# A feature selection strategy for the analysis of spectra from a photoacoustic sensing system

Alessandro Ulrici*[a,b], Renato Seeber[b,c], Marco Calderisi[a,b], Giorgia Foca[a], Juho Uotila[d], Mathieu Carras[e], Anna Maria Fiorello[f]

[a]Dipartimento di Scienze della Vita, Università di Modena e Reggio Emilia, Via Amendola 2, 42122 Reggio Emilia, Italy; [b]Consorzio INSTM, Via G. Giusti 9, 50121 Firenze, Italy; [c]Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi 183, 41125 Modena, Italy; [d]Gasera Ltd., Tykistökatu 4, 20520 Turku, Finland; [e]III-V Lab, avenue Augustin Fresnel 1, Palaiseau, France; [f]Selex-SI, Via Tiburtina, Km 12,400, 00131, Rome, Italy

## ABSTRACT

In the frame of the EU project CUSTOM, a new sensor system for the detection of drug precursors in gaseous samples is being developed, which also includes an External Cavity-Quantum Cascade Laser Photo Acoustic Sensor (EC-QCLPAS). In order to define the characteristics of the laser source, the optimal wavenumbers within the most effective 200 $cm^{-1}$ range in the mid-infrared region must be identified, in order to lead to optimal detection of the drug precursor molecules in presence of interfering species and of variable composition of the surrounding atmosphere. To this aim, based on simulations made with FT-IR spectra taken from literature, a complex multivariate analysis strategy has been developed to select the optimal wavenumbers. Firstly, the synergistic use of Experimental Design and of Signal Processing techniques led to a dataset of 5000 simulated spectra of mixtures of 33 different gases (including the 4 target molecules). After a preselection, devoted to disregard noisy regions due to small interfering molecules, the simulated mixtures were then used to select the optimal wavenumber range, by maximizing the classification efficiency, as estimated by Partial Least Squares – Discriminant Analysis. A moving window 200 $cm^{-1}$ wide was used for this purpose. Finally, the optimal wavenumber values were identified within the selected range, using a feature selection approach based on Genetic Algorithms and on resampling. The work made will be relatively easily turned to the spectra actually recorded with the newly developed EC-QCLPAS instrument. Furthermore, the proposed approach allows progressive adaptation of the spectral dataset to real situations, even accounting for specific, different environments.

**Keywords:** Signal Processing, Feature Selection, Classification, Wavelet Transform, Genetic Algorithms, mid-infrared spectra, drug precursors.

## 1. INTRODUCTION

The EU FP7 Collaborative Project (Theme: Security) CUSTOM (Drugs and Precursor Sensing by Complementing Low Cost Multiple Techniques)[1] is focused on the development of a new sensor system for the identification of vapors of chemicals that are used as drug precursors (target molecules) in indoor and outdoor environments. This means that they are in presence of a relatively high number of possibly interfering species (pollutants) and of background air components with variable concentrations. This new sensing system makes use of two integrated systems: an External Cavity-Quantum Cascade Laser Photo Acoustic Sensor (EC-QCLPAS)[2] and a Led Induced Fluorescence Optical Sensor (FLUO). The EC-QCLPAS system works in the mid-infrared region, and is capable to cover a 200 $cm^{-1}$ range within the 1000-2500 $cm^{-1}$ spectral range. Key points for the optimal performance of this sensor lies therefore in the definition both of the optimal wavenumber value where the 200 $cm^{-1}$ working range must be centered, and of the single wavenumber values that must be considered within the selected range. These conditions are essential in order to achieve optimal detection of the target molecules in presence of many possible pollutants and of air components. Since the final version of the sensing system was still under development, one of the main problems to face consisted in the necessity to estimate in advance the spectral response of the final device in a large spectral range (1000-2500 $cm^{-1}$), though in absence of real experimental data, except for few spectra acquired with a prototype version of the sensor, operating in a

*alessandro.ulrici@unimore.it; phone +39 0522 522043; +39 0522 522027; www.dipsaa.unimore.it/eng/staff/ulrici

much narrower wavelength interval.

