Informed Feature Selection Algorithm for Optimizing Instrumental Odour Monitoring Systems in Drones

Javier Alonso-Valdesueiro*[a], Alessandro Benegiamo*[b], Javier Burgues[a], Albert Vidal[b], Agustin Gutierrez-Galvez[a], Santiago Marco[a,b]

[a]*Department of Electronic and Biomedical Engineering, University of Barcelona, Martí i Franquès 1,, Barcelona, 08028, Catalonia, Spain*
[b]*Signal and Information Processing for Sensing Systems Group, Institute for Bioengineering of Catalonia, Baldiri i Reixac, 4 Torre I, Barcelona, 08028, Catalonia, Spain*

Abstract

This study demonstrates how advanced feature selection significantly enhances the performance of Instrumental Odour Monitoring Systems (IOMS) on-boarded in drones. By employing the Nested Sequential Forward Feature Selection algorithm in conjunction with the Interval Partial Least Squares (NiPLS) regression model, a notable reductions in the Root Mean Square Error of Prediction (RMSEP) and the Limits of Agreement (LoA) at a 95% Confidence Interval (CI) are achieving during test of the quantification model. The feature selection process is applied in two stages: first, to identify the most relevant sensors, and second, to determine the optimal acquisition periods. This method was validated using data from an IOMS mounted on an octocopter drone over a Wastewater Treatment Plant (WWTP). Results show a reduction in RMSEP from 2.1x to 1.8x and a decrease in maximum LoA at 95% CI from 5x to 3.4x, with a correlation coefficient of approximately 0.9 for measurements from various sources within a particular WWTP.

Keywords: Instrumental Odour Monitoring System, Chemical sensors, Feature selection, Odour quantification, On-boarded system, Drone Instrumentation

1. Introduction

Odour quantification in industrial environments has gained attention due to legal regulations [1, 2, 3]. The most common method, dynamic olfactometry, involves a human panel that evaluates the odor levels of air samples [4]. However, this method has limitations, including a high Limit of Agreement (LoA) and significant costs.

Instrumental Odour Monitoring Systems (IOMSs) have been deployed as alternatives, particularly in Wastewater Treatment Plants (WWTPs) [5, 6]. IOMSs onboard drones can create 3D odour maps and further reduce costs [7, 8]. These systems use chemical sensors to measure gas concentrations, and dynamic olfactometries are performed on collected samples to calibrate multivariate prediction models.

Partial Least Squares (PLS) [9] is commonly used for odour prediction, with performance improved by feature reduction methods such as Variable Importance in Projection (VIP) [10]. VIP scores help select relevant features for model training, followed by validation using Cross-Validation (CV) [11].

Despite these methods, the Root Mean Square Error of Prediction (RMSEP) remains high, limiting the use of IOMSs for regulatory purposes. Recent efforts focus on reducing RMSEP with new algorithms for small datasets [12, 13, 14, 15].

This paper presents a new approach that uses a small dataset from an IOMS onboard a DJI drone over a WWTP. A two-step feature selection process, combining Nested Sequential Forward Selection [16, 17] and Interval Partial Least Squares [18], is proposed to improve model performance.

The presented contribution is organized as follows. Firstly, in the Methods section, the onboarded IOMS and the measurement campaign are presented. In this section, the data processing workflow is also described in detail. Secondly, in the Results and Discussion section, the performance figures of the proposed workflow are summarized when applied to the dataset obtained in the measurement campaign. Finally, in the Conclusions section, a summary of the NiPLS performance is presented, and its future applications in the field of on-

boarded IOMSs are described.

2. Methods

2.1. Onboarded IOMS

Fig.1(a) shows the block diagram of the onboarded IOMS performing measurements during flights.

