finally, a Kaiser-Meyer-Olkin (KMO) test was run to ensure suitability of the data for the application of PCA.

2.3. Data reduction

PCA is a multivariate statistical method for data mining and is often used for process understanding, monitoring (fault detection), and control of industrial processes such as WWTPs (Gernaey et al., 2004; Villez et al., 2008). The principle of PCA is to reduce the amount of information available to a smaller number of variables, or principal components (PCs), capable of explaining most of the variance of the dataset. In this way, it is possible to unravel hidden dependencies among known key variables.

A set of variables describing a certain process can be represented by a two dimensional matrix Z composed of N samples and M variables (NxM) (Equation (1).

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & z_{1,j} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & \dots & x_{2,M} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i,1} & x_{i,2} & \dots & x_{i,j} & \dots & x_{i,M} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{\cdot N,j} & \dots & x_{N,M} \end{bmatrix} \quad (1)$$

By calculating the scatter matrix (Equation (2) of this dataset, or the covariance matrix (Equation (3) of the standardized data $\tilde{X}$, it is possible to generate a newly referred dataset expressed by a new set of variables which are linear combinations of the original variables (Equation (4).

$$S = \sum_{1}^{N} (x_j - \bar{x})(x_j - \bar{x})^T \quad (2)$$

$$\sum = cov(\tilde{X}) = \frac{\tilde{X}^T \cdot \tilde{X}}{N} \quad (3)$$

Of this new set of variables, the coefficients are the principal components, i.e. the new reference system is defined.

$$\begin{aligned} t_{i,1} &= \tilde{\boldsymbol{X}}_{i,\cdot} \cdot \boldsymbol{p}_{\cdot,1} = \tilde{x}_{i,1} \cdot p_{1,1} + \tilde{x}_{i,2} \cdot p_{2,1} + \dots + \tilde{x}_{i,2} \cdot p_{M,1} \\ t_{i,2} &= \tilde{\boldsymbol{X}}_{i,\cdot} \cdot \boldsymbol{p}_{\cdot,2} = \tilde{x}_{i,1} \cdot p_{1,2} + \tilde{x}_{i,2} \cdot p_{2,2} + \dots + \tilde{x}_{i,2} \cdot p_{M,2} \\ t_{i,C} &= \tilde{\boldsymbol{X}}_{i,\cdot} \cdot \boldsymbol{p}_{\cdot,C} = \tilde{x}_{i,1} \cdot p_{1,C} + \tilde{x}_{i,2} \cdot p_{2,C} + \dots + \tilde{x}_{i,2} \cdot p_{M,C} \end{aligned} \quad (4)$$

An interesting property of the $\boldsymbol{p}_{\cdot,c}$ vectors is that they are the eigenvectors of the covariance matrix S and the corresponding eigenvalues λ_c (Equation (5) are equal to the variance of the corresponding linear combinations (Johnson and Wichern, 1992).

$$\lambda_c = var(\boldsymbol{t}_{\cdot,c}) = var(\boldsymbol{p}_{\cdot,c} \cdot \tilde{\boldsymbol{X}}) \quad (5)$$

Thus, by sorting the eigenvectors according to their eigenvalues and

selecting the first c of them, one has exactly determined the order of the PCs. Another important characteristic of the eigenvalues is that the relative variance (RV) captured by the c^{th} component can be expressed as in (Equation (6):

$$RV = \frac{var(\boldsymbol{t}_{.c})}{tr(\boldsymbol{S})} = \frac{\lambda_c}{\sum_{b=1}^{\max(M,N)} \lambda_b} \quad (6)$$

The equation can be explained as the proportional amount of variance captured by the c[th] PC, and is equal to the ratio of its corresponding eigenvalue to the sum of all eigenvalues (Johnson and Wichern, 1992). The relative cumulative variance (RCV) of all components is the sum of the relative variances of each component (Equation (7).

$$RCV = \frac{\sum_{c=1}^{C} var(\boldsymbol{t}_{.c})}{tr(\boldsymbol{S})} = \frac{\lambda_b}{\sum_{b=1}^{\max(M,N)} \lambda_b} \quad (7)$$

Once the PCs are identified there are different methods used for data reduction, but the common target is to capture a maximal amount of variance with a minimum number of dimensions. It is generally recommended to select the number of PCs explaining at least 70% of the variance of the original dataset (Villez et al., 2008).

2.4. Clustering

Most commonly applied clustering techniques are based on two popular methods, i.e. the iterative square-error partitional clustering and the agglomerative hierarchical clustering. Clustering algorithms in literature can generally be classified into two types: hierarchical clustering and partitional clustering. Hierarchical clustering methods include agglomerative algorithms and are more efficient in handling noise and outliers than partitional algorithms. On the other hand, partitional clustering admit relocation of points from a different cluster thus allowing to correct initial partitions in later stages.