For this reason, it has been necessary to make use of simulated responses of gas mixtures, derived from literature spectra measured with different FT-IR instruments on the pure chemical components, at unit concentration. Since the new EC-QCLPAS must be suitable to work in different environments, with varying types and amounts of chemical species, a proper simulation of a wide variety of gas mixtures was made, in order to build a spectral dataset as much representative and "realistic" as possible. To this aim a complex strategy has been developed, mainly involving preprocessing of spectral data and experimental design techniques for the simulation of the concentrations of the various gaseous species involved in the study[3]. This led to a dataset of 5000 simulated spectra of mixtures of 33 different gases (including the 4 target molecules, 9 air components and 20 interfering species), each one considered at varying concentration levels.

The dataset of simulated mixtures was then used to select the optimal 200 $cm^{-1}$ wavenumber range and the single wavenumbers therein. More in detail, before considering the actual feature selection phase, a thoughtful preselection was made on the whole 1000-2500 $cm^{-1}$ spectral range, in order to cancel out the regions containing sharp and intense absorption peaks of abundant interfering chemical species with a low number of atoms (e.g. $H_2O$, $CO_2$). The problem eventually arising from sharp peaks of small molecules is due to the relatively low accuracy of the EC-QCLPAS intensity values measured in correspondence with these spectral regions, which in turn is due to the low hypothesized final instrument precision in repositioning, when compared to the extremely narrow bandwidth. To this aim, by means of a Wavelet Transform (WT)[4-6] based algorithm developed *ad-hoc* and following previous literature suggestions[7], a preselection procedure named *SMARTGRID* was implemented, which makes use of a *sharpness function* in order to identify and exclude the spectral regions containing the sharp peaks.

After this preselection step, the simulated spectra of mixtures were used to select the optimal 200 $cm^{-1}$ wavenumber range. This goal was reached by estimating the Classification Efficiency (EFF) values calculated with Partial Least Squares - Discriminant Analysis (PLS-DA)[8-9] for each position of a moving window 200 $cm^{-1}$ wide, covering the whole 1000-2500 $cm^{-1}$ spectral range by steps of 1 $cm^{-1}$. In particular, for each target molecule and for each position of the moving window, a PLS-DA model was calculated to discriminate the presence/absence of the considered target within the mixtures spectra, and the corresponding EFF was estimated both in cross-validation and in prediction of an external test set. Then, the window position leading to the maximum value of the overall EFF estimated in cross-validation was chosen as the optimal wavenumber range.

Finally, the optimal wavenumber values were identified within the selected range using a feature selection approach based on Genetic Algorithms (GA)[10-11] and on resampling. This approach consisted in performing a series of 100 random subsamplings of 2000 spectra from a training set of 3000 spectra, and in applying GA to each subsample for the classification of each target molecule considered separately. The global frequency of selection in correspondence to each wavenumber was calculated, and the spectral variables were then ranked accordingly. Finally, starting from the most frequently selected wavenumber and adding each time a further variable, a series of PLS-DA models was calculated using all 5000 samples of the simulated spectral database (3000 spectra in the training set and 2000 in the test set), and the optimal number of single wavenumbers to be kept was defined on the basis of the maximum value of the overall EFF estimated in cross-validation. This final selection allowed us to define a restricted number of wavenumbers leading to effective classification models for each one of the considered target molecules.

Using this approach, even in absence of true experimental spectra, we were able to identify the optimal spectral variables for the detection of the 4 drug precursors (acetic anhydride, ephedrine, 1-phenyl-2-propanone and safrole) in presence of 20 possible pollutants and of 9 air components. The choice was the guide for the definition of the proper working conditions of the laser source. The work made will be relatively easily turned to the spectra actually recorded with the newly developed EC-QCLPAS instrument. Furthermore, the proposed approach allows one to progressively adapt the spectral dataset to real situations, even accounting for specific, different environments.

## 2. DATASET OF SIMULATED GAS MIXTURES SPECTRA

The analyzed dataset of the simulated gas mixture spectra contains all the considered chemical species (target molecules, pollutants and air components) at the chosen concentration levels, in the suitable mixing proportions[3]. The 5000 spectra cover the 1000-2500 $cm^{-1}$ range with a 0.5 $cm^{-1}$ spectral resolution; thus, each spectrum is a vector composed of 3001

spectral variables. The procedure followed to build this dataset consisted in repeating 5 times the simulation of 1000 spectra. Each block of 1000 spectra was different from the previous one, and was composed by randomly merging:

- 499 mixtures of target molecules, each one containing from 1 to 4 targets, whose concentrations varied according to a set of Full Factorial Designs (FFDs)[12];
- 750 mixtures of pollutants, sampled randomly from the whole pollutants matrix composed by 34080 different mixtures including a number of pollutants raging from 1 to 3, whose concentrations varied according to a set of FFDs;
- 1000 mixtures of air components, where all the 9 components have always been included, and where the concentration values for each component were randomly selected from a lognormal distribution specifically built on the basis of the corresponding average and maximum concentration values.