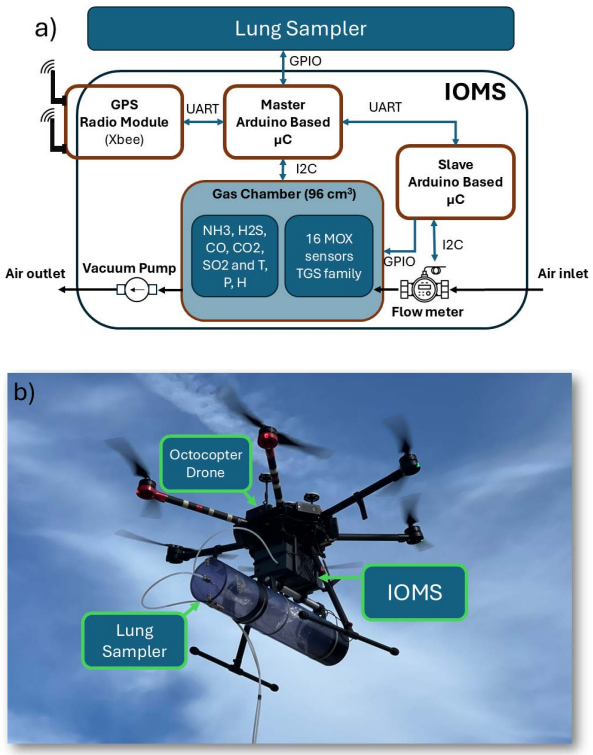

Figure 1: IOMS onboarded in the flying drone. (a) Block Diagram of the onboarded IOMS. The drawing includes the communications modules, the lung sampler and the batteries. (b) Onboarded Hardware in the DJI octocopter. The figure shows the drone, the IOMS and the lung sampler.

The Instrumental Odour Monitoring System (IOMS) features a gas chamber containing 5 electrochemical sensors and 16 MOX sensors within a 96 cm^3 volume, as detailed in Tables 1 and 2. The electrochemical sensors, sourced from Alphasense Inc., are connected to a Libellium microcontroller via proprietary analog frontends. The MOX sensors, part of the Figaro TGS series, include the models TGS2600, TGS2602, TGS2611, and TGS2620. These sensors are heated to various temperatures by applying different voltages through PWM signals, as described in Table 1 [19, 20]. The PWM signals are generated by an Arduino device and conditioned

	Technology	Range	Accuracy	Resp. time (T_{90})
Flow rate	Ultrasonic	-33 to +33 L/min	±3% m.v.	<1 s
CO_2	NDIR	0 to 5000 ppm	±100 ppm	<60 s
CO	Electrochemical	0 to 100 ppm	±0.5 ppm	<20 s
H_2S	Electrochemical	0 to 20 ppm	±0.1 ppm	<20 s
NH_3	Electrochemical	0 to 100 ppm	±0.5 ppm	<90 s
SO_2	Electrochemical	0 to 20 ppm	±0.1 ppm	<45 s

Table 1: Specifications of electrochemical, NDIR and Flow sensor

Sensor	Model	Target gases	Heater voltage (V)
M1,2,3,4	TGS 2600	H_2, CO, Ethanol	1.6,3.2,4.0,4.9
M5,6,7,8	TGS 2602	H_2S, NH_3, Toluene	1.6,3.2,4.0,4.9
M9,10,11,12	TGS 2611	CH_4, Hydrocarbons	1.6,3.2,4.0,4.9
M13,14,15,16	TGS 2620	Alcohols, ketones	1.6,3.2,4.0,4.2

Table 2: Specifications of the MOX sensors included in the slave board

for power transfer using a custom-made electronic platform.

Figure 1(b) illustrates all the equipment mounted on a DJI octocopter during flight. This setup includes the IOMS and a lung sampler for collecting sample bags filled with environmental air. Both the IOMS and the lung sampler are connected to their inputs via 10-meter PTFE tubes. This length is sufficient to prevent the downwash effect from the drone [21, 22]. The tubes ensure that air is collected from the same environment as the drone flies over the plant.

Air is drawn in using a vacuum pump, and the airflow is measured by a flow sensor, as detailed in Table 1. Data from electrochemical sensors, MOX sensors, and the flow sensor are collected by the Libellium microcontroller. This data, along with battery status, GPS location, and other system status information, is framed and transmitted via an Xbee module to a ground-based laptop.

