UNIVERSITAT ROVIRA I VIRGILI

Departament de Química Analítica i Química Orgànica

METODOLOGÍAS ANALÍTICAS BASADAS EN ESPECTROSCOPIA DE INFRARROJO Y CALIBRACIÓN MULTIVARIANTE. APLICACIÓN A LA INDUSTRIA PETROQUÍMICA

Tesis Doctoral

Santiago Macho Aparicio

Tarragona, 2002

Metodologías analíticas basadas en espectroscopia de infrarrojo y calibración multivariante. Aplicación a la industria petroquímica.

Tesis Doctoral

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI
Departament de Química Analítica i Química Orgànica
Àrea de Química Analítica

METODOLOGÍAS ANALÍTICAS BASADAS EN ESPECTROSCOPIA DE INFRARROJO Y CALIBRACIÓN MULTIVARIANTE. APLICACIÓN A LA INDUSTRIA PETROQUÍMICA.

Memoria presentada por
Santiago Macho Aparicio
para conseguir el grado de
Doctor en Química
Tarragona, 2002

UNIVERSITAT
ROVIRA I VIRGILI

DEPARTAMENT DE QUÍMICA ANALÍTICA
I QUÍMICA ORGÀNICA

Plaça Imperial Tàrraco, 1
43005 Tarragona
Tel. 34 977 55 81 37
Fax 34 977 55 95 63
e-mail: secqaqo@quimica.urv.es

Dra. MARIA SOLEDAD LARRECHI GARCÍA, profesora titular del Departament de Química Analítica i Química Orgànica de la Facultat de Química de la Universitat Rovira i Virgili,

CERTIFICA: Que la presente memoria que tiene por título: "METODOLOGÍAS ANALÍTICAS BASADAS EN ESPECTROSCOPIA DE INFRARROJO Y CALIBRACIÓN MULTIVARIANTE. APLICACIÓN A LA INDUSTRIA PETROQUÍMICA", ha sido realizada por SANTIAGO MACHO APARICIO bajo mi dirección en el Àrea de Química Analítica i Química Orgànica de esta Universidad y que todos los resultados presentados son fruto de las experiencias realizadas por dicho doctorando.

Tarragona, abril de 2002

Dra. Mª Soledad Larrechi

AGRADECIMIENTOS

Quisiera que las próximas líneas sirvan como reconocimiento a las personas e instituciones que han hecho posible la realización de esta tesis. A todas ellas, quiero agradecer de todo corazón la ayuda dispensada:

A mi directora, Mª Soledad Larrechi García, por su guía y sus valiosas aportaciones, sin las que este trabajo no hubiera sido posible.

Al director del grupo de Quimiometría y Cualimetría, F. X. Rius Ferrús, por la oportunidad que me brindó de realizar este trabajo, y lo mucho que me ha enseñado.

A los miembros del grupo con los que he trabajado de forma más estrecha, que tanto me han aportado: Pili, Antoni, Florenci y Ricard.

Al resto de miembros actuales del Grupo, así como a los antiguos miembros, que me han ayudado y ofrecido su amistad, afortunadamente la lista es larga, pero sentiría mucho dejarme a alguien: Itziar, Jordi, Joan, Jaume, Javi, Angel, Alicia P., Alicia M., Josep Lluís, Mari, Enric, Esther, Paquita, Gemma, Núria, Alberto, Pablo, Sara, Barbara...

A los compañeros y laborantes del Área de Química Analítica, del resto del Departamento y del Departamento de Física i Inorgànica, especialmente a mis compañeros de promoción: Alex, Pere y Ramón.

A la Universitat Rovira i Virgili, a la Fundación REPSOL y al Departament d'Universitats Recerca i Societat de la Informació de la Generalitat de Catalunya por el soporte económico que ha permitido realizar esta tesis.

Mi más sincero agradecimiento a todos aquellos miembros de las empresas con las que he tenido contacto en el transcurso de esta tesis: Repsol Petróleo Tarragona y Transformadora de Propileno, muchas son las personas han ayudado a hacer posible este trabajo, pero quisiera nombrar a E. Moratinos y C. García.

A mi familia y amigos, que también han puesto su grano de arena y me han apoyado incondicionalmente.

A Inma, muy especialmente, siempre a mi lado y dispuesta a ayudarme. Gracias por el ánimo y el apoyo que me has dado.

A Inma

ÍNDICE

Objetivo

OBJETIVO

El objetivo general de la presente tesis doctoral ha sido el desarrollo y aplicación de metodologías analíticas basadas en la combinación de medidas espectroscópicas de infrarrojo medio y cercano con métodos quimiométricos de análisis multivariante. Las aplicaciones desarrolladas están enfocadas a la determinación cuantitativa de diferentes propiedades físicas y químicas de interés en la industria petroquímica.

Este objetivo general, enunciado arriba, se puede desglosar en los siguientes objetivos más detallados:

1 Desarrollo de modelos de calibrado multivariantes basados en medidas espectroscópicas.
2 Estudio de la estabilidad en el tiempo de los modelos desarrollados y su transferencia a nuevas situaciones experimentales.
3 Detección de situaciones anómalas, no contempladas en el proceso de calibrado (detección de muestras discrepantes).

Las aplicaciones sobre las que se ha trabajado han sido:

a) La determinación del contenido de las diferentes familias de hidrocarburos que constituyen la nafta.
b) El análisis del contenido en etileno y otras propiedades mecánicas, como la viscosidad o el índice de fluidez, de diferentes tipos de polímeros y copolímeros de polipropileno.

Estos dos tipos de muestras, la nafta y el polipropileno, constituyen dos problemas industriales diferentes entre sí, tanto por el diferente estado físico de la muestra como por las diferencias en los procesos industriales involucrados. La

nafta es una fracción líquida del petróleo, obtenida por destilación, cuya composición depende de la procedencia del petróleo de partida así como de las operaciones de separación que se han aplicado a este crudo. Los copolímeros de polipropileno son sólidos, producto de la polimerización de propileno en presencia de otros polímeros, y sus características finales dependerán de la naturaleza y proporción de los polímeros añadidos.

1. Introducción

1.1 PRESENTACIÓN DE LOS CONTENIDOS

Desde el punto de vista industrial existe gran interés en el desarrollo de metodologías analíticas que proporcionen gran cantidad de información, que ésta sea de calidad y que además lo hagan en el menor tiempo posible. Esta idea, que es aplicable a cualquier proceso industrial, cobra especial relevancia en un sector como el de la industria petroquímica, en la que es habitual la monitorización y control de los procesos basándose en la información que proporciona el laboratorio de análisis.

Normalmente los métodos de análisis utilizados en estos laboratorios son métodos normalizados por organismos oficiales, específicos para cada propiedad o analito a determinar. Frecuentemente tienen ciertas características que los hacen poco eficaces en el entorno industrial, como puede ser un elevado tiempo de análisis o la necesidad de una manipulación intensiva de la muestra previa al análisis. Estas características hacen que sean técnicas poco adecuadas para un control en línea de los procesos industriales, tanto por la demora en obtener el resultado, como porque es técnicamente difícil acoplarlas a la línea de producción.

La espectroscopia de infrarrojo, tanto en la zona del infrarrojo medio como el infrarrojo cercano, resulta una técnica atractiva en el ámbito industrial porque:

a) Proporciona una información química muy versátil. La espectroscopia molecular, como consecuencia del tipo de transiciones propias de esta zona, aporta información de una gran cantidad de compuestos químicos. Esta propiedad, combinada a las técnicas multivariantes, permite considerar la determinación simultánea de más de un analito o propiedad de interés a partir del espectro registrado de la muestra.

b) La obtención del espectro se hace de una forma rápida y no es necesario un pretratamiento de la muestra, lo que facilita el acoplamiento de esta técnica a la línea de producción (análisis en línea).

La falta de especificidad de la señal analítica registrada, característica ventajosa desde el punto de vista de la versatilidad, tiene como inconveniente que obliga al empleo de técnicas de análisis multivariante, en concreto técnicas de calibrado multivariante. Estos métodos extraen eficazmente la información de los datos, aunque necesitan un adecuado estudio del problema abordado y una etapa de desarrollo de la metodología más laboriosa que los análisis clásicos.

En esta tesis doctoral se ha abordado la determinación del contenido total en parafinas, isoparafinas, naftenos y aromáticos, así como del contenido desglosado por número de átomos de carbonos de cada una de estas familias de hidrocarburos, en muestras de nafta, utilizando la señal registrada tanto en la zona del infrarrojo medio como en el infrarrojo cercano. También se ha abordado la determinación del contenido en etileno, viscosidad e índice de fluidez de muestras de polipropileno a partir de la señal registrada en la zona del infrarrojo cercano.

La técnica de calibración multivariante que se ha utilizado ha sido la calibración multivariante por mínimos cuadrados parciales (PLS). Las diferentes propiedades de las muestras se han determinado en los propios laboratorios de las industrias colaboradoras, utilizando los métodos de análisis habituales y que son la cromatografía de gases en el caso de la nafta y la espectroscopia IR, así como otros análisis físicos, en el caso de los polímeros. Los resultados de estos análisis se han utilizado para construir y comprobar la validez de los modelos de calibración desarrollados.

En la etapa de validación de los modelos de calibrado se realiza una estimación del error de predicción futuro. Habitualmente esta estimación consiste en un único

valor de error promedio que da una información de las características predictivas globales del modelo, aunque en algunas situaciones, como por ejemplo el ámbito industrial, puede ser preferible proporcionar un error de predicción específico para cada muestra, equivalente a los intervalos de predicción habituales en la calibración univariante. En el campo de la calibración multivariante por métodos que utilizan la descomposición en factores existen varias expresiones propuestas para calcular el error de predicción específico, aunque no existe una comúnmente aceptada, por lo que se ha estudiado el uso de diferentes expresiones en algunas de las aplicaciones tratadas en esta tesis.

El establecimiento de esta metodología analítica no acaba con el desarrollo y posterior validación del modelo de calibrado propiamente dicho, si no que es necesario establecer un procedimiento de control de su validez con el tiempo. Con esta finalidad se ha abordado el tema de la detección de situaciones anómalas de utilización de los modelos, no contempladas en el proceso de calibrado, así como la estabilidad de los sistemas con el tiempo. Estos dos aspectos se han abordado tanto mediante técnicas de control de tipo univariante, como los gráficos de Shewhart, cómo mediante técnicas de control de tipo multivariante, como el estadístico T^2 de Hotelling o el estadístico Q.

Cuando se detecta la perdida de validez del modelo, es posible la aplicación de métodos de corrección que permiten seguir utilizando el modelo. Esta modificación del modelo requiere una cierta experimentación, pero que sería mucho menor de la que comportaría un proceso de recalibrado completo, lo que se traduce en un importante ahorro de trabajo y tiempo. Con esta finalidad se han aplicado diferentes técnicas de transferencias de modelos, como la corrección de la pendiente y el sesgo (*slope-bias correction* , SBC) o la estandarización directa por partes (*piecewise direct standardisation*, PDS).

La aplicación práctica de estas metodologías ha permitido comprobar que en ocasiones no se tienen en cuenta algunas consideraciones básicas que ayudan a evaluar las posibilidades de éxito de la aplicación propuesta, como por ejemplo la relación entre el espectro infrarrojo y la propiedad a determinar, la existencia de agrupaciones en las muestras o el hecho de trabajar en intervalos de la propiedad de interés muy estrechos en comparación con la reproducibilidad de la técnica de referencia. Otros aspectos como las distintas estrategias para estandarizar o actualizar el modelo, son de vital importancia en el caso de la aplicación de la calibración multivariante a problemáticas industriales y que hay que planificar desde el mismo momento del desarrollo del modelo. Todos estas consideraciones han sido aplicadas a las aplicaciones desarrolladas en la presente tesis doctoral.

Otro aspecto importante de la calibración multivariante es la selección de variables, en contraposición a la utilización del espectro completo para construir el modelo. Cuando el número de longitudes de onda es muy grande se pueden introducir mejoras en el modelo de calibrado reduciendo el número de éstas, eliminando las que contienen un mayor grado de ruido, información no relevante o términos no lineales y trabajando únicamente con aquellas que están directamente relacionadas con el analito o la propiedad de interés. La selección de longitudes de onda que permite construir modelos más simples y por tanto más sencillos de interpretar ha sido aplicada al desarrollo de modelos en naftas.

1.2 CASOS PRÁCTICOS ESTUDIADOS

1.2.1 Determinación de PIONA en nafta

La nafta es el producto de partida para la producción de alquenos (olefinas) que se realiza a través del proceso de *cracking* (Fig. 1), base de la industria petroquímica [1]. Como su nombre indica (en inglés rotura, *cracking*) consiste en la reducción del tamaño de las moléculas del producto de partida, para obtener productos más ligeros. La nafta se introduce en el horno en presencia de vapor para

evitar la formación de coque, por lo que estas unidades se denominan *steam crackers* (crackers con vapor). Las condiciones de operación de los hornos de la unidad deben adaptarse a las características químicas de la nafta de alimentación para obtener un resultado óptimo. Estas características químicas de la nafta varían debido a cambios en el origen del producto, cambios en las etapas productivas previas (destilación) o en las condiciones de almacenaje, etc.

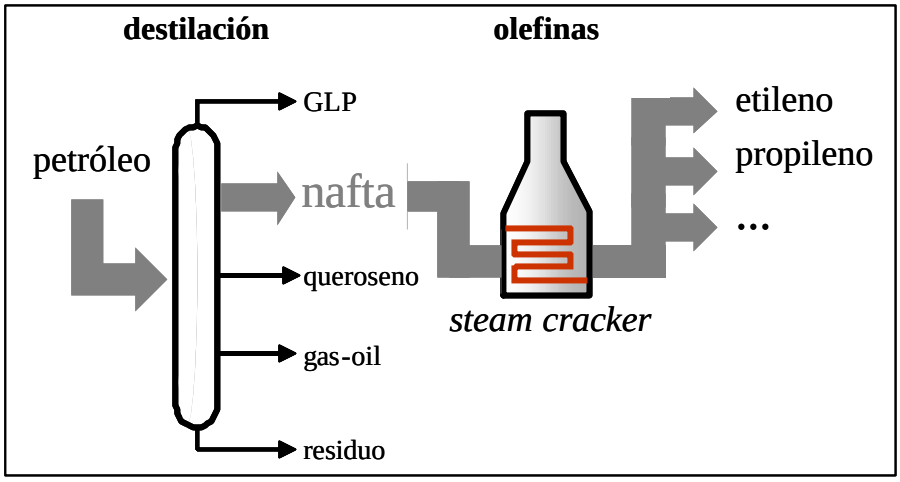


Fig. 1. Esquema general de la producción de olefinas a partir de la nafta del petróleo.

Para caracterizar la nafta de alimentación se debe determinar el contenido en cada una de las familias de los diferentes hidrocarburos presentes. Estos hidrocarburos son alcanos lineales (n-parafinas), alcanos ramificados (isoparafinas), alcanos cíclicos (naftenos), alquenos (olefinas) y aromáticos. Este análisis se conoce comúnmente como determinación de PIONA.

El número de constituyentes de la nafta es tan amplio que un sistema cromatográfico con una única columna no puede separarlos todos en un tiempo razonable. Para realizar este análisis, en la mayoría de laboratorios de refinerías se utiliza un sistema cromatográfico multicolumna comercial [2]. Este sistema utiliza diversas columnas acopladas mediante válvulas para separar secuencialmente todos

los componentes de la muestra. Un esquema de un sistema de este tipo se puede observar en la Fig. 2.

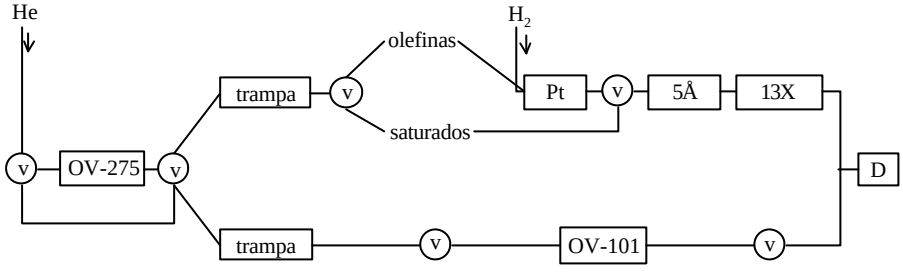


Fig. 2. Esquema básico del sistema multicolumna para análisis de PIONA (Adaptado de Beens y col. [3]).

Este sistema proporciona la composición (% en peso) de n-parafinas, isoparafinas, naftenos, olefinas y aromáticos, según el número de carbonos de cada componente. Este análisis está ampliamente extendido en la industria petroquímica para el análisis de composición de gasolina y nafta. Un importante inconveniente que presenta esta técnica es el elevado tiempo de análisis, ya que son necesarias entre 3 y 4 horas para llevarlo a cabo. La relativa lentitud del análisis no permite una optimización adecuada del *cracker*, por lo que la posibilidad de emplear una técnica más rápida, como la espectroscópica de infrarrojo produciría importantes beneficios en el control y optimización de la producción de olefinas.

1.2.2 Determinación de etileno, viscosidad e índice de fluidez en polipropileno

El polipropileno (PP) es uno de los polímeros termoplásticos comerciales más importantes del mercado, con una demanda en continua expansión. Se obtiene por polimerización de propileno, obtenido a su vez como subproducto en el proceso de obtención de etileno por craqueo con vapor, o bien del proceso de *cracking* catalítico para la obtención de gasolina. Es un producto de coste relativamente bajo, que puede ser fácilmente modificado mediante la copolimerización para adecuar sus propiedades físicas a diferentes aplicaciones.

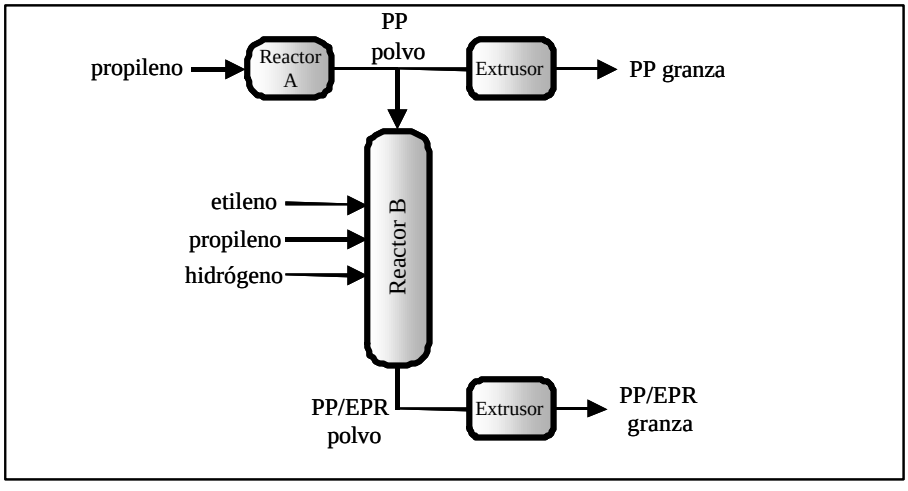


Fig. 3. Esquema de la producción de polipropileno homopolímero (PP) y copolímero de etileno-propileno (PP/EPR o iPP).

La copolimerización más corriente consiste en añadir un elastómero al polipropileno para modificar su cristalinidad, ya que se introducen irregularidades en la macromolécula. El producto resultante de la copolimerización tiene una mayor dureza a bajas temperaturas y una menor tendencia a la fractura. El elastómero más utilizado es a su vez un copolímero, de etileno-propileno (EPR, del inglés *ethylene-propylene rubber*). El polipropileno producido se denomina *impact resistant* (resistente al impacto), y contiene entre un 10-30% de EPR. Aunque este producto se denomina habitualmente copolímero heterofásico, debido a la convivencia de una fase cristalina y otra amorfa, no existen enlaces químicos entre el PP y el EPR añadido, por lo que en realidad se trata de una mezcla física de los dos polímeros y no de un verdadero copolímero. En la actualidad este copolímero de PP se produce directamente en un proceso multietapa, consistente en una copolimerización *in situ* de propileno y etileno en presencia del homopolímero de PP con un sistema catalítico de $TiCl_3$-$AlEt_3$ (Fig. 3).

El copolímero de impacto formado (*impact* PP o PP/EPR) posee una matriz de homopolímero de PP en la que se halla disperso el EPR en forma de pequeñas partículas discretas. Estas partículas protegen la matriz contra la propagación de

grietas disipando grandes cantidades de energía en el material que rodea la partícula [4]. La caracterización de la compleja estructura de este material se ha llevado a cabo por diferentes técnicas: resonancia magnética nuclear (RMN), fraccionamiento por elución a temperatura variable – cromatografía de exclusión por tamaños (TREF-SEC, del inglés *Temperature rising elution fraction - Size Exclusión Chromatography*) y la microscopia electrónica (TEM) [5, 6].

En la planta de producción de polipropileno se llevan a cabo análisis más sencillos para comprobar la calidad del producto, como por ejemplo la determinación del % de etileno añadido, del índice de fluidez o la viscosidad. En el caso del % de etileno, el método de determinación es la espectroscopia de infrarrojo medio. Para llevar a cabo este análisis, la muestra debe ser calentada a 150 ºC y prensada para obtener una lámina fina de material, adecuada para un análisis por transmisión de infrarrojo. El índice de fluidez (*melt flow index*, MFI, en inglés) se mide por la cantidad (peso en gramos) de polímero fundido y extruido durante 10 min. a través de un orificio de longitud y diámetro interno normalizado, en unas condiciones específicas de presión y temperatura [7] (Fig. 4). La viscosidad intrínseca (o *limiting viscosity number* según la nomenclatura IUPAC) se obtiene extrapolando la viscosidad reducida a concentración cero y requiere la disolución del polímero en tetrahidronaftaleno a 135ºC [8].