From the simulated spectra it is clear that the sharp peaks mainly due to $H_2O$ and $CO_2$ have very high intensities, completely overwhelming the spectral bands of actual interest. The classification of the gas mixtures spectra depending on the presence / absence of each single target molecule is not a trivial task, due both to the presence of very intense sharp peaks of small molecules and to the lack of spectral regions where the contribution of the target molecules is clearly distinguishable from that of pollutants and air components. For this reason, a thoughtful feature selection procedure had to be implemented.

## 3. SPECTRA PREPROCESSING

The first step of the feature selection procedure consisted in a careful preprocessing of the analyzed dataset, aimed at discarding the signal portions containing sharp intense peaks of small interfering chemical components, since their contribution to the final mixtures spectra is highly irreproducible and could therefore negatively affect the final classification models. At the same time, however, utmost attention was paid to include all the neighboring regions, which could potentially bring useful information for the identification of the target molecules.

### 3.1 Problems in detection of sharp peaks with EC-QCLPAS

Small molecules, such as $H_2O$ and $CO_2$, possess typical spectra showing many very narrow and high intensity peaks, as it can be seen in the example of Figure 1, which reports the spectrum of $CO_2$ measured at a 1 ppm concentration. Considering that the instrumental resolution is related to a error in repositioning of about 0.5 $cm^{-1}$, and that the FWHM of the laser line is approximately equal to 0.1 $cm^{-1}$, it is obvious that, as it is shown in Figure 2, the error in repositioning heavily affects the accuracy of the spectral intensity measurement. For a 1 ppm concentration, the signal intensity could range between about $1 \times 10^{-2}$ absorbance units to $4 \times 10^{-4}$ absorbance units, so that the $CO_2$ contribution could vary of a factor of 25.

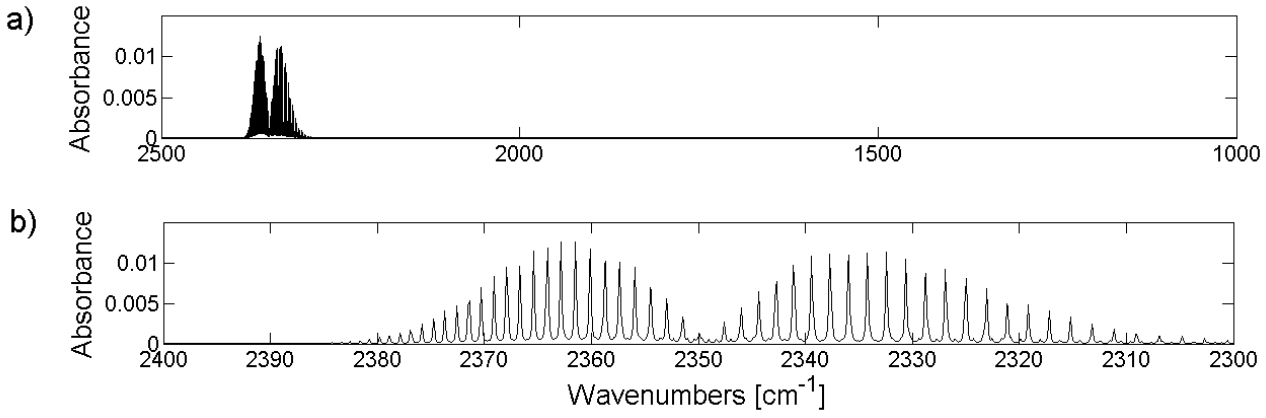


Figure 1. a) spectrum of CO2 in the whole considered range and b) zoom of the high intensity sharp peaks in the 2300-2400 $cm^{-1}$ spectral range.

