IMPROVEMENT OF MS BASED E-NOSE PERFORMANCES BY INCORPORATION OF CHROMATOGRAPHIC RETENTION TIME AS A NEW DATA DIMENSION

Cosmin Burian

ISBN: 978-84-694-0293-1
Dipòsit Legal: T-202-2011

Improvement of MS-based e-nose performances by incorporation of chromatographic retention time as a new data dimension

PhD Thesis
Cosmin Burian

UNIVERSITAT ROVIRA I VIRGILI

Cosmin Burian

Improvement of MS-based e-nose performances by incorporation of chromatographic retention time as a new data dimension

DOCTORAL THESIS

Supervisors

Jesus Brezmes Llecha

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2010

Universitat Rovira i Virgily
Department d'Enginyeria Electronica,
Electrica i Automatica

I, **Dr. Jesus Brezmes Llecha**, Associate professor in the Department d'Enginyeria Electronica, Electrica i Automatica at Universitat Rovira i Virgily,

CERTIFY:

That the Doctoral Thesis, entitled "**Improvement of MS-based e-nose performances by incorporation of chromatographic retention time as a new data dimension**", presented by **Burian Cosmin** for the award of the degree of Doctor of Rovira i Virgili University, has been carried out under my supervision at the Department Department d'Enginyeria Electronica, Electrica i Automatica of this university, all the results presented in this thesis were obtained in experiments conducted by the above mentioned student and that it fulfils all the requirements to be eligible for the European Doctorate Label.

Tarragona, 20 May 2010

Acknowledgements

Usually it is customary to write an acknowledgements page. As one of the last things to do for a thesis, one, fed up with it and with strong wishes to finish, tends to skip it. But that would be a mistake. It takes a long time to write a PhD thesis, though not as long as it takes to lay some rail track, surprisingly, and there are many people in the background who support the doctorate student. I would like to express here my thanks to the people who have been very helpful during the time it took me to write this thesis.

First of all I would like to thank Jesus for all the support he gave me all these years, for being there when I needed an advice, for all the great ideas, and for his never ending patience. To Mariona, who had open the world of data analysis to me and spared her time to help me in this wonderful field and in the field of gas chromatography. To both of them many thanks for the constructive criticism that I always appreciated even more so than praises. To Xavier for all the meetings planning experiments and discussing the results, and for the very pleasant paella group reunions.

To Rosa for her help while measuring los oils d'oliva i d'avellana. For the nice moments and the fun we had working in Servei de Recursos Cientificos.

To all the people in the department that had helped me with any questions that I had, to Xavier and Eduard who taught the doctorate classes. To professors Motzalev and Vasiliev for all the interesting discussion, and to all the teachers in the department that I had the pleasure to talk to.

To the good friends that I made here. To Johann, first flat mate and then a very good friend for life, for all the fun we had these years, for the good and the bad jokes. To Roman, with whom even though I did not spent a lot of time, for the animating discussions. To both of them for their way of thinking.

To the Romanian friends that I made here, Oana and Bogdan, for their friendship, the pleasant times we spent together watching a movie, sharing recipes, or cooking.

To Diana and Anca, my last year flat mates, for the fun we had playing sheptica and rentz, remembering the nice times in our lives, and poking fun at each other.

To Alfonso, my Spanish friend, for his love of life, true friendship, and for all our exciting battles we had in the black and white jungle.

To the friends in 334 and the pilot plant, to Raul for his help with Matlab when I began programming, to Lukas, Stela, Edwin, Roger, Julia and many more. Thanks to all of them for listening to me and the time spent together.

To Luminita, who has been by my side this last year, for all that she had given me, for her never ending support (with both meanings in English and Romanian languages). For her delicious food and closing an eye when I was using the thesis as an excuse not to cook or wash the dishes when it was my turn.

To my parents and sister for being close to me even though far away, for their support and understanding in my choices.

And lastly to the higher power: Thank God I finally finished.

Table of Contents:

Summary

The importance of the sense of smell in nature and human society can be seen in the devoted interest shown to the analysis of odor and flavor in food industry. Although food and beverages are one of the biggest application areas for odor sensing, other fields have shown the need for this technology. Unfortunately using human sensory test panels or dogs are costly, prone to fatigue, subjective, unreliable and not good at quantifying, while laboratory chemical analysis despite accurate, unbiased, and quantitative, is labor intensive, requires trained specialists, and is time consuming. Because of these drawbacks artificial olfaction came into attention.

The term "electronic nose" is associated with an array of chemical gas sensors with a broad and overlapping selectivity for measurements of volatile compounds combined with computerized data analysis tools. The electronic nose is used to provide comparative rather than qualitative information in an analysis, and because the interpretation can be automated, the device is suited for quality control and analysis. Despite some promising achievements, solid-state gas sensors have not lived up to their expectations. Low sensitivity and selectivity, sensor short life span, difficult calibration and drift problems have proven serious limitations. In an effort to improve the solid-state sensors drawbacks new approaches have been taken, using different sensors for the electronic nose. Optical sensor systems, ion mobility spectrometry and infrared spectrometry approaches are examples of techniques that have been tested.

Mass spectrometry (MS) based electronic noses were first reported in 1998 [B. Dittmann, S. Nitz and G. Horner. Adv. Food Sci. 20 (1998), p. 115], and they represented a major leap in sensibility, challenging the chemical sensor-based electronic nose. This new approach to the concept of an electronic nose uses virtual sensors in the form of m/z ratios. A complex and very reproducible fingerprint is obtained in the form of a mass spectra, which is processed by pattern recognition algorithms towards classification and even quantification. Even though the MS-based e-nose outperforms the classic solid-state sensors-based electronic nose, its use is currently limited to laboratory desktop

instrumentation. The lack of portability may not be for long an issue, as miniature mass spectrometer have been produced at a proof-of-concept stage.

A more critic drawback of the MS electronic nose consists in the way samples are analyzed. Simultaneous fragmentation of complex isomer mixtures can produce very similar results following this approach. A better electronic nose would be one that combines the sensibility and power of identification of the mass detector with the separation ability of the gas chromatography. The main drawback to this approach is again the cost and lack of portability of the equipment. In addition to the problems already encountered in mass spectrometry, gas chromatography analysis take are time consuming.

To address these issues, miniaturizations in capillary gas chromatography (GC) which make possible a GC-on-a-chip, fast-GC and flash-GC that make use of short columns, reducing the time of analysis to elution times as short as seconds have been reported and, in some cases, have been commercialized. The miniaturization of mass spectrometry and gas chromatography has a great potential to improve the performance, usefulness and affordability of the new generation of electronic noses.

This thesis is aimed to the study and evaluation of the GC-MS approach for the electronic nose as a previous step to the development of the above-mentioned technologies. Thus the main objective of the thesis is to study if the retention time of a chromatographic separation can improve the mass sensor-based electronic nose performance by showing that the addition of the third dimension is useful, bringing extra information, helping in the classification of samples. This can be done in two ways:

- comparing two-way data analysis of mass spectra data with two-way data analysis of unfolded or concatenated matrices for the three-way data.
- comparing two-way data analysis of mass spectra with three-way data analysis of the whole tri-dimensional dataset.

From the chromatographic point of view, the goal will be to optimize the chromatographic method in order to shorten the analysis to a minimum while still having acceptable results.

An important step in multi-way multivariate data analysis is data preprocessing. Because of this the last goal will be to determine which preprocessing techniques are best for two and three-way data analysis.

In order to reach the proposed goals two datasets were created. The first one consisted of mixtures of 9 isomers of dimethylphenol and ethylphenol. The reason for this choice was the similarity of the mass spectra between each other. In this way the mass spectra-based electronic nose would be challenged by the dataset. Also by measuring the retention time of the nine isomers alone, the solutions were made as such as the dataset would prove a challenge if just the retention time would be used. Therefore this "artificial" dataset holds our hopes in showing the improvements of using both dimensions, the MS (mass spectra) and GC (retention time).

Twenty classes, representing solutions of the nine isomers were measured in ten repetition each, for three chromatographic methods, giving a total of 600 measurements. The chromatographic methods were designed to give a fully resolved chromatogram, a coeluted peak and an intermediary situation with a partially resolved chromatogram. The data was recorded in a three dimensional matrix with the following directions: (samples measured) x (m/z ratio) x (retention time). By "collapsing" the x and y axis of the chromatographic retention time and m/z fragments respectively we obtained two matrices representing the average mass spectra and the total ion chromatogram (TIC) respectively. These approaches loose the information brought by the third dimension and so unfolding the 3D original matrix and concatenation of TIC and average mass spectra were taken into consideration as a way to preserve the extra information of the third dimension in a two dimensional matrix.

The data was pretreated by peak alignment, mean centering and normalization by peak height and peak area; preprocessing tools that were also evaluated for their performance.

For analyzing the two-way data PCA, PLS-DA and fuzzyARTMAP were used. The PCA and PARAFAC clustering were evaluated by the intervariance/intravariance ratio, while the fuzzyARTMAP gave classification success rates percentages.

When PCA and PARAFAC were used, as expected, the resolved chromatographic method (method 1) gave the best results overall, where 2D algorithms work better, whereas in a more complicated case (more coeluted peaks of method 3) they lose effectiveness in front of 3D methods.

In the case of PLS-DA and n-PLS, even though the results are not as conclusive as the PCA and PARAFAC results, the differences being minimal, the multi-way PLS-DA model yields a high prediction success rate in both datasets. It is also recommended over the unfolded and concatenated data because it builds a more parsimonious model.

For the fuzzyARTMAP analysis, the voting strategy used proved that when using the average mass spectra and TIC information together the most consistent results are obtained.

The second dataset addresses the issue of extra virgin olive oil adulteration with hazelnut oil, which because of the similarities between the two oils is one of the most difficult to detect. Four extra virgin olive oils and two hazelnut oils were measured pure and as mixtures of 30%, 10%, 5% and 2% with the same goals of proving that the addition of the extra dimension improves results. Five repetitions were made for each preparation, giving a total of 190 samples: 4 pure olive oils, 2 pure hazelnut oils, 32 olive oil - hazelnut oil adulterations, giving a total of 38 classes. Two chromatographic methods were used. The first one was aimed at a complete separation of olive oil constituents and employed a temperature-programmed separation, while the second method's objective was a coeluted peak, thus a constant temperature separation was employed. The data was analyzed by means of PCA, PARAFAC, PLS-DA and n-PLS.

As in the "artificial" dataset, the PCA and the PARAFAC were analyzed by means of the clusterization capability, which showed that the best results are obtained by the unfolded data followed by PARAFAC.

From the column optimization point of view, the performance of the short column is below the long column approach, but this case proves once more that the addition of the third dimension increases the performance of the MS-based electronic nose.

For the PLS-DA and n-PLS the success rates were compared for both the long and the short chromatographic runs. While for the long column the best performance is for the TIC data, the short column shows better performance for the concatenated data of average mass spectra and TIC. Moreover the prediction success rates are the same for the long column TIC data as for the short column concatenated data. This case is very interesting because PLS approach proves that the third dimension improves the results and, moreover, by using the short column the analysis time is considerably shortened.

Certain performances are expected from an electronic nose. For the time being neither approach got close enough to them in order to produce a hit for the markets. Solid state sensors have drawbacks almost impossible to overcome. The mass spectra-based electronic nose lacks portability and sometimes its performances have shortcomings, and the gas chromatograph-mass spectrometer apparatus suffers from the same portability problem as MS and is time consuming. The development of powerful mathematical algorithms over the last few years along with advances in miniaturization for both MS and GC and fast chromatography show some hope for a much better electronic nose.

Through this work we can confirm that the addition of the chromatographic retention time as an extra dimension brings an edge over existing electronic nose technologies. While for the fully resolved chromatograms there is no performance gained, or the gain is minimal, for a short column the extra information improves the results, in some cases making them as good as when a long column is used. This is very important because the measurements on a gas chromatograph - mass spectrometer can be optimized for very short runs, a very important characteristic for an electronic nose. This would allow the design of a higher throughput instrument suitable, for example, for quality control in product lines.

Resumen

La importancia del sentido de olor en la naturaleza y en la sociedad humana queda latente con el gran interés que se muestra en el análisis del olor y el gusto en la industria alimentaria. Aunque las aéreas más interesadas son las de la alimentación y bebida, también se ha mostrado la necesitad para esta tecnología en otros campos como en el de la cosmética. Lamentablemente, el uso de los paneles sensoriales humanos o paneles caninos son costosos, propensos al cansancio, subjetivos, poco fiables e inadecuados para cuantificar, mientras que el análisis de laboratorio, a pesar de la precisión, imparcialidad y capacidad cuantitativa, necesita una labor intensa, con personal especializado y requiere de mucho tiempo. Debido a estos inconvenientes el concepto de olfato artificial generó un gran interés en entornos industriales.

El término "nariz electrónica" se asocia con una serie de sensores de gases químicos, con una amplia superposición de selectividad para las mediciones de compuestos volátiles en combinación con los instrumentos informáticos de análisis de datos. La nariz electrónica se utiliza para proporcionar una información comparativa en vez de una cualitativa en un análisis, y porque la interpretación puede ser automatizada, el dispositivo es adecuado para el control de calidad y análisis. A pesar de algunos logros prometedores, los sensores de estado sólido de gas no han cumplido con sus expectativas. La baja sensibilidad y selectividad, la corta vida del sensor, la calibración difícil y los problemas de deriva han demostrado serias limitaciones. En un esfuerzo para mejorar los inconvenientes de los sensores de estado sólido, se han adoptado nuevos enfoques, utilizando diferentes sensores para la nariz electrónica. Sistemas de sensores ópticos, la espectrometría de movilidad iónica y la espectrometría infrarroja son ejemplos de técnicas que han sido probadas.

Las narices electrónicas basadas en la espectrometría de masas (MS) aparecieron por primera vez en 1998 [B. Dittmann, S. y G. Nitz Horner. Adv. Food Sci. 20 (1998), p. 115], y representan un salto importante en la sensibilidad, retando a la nariz electrónica basada en sensores químicos. Este nuevo enfoque del concepto de una nariz electrónica usa sensores virtuales en forma de

proporciones m/z. Una huella digital compleja y muy reproducible se obtiene en forma de un espectro de masas, que se procesa mediante algoritmos de reconocimiento de patrones para la clasificación y cuantificación. A pesar de que la nariz electrónica basada en la espectrometría de masas supera a la nariz electrónica clásica de sensores de estado sólido en muchos aspectos, su uso se limita actualmente a la instrumentación de laboratorio de escritorio. La falta de portabilidad no representará necesariamente un problema en el futuro, dado que espectrómetros de masas en miniatura se han fabricado ya en una fase de prototipito.

Un inconveniente más crítico de la nariz electrónica basada en MS consiste en la manera en la que se analizan las muestras. La fragmentación simultánea de mezclas complejas de isómeros pueden producir resultados muy similares a raíz de este enfoque. Una nariz electrónica mejor sería la que combina la sensibilidad y el poder de identificación del detector de masas con la capacidad de separación de la cromatografía de gases. El principal inconveniente de este enfoque es de nuevo el coste y la falta de portabilidad de los equipos. Además de los problemas anteriores con la espectrometría de masas, el análisis de cromatografía de gases requiere mucho tiempo de medida.

Para abordar estas cuestiones, se han reportado miniaturizaciones en cromatografía capilar de gases (GC) que hacen posible el GC-en-un-chip, CG-rápido y CG-flash que hacen uso de columnas cortas, reduciendo el tiempo de análisis a los tiempos de elución como segundos y, en algunos casos, se han comercializado. La miniaturización de la espectrometría de masas y cromatografía de gases tiene un gran potencial para mejorar el rendimiento, la utilidad y la accesibilidad de la nueva generación de narices electrónicas.

Esta tesis se dedica al estudio y a la evaluación del enfoque del GC-MS para la nariz electrónica como un paso anterior al desarrollo de las tecnologías mencionadas anteriormente. El objetivo principal de la tesis es de estudiar si el tiempo de retención de una separación de cromatografía puede mejorar el rendimiento de la nariz electrónica basada en MS, mostrando que la adición de una tercera dimensión trae más información, ayudando a la clasificación de las pruebas. Esto se puede hacer de dos maneras:

- comparando el análisis de datos de dos vías de espectrometría de masas con análisis de datos de dos vías de matrices desplegadas y concatenadas para los datos de tres vías y

- comparando el análisis de datos de dos vías del espectrometría de masas con el análisis de datos de tres vías para el conjunto de datos tridimensionales.

Desde el punto de vista de cromatografía, la meta será la de optimizar el método cromatográfico con el fin de reducir el tiempo de análisis a un mínimo sin dejar de tener resultados aceptables.

Un paso importante en el análisis de datos multivariados de vías múltiples es el preprocesamiento de datos. Debido a este objetivo, el último objetivo será el de determinar qué técnicas de preprocesamiento son las mejores para y el análisis de dos y tres vías de datos.

Con el fin de alcanzar los objetivos propuestos se crearon dos grupos de datos. El primero consiste en las mezclas de nueve isómeros de dimetilfenol y etilfenol. La razón de esta elección fue la similitud de los espectros de masas entre sí. De esta manera la nariz electrónica basada en espectrometría de masas sería retada por el conjunto de datos. También teniendo en cuenta el tiempo de retención de los nueve isómeros solos, las soluciones se hicieron, como si el conjunto de datos demostraría el reto si se usaría sólo el tiempo de retención. Por tanto, este conjunto de datos "artificiales" sostiene nuestras esperanzas en mostrar las mejoras de la utilización de ambas dimensiones, la MS (espectros de masas) y la GC (tiempo de retención).

Veinte clases, representando las soluciones de los nueve isómeros se midieron en diez repeticiones cada una, por tres métodos cromatográficos, dando un total de 600 mediciones. Los métodos cromatográficos fueron diseñados para dar un cromatograma resuelto por completo, un pico coeluido y una situación intermediaria con un cromatograma resuelto parcialmente. Los datos fueron registrados en una matriz de tres dimensiones con las siguientes direcciones: (muestras medidas) x (proporción m/z) x (tiempo de retención). Por "colapsar" los ejes X e Y del tiempo de retención cromatográfica y los fragmentos m/z, respectivamente, se obtuvieron dos matrices que representan los espectros de masa regular y el cromatograma de iones totales, respectivamente. Estos enfoques sueltan la información traída por la tercera dimensión y el despliegue por lo que la matriz original 3D y la concatenación de las TIC y el espectro de masa media se han tenido en consideración como una forma de preservar la información adicional de la tercera dimensión en una matriz de dos dimensiones.

Los datos fueron tratados mediante la alineación de picos, con una media de centrado y la normalización por la altura máxima y el área del pico, los instrumentos de pre-procesamiento que también fueron evaluados por sus logros.

Para el análisis de datos de dos vías fueron utilizados el PCA, PLS-DA y fuzzyARTMAP. La agrupación de PCA y PARAFAC fueron evaluados por la relación intervariedad - intravariedad, mientras que los resultados mediante fuzzyARTMAP fueron dados como el éxito de la las tasas de clasificación en porcentajes.

Cuando PCA y PARAFAC se utilizaron, como era de esperar, el método de cromatografía resuelto (método 1) dio los mejores resultados globales, donde los algoritmos 2D funcionan mejor, mientras que en un caso más complicado (picos más coeluidos del método 3) pierden eficacia frente a métodos 3D.

En el caso de PLS-DA y n-PLS, aunque los resultados no son tan concluyentes como los resultados del PCA y PARAFAC, tratándose de las diferencias mínimas, el modelo de vías múltiples PLS-DA ofrece un porcentaje de éxito en la predicción de ambos conjuntos de datos. También se recomienda el n-PLS en vez de utilizar datos desplegados y concatenados, ya que construye un modelo más parsimonioso.

Para el análisis fuzzyARTMAP, la estrategia de votación empleada ha demostrado que al usar los espectros de masa media y la información del cromatograma de iones totales juntos se obtienen resultados más consistentes.

En el segundo conjunto de datos se aborda el problema de la adulteración del aceite de oliva extra virgen con aceite de avellana, que debido a las similitudes entre los dos aceites es una de las más difíciles de detectar. Cuatro aceites extra virgen de oliva y dos aceites de avellana se midieron puros y en mezclas de 30%, 10%, 5% y 2% con los mismos objetivos mostrando que la adición de la extra dimensión mejora los resultados. Se han hechos cinco repeticiones para cada preparación, dando un total de 190 muestras: 4 aceites puros de oliva, 2 aceites puros de avellana y 32 adulteraciones de aceite de avellana en aceite de oliva, dando un total de 38 clases. Dos métodos cromatográficos fueron utilizados. El primero estaba dirigido a una completa separación de los componentes del aceite de oliva y empleó una separación con temperatura programable, mientras que el objetivo del segundo método fue un pico coeluido, por lo tanto fue contratada una temperatura constante de separación. Los datos fueron analizados por medio de la PCA, PARAFAC, PLS-DA y PLS-n.

Como en el conjunto "artificial" de datos, el PCA y PARAFAC se analizaron por medio de la capacidad de clusterización, que mostró que los mejores resultados se obtienen con los datos desplegados seguido por los datos 3D tratados con el PARAFAC.

Desde el punto de vista de optimización de la columna, los logros obtenidos por la columna corta está por debajo del enfoque de la columna larga, pero este caso demuestra una vez más que la adición de los incrementos de tercera dimensión mejoran la nariz electrónica basada en MS.

Para el PLS-DA y n-PLS se evaluaron las tasas de éxito comparativamente, tanto para las corridas cromatográficas largas como para las cortas. Mientras que para la columna larga el mejor rendimiento es para los datos del cromatograma de iones totales (TIC), la columna corta muestra mejor rendimiento para los datos concatenados de los espectros de masa media y TIC. Además, la predicción de las tasas de éxito son las mismas para los datos TIC de columna larga como para los datos concatenados de la columna corta. Este caso es muy interesante porque demuestra que el enfoque PLS de la tercera dimensión mejora los resultados y, por otra parte, mediante el uso de la columna corta el tiempo de análisis se acorta considerablemente.

Se esperan ciertos logros de la nariz electrónica. Por el momento, ninguno de esos enfoques se acercó lo suficiente para producir una respuesta positiva en los mercados. Los sensores de estado sólido tienen inconvenientes casi imposibles de superar. La nariz electrónica basada en espectrometría de masas tiene una falta de portabilidad y a veces sus logros son insuficientes, y el aparato del cromatógrafo de gases-espectrómetro de masas sufre problemas de portabilidad igual que espectrómetro de masas y toma mucho tiempo. El desarrollo de potentes algoritmos matemáticos durante los últimos años, junto con los avances en la miniaturización, tanto para MS y GC y mostrar cromatografía rápida cierta esperanza de una nariz electrónica mucho mejor.

A través de este trabajo podemos afirmar que la adición del tiempo de retención cromatográfica como una dimensión extra aporta una ventaja sobre las actuales tecnologías de la nariz electrónica. Mientras que para los cromatogramas totalmente resueltos no se logran mejoras o la ganancia es mínima, sobre todo en la predicción, para una columna corta la información adicional mejora los resultados, en algunos casos, hacerlos tan bien como cuando una larga columna se utiliza. Esto es muy importante ya que las mediciones en un cromatógrafo de gases - espectrometro de masas se pueden optimizar para tramos muy cortos, una característica muy importante para una nariz electrónica. Esto permitiría el diseño de un instrumento de mayor rendimiento, adecuado para el control de calidad en líneas de productos.

Chapter 1

Introduction

1. Introduction

1.1. Preface

The importance of the sense of smell in nature and human society can be seen in the devoted interest shown to the analysis of odor and flavor in food industry, where traditionally trained sensory panel and chemical analysis had been used to characterize production form raw materials to final products. Although food and beverages are one of the biggest application areas for odor sensing, other fields have shown the need for its analysis. Because of their higher sensibility nose, dogs have been used in modern times for detecting drugs and explosives in trace elements in airports and customs.

Unfortunately these methods of using human sensory test panels, dogs or laboratory chemical analysis have a series of drawbacks. Human test panels and/or dogs are costly, prone to fatigue, subjective, unreliable and not good at quantifying. Chemical analysis, on the other hand, while is accurate, unbiased, and quantitative, is labor intensive, requires trained specialists, and is time consuming. Because of these drawbacks another method had to be found, and it came in the form of artificial olfaction.

In the almost thirty years that passed since the first reported design of the "electronic nose", in 1982, by Persaud and Dodd, we have seen rapid development of the concept with applications in many fields from public safety, environmental, medicine to food industry. The term "electronic nose" is associated with an array of chemical gas sensors with a broad and overlapping selectivity for measurements of volatile compounds combined with computerized data analysis tools.

The electronic nose is used to provide comparative rather than qualitative information in an analysis, and because the interpretation can be automated, the device is suited for quality control and analysis.

Despite some promising achievements, solid-state gas sensors have not lived up to their expectations. They solved a number of drawbacks related to sensory test panels and laboratory analysis, but their relatively low sensitivity and selectivity, sensor short life-span, difficult calibration and drift problems have proven serious limitations.

In an effort to improve the solid-state sensors drawbacks new approaches have been taken, using different sensors for the electronic nose. Optical sensor systems, ion mobility spectrometry, infrared spectrometry approaches are examples of techniques that have been tried.

Mass spectrometry based electronic noses were first reported in 1998, and they represented a major leap in sensibility, challenging the chemical sensor-based electronic nose. This new approach to the concept of an electronic nose uses virtual sensors in the form of m/z ratios. Volatiles are introduced in the fragmentation chamber of the mass spectrometer resulting in simultaneous ionization and fragmentation. Thus a complex and very reproducible fingerprint is obtained in the form of a mass spectra. The mass spectra output data is then processed by pattern recognition algorithms towards classification, recognition and even quantification.

Even though the MS-based electronic nose outperforms the classic solid-state sensors-based electronic nose, its use is currently limited to laboratory desktop instrumentation. The lack of portability may not be for long an issue though, as miniature mass spectrometer have been produce at a proof-of-concept stage. With miniaturization the cost is expected to drop as well.

Another, more critic, drawback of the MS electronic nose consists in the way samples are analyzed. Simultaneous fragmentation of complex isomer mixtures can produce very similar results following this approach. A better electronic nose would be one that combines the sensibility and power of identification of the mass detector with the separation ability of the gas chromatography. The main drawback to this approach is again the cost and lack of portability of the equipment. In addition to the problems already encountered in mass spectrometry, in GC analysis we have the drawback of long analysis times, which can take as long as 1.5 hours.

To address these issues, miniaturizations in capillary gas chromatography (GC) which make possible a GC-on-a-chip, fast-GC and flash-GC that make use of short columns, reducing the time of analysis to elution times as short as seconds have been reported and, in some cases, have been commercialized.

The miniaturization of mass spectrometry and gas chromatography has a great potential to improve the performance, usefulness and affordability of the new generation of electronic noses. This thesis is aimed to the study and evaluation of the GC-MS approach for the electronic nose as a previous step to the development of the above mentioned technologies.

1.2. Objectives

The main objective of the thesis is to study if the retention time of a chromatographic separation can improve the mass sensor-based electronic nose performance. In order to see if this is the case several goals have to be achieved:

1. to show that the addition of the third dimension is useful, bringing extra information, helping in the classification of samples. This can be done in two ways:
   a. comparing two-way data analysis of mass spectra data with two-way data analysis of unfolded or concatenated matrices for the three-way data.
   b. comparing two-way data analysis of mass spectra with three-way data analysis of the whole tri-dimensional dataset.
2. to optimize the chromatographic method in order to shorten the time-consuming chromatographic analysis to a minimum, while still having acceptable results
3. to determine the best preprocessing technique for two and three-way analysis

In order to reach the proposed goals two datasets were created. The first one consisted of mixtures of 9 isomers of dimethylphenol and ethylphenol. The reason for this choice was the similarity of the mass spectra between each other. In this way the mass spectra-based electronic nose would be challenged by the dataset. Also, by measuring the retention time of the nine isomers alone, the solutions were made as such as the dataset would prove a challenge if just the retention time would be used. Therefore this artificial dataset holds our hopes in showing the improvements of using both dimensions, the MS and GC.

The second dataset addresses the issue of extra virgin olive oil adulteration with hazelnut oil. Four extra virgin olive oils and two hazelnut oils were measured pure and as mixtures with the same goals of proving that the addition of the extra dimension improves results. Also the application of these approaches to the more complex real world data would prove another challenge for tow-way data analysis approaches.

1.3. Organization of the document

The thesis at hand is made up of 5 chapters described as follows:

The first chapter presents a short description of the work done in the thesis as a continuation of the research done in the field of electronic noses in the research group in the department. In other words, the section tries to locate this work in the global effort made by the research group towards the design of artificial olfaction systems. To end up this section, the objectives of the thesis are enumerated and explained.

Chapter 2 presents the state of the art in electronic nose technology and data analysis for electronic noses. The first half describes the importance of smell and monitoring certain volatiles in human life, and then continues with a short exposition of various chemical sensor-based electronic noses detection technologies. Data treatment for these approaches is shortly reviewed. Continuing with the chemical sensor-based electronic nose, some past and present commercially available e-noses are mentioned. As an improvement to the chemical sensor-based e-nose the mass spectrometry-based electronic nose (MS e-nose) is presented with its improvements in performance. Applications and commercially available MS e-noses are mentioned. The chapter continues then pointing out the drawbacks of MS e-noses and the proposed solution of incorporating the retention time dimension of a chromatographic separation. The second half of chapter 2 addresses data analysis for MS-based electronic noses as well as three-way data analysis used for the analysis of data from the gas chromatograph-mass spectrometer as three-dimensional data.

Chapter 3 presents the experimental work which consist of an "artificial" dataset of mixtures of nine isomers and the olive oil dataset; each experiment is described, and results are presented and commented.

Chapter 4 presents a discussion of the results obtained in the two experiments, while chapter 5 draws the conclusions from the work done in these 2 experiments.

Chapter 2

State of the art

2. State of the art

This chapter is going to take a look at the history of artificial olfaction research and how it is related to this thesis. We are going to start by looking at the fundamental role of the sense of smell in nature and human society, and how it led to the implementation of human sensory panels at first and the development and evolution of artificial olfaction devices. The classical electronic nose, consisting of an array of chemical sensors, will be reviewed highlighting the conception, construction, commercially available devices and its applications in various markets and industries. The new approaches to the electronic smell will be shortly presented, mentioning the different types of sensors used, and how the chemical sensor electronic nose was not only enhanced but even complemented by them. As a precursor for this research the mass spectrometry electronic nose will be looked in detail and see how it differs from the classical electronic nose and what are the main advantages that brings to the field.

After reviewing the MS-based e-nose state of the art and the flash gas chromatography, we'll try to look a little further and investigate the potential of an electronic nose based on a gas chromatograph mass spectrometer(GC-MS), its particular three-dimensional data output and the software to handle it.

As the third element in any electronic nose, the data analysis system will not be left out. The two types of data from the classic electronic nose and from the GC-MS, two dimensional data and three-dimensional data respectively, will be presented, briefly describing the two-way data processing, and making more emphasis on the three-way data analysis. The importance of data pretreatment and the usual pre-processing will be discussed.

2.1. Electronic nose

2.1.1. The role of aroma in society

The sense of smell has long played a fundamental role in human development and biosocial interactions. Consequently, the olfactory sense has become a key element in the development of many commercial industries that manipulate the aroma properties of their manufactured goods in order to improve product appeal, quality, and consistency so that consumers quickly identify unique scents with individual brands. A wide diversity of examples ranging from wines and cuisine, perfumes and colognes added to personal health-care products, and scents applied to product packaging are demonstrating the importance of aroma qualities. Similarly, spices, once among the most valued commodities for trade in ancient times, have been used throughout human history to enhance the flavor of foods and air.

Thus, aroma characteristics have contributed immensely to the value of many commercial products. As a result, research and quality control of aroma characteristics during product manufacturing has become of utmost importance in industrial production operations because product consistency is essential for maintaining consumer brand recognition and satisfaction.

Despite the importance of the olfactory sense to mankind, the sense of smell in man is often considered the least refined of the human senses, far less sensitive than that of other animals. For example, the human nose possesses only about one million aroma receptors that work in tandem to process olfactory stimuli whereas dogs have about 100 million receptors that distinguish scents at least 100 times more effectively than the average human [1].

Furthermore, the ability to detect chemicals in the environment is critical to the survival of most prokaryotic and eukaryotic organisms. A clear indication of the importance of olfactory systems in higher eukaryotes is the significant proportion (up to 4%) of the genome that is devoted to encoding products used in building olfactory sensory tissues [2].

The relatively low sensitivity and discrimination capabilities of the human nose, coupled with the common occurrence of olfactory fatigue, has led to the need for electronic instruments with sensors capable of performing repeated discriminations with high precision to eliminate human fatigue.

2.1.2. Development of the electronic nose: A historical view

The first studies involving aroma measurements were done in the 1920s by Zwaardemaker and Hogewind [3] who focused on measuring the electricity of a fine spray of water. They found that the addition of volatile substances to the water increased the spray-electricity that could be used to detect “the presence of small amounts of aromatic compounds by means other than through the sense of smell.” The first real tool for measuring aromas was developed by Hartman [4] in 1954. The sensing element was a microelectrode, a simple platinum wire of 0.8 mm in diameter, which measured the flow of current by a sensitive millivoltmeter. Hartman also was the first to propose the idea that the apparatus could operate with several different coated sensitive elements, and that some different electrode-coating substances could be capable of giving differential responses to different compounds [5].

Moncrieff [6] worked on the concept that different coatings materials, such as polyvinyl chloride, gelatin, and vegetable fats could be capable of providing different and complementary data for the discrimination of simple and complex aromas. His studies were limited to the use of a single temperature-sensitive resistor, but postulated that an array with six thermistors, provided with six different coatings, could discriminate large numbers of different aromas. In 1965, two other groups published studies and experiments on olfaction devices: Buck et al. [7] studied the modulation of conductivity as an answer to differentiating aromas bouquets, while Dravnieks and Trotter [8] used the modulation of contact potential to monitor aromas. These studies have been considered only a first approach to aromas evaluation because of the lack of analytical instruments. However, about 20 years later (1982), the idea of an electronic-nose instrument with an intelligent, chemical array sensor system for aroma classification resulted from studies of Persaud and Dodd [9] and Ikegami and Kaneyasu [10]. By that time, the development of computers and electronic sensors made it conceptually possible to obtain an electronic device capable of imitating the mammalian olfactory system.

The term “electronic nose” was coined in 1988 by Gardner and Bartlett, who later defined it as “an instrument which comprises an array of electronic chemical sensors with partial specificity and an appropriate pattern recognition system, capable of recognizing simple or complex odors” [11]. In 1991, scientific interest in the use and applications of electronic noses was sanctioned by the first advanced workshop on chemosensory information processing during a session of the North Atlantic Treaty Organization (NATO) that was entirely dedicated to the topic of artificial olfaction. Since 1991, interest in biological sensors technology has grown dramatically as is evident by numerous scientific articles on the subject and commercial efforts to develop and improve sensor technologies and tools of greater sophistication and improved capabilities, diverse sensitivities and with ever-expanding applications.

2.1.3. Short description of the electronic nose concept and modules

An electronic nose is a machine that is designed to detect and discriminate among complex odors using a sensor array. The sensor array consists of broadly tuned (non-specific) sensors that are treated with a variety of odor-sensitive biological or chemical materials. An odor stimulus generates a characteristic fingerprint (or smellprint) from the sensor array. Patterns or fingerprints from known odors are used to construct a database and train a pattern recognition system so that unknown odors can subsequently be classified and identified. This is the classical concept of an e-nose; however, in recent years, the classical sensor types used for e-noses have been enhanced and complemented by other technologies introduced in this field. Nevertheless, in a broader sense, electronic nose instruments are composed of three elements: *a sample handling system*, *a detection system*, and *a data processing system*.

2.1.4. Electronic nose sampling systems

Sample handling is a critical step affecting the analysis by e-noses whose importance is very often ignored. The quality of the analysis can be greatly improved by adopting an appropriate sampling technique. To transport the volatile compounds present in the headspace (HS) of the sample into the e-nose detection system, several sampling techniques have been used:

The Static headspace (SHS) technique consists of placing the sample in a hermetically sealed vial and then, once equilibrium has been established between the matrix and the gaseous phase, sampling the HS is executed. Sample temperature, equilibration time, vial size and sample quantity are the main parameters that have to be optimized. Because of the poor repeatability of manual HS injection, the use of an automatic HS sampler is highly recommended. In some applications a vapor-flow system has been used, and has provided better control than manual headspace injection of the operating temperature and the amount of analyte that is introduced into the detector.

Purge and trap (P&T) and *dynamic headspace* (DHS) techniques have been used in some applications to increase sensitivity, since they provide a volatile compounds pre-concentration step. In these systems, the volatile components are purged by a stream of inert gas and trapped onto an adsorbent. In the case of P&T, the gas flow is injected through the sample, whereas, in the case of DHS, only the HS is purged with the gas. The constant depletion of the HS leads to a displacement of the equilibrium in favor of the desorption of these molecules from the matrix. The trapped molecules are subsequently desorbed by heating and introduced into the detection system. Apart from the choice of trap, the main parameters to optimize are the temperature of the sample, the equilibration time, the flow rate of the extractor gas and the purge time of the HS.

Solid-phase microextraction (SPME) is a user-friendly preconcentration method. The principle involves exposing a silica fiber covered with a thin layer of adsorbent in the HS of the sample in order to trap the volatile components onto the fiber. The adsorbed compounds are then desorbed by heating and introduced into the detection system. Apart from the nature of the adsorbent deposited on the fiber, the main parameters to optimize are the equilibration time, the sample temperature and the duration of the extraction. This technique has a considerable concentration capacity and it is very simple because, unlike P&T or DHS, it does not require especial equipment.

Stir bar sorptive extraction (SBSE) is based on a magnetic bar coated with polymers, which can be held in the HS for sampling. Its loading capacity is much higher than that of SPME. Even though it has been developed only recently, SBSE is a promising extraction technique when very high sensitivity is required [12].

Inside-needle dynamic extraction (INDEX) is also a preconcentration technique. The INDEX needles contain an absorbing polymer phase like a fixed bed. The volatile compounds are forced through the needle by repeated aspiration/ejection motions of the syringe plunger. The potential advantage of this system compared to SPME is its mechanical robustness and the possibility of increasing the amount of absorbing polymer as well as the surface area available for adsorbing volatile compounds.

Although, any sampling headspace technique can be used as the sample-handling part of an e-nose, the choice must be made with care and take into account the type of sample and the specifications required by each method. SHS is the most common technique because it is very simple to use. However, for some applications, the SHS technique has the drawback of low sensitivity because the volatile compounds are not pre-concentrated. On the other hand, pre-concentration systems improve the sensitivity making detection easier and extracting semi-volatiles which otherwise would not be detected. However, they introduce a supplementary step in the method, which increases the time of analysis. Moreover, analytical artifacts (memory effects, bleeding or irreversible adsorption) are generated in some cases. In this respect, the pre-concentration media must be carefully chosen.

2.1.5. Electronic nose detection technologies

The sensor array in an electronic nose performs very similar functions to the olfactory receptors in the human olfactory system. Thus, the sensor array may be considered the heart and most important component of the electronic nose. The instrument is completed by an interface with the computer processing unit (CPU), recognition library and recognition software that serve as the brain to process input data from the sensor array for subsequent data analysis.

A good sensor should fulfill a number of criteria. First, the sensor should have the highest sensitivity to the target group of chemical compound(s) intended for detection with a threshold of detection similar to that of the human nose, down to about 10^{-12} g mL^{-1} [13]. To be most useful with diverse detection capabilities, e-nose sensors should have relatively low selectivity in order to be sensitive to a wide number of different chemical compounds. Also the sensor array must have low sensitivity to variable environmental parameters, in particular to temperature and air humidity. Sensors should be capable of operating at relatively low temperatures when necessary, have short calibration and training requirements, fast recovery time between runs and maintenance procedures to maintain low operating costs. They must also have short recording and analysis times, particularly when used as on-line systems, and high sensor array stability. Adding to all these it ultimately must be very portable and small for convenient diverse operations and with built-in recording and analysis capabilities.

The basis of electrochemical gas sensor operation involves interactions between gaseous molecules and sensor-coating materials that modulate electrical current passing through the sensor, detectable by a transducer that converts the modulation into a recordable electronic signal [14]. There are many different types of electrochemical sensors (e.g. metal-oxide gas sensors, metal-oxide semiconductor field effect transistors, conducting polymer gas sensors, acoustic wave gas sensors, quartz crystal microbalance sensors, surface acoustic wave devices, field-effect gas sensors, electrochemical gas sensors, fiber-optic gas sensors) and many different types of sensor-coating materials which are classified according to additive doping materials, the type and nature of the chemical interactions, the reversibility of the chemical reactions and running temperature. A summary of the types and mechanisms involved with some common gas sensor technologies are listed in **Table 2.1**.

Sensor type	Sensitive material	Detection principle
Acoustic sensors: Quartz crystal microbalance (QMB); surface & bulk acoustic wave (SAW, BAW)	organic or inorganic film layers	mass change (frequency shift)
Calorimetric; catalytic bead (CB)	pellistor	temperature or heat change (chemical reaction)
Catalytic field-effect sensors (MOSFET)	catalytic metals	electric field change
Colorimetric sensors	organic dyes	color changes, absorbance
Conducting polymer sensors	modified conducting polymers	resistance change
Electrochemical sensors	solid or liquid electrolytes	current or voltage change
Fluorescence sensors	Fluorescence-sensitive detector	fluorescent-light emissions
Infrared sensors	IR-sensitive detector	Infrared-radiation absorption
Metal oxides semi-conducting (MOS, Taguchi)	doped semi-conducting metal oxides (SnO2, GaO)	resistance change
Optical sensors	photodiode, light-sensitive	light modulation, optical changes

Table 2.1: Types and mechanisms of common electronic-nose gas sensors

Transducer recording devices of various types in electronic-nose sensors are categorized according to the nature of the physical signal they measure. The most common methods utilize transduction principles based on electrical measurements, including changes in current, voltage, resistance or impedance, electrical fields and oscillation frequency. Others involve measurements of mass changes, temperature changes or heat generation. Optical sensors measure the modulation of light properties or characteristics such as changes in light absorbance, polarization, fluorescence, optical layer thickness, color or wavelength (colorimetric) and other optical properties.

The most widely used class of gas sensors in the electronic nose field is the metal-oxide gas sensor. They were first used commercially in the 1960s as household gas alarms in Japan [13]. More recent uses include applications in many different industrial processes. Basically, a metal-oxide sensor consists of a ceramic support tube containing a heater spiral, usually composed of platinum. The most widely used coating material is tin-dioxide (SnO_2), doped with small amounts of catalytic metal additives. The sorption of gas molecules provokes changes in conductivity brought about by combustion reactions with oxygen species on the surface of the tin-dioxide particles.

The metal-oxide semiconductor field effect transistors (MOSFET) were firstly reported by Lundström *et al.* [15] in 1975 based on the tendency of a number of metals to adsorb and dissolve hydrogen [16]. Metal oxide semiconductor (MOS) sensors consist of three layers: a silicon semiconductor, a silicon oxide insulator and a catalytic metal through which the applied voltage creates an electric field. When polar compounds interact with the metal, the electric field is modulated and recorded by the transistor [13].

Conducting or conductive polymer gas sensors operate based on changes in electrical resistance caused by adsorption of gases onto the sensor surface. Conductive electroactive polymers have attracted much interest for use as electronic noses since the early 1980s [17], particularly because they have high sensitivities, short response times, they are easily synthesized, have good mechanical properties and are particularly useful because they operate at room temperature [18]. Conductive polymer gas sensors consist of a substrate, usually silicon, a pair of gold-plated electrodes and a conducting organic polymer coating as the sensing element [13]. The sensitivity of conductive polymers to VOCs is measured as changes in electrical resistance. Conducting polymers are usually synthesized by chemical or electrochemical oxidizing of the corresponding monomers. The most widely used sensor coating monomers are polypyrrole, polyaniline and polythiophene [14]. The common feature of conductive polymer materials is the presence of a conjugated pi-electron system that extends over the whole polymer.

Acoustic wave gas sensors use a mechanical (acoustic) wave as the sensing mechanism. As the acoustic wave propagates through or on the surface of the sensor coating material, any changes to the characteristics of the propagation path, due to the sorption of VOCs, affect the velocity and/or amplitude of the wave [19]. They consist of a piezoelectric substrate, usually quartz (SiO2), lithium niobate (LiNbO3), lithium tantalite (LiTaO3) or zinc oxide, doped with a suitable surptive material.

Electrochemical gas sensors operate at room temperature, have low power consumption and their sensing methodology is based on the electrochemical oxidation or reduction of volatile molecules at a catalytic electrode surface [20]. This technology has a good relevance when applied to the detection and measurement of electrochemically active gases, but they are not very sensitive to a wide diversity of compounds, especially aromatic hydrocarbons [21].

There are a variety of advantages and disadvantages of using various e-nose sensors based on their response and recovery times, sensitivities, detection range, operating limitations, physical size, inactivation by certain poisoning agents, and other limitations that are specific to individual sensor types. The types and categories of advantages and limitations associated with individual e-nose sensor types are closely linked with the nature of the technology that determines the principle for detection and the types of gas analytes that may be detected with each sensor type. A listing of some of the major advantages and disadvantages associated with each e-nose sensor type are summarized in **Table 2.2**

Sensor type	Advantages	Disadvantages
Calorimetric or catalytic bead (CB)	Fast response and recovery time, high specificity for oxidized compounds	High temperature operation, only sensitive to oxygen-containing compounds
Catalytic field-effect sensors (MOSFET)	Small sensor size, inexpensive operating costs	Requires environmental control, baseline drift, low sensitivity to ammonia and carbon dioxide
Conducting polymer sensors	Ambient temperature operation, sensitive to many VOCs, short response time, diverse sensor coatings, inexpensive, resistance to sensor poisoning	Sensitive to humidity and temperature, sensors can be overloaded by certain analytes, sensor life is limited
Electrochemical sensors (EC)	Ambient temperature operation, low power consumption, very sensitive to diverse VOCs	Bulky size, limited sensitivity to simple or low mol. wt. gases
Metal oxides semiconducting (MOS)	Very high sensitivity, limited sensing range, rapid response and recovery times for low mol. wt. compounds (not high)	High temperature operation, high power consumption, sulfur & weak acid poisoning, limited sensor coatings, sensitive to humidity, poor precision
Optical sensors	Very high sensitivity, capable of identifications of individual compounds in mixtures, multi-parameter detection capabilities	Complex sensor-array systems, more expensive to operate, low portability due to delicate optics and electrical components
Quartz crystal microbalance (QMB)	Good precision, diverse range of sensor coatings, high sensitivity	Complex circuitry, poor signal-to-noise ratio, sensitive to humidity and temperature
Surface acoustic wave (SAW)	High sensitivity, good response time, diverse sensor coatings, small, inexpensive, sensitive to virtually all gases	Complex circuitry, temperature sensitive, specificity to analyte groups affected by polymeric- film sensor coating

Table 2.2: Summary of advantages and disadvantages of e-nose sensor types

Conducting polymer and electrochemical sensors are probably the most versatile e-nose sensor types available due to operation at ambient or room temperature, low power consumption, good sensitivity to a wide range of gas or volatile analytes, and inexpensive operating costs. Conducting polymers are available in a very large and diverse range of sensor coating types providing almost unlimited combinations of sensors in the array for analysis of any specific organic chemical classes or VOC mixture types possible in any particular application. This versatility of conducting polymer sensors is especially true as the number of sensors in the array increases although more sensors does not necessarily mean better efficiency of detection, portability, or operating costs. By contrast, metal oxide and calorimetric or catalytic bead sensors must operate at high temperatures, resulting in greater operating costs, and have much more limited range of detectable analytes. Nevertheless, certain analytes always require high temperature sensors for effective detection and sensitivity.

2.1.6. Data analysis approaches for electronic nose systems

The analogue outputs generated by e-nose sensors have to be analyzed and interpreted in order to provide useful information to the operator. Commercially available analysis techniques fall into three main categories as follows [13]:

1. Graphical approaches: bar charts, profiles, polar and offset polar plots
2. Multivariate data analysis (MDA): principal component analysis (PCA), canonical discriminate analysis (CDA), featured within (FW) and cluster analysis (CA)
3. Artificial intelligence approaches: artificial neural networks (ANN) and radial basis functions (RBF)

The choice of the method used depends on the type of available input data acquired by the sensors and the type of information that is sought. The simplest form of data reduction is the graphical analysis approach, useful for comparing samples or comparing aroma identification elements of unknown analytes relative to those of known sources in reference libraries. Multivariate data analysis comprises a set of techniques for the analysis of data sets with more than one variable by reducing high dimensionality in a multivariate problem when variables are partly correlated, so they can be displayed in two or three dimensions. For electronic-nose data analysis, MDA is very useful when sensors have partial-coverage sensitivities to individual compounds present in the sample mixture. Multivariate analysis can be divided into untrained or trained techniques. Untrained techniques are used when a database of known samples has not been previously built, therefore it is neither necessary nor intended for recognizing the sample itself, but for making comparisons between different unknown samples, to discriminate between them. The simplest and most widely used untrained MDA technique is principal component analysis. PCA is most useful when no known sample is available, or when hidden relationships between samples or variables are suspected. On the contrary, trained or supervised learning techniques classify unknown samples on the basis of characteristics of known samples or sets of samples with known properties that are usually maintained in a reference library that is accessed during analysis.