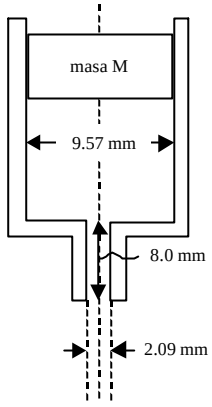


Fig. 4. Esquema del sistema para la medición del índice de fluidez. Adaptado de ref. 9

Como se puede ver, los diferentes análisis del polipropileno requieren pretratamientos de la muestra: fundido y prensado para obtener una lámina, fundido y extrusión o disolución a altas temperaturas. Estos pretratamientos representan una porción considerable del tiempo total de análisis, plantean problemas de reproducibilidad y son una fuente de errores sistemáticos si no se realizan correctamente, ya que pueden alterar fácilmente la muestra y afectar al resultado del análisis. Por estos motivos, representa una alternativa muy atractiva el uso de un único análisis espectroscópico (un solo espectro) por muestra, para determinar rápidamente las múltiples propiedades que se analizan rutinariamente en este tipo de polímeros.

1.3 ESTRUCTURA DE LA TESIS

El trabajo desarrollado en la presente tesis doctoral, ha sido consecuencia de la interacción entre el grupo de investigación en el cual he trabajado con dos empresas del polígono industrial de Tarragona; la empresa Repsol-YPF y la empresa Transformadora de propileno (TDP), lo que ha permitido un conocimiento de los aspectos de carácter práctico que entraña el desarrollo de esta metodología.

La secuencia del trabajo efectuado responde a la necesidad de compaginar las actividades de organismos tan diferenciados como son la Empresa y la Universidad. Por este motivo y con la finalidad de exponer de la forma más sencilla posible el trabajo efectuado, la presente tesis se ha estructurado de la forma que se detalla a continuación:

- En primer lugar se describen los principales objetivos de la tesis así como su justificación.
- En el capítulo 1 se presentan los contenidos de la tesis y el interés de las aplicaciones industriales sobre las que se ha trabajado. Por último se expone un esquema de la tesis.

- En el capítulo 2 se hace una introducción a los fundamentos teóricos, espectroscópicos y quimiométricos, utilizados.
- En el capítulo 3 se presenta la parte experimental de este trabajo.
- En los capítulos 4 a 8 se presentan los resultados obtenidos.
- En el capítulo 9 se presentan las conclusiones generales de la tesis.

El capítulo 2, en el que se presentan los fundamentos teóricos, y el capítulo 3, donde se presentan aspectos experimentales, tienen contenidos que también se incluyeron en algunas de las publicaciones presentadas, aunque se ha mantenido esta duplicidad por motivos de claridad y para tener una visión conjunta del trabajo.

Los capítulos 4 a 8 tienen una estructura similar: un apartado introductorio que da paso a la publicación realizada, con una ampliación de algunos aspectos experimentales o de los resultados que no están suficientemente tratados en la publicación, para terminar con las conclusiones de cada apartado.

El capítulo 4 contiene un apartado introductorio más extenso que el resto de capítulos de resultados, en el que se presentan los modelos de calibrado desarrollados para el polipropileno, ya que ciertos aspectos de este modelo se tratan en sucesivos capítulos (cap. 5 y 6). Los modelos de nafta utilizando la espectroscopia de infrarrojo medio, fueron objeto de una publicación que se presentada en este capítulo.

En los capítulos 5 y 6 se trata el tema del control y transferencia (cap. 5) y de la detección de muestras discrepantes (cap. 6) del modelo desarrollado para determinar el contenido en etileno en muestras de polipropileno. Estos aspectos fueron objeto de sendas publicaciones que se presentan en estos capítulos.

El capítulo 7 presenta consideraciones de carácter general y práctico de la utilización de la calibración multivariante y los métodos espectroscópicos en la industria petroquímica, que también fue objeto de una publicación.

En el capítulo 8 se presenta un aspecto importante de la calibración multivariante que no se había desarrollado antes en esta tesis, como es la selección de variables. La selección de las variables (longitudes de onda en este caso) que aportan más información de la propiedad de interés, si bien seguramente no comportará una mejora sensible de la exactitud de los modelos, puede aportar otras posibles mejoras como podría ser una mayor sencillez, robustez y facilidad de actualización y estandarización del modelo.

Por último, en el capítulo 9 se presentan las conclusiones generales de la tesis y las perspectivas de futuro de estas metodologías analíticas en el campo de la industria petroquímica.

BIBLIOGRAFÍA

1 R. E. Kirk and D. F. Othmer, *Kirk-Othmer concise encyclopedia of chemical* technology, Wiley, New York, **1985**.

2 P. van Arkel, J. Beens, J. Spaans, D. Grutterink and R. Verbeek, *J. Chromatogr. Sci. 26*, **1988**, 267.

3 J. Beens, U.A.Th. Brinkman *Trends Anal. Chem. 19*, **2000,** 260.

4 M. Francis, Jr. Mirabella, *J. Appl. Polym. Science: Appl. Polym. Symp. 51*, **1992**, 117.

5 T. Usami, Y. Gotoh, H. Umemoto, and S. Takayama, *J. Appl. Polym. Science: Appl. Polym. Symp. 52*, **1993**, 145.

6 Z. S. Petrovic, J. Budinski-Simendic, V. Divjakovic, and Z. Skrbic, *J. Appl. Polym. Science 59*, **1996**, 301.

7 The American Society for Testing and Materials (ASTM), Test Method D1238-01. ASTM Annual Book of Standards, vol. 08.01, West Conshohocken, PA, USA, **2001**.

8 The American Society for Testing and Materials (ASTM), Test Method D2857-95(2001). ASTM Annual Book of Standards, vol. 08.02, West Conshohocken, PA, USA, **2001**.

9 N.G. McCrum, C.P. Buckey and C. B. Bucknall, *Principles of Polymer Engineering*, 2nd Ed. Oxford University Press, Oxford, **1997**.

2. Fundamentos teóricos

2.1 ESPECTROSCOPIA DE INFRARROJO

2.1.1 Aspectos fundamentales

La espectroscopia molecular se basa en la interacción entre la radiación electromagnética y las moléculas. Dependiendo de la región del espectro en la que se trabaje y por tanto de la energía de la radiación utilizada (caracterizada por su longitud o número de onda), esta interacción será de diferente naturaleza: excitación de electrones, vibraciones moleculares y rotaciones moleculares [1]. La molécula, al absorber la radiación infrarroja, cambia su estado de energía vibracional y rotacional. Las transiciones entre dos estados rotacionales requiere muy poca energía, por lo que solo es posible observarlas específicamente en el caso de muestras gaseosas. En el caso del estudio del espectro infrarrojo (IR) de muestras sólidas y líquidas sólo se tienen en cuenta los cambios entre estados de energía vibracional [2].

Utilizando la mecánica quántica y el modelo del oscilador anarmónico para representar los enlaces, se demuestra que las bandas en el infrarrojo se producen como consecuencia de transiciones entre niveles de energía en los que el número cuántico vibracional (υ) cambia en una unidad ($\Delta\upsilon=\pm1$), denominada banda fundamental, o en más de una unidad ($\Delta\upsilon=\pm2, \pm3, \ldots$), que se denominan sobretonos [3]. Aunque teóricamente son posible $\Delta\upsilon$ superiores, en la práctica sólo se observan estas tres transiciones. Las bandas de absorción aparecen aproximadamente (existen otros términos despreciables) a frecuencias: $\boldsymbol{u}$ (la banda fundamental), $2\boldsymbol{u}$ y $3\boldsymbol{u}$ (los sobretonos) [2]. Estos últimos tienen una menor intensidad que la banda fundamental. También se producen bandas como consecuencia de la interacción de dos vibraciones diferentes:

$$\boldsymbol{u}_{comb} = n_1\boldsymbol{u}_1 \pm n_2\boldsymbol{u}_2 \pm \ldots$$

Una molécula poliatómica (*n* átomos) tiene 3*n*-6 modos de vibración diferentes (3*n*-5 si la molécula es lineal). Cada uno de estos modos de vibración viene representado por una curva de energía potencial diferente y da lugar a una banda fundamental y sus correspondientes sobretonos en el infrarrojo. Los modos de vibración que se pueden producir incluyen: cambios en la distancia de enlace (elongaciones o *stretching*, que pueden ser simétricas o asimétricas) y cambios en el ángulo de enlace, o *bending* (simétricas en el plano, asimétricas en el plano, simétricas fuera del plano y asimétricas fuera del plano) [4].

2.1.2 Regiones espectrales

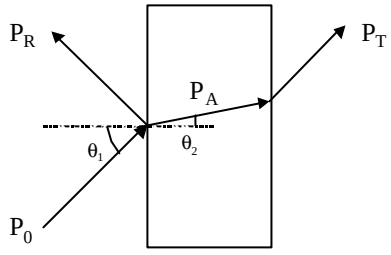
Aunque el espectro infrarrojo se extiende desde 10 a 14 300 cm^{-1}, desde un punto de vista funcional se divide en tres zonas: IR lejano, donde se producen las absorciones debidas a cambios rotacionales, el IR medio (MIR o simplemente, IR), donde tienen lugar las vibraciones fundamentales y el IR cercano (*near infrared*, NIR), donde se producen absorciones debidas a sobretonos y combinaciones de las bandas fundamentales.

Tabla 1. División del espectro IR, adaptado de ref. 2, pag 21.

Región	Transición característica	Longitud de onda (nm)	Número de onda (cm^{-1})
Infrarrojo cercano (NIR)	Sobretonos y combinaciones	700-2500	14300-4000
Infrarrojo medio (IR)	Vibraciones fundamentales	$2500 – 5 \times 10^4$	4000-200
Infrarrojo lejano	Rotaciones	$5x10^4 – 10^6$	200-10

2.1.3 Tipos de medidas en infrarrojo

Cuando la radiación incide en la muestra (Fig. 5), ésta puede sufrir diferentes fenómenos: absorción, transmisión y reflexión. La intensidad de la luz transmitida a través de la muestra (P_T) es menor que la intensidad incidente (P_0). Una parte de esta intensidad incidente se ha reflejado (P_R), mientras que otra parte ha sido absorbida por la sustancia (P_A).



$$P_0 = P_A + P_T + P_R$$

Fig. 5 Fenómenos de absorción, transmisión y reflexión de la radiación electromagnética al interaccionar con la materia.

La medida más común en el infrarrojo es la que se basa en la absorción (o la intensidad transmitida), aunque también se han desarrollado espectroscopias basadas en el fenómeno de la reflexión como son la reflectancia total atenuada y la reflectancia difusa. A continuación se describen estas técnicas, que son las que se han utilizado en el transcurso de la presente tesis.

2.1.3.1 Absorción/Transmisión

El espectro por transmisión a través de la muestra determina P_A, ya que esta pérdida en la intensidad luminosa incidente está relacionada con la concentración de la muestra. La transmitancia, *T*, se calcula como la fracción de radiación (P_T/P_0) transmitida a través de la muestra. La intensidad de absorción de la luz, absorbancia (*A*) se calcula como:

$$A = -\log T = \log \frac{P_0}{P_T}$$

La representación de la transmitancia o de la absorbancia como una función de la longitud de onda, λ, o del número de onda , $\bar{n}$, es lo que conforma el espectro de la muestra. La relación que existe entre la concentración y la absorbancia está descrita por la ley de Lambert-Beer:

$$A = \varepsilon c l$$

La absorción de la radiación por parte de la muestra es proporcional a la longitud del camino óptico (anchura de la celda, l) [en cm], a la concentración de la solución [mol/L] y a una constante proporcional específica de cada muestra, e, denominada absortividad molar, [L mol^{-1} cm^{-1}]. Esta ley lineal se cumple únicamente para disoluciones diluidas ($c \leq$ a 0.1 M), pudiéndose producir desviaciones de la linealidad a concentraciones más elevadas al variar e como consecuencia de cambios en el índice de refracción de la disolución. Otras desviaciones de la linealidad tienen su origen en la propia instrumentación: presencia de luz reflejada y dispersada, luz no estrictamente monocromática o detectores de respuesta no lineal [5].

2.1.3.2 Reflectancia total atenuada (attenuated total relflectance, ATR)

El principio de esta medida se basa en el fenómeno de la reflexión total interna y la transmisión de la luz a través de un cristal con un elevado índice de refracción (Fig. 6). La radiación penetra (unos μm) más allá de la superficie del cristal donde se produce la reflexión total, en forma de onda evanescente [6]. Si en el lado exterior del cristal se coloca un material absorbente (muestra), la luz que viaja a través del cristal se verá atenuada (de ahí el nombre de la técnica) y se puede registrar el espectro de la muestra.

El ángulo de la luz incidente y la geometría del cristal facilitan que se produzcan sucesivas reflexiones en sus caras internas. El espectro medido tiene una apariencia similar al espectro de transmisión, excepto por ciertas variaciones en la intensidad en función de la longitud de onda que se producen.

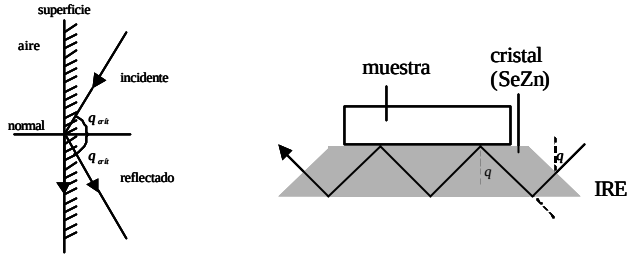


Fig. 6. Reflexión total interna y elemento de reflexión interna (IRE) utilizado en el sistema ATR. Adaptado de ref. 6 pag. 321 y 324.

La profundidad de penetración [6], d_p, depende de la longitud de onda de la radiación, λ, del índice de refracción del cristal, n_p, del índice de refracción de la muestra, n_s, y del ángulo de incidencia del haz de radiación del cristal, θ, de acuerdo con la ecuación:

$$d_p = \frac{\lambda}{2\pi n_p (sen^2\theta - n_{sp}^2)^{1/2}}$$

donde $n_{sp}=n_s/n_p$ ($n_s<n_p$). El camino óptico total en la muestra se obtiene multiplicando d_p por el número de reflexiones que se hayan producido a través de la muestra. Esta técnica de muestreo es muy efectiva para el análisis de sólidos y líquidos, especialmente en las regiones del infrarrojo medio y del infrarrojo cercano. Para obtener medidas adecuadas es necesario que exista un contacto íntimo entre la muestra y el cristal del ATR, por lo que esta técnica se utiliza sobretodo en líquidos o en sólidos que se puedan compactar contra el cristal aplicando presión. Esta técnica es especialmente útil, por ejemplo, en el caso de medir muestras viscosas utilizando caminos ópticos muy cortos: una celda de transmisión de estas características sería muy difícil de llenar y limpiar debido a la consistencia de la muestra.

2.1.3.3 Reflectancia difusa

Otra medida que se basa en el fenómeno de la reflexión es la reflectancia difusa. Cuando la luz incide sobre una muestra opaca y no absorbente tiene lugar el fenómeno de la reflexión especular regido por las ecuaciones de Fresnel (Fig. 7a).

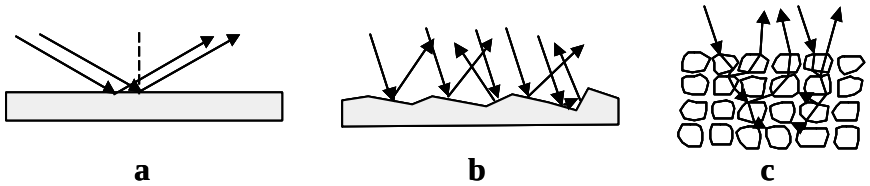


Fig. 7. Procesos de reflexión en un material especular (a) o irregular (b). Fenómeno de reflectancia difusa (c). Adaptado de ref 2, pag. 44-45.

La intensidad reflejada sobre el total incidente depende de los índices de refracción del aire y la muestra (n_1, n_2). Para el caso de un ángulo de incidencia igual a cero la expresión es [2]:

$$\frac{P_R}{P_O} = \frac{(n_2 - n_1)^2}{(n_2 + n_1)^2}$$

Cuando la luz incide sobre una superficie irregular, se puede considerar que la frontera entre el medio y la muestra está formado por una serie de pequeñas interfases orientadas en todos los ángulos posibles (Fig. 7b). De esta forma aunque cada una de estas pequeñas interfases refleja la luz siguiendo la ley de Fresnel, el efecto global es una reflexión de la luz a cualquier ángulo (reflectancia difusa).

La radiación que se transmite a través de la primera interfase (Fig. 7c) puede sufrir absorción por parte de la muestra, por lo que la intensidad de la luz se verá atenuada según la ley de Beer. Esta radiación que ha atravesado la primera capa de partículas se difunde a las siguientes capas a través de reflexiones aleatorias, refracciones y dispersión y puede sufrir nuevos fenómenos de atenuación. La longitud de camino óptico seguido por la luz es muy difícil de describir

matemáticamente, sobre todo si las partículas de la muestra tienen tamaños heterogéneos, por lo que no se ha desarrollado una teoría rigurosa de la reflectancia difusa. Sí que existen teorías basadas en la práctica, la más conocida la de Kubelka-Munk [7]. Esta teoría propone para una capa completamente opaca y de grosor infinito:

$$\frac{(1-R_\infty)^2}{2R_\infty} = \frac{k}{s} = \frac{\varepsilon c}{s}$$

donde R_∞ es la reflectancia de la capa infinitamente gruesa, y k y s son las constantes de *scattering* y absorción, respectivamente. El coeficiente de absorción (k), es igual a la concentración multiplicada por la absortividad definida en la ley de Beer (εc).

En la práctica la reflectancia difusa se mide respecto a un estándar no absorbente y a continuación se calcula el logaritmo para llegar a una relación lineal con la concentración [8]:

$$\log \frac{R_{\text{stand}}}{R} = \log \frac{1}{R} + \log R_{\text{stand}} \propto \frac{\varepsilon c}{s}$$

R y R_{stand} representan la reflectancia de la muestra y del estándar respectivamente (siendo mayor la reflectancia del estándar que la de la muestra). Para la luz monocromática $\log R_{\text{stand}}$ es constante y puede ser ignorado. Si se cumple la condición de aditividad del espectro, la expresión anterior puede ser reescrita como:

$$c = k + \frac{s}{\varepsilon} \log \frac{1}{R}$$

Con lo que, cómo en el caso de la ley de Beer, existe una relación entre la concentración y la radiación medida en el espectro. Aunque existe el inconveniente de que *s* no es constante y depende de varias propiedades de la muestra, como el tamaño de partícula y el contenido de humedad. Al aumentar el tamaño medio de partícula ($s \propto 1/d$) se produce una reducción del *scattering* y la radiación penetra más profundamente en la muestra, con lo que el log(1/R) aumenta. El efecto es un desplazamiento a lo largo del eje de ordenadas como función del tamaño de partícula. Además el *scattering* de las partículas pequeñas depende de la longitud de onda, lo que provoca que el desplazamiento debido al tamaño de partícula no sea constante en todo el espectro.

La presencia de agua en el espectro provoca la aparición de bandas características, y además afecta a la totalidad del espectro debido a la variación que se introduce en el índice de refracción del medio (n_o), provocando un aumento de log(1/R). Puesto que el grado de humedad puede variar entre muestras, en la práctica la constante de *scattering* (*s*) se convierte en una incógnita para cada una de las nuevas muestras, por lo que no se puede llevar a cabo el análisis cuantitativo.

2.2 INTERPRETACIÓN DE ESPECTROS

2.2.1 Asignación de bandas

En el espectro infrarrojo medio, entre 4000 y 1300 cm^{-1} (región de frecuencias de grupo) se observan una serie de bandas asignadas a vibraciones de sólo dos átomos de la molécula. En este caso la banda de absorción se asocia únicamente a un grupo funcional y a la estructura molecular completa, aunque hay influencias estructurales que provocan desplazamientos significativos en la frecuencia de la vibración. Esta vibraciones derivan de grupos que contienen hidrógeno (C-H, O-H, y N-H) o grupos con dobles y triples enlaces aislados. Entre 1300 y 400 cm^{-1} (*fingerprint region*) la asignación a grupos funcionales determinados es más difícil debido a la multiplicidad de bandas, pero es una zona de espectro muy útil para la

identificación de compuestos específicos [9]. La Fig. 8 muestra un cuadro resumen de las frecuencias de absorción de los grupos funcionales más comunes en el IR medio. En el espectro de infrarrojo cercano, predominan las bandas debidas a sobretonos y combinaciones de enlaces en los que participa el hidrógeno (debido a que aumenta el grado de anarmonicidad de la vibración), en la Fig. 9 se muestra un cuadro resumen de las absorciones más habituales en el infrarrojo cercano.

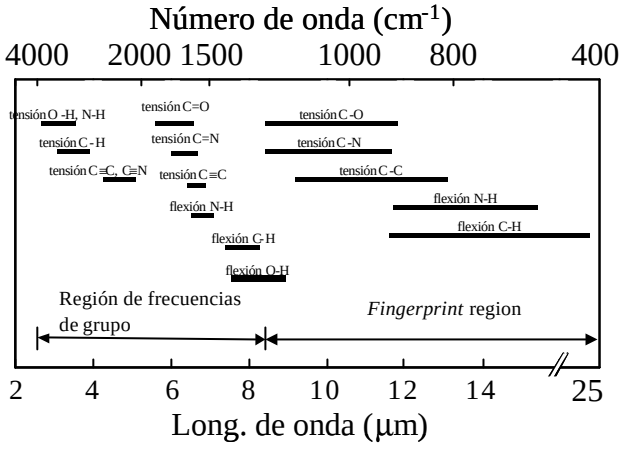


Fig. 8. Frecuencias de vibración en el infrarrojo medio. Adaptado de ref. 9, pag. 556.