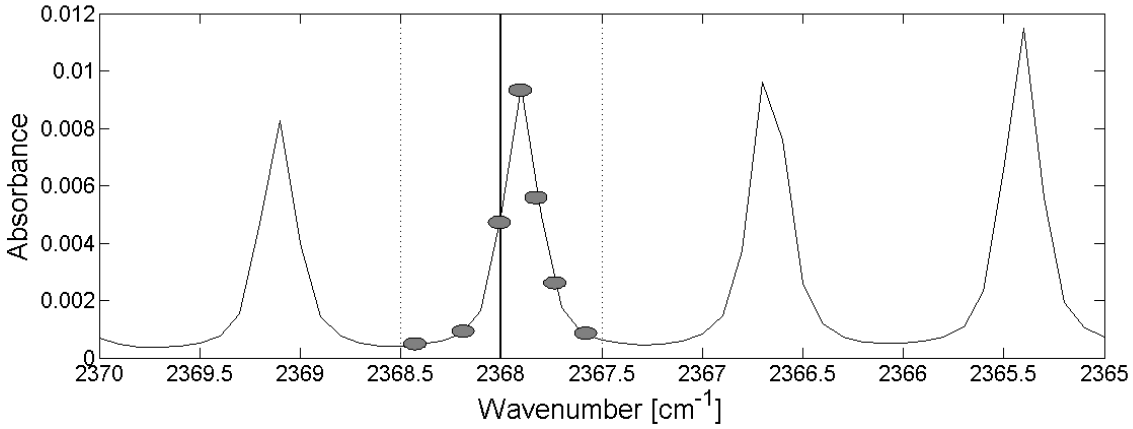


Figure 2. Example of the possible results of the measurement of $CO_2$ at the nominal wavenumber of 2368 $cm^{-1}$ (solid vertical line): the gray circles represent some possible measurement results made with a bandwidth equal to 0.1 $cm^{-1}$ and with a repositioning error equal to ±0.5 $cm^{-1}$ (vertical dotted lines).

For this reason, an algorithm was developed *ad-hoc* in order to detect the extent, intensity and position of the sharp peaks deriving from all the small molecules of the dataset that could generate this effect, and to discard all the corresponding spectral regions in an interval equal to the error in repositioning.

### 3.2 The SMARTGRID approach

The implemented approach, named SMARTGRID, makes use of the Wavelet Transform (WT)[4-6] to separate the high frequency content of the signal, corresponding to the sharp peaks, from the smooth variation due to the presence of large absorption bands and signal background. The idea to define a smart grid to discard the regions containing sharp peaks from the remainder part of the signal was formerly presented in the literature by Dunayevskiy et al.[7], but here this method is extended to complex cases, i.e. involving more than one target and one interfering molecule, which required the development of the WT-based algorithm.

The SMARTGRID procedure starts from simulations of all the previously denoised signals[3] of the small molecules included in the dataset sampled at a 0.1 $cm^{-1}$ resolution in the 1000–2500 $cm^{-1}$ range at a 1 ppm concentration and, by means of a graphical interface based on the Fast Wavelet Transform decomposition[4,6] of each single signal, allows the user to interactively estimate the best parameter settings. These consist of the type of wavelet function and level of decomposition that lead to best separation between the smooth component (i.e. the low frequency variations due to large absorption bands and to the background, which are collected into the wavelet approximation vector) and the "sharp" component (i.e. the variations connected to the sharp peaks, which are collected into the wavelet detail vectors) of the signal.

In order to quantify the position and the intensity of the sharp peaks, for each molecule *i* of the considered dataset that shows sharp peaks, a sharpness profile was defined as:

$$S_{i,\nu} = \left| I_{i,\nu} \times \frac{D_{i,\nu}}{A_{i,\nu}} \right| \times C_i \qquad (1)$$

where $I_{i,\nu}$ is the intensity of the denoised spectrum of the *i*-th molecule at the ν-th wavenumber, considered at a 1 ppm concentration, $A_{i,\nu}$ is the corresponding intensity of its smooth part, $D_{i,\nu}$ is the corresponding intensity of the sharp part, and $C_i$ is the concentration (ppm) of the *i*-th molecule, estimated by considering the worst scenario, i.e. the case where it is present at the maximum possible concentration level[3,7]. This function reflects how intense, for each single wavenumber, the sharp variations are with respect to the neighboring parts of the signal, i.e. with respect to the smooth variations, and weights this ratio by the corresponding intensity of the 1 ppm absorbance spectrum and by the maximum possible concentration of the *i*-th molecular species.

