REPORT FOR ESF-INTROP EXCHANGE GRANTS

IMPROVEMENTS IN SPECTROSCOPY DATA PROCESSING : FASTER PRODUCTION AND BETTER RELIABILITY OF LAB DATA

M. Ródenas, FUNDACIÓN CEAM (Paterna- Valencia, Spain)

1. PURPOSE OF THE VISIT

Spectroscopic techniques are widely used in many research centres, including simulation chambers sites, due to its high temporal resolution and its versatility in the simultaneous characterisation of a considerable number of gas compounds, what is especially important in complex mixtures. Besides, these techniques are known to be very reliable because they do not imply any manipulation of the sample that could perturb the reactive mixture and therefore the results.

The groups which are working on spectroscopy often use very different techniques to analyse their data: subtraction from commercial programs, home-made automated analysis software or even manual subtraction of reference spectra.

For example, the typical method for infrared (IR) spectra analysis used is a manual process, which can be a very cumbersome task and affected by the subjectivity of the analyst. In contrast, automated methods would considerably reduce the analysis time, allowing all those compounds absorbing in the same wavelength range to be analysed at once. Furthermore, the error coming from the difficulty when retrieving by hand a compound that interferes with others would be reduced, what would improve the quality of the data.

In the case of UV-Vis, different programs already exist for the automatic analysis of DOAS data. An effort must be done in terms of software improvement with the aim of decreasing the number of freedom degrees to choose during the analysis (what makes a difference in the results) and therefore the analyst intervention.

In order to improve the quality insurance of these data it is of primary importance to share knowledge in order to establish a discussion on the best and common practices to adopt. This was precisely the aim of the stay, where the general objective was to acquire expertise in spectroscopic techniques, working with the research group of Dr. Doussin at LISA. The work was closely related to the topic treated in the PhD thesis of M. Ródenas on the improvement of analysis tools in spectroscopy.

LISA was chosen because of their expertise in this field. Indeed, the research group leaded by JF Doussin runs since 1996 two analytical in-situ spectrometric pathways (in the UV and in the IR ranges) and has developed in 1999 his own spectrometric software. Furthermore, LISA comprises two other groups where similar algorithms are used to process satellite data or remote spectrometric sounding of other planet atmospheres.

Therefore, the aim of the visit was to improve, with the advice of LISA people, and to test a home-made software developed at CEAM for the automatic analysis of IR data. This software was compared with the one developed at LISA and with an automated classic fitting routine implemented in commercial programmes. Furthermore, the effect of applying a pre-processing of the spectra to analyse was tested. An adapted version of the CEAM's software for the analysis of UV-Vis data that includes a new algorithm for the filtering process, what would reduce the required intervention of the analyst, was also discussed.

To make the study a set of samples accounting for complex gas mixtures were different gases interfere each other (IR range) was analysed. Particular cases were studied.

The last part of the stay was reserved to visit other groups of LISA or other institutes working closely to LISA were research on developing and enhancing spectroscopy-related instruments is being done. The objective here was to establish lines for future collaborations and to facilitate the transfer of knowledge in spectrometric data treatment.

2. DESCRIPTION OF THE WORK CARRIED OUT DURING THE VISIT

The 6 week stay was done at LISA Labs. (París), under the supervision of Dr. Jean Francois Doussin who leads the group of Measurement and Reactivity of Species of Atmospheric Interest. The work was also supervised by Dr. Bénédicte Picquet-Varrault.

According to the working plan proposed in the report presented to apply for the ESF INTROP grant, the work was organized as follows:

1. **State of the art.**

   During the first stage of the stay, there was a review and a discussion on the state-of-the-art concerning software for IR analysis. In fact, some automatic procedures exist for the analysis of IR data. Classic automated techniques such us linear or non-linear fitting routines are implemented in commercial software packages. Nevertheless, many of the groups who work in the characterisation of data from IR spectra still use a manual procedure arguing that it better allows having an overview of the analysis process carried out and to check that interfering products formed are not confused with the compound being analysed. In that sense, througout this work, there was an agreement that automatic methods are not only quicker, but also the subtraction of different compounds absorbing in the same spectral range is more reliable than when done "at a guess" (manual analysis), especially when they show broad features and there is not a sharp structure that the eye can recognise easily.

   In recent years, some sophisticated techniques like Principal Component Analysis (PCA), Independent Component Analysis (ICA) or Neural Networks (NN) are been developed and applied to analyse IR spectra. Examples can be found in satellite data retrievals (Aires, 2002), mainly focused in a couple of compounds (normally CO or Ozone) or in the analysis of synthetic samples of gasoline constituents (Pasadakis,2006). These procedures are not, to our knowledge, of extensive use to the analysis of the characterisation of a real complex mixture of compounds, as is the case of interest in this work. Other classic methods used in other applications (UV) like minimum least squares can be found in commercial programmes recently enhanced (eg. OMNIC). Nevertheless, among the methods available to remove the baseline (formed due to aerosols, equipment instabilities or unknown broadband shaped products), it allows polynomials up to a degree of 3. In some cases, it must be said that wave shaped baselines due to unknown broadband compounds can not be removed by these methods since such polynomial do not fit their shape properly, appearing in the residual spectra and interfering with the compounds of interest.

2. Overview of the software developed at CEAM and at LISA: similarities and differences.

Both software calculate the concentration of the compounds in an automatic way, and are different in terms of algorithms used for the data evaluation.

LISA programme uses a scalar product decomposition method where the spectra are multiplied by the references of the pure compounds measured in advance. In this way, the projection of each compound reference on the sample through the scalar product, accounts for the concentration of such compound in the sample analysed. The equations are written as a matrix from which, once solved, the concentration can be obtained.

Oppositely, the method implemented at CEAM is based on a classic minimum square fitting algorithm. The improvement added is a new filter algorithm developed to remove the broadband contributions in the spectra due to a not flat baseline or to unknown wide absorption compounds. This is an important issue when dealing with IR spectra, since the high number of compounds absorbing in the same range not always allows having references of all of them, what would result in wrong retrievals. The filter method is based on splitting the spectral range in overlapping windows, where low degree polynomials are adjusted, while taking into account the contribution of the pure compounds on the spectrum shape, to remove the broadband. All the corresponding equations are introduced in a matrix from where the concentrations can be retrieved. Since the width of the windows is small enough, each polynomial allows to model, and therefore to filter, properly the broadband or even the unknown compounds, allowing a better identification of the compounds of interest with less interferences than with other methods.

The program can be used to analyse both, FTIR and DOAS spectra, with slight changes to deal with the format of the data. Parameters to be used with FTIR data were optimised during this work.