In addition to hierarchical and partitional clustering, a large number of methods are available from the literature (Han and Kamber, 2001). One example among the most implemented solutions alternative to hierarchical and partitional clustering, is the density-based clustering. This method groups a dataset based on specific criterion of the density functions, defining density as the number of objects in a particular neighborhood of a dataset.

Three clustering techniques were chosen to be applied to the results of this work, being among the most widely accepted by the scientific community (Pedregosa et al., 2011), in order to evaluate the capabilities of grouping relevant information isolated by the PCs. In particular, *K*-means and the agglomerative clustering are two well-known algorithms already tested in wastewater treatment (Dürrenmatt and Gujer, 2011; López García and Machón González, 2004), while, to the author's best knowledge, recent improvements of density based clustering methods, i.e. hierarchical methods, have never been used in wastewater treatment applications.

2.4.1. K-means

The *K*-means algorithm has been traditionally used as a non-hierarchical method for the analysis of the data prior to more rigorous methods such as hierarchical methods or PCA. *K*-means normally divides the dataset in a number of pre-defined clusters (K) and, as a result of the Ward's method, iteratively minimizes the sum of squared errors within the cluster. For doing so, at the ith iteration each point x is assigned to a cluster based on the following relation.

$$x \in c_j(k) \quad if \quad \|x - z_j(k)\| < \|x - z_i(k)\| \quad (8)$$

With $c_j(k)$ the set of samples with center $z_j(k)$. At this point, the sum of squared distances for all points belonging to the new cluster center is minimized with the sample mean of $c_j(k)$ (*inter alia:* Han and Kamber, 2001).

2.4.2. Agglomerative

This algorithm uses a bottom-up approach, therefore starting with each sample being a separate cluster itself. Successively, groups are merged according to a distance measure, similarly to the *K*-means case, this is done minimizing the sum of squared differences between two clusters (Ward's method) or using the maximum distances between all observations of the different sets (maximum linkage method), but tackling the objective with a hierarchical approach. This recursively merges the pair of clusters that minimally increase a given linkage distance (Murtagh and Legendre, 2014). The classification may stop when all samples are in a single group or when the required number of clusters is reached. Nonetheless, with this method the statistical distance between each cluster can be visualized.

2.4.3. HDBSCAN

Campello et al. (2013), demonstrated that extending the original density based method (DBSCAN) with a hierarchical clustering algorithm, it was possible to achieve an improved application of the DBSCAN. This is one of the latest developments in clustering algorithms providing improvements in the results of a wide variety of data (McInnes et al., 2017; Melvin et al., 2016). HDBSCAN has been observed to be useful for determining a system's stability by grouping stable systems in few bins (Melvin et al., 2016). In this work, HDBSCAN is considered for classification of the PCA output given its exceptional results reported in the literature.

3. Results and discussion

Fig. 2 shows the time series of the variables acquired from the WWTP of Eindhoven for this study. It is noticeable how the peaks in N_2O concentration in the liquid phase (and therefore its actual production) correspond to peaks in NH_4, however, the contrary cannot be stated. The production of N_2O is in fact related to multiple interchanging factors and therefore qualifies as a multivariate problem.

The N_2O sensor 1 (Fig. 2, top graph), located at the beginning of the aerated compartment and the first sensor according to the flow direction, always shows a higher concentrations compared to the N_2O sensor 2. This is mostly due to the stripping effect of the aeration package, but the high concentration of N_2O at the end of the anoxic zone confirms its production prior to entering the aerobic zone.

Given the relevant concentrations observed by the N_2O sensor 2 about 70 m downstream at the end of the aerated compartment despite stripping effect by diffused aeration, it can be stated that N_2O production also occurs within the aerated zone. This means that multiple pathways of N_2O production occur in the different zones of a biological tank. The concentrations recorded by the N_2O sensor 1 are most likely caused by the activity of AOB in DO limiting conditions, while the signal recorded

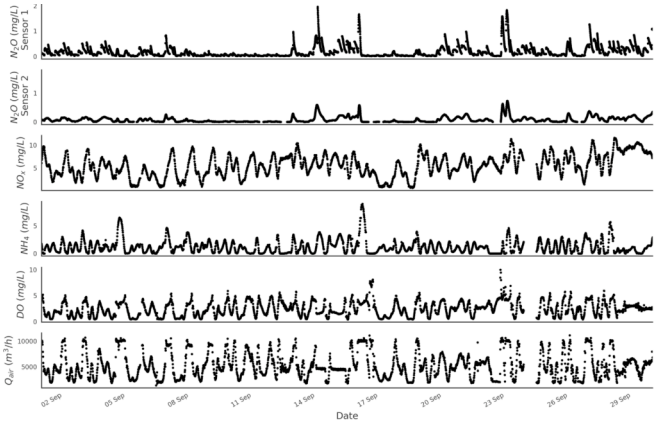


Fig. 2. Dataset of an entire month for a bioreactor of the WWTP of Eindhoven.

by the N_2O sensor 2, given the non-limiting levels of DO also during N_2O peaks, are most likely to be caused by autotrophic NH_2OH oxidation. In this view, another relevant element is the NO_2 concentration, reported here together with NO_3 as NO_x, which is strongly influencing the N_2O production (Peng et al., 2015).