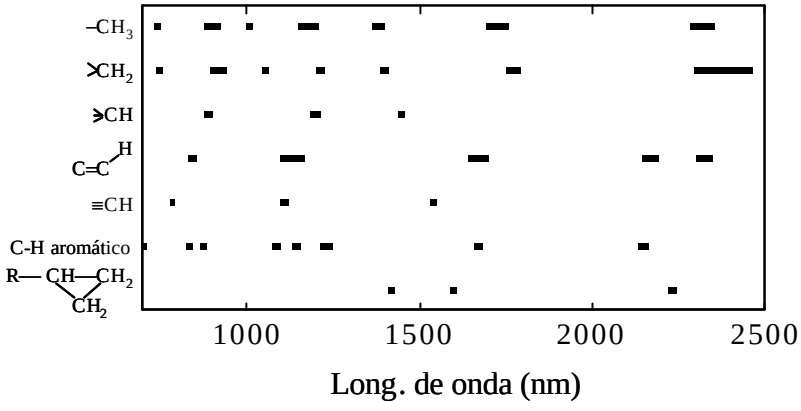


Fig. 9. Frecuencias de vibración en el infrarrojo cercano. Adaptado de ref. 2, pag. 28.

En el NIR las bandas de absorción no están tan bien definidas como en el infrarrojo medio, apareciendo en forma de bandas ancha y solapadas entre si, por lo

que es más difícil realizar una asignación a un componente o grupo funcional concreto de la muestra. Las bandas tienen una menor intensidad (la absortividad de la muestra es menor), por lo que se debe trabajar con caminos ópticos más largos, de 0.5 a 10 mm, frente a los 10-50 μm utilizados en el infrarrojo medio.

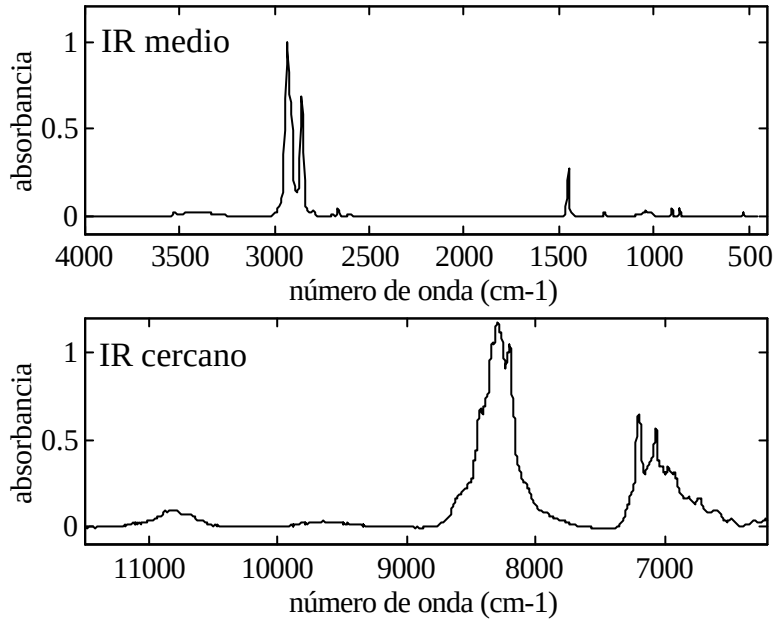


Fig. 10. Comparación entre el espectro IR y NIR del ciclohexano.

2.2.2 Segunda derivada de los espectros

Una aproximación alternativa al problema de la superposición de picos y a la corrección de la línea base es el uso de la segunda derivada del espectro. La primera derivada del espectro se puede calcular muy fácilmente restando las absorbancias a longitudes de onda adyacentes (en el caso de longitudes de onda equidistantes). La segunda derivada se obtiene aplicando de nuevo este proceso. La segunda derivada tiene ciertas características muy interesantes: tiene mínimos en la posición de las bandas de absorción del espectro original, facilitando en gran medida la resolución de los picos solapados. Además, también elimina problemas

en la línea de base. Sus principales desventajas son que disminuye la relación señal/ruido y aumenta la complejidad del espectro. Debido a que la diferenciación es una operación lineal, la ley de Lambert-Beer sigue siendo válida en el espectro derivado, por lo que éste puede ser utilizado para determinaciones cuantitativas.

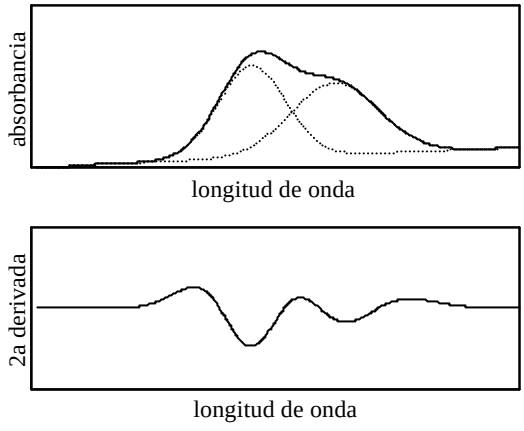


Fig. 11. Resolución de dos bandas solapadas a través de la segunda derivada**.**

2.3 FUNDAMENTOS QUIMIOMÉTRICOS

2.3.1 Introducción

El progresivo aumento de la complejidad de la instrumentación analítica ha permitido obtener volúmenes de datos cada vez mayores. La conversión de estos datos en información útil requiere del uso de herramientas matemáticas y estadísticas, que se han agrupado en la disciplina denominada Quimiometría [10, 11, 12, 13, 14]. Este apartado presenta brevemente la base teórica de las herramientas quimiométricas utilizadas en la presente tesis.

En primer lugar se introduce la descomposición en componentes principales (o PCA, del inglés *principal components analysis*) ya que es una técnica con múltiples aplicaciones quimiométricas, como la calibración (regresión sobre componentes principales, PCR), el preprocesado de la respuesta, el análisis de agrupaciones, etc. Después se hace un repaso de los pretratamientos de los espectros utilizados a lo largo de la tesis y por últimos se presentan los fundamentos así como otros aspectos de carácter más práctico de la calibración multivariante, como el control estadístico o la estandarización de los modelos desarrollados.

2.3.2 Descomposición en componentes principales (PCA)

Al utilizar métodos espectroscópicos se obtienen respuestas de cada muestra para cientos o miles de variables, en este caso longitudes de onda. La matriz, $\mathbf{R}_{IxJ}$ (I filas por J columnas) representa las respuestas de I muestras analizadas a J longitudes de onda diferentes. El elevado número de variables J que caracterizan cada muestra impiden un análisis y representación gráfica sencillos de las muestras.

En este caso, el método de descomposición en componente principales [15] es muy útil, porque permite representar la variabilidad presente en **R**, en unos pocos

factores (o componentes principales) que son combinaciones lineales de las variables originales.

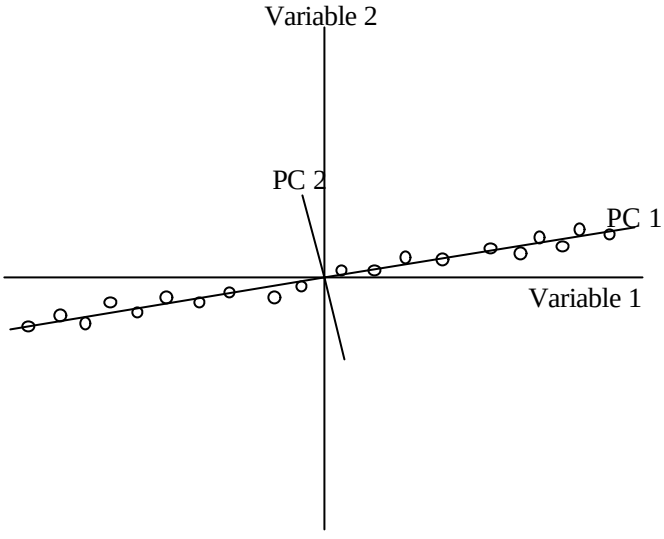


Fig. 12. Representación gráfica de la descomposición en componentes principales de un conjunto de muestras definidas por dos únicas variables. En este sencillo ejemplo la descomposición consiste en un simple cambio de ejes (componentes principales).

El análisis en componentes principales proporciona una aproximación a la matriz **R** como un producto de dos matrices: la matriz de *scores*, **T** y la matriz de *loadings*, **P**, que capturan la estructura de los datos de **R**. Los *scores* capturan la estructura de las filas o lo que es lo mismo, las relaciones entre objetos (muestras) y los *loadings* retienen la relación existente entre las variables.

$$\mathbf{R} = \mathbf{T}\mathbf{P}^{T} + \mathbf{E}$$

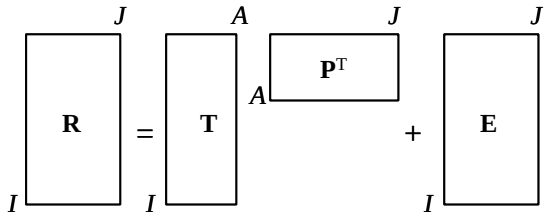


Fig. 13. Notación matricial de la descomposición en componentes principales.

El análisis de componentes principales (PCA) se aplica frecuentemente cuando se trabaja con datos colineales. Esta colinealidad en los datos significa que la información principal de las variables **R** se puede condensar en un conjunto más pequeño de *A* variables. Cada una de estas nuevas *A* variables se denomina componente principal o factor. El conjunto de factores se puede ver más claramente si se representa el producto $\mathbf{TP}^{\mathrm{T}}$ como la suma de *A* términos de la forma $\mathbf{t}_a\mathbf{p}_a^{\mathrm{T}}$, que corresponden a cada una de las *A* columnas de las matrices **T** y **P**.

$$\mathbf{R} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \ldots + \mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \mathbf{E}$$

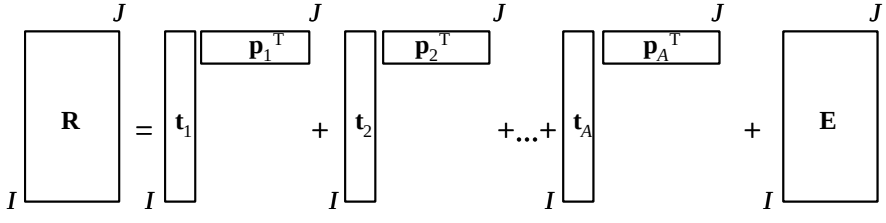


Fig. 14. Notación matricial extendida de la descomposición en componentes principales.

El primer componente principal es aquél que explica una mayor cantidad de la información contenida en **R** (maximiza $\mathbf{p}_1^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{p}_1 = \mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1$). El siguiente factor $\mathbf{p}_2$, explica la máxima información de **R** no contenida en $\mathbf{p}_1$ (maximiza

$\mathbf{p}_2^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{p}_2 = \mathbf{t}_2^{\mathrm{T}}\mathbf{t}_2$ y es ortogonal al primer componente principal, esto es $\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_2 = 0$). Los sucesivos factores explican cada vez menos información y son ortogonales a los anteriores. Las condiciones de ortogonalidad de *scores* y *loadings* se pueden resumir como:

$$\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$$

$$\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathrm{diag}(\boldsymbol{t}_a) \qquad \text{o bien} \qquad \boldsymbol{t}_a = \mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a \quad a = 1,2...A$$

Si la matriz $\mathbf{R}$ está centrada, entonces $\boldsymbol{t}_a$ representa los valores propios de la matriz $\mathbf{R}^{\mathrm{T}}\mathbf{R}$, y los vectores $\mathbf{p}_a$ representan sus vectores propios. Esto significa que los *loadings* satisfacen la ecuación:

$$\mathbf{R}^{\mathbf{T}}\mathbf{R}\mathbf{p}_a = \mathbf{p}_a\hat{o}_a$$

La magnitud de los valores propios indica la cantidad de variabilidad (información) que retiene cada uno de los componentes principales.

El análisis en componentes principales es muy útil para la interpretación de datos multivariantes. Por un lado, la representación de los *scores* permite establecer relaciones entre las muestras, permitiendo así la detección de muestras discrepantes y agrupaciones. Por otra parte, los *loadings* permiten comparar y estudiar la influencia de las distintas variables (longitudes de onda en este caso).

2.3.3 Técnicas de pretratamiento de datos

El pretratamiento de la señal es uno de los primeros pasos que se lleva a cabo en el análisis de datos multivariantes. Consiste en manipulaciones matemáticas que se aplican antes de cualquier otro tipo de análisis y tratan de anular o, al menos reducir, fuentes de variabilidad en la señal, ya sea de carácter aleatorio (como el

ruido) o de carácter sistemático (variaciones en la línea base, etc.), que no están relacionadas con el analito o la propiedad de interés.

2.3.3.1 Centrado

El centrado de una variable consiste en la sustracción del valor medio a todos sus elementos. En el caso de los datos espectrales el centrado consiste en restar al espectro de cada una de las muestras el espectro medio. Este pretratamiento pone de relieve las diferencias entre espectros, al haber eliminado la tendencia común (el espectro medio). Se utiliza tanto en PCA como en PCR o PLS.

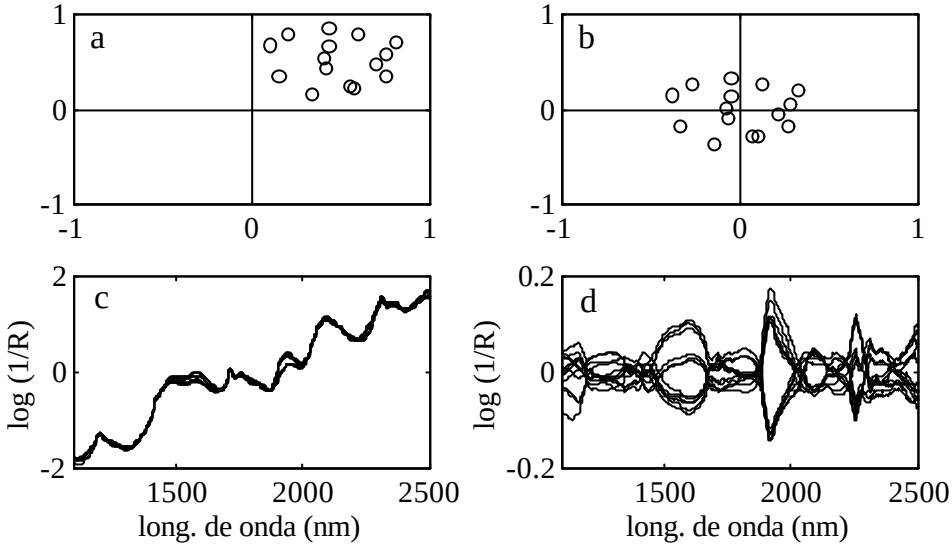


Fig. 15. Centrado por columnas. a) Objetos definidos por dos variables. b) Los mismos objetos centrados. c) Espectro de 12 muestras de alcohol polivinílico (PVA). d) Los mismos espectros de PVA centrados.

2.3.3.2 Suavizado

La aplicación de técnicas de suavizado tienen el objeto de reducir matemáticamente el ruido aleatorio que acompaña a la señal analítica. Aunque existen otros métodos más sencillos como el de la media móvil, en esta tesis se ha

utilizado una herramienta basada en un ajuste polinómico móvil, como es el filtro de Savitzki-Golay [16]. Este suavizado consiste en interpolar un polinomio de grado *n* cada serie de *m* puntos de la señal, de forma que el valor de la respuesta en cada punto se reemplaza por una combinación lineal de los puntos vecinos. Este método, al requerir el uso de 2*m*+1 valores para el cálculo de cada valor de la señal corregida, provoca el truncamiento del espectro en sus extremos, perdiendo 2*m*+1 valores en cada aplicación. Este problema se resuelve en la generalización presentada por P. A. Gorry [17], que es el método utilizado en el transcurso de esta tesis.

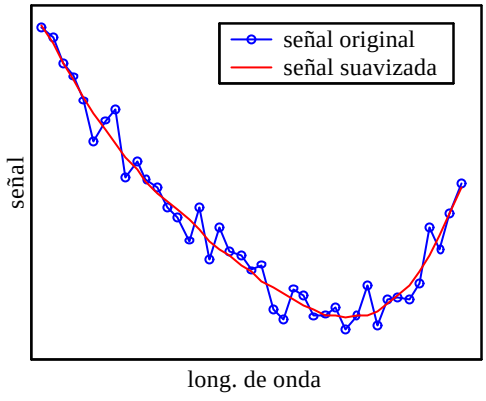


Fig. 16. Suavizado de una señal espectroscópica utilizando una ventana de 21 puntos y un polinomio de 3er grado.

2.3.3.3 Corrección de la línea base

Aparte de fuentes de variabilidad de alta frecuencia, como sería el ruido, la señal medida puede contener fuentes de variabilidad de baja frecuencia, no relacionadas con la propiedad de interés, y que se denominan variaciones de la línea base. A continuación se presentan las diferentes herramientas que se han utilizado en la presente tesis para corregir la variación de la línea base.

Modelado explícito de la línea base

Este método consiste en aproximar la línea de base del espectro con una función polinómica, que se resta posteriormente al espectro para obtener el espectro corregido. El caso más sencillo sería el polinomio de grado 0 (una constante), también denominada *offset correction* y consiste en restar a cada espectro la absorbancia medida a una determinada longitud de onda. La selección de esta longitud de onda se puede hacer en base al conocimiento químico de la muestra (debe ser una zona sin absorción de ningún componente que varíe en la muestra), tomando la longitud de onda de menor variabilidad entre muestras [18] o la longitud de onda que al realizar un modelo PLS tiene un valor en los *loadings* próximo a cero [19].

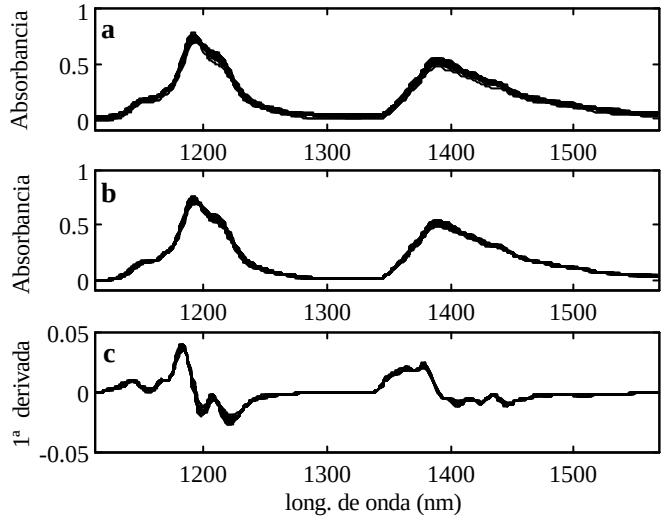


Fig. 17. Espectros NIR de muestras de nafta sin pretratamiento (a), offset correction a 1100 nm (b) y primera derivada (c).

Se pueden utilizar polinomios de grado superior, para adaptar la corrección a las características de la variación de la línea base. Los más habituales son el uso de una línea recta (polinomio de grado 1) o de una curva (polinomio de grado 2). Cuando se utilizan polinomios de grado 1 o 2, esta corrección se denomina *detrending*. En

la Fig. 18 se muestra un ejemplo esquemático de una corrección utilizando un polinomio de 1er grado.

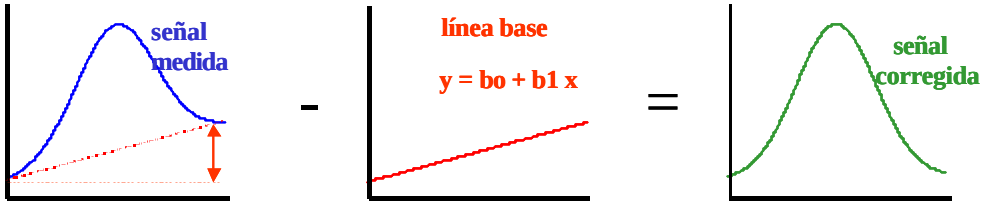


Fig. 18. Esquema del proceso de corrección de la línea base utilizando un modelo explícito lineal (*detrendig*).

Derivadas

El uso de derivadas [20] consigue diferenciar mejor picos solapados y elimina desplazamientos lineales y cuadráticos de la línea base. La primera derivada se obtiene como la diferencia entre las absorbancias, *A*, a dos longitudes de onda consecutivas (λ_1 y $\lambda_{1+\delta}$):

$$dA = \frac{A_{l1} - A_{l1+d}}{\Delta l}$$

La derivada reduce la correlación entre variables y el efecto la dispersión debido al tamaño de las partículas. Presenta el inconveniente de magnificar el ruido en la señal, por lo que habitualmente se utiliza en combinación con métodos de suavizado de la señal, como el del ajuste polinómico móvil, presentado en el apartado 2.3.3.2. Derivadas de orden superior se obtienen aplicando sucesivamente la derivada, aunque no suelen utilizarse superiores a orden 2 (segunda derivada), ya que decrece la relación señal-ruido (S/N).

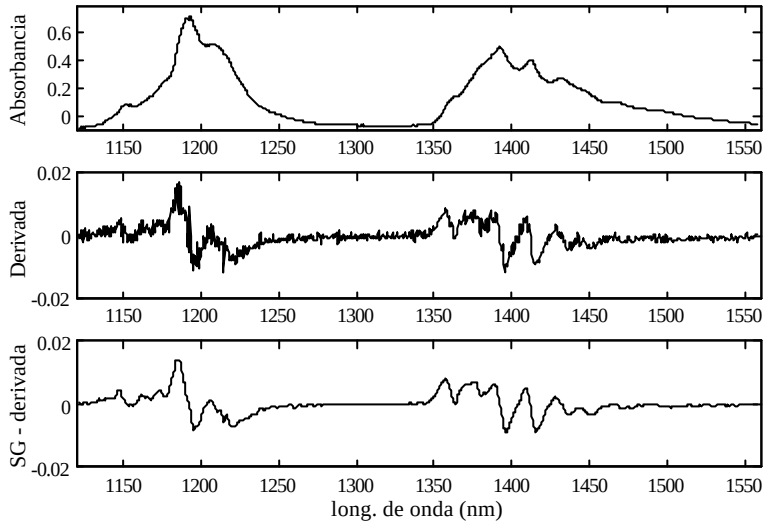


Fig. 19. Ejemplo de la derivada de un espectro de nafta (arriba), sin utilizar un suavizado (en medio) y utilizando un suavizado de Savitzki-Golay (abajo).