3. Tests on IR spectra.

It was accorded to focus the present work on the analysis of IR data. The study was focused on FTIR data because the DOAS device available at LISA has a configuration that allows getting simultaneously two beams: one passing through the sample and a second beam directly collected from the lamp. This set-up that uses a 2-dimension CCD avoids broadbands in DOAS spectra resulting from changes in the lamp because they are continuously monitored and corrected for. In that sense, the algorithm for filtering the spectra is not as necessary as in many other typical DOAS configurations, as is the case of the DOAS system at EUPHORE. Nevertheless, the software was discussed and also its applicability in both set-ups to the analysis of spectra with broadbands coming from aerosols or compounds with interfering broad absorptions in the UV.

Therefore, FTIR spectra were used to perform the tests. Best practices and the most convenient algorithms for IR analysis were reviewed and discussed. The convenience of adding new tools were summarised: eg., software to adapt the spectra (UV and IR) to the analysis conditions of the groups involved (change of resolution, apodization, etc) could be of interest for future collaborations.

4. **Selection of spectra in the tests.**

Spectra belonging to 3 experiments were selected for the analysis. The aim was to test the behaviour of the LISA and CEAM's software under different conditions:

a. A mixture including 16 compounds to check the software on samples with high degree of interference among the compounds.

b. Spectra with CO, CO_2 and H_2O, to check how the softwares analyse sharp structures.

c. Spectra containing HONO + nitrite, to check the software when one of the compounds absorbing is unknown, i.e., no reference of the pure compound is available.

A comparison of the software was done through its application to the analysis of the spectra of such experiments. Also, there was a comparison to the classic manual method. Furthermore, there were data available obtained through a manual analysis for some of the compounds. The LISA group provided "real concentrations" for the tests *a* and *b* as result of the average of several analysis performed applying the LISA software with different parameters (width of wavelength range and spectral regions). These were used as reference for quantifying the goodness of the data results obtained with the softwares. Data from other methods were not used in the average since at that time, CEAM software was being enhanced to deal with IR data.

This was a tricky point since there was a discussion on whether those "real concentrations" would be favouring the LISA results as they were obtained from its software. Nevertheless it was accorded to be the best option in the absence of other data.

For the third experiment, DOAS data were available and they were used as "real concentrations" for comparing the FTIR data obtained.

5. **Visit to other groups.**

Taking the advantage that LISA is a recognised group with an important trajectory in spectroscopy, there was a visit to the facilities and to people who are making important efforts in this field. The aim was to visit and know the work of these groups, not only on analysis software but also in terms of instrumentation development. In particular, the LISA simulation chamber as well as the new CESAM chamber were visited.

The DOAS set-up mentioned above used at LISA was seen. This set-up would be very useful at CEAM to measure absorption cross sections. The reason is that, given the current DOAS configuration at CEAM, changes in the lamp with time influence the low frequency contribution of the absorption cross sections. This is especially important when dealing with non- or semi-volatile compounds since it takes time to introduce them into the chamber. Nevertheless, with the LISA set-up, there are simultaneous measurements of I and Io, therefore according to the formula ln(I/I_o) from where the cross section is obtained, changes in the lamp affect both in the same way, and are compensated, allowing to get more reliable absolute cross sections.

In the time of the stay, a new FTIR system was being installed in the CESAM chamber, what was interesting because, unlike CEAM, no laser is used for the alignment.

Apart of knowing the work of Dr. Doussin group members, there was a visit to Dr. Maxim Eremenko, who works in satellite data and to Dr. Johannes Orphal from LISA. The second one commented his familiarity with the UV data analysis more than with the IR analysis. He found interesting the application of the filter algorithm to UV spectra, what maybe would allow to increase the analysis evaluation region or to obtain optical properties in aerosols, what should be studied further. In satellite data, small UV evaluation regions are used, what means that the broadband can be removed by fitting a single polynomial to the whole region. There was a discussion on this issue because the CEAM algorithm would be useful when using wider regions where a sole polynomial would not be enough to model the broadband. Dr. Orphal argued that it is not clear why the community is not using such wider ranges. One of the reasons could be that, with classical methods the broadband is not well removed in wide regions and it implies an interference in the analysis. It must be noted that as long as the analysis region increases, the detection limit improves.

Regarding the thesis work of M. Ródenas, where apart of the software tested during the stay, other algorithms have to be investigated, Dr. Orphal suggested to work on PCA methods, already used in (Gomez et al., 2004) where he was co-author.

~~3.~~ DESCRIPTION OF THE MAIN RESULTS OBTAINED

On the first hand, some enhancements were done on the software allowing dealing with the high number of compounds analysed at the same time. Note that the algorithm had been programmed to work with UV data, where the compounds in the sample are less than in IR. For example, problems of deficient rank in the matrix used in the algorithm were corrected for and data formats were treated.

The algorithms tested in this study were the following: manual analysis, classic fitting (minimum least squares) and the algorithm proposed in this work, based in a home-made modified classic fitting that includes a novel filter process as explained above (referenced as polynomial-windows in this work). This software also allows making the derivative of the spectra prior to its analysis. This simple pre-processing of the spectra that enhances their absorption features can be especially useful with compounds that do not show sharp shapes.

The graphs below show estimates of the mean and standard deviation of the normal distribution of the error values (error = theoretical concentration – calculated concentration). Also a quantification of the errors in percentage is done, according to the formula:

$$\text{ERROR} = \sqrt{\frac{\sum (C_{Theoretical} - C_{Calculated})^2}{n}} \cdot 100 \Big/ C_{Max\_Theoretical}$$

1. ANALYSIS OF A COMPLEX MIXTURE

A total of 23 spectra obtained in an experiment carried out at the LISA chamber were analysed. The mixture contained 16 compounds, and all of them were analysed at the same run of the software: Acetaldehyde, Ethyl Acetate, Acetoxyacetaldehyde, Acetic acid, Formic acid, anhydride acetic, anhydride formic acid, formaldehyde, nitric acid, nitrous acid, dinitrogen pentoxide, methyl nitrate, methyl nitrite, nitric oxide, nitrogen dioxide and PAN. Next figure shows an example of a spectrum analysed, in red. The other curves are the reference spectra of the pure compounds.

