AUTOMATIC BLOB FITTING IN COMPREHENSIVE TWO-DIMENSIONAL GAS CHROMATOGRAPHY IMAGES

B. Celse(1), Stéphane Bres(3), F. Adam(4), F. Bertoncini(4), L. Duval(2)

(1) IFP, Technology, Computer Science and Applied Maths Division, BP3, 69390 Vernaison, France
(2) IFP, Technology, Computer Science and Applied Maths Division, 92852 Rueil-Malmaison, France
(3) LIRIS, UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard, 69000 Lyon, France
(4) IFP, Physics and Analysis Division, BP3, 69390 Vernaison, France

ABSTRACT

Two-dimensional gas chromatography is a recent technology which is particularly efficient for detailed molecular analysis. However, due to the novelty of the method and the lack of automated analysis tools, quantitative data processing is often performed manually. Hence, results are strongly user-dependent, time consuming and, consequently, relatively inaccurate In this paper, we extend conventional techniques for signal analysis by utilizing specific characteristics of chromatographic data and by developing new methods for estimating the quantitative contribution of chemical regions from the produced images. Data-driven information is retrieved from chemical quantitative analysis based on Savitzky-Golay automatic peaks location determination, which increases both the processing speed and the analysis efficiency and improves our confidence in experimental repeatability.

1. INTRODUCTION

Comprehensive two-dimensional gas chromatography (GC×GC) is a promising new technology to unravel complex mixtures such as petroleum samples [17], [1]. In GC×GC, the entire chemical sample is submitted to two one-dimensional GC separations involving different properties of analytes such as volatility (*i.e.* separation according to boiling points) and polarity (*i.e.* the class of compounds). The separation is achieved using two columns with different selectivities connected together through a modulator [11] that traps, focuses and re-injects periodically (each modulation period, *typically lasting between* 4 and 10 s) the effluent from the first to the second column. An appropriate column association results in highly organized 2D chromatograms with several thousands of peaks, which are arranged in the form of bands [11].

Detection occurs at the outlet of the second column and is recorded as a function of the elution time. The 2D chromatogram consists into slices (as wide as the modulation period) of the raw data which are stacked side by side. The different steps of a GC×GC analysis are presented on Figure 1 (cf. [3]). Figure 2 represents the 2D chromatogram obtained for the separation of nitrogen compounds contained into a middle distillate sample.

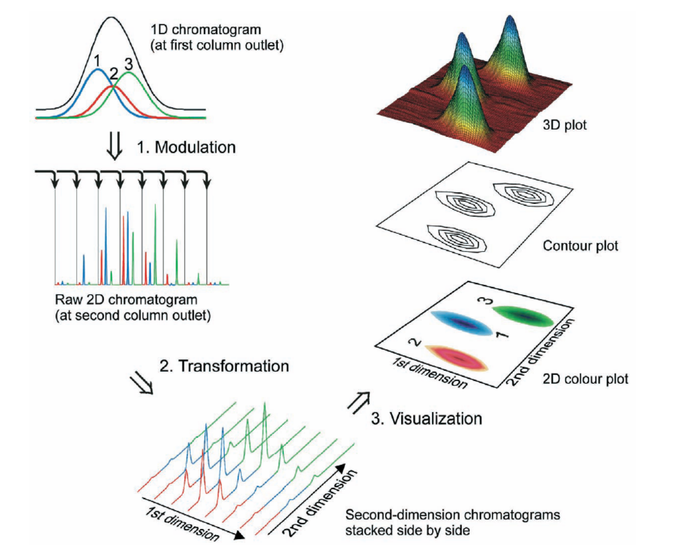


Figure 1: Generation and visualization of GC×GC image.

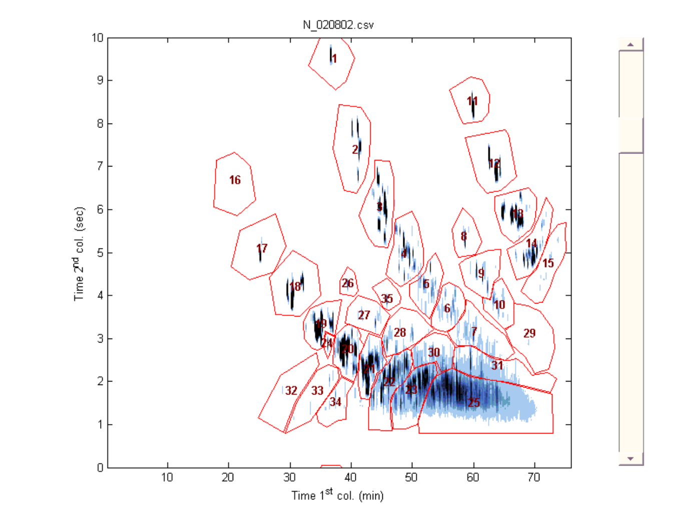


Figure 2: GC×GC image (2D chromatogram) for the separation of nitrogen compounds.