2.3.3.4 Corrección del "ligth scatteing"

En la familia de preprocesados que corrigen los efectos multiplicativos (*ligth scatteing*), que se producen en la espectroscopia por reflectancia como consecuencia de la variación en el tamaño de partícula, se encuentran dos transformaciones, la *standard normal variate* (SNV) [21] y la *multiplicative scatter correction* (MSC) [22, 12, 23, 24]. La principal diferencia entre ambos métodos es que el SNV se aplica a cada espectro independientemente, mientras el MSC requiere que se defina un espectro de referencia. La transformación SNV está definida como:

$$r_{i,j}^{SNV} = \frac{r_{i,j} - \overline{r}_i}{\sqrt{\sum_j (r_{i,j} - \overline{r}_i)^2}} \sqrt{J-1}$$

donde $r_{i,j}$ representa la respuesta (absorbancia) de la muestra i a la longitud de onda j, $\bar{r}_i$, representa la media de las respuestas de la muestra i, J el número total de longitudes de onda y $r_{i,j}^{SNV}$ representa el espectro corregido.

La transformación MSC se basa en el hecho empírico que, para un conjunto de muestras similares, el valor de la absorbancia está linealmente relacionado con la media de los valores espectrales (con la media tomada sobre todas las muestras). Primero se realiza una regresión de la respuesta de cada muestra a la longitud de onda j, $r_{i,j}$, sobre la respuesta media de todas las muestra a esa longitud de onda, $\bar{r}_j$:

$$r_{i,j} = c_i + b_i \bar{r}_j + e_{i,j} \qquad \text{con} \quad \bar{r}_j = \sum_i \frac{r_{i,j}}{I}$$

Y la señal transformada final, $r_{i,j}^{MSC}$, se obtiene como:

$$r_{i,j}^{MSC} = \frac{r_{i,j} - \hat{c}_i}{\hat{b}_i}$$

Helland *et al.* [25] realizan una comparación de estos dos métodos y concluyen que están íntimamente relacionados, y que los resultados de aplicar una u otra corrección son muy similares en términos de habilidad de predicción.

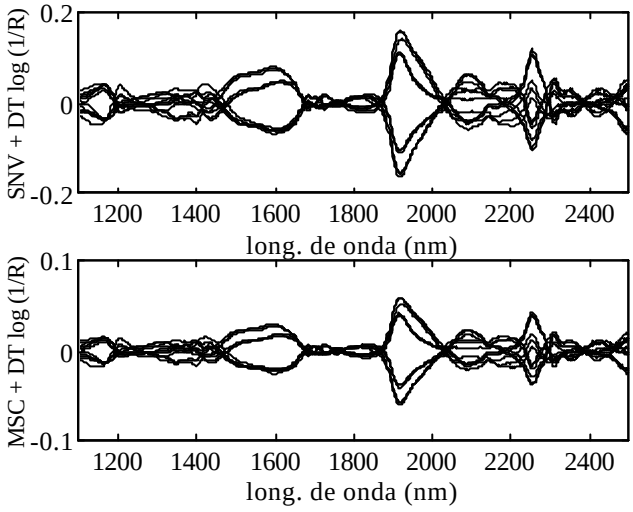


Fig. 20. Comparación de las transformaciones SNV (arriba) y MSC (abajo), en las muestras de alcohol polivinílico.

2.3.4 Etapas de la calibración multivariante

Entre los métodos quimiométricos que mayor éxito han tenido en las aplicaciones industriales se encuentra los dirigidos a cuantificar (calibración multivariante). Las etapas para llevar a cabo una regresión multivariante a partir de datos espectroscópicos propuestas por la ASTM en sus prácticas estándar [26] son:

1. Selección del conjunto de muestras de calibración.
2. Establecimiento del modelo.
3. Validación del modelo de calibración.
4. Predicción de muestras desconocidas.
5. Monitorización y transferencia.

2.3.4.1 Selección del conjunto de calibración

El conjunto de calibración, también denominado conjunto de entrenamiento debe contemplar todas las fuentes de variabilidad del sistema [27], tanto físicas como químicas. Para conseguir este objetivo de una forma rigurosa, se ha

propuesto el uso de estrategias de diseño de experimentos [27], aunque la complejidad de las muestras reales raramente permite aplicar esta aproximación.

La situación más común es contar con un amplio conjunto de muestras candidatas a las que se ha medido el espectro y la propiedad de interés. En este caso se han propuesto técnicas de selección de muestras basadas en el análisis de agrupaciones (*cluster analysis*) [28], o algoritmos de selección de muestras como el de Kenard-Stone [29], para decidir qué espectros NIR representan mejor la población total.

La presencia de agrupaciones destacadas en el conjunto de calibración puede conducir a errores, por lo que es importante analizar los datos y si se detecta una agrupación severa se debe evaluar si es más adecuado construir un modelo global para todas las muestras, o bien es necesario dividir en grupos las muestras y crear modelos diferentes para cada grupo. La detección de agrupaciones se puede realizar mediante una simple inspección visual de los *scores* PCA. También se han propuesto criterios gráficos para detectar agrupaciones, como las curvas de distancias (*distance curves*) [30] o criterios numéricos, como el estadístico de Hopkins, *H*, [31, 32], que se aplica a los *scores* de la descomposición PCA y se basa en la comparación de la distancia euclidiana entre un objeto y su vecino más próximo (*W*) y la distancia entre un objeto artificial, distribuido aleatoriamente en el espacio, y el objeto real más próximo (*U*). La elección del objeto real y artificial se realiza varias veces (iteraciones) para un porcentaje de la población total de objetos.

$$H = \frac{\sum_{e=1}^{N} U_e}{\sum_{e=1}^{N} U_e + \sum_{e=1}^{N} W_e}$$

El valor de H oscila entre 0.5 para un conjunto de objetos distribuidos homogéneamente (las distancias U y W son muy parecidas) hasta 1 para un conjunto con agrupaciones muy marcadas (U>>W).

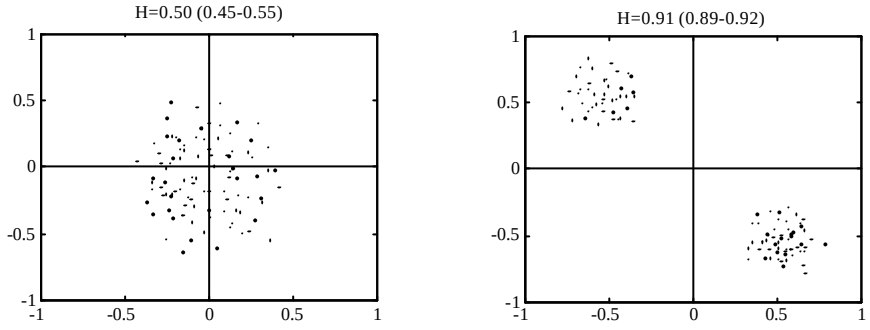


Fig. 21. Estadístico de Hopkins en el caso de un conjunto de muestras homogéneo (H=50) y en el caso de un conjunto de muestras agrupadas H próximo a 1.

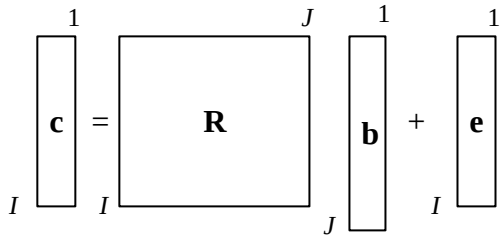
2.3.4.2 Establecimiento del modelo

En los modelos multivariantes inversos, la concentración (o cualquier otra propiedad de la muestra) se modela en función de la respuesta (en este caso el espectro IR) de la forma:

$$c_k = \sum_{j=1}^{J} r_j b_{j,k} + e_k$$

donde c_k es la concentración del analito k en la muestra, r_j es la respuesta de la muestra en la variable j, $b_{j,k}$ es el coeficiente que relaciona la variable j con la concentración del analito k y e_k es el término del error no modelado por el modelo. Utilizando una notación matricial, en la etapa de calibración con I muestras, este modelo se puede escribir como:

$$\mathbf{c} = \mathbf{Rb} + \mathbf{e}$$



donde **c** es el vector de las concentraciones del analito para las *I* muestras de calibración, **R** es la matriz de las respuestas de las *I* muestras en las *J* variables, **b** el vector de los coeficientes de regresión y **e** el vector de los errores o residuales.

El vector de coeficientes de regresión se obtiene en la etapa de calibración como:

$$\hat{\mathbf{b}} = \mathbf{R}^{+}\mathbf{c}$$

donde $\mathbf{R}^{+}$ es la matriz pseudoinversa de **R**. La forma en la que se calcula esta matriz pseudoinversa difiere según el método de calibración utilizado. Por ejemplo, en el caso de la calibración multivariante inversa (ILS), $\mathbf{R}^{+}$ se obtiene por la solución de mínimos cuadrados clásicos, $\mathbf{R}^{+} = (\mathbf{R}^{T}\mathbf{R})^{-1}\mathbf{R}^{T}$.

Los métodos de descomposición sobre factores o componentes principales, como la regresión por mínimos cuadrados parciales (*Partial Least Squares*, PLS), solucionan el problema de la colienealidad en los datos (muy habitual en datos espectroscópicos) ya que descomponen la matriz de respuestas, **R**, en una serie de factores ortogonales entre sí, lo que evita los problemas de la inversión de la matriz $\mathbf{R}^{T}\mathbf{R}$.

A continuación se presenta muy brevemente el algoritmo de obtención del modelo PLS, la forma en que se selecciona el número óptimo de factores así como las herramientas para detectar observaciones anómalas (detección de *outliers*).

Regresión por mínimos cuadrados parciales (PLS)

La regresión por mínimos cuadrados parciales realiza una descomposición en factores, como el análisis de componentes principales (PCA), pero utiliza la información de la concentración en la descomposición de la matriz de respuestas, **R**. El primer factor describe la dirección de máxima varianza en **R**, que al mismo tiempo se correlaciona con la concentración.

El procedimiento para realizar la descomposición de la matriz de las respuestas, consta de los pasos siguientes:

1) Calcular los, *loadings* y *scores*,:

$$\mathbf{p}_{h+1} = \frac{\mathbf{R}_h^{\mathrm{T}}\mathbf{c}}{\left\|\mathbf{R}_h^{\mathrm{T}}\mathbf{c}\right\|}$$

$$\mathbf{t}_{h+1} = \frac{\mathbf{R}_h\mathbf{p}_{h+1}}{\left\|\mathbf{R}_h\mathbf{p}_{h+1}\right\|}$$

donde $\| \ \|$ representa la norma del vector.

2) Calcular la matriz residual $\mathbf{R}_{h+1}$ según:

$$\mathbf{R}_{h+1} = (\mathbf{I}_I - \mathbf{t}_{h+1}\mathbf{t}_{h+1}^{\mathrm{T}})\mathbf{R}_h(\mathbf{I}_J - \mathbf{p}_{h+1}\mathbf{p}_{h+1}^{\mathrm{T}})$$

donde **I** es la matriz identidad de dimensión *I*.

Estos dos pasos iniciales se calculan desde $h = 0$ hasta el número de componentes principales que se quiere calcular. Para el caso $h = 0$, $\mathbf{R}_0$ es la matriz de datos original.

Los vectores **t** y **p** calculados para cada valor de *h* forman, respectivamente, las columnas de las matrices ortonormales **T** y **P**, que se utilizan a continuación para invertir la matriz **R**.

3) Calcula una nueva matriz, **Q**, de la forma siguiente:

$$\mathbf{Q} = \mathbf{T}^T \mathbf{R} \mathbf{P}$$

a la que se le aplica la descomposición en valores singulares (SVD) [33], de manera que:

$$\mathbf{Q} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

Como **T** y **P** son ortonormales se cumple también que:

$$\mathbf{R} = \mathbf{T} \mathbf{Q} \mathbf{P}^T$$

Por tanto, la matriz original **R** se puede reconstruir como:

$$\mathbf{R} = \mathbf{T}\mathbf{U}^T \mathbf{S} \mathbf{V}\mathbf{P}^T$$

y la pseudoinversa se obtiene a partir de la ecuación siguiente:

$$\mathbf{R}^+ = \mathbf{P}\mathbf{V} \mathbf{S}^{-1} (\mathbf{T}\mathbf{U})^T$$

Factores a incluir en el modelo

Se debe decidir el número de factores o componentes principales que será utilizado en el modelo final, o lo que es lo mismo, determinar el tamaño óptimo del modelo. Esta elección se basa en el cálculo de un error de predicción medio para modelos que incluyen cada vez más factores (1, 2 ... *A*) y en el estudio de la evolución de este error de predicción medio. Si se dispone de un conjunto independiente de muestras, no utilizado en la calibración, se puede calcular la raíz cuadrada del error medio de predicción (*Root-Mean-Square Error of Prediction*, RMSEP), para el conjunto de I_P muestras que no han participado en la calibración:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I} (\hat{c}_i - c_i)^2}{I_P}}$$

donde c_i y $\hat{c}_i$ representan la concentración determinada de forma independiente y la concentración predicha por el modelo respectivamente. Si no se dispone de un conjunto independiente de muestras se puede utilizar el método de la validación cruzada o *cross-validation* [34, 35], en la que sucesivamente se va dejando una parte de las muestras fuera del conjunto de calibración, se realiza el modelo con las muestras restantes y se predicen las muestras descartadas.

Este proceso permite obtener predicciones independientes sin renunciar al uso de toda la información (muestras) disponible en el conjunto de calibración. En este caso se obtiene un error de predicción similar al RMSEP, que se denomina raíz cuadrada del error medio de validación cruzada (o *Root-Mean-Square Error of Cross-validation*, RMSECV) :

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{I}(\hat{c}_i - c_i)^2}{I}}$$

Existen en la literatura diferentes criterios para la selección del número óptimo de factores en los modelos de calibración. Un primer criterio muy simple consiste en representar el RMSEP (o el RMSECV) frente al número de factores del modelo y tomar como óptimo el primer mínimo local o el mínimo absoluto. Este es un criterio que puede conducir fácilmente a sobreajustes y subajustes, por lo que no es recomendable usarlo. Existen otros criterios de selección más formales [36] basados en comparaciones F [37, 38] que proporcionan una base más fiable para la selección del número de factores a incluir en el modelo.

Detección de outliers

Una de las ventajas de los métodos multivariantes sobre los tradicionales univariantes, es la capacidad que tienen de detectar la observación u observaciones inconsistentes con el resto de los datos [39]. En la etapa de establecimiento del modelo se puede utilizar información de la influencia de los objetos en el conjunto de calibración (*leverage*) y de los residuales, tanto en la propiedad de interés como en la respuesta instrumental [12]. La detección de los *outlier* en esta etapa es importante porque la inclusión de estas muestras discrepantes en el modelo degrada su capacidad predictiva.

Leverage. Es una medida de la posición (o influencia) de una muestra en relación al modelo. Muestras con un elevado valor de *leverage* están muy alejadas del centro del modelo, por lo que tendrán una influencia muy alta sobre el mismo. Este valor se calcula como:

$$h_i = \frac{1}{I} + \mathbf{t}_i^{\mathrm{T}} (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1} \mathbf{t}_i$$

donde $\mathbf{t}_i$ representa el vector de *scores* de la muestra *i*, **T**, la matriz de *scores* del modelo y *I* el número de muestras de calibración. Se proponen diferentes niveles umbral, los más aceptados son dos o tres veces el *leverage* medio de calibración, que es igual a 1+*A*/*I* [12], siendo *A* el número de componentes principales o factores utilizados en el modelo.

Residuales en la respuesta instrumental. Los residuales en la respuesta (o residuales espectrales) reflejan la falta de ajuste entre las respuestas experimentales utilizadas en la calibración, **R**, y las respuestas reconstruidas por el modelo con *A* factores ($\hat{\mathbf{R}} = \mathbf{TP}^{\mathrm{T}}$).

$$\mathbf{E} = \mathbf{R} - \mathbf{TP}^{\mathrm{T}}$$

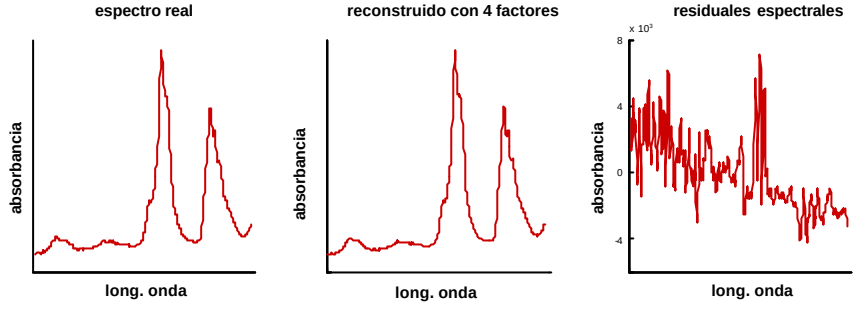


Fig. 22. Ejemplo del cálculo del residual de un espectro NIR. Al espectro original se le resta el espectro reconstruido con 4 factores para obtener el residual espectral.

Los residuales en la respuesta se pueden utilizar de diferentes formas. La más habitual es, para el error en la respuesta de la muestra *i*, $\mathbf{e}_i$, realizar una suma de cuadrados extendida a las *J* longitudes de onda y dividir por los grados de libertad (df) adecuados, para obtener una desviación estándar de la muestra *i*, $s(\mathbf{e}_i)^2$.

$$s(\mathbf{e}_i)^2 = \frac{\sum_{j=1}^{J} \hat{e}_{ij}^2}{\mathrm{df}\ / I} \qquad \text{con df} = I\,J - J - A\ (\max(I,J))$$

También se utilizan los residuales en la respuesta para realizar distintos test *F*, que comparan la suma de cuadrados de los residuales para el conjunto de calibración y para la muestra *i* [37,14]

Residuales en la concentración. En la etapa de establecimiento del modelo se dispone del valor de la concentración (o la propiedad de interés) determinado por el método de referencia. Los residuales en la concentración comparan el valor predicho por el modelo multivariante $\hat{\mathbf{c}}$ con el valor considerado verdadero, **c**, que proporciona el método de referencia.

$$\mathbf{f} = \mathbf{c} - \hat{\mathbf{c}}$$

Muchas veces la detección de *outliers* se realiza combinando estas herramientas, como en el gráfico que se representa el residual (espectral o de concentraciones) frente al *leverage* de las muestras [40, pag. 114].

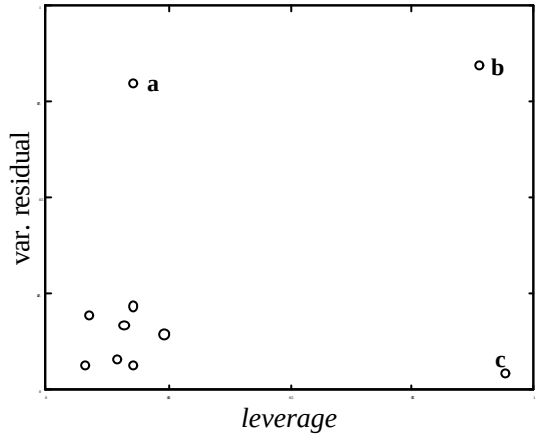


Fig. 23. Gráfico del residual frente al *leverage*. (**a**) Objetos con una varianza residual elevada se consideran *outliers*, (**b**) si además tienen un *leverage* alto son *outliers* peligrosos para el modelo, debido a que tienen mucha influiencia sobre él. Las muestras con un *leverage* alto (**c**) son muestras influyentes y no necesariamente *outliers*.

2.3.4.3 Validación del modelo

Los métodos de calibración sesgados, como PCR o PLS, no se apoyan directamente en un modelo teórico y pueden incorporar variabilidad de los datos no necesariamente relacionada con la propiedad de interés, por lo que deben ser cuantitativa o cualitativamente validados.

La validación consiste en el análisis de un grupo de muestras independiente al utilizado en la calibración y comprueba que no existe un error sistemático (*bias*) entre las predicciones que realiza el modelo y los valores proporcionados por el método de referencia. También se mide el grado de concordancia entre las predicciones del modelo y los valores del método de referencia.

Test de la ausencia de sesgo

La presencia de error sistemático se comprueba mediante una regresión lineal entre las concentraciones (u otra propiedad) reales, medidas por la técnica de referencia y las concentraciones predichas por el modelo multivariante. Si no existe error sistemático el resultado de esta regresión debería ser una recta de pendiente (b_1) igual a 1 y ordenada en el origen (b_o) igual a 0. Debido a la existencia de los errores aleatorios se obtendrán valores de b_1 y b_0 ligeramente diferentes a los esperados. El test conjunto de la pendiente y la ordenada en el origen [41] comprueba si hay diferencia significativa entre los valores de b_1 y b_0 encontrados y los valores esperados (1 y 0). Como tanto la concentración determinada por el método de referencia como la determinada por el modelo de calibración está sujeta a un cierto error se puede modificar el test conjunto para que tenga en cuenta la presencia de errores en ambos ejes [42].