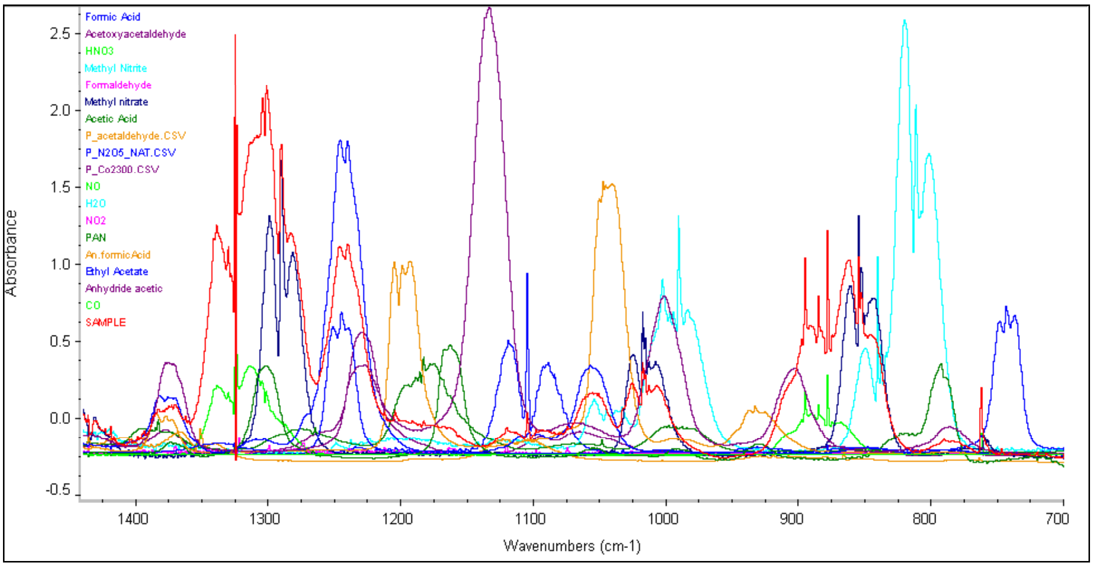

Fig. 1. Sample to analyse (red) and reference spectra of the pure compounds contained in the sample.

Several tests were made in order to optimise the algorithm. In the first one, the effect of saturation in spectra was studied. Next figure shows a reference spectrum of HNO3 (in red). Green line is a spectrum showing saturation of this compound. It implies that the shape of the absorption is not maintained. Blue line shows a higher concentration and higher saturation in the sample. In this case the typical features of HNO_3 can not be recognised.

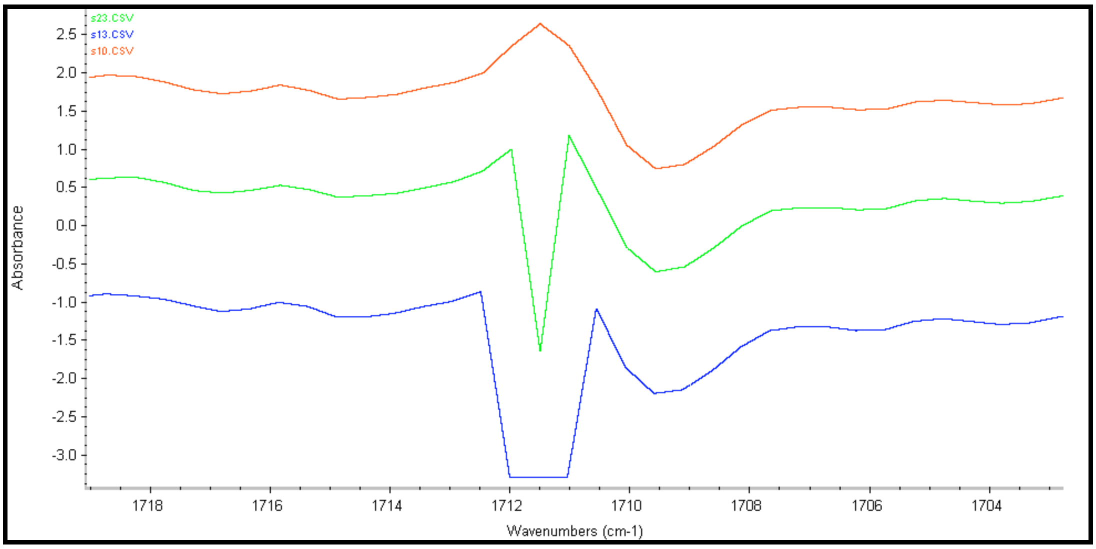

Fig 2. Temporal evolution of the samples when HNO_3 is introduced in amounts that reach saturation of spectra.

In this case, HNO_3 can not be analysed properly. Lisa data show the best results under these conditions, while classical methods and the proposed ones work worse. This is due to the fact that in the case of the LISA algorithm, all the pixels in the x-axis accounts with the same weight, while in the other methods, the best fitting is found when the residual (difference between the spectra modelled according to the calculated concentration and the spectra analysed) is minimised. This means that, for example, in the first case 10 pixels saturated in a sample of 100 pixels account exactly for an error of 10%. In the other cases, due to the algorithm used, if those 10 pixels show a strong absorption (y-axes), the fitting routine will try to compensate that negative peak by increasing the concentration of other compounds that show a similar negative peak (or will calculate negative concentrations if the peaks are positive) yielding in a lower concentration of HNO_3, so that the sum of both results in a minimum residual. It could happen that the wrong concentration of the second compound will force to compensate the residual in another part of the spectrum by calculating the concentration of third compounds absorbing in another common region (note that a compound can show several absorptions in different regions), and so on. Depending on the size of the anomalous peak, this effect will be more or less important, and the values of the concentrations retrieved may vary. In the case studied, the amplitude of the saturated peak is strong, therefore the error is high.

To avoid this effect, the method proposed has been improved by using a corrected fitting procedure, where those erroneous pixels, which are seen as a deviation of the linearity by the algorithm, have less weight and account less to calculate the concentration factors. Next figure shows the results obtained by the LISA algorithm, the classic method and those proposed in this work: with and without such improvements in the analysis of the derivative method and of the use of the filtered method described in this work (Pol-windows).

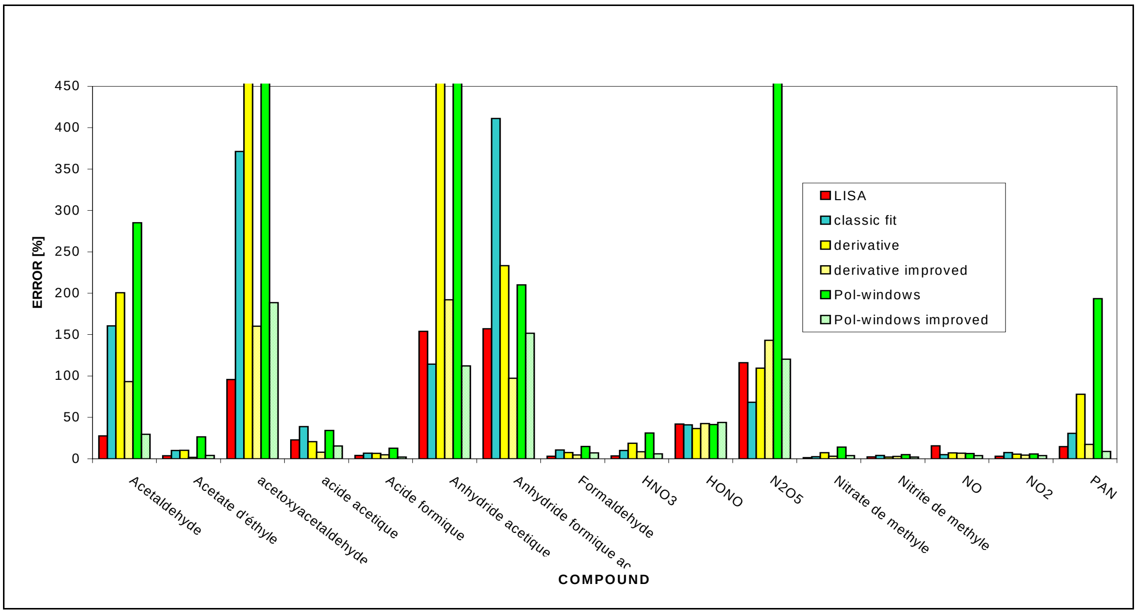

Fig 3: Analysis using the algorithms tested. Rows with lines show the results using the improved algorithm that gives less weight to the anomalous – saturated pixels.