2. GC×GC Analysis

In the literature, several approaches are reported to perform peak quantification in GC×GC. The most common one integrates all individual second-dimension peaks by means of conventional integration algorithms and, next, sums all peak areas belonging to one 2D peak [7], [1]. This type of processing is generally performed either by using two software programs, *i.e.* using conventional 1D GC software programs for peak integration and another program for the subsequent combination of peak contributions.

In a second approach, first a so-called base plane (corresponding to non chemically significant background variations) is subtracted, and subsequently three dimensional peak volumes are calculated by means of imaging procedures [13]. There is an on-going debate on whether this approach can also be applied to the quantification of analytes in complex samples with little or not structured chromatograms. In theses samples, the base plane correction may fail, resulting in illogical negative peaks areas or volumes.

There exists three generic types of applications in chromatography [17].

- The most common type of application is based on converting retention times into peak identities and the corresponding peaks areas into amounts or concentrations. The desired actual information is the concentrations of a limited number of prespecified components. This strategy is usually referred to as "target-compound analysis".
- In the second type of application, there is either not the possibility or not the need to identify all individual peaks. Visualizing a limited number of groups of analytes (*e.g.* acids, ketones, phtalate esters, aromatic hydrocarbons) in a sample of largely unknown composition is the main aspect of interest. Instead of "component groups", the denomination "pseudo-components" is also used. Pseudo-components often have structural properties in common, such as specific groups, an identical number of aromatics rings, a specific configuration of double bonds, etc. Separation of the samples into individual component groups provides valuable information.
- The third type of application ("non target analysis") is performed to obtain an overview of the sample's constituents. In other words, an attempt is made to identify "all peaks" above a certain signal-to-noise ratio in the chromatogram.

The present work presents techniques for the first two applications. Classical data processing steps for these kind of application are [12] (*cf.* Figure 3) :

- background or base plane removal.
- blob detection that is the process of aggregating clusters of pixels that form distinct peaks. This operation is generally performed automatically using a previously generated template (*i.e.* a list of polygonal zones, each one encompassing several peaks). This template includes metadata such as compound names.
- template matching that is the process of moving shifting the corner of the polygonal zone to adapt them to the new analysis.

[17] describes main requirements for these type of applications. In particular, it focuses on quantitative detection and group identification. Therefore this type of application requires group-wise integration and quantification methods.

The template matching step is crucial. It is often user-dependent. Hence, a peak detection algorithm is proposed in the present paper to automate the template matching step and to reduce the analysis' user-dependency. Because blobs are related to the presence of peaks, the main idea of the algorithm is to find peaks inside blobs and then to fit blob frontiers to the start or the stop of each peak. In this paper, we provide then a method to:

- Load a pattern on an new analysis,
- Detect peaks in each column of the image,
- Fit blobs with respect to the start and stop of each peak.

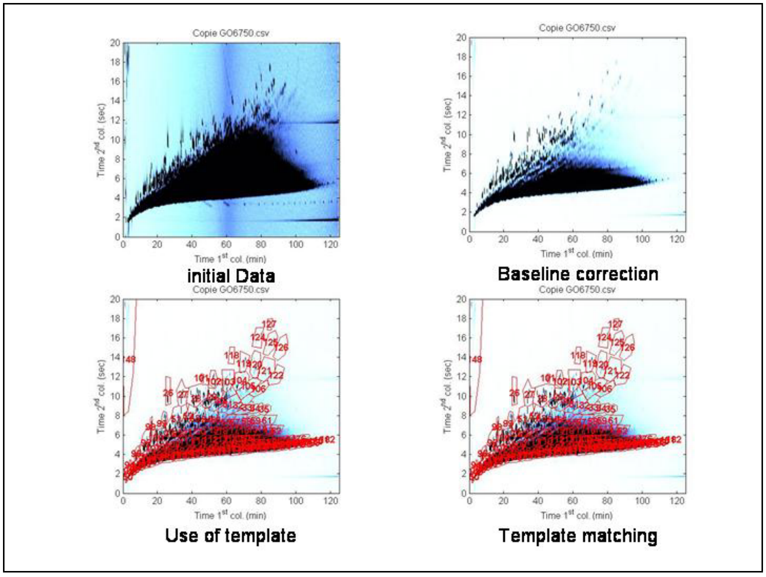


Figure 3: GC×GC data processing steps.