Artificial neural networks (ANN) are the newest analysis techniques incorporated to statistical software packages for commercially available electronic noses. Mimicking the cognitive processes of the human brain, neural network approaches are based on interconnected data processing algorithms that work in parallel. Various instrument-training methods are employed through pattern-recognition algorithms that look for similarities and differences between identification elements of known aroma patterns found in an analyte-specific reference library. The training process requires a discrete amount of known sample data to train the system and is very efficient in comparing unknown samples to known references [24]. The result of ANN data analysis usually is in the form of a percentage match of identification elements in the sample with those of aroma patterns from known sources in the reference library.

2.1.7. Commercial electronic noses

Following the Gardner-Bartlett definition, a detection device to be considered an electronic nose must contain an intelligent chemical-array sensor system that mimics the mammalian olfactory system and is used specifically to sense aromatic VOCs. The implication is that all sensing devices that have only one sensor or can detect only one compound or aroma (electronic aroma monitors) cannot by definition be considered electronic noses. Thus, electrochemical cells (ECs) that detect only one specific gas are not electronic noses according to the Garner-Bartlett definition.

Being around for almost 30 years as a concept and 20 as a commercial instrument, the electronic predictions regarding market sizes did not live to the previous expectations. Thus many electronic noses didn't survive the competitive economy. Some reasons for which this happened are high apparatus costs and yet an immature technology.

Some electronic noses are still commercially available today and have a wide range of applications in various markets and industries ranging from food processing, industrial manufacturing, quality control, environmental protection, security, safety and military applications to various pharmaceutical, medical, microbiological and diagnostic applications. A summary of some of the most widely used electronic noses with manufacturers, models available and technological basis are listed in **Table 2.3**.

Instrument type	Manufacturer	Models	Technology basis
Single technology (e-nose sensors only)	Airsense Analytics	i-Pen, PEN2, PEN3	MOS sensors
	Alpha MOS	FOX 2000, 3000, 4000	MOS sensors
	Applied Sensor	Air quality module	MOS sensors
	Chemsensing	ChemSensing Sensor array	Colorimetric optical

	CogniScent Inc.	ScenTrak	Dye polymer sensors
	CSIRO	Cybernose	Receptor-based array
	Dr. Födisch AG	OMD 98, 1.10	MOS sensors
	Forschungszentrum Karlsruhe	SAGAS	SAW sensors
	Gerstel GmbH Co.	QSC	MOS sensors
	GSG Mess- und Analysengeräte	MOSES II	Modular gas sensors
	Illumina Inc.	oNose	Fluorescence optical
	Microsensor Systems Inc	Hazmatcad, Fuel Sniffer, SAW MiniCAD mk II	SAW sensors
	Osmetech Plc	Aromascan A32S	Conducting polymers
	Sacmi	EOS 835, Ambiente	Gas sensor array
	Scensive Technol.	Bloodhound ST214	Conducting polymers
	Smiths Group plc	Cyranose 320	Carbon black-polymers
	Sysca AG	Artinose	MOS sensors
	Technobiochip	LibraNose 2.1	QMB sensors
Combined technology (e-nose + other types)	Airsense Analytics	GDA 2	MOS, EC, IMS, PID
	Alpha MOS	RQ Box, Prometheus	MOS, EC, PID, MS
	Electronic Sensor Technology	ZNose 4200, 4300, 7100	SAW, GC
	Microsensor Syst.	Hazmatcad Plus, CW Sentry 3G	SAW, EC
	Rae Systems	Area RAE monitor	CB, O_2, EC, PID
		IAQRAE	Thermistor, EC, PID, CO_2, humidity
	RST Rostock	FF2, GFD1	MOS, QMB, SAW

Table 2.3 Commercially available electronic noses

These represent a wide diversity of sensor types based on uniquely different technologies. The list includes instruments with single-technology sensor arrays and combined-technology instruments that consist of e-noses working in tandem with classical analytical systems. The additional need to identify individual chemical species or components within sample mixtures, beyond the identity of the sample (source) as a whole, recently has caused a necessary merger of electronic nose technologies with purely analytical instruments. These technological mergers have resulted in new instrumentations that have diffused the border between electronic noses and conventional analytical instruments. Nevertheless, the division between pure electronic-nose instruments based on collective sensor-array outputs and classical analytical instruments with single-detector outputs is clear. Classical analytical instruments and detectors such as electron capture detectors (ECD), flame ionization detectors (FID), flame photometry detector (FPD), gas chromatographs (GC), infrared spectrometers (IRS), ion mobility spectrometers (IMS), mass spectrometers (MS), nuclear magnetic resonance spectrometers (NMRS), photoionization detector (PID) and quadrupole fingerprint mass spectrometers (QFMS) are not considered electronic noses in the strictest sense because they do not provide a collective data output from a sensor array and are designed to detect and identify individual components of a gas mixture.

The uses of electronic noses have grown rapidly as new applications have been discovered. The numbers of e-noses sold by various manufacturers has largely depended on the technology basis of individual instruments, costs per unit, and specific application needs. In 1997, there were about 500 total desk-top analytical instruments units sold worldwide with an approximate market value of $30 million Euros [14]. Within the past ten years, the Applied Sensor Company has sold the highest number of units (> 100,000) of their e-nose (the Air Quality Module electronic nose). Their system is primarily used to maintain ambient or environmental air quality by detection of odors, VOCs and carbon dioxide within living spaces [22].

The Alpha-MOS Fox electronic nose was designed in collaboration with the Universities of Warwick and employs either six (Fox 2000), 12 (Fox 3000) or 18 (Fox 4000) metal oxide gas sensors and can be used with external carrier gas bottles in a flow-injection system, or with an internal pump and mass-flow controller.

Among the electronic noses that did not survived the economy we shortly mention: Aromascan A32S, Cyranose 320 , Airsense PEN2. The Aromascan A32S was an organic matrix-coated polymer-type 32-detector e-nose based on an earlier design using technology arising from the University of Manchester, Institute of Science and Technology. Aromascan was acquired by the Osmetech company, which currently produces the eSensor XT-8 system, a diagnosis platform by means of electrochemical detection technology to detect nucleic acids on microarrays. The Cyranose 320 was a portable electronic-nose system whose component technology consists of 32 individual polymer sensors blended with carbon black composite and configured as an array [23]. The company was acquired by Smiths Detection, whose products are oriented towards surveillance, explosive and narcotics detection. Airsense's PEN2 and PEN3 e-noses were unique on the electronic nose market in the sense that they were a very small and portable 10 metal-oxide semiconductor (MOS) gas sensor array with a small-volume measuring chamber. It can be linked with an adsorbent trapping unit or a headspace auto sampler for laboratory analyses. Now the company does not market the product anymore, being oriented towards fire detection systems.

2.1.8. Electronic nose applications

Electronic-nose systems have been specifically designed to be used for numerous applications in many different industrial production processes. A wide variety of industries based on specific product types and categories, such as the automobile, food, packaging, cosmetic, drug and biomedical industries use e-noses for a broad and diverse range of applications including quality control of raw and manufactured products, process design, freshness and maturity (ripeness) monitoring, shelf-life investigations, authenticity assessments of premium products, classification of scents and perfumes, microbial pathogen detection and environmental assessment studies (**Table 2.4**).

In the fruit industry, the age of fruits (ripeness or maturity level) determines the shelf life and future rate of quality loss due to changes in flavor, firmness and color. Harvesting fruits at an optimal physiological condition ensures good quality at a later stage (when evaluated by the consumer) by enhancing a number of quality characteristics that extend the shelf life, slow the rate of decline in firmness or texture, and maintain a preferred level of flavor and overall appearance.

Industry sector	Application area	Specific use types and examples
Agriculture	crop protection	homeland security, safe food supply
	harvest timing & storage	crop ripeness, preservation treatments
	meat, seafood, & fish products	freshness, contamination, spoilage
	plant production	cultivar selection, variety characteristics
	pre- & post-harvest diseases	plant disease diagnoses, pest identification detect non-indigenous pests of food crops
Airline transportation	public safety & welfare passenger & personnel security	explosive & flammable materials detection
Cosmetics	personal application products	perfume & cologne development
	fragrance additives	product enhancement, consumer appeal
Environmental	air & water quality monitoring	pollution detection, effluents, toxic spills
	indoor air quality control	malodor emissions, toxic/hazardous gases
	pollution abatement regulations	control of point-source pollution releases
Food & beverage	consumer fraud prevention	ingredient confirmation, content standards
	quality control assessments	brand recognition, product consistency
	ripeness, food contamination	marketable condition, spoilage, shelf life
	taste, smell characteristics	off-flavors, product variety assessments

Manufacturing	processing controls	product characteristics & consistency
	product uniformity	aroma and flavor characteristics
	safety, security, work conditions	fire alarms, toxic gas leak detection
Medical & clinical	pathogen identification	patient treatment selection, prognoses
	pathogen or disease detection	disease diagnoses, metabolic disorders
	physiological conditions	nutritional status, organ failures
Military	personnel & population security	biological & chemical weapons
	civilian & military safety	explosive materials detection
Pharmaceutical	contamination, product purity	quality control of drug purity
	variations in product mixtures	formulation consistency & uniformity
Regulatory	consumer protection	product safety, hazardous characteristics
	environmental protection	air, water, and soil contamination tests
Scientific research	botany, ecological studies	chemotaxonomy, ecosystem functions
	engineering, material properties	machine design, chemical processes
	microbiology, pathology	microbe and metabolite identifications

Table 2.4: Industry-based applications for electronic noses

Traditional measuring techniques are destructive and involve random sampling to assess fruit quality. Consequently, individual fruits or fruit clusters are not graded for quality assessments needed for optimizing treatments and marketing strategies. There is a need for non-destructive techniques to assess fruit quality based on aroma characteristics that are highly correlated with all of the factors that affect shelf life and future marketability.

Several studies have demonstrated that the aroma emitted by fruits can indicate the maturity level and thus quality and shelf life of the marketed product. Pathange *et al.* [25] used maturity indices to categorize fruit of the "Gala" apple variety into three maturity groups referred to as immature, mature and over-mature fruits. Multivariate analysis of variance of the e-nose sensor data indicated that the instrument could classify the fruit into the correct group in 87% of the samples. Oshita *et al*. [26] examined the shelf life of "La France" pears and judged them using e-nose data as either immature, mature or over-mature. The aromas of pears were classified based on their physiological states determined from distinct aroma patterns derived from a 32-sensor array output.

Predictions of fruit maturity level and shelf life have been done on various fruits. For example, fried mango chips were evaluated for the presence of deteriorative aromas [27]. Fuji apples were evaluated for the effects of different storage conditions, storage periods and days of shelf life on ripening and condition [28]. Supriyadi [35] investigated the specific aroma of a pentane extract in snake fruit. Others have examined the pre- and post-harvest characteristics of kiwifruit [30], and fresh-cut vegetables like chicory [31].

Using e-noses as a means of monitoring fruit freshness and shelf-life prior to marketing can have a number of benefits that maximize profits and optimize customer satisfaction. Information from e-noses on fruit physiological states, based on changes in released volatiles, can be applied to retard the ripening process through exposure of the fruit to ripening inhibitors at the appropriate time, adjustments in fruit storage conditions, and removal of bruised or damaged fruits that enhance ripening of surrounding fruits and contribute to storage losses due to rots, decays, and various fruit diseases.

Dairy products contain off-flavor compounds created by a variety of mechanisms such as through the action of natural and microbial enzymes and chemical changes catalyzed by light or heavy metals. In cheese, quality, flavor and taste are closely connected to the ripening process, which depends on the growth of bacteria, lipid degradation and oxidation, and proteolysis. Traditionally, sensory analysis was used to determine the product identity of cheese. However, detection of aroma compounds using electronic noses has become more and more important.

Much work has been done in the electronic detection of quality characteristics of meat products within the food industry. Berdagué and Talou [32] studied instruments based on MOS sensors, starting in 1993 with the Alabaster-UV. It was shown that it could differentiate maxima in aroma perception resulting from the maturation of dry non-spiced meats, and rapidly detected sex-linked differences in meat product composition. Vernat-Rossi *et al*. [33] demonstrated that non-controlled ambient air that simulated on-line quality control could be used in the rapid discrimination of food products. Vestergaard *et al.* [34] found the storage time of a pork-meat pizza-topping product was predictable using an electronic nose.

Olive oil, and in particular virgin olive oil is highly appreciated by consumers due to its nutritional benefits and is usually considerably more expensive than other edible vegetable oils. Hence, adulteration with cheaper or lower quality oils may afford important benefits from the economic point of view. Among the most frequent adulterations are those carried out with sunflower oil, maize oil, olive–pomace oil [35], and even with hazelnut oil [36]. Thus, continuous vigilance is required to control the adulteration of olive oil products and to protect the interests of consumers.

The development of new techniques for the detection of adulteration of olive oil is one of the most fundamental subjects for many researches. According to the International Olive Oil Council, olive oil has been classified into various grades, namely virgin olive oils (including extra virgin olive oil, virgin olive oil and ordinary virgin olive oil), refined olive oils and olive–pomace oils (including crude olive–pomace oil, refined olive–pomace oil and olive–pomace oil) [37]. Virgin olive oil is obtained from the fruit of the olive tree only by mechanical pressing. It is the highest quality oil and has a free acidity, expressed as percentage of oleic acid, of no more than 0.8 grams per 100 grams. Because of the high price of virgin olive oil, adulteration of olive oil with low-grade olive oils – olive–pomace oil or refined olive oil and other cheaper vegetable oils such as hazelnut oil, sunflower (SF) oil, soybean oil and maize oil, has taken place[38].

Various analytical methods have been used to detect virgin olive oil adulteration. Most of these are based on spectroscopy (ultraviolet, near-infrared [NIR], mid-infrared, visible, Raman), isotopic analysis, chromatography (gas chromatography, high performance liquid chromatography) and electronic nose systems [39, 40]. However, most of these techniques require too much time for routine use in the food industry.

New techniques based on the generation of a headspace have been developed. These are particularly attractive in the sense that they measure volatile substances in the same way that they are detected by the human olfactory system. Among the different headspace methods available, those employing a dynamic headspace have the advantage of including a pre-concentration step, thereby improving detection limits [41]

Several examples can be found in the literature that demonstrate the success of using an electronic nose for the quality evaluation of olive oils. Most of them use an array of conducting polymer sensors [42] for the discrimination of quality, variety of olive or geographic origin. The system has been applied to detect the rancid defect in virgin olive oil [43].

Among the frequent adulterations encountered one of the most hard to detect is with hazelnut oil. Because of the similarities of the two oils, concentrations below 20% of hazelnut oil in olive oil has been difficult to confirm. Analytical methods for monitoring olive oil adulteration with hazelnut oil include determination of filbertone ((E)-5-methylhept-2-en-4-one) by liquid chromatography–gas chromatography (LC–GC) and isotope dilution [44], composition of sterols and triacylglycerols by GC and high-field ^{1}H NMR [45],non-volatile marker components present in the polar fraction of hazelnut oils by RP-HPLC [46]. However, the amount and composition of sterols cannot detect levels of hazelnut oil below 30% [47].

Profile of volatile compounds of oils, especially those obtained by cold pressing and unrefined could serve as a valuable tool in the identification of their origin and adulteration. Analysis of volatile compounds can be performed by GC/MS, based on their separation on analytical column and subsequent identification using mass spectrometry or other detector.

The alternative approach to this method will be to use the electronic nose technology. In recent years e-noses have been applied to the detection of rancidity [48], adulterants [49] and determination of geographical origin [50]. Two types of electronic noses can be found on the market, sensor based and mass spectrometer based, from which the predominantly used one is the metal oxide semiconductor based electronic nose.

Modern medicine faces the problem and challenge of achieving effective disease diagnoses through early detections in order to facilitate the application of rapid treatments, but at the same time dramatically reducing the invasiveness of diagnostic treatments. Many medical researchers have published experimental data in the last ten years to demonstrate the feasibility of using the electronic nose to diagnose human diseases and to identify many different pathogenic microorganisms through the detection of the VOCs they emit both in vitro and in vivo [51].

Some highly pathogenic gastroesophageal bacteria were correctly discriminated by Pavlou *et al.* [52]. More recently, other in vitro studies have reported the feasibility of recognizing several strains of two anaerobic bacteria and two fecal pathogens (*E. coli* and *Salmonella typhimurium*) by the use of an electronic nose [53]. Urinary tract infections have been thoroughly investigated by Di Natale *et al.* [54] and Aathithan *et al.* [55]. Pavlou *et al*. [56] proposed the use of the electronic nose as a diagnostic tool for patients affected with kidney diseases, by distinguishing traces of blood in urine samples, and for the rapid identification of *E. coli*, *Proteus* spp. and *Staphylococcus* spp. infections at very high levels of confidence. The electronic nose has interesting applications in the analysis of human breath to potentially provide quick diagnosis of many diseases. In the case of pneumonia diagnosis, Hockstein *et al.* [57] discriminated between diseased and non-diseased patients with accuracy as high as 91.6%.

One of the most disputed and promising applications of electronic nose technologies is for the early detection and diagnosis of oncologic diseases, in particular lung cancer. The initial work to evaluate the feasibility of the electronic nose in detecting these cancer-marker compounds was done by Di Natale *et al.* [58]. The electronic nose could successfully detect 100% of lung cancer affected patients, 94% of controls and 44% of post-surgery patients.

2.1.9. Electronic nose terminology

The electronic nose was born from the real need in the food industry for objective, automated quality-monitoring sampling systems that can characterize the odor of a product and determine if the production is running according to standards without the need of a human sensory panelist, or time consuming analytical methods and data interpretation.

The term electronic nose came from the similarities with the mammalian olfaction system. They both share the same principle. Human olfactory receptors or chemical / physical sensors transmit their signals to the brain or to a pattern recognition system respectively.

However the similarities stop here. An electronic nose does not exactly replicate the mammalian nose. Not everything that smells can be a good e-nose application and not everything that can be measured with the electronic nose has a smell. The olfactory receptors have evolved over many years of evolution to specific molecules, being optimized towards food analysis to ensure survival. They are good to indicate rotten food, sexual pheromones, etc. On the other hand sensors work very different. An electronic nose has no problem in detecting gases that a mammalian nose cannot, as long as they are in sufficient quantity. Gases like carbon monoxide and hydrogen cyanide, even though present a danger to life are not detectable by a mammalian nose, being around for too few years for the nose to adapt to them. Also, form the odorous point of view; electronic noses not always can succeed in detecting them. Evolution has blessed mammals with low limit of detection for certain gases, and with high receptor specificity, while the principles of the electronic nose do not allow the same performance. Thus an odor description form an electronic nose is not always possible.

Because of these differences in measuring the molecules, odorant or not, the term electronic nose should not be stressed too much. Some authors [59] point out the fact that the two noses do not work in the same way and do not detect the same constituents, and therefore suggest the use of the term electronic nose between quotation marks. Mielle P. also suggests the use of other terms like: flavor sensor, odor sensor, aroma sensor or gas sensor, which do not suggest such strong links with the mammalian olfaction.

Because an electronic nose analyses the whole volatile species, odor and odorless, while the mammalian nose focuses only on the odorant volatiles the real problem with the electronic nose definition arises from the subjectivity of odors; molecules that are predominantly detected by mammals. Therefore, the emphasis is on the mammalian nose not on the volatiles to be measured. By making the electronic nose objective, and not taking into consideration if the sample to be analyzed can or not be smelled, the electronic nose could become the better nose. It could recognize and explain to the rational human brain why it makes the classifications it does. In order to do so, some drawbacks are to be overcome.

2.1.10. Advantages and drawbacks of the chemical sensor electronic nose

As explained, gas-sensor e-noses operate on the basis of physical and/or chemical interactions of the volatile compounds with an array of solid-state gas sensors, each of which has a partial specificity. Ideally, the sensors should fulfill a number of requirements, such as broad selectivity, high sensitivity, robustness, stability, rapid response and rapid recovery times. Moreover, if they are used in poorly controlled environments, they should be required to have a low sensitivity to environmental variables, such as temperature and humidity. A range of sensor technologies has been reported for e-nose applications [60], although no single gas sensor technology fulfils all these requirements.

Metal-oxide semiconductors (MOS), probably the most widely used in e-nose applications today, display a high level of sensitivity for a wide range of organic vapors and provide perhaps the best balance between drift, lifetime and sensitivity. However they have important disadvantages like sensitivity to humidity, poor precision, high temperature operation, etc. Also, MOS devices have a logarithmic dependence of the sensor response on the gas concentration. This causes problems in the presence of high concentrations of detectable species (e.g., ethanol in alcoholic beverages). Other potential problems have also been reported when MOS devices are used with food products (e.g., the baseline recovery is slow when they are exposed to high molecular weight compounds or they are susceptible to poisoning by sulphur-containing species or by weak acids) [61].

Conducting polymer (CP) sensors have the advantage of operating at room temperature, and their selectivity is generally better than that of MOS sensors. They are also relatively resistant to poisoning. But the polymer response is sensitive to humidity. This is an important drawback when water is a major component of the sample headspace.

The quartz crystal microbalances (QMBs) and the surface acoustic wave transducers (SAWs), both of which are acoustic wave devices, show great promise. Their sensitivity is at the µg/l level whereas the sensitivities of MOS and CP sensors are at the mg/l level. However, poor batch-to-batch reproducibility during manufacture and dependence of the response on temperature [60] are problems that still have to be addressed.

At first, e-noses tended to be based on an array of gas sensors with the same technology, but practical experience has shown that often this does not produce enough information for many real-world problems. Increasingly, the tendency is to combine different technologies of gas sensors to produce hybrid systems. However, this involves more complex electronics and it is then necessary to normalize or standardize the different sensor outputs [60].

Despite the considerable number of applications that appear in food-analysis literature [60, 62, 63], much development is still required before e-noses based on gas sensors can reach their full potential in real, commercially ready applications.

2.1.11. The mass spectrometry based electronic nose

Mass spectrometry has been described as the smallest scale in the world, not because of the mass spectrometer’s size but because of the size of what it weighs: molecules. Over the past decade, mass spectrometry has undergone tremendous technological improvements allowing a wide range of application from steroid detection in anti-doping for athletes, to petroleum, comet dust and biological analysis of proteins, peptides, carbohydrates, DNA, drugs, and many other relevant molecules.

A mass spectrometer determines the mass of a molecule by measuring the mass-to-charge ratio (*m/z*) of its ion. Ions are generated by inducing either the loss or gain of a charge from a neutral species. Once formed, ions are electrostatically directed into a mass analyzer where they are separated according to *m/z* and finally detected. The result of molecular ionization, ion separation, and ion detection is a spectrum that can provide molecular mass and even structural information [64].

2.1.11.1. A short history of mass spectrometry

The foundations of mass spectrometry date back nearly 200 years to John Dalton, who proposed that all matter is composed of elements, that the number and types of atoms corresponded to the exact number and types of elements, and that all elements have a measurable physical property, the atomic weight. Later, J.J. Thompson exploring the composition of canal rays channeled a stream of ionized neon through an electric field and measured its deflection by placing a photographic plate in its path. Two patches of light were observed on the photographic plate, and Thompson concluded that neon is composed of atoms of two different atomic masses (isotopes neon-20 and neon-22). This observation that lighter atoms responded differently than heavier atoms in magnetic and electric fields led to a better understanding of elements and their physical properties, in particular mass, and set the scene for the design of instruments to measure it.

The separation of neon isotopes by their mass is the first example of mass spectrometry. F. W. Aston and A. J. Dempster continued the research, building the first fully functional mass spectrometer in 1919.

2.1.11.2. Instrument description

Four basic components are, for the most part, standard in all mass spectrometers (**Figure 2.1**.): a sample inlet, a ionization source, a mass analyzer and an ion detector. Some instruments combine the sample inlet and the ionization source, while others combine the mass analyzer and the detector. However, all sample molecules undergo the same processes regardless of the configuration of the instrument. Sample molecules are injected into the instrument through a sample inlet. Once inside the instrument, the sample molecules are converted to ions in the ionization source, before being electrostatically propelled into the mass analyzer. Ions are then separated according to their *m/z* within the mass analyzer. The detector converts the ion energy into electrical signals, which are then transmitted to a computer.

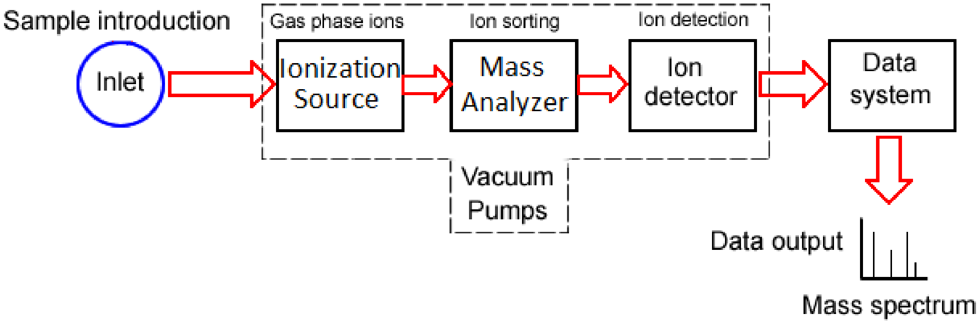


Figure 2.1: Components of a mass spectrometer. The ion source does not have to be within the vacuum of the mass spectrometer. For instance, ESI and APCI are at atmospheric pressure and are known as atmospheric pressure ionization (API) sources

Sample introduction was an early challenge in mass spectrometry. Nowadays, in order to perform mass analysis on a sample, which is initially at atmospheric pressure, it must be introduced into the instrument in such a way that the vacuum inside the instrument remains relatively unchanged. The most common methods of sample introduction are direct insertion with a probe or plate commonly used with MALDI-MS (Matrix-assisted laser desorption/ionization-Mass Spectrometry), direct infusion or injection into the ionization source such as ESI-MS (electrospray ionization mass spectrometry).

The Ionization method refers to the mechanism of ionization while the ionization source is the mechanical device that allows ionization to occur. The different ionization methods work by either ionizing a neutral molecule through electron ejection, electron capture, protonation, cationization, or deprotonation, or by transferring a charged molecule from a condensed phase to the gas phase. Each method has its advantages and disadvantages.

Prior to the 1980s, electron ionization (EI) was the primary ionization source for mass analysis. However, EI limited chemists and biochemists to small molecules well below the mass range of common bio-organic compounds. Also EI often generates too much fragmentation and it can be unclear if the highest mass ion is a molecular ion or a fragment. This limitations motivated scientists to develop the new generation of ionization techniques, including fast atom/ion bombardment (FAB), matrix-assisted laser desorption/ionization (MALDI), and electrospray ionization (ESI) (**Table 2.5**.). These techniques have revolutionized biomolecular analyses, especially for large molecules. Among them, ESI and MALDI have clearly evolved to be the methods of choice when it comes to biomolecular analysis, offering excellent mass range and sensitivity.

Ionization source	Acronym	Event
Electrospray ionization	ESI	evaporation of charged droplets
Nanoelectrospray ionization	nanoESI	evaporation of charged droplets
Atmospheric pressure chemical ionization	APCI	corona discharge and proton transfer
Matrix-assisted laser desorption/ionization	MALDI	photon absorption/proton transfer
Desorption/ionization on silicon	DIOS	photon absorption/proton transfer
Fast atom/ion bombardment	FAB	ion desorption/proton transfer
Electron ionization	EI	electron beam/electron transfer
Chemical ionization	CI	proton transfer

Table 2.5: Ionization techniques

All mass spectrometers rely on a *mass analyzer*, but not all analyzers operate in the same way. Some separate ions in space while others separate ions by time. In the most general terms, a mass analyzer measures gas phase ions with respect to their mass-to-charge ratio (*m/z*), where the charge is produced by the addition or loss of a proton(s), cation(s), anion(s) or electron(s). The addition of charge allows the molecule to be affected by electric fields thus allowing its mass measurement. It is important to remember that a mass spectrometer measures the m/z ratio, not the mass. If an ion has multiple charges, the m/z will be less.

The first mass analyzers, made in the early twentieth century, used magnetic fields to separate ions according to their radius of curvature through the magnetic field. The design of modern analyzers has changed significantly in the last five years, now offering much higher accuracy, increased sensitivity, broader mass range, and the ability to give structural information. Because ionization techniques have evolved, mass analyzers have been forced to change in order to meet the demands of analyzing a wide range of biomolecular ions with part per million mass accuracy and sub fentomole sensitivity. (**Table 2.6**)

Mass analyzer	Event
Quadrupole	scan radio frequency field
Quadrupole Ion Trap	scan radio frequency field
Time-of-Flight (TOF)	time-of-flight correlated directly to ion's *m/z*
Time-of-Flight Reflectron	time-of-flight correlated directly to ion's *m/z*
Quad-TOF	radio frequency field scanning and time-of-flight
Magnetic Sector	magnetic field affects radius of curvature of ions
Fourier Transform Ion Cyclotron Resonance MS	translates ion cyclotron motion to *m/z* (FTMS)

Table 2.6: Mass analyzers

The performance of a mass analyzer can typically be defined by the following characteristics: accuracy, resolution, mass range, tandem analysis capabilities, and scan speed.

Once the ions are separated by the mass analyzer, they reach the *ion detector*, which generates a current signal from the incident ions. The most commonly used detector is the electron multiplier, which transfers the kinetic energy of incident ions to a surface that in turn generates secondary electrons. However, a variety of approaches are used to detect ions depending on the type of mass spectrometer.

The most common means of detecting ions involves an electron multiplier, which is made up of a series of aluminum oxide dynodes maintained at ever increasing potentials. Ions strike the first dynode surface causing an emission of electrons. These electrons are then attracted to the next dynode held at a higher potential and therefore more secondary electrons are generated. Ultimately, as numerous dynodes are involved, a cascade of electrons is formed that results in an overall current gain on the order of one million or higher.

All mass spectrometers need a vacuum to allow ions to reach the detector without colliding with other gaseous molecules or atoms. If such collisions did occur, the instrument would suffer from reduced resolution and sensitivity.

2.1.11.3. Sampling system

The mass spectrometry based electronic nose, as in the case of chemical sensor e-noses is a union of three elements: *the sample handling system*, *the detection system*, and *the data processing system*.

The only difference between the two types of electronic noses is the detection system, and therefore the sample handling systems used in chemical sensors based artificial olfaction can also be used in MS-based e-noses. Based on the sample to be analyzed we can choose between: *static headspace*, *purge and* *trap*, *dynamic headspace*, *solid-phase microextraction, stir bar sorptive extraction*, and i*nside-needle dynamic extraction*.

To these methods we could add *membrane introduction mass spectrometry* (MIMS), which is a sample handling system used in mass spectrometry based e-noses. This technique allows the direct introduction of specific compounds of a liquid or gas sample into a mass spectrometer. A thin membrane is installed between the sample and the ion source of a mass spectrometer in such a way that some compounds dissolve in the membrane, diffuse through it and, finally, evaporate directly into the ion source [65].

2.1.11.4. History of MS-based e-noses

Developed only about ten years ago [66], the e-noses based on MS represent a new approach to the electronic nose technology. This new instrument introduces the volatile compounds into the ionization chamber of a MS instrument without prior chromatographic separation [67, 68]. The mass spectrum obtained results from the simultaneous ionization and fragmentation of all the volatile compounds coming from the sample. Each fragment ion (*m/z* ratio) represents a “pseudo sensor” and its abundance is equivalent to a sensor signal. Therefore, the number of “pseudo sensors” in MS-based e-noses is much larger than in gas-sensor-based e-noses. Moreover, by selecting the optimum set of fragment ions (the optimum “sensor array”), the instrument can be tailored to particular applications leading to a successful discrimination. The selection of particular fragment ions may be based on a prior knowledge of the composition to be analyzed – analytical chemistry can provide the necessary tools to characterize the samples – or on the results of mathematical feature extraction algorithms. Also the sensitivity and the selectivity of the instrument can be improved using a Mass Spectrometer in the selected ion monitoring (SIM) mode.

Another important advantage is that these “pseudo sensors” contain chemical information about the sample. Therefore, information about what kinds of compounds are responsible for the differences between samples can be obtained from the ion-fragmentation patterns directly.

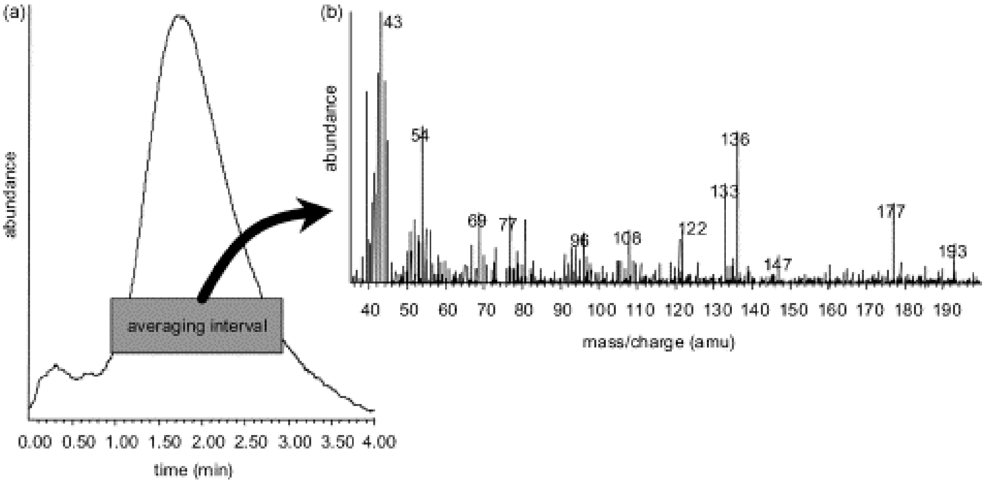


Figure 2.2: mass spectrometry based electronic nose signal

The gas chromatograph mass spectrometer (GC-MS) as a desk operated analytical apparatus can be used as a mass spectrometry-based electronic nose in two ways. First by simply using an uncoated deactivated retention gap as a transfer line between the sample-handling device and the detector instead of the chromatographic column. In this way, all the compounds are introduced into the MS rapidly and simultaneously, because there is no chromatographic separation. And second approach, by applying strong chromatographic conditions (high temperature and high carrier gas flow) to an analytical capillary column (**Figure 2.2**). With the later approach, it is easy to switch from a GC–MS working as an e-nose to a conventional GC–MS, simply by changing the temperature program and the column-gas flow.

2.1.11.5. Data analysis for mass spectrometry based electronic noses

The first step in data analysis is to pre-process the signals generated by the gas sensors or the MS spectra. This process transforms the data into the most appropriate form and enhances the features within the data that are useful in the subsequent steps. The pre-processing steps include mass-spectra normalization, baseline correction, and noise reduction or variable weighing, among others [69]. Afterwards, pre-processed data are analyzed by various chemometric techniques, which are available in statistical software packages that are usually bundled with the instrument.

The chemometric techniques used include unsupervised and supervised pattern recognition (PR) techniques. The former reveal natural groupings of the samples in the data set and also detect outlying samples. Hierarchical cluster analysis (HCA) and principal component analysis (PCA) are the most commonly used unsupervised PR techniques.

When the purpose of the analysis is to predict a continuous property (analyte concentration or aging time), regression methods are required, such as partial least squares (PLS) or principal component regression (PCR).

When the instrumental responses recorded are linear, as in MS-based e-noses, the mentioned statistical methods give very good results. However, in gas-sensor-based e-noses, instrumental responses are essentially non-linear. In such cases, input data must be transformed into linear responses before applying the chemometric techniques described above. The use of artificial neural networks (ANNs) is another alternative for analyzing e-nose non-linear data.

All these data analysis techniques will be treated in detail in 2.2.

2.1.11.6. Commercially available MS-based e-noses

Some commercial electronic noses based on mass spectrometry exist today on the market, and a list of them can be seen in **Table 2.7**:

Manufacturer	Model	Technology
Agilent www.chem.agilent.com	4440A	quadrupole fingerprint mass spectrometry
Alpha MOS www.alpha-mos.com	Kronos	quadrupole fingerprint mass spectrometry
	Prometheus	MS and 18 MOX sensors
Owlstone Nanotech, Inc. www.owlstonenanotech.com	Tourist	field asymmetric ion mass spectrometry
	Lonestar	field asymmetric ion mass spectrometry
SMart Nose smartnose.com	SMart Nose 2000	quadrupole fingerprint mass spectrometry

Table 2.7: spectrometry-based electronic noses

2.1.11.7. Drawbacks of MS e-noses

Even though the mass spectra based electronic nose presents major improvements over solid state-based electronic noses, the lack of portability of the desk instrument impairs its spread as a device. The high cost also limits the availability to only a few researchers. Miniaturization of the device is a must in order to classify this approach as a robust device, one fundamental requests of any electronic nose equipment.

Among these drawbacks a major one is related to the fact that in the case of a mixture of components the mass spectra electronic base has limitations, not being able to separate between small variations in the concentrations of the measured solution constituents. Introducing a chromatographic separation instead of using no separation improves the performances of the device by being able to identify small variations in components variation.

2.1.12. The gas chromatograph mass spectrometer as an electronic nose

Compared to sophisticated analytical chemistry, the claim of electronic noses is to be simple and to have a high throughout trough fast measurements. Therefore, gas chromatography equipment in order to approach the electronic nose philosophy, has to be used in the fast or ultrafast mode. In order to increase the separation speed during analysis, different parameters have to be adapted. This can be an increase of the carrier gas flow rate, an increase of the temperature-program heating rates, a reduction of the column length, a reduction of the column diameter, a reduction of the thickness of the stationary phase, and the use of a faster carrier gas. Depending on the sample, it is important to avoid using all these possibilities at once, because this always results in a resolution decrease, the sample capacity, or both. It is also important to note that these optimizations increase the demands on the detector technology used in terms of sensitivity, speed, and dead volume.

Even fast GC systems do exist, this technique still has to be more portable for onsite deployment. If the technological development can achieve miniaturization, GC or GC-MS micro instruments could easily replace traditional solid state electronic noses.

A short history of chromatography, the coupling with the mass spectrometry and recent advances in the field will be described in the following subchapters, as the technology is directly involved in the improvement of the mass spectrometer based electronic nose.

2.1.12.1. History of chromatography

Our world is one of complex mixtures. Petroleum is thought to contain over 100.000 components, and the number of different proteins in the human body are estimated at around the same number. Thus chromatography was born from the need to separate and analyze the mixtures of substances around us.

The history of chromatography spans from the mid-19th century until today, since progress on the technique is still evolving. The word chromatogram comes from the Greek words *chroma* = *color* and *graphein* = *to write*, which together, literally means "color writing". Even though the term originates from a study in the early 1900s by Russian botanist Mikhail Tsvet in separation of plant pigments, the earliest use of chromatography is attributed to the German chemist Friedrich Runge, who in 1855 described the use of paper in dye analysis by passing through it a mixed solution to create separation through absorption.

Chromatography methods changed little after Tsvet's work until the explosion of mid-20th century research in new techniques, particularly thanks to the work of A. J. P. Martin and R. L. M. Synge. They developed partition chromatography to separate chemicals with only slight differences in partition coefficients between two liquid solvents.

The invention of gas chromatography is generally attributed to A.T. James and A.J.P. Martin in their 1952 paper [70]. The novel aspect of Martin’s work was the employment of partition as the separation principle. In the decade after its discovery, gas chromatography showed rapid growth in biochemistry, food and flavor studies.

2.1.12.2. Modular description

The principal part of the gas chromatograph is the column, originally a tube packed with a solid support coated with the stationary liquid, but now a fine tube with the liquid coated on the inner surface. The carrier inert gas, which could be nitrogen, helium or hydrogen, passes from a cylinder through a pressure / flow-rate-controlling device to the sample injector at the column inlet. After introducing the mixture sample into the injector separation is achieved by a series of partitions between a moving gas phase and the stationary liquid phase. When the separated mixture components of the mixture emerge (‘’are eluted’’) from the column, they are detected by the measurement of some chemical or physical property. An important factor influencing column performance is its temperature and pressure. Gas chromatography can be applied to the analysis of mixtures, which contain compounds with boiling points from near zero to over 700 K, or which can be heated sufficiently without decomposition to give a vapor pressure of a few mmHg. Good temperature control is very important. This was originally achieved by enclosing the column in a vapor jacket or thermos tatted block, but now is done by placing the column in a hot-air oven. (**Figure 2.3**)

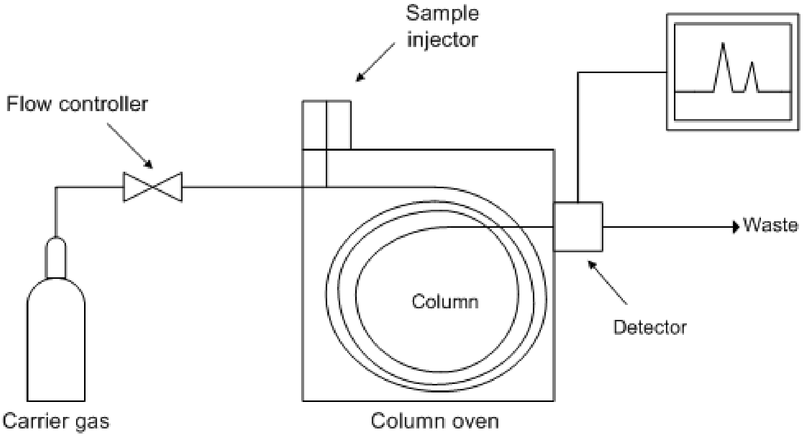


Figure 2.3: Diagram of a gas chromatograph

The *column* is at the centre of the analytical gas chromatograph, and the quality of the separation can be that of the column only. Early GC was carried out on packed columns filled with particles each of which was coated with a liquid or elastomeric stationary phase. The resolution of packed columns is limited by their length, itself restricted by the pressure drop consequent on the resistance to gas flow. This restriction was removed by the invention of the capillary column [71] which greatly increased separation efficiency, and by working at a lower temperature, give much better separation in equal times, or the same separation in shorter time (up to 10 times). On the other hand, because of the smaller amount of stationary phase, the capacity of capillary columns is limited. Also special sample-introduction methods and more sensitive detectors are needed.

The choice of the appropriate column for a given required separation depends on the chemical nature of the analyte, the sample matrix, and the solvent. Based on the nature of the molecular interactions between analyte and stationary phase the type of column must be chosen from the wide range available today. Non-polar interactions results in separations based on analyte volatility, while a polar stationary phase man induce a dipole moment resulting in increased retention and selectivity. A column coated with hydroxyl groups may separate analytes capable of hydrogen bonding. Shape-selective separations are also possible with chiral or liquid-crystal stationary phases. Other choices in capillary column selection include column internal diameter, film thickness and length.

A very important part of the gas chromatograph is the *sample introduction*. In the early days this was carried out by means of a syringe and a hypodermic needle. At first re-sealable rubber cap was employed, but this was replaced by a heat-resistant elastomeric septum compressed in a metal fitting. A procedure that has persisted until today. More recently spring controlled septum have been introduced, like the Merlin Microseal Septum from Agilent [72], which extends the seal life to up to a year and 5000 injections, resulting in time savings in changing septa incurring in less instrument downtime. Moreover, Chromatographic performance is improved due to less bleed and fewer ghost peaks, improving quantification and reliability. The device provides two distinct sealing mechanisms. The first is a double O-ring seal around the syringe that ensures gas leak integrity during the time of injection. There is no traditional septum to core or flake, responsible for bleeding and ghost peaks, and less force is required for the manual injection. The second seal is a spring-assisted duckbill that reliably maintains a high-pressure seal within the injection port at all times.

Because injection of a representative part of the sample as a narrow band in a quantity consistent with the capacity of the capillary column is often a limiting factor, a variety of injection methods have been developed. The basic splitter injector, which allows a small fraction of a rapidly volatilized sample to enter the column while the major portion is vented to waste, was introduced, which allowed for larger volumes to be injected. More specialized sample-introduction systems became available for capillary GC, including the programmed temperature vaporization (PTV) injector, and a variety of pyrolyzer systems [73]. PTV injectors can be used in split, splitless or direct mode. Here, the inlet temperature is maintained below the boiling points of solvent and solutes, but is rapidly programmed to vaporize each component in turn. Solutes are then exposed to less thermal stress and large volumes can also be injected. Low levels of volatile organics in environmental matrices may be analyzed by headspace, dynamic stripping or purge-and-trap sampling. In purge-and-trap, the sample is purged with helium and volatile analytes collected on a trap of adsorbent material from which they are released by rapid heating.

With the introduction of capillary columns, greater precision was required in the *pneumatic control systems*. Control of the gases required to run a GC has been through a combination of on-off valves. These have evolved together with the instrumentation, to the point that today we have total feedback controls to maintain constant flow rates of the carrier gases by monitoring the gas pressures and flow rates that, in turn, control electronic regulators.

Much of the flow regime in capillary instrumentation is closely associated with the requirements of sample injection and this is now possible through the automatic computer control of the pneumatics. As the instruments have evolved, a trend towards the use of keypads to set the conditions of the instrument is observed. And with the rapid spread of the modern PC, control and data acquisition programs tended to move on the PC. Now, it is possible with the automatic flow-control modules to function under mass control or pressure control of the carrier gas. This allows a choice of constant pressure, constant flow or even pressure programming.

The next important part after the column and the injector in the gas chromatograph is *the detector*. A number of detectors were employed in chromatography. The first gas chromatograms were generated by an automated titration system. But this was a rudimentary system, and conductivity measuring devices started to be used as detectors. The first one to be employed was the catharometer, which used the temperature (electrical resistance) charge of a filament. It responded to most analytes, but the requirements for trace analytes and the development of capillary columns resulted in new approaches. The most detectors are the flame ionization detector (FID) and the thermal conductivity detector (TCD). Both are sensitive to a wide range of components, and both work over a wide range of concentrations. While TCDs are essentially universal and can be used to detect any component other than the carrier gas (as long as their thermal conductivities are different from that of the carrier gas, at detector temperature), FIDs are sensitive primarily to hydrocarbons, and cannot detect water.

Pretty soon it became clear that the quantitative analysis of the gas chromatograph made the perfect combination with the qualitative analysis of the mass spectrometer, being the most effective technique for the analysis of complex mixtures. At first, the high flow rates of packed chromatographic columns were incompatible with the vacuum requirements of the mass spectrometer and hindered the progress of the GC-MS. However, with the introduction of capillary columns and their smaller flow rates made the GC-MS a viable technique.

In all types of GC-MS, as in other GC ionization detectors, ions are produced by electron or chemical ionization but are now sorted according to molecular weight (or mass-to-charge ratio) by one of a range of analyzers: magnetic sector, quadrupole, or time-of-flight. The resulting mass spectrum is related to the molecular weight and structure of the analyte, and allows identification through comparison with a library, or through a-priori interpretation. The MS detector can be operated in either scanning mode or in selected-ion monitoring mode. Thus, a range of m/z values may be scanned (say

35–300) in a bench-top instrument in time intervals as short as 0.5 to 1 s. A computer is then used to plot either the total-ion current (analog to the response of a universal thermal conductivity detector) or individual ion currents (‘‘mass chromatograms’’), which may be specific to selected compound types. Much greater sensitivity can be achieved by monitoring only one ion or a few ions.

As a performance increase of the mass detector, time-of-flight mass detectors were proposed by Martin [74]. By accelerating the ions in an electric field of known strength all the ions with the same charge acquire the same kinetic energy. The velocity of the ion depends on the mass-to-charge ratio (heavier particles reach lower speeds), and by measuring the time that takes for each particle to reach a detector one can find the mass-to-charge ratio of the ion. This results in accurate mass measurements (in contrast to the unit m/z resolution of ‘‘bench-top’’ MS), which allow the molecular formulae of ions to be determined, and rapid rates of accumulation of spectra (up to 500 Hz), which allow GC peaks with widths as narrow as 12 ms to be identified [75].

2.1.12.3. Future portable GC-MS systems as portable electronic noses

Gas chromatography expanded with great speed over the two decades following its invention in 1952, and much current practice has its roots in that period. The introduction of robust, efficient and reproducible fused-silica capillary columns and the availability of relatively inexpensive but reliable equipment for GC-MS provided a crucial stimulus in the 1980s. High-speed GC, time-of-flight MS, and even comprehensive GCxGC now promise further expansions. The versatility of modern GC will expand its application areas. In particular, chromatography has a vital role to play in process control in chemical industry, and on-line and at-line GC methods are replacing analysis in a central laboratory.

As mentioned before, if fast and ultrafast chromatograms are within reach with the modification of the analysis parameters, the miniaturization is not that easily achievable.