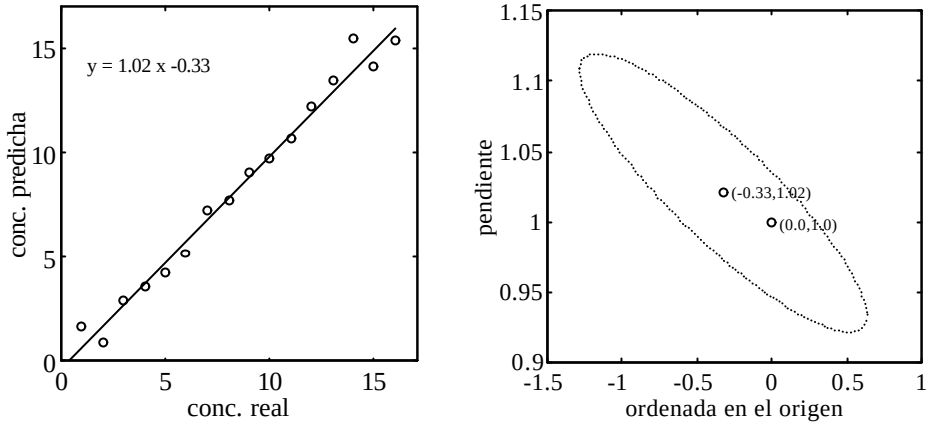


Fig. 24. Test conjunto de la pendiente y ordenada en el origen.

Concordancia con el método de referencia

En la práctica, esta comparación se realiza a través de un error medio de predicción, como el RMSEP. En caso de no disponer de un conjunto independiente de muestras se puede utilizar el error obtenido por validación cruzada, RMSECV. El valor de RMSEP o el RMSECV proporciona una estimación de la exactitud

(*accuracy*) del modelo. Tanto el RMSEP como el RMSECV proporcionan una estimación del error que se comete al predecir muestras futuras [43].

Error de predicción individual o específico

Aunque en la mayoría de aplicaciones se utiliza un único valor de error (RMSEP o RMSECV) para representar la habilidad predictiva global del modelo, en la práctica, es interesante poder proporcionar una estimación del error específico de cada determinación, tal como se hace en la calibración únivariante con los intervalos de predicción. Se han propuesto diferentes aproximaciones para el cálculo de este error específico, como la utilizada en el programa Unscrambler [40, pag. 436, 44], que fue corregida por De Vries y col. [45] o la propuesta por Faber [46]:

$$\hat{s}_{c_u} = \left(\frac{1}{I} + h_u\right)\left(s_e^2 + s_{\Delta\mathbf{c}}^2 + \|\mathbf{b}_A\|^2 s_{\Delta\mathbf{R}}^2\right) + s_{e_u}^2 + \|\mathbf{b}_A\|^2 s_{\Delta\mathbf{R}_u}^2$$

donde I es el número de muestras de calibración, h_u el *leverage* de la muestra desconocida. La varianza del análisis de referencia ($s_{\Delta\mathbf{c}}^2$), la varianza de la respuesta en calibración ($s_{\Delta\mathbf{R}}^2$) y para la nueva muestra ($s_{\Delta\mathbf{R}_u}^2$) se pueden calcular a partir de repeticiones; $\|\mathbf{b}_A\|$ es la norma euclidiana del vector de regresión para el modelo PCR o PLS de *A* factores; s_e^2 y $s_{e_u}^2$ representan la varianza de la parte no modelada de c (los residuales del modelo) para la calibración y la predicción de nuevos objetos respectivamente, y que pueden ser determinados a partir del error al cuadrado medio de calibración (MSEC):

$$\text{MSEC} = s_e^2 + s_{\Delta\mathbf{c}}^2 + \|\mathbf{b}_A\|^2 s_{\Delta\mathbf{R}}^2$$

con

$$\mathrm{MSEC} = \frac{\sum_{i=1}^{I} (c_i - \hat{c}_i)^2}{I - A - 1}$$

Faber y col. [47] utilizan esta expresión en una aplicación práctica para la determinación de compuestos oxigenados en gasolina. Algunos autores han criticado esta expresión, planteando problemas como la falta de términos específicos de la muestra (sólo el leverage, h_u, lo es) [44], o que no tiene en cuenta el sesgo [48]. Estas críticas han sido rebatidas por Faber [49].

2.3.4.4 Predicción de muestras desconocidas

Una vez el modelo ha sido aceptado, ya puede ser utilizado para el análisis de nuevas muestras. En esta etapa se deben seguir utilizando los test para detectar muestras discrepantes, *outliers*, con el fin de detectar la presencia de extrapolaciones al modelo, presencia de nuevos interferentes, fallos instrumentales, etc. En este caso se pueden utilizar medidas del *leverage* de las muestras, y del residual espectral. Herramientas para el control estadístico multivariante, como el estadístico T^2 de Hotelling y el estadístico Q, que serán introducidas a continuación para su uso en el control estadístico del modelo multivariante, se pueden utilizar también para la detección de *outliers*, ya que proporcionan una información similar al *leverage* (el T^2) y al residual espectral (el estadístico Q). Los residuales de la concentración (u otra propiedad de interés) no están disponibles ya que estas muestras no han sido analizadas por el método de referencia. La detección de los *outliers* en esta etapa es muy importante porque la predicción de estas muestras puede diferir significativamente del valor verdadero.

2.3.4.5 Monitorización y transferencia

Control de la estabilidad con el tiempo

Para que este método proporcione resultados válidos no es suficiente con

asegurar la exactitud del modelo en el momento de su desarrollo, sino que es necesario controlar su estabilidad con el tiempo. Para llevar a cabo este control, una posible estrategia es realizar periódicamente comparaciones entre las predicciones del modelo y los valores de referencia determinados por la técnica de referencia.

Se pueden utilizar otras estrategias, que no requieren utilizar el método de referencia, si se dispone de una muestra estable. En ese caso se puede analizar periódicamente una muestra de control [50] y utilizar herramientas de control estadístico de procesos (*statistical process control*, SPC) [51] en el caso de medidas univariantes o de control estadístico de procesos multivariantes (*multivariate statistical process control*, MSPC) [52] en el caso de múltiples medidas.

La base de cualquier gráfico de control es la representación del valor medido (directamente o después de algún tipo de transformación) respecto al tiempo. La decisión de si el sistema está bajo control estadístico o no, se realiza con la ayuda de unos límites de control, representados también en el gráfico.

SPC. Gráficos de control de Shewhart

Un gráfico de control muy sencillo para respuestas univariantes es el gráfico X de control de Shewhart [53]. En este gráfico se representan las observaciones (x_i) en el orden que se obtienen, el valor verdadero (μ) o el $\bar{x}$ en una línea central, y los límites de control y aviso se sitúan en $\pm 3\,\sigma$ y $\pm 2\,\sigma$ respectivamente, donde σ representa la desviación estándar de la población . Este gráfico puede ser utilizado, por ejemplo, para monitorizar la predicción de una muestra de control por el modelo multivariante.

MSPC. Estadísticos T^2 y Q

Los gráficos de control univariantes no aprovechan la potencialidad de la instrumentación analítica ya que no sirven para monitorizar al mismo tiempo múltiples respuestas. Los gráficos de control multivariante proporcionan información adicional y son capaces de detectar situaciones de fuera de control donde los gráficos univariantes no pueden hacerlo. Esta metodología dispone las respuestas multivariante de los sucesivos análisis de la muestra en forma de matriz (**X**), cada fila representa una muestra y las columnas son las variables. Esta matriz se descompone mediante un análisis de componentes principales (PCA). A partir de las matrices de los *scores* y los *loadings* se calculan diferentes estadísticos para cada muestra.

Estadístico T^2 de Hotelling. Fue propuesto originariamente por Hotelling [54] y mide la variación de cada muestra dentro del modelo PCA. Se calcula como la suma de los cuadrados de los *scores* según la ecuación siguiente:

$$T_i^2 = \mathbf{x}_i \mathbf{P}_k \, l^{-1} \mathbf{P}_k^{\mathrm{T}} \mathbf{x}_i^{\mathrm{T}}$$

donde $\mathbf{x}_i$ es el vector de respuestas para la muestra i, $\mathbf{P}_k$ es la matriz de *loadings* y λ es una matriz diagonal que contiene los valores propios asociados a los vectores propios incluidos en el modelo PCA. El límite de control del gráfico T^2 se calcula según:

$$T_{k,m,a}^2 = \frac{k(m-1)}{m-k} F_{k,m-k,a}$$

donde m es el número de análisis de la muestra de control que se han utilizado para construir el modelo PCA, k el número de componentes principales o factores incluidos en el modelo PCA y $F_{k,m-k,a}$ el valor de la distribución F para k y m-k gados de libertad y una probabilidad a de cometer un error de tipo I.

El gráfico T^2 monitoriza la distancia de una nueva medida al valor de referencia en el espacio reducido de los factores PCA. Permite detectar si la variación incluida en los componentes principales considerados es más grande que la que le correspondería si solo influyeran variaciones aleatorias. La interpretación de este gráfico es la misma que cualquier gráfico univariante; las muestras fuera de control poseen un valor de T^2superior al límite, y aparecen más allá de la línea de control.

Estadístico Q Mide la falta de ajuste de la nueva muestra al modelo PCA desarrollado. Se calcula mediante la ecuación siguiente [55]:

$$Q_i = \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} = \mathbf{x}_i (\mathbf{I} - \mathbf{P}_k \mathbf{P}_k^{\mathrm{T}}) \mathbf{x}_i^{\mathrm{T}}$$

donde $\mathbf{e}_i$ es la fila i de la matriz de los residuales (**E**), **I** es la matriz identidad, $\mathbf{P}_k$ la matriz de los k vectores de *loadings* incluidos en el modelo PCA y $\mathbf{x}_i$ es el vector fila de la respuesta de la muestra i.

El límite de control del gráfico Q [56] se calcula según:

$$Q_a = \Theta_1 \left[\frac{c_a \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}}$$

donde

$$\Theta_i = \sum_{j=k+1}^{n} l_j^i \quad \text{para} \quad i = 1, 2, 3$$

y

$$h_0 = 1 - \frac{2\Theta_1 \Theta_3}{3\Theta_2^2}$$

k es el número de componentes principales retenidos en el modelo, λ_j el valor propio asociado al componente principal j, n el número total de componentes principales y c_a la desviación estándar normal correspondiente a un error de tipo I (a) fijado. El valor de Q indica el grado de ajuste de cada muestra al modelo PCA y da una medida de la variación de la muestra no incluida en el modelo.

Algoritmo de identificación de sensores por eliminación en sentido inverso (BESI). Una desventaja de la monitorización multivariante frente a la univariante es la dificultad que existe para determinar, en caso de una muestra fuera de control, que respuestas han sido las responsables de este comportamiento. Una posible solución es el estudio de gráficos de contribución [57] o del vector de residuales. El método BESI, publicado por Stork y col. [58], es un algoritmo que detecta las variables con respuesta anómala que provocan el fuera de control en la variación no explicada por el modelo. Este método calcula una relación entre valores de Q eliminando las variables sospechosas y el valor Q del límite de control. Las variables sospechosas son eliminadas secuencialmente hasta que la muestra vuelve a estar bajo control estadístico.

Mantenimiento de los modelos multivariantes

El desarrollo de modelos de calibración multivariantes supone un esfuerzo experimental muy grande, ya que se requiere un elevado número de muestras representativas que deben ser analizadas por el método instrumental y por el de referencia. Por este motivo el periodo de validez y la sensibilidad a los cambios en las condiciones experimentales de un modelo son parámetros muy importantes. El modelo puede producir predicciones erróneas por diferentes razones: cambios en la respuesta instrumental (variaciones de la línea base, cambios en el camino óptico), cambios en las condiciones experimentales (temperatura, presión), o cambios en la matriz química de la muestra, producidos por ejemplo por alteraciones en el proceso de producción, cambios en los materiales de origen, etc. Se puede

solucionar la perdida de validez del modelo utilizando técnicas de transferencia de modelos, lo que evita un proceso de recalibrado completo.

Estas técnicas transforman el espectro medido, los coeficientes del modelo de predicción o las propias predicciones, para conseguir que el modelo siga siendo válido. Estas técnicas son útiles en caso de un cambio instrumental discreto o para transferir un modelo entre dos instrumentos distintos. Requieren contar con un subgrupo de muestras, en número sensiblemente menor al número de muestras de calibración, medidas en las dos condiciones experimentales (por ejemplo: antes y después del cambio instrumental, o en cada uno de los dos instrumentos, según sea el caso). Son ejemplos de este tipo de técnicas la corrección del sesgo y la pendiente (*bias/slope correction*) [59] o la *piece wise standardisation* [60]. La primera técnica transforma las predicciones y la segunda transforma los espectros.

Cuando no es posible contar con muestras medidas en ambas condiciones experimentales o en el caso de variaciones instrumentales o ambientales de carácter no discreto (ej. un cambio aleatorio de la temperatura), se pueden utilizar métodos alternativos, como los modelos de calibración robustos, menos sensibles a las variaciones instrumentales, ambientales o a la presencia de nuevos interferentes. Estos modelos incluyen implícitamente las variaciones (modelos globales), hacen un pretratamiento de los espectros o usan técnicas de calibración más robustas como las redes neuronales [61], el PLS con selección de variables (IVS-PLS) [62] o la regresión local ponderada (LWR) [63].

Corrección de la pendiente y el sesgo (slope/bias correction, SBC). En esta técnica las respuestas del subconjunto de muestras de estandarización se miden en las dos condiciones experimentales, $\mathbf{R}_1$ y $\mathbf{R}_2$ (T muestras por J variables). Con el modelo inicial, $\mathbf{b}$, se obtienen las concentraciones predichas, $\hat{\mathbf{c}}_{\mathbf{R}_1}$ y $\hat{\mathbf{c}}_{\mathbf{R}_2}$:

$$\hat{\mathbf{c}}_{\mathbf{R}_1} = \mathbf{R}_1\mathbf{b}$$

$$\hat{\mathbf{c}}_{\mathbf{R}_2} = \mathbf{R}_2\mathbf{b}$$

A continuación se realiza una regresión lineal entre las concentraciones predichas en las primeras condiciones frente a la predicción en las segundas condiciones:

$$\hat{\mathbf{c}}_{\mathbf{R}_1} = b_0 + b_1\,\hat{\mathbf{c}}_{\mathbf{R}_2}$$

para proporcionar así los parámetros de corrección del sesgo (b_0) y de la pendiente (b_1). El método de regresión más adecuado para hacer esta regresión es el de los mínimos cuadrados ortogonales [59], que considera errores de similar magnitud en ambos ejes. Las predicción de una muestra medida en las segundas condiciones experimentales, $\hat{c}_{2,un}$ se corrige de acuerdo a:

$$(\hat{c}_{2,un})_{std} = b_0 + b_1\hat{c}_{2,un}$$

para obtener la concentración estandarizada, $(\hat{c}_{2,un})_{std}$. Esta técnica es muy simple y rápida, aunque solo es aplicable si las diferencias en las respuestas instrumentales son pequeñas.

Estandarización directa por partes (*Piecewise Direct Standardization*). El método PDS [60] relaciona de forma multivariante las respuestas de las muestras del subconjunto de estandarización. La técnica considera que existen correlaciones entre los valores de la respuesta por grupos de variables. Así, las respuestas en la variable *i* de las muestras en las primeras condiciones, $\mathbf{r}_{1i}$, se relacionan con una

ventana de respuestas, centrada en la variable *i*, para las mismas muestras pero en las segundas condiciones.

$$\mathbf{r}_{1i} = \mathbf{X}_i^{\mathrm{T}}\mathbf{f}_i + f_{0i}$$

donde $\mathbf{f}_i$ son los coeficientes de corrección y f_{0i} el término independiente o corrección aditiva de fondo (*additive background correction*, ABC) [64] y se obtienen por una regresión multivariante PCR o PLS. Para una regresión multivariante (*J* variables), se obtendrán *J* vectores $\mathbf{f}_i$ y *J* valores f_{oi}, que se pueden agrupar en la matriz **F**:

$$\mathbf{F} = diag\left(\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_J\right)$$
$$\mathbf{f}_0^{\mathrm{T}} = \left(f_{01}, f_{02}, \ldots, f_{0i}, \ldots, f_{0J}\right)$$

Las respuestas en las nuevas condiciones, $\mathbf{r}_{2,un}^{\mathrm{T}}$ se estandarizan según:

$$\left(\mathbf{r}_{2,un}^{\mathrm{T}}\right)_{std} = \mathbf{r}_{2,un}^{\mathrm{T}}\mathbf{F} + \mathbf{f}_0^{\mathrm{T}}$$

y el vector de respuestas estandarizado, $\left(\mathbf{r}_{2,un}^{\mathrm{T}}\right)_{std}$, se utiliza en el modelo para obtener las predicciones finales.

Esta técnica consigue corregir cambios más complejos que otras más simples, como la corrección de la pendiente y el sesgo, aunque tiene la desventaja del mayor número de parámetros que se deben establecer: el tamaño de la ventana, el número de factores que intervienen en los modelos locales (ventanas) o el número de muestras de estandarización a utilizar. Si estos parámetros se escogen incorrectamente se pueden producir errores en la transferencia, o artefactos [65],

producidos por una selección de muestras poco representativas o por una incorrecta selección del número óptimo de factores en los modelos locales.

Tanto en la técnica de transferencia PDS, como en la corrección de la pendiente y el sesgo, hay que seleccionar un subconjunto de muestras para realizar la transferencia. Un método de asegurar la representatividad de las muestras seleccionadas es utilizar un algoritmo de selección de muestras como el de Kenard-Stone [66], que selecciona secuencialmente las muestras de forma que se extiendan uniformente en el espacio experimental.

BIBLIOGRAFÍA

1 G. Schwedt, *The essential guide to analytical chemistry*, John Wiley & Sons, Chichester, **1997**.

2 B. G. Osborne, T. Fearn and P. H. Hindle , *Practical NIR spectroscopy with applications in food and beverage analysis*, Longman Scientific & Technical, 2nd ed. Harlow, England, **1993**.

3 J. M. Hollas, *Modern Spectroscopy*, John Wiley & Sons, 2nd ed. Chichester, England, **1992**.

4 R. M. Silverstein and F. X. Webster, *Spectrometric Identification of Organic Compounds*, 6a ed. John Wiley & Sons, New York, **1998**.

5 C. E. Miller, *NIR news 4*, **1996**, 3.

6 K. A. Rubinson and J. F. Rubinson *Análisis instrumental*, Pretice Hall, Madrid, **2001**.

7 V. P. Kubelka and F. Munk, *Z. Tech. Physik. 12*, **1931**, 593.

8 K. H. Norris, *Multivariate analysis of raw materials, in Chemistry and World Food Supplies: the New Frontiers, Chemrawn II* L. W. Shemilt ed. Pergamon Press, Oxford, **1983**.

9 R. Keller, J. M. Mermet, M. Otto and H. M. Widmer (ed). *Analytical Chemistry*, John Wiley & Sons, New York, **1998**.

10 R. Kramer, *Chemometric Techniques for Quantitative Analysis*, Marcel Dekker, New York, **1998**.

11 D. L. Massart, B. G. M. Vandegiste, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke, *Haandbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam, **1997**.

12 H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, Chichester, **1989**.

13 K. Esbensen, S. Schönkopf and T. Midtgaard, *Multivariate Analysis in Practice*, Camo, Trondheim, **1996**.

14 K. R. Beebe, R. J. Pell and M. B. Seasholtz, *Chemometrics. A Practical Guide*, John Wiley & Sons, New York, **1998**.

15 K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, **1980**.

16 A. Savitzky and M. J. E. Golay, *Anal. Chem. 36*, **1964**, 1267.

17 P. A. Gorry, *Anal. Chem. 62*, **1990**, 571.

18 J. J. Kelly, C. H. Barlow, T. M. Jinguji and J. B. Callis, *Anal. Chem. , 61*, **1989**, 313.

19 A. F. Parisi, L. Nogueiras and H. Prieto, A*nal. Chem. Acta. 238*, **1990**, 95.

20 J. Steiner, Y. Termonia and J. Deltour, *Anal. Chem. 44*, **1972**, 1906.

21 R. J. Barnes, M. S. Dhanoa and S. J. Lister, *Appl. Spectrosc. 43*, **1989**, 772.

22 P. Geladi, D. MacDougall and H. Martens, *Appl. Spectrosc. 39*, **1985**, 491.

23 T. Næs, *NIR news 5*, **1994**, 4.

24 T. Næs, *NIR news 5*, **1994**, 13.

25 I. S. Helland, T. Naes and T. Isakson, *Chemom. Intel. Lab. Syst. 29*, **1995**, 233.

26 The American Society for Testing and Materials (ASTM), Practice E1655-00. *ASTM Annual Book of Standards*, vol. 03.06, West Conshohocken, PA, USA, **2001**.

27 T. Næs and T. Isaksson, *Appl. Spectrosc. 43*, **1989**, 328.

28 T. Næs, *J. Chemom. 1*, **1987**, 121.

29 R. W. Kennard and L. A. Stone, *Technometrics 11*, **1969**, 137.

30 K. Szczubialka, J. Verdú-Andrés and D. L. Massart, *Chemom. Intell. Lab. Syst. 41*, **1998**, 145.

31 B. Hopkins, *Ann. Bot. 18*, **1954**, 213.

32 R. G. Lawson and P. Jurs, *J. Chem. Inf. Comp. Sci. 30*, **1990**, 36.

33 A. Lorber, L. E. Wangen and B.R. Kowalski, *J. Chemom. 1*, **1987**, 19.

34 M. J. Stone, *Roy. Staist. Soc. B 36*, **1973**, 111.

35 S. Wold, *Technometrics*, 20, **1978**, 397.

36 K. Esbensen, *et. al. Multivariate Analysis - in practice*, Trondheim, **1994** pp. 111-112.

37 D. M. Haaland and E. V. Thomas, *Anal. Chem, 60*, **1988**, 1193.

38 D. W. Osten, *Journal of Chemometrics, 2*, **1988**, 39.

39 W. J. Egan and S. L. Morgan, *Anal. Chem. 70*, **1998**, 2372.

40 *Unscrambler User Manual*, CAMO A/S, Trondheim, **1998**.