The paper is organized as follows:

- Section 2 presents the peak detection algorithm developed. The use of high-order derivatives was shown to be very efficient for peak finding. However, since the noise is amplified by derivative computation, we apply the Stavitzky–Golay [14] smoother. This strategy allows noise removal without loosing valuable information.
- Section 3 details the algorithm used to fit blobs to chemically related compounds.
- Section 4 provides results obtained from real data. The use of automatic blob fitting considerably improves the results. All these features are implemented in an industrial software named *Polychrom*.

3. PEAK DETECTION ALGORITHM

Several deconvolution techniques have been developed for chromatography. They rely on the assumption that the underlying individual peak profiles (intermingled) within the gross chromatographic signal can be described through mathematical peaks models. This assumption has driven an increased interest in the development of improved peak models ([15], [8], [10], [9]).

Peak detection algorithms often have difficulties in detecting the presence of more than one peak when several compounds coelute, yielding shoulders on main peaks ([9], [4]). To detect peaks, derivatives of the second dimension signal are inspected. The n-order derivatives are computed through the well-known Stavitzky–Golay (SG) algorithm [14]. This technique determines smoothed derivatives on the chromatographic signal based on least-squares polynomial fitting, to compensate for the effect of noise amplification, while preserving the peak's shape.

If we assume peaks as a approximately Gaussian, derivatives of the signal can be used as follows:

- Peak extrema correspond to the root of the first derivative.
- Start and Stop times of the peak correspond to roots of the first, second and third derivative.
- Peak extrema correspond to minima of the second derivative
- Peak extrema correspond to a root of the third derivatives.

The peak detection algorithm is based on root finding in the first and third derivative and negative regions in the second derivative. It is similar to the algorithm proposed by [16].

In the case of weak interference of elution peaks (cf. Figure 4), a peak is detected at time t, when following constraints are fulfilled:

1. The first derivative is close to zero. It should correspond to a sign change from negative to positive regions;
2. The second derivative must be a minimum (negative one);
3. The value of signal must be superior to a threshold.

The start time of a peak (respectively the stop time) is detected a time t which corresponds to one root on the first derivative before (respectively after) the maximum of the peak. Figure 6 presents an example of peaks detection in a real signal a exhibiting partial co-elution of peaks. It is obvious that the peaks detection is rather accurate.

In the case of strong interference of elution peaks (cf. Figure 5, bottom left); there are no roots in the first derivative between two peaks (figure in the left). A peak is detected at time t, when following constraints are fulfilled:

1. The third derivative is close to zero. It should correspond to a sign change from negative to positive regions;
2. The second derivative must be a minimum (negative one);
3. The value of signal must be superior to a threshold.

The time start of a peak (respectively time end) is detected at time t which corresponds to two roots on the third derivative before (respectively after) the maximum of the peak.

Figure 8 shows an example of strong co-elution. In this case, simple integration fails to detect properly individual peaks (cf. Figure 7). Peak does not match with root on first derivative. Complex integration is then required to detect peak.

The second algorithm is more sensitive than the first one but require a more complex parameter selection and tuning.

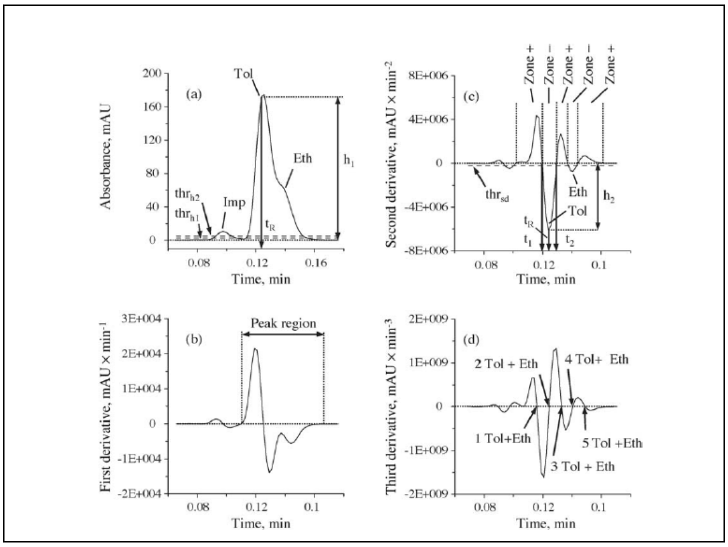


Figure 4: Use of derivative in the case of partial co-elution (top left : signal, bottom left : first derivative, top right : second derivative, bottom right : third derivative). First and second derivatives achieve to detect individual peaks.