The LISA algorithm is not so sensitive to the problem of saturation as the other methods, as explained above. The improved method clearly enhances the results and the goodness of the methods is similar.

On the contrary, there is an important increase of the time consumed to analyse the samples if the improved method is used, reaching even 100 seconds per sample. LISA software is a very fast method that just takes ca. 1 second per sample. Looking at the graph above where there is no clear goodness of one of the methods compared to the others, the LISA method would be proposed to be used to analyse the data. Nevertheless, it must be said that in normal conditions the saturation of the spectra should be avoided to get reliable concentration profiles, not only because wrong effects as those described above could appear but also because non-linearities of the absorbance with the concentration exist, yielding in erroneous results. In fact, a general rule accepted is that working with absorbances above 1 must be avoided, while in the region analysed, values of up to 2.5 are reached.

The same evaluation in a region that avoids the saturation in the spectra (1272-767 cm-1) did not require the use of the improved method. Next figure shows the results obtained:

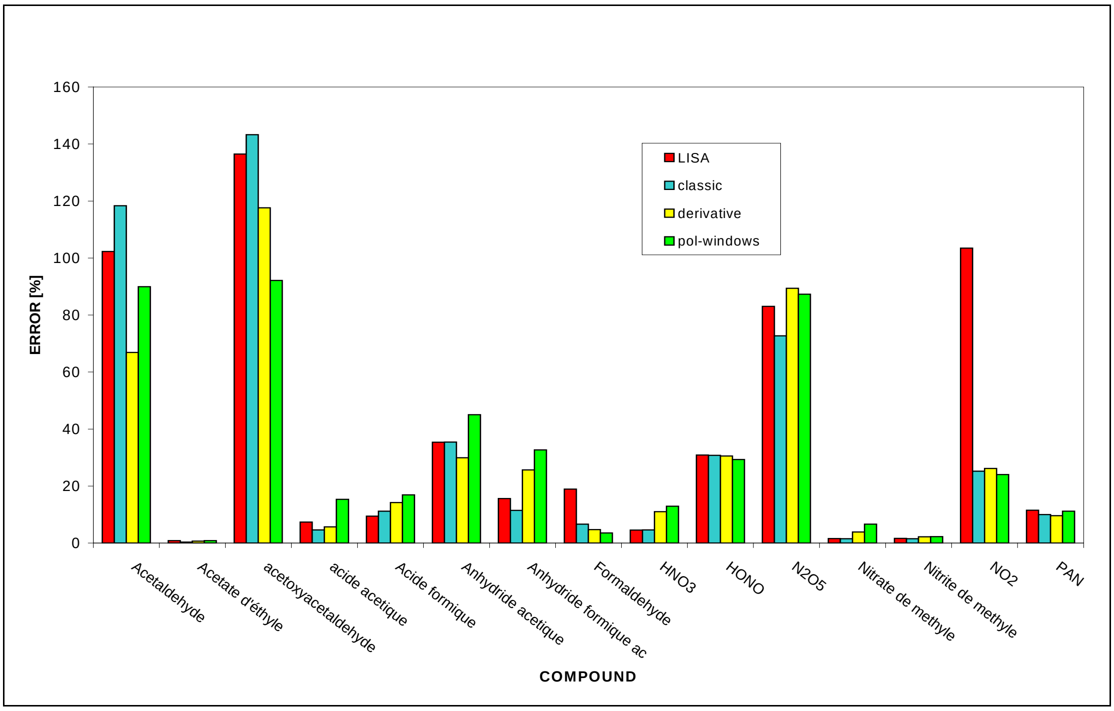

Fig. 4. Analysis in a non-saturated region.

Note that errors obtained in all the analysis are lower than if the saturated region is used, as expected.

From the data obtained, again it can not be said that any of the methods employed work clearly better than the others. Therefore, again an important parameter to take into account is the time consumed by the algorithms. Next, a table showing a comparative of the time used to analyse the set of 23 spectra belonging to the experiment studied.

Algorithm	Time to analyse 23 spectra (sec)
LISA	11
Classic fit	23
Derivative	34
Polynomial-windows	310

Table 1. Comparative of the time consumed to analyse 23 spectra.

In order to compare these evaluation times, it must be taken into account that the LISA programme runs in a computer working with Window-95, while the other softwares have been tested using a PC under Windows-XP with 2.4Ghz and 539MB RAM. Nevertheless, at the moment it is not possible to install the LISA programme in a higher platform than W-95, although it is foreseen to reprogram it.

2.ANALYSIS OF NO, CO AND CO$_2$ (2322-1878 CM^{-1})

In this second test, the analysis of a different kind of compounds has been done. NO, CO and CO$_2$ show high frequency structures and the behaviour of the algorithms tested was studied. The same spectra as those used in the previous test were analysed, although in the corresponding spectral region.

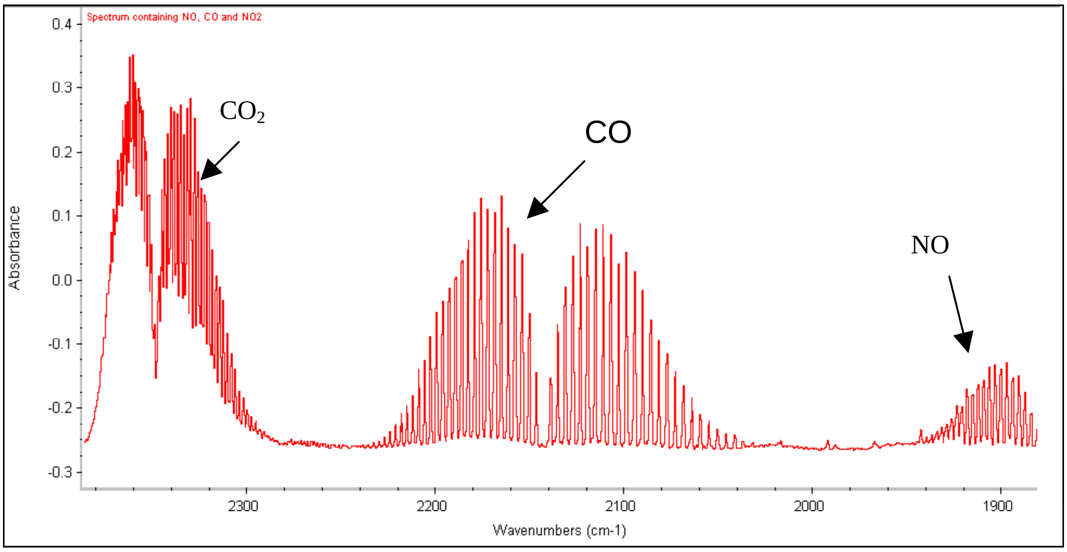

Fig. 5. Spectrum containing the typical sharp absorptions of CO$_2$, CO and NO.