Through miniaturization the gas chromatograph, as the mass spectrometer have the potential to be used fully as electronic noses, with on-field deployment, not being grounded on the laboratory bench. James Lovelock made the first known reference to a GC-on-a-chip verbally in the early 1970s. It was made on a magnesium oxide chip, less than 50 mm x 20 mm. It used electrolytic and coulometric methods to produce the gas. The Stanford GC designed by Terry [76] was also reported in the 1970s. It was a complete, working GC on a silicon wafer (5 cm in diameter), including column, injector and detector. A simple separation of gaseous hydrocarbon mixtures using this solid-state chromatograph involved injections of 4nL interval volume and took place in a 2 ms time period. The separation was performed in 5 s, with observed retention times of 2.6 s for the unretained component, nitrogen. The detector, a TCD, was based on a nickel film resistor and was batch-fabricated on a separate silicon wafer and then mounted by mechanical clamping on the wafer containing the column. Because capillary technology was still in its infancy and the lack of technological experience to use the device, the attempts to commercialize it failed.

One of the first attempts to base an electronic nose using chromatography technology was presented by Zampolli [77]. The selective hybrid microsystem consists of a zero grade air unit, a commercial minipump, a minivalve, a silicon micromachined packed GC column, and an MOX sensor as the detector. The analysis time of a certain mixture of volatiles depends on the type of stationary phase, gas flow rate, column length, and temperature of the GC column. The authors have shown that within 15 min the complete separation of benzene, toluene, and *m*-xylene is possible, with an excellent detection limit, well below the expected indoor threshold values.

The use of chromatographic separation techniques together with micromachining technology and miniature mass spectrometers will allow developing handheld portable devices for selective VOC monitoring. The power of the MS-based electronic nose applications would no longer be confined to the laboratory, but on site implementations would be possible, and by coupling the micro GC with the miniaturized MS, the power of the electronic nose will grow even more.

2.2. Data analysis

The third part of the electronic nose is the data processing system. This is considered to be the most important part of the electronic nose. In fact, an essential step in modern chemical analysis is pattern recognition. Together with the progress in electronics, the high performance attained by statistical programs helped to made possible the introduction of the artificial olfaction concept.

Pattern recognition methods are most widely classified by means of their learning approach (either supervised or unsupervised) even though they can also be classified in terms of linearity or parametric basis, and also by the dimensionality of data they deal with.

During supervised learning the data vectors are tagged with a descriptor; the classes are learned and grouped according to their description. After learning is complete an unknown vector may be classified using the relationships learnt a priori from the known vectors. In unsupervised learning there are no descriptors, the measurements are categorized based on a predefined similarity index.

Statistical methods are parametric as they assume the data may be described in terms of probability density functions. Examples of these methods would include Principal Component Analysis (PCA), hierarchical cluster analysis (HCA), principal component regression (PCR), discriminant factor analysis (DFA), analysis of variance between groups (ANOVA), partial least squares regression (PLS) , principal components regression (PCR), soft independent modeling class analogy (SIMCA) and clustering algorithms such as k-means.

Artificial intelligence (AI) techniques are generally non-conventional, intuitive approaches for problem solving, often biologically inspired and may be split into three sub-groups:

1. *Artificial neural networks* (ANN), which include multi layer perceptrons (MLP), radial basis function networks (RBF), self organizing maps (SOM), learning vector quantization (LVQ) and self organizing competitive

systems such as the adaptive resonance theory (ART) or the family growing cell of algorithms.

2. *Fuzzy logic* and fuzzy rules based algorithms
3. *Genetic Algorithms* (GA) used for feature selection

Based on the dimensionality of the multivariate data to be analyzed, we can classify the pattern recognition methods into two-way data analysis (PCA, PLS, etc.), three-way data analysis (PARAFAC, n-PLS, Tucker) and even multi-way data analysis, where the data exceeds three dimensions.

In this section we are going to take a look at the types of data generated by the mass spectra and the gas chromatograph-mass spectra, the preprocessing techniques required for a proper data analysis and the data analysis techniques used for each of them

2.2.1. Types of data

A distinction should be made between one-way, two-way, three-way, and multi-way data. If a single measurement on a chemical system generates a single number and this measurement is repeated, then a sequence of numbers is the result. For example a temperature measurement, repeated 3 times gives the temperatures t_1, t_2 and t_3. This data set is a sequence of numbers that has to be analyzed with one-way tools, like calculating the mean, median, standard deviation, variance, etc.

Going up one level, we find instruments that generate sequence of numbers, like ultraviolet-visible spectroscopy, infrared spectroscopy, chromatography with univariate detectors like flame ionization detection (FID), and many more. By measuring one sample with this type of instruments we obtain a vector. A measurement form this type of instrument is a vector. Taking I of these measurements results in a matrix $\boldsymbol{X}$ of I x J. This is a two-way array, and such a matrix can be analyzed with two-way analysis tools like principal component analysis (PCA), Partial Least Squares Regression (PLS), etc.

If a single measurement generates a table of numbers for each sample we have instruments like a fluorescence emission-extraction landscape, or a chromatogram where the detector is a mass spectrometer. This results in a matrix $\boldsymbol{X}$ (J x K), where J is the time axis of the chromatographic separation and K is the mass fragment axis. The matrix can be seen in any of the two ways, as a series of mass spectra for each point of the chromatogram, or as series of chromatograms for each mass-charge ratio. Taking I of these measurements generates a three-way array of size I x J x K. Such arrays can be analyzed with three-way analysis methods, like PARAFAC, n-PLS, etc.

It is even possible to generate higher-way data, like four-way (GCxGC-MS), five-way data (GCxGCxGC-MS), or even higher, but these instruments are less common.

With the advance in analytical instruments and the arrival of hyphenated techniques the dimensionality of the provided data has increased form univariate data (a single observation: temperature measuring, pH analysis, etc) to multivariate data, in which more than one reading is collected at one time. Multivariate data, based on dimensionality, can be classified into two-way (a chromatogram, a mass spectra, etc.), three-way (a 3D data from a gas chromatograph-mass spectrometer) or multi-way data (the output of a GCxGC-MS)(see **Table 2.8**).

Due to the more complex nature of the three-way data, which is used very often in this work, compared to the two-way data , some notations and terminology used in multi-way analysis need to be discussed.

Kiers [78] describes the first suggestion for a standardized notation for multi-way analysis not only applicable in chemistry, but in data analysis in general. It is advised to use these suggestions for notation in order to make communication across and within different disciplines more straightforward.

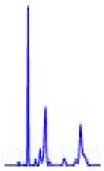
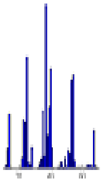
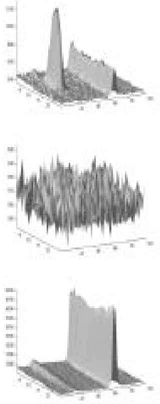
Data type	**Univariate**	**Multivariate**			
		Two-way		**Three-way**	**Multi-way**
Instrument	Thermometer, pH, etc.	GC	MS	GC-MS	GCxGC-MS
Example					
Dimen-sions	One observation	Reten-tion time	Mass channel	Retention time x Mass channel	Ret. time 1 x Ret. time 2 x Mass channel
Sample size	Scalar	Vector (N=1)		Matrix (N=2)	Three-way array (N=3)
Dataset size	Vector	Matrix		Three-way	(N+1)-way

Table 2.8: Types of data

Multi-way arrays, also referred to as tensors, are higher-order generalizations of vectors and matrices. Higher-order arrays are represented $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, where the order of $\underline{\mathbf{X}}$ is N (N>2) while a vector and a matrix are arrays of orders 1 and 2, respectively. Higher-order arrays have a different terminology compared to two-way data sets. Each dimension of a multi-way array is called a mode introduced in [79] and [80] (or way), and the number of variables in each mode is used to indicate the dimensionality of a mode. For instance, $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is a multi-way array with N modes (called N-way array or Nth-order tensor) with I_1, I_2, . . . , I_N dimensions in the first, second, and Nth mode, respectively. Each entry of $\underline{\mathbf{X}}$ is denoted by $x_{i1\ i2\ \dots\ iN}$. For a special case, where N = 3, let $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a three-way array. Then, $x_{i1\ i2\ i3}$ denotes the entry in the i_1^{th} row, i_2^{th} column, and i_3^{th} tube of $\underline{\mathbf{X}}$ (**Figure 2.4**. A, B, C) [81]. When an index is fixed in one of the modes and the indices vary in the two other modes, this data partition is called a slice(or a slab). For example, when the i_1^{th} row of $\underline{\mathbf{X}}$ is fixed, then it is a horizontal slice of size I_2 x I_3 (**Figure 2.4** D).

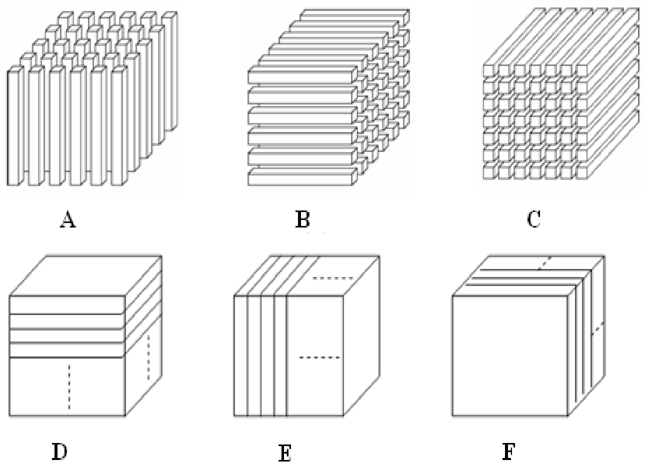


Figure 2.4: Nomenclature in 3D matrices: (A) Columns, (B) Rows, (C) Tubes, (D) Horizontal Slices, (E) Vertical Slices, (F) Frontal Slices

We denote higher-order arrays using underlined boldface letters, e.g., $\underline{\mathbf{X}}$, following the standard notation in [78]. Matrices and vectors are represented by boldface capital, e.g., $\mathbf{X}$, and boldface lowercase letters, e.g., x, respectively. Scalars are denoted by italic letters, x or X. Matrix and tensor entries are represented by lowercase letters with subscripts, e.g., x_{ij} or x_{ijk}.

2.2.2. Data preprocessing

Preprocessing the data prior to the analysis is often a very important step. In fact sometimes proper preprocessing can make the difference between a useful model and no model at all. The purpose of preprocessing is to try to transform the data into the most suitable form for the analysis. Sometimes there may be background effects we wish to remove by some suitable transformation, or the variables may be measured in different units that require scaling to obtain equally important mathematical values.

Once a chromatographic data set has been collected the first step is always to examine the data. The primary purpose of this step is to use the eye to look for the obvious features and errors. Errors can occur in both the measured variables and particular values. This initial view of the data may indicate the need for preprocessing.

2.2.2.1. Pretreatment of chromatographic data

Despite all the advanced technology, chromatograms present today the same drawbacks as half a century ago. Chromatograms still need baseline correction; peaks suffer from drift, and present noise.

Analyzing a chromatogram, wherever is measured with a univariate detector (FID) or a total ion chromatogram from a GC-MS a chromatogram can be divided into three constituting parts, which by summing will result in the overall chromatographic signal (**Figure 2.5**):

a. the relevant signal
b. the background baseline
c. the noise

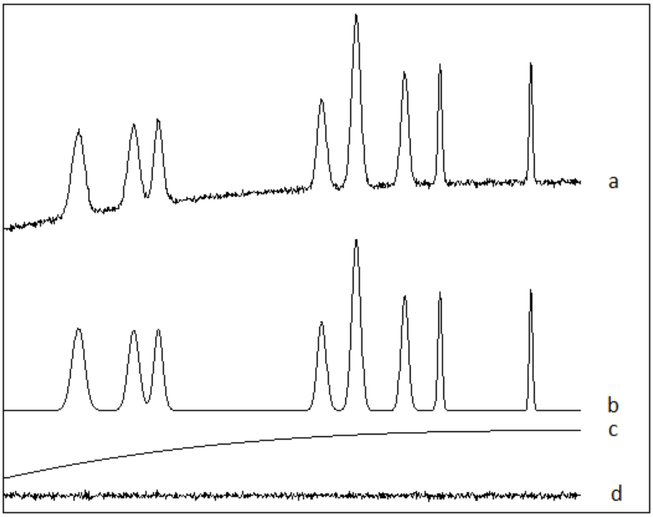


Figure 2.5: Components of a chromatogram:
a) overall signal, b) relevant signal, c) baseline and d) noise

There is a tendency to use the entire chromatographic signal for data analysis, instead of using only the integrated peaks of interest. The chromatographic profile is then considered a fingerprint that contains features unique to a given sample. The idea is to include as much information as possible instead of setting pre-established criteria for where in the chromatogram the information is placed. But prior to chemometric data analysis, some important issues of the chromatographic signal must be taken into consideration. Peak alignment, overlapped peaks, baseline correction and in some cases signal enhancement are the most important.

Signal enhancement techniques such as noise filtering, smoothing and other techniques are performed to increase the signal-to-noise ratio for the peaks of interest and to remove the noise. This may confirm the presence of a peak in a noisy baseline, but it will not improve the precision or accuracy of peak integration. Although signal enhancement techniques are efficient, experimental conditions should always be considered as a first attempt to eliminate excessive noise.

Several signal enhancement techniques are available and they can be divided into digital filtering in the frequency domain and in time domain. Filtering the signal in the frequency domain separates noise from signal by dividing the overall signal into high frequency components and low frequency components. As the noise has a much higher frequency than the signal the use of a low-pass filter allows the separation of noise from the usefull signal. For the time domain the filters usually work within a finite-width moving window of the chromatogram. The most widely used technique in analytical chemistry is the polynomial filter suggested by Savitzky and Golay [82; 83], which uses a least squares fit of a polynomial of a given order to a certain window size in the chromatogram.

Baseline correction has been an issue in chromatography for decades. Several baseline methods are available in the literature and one of the first descriptions of the importance of removing baseline drifts was presented by Wilson [84]. They used a simple approach of integrating the area below the peak profile and then subtracting this from the overall area (peak plus baseline). Nowadays most methods are based on subtracting a fitted polynomial following the baseline curvature or by separating baseline and signal in a low-rank factor model.

One of the most cited baseline correction tools is the asymmetric least squares smoothing by Eilers [85]. This method works by fitting an initial polynomial of a certain order to all data points in the chromatogram. By iteratively penalizing (weighting) positive deviations (signal above the fitted polynomial) more than negative deviations (lower intensity signal plus baseline points) the polynomial will at some point approximate the baseline within a predefined limit and the resulting polynomial can be subtracted from the overall signal.

The problem of *overlapped peaks* is one of the most encountered in chromatography, and there are two ways to approach it [86]. The first is related to the chromatographic instrument, and involves the chromatographic separation parameters, like column characteristics, temperature ramp, mobile / stationary phase, flow rate, etc, or the addition of a second column in a GC x GC system. And the second approach is a mathematical one. Software can be used to resolve the overlapping peaks into pure peak profiles. This method is often called peak deconvolution. Based on the chromatogram type we can treat the data in two ways.

If the chromatographic detector is a univariate detector and the chromatogram is a vector, a one-dimensional sample, no additional information about the peak constituents is available. In this case a peak fitting approach can be used. A Gaussian shape is assumed for each peak, and by changing the parameters for the involved peak functions, the overall signal can be approximated by fitting an experimental chromatogram to a linear combination of individual chromatographic peaks. Thus a mathematical peak model is needed to describe each elementary peak.

On the other hand if the chromatographic detector is multivariate, and each peak is characterized by an additional multivariate dimension, then the additional information about each analyte can be used to deconvolute the overlapping peak.

One other major problem in chromatography is *elution time variations*, which are caused by unavoidable variations in the instrument parameters. It is thus necessary to align the chromatographic data in the time dimension so that the same peak is located at the same elution time for all samples. Analyzing individual samples can be achieved by using a proper indexing system, but when the samples are stacked to give higher order arrays this is no longer a trivial problem.

In order to perform a chromatographic alignment in a data set five things must be considered:

I. The type of data
II. Transformation of separation time axis
III. Alignment quality
IV. Optimization of alignment parameters
V. A reference chromatogram

I. Depending on the data type (two-way or three-way) the extra information from a multivariate detector can be used actively to guide the peak alignment.

II. Elution time variations often need only a constant, linear shift correction by stretching or shrinking of the whole elution time axis, or simply a movement of the whole chromatogram a certain integer sideways for proper alignment [87]. This is known as a systematic shift. However, if the column is changed between runs, if different chromatographic columns are used or if samples are measured over a long time then a more complex shift correction is needed. This unsystematic shift is characterized by a different degree of shift for multiple peaks across samples and can be seen as peaks shifting independent of one another in the same chromatogram.

III. Quality of alignment is judged by several methods, like the sum of distances between paired features, the Pearson's correlation coefficient [88; 89], the Euclidean distance for the similarity of two chromatographic signals [90]

IV. Most alignment methods optimize the alignment by the quality of alignment. In order to keep the computation time to a minimum some restrictions are imposed.

V. Choosing the reference chromatogram is a very important step of the alignment method. A poorly chosen reference sample can lead to a failure. Suggestions for peaking the reference are given by Daszykowski and Skov [91, 92]. Among these is the average chromatogram, the first loading of the PCA model. However, the choice depends on the homogeneity of the samples, on the degree of missing peaks across chromatograms and many other things.

Several alignment techniques for fingerprint chromatographic data are presented in the literature: Correlation Optimized Warping (COW), Dynamic Time Warping (DTW) [89], Parametric Time Warping (PTW) and Semi-parametric Time Warping (STW) [93] and Target Peak Alignment (TPA) [94]. These algorithms differ in four aspects:

A. the way the warping function/path is defined (non-parametric or parametric),
B. whether landmarks are used to guide the alignment,
C. the similarity measure that must be optimized (e.g. Euclidean distance, correlations coefficient, sum of squares of the difference between sample and reference chromatogram)
D. the algorithmic technique that is used to find the optimal warping function or path

If, on the other hand, the chromatogram comes from a hyphenated technique (GC-MS landscape) additional information is available. This can be used actively or passively in the alignment.

Passive means that the spectral dimension is summed and vector based alignment techniques can be used. Doing this it is assumed that no shifts are observed in the spectral dimension, which is a valid assumption for most desktop analytical low-resolution instruments [95].

Active means that the spectral dimension is used when aligning, either as an additional tool in vector based alignment or included in the alignment quality measure. In some alignment methods the spectral information can be used to select similar features from reference and sample chromatogram, but the alignment is still performed on the vectorized signal. Comparing the mass spectrum at certain times does this. Krebs [96] presented a method for alignment of GC–MS data that identifies landmarks (peaks in the TIC chromatogram that are above a selected threshold) in the data, comparing them across samples, and aligning only those determined to be the same by a high correlation of the peaks in the m/z dimension. The landmarks are defined as peaks above a threshold, and match these landmarks between two measurements if the correlation score of the corresponding mass spectra exceeds 0.99. In another approach, by Gong [97], each peak in individual sample chromatograms was resolved into pure chromatogram and spectra, providing qualitative and quantitative measures of each peak in each sample. By selecting marker compounds known to be a specific chemical compound present in all samples they used these actively in the alignment by interpolation between these landmarks. Xu [98] used a similar approach based on local factor analyses to find similar target peaks between chromatograms. Dividing the chromatogram into segments they aligned the most intense features in these segments.

2.2.2.2. Normalization

Normalization of a sample vector is done by dividing each variable by a constant, and different constants can be used.

The reason behind normalization is to remove systematic variation, thus is associated with the total amount of the sample. In mass spectrometry it is used to normalize to the largest m/z peak by dividing the whole spectra to the maximum number. In chromatography normalizing the entire chromatogram can remove the effect of variable injection volume. Usually an internal standard is used, and the normalization is done by dividing each sample chromatogram to a constant, the area of the internal standard peak.

Depending on the constant used in the normalization several results can be obtained. For instance by normalizing by the sum of all the elements gives normalization to unit area. Normalizing to the unit length is done by dividing each element to the square root of the sum of all the squared values in the vector.

Depending of what systematic variation is removed one normalization constant may be more appropriate than the other. Care should be taken in order not to remove important information like concentration.

Two-way data normalization is very similar to three-way and even multi-way normalization. If in the two-way case each element of the sample vector was divided by a constant, for the three-way case each element of the horizontal slab (a sample matrix) is divided by the chosen constant. Even though visualizing can be difficult, for the multi-way case normalization is straightforward. the important part is choosing the constant, and knowing which systematic variation we want to suppress.

2.2.2.3. Centering and scaling

Centering and scaling the data are preprocessing tools that operate on each variable. Their role is different though. While centering is applied to account for an offset in the data, scaling is used to emphasize some variables over others to increase their influence on the model. Usually the emphasized variables are the ones whose scale is somewhat lower, and thus this process makes all the variable variances comparable [99].

Mean centering a variable is accomplished by subtracting the mean of that variable vector from all of its elements. Performing the mean centering over multiple variables results in the removal of the mean vector from all sample vectors in the data set. In the mean centering of a chromatogram, for example, the mean chromatogram is removed from each sample (chromatogram) and thus all that remains is the relative differences in intensity, making them easier to discern.

Mean centering generally does not hurt and often helps, therefore is usually used as a default.

Scaling is a preprocessing technique by variable weighting. This is done by multiplying all elements in a variable vector by a number (weight). If the weights are chosen wisely, this step can improve the analysis results. There are two types of weighting. One emphasizes some variables over others and the other one does the contrary by placing all the variables on equal footing.

Among the weighting methods that emphasize some variables over others we briefly remind *a priori information weighting* and *variable selection*. The first case should be driven by accurate a priori information, which may come from theory, prior experience or experimentation. The case of variable selection on the other hand emphasizes some variables over other by using a weight of zero. The selection of variables can be driven by statistical methodology or chemical information.

On the other hand, when variables need to be given the same importance *variance scaling* comes into play. This is achieved by dividing each element in a variable vector by the standard deviation of that variable. The primary reason for variance scaling is to remove artificially imposed weighting by the scale of the variables. This influence is arbitrary if the units of the variables are different.

In practice variance scaling is often performed in conjunction with mean centering and is called *autoscaling*. Interpreting autoscaled data may be difficult because the units have been removed. Even though comfortable seeing patterns are not there anymore, this preprocessing step can in fact improve the analysis.

Bro and Smilde [100] show that often when centering is performed subjective and qualitative reasons are given, thus situations may occur where subtraction averages does not work and will lead to models that fit the original data more poorly than if the data had not been preprocessed. Basically centering should be performed only if there are common offsets in the data or if modeling such offset provides a reasonable model. Thus centering is performed to make interval-scaled data behave as ratio-scale data. Centering should make a difference, and it should manifest as:

a. reduced rank of the model
b. increase fit of data
c. specific removal of offset
d. avoidance of numerical problems

Centering is most conveniently seen as a projection step for removing offsets, which for bilinear models can be of two types: constant across one mode or constant across both modes. Centering across one mode is called *single centering*, while performing a centering across the first mode and then another centering across the second mode is called *double centering*. Slab centering refers to centering by subtracting from each horizontal slab in the three-way array the overall average of that slab.

For scaling another terminology is used. If a matrix is scaled such that each row is multiplied by a scalar, the we call this *scaling within the first mode*. If each column is multiplied by a certain scalar as in auto scaling it is referred as *scaling within the second mode*.

Unlike centering, scaling does not change the structure of the model. Scaling is used to change the weights given to different parts of the data in fitting the model. Although scaling is important it usually has less dramatic influence on the fitted model than centering.

Scaling is used for several reasons, from which we highlight:

a. to adjust the scale differences
b. to allow for different size of subsets of data (block scaling)

When scaling within several modes the desired situation becomes more complicated because scaling one mode affects the scale of the other mode. For example, scaling to a standard deviation of one within both the first and second modes will not be possible, not even using iterative scaling. If scaling to mean squares of one are desired within both modes, this has to be done iteratively until convergence. Using mean squares instead of standard deviation for scaling has the property that iterative scaling is guaranteed to converge when no centering is included. The purpose of scaling is mainly to bring the levels of variation of variables to equivalent levels.

The interdependence of centering and scaling is a complicated issue in preprocessing, where few standardized ways exist in two-way analysis. It is important to be aware that not all combinations of centering and scaling will work. Generally only centering across both modes is straightforward, or scaling with one mode combined with centering across the other mode (like in auto scaling).

Proper centering will correctly remove the presupposed offsets and will not introduce other offsets into the data. Likewise proper scaling introduces the correct weights into the loss function.

Different words are used for centering (across) and scaling (within). This is because Centering is performed across a mode in the sense that one offset is subtracted from every element in a specific vector, the data is centered across the elements of one mode, while scaling is performed by multiplying all elements in the array by the same scalar. For three-way data centering works by subtracting the average value from each element of the slab, while scaling is done by multiplying the whole slab by the same scalar. Thus scaling is performed within the elements of one mode.

The particularity of preprocessing of three-way arrays by centering and scaling rise from the fact that centering has to be performed across a specific mode, and scaling has to be done by transformation within a specific mode. If the basis of two-way preprocessing is understood, three-way preprocessing is simple. The difficulties in three-way preprocessing arise because of the problems mentioned above. Also because of three-way data is often rearranged by unfolding to two-way arrays preprocessing problems are enhanced if applied to the matricide data.

For a three-way array offsets may occur in three different ways: constant across all modes, constant across two modes and constant across one mode only. Regardless of the offset structure, by centering the data must be preprocessed so that the offsets are projected onto the nullspace of vectors of ones in a particular mode.

As for two-way case, only single centering leads to the properties sought in centering (removal of offsets). The other types of centering, such as subtracting the overall mean, will introduce artifacts that have to be modeled additionally.

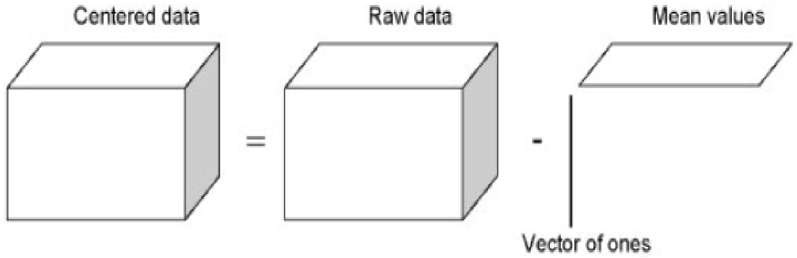


Figure 2.6: Centering across the first mode in three-way arrays. For each column of the raw data a mean value is calculated and subtracted from each element in the column. A two-way matrix of mean values is obtained

As in the two-way case, scaling is a transformation of a particular object. For the three-way arrays whole matrices, rather than columns, have to be scaled by the same value.

Scaling has to be applied by transforming the data within a given mode. It is not appropriate to scale an array within two combined modes, which can happen when autoscaling an unfolded array.

2.2.2.4. Data arrangement

A three-way data can be analyzed with multi-way and two-way data analysis techniques.

The simplest way to look at information locked up in a three-way array is to transform it in order to be analyzed by traditional two-way methods. This can be done by unfolding, summation, or concatenation.

Unfolding (matricization, flattening) means transforming a third- or higher-order array into a two-way data set and has multiple definitions in the literature [101]. Two different definitions of unfolding in the first mode are illustrated in **Figure 2.7**. Once a three-way array is flattened and arranged as a two-way data set, two-way analysis methods can be employed in understanding the structure in data. For instance the signal from a gas chromatograph-mass spectrometer can be unfolded in 2 different ways. First we can matricide the data by pasting one mass spectra one after the other for each point obtained in the retention time. And secondly we can flatten the data by pasting one after the other chromatograms for each mass-charge ratio. Tucker1 [102] is one such multi-way method that that is based on matricization the multi-way array.

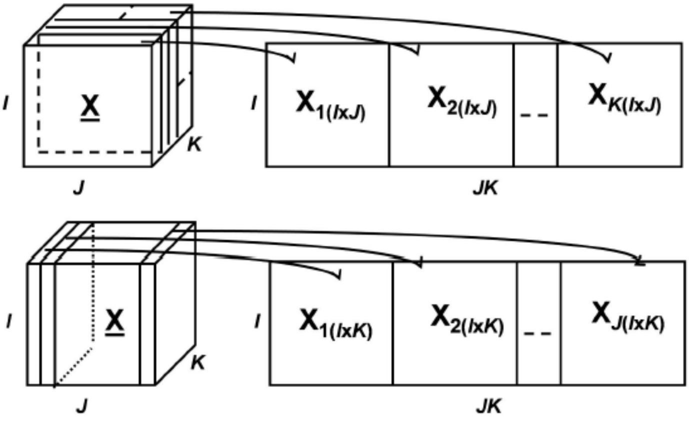


Figure 2.7: Matricization of a three-way array in the first mode

Another way to analyze the data is by adding one dimension of the three dimensional data and "collapsing" the respective axis, thus obtaining a two dimensional matrix. This can also be done in two ways, by summing each of the variable axis (J and K) independently (**Figure 2.8** A and B). Taking the example of the data from the gas chromatograph-mass spectrometer if we add the time dimension of the chromatographic separation we are going to get the average mass spectra. On the other hand if we add the m/z dimension we are going to obtain the total ion chromatogram (TIC).

Compared to unfolding the data, these methods of summation show some advantages and also some drawbacks. On one size of the data is much reduced, making the processing easier, but on the other hand the information brought by the extra dimension is lost in the process.

Another way to transform the three-way data in order to be analyzed by classic two-way methods is a compromise, a middle way between unfolding and summation. This method takes into account both dimensions by concatenation of the two matrices resulted by summing the variable dimension of the three-way array (**Figure 2.8** C).

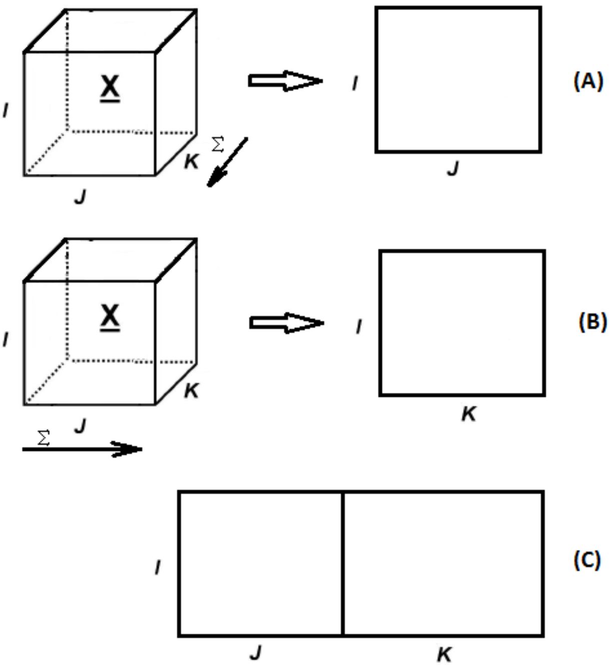


Figure 2.8: Data reduction of a three-way array by (A) summing the K variables, (B) summing the J variables, and (C) concatenation of the two matrices obtained by summation

Even though rearranging multi-way arrays as two-way data sets and analyzing them with two-way methods may prove easier, this usually results in information loss and misinterpretation especially if the data are noisy.

Thus, multi-way models are more advantageous in terms of interpretation and robustness to noise compared to two-way models. In addition to these, some multi-way models such as PARAFAC are also unique under mild conditions given by the well-known result of Kruskal [103] in contrast to two-way factor models suffering from rotational ambiguity. Finally, it is always desirable to choose the simplest model for the data and multi-way models may be the simplest possible models.

2.2.3. Pattern recognition

2.2.3.1. Two-way pattern recognition

Principal component analysis (PCA), the basic "work horse" of multivariate data analysis, is a well-known multivariate analysis technique aiming at summarizing data reducing the original dimensions. Its focus is as much on the interrelationships among variables as on the observations [104]. The idea of PCA is based on the assumption that the direction of the largest variance in the data carries most of the information. Therefore the n-dimensional data space is rotated in such a way that the directions of the largest variance become the coordinate axes of the data space. The resulting new axes (principal components) are normal to each other and are sorted by decreasing variance. Thus the first principal component shows the maximum variance in the data.

PCA is performed with no information on the classification of samples, being based solo on the variance of the data-set. Therefore it is an unsupervised method.

Mathematically, PCA is a process that decomposes the covariance matrix of a matrix into two parts: eigenvectors (score matrix) and eigenvalues (loadings matrix), whereas Singular Value Decomposition (SVD) decomposes a matrix per se into three parts: singular values, column eigenvectors, and row eigenvectors. The relationships between PCA and SVD lie in that the eigenvalues are the square of the singular values and the column vectors are the same for both. An extra matrix is collected, the residual matrix, which is the difference between the original and the reconstructed matrix of the calculated model.

The loadings (**V**) and scores (**U**) matrices are orthogonal. The loadings can be understood as the weights for each original variable when calculating the principal component, whereas the score matrix contains the original data in a rotated coordinate system (**Figure 2.9**). Another two properties of the principal components (PC) is that they are hierarchical and sequential, which means that the first PC holds the maximum information of the data, followed by the second

PC which retains the maximum information that was not included in the first PC and so on until the last PC.

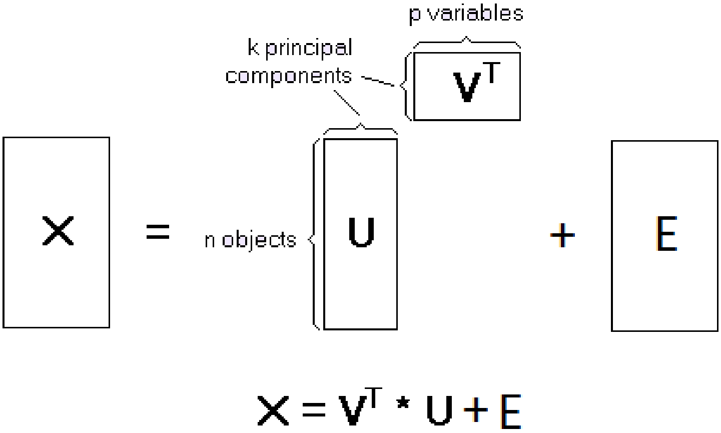


Figure 2.9: PCA model - graphical explanation

It should be emphasized that PCA is a visualization and a reduction tool, and it should not be used as an objective classifier.

The other most used multivariate data analysis tool is Partial Least Square Regression (PLS). PLS regression is a technique that generalizes and combines features from principal component analysis and multiple regression. It is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (predictors). It originated in the social sciences (economy) [105] but became popular first in chemometrics due in part to Herman's son Svante [106], and in sensory evaluation [107]. It was first presented as an algorithm akin to the power method (used for computing eigenvectors) but was rapidly interpreted in a statistical framework [108].

Research in science and engineering often involves using controllable variables (factors) to explain, regulate, or predict the behavior of other variables (responses). When the factors are few in number, are not significantly redundant (collinear), and have a well-understood relationship to the responses, then

Multiple Linear Regression (MLR) can be a good way to turn data into information. However, if any of these three conditions breaks down, MLR can be inefficient or inappropriate.

MLR can be used with many factors. However, if the number of factors get larger than the number of observations, the model is likely to fit the sampled data perfectly but fail to predict new data correctly. This phenomenon is called over-fitting. In such cases, although there are many manifest factors, there may be only a few underlying or latent factors that account for most of the variation in the response. The general idea of PLS is to try to extract these latent factors, accounting for as much of the manifest factor variation as possible while modeling the responses well. For this reason, the acronym PLS has also been taken to mean ‘‘projection to latent structure.’’

PLS is a method for constructing predictive models when the factors are many and highly collinear. The emphasis is on predicting the responses and not necessarily on trying to understand the underlying relationship between the variables. For example, PLS is not usually appropriate for screening out factors that have a negligible effect on the response. However, when prediction is the goal and there is no practical need to limit the number of measured factors, PLS can be a useful tool. In other words, the PLS model fitness is somewhat sacrificed in favor of its prediction ability.

The PLS model can be viewed as consisting of outer relations of individual X and Y blocks and an inner relation linking both X and Y. This can be represented graphically as in **Figure 2.10**. The goal is to minimize the error matrix F while maintaining the correlation between X and Y through the inner relation U = BT.

The matrix X is decomposed into a scores matrix, T, loadings matrix, P', and an error matrix, E. The matrix, Y, is decomposed into a scores matrix, U, loadings matrix, Q', and an error matrix, F. The extracted factors, T, are used to predict the Y scores, U, and these predictions are then used to construct predictions for the responses.

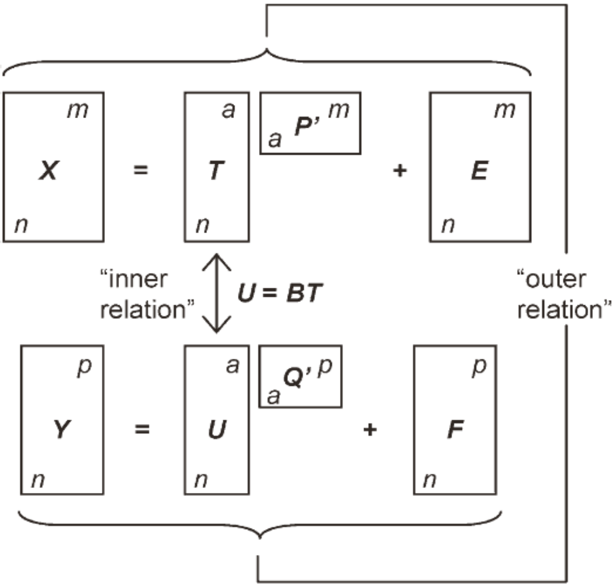


Figure 2.10: PLS model consists of outer relations between X and Y and an inner relation both blocks

2.2.3.2. Multi-way pattern recognition

Multi-way data analysis, dating back to 1920s to the studies of tensor decompositions by Hitchcock [109], [110], is the extension of two-way data analysis to higher-order data sets. Multi-way analysis is often used for extracting hidden structures and capturing underlying correlations between variables in a multi-way array.

It has been shown in numerous research areas, including social networks [111], neuroscience [112] and process analysis [113] that underlying information content of the data may not be captured accurately or identified uniquely by two-way data analysis. Two-way analysis methods, by which we refer to those based on factor models here, suffer from rotational freedom unless specific constraints such as statistical independence and orthogonality are enforced. On the other hand, these constraints requiring prior knowledge or unrealistic assumptions are not often necessary for multi-way models. For example, in fluorescence spectroscopy, one of the most common multi-way models, i.e., Parallel Factor Analysis (PARAFAC), can uniquely identify the pure spectra of chemicals from measurements of mixtures of chemicals. Consequently, multi-way analysis with advantages over two-way analysis in terms of uniqueness as well as robustness to noise and ease of interpretation has been a popular exploratory analysis tool in a variety of application areas, which we discuss throughout this survey.

The most well known multi-way models in literature are Tucker models and the PARAFAC model, which is also called Canonical Decomposition (CANDECOMP) [114]. This section will briefly describe these models as well as recent models built on the principles of PARAFAC and Tucker , which have relaxed the restrictions enforced to capture data-specific structures.

2.2.3.2.1. PARAFAC family

PARAFAC [115], which has been originally introduced as the polyadic form of a tensor in [116], is an extension of bilinear factor models to multilinear data. PARAFAC is based on Cattell's principle of Parallel Proportional Profiles [117]. The idea behind Parallel Proportional Profiles is that if the same factors are present in two samples under different conditions, then each factor in the first sample is expected to have the same pattern in the second sample but these patterns will be scaled depending on the conditions. Mathematically, a PARAFAC model can be represented as the decomposition of a tensor as a linear combination of rank-one tensors.

The motivation behind PARAFAC is to obtain a unique solution such that component matrices are determined uniquely up to a permutation (rank-one tensors can be arbitrarily reordered), and scaling of columns. It is this uniqueness property that makes PARAFAC a popular technique in various fields. For example, in fluorescence spectroscopic data analysis [118], a unique PARAFAC model allows us to find physically and chemically meaningful factors directly from measurements of mixtures of chemicals. Uniqueness is achieved by the restrictions imposed by the model. The most significant restriction is that factors in different modes can only interact factor wise. The interaction between factors in different modes is represented by a core array in multi-way models.

PARAFAC can be considered as a constrained version of Tucker3 model (the number of factors in all modes is equal and the diagonal elements of the core matrix are equal to 1). The decomposition of three-way data is made into triads or tri-linear components (multi-linear if N >3), but instead of one score vector and one loading vector as in bilinear PCA, each component consists of one score vector and two loading vectors. In multi-way analysis the distinction between score and loading vectors is often not made and one also uses the term loadings for all modes. A PARAFAC model of a three-way array is then given by three loading matrices (A, B and C) with elements a_{if} , b_{jf} and c_{kf} (**Figure 2.11**).

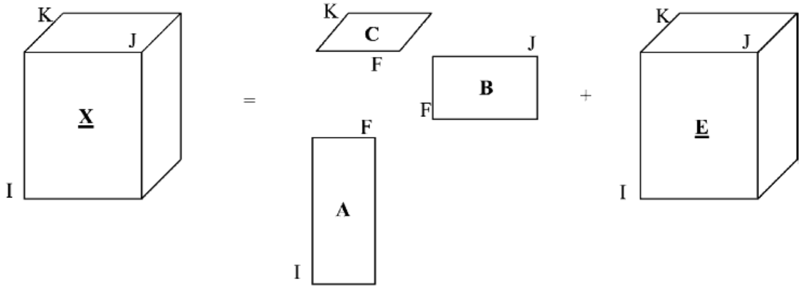


Figure 2.11: PARAFAC model - graphical explanation

Generally, the PARAFAC model is easier to interpret and many authors prefer to apply PARAFAC when it is possible (i.e. the complexity in all dimensions is equal). In this work the core consistency tool was used to obtain the optimal complexity of the PARAFAC model.

Among the PARAFAC model extensions we can mention PARAFAC2, Shifted PARAFAC (S-PARAFAC), Convolutive PARAFAC (cPARAFAC) and Parallel Factors with Linear Dependency (PARALID).

PARAFAC2 [119] is a less restrictive model that relaxes a PARAFAC model by requiring the invariance of the multiplication of a component matrix with its transpose in one mode rather than the invariance of the components themselves.

Component matrix in the first mode (or one of the modes) can vary across slices in a PARAFAC2 model. This relaxation enables the use of multi-way models in the cases, where a PARAFAC model cannot fully recover the underlying structure. Furthermore, PARAFAC2 solves the problem of modeling three-way arrays with slices of different dimensionality (if the dimensionality differs only in one mode). An example of such a multi-way array is an environmental data set that contains the concentrations of some chemical compounds measured at certain time periods across several sampling sites (sampling sites x parameters x time) [120]. It is quite common to have measurements from sampling sites for varying time periods, which would result in a three-way array with different dimensionality in one of the modes (e.g., time mode in this case).

S-PARAFAC [121] has been introduced in order to deal with shifting factors in sequential data such as time series or spectral data. While PARAFAC restricts the data to have the same factor in various proportions in all samples based on Cattell's idea, S-PARAFAC relaxes this restriction by incorporating shifting information into the model and capturing the factors even if they are available in shifted positions in different samples. S-PARAFAC and PARAFAC2 are quite similar; both models are less-constrained versions of a PARAFAC model and can model data with shifting factors. However, they also have their differences since PARAFAC2 can only capture shifts that maintain the inner product of the factors, while S-PARAFAC can model independent shifts at each factor.

Another extension of a PARAFAC model is cPARAFAC [122], which is a generalization of Nonnegative Matrix Factor Deconvolution (NMFD) to multi-way spectral data. cPARAFAC, closely related to S-PARAFAC, has been proposed for multichannel spectral data analysis in order to model convolutive mixtures. Convolution basically means generating a mixture by sending the sources through a filter. When convolution filter is sparse, cPARAFAC becomes equivalent to S-PARAFAC.

PARALIND [123] A common problem that arises in real data analysis is that ranks of the component matrices may not be the same (called rank deficiency). That would require extracting different number of factors in different modes. In that case, fitting a PARAFAC model would give rank deficient solutions and would not guarantee meaningful uniqueness. PARALIND is proposed as an approach for modeling such cases. This model introduces dependency (or interaction) matrices among component matrices to enable the modeling of the data with component matrices with different ranks and capture the dependency between components. Besides, via dependency matrices, prior knowledge about the data and constraints can be incorporated into the model.

Model	**Mathematical Formulation**	**Handles Rank-deficiency**	**Extended to N-way data**
PARAFAC	$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}$	x	✓
PARAFAC2	$\mathbf{X}_k = \mathbf{A}_\mathbf{k} \mathbf{D}_\mathbf{k} \mathbf{B}^T + \mathbf{E}_k$	x	✓
S-PARAFAC	$x_{ijk} = \sum_{r=1}^{R} a_{(i+s_{jr})r} b_{jr} c_{kr} + e_{ijk}$	x	✓
cPARAFAC	$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{(j-\theta)r} c_{kr}^{\theta} + e_{ijk}$	x	✓
PARALIND	$\mathbf{X}_k = \mathbf{AHD}_\mathbf{k} \mathbf{B}^T + \mathbf{E}_k$	✓	✓

Table 2.9: Models form PARAFAC family

2.2.3.2.2. Tucker family

Tucker models [124] are less restricted multi-way models than the models in PARAFAC family. They are also called three-mode factor analysis for three-way arrays or N-mode component analysis [125] for higher-order generalizations.

Similar to PARAFAC, Tucker is an extension of bilinear factor analysis to tensors.

Tucker3 method decomposes the 3-way data arrays $\underline{\boldsymbol{X}}$ (see **Figure 2.12**) into three orthonormal loading matrices, denoted as $\boldsymbol{A}$ (I×L), $\boldsymbol{B}$ (J ×M), $\boldsymbol{C}$ (K×N) and the core matrix $\underline{\boldsymbol{Z}}$ (L×M ×N), which describes the interactions among $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$. The largest squared elements of the core matrix $\underline{\boldsymbol{Z}}$ indicate the most important factors that describe $\underline{\boldsymbol{X}}$. The number of factors in each mode is not necessarily the same ($L \neq M \neq N$). Because of rotational freedom of the model, the core matrix is needed for their interpretation [126]. This flexibility allows an interaction between a factor with any factor in the other modes. While the core array enables us to explore the underlying structure of a multi-way data set much better than a restricted PARAFAC model, the full core array structure in Tucker3 has some drawbacks. First, this property is the reason for rotational indeterminacy in Tucker3 models. Unlike PARAFAC, a Tucker3 model cannot determine component matrices uniquely. When a component matrix is rotated by a rotation matrix, it is possible to apply the inverse of the rotation matrix to the core and still obtain the same model fit. Therefore, a Tucker3 model can determine component matrices only up to a rotation. Second, the interpretation of Tucker3 models is much more difficult compared to PARAFAC models.

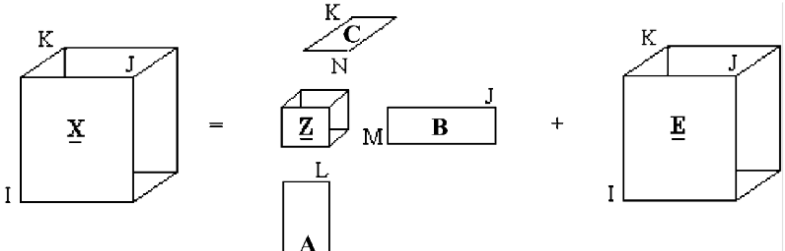


Figure 2.12: Tucker model - graphical explanation

Originally, Tucker family contains Tucker1, Tucker2, and Tucker3 models (Table 2.10). Tucker1 is based on the simple idea of rearranging multiway data as a matrix and decomposing the unfolded data using Singular Value Decomposition (SVD) [127]. Tucker2 and Tucker3 models allow rank reduction in more than one mode and are named after the number of modes rank reduction is applied [128]. Desired rank reduction in each mode is a user-specified parameter, and determining these parameters in Tucker models is a tedious task. While using ranks indicated by SVD on unfolded data in each mode is a practical option, systematic methods (ex: cross validation and Difference in Fit (DIFFIT) [129] have also been developed. DIFFIT enumerates all possible models and uses the differences between model fits to determine the number of components. However, high computational complexity of DIFFIT makes it inefficient. Therefore, it has later been improved by comparing approximate model fit values rather than exact model fits [130]. The most recent work in finding the number of components is based on searching for the convex hull on the plot of model fit values versus number of free parameters [131]. This approach is more general than previously proposed methods and helps in determining the model parameters not only in Tucker3 but also in Tucker1, Tucker2, and PARAFAC models. Nevertheless, there is no straightforward way to find the optimal number of components [131], and several diagnostics should be used to have a true understanding of the structure of a multi-way data set.

Model	**Mathematical Formulation**	**Handles Rank-deficiency**	**Extended to N-way data**
Tucker1	$x_{ijk} = \sum_{p=1}^{P} g_{pik} a_{ip} + e_{ijk}$	✓	✓
Tucker 2	$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} g_{pqk} a_{ip} b_{jp} + e_{ijk}$	✓	✓
Tucker3	$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqk} a_{ip} b_{jp} c_{kr} + e_{ijk}$	✓	✓
S-Tucker3	$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqk} a_{(i+s_{jp})p} b_{jp} c_{kr} + e_{ijk}$	✓	✓

Table 2.10: Models from Tucker family

In order to capture shifting factors, similar extensions as in PARAFAC models have also been studied for Tucker models. Shifted Tucker3 (S-T3) and Shifted Tucker2 (S-T2) are the combinations of Shifted Factor Analysis with Tucker models. Although it is not proven formally, it has been discussed in [132] that incorporating shifting information in S-T3 suggests the uniqueness of an S-T3 model.