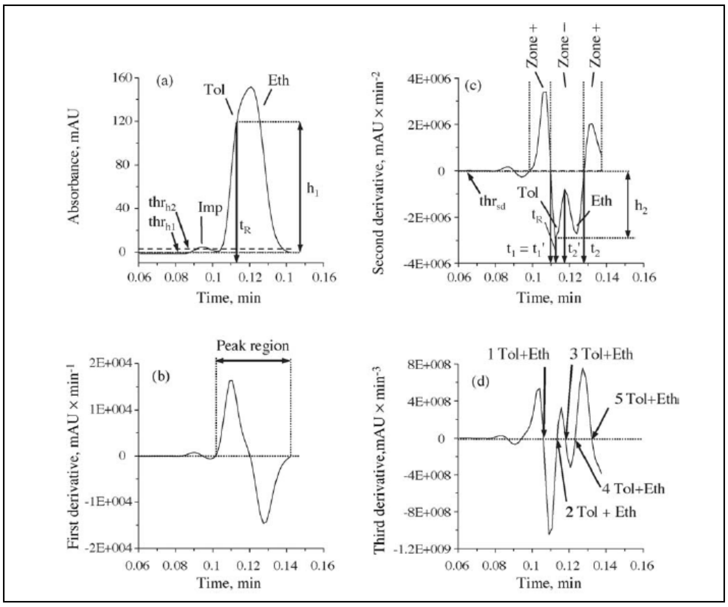


Figure 5: Use of derivative in the case of strong co-elution (top left : signal, bottom left : first derivative, top right : second derivative, bottom right : third derivative). Third derivative must be used in order to detect peaks.

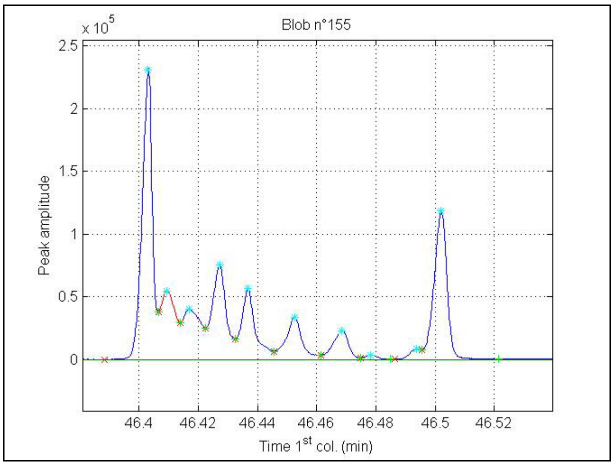


Figure 6: Example of detected peaks (red stars correspond to start time, green stars correspond to stop time, blue stars correspond to peaks)

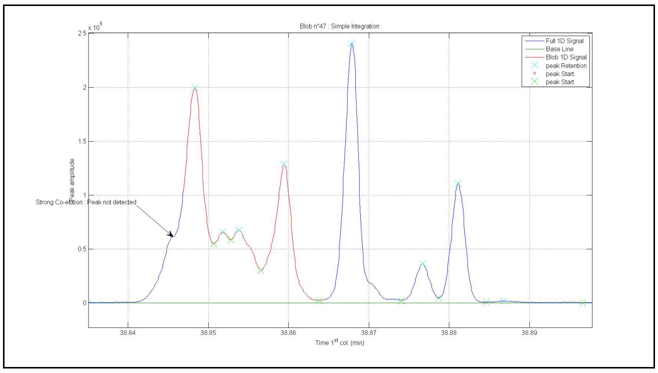


Figure 7: Strong co-elution : peaks are not detected by simple integration.