The laptop runs custom-made software that collects data sent by the IOMS via Xbee, records it, and displays the results on the screen. The software also allows the user to command the IOMS to activate the lung sampler, synchronizing sensor readings with the air filling the bag inside the lung sampler. Bags are typically collected at various locations within the WWTP and transported to a certified company for dynamic olfactometric analysis. Once the odour quantification is provided by the company, the sensor data from each location and the odour quantification in $^{OU_E}/_{m^3}$ are used to calibrate an odour prediction model. For more information about the IOMS, consult previous work [7, 23].

2.2. Measurement Campaign

Fig.2 shows the WWTP where the measurement campaign was performed. It is located in the south-west of

Spain and presents a typical topology of a WWTP.

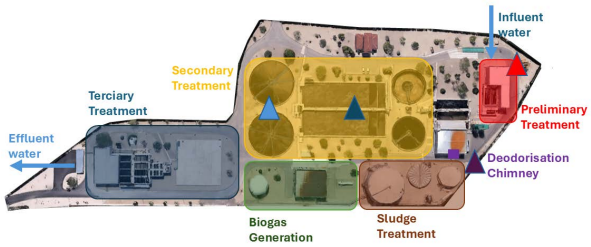

Figure 2: Waste Water Treatment Plant where the measurement campaign was performed. ▲ Primary Settler, ▲ Bioreactor, ▲ Preliminary Treatment Building, ▲ Deodorisation Chimney.

Day	Date	Settler▲	Bioreactor▲	Pretreatment▲	Chimney▲	Total (odour)	Blanks	Total
1	24/06/2020	3	3	2	2	10	7	17
2	25/06/2020	2	2	2	2	8	6	14
3	14/07/2020	3	3	3	3	12	11	23
4	15/07/2020	3	3	3	3	12	7	19
Total		11	11	10	10	42	31	73

Table 3: Number of samples collected in each source during the four measurement days

The primary factor influencing odour levels at each source is the combination of gases emitted by the source itself [6]. While environmental factors such as wind, temperature, and humidity also affect odour levels, the source emissions are the main determinant under consistent conditions when performing dynamic olfactometries [24]. The Primary Settler (▲ in Fig.2) typically emits gases such as hydrogen sulphide (H_2S) and methane (CH_4) [25]. The Bioreactor and the Preliminary Treatment Building (▲ and ▲ in Fig.2) are common sources of carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) [24]. Finally, the Deodorisation Chimney (▲ in Fig.2) usually emits volatile organic compounds (VOCs) and hydrogen sulphide (H_2S) [26].

The campaign spanned four days, during which the drone conducted measurements over four sources. Table 3 provides a summary of these measurements. Each day, at least two sample bags were collected from each source at different heights. Additionally, locations with low odour levels (referred to as blanks in Table 3) were also sampled. Figure 3 displays the sensor readings obtained on the first day of the measurement campaign. For more information about the IOMS, please refer to previous work [7, 23].

2.3. General Workflow

The data analysis process for developing an odour level prediction model is illustrated in Fig.4(a). This figure provides a general overview of the training and validation process, consistent with previous publications [7, 8, 23].

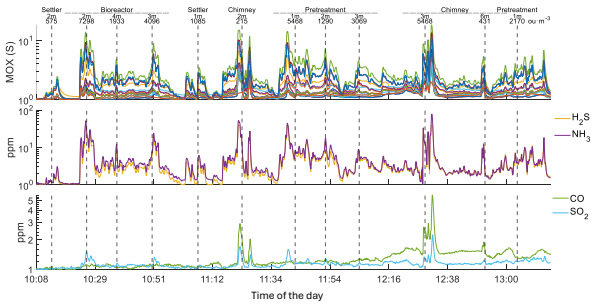

Figure 3: Sensor readings in the first day of campaign. The figure includes the response of the 16 MOX sensors, the H_2S and the NH_3 sensors, the CO and the SO_2 sensors, and the CO_2 sensor. The vertical pink bars indicate the moment when the lung sampler was on.

Firstly, the sensor readings are pre-processed using different techniques for MOX sensors and Electrochemical sensors. After a baseline correction for each sensor, a log-transformation was applied at every data sequence recorded with the IOMS. The Feature matrix was organized forming a total matrix of *nxm* rows and columns, where each row corresponds to a sampling bag collected with the lung sampler and each column to a single value of the sensor readings.