Next, the standard deviation and the mean of the gaussian distribution of the error function (theoretical value minus calculated).

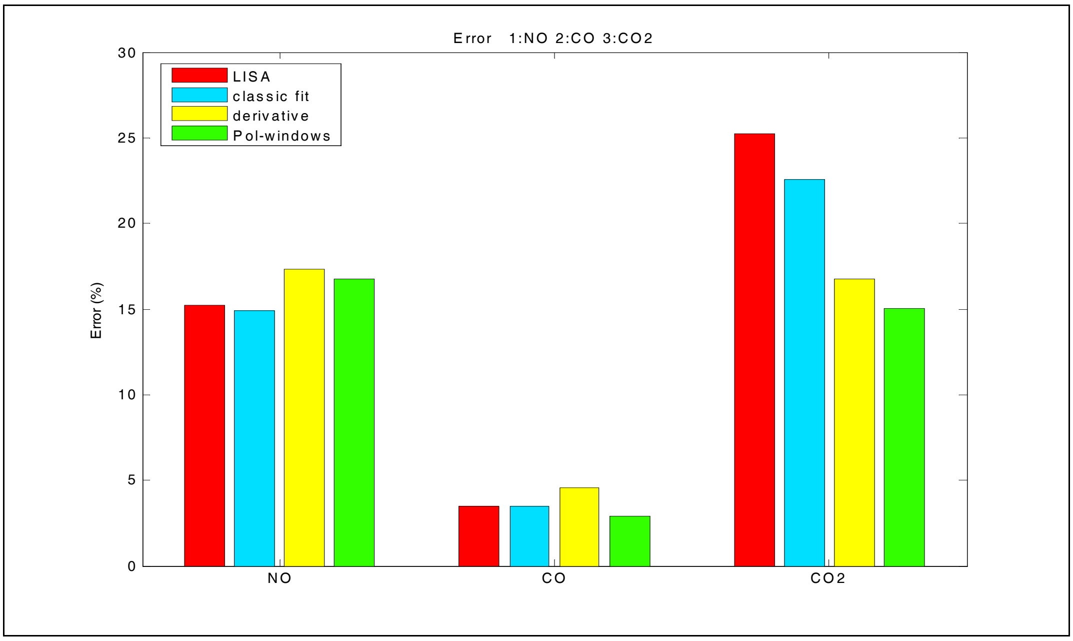

Fig. 6. NO, CO CO_2 error with different analysis methods.

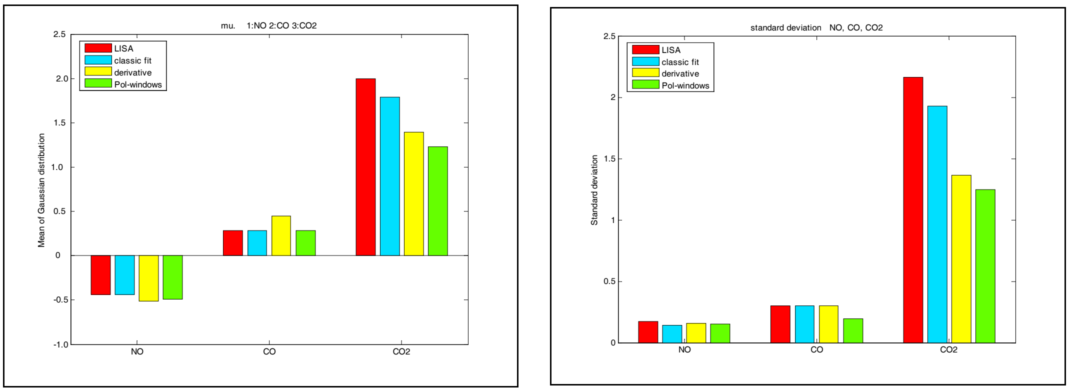

Fig. 7. Mean and standard deviation of gaussian distribution

Results are similar. The use of polynomial-windows proposed in this word show slightly better results. The computing time in this analysis for 23 spectra is shown in the next table.

Algorithm	**Time to analyse 23 spectra (sec)**
LISA	10
Classic fit	10
Derivative	30
Polynomial-windows	406

Table 2. Comparative of the time consumed to analyse 23 spectra.

3.EXPERIMENT WITH HONO + NITRITE

The aim was to check the behaviour of the methods when some of the compounds in the mixture was unknown, and therefore its reference was not used in the evaluation. Spectra were analysed using the same spectral range (745 – 1100 cm-1). In this case, the mixture contained SF_6, HONO and two Nitrites. Concentration profiles obtained with the different methods used are shown. Data obtained with DOAS were used as “real concentration” to compare the results.

The improved method in the CEAM’s software was not used in the following data.

3.3.1. SF_6 + HONO as references

The range analysed contained SF_6, HONO and Propyl Nitrite. In the fitting routine, only the references of the two first compounds were used. HONO and Propyl Nitrite show their main absorptions interfering each other as can be seen in Fig. 8.

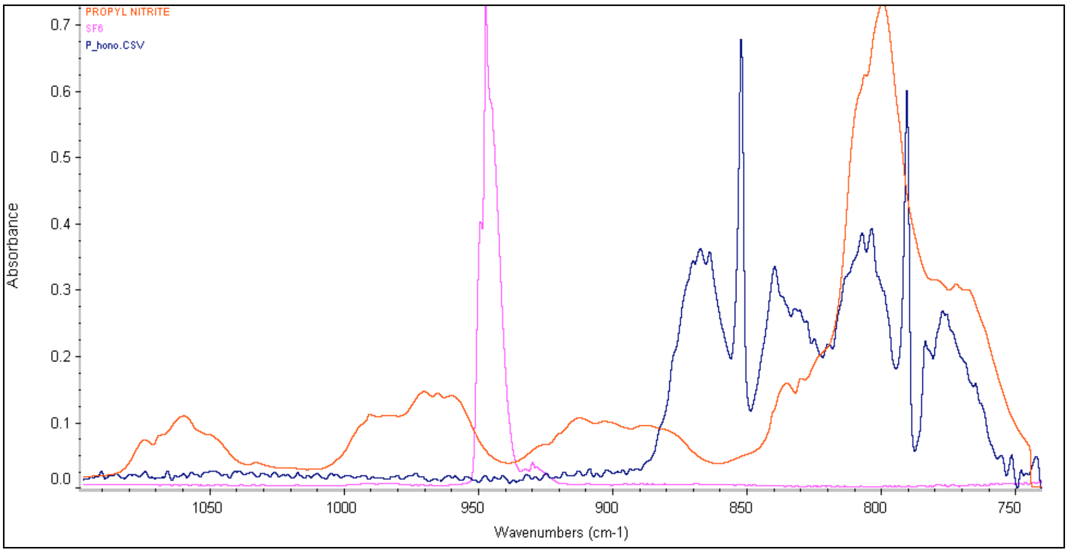

Fig 8. IR absorption of SF6, HONO and Propyl Nitrite

The figure above shows the concentration retrieved by the different methods tested. HONO data analysed by DOAS is shown for comparison. Note that the addition of Propyl Nitrite causes an interference on the HONO retrieval. Although not shown here, it was observed that the analysis of SF_6 does not suffer this effect and all the methods show similar results, as expected, since its sharp absorption is not in the same spectral region as the main absorptions of the other compounds.