2.2.3.2.3. Multi-way calibration. Multi-linear PLS

N-PLS regression is the extension of ordinary PLS (two-way PLS) to three or more (N) kinds of objects [133]. Similar to standard PLS regression, n-PLS builds a model incorporating a relationship between a set of independent variables X and a dependent variable set Y. The extension provided in the multi-way version of PLS regression is that independent data is not limited to having only one mode of variation.

2.2.3.2.4. Alternative methods

There are several other models based on different approaches than PARAFAC and Tucker models for multi- way data analysis.

Multilinear Engine (ME) [134] is a program that is capable of fitting different models including PARAFAC and PARAFAC2 on multi-way arrays using a general-purpose optimization/curve fitting approach. Although models mentioned so far are only capable of modeling multi-linearity in data, structure tables created by specified variables and functions enable ME to fit multilinear as well as quasi-multilinear models. Multilinear models are based on mathematical expressions, which are linear with respect to each set of variables corresponding to different modes whereas quasi-multilinear models contain nonlinearity in the sense of polynomials. Therefore, the ME can explore a wider range of structures in data compared to PARAFAC and Tucker3.

Another model focusing on three-way data analysis is STATIS [135], which was originally studied in [136]. This model explores each mode separately, rather than simultaneously like in the N-way analysis methods. It considers each observation/sample as a slice of a three-way array and computes the covariance matrix corresponding to that slice. The basic principle in the model is to apply Principal Component Analysis (PCA) on a global covariance matrix formed as a linear combination of covariance matrices corresponding to individual slices. Using STATIS, it is possible to analyze three-way arrays with slices of different sizes. One disadvantage of STATIS is that it cannot be generalized to N-way arrays.

Methods, referenced so far, focus on the analysis of a single multi-way array. On the other hand, multiple multi-way arrays are also encountered in various studies such as in batch process control, where multiple multi-way arrays need to be analyzed simultaneously. One approach to deal with such problems is to analyze each multi-way array using a certain model such as a Tucker or a PARAFAC model and then combine summaries of information from different multi-way arrays in a single matrix [137]. The matrix containing summaries from different arrays can then be analyzed using bilinear factor models. This approach can be considered as a generalized version of Collective PCA [138] to higher-order data sets.

2.2.3.2.5. Data types and application areas of multi-way methods

Fluorescence spectroscopy is by far the most abundant type of data used for multi-way analysis [139]. As one of the most popular application of a PARAFAC model, fluorescence excitation-emission data is a commonly used data type in chemistry, medicine, and food science. Such data typically consist of samples containing different concentrations of several chemical compounds. The goal of PARAFAC analysis on this data type is to determine the compounds found in each sample as well as the relative concentrations of compounds. Fluorescence spectroscopy enables the generation of three-way data sets with the following modes: samples, emission, and excitation wavelengths. An example of a PARAFAC model on a fluorescence spectroscopic data set is given as an in-depth study by Andersen and Bro [140]. This study is an important resource demonstrating the underlying idea of the structural model of PARAFAC, its benefits and limitations.

A close relationship between the unique PARAFAC model and the fundamental structure of common fluorescence spectroscopic data exists. The reason why the PARAFAC model is the most appropriate for modeling such a data is the each one of the fluorescence landscapes is a rank-one matrix and a particular analyte has specific signatures in emission and excitation modes. There is no need to model such a data set using a Tucker3 model since the components in each mode only interact with components with the same ID in other modes, indicating a superdiagonal core structure as in the case of a PARAFAC model.

Although fluorescence excitation-emission data is ideal for PARAFAC modeling, there are a number of problems with typical fluorescence data that complicates the application of PARAFAC. One of the most notable problems is the different kind of scattering, which results in data that does not follow the PARAFAC well, and does not contain any chemical information. Several applications have been tested to remedy the scattering influence by downweighting the scatter, setting it to zero or modeling it. [140, 141, 142]

Chromatography is another area where multi-way analysis has been used extensively. Ideally, a chromatographic experiment leads to perfect separation of chemical analytes and the resulting chromatograms can be directly converted to peak areas (concentrations) and used for further analysis. Sometimes, though, it is not possible to achieve perfect separation, either because of the complexity of the samples or because faster chromatographic runs are preferred. In such situations, overlapping peaks result and data analysis can be used for further achieving selectivity by mathematical means or simply for fingerprinting the elution profiles [143].

In chromatography there are multiple types of multi-way data, depending on the type of instrument used. The most common is the gas chromatograph-mass spectrometer, which is the subject of this work, and is therefore described in more detail. Some other multi-way chromatographic data comes from the two dimensional chromatography: comprehensive two-dimensional gas chromatography (GC x GC) [144], comprehensive two-dimensional liquid chromatography (LC x LC) with a univariate detector. If the detector in these two dimensional apparatus are multivariate like the mass spectrometer, then the data will have even higher order (four-way).

Flow injection is another application where three-way data occur [145, 146],and there are often problems with so-called rank overlap because of the lack of separation in the time mode. Least squares-based and eigenvalue-based algorithms were tested for calibration in FIA (flow injection analysis) systems with a pH profile induced and with rank overlap [147].

Magnetic resonance spectroscopy is gaining momentum in new types of applications, and multi-way analysis has been extensively used for both low-field and high-field NMR types of data. In low-field NMR, the use of multi-way analysis arises from the often multi-exponential nature of the data. Such data can be modeled by rearranging the data appropriately (slicing) and then fit a PARAFAC model to the rearranged data [148]. The resulting models provide direct access to the quantitative and qualitative information of the underlying exponentials. Often, the use of least squares fitting through alternating least squares is replaced by faster approximate methods based on eigenvalue decompositions. These methods are particularly useful for low-field NMR because the signal to noise ratio is high and the model error low leading to excellent performance of these methods.

Multichannel *electroencephalogram* (EEG) data enables the capture of the correlation between the channels by representing signals in both time and frequency domains. The recordings are commonly represented as an I x J matrix containing signals recorded for I time samples at J channels. In order to discover the brain dynamics, often frequency content of the signals, for instance signal power at K particular frequencies, also needs to be considered. In that case, EEG data can be arranged as an I x J x K three-way data set [149]. Multiway analysis of a three-way EEG array can then be used to extract the signatures of brain dynamics in time, frequency, and electrode domains.

To this data type briefly presented above we could ad *near infrared spectroscopy* (NIR) batch data, modeled using multi-way analysis by Geladi [150, 151]. *Ultraviolet-visible spectroscopy* (UV/Vis) [152, 153], *X-ray fluorescence* (XRF) [154], *Inductively coupled plasma atomic emission spectroscopy* (ICP-AES) [155] and *image analysis* [156] are just another type of data where multi-way models have been employed to reveal the hidden information within the data.

2.2.3.2.6. Application areas of multi-way data analysis

Multi-way models are employed in numerous disciplines addressing the problem of finding the multilinear structure in multi-way data sets. There are many applications in various fields, and this survey offers some representative examples from different research areas.

Calibration is one of the major fields of application in chemometrics and also using multi-way methods [157 - 159]. Several papers describe and compare different theoretical approaches to calibration [160, 161] and a large number of applications especially within analytical chemistry have been published.

Multivariate statistical process control (MSPC) often involves multi-way data and hence applications of multi-way analysis have appeared especially in recent years [162 - 164]. Experimental results from a pilot-scale paper mill filtration application was modeled using PARAFAC [165], PARAFAC2 was used for fault detection and diagnosis in a semiconductor etching [166], n-PLS was used to model the performance of an industrial fedbatch fermentation process [167], while a modified Tucker3 model was used for modeling first-order chemical batch reactions monitored by UV/Vis spectroscopy [168]. PARAFAC and N-PLS was used to model dynamic behavior in an on-line batch process [169]. Finally, PARAFAC was developed for on-line monitoring of batch processes [170]. Other models based on multidimensional scaling have also been proposed [171].

Metabonomic data are often multi-way of structure. Dyrby [172] showed how a traditional dose-response experiment could be properly explored using Tucker3 models and PARAFAC was used for similar data [173]. Idborg [174] exemplified the use of PARAFAC and N-PLS in handling metabolite screening using liquid chromatography-electrospray ionization mass spectrometry. Lipoprotein characterization using 2D diffusion-edited NMR spectroscopy was analyzed with PARAFAC providing chemically meaningful spectra [175].

Environmental analysis is another field where the multi-way analysis has been used extensively. The structure of the data collected in these studies is usually three-way, representing sites, variables and depths, like in this Indian study of soils irrigated with waste water [176]. In a study monitoring the Venice lagoon a Tucker3 model has been used leading to the easy identification of the effects present in each of the three modes (sampling sites x variables x sampling times) of the data [177]. N-PLS was used for predicting average air temperature, dew temperature and precipitation [178] while PARAFAC was used to characterize polycyclic aromatic hydrocarbons [179]. The availability and mobility of trace elements and heavy metals were analyzed using PARAFAC [180].

Kinetics studies have been studied with multi-way methods. PARAFAC-like models with kinetic constrains incorporated have been published [181], as well as reports on suitability of three-way methods for kinetic modeling [182, 183]. Sometimes, direct application of multi-way methods such as PARAFAC are hampered by the changes in actual shapes of kinetic profiles, yet there have been experiments where PARAFAC2 was used to overcome this problem [184].

Classification is an area where little work has been done with dedicated multi-way methods, although new types of classification have been suggested [185]. In one study PARAFAC fluorescence analysis of olive oils was used to discriminate between non-adulterated and adulterated samples [186], while in another one the chewing sound was used to detect differences in foods [187]. Tucker3 modeling was used to study differences in patterns of nutritional content of foodstuffs [188].

2.2.3.2.7. Available software

As multi-way analysis is spreading from chemometrics and psychometrics to other fields, software tools have also been developed and improved. Some available software for multi-way data analysis are the N-way Toolbox [189], Tensor Toolbox [190], [191], PLS_Toolbox [192], and CuBatch [193], which all run under Matlab. The N-way toolbox [194] is the original toolbox, which has combined multi-way analysis techniques such as PARAFAC and Tucker models in a software package and enabled the application of these models in different fields. The Tensor Toolbox was initially introduced as a TensorClass, to handle mathematical operations on multi-way and later has been extended to manipulate efficiently multi-way arrays and sparse tensors. CuBatch is another software package recently introduced as a multi-way analysis toolbox, originally built for analyzing batch process data, but it is applicable on multi-way data sets in general. Preprocessing techniques such as centering and scaling and different techniques for identifying outliers are also included in this toolbox. CuBatch contains Nway Toolbox functions and it is a more developed version of the initial toolbox. Apart from these freely available toolboxes, there is also a commercial toolbox called PLS_Toolbox, which can be licensed, giving access to analysis of multi-way arrays with numerous multi-way models providing visual analysis tools.

In addition to software running under Matlab, there is also another software package called The Multilinear Engine [195] implemented in Fortran. There are also other software packages for manipulating multiway arrays but they do not particularly focus on multi-way data analysis or multi-way models. Among these we remind DATAN [196], which contains PARAFAC and The Unscrambler [197] which contains n-PLS. Also, the Three-Mode Company [198], devoted to creating and promoting three-mode data analysis, is a good resource, giving a extensive list of available software.

More information on these software packages is given by Bader [199].

2.2.3.3. Neural networks

Artificial neural networks (ANN) are computational systems that simulate the microstructure (neurons) of a biological nervous system. The most basic components of ANNs are modeled after the structure of the brain, and therefore the terminology is borrowed from neuroscience. They can be defined as a structure composed of a number of interconnected units [200]. Each unit has an input/output (I/O) characteristic and implements a local computation or function. The output of each unit is determined by its I/O characteristic, its interconnection to other units and (possibly) external inputs, and its internal function. The network usually develops an overall functionality through one or more forms of training. The fundamental unit or building block of the ANN is called artificial neuron [201]. The neuron has a set of inputs (X_i) weighted before reaching the main body of the processing element. In addition, it has a bias term, a threshold value that has to be reached or exceeded for the neuron to produce a signal, a non-linearity function (f_i) that acts on the produced signal (R_i), and an output (O_i). The basic model of a neuron is illustrated in **Figure 2.13**.

Learning is the process by which the neural network adapts itself to a stimulus and eventually (after making the proper adjustments by adjusting its synaptic weights) produces the desired response. The two major categories of learning in ANN are the supervised and the unsupervised. In the supervised learning, the output response is compared to a desired target response. If the actual response differs from the target response, the network generates an error signal, which is then used to calculate the adjustment that should be made to the synaptic weights, so that the actual output matches the target output. In contrast, unsupervised learning does not require a target output. During the training session, as the neural net receives input excitations or patterns, it arbitrary organizes them into categories. When a stimulus is later applied, the network provides an output response indicating the class to which the stimulus belongs. If a class cannot be found for the input stimulus, a new class is generated.

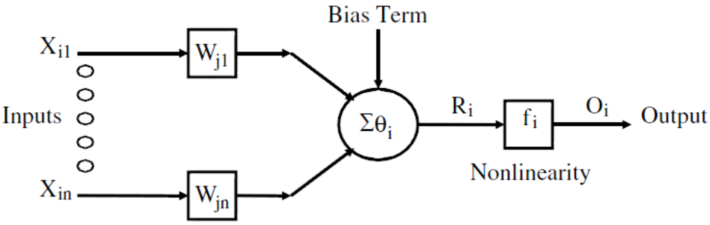


Figure 2.13: model of artificial neuron

ANN generally consist of a number of layers: the layer where the input patterns are applied is called the input layer, the layer where the output is obtained is the output layer, and the layers between the input and output layers are the hidden layers (**Figure 2.14**). There may be one or more hidden layers, which are so named because their outputs are not directly observable. The addition of hidden layers enables the network to extract higher-order statistics which are particularly valuable when the size of the input layer is large [202]. Neurons in each layer are fully or partially interconnected to preceding and subsequent layer neurons with each interconnection having an associated connection strength (or weight). The input signal propagates through the network in a forward direction, on a layer-by-layer basis. These networks are commonly referred to as multilayer perceptrons (MLP).

The back-propagation training algorithm is commonly used to iteratively minimize a function with respect to the interconnection weights and neurons thresholds. The training process is terminated either when the mean-square-error (MSE), root-mean-square-error (RMSE), or normalized-mean-square-error (NMSE), between the observed data and the ANN outcomes for all elements in the training set has reached a pre-specified threshold or after the completion of a pre-specified number of learning epochs.

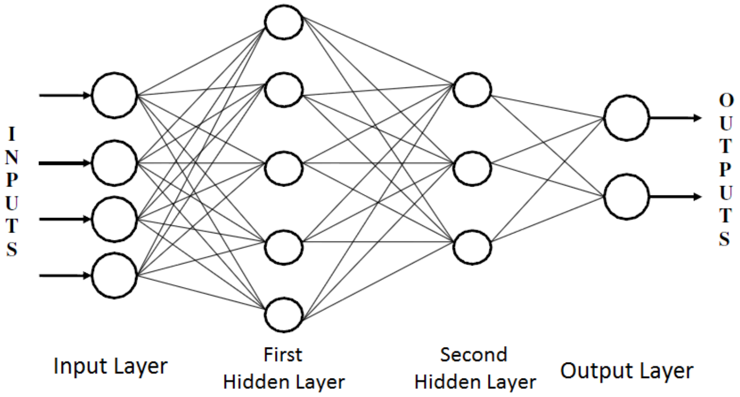


Figure 2.14: General configuration of an artificial neural network

Many types of artificial neural networks exist, like feed forward neural networks, radial basis function network, Kohonen self-organizing maps, and many more. Despite many different neural network paradigms do exist, we centered on using fuzzy ARTMAP neural networks in this thesis.

The Fuzzy ART is a neural network introduced by Carpenter, Grossberg and Rosen in 1991 [203]. It is a modified version of the binary ART1 [204], which is notably able to accept analog fuzzy input patterns (vectors whose components are real numbers between 0 and 1). The Fuzzy ART is an unsupervised neural network capable of incremental learning, that is, it can learn continuously without forgetting what it has previously learned.

A Fuzzy ART network is formed of two layers of neurons, the input layer F_1 and the output layer F_2, as illustrated in **Figure 2.15**. Both layers have an activity pattern, schematized on the figure with vertical bars of varying height. The layers are fully interconnected, each neuron being connected to every neuron on the other layer. Every connection is weighted by a number lying between 0 and 1. A neuron of F_2 represents one category formed by the network and is characterized by its weight vector w_j (j is the index of the neuron) . The weight vector's size is equal to the dimension M of layer F_1. Initially all the weight vectors' components are fixed to 1. Until the weights of a neuron are modified, we say that it is uncommitted. Inversely, once a neuron's weights have been modified, this neuron is said to be committed.

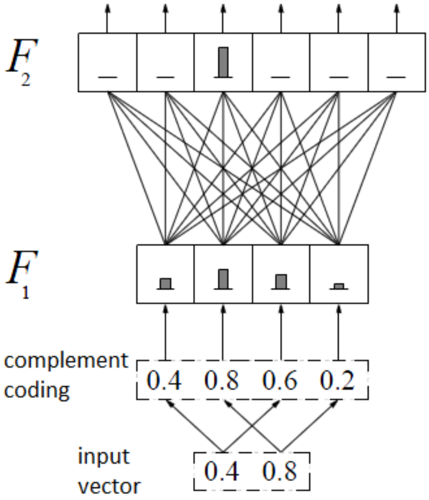


Figure 2.15: Fuzzy ART network

The network uses a form of normalization called complement coding. The operation consists on taking the input vector and concatenating it with its complement. The resulting vector is presented to layer F_1. Therefore, the dimension M of layer F_1 is the double of the input vector's dimension. Complement coding can be deactivated, in this case layer F_1 will have the same dimension as the input vector. Unless specified otherwise, we will always suppose that complement coding is active.

The Fuzzy ART learns by placing hyperboxes in the M/2 - dimensions hyperspace, M being the size of layer F_1. As said earlier, each neuron of layer F_2 represents a category formed by the network, and this category is defined by a box. The position of the box in the space is encoded in the weight vector of the neuron. Because of complement coding and of the learning process we will explain later, the first half of the weight vector memorizes one corner of the box (the closest to the origin) and the other half memorizes the opposite one.

The Fuzzy ARTMAP, introduced by Carpenter et al. in 1992 [205] is a supervised network which is composed of two Fuzzy ARTs. The Fuzzy ARTs are identified as ART_a and ART_b. The parameters of these networks are designated respectively by the subscripts a and b. The two Fuzzy ARTs are interconnected by a series of connections between the F_2 layers of ART_a and ART_b. The connections are weighted (a weight w_{ij} between 0 and 1 is associated with each one of them). These connections form what is called the map field F^{ab}. The map field has two parameters (β_{ab} and ρ_{ab}) and an output vector x^{ab}.

The input vector a of ART_a is put in complement coding form, resulting in vector A. Complement coding is not necessary in ART_b so we present the input vector B directly to this network. **Figure 2.16** presents the structure of the Fuzzy ARTMAP. The weights of the map field's connections are illustrated by a vertical bar with a height proportional to the size of the weight. The weights of the map field are all initialized to 1.

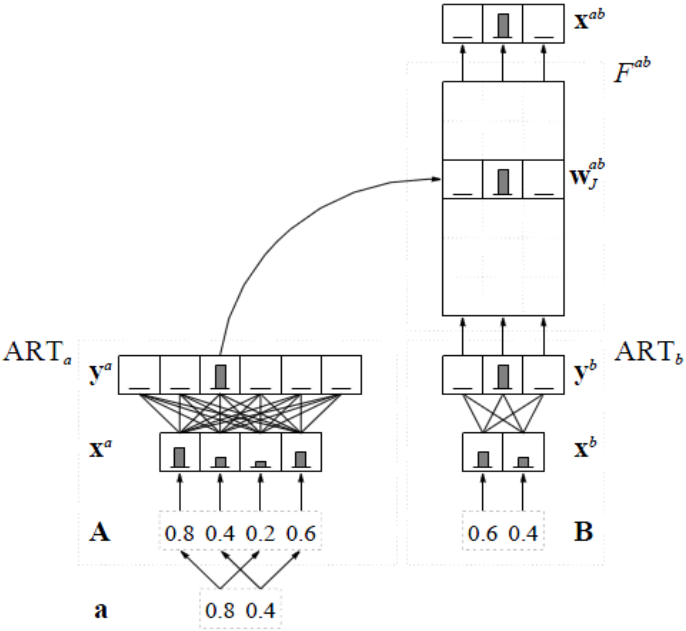


Figure 2.16: Fuzzy ARTMAP network

In order to train the Fuzzy ARTMAP, we present to ART_a a vector representing a data pattern and to ART_b a vector that is the desired output corresponding to this pattern. The network uses a form of hypothesis testing. When it receives the first vector, it deduces to which category it should belong. With the second vector, the Fuzzy ARTMAP can either confirm or reject the hypothesis, in which case the process is repeated with a new one.

Chapter 3

Experimental design

3. Experimental design

3.1. Dimethylphenol isomers mixtures to challenge the MS-based electronic nose

3.1.1. Introduction

Initially, the field of machine olfaction and electronic noses was developed based on solid-state gas sensors. Because of well-known sensor drawbacks such as reproducibility and short lifetime, in the last few years mass spectrometry-based E-noses are becoming an increasingly used alternative. Among the well-known advantages of MS-based electronic noses the most important is a better reproducibility. The main drawback of MS-based electronic noses is that they are not portable for in-field applications.

Although not precisely being a gas sensor, Mass Spectra-based electronic noses can be used together with chemometric methods to obtain a fingerprint of the aroma of a product and classify samples accordingly.

In scientific fields such as medicine, where the complexity of the data challenges the sensor array even when using the MS–sensor approach, the combination of gas chromatography and mass spectroscopy coupled to sophisticated pattern recognition techniques like multi-way data analysis can prove a very useful approach, increasing resolution and reliability.

In this work, it is our intention, to evaluate whether a GC-MS configuration can improve the results of a standard MS-based E-nose without extending the measurement time to that used in a typical GC run using 2-D and/or 3D pattern recognition algorithms.

We base our hypothesis on the fact that the pure MS signal of complex mixtures generates a great mix of mass fragments from all the components. These simultaneous fragmentations generate non-linear behaviors and interferences that make identification a very complex problem. On the other hand, in GC-MS, each component is subject to fragmentation at a different time, generating less mass fragment mixtures and a more linear behavior. In this approach, because the information is not coming mixed at a single time, the classification ability is improved and identification seems more plausible.

Based on the fact that the electronic noses are expected to give fast answers, the chromatographic methods have to be optimized in order to require a short measurement time, while keeping the advantages brought by the retention time information, the third dimension of the data. That can be achieved using short length columns or fast chromatographic methods.

Because the nature of the output data of a gas chromatograph mass spectrometer exceeds the direct applicability of two-way data analysis methods like PCA or PLS, we are going to use three dimensional data analysis methods like PARAFAC and n-PLS as well, and see which approach is more suitable to this new structure of data.

Multi-way data analysis, originating in psychometrics back in the sixties, is the extension of two-way data analysis to higher-order datasets. Multi-way analysis is often used for extracting hidden structures and capturing underlying correlations between variables in a multi-way array. The difference between two-way and multi-way data analysis is the format of the data being analyzed. Multi-way arrays, often referred to as tensors, are higher-order generalizations of vectors and matrices. More on these methods can be found on previous sections.

3.1.2. Goals

The main goals of this experiment are the following:

- to show that the addition of the third dimension is useful, and brings extra information, helping in the classification of samples

- to optimize the chromatographic method in order to shorten the time consuming chromatographic analysis to a minimum, while still having acceptable results
- to evaluate the best preprocessing techniques for the application of two-way (PCA, PLS) and multi-way data analysis (PARAFAC, n-PLS)
- To compare 2-way and multi-way methods taking into account than 3-way data has to be re-arranged to be used by 2-way methods

3.1.3. Experimental

In order to improve the Mass Spectra-based electronic nose (MS-EN) performance by taking also into account the chromatographic separation of a typical GC-MS run, , we looked for an experiment to challenge the MS-based electronic nose in which we would include the whole three dimensional data output of a gas chromatograph/mass spectrometer.

Based on their physical and chemical characteristics, the most suited components to prove difficult to separate and classify by the MS-based e-nose are the organic isomers. With the same molecular mass and formula, and the only differences based on the relative position of the atoms in the structure, these organic compounds, when introduced in the ionization chamber, brake down in more or less the same fragments, thus giving very similar mass spectra.

Therefore we choose 9 isomers of dimethylphenol and ethylphenol as follows: 1) 2,3-Dimethylphenol; 2) 2,4-Dimethylphenol; 3) 2,5-Dimethylphenol; 4) 2,6-Dimethylphenol; 5) 3,4-Dimethylphenol; 6) 3,5-Dimethylphenol; 7) 2-Ethylphenol; 8) 3-Ethylphenol and 9) 4-Ethylphenol. From **Figure 3.1** it can be seen that they have very similar mass spectra:

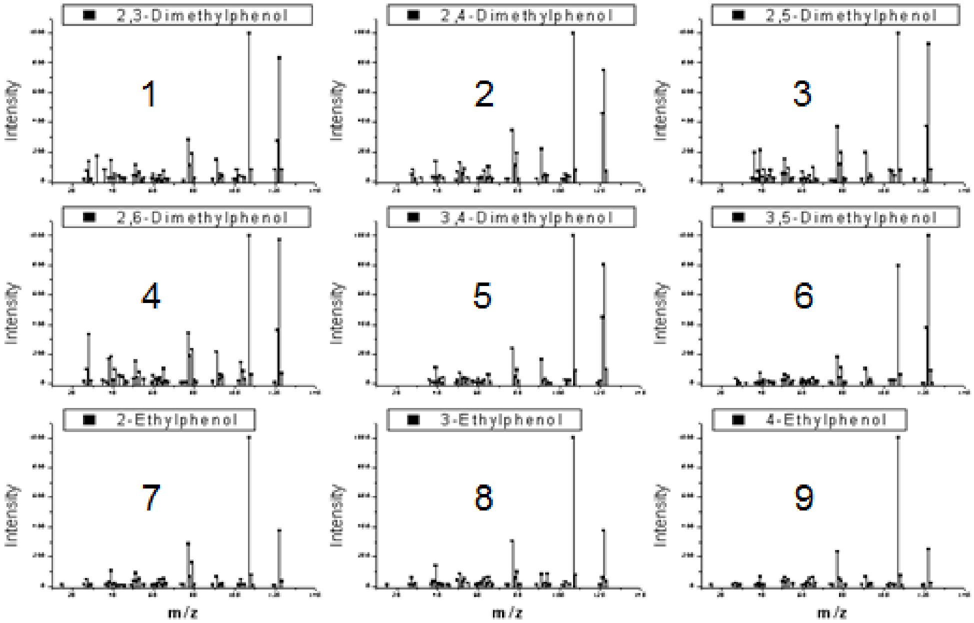


Figure 3.1: Theoretical mass spectra (from the NIST 62 library, National Institute of Standards and Technology MS Database) of the nine isomers of dimethylphenol and ethylphenol

The idea of the experiment is to generate and later identify different solutions of the nine isomers mixed with methanol. In order to have a challenging data set, we designed the experiment looking at 2 issues. First, a PCA of theoretical mass spectra of the nine isomers was performed. These spectra were obtained from the NIST 62 library. Based on these PCA it was decided which isomers should be combined into solutions in order for the final mixtures to have a high degree of similarity, thus challenging the performance of the MS-based electronic nose. The projection revealed that 2,3-Dimethylphenol and 2,4-Dimethylphenol and 2-Ethylphenol and 3-Ethylphenol had very similar mass spectra and therefore it was difficult to distinguish solutions which contained one of them (**Figure 3.2**) looking at mass spectra alone.

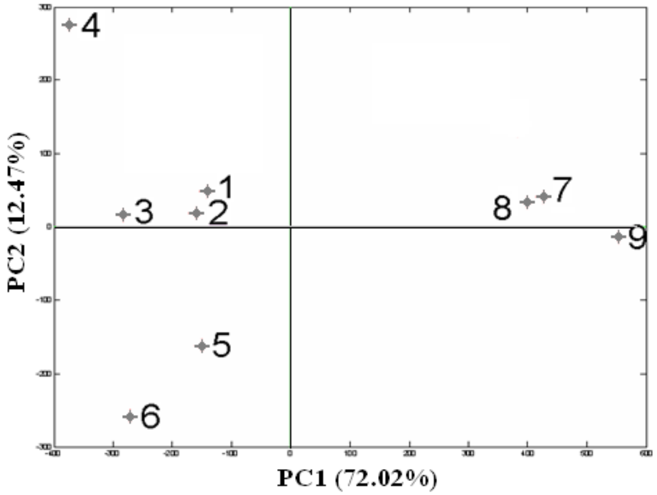


Figure 3.2: Principal Component Analysis with 2 principal components of the isomers' theoretical mass spectra as follows: 1) 2,3-Dimethylphenol; 2) 2,4-Dimethylphenol; 3) 2,5-Dimethylphenol; 4) 2,6-Dimethylphenol; 5) 3,4-Dimethylphenol; 6) 3,5-Dimethylphenol; 7) 2-Ethylphenol; 8) 3-Ethylphenol and 9) 4-Ethylphenol

Secondly, we looked at the isomers from the chromatographic point of view (Figure 3). To do this, we made solutions containing just one isomer in methanol in a concentration of 1% and one methanol solution containing all the nine isomers in a concentration of 1% each. One µl of each of the nine solutions was injected into the injection port of a Shimadzu QP 5000 gas chromatograph-mass spectrometer equipped with an Equity-5 poly (5%defhenyl.95%dimethylsiloxane) (30m x 0.25mm x 0.25mm) capillary column, purchased from Supelco Inc. A temperature-programmed separation was employed. The GC oven was held at 50^{o}C for one minute after which the temperature was raised with a rate of 10^{o}C/minute until reaching 180ºC.

First we injected the solution containing all nine isomers, and the chromatographic run showed that some of the nine isomers peaks were overlapping (**Figure 3.3**). This fact indicated that the experiment would be challenging also from a chromatographic point of view, and giving the opportunity for the combined information of the mass spectra and chromatographic runs to improve both the more classical isolated MS or GC approaches.

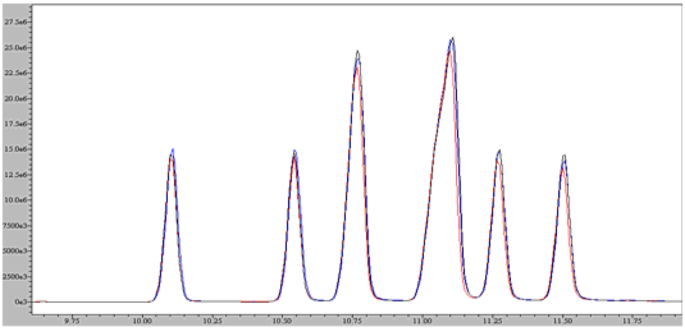


Figure 3.3: Chromatogram of the solution containing all the nine isomers

In order to see where on the time line of the chromatographic separation each isomer injected would coelute following the same chromatographic method, one µl of each solution containing just one isomer was analyzed. In this way we identified all the nine components in the chromatographic separation (**Figure 3.4**).

The chromatogram shows that isomers 2 and 3, and 6, 8 and 9 are difficult to isolate. This situation would prove a challenge for any GC-MS system if the classification were to be based on the chromatographic signal alone.

In order to have a stable reference we added benzene as an internal standard. Benzene concentration was set higher than the rest of the isomers in order to have the highest peak on the chromatogram, making the pretreatment of the data through normalization easier.

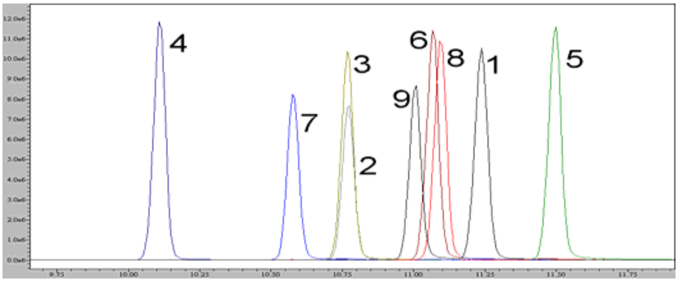
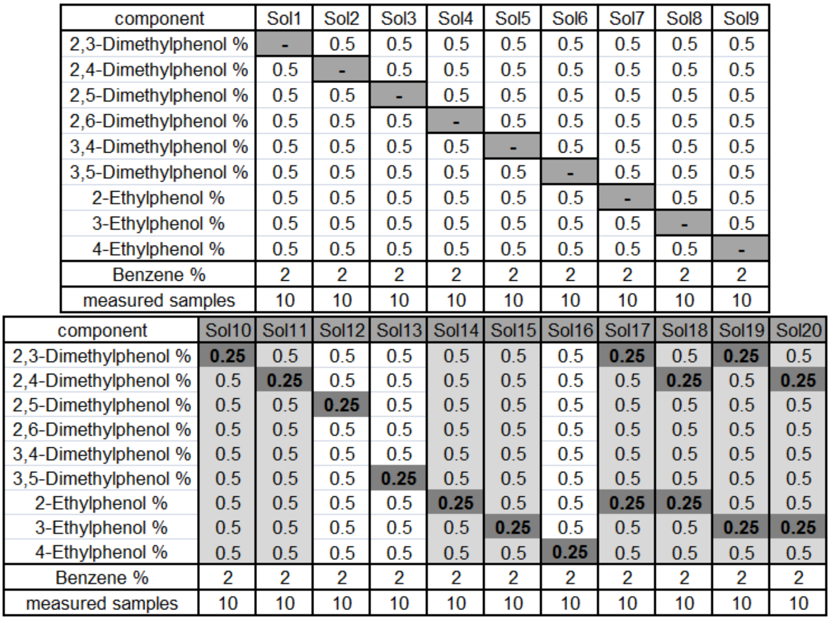


Figure 3.4: Superimposed individual chromatographic retention times for each of the 9 isomers

The experiment, based on the PCA of the theoretical mass spectra and the retention time of the nine isomers was designed as having two main parts (see **Table 3.1**). In the first part, from solution 1 to 9, the solutions have all the isomers in a concentration of 0.5%, except for one that is absent. All of this diluted in ethanol plus the internal standard (benzene) in a concentration of 2%.

The second one, which involves solutions ten to twenty, was designed to increase the difficulty for the MS and GC analysis. Here all the solutions had 2% of benzene, 0.5% of most of the isomers while certain isomers had half of that concentration (0.25%). These isomers where chosen based on the PCA of their theoretical mass spectra and on their retention time to increase complexity. Isomers 1 and 2 and 7 and 8, being so close in a PCA, would prove to be challenging for the mass spectra analysis. From the chromatogram point of view, isomers 2 and 3 and 6, 8 and 9, having almost the same retention times, would challenge the chromatographic analysis.



component	Sol1	Sol2	Sol3	Sol4	Sol5	Sol6	Sol7	Sol8	Sol9
2,3-Dimethylphenol %	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2,4-Dimethylphenol %	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2,5-Dimethylphenol %	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5
2,6-Dimethylphenol %	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5
3,4-Dimethylphenol %	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5
3,5-Dimethylphenol %	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5
2-Ethylphenol %	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5
3-Ethylphenol %	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5
4-Ethylphenol %	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-
Benzene %	2	2	2	2	2	2	2	2	2
measured samples	10	10	10	10	10	10	10	10	10

component	Sol10	Sol11	Sol12	Sol13	Sol14	Sol15	Sol16	Sol17	Sol18	Sol19	Sol20
2,3-Dimethylphenol %	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	**0.25**	0.5
2,4-Dimethylphenol %	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	**0.25**
2,5-Dimethylphenol %	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2,6-Dimethylphenol %	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,4-Dimethylphenol %	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,5-Dimethylphenol %	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2-Ethylphenol %	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	**0.25**	**0.25**	0.5	0.5
3-Ethylphenol %	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	**0.25**	**0.25**
4-Ethylphenol %	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5
Benzene %	2	2	2	2	2	2	2	2	2	2	2
measured samples	10	10	10	10	10	10	10	10	10	10	10

Table 3.1: Isomer mixtures experimental design

In order to study time optimization in the time axis (chromatographic separation) and the influence of more coeluted peaks (equivalent to faster GC measurements and/or shorter columns), three different chromatographic methods were programmed.

In the first method all the isomers in the TIC were resolved to the maximum possible, and therefore a temperature-programmed separation was employed. The GC oven was held at 50°C for one minute after which the temperature was raised with a rate of 10°C/minute until 180°C, giving a chromatographic run of 20 minutes.

Method two and three were designed to give more coeluted peaks, and therefore isothermal separations at 175 and 190°C were used, reducing the retention time to no more than 5 min. These methods gave more coeluted peaks and we expected them to generate more challenging datasets for our system (**Figure 3.5**). For all three methods the mass detector was operating in the electron impact ionization mode with a scan range from m/z 40 to m/z 200 at 0.5 scan/s. The ion source temperature was kept at 250°C.

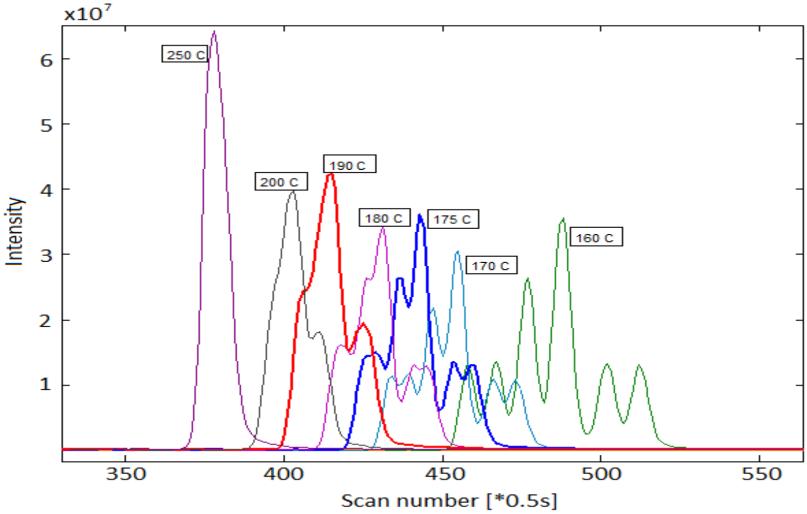


Figure 3.5: Isothermal chromatographic runs from which we choose method 2 (175°C) and method 3 (190°C)

The measurements were conducted through syringe injection, introducing 1 µl per injection. Ten repetitions were made per solution and chromatographic method. Two blanks were also measured, one consisting of just methanol and the other consisting on methanol and 2% of benzene.

For each measurement, the data was collected in 3-D format. Separation time was plotted against the x-axis, the m/z ratio (mass spectra) against the y-axis and the intensity against the z-axis. This format gave us a regular 2D matrix for each measurement so that each element value of the matrix represented intensity in a given coeluted moment and m/z ratio. Grouping all the measurements resulted in a three-dimensional matrix in which each measurement was represented in a horizontal plane of the 3D matrix. Therefore, the 3-D matrix (I x J x K) was organized with the following directions: samples measured (I, rows), m/z ratio (J, columns), and retention time (K, tubes) (**Figure 3.6**).

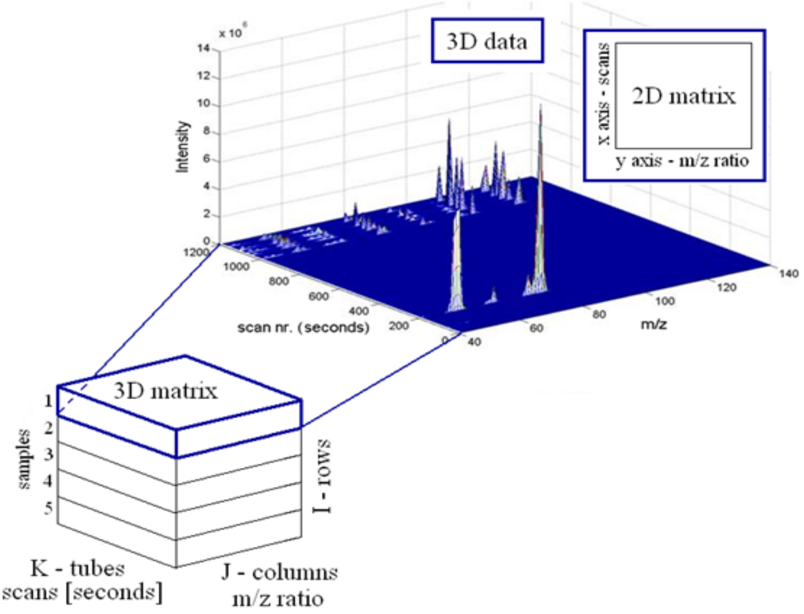


Figure 3.6: Graphical representation of the 3D data of one GC-MS measurement and the 3D matrix containing several measurements (I-rows) of 2D GC-MS matrices (K-tubes and J-columns)

The original 3D matrix containing the data from the measurements was analyzed using two and three way methods. In order to analyze 3D data with two-way methods (PCA, PLS-DA) we converted the 3D data to 2D data following different approaches.

In a first strategy, we "collapsed" the x-axis of the chromatographic separation time obtaining a 2D matrix where each file contained the average mass spectra of a single measurement, as it can be seen on **Figure 3.7** left. The second option of the first strategy is to collapse the y-axis of the m/z fragments, obtaining a matrix with the total ion chromatogram or TIC (**Figure 3.7** right).

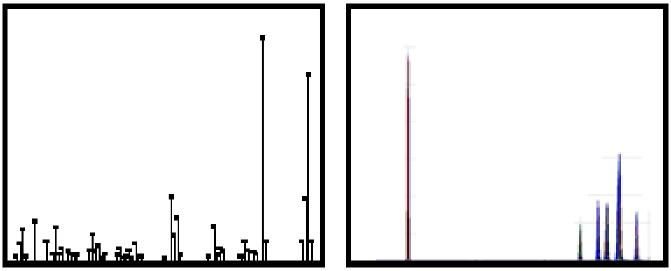


Figure 3.7: Average mass spectra (left) and total ion chromatogram (right)

By collapsing (summing) in either axis the extra information brought by the third dimension is lost. In order to compensate for this and still allow the data to be treated with 2D methods, two additional approaches were used: unfolding (**Figure 3.8**) and concatenation (**Figure 3.9**).

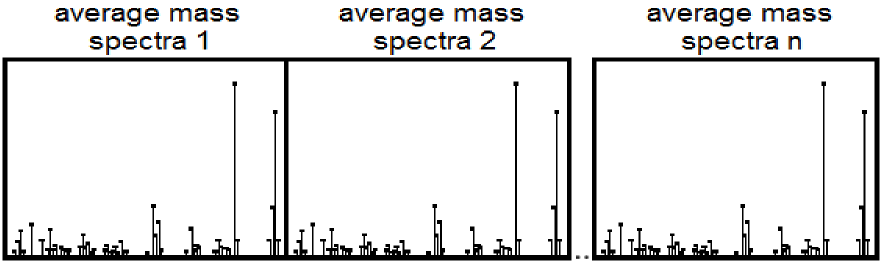


Figure 3.8: Graphical representation of unfolded data

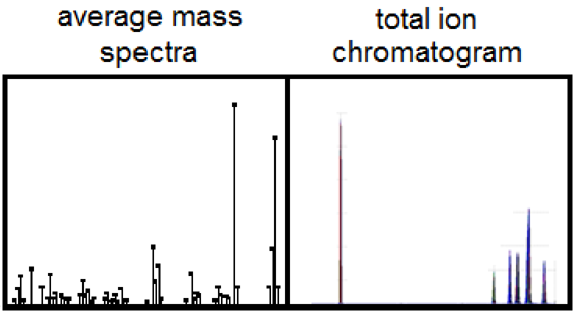


Figure 3.9: Graphical representation of concatenated data

Unfolding the three dimensional data generates a 2 dimensional matrix where all the original data is maintained in relation to their value but the 3D structural nature of the dataset is lost. This approximation consists on taking the mass spectra of each chromatographic scan and pasting it where the previous scan ended, in the same manner as you would put a deck of cards side by side on the table.

With the concatenation approach, the extra information brought by the chromatographic separation is kept, but not in the same way. The three dimensional matrix is collapsed in both directions, first giving the average mass spectra and then the total ion chromatogram. Once both vectors are obtained, the chromatogram vector for each measurement is concatenated with the mass spectra vector of the same measurement, giving a new 2D matrix where each file is composed by the mass spectra and TIC of any given measurement. To avoid a greater numerical influence from either type of data (chromatogram or mass spectra) they are both normalized between 0 and 1 prior to the concatenation.

3.2. The Olive oil dataset

3.2.1. Introduction

Virgin olive oil has a characteristic flavor that distinguishes it from other edible vegetable oils. It is obtained from the olive fruit by mechanical processes only and no further treatments are required before consumption. The absence of refining processes helps preserve the organoleptic characteristics and the nutritional properties that distinguish virgin olive oil from other edible oils. Olive oil's characteristic aroma and, in particular, its green and fruity attributes depend on many volatile compounds derived from the degradation of polyunsaturated fatty acids through a chain of enzymatic reactions known as the lipoxygenase pathway, which occurs during the oil extraction process. [206]

Adulteration of olive oil with hazelnut oil has been a serious problem for regulatory agencies, oil suppliers and consumers. This fraudulent practice, arising due to the difference in economic value between the two oils, causes loses of millions of euros per year. Detection of hazelnut oil in admixtures with olive oil has always been very difficult to confirm using conventional approaches, especially at adulteration levels below 20%, due to the similarity of the two oils in parameters such as fatty acid and sterol content. [207].

Because of these illegal practices and the need to analyze the quality of the virgin olive oil, recent studies about various analytical methods for the examination of the volatile compounds of olive oils have emerged. In this way, a large number of components that contribute to the aroma of olive oil have been identified. Distillation methods have traditionally been applied in the analysis of plant materials. Steam distillation (SD), simultaneous distillation/extraction (SDE) and microwave-assisted extraction (MAE) were used for this purpose. [208]. Dynamic headspace techniques have been used to correlate the composition of the olive oil headspace to sensory attributes.

More recently, the solid-phase microextraction (SPME) technique has been introduced as an alternative to the dynamic headspace technique as a sample pre-concentration method prior to chromatographic analysis. This allowed the characterization of virgin olive oils form different varieties and regions. [209]

3.2.2. Goals

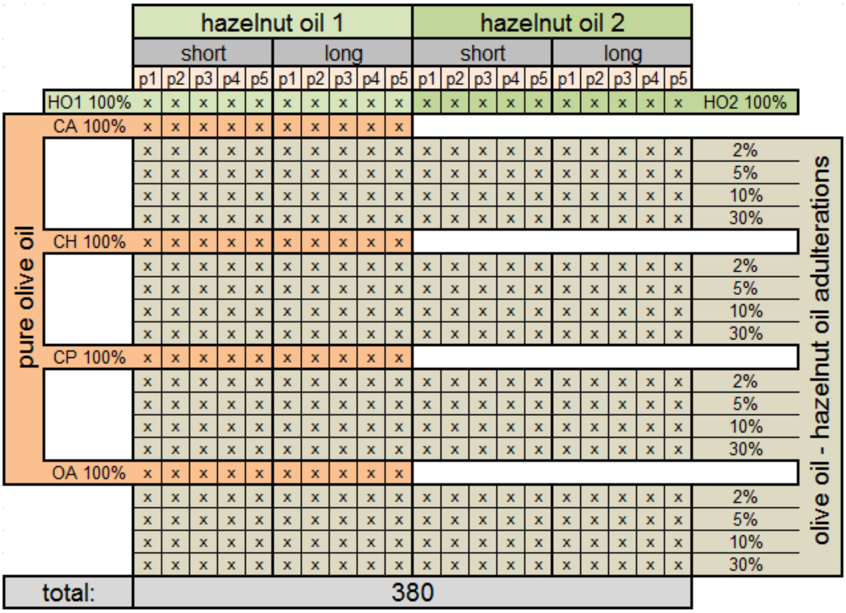
The main goals of this experiment are the following:

- to show that the addition of the third dimension is useful, and brings extra information, helping in the classification of pure and adulterated olive oil samples
- to optimize the chromatographic method in order to shorten the time consuming chromatographic analysis to a minimum, while still having acceptable results in fraud detection
- To compare 2-way and multi-way methods taking into account than 3-way data has to be re-arranged to be used by 2-way methods
- to go beyond the reported 20% detectable limit for the extra virgin olive oil - hazelnut oil adulterations

3.2.3. Materials and methods

Four commercial virgin olive oils, denoting the company and the olive fruit variety, were used in the analysis: Carrefour Arbequina, Carrefour Hojiblanca, Carrefour Picual and Oleaurum Arbequina. Two types of hazelnut oil were used to prepare adulterations of 30%, 10%, 5% and 2% of hazelnut oil in olive oil. The first hazelnut oil was regular hazelnut oil and the second one was ecologic hazelnut oil. Five repetitions were made for each preparation, giving a total of 190 samples per chromatographic method: 4 pure olive oils, 2 pure hazelnut oils and 32 olive oil - hazelnut oil adulterations, giving a total of 38 classes (

Table 3.2).



		hazelnut oil 1										hazelnut oil 2											
		short					long					short					long						
		p1	p2	p3	p4	p5	p1	p2	p3	p4	p5	p1	p2	p3	p4	p5	p1	p2	p3	p4	p5		
	HO1 100%	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	HO2 100%	
pure olive oil	CA 100%	x	x	x	x	x	x	x	x	x	x												
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	2%	olive oil - hazelnut oil adulterations
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	5%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	10%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	30%	
	CH 100%	x	x	x	x	x	x	x	x	x	x												
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	2%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	5%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	10%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	30%	
	CP 100%	x	x	x	x	x	x	x	x	x	x												
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	2%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	5%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	10%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	30%	
	OA 100%	x	x	x	x	x	x	x	x	x	x												
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	2%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	5%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	10%	
		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	30%	
total:		380																					

Table 3.2: Olive oil experimental data set

A particularity of the dataset is that in the second hazelnut oil adulterations and the pure second hazelnut oil dimethylphenol was introduced as internal standard.