The vertical lines in Fig.3 indicate when the lung sampler is activated. Around these marks, 5 minutes are considered relevant for the analysis, and the rest of the time series is excluded from the features. The initial feature matrix has dimensions of $73x945$, including blanks for pre-processing purposes (73 samples by 21 sensors multiplied by 45 samples per sensor). After removing the blanks and excluding non-useful samples, the final feature matrix consists of $40x945$, where each sensor has 45 features per sample.

Once the feature matrix is constructed, a Partial Least Squares (PLS) model is trained and tested as described in [13]. A double leave-one-block-out cross-validation (CV) scheme is employed for model building. This approach uses data from three days for model training and reserves one day for validation, resulting in four different models. The model with the best performance is then selected.

In order to avoid complexity grow uncontrolled at each of the proposed models, within each calibration set the "leave one sample out" (LOO) scheme is used. N PLS models are built with different Latent Variables (LVs) leaving a sample for test out of the building process. The number of LVs is therefore optimized by evaluating the Root Mean Square Error in Cross Validation (RMSECV). Finally, the models are refit using the cal-

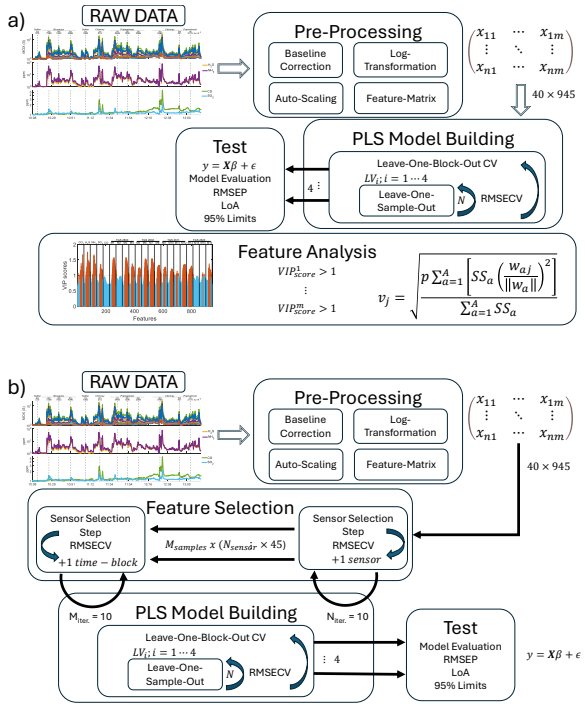

Figure 4: Data analysis Workflow. (a) General Workflow with Feature Selection based on VIP calculation and discrimination. (b) Modified Workflow with NiPLS as Feature Selection methodology.

ibration samples used during training and used to produce prediction plots on blind samples. The bias of the prediction and 95% limits of agreement (LoA) are computed using the Bland-Altman methodology [27].

Whenever the model is ready for predictions, the Variable Importance in Projection (VIP) method is used to identify the most relevant features during training. The VIP scores of each sensor are averaged per day for each feature (45 per sample for each sensor), leading to an averaged VIP score distribution depicted in Fig. 6 (b).

2.4. Nested Sequential Forward Selector using iPLS for Odour Quantification

The presented workflow is optimal for datasets with a low number of samples. In the performed measurement campaign, this number reduces to 73, with almost half of the samples corresponding to blank samples used for signal correction. Its strength lies in the cross-validation technique, which performs nested leave-one-block-out and leave-one-sample-out procedures to optimize models. However, no feature selection was performed in the workflow, leaving the information from the extracted features unexploited.

Figure 4(b) shows the modified data analysis workflow proposed to optimize the model training presented in Section 2.3 and depicted in Fig. 4(a). The proposed methodology is very similar to the one applied for cross-validation (CV) in the general workflow but incorporates a two-step feature selection process. This feature selection stage is based on the Nested Sequential Forward Algorithm.