The overall *sharpness function* is the mean of the sharpness profiles of all the molecules generating sharp peaks, thus representing the positions and intensities of all the sharp peaks that can be present in the final mixtures. Then, all the wavenumber values where the sharpness function assumes values higher than a fixed threshold are discarded, together with all the wavenumbers whose distance from these ones is lower than the estimated error in repositioning. Finally, the selected wavenumbers are downsampled, in order to match with the desired spectral resolution of 0.5 $cm^{-1}$.

Of course, the size of the so defined SMARTGRID, i.e. the number of spectral variables that are preselected, heavily depends upon the fixed threshold value: the higher the threshold value, the higher the number of included variables, possibly still containing "noisy" variables. The optimal threshold value was therefore defined using an empirical approach, i.e. calculating PLS-DA models on sets of variables selected using different threshold values, and then keeping the one leading to the maximum value of Classification Efficiency in cross-validation.

## 4. SELECTION OF THE OPTIMAL SPECTRAL RANGE

The simulated mixtures spectra were then used to select the optimal 200 $cm^{-1}$ wavenumber range, considering only the spectral variables retained by the SMARTGRID. To this aim, a moving window with the same size of the final spectral range (200 $cm^{-1}$) was used, which covered the whole 1000-2500 $cm^{-1}$ range moving with steps of 1 $cm^{-1}$. Data were pre-processed using autoscaling (AUTO) and Pareto scaling plus mean centering (PARETOMNCN)[8]. Models were validated using both Cross-Validation (CV)[8], and an external Test Set (TS). In correspondence to each position of the moving window, a PLS-DA model was calculated to discriminate the presence/absence of each one of the 4 target molecules, and the corresponding Classification Efficiency (EFF), defined as the geometric mean of the two parameters Sensitivity (the percentage of mixtures correctly assigned to the proper class) and Specificity (the percentage of objects of the other class correctly rejected by the class model), was estimated. The Global Classification Efficiency (GEFF), calculated as the geometric mean of the EFF values of the four targets, was then calculated for each window position both on the CV and on the TS prediction results. Finally, the window position leading to the maximum value of GEFF estimated in CV was selected as the optimal wavenumber range.

The overall results indicate that the classification performance is satisfactory: not only there are regions where the GEFF values are quite high, but a good agreement was also observed between the CV and TS results, in addition to a good coherence between the GEFF values obtained with AUTO and PARETOMNCN. The window leading to the best results in CV was identified and set as the center of the optimal wavenumber range for further feature selection. This wavenumber region includes 364 wavenumbers (out of a total of 399 variables, 35 wavenumbers were previously deleted with the SMARTGRID approach). In general, the classification performance was quite similar for the different targets. The only exception was observed for acetic anhydride, for which better results could be obtained in a different wavelength region. However, also this target molecule is identified quite well with the window selected for the other targets.

In all the cases, values higher than 85% were obtained, and the GEFF values for the prediction of TS mixtures resulted higher than 86%. Almost all samples without target molecules were classified correctly. The most part of false positives were in fact due to samples containing a target molecule different from the one considered in the classification model: for example, in the classification model for target 1 the most part of false positives corresponded to mixtures including some of the other target molecules. As for the false negatives are concerned, their evaluation revealed that they correspond to the mixtures with the lowest concentration values of the analyzed target. The evaluation of the target concentration values in the false negatives allowed us to obtain a first rough estimate of the detection limit obtained with this classification models, which resulted equal to about 200 ppb. This value is higher than the final expected one (50 ppb); however, it must be underlined that further elaborations, including the optimization of the discriminant threshold value could reasonably lead to notable decrease of this value.

## 5. SELECTION OF THE SINGLE WAVENUMBERS

The further step consisted in the selection of the single wavenumbers within the optimal spectral range. To this aim, a feature selection method based on GA[10-11] and on resampling was implemented.

GA are search algorithms based on the principles of natural evolution, that find optimal solutions to a given problem (in this case, the identification of the wavenumbers leading to PLS-DA model with maximum EFF values), by simulation of a natural evolution process of chromosomes. In the present case, each chromosome corresponds to a given subset of wavenumbers, that is initially chosen randomly within the 364 variables belonging to the optimal spectral range. The quality of each chromosome is quantified by the EFF value of the corresponding PLS-DA model. Through the simulation of an evolutionary process including competition between chromosomes, reproduction to create new chromosomes that replace the ones that have reached the end of their life-time, and semi-random changes that might lead to fitter chromosomes, the population of chromosomes is updated step-by step, generally converging towards a sub-optimal solution. This means that, although it cannot be proven that the GA run has really found the optimal solution, some of the results obtained by the GA are better than any previously known solutions.