Also included in this dataset is Q_{air} supplied as m^3/h over the aeration package surface (Fig. 2, bottom). Q_{air} is obviously strongly linked primarily with DO, and then with NH_4, given that the aeration control is primarily based on the NH_4 concentration in the tank.

These variables potentially contain most of the information required to develop a monitoring tool for N_2O aimed at minimizing its emission.

The dataset reported in Fig. 2 contains high quality data in terms of time frequency and sensor status. This is not common for a general WWTPs data stream since periods of missing data for maintenance or failures of probes are rather frequent. The dataset shows a period of regular operation of the plant, good data quality of the sensors without major failures (only exception between 24 and 25th September), and was acquired during a month of good performance in dry weather. This represents a good example of training dataset for the application of a data mining technique.

3.1. Pre-processing

A generally applicable smoothing function including a moving average and the extraction of the 70th percentile from the time-wise distribution was implemented. This approach was observed to remove outliers and those variabilities in the dataset having higher frequency than the plant's biological dynamics which are interesting for the study. The resulting dataset was composed of representative data describing a characteristic day for each of the variables (Fig. 3).

A KMO test on the resulting daily pattern, returned a score of 0.53, just coping with the minimum acceptable requirements (Kaiser, 1974) for the application of the PCA. On the other hand, despite the high volume of data, the KMO scored 0.41 on the raw dataset, confirming its unsuitability for direct PCA application without preprocessing.

By definition, the 70th percentile of a distribution returns the value below which can be found 70% of the observations. This eliminates the most infrequent absolute daily peaks and valleys, but leaves the general daily pattern of the dataset and its internal variability. This is the reason why the relative concentrations of NH_4, NO_x and DO in Fig. 3 are somewhat higher than one would normally expect in a correctly managed bioreactor, i.e. concentrations of NO_x and NH_4 peak above 8 and 2 mg/L. In this way, treating all variables the same, means maintaining the intrinsic information of a daily pattern for all variables even though relative values are slightly higher than in reality. However, this raises no concern in terms of the application of the PCA since this technique uses the correlation matrix (or the scatter matrix) to derive relations between standardized variables and therefore is not affected by the relative value of a variable.

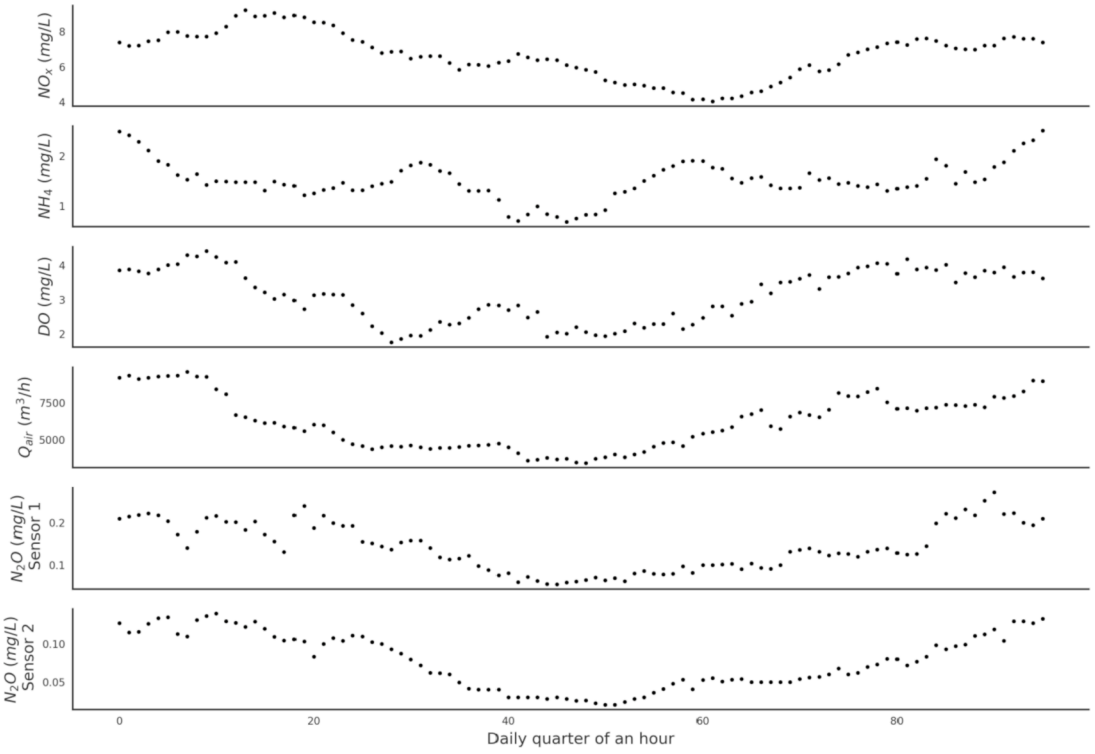
3.2. Application of the Principal Component Analysis