In the first step of the feature selection stage, a PLS model is built using the general workflow but including only one sensor of the sequence of 945 features of each sample. The RMSECV is calculated and then another PLS model is built adding a different sensor to the previous one. The process is iterated 10 times looking for a minimum in the RMSCEV. As depicted in Fig.8(a), at each iteration, a table informing the configuration of the sensors is created. The table provides information about which sensors are included in the feature matrix (■) and which are not (■).

In the second step of the feature selection stage, a time-slot of each sensor around the activation of the lung sampler (the 5 minutes window corresponding to the vertical bar in Fig.3) is divided in 15 blocks, which corresponds to time-blocks of 20 seconds. With this partition, a similar procedure than the one described in the Sensor step is performed. Time-blocks are added to the model building process and the RMSECV is observed. After 10 iterations a minimum appears to be clear for a selection of time-blocks.

3. Results and Discussion

3.1. Gas Concentration Measurements

Figure 3 shows an example of data collected by the IOMS during a day of the campaign. Data from the different sensors is pre-processed using the blank samples, and the baseline level is corrected with the lowest value recorded for each sensor that day. After this adjustment, the 5-minute time slot around the sampling bag filling (■ rectangle in Fig.3, corresponding to 45 samples) is included in the feature matrix as explained in Fig.4. The final appearance of the feature matrix after pre-processing according to the general workflow is depicted in Fig. 5.

A more detailed analysis is presented in Fig.6. This figure shows the data from the H_2S and NH_3 sensors, with the 5-minute time slot marked by a pink rectangle (■). This time slot corresponds to the data included in the feature matrix for each sensor (see Fig.6 (a)).

Figure 6 (b) shows the VIP for each sensor averaged per day, computed as explained in Section 2.3. The VIP

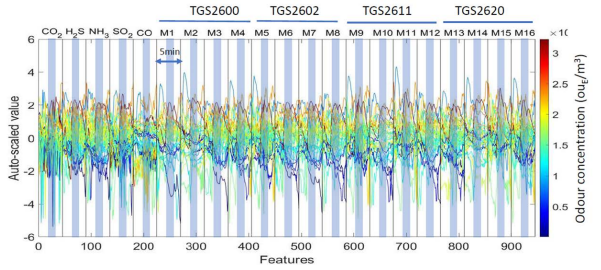

Figure 5: Preprocessed data of the different sources for every day of the measurement campaign. The 5 minutes of valid data have been marked in the plot.

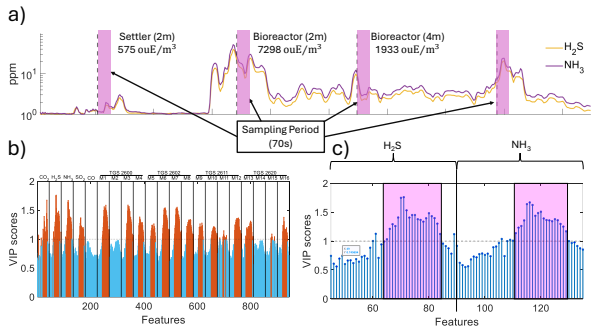

Figure 6: Detail of the signal acquired from two of the most significant sensors (H_2S and NH_3) when the drone is flying over several sources and VIP scores. (a) Detail of the signal acquired and sampling period (□). (b) Distribution of the VIP scores of every sensor when averaged over the four days of campaign at every time-slot. (c) Detail of the averaged VIP scores for H_2S and NH_3. The sampling period is also marked (□).

analysis indicates that within the 5-minute time slot considered for each sensor, the most relevant features are typically located at the end of the sampling bag filling process. Specifically, there are 21 samples with VIPs greater than 1. Additionally, the highest VIP scores are observed for the H_2S and NH_3 electrochemical sensors, as well as some of the MOX sensors. Therefore, it should be possible to reduce the features used to train the prediction PLSR model by selecting the relevant sensors and the features of each relevant sensor within the 5-minute time slot.

3.2. General Workflow Performance

In orther to compare the performance of the NiPLS workflow with workflows presented in the literature [23, 8] the performance of the workflow depicted in Fig. 4 (a) is presented in Fig. 7.