41 N. Draper and H. Smith, *Applied Regression Analysis*, 2nd ed. Wiley, New York, **1981**.

42 J. Riu, F. X. Rius, *Anal. Chem. 68*, **1996**, 1851.

43 E. V. Thomas, *Anal. Chem. 66*, **1994**, 795.

44 M. Høy, K. Steen and H. Martens, *Chemom. Intell. Lab. Syst. 44*, **1998**, 123.

45 S. De Vries and C. J. F. Ter Braak, *Chemom. Intell. Lab. Syst. 30*, **1995**, 239.

46 N. M. Faber, B. R. Kowalski, *Chemom. Intell. Lab. Syst. 34*, **1996**, 283.

47 N. M. Faber, D. L. Duewer, S. J. Choquette, T. L. Gree and S. N. Chesler. *Anal. Chem. 70*, **1998**, 2972.

48 T. Morsing and C. Ekman, *J. Chemom. 12*, **1998**, 295.

49 N. Faber, *Chemom. Intell. Lab. Syst. 52*, **2000**, 123.

50 K. Doerffel, G. Herfurth, V. Liebich, and E. Wendlandt, *Fresenius J. Anal. Chem. 341*, **1991**, 519.

51 E. L Grandt and R. S. Leavenworth, *Statistical Process Control*, McGraw-Hill, Madrid, **1993**.

52 T. Kourti and J.F. MacGregor, *Chemom. Intell. Lab. System 28*, **1995**, 3.

53 W. A. Shewhart, *Economic control of Quality*, Van Nostrand, **1931**.

54 H. Hotelling, *Multivariate Quality Control*, in C. Eisenhart, M. W. Hastay and W. A. Wallis (Eds.); *Techniques of Statistical Analysis*, McGraw-Hill, New York, **1947**.

55 J. E. Jackson, *A User's guide To Principal Components*, John Wiley and Sons, New York, **1991**.

56 J.E. Jackson, G.S. Mudholkar, *Technometrics, 21*, **1979**, 341.

57 P. Miller, R.E. Swanson and C. E. Heckler, *Appl. Math. and Comp. Sci. 8*, **1998**, 775.

58 C. L. Stork, D. J. Veltkamp and B.R. Kowalski, *Anal. Chem. 69*, **1997**, 5031.

59 E. Bouveresse, C. Hartmann, D.L. Massart, I.R. Last and K. A. Prebble, *Anal. Chem. 68*, **1996**, 982.

60 Y. Wang, D. J. Veltkamp and B.R. Kowalski, *Anal. Chem. 63*, **1991**, 2750.

61 P. J. Gemperline, *Chemom. Lab. Int. Syst. 39*, **1997**, 29.

62 H. Swierenga, F. Wülfert, O. E. de Noord, A. P. de Weijer, A. K. Smilde, and L. M. C. Buydens, *Anal. Chim. Acta 411*, **2000**, 121.

63 S. Y. Chang, E. H. Baughman and B. C. McIntosh. *Appl. Spectrosc. 55*, **2001**, 1199.

64 Z. Wang, T. Dean and B. R. Kowalski, *Anal. Chem. 67*, **1995**, 2379.

65 O. E. de Noord, *International Chemometrics Research Meeting*, Veldhoven, The Netherlands, 3-7 July, **1994**.

66 R. W. Kennard, L. A. Stone, *Technometrics 11*, **1969**, 137.

3. PARTE EXPERIMENTAL

3.1 PRESENTACIÓN DE LOS CONJUNTOS DE MUESTRAS ESTUDIADOS

3.1.1 Muestras

Nafta

Se estudió un primer conjunto de treinta y seis muestras de nafta procedentes de la alimentación de la unidad de producción de alquenos (olefinas) que la refinería de la empresa Repsol Petróleo tiene en Tarragona. Estas muestras fueron recogidas entre julio y octubre de 1998, y son representativas de la nafta de alimentación de esta unidad durante estos cuatro meses de producción. Un segundo conjunto de 132 muestras de nafta procedentes de la alimentación de la misma unidad fue recogido entre julio de 1998 y marzo de 2000. La mayor parte de las muestras de este segundo conjunto fueron recogidas entre agosto de 1999 y febrero de 2000, aunque también se incluyeron muestras anteriores para aumentar la representatividad de las muestras.

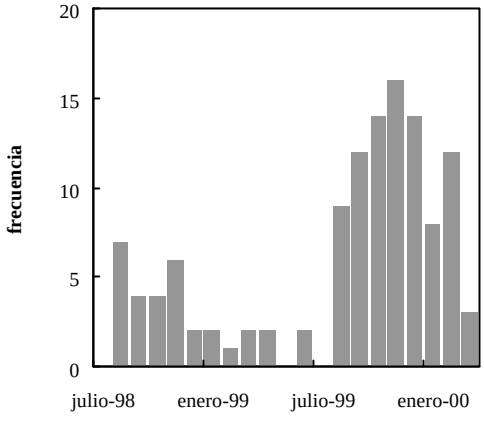


Fig. 25. Frecuencia de muestreo (por meses) del conjunto de 132 naftas analizadas por infrarrojo cercano.

Polipropileno

Se analizaron diferentes tipos de polímeros de propileno procedentes de la producción, en un periodo de 9 meses, de la planta que la empresa Transformadora de Propileno (TDP) tiene en Tarragona. Estos polímeros fueron:

a) homopolímero de polipropileno (PP) (69 muestras).
b) copolímero de etileno-propileno al azar, también denominado goma de etileno-propileno o *ethylene-propylene rubber*, EPR, (70 muestras).
c) copolímero de PP/EPR, también denominado polipropileno de impacto (iPP), formado por la polimerización *in situ* de EPR en presencia de polipropileno (177 muestras).

3.1.2 Análisis de referencia

Naftas

Las muestras de nafta se caracterizaron con un sistema cromatográfico multicolumna en el laboratorio de análisis de Repsol Petróleo. Este sistema cromatográfico proporciona una información muy detallada de la composición, ya que determina el porcentaje en peso de cada una de la familia de hidrocarburos: alcanos lineales, ramificados y cíclicos, alquenos y aromáticos, en función del número de carbonos de los compuestos. En la tabla siguiente se muestra un ejemplo de los resultados que proporciona el análisis cromatográfico.

C-num	N	I	P	O	A	Total
3	0,00	0,00	0,00	0,00	0,00	0,00
4	0,00	0,05	0,23	0,00	0,00	0,28
5	0,45	4,23	4,51	0,00	0,00	9,19
6	4,13	6,97	5,29	0,01	0,46	16,86
7	7,99	7,46	5,88	0,01	0,90	22,24
8	7,51	8,07	4,74	0,02	1,97	22,31
9	6,03	6,46	3,45	0,05	2,04	18,03
10	2,59	4,40	1,45	0,01	0,32	8,76
11	0,14	0,47	0,68	0,04	0,00	1,39
Total	28,83	38,11	26,23	0,14	5,69	99,07

Tabla 2. Ejemplo del resultado de análisis por GC de las muestras de nafta. El análisis proporciona el % (p/p) para cada una de las familias de hidrocarburos: naftenos (N), isoparafinas (I), parafinas (P), olefinas (O) y aromáticos (A), y en función del número de carbonos (de 3 a 11) de los compuestos.

Polipropileno

Los valores de referencia de cada una de las muestras de polipropileno (PP, EPR y EPR/PP) fueron analizadas en el laboratorio de control de calidad de la empresa TDP, y se determinaron las propiedades: % etileno, índice de fluidez y viscosidad. El % de etileno en las muestras fue determinado a partir del espectro IR de las muestras, por cociente de áreas. El índice de fluidez se determinó mediante un ensayo físico según la norma ASTM D1238 y la viscosidad intrínseca se determinó utilizando un viscosímetro de Ubbelhode según la norma ASTM D2857.

3.1.3 Medidas espectroscópicas

Infrarrojo medio

Se analizó por espectroscopia de infrarrojo medio el conjunto de treinta y seis muestras de nafta. El espectro de las muestras fue medido entre 4000 y 650 cm^{-1}, en un espectrómetro por transformada de Fourier, FT-IR Midac Prospect, equipado con un sistema de muestreo por reflectancia total atenuada horizontal (HATR), que utiliza un elemento de reflexión de seleniuro de zinc (Pike technologies, Madison, USA).

Infrarrojo cercano

Se analizó por infrarrojo cercano el conjunto de 136 muestras de nafta. Los espectros fueron medidos por reflectancia difusa (log 1/R) utilizando una celda para líquidos (o celda de transflectancia) con control de temperatura, en un espectrofotómetro dispersivo Bran-Luebbe IA500. Los espectros se realizaron a 20 ºC de temperatura y se registró el espectro entre 1100 y 2500 nm, en intervalos regulares de 2 nm.

Los diferentes tipos de polímeros de propileno (PP, EPR y EPR/PP) fueron analizados por reflectancia difusa (log 1/R) utilizando una celda rotatoria de gran capacidad, en un espectrofotómetro dispersivo Bran-Luebbe IA500. Se registró el espectro entre 1100 y 2500 nm, en intervalos regulares de 2 nm. Un primer conjunto de espectros de copolímeros EPR/PP (88 espectros) fue medido en condiciones instrumentales inadecuadas. Este conjunto de espectros se utilizó en el capítulo 6, dedicado a la detección de muestras *outlier* en la etapa de predicción del modelo. Para comprobar la estabilidad del modelo desarrollado para la predicción del % de etileno en EPR/PP, se registraron repeticiones periódicas del espectro de nueve de las 177 muestras durante tres meses. Cada una de las nueve muestras fue analizada dos veces diariamente (mañana y tarde), por dos analistas diferentes, con lo que se obtuvo un total de 150 espectros para cada una de las nueve muestras.

3.1.4 Análisis de los espectros obtenidos

Nafta

En la Fig. 26 se muestra un IR medio de una de las muestras de nafta. Las bandas más intensas se encuentran cerca de 3000 cm^{-1}: la tensión de C-H en CH_3 a 2956 cm^{-1} y 2872 cm^{-1} y la tensión de C-H en CH_2 a 2926 cm^{-1} y 2853 cm^{-1}. También tienen una intensidad importante las bandas a 1495 cm^{-1} y 1385 cm^{-1}, que corresponden a torsiones de tijera (simétrica y asimétrica) del grupo CH_2. A números de onda más pequeños se encuentran otras bandas aunque de asignación

más compleja como las bandas de deformación de C-H aromático en el plano y fuera de él.

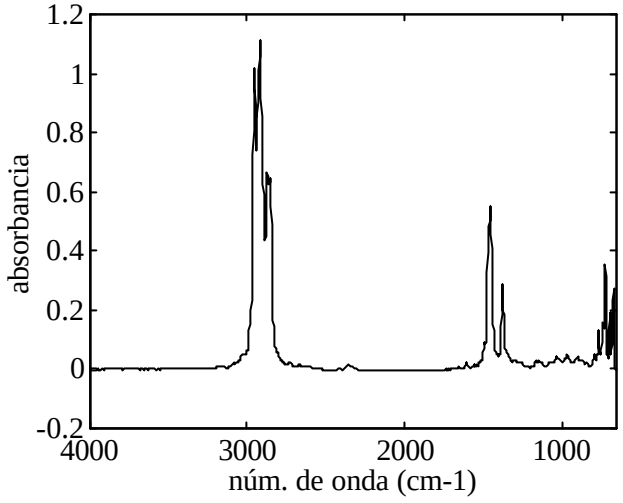


Fig. 26. Espectro en el infrarrojo medio (4000-650 cm^{-1}) de una de las muestras de nafta, medido utilizando un sistema de reflectancia total atenuada horizontal (HATR).

En la Fig. 27 se muestra el espectro en el infrarrojo cercano de una muestra de nafta. La zona de trabajo del espectrofotómentro (1100-2500 nm) permite registrar las zonas del primer sobretono del enlace C-H en CH_2 y CH_3 (1695-1765 nm), del segundo sobretono (1150-1210 nm), así como bandas de combinación a 2150-2450 nm y 1340-1435 nm. Las bandas de combinación por encima de 2300 nm poseen intensidades muy elevadas y finalmente se utilizó el intervalo de longitudes de onda entre 1100 y 2230 nm, que comprende el primer y segundo sobretono (mucho menos intenso que el primero), y las zonas de bandas de combinación a 1400 y 2220 nm.

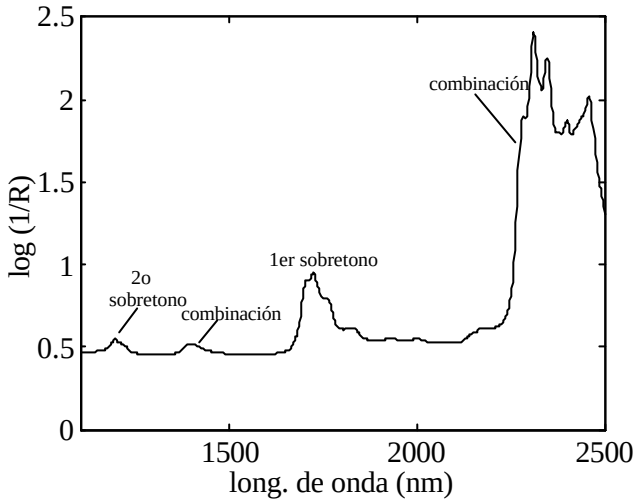


Fig. 27. Espectro NIR (1100-2500 nm) de una muestra de nafta. Se han señalado el origen de las bandas (C-H en metilo, metileno y aromático) en cada una de las zonas del espectro.

Polipropileno

En la Fig. 28 se muestra el espectro en el infrarrojo cercano tres muestras de polímeros: EPR/PP, EPR y PP. Las diferencias en las posiciones de la línea base entre los tres polímeros está motivada por la diferencia en los tamaños de las partículas entre las muestras, y es común en la espectroscopia por reflectancia. La corrección de este efecto se trata en el apartado 6.3.

Los polímeros producidos a partir de la polimerización de propileno y etileno están formados por grupos metilo (CH_3) y metileno (CH_2) y las bandas que se observan son las misma que ya se han comentado en el caso de las muestras de nafta. Con la finalidad de facilitar la comparación de bandas en la Fig. 29 se puede observar un espectro de nafta y un espectro del copolímero EPR/PP.

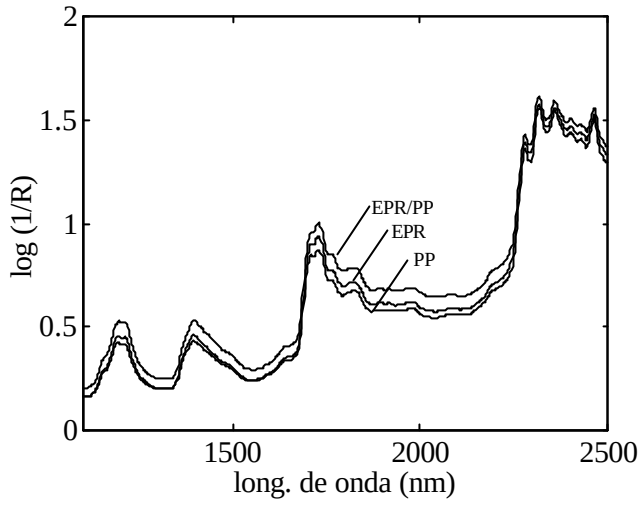


Fig. 28. Espectro NIR (1100-2500 nm) de muestras de polímeros: EPR/PP, EPR y PP.

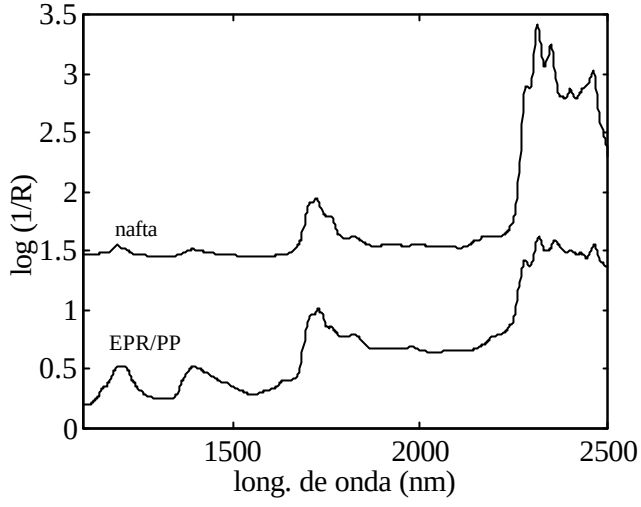


Fig. 29. Espectro NIR (1100-2500 nm) de una muestra de nafta (arriba) y de una muestra EPR/PP. El espectro de nafta se ha desplazado hacia arriba para facilitar la comparación.

4. Modelos de calibrado

4.1 INTRODUCCIÓN

En el presente capítulo se describe brevemente el desarrollo de los modelos de calibración realizados en el transcurso de la presente tesis doctoral. En primer lugar se presentan los modelos de calibración de las propiedades de los polímeros (apartado 4.2). Algunos aspectos relacionados con el modelo de predicción del % de etileno en copolímeros EPR/PP (o iPP) fueron estudiados en profundidad y publicados como trabajos independientes y se presentan en próximos capítulos: el control y transferencia del modelo o la detección de *outliers* en predicción (capítulos 5 y 6). Otros modelos referidos a propiedades físicas del polímero fueron menos satisfactorios y sólo se presentan brevemente.

En el siguiente apartado del capítulo (apartado 4.3) se presentan los modelos para la determinación del PIONA desglosado en muestra de nafta por espectroscopia de infrarrojo medio y que fueron presentados en el artículo "*Multivariate determination of several compositional parameters related to the content of hydrocarbon in naphtha by MIR*" publicado en la revista *Analyst*.

En el desarrollo de los modelos se ha seguido siempre los siguientes pasos:

a.- Estudio previo de los espectros a través de una descomposición en componentes principales, que permite estudiar la presencia de agrupaciones y muestras discrepantes muy claras.

b.- Modelo de calibrado

b.1.- Selección de las muestras a usar en el modelo.

b.2.- Selección del número óptimo de factores a incluir en el modelo.

b.3.- Detección de la presencia de muestras discrepantes (*outliers*).

b.4.- Cálculo del error medio de predicción.

b.5.- Comprobación de la ausencia de sesgo.

b.6.- Cálculo de errores de predicción específicos de cada muestra.

4.2 MODELOS DE CALIBRACIÓN DE POLÍMEROS DERIVADOS DEL POLIPROPILENO

4.2.1 Agrupación de las muestras

En primer lugar se realizó una descomposición en componentes principales del conjunto total de espectros, corregidos por un pretratamiento SNV. Este conjunto comprende 179 muestras de EPR/PP, 69 muestras de EPR y 70 muestras de PP. La representación de los dos primeros componentes principales se muestra en la Fig. 30.

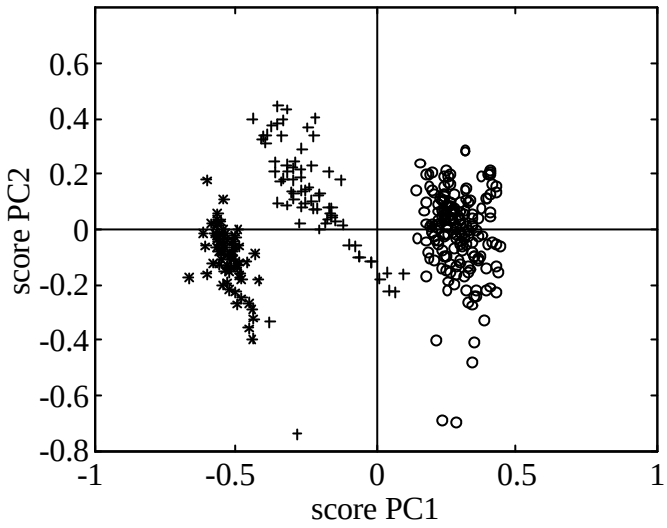


Fig. 30. *Scores* en los dos primeros componentes principales de las muestras de polímeros: EPR/PP (o), EPR (+) y PP (*). Los dos primeros componentes principales retienen un 75.9% y un 17.5% de la variabilidad total respectivamente. Estadístico de Hopkins, con el 100% de la población y 20 objetos, es igual a H=0.93 (H_{min}=0.92, H_{max}=0.94), el valor próximo a 1 indica agrupación en los datos.

Debido a la diferente naturaleza de los polímeros, los espectros de las muestras se agrupan según el tipo de polímero, y la agrupación observada se ve corroborada por el estadístico de Hopkins (H=0.93). Por este motivo, *a priori*, es recomendable

utilizar modelos diferentes para predecir el % de etileno en copolímeros EPR y EPR/PP.

4.2.2 Determinación del porcentaje de etileno en EPR/PP

Análisis de agrupaciones

Las 179 muestras de EPR/PP se dividieron en dos conjuntos (calibración y validación), ya que se consideró que no era necesario un número tan elevado de muestras para desarrollar el modelo. Se escogieron 80 muestras de copolímero utilizando el algoritmo de Kenard-Stone aplicado sobre los *scores* de una descomposición PCA de los espectros. La descomposición PCA se utiliza también para cuantificar el nivel de agrupación de las muestras y para la detección de muestras discrepantes. La representación gráfica de la dispersión de las muestras en los dos primeros componentes principales de la descomposición se muestra en la Fig. 31.