All variables were fed to the PCA except for N_2O measurements. In this way, the information contained in those variables can be effectively tested for its capability of describing N_2O production.

The first two PCs were selected as they explained more than 90% of the variance of the entire dataset (i.e. the first, second, and third PC explained 68%, 23%, and 9% of the total variance respectively). Therefore, two PCs can be considered to describe most of the variability of the original variables.

The results of the first two PCs are reported in Fig. 4. It is interesting to notice that two main groups of data points can be already distinguished at the positive and negative sides of the x axis. The measurements of the liquid concentrations of N_2O were used to color the data points according to the concentration measured, so to ease the visualization of highly emitting clusters. The left graph reports the values of the PCs colored according to the N_2O sensor 1, while the data points in the right graph were colored according to the concentration measured by the N_2O sensor 2.

The red vectors reported on the scatterplot indicate the degree of



Figs. 3. 15 min 70th percentile of the raw dataset. This represents a typical daily 24 h pattern of the WWTP dynamics, and is the input of the PCA.

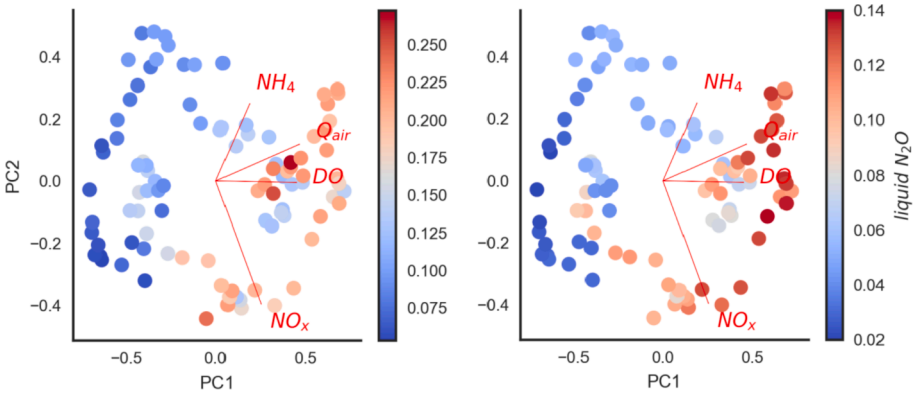


Fig. 4. Scatterplot of the first two PCs labeled according to the N_2O concentrations in the liquid of sensor 1 (left) located at the beginning of the aeration compartment and sensor 2 (right) located at the end of the aeration compartment.

correlation (variance explained) by a certain PC with respect to each original variable. PC1 is more correlated with Q_{air} and DO while PC2 describes the behavior of the nitrogen species. A small correlation of PC1 with NH_4 and NO_x was expectable, given the effect of DO on the nitrogen transformation, as well as the small correlation between Q_{air} and PC2 in the direction of NH_4, as a result of the NH_4 based control. Finally, it is important to mention that in the direction of each vector the values of the respective variable increase, hence, the expectable opposite directions of NH_4 and NO_x. The vicinity of the vectors relative to Q_{air} and DO suggests that they bring very similar information to the results.

Both N_2O sensors' highest concentrations are mostly clustered on the positive side of PC1, indicating that Q_{air}, and ultimately DO, has a high impact on N_2O production. However, the cluster forming for both graphs in Fig. 4 at the negative side of PC2 indicates that NO_x has also a strong importance on N_2O formation. These two groups of data are already suggesting two main types of pathways possible for N_2O production. Interestingly, the N_2O sensor 2, being physically closer to the rest of the sensors in the tank considered for the PCA, is returning a better defined separation between high and low N_2O concentrations (Fig. 4, right).

The two groups of data points identifiable for high N_2O concentrations, can be interpreted as the interchange of the two main pathways already observed to be occurring in this plant (Bellandi et al., 2018; Porro et al., 2017), i.e. AOB denitrification and incomplete NH_2OH oxidation pathways. The data grouping close to the tip of the DO and Q_{air} vectors are related to the highest DO concentrations observed in the time series, and therefore most likely to be linked to the incomplete NH_2OH oxidation pathway.