Since GA is an essentially stochastic algorithm, it is obvious that the final solutions of different GA runs will not be exactly the same. By comparing the solutions from different GA runs, where each time the procedure starts with a new randomly generated initial population of chromosomes, it is therefore possible to have an idea of the robustness of the method: this is the key principle of the GA version developed by Leardi[10], that was used in the present work.

Moreover, in order to make the feature selection as much as possible independent of the specific set of training set objects, thus further minimizing the risk of obtaining chance correlations, the GA-based method developed here for the selection of the optimal wavenumbers was repeatedly applied 100 times, each time on a different subsample of the whole training set. More in detail, a series of 100 random subsamples, each one containing 2000 spectra, was created starting from a training set of 3000 spectra, taking care to have a balanced number of target molecules within each subset. GA was then applied to each subsample for the classification of each target molecule considered separately, performing 100 different GA runs, each run consisting of 100 evaluations, i.e. 100 subsequent generations during which the "evolution" of the EFF value estimated in CV drives the selection of the optimal variables. For each run, the frequency of selection of each variable (wavenumber) is then kept, which expresses the contribution of that variable to EFF. Then, for each subsample, the mean frequency of selection of each variable among the 100 runs is calculated and stored in the corresponding frequency vector. This process is repeated for each one of the 100 subsamples, thus generating 100 frequency vectors. The overall results can be represented as an image, where each pixel row corresponds to the frequency vector of a subsample, and where the color of each pixel with ($i$, $j$) coordinates corresponds to the frequency of selection of the $j$-th spectral variable using the $i$–th subset of objects. The values of the frequencies of selection can be obtained by comparison with a color scale.

Based on such a representation of the data from GA selection, it is not only possible to deduce which wavenumbers are the most promising ones for the detection of drug precursors, but also how much stable (robust) the predictions are, independently of the set of mixtures that are used. The presence of vertical series of pixels with colors corresponding to high frequency values, in fact, indicates that the corresponding variables are repeatedly selected, independently of the particular subset that has been considered. A good convergence in the variables selection was obtained for targets 2, 3 and 4, while higher variations were obtained for target 1 (acetic anhydride), which also shows lower selection frequency values. This result is probably ascribable to the fact that the mixtures containing acetic anhydride are more easily identifiable; thus, there could be a wider choice in the selection of variables that lead to equivalent EFF values. The described procedure was repeated twice, by applying the GA selection on the variables pretreated using both AUTO and PARETOMNCN, leading to very similar results.

Based on these results, for each target molecule and each variable, the corresponding target frequency selection was calculated as the median value of the corresponding frequencies over all the 100 subsamplings (i.e. over all the 100 rows of each image). After normalizing each target frequency selection vector with respect to its mean value, the overall frequency selection vector was then obtained as the mean of the four target frequency selection vectors. Finally, the 364 variables were sorted in decreasing order of the values of the overall frequency selection vector, i.e. starting from the variable that globally obtained the higher number of selections and ending with the less frequently selected variable. In this way, a dataset with the 364 sorted spectral variables was built, using the 3000 mixtures used in the GA selection as the training set, and the remainder 2000 mixtures at the external test set (TS).

PLS-DA models (using both AUTO and PARETOMNCN pretreatments) were then calculated for each target molecule using the sorted dataset, starting from the first two variables and adding iteratively one variable at each cycle, up to a total of 100 variables. The optimal number of LVs for each cycle and each target was selected on the basis of the geometric average of the EFF values calculated in CV.

Since two operating modes are planned for the final device using the EC-QCLPAS[1-2], two final classification models can be developed, one to be used for the "High Probability Of Detection" (HiPOD) operating mode and one for the "Low Probability of False Alarms" (LoPFA) operating mode. The HiPOD mode is planned for a first fast screening; thus, the corresponding classification model must be performed on few (maximum 40) wavenumbers. The LoPFA mode is activated in case of positive response of the HiPOD mode. In this case a longer and more refined analysis can be performed, allowing the inclusion of up to 100 wavenumbers. For this reason, based on the classification results, two sets of variables are chosen: the first one, for the HiPOD mode, includes those leading to the best GEFF values by considering up to 40 variables, while the second one, for the LoPFA mode, includes the overall best performance, within the first 100 variables.