In this case, when 40 samples are considered (from the 73 acquired, after removing blanks), the best PLSR model presents 2 latent variables. This model achieves

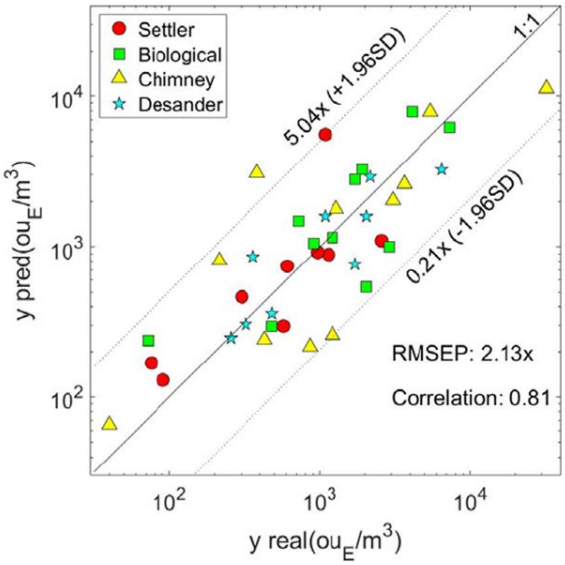

Figure 7: Performance of the general workflow presented in section 2.3. The number of useful samples have been reduced to 40 plus blanks. Prediction plot for the best PLS model built with LoA limits at 95% CI of $[0.2x - 5.0x]$ and correlation coefficient of 0.81.

an RMSEP of 2.13x for the 4 sources considered at different heights: the Settler (▲), the Bioreactor (▲), the Chimney (▲), and the Pretreatment Building (▲). The calculated LoA at 95% CI ranges from 0.2x to 5x of the prediction.

3.3. NiPLS Performance

With the NiPLS presented in Fig. 4 (b) the results of the novel feature selection algorithm are presented. When the iPLS achieves the sixth and fifth iteration, the RMSECV goes down in both steps of the selection process (sensor step and time-slot step).

In the sensor selection step, six sensors are identified as the most relevant: the NH3, SO2, and CO electrochemical sensors, and the MOX sensors M6, M7, and M11. These MOX sensors belong to the second (TGS2602) and third (TGS2611) groups of MOX sensors (see Table 2), operating at 3.2 V, 4.0 V, and 4.0 V, respectively.

In the time-slot selection step, five time slots of the preselected sensors are chosen as the most relevant features. However, the features from the M11 MOX sensor are excluded. For the NH_3, M7, and M6 sensors, the selected features correspond to the central part of the 45 features, while for the SO_2 and CO electrochemical sensors, the initial features appear to be more relevant.

With this features, the resulted PLSR model shows the performance depicted in Fig. 9

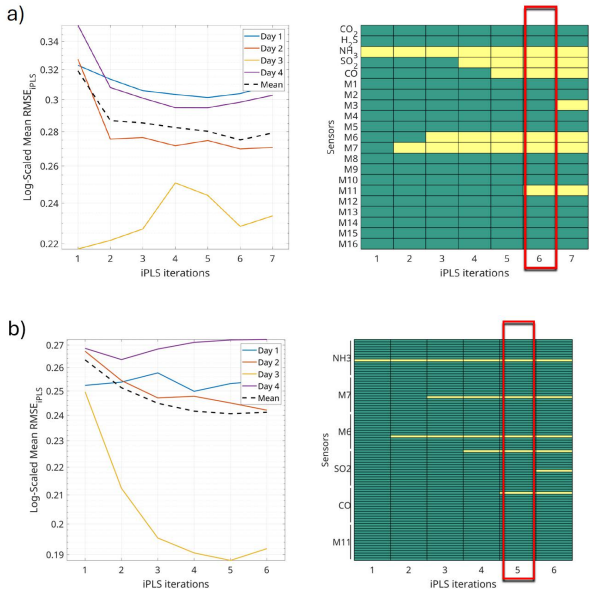

Figure 8: Nested Sequential Forward Selector. The graphs show the evolution of the RMSECV for each day (and mean) during the block selection process. ■ interval included in the feature matrix, ■ interval not included in the feature matrix (a) In the Sensor Step three electrochemical sensors (NH_3, SO_2 and CO) and three MOX sensors (M6, M7 TGS2602 and M11 TGS2611) show the maximum improvement in RMSECV. (b) In Time-slot Step the mid time-slot (15 seconds) of the NH_3, CO, M6 and M7 sensors improve even further the RMSECV.