The SPME extraction was carried out with two Divinylbenzene/Carboxen/ Polydimethylsiloxane (2cm-50/30um DVB/CAR/PDMS) (from Supelco) extraction fibers, one for each chromatographic method. Before use, the fiber was conditioned as recommended by the manufacturer. For the analysis 5 ml of olive oil - hazelnut oil mixture were placed into a 20 ml vial. One aliquot of one ml was put aside for the NMR analysis, and in the remaining 4 ml an internal standard in the form of di-methyl-phenol was added. The vials were sealed with teflon septum, and placed in a water bath held at 50°C under magnetic stirring. After the temperature equilibrium was reached the SPME fiber was exposed for 30 min to the sample headspace and immediately desorbed in the gas chromatograph injector.

Identification of compounds was performed by an Agilent 6890 GC System gas chromatograph coupled to quadrupole mass selective spectrometry using Agilent 5973 Mass Selective Detector. Analytes were separated on a Supelcowax-10 (Supelco) 30 m x 0.25 mm diameter, 0.25 μm film thickness column.

Two chromatographic methods were used. The first one was aimed at the complete separation of olive oil components and employed a temperature-programmed separation: The temperature was maintained at 40^{o}C for one minute, and then increased to 75^{o}C at 4^{o}C/min. At 75^{o}C the rate of climb was set to 8^{o}C/min until 140^{o}C, where the rate of climb was set one again to 20^{o}C/min until 240^{o}C where the chromatographic run was over, taking 22.88 minutes. The second chromatographyc method's objective was to obtain a coeluted peak, and that is why a 240^{o}C constant temperature separation was employed.

GC‑MS analysis in the complete scanning mode, within the 35-300 mass range, was performed to allow the identification of compounds in oil samples.

Chapter 4

Results and Discussion

4. Results and Discussion

4.1. Dimethylphenol isomers experiment

4.1.1. A fuzzy ARTMAP preview of the data

To test if any useful extra information is brought by the GC retention time we worked first with the mass spectra data and the chromatographic data. These two sets of data were pretreated by normalization, and fed into an in-house developed fuzzy ARTMAP algorithm, which tried to classify each solution. The network was tested using the leave-one-out cross-validation method: given n measurements, the network was trained n times using n-1 training vectors. The vector left out in the training phase was then used for testing. Performance was estimated as the average performance over the n tests. Therefore, for each iteration of the cross-validation process, a different measurement was left out. Both data matrices, the MS data and GC data, were normalized because the fuzzy ARTMAP network needs the input data to be between 0 and 1.

The results of the fuzzy ARTMAP neural network were displayed as a percentual success rate (**Figure 4.1**) and in the form of a confusion matrix.

Because for the MS approach we coeluted the chromatographic peaks through software by adding the time dimension, we would expect a similar result of the percentual success rate for the three chromatographic methods. The fact that method one gives the worst results, even though it is the most resolved chromatographic method can be attributed to the fact that the molecules fragmentate differently for a resolved chromatogram, where each separated isomer fragmentates, compared to a coeluted chromatogram where all the isomers arrive to the ionization chamber and fragmentate all together. Because of this summing the retention time is not the same as having a coeluted chromatogram in the first place. This fact explains why the results of the 2 isothermal methods give similar results.

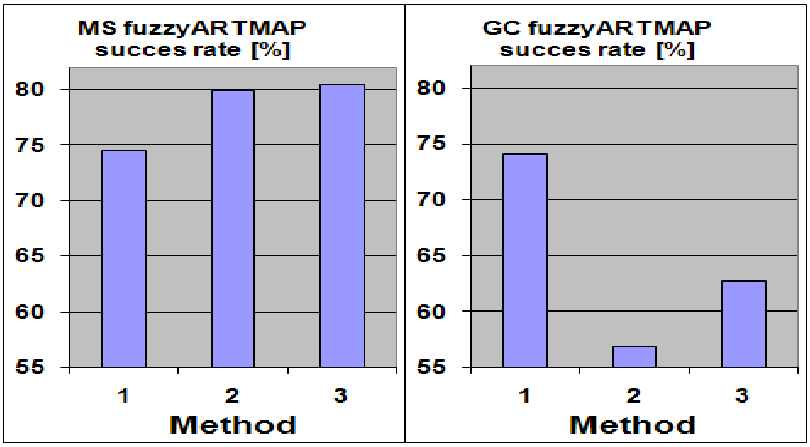
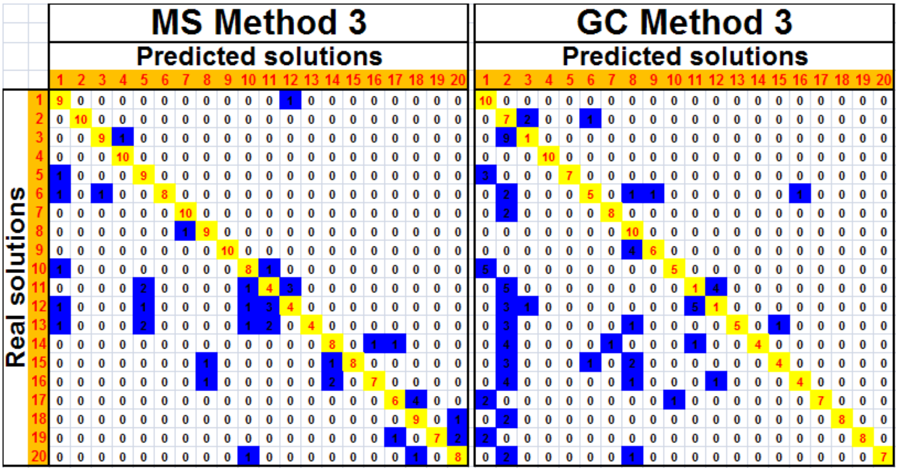


Figure 4.1: MS and GC fuzzy ARTMAP success rates for method 1 (least coeluted), method 2 (medium coeluted) and method 3 (most coeluted)

From the GC point of view, the first method, the one with a temperature programmed separation, in which almost all the isomers were separated gives better results than the isothermal methods 2 and 3, because the signal output of method one provides more information, allowing for a greater variability between samples. Again the classification success rates of the more coeluted methods 2 and 3 are closer between them.

The confusion matrix (**Figure 4.2**) shows the results of the fuzzy ARTMAP neural network as real solutions (rows) vs. the results predicted (columns) by the neural network. In a case of 100% success rate the diagonal of the matrix should present all 10 repetitions. In the case of the 20 solutions we can see that not all the solutions are predicted correctly, proving that this experiment is a challenge for the MS system.



MS Method 3

Predicted solutions

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Real solutions	1	9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	2	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	1	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	1	0	1	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0
	10	1	0	0	0	0	0	0	0	0	8	1	0	0	0	0	0	0	0	0	0
	11	0	0	0	0	2	0	0	0	0	1	4	3	0	0	0	0	0	0	0	0
	12	1	0	0	0	1	0	0	0	0	1	3	4	0	0	0	0	0	0	0	0
	13	1	0	0	0	2	0	0	0	0	1	2	0	4	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	1	1	0	0	0
	15	0	0	0	0	0	0	0	1	0	0	0	0	0	1	8	0	0	0	0	0
	16	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	7	0	0	0	0
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	4	0	0
	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	1
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	7	2
	20	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	8	

GC Method 3

Predicted solutions

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Real solutions	1	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	7	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	3	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	2	0	0	0	5	0	1	1	0	0	0	0	0	0	1	0	0	0	0
	7	0	2	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	4	6	0	0	0	0	0	0	0	0	0	0	0
	10	5	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
	11	0	5	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0
	12	0	3	1	0	0	0	0	0	0	0	5	1	0	0	0	0	0	0	0	0
	13	0	3	0	0	0	0	0	1	0	0	0	0	5	0	1	0	0	0	0	0
	14	0	4	0	0	0	0	1	0	0	0	1	0	0	4	0	0	0	0	0	0
	15	0	3	0	0	0	1	0	2	0	0	0	0	0	0	4	0	0	0	0	0
	16	0	4	0	0	0	0	0	1	0	0	0	1	0	0	0	4	0	0	0	0
	17	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7	0	0	0
	18	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0
	19	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0
	20	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	7

Figure 4.2: Fuzzy ARTMAP results for the shortest method 3 mass spectra approach (left) and chromatographic approach (right) (real (y) vs. predicted (x) solution)

The results show a better performance for the MS approach than for the GC data (**Figure 4.2**), which happens because the MS fingerprint is more specific than the GC fingerprint. We also see a difference between the chromatographic methods, which indicates that a more coeluted pick is performing better for the MS sensor.

Comparing the fuzzy ARTMAP results we can see (**Figure 4.2**) that the errors on the confusion matrix are different in the MS dimension than in the GC dimension. This is a clear sign that using both information dimensions (time and m/z) should improve the classifying ability of the MS-based electronic nose.

From this point the data can be analyzed in 2 formats: the two dimensional data through multivariate data analysis like PCA, PLS and fuzzy ARTMP, and three-dimensional data through multi-way data analysis like PARAFAC and n-PLS.

4.1.2. Data preprocessing

Prior to these steps we preprocessed the data through peak alignment, normalization and mean centering. These preprocessing methods were evaluated to see which one gave best results with the data generated by the GC-MS configuration.

For the peak alignment Recursive Alignment through Fast Fourier Transform (RAFFT) was employed, aligning the data in the three dimensional matrix. The alignment was done for each m/z chromatogram separately (**Figure 4.3**). After peak alignment the 3D data matrix was transformed in five different matrices: 3d, MS, GC, unfolded and concatenated.

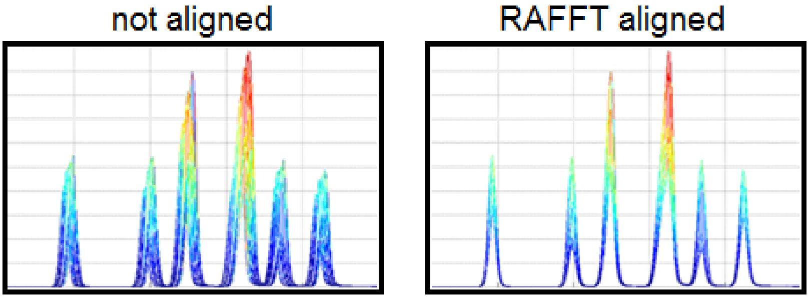


Figure 4.3: TIC of several measurements: unaligned and RAFFT aligned

Normalization was applied using three different approaches:

- In the first approach, the intensity points of each measurement were normalized between 0 and 1.
- In order to avoid instrumental and injection error, a second normalization was also evaluated by an internal standard, the benzene peak area. The concentration of benzene being the same in each sample, the intensities of the other compounds are relative to benzene.
- Finally, in a third approach, the data was mean centered following the sample direction (columns) in all 5 data sets. At each column, the mean value was calculated and subtracted from all the points of the column.

These approaches allowed us to perform a complete study and comparison between each type of pre-processing used: Alignment vs. raw data, Unity normalization vs. benzene normalization and raw data vs. mean centering.

From the three chromatographic methods made we concentrated our attention on the most resolved and the most coeluted one (methods 1 and 3, respectively). Both had some advantages and some drawbacks. Method one, aimed at obtaining resolved chromatograms, extracts more information at the expense of analysis time, while method three, aimed at saving time, generated more coeluted peaks, because of a much shorter time of analysis and therefore less data, presents a higher challenge to the classification algorithms.

2-D datasets were analyzed by Principal Component Analysis (PCA), Partial Least Squares Discriminant Analysis (PLS-DA) and with an in-house developed fuzzy ARTMAP algorithm, while the 3-D dataset was analyzed by Parallel Factor Analysis (PARAFAC), and multi-way PLS-DA (n-PLS-DA).

4.1.3. PCA and PARAFAC

Different PCA and PARAFAC projections were performed using the multivariate GC-MS data obtained. The goal was to see if measurements from different solutions were separated and repetitions of each mixture were clustered together.

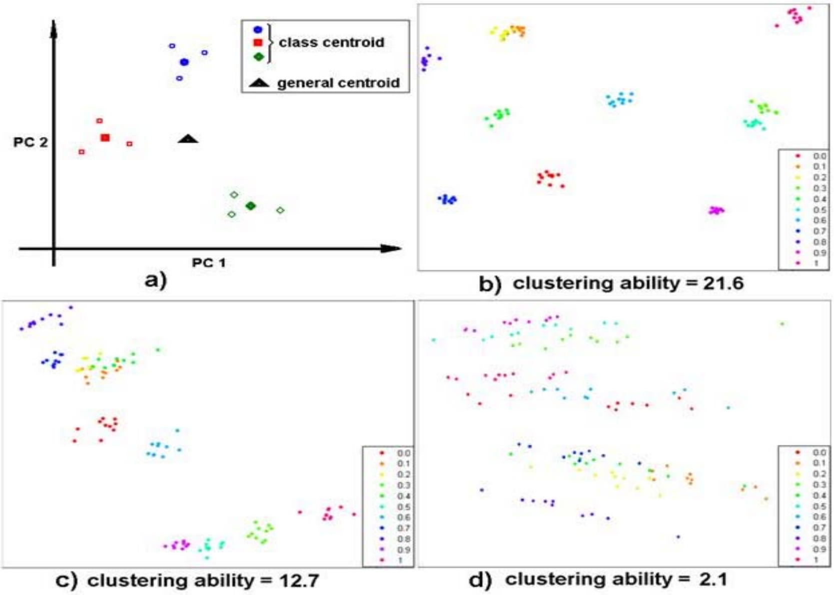


Figure 4.4: a) clustering ability; b), c) and d) show how cluster ability works. The cluster ability index increases as cluster are better defined in PCA or PARAFAC

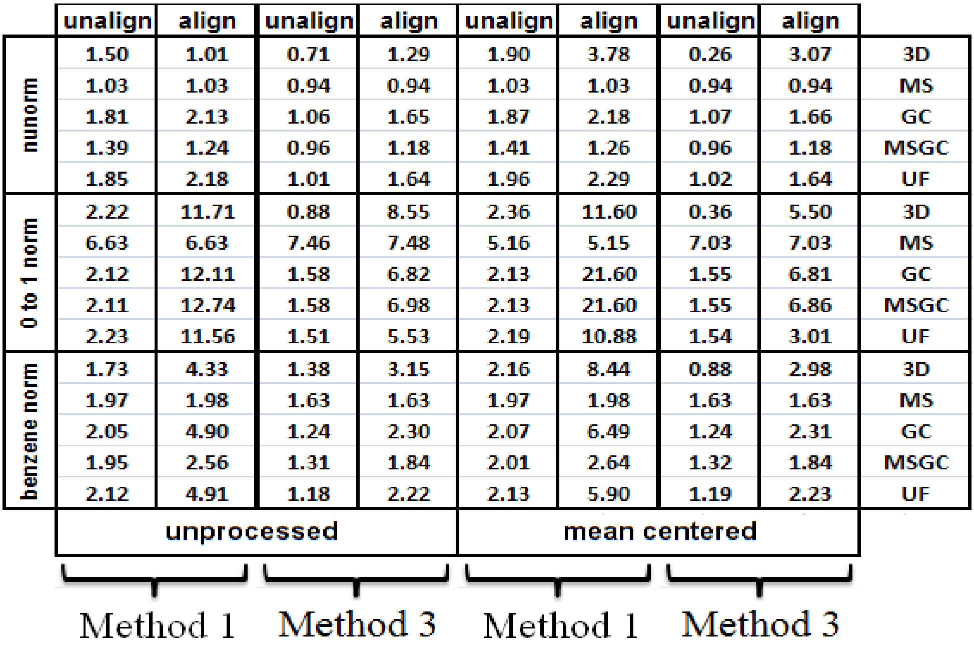
Since these projections are represented with graphics, an objective parameter to see how well the measurements were separated and clustered was defined (**Figure 4.4**). This parameter was a ratio between two concepts; the intravariance (Eq. 4.3), being the first, is a measure of how close are the points of one particular class of objects to each other. It represents the mean distance from each measurement point to the center of its class (class centroid (Eq. 4.1)). A lower value implies good repetivity between measurements, whereas higher values would indicate drift or noise. The second factor is the intervariance (Eq. 4.6). This value is calculated for the entire measurement set, giving the mean distance between the centroids of each class and the centroid for the entire dataset (general centroid (Eq. 4.4)). Higher values imply better discrimination performance. With these two values, the measure of how well the PCA or PARAFAC clusters measurements into different clusters for different solutions was calculated as the ratio of intervariance to mean intravariance. Higher values mean better resolution and/or reproducibility, while lower values means poorer resolution and/or reproducibility (**Table 4.2**).

$\boldsymbol{C_c = \frac{\sum_{i=1}^{R} S_i}{R}}$	Eq. 4.1	Legend:
		S = scores
		R = repetitions
		C_c = class centroid
		d = distance point to centroid
		C = number of classes
		Intra = intravariance
		C_g = general centroid
		D = distance C_c to C_g
		Inter = intervariance
		n = number of principal components

$\boldsymbol{d = \sum_{i=1}^{n} (\lvert S_i - C_i \rvert)}$	Eq. 4.2	
$\boldsymbol{Intra = \frac{\sum_{i=1}^{M} S_i}{M}}$	Eq. 4.3	
$\boldsymbol{C_g = \frac{\sum_{i=1}^{M} S_i}{M}}$	Eq. 4.4	
$\boldsymbol{D = \sum_{i=1}^{n} \left(\left\lvert C_{c_i} - C_{g_i} \right\rvert \right)}$	Eq. 4.5	
$\boldsymbol{Inter = \frac{\sum_{i=1}^{C} D_i}{C}}$	Eq. 4.6	
$\textbf{cluster ability} = \frac{\textbf{Inter}}{\textbf{mean(Intra)}}$	Eq. 4.7	

Table 4.1: Cluster ability equation table

As expected, method one, where the chromatogram was well resolved gave better values than method three, the shorter GC time and more coeluted analysis. Best classification results were obtained for method one, with the GC and MSGC concatenated matrix when the data is aligned, 0 to 1 normalized and mean centered. By contrast method 3, the most difficult one from the point of view of chromatographic separation, gives better results using PARAFAC when the data is aligned, normalized between 0 and 1 and not mean centered.



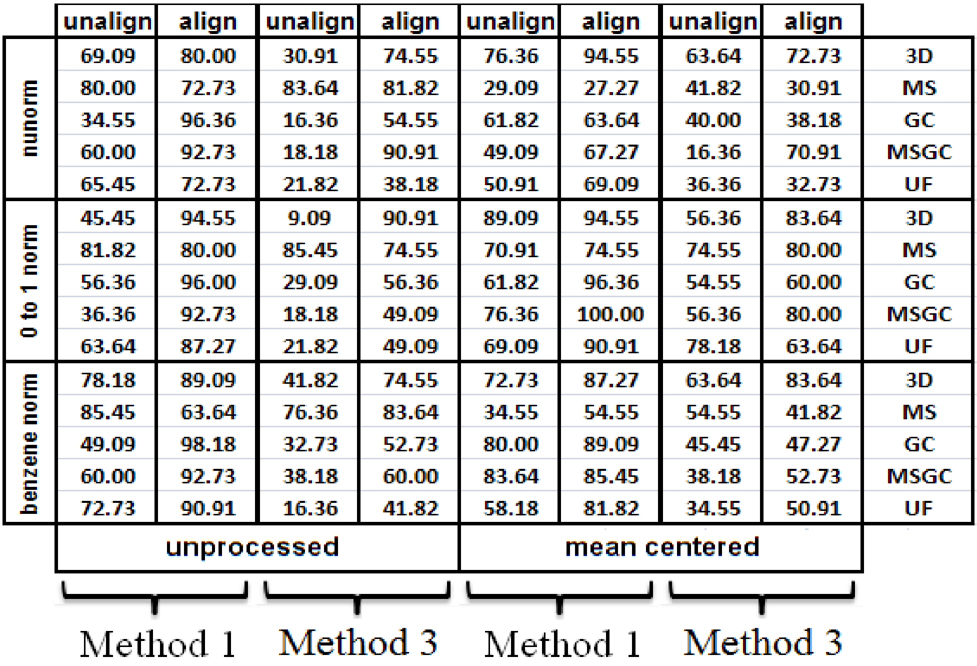
	unprocessed				mean centered				
	Method 1		Method 3		Method 1		Method 3		
	unalign	align	unalign	align	unalign	align	unalign	align	
nunorm	1.50	1.01	0.71	1.29	1.90	3.78	0.26	3.07	3D
	1.03	1.03	0.94	0.94	1.03	1.03	0.94	0.94	MS
	1.81	2.13	1.06	1.65	1.87	2.18	1.07	1.66	GC
	1.39	1.24	0.96	1.18	1.41	1.26	0.96	1.18	MSGC
	1.85	2.18	1.01	1.64	1.96	2.29	1.02	1.64	UF
0 to 1 norm	2.22	11.71	0.88	8.55	2.36	11.60	0.36	5.50	3D
	6.63	6.63	7.46	7.48	5.16	5.15	7.03	7.03	MS
	2.12	12.11	1.58	6.82	2.13	21.60	1.55	6.81	GC
	2.11	12.74	1.58	6.98	2.13	21.60	1.55	6.86	MSGC
	2.23	11.56	1.51	5.53	2.19	10.88	1.54	3.01	UF
benzene norm	1.73	4.33	1.38	3.15	2.16	8.44	0.88	2.98	3D
	1.97	1.98	1.63	1.63	1.97	1.98	1.63	1.63	MS
	2.05	4.90	1.24	2.30	2.07	6.49	1.24	2.31	GC
	1.95	2.56	1.31	1.84	2.01	2.64	1.32	1.84	MSGC
	2.12	4.91	1.18	2.22	2.13	5.90	1.19	2.23	UF

Table 4.2: PARAFAC (3D) and PCA (MS, GC, MSGC, and UF) Cluster capability. The merit figure is the ratio intervariance/intravariance. Higher values mean better reproducibility and/or better selectivity

A first conclusion is that aligning the data to counteract chromatographic shift and normalizing it between 0 and 1 improves classification accuracy in most of the possible scenarios. An interesting result is that in the case of the most resolved chromatograms, 2D algorithms work better, whereas in a more complicated case (more coeluted peaks) they lose effectiveness in front of 3D methods, which give better results.

For the Fuzzy ARTMAP classification using PCA or PARAFAC coordinates the data was separated into training and testing data sets, using a 50-50 training-testing ratio. After applying PCA and PARAFAC for the training set the scores were normalized between 0 and 1 and fed into a fuzzy ARTMAP neural network for training. The test set was then projected on to the PCA/PARAFAC model, and the scores were also normalized between 0 and 1 and used as the test set for the already trained fuzzy ARTMAP neural network.

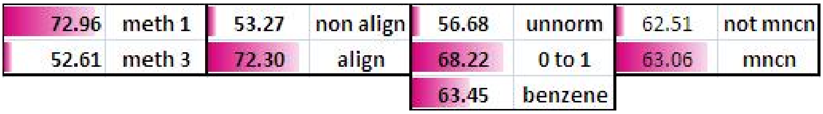
The results were judged by means of success rate of the confusion matrix (**Table 4.3**). The results are in good agreement with the cluster analysis, showing that method 1 presents better results when two-way analysis methods are employed while method 3 gives better results with three-way data analysis is used. For method 1 the best results were obtained by the PCA projection of the MSGC concatenated matrix when the data was aligned, mean centered and normalized between 0 and 1, scoring 100% success rate, followed by GC data with a 96.36% success rate. In the case of the shortest method the best result was obtained using PARAFAC when the data was aligned, normalized from 0 to 1 and not mean centered, as well as using the PCA projection of the MSGC matrix after alignment.



	unalign	align	unalign	align	unalign	align	unalign	align	
nunorm	69.09	80.00	30.91	74.55	76.36	94.55	63.64	72.73	3D
	80.00	72.73	83.64	81.82	29.09	27.27	41.82	30.91	MS
	34.55	96.36	16.36	54.55	61.82	63.64	40.00	38.18	GC
	60.00	92.73	18.18	90.91	49.09	67.27	16.36	70.91	MSGC
	65.45	72.73	21.82	38.18	50.91	69.09	36.36	32.73	UF
0 to 1 norm	45.45	94.55	9.09	90.91	89.09	94.55	56.36	83.64	3D
	81.82	80.00	85.45	74.55	70.91	74.55	74.55	80.00	MS
	56.36	96.00	29.09	56.36	61.82	96.36	54.55	60.00	GC
	36.36	92.73	18.18	49.09	76.36	100.00	56.36	80.00	MSGC
	63.64	87.27	21.82	49.09	69.09	90.91	78.18	63.64	UF
benzene norm	78.18	89.09	41.82	74.55	72.73	87.27	63.64	83.64	3D
	85.45	63.64	76.36	83.64	34.55	54.55	54.55	41.82	MS
	49.09	98.18	32.73	52.73	80.00	89.09	45.45	47.27	GC
	60.00	92.73	38.18	60.00	83.64	85.45	38.18	52.73	MSGC
	72.73	90.91	16.36	41.82	58.18	81.82	34.55	50.91	UF
	unprocessed				mean centered				
	Method 1		Method 3		Method 1		Method 3		

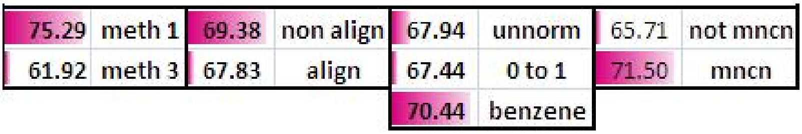
Table 4.3: Fuzzy ARTMAP prediction success rate for PARAFAC (3D) and PCA (MS, GC, MSGC and UF) projections

Averaging the success rate for each method shows that aligning, mean centering and normalizing the data between 0 and 1 gives the best results, indicating that this is the best way to preprocess the data.(**Table 4.4** and **Table 4.5**)



72.96	meth 1	53.27	non align	56.68	unnorm	62.51	not mncn
52.61	meth 3	72.30	align	68.22	0 to 1	63.06	mncn
				63.45	benzene		

Table 4.4: : averages for PCA and PARAFAC data showing the best performances



75.29	meth 1	69.38	non align	67.94	unnorm	65.71	not mncn
61.92	meth 3	67.83	align	67.44	0 to 1	71.50	mncn
				70.44	benzene		

Table 4.5: averages for PLS-DA and n-PLS results showing the best performances

4.1.4. PLS-DA and n-PLS-DA

For PLS-DA, and n-PLS-DA analysis, each dataset was separated into training and testing sets by a chosen training/test ratio of 7/3, rather than 5/5 like in the PCA – PARAFAC analysis, since this split gave better results on the PLS-DA – n-PLS-DA. The training set was used to train the PLS-DA and n-PLS-DA models. Onto these models the test set was projected.

The model and prediction performance were evaluated by means of the Root Mean Square Error of Cross-Validation (RMSECV) calculated using the training measurements, and the Root Mean Square Error of Prediction (RMSEP), respectively.

The data proved difficult to classify and predict if all the 20 different mixtures were analyzed. Because of this, the first 9 solutions were left out and we used the most challenging, mixtures 10 to 20. The overall prediction performance was assessed by means of sample prediction success rate.

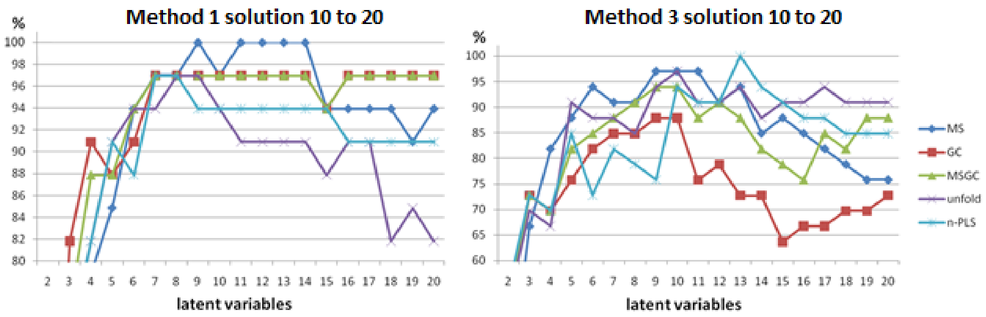


Figure 4.5: PLS-DA and n-PLS-DA prediction success rate for method 1 (left) and method 3 (right)

In **Figure 4.5** (left) and (right) the graphs represent results present the sample prediction success rate for methods one and three for solutions 10 to 20. It can be noticed that method 3, in which peaks are more coeluted, gives worse results than method 1, where a temperature programmed separation was employed in order to achieve a better separation. We can clearly notice this in the chromatographic (GC) success rate line.

In method 1, the time retention time does not increase the resolution of the device compared to the regular chromatogram or even the mass spectra. Therefore, in well-resolved chromatograms (that are also time-consuming) 3D data methods or 2D data are not increasing the resolution of the system, we can even say that the third dimension, in general, does not help to improve results.

On the other hand, in the more challenging and coeluted method 3, where measurement time is kept to a minimum, the retention time information gives additional information and the n-PLS-DA multi-way method presented the best results giving 100% success rate classification for 13 latent variables (**Figure 4.5** (right)). As expected, GC information alone does not give enough information.

Even though the PLS-DA and n-PLS-DA results are not as conclusive as the PCA and PARAFAC results, the differences being minimal, the multi-way PLS model yields a high prediction success rate in both datasets. It is also recommended over the unfolded and concatenated data because it builds a simpler and more parsimonious model.

4.1.5. Incorporating the third dimension using a novel Fuzzy ARTMAP approach

Fuzzy ARTMAP neural networks present a different approach to the question if the third dimension of a chromatographic separation brings new information towards the improvement of the MS-based electronic nose. In this section we left out the traditional two and three-way approaches and we focused on neural networks.

To improve the results of the fuzzy ARTMAP neural network, which was applied on specific mass spectra and chromatographic data, a voting strategy was implemented. The 10 repetitions of each solution were scrambled, and each solution group was divided into training and testing measurements by a given training-evaluation ratio. The number of measurements used for training ranged between 1 and 9. These data was fed into a fuzzy ARTMAP neural network by a given number (10) of times (votes) and results were recorded. The winning class was decided by the total number of votes each solution received, so that the one with more votes was selected as the output since it was the most probable prediction. Initially, the input data for the fuzzy ARTMAP voting strategy were the average mass spectra (MS) and total ion chromatogram (GC) matrices. The matrices were normalized between 0 and 1.

In order to take advantage of both data types we followed 2 approaches: 1) Concatenation of the MS and GC matrices into a single matrix called MSGC and 2) by summing the MS votes with the GC votes for a given measurement.

By combining the mass spectra and chromatographic information through the voting approach, the MS and GC matrix were fed into the fuzzy ARTMAP independently and then the votes that they had collected were added and the winning class was decided.

Possible outputs from the algorithm were correctly classified, wrongly classified or unclassified, when the number of votes for one solution was the same with the number of votes for another solution. Because of the high number of unclassified votes in the combined MS and GC voting method, and because the MS proved to give better results, we decided that in case of a misclassification (tie) in the voting strategy of the MS and GC matrices the winner should be the MS winner (MSGC vote MS improved).

The results were monitored by the success rate based on the confusion matrix (real solution vs. predicted solution).

Looking at **Figure 4.6** we can see that the MSGCvoteMS results are always on top of the other approaches if 1 to 8 measurements are used for training. Because in practice a high number a measurements is hard to obtain this shows that the MSGCvoteMS approach is the best one, giving the best results when fewer measurements are used for training.

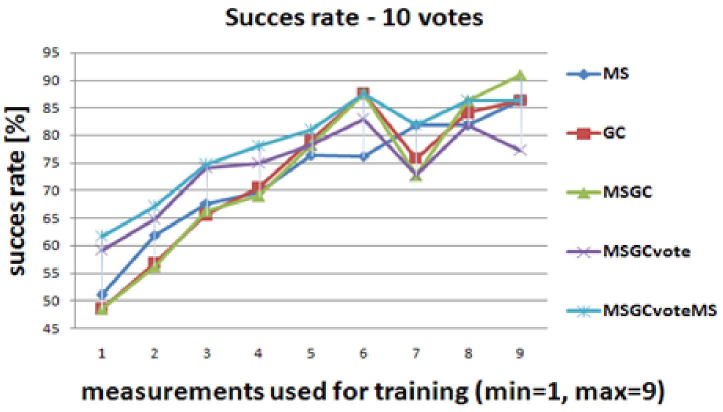


Figure 4.6: Voting strategy Fuzzy ARTMAP success rates

Using 9 of the 10 measurements for training leads to worse results because there is only one measurement left for evaluation, and even a few misclassified samples would have a great impact on the results, leading to a lower success rate. That's why we recommend that the measurements used for training should be between 50 to 60 percent.

The highest success rates, in the range of 90%, were obtained by the MSGC vote MS improved method, which takes advantage of both methods, but does not have the drawback of the MSGC vote method, that misclassifies too many samples.

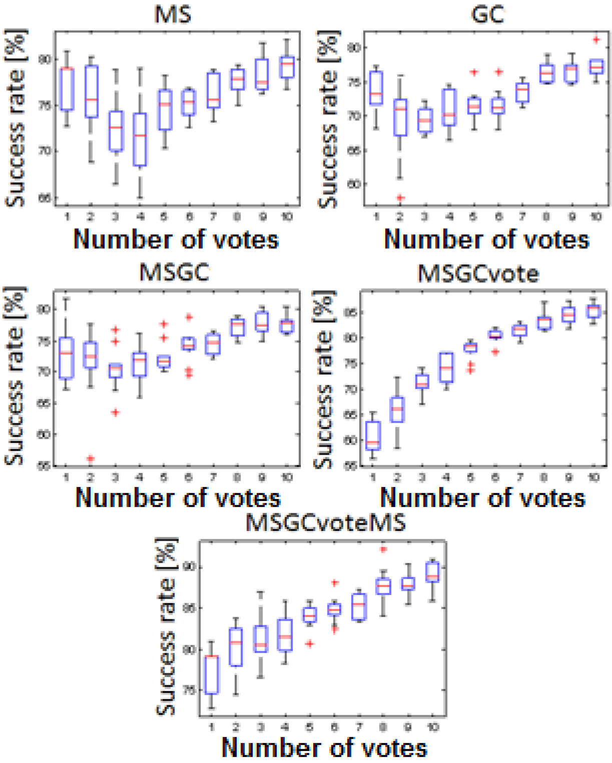


Figure 4.7: Success rates for MS, GC, MSGC, MSGCvote and MSGCvoteMS for 5 train – 5 evaluation measurements, box plot for 10 repetitions

Repeating the analysis 10 times for each approach, using 5 measurements for training and 5 for evaluation gives us the chance to analyze the consistency of the results through box and whisker plots. This statistic analysis (**Figure 4.7**) shows that the MS and GC vote approaches are not very consistent, giving a wide range of success rates. The most consistent method is MSGCvote, because by improving the MSGCvote results by choosing the MS result in case of a misclassification, errors can be introduced by a misclassification of the MS approach. Even though not as consistent as the MSGCvote approach, overall, the best success rates are obtained by the MSGCvoteMS method.

The variation of the vote number and train-evaluation ratio parameters showed that the results are better with an increase in the number of votes.

4.2. The Olive oil adulteration experiment

The data collected in the experiment detailed in section 3.2.2 was analyzed by means of PCA, PARAFAC, PLS-DA and n-PLS and its performance displayed as a clusterization capability merit figure and classification success rates respectively.

4.2.1. Data pretreatment

In order to be treated with PCA and PLS-DA (both 2-way methods) the original 3D data was arranged in 2D format by summing the time dimension, giving the average mass spectrum (the MS spectra) or by summing the m/z dimensions giving the average retention time (the time-dependent "TIC" chromatogram o GC). In order to use the two-way data analysis algorithms and to use the information brought by the retention time and m/z dimensions at the same time, the two previous two-dimensional sets were concatenated giving the MSGC set. Another way of having all the 3D information for the two-dimensional methods (PCA and PLS-DA) was unfolding (UF) the 3D matrix by pasting one mass spectra one after the other for each point obtained in the retention time.

First we looked at the data normalization issue, and for that we first applied a PCA algorithm (results are summarized in **Figure 4.8** and **Figure 4.9**) and then the PLS-DA (**Figure 4.10**) algorithm to the measurements that had an internal standard (the measurements done with the ecologic hazelnut oil). A total of 17 classes were considered: pure hazelnut oil and adulteration mixtures containing extra virgin olive oil (Carrefour Arbequina, Carrefour Hojiblanca, Carrefour Picual, Oleaurum Arbequina) and 30%, 10%, 5%, and 2% of hazelnut oil .

Three normalization approaches were evaluated: no normalization, normalization by the internal standard peak height, and normalization by the internal standard peak volume.

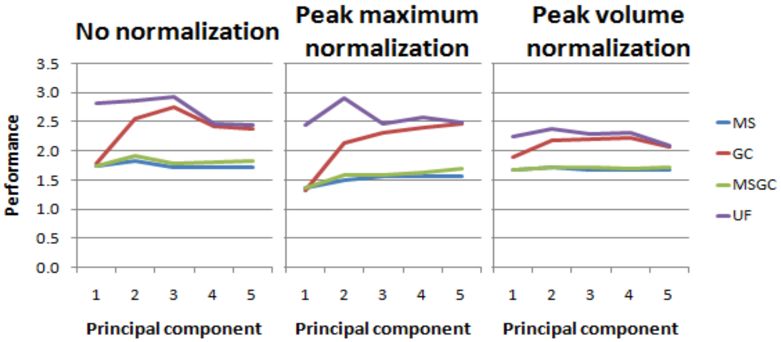


Figure 4.8: Comparison between different normalization strategies for MS, GC, MSGC and UF data treated with PCA (Performance indices - higher means better) - Long CG method

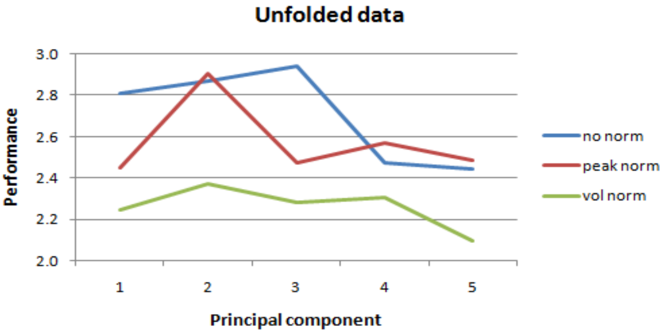


Figure 4.9: Comparison between normalization strategies for unfolded data treated with PCA (Performance indices - higher means better) - Long method

Both Normalizations, by internal standard peak maximum and internal standard peak volume were done using the three dimensional matrix. For the maximum peak normalization the 3d data of each measurement was divided by the peak maximum, while when normalizing by peak volume we divided it to the whole three dimensional peak. This is the same as normalizing by the total ion chromatogram peak area. Only after normalization was done the data was transformed into MS and GC matrices.

The PCA projections were assessed by means of clusterization values described in section 4.1.3, while the performance of PLS-DA predictions were described by their success rate.

Looking at the PCA clusterization indices we can see very little change for the MS and MSGC data regardless of the normalization strategy used (**Figure 4.8**). On the other hand, for the GC and even more for the unfolded data (**Figure 4.9**) normalizing by the internal standard peak volume yields the worst results, affirmation sustained also by the PLS-DA analysis (**Figure 4.10**). Choosing between peak maximum normalization and no normalization seemed to have little effect for PCA analysis.

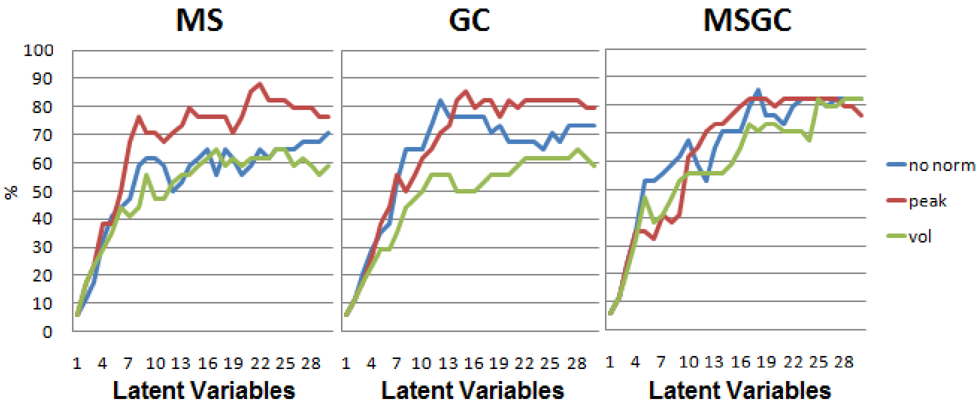


Figure 4.10: Comparison between normalization strategies for MS, GC and MSGC data treated with PLS-DA - Long method

Using PLS-DA as the assessment tool for different normalization approaches we can see that for the MS data the best strategy would be to normalize by the peak maximum, while for GC data there isn't much difference between peak maximum normalization and no normalization approaches. By combining both the MS and GC information through concatenation we see that all normalizations approaches yield similar results.

Figure 4.10 clearly show that taking into consideration all the mass spectra and chromatographic information improves the results. As the normalizations are made when the data is in three dimensional format, "collapsing" the all the data into the MS and GC matrices through summation leads to a loss of information. We sustain our affirmation pointing out that, when the data is normalized by peak volume or normalization is not applied, the prediction success rates improve for the MSGC concatenated matrix, compared to the MS and GC matrices alone, where the data is incomplete.

4.2.2. PCA and PARAFAC analysis

The PCA and the PARAFAC results were evaluated by means of the clusterization capability described in previous sections.

First, we analyzed all the 38 classes (two pure hazelnut oils, four pure olive oils and 32 adulterations between them). Since only the ecologic hazelnut oils contained the internal standard, all the data was cropped just before its peak (**Figure 4.11**).

We can clearly see from this analysis that while the MS and MSGC have the worst performance, the three dimensional data (PARAFAC projection) has the best performance, followed closely by the 2D unfolded data, outperforming the GC data. This holds true for both the long and the short chromatographic methods. This proves that there is a lack of information when looking just at the mass spectra data as in the MS-based electronic nose, compared with the complete information of three dimensional data or its unfolded version.

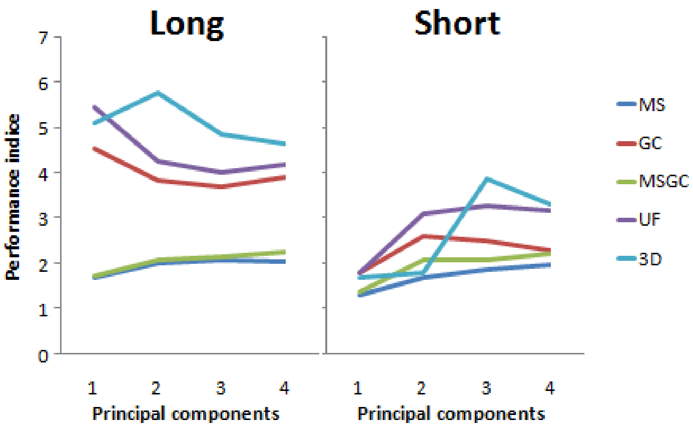


Figure 4.11: PCA and PARAFAC performance indices for all the 38 classes for the Long and Short chromatographic method

From the optimization point of view, we can say that, for the short method, working with all the data, in its three dimensional or its unfolded form, yields better results than the mass spectra data for the large chromatographic method. This is an important conclusion as a shorter chromatographic method can be used to improve the mass spectrometry-based electronic nose at the expense of only 5 minutes worth of chromatographic partial separation.

The data was also analyzed in smaller subsets to detail how the adulteration was detected with different types of olive oil.

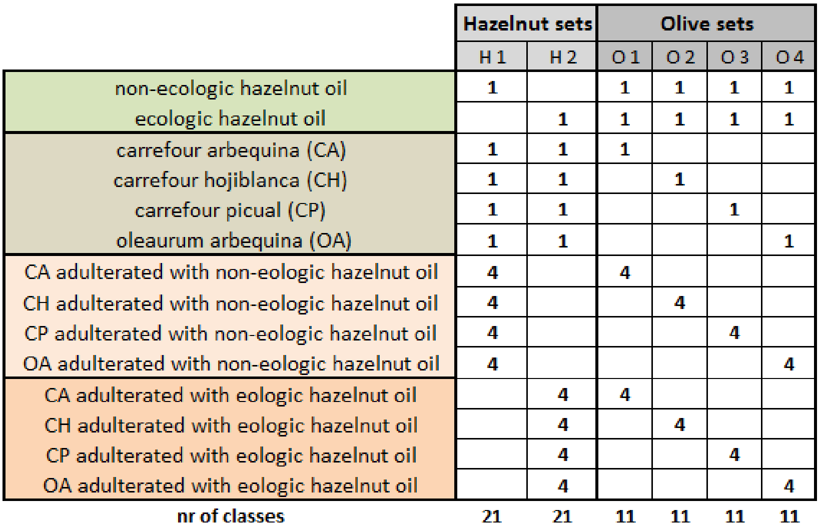
The first subset (the Hazelnut oils subset) looked at the samples involving each of the two types of hazelnut oil, while the second subset took into consideration just the measurements involving each type of olive oil (**Table 4.6**).

The two hazelnut oil sets contained 21 classes each:

- H1 - the pure non-ecologic hazelnut oil, the four pure extra virgin olive oils and their adulterations of 2%,5%,10% and 30% with the non-ecologic hazelnut oil
- H2 - the pure ecologic hazelnut oil, the four pure extra virgin olive oils and their adulterations of 2%,5%,10% and 30% with the ecologic hazelnut oil

The four olive oil sets contained fewer measurements to a total of 11 classes per set:

- O1: - the pure ecologic and non-ecologic hazelnut oil, the pure Carrefour Arbequina olive oil, and its adulterations with both hazelnut oils
- O2: - the pure ecologic and non-ecologic hazelnut oil, the pure Carrefour Hojiblanca olive oil, and its adulterations with both hazelnut oils
- O3: - the pure ecologic and non-ecologic hazelnut oil, the pure Carrefour Picual olive oil, and its adulterations with both hazelnut oils
- O4: - the pure ecologic and non-ecologic hazelnut oil, the pure Oleaurum Arbequina olive oil, and its adulterations with both hazelnut oils



	Hazelnut sets		Olive sets			
	H1	H2	O1	O2	O3	O4
non-ecologic hazelnut oil	1		1	1	1	1
ecologic hazelnut oil		1	1	1	1	1
carrefour arbequina (CA)	1	1	1			
carrefour hojiblanca (CH)	1	1		1		
carrefour picual (CP)	1	1			1	
oleaurum arbequina (OA)	1	1				1
CA adulterated with non-eologic hazelnut oil	4		4			
CH adulterated with non-eologic hazelnut oil	4			4		
CP adulterated with non-eologic hazelnut oil	4				4	
OA adulterated with non-eologic hazelnut oil	4					4
CA adulterated with eologic hazelnut oil		4	4			
CH adulterated with eologic hazelnut oil		4		4		
CP adulterated with eologic hazelnut oil		4			4	
OA adulterated with eologic hazelnut oil		4				4
nr of classes	21	21	11	11	11	11

Table 4.6: The six subsets (two for each hazelnut oil and four for each olive oil) were independently analyzed with PCA and PARAFAC

The results show the same trend as seen when the PCA and PARAFAC was applied for the whole dataset (**Figure 4.12**,**Figure 4.13** and **Figure 4.14**). The methods that take into consideration all the information available in the data outperform the ones that do not. Between applying the PARAFAC for the three-dimensional data and applying PCA for the unfolded 3d data, PARAFAC (“3D”) would be the better choice, based on the results and the fact that the interpretation is easier.