The data grouping close to the tip of the NO_x axis and closer to the zero of PC1, are more likely to correspond to high NO_2 concentrations as NO_2 is also inherently linked to DO (Peng et al., 2015) since lower DO concentrations can lead to higher NO_2 concentrations due to the difference in oxygen half-saturation index between AOB and NOB (Hanaki et al., 1990; Mota et al., 2005). This suggests a possible AOB preference of NO_2 as the electron acceptor over DO (Bock et al., 1995; Kampschreur et al., 2009) and the production of N_2O due to AOB denitrification. In addition to this, since red dots of N_2O sensor 2 reach to the negative side of PC2, this can correspond to more limiting DO concentrations associated with the AOB denitrification pathway.

3.3. Clustering

In view of applying the PCA results on a full-scale control, a clustering technique is necessary for automating the recognition of the different N_2O production pathways. The three clustering methods introduced were applied to the PCA results with the aim of recognizing the different clusters in terms of N_2O formation and possibly extract more information.

The different clusters are colored differently to distinguish the different groups. Colors are not specific of a single cluster.

3.3.1. *K-means*

The main input of the *K*-means clustering method is the number of clusters. The minimum number of interesting groups for the purpose of this study is 3 if we focus on the recognition of the two main N_2O production pathways (i.e. AOB denitrification and incomplete NH_2OH oxidation) and the zone of low N_2O production. A number of 4 clusters was also used to further test the algorithm.

Initializing *K*-means with 3 clusters (Fig. 5, left), it is interesting to see how the resulting clusters at equilibrium are already nicely divided among the groups previously indicated, corroborating with the initial interpretation of the raw PCA results. However, the points closer to the NH_4 vector should not belong to the cluster of high emissions for incomplete NH_2OH oxidation. The cluster relative to AOB denitrification is instead rather well defined, including also one of the points in the negative side of PC1 known to have elevated N_2O concentration.

When 4 clusters were used for initialization (Fig. 5, right), the resulting cluster responsible for N_2O formation due to incomplete NH_2OH oxidation was defined better than in the previous case, although some of the points close to the NH_4 vector are still included. The cluster attributable to AOB denitrification remains the same, while the low emission cluster, closer to the PC2 axis is divided in two as expected from the need of dividing the space in 4.

3.3.2. *Agglomerative*

Using the Ward's method, four clusters were needed for initialization in order for the algorithm to distinguish the two groups of data known to describe the two main N_2O production pathways, i.e. AOB denitrification and incomplete NH_2OH oxidation (Fig. 6, left). Without the initialization of the algorithm to target four final clusters, it was not possible to achieve this distinction. This resulted also in the division in two clusters of the group of data linked to small N_2O concentrations in the same fashion as for the *K*-means method.

Initializing the agglomerative clustering to target 3 clusters with the maximum linkage method for the iterated merging of initial clusters, the algorithm isolated all three main groups of data, i.e. the AOB denitrification, the incomplete NH_2OH oxidation and the area of low N_2O production (Fig. 6, right).

The Ward's algorithm performed better in terms of time, taking only 1/3 of the time needed for the maximum linkage method. The difference in time is probably due to the fact that the Ward's method was able stop one iteration earlier (ending with 4 clusters instead of 3). However, although each algorithm performs in the order of few milliseconds, in terms of efficiency for future implementation this can be a useful selection criterion to choose between the algorithms.

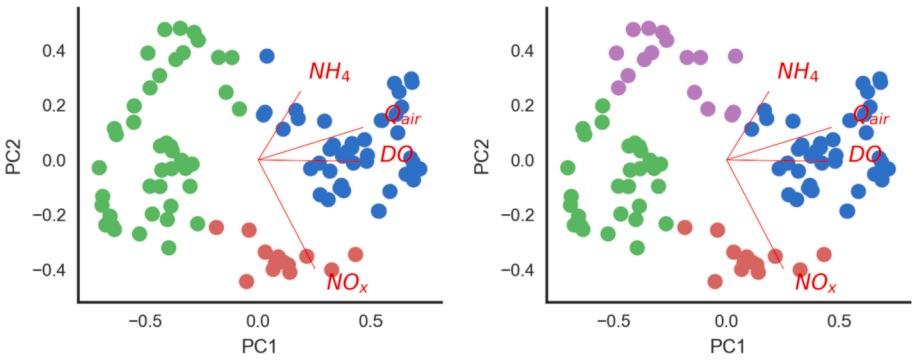


Fig. 5. *K*-Means with 3 clusters (left) and 4 clusters (right). Colors are randomly assigned only to distinguish clusters.