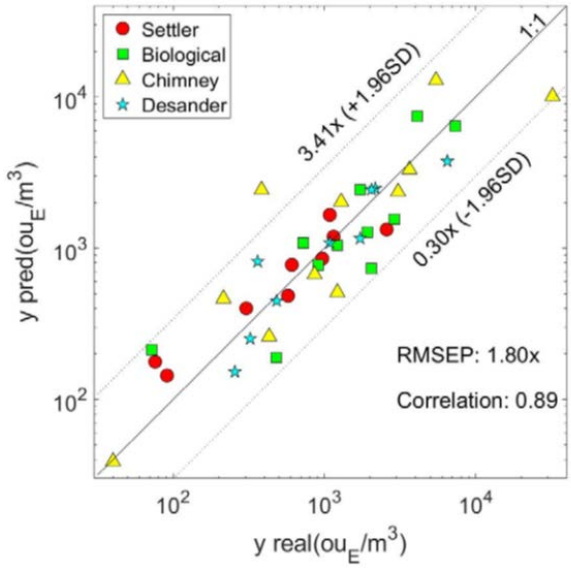

Figure 9: Performance of the optimized model using NiPLS. The number of useful samples have been reduced to 40 plus blanks. The model shows LoA limits at 95% CI of $[0.3x - 3.4x]$ and a correlation coefficient of 0.89.

A reduction in RMSEP is observed when the same set of samples used in Section 3.2 is considered. The NiPLS model shows an RMSEP of 1.8x, and the LoA at 95% CI lies in the 0.3x to 3.4x range. This represents a considerable improvement compared to traditional workflow applied to the dataset (see Fig. **??**).

4. Conclusions

This contribution demonstrates how obtaining information on the importance of sensors and features enables the implementation of an algorithm to automatically select the most relevant information, resulting in improved model performance. By combining this approach with the Interval Partial Least Squares Regression model, specifically tailored for IOMSs studying odour concentrations in Wastewater Treatment Plants (WWTPs), significant enhancements are achieved compared to traditional methods. For the measurement campaign conducted at a WWTP in Spain, this novel workflow shows an improvement in RMSEP, but its major contribution is the reduction in the Limit of Agreement at 95% CI of the prediction from 5x to 3.4x.

These results enhance the performance of IOMSs onboarded in drones for evaluating odour emissions, allowing them to be used as a complementary odour analysis tool alongside dynamic olfactometry in WWTPs.

Furthermore, the insights gained from identifying relevant sensors and features can be used to optimize IOMSs, reducing system size and data load transferred from the IOMS to the ground PC.

References

[1] T. Brinkmann, R. Both, B. Scalet, S. Roudier, L. Sancho, Jrc reference report on monitoring of emissions to air and water from ied installations, EUR 29261 EN. Eur. IPPC Bur. Eur. Comm. Jt. Res. Cent. (2018).

[2] T. Group, Integrating climate into our strategy [www document], `https://www.total.com/sites/default/files/atoms/files/total_rapport_climat_2019_en.` (2019).

[3] F. Zhou, S. Pan, W. Chen, X. Ni, B. An, Monitoring of compliance with fuel sulfur content regulations through unmanned aerial vehicle (uav) measurements of ship emissions, Atmos. Meas. Tech. 12 (2019).

[4] S. Sironi, L. Capelli, P. Céntola, R. Del Rosso, S. Pierucci, Odour impact assessment by means of dynamic olfactometry, dispersion modelling and social participation, Atmospheric Environment 44 (3) (2010) 354–360.

[5] F. Cangialosi, G. Intini, D. Colucci, On line monitoring of odour nuisance at a sanitary landfill for non-hazardous waste, Chem. Eng. Trans 68 (2018) 127–132.