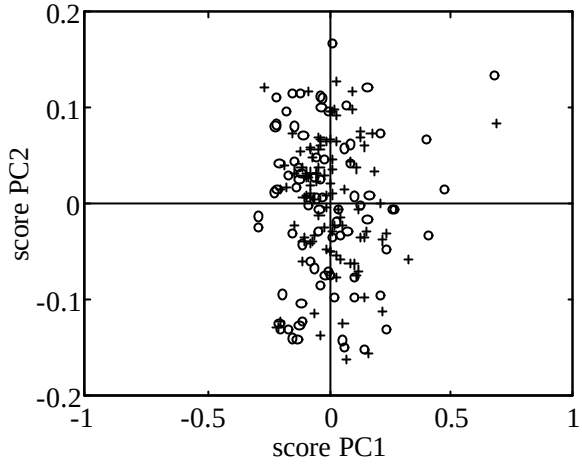


Fig. 31. Distribución en el espacio de los dos primeros componentes principales de las muestras de EPR/PP. El conjunto de calibración de 80 muestras (+) y el conjunto de validación de 99 muestras (o). Los dos primeros componentes principales explican respectivamente, un 72.8% y un 16.4% de la variabilidad en los espectros.

A partir del estudio de este gráfico (y de los *scores* en PC superiores) no parece haber una agrupación severa de las muestras, aunque el estadístico de Hopkins, que oscila entre 0.5 (conjunto homogéneo) y 1 (agrupaciones), con el 100% de la población y 20 iteraciones tiene un valor de H=0.64 (H_{min}=0.62-H_{max}=0.68), ligeramente por encima del 0.5.

Sobre los *scores* de la descomposición PCA se pueden aplicar estrategias de detección de muestras discrepantes, como el test de *outliers* de Grubb's sobre el estadístico de Rao, que señala las siguientes muestras.

Tabla 3. Muestras señaladas en el test simple de Grubbs aplicado sobre el estadístico de Rao.

D^2_{PC1}	D^2_{PC2}	D^2_{PC3}	D^2_{PC4}
-	176C, 62C	164C, 40C, 118C	87C, 176C, 143C, 47C

Modelo de calibración

Con las 80 muestras del conjunto de calibración se construyó un modelo para la predicción del contenido en etileno. El número de factores a incluir en el modelo se determinó mediante una validación cruzada completa, utilizando diferentes criterios (Haaland [1], Unscrambler [2] y Osten [3]) aplicado sobre el PRESS (Fig. 32).

Se analizó la presencia de muestras discrepantes utilizando medidas de la importancia de los objetos (*leverage*) en el conjunto de calibración, del residual espectral y del residual en la concentración (% de etileno en este caso). En el conjunto de calibración, una de las muestras (176C) presenta un *leverage* elevado, aunque el modelo la predice correctamente (residual pequeño), por lo que no fue eliminada del conjunto de calibración. Otra de las muestras (62C), que ya fue señalada en la descomposición PCA inicial de los espectros, también tiene un *leverage* alto. Esta muestra fue eliminada del conjunto de calibración, ya que era

una muestra única, con un contenido en etileno muy bajo (5%), en comparación con el resto de muestras del conjunto. En la predicción de las muestras no incluidas en la calibración (*test set*) se observó la presencia de dos muestras *outlier* (27C, 93C), que se pueden observar claramente en la Fig. 33, donde se muestra la varianza residual en la predicción del % de etileno frente al *leverage*, en las muestras de calibración y en las de validación.

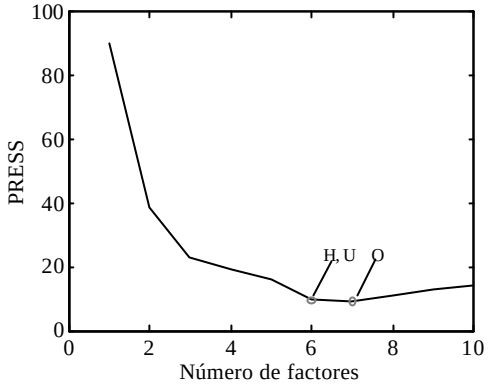


Fig. 32. PRESS en función del número de factores incorporados al modelo. Se señala el número óptimo que proponen diferentes criterios (H: Haaland, U: Unscrambler y O: Osten).

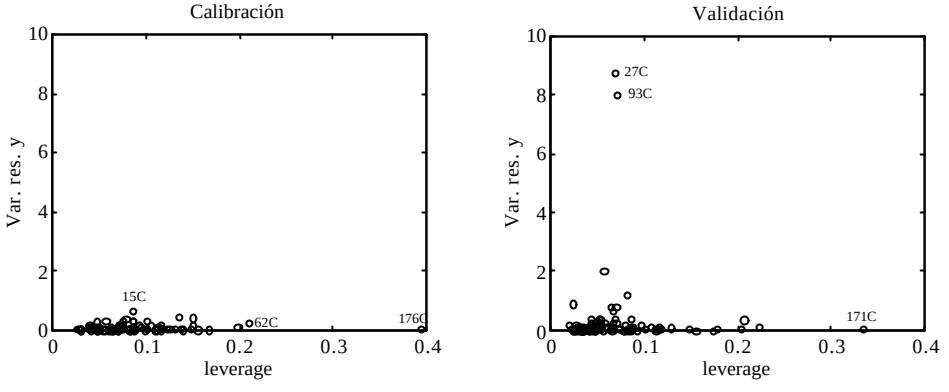


Fig. 33. Varianza residual en y frente al *leverage* en la etapa de calibración (izq.) y en la etapa de validación (der.). Se ha mantenido la misma escala para facilitar la comparación.

La presencia de error sistemático en la calibración se comprobó con el test conjunto de la pendiente y la ordenada en el origen. En la Tabla 4 se muestra el valor del nivel de significancia (a) asociado con la F calculada.

La habilidad de predicción global del modelos se midió a través de los valores RMSECV, calculado para las muestras de calibración en la validación cruzada completa y el RMSEP, calculado con las muestras no incluidas en el modelo. Ambos valores se muestran en la Tabla 6.

En la Fig. 34 se muestra el valor de % de etileno medido por el método de referencia frente al % de etileno predicho por el modelo multivariante en la validación cruzada completa. Se observa que existen dos grupos de muestras, uno con muestras que contienen un porcentaje de etileno entre el 8 y el 10%, mientras el segundo grupo de muestras contiene entre el 12 y el 13% de etileno. Esta agrupación se produce como consecuencia de las características de la producción del polipropileno.

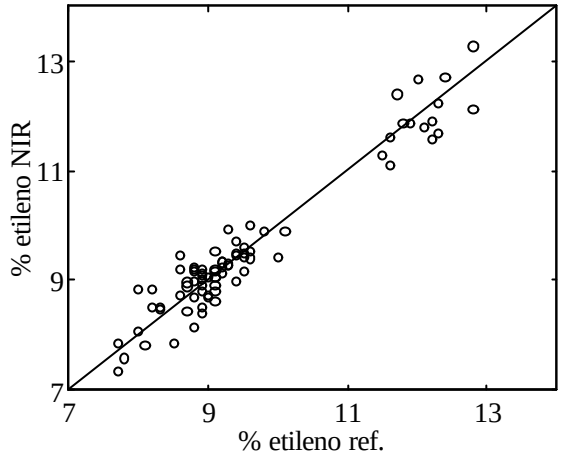


Fig. 34. Porcentaje de etileno medido por el método de referencia frente al predicho por el modelo multivariante.

En la Tabla 4 se muestra el resumen del modelo de calibración para la predicción del % de etileno en EPR/PP. Se indican el número de factores utilizados por el modelo, las muestras que fueron detectadas como *outlier* (tanto en el conjunto de calibración como en el de validación), el resultado del test de detección del error sistemáticos, así como el error global de predicción del modelo (RMSCV y RMSEP) y sus valores relativos.

Tabla 4. Tabla resumen del modelo de predicción de % de etileno en EPR/PP.

Propiedad	**Reproducibilidad técnica ref.**	**Núm. factores**	***Outliers***	**Test de *bias***	**Error de predicción (% etileno)**	**Error relativo**
% etileno	0.10-0.15 % (p/p)	6	62C 27C, 93C (val.)	89.2%	RMSCV=0.35 RMSEP=0.34	RRMSCV=3.7% RRMSEP=3.7%

4.2.3 Determinación del porcentaje de etileno en EPR

Análisis de agrupaciones

Se realizó un análisis de agrupaciones con los 70 espectros de copolímeros EPR. La dispersión de las muestras en los *scores* de los dos primeros componentes principales se muestra en la Fig. 35.

Cuatro de las muestras de EPR tienen un comportamiento muy diferente al resto. Tres de estas muestras (3R, 28R y 29R) tienen un contenido en etileno muy bajo (1%), mientras el resto de las muestras contienen entre 3.5% y 4.5%. Esta muestras son señaladas en los primeros componentes principales cuando se realiza el test simple de Grubbs sobre el estadístico de Rao.

Tabla 5. Muestras señaladas en el test simple de Grubbs aplicado sobre el estadístico de Rao.

D^2_{PC1}	D^2_{PC2}
3R, 28R, 29R, 51R	51R, 29R, 20R

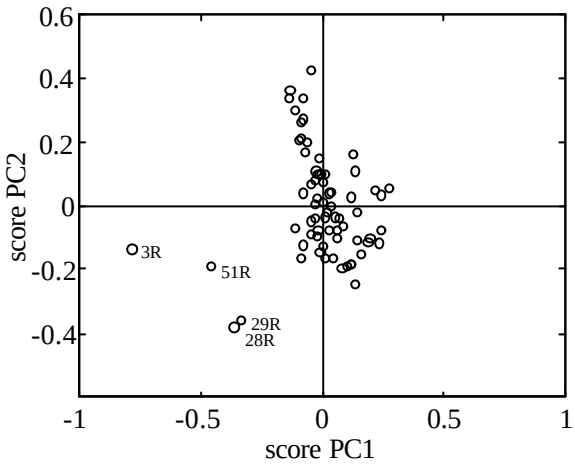


Fig. 35. Distribución en el espacio de los dos primeros componentes principales de las muestras de EPR. Los dos primeros componentes principales explican respectivamente, un 75.3% y un 22.5% de la variabilidad total de las muestras.

Debido al comportamiento tan diferente de estas muestras, no se podrán utilizar al construir el modelo de calibración par la determinación del % de etileno en EPR. Eliminando estas cuatro muestras el grado de agrupación, medido por el estadístico de Hopkins, es de H=0.62 (H_{min}=0.59, H_{max}=0.66), por lo que no parece haber diferentes grupos de muestras.

Modelo de calibración

Debido al número de muestras disponible (66 muestras), se utilizó el conjunto completo para construir el modelo y la validación cruzada completa para la selección del número óptimo de factores y la determinación de la habilidad predictiva global del modelo. En la selección del número óptimo de factores se utilizaron los mismos criterios que en el modelo anterior.

La presencia de muestras discrepantes se analizó mediante medidas de la importancia de los objetos (*leverage*) en el conjunto de calibración y mediante los residuales.

Se comprobó la presencia de error sistemático con el test conjunto de la pendiente y la ordenada en el origen. En la Tabla 6 se muestra el valor del nivel de significancia (*a*) asociado con la *F* calculada. En este caso el valor encontrado es inferior al 5%, por lo que se deduce la presencia de error sistemático. Efectivamente, en el gráfico de valores de referencia frente a predicciones NIR (Fig. 36), se observa como los errores de predicción son grandes en comparación con el intervalo de % de etileno en las muestras. Probablemente, el intervalo de concentraciones de etileno es demasiado estrecho comparado con la reproducibilidad de la técnica de referencia como para obtener una calibración adecuada. De hecho, en el conjunto de calibración existen únicamente siete niveles de % de etileno: 3.6%, 3.7% ... hasta 4.2%.

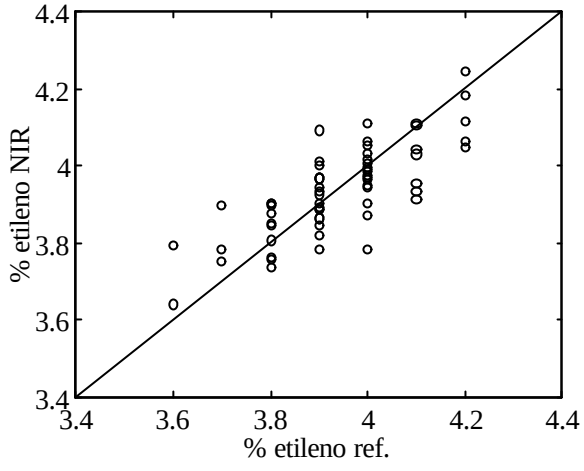


Fig. 36. Valor de % de etileno de referencia frente al predicho por el modelo multivariante con todas las muestras, 5 factores, RMSECV=0.09%.

La Tabla 6 resume los resultados obtenidos en el modelo de calibración para la predicción del % de etileno en EPR. Se indica el número de factores utilizado, los *outlier*, el resultado del test de detección del error sistemático (*bias*), así como el error global de predicción del modelo (RMSECV) y su valor relativo.

Tabla 6. Tabla resumen del modelo de predicción de % de etileno en EPR.

Propiedad	Reproducibilidad técnica ref.	Núm. factores	*Outliers*	Test de *bias*	Error de predicción (% etileno)	Error relativo
% etileno	0.10-0.15 % (p/p)	5	3R, 51R, 28R, 29R	1.6%	RMSCV=0.09	RRMSCV=2.3%

4.2.4 Determinación del índice de fluidez en copolímeros de EPR/PP, EPR y PP y viscosidad en EPR/PP y EPR.

Las propiedades mecánicas de los polímeros como el índice de fluidez o la viscosidad, que no tienen una relación directa con el espectro, son más difíciles de modelar. Se desarrollaron modelos para la determinación del índice de fluidez en EPR/PP, EPR y PP y modelos para la determinación de la viscosidad en EPR/PP y EPR. En todos los casos se siguieron los mismos pasos indicados en los modelos comentados anteriormente, aunque debido a los resultados finales obtenidos únicamente se mostrará aquí el modelo final, en forma de gráficos de la propiedad medida por el método de referencia frente al valor predicho por el modelo multivariante.

En la Fig. 37 se presentan los resultados de la validación cruzada completa para los modelos de índice de fluidez. Únicamente en el caso de los copolímeros EPR/PP se consigue un modelo que sigue aproximadamente la tendencia del índice de fluidez, aunque seguramente este hecho se produce por la correlación que existe entre el índice de fluidez y el % de etileno en las muestras. En cambio para los modelos de EPR y especialmente para el de PP homopolímero, los modelos desarrollados son insatisfactorios.

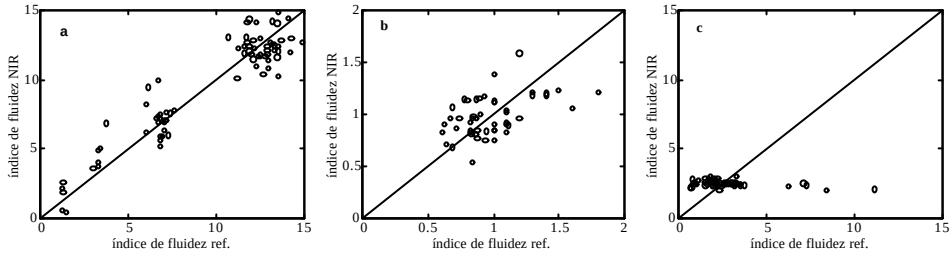


Fig. 37. Índice de fluidez medido por el método de referencia frente a la predicción del modelo multivariante para el copolímero EPR/PP, 8 factores, RMSECV=1.37% (**a**), el copolímero EPR, 5 factores, RMSECV=0.23% (**b**) y para el homopolímero de PP, 1 factor, RMSECV=1.8% (**c**).

En la Fig. 38 se muestran los resultados de los modelos de viscosidad en EPR/PP y EPR. En ninguno de los casos la técnica de calibrado no fue capaz de establecer una relación matemática satisfactoria entre el espectro de las muestras y su viscosidad.

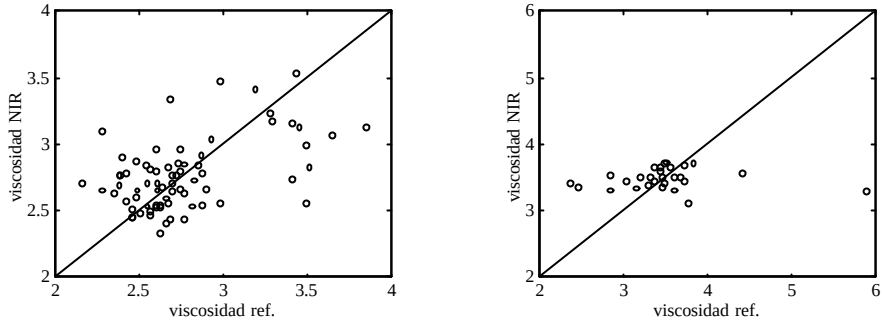


Fig. 38. Viscosidad medida por el método de referencia frente a la predicción del modelo multivariante para el copolímero EPR/PP (izq.) y para el copolímero EPR (der.).

4.3 MULTIVARIATE DETERMINATION OF SEVERAL COMPOSITIONAL PARAMETERS RELATED TO THE CONTENT OF HYDROCARBON IN NAPHTHA BY MIR SPECTROSCOPY

Analyst, 1999, **124**, 1827-1831

Santiago Macho, Ricard Boqué, Mª Soledad Larrechi, F. Xavier Rius

Departament de Química Analítica i Química Orgànica.
Universitat Rovira i Virgili. Pl. Imperial Tàrraco, 1, 43005-Tarragona, Spain

Abstract

Several compositional parameters of the main hydrocarbon families are determined in naphtha. This hydrocarbon mixture is traditionally analysed by gas chromatography (GC) but there are several alternative spectroscopic method such as near-infrared spectroscopy combined with multivariate calibration. Spectroscopic methods have the advantage that they are faster and more suitable for on-line analysis. Previous studies have determined the global percentage of each hydrocarbon family: linear and branched paraffins, naphthenes and aromatic compounds. Here we present the determination by mid infrared (MIR) spectroscopy and multivariate calibration of the most detailed compositional parameters provided by gas chromatography. Results were good, that is to say there was no bias and the root mean square error of cross validation (RMSECV) was low, for the determination of all the detailed naphthenes and aromatic compounds, and most n-paraffins and isoparaffins. Several methods of detecting outliers are discussed and the quality of the models is evaluated by sample specific error predictions calculated by two approaches: the Unscrambler expression and the Faber-Kowalski expression. The proposed method enables a more specific analysis of naphtha than other methods and provides the same information as GC.

Received 14th July 1999, Accepted 3rd November 1999

ver el artículo cortesía de The Royal Society of Chemistry.

4.4 AMPLIACIÓN DE ASPECTOS EXPERIMENTALES

Pretratamiento de los espectro de IR medio

Se observó que en algunos de los espectros de las muestras de nafta, medidos en el infrarrojo medio con el sistema de muestreo por reflectancia total atenuada horizontal (HATR), se produjo un desplazamiento lineal de la línea base, sobre todo entre 4000 y 2200 cm^{-1} (Fig. 39), que se corrigió mediante un *detrending* utilizando una línea recta para modelar el desplazamiento de la línea base. A continuación a los espectros también se les aplicó un pretratamiento SNV.

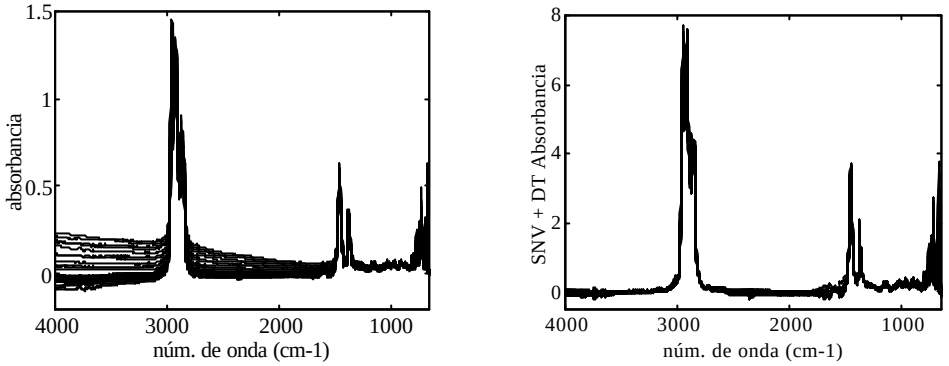


Fig. 39. Espectros de las muestras de nafta en el infrarrojo medio sin tratar (izq.) y después de la aplicación de los pretratamientos SNV y *detrending* (der.).

Determinación de los parámetros de composición en el infrarrojo cercano

La determinación de la composición desglosada de las muestras nafta también se llevó a cabo con las 132 muestras a las que se midió el espectro en el infrarrojo cercano (NIR), obteniéndose resultados muy similares a los obtenidos en el infrarrojo medio. En la Fig. 40 se muestra la comparación de los modelos para la determinación del benceno (aromáticos C_6).

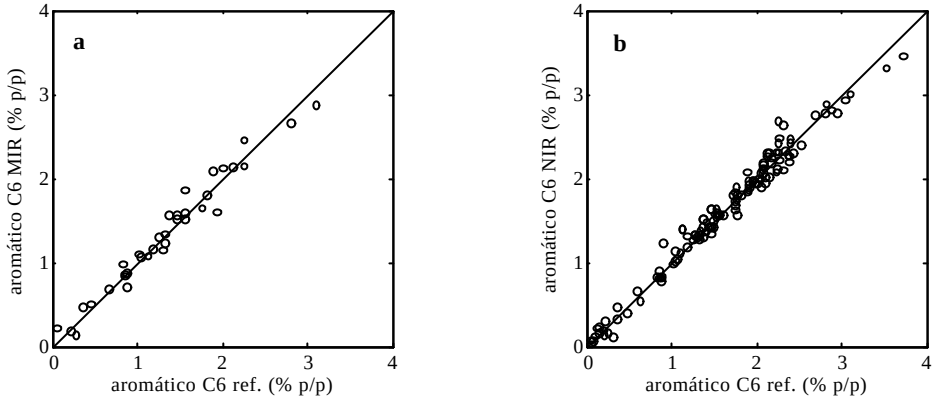


Fig. 40. Representación del % de aromáticos C6 determinado por el método de referencia frente al % aromáticos C6 predicho por el modelo PLS en el infrarrojo medio (a) y el modelo PLS en el infrarrojo cercano (b). En ambos casos se utilizaron 5 factores y los errores medios de predicción fueron RMSECV=0.13, en el caso del MIR y RMSECV=0.11, en el caso del NIR.