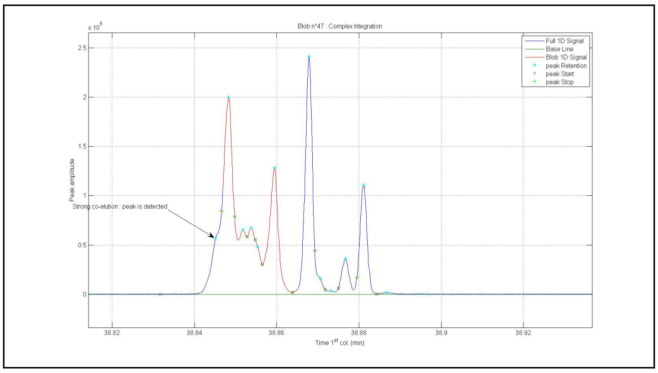


Figure 8: Strong co-elution : peaks are successfully detected by more complex integration procedure.

4. BLOB FITTING

If start time and stop times of each peak are known, the following algorithm is implemented in order to fit blob.

For each blob :
1. Determine the intersection between each column of the image and the blob; let P be this point.
2. Find the nearest peak to P;
3. If P is below the peak, move it down toward the nearest end of peak;
4. If P is above the peaks, move it up toward the nearest end of peak;

For instance, Figure 9-left displays blobs (red plot) obtained manually from well-separated peaks. Figure 9-right represents the contour plot for the same blobs obtained after automatic fitting leading to more accurate results.

The same experience is carried out within a middle distillate analysis (cf. Figure 10). This figure presents peaks obtained by the previous algorithm. Blob location appears as not accurate (e.g. frontier points do not correspond to peak starts or peak stops). Figure 11 shows new blobs location using automatic blob fitting. Obviously, better-defined blobs have been successfully obtained without user action.

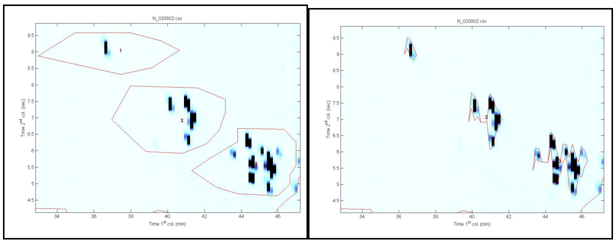


Figure 9: Blobs contour plots without (left) and with (right) automatic fitting for individual peaks.

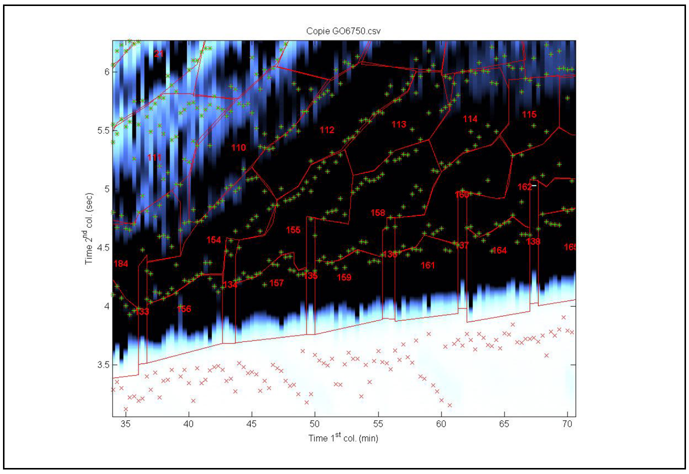


Figure 10: Blobs contour plots (manually determined) for middle distillate.

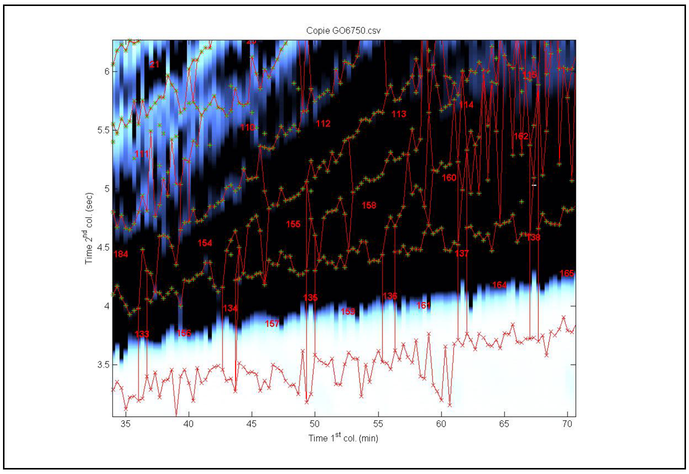


Figure 11: Blobs contour plots after automatic fitting for middle distillate.