[6] L. Capelli, S. Sironi, P. Céntola, R. Del Rosso, M. Grande, Electronic noses for the continuous monitoring of odours from a wastewater treatment plant at specific receptors: Focus on training methods, Sensors and Actuators B: Chemical 131 (1) (2008)

53–62, special Issue: Selected Papers from the 12th International Symposium on Olfaction and Electronic Noses.

[7] J. Burgués, S. Marco, Environmental chemical sensing using small drones: A review, Science of The Total Environment 748 (2020) 141172.

[8] J. Burgués, M. D. Esclapez, S. Doñate, L. Pastor, S. Marco, Aerial mapping of odorous gases in a wastewater treatment plant using a small drone, Remote Sensing 13 (9) (2021).

[9] S. Wold, A. Ruhe, H. Wold, W. J. Dunn, III, The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses, SIAM Journal on Scientific and Statistical Computing 5 (3) (1984) 735–743.

[10] I. Chong, J. C.H., Performance of some variable selection methods when multicollinearity is present, Chemometrics and Intelligent Laboratory Systems 78 (1) (2005) 103–112.

[11] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double cross validation, Journal of Chemometrics 23 (4) (2009) 160–171.

[12] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer, New York, 2009.

[13] S. Wold, M. Sjöström, L. Eriksson, Pls-regression: A basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2) (2001) 109–130.

[14] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, Chichester, 1989.

[15] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Statistics Surveys 4 (2010) 40–79.

[16] A. K. Jain, D. Zongker, Feature selection: Evaluation, application, and small sample performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2) (1997) 153–158.

[17] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[18] L. Nø rgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. B. Engelsen, Interval partial least-squares regression (ipls): A comparative chemometric study with an example from near-infrared spectroscopy, Applied Spectroscopy 54 (3) (2000) 413–419.

[19] J. Fonollosa, L. Fernández, R. Huerta, A. Gutiérrez-Gálvez, S. Marco, Temperature optimization of metal oxide sensor arrays using mutual information, Sensors and Actuators B: Chemical 187 (2013) 331–339, selected Papers from the 14th International Meeting on Chemical Sensors.

[20] J. Burgués, S. Marco, Low power operation of temperature-modulated metal oxide semiconductor gas sensors, Sensors 18 (2) (2018) 339.

[21] Y. Zheng, S. Yang, X. Liu, J. Wang, T. Norton, J. Chen, Y. Tan, The computational fluid dynamic modeling of downwash flow field for a six-rotor uav, Frontiers of Agricultural Science and Engineering 5 (2) (2018) 159.

[22] Y. Zhu, Q. Guo, Y. Tang, X. Zhu, Y. He, H. Huang, S. Luo, Cfd simulation and measurement of the downwash airflow of a quadrotor plant protection uav during operation, Computers and Electronics in Agriculture 201 (2022) 107286.

[23] J. Burgués, M. D. Esclapez, S. Doñate, S. Marco, Rhinos: A lightweight portable electronic nose for real-time odor quantification in wastewater treatment plants, iScience 24 (12) (2021) 103371.

[24] J. L. Campos, D. Valenzuela-Heredia, A. Pedrouso, A. Val del Río, M. Belmonte, A. Mosquera-Corral, Greenhouse gases emissions from wastewater treatment plants: Minimization, treatment, and prevention, Journal of Chemistry 2016 (1) (2016) 3796352.

[25] U. S. E. P. Agency, Primer for municipal wastewater treatment systems, Tech. rep., EPA (2004).
URL https://www3.epa.gov/npdes/pubs/primer.pdf

[26] C. S. S. L. Casaca, N. P. F. Henriques, E. A. V. Nunes, T. A. P. S. Simões, Ventilation and deodorization system optimization of a wastewater treatment plant, in: XI Congreso Ibérico IX Congreso Iberoamericano de Ciencias y Técnicas del Frío, 2022.

[27] P. Taffé, When can the bland & altman limits of agreement method be used and when it should not be used, Journal of Clinical Epidemiology 137 (2021) 176–181.