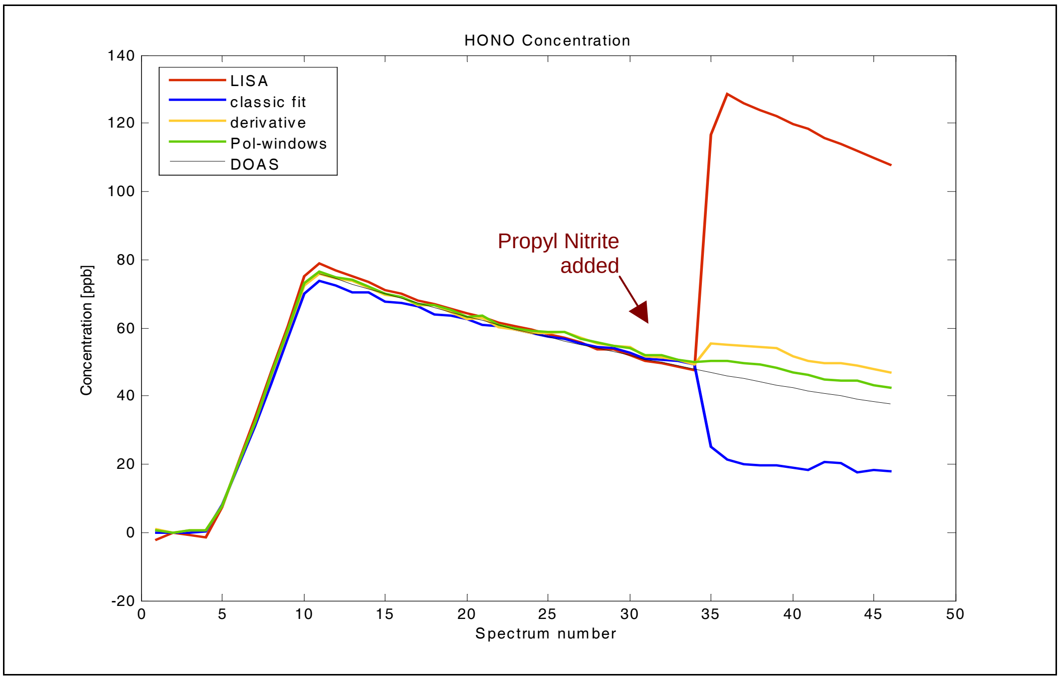

Fig 9. HONO concentration analysed by different methods

The corresponding error compared to the DOAS and the results of the gaussian distribution are shown in the next figures:

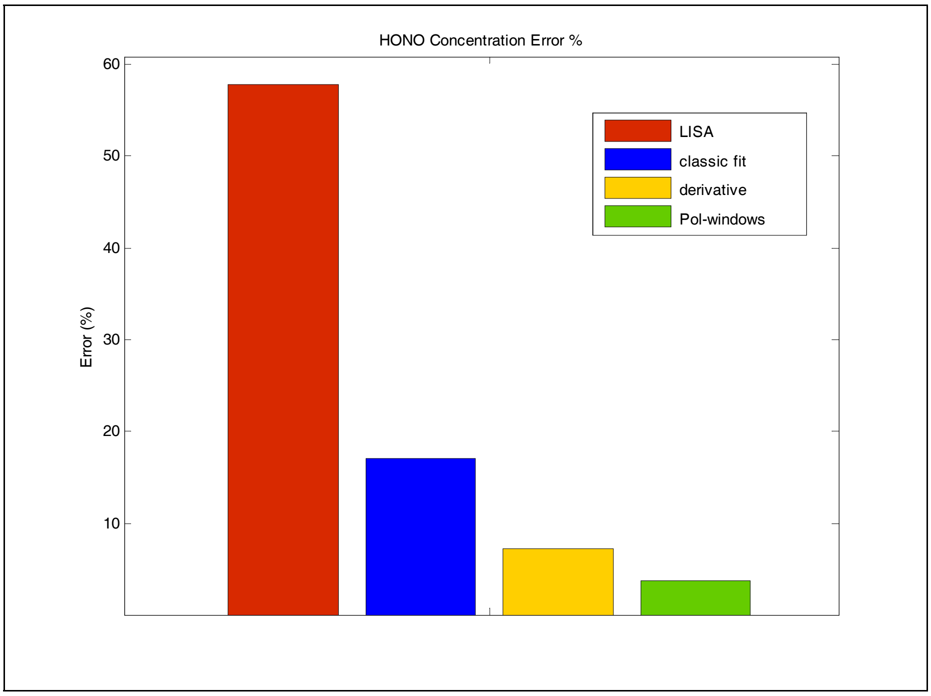

Fig. 10. HONO error with different analysis methods.

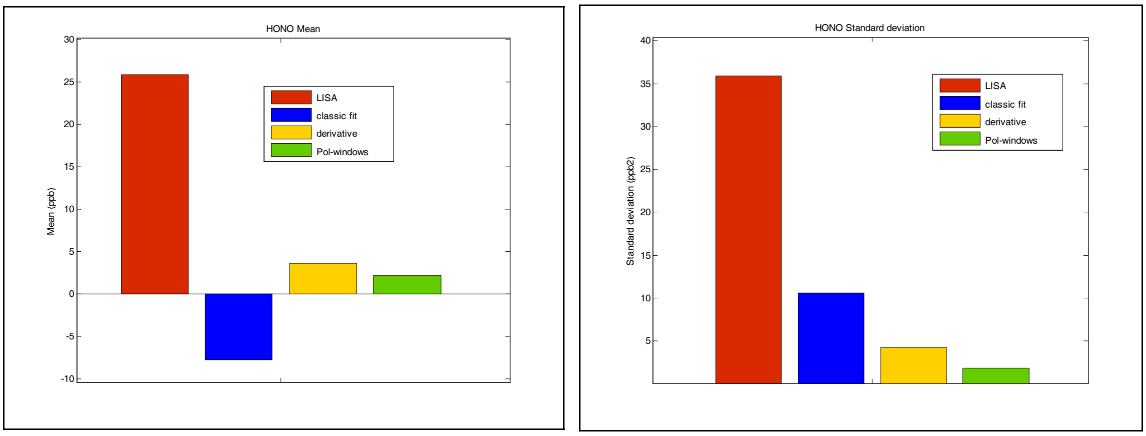

Fig. 11. Mean and standard deviation of gaussian distribution

As seen in the results, the use of the polynomial-window method proposed in this work clearly reduces the error committed. Therefore, in the analysis of samples where a compound is contained, but which reference is not available, this method would be recommended.