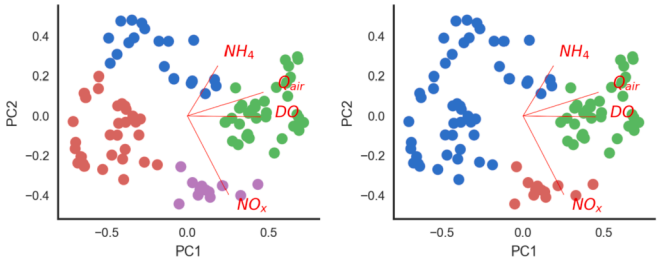


Fig. 6. Agglomerative clustering with Ward's (left) and maximum linkage (right) methods. Colors are randomly assigned only to distinguish clusters.

Interestingly, for both merging algorithms, the data points close to the NH_4 vector are correctly grouped with the cluster of data relative to low N_2O concentration, corroborating with the results discussed in the PCA section. However, few data points corresponding to the negative part of PCs and characterized with a rather high N_2O concentration by the N_2O sensors, were included in the low N_2O concentration cluster by both algorithms. Finally, the two clusters relative to high N_2O concentrations coincide for the two methods.

3.3.3. *HDBSCAN*

This clustering method, distinct from the former ones, requires as input the minimum number of points to be considered as a cluster. With this feature, the HDBSCAN output can also consider the existence of data points not belonging to any of the clusters (reported in black).

With a minimum cluster size of 4 data points (Fig. 7, left) the HDBSCAN distinguishes between the two clusters of known high N_2O concentration. Interestingly, between these two clusters there are two black data points not belonging to either of the clusters. This is an interesting result since it allows for the existence of points of transition between one cluster and another. However, the data points close to the NH_4 vector, known to belong to low N_2O concentrations or at least expected to be classified in a transition zone, are instead grouped with the high N_2O concentration due to the incomplete NH_2OH oxidation pathway.

In the negative part of both PC1 and PC2, the high N_2O concentration cluster linked to AOB denitrification, and the rest of the clusters close to the PC2 axis, are divided by three black data points that the previous clustering methods grouped uncertainly. In fact, these three data points seem to lay in a transition zone that only the HDBSCAN is able to detect.

The low N_2O concentration zone, in the negative side of PC1, is divided into four clusters (Fig. 7, left). Although this subdivision is allowed by the minimum cluster size, no physical meaning could be found for these different clusters. This granularity of clusters in this part of the graph disappears when increasing the minimum cluster size to 9 (Fig. 7, right).

Increasing the minimum cluster size to 9 (Fig. 7, right), sensibly decreases the number of clusters on the left part of the plot (corresponding to the lowest concentrations of N_2O) while maintaining the two clusters relative to high N_2O concentration (center and right of PC1). Interestingly, the points close to the NH_4 vector are still classified within the cluster of high N_2O concentration due to incomplete NH_2OH oxidation, but more data points were addressed (in black) to the transitional points. Therefore, this initialization performed slightly better than the minimum cluster size of 4.

3.4. *Overall evaluation*

All clustering methods were able to recognize differences among those clusters generated by the two PCs resulting from the application of the PCA. The *K*-means method could sufficiently isolate the main clusters known to be linked to specific N_2O production pathways. However,

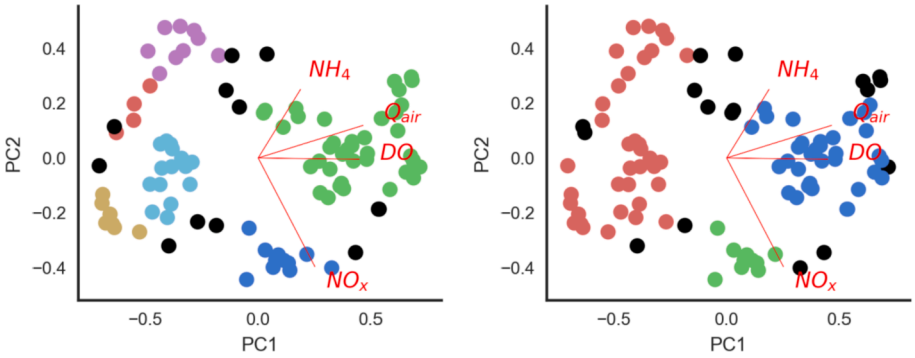


Fig. 7. HDBSCAN with minimum clusters size 4 (left) and minimum clusters size 9 (right). Colors are randomly assigned only to distinguish clusters. In black the data points not attributed to any cluster.

some imprecisions in the classification of data points close to the edge of two neighboring clusters were observed. In this view, the agglomerative method was able to identify with more precision those data points that were erroneously addressed by *K*-means to the clusters of higher emission. On the other hand, the HDBSCAN method does not coerce the attribution of boundary data points to a cluster and allows to consider the existence of transitional zones. This is an important point in monitoring full-scale WWTPs as conditions in AS tanks are highly dynamic and transitions from one state to another are continuously happening.