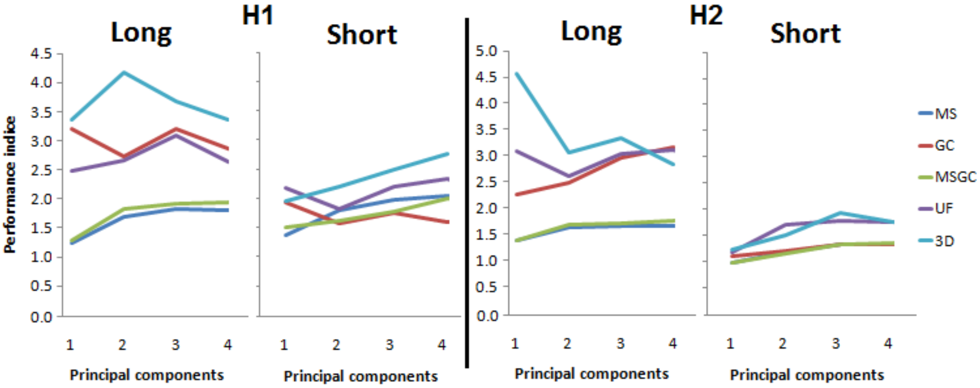


Figure 4.12: PCA and PARAFAC performance indices for H1 - non-ecologic hazelnut oil subset (left) and H2 - ecologic hazelnut oil subset (right)

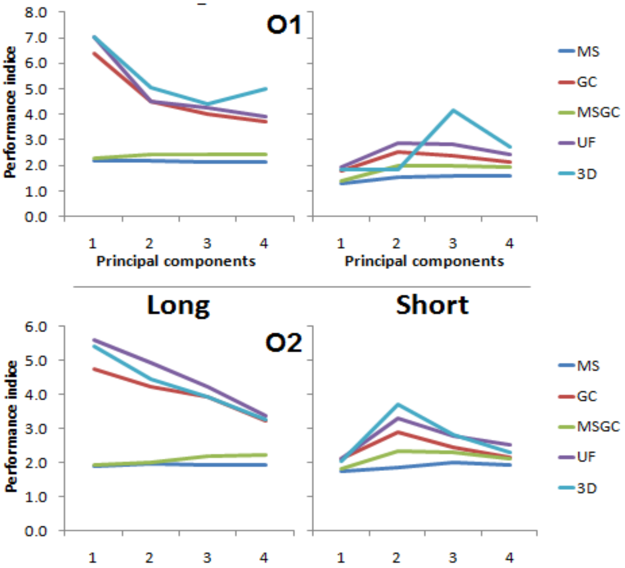


Figure 4.13: PCA and PARAFAC performance indices for O1 - Carrefour Arbequina oil subset (up) and O2 - Carrefour Hojiblanca oil subset (down)

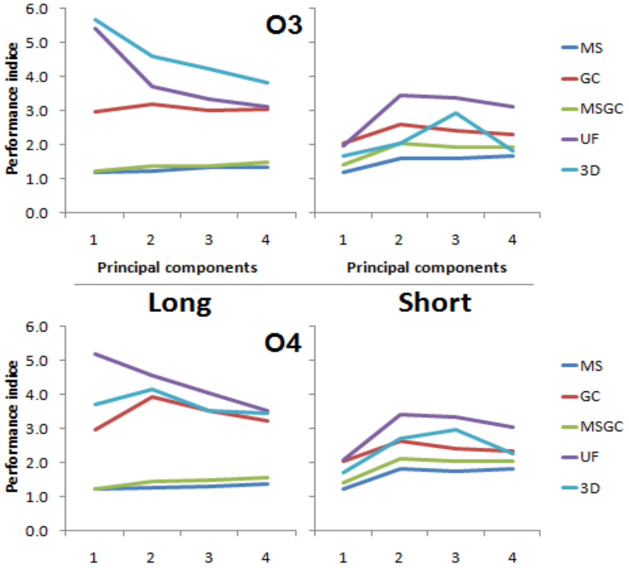


Figure 4.14: PCA and PARAFAC performance indices for O3 - Carrefour Picual oil subset (up) and O4 - Oleaurum Arbequina oil subset (down)

For all the PCA and PARAFAC data analysis peak normalization was not used, and since not all the samples contained it, the internal standard was left out.

4.2.3. PLS-DA and n-PLS analysis

Because of time consuming computations for the PLS-DA and n-PLS we looked at just the 21 classes associated with the ecologic hazelnut oil: the four pure extra virgin olive oils, the pure hazelnut oil and all their 16 adulterations (four pear each olive oil: 2%, 5%, 10% and 30%). The success rates of PLS-DA and n-PLS were compared for both the long and the short chromatographic run.(

Figure 4.15)

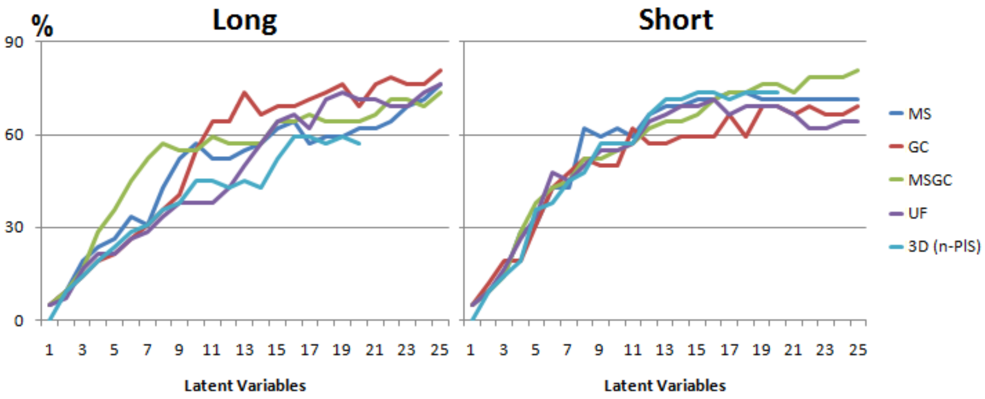


Figure 4.15: PLS-DA and n-PLS success rates for the Long and Short column for all the 21 classes.

The data was not treated for normalization and the internal standard was not taken into consideration for the same reasons stated before.

For the long method we can see that the best results are given by the analysis of just the chromatographic information from the data, while the mass spectra information gives weaker prediction success rates, with the MSGC falling in between. For this analysis the methods that traditionally combine the whole data (the three-way and the unfolded data) give the worst results suggesting that when the chromatogram is resolved two-way data methods give the best results and 3D information does not improve prediction.

On the other hand, looking at the short method, because of the lack of information found in the coeluted peak, the analysis of the GC data gives the worst prediction success rates, while mass spectrometry information outperforms it. The MSGC concatenated information on the other hand shows results better than both of them. Treating the data as a whole, n-PLS gives similar results with the PLS-DA analysis of MSGC concatenated data for the same number of latent variables. Unfolding the data, and analyzing it by means of PLS-DA, even though follows MS, MSGC and 3D data quite close until 16 latent variables does not keep the high prediction success rates for higher latent variables.

Comparing the two chromatographic methods best results we can see that the chromatographic information of the long column gives similar results with the three dimensional and concatenated matrix analysis for the short column.

This case is very interesting because the PLS approach proves that the addition of the extra information improves the results in the short method and, moreover, by using the short column and the both GC and MS information, concatenated or as a three dimensional matrix, we can achieve the same results as using the long chromatographic method (**Figure 4.16**) which takes much more time (22.9 minutes compared to just 5 minutes).

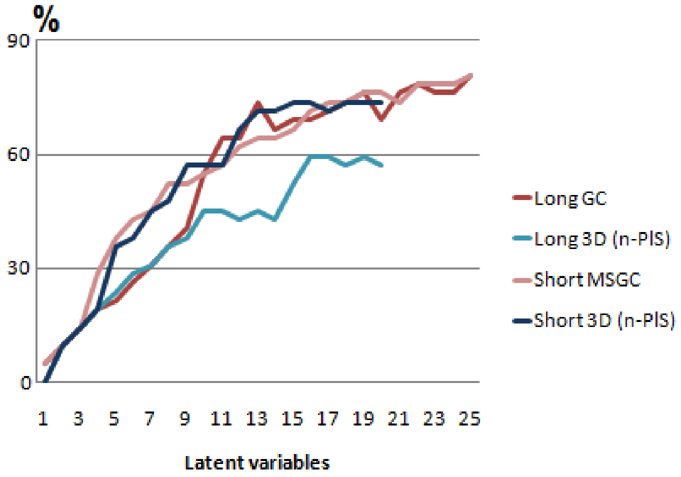


Figure 4.16: PLS-DA success rate for the Long method (22.9 min) using GC data and for the Short method (5 min) using MSGC data

Chapter 5

Conclusions

5. Conclusions

In order to be a commercial success, Electronic Nose systems should show certain operational advantages over traditional methods. For the time being there is no approach in the electronic nose field that justifies this technique over classical approaches in order to generate a clear demand hit in the market.

While low sensitivity and selectivity, short-term life span, difficult calibration and drift problems are important drawbacks to overcome for solid state gas sensors, the improvements brought by the mass spectrometer sensor are counterbalanced by its lack of portability. Even more, simultaneous fragmentation of complex mixtures can produce similar results. The gas chromatograph - mass spectrometer, while still a desktop apparatus, improves the MS based system by increasing the amount of data per sample and thus the sensitivity and selectivity. One drawback brought by the GC-MS is the long analysis time, which eliminates a crucial characteristic of an electronic nose system which is supposed to have higher throughout compared to classical approaches.

To address these issues, the miniaturization of mass spectrometry and gas chromatography has a great potential to improve the performance, usefulness and affordability of the new generation of electronic noses. On the other hand, to speed up the analysis time more powerful mathematical algorithms, that take advantage of the extra information brought by the retention time dimension, should be developed, and this has been the main focus of this thesis.

One of the first conclusions drawn from the work presented in this thesis is that data preprocessing is a must for multivariate and multi-way data analysis. The results prove this affirmation. Aligning the data to counteract chromatographic shift and normalizing it between 0 and 1 improves classification accuracy. Also normalization and mean centering are important leading to better results.

In order to improve the MS-based electronic nose, we have shown, using a fuzzy ARTMAP neural network, that the information brought by the mass spectra is not the same as the information obtained by performing the chromatographic separation. The confusion matrix obtained clearly show this, and points out that by combining the spectral and chromatographic data better results could be obtained. This affirmation was proven by applying the PCA, PLS-DA and fuzzy ARTMAP analysis of the concatenated and unfolded data, and PARAFAC and n-PLS analysis of the three dimensional data.

Also the fuzzy ARTMAP indicates that different fragmentation patterns of individual versus grouped mixtures constituents has an influence on the results, but is difficult to say for sure because "collapsing" the three dimensional data into the MS matrix through summation introduces artifacts.

Applying PCA and PARAFAC for both the "synthetic" and olive oils datasets showed that the addition of extra information in the form of a third dimension represented by the chromatographic separation improves the mass spectrometry-based electronic nose. In both cases the results are very similar and therefore, it is a consistent conclusion. For the same kind of data preprocessing applied, the PARAFAC results of the complete three-dimensional data and the PCA of the unfolded data usually provide the best results, with occasionally GC data challenging and outperforming them.

The PLS-DA and n-PLS analysis of the data, even if not as conclusive as the PCA and PARAFAC proves that if a full chromatographic separation is performed the preferred methods of analysis should be the two-way methods on GC data. On the other hand if a short column is used, leading to a coeluted peak, the two-way PLS methods that combine both the MS and GC information and the three-way n-PLS analysis of the whole 3D data improve the results over mass spectra or chromatographic data alone.

With both the "synthetic" and the "real-world" datasets we can confirm that the addition of the chromatographic retention time as an extra dimension brings an improvement over existing electronic nose technologies. While for the fully resolved chromatograms there is no performance gained, or the gain is minimal, for a short column the extra information improves the results, in some cases making them as good as when a long column is used. This is very important because the measurements on a gas chromatograph - mass spectrometer can be optimized for very short runs, a very important characteristic for an electronic nose. This would allow the design of a higher throughput instruments suitable, for example, for quality control in product lines.

References

References

1 Ouellette, J. Electronic noses sniff out new markets. Indust. Physic. 1999, 5, 26-29.

2 Firestein, S. How the olfactory system makes sense of scents. Nature 2001, 413, 211-218.

3 Zwaardemaker, H.; Hogewind, F. On spray-electricity and waterfall-electricity. KNAW Proceedings, Amsterdam, Netherlands, 1920; Volume 22, pp. 429-437

4 Hartman, J.D. A possible method for the rapid estimation of flavors in vegetables. Proc. Amer. Soc. Hort. Sci. 1954, 64, 335-342

5 Castro, R.; Mandal, M.K.; Ajemba, P.; Istihad, M.A. IEEE Transact. 2003, 49, 1431-1437.

6 Moncrieff, R.W. An instrument for measuring and classifying odors. J. Appl. Physiol. 1961, 16, 742-749

7 Buck, T.M.; Allen, F.G.; Dalton, M. Detection of chemical species by surface effects on metals and semiconductors. In Surface effects in detection; Bregmand, J.I., Dravnieks, A., Eds.; Spartan Books Inc.: Washington, D.C., USA, 1965; pp. 1-27

8 Dravnieks, A.; Trotter, P.J. Polar vapour detector based on thermal modulation of contact potential. J. Sci. Instrum. 1965, 42, 624-627

9 Persaud, K.C.; Dodd, G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. Nature 1982, 299, 352-355

10 Ikegami, A.; Kaneyasu, M. Olfactory detection using integrated sensors. Proceedings of the 3rd international conference on solid-state sensors and actuators, New York, NY, USA, 1985; pp. 136-139.

11 Gardner, J.W.; Bartlett, P.N. A brief history of electronic noses. Sens. Actuat. B: Chem. 1994, 18, 211-220.

12 David, F., B. Tienpont, et al. (2003). "Stir-bar sorptive extraction of trace organic compounds from aqueous matrices." Lc Gc Europe 16(7): 410

13 Schaller, E.; Bosset, J.O.; Esher F. Electronic noses and their application to food. Lebensm.-Wiss. Ul.-Technol. 1998, 31, 305-316.

14 Gardner, J.W.; Bartlett, P.N. Electronic Noses. Principles and Applications; Oxford University Press: Oxford, UK, 1999; pp. 221-245.

15 Lundström, K.I.; Shivaraman, M.S.; Svenson, C.M. A hydrogen-sensitive Pd gate MOS transistor. J. App. Phys. 1975, 46, 3876-3881

16 Lundström, I.; Shivaraman, M.S.; Svenson, C.S.; Lundkvist, L. Hydrogen sensitive MOSFET. Appl. Phys. Lett. 1975, 26, 55-57.

17 Nylander, C.; Armgarth, M.; Lundström, I. An ammonia detector based on a conducting polymer. Anal. Chem. Symp. Ser. 1983, 17, 203-207

18 Bai, H.; Li, C.; Chen, F.; Shi, G. Aligned three-dimensional microstructures of conducting polymer composites. Polymer 2007, 48, 5259-5267

19 Draft, B. Acoustic wave technology sensors. IEEE Trans. 2001, 49, 795-802

20 Hobbs, B.S.; Tantrum, A.D.S.; Chan-Henry, R. Liquid electrolyte fuel cells. In Techniques and Mechanisms in Gas Sensing; Moseley, P.T., Norris, I.O.W., Williams, D.E., Eds.; Adam Hilger: Bristol, UK, 1991; pp. 161-188

21 Strike, D.J.; Meijerink, M.G.H.; Koudelka-Hep, M. Electronic noses - A mini-review. Fres. J. Anal. Chem. 1999, 364, 499-505

22 Applied Sensor Co. Air Quality Module electronic nose; www.appliedsensor.com.

23 Dutta, R.; Hines, E.L.; Gardner, J.W.; Boilot, P. Bacteria classification using Cyranose 320 electronic nose. Biomed. Eng. Online 2002, 1, 1-4

24 Hodgins, D. The electronic nose: sensor array-based instruments that emulate the human nose. Techniques for analyzing food aroma; Marsili, R., Ed., Marcel Dekker Inc.: New York, USA, 1997; pp. 331-371

25 Pathange, L.P.; Mallikarjunan, P.; Marini, R.P.; O'Keefe, S.; Vaughan, D. Non-destructive evaluation of apple maturity using an electronic nose system. J. Food Engin. 2006, 77, 1018-1023

26 Oshita, S.; Shima, K.; Haruta, T.; Seo, Y.; Kawagoe, Y.; Nakayama, S.; Takahara, H. Discrimination of odors emanating from "La France" pear by semi-conducting polymer sensors. Computers Electr. Agric. 2000, 26, 209-216

27 Niruntasuk, K.; Innawong, B.; Parakulsulsatid, P. Shelf life determination of vacuum fried mango chips using electronic nose. In The Proceedings of the 44th Kasetsart University Annual Conference, Kasetsart University, Kasetsart, Thailand, 2006; pp. 200-209

28 Echeverria, G.; Graell, J.; Lopez, M.L.; Brezmes, J.; Correig, X. Volatile production in "Fuji" apples stored under different atmospheres measured by headspace/gas chromatography and electronic nose. Acta Hort. 2005, 682, 1465-1470

29 Supriyadi, S.; Shimuzu, K.; Suzuky, M.; Yoshida, K.; Muto, T.; Fujita, A.; Tomita, N.; Watanabe, N. Maturity discrimination of snake fruit (Salacca edulis Reinw.) cv. Pondoh based on volatiles analysis using an electronic nose device equipped with a sensor array and fingerprint mass spectrometry. Flavour & Fragr. J. 2004, 19, 44-50

30 Costa, G.; Noferini, M.; Montefiori, M.; Brigati, S. Non-destructive assessment methods of kiwifruit quality. Acta. Hort. 2003, 610, 179-189

31 Riva, M.; Benedetti, S.; Mannino, S. Shelf life of fresh cut vegetables as measured by an electronic nose: preliminary study. Ital. Food Techn. 2002, 27, 5-11

32 Berdagué, J.L.; Talou, T. Examples of applications for meat products of semiconductor gas sensors. Sci. Alim. 1993, 13, 141-148

33 Vernat-Rossi, V.; Vernat, G.; Berdagué, J.L. Discrimination of agroalimentary products by gas sensors with semiconductors functioning with ambient air of the laboratory. Various approaches of signal treatment. Analysis 1996, 24, 309-315

34 Vestergaard, J.S.; Martens, M.; Turkky, P. Application of an electronic nose system for prediction of sensory quality changes of a meat product (pizza topping) during storage. LWT - Food Sci. Technol. 2007, 40, 1095-1101

35 V. Baeten, M. Meurens, M.T. Morales and R. Aparicio J. Agric. Food Chem. 44 (1996), p. 2225, I.J. Wesley, F. Pacheco and A.E.J. McGill J. Am. Oil Chem. Soc. 73 (1996), p. 515

36 M.L. Ruiz del Castillo, M.M. Caja, M. Herraiz and G.P. Blanch J. Agric. Food Chem. 46 (1998), p. 5128

37 Trade Standard Applying to Olive Oil and Olive-Pomace Oil. International Olive Oil Council, Madrid, Spain, 24 November 2006

38 Oliveros, M.C.C., Pavon, J.L.P., Pinto, C.G., Laespada, M.E.F., Cordero, B.M. and Forina, M. 2002. Electronic nose based on metal oxide semiconductor sensors as a fast alternative for the detection of adulteration of virgin olive oils. Anal. Chim. Acta 459, 219-228

39 CHRISTY, A.A., DU, Y.P. and OZAKI, Y. 2004. The detection and quantification of adulteration in olive oil by near-infrared spectroscopy and chemometrics. Jpn. Soc. Anal. Chem. 20, 935-940

40 Flores, G., Ruiz Del Castillo, M., Blanch, G.P. and Herraiz, M. 2006a. Detection of the adulteration of olive oils by solid phase microextraction and multidimensional gas chromatography. Food Chem. 97, 336-342

41 M.T. Morales, M.V. Alonso, J.J. Rios and R. Aparicio " Virgin Olive Oil Aroma: Relationship between Volatile Compounds and Sensory Attributes by Chemometrics" J. Agric. Food Chem. 43 (1995), p. 2925

42 A. Guadarrama, M.L. Rodríguez-Méndez, J.A. de Saja, J.L. Ríos and J.M. Olías " Array of sensors based on conducting polymers for the quality control of the aroma of the virgin olive oil" Sens. Actuators B-Chem. 69 (2000), p. 276

43 R. Aparicio, S.M. Rocha, I. Delgadillo and M.T. Morales " Detection of Rancid Defect in Virgin Olive Oil by the Electronic Nose" J. Agric. Food Chem. 48 (2000), p. 853

44 Blanch, P. G., Caja, M. M., Ruiz del Castillo, M., Herraiz, M. (1998). Journal of Agricultural and Food Chemistry, 46, 3153-3157

45 C. Mariani, G. Bellan, E. Lestini and R. Aparicio, The detection of the presence of hazelnut oil in olive oil by free and esterified sterols, European Food Research and Technology 223 (2006), pp. 655-661

46 D. Zabras and M.H. Gordon, Detection of pressed hazelnut oil in virgin olive oil by analysis of polar components: Improvement and validation of the method, Food Chemistry 84 (2004), pp. 475-483

47 L. Cercaci, M.T. Rodriguez-Estada and G. Lercker, Solid-phase extraction-thin-layer chromatography-gas chromatography method for the detection of hazelnut oil in olive oils by determination of esterified sterols, Journal of Chromatography A 985 (2003), pp. 211-220

48 M.S. Cosio, D. Ballabio, S. Benedetti and C. Gigliotti, Evaluation of different storage conditions of extra virgin olive oils with an innovative recognition tool built by means of electronic nose and electronic tongue, Food Chemistry 101 (2007), pp. 485-491

49 M.S. Cosio, D. Ballabio, S. Benedetti and C. Gigliotti, Geographical origin and authentication of extra virgin olive oils by an electronic nose in combination with artificial neural networks, Analytica Chimica Acta 567 (2006), pp. 202-210

50 A. Cimato, D. Dello Monaco, C. Distante, M. Epifani, P. Siciliano and A.M. Taurino et al., Analysis of single-cultivar extra virgin olive oils by means of an electronic nose and HS-SPME/GC/MS methods, Sensors and Actuators B 114 (2006), pp. 674-680

51 Casalinuovo, I.A.; Di Pierro, D.; Coletta, M.; Di Francesco, P. Application of electronic noses for disease diagnosis and food spoilage detection. Sensors 2006, 6, 1428-1439.

52 Pavlou, A.K.; Magan, N.; Sharp, D.; Brown, J.; Barr, H.; Turner, A.P. An intelligent rapid odour recognition model in discrimination of Helicobacter pylori and other gastroesophageal isolates in vitro. Biosens. Bioelectron. 2000, 15, 333-342

53 Siripatrawan, U. Rapid differentiation between E. coli and Salmonella typhimurium using metal oxide sensors integrated with pattern recognition. Sens. Actuat. B: Chem. 2008, 133, 414-419

54 Di Natale, C.; Mantini, A.; Macagnano, A.; Antuzzi, D.; Paolesse, R.; D'Amico, A. Electronic nose analysis of urine samples containing blood. Physiol. Meas. 1999, 20, 377-384

55 Aathithan, S.; Plant, J.C.; Chaudry, A.N.; French, G.L. Diagnosis of bacteriuria by detection of volatile organic compounds in urine using an automated headspace analyzer with multiple conducting polymer sensors. J. Clin. Microbiol. 2001, 39, 2590-2593

56 Pavlou, A.; Turner, A.P.F.; Magan, N. Recognition of anaerobic bacterial isolates in vitro using electronic nose technology. Lett. Appl. Microbiol. 2002, 35, 366-369

57 Hockstein, N.G.; Thaler, E.R.; Torigian, D.; Miller, W.T., Jr.; Deffenderfer, O.; Hanson, C.W. Diagnosis of pneumonia with an electronic nose: correlation of vapor signature with chest computed tomography scan findings. Laryngoscope 2004, 114, 1701-1705

58 Di Natale, C.; Macagnano, A.; Martinelli, E.; Paolesse, R.; D'Arcangelo, G.; Roscioni, C.; Finazzi-Agro, A.; D'Amico, A. Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. Biosens. Bioelectron. 2003, 18, 1209-1218

59 Mielle Patrick, " 'Electronic noses': Towards the objective instrumental characterization of food aroma", Trends in Food Science & Technology Vol. 7, issue 12, Pg 432-438

60 E. Zubritsky. Anal. Chem. 72 (2000), p. 421A.

61 D.J. Strike, M.G.H. Meijerink and M. Koudelka-Hep. Fresenius' J. Anal. Chem. 364 (1999), p. 499

62 J.W. Gardner and K.C. Persaud, Editors, Electronic Noses and Olfaction 2000, IOP Publishing, Bristol, UK (2000)

63 T.A. Dickinson, J. White, J.S. Kauer and D.R. Walt. Current trends in 'artificial-nose' technology Tibtech 16 (1998), p. 250

64 Sparkman, O. David "Mass spectrometry desk reference" Pittsburgh: Global View Pub, 2000

65 Kotiaho, T., F. R. Lauritsen, et al. (1991). "MEMBRANE INTRODUCTION MASS-SPECTROMETRY." Analytical Chemistry 63(18): A875

66 B. Dittmann, S. Nitz and G. Horner. Adv. Food Sci. 20 (1998), p. 115

67 C. Pèrés, F. Begnaud, L. Eveleigh and J.L. Berdagué. Fast characterization of foodstuff by headspace mass spectrometry Trends Anal. Chem. 22 (2003), p. 858

68 B. Dittmann and S. Nitz. Strategies for the development of reliable QA/QC methods when working with mass spectrometry-based chemosensory systems Sens. Actuators, B 69 (2000), p. 253

69 F. Brakstad. Chemom. Intell. Lab. Syst. 29 (1995), p. 157

70 A.T. James, A.J.P. Martin, "Gas-Liquid Partition Chromatography - The Separation And Micro-Estimaton Of Volatile Fatty Acids From Formic Acid To Dodecanoic Acid", Biochem. J. 50 (1952) 679

71 A.J.P. Martin, D.H. Desty (Editors), Vapour Phase Chromatography, Butterworths, London, 1957

72 www.agilent.com

73 M.L. Lee, F.J. Yang, K.D. Bartle, Open Tubular Column Gas Chromatography, Wiley, New York, USA, 1984

74 A.J.P. Martin, in: M. van Swaay (Editor), Proceedings 4th International Symposium on GC, Hamburg, 1962, Butterworths, London, 1962

75 M. van Deursen, J. Beens, H.-G. Janssen, P.A. Leclercq, C.A. Cramers, " Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography" J. Chromatogr. A 878 (2000) 205

76 Phyllis R. Brown, Eli Grushka, Advances in Chromatography vol. 33, 1993

77 Zampolli, S.; Elmi, I.; Sturmann, J.; Nicoletti, S.; Dori, L.; Cardinali, " Selectivity enhancement of metal oxide gas sensors using a micromachined gas chromatographic column" C. Sens. Actuators, B2005, 105, 400

78 H.A.L. Kiers, "Towards a Standardized Notation and Terminology in Multi-way Analysis," J. Chemometrics, vol. 14, no. 3, pp. 105-122, 2000

79 L.R. Tucker, "Implications of Factor Analysis to Three-Way Matrices of Measurement of Change," Problems in Measuring Change; pp. 122-137, The Univ. of Wisconsin Press, 1963

80 L.R. Tucker, "The Extension of Factor Analysis to Three-Dimensional Matrices," Contributions to Math. Psychology; pp. 110-182, Holt, Rinehart and Winston, 1964

81 B.W. Bader and T.G. Kolda, "Algorithm 862: MATLAB Tensor Classes for Fast Algorithm Prototyping," ACM Trans. Math. Software, vol. 32, no. 4, pp. 635-653, 2006

82 Savitzky, A. and Golay,M.J.E. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, 36 (8): 1627-1639

83 Steinier, J., Termonia,Y., and Deltour,J. 1972. Comments on smoothing and differentiation of data by simplified least squares procedure. Analytical Chemistry, 44 (11): 1906-1909

84 Wilson, J.D. and McInnes,C.A.J. 1965. The elimination of errors due to baseline drift in the measurement of peak areas in gas chromatography. Journal of Chromatography A, 19: 486-494

85 Eilers, P.H.C. 2004. Parametric time warping. Analytical Chemistry, 76 (2): 404-411

86 Maeder, M. and Zilian,A. 1988. Evolving Factor-Analysis, a New Multivariate Technique in Chromatography. Chemometrics and Intelligent Laboratory Systems, 3 (3): 205-213

87 van den Berg, F., Tomasi,G., and Viereck,N. 2005. "Warping: investigation of NMR pre-processing and correction." Magnetic Resonance in Food Science: The Multivariate Challenge. The Royal Society of Chemistry. Cambridge. 131-138

88 Nielsen, N.P.V., Carstensen,J.M., and Smedsgaard,J. 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. Journal of Chromatography A, 805 (1-2): 17-35

89 Tomasi, G., van den Berg,F., and Andersson,C. 2004. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. Journal of Chemometrics, 18 (5): 231- 241

90 Pravdova, V., Walczak,B., and Massart,D.L. 2002. A comparison of two algorithms for warping of analytical signals. Analytica Chimica Acta, 456 (1): 77-92

91 Daszykowski, M. and Walczak,B. 2007. Target selection for alignment of chromatographic signals obtained using monochannel detectors. Journal of Chromatography A, 1176 (1-2): 1-11

92 Skov, T., van den Berg, F., Tomasi, G., and Bro, R. 2006. Automated alignment of chromatographic data. Journal of Chemometrics, 20 (11-12): 484-497

93 van Nederkassel, A.M., Daszykowski,M., Eilers,P.H.C., and Heyden,Y.V. 2006a. A comparison of three algorithms for chromatograms alignment. Journal of Chromatography A, 1118 (2): 199-210

94 van Nederkassel, A.M., et all. 2006. Chemometric treatment of vanillin fingerprint chromatograms - Effect of different signal alignments on principal component analysis plots. Journal of Chromatography A, 1120 (1-2): 291-298

95 Bylund, D., Danielsson,R., Malmquist,G., and Markides,K.E. 2002. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modeling of liquid chromatography-mass spectrometry data. Journal of Chromatography A, 961 (2): 237-244

96 Krebs, M.D., Tingley,R.D., Zeskind,J.E., Holmboe,M.E., Kang,J.M., and Davis,C.E. 2006. Alignment of gas chromatography-mass spectrometry data by landmark selection from complex chemical mixtures. Chemometrics and Intelligent Laboratory Systems, 81 (1): 74-81

97 Gong, F., Liang,Y.Z., Fung,Y.S., and Chau,F.T. 2004. Correction of retention time shifts for chromatographic fingerprints of herbal medicines. Journal of Chromatography A, 1029 (1-2): 173-183

98 Xu, C.J., Liang,Y.Z., Chau,F.T., and Vander Heyden,Y. 2006. Pretreatments of chromatographic fingerprints for quality control of herbal medicines. Journal of Chromatography A, 1134 (1-2): 253-259

99 Beebe KR, Randy J, "Chemometrics: a practical guide" Wiley & Sons 1998

100 Bro R, Smilde A. Centering and scaling in component analysis, J. Chemometrics 2003; 17: 16-33

101 L. de Lathauwer, B. de Moor, and J. Vandewalle, "A Multilinear Singular Value Decomposition," SIAM J. Matrix Analysis and Applications, vol. 21, no. 4, pp. 1253-1278, 2000

102 P.M. Kroonenberg and J. de Leeuw, "Principal Component Analysis of Three-Mode Data by Means of Alternating Least Squares Algorithms," Psychometrika, vol. 45, no. 1, pp. 69-97, 1980

103 J.B. Kruskal, "Rank Decomposition, and Uniqueness for 3-way and n-Way Arrays," Multi-way Data Analysis, pp. 8-18, Elsevier, 1989

104 Esbensen K. H. "Multivariate data analysis in practice", CAMO, 2002

105 Wold, H. Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiaah (Ed.). Multivariate Analysis. (pp.391-420) New York: Academic Press, 1966

106 Geladi, P., & Kowlaski B. Partial least square regression: A tutorial. Analytica Chemica Acta, 35, 1-17, 1986

107 Martens, H, & Naes, T. Multivariate Calibration. London: Wiley, 1989

108 Helland I.S. Pls regression and statistical models. Scandivian Journal of Statistics, 17, 97-114, 1990

109 F.L. Hitchcock, "The Expression of a Tensor or a Polyadic as a Sum of Products," J. Math. and Physics, vol. 6, no. 1, pp. 164-189, 1927

110 F.L. Hitchcock, "Multiple Invariants and Generalized Rank of a p-Way Matrix or Tensor," J. Math. and Physics, vol. 7, pp. 39-79, 1927

111 E. Acar, S.A. Camtepe, M. Krishnamoorthy, and B. Yener, "Modeling and Multiway Analysis of Chatroom Tensors," Proc. IEEE Int'l Conf. Intelligence and Security Informatics (ISI '05), pp. 256-268, 2005

112 F. Estienne, N. Matthijs, D.L. Massart, P. Ricoux, and D. Leibovici, "Multi-Way Modelling of High-Dimensionality Electroencephalographic Data," Chemometrics and Intelligent Laboratory Systems, vol. 58, no. 1, pp. 59-72, 2001

113 S. Gourve´nec, I. Stanimirova, C.A. Saby, C.Y. Airiau, and D.L. Massart, "Monitoring Batch Processes with the STATIS Approach," J. Chemometrics, vol. 19, no. 5-7, pp. 288-300, 2005

114 J.D. Carroll and J. Chang, "Analysis of Individual Differences in multidimensional Scaling via an n-Way Generalization of "Eckart-Young" Decomposition," Psychometrika, vol. 35, no. 3, pp. 218-319, 1970

115 R.A. Harshman, "Foundations of the PARAFAC Procedure: Models and Conditions for an 'Explanatory' Multi-Modal Factor Analysis," UCLA Working Papers in Phonetics, no. 16, pp. 1-84, 1970

116 F.L. Hitchcock, "The Expression of a Tensor or a Polyadic as a Sum of Products," J. Math. and Physics, vol. 6, no. 1, pp. 164-189, 1927

117 R.B. Cattell, "Parallel Proportional Profiles and Other Principles for Determining the Choice of Factors by Rotation," Psychometrika, vol. 9, no. 4, pp. 267-283, 1944

118 C.M. Andersen and R. Bro, "Practical Aspects of PARAFAC Modelling of Fluorescence Excitation-Emission Data," J. Chemometrics, vol. 17, no. 4, pp. 200-215, 2003

119 R.A. Harshman, "PARAFAC2: Mathematical and Technical Notes," UCLA Working Papers in Phonetics, vol. 22, pp. 30-44, 1972

120 I. Stanimirova, B. Walczak, D.L. Massart, V. Simeonov, C.A. Saby, and E. di Crescenzo, "STATIS, A Three-Way Method for Data Analysis. Application to Environmental Data," Chemometrics and Intelligent Laboratory Systems, vol. 73, no. 2, pp. 219-233, 2004

121 R.A. Harshman, S. Hong, and M.E. Lundy, "Shifted Factor Analysis-Part I: Models and Properties," J. Chemometrics, vol. 17, no. 7, pp. 363-378, 2003

122 M. Morup and M.N. Schmidt, "Sparse Non-Negative Tensor 2D Deconvolution (SNTF2D) for Multichannel Time-Frequency Analysis," technical report, Technical Univ. of Denmark, DTU, 2006

123 R. Bro, R.A. Harshman, and N.D. Sidiropoulos, "Modeling Multi-Way Data with Linearly Dependent Loadings," Technical Report 2005-176, KVL, 2005

124 L.R. Tucker, "Some Mathematical Notes on Three-Mode Factor Analysis," Psychometrika, vol. 31, pp. 279-311, 1966

125 A. Kapteyn, H. Neudecker, and T. Wansbeek, "An Approach to n-Mode Components Analysis," Psychometrika, vol. 51, no. 2, pp. 269-275, 1986

126 R. Henrion, N-way principal component analysis. Theory, algorithms and applications, Chemomet. Intell. Lab. Syst. 25 (1994) 1-23

127 G.H. Golub and C.F. van Loan, Matrix Computations. The Johns Hopkins Univ. Press, 1996

128 P.M. Kroonenberg and J. de Leeuw, "Principal Component Analysis of Three-Mode Data by Means of Alternating Least Squares Algorithms," Psychometrika, vol. 45, no. 1, pp. 69-97, 1980

129 M.E. Timmerman and H.A.L. Kiers, "Three Mode Principal Components Analysis: Choosing the Numbers of Components and Sensitivity to Local Optima," British J. Math. and Statistical Psychology, vol. 53, no. 1, pp. 1-16, 2000

130 H.A.L. Kiers and A. der Kinderen, "A Fast Method for Choosing the Numbers of Components in Tucker3 Analysis," British J. Math. and Statistical Psychology, vol. 56, no. 1, pp. 119-125, 2003

131 E. Ceulemans and H.A.L. Kiers, "Selecting among Three-Mode Principal Component Models of Different Types and Complexities: A Numerical Convex-Hull Based Method," British J. Math. and Statistical Psychology, vol. 59, no. 1, pp. 133-150, 2006

132 R.A. Harshman, S. Hong, and M.E. Lundy, "Shifted Factor Analysis-Part I: Models and Properties," J. Chemometrics, vol. 17, no. 7, pp. 363-378, 2003

133 R. Bro, J. Chemom., 1996, 10, 47-62

134 P. Paatero, "The Multilinear Engine-A Table-Driven, Least Squares Program for Solving Multilinear Problems, Including the n-Way Parallel Factor Analysis Model," J. Computational and Graphical Statistics, vol. 8, no. 4, pp. 854-888, 1999

135 I.Stanimirova, B. Walczak, D.L. Massart, V. Simeonov, C.A. Saby, and E. di Crescenzo, "STATIS, A Three-Way Method for Data Analysis. Application to Environmental Data," Chemometrics and Intelligent Laboratory Systems, vol. 73, no. 2, pp. 219-233, 2004

136 A. Carlier, C. Lavit, M. Pages, M. Pernin, and J. Turlot, "A Comparative Review of Methods, Which Handle a Set of Indexed Data Tables," Multiway Data Analysis, pp. 85-101, Elsevier, 1989

137 A.K. Smilde, J.A. Westerhuis, and R. Boque´, "Multi-way Multi-block Component and Covariates Regression Models," J. Chemometrics, vol. 14, no. 3, pp. 301-331, 2000

138 H. Kargupta, W. Huang, K. Sivakumar, and E. Johnson, "Distributed Clustering Using Collective Principal Component Analysis," Knowledge and Information Systems J., vol. 3, no. 4, pp. 422-448, 2001

139 F. Vogt, B. Dable, J. Cramer, and K. Booksh, Recent advancements in chemometrics for smart sensors. Analyst 129, 6 (2004):492-502

140 C.M. Andersen and R. Bro, "Practical Aspects of PARAFAC Modeling of Fluorescence Excitation-Emission Data," J. Chemometrics, vol. 17, no. 4, pp. 200-215, 2003

141 R. D. Jiji and K. S. Booksh, Mitigation of Rayleigh and Raman spectral interferences in multiway calibration of excitation-emission matrix fluorescence spectra. Analytical Chemistry 72, 4 (2000):718-725

142 Rinnan, K. Booksh, and R. Bro, First order Rayleigh as a separate component in the decomposition of fluorescence landscapes. Analytica Chimica Acta 537 (2005):349-358

143 E. R. Malinowski, Factor Analysis in Chemistry (JohnWiley and Sons, New York, 2002

144 Jan Blomberg, Peter J. Schoenmakers, Jan Beens, Robert Tijssen (1997). "Compehensive two-dimensional gas chromatography (GC×GC) and its applicability to the characterization of complex (petrochemical) mixtures". Journal of High Resolution Chromatography 20 (10): 539-544

145 J. Saurina, S. Hernandez-Cassou, A. Izquierdo-Ridorsa, and R. Tauler, pH-Gradient spectrophotometric data files from flowinjection and continuous flow systems for two- and three-way data analysis. Chemometrics and Intelligent Laboratory Systems 50, 2 (2000):263-271

146 B. Karlberg and R. Torgrip, Increasing the scope and power of flow-injection analysis through chemometric approaches. Analytica Chimica Acta 500, 1-2 (2003):299-306

147 M. M. Reis, S. P. Gurden, A. K. Smilde, and M. M. C, Ferreira, Calibration and detailed analysis of second-order flow injection analysis data with rank overlap. Analytica Chimica Acta 422, 1 (2000):21-36

148 W. Windig and B. Antalek, Direct exponential curve resolution algorithm (DECRA): A novel application of the generalized rank annihilation method for a single spectral mixture data set with exponentially decaying contribution profiles. Chemometrics and Intelligent Laboratory Systems 37, 2 (1997):241-254

149 F. Miwakeichi, E. Mart?´nez-Montes, P. Valde´s-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, "Decomposing EEG Data into Space-Time-Frequency Components Using Parallel Factor Analysis," NeuroImage, vol. 22, no. 3, pp. 1035-1045, 2004

150 P. Geladi and P. A° berg, Three-way modeling of a batch organic synthesis process monitored by near infrared spectroscopy. Journal of Near Infrared Spectroscopy 9, 1 (2001):1-9

151 P. Geladi and J. Forsstr¨om, Monitoring, of a batch organic synthesis by near-infrared spectroscopy: modeling and interpretation of three-way data. Journal of Chemometrics 16, 7 (2002):329-338

152 H. Abdollahi and F. Nazari, Rank annihilation factor analysis for spectrophotometric study of complex formation equilibria. Analytica Chimica Acta 486, 1 (2003):109-123

153 J. M. M. Leitao and J. C. G, E. da Silva, PARAFAC and PARAFAC2 calibration models for antihypertensor Nifedipine quantification. Analytica Chimica Acta 559, 2 (2006):271-280

154 E. R. Pereira, M. M. Sena, M. A. Z. Arruda, and R. J. Poppi, Exploratory analysis ofL'vov platform surfaces for electrothermal atomic absorption spectrometry by using three-way chemometric tools. Analytica Chimica Acta 495, 1-2 (2003):177-193

155 A. Moreda-Pineiro, A. Marcos, A. Fisher, and S. J. Hill, Parallel factor analysis for the study of systematic error in inductively coupled plasma atomic emission spectrometry and mass spectrometry. Journal of Analytical Atomic Spectrometry 16, 4 (2001):360-369

156 J. Huang, H. Wium, K. B. Qvist, and K. H. Esbensen, Multi-way methods in image analysis-relationships and applications. Chemometrics and Intelligent Laboratory Systems 66, 2 (2003):141-158

157 A. K. Smilde, R. Bro, and P. Geladi, Multi-way Analysis. Applications in the Chemical Sciences (Chichester, Wiley, 2004)

158 R. G. Brereton, Introduction to multivariate calibration in analytical chemistry. Analyst 125, 11 (2000):2125-2154

159 R. Bro, Multivariate calibration-What is in chemometrics for the analytical chemist? Analytica Chimica Acta 500, 1-2 (2003):185- 194

160 Z. Zeng, Y. Y. Cheng, and G. F. Shen, A computational method for quantitative determining the active component of chinese medicine integrating PCA with GRAFA. Acta Chimica Sinica 61, 1 (2003):84-88

161 J. A. Arancibia, A. C. Olivieri, and G. M. Escandar, First- and second-order multivariate calibration applied to biological samples: determination of anti-inflammatories in serum and urine. Analytical and Bioanalytical Chemistry 374, 3 (2002):451-459

162 A. K. Smilde, Comments on three-way analyses used for batch process data. Journal of Chemometrics 15, 1 (2001):19-27

163 S. P. Gurden, J. A. Westerhuis, and A. K. Smilde, Monitoring of batch processes using spectroscopy. AIChE Journal 48, 10 (2002):2283-2297

164 D. J. Louwerse and A. K. Smilde, Multivariate statistical process control of batch processes based on three-way models. Chemical Engineering Science 55, 7 (2000):1225-1235

165 M. Kallioinen, S. P. Reinikainen, J. Nuortila-Jokinen, M. Manttari, T. Sutela, and P. Nurminen, Chemometrical approach in studies of membrane capacity in pulp and paper mill application. Desalination 175, 1 (2005):87-95

166 B. M. Wise, N. B. Gallagher, and E. B. Martin, Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. Journal of Chemometrics 15, 4 (2001):285- 298

167 J. A. Lopes and J. C. Menezes, Industrial fermentation end product modeling with multilinear PLS. Chemometrics and Intelligent Laboratory Systems 68, 1-2 (2003):75-81

168 S. P. Gurden, J. A. Westerhuis, S. Bijlsma, and A. K. Smilde, Modelling of spectroscopic batch process data using grey models to incorporate external information. Journal of Chemometrics 15, 2 (2001):101-121

169 J. H. Chen and J. H.Yen, Three-way data analysis with time lagged window for on-line batch process monitoring. Korean Journal of Chemical Engineering 20, 6 (2003):1000-1011

170 X. Meng, A. J. Morris, and E. B. Martin, On-line monitoring of batch processes using a PARAFAC representation. Journal of Chemometrics 17, 1 (2003):65-81

171 T. F. Cox, Multidimensional scaling used in multivariate statistical process control. Journal of Applied Statistics 28, 3-4 (2001):365- 378

172 M. Dyrby, D. Baunsgaard, R. Bro, and S. B. Engelsen, Multi-way chemometric analysis of the metabolic response to toxins monitored by NMR. Chemometrics and Intelligent Laboratory Systems 76, 1 (2005):79-89

173 S. C. Connor,W.Wu, B. C. Sweatman, J. Manini, J. N. Haselden, D. J. Crowther, and C. J. Waterfield, Effects of feeding and body weight loss on the H-1-NMR-based urine metabolic profiles of male Wistar-Han rats: Implications for biomarker discovery. Biomarkers 9, 2 (2004):156-179

174 H. Idborg, P. O. Edlund, and S. P. Jacobsson, Multivariate approaches for efficient detection of potential metabolites from liquid chromatography-mass spectrometry data. Rapid Communications In Mass Spectrometry 18, 9 (2004):944-954

175 M. Dyrby, M. Peteresen, A. D. Whittaker, L. Lambert, L. Nørgaard, R. Bro, and S. B. Engelsen, Analysis of lipoproteins using 2D diffusion-editedNMRspectroscopy and multi-way chemometrics. Analytica Chimica Acta 531 (2005):209-216

176 K. P. Singh, A. Malik, V. K. Singh, S. Sinha, Multi-way data analysis of soils irrigated with wastewater-A case study. Chemometrics and Intelligent Laboratory Systems 83 (2006) 1-12

177 R. Leardi, C. Armanino, S. Lanteri, and L. Alberotanza, Three-mode principal component analysis of monitoring data from Venice lagoon. Journal of Chemometrics 14, 3 (2000):187-195

178 K. Bergant and L. Kajfez-Bogataj, N-PLS regression as empirical downscaling tool in climate change studies. Theoretical and Applied Climatology 81, 1-2 (2005):11-23

179 P. D. Wentzell, S. S. Nair, and R. D. Guy, Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. Analytical Chemistry 73, 7 (2001):1408-1415

180 R. Pardo, B. A. Helena, C. Cazurro, C. Guerra, L. Deban, C. M. Guerra, and M.Vega, Application of two- and three-way principal component analysis to the interpretation of chemical fractionation results obtained by the use of the BCR procedure. Analytica Chimica Acta 523, 1 (2004):125-132

181 Y. M. Neuhold and M. Maeder, Hard-modelled trilinear decomposition (HTD) for an enhanced kinetic multicomponent analysis. Journal of Chemometrics 16, 5 (2002):218-227

182 R. Dyson, M. Maeder, Y. M. Neuhold, and G. Puxty, Analyses of three-way data from equilibrium and kinetic investigations. Analytica Chimica Acta 490, 1-2 (2003):99-108

183 S. R. Crouch, J. Coello, S. Maspoch, and M. Porcel, Evaluation of classical and three-way multivariate calibration procedures in kinetic-spectrophotometric analysis. Analytica Chimica Acta 424, 1 (2000):115-126

184 J. M. D. Cueva, A. V. Rossi, and R. J. Poppi, Modeling kinetic spectrophotometric data of aminophenol isomers by PARAFAC2. Chemometrics and Intelligent Laboratory Systems 55, 1-2 (2001):125-132

185 E. Ceulemans and I. Van Mechelen, Tucker2 hierarchical classes analysis. Psychometrika 69, 3 (2004):375-399

186 F. Guimet, J. Ferre, and R. Boque, Rapid detection of olivepomace oil adulteration in extra virgin olive oils from the protected denomination of origin "Siurana" using excitationemission fluorescence spectroscopy and three-way methods of analysis. Analytica Chimica Acta 544, 1-2 (2005):143-152

187 N. De Belie, M. Sivertsvik, and J. De Baerdemaeker, Differences in chewing sounds of dry-crisp snacks by multivariate data analysis. Journal of Sound and Vibration 266, 3 (2003):625-643

188 V. Pravdova, B. Walczak, D. L. Massart, H. Robberecht, R. Van Cauwenbergh, P. Hendrix, and H. Deelstra, Three-way principal component analysis for the visualization of trace elemental patterns in vegetables after different cooking procedures. Journal of Food Composition and Analysis 14, 2 (2001):207-225

189 C.A. Andersson and R. Bro, "The n-Way Toolbox for MATLAB," Chemometrics and Intelligent Laboratory Systems, vol. 52, no. 1, pp. 1-4, 2000

190 B.W. Bader and T.G. Kolda, "Algorithm 862: MATLAB Tensor Classes for Fast Algorithm Prototyping," ACM Trans. Math. Software, vol. 32, no. 4, pp. 635-653, 2006

191 B.W. Bader and T.G. Kolda, MATLAB Tensor Toolbox Version 2.2, http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox, 2007

192 PLS_Toolbox, Eigenvector Research Inc., http://www.eigenvector.com/, 2007

193 S. Gourve´nec, G. Tomasi, C. Durvillec, E. di Crescenzo, C.A. Saby, D.L. Massart, R. Bro, and G. Oppenheim, "CuBatch, A MATLAB Interface for n-Mode Data Analysis," Chemometrics and Intelligent Laboratory Systems, vol. 77, nos. 1-2, pp. 122-130, 2005

194 www.models.kvl.dk

195 P. Paatero, "The Multilinear Engine-A Table-Driven, Least Squares Program for Solving Multilinear Problems, Including the n-Way Parallel Factor Analysis Model," J. Computational and Graphical Statistics, vol. 8, no. 4, pp. 854-888, 1999

196 www.multid.se

197 www.camo.com

198 http://three-mode.leidenuniv.nl

199 B.W. Bader and T.G. Kolda, "Efficient MATLAB Computations with Sparse and Factored Tensors," SIAM J. Scientific Computing, vol. 30, no. 1, pp. 205-231, 2006

200 Skapura D. Building neural networks. New York: ACM Press, Addison-Wesley; 1996

201 Kartalopoulos SV. Understanding neural networks and fuzzy logic: basic concepts and applications. IEEE Press; 1996

202 Schalkoff RJ. Artificial neural networks. McGraw-Hill; 1997

203 Gail A. Carpenter, Stephen Grossberg, and David B. Rosen. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks, 4:759-771, 1991

204 Gail A. Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics and Image Processing, 37:54-115, 1987

205 Gail A. Carpenter, Stephen Grossberg, Natalya Markuzon, John H. Reynolds, and David B. Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on Neural Networks, 3 : 698-713, 1992

206 Angerosa, F., Mostallino, R., Basti, C., & Vito, R. (2000). Virgin olive oil odor notes: their relationship with volatile compounds from lipoxygenase pathway and secoiridoid compounds. Food Chemistry, 68, 283-287

207 Parcerisa, J., Richardson, D. G., Rafecas, M., Codony, R., & Boatella, J. (1998). Fatty acid, tocopherol, and sterol content of some hazelnut varieties (Corylus avellana L.) harvested in Oregon (USA). Journal of Chromatography A, 805, 259-268

208 Marriott, P. J., Shellie, R., & Cornwell, C. (2001). Gas chromatographic technologies for the analysis of essential oils. Journal of Chromatography A, 936, 1-22

209 Vichi, S., Pizzale, L., Conte, L. S., Buxaderas, S., & Lopez-Tamames, E. (2005). Simultaneous determination of volatile and semi-volatile aromatic hydrocarbons in virgin olive oil by headspace solid-phase microextraction coupled to gas chromatography/mass spectrometry

Appendix

Paper

Cosmin Burian, Jesus Brezmes, Maria Vinaixa, Eduard Llobet, Xavier Vilanova, Nicolau Cañellas and Xavier Correig, "Ms-electronic nose performance improvement using the retention times dimension and 2-way and 3-way data processing methods", Sensors and Actuators B 143 (2010) 759–768

Sensors and Actuators B 143 (2010) 759–768

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

journal homepage: www.elsevier.com/locate/snb

MS-electronic nose performance improvement using the retention time dimension and two-way and three-way data processing methods

Cosmin Burian[a], Jesus Brezmes[a,b,*], Maria Vinaixa[a,b], Nicolau Cañellas[a,b], Eduard Llobet[a], Xavier Vilanova[a], Xavier Correig[a,b]

[a] Departament de Enginyeria Electrònica, Elèctrica I Automatica, Universitat Rovira i Virgili, URV-IISPV, Av. Paisos Catalans 26, 43007 Tarragona, Spain
[b] CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), c. Mallorca 183, 08036 Barcelona, Spain

ARTICLE INFO

Article history:
Received 17 March 2009
Received in revised form
16 September 2009
Accepted 7 October 2009
Available online 27 October 2009

Keywords:
Multi-wall
PARAFAC
n-PLS
Electronic nose
GC–MS
Fuzzy Artmap

ABSTRACT

In order to improve the mass spectra (MS)-based electronic nose (E-nose) performance, we have included the retention time data given by a new E-nose configuration based on a gas chromatograph–mass spectrometer (GC–MS) as a third dimension. The primary aim of this work is to show that the addition of the third dimension is useful, and brings extra information, helping in the classification of samples. By using this extra information our second goal is to optimize the chromatographic method in order to shorten the time of the chromatographic analysis to a minimum, while still having acceptable results. An experiment was designed in the form of 20 solutions with a high degree of similarity in mass spectra and chromatographic retention times. In order to optimize the system performance and reduce the time of the measurements to a minimum two different chromatographic methods were evaluated. By analyzing these solutions with two-way and three-way PCA, PARAFAC, PLS-DA and n-PLS-DA, and concatenated with supervised Fuzzy Artmap paradigms, we show that the addition of the extra information in the form of the chromatographic separation, even when using a short chromatographic separation, improves the results obtained, compared to the two-way analysis of the mass spectra or total ion chromatogram (TIC) alone. A third goal was to see which signal processing approach was the best suited. We found that when the retention time is used as a third dimension when chromatographic peaks are well resolved, two-way methods work better than their three-way counterparts, whereas in the case of a more challenging situation (a more coeluted chromatogram, with a much shorter measurement time) three-way methods perform better than classic two-way approaches.