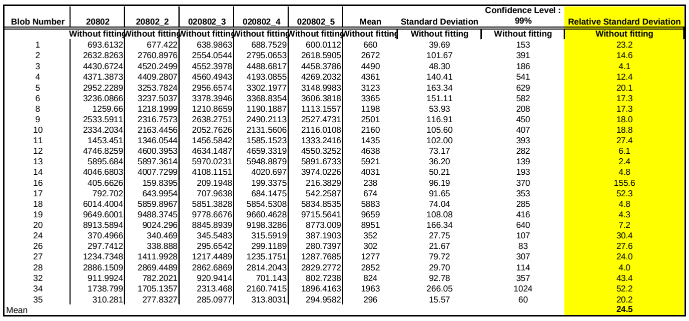
4. RESULTS

Quantitative experiments have been performed with data obtained for the analysis of nitrogen compounds in middle distillates (typical 2D-chromatogram reported in Figure 2). In order to determine the repeatability of the process, five replicate experiments have been carried out. The statistic dispersion of blob areas was measured using the Student's test with a confidence level of 99% by:

$$Err = 100 * 4.03 * \sigma / \mu \qquad (1)$$

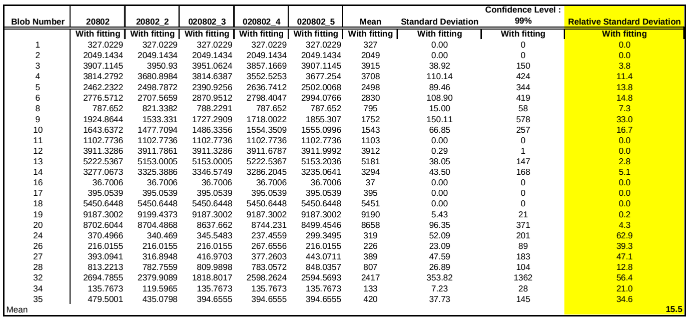
with σ denoting the standard deviation of the blob and μ its area.

Figure 12 gathers results manually obtained. Figure 13 shows results obtained after automatic fitting. Without the automated blob fitting, the statistics dispersion was measured as 25%. Thanks to the automated fitting process, it was reduced to 15%, which is a significant gap for performing routine type analysis in industrial laboratories.



Blob Number	20802	20802_2	020802_3	020802_4	020802_5	Mean	Standard Deviation	Confidence Level : 99%	Relative Standard Deviation
	Without fitting	Without fitting	Without fitting	Without fitting	Without fitting	Without fitting	Without fitting	Without fitting	Without fitting
1	693.6132	677.422	638.9863	688.7529	600.0112	660	39.69	153	23.2
2	2632.8263	2760.8976	2554.0544	2795.0653	2618.5905	2672	101.67	391	14.6
3	4430.6724	4520.2499	4552.3978	4488.6817	4458.3786	4490	48.30	186	4.1
4	4371.3873	4409.2807	4560.4943	4193.0855	4269.2032	4361	140.41	541	12.4
5	2952.2289	3253.7824	2956.6574	3302.1977	3148.9983	3123	163.34	629	20.1
6	3236.0866	3237.5037	3378.3946	3368.8354	3606.3818	3365	151.11	582	17.3
8	1259.66	1218.1999	1210.8659	1190.1887	1113.1557	1198	53.93	208	17.3
9	2533.5911	2316.7573	2638.2751	2490.2113	2527.4731	2501	116.91	450	18.0
10	2334.2034	2163.4456	2052.7626	2131.5606	2116.0108	2160	105.60	407	18.8
11	1453.451	1346.0544	1456.5842	1585.1523	1333.2416	1435	102.00	393	27.4
12	4746.8259	4600.3953	4634.1487	4659.3319	4550.3252	4638	73.17	282	6.1
13	5895.684	5897.3614	5970.0231	5948.8879	5891.6733	5921	36.20	139	2.4
14	4046.6803	4007.7299	4108.1151	4020.697	3974.0226	4031	50.21	193	4.8
16	405.6626	159.8395	209.1948	199.3375	216.3829	238	96.19	370	155.6
17	792.702	643.9954	707.9638	684.1475	542.2587	674	91.65	353	52.3
18	6014.4004	5859.8967	5851.3828	5854.5308	5834.8535	5883	74.04	285	4.8
19	9649.6001	9488.3745	9778.6676	9660.4628	9715.5641	9659	108.08	416	4.3
20	8913.5894	9024.296	8845.8939	9198.3286	8773.009	8951	166.34	640	7.2
24	370.4966	340.469	345.5483	315.5919	387.1903	352	27.75	107	30.4
26	297.7412	338.888	295.6542	299.1189	280.7397	302	21.67	83	27.6
27	1234.7348	1411.9928	1217.4489	1235.1751	1287.7685	1277	79.72	307	24.0
28	2886.1509	2869.4489	2862.6869	2814.2043	2829.2772	2852	29.70	114	4.0
32	911.9924	782.2021	920.9414	701.143	802.7238	824	92.78	357	43.4
34	1738.799	1705.1357	2313.468	2160.7415	1896.4163	1963	266.05	1024	52.2
35	310.281	277.8327	285.0977	313.8031	294.9582	296	15.57	60	20.2
Mean									**24.5**