For an online application, based on their good performances, at the moment both the agglomerative method and the HDBSCAN are equally applicable. For discriminating between one method or the other would need more testing.

For a practical online application, the clustering method chosen, could be initially integrated in a supervisory system to alert operators on the possibility of an important N_2O production. Based on the PCA model built with the training dataset, the online data stream can be projected on the PCs space, thus, potentially revealing in which of the clusters related to N_2O production the system is. Based on the cluster, specific instructions can be proposed. For instance, in the case that the system would be directing to the cluster responsible for N_2O production due to incomplete NH_2OH oxidation, the operator could evaluate the option of reducing the DO, thus limiting this reaction. On the other hand, if the system would reveal to be shifting towards the cluster responsible for N_2O production due to AOB denitrification, the operator could be prompted to evaluate the possibility of increasing the DO concentration. Simple instructions or suggestions deriving from a thorough analysis of WWTP data in real time.

4. Conclusions

PCA was applied to a dataset of the WWTP of Eindhoven for detecting a possible relation between variables known to be highly related to N_2O production. A PCA model was defined after a small pre-processing step defining the most typical behavior observed in one entire month for all variables. The PCA model could separate the two main N_2O production pathways by using two PCs. The results show that the two PCs could isolate the main known relations between N_2O production and plant operation. Both the AOB denitrification and the incomplete NH_2OH oxidation N_2O production pathways were nicely identifiable.

In view of applying these results to full-scale, three clustering methods were tested for automating the identification of the different regions of the PCA scatterplot. The *K*-means method could sufficiently separate between the two main N_2O production pathways, although some of the edges of the clusters included data points that could be questionable. Both the HDBSCAN and the agglomerative methods successfully differentiated between the two N_2O production pathways excluding irrelevant points that were difficult to detect.

Results confirm the potential for defining a new monitoring system for N_2O emissions based on historical plant data. Operators could be provided with important information deriving from a thorough analysis of the AS tank, this in view of a full integration in a SCADA system. Future implementations should consider the introduction of MPCA to increase the informative content of the original dataset and limit the loss of information in the pre-processing step.

Author contributions section

Giacomo Bellandi: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing, visualization; Stefan Weijers: conceptualization, resources, project administration, funding acquisition, writing review; Riccardo Gori: resources, project administration, funding acquisition, writing review; Ingmar Nopens: supervision, conceptualization, resources, project administration, funding acquisition, writing review;

References

Bellandi, G., Porro, J., Senesi, E., Caretti, C., Caffaz, S., Weijers, S., Nopens, I., Gori, R., 2018. Multi-point monitoring of nitrous oxide emissions in three full-scale conventional activated sludge tanks in Europe. Water Sci. Technol. 77, 880–890.

Bock, E., Schmidt, I., Stuven, R., Zart, D., 1995. Nitrogen loss caused by denitrifying Nitrosomonas cells using ammonium or hydrogen as electron donors and nitrite as electron acceptor. Arch. Microbiol. 163, 16–20.

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates, pp. 160–172.

Daelman, M.R.J., van Voorthuizen, E.M., van Dongen, L.G.J.M., Volcke, E.I.P., van Loosdrecht, M.C.M., 2013. Methane and nitrous oxide emissions from municipal wastewater treatment – results from a long-term study. Water Sci. Technol. 67, 2350.

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 N2O emissions from a full-scale municipal wastewater treatment plant. Sci. Total Environ. 536, 1–11.

Dürrenmatt, D.J., Gujer, W., 2011. Identification of industrial wastewater by clustering wastewater treatment plant influent ultraviolet visible spectra. Water Sci. Technol. 63, 1153.

Gernaey, K.V., Van Loosdrecht, M.C.M., Henze, M., Lind, M., Jørgensen, S.B., 2004. Activated sludge wastewater treatment plant modelling and simulation: state of the art. Environ. Model. Software 19, 763–783.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques. Los Atos, CA.

Hanaki, K., Chalermraj, W., Shinichiro, O., 1990. Nitrification at low levels of dissolved oxygen with and without organic loading in a suspended-growth reactor. Water Res. 24, 297–302.

IPCC, 2013. Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (Cambridge).

IPCC, 2014. IPCC 2014, Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Johnson, R.A., Wichern, D.W., 1992. Applied Multivariate Statistical Analysis (New York).

Kaiser, H.F., 1974. An index of factorial simplicity. Psychometrika 39, 31–36.