1. Introduction

Odor and flavor have always been an important attribute in many products from different fields such as cosmetics and the food industry. Increasing pressure from the consumer has created the need to analyze and classify these attributes in a more elaborate, objective and analytical way. The field of machine olfaction and electronic nose was developed as a consequence. Solid-state gas sensors in combination with pattern recognition were employed since they offered the advantage of objectivity compared to human sensory panels. But because of well-known sensor drawbacks such as reproducibility and short lifetime, in the last few years mass spectrometry-based E-noses are becoming an increasingly used alternative. Among the well-known advantages of MS-based electronic noses the most important is a better reproducibility. The main drawback of MS-based electronic noses is that they are not portable for in-field applications [1–3].

Although not precisely being a gas sensor, mass spectra-based electronic noses can be used together with chemometric methods to obtain a fingerprint of the aroma of a product and classify samples accordingly [1]. Even though the MS system, because of its lack of portability, complexity and expensiveness, fails to comply with the original electronic nose definition, it is a powerful tool that can be used where sensor array electronic noses fail due to the lack of reliability.

Anyway, in scientific fields such as medicine, where the complexity of the data challenges the sensor array and even the MS-sensor approach, the combination of gas chromatography and mass spectroscopy coupled to sophisticated pattern recognition techniques like multi-way data analysis can prove a very useful approach, increasing resolution and reliability.

It is our intention, in this paper, to evaluate whether a GC–MS configuration can improve the results of a standard MS-based E-nose without extending the measurement time to that used in

[*] Corresponding author at: Departament d'Enginyeria Electronica, Elèctrica i Automàtica, Universitat Rovira i Virgili (URV-IISPV), Av. Paisos Catalans, 26. 43007 Tarragona, Spain. Tel.: +34 656655172; fax: +34 977559605.
E-mail address: jesus.brezmes@urv.cat (J. Brezmes).

0925-4005/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.snb.2009.10.015

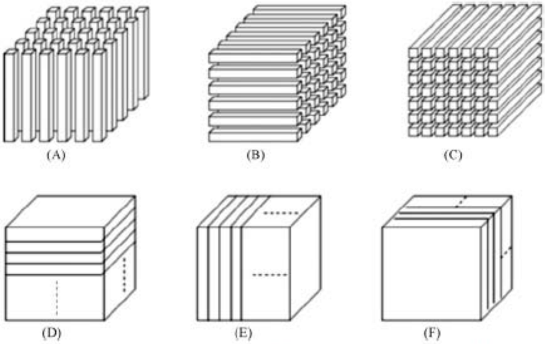


Fig. 1. Nomenclature in 3D matrices: (A) columns, (B) rows, (C) tubes, (D) horizontal slices, (E) vertical slices, (F) frontal slices.

a typical GC run using 2D and/or 3D pattern recognition algorithms.

We base our hope on the fact that the pure MS signal of complex mixtures generates a great mix of mass fragments from all the components. These simultaneous fragmentations generate non-linear behaviors and interferences that make identification a more complex problem. On the other hand, in GC–MS, where each separated component is subject to fragmentation, generating less mass fragments mixtures, which are separated between them in time as well. In this approach, because the information is not coming mixed at a single time, the classification ability is improved.

Based on the fact that the electronic noses are expected to give a response in a relatively short time, the chromatographic methods have to be optimized in order to reduce the measurement time as much as possible, while keeping the advantages brought by the third dimension. It can be achieved using short length columns or fast chromatographic methods. In this paper we will start using the later approach.

3D data analysis methods like PARAFAC [4–6], Tucker [7,8] and n-PLS have proven their advantages in different areas, such as spectroscopy [4], food chemistry [9] and environmental studies [10–12] and they have successfully been employed to interpret multi-way data sets. It is our intention to compare, study and evaluate these algorithms in an initial experiment that could be used as a case study for an electronic nose based on a GC–MS configuration.

Multi-way data analysis, originating in psychometrics back in the sixties [13], is the extension of two-way data analysis to higher-order datasets. Multi-way analysis is often used for extracting hidden structures and capturing underlying correlations between variables in a multi-way array. The difference between two-way and multi-way data analysis is the format of the data being analyzed. Multi-way arrays, often referred to as tensors, are higher-order generalizations of vectors and matrices.

Higher-order arrays have a different terminology compared to that of two-way datasets. Each dimension of a multi-way array is called a mode (or a way) and the number of variables in each mode is used to indicate the dimensionality of a mode. For a three-way array the modes are called row, column and tube [14]. When an index is fixed in one mode and the indices vary in the two other modes, this data partition is called a slice (or slab) in higher-order terminology (Fig. 1). A common trend in multi-way terminology and notation is to follow the guidelines outlined in Ref. [15].

In our configuration (gas chromatography + mass spectrometry) the instrument returns three-way data, i.e., a cube matrix. For each sample we have a matrix containing a mass spectrum for each scan. The GC–MS takes a "mass picture" of the substance being analyzed with each scan. On the other hand, a simple chromatogram or mass spectrum is described by two-way data: for each sample we have a single file with a vector with mass or chromatogram abundances. Just for illustration, we can affirm that a temperature measurement during a day is a one-way data, since for each sample we have one value, a scalar.

In chemistry, one of the most popular applications of a PARAFAC model is modeling fluorescence excitation–emission data, which is a commonly used data type in chemistry, medicine and food science. The data is arranged in an I × J × K three-way matrix, where the first index (I) refers to the samples, the second one (J) refers to the emission wavelengths and the third (K) refers to the excitation wavelengths. Applying the multi-way modeling like PARAFAC, several studies have explored and described the underlying chemical phenomena in fluorescence spectral data obtained from pigment complexes in pea thylakoids [16], form sugar solutions investigating quality [17] and process parameters [18]. Another example of a PARAFAC model on a fluorescence spectroscopic dataset is given as an in-depth study on a fish dataset and data with known fluorophores in Ref. [19].

Multi-way data analysis has also often been employed in environmental studies. Leardi et al. [20] studied the environmental information of the Venice lagoon over a period of four years using the Tucker3 model on a three-way dataset consisting of 13 sampling sites (I), 11 variables (J) and 44 sampling times (K). The three-mode PCA allowed an easier interpretation of spatial and temporal phenomena taking place in the region. In another study [21] soils irrigated with wastewater were analyzed using both two-way and three-way PCA with the aim to assess the soil contamination level and impact of wastewater irrigation. The three-way data analysis was performed with PARAFAC and Tucker3 models. The three-way PCA taking into account the true three-dimensionality structure of the dataset (23 sites × 24 variables × 5 depths) allowed visualization of the hidden information, which could be jointly interpreted.

Dyrby et al. [22] applied the multi-way chemometric model Tucker3, for the first time, to nuclear magnetic resonance (NMR) time series data from studies of the metabolic response to toxins. The observed major metabolic perturbations related to toxicity were in good agreement with those found in previous studies using two-way chemometric methods. The Tucker analysis had the additional advantage of producing easily interpretable time profiles and

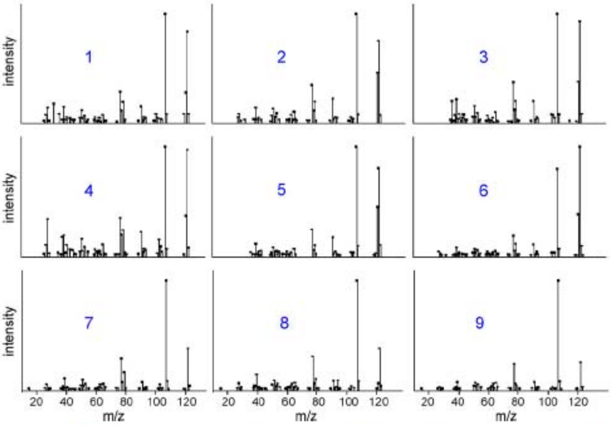


Fig. 2. Theoretical mass spectra (from NIST 62 (National Institute of Standards and Technology MS Database) library) of the nine isomers of dimethylphenol and ethylphenol (2,3-Dimethylphenol; 2,4-Dimethylphenol; 2,5-Dimethylphenol; 2,6-Dimethylphenol; 3,4-Dimethylphenol; 3,5-Dimethylphenol; 2-Ethylphenol; 3-Ethylphenol and 4-Ethylphenol).

extraction of metabolic perturbations with common time profiles only.

In this paper, a study on whether the addition of a third dimension to the measurement data obtained by a MS-electronic nose using chromatographic separation is able to improve the performance of such a system is envisaged.

2. Experimental

2.1. Methods

Twenty mixtures of nine isomers of dimethylphenol and ethylphenol (2,3-Dimethylphenol; 2,4-Dimethylphenol; 2,5-Dimethylphenol; 2,6-Dimethylphenol; 3,4-Dimethylphenol; 3,5-Dimethylphenol; 2-Ethylphenol; 3-Ethylphenol and 4-Ethylphenol) were measured and analyzed by means of gas chromatography mass spectrometry. The nine isomers were chosen based on their theoretically similar mass spectra (Fig. 2).

In order to have a challenging data set, we designed the experiment looking at two issues. First, a PCA of all the 123 masses of theoretical mass spectra of the nine isomers was performed. These spectra were obtained from the NIST 62 library. Based on these PCA it was decided which isomers should be combined into solutions in order for the final mixtures to have a high degree of similarity, thus challenging the performance of MS-based electronic nose. The projection revealed that 2,3-Dimethylphenol and 2,4-Dimethylphenol and 2-Ethylphenol and 3-Ethylphenol had very similar mass spectra and therefore it was difficult to distinguish solutions which contained one of them (Fig. 3) looking at mass spectra alone.

Secondly, we looked at the isomers from the chromatographic point of view (Fig. 4). To do this, we made solutions containing just one isomer in methanol in a concentration of 1%. One millilitre of each of the nine solutions was injected into the injection port of a Shimadzu QP 5000 gas chromatograph–mass spectrometer equipped with an Equity-5 poly(5%defhenyl-95%dimethylsiloxane) (30 m × 0.25 mm × 0.25 mm) capillary column, purchased from Supelco Inc. A temperature-programmed separation was employed. The GC oven was held at 50 °C for 1 min after which the temperature was raised with a rate of 10 °C/min until reaching 180 °C.

The chromatogram shows that isomers 2 and 3, and 6, 8 and 9 are difficult to isolate. This situation would prove a challenge for any GC–MS system if the classification is based on the chromatographic signal alone.

In order to have a stable reference we added benzene as an internal standard. Benzene concentration was set higher than the rest of the isomers in order to have the highest peak on the chromatogram, making the pretreatment of the data through normalization easier.

The experiment, based on the PCA of the theoretical mass spectra and the retention time of the nine isomers was designed as having two main parts (see Table 1). In the first part, from solutions

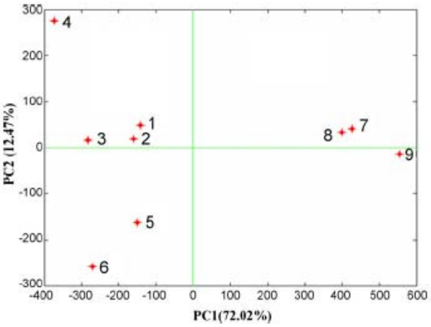


Fig. 3. Principal Component Analysis with 2 PC of the isomers' theoretical mass spectra as follows: (1) 2,3-Dimethylphenol; (2) 2,4-Dimethylphenol; (3) 2,5-Dimethylphenol; (4) 2,6-Dimethylphenol; (5) 3,4-Dimethylphenol; (6) 3,5-Dimethylphenol; (7) 2-Ethylphenol; (8) 3-Ethylphenol and (9) 4-Ethylphenol.

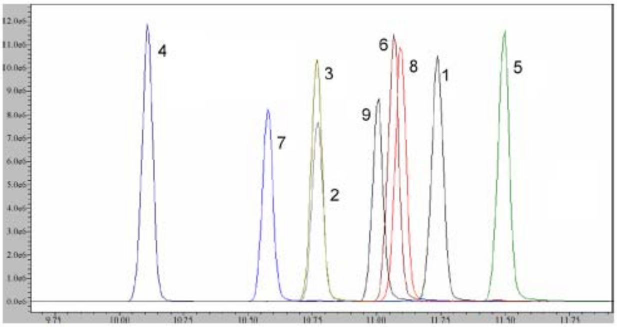


Fig. 4. Superimposed individual chromatographic retention times for each of the nine isomers used in our study.

1 to 9, the solutions have all the isomers in a concentration of 0.5%, except for one that is absent (plus the internal standard in a concentration of 2%). The second one, which involves the solutions 10–20, was designed to increase the difficulty for the MS and GC analysis. Here all the solutions had 2% of benzene, 0.5% of most of the isomers while certain isomers had half of that concentration (0.25%). These isomers were chosen based on the PCA of their theoretical mass spectra and on their retention time to increase complexity. Isomers 1 and 2 and 7 and 8, being so close in a PCA would prove to be challenging for the mass spectra analysis. From the chromatogram point of view, isomers 2 and 3 and 6, 8 and 9, having almost the same retention times, would challenge the chromatographic analysis.

In order to study time optimization in the time axis (chromatographic separation) and the influence of more coeluted peaks (equivalent to faster GC measurements and/or shorter columns), two different chromatographic methods were programmed. In the first method all the isomers in the TIC were resolved to the maximum possible. A temperature-programmed separation was employed. The GC oven was held at 50 °C for 1 min after which the temperature was raised with a rate of 10 °C/min until 180 °C. For the second method an isothermal separation at 190 °C was chosen. This method gave more coeluted peaks and we expected them to generate more challenging datasets for our system. For both methods the mass detector was operating in the electron impact ionization mode with a scan range from *m*/*z* 40 to *m*/*z* 200 at 0.5 scan/s. The ion source temperature was kept at 250 °C.

The measurements were conducted through syringe injection, introducing 1 ml per injection. Ten repetitions were made per solution per chromatographic method. Two blanks were also measured, one consisting of just methanol and the other consisting on methanol and 2% of benzene.

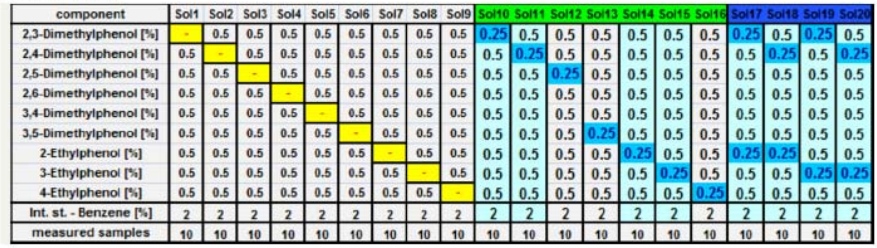
For each measurement, the data collected was arranged into a 3D format. Separation time was plotted against the *x*-axis, the *m*/*z* ratio (mass spectra) against the *y*-axis and the intensity against the *z*-axis. This format gave us a regular 2D matrix for each measurement so that each element value of the matrix represented intensity.

The grouping of all the measurements was arranged in a three-dimensional matrix in which each horizontal plane of the 3D matrix was represented by one measurement. Therefore, the 3D matrix (I × J × K) was organized with the following directions: samples measured (I, rows), *m*/*z* ratio (J, columns), and retention time (K, tubes) (Fig. 5).

The original 3D matrix containing the data from the measurements was analyzed using two-way and three-way methods. In order to analyze 3D data with two-way methods (PCA, PLS-DA) we converted the 3D data to 2D data following different approaches.

First, we "collapsed" the *x*-axis of the chromatographic separation time obtaining a 2D matrix where each file contained the average mass spectra of each measurement. We "collapsed" the *y*-axis of the *m*/*z* fragments, obtaining a matrix with the total ion chromatogram or TIC (Fig. 6).

Table 1
Experimental design.



component	Sol1	Sol2	Sol3	Sol4	Sol5	Sol6	Sol7	Sol8	Sol9	Sol10	Sol11	Sol12	Sol13	Sol14	Sol15	Sol16	Sol17	Sol18	Sol19	Sol20
2,3-Dimethylphenol [%]	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.25	0.5
2,4-Dimethylphenol [%]	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.25
2,5-Dimethylphenol [%]	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2,6-Dimethylphenol [%]	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,4-Dimethylphenol [%]	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,5-Dimethylphenol [%]	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2-Ethylphenol [%]	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.25	0.25	0.5	0.5
3-Ethylphenol [%]	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.25	0.25
4-Ethylphenol [%]	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5
Int. st. - Benzene [%]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
measured samples	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

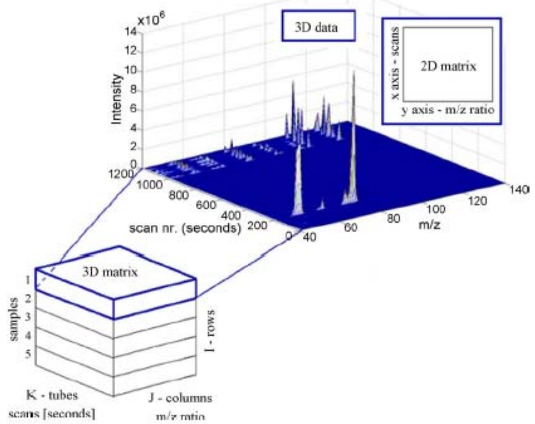


Fig. 5. Graphical representation of the 3D matrix containing several measurements (I-rows) of 2D GC–MS matrices (K-tubes and J-columns), and the 3D data of one GC–MS measurement.

By collapsing (summing) in either axis the extra information brought by the third dimension is lost. In order to compensate for this and still allow the data to be treated with 2D methods two additional approaches were used: unfolding (Fig. 7) and concatenation (Fig. 8).

Unfolding the three-dimensional data generates a two-dimensional matrix where all the original data is maintained in relation to their value but the 3D structural nature of the dataset is lost. This approximation consists on taking the mass spectra of each chromatographic scan and pasting it where the previous scan ended, in the same manner as you would put a deck of cards side by side on the table.

With the concatenation approach, the extra information brought by the chromatographic separation is kept, but not in the same way. The three-dimensional matrix is collapsed in both directions, first giving the average mass spectra and then the total ion chromatogram. Once both vectors are obtained, the chromatogram vector for each measurement is concatenated with the mass spectra vector of the same measurement, giving a new 2D matrix where each file is composed by the mass spectra and TIC of any given measurement. To avoid a greater numerical influence from either type of data (chromatogram or mass spectra) they are both normalized between 0 and 1 prior to the concatenation.

2.2. *Data preprocessing*

Peak alignment, normalization and mean centering preprocessing algorithms were also evaluated to see which one gave best results with the data generated by the GC–MS configuration.

For the peak alignment, Recursive Alignment through Fast Fourier Transform (RAFFT) [23,24] was employed, aligning the data in the three-dimensional matrix. The alignment was done for each *m/z* chromatogram separately (Fig. 9). After peak alignment the 3D data matrix was transformed in five different matrices: 3d, MS, GC, unfolded and concatenated.

Normalization was applied using three different approaches:

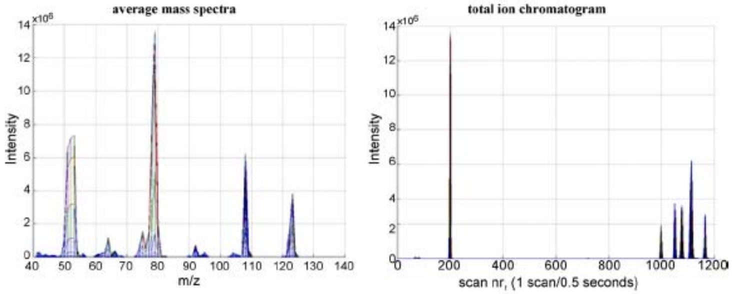


Fig. 6. Average mass spectra and total ion chromatogram.

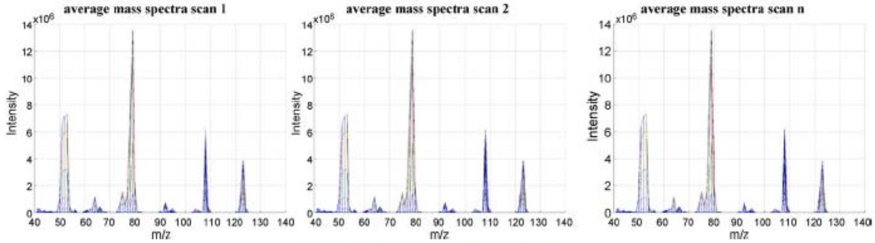


Fig. 7. Graphical representation of unfolded data.

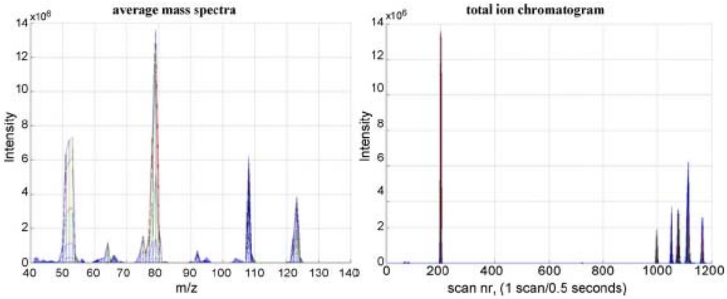


Fig. 8. Graphical representation of concatenated data.

- In the first approach, the intensity points of each measurement were normalized between 0 and 1.
- In order to avoid instrumental and injection error, a second normalization was also evaluated by an internal standard, the benzene peak area. The concentration of benzene being the same in each sample, the intensities of the other compounds are relative to benzene.
- Finally, in a third approach, the data was mean centered following the sample direction (columns) in all five data sets. At each column, the mean value was calculated and subtracted from all the points of the column.

These approaches allowed us to perform a complete study and comparison between each type of preprocessing used: alignment vs. raw data, unity normalization vs. benzene normalization and raw data vs. mean centering.

As mentioned before, two different chromatographic methods were applied for the study. Each one had some advantages and some drawbacks. Method 1, aimed at obtaining resolved chromatograms, extracts more information at the expense of analysis time, while method 2, aimed at saving time, generated more coeluted peaks, because of a much shorter time of analysis and

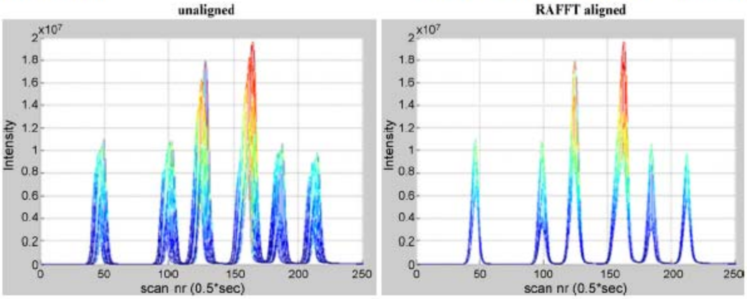
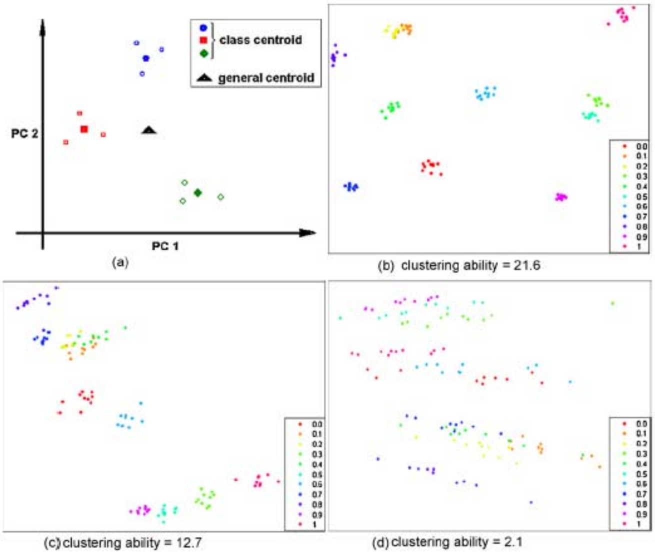


Fig. 9. TIC of several measurements: unaligned and RAFFT aligned.



(a)

(b) clustering ability = 21.6

(c) clustering ability = 12.7

(d) clustering ability = 2.1

$$C_c = \frac{\sum_{i=1}^{R} S_i}{R} \quad \text{eq. 1}$$

$$d = \sum_{i=1}^{n} (|S_i - C_i|) \quad \text{eq. 2}$$

$$\text{Intra} = \frac{\sum_{i=1}^{R} d_i}{R} \quad \text{eq. 3}$$

$$C_g = \frac{\sum_{i=1}^{M} S_i}{M} \quad \text{eq. 4}$$

$$D = \sum_{i=1}^{n} (|C_{c_i} - C_{g_i}|) \quad \text{eq. 5}$$

$$\text{Inter} = \frac{\sum_{i=1}^{C} D_i}{C} \quad \text{eq. 6}$$

S= scores
R= repetitions
C_c = class centroid
Intra= intravariance
C_g = general centroid
M= measurements
d= distance point to centroid
D= Distance C_c to C_g
C= number of classes
n= nr of principal components

(e)

$$\text{cluster ability} = \frac{\text{Inter}}{\text{mean(Intra)}} \quad \text{eq. 7}$$

Fig. 10. (a) Clustering ability; (b)–(d) show how cluster ability works. The cluster ability index increases as cluster is better defined in PCA or PARAFAC.

therefore less data, presents a higher challenge to the classification algorithms.

2D datasets were analyzed by Principal Component Analysis (PCA), Partial Least Squares Discriminant Analysis (PLS-DA), while the 3D dataset was analyzed by Parallel Factor Analysis (PARAFAC), and multi-way PLS-DA.

3. Results

3.1. PCA and PARAFAC

Different PCA and PARAFAC projections were performed using the multivariate GC–MS data obtained. The goal was to see if mea-

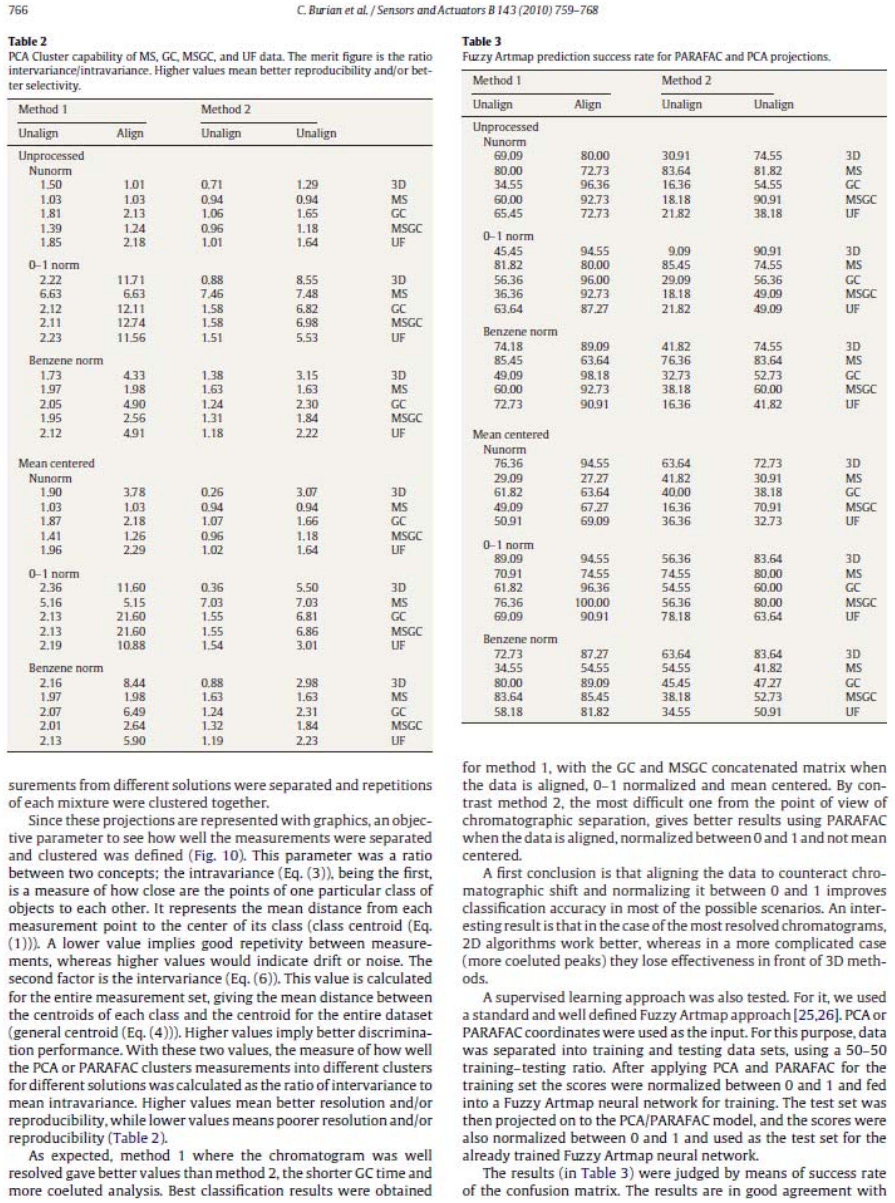


Table 2
PCA Cluster capability of MS, GC, MSGC, and UF data. The merit figure is the ratio intervariance/intravariance. Higher values mean better reproducibility and/or better selectivity.

Method 1		Method 2		
Unalign	Align	Unalign	Unalign	
Unprocessed				
Nunorm				
1.50	1.01	0.71	1.29	3D
1.03	1.03	0.94	0.94	MS
1.81	2.13	1.06	1.65	GC
1.39	1.24	0.96	1.18	MSGC
1.85	2.18	1.01	1.64	UF
0–1 norm				
2.22	11.71	0.88	8.55	3D
6.63	6.63	7.46	7.48	MS
2.12	12.11	1.58	6.82	GC
2.11	12.74	1.58	6.98	MSGC
2.23	11.56	1.51	5.53	UF
Benzene norm				
1.73	4.33	1.38	3.15	3D
1.97	1.98	1.63	1.63	MS
2.05	4.90	1.24	2.30	GC
1.95	2.56	1.31	1.84	MSGC
2.12	4.91	1.18	2.22	UF
Mean centered				
Nunorm				
1.90	3.78	0.26	3.07	3D
1.03	1.03	0.94	0.94	MS
1.87	2.18	1.07	1.66	GC
1.41	1.26	0.96	1.18	MSGC
1.96	2.29	1.02	1.64	UF
0–1 norm				
2.36	11.60	0.36	5.50	3D
5.16	5.15	7.03	7.03	MS
2.13	21.60	1.55	6.81	GC
2.13	21.60	1.55	6.86	MSGC
2.19	10.88	1.54	3.01	UF
Benzene norm				
2.16	8.44	0.88	2.98	3D
1.97	1.98	1.63	1.63	MS
2.07	6.49	1.24	2.31	GC
2.01	2.64	1.32	1.84	MSGC
2.13	5.90	1.19	2.23	UF

Table 3
Fuzzy Artmap prediction success rate for PARAFAC and PCA projections.

Method 1		Method 2		
Unalign	Align	Unalign	Unalign	
Unprocessed				
Nunorm				
69.09	80.00	30.91	74.55	3D
80.00	72.73	83.64	81.82	MS
34.55	96.36	16.36	54.55	GC
60.00	92.73	18.18	90.91	MSGC
65.45	72.73	21.82	38.18	UF
0–1 norm				
45.45	94.55	9.09	90.91	3D
81.82	80.00	85.45	74.55	MS
56.36	96.00	29.09	56.36	GC
36.36	92.73	18.18	49.09	MSGC
63.64	87.27	21.82	49.09	UF
Benzene norm				
74.18	89.09	41.82	74.55	3D
85.45	63.64	76.36	83.64	MS
49.09	98.18	32.73	52.73	GC
60.00	92.73	38.18	60.00	MSGC
72.73	90.91	16.36	41.82	UF
Mean centered				
Nunorm				
76.36	94.55	63.64	72.73	3D
29.09	27.27	41.82	30.91	MS
61.82	63.64	40.00	38.18	GC
49.09	67.27	16.36	70.91	MSGC
50.91	69.09	36.36	32.73	UF
0–1 norm				
89.09	94.55	56.36	83.64	3D
70.91	74.55	74.55	80.00	MS
61.82	96.36	54.55	60.00	GC
76.36	100.00	56.36	80.00	MSGC
69.09	90.91	78.18	63.64	UF
Benzene norm				
72.73	87.27	63.64	83.64	3D
34.55	54.55	54.55	41.82	MS
80.00	89.09	45.45	47.27	GC
83.64	85.45	38.18	52.73	MSGC
58.18	81.82	34.55	50.91	UF

surements from different solutions were separated and repetitions of each mixture were clustered together.

Since these projections are represented with graphics, an objective parameter to see how well the measurements were separated and clustered was defined (Fig. 10). This parameter was a ratio between two concepts; the intravariance (Eq. (3)), being the first, is a measure of how close are the points of one particular class of objects to each other. It represents the mean distance from each measurement point to the center of its class (class centroid (Eq. (1))). A lower value implies good repetivity between measurements, whereas higher values would indicate drift or noise. The second factor is the intervariance (Eq. (6)). This value is calculated for the entire measurement set, giving the mean distance between the centroids of each class and the centroid for the entire dataset (general centroid (Eq. (4))). Higher values imply better discrimination performance. With these two values, the measure of how well the PCA or PARAFAC clusters measurements into different clusters for different solutions was calculated as the ratio of intervariance to mean intravariance. Higher values mean better resolution and/or reproducibility, while lower values means poorer resolution and/or reproducibility (Table 2).

As expected, method 1 where the chromatogram was well resolved gave better values than method 2, the shorter GC time and more coeluted analysis. Best classification results were obtained for method 1, with the GC and MSGC concatenated matrix when the data is aligned, 0–1 normalized and mean centered. By contrast method 2, the most difficult one from the point of view of chromatographic separation, gives better results using PARAFAC when the data is aligned, normalized between 0 and 1 and not mean centered.

A first conclusion is that aligning the data to counteract chromatographic shift and normalizing it between 0 and 1 improves classification accuracy in most of the possible scenarios. An interesting result is that in the case of the most resolved chromatograms, 2D algorithms work better, whereas in a more complicated case (more coeluted peaks) they lose effectiveness in front of 3D methods.

A supervised learning approach was also tested. For it, we used a standard and well defined Fuzzy Artmap approach [25,26]. PCA or PARAFAC coordinates were used as the input. For this purpose, data was separated into training and testing data sets, using a 50–50 training–testing ratio. After applying PCA and PARAFAC for the training set the scores were normalized between 0 and 1 and fed into a Fuzzy Artmap neural network for training. The test set was then projected on to the PCA/PARAFAC model, and the scores were also normalized between 0 and 1 and used as the test set for the already trained Fuzzy Artmap neural network.

The results (in Table 3) were judged by means of success rate of the confusion matrix. The results are in good agreement with

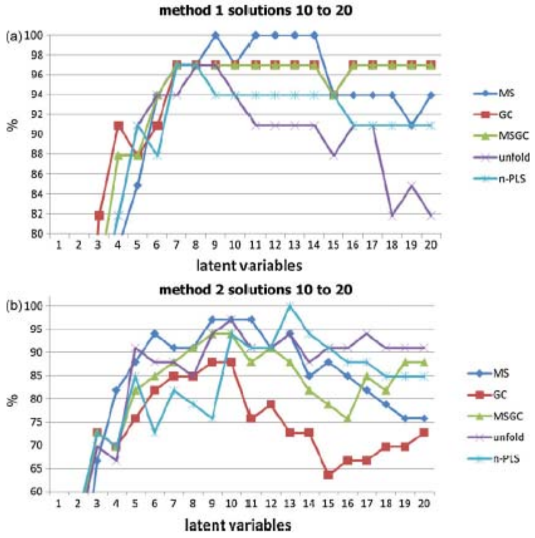


Fig. 11. PLS-DA and n-PLS-DA prediction success rate for method 1 (a) and method 2 (b).

the cluster analysis, showing that method 1 presents better results when two-way analysis methods are employed while method 2 gives better results with three-way data analysis is used. For method 1 the best results were obtained by the PCA projection of the MSGC concatenated matrix when the data was aligned, mean centered and normalized between 0 and 1, scoring 100% success rate, followed by GC data with a 96.36% success rate. In the case of the second method the best result was obtained using PARAFAC when the data was aligned, normalized from 0 to 1 and not mean centered, as well as using the PCA projection of the MSGC matrix after alignment.

Averaging the success rate for each method shows that aligning, mean centering and normalizing the data between 0 and 1 gives the best results, indicating that this is the best way to preprocess the data.

3.2. PLS-DA and n-PLS-DA

For PLS-DA, and n-PLS-DA analysis, each dataset was separated into training and testing sets by a chosen training/test ratio of 7/3, rather than 5/5 like in the PCA–PARAFAC analysis, since this ratio gave better results on the PLS-DA–n-PLS-DA. The training set was used to train the PLS-DA and n-PLS-DA models. Onto these models the test set was projected.

The model and prediction performance were evaluated by means of the Root Mean Square Error of Cross-Validation (RMSECV) calculated using the training measurements, and the Root Mean Square Error of Prediction (RMSEP), respectively.

The data proved difficult to classify and predict if all the 20 different mixtures were analyzed. Because of this, the first 9 solutions were left out and we used the most challenging 10–20. The overall prediction performance was assessed by means of sample prediction success rate.

In Fig. 11(a) and (b) the graphs represent results present the sample prediction success rate for methods 1 and 2 for solutions 10–20. It can be noticed that method 2, in which the peaks are more coeluted, gives worse results than method 1, where a temperature-programmed separation was employed in order to achieve a better separation. We can clearly notice this in the chromatographic (GC) success rate line.

In method 1, the time retention time does not increase the resolution of the device compared to the regular chromatogram or even the mass spectra. Therefore, in well-resolved chromatograms (that are also time-consuming) 3D data methods or 2D data are not increasing the resolution of the system.

On the other hand, in the more challenging and coeluted method 2 retention time information gives additional information and the n-PLS-DA multi-way method presented the best results giving 100% success rate classification for 13 latent variables (Fig. 11(b)). As expected, GC information alone does not give enough information.

Even though the PLS-DA and n-PLS-DA results are not as conclusive as the PCA and PARAFAC results, the multi-way PLS-DA model yields a high prediction success rate in both datasets. It is also recommended over the unfolded and concatenated data because it builds a simpler and more parsimonious model.

4. Conclusion

This study has proven that for the two chromatographic methods studied (the most resolved and the shortest, more coeluted one), the addition of the third dimension based on the chromatographic retention time improves the performance compared to a MS-based electronic nose using the mass spectra or total ion chromatogram alone.

Treating the two-dimensional data with two-way methods (PCA and PLS-DA) and three-way algorithms (PARAFAC and n-PLS-DA)

showed that for the first chromatographic method, where the peaks are well defined and separated, the two-way methods work best, whereas for the more complicated signal of the third method, where the peaks are coeluted, three-way methods are better suited.

From the point of view of data pretreatment, it is clear that alignment is an important step, improving the results obtained in most of the cases, even for the more coeluted method 2. From the point of view of mean centering and normalization, the recommendation for this dataset is to normalize data between 0 and 1 and to mean center the data to start data processing.

In any case, the statistical significance of these results and conclusions are related to our experiment only and other types of experiments should be designed to validate our observations as a general trend. We encourage scientific readers to design new synthetic or real-world applications to validate the conclusions obtained in this work using the same approaches we have shown in this paper.

Acknowledgment

This work has been funded by Ciberdem, "CIBER de Diabetes y Enfermedades Metabólicas Asociadas", The "CIBER de Diabetes y Enfermedades Metabólicas Asociadas" is an ISCIII iniciative.

References

[1] S. Ampuero, J.O. Bosset, The electronic nose applied to diary products: a review, Sensors and Actuators B 94 (2003) 1–12.

[2] M. Vinaixa, A. Vergara, C. Duran, E. Llobet, C. Badia, J. Brezmes, X. Vilanova, X. Correig, Fast detection of rancidity in potato crisps using e-nosesbased on mass spectrometry or gas sensors, Sensors and Actuators B 106 (2005) 67–75.

[3] C. Peres, F. Begnaud, L. Eveleigh, J.-L. Berdague, Fast characterization of foodstuff by headspace mass spectrometry (HS-MS), Trends in Analytical Chemistry 22 (11) (2003).

[4] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.

[5] H. Kiers, J.T. Berge, R. Bro, Parafac2—Part I. A direct fitting algorithm for the Parafac2 model, Journal of Chemometrics 13 (1999) 275–294.

[6] N. Fabera, R. Brob, P. Hopke, Recent developments in ANDECOMP/PARAFAC algorithms: a critical review, Chemometrics and Intelligent Laboratory Systems 65 (2003) 119–137.

[7] P. Geladi, Analysis of multi-way (multi-mode) data, Chemometrics and Intelligent Laboratory Systems 7 (December (1–2)) (1989) 11–30.

[8] R. Henrion, N-way principal component analysis—theory, algorithms and applications, Chemometrics and Intelligent Laboratory Systems 25 (September (1)) (1994) 1–23.

[9] V. Pravdova, C. Bouconb, S. de Jong, B. Walczak, D.L. Massart, Three-way principal component analysis applied to food analysis: an example, Analytica Chimica Acta 462 (2002) 133–148.

[10] P. Barbieri, G. Adami, S. Piselli, F. Gemiti, E. Reisenhofer, A three-way principal factor analysis for assessing the time variability of freshwaters related to a municipal water supply, Chemometrics and Intelligent Laboratory Systems 62 (2002) 89–100.

[11] G.R. Flaten, B. Grung, O.M. Kvalheim, Multi-way exploration of regular environmental monitoring surveys, Chemometrics and Intelligent Laboratory Systems 77 (2005) 104–114.

[12] I. Stanimirova, V. Simeonov, Modeling of environmental four-way data from air quality control, Chemometrics and Intelligent Laboratory Systems 77 (2005) 115–121.

[13] L.R. Tucker, The extension of factor analysis to three-dimensional matrices, in: Contributions to Mathematical Psychology, Holt, Rinehart and Winston, New York, 1964, pp. 110–182.

[14] B.W. Bader, T.G. Kolda, Algorithm 862: Matlab tensor classes for fast algorithm prototyping, ACM Transactions on Mathematical Software 32 (4) (2006) 635–653.

[15] H.A.L. Kiers, Towards a standardized notation and terminology in multiway analysis, Journal of Chemometrics 14 (3) (2000) 105–122.

[16] R.T. Ross, C.-H. Lee, C.M. Davis, B.M. Ezzeddine, E.A. Fayyad, S.E. Leurgans, Resolution of fluorescence spectra of plant pigment complexes using trilinear models, Biochimica et Biophysica Acta 1056 (1991) 317–320.

[17] D. Baunsgaard, C.A. Andersson, A. Arndal, L. Munck, Multi-way chemometrics of mathematical separation of fluorescent colorants and colour precursors form spectrofluorimetry of beet sugar and beet sugar thick juice as validated by HPLC analysis, Food Chemistry 70 (2000) 113–121.

[18] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, Chemometrics and Intelligent Laboratory Systems 46 (1999) 133–147.

[19] C.M. Andersen, R. Bro, Practical aspects of PARAFAC modeling of fluorescence excitation–emission data, Journal of Chemometrics 17 (4) (2003) 200–215.