3.3.2. SF_6 + HONO + Propyl-Nitrite as references

In this test, the same experiment with samples containing SF6, HONO and Propyl Nitrite was analysed. The references of these compounds were used in the fitting. A second nitrite (Pentyl) was also added to the samples after the spectrum number 46 as can be seen in the figure below. Unlike the Propyl Nitrite, there was no pure reference of Pentyl Nitrite, although it could be observed in the samples that their main absorptions are very similar, and therefor interfere.

The aim of this test was the analysis of HONO in the presence of a compound interfering which reference is used in the fitting, and in the presence of another compound which reference is not used.

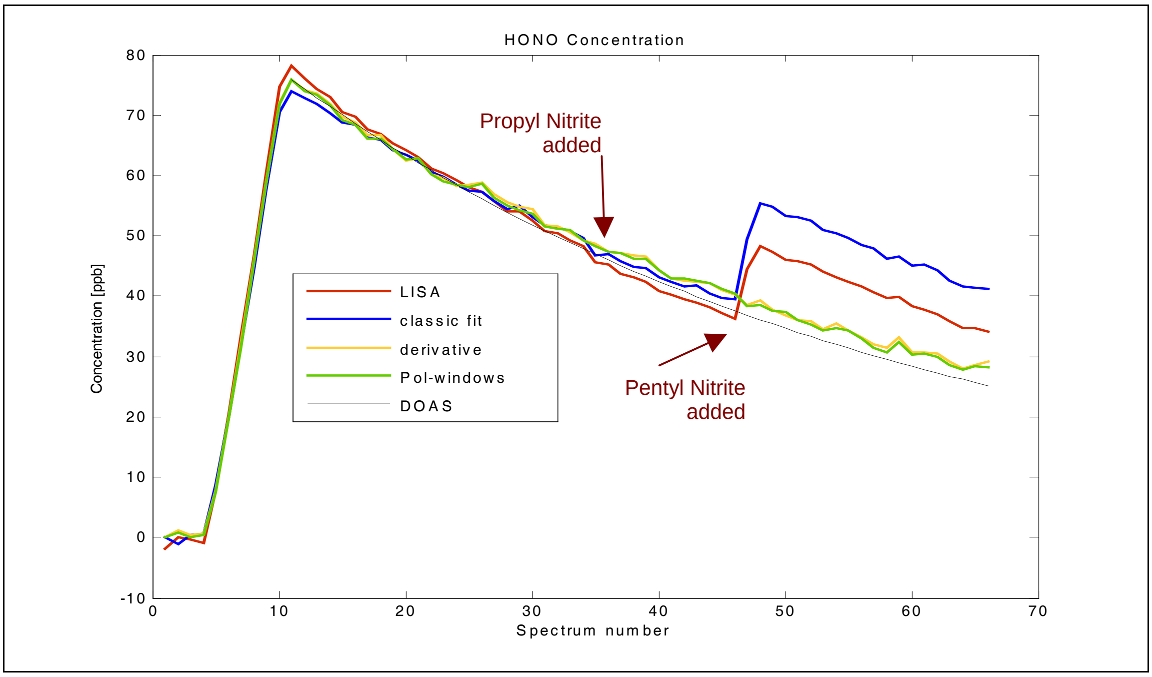

Fig 12. HONO concentration.

As in the previous case, the presence of an unknown compound in the sample implies errors in the determination of HONO. The use of derivatives and of polynomial-window shows similar and better results than the other methods.

The corresponding error concentration results and the mean and standard distribution are show:

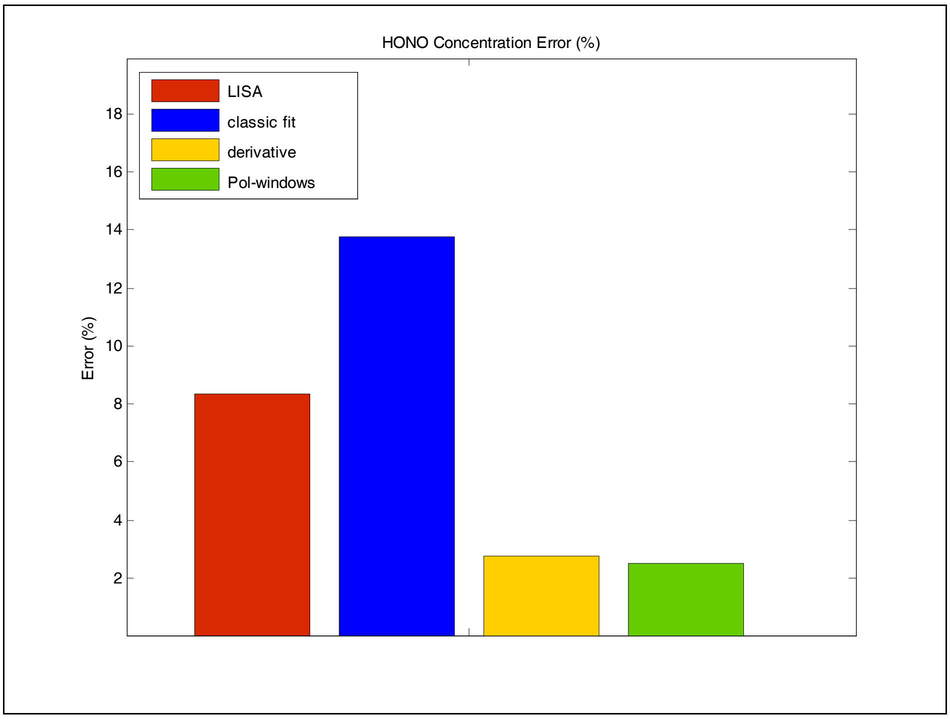

Fig 13. HONO concentration error compared to DOAS data.