Kampschreur, M.J., Temmink, H., Kleerebezem, R., Jetten, M.S.M., van Loosdrecht, M.C.M., 2009. Nitrous oxide emission during wastewater treatment. Water Res. 43, 4093–4103.

Kampschreur, M.J., van der Star, W.R.L., Wielders, H.a., Mulder, J.W., Jetten, M.S.M., van Loosdrecht, M.C.M., 2008. Dynamics of nitric oxide and nitrous oxide emission during full-scale reject water treatment. Water Res. 42, 812–826.

Law, Y., Ni, B.J., Lant, P., Yuan, Z., 2012a. N2O production rate of an enriched ammonia-oxidising bacteria culture exponentially correlates to its ammonia oxidation rate. Water Res. 46, 3409–3419.

Law, Y., Ye, L., Pan, Y., Yuan, Z., 2012b. Nitrous oxide emissions from wastewater treatment processes. Philos. Trans. R. Soc. B Biol. Sci. 367, 1265–1277.

López García, H., Machón González, I., 2004. Self-organizing map and clustering for wastewater treatment monitoring. Eng. Appl. Artif. Intell. 17, 215–225.

Mampaey, K.E., Beuckels, B., Kampschreur, M.J., Kleerebezem, R., van Loosdrecht, M.C.M., Volcke, E.I.P., 2013. Modelling nitrous and nitric oxide emissions by autotrophic ammonium oxidizing bacteria. Environ. Technol. 34, 1555–1566.

McInnes, L., Healy, J., Astels, S., 2017. HDBSCAN: hierarchical density based clustering. J. Open Source Softw. 2, 205.

Melvin, R.L., Godwin, R.C., Xiao, J., Thompson, W.G., Berenhaut, K.S., Salsbury, F.R., 2016. Uncovering large-scale conformational change in molecular dynamics without prior knowledge. J. Chem. Theor. Comput. 12, 6130–6146.

Monteith, H.D., Sahely, H.R., MacLean, H.L., Bagley, D.M., 2005. A rational procedure for estimation of greenhouse-gas emissions from municipal wastewater treatment plants. Water Environ. Res. a Res. Publ. Water Environ. Fed. 77, 390–403.

Mota, C., Head, M., Ridenoure, J., Cheng, J., de los Reyes, F., 2005. Effects of aeration cycles on nitrifying bacterial populations and nitrogen removal in intermittently aerated reactors. Appl. Environ. Microbiol. 71, 8565–8572.

Murtagh, F., Legendre, P., 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. 31, 274–295.

Ni, B.J., Peng, L., Law, Y., Guo, J., Yuan, Z., 2014. Modeling of nitrous oxide production by autotrophic ammonia-oxidizing bacteria with multiple production pathways. Environ. Sci. Technol. 48, 3916–3924.

Olsson, G., Carlsson, B., Comas, J., Copp, J., Gernaey, K.V., Ingildsen, P., Jeppsson, U., Kim, C., Rieger, L., Rodríguez-Roda, I., Steyer, J.-P., Takács, I., Vanrolleghem, P.A., Vargas, A., Yuan, Z., Åmand, L., 2014. Instrumentation, control and automation in wastewater – from London 1973 to Narbonne 2013. Water Sci. Technol. 69, 1373.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peng, L., Ni, B.J., Ye, L., Yuan, Z., 2015. The combined effect of dissolved oxygen and nitrite on N2O production by ammonia oxidizing bacteria in an enriched nitrifying sludge. Water Res. 73, 29–36.
Porro, J., Bellandi, G., Rodriguez-Roda, I., Deeke, A., Weijers, S., Vanrolleghem, P.A., Comas, J., Nopens, I., 2017. Developing an artificial intelligence-based WRRF nitrous oxide mitigation road map: the Eindhoven N2O mitigation case study. In: WEFTEC 2017, pp. 1703–1715. Chicago, IL (USA).
Ravishankara, A.R., Daniel, J.S., Portmann, R.W., 2009. Nitrous oxide (N2O): the dominant ozone-depleting substance emitted in the 21st century. Science 326, 123–125.
Schreiber, F., Wunderlin, P., Udert, K.M., Wells, G.F., 2012. Nitric oxide and nitrous oxide turnover in natural and engineered microbial communities: biological pathways, chemical reactions, and novel technologies. Front. Microbiol. 3.
Villez, K., Ruiz, M., Sin, G., Colomer, J., Rosén, C., Vanrolleghem, P.A., 2008. Combining multiway Principal Component Analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR processes. Water Sci. Technol. 57, 1659–1666.
Villez, K., Vanrolleghem, P.A., Corominas, L., 2016. Optimal flow sensor placement on wastewater treatment plants. Water Res. 101, 75–83.