The GEFF values obtained with the PLS-DA models for the prediction of the TS mixtures using both AUTO and PARETOMNCN were always higher than 85% and 87% for the HiPOD and for the LoPFA operating modes, respectively. The obtained results are quite comparable with those obtained using all the 364 variables of the optimal spectral range, and the prediction of TS data is always better for acetic anhydride and ephedrine.

A further refinement of these models will be made by evaluating the possibility to perform a further selection of the selected wavenumbers by GA, focusing on the single targets separately. Not all the selected wavenumbers, in fact, could be useful for the identification of a single target molecule. Moreover, a careful setting of the threshold limit for the detection of each single target will presumably allow us to further increase the probability of detection in the HiPOD mode and to lower the probability of false alarms in the LoPFA mode.

## 6. CONCLUSIONS

A complex strategy, based on the use of various signal processing and multivariate methods, allowed us to estimate the possibility to detect the presence of four drug precursors in vapor phase at low concentrations and in presence of a multitude of interfering chemical species. In absence of real experimental data, these results are driving the construction of the final EC-QCLPAS device. The work made will be relatively easily turned to the spectra now recorded with the newly developed EC-QCLPAS instrument, which in turn will help check the correctness of the classification models and update it, leading to a continuous improvement of the detector classification performance. Moreover, model improvements on locally acquired data will allow a better (more precise) tuning of the single instruments to the requirements of specific applications: for example, it will be possible to tune an instrument working in a specific airport custom with respect to this environment (in terms of gas mixtures variability), which is different from the environment that can be found within containers in a freight port.

## REFERENCES


[1] Secchi, A., Fiorello, A. M., D'Auria, S., Varriale, A., Ulrici, A., Seeber, R., Uotila, J., Venditto, V., Estensoro, P., Colao, F., "Drugs and precursor sensing by complemenTing low cost multiple techniques: Overview of the European FP7 Project CUSTOM", Proc. SPIE 8545-17 (2012).

[2] Uotila, J., Lehtinen, J., Kuusela, T., Sinisalo, S., Maisons,G., Terzi, F., Tittonen, J., "Drug precursor vapor phase sensing by cantilever enhanced photoacoustic spectroscopy and quantum cascade laser", Proc. SPIE 8545-8 (2012).

[3] Calderisi, M., Ulrici, A., Pigani, L., Secchi, A., Seeber, R., "Experimental design-based strategy for the simulation of complex gaseous mixture spectra to detect drug precursors", Proc. SPIE 8545-23 (2012)

[4] Walczak, B. (ed.), [Wavelets in Chemistry], Elsevier, Amsterdam (2000).

[5] Cocchi, M., Seeber, R., Ulrici, A., "WPTER: Wavelet Packet Transform for Efficient pattern Recognition of signals", Chemom. Intell Lab. Syst. 57(2), 97-119 (2001).

[6] Cocchi, M., Seeber, R., Ulrici, A., "Multivariate calibration of analytical signals by WILMA (Wavelet Interface to Linear Modelling Analysis)", J. Chemometrics 17 (8-9), 512-527 (2003).

[7] Dunayevskiy, I. Tsekoun, A., Prasanna, M., Go, R., Patel, C. K. N., "High-sensitivity detection of triacetone triperoxide (TATP) and its precursor acetone", Appl. Optics 46 (25), 6397-6404 (2007).

[8] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., Wold, S., [Multi- and Megavariate Data Analysis. Part I], Umetrics Academy, Umea (2001).
[9] Pigani, L., Foca, G., Ulrici, A., Ionescu, K., Martina, V., Terzi, F., Vignali, M., Zanardi, C., Seeber, R., "Classification of red wines by chemometric analysis of voltammetric signals from PEDOT-modified electrodes", Anal. Chim. Acta, 643(1-2), 67–73 (2009).
[10] Leardi, R., [Nature inspired methods in chemometrics: genetic algorithms and artificial neural networks], Elsevier, Amsterdam, 169-196 (2003).
[11] Leardi, R. , Lupiáñez Gonzales, A., "Genetic algorithms applied to feature selection in PLS regression: how and when to use them", Chemom. Intell. Lab. Syst. 41, 195-207 (1998).
[12] Box, G. E. P., Hunter, W. G., Hunter, J. S., [Statistics for Experimenters], John Wiley& Sons Inc., New York, NY, 291-342 (1978).