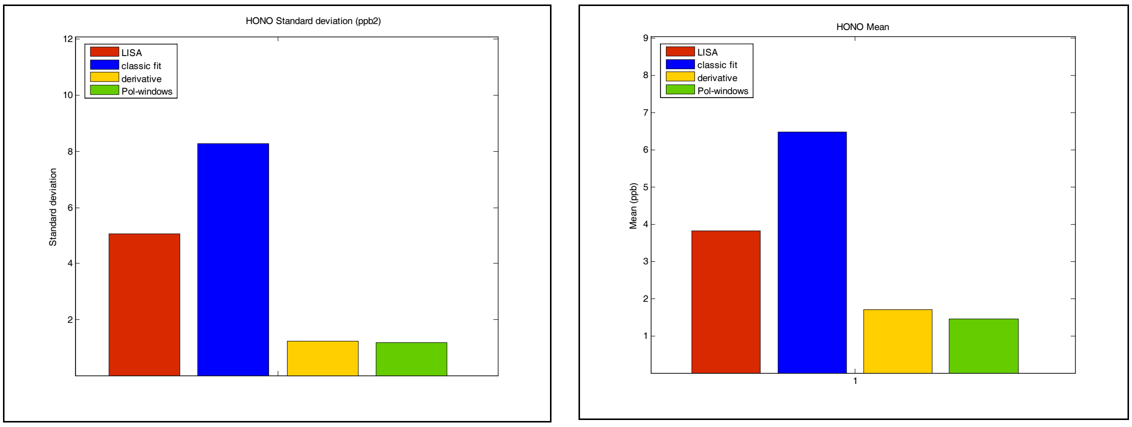

Fig 14. HONO mean and standard deviation. Errors compared to DOAS data.

The concentration of Propyl Nitrite in the presence of Pentyl Nitrite which absorptions are very similar is shown in the next figure.

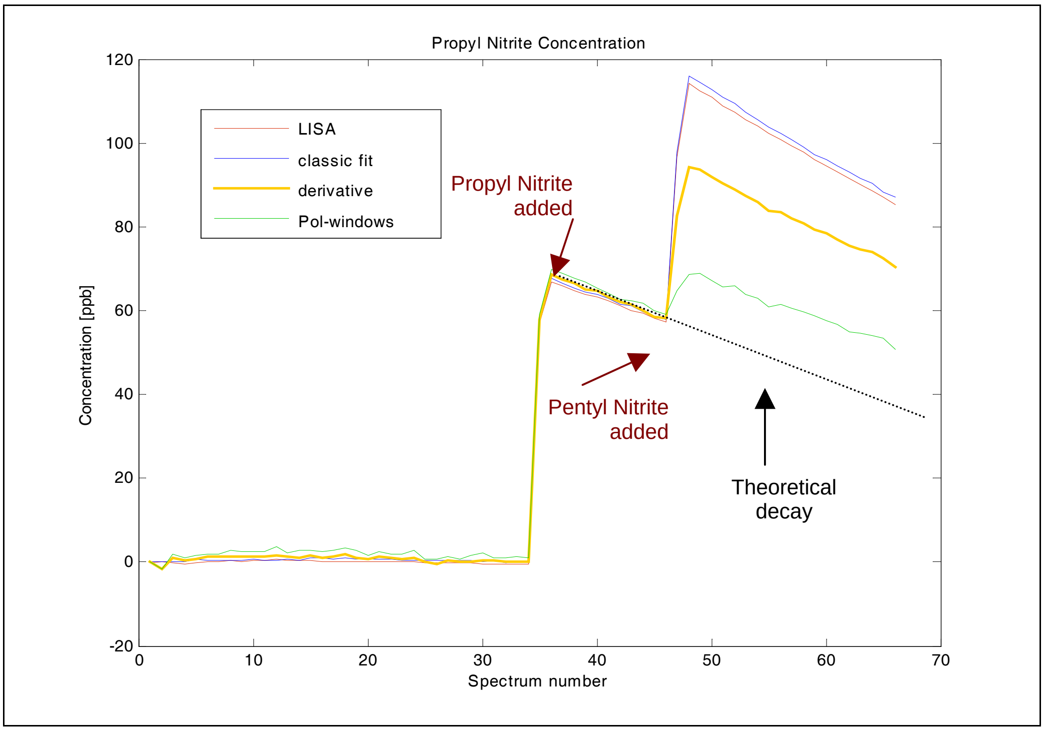

Fig 15. Propyl Nitrite concentration

Although there are no available data on the real concentration of Propyl Nitrite, theoretically, the addition of Pentyl Nitrite to the mixture must not cause a reaction among them because the compound was introduced with the chamber closed and in the absence of other reactants. Therefore, there should not be a step of the concentration in the analysis of Propyl Nitrite. Nevertheless, all the methods see such step as a result of an interference with Pentyl Nitrite (see Fig. 15) , which reference is not available and therefore is not used in the fitting. If we follow the behaviour of the Propyl Nitrite decay (dotted line), the deviation of the data analysis – after the addition of Pentyl Nitrite - by the methods tested with respect to it, results in errors of roughly 122%, 124%, 78% and 28% for the methods of LISA, classic fit, use of derivative and polynomial-windows, respectively.

Therefore, it can be concluded that in the presence of an unknown compound that exists in the sample to analyse, what is a very common issue in FTIR data given the huge number of compounds that can absorb in this region, the use of polynomial-windows decreases the error of the interference. On the contrary, the analysis time is much higher than when using the LISA method or the classic fit. The use of derivatives as a pre-processing of the data can improve the results, reducing the interferences among the compounds.

In the other cases, when references of all the compounds present in a sample to analyse are available, all the methods tested yield in similar results, although the LISA method is the faster.

4. FUTURE COLABORATIONS WITH HOST INSTITUTE

LISA and CEAM are two of the institutes that are involved in the Eurochamp project. For the EUROCHAMP-2, Lisa has proposed an issue on “Development of software and tools”. CEAM will contribute with sharing among the partners involved routines for dealing with spectra. These routines will be used to change the resolution, apodization, etc, of IR spectra so that these can be adapted and used by the different groups. CEAM already has developed software with this purpose for UV spectra. As a result of the stay, LISA will provide the algorithms to CEAM for working with IR spectra and CEAM will contribute programming a tool that can be used as exe file, accordingly.

Also, on the course of the project, the software developed at CEAM and tested during the stay, will be enhanced by means of a user-friendly environment and provided for its use among the groups involved. In any case, LISA can access the software independently of the successful of the project application.

5. PROJECTED PUBLICATIONS/ARTICLES RESULTING OR TO RESULT FROM YOUR GRANT

There is already a contribution to the HONO invited expert workshop, with the poster “HONO measurement inter-comparison by different techniques at the EUPHORE chambers”, where INTROP was acknowledged. The data shown are those studied in the 3rd set of spectra analysed in this work. On the other hand, it is foreseen to publish these results in a journal (e.g. Applied Optics or Anal. Chim. Acta) in a future.

Bibliography:

Pasadakis et al. Identifying constituents in commercial gasoline using Fourier transform-infrared spectroscopy and independent component analysis. Analytica Chimica Acta 578 (2006) 250–255.

F. Aires et al. Remote sensing from the infrared atmospheric sounding interferometer instrument 2. Simultaneous retrieval of temperature, water vapor, and ozone atmospheric profiles Journal of Geophysical Research (2002), Vol. 107, NO. D22, 4620,