Figure 12: Manual analysis for 5 replicates.



Blob Number	20802	20802_2	020802_3	020802_4	020802_5	Mean	Standard Deviation	Confidence Level : 99%	Relative Standard Deviation
	With fitting	With fitting	With fitting	With fitting	With fitting	With fitting	With fitting	With fitting	With fitting
1	327.0229	327.0229	327.0229	327.0229	327.0229	327	0.00	0	0.0
2	2049.1434	2049.1434	2049.1434	2049.1434	2049.1434	2049	0.00	0	0.0
3	3907.1145	3950.93	3951.0624	3857.1669	3907.1145	3915	38.92	150	3.8
4	3814.2792	3680.8984	3814.6387	3552.5253	3677.254	3708	110.14	424	11.4
5	2462.2322	2498.7872	2390.9256	2636.7412	2502.0068	2498	89.46	344	13.8
6	2776.5712	2707.5659	2870.9512	2798.4047	2994.0766	2830	108.90	419	14.8
8	787.652	821.3382	788.2291	787.652	787.652	795	15.00	58	7.3
9	1924.8644	1533.331	1727.2909	1718.0022	1855.307	1752	150.11	578	33.0
10	1643.6372	1477.7094	1486.3356	1554.3509	1555.0996	1543	66.85	257	16.7
11	1102.7736	1102.7736	1102.7736	1102.7736	1102.7736	1103	0.00	0	0.0
12	3911.3286	3911.7861	3911.3286	3911.6787	3911.9992	3912	0.29	1	0.0
13	5222.5367	5153.0005	5153.0005	5222.5367	5153.2036	5181	38.05	147	2.8
14	3277.0673	3325.3886	3346.5749	3286.2045	3235.0641	3294	43.50	168	5.1
16	36.7006	36.7006	36.7006	36.7006	36.7006	37	0.00	0	0.0
17	395.0539	395.0539	395.0539	395.0539	395.0539	395	0.00	0	0.0
18	5450.6448	5450.6448	5450.6448	5450.6448	5450.6448	5451	0.00	0	0.0
19	9187.3002	9199.4373	9187.3002	9187.3002	9187.3002	9190	5.43	21	0.2
20	8702.6044	8704.4868	8637.662	8744.231	8499.4546	8658	96.35	371	4.3
24	370.4966	340.469	345.5483	237.4559	299.3495	319	52.09	201	62.9
26	216.0155	216.0155	216.0155	267.6556	216.0155	226	23.09	89	39.3
27	393.0941	316.8948	416.9703	377.2603	443.0711	389	47.59	183	47.1
28	813.2213	782.7559	809.9898	783.0572	848.0357	807	26.89	104	12.8
32	2694.7855	2379.9089	1818.8017	2598.2624	2594.5693	2417	353.82	1362	56.4
34	135.7673	119.5965	135.7673	135.7673	135.7673	133	7.23	28	21.0
35	479.5001	435.0798	394.6555	394.6555	394.6555	420	37.73	145	34.6
Mean									**15.5**

Figure 13: Automatic analysis for 5 replicates.

5. CONCLUSION

GC×GC is an efficient technology for the analysis of complex mixture such as petroleum samples but it still suffers from its user-dependency involving time-consuming and inaccurate post-processing. To overcome this limitation, an automatic fitting procedure of blob based on a filtered derivation has been implemented. It is based on accurate determination of peak positions in signal in the second separation column. The proposed method was demonstrated to be able to improve analysis repeatability and to reduce the processing time. It is now implemented in the industrial software *Polychrom*.