[20] R. Leardi, C. Armanino, S. Lanteri, L. Alberotanza, Three-mode principal component analysis of monitoring data from Venice lagoon, Journal of Chemometrics 14 (2000) 187–195.

[21] K. Singh, A. Malik, V. Singh, S. Sinha, Multi-way data analysis of soils irrigated with wastewater—a case study, Chemometrics and Intelligent Laboratory Systems 83 (2006) 1–12.

[22] M. Dyrby, D. Baunsgaard, R. Bro, S.B. Engelsen, Multiway chemometric analysis of the metabolic response to toxins monitored by NMR, Chemometrics and Intelligent Laboratory Systems 75 (2005) 79–89.

[23] J.W.H. Wong, G. Cagney, H.M. Cartwright, SpecAlign—processing and alignment of mass spectra datasets, Bioinformatics 21 (2005) 2088–2090.

[24] J.W.H. Wong, H.M. Cartwright, An application of Fast Fourier Transform cross-correlation for the alignment of large chromatographic and spectral datasets, Analytical Chemistry 77 (2005) 5655–5661.

[25] G.A. Carpenter, S. Grossberg, Adaptive resonance theory, in: C. Stuart, Shapiro (Eds.), Encyclopedia of Artificial Intelligence, second ed., Wiley and Sons, New York, 1992, pp. 13–21.

[26] G.A. Carpenter, S. Grossberg, N. Markuzon, J.H. Reynolds, D.B. Rosen, Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps, IEEE Transactions on Neural Networks 3 (5) (1992) 698–713.

Biographies

Cosmin Burian obtained his chemical engineering bachelor grade in 2004 at the Faculty of Chemistry and Chemical Engineering of the "Babes-Bolyai" University of Cluj-Napoca, Romania, and the Master grade in 2005 in Applied Electrochemistry at the same university. At present he is a PhD in the SIPOMICS research group.

Jesús Brezmes graduated in telecommunication engineering from the Universitat Politècnica de Catalunya (UPC) (Barcelona, Spain) in 1993, and received his PhD in 2001 from the same university. He is currently an associate professor in the Electronic Engineering Department at the Universitat Rovira i Virgili (Tarragona, Spain). His main areas of interest are in bioinformatics, chemometrics, multivariate and statistical análisis and other signal processing algorithms to be applied in omic sciences.

Maria Vinaixa graduated in chemistry from the Universitat Rovira i Virgili of Tarragona (Catalonia, Spain) in 2002. In 2008 she earned her PhD in the Electronic Engineering Department at the same university working on the application of signal processing and pattern recognition data analysis techniques to chemical sensor arrays and mass spectrometry for complex aroma analysis. At present she is involved in an interdisciplinary metabolomics research project which deals on the use of NMR and mass spectrometry as a tool in biomedical diagnosis. Her main activity is focused on the use of signal processing techniques for chemical and biomedical data analysis of several biofluids; and the comprehensive and quantitative analysis of the pool of metabolites to give insight into diabetes and other metabolic associated disorders.

Nicolau Cañellas graduated in telecommunication engineering from the Universitat Politècnica de Catalunya (UPC) (Barcelona, Spain) in 1991, and received his PhD in 2006 from the same university. He is currently an associate professor in the Electronic Engineering Department at the Universitat Rovira i Virgili (Tarragona, Spain). His main areas of interest are in bioinformatics and signal processing for omic sciences.

Eduard Llobet graduated in telecommunication engineering from the Universitat Politecnica de Catalunya (UPC) (Barcelona, Spain) in 1991, and received his PhD in 1997 from the same university. During 1998, he was a visiting fellow at the School of Engineering, University of Warwick (UK). He is currently full professor of Electronic Engineering in the Electronic Engineering Department at the Universitat Rovira i Virgili (Tarragona, Spain). His main areas of interest are in the fabrication, and modeling, of semiconductor gas sensors and in the application of intelligent systems to complex odor analysis.

Xavier Vilanova graduated in telecommunication engineering from the Universitat Politècnica de Catalunya (UPC) (Barcelona, Spain) in 1991, and received his PhD in 1998 from the same university. He is currently an associate professor in the Electronic Engineering Department at the Universitat Rovira i Virgili (Tarragona, Spain). His main areas of interest are in semiconductor chemical sensors modeling and simulation.

Xavier Correig graduated in telecommunication engineering from the Universitat Politècnica de Catalunya (UPC) (Barcelona, Spain) in 1984, and received his PhD in 1988 from the same university. He is a full professor of Electronic Technology in the Electronic Engineering Department at the Universitat Rovira i Virgili (Tarragona, Spain). His research interests include heterojunction semiconductor devices and solid-state gas sensors. Dr. Correig is a member of the Institute of Electrical and Electronic Engineers.

Congress contribution I

Abstract I

Oral Presentation at:

ISOEN 2009
University of Brescia - Italy
15-17 April

Cosmin Burian, **Jesus Brezmes**, Maria Vinaixa, Eduard Llobet, Xavier Vilanova, Nicolau Cañellas and Xavier Correig, "Ms-electronic nose performance improvement using gc retention times and 2-way and 3-way data processing methods"

MS-ELECTRONIC NOSE PERFORMANCE IMPROVEMENT USING GC RETENTION TIMES AND 2-WAY AND 3-WAY DATA PROCESSING METHODS

Cosmin Burian[1], Jesus Brezmes [1,2], Maria Vinaixa [1,2], Eduard Llobet[1], Xavier Vilanova[1], Nicolau Cañellas[1] and Xavier Correig [1,2]
[1]Department of Electronic Engineering, Universitat Rovira i Virgili,
Avenida Paisos Catalans 26, 43007 Tarragona, Spain
[2]CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM)
cosmin.burian@urv.cat, jesus.brezmes@urv.cat,

Abstract

We have designed a challenging experimental sample set in the form of 20 solutions with a high degree of similarity in order to study whether the addition of chromatographic separation information improves the performance of regular MS based electronic noses. In order to make an initial study of the approach, two different chromatographic methods were used. By processing the data of these experiments with 2 and 3-way algorithms, we have shown that the addition of chromatographic separation information improves the results compared to the 2-way analysis of mass spectra or total ion chromatogram treated separately. Our findings show that when the chromatographic peaks are resolved (longer measurement times), 2-way methods work better than 3-way methods, whereas in the case of a more challenging measurement (more coeluted chromatograms, much faster GC-MS measurements) 3-way methods work better.

1. Introduction

In this paper it is our intention, to evaluate whether a GC-MS configuration can improve the results of a standard MS-Enose using only a small fraction of the measurement time required in a typical GC-MS run (2-3 minutes versus 30 minutes or more).

To do so, m/z (mass spectra information) variation along a retention time axis is collected so that a final 3D matrix is created. In this matrix, each file is related to the measurement of a sample. Each sample measurement data is laid out in a plane with 2 axis: the time axis (columns) and the m/z axis (the so called "tubes").

To process the 3D data matrix obtained, new 3D data analysis methods like PARAFAC [1,2], Tucker [3] , and N-PLS have been considered due to their proven advantages in different areas, such as spectroscopy [1], food chemistry [4] and environmental studies [5]. In these applications, these algorithms have been successfully employed to interpret multi-way data sets. It is our intention to compare, study and evaluate these algorithms in the experimental setup we have designed so that this work could be referenced as a case study for an electronic nose based on a GC-MS configuration.

Multiway data analysis, originating in psychometrics back in the sixties [6], is the extension of two-way data analysis to higher-order datasets. Multiway analysis is often used for extracting hidden structures and capturing underlying correlations between variables in a multiway array.

The difference between two-way and multiway data analysis is the format of the data being analyzed. Multiway arrays, often referred to as tensors, are higher-order generalizations of vectors and matrices.

2. Experimental

Twenty mixtures of nine isomers of dimethylphenol and ethylphenol were measured and analyzed by means of gas chromatography mass spectrometry (GC-MS). The nine isomers were chosen based on their theoretically similar mass spectra in order to have a challenging data set.

To design the experiment we looked at 2 key issues: the PCA of the 9 isomers mass spectra and their chromatographic retention times. To obtain the chromatographic retention time (Fig.1.), 9 solutions of 1% isomer in methanol containing only one isomer at a time were prepared and analyzed. In order to see which of the 9 isomers has most alike mass spectra we calculated a PCA of the theoretical mass spectra of each isomer (Fig. 2).

Based on this information we designed the experiment as shown in **Table 1.** Benzene acts as an internal standard, and having the highest concentration it is used in the normalization pretreatment of the data. The experiment was designed in order to be a challenging sample set for any mass spectra-based electronic nose.

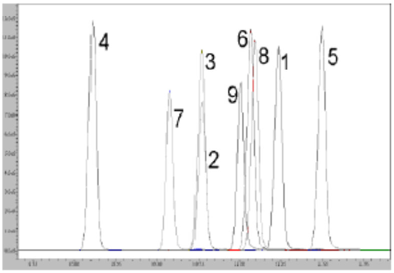


Fig.1 Chromatographic retention times for the 9 isomers as follows: 1) 2,3-Dimethylphenol;2) 2,4-Dimethylphenol;3) 2,5-Dimethylphenol;4) 2,6-Dimethylphenol;5) 3,4-Dimethylphenol;6) 3,5-Dimethylphenol;7) 2-Ethylphenol;8) 3-Ethylphenol and 9) 4-Ethylphenol; Isomers 2 and 3 and 6,8,9 are the most similar among them.

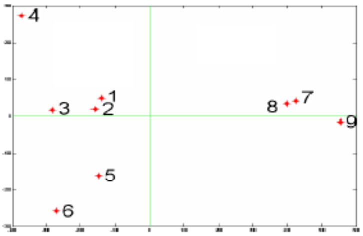


Fig.2 Principal Component Analysis of the mass spectra of the 9 isomers. Isomers 7, 8 ,9 and 1, 2, 3 are the most similar among them

Two chromatographic methods were studied. In method one, we tried to separate the isomers as much as possible, trying to get a well-resolved chromatogram. This method is supposed to be easier to analize. To achieve this, a temperature programmed separation was used, starting at 50°C, where the temperature was kept constant for one minute, until 180°C, where almost all the isomers were separated. Method two was designed to return coeluted peaks and therefore it was performed with an isothermal temperature of 190°. This method was supposed to give a more challenging dataset and to be executed in a much shorter time. The measurements were conducted through syringe injection of 1 µl per measurement, and ten repetitions for each method and solution were made. Two blanks were also measured, one consisting of just methanol and the other containing methanol and 2% of benzene.

component	Sol1	Sol2	Sol3	Sol4	Sol5	Sol6	Sol7	Sol8	Sol9	Sol10	Sol11	Sol12	Sol13	Sol14	Sol15	Sol16	Sol17	Sol18	Sol19	Sol20
2,3-Dimethylphenol	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	**0.25**	0.5
2,4-Dimethylphenol	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	**0.25**
2,5-Dimethylphenol	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2,6-Dimethylphenol	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,4-Dimethylphenol	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,5-Dimethylphenol	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	**0.25**	**0.25**	0.5	0.5
3-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	**0.25**	**0.25**
4-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5
B/Et-OH/acetone	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%

Table1. Experiment design (the numbers represent percentage of substance against methanol)

3. Results

The two resulting three-way (3D) data matrices (one for each method) were used for data analysis.

3.1. Data processing

Peak alignment, mean centering and normalization were the pre-processing steps applied to the dataset matrices generated. For peak alignment, Recursive Alignment through Fast Fourier Transform (RAFFT) [7, 8] was employed, aligning the data in the initial three dimensional matrix. The alignment was done for each m/z chromatogram separately

The normalization was made in two ways: between 0 and 1 and by an internal standard, the benzene peak area. Finally the data was mean centered following the sample direction in all 5 sets.

These approaches allowed us to perform a complete study and comparison between each type of preprocessing used: Alignment or not aligned, 0 to 1 normalization, benzene normalization or not normalized and mean centering or not mean centered.

In order to analyze the 3D data matrix with two-way methods (such as PCA, PLS and PLS-DA), different ways of converting the dataset into a 2D matrix were studied. In 2D matrices sample measurement data is arranged in different files and each column represents different descriptors or variables.

First, to obtain the mass spectra of each sample (MS matrix) from the 3D matrix we added the x axis (columns) of the chromatographic separation time for each m/z. On the other hand, summing all the m/z hits at a given scan (adding in the tube direction, y) we built the total ion chromatogram or TIC and its corresponding matrix, the GC matrix.

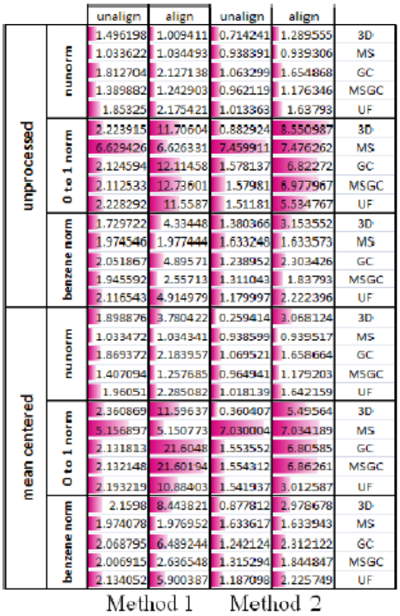
Anyway, by summation, the extra information brought by the third dimension (the time or the m/z axis) is lost. In order to compensate for this and still allow the data to be treated with two-way methods two additional approaches were used: unfolding (the UF matrix) and concatenation (the MSGC matrix).

Unfolding is done by taking the mass spectra of each chromatographic scan and pasting it where the previous scan ended. On the other hand concatenation keeps the extra information by concatenating the average mass spectra with the total ion chromatogram in a single file of the 2D matrix, with the variables of the TIC and the m/z spectra concatenated in consecutive columns.

2-D datasets were analyzed by Principal Component Analysis (PCA), Partial Least Squares Discriminant Analysis (PLS-DA), fuzzy ARTMAP and fuzzy voting, while the 3-D dataset was analyzed by Parallel Factor Analysis (PARAFAC), and multi-way PLS-DA.

3.2. PCA and PARAFAC results

The PCA and PARAFAC results were evaluated using a clusterization merit figure (based on the relationship between intra-class and inter-class distances, higher meaning better clustering in the PCA or PARAFAC scores graph) and by success classification rates using a fuzzy ARTMAP neural network.



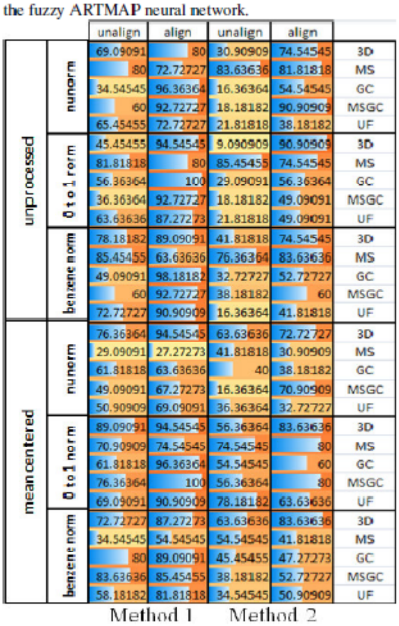
		Method 1		Method 2		
		unalign	align	unalign	align	
unprocessed	nunorm	1.496198	1.009411	0.714241	1.289555	3D
		1.033622	1.034493	0.938391	0.939306	MS
		1.812704	2.127138	1.063299	1.654868	GC
		1.389882	1.242903	0.962119	1.176346	MSGC
		1.85325	2.175421	1.013363	1.63793	UF
	0 to 1 norm	2.223915	11.70604	0.882924	8.550987	3D
		6.629426	6.626331	7.459911	7.476262	MS
		2.124594	12.11458	1.578137	6.82272	GC
		2.112533	12.73601	1.57981	6.977967	MSGC
		2.228292	11.5587	1.51181	5.534767	UF
	benzene norm	1.729722	4.33448	1.380366	3.153552	3D
		1.974546	1.977444	1.633248	1.633573	MS
		2.051867	4.89571	1.238952	2.303426	GC
		1.945592	2.55713	1.311043	1.83793	MSGC
		2.116543	4.914979	1.179997	2.222396	UF
mean centered	nunorm	1.898876	3.780422	0.259414	3.068124	3D
		1.033472	1.034341	0.938599	0.939517	MS
		1.869372	2.183957	1.069521	1.658664	GC
		1.407094	1.257685	0.964941	1.179203	MSGC
		1.96051	2.285082	1.018139	1.642159	UF
	0 to 1 norm	2.360869	11.59637	0.360407	5.49564	3D
		5.156897	5.150773	7.030004	7.034189	MS
		2.131813	21.6048	1.553552	6.80585	GC
		2.132148	21.60194	1.554312	6.86261	MSGC
		2.193219	10.88403	1.541937	3.012587	UF
	benzene norm	2.1598	8.443821	0.877812	2.978678	3D
		1.974078	1.976952	1.633617	1.633943	MS
		2.068795	6.489244	1.242124	2.312122	GC
		2.006915	2.636548	1.315294	1.844847	MSGC
		2.134052	5.900387	1.187098	2.225749	UF

Fig.3. Cluster analysis of 3D PARAFACT and 2D PCA of MS, GC, MSGC, and UF data. Intervariance/intravariance rapport. Higher values represent better clustering in the PCA/PARAFAC graph

As expected, method one proved easier to classify than method three. Best classification results were obtained for method one (where the peaks are more resolved), with the MSGC concatenated matrix when the data is aligned, 0 to 1 normalized and mean centered. By contrast method 3, the most difficult one from the point of view of chromatographic separation (peaks more coelluted), gives better results using PARAFAC when the data is aligned, normalized between 0 and 1 and not mean centered.

Aligning the data for chromatographic shift and normalizing it between 0 and 1 improves classification accuracy almost in all of the cases. An interesting result is that in the case of a less challenging method (method 1, the most resolved GC) 2D algorithms works best, whereas in a more complicated case such as method 2 (coeluted peaks, shortest measurement time) they lose effectiveness in front of 3D methods.

For the Fuzzy Artmap classification of the PCA (2D) and PARAFAC (3D) coordinates the data was separated into training and testing sets, using a 50-50 training-testing ratio. After applying PCA and PARAFAC for the training set the scores were normalized between 0 and 1 and fed into a fuzzy ARTMAP neural network for training. The test set was then projected on to the already created PCA/PARAFAC model, and the scores were also normalized between 0 and 1 and used as the test set for the fuzzy ARTMAP neural network.



		Method 1		Method 2		
		unalign	align	unalign	align	
unprocessed	nunorm	69.09091	80	30.90909	74.54545	3D
		80	72.72727	83.63636	81.81818	MS
		34.54545	96.36364	16.36364	54.54545	GC
		60	92.72727	18.18182	90.90909	MSGC
		65.45455	72.72727	21.81818	38.18182	UF
	0 to 1 rorm	45.45455	94.54545	9.090909	90.90909	3D
		81.81818	80	85.45455	74.54545	MS
		56.36364	100	29.09091	56.36364	GC
		36.36364	92.72727	18.18182	49.09091	MSGC
		63.63636	87.27273	21.81818	49.09091	UF
	benzene norm	78.18182	89.09091	41.81818	74.54545	3D
		85.45455	63.63636	76.36364	83.63636	MS
		49.09091	98.18182	32.72727	52.72727	GC
		60	92.72727	38.18182	60	MSGC
		72.72727	90.90909	16.36364	41.81818	UF
mean centered	nunorm	76.36364	94.54545	63.63636	72.72727	3D
		29.09091	27.27273	41.81818	30.90909	MS
		61.81818	63.63636	40	38.18182	GC
		49.09091	67.27273	16.36364	70.90909	MSGC
		50.90909	69.09091	36.36364	32.72727	UF
	0 to 1 norm	89.09091	94.54545	56.36364	83.63636	3D
		70.90909	74.54545	74.54545	80	MS
		61.81818	96.36364	54.54545	60	GC
		76.36364	100	56.36364	80	MSGC
		69.09091	90.90909	78.18182	63.63636	UF
	benzene norm	72.72727	87.27273	63.63636	83.63636	3D
		34.54545	54.54545	54.54545	41.81818	MS
		80	89.09091	45.45455	47.27273	GC
		83.63636	85.45455	38.18182	52.72727	MSGC
		58.18182	81.81818	34.54545	50.90909	UF

Fig.4. Fuzzy Artmap classification success rate for PARAFAC (3D) and PCA (MS, GC, MSGC, UF)

The results were judged by means of success rate of the confusion matrix. (Fig.4.).The results show that method 1 reaches better results when two-way analysis methods are employed while method 2 gives better results with three-way data analysis. For method 1, the best results were obtained by the PCA of the MSGC concatenated matrix when the data was aligned, mean centered and normalized between 0 and 1, scoring 100% success rate, followed by GC data with 96.36% success rate. In the case of the second method the best result was obtained using PARAFAC when the data was aligned, normalized from 0 to 1 and not mean centered, as well as using a PCA projection of the MSGC matrix when the data was just aligned.

3.3. PLS-DA and n-PLS-DA

For PLS and n-PLS-DA analysis, each dataset was separated into training and testing sets by a chosen training/test ratio of 7/3. The training set was used to train the PLS and n-PLS models. Onto these models the test set was projected.

The model and prediction performance were evaluated by means of the Root Mean Square Error of Cross-Validation (RMSECV), and the Root Mean Square Error of Prediction (RMSEP), respectively. RMSECV and RMSEP represent cross-validation error and prediction error respectively.

The results present the sample prediction success rate for methods one and two for solutions 10 to 20, and are shown in Fig.5. and Fig.6.

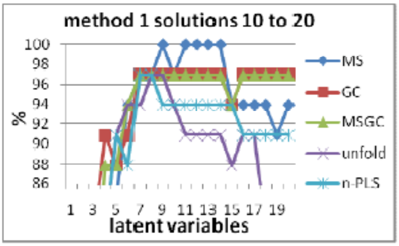


Fig.5. Prediction success rate for method 1

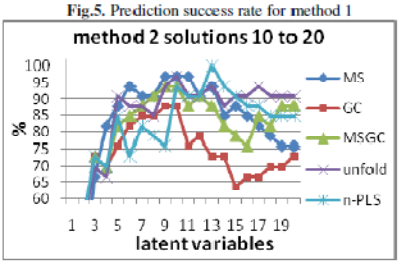


Fig.6. Prediction success rate for method 2

We notice that method 2, in which peaks are more coeluted, gives worse results than method 1, where a temperature programmed separation was employed in order to achieve a better separation. We can clearly notice this in the chromatographic data sample success rate.

Even though the average mass spectra gives the best performance when classifying the samples, the combination of this information with the total ion chromatogram by means of unfolding or unifying the MS data with the TIC data almost always gives better results than MS or TIC data alone (specially in lower latent variable models). Between the unfolded data and the unified MSGC data the later one is appearing to give better results and presents the advantage of an easier model interpretation.

The multi-way PLS model seems to present consistent results on both methods, yielding a high prediction success rate in both datasets. The n-PLS algorithm presented the best results in the most difficult case, the second method, solutions 10 to 20, giving 100% success rate classification for 13 latent variables. (Fig.6.)

4. Conclusions

The experiment has shown that the addition of data from a chromatographic separation improves the results compared to using the Mass Spectra alone in most of the cases.

Treating the data with 2-way (PCA and PLS-DA) and 3-way (PARAFAC and n-PLS-DA) methods showed that working with a well resolved chromatogram 2-way methods work best, whereas for coeluted peaks (shorter GC runs) three-way methods are better suited.

From the point of view of data pretreatment, alignment is an important step, removing the possibility of sample classification based on signal drift or injection error.

Acknowledgments

This work has been funded by Ciberdem, The CIBER de Diabetes y Enfermedades Metabólicas Asociadas es an ISCIII initiative.

References

1. R Bro, "PARAFAC. Tutorial and applications", Chemometrics and Intelligent Laboratory Systems 38 (1997) 149-171
2. N.Fabera, R.Brob, P.Hopke, "Recent developments in ANDECOMP/PARAFAC algorithms:a critical review", Chemometrics and Intelligent Laboratory Systems 65 (2003) 119– 137
3. P. Geladi, "Analysis Of Multi-Way (Multi-Mode) Data", Chemometrics And Intelligent Laboratory Systems 7 (1-2): 11-30 Dec 1989
4. V. Pravdova, C. Bouconb, S. de Jong, B. Walczak, D.L. Massart, "Three-way principal component analysis applied to food analysis: an example", Analytica Chimica Acta 462 (2002) 133–148
5. P. Barbieri, G. Adami, S. Piselli, F. Gemiti, E. Reisenhofer, "A three-way principal factor analysis for assessing the time variability of freshwaters related to a municipal water supply", Chemometrics and Intelligent Laboratory Systems 62 (2002) 89–100
6. Tucker, L. R. 1964. The extension of factor analysis to three-dimensional matrices. In Contributions to Mathematical Psychology. Holt, Rinehart and Winston, New York, 110-182
7. Wong, J.W.H, Cagney, G. and Cartwright, H.M. (2005) SpecAlign - processing and alignment of mass spectra datasets. Bioinformatics, 21: 2088-2090
8. Wong, J.W.H. and Cartwright, H.M. (2005) - An application of Fast Fourier Transform cross-correlation for the alignment of large chromatographic and spectral datasets. Analytical Chemistry, 77: 5655-5661

Congress contribution II

Abstract II

Poster Presentation at:

ISOEN 2009
University of Brescia - Italy
15-17 April

Cosmin Burian, **Jesus Brezmes**, Maria Vinaixa, Eduard Llobet, Xavier Vilanova, Nicolau Cañellas and Xavier Correig, "A fuzzy ARTMAP approach to the incorporation of chromatographic retention timeinformation to a MS based E-nose"

A fuzzy ARTMAP approach to the incorporation of chromatographic retention time information to a MS based E-nose

Cosmin Burian[1], Jesus Brezmes [1,2], Maria Vinaixa [1,2], Eduard Llobet[1], Xavier Vilanova[1], Nicolau Cañellas[1] and Xavier Correig [1,2]
[1]Department of Electronic Engineering, Universitat Rovira i Virgili,
Avenida Paisos Catalans 26, 43007 Tarragona, Spain
[2]CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM)
cosmin.burian@urv.cat, jesus.brezmes@urv.cat,

Abstract

This paper presents the work done with Fuzzy ARTMAP neural networks in order to improve the performance of mass spectrometry-based electronic noses using the time retention of a chromatographic column as additional information. Solutions of nine isomers of dimethylphenols and ethylphenols were used in this experiment. The gas chromatograph mass spectrometer response was analyzed with an in-house developed Fuzzy ARTMAP neural network, showing that the combined information (GC plus MS) gives better results than MS information alone in most of the cases.

1. Introduction

The goal of this work is to improve the performance of an MS based Electronic Nose by using the combination of Gas Chromatography and Mass Spectrometry. This approach adds extra information to the dataset, since both the average mass spectra (MS) and Total Ion Chromatogram (TIC) of the GC/MS measurements are recorded. Fuzzy ARTMAP Neural Networks [1] are used to combine both data dimensions. To generalize results, three different chromatographic retention times were investigated.

Even though in many experiments the MS electronic nose has proven better than gas sensor based multi-sensor systems [2], some complex mixtures have proven difficult to classify with any E-nose technique. In most of the cases the methods used for the classification of this kind of data are two and three-way methods such as PCA, PLS, PARAFAC and n-PLS. By using a fuzzy ARTMAP neural network combining the chromatographic and mass spectra data in two different ways (concatenation and by addition of the neural network results) we want to prove that the extra information added by the chromatographic separation is improving the results in a challenging experiment.

2. Experimental

Twenty mixtures of nine isomers of dimethylphenol and ethylphenol were measured and analyzed by means of gas chromatography mass spectrometry (GC-MS). In order to have a challenging data set, the nine isomers were chosen based on their theoretically similar mass spectra.

For the experiment design we looked at 2 key observations: the PCA of the 9 isomers mass spectra and their chromatographic retention time. To obtain the chromatographic retention time (Fig.1.), 9 solutions of 1% isomer in methanol containing only one isomer were prepared and analyzed. In order to see which of the 9 isomers had most alike mass spectra we performed a PCA of the theoretical mass spectra of each isomer (Fig.2.).

Based on this information we designed the experiment as shown in **Table 1.** Benzene acts as an internal standard, and having the highest concentration it is used in the normalization pretreatment of the data. The experiment was designed in order to having a challenging sample set for the mass spectra-based electronic nose.

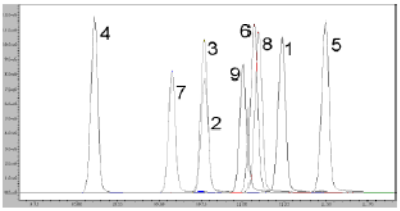


Fig.1. Chromatographic retention times for the 9 isomers as follows: **1)** 2,3-Dimethylphenol;**2)** 2,4-Dimethylphenol;**3)** 2,5-Dimethylphenol;**4)** 2,6-Dimethylphenol;**5)** 3,4-Dimethylphenol;**6)** 3,5-Dimethylphenol;**7)** 2-Ethylphenol;**8)** 3-Ethylphenol and **9)** 4-Ethylphenol;

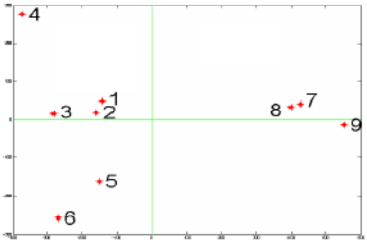


Fig.2. Principal Component Analysis of the isomers mass spectra

Three chromatographic methods were used. In method one, we tried to separate as much as possible the isomers. To achieve this, a temperature programmed separation was used, starting at 50°C, where the temperature was kept constant for one minute, until 180°C, where almost all the isomers were separated, giving a retention time of 20 min. Method two and three were designed to give more coeluted peaks, and therefore isothermal separations at 175 and 190°C were used, reducing the retention time to no more than 5 min. The measurements were conducted through syringe injection of 1 µl per measurement, and ten repetitions for each method and solution were made. Two blanks were also measured, one consisting of just methanol and the other with methanol and 2% of benzene.

component	Sol1	Sol2	Sol3	Sol4	Sol5	Sol6	Sol7	Sol8	Sol9	Sol10	Sol11	Sol12	Sol13	Sol14	Sol15	Sol16	Sol17	Sol18	Sol19	Sol20
2,3-Dimethylphenol	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.25	0.5
2,4-Dimethylphenol	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.25
2,5-Dimethylphenol	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2,6-Dimethylphenol	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,4-Dimethylphenol	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,5-Dimethylphenol	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.25	0.25	0.5	0.5
3-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.25	0.25
4-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.5	0.5	0.5	0.5
BIEt-OH/acetone	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%

Table1. Experiment design (the numbers represent percentage of substance into methanol)

3. Results

Having the data in a 3D format, we added the x axis of the chromatographic separation time obtaining the average mass spectra, which represented the MS approach. We also added the y axis of the m/z fragments, obtaining the total ion chromatogram, giving the GC approach. These were the input matrices that were fed into the fuzzy ARTMAP routines.

The two sets of data (MS and GC) were individually pretreated by normalization, and fed into an in-house developed fuzzy ARTMAP algorithm which tried to classify each solution. The network was tested using the leave-one-out cross-validation method: Given n measurements, the network was trained n times using n-1 training vectors. The vector left out in the training phase was then used for testing. Performance was estimated as the average performance over the n tests. Both data matrices, the MS data and GC data, were normalized because the fuzzy ARTMAP network needs the input data to be between 0 and 1.

The results of the fuzzy ARTMAP neural network were displayed as a percentual success rate and in the form of a confusion matrix.

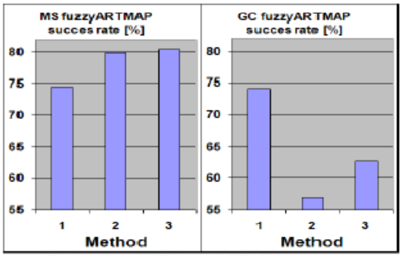


Fig.3. MS and GC fuzzy ARTMAP success rates

Because for the MS approach we coeluted the chromatographic peaks through software by adding the time dimension, we would expect a similar result of the percentual success rate for the three chromatographic methods. The fact that method one gives the worst results, even though it is the most resolved chromatographic method can be attributed to the fact that the molecules fragmentate differently for a resolved chromatogram, where each separated isomer fragmentates, compared to a coeluted chromatogram where all the isomers arrive to the ionization chamber and fragmentate all together. Because of this summing the retention time is not the same as having a coeluted chromatogram in the first place. This fact explains why the results of the 2 isothermal methods give similar results.

From the GC point of view, the first method of a temperature programmed separation, in which almost all the isomers were separated, gives better results than the isothermal methods 2 and 3, because the signal output of method one contains more information that discriminates better different type of samples. Again the classification success rates of the more coeluted methods 2 and 3 are closer.

The confusion matrix (Fig. 4) shows the results of the fuzzy ARTMAP neural network as real solutions (rows) vs. the results predicted (columns) by the neural network. In a case of 100% success rate the diagonal of the matrix should present all 10 repetitions. In the case of the 20 solutions we can see that not all the solutions are predicted correctly, proving that this experiment is a challenge for the MS system

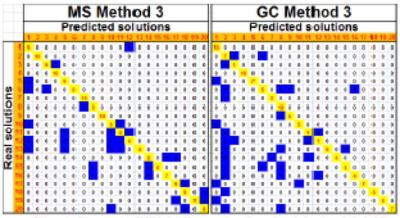


Fig.4. Fuzzy ARTMAP results for the Method 3 MS and GC (real (y) vs. predicted (x) solution)

The results show a better performance for the MS approach than for the GC data (Fig. 3), which happens because the MS fingerprint is more specific than the GC fingerprint. We also see a difference between the chromatographic methods which indicates that a more coeluted pick is performing better for the MS sensor.

Comparing the fuzzy ARTMAP results we can see (fig. 4) that the errors on the confusion matrix are different in the MS dimension than in the GC dimension. This is a clear sign that using both information dimensions (time and m/z) should improve the classifying ability of the MS-based electronic nose.

To improve the results of the fuzzy ARTMAP neural network a voting strategy was implemented [1]. The 10 repetitions of each solution were scrambled, and each solution group was divided into training and testing measurements by a given training-evaluation ratio. The number of measurements used for training ranged between 1 and 9. These data was fed into a fuzzy ARTMAP neural network by a given number (10) of times (votes) and results were recorded. The winning class was decided by the total number of votes each solution received, so that the one with more votes was selected as the output since it was the most probable prediction. Initially, the input data for the fuzzy ARTMAP voting strategy were the average mass spectra (MS) and total ion chromatogram (GC) matrices. The matrices were normalized between 0 and 1.

In order to take advantage of both data types we followed 2 approaches: 1) Concatenation of the MS and GC matrices into a single matrix called MSGC and 2) by summing the MS votes with the GC votes for a given measurement.

By combining the mass spectra and chromatographic information through the voting approach, the MS and GC matrix were fed into the fuzzy ARTMAP independently and then the votes that they had collected were added and the winning class was decided.

Possible outputs from the algorithm were correctly classified, wrongly classified or unclassified, when the number of votes for one solution was the same with the number of votes for another solution. Because of the high number of unclassified votes in the combined MS and GC voting method, and because the MS proved to give better results, we decided that in case of a misclassification (tie) in the voting strategy of the MS and GC matrices the winner should be the MS winner (MSGC vote MS improved).

The results were monitored by the success rate based on the confusion matrix (real solution vs. predicted solution).

Looking at Fig.5 we can see that the MSGCvote MS results are always on top of the other approaches if 1 to 8 measurements are used for training. Because in practice a high number a measurements is hard to obtain this shows that the MSGCvoteMS approach is the best one, giving the best results when fewer measurements are used for training.

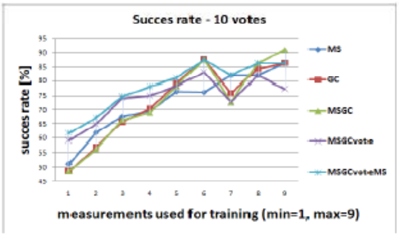


Fig.5. Voting strategy Fuzzy ARTMAP success rates

Using 9 of the 10 measurements for training leads to worse results because there is only one measurement left for evaluation, and even a few misclassified samples would have a great impact on the results, leading to a lower success rate. That's why we recommend that the measurements used for training should be between 50 to 60 percent.

The highest success rates, in the range of 90%, were obtained by the MSGC vote MS improved method, which takes advantage of both methods, but does not have the drawback of the MSGC vote method of misclassification.

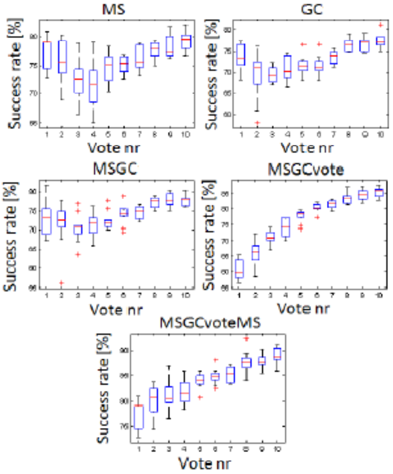
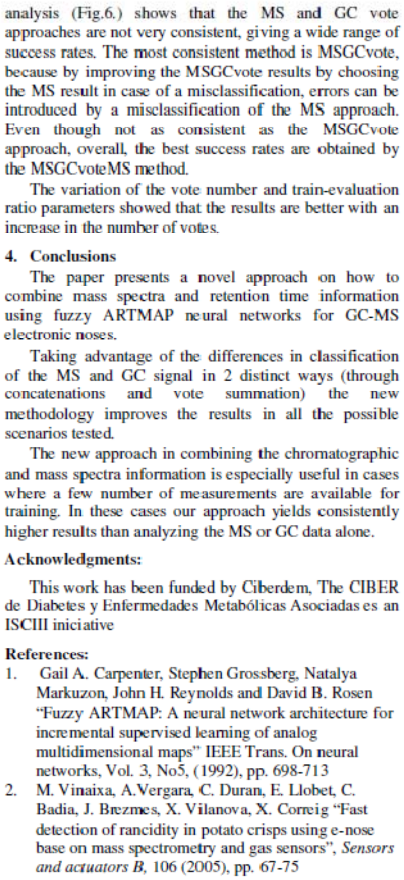


Fig.6. Success rates for MS, GC, MSGC, MSGCvote and MSGCvoteMS for 5 train – 5 evaluation measurements, box plot for 10 repetitions

Repeating the analysis 10 times for each approach, using 5 measurements for training and 5 for evaluation gives us the chance to analyze the consistency of the results through box and whisker plots. This statistic



analysis (Fig.6.) shows that the MS and GC vote approaches are not very consistent, giving a wide range of success rates. The most consistent method is MSGCvote, because by improving the MSGCvote results by choosing the MS result in case of a misclassification, errors can be introduced by a misclassification of the MS approach. Even though not as consistent as the MSGCvote approach, overall, the best success rates are obtained by the MSGCvoteMS method.

The variation of the vote number and train-evaluation ratio parameters showed that the results are better with an increase in the number of votes.

4. Conclusions

The paper presents a novel approach on how to combine mass spectra and retention time information using fuzzy ARTMAP neural networks for GC-MS electronic noses.

Taking advantage of the differences in classification of the MS and GC signal in 2 distinct ways (through concatenations and vote summation) the new methodology improves the results in all the possible scenarios tested.

The new approach in combining the chromatographic and mass spectra information is especially useful in cases where a few number of measurements are available for training. In these cases our approach yields consistently higher results than analyzing the MS or GC data alone.

Acknowledgments:

This work has been funded by Ciberdem, The CIBER de Diabetes y Enfermedades Metabólicas Asociadas es an ISCIII iniciative

References:

1. Gail A. Carpenter, Stephen Grossberg, Natalya Markuzon, John H. Reynolds and David B. Rosen "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps" IEEE Trans. On neural networks, Vol. 3, No5, (1992), pp. 698-713
2. M. Vinaixa, A.Vergara, C. Duran, E. Llobet, C. Badia, J. Brezmes, X. Vilanova, X. Correig "Fast detection of rancidity in potato crisps using e-nose base on mass spectrometry and gas sensors", *Sensors and actuators B,* 106 (2005), pp. 67-75

Congress contribution II

Poster

Poster Presentation at:

ISOEN 2009
University of Brescia - Italy
15-17 April

Cosmin Burian, **Jesus Brezmes**, Maria Vinaixa, Eduard Llobet, Xavier Vilanova, Nicolau Cañellas and Xavier Correig, "A fuzzy ARTMAP approach to the incorporation of chromatographic retention timeinformation to a MS based E-nose"

A Fuzzy ARTMAP Approach To The Incorporation Of Chromatographic Retention Time Information To A MS Based E-Nose

C. Burian, M. Vinaixa, J. Brezmes, Eduard Llobet, Xavier Vilanova, Nicolau Cañellas, X. Correig
DEEEA, ETSE, Universitat Rovira i Virgili, Avda. Països Catalans, 26, 43007 Tarragona

INTRODUCTION

- Even if the MS sensor has proven itself as a valuable tool in the electronic nose field, some complex mixtures have shown more difficult to be classified. In most of the cases the methods used for the classifications were two-way methods. By adding the time dimension to the m/z fractions and their intensity we hope to improve the performance of the electronic nose.
- This paper presents the work done with Fuzzy ARTMAP neural networks in order to improve the performance of mass spectrometry-based electronic noses using the time retention of a chromatographic column as additional information.
- Solutions of nine isomers of dimethylphenols and ethylphenols were used in this experiment.
- The gas chromatograph mass spectrometer response was analyzed with a Fuzzy ARTMAP neural network, showing that the combined information (GC plus MS) gives better results than MS information alone in most of the cases.

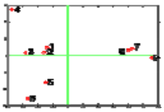
OBJECTIVES

- To improve the performance of the mass spectrometry-based electronic nose by using the combination of gas chromatography and mass spectrometry adding the third dimension representing the time parameter of the chromatographic separation
- To use the extra information of a chromatographic separation through another approach than the three-way methods
- To optimize the analysis through different chromatographic retention times

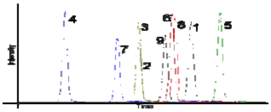
EXPERIMENT

In order to challenge the mass spectra-based electronic nose we chose for the experiment 9 organic isomers of dimethylphenol and ethylphenol

For the experiment design we looked at two aspects:
1) the PCA of the 9 isomers mass spectra
2) their chromatographic separation



Principal Component Analysis of the isomers mass spectra



Chromatographic retention times for the 9 isomers

Based on this information we designed the experiment as follows:

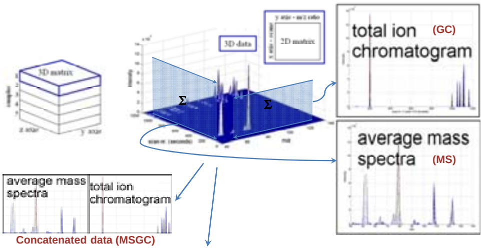
component	Sol1	Sol2	Sol3	Sol4	Sol5	Sol6	Sol7	Sol8	Sol9	Sol10	Sol11	Sol12	Sol13	Sol14	Sol15	Sol16	Sol17	Sol18	Sol19	Sol20
2,3-Dimethylphenol	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	**0.25**	0.5
2,4-Dimethylphenol	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	**0.25**
2,5-Dimethylphenol	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2,6-Dimethylphenol	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,4-Dimethylphenol	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3,5-Dimethylphenol	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	**0.25**	**0.25**	0.5	0.5
3-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	**0.25**	**0.25**
4-Ethylphenol	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-	0.5	0.5	0.5	0.5	0.5	0.5	**0.25**	0.5	0.5	0.5	0.5
Benzene	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%

To optimize the analysis three chromatographic methods were used. In method one we used a temperature programmed separation, and separated almost all the isomers. Method 2 and 3 were isothermal separations designed to give more coeluted peaks.

RESULTS

The data was recorded from the GC/MS in a 3D format and separated into two groups by adding the time and the m/z dimension respectively

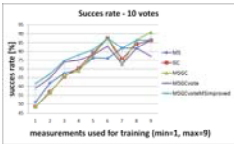
The two data groups, the average mass spectra group (MS) and the total ion chromatogram group (GC) were fed into a leave one out fuzzy ARTMAP algorithm:



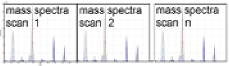
Concatenated data (MSGC)

Errors on the confusion matrix are different in the MS dimension than in the GC dimension. Using both information dimensions (time and m/z) should improve the classifying ability of the MS-based electronic nose.

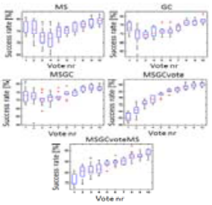
To improve the results of the fuzzy ARTMAP neural network a voting strategy was implemented



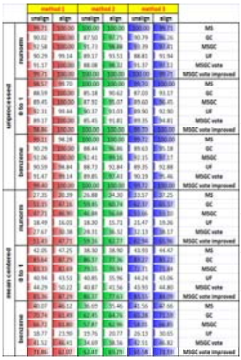
Voting strategy Fuzzy ARTMAP success rates shows that using from 5 to 8 measurements for training is the best choice



Unfolded data (UF)



Box plot for success rates for MS, GC, MSGC, MSGCvote and MSGCvoteMS for 5 train measurements, 10 repetitions



		method 1		method 2		method 3		
		unalign	align	unalign	align	unalign	align	
unprocessed	maximum	99.71	100.00	100.00	100.00	100.00	99.71	MS
		90.02	100.00	87.50	97.75	90.79	96.26	GC
		92.58	100.00	91.73	98.88	93.79	97.81	MSGC
		90.29	99.14	89.17	93.53	88.81	91.94	UF
		91.17	100.00	88.08	98.32	91.37	97.11	MSGC vote
		99.71	100.00	100.00	100.00	100.00	99.71	MSGC vote improved
	0 to 1	98.57	99.70	100.00	100.00	99.70	100.00	MS
		88.59	100.00	85.18	90.62	87.00	93.17	GC
		89.45	100.00	87.92	95.07	89.60	96.45	MSGC
		92.31	99.44	90.37	91.03	89.90	92.90	UF
		89.17	100.00	85.45	91.81	89.35	94.81	MSGC vote
		98.86	100.00	100.00	100.00	99.70	100.00	MSGC vote improved
	benzene	99.11	98.28	100.00	100.00	99.72	100.00	MS
		90.29	100.00	88.44	96.86	89.63	95.18	GC
		92.06	100.00	91.41	99.16	92.15	97.17	MSGC
		90.59	98.84	88.73	92.84	89.35	92.88	UF
		91.47	99.14	89.85	97.41	90.19	95.46	MSGC vote
		98.40	100.00	100.00	100.00	99.72	100.00	MSGC vote improved
mean centered	maximum	27.35	20.39	26.88	34.30	33.57	37.25	MS
		51.15	47.16	59.85	60.74	62.37	60.57	GC
		47.71	46.90	46.89	56.69	53.66	53.10	MSGC
		18.49	16.01	18.30	15.71	21.47	19.26	UF
		27.67	30.38	28.31	36.52	32.13	38.17	MSGC vote
		51.43	47.71	59.36	62.77	62.94	65.76	MSGC vote improved
	0 to 1	42.05	47.25	38.30	38.90	43.93	44.47	MS
		85.64	87.25	86.37	77.36	80.27	81.22	GC
		80.33	82.69	79.55	78.94	73.71	73.84	MSGC
		40.94	43.53	40.85	35.96	44.24	43.06	UF
		44.29	50.22	40.87	41.56	43.93	44.80	MSGC vote
		85.36	87.29	86.37	77.63	83.55	84.29	MSGC vote improved
	benzene	40.07	46.12	36.69	35.46	41.56	47.56	MS
		70.74	81.49	62.45	64.76	65.28	71.59	GC
		66.71	83.80	57.87	62.96	58.01	66.85	MSGC
		18.77	23.90	19.76	20.77	26.13	30.65	UF
		41.52	46.41	34.69	38.36	42.51	46.82	MSGC vote
		71.86	83.07	62.87	65.79	60.58	73.59	MSGC vote improved

Average of 3 repetitions of success rates for MS, GC, MSGC, UF, MSGCvote and MSGCvote improved shows that the success rates for the MSGCvote improved are always the best ones.

In case of an indecision in the voting strategy MSGCvote improved chooses the best result between the MS and GC

It also presents the results for three types of data pretreatment:
1.Alignment with RAFFT algorithm
2.Normalization between 0 and 1 and by the internal standard (benzene)
3.Mean centering

CONCLUSIONS

- a novel approach on how to combine mass spectra and retention time information using fuzzy ARTMAP neural networks has been shown
- taking advantage of the differences in classification of the MS and GC signal in 2 distinct ways (through concatenations and vote summation) the new methodology improves the results in all the possible scenarios tested.
- the new approach in combining the chromatographic and mass spectra information is especially useful in cases where a few number of measurements are available for training. In these cases our approach yields consistently higher results than analyzing the MS or GC data alone

Acknowledgments
This work has been funded by Ciberdem, The CIBER de Diabetes y Enfermedades Metabólicas Asociadas is an ISCIII iniciative