Additional experiments are conducted with active contour methods in order to improve the fidelity and accurateness of image post -processing as far as possible.

6. REFERENCES

[1] J. Beens, H. Boelens, R. Tijssen, J. Blomberg, 1998, *Quantitative aspects of comprehensive two-dimensional gas chromatography (GC x GC)*, J. High Resolut. Chromatogr. Vol. 21 No 47.

[2]F. Bertoncini, C. Vendeuvre, D. Thiébaut, 2005, *Interest and Applications of Multidimensional Gas Chromatography for Trace Analysis in the Petroleum Industry*, Oil and Gas, Science and Technology, Vol. 60, No. 6, pp. 937-950

[3]Jens Dalluge, Jan Beens, Udo A.Th. Brinkman, 2003, *Comprehensive two-dimensional gas chromatography: a powerful and versatile analytical tool*, Journal of Chromatography A, Vol. 1000 pp. 69–108

[4]J.L. Excoffier, G. Guiochon, 1982, *Automatic Peak Detection in Chromatography*, Chromatographia, Vol. 15 No 9.

[5]P.J.P. Cardot, P. Trolliard, S. Tembely, 1990, A fully automated chromatographic peak detection and treatment software for multi-user multi-task computers, J. Pharm. Biomed. Anal., Vol 8.

[6] Frysinger G. S., Gaines R. B., Reddy C. M., 2002, *GC×GC A new analytical tool for environmental forensics*, Environ. Forens, Vol. 3, pp. 27–34.

[7]P. Korytar, P.E.G. Leonards, J. de Boer, U.A.Th. Brinkman, 2002, *High-resolution separation of polychlorinated biphenyls by comprehensive two-dimensional gas chromatography*, Journal of

Chromatography A, Volume 958, Issues 1-2, 7 June 2002, Pages 203-218 .

[8]K. Lan, W. Jorgenson, 2001, *A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks***,** Journal of Chromatography A, Volume 915, Issues 1-2, pp. 1-13.

[9]J. Li, 2002, *Comparison of the capability of peak functions in describing real chromatographic peaks*, Journal of Chromatography A, **Volume 952,** Issues 1-2, pp. 63-70.

[10]T.L. Pap, Zs. Papai, 2001, *Application of a new mathematical function for describing chromatographic peaks*, Journal of Chromatography A, Volume 930, Issues 1-2, Pages 53-60.

[11] Philipps J. B., Beens J., 1999, *Comprehensive two-dimensional gas chromatography: A hyphenated method with strong coupling between the two dimensions*, J. Chromatography A, Vol. 856, pp. 331–347.

[12] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, 2004, *Information technologies for comprehensive two-dimensional gas chromatography*, Chemometrics and Intelligent Laboratory Systems, vol. 71, n^{o}2, pp. 107-120

[13] S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford, in Oroc. 25th Inter. Symp. Cap. Chromatogr., Riva del Garda, Italy, 2002.

[14]Savitzky A., Golay M. J. E., *Smoothing and differentiation of data by simplified least squares procedures*, *Anal. Chem.*, vol. 36, pp. 1627–1639, 1964.

[15] A. G. Stromberg, S. V. Romanenko, E. S. Romanenko, 2000, *Systematic study of elementary models of analytical signals in the form of peaks and waves*, Journal of analytical chemistry (J. anal. chem.) , vol. 55, no7, pp. 615-625.

[16] G. Vivo-Truyols, J.R. Torres-Lapasi, A.M. van Nederkassel, Y. Vander Heyden, D.L. Massart, 2005, *Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals*, Journal of Chromatography A, 1096 (2005) 133–145

[17]Van Mispelaar V.G., Janssen H.G., Tas A.C., Schoenmakers P.J., 2005, *Novel system for classifying chromatographic applications, exemplified by comprehensive two dimensional gas chromatography and multivariate analysis*, Journal of Chromatography A., 1071 (2005) pp. 229-237.

[18]Vendeuvre C., Bertoncini F., Duval L., Duplan J.L., Thiébaut D., Hennion M.C., 2004, *Comparison of conventional gas chromatography and comprehensive two-dimensional gas chromatography for the detailed analysis of petrochemical samples*, Journal of Chromatography A, Volume 1056, Issues 1-2, pp. 155-162 .