4.5 CONCLUSIONES

El desarrollo de los diferentes modelos de calibración para las muestras de polímero y para las de nafta ha demostrado que es posible obtener modelos de calibración útiles cuando se intenta predecir propiedades relacionadas con la composición química, como concentraciones o porcentajes de compuestos que absorben en el infrarrojo. En este caso existe una relación directa entre el espectro y la propiedad a determinar: el % de etileno en el copolímero de EPR/PP o el porcentaje en aromáticos, naftenos, etc. en nafta.

En ocasiones, los modelos de calibración para estas mismas propiedades relacionadas con la composición, no son capaces de obtener una relación adecuada debido a que el intervalo de concentraciones de las muestras de calibración es muy estrecho, con lo que el error experimental de la técnica de referencia se convierte en el término de error más importante del proceso de calibración. Tal sería el caso del % de etileno en muestras de EPR o los compuestos minoritarios en la nafta C_{10}

y C_{11}. En estos casos sería deseable aumentar el intervalo de trabajo de la propiedad de interés, utilizando muestras de EPR con valores de % de etileno por debajo de 3% y por encima de 5%. Incluso una calibración conjunta de muestras de EPR y EPR/PP, que tiene valores de % de etileno más altos, podría es útil, a pesar de las diferencias entre las muestras. El uso de técnicas de calibración basada en agrupaciones, como el método de *k* vecinos más próximos (*k-nearest neighbours*, KNN), sería otra alternativa que podría dar buenos resultados.

En un tercer grupo, se podrían agrupar todos aquellos modelos de propiedades cuya relación con el espectro NIR no es tan directa como en el caso de los compuestos químicos que absorben en el infrarrojo próximo. Este sería el caso de la determinación del índice de fluidez o la viscosidad en polímeros de polipropileno, donde no ha sido posible desarrollar modelos de calibración útiles, utilizando modelos lineales como el PLS. El uso de métodos de calibración no lineales como las redes neuronales o el PLS-no lineal, o el uso de transformaciones de la señal como la corrección ortogonal de la señal (*orthogonal signal correction*, OSC) [4] podría mejorar los resultados.

BIBLIOGRAFÍA

1 D.M. Haaland and E.V. Thomas, *Anal. Chem. 60*, **1988**, 1193.

2 K. Esbensen, S. Schönkopf and T. Midtgaard, *Multivariate Analysis - in practice*, ed. Camo A/S, Trondheim, **1996**, p. 111.

3 D.W. Osten, *J. Chemom. 2*, **1988**, 39.

4 S. Wold, H. Antti, F. Lindgren and J. Ohman, *Chemom. Intell. Lab. System. 44*, **1998**, 175.

5. Control y estandarización

5.1 INTRODUCCIÓN

El desarrollo de los modelos multivariantes requiere un elevado esfuerzo debido a que es necesario analizar un conjunto de muestras amplio. Al problema del gran número de análisis a realizar, se une la dificultad de obtener un conjunto representativo de muestras, lo que en un ambiente industrial puede significar tener que alargar en el tiempo la etapa de muestreo para conseguir esta representatividad.

Por tanto, una vez el modelo ha sido desarrollado, es deseable que éste sea válido por un periodo de tiempo cuanto más largo mejor. Esto significa que durante este periodo el error de predicción del modelo no es significativamente diferente del error de predicción obtenido en el momento de la calibración. De todas formas hay numerosas razones que pueden provocar que el modelo deje de ser válido y produzca predicciones erróneas: cambio del espectrofotómetro o de alguno de sus componentes, cambios en el ambiente como la temperatura o cambios en la propias características físicas o químicas de la muestra. Si se produce esta perdida de validez del modelo, es importante detectarla rápidamente y aplicar estrategias de estandarización de modelos que lo corrigen sin la necesidad de un proceso de recalibrado completo, con lo que se ahorra un gran cantidad de tiempo y trabajo.

En el presente capítulo se ha prestado un especial interés a la detección de la pérdida validez del modelo, concretamente cuando se producen cambios puntuales en el sistema de medida (espectrofotómetro y ambiente), así como a su posterior estandarización para conseguir de nuevo la validez del modelo. Se ha realizado una monitorización exhaustiva del modelo para la predicción del % de etileno en muestras de EPR/PP. Para esta monitorización se han utilizado técnicas de control, aplicadas a la medida periódica de un pequeño conjunto de muestras de control, representativas del conjunto de calibrado. Se han utilizado gráficos univariantes, como el gráfico de Shewhart, o gráficos basados en medidas multivariantes, como

el T^2 y Q, utilizados en el control estadístico multivariante de procesos (MSPC). Cuando se detectó la pérdida de validez del modelo multivariante, y aprovechando las experiencias realizadas para su control, se aplicaron diferentes estrategias de estandarización del modelo multivariante, todas ellas basadas en la medida de un pequeño conjunto de muestras representativas, antes y después de haberse producido el cambio instrumental. Entre las distintas técnicas de estandarización existentes se han utilizado dos, una muy simple, como es la corrección de la pendiente y el sesgo (SBC), y una segunda técnica más elaborada, como sería la estandarización directa por partes (PDS).

En el apartado siguiente (5.2) se presenta la publicación del trabajo realizado: "*Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy. Standardisation of the calibration model"* publicado en la revista *Analytica Chimica Acta*. A continuación se presenta una ampliación de algunos de los resultados obtenidos (apartado 5.3) y la conclusiones de este capítulo (apartado 5.4).

5.2 MONITORING ETHYLENE CONTENT IN HETEROPHASIC COPOLYMERS BY NEAR-INFRARED SPECTROSCOPY. STANDARDISATION OF THE CALIBRATION MODEL

Anlytica Chimica Acta, 2001, **445**, 213-220

S. Macho*, A. Rius, M.P. Callao and M.S. Larrechi

Departament de Química Analítica i Q. Orgànica.
Universitat Rovira i Virgili. Pl. Imperial Tarraco, 1. 43005 Tarragona, Spain

Abstract

The concentration of ethylene in samples of heterophasic copolymers was monitored for three months by NIR spectroscopy and partial least-squares (PLS) multivariate calibration in an attempt to assess the validity over time of the calibration model. Assuming that the model would cease to be valid, we selected and monitored sufficient samples so as to initiate a process of standardisation. The samples were selected with the Kennard-Stone algorithm and according to the representivity of their ethylene content. The monitoring techniques were univariate (Shewhart) and multivariate (T^2 and Q) control plots. The results show that the system was stable, until the point at which there was a cause in variation which gave rise to an erroneous prediction of the ethylene content. This change did not vary over time. We used two techniques to standardise the calibration model: slope bias correction (SBC) and piecewise direct standardisation (PDS). They both corrected the error detected.

Keywords: NIR; SPC; MSPC; Standardisation; SBC; PDS
Received 16th March 2001; accepted 12th July 2001

1. Introduction

The main aim of this study is to monitor the validity over time of a multivariate calibration model developed to determine ethylene in samples of heterophasic copolymers using near-infrared spectroscopy (NIR) and partial least-squares (PLS) multivariate calibration [1]. This technique is sensitive to a variety of factors such as temperature, the ageing of the source, etc. which suggests that the model may cease to be valid after some time. Therefore, a second aim of this study is to draw up a protocol that shall enable it to be used by applying standardisation techniques.

Near-infrared (NIR) spectroscopy is a promising tool for process analysis of synthetic polymers. Real time monitoring will reduce the amount of product manufactured under "outside specifications" [2-4]. Propylene/ethylene is an extremely popular polymer that serves as a basic feed material for injection moulding industries. NIR spectroscopy has been successfully implemented to determine these polymers in conjunction with multivariate calibration techniques [5]. However, chemical process monitoring do not use multivariate calibration models enough because small instrumental disturbances may affect the validity of the models over time. Therefore, for an on-line system to be developed, the model for the process must be accurate for a long time [6].

Control charts are widely used in industry to help control the quality of products. The most frequent type is univariate graphs, which represent a parameter of interest against time. If the property of interest is derived from a multivariate signal, multivariate statistical process control (MSPC) methods, control charting of T^2 and the residual measure Q can be applied [7]. The main drawback of these graphs is that the parameter being monitored is a statistic that has no chemical meaning and they are therefore, more difficult to interpret. On the other hand, unlike the univariate graphs, with subsequent processing they can show the causes of change, because it is the signal that is being processed and it does not interact with the calibration model [8].

Developing a multivariate calibration model is a long process and having to go back and start from scratch when the model ceases to be valid can present a problem. For this reason, it would be desirable to have the chance to standardise the model. Standardisation strategies have been studied in depth in recent years [9-11], particularly, in the area of NIR instrumentation. The complexity of standardisation techniques varies significantly and their suitability depends on the problem that causes the invalidity of the model. However, we have found no references to their being applied in conjunction with monitoring processes and real data.

This study starts by selecting the samples to be measured over time. Two criteria were used to make the selection. First, the Kennard-Stone criterion was applied to the complete sample set to choose representative samples as a function of the signal. Second, the concentration of analyte in these samples was taken into account. Monitoring several samples will show whether instrumental changes equally affect samples from the centre and the edges of the model. If it is detected that the calibration model is not valid, monitoring will also enable previous work to be used to adjust the model to new situations. Carrying out the monitoring process on various samples is not a considerable experimental cost because NIR is fast and the sample does not need to be pre-treated.

We used three types of graph: Shewhart graphs [13], which monitor the ethylene concentration, and the T^2 and Q graphs [7], which were obtained from the signals and used to compare the information that they supplied.

During the monitoring process, there was a cause of variability that prompted us to carry out a process of standardisation, for which two techniques were used and compared: slope bias correction (SBC) [14-16], whose theoretical basis and mathematical algorithm are straightforward, and piecewise direct standardisation

(PDS) [17-19], which is widely used and gives good results in many problems of this type.

2. Theory

2.1. Sample selection

The Kennard-Stone algorithm selects the samples which are furthest from each other in the group one by one [12]. The term used to measure the distance is the Euclidean distance. For a response matrix with N samples (rows) and K variables (columns), the multivariate Euclidean distance between samples i and j is:

$$D_{ij} = \left\| \mathbf{x}_i - \mathbf{x}_j \right\| = \sqrt{\sum_{v=1}^{K} (x_{iv} - x_{jv})^2} \qquad (1)$$

The first step is to choose the two furthest samples (maximum D_{ij}). The third sample is selected by carrying out the following steps: the distance between each object and the two furthest samples is calculated; the shortest of each of these pairs of distances is chosen and the sample with the maximum value in this group of minimal distances is selected.

Generalising, if M samples from the total N have been chosen, the next sample $M+1$ is selected by calculating:

$$d_i\,(M) = \min \{ D_{i1}, D_{i2}, ..., D_{iM} \}$$

for the N-M samples that have not been selected previously. Of these, the one which complies with the following equation is chosen:

$$d_i\,(M+1) = \max \{ d_i\,(M) \}$$

The algorithm does not indicate how many of them are required. However, if a standardisation is to be more advantageous than a new calibration process, the number of samples should not be very high.

2.2. Statistical control algorithms

2.2.1 Shewhart charts

These were the first control charts to be developed and are very simple to use. They plot the measurements against time or against the order in which they were made [13].

X Sewhart graphs control the mean of the process. These charts are built with m as a central line and usually have warning limits at $\pm 2s$ and control limits at $\pm 3s$. These limits have probabilities of error of approximately 5 and 0.2%, respectively, that a point falling beyond the limits is in control. The parameters m and s are usually estimated as follows:

$$m = \bar{x} \; ; \qquad s = \frac{s}{c_4} \tag{2}$$

where $\bar{x}$ is the mean of the data, s the standard deviation, and c_4 a corrector value, which depends on the number of measurements [7].

2.2.2 T^2 and Q statistics:

The results of a multivariate measurement of a control sample can be condensed in a matrix **X** with dimensions m (number of measurements recorded on the control samples) by n (variables). This matrix can then be reduced by principal components analysis to a new matrix **T**, m rows by k columns (number of factors),

which contains the maximum information about the original data, in accordance with the following expression [20]:

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad (3)$$

where **T** is the matrix of scores (coordinates of the samples in the new space of the principal components), **P** is the matrix of loadings (contributions of each variable to the principal components) and **E** is the matrix of the residuals of the model (information from the original matrix that the principal components model does not consider).

For a sample *i*, the statistic Q_i is calculated from the residual $\mathbf{e_i^T}$ (row *i* of matrix **E**) and its transposed vector $\mathbf{e_i}$ by:

$$Q_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{e}_i \qquad (4)$$

The Q parameter indicates the lack of fit of the data to the PCA model, i.e. the causes of variability that are not included in the model.

The limit value Q_{lim} [21] is calculated as follows:

$$Q_{lim} = \Theta_1\left[\frac{c_a\sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0(h_0-1)}{\Theta_1^2}\right]^{\frac{1}{h_0}} \qquad (5)$$

where h_0 and Θ_i are different combinations of the eigenvalues of the principal components not contained by the model and c_a is the standard normal deviation for the level of significance α chosen .

From the measurement of a sample *i*, $\mathbf{x_i^T}$ (row *i* of matrix **X**), T_i^2 is calculated as:

$$T_i^2 = \mathbf{x}_i^{\mathrm{T}}\mathbf{P}_k^{\mathrm{T}}\,\boldsymbol{l}^{-1}\mathbf{P}_k\mathbf{x}_i \qquad (6)$$

where $\mathbf{P_k}$ is the loading matrix of the model for *k* factors and *l* the diagonal matrix of the eigenvalues.

Hotelling's T^2 statistic includes the variation of the samples in the PCA model. The T^2 charts have the control limit $T^2_{\lim}$ [22]:

$$T^2_{\lim} = \frac{k(m-1)}{m-k} F_{k,m-k,a} \qquad (7)$$

2.3. Standardisation algorithms

2.3.1. Slope bias correction (SBC)

This is a simple univariate standardisation technique which establishes a linear regression between $\hat{\mathbf{c}}_{\mathbf{T1}}$ and $\hat{\mathbf{c}}_{\mathbf{T2}}$. These are the concentrations predicted by the model with the responses of the selected samples in the first and the second experimental conditions, respectively [14]:

$$\hat{\mathbf{c}}_{\mathbf{T1}} = bias + slope \times \hat{\mathbf{c}}_{\mathbf{T2}} \qquad (8)$$

Both the *slope* and the intercept (so-called *bias* here) of the regression are applied to obtain the standardised concentration $\left(\hat{c}_{2,unk}\right)^{std}$ as follows:

$$\left(\hat{c}_{2,unk}\right)^{std} = bias + slope \times \hat{c}_{2,unk} \qquad (9)$$

2.3.2. Piecewise direct standardisation (PDS)

This is a multivariate standardisation technique in which the response of each variable in the first experimental conditions r_{1i} is related to the response of a group

of variables $\mathbf{R}_2 = [r_{2i-j} \ldots r_{2i} \ldots r_{2i+j}]$ in the second experimental conditions. By extending this to all the transfer samples, the relation can be written as [17]:

$$\mathbf{r}_{1i} = \mathbf{R}_{2i}\mathbf{b}_i + \mathbf{b}_{0i} \qquad (10)$$

where a multivariate regression (e.g. PCR) can calculate the regression coefficients $\mathbf{b}_i$ and the offset term $\mathbf{b}_{0i}$. The number of columns in matrix $\mathbf{R}_{2i}$ is called the window size i.e. the number of variables from second experimental conditions involved in the relation.

If the process is repeated for all the wavelengths, *n* vectors of *b* coefficients and *n* values of b_o are obtained. These are grouped in the corresponding matrix and vector:

$$\mathbf{F} = \mathbf{diag}\left(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_i, \ldots, \mathbf{b}_n\right)$$
$$\mathbf{b}_{\mathrm{O}}^{\mathrm{T}} = \left(b_{\mathrm{O}1}, b_{\mathrm{O}2}, \ldots, b_{\mathrm{O}i}, \ldots, b_{\mathrm{O}n}\right)$$

This matrix and this vector are used to correct $\mathbf{r}_{2,\mathrm{unk}}^{\mathrm{T}}$, a sample measured in the new experimental conditions, to the standardised $\left(\mathbf{r}_{2,\mathrm{unk}}^{\mathrm{T}}\right)^{\mathrm{std}}$:

$$\left(\mathbf{r}_{2,\mathrm{unk}}\right)^{\mathrm{std}} = \mathbf{r}_{2,\mathrm{unk}}^{\mathrm{T}}\mathbf{F} + \mathbf{b}_0^{\mathrm{T}} \qquad (11)$$

Once the spectrum has been corrected, the concentration can be predicted with the initial model, and the concentration can be obtained in two standardised conditions $(\mathbf{c}_{2,\mathrm{unk}})^{\mathrm{std}}$.

3. Experimental

3.1 Samples

We obtained 177 samples of impact-resistant poly(propylene-ethylene) copolymer (impact PP) from TPD, Tarragona, Spain, over a 9-month period of production. With these samples, we constructed and validated a PLS calibration model to determine their ethylene content. Nine of these samples were analysed twice a day for three months by two different analysts and 150 spectra were generated for each sample.

3.2. Instrumentation

Spectroscopic data between 1100 and 2100 nm were collected by diffuse reflectance on a Bran-Luebbe IA500 spectrometer, equipped with a rotating cup drawer. This system can scan a larger area of the sample, which improves the repeatability of the spectrum.

3.2.1. Software

The results were processed with calculation routines programmed with the MATLAB 4.0 software [23].

4. Results and discussion

Fig. 1 shows how the 177 samples are distributed in the space of the first two principal components of the spectral matrix. These two components retain 98% of the variance. We have numbered the first six samples (circles) in accordance with the order in which they were selected by the Kennard-Stone algorithm. The following three samples (squares) were selected so that they would cover the whole concentration range.

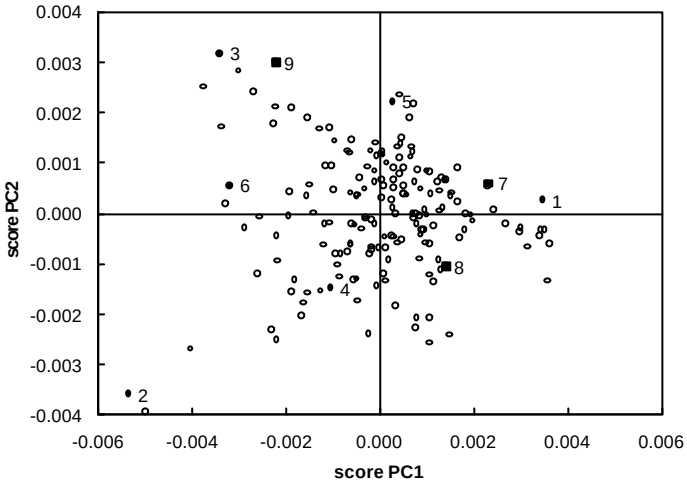


Fig. 1. Distribution of the calibration samples (○) in the space defined by the first two principal components (98% of variance). Nine samples were selected for the control and later standardization of the model. Of these, six were selected consecutively by the Kennard-Stone algorithm (●). The adjacent number indicates the order of selection. The three remaining samples were selected so as to cover the whole range of concentrations (■).

A previously validated PLS calibration model was used to predict the ethylene content of the nine samples selected. The Shewhart graph of each of the samples was constructed by monitoring the percentage of ethylene predicted by the model. The mean and the standard deviations of the first 40 measurements were used to assess the limits. All the graphs are similar and, by way of example, the results for sample 8 are shown in Fig. 2. The ethylene content seems to be in control until approximately measurement 50 when it was systematically increased by an unexpected cause of variation. All the remaining values were over the warning limit. One reason for the change in the ethylene prediction may be a wavelength instrumental shift that could not be corrected by the automatic wavelength calibration procedure with polystyrene. Other sources of variation cannot be discarded.

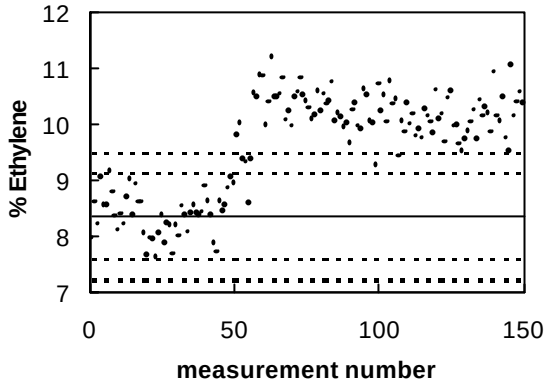


Fig. 2. Shewhart graph for the ethylene content of the sample.

T^2 and Q control charts were constructed by using a PCA model of the first 40 measurements of each sample. Limits were calculated for α=0.05 and α=0.002. As an example, Figure 3 shows the T^2 and Q graphs for the sample mentioned above. The results shown by the Shewhart chart can also be seen in the multivariate control charts. After approximately the 50th measurement all the samples are out of control in both charts. This is more apparent in the Q chart than in the T^2 chart, which suggests that the change in the signal is greater in the part that is not considered in the established PCA model than in the part that is. An analysis of the spectra collected for this sample at the beginning of the control process (recording order 1), in a control (recording order 2) and out-of-control (recording order 57) situation shows that there is no apparent difference between them.

So the difference between the values of the spectra in the two conditions was determined and the results are shown in Fig. 4. It can be seen that some areas of the spectrum are more sensitive to random variations and it is precisely these areas that are affected by the change. This may be why the change in graph Q which represents the information not considered in the PCA model, and therefore, the randomness of the signal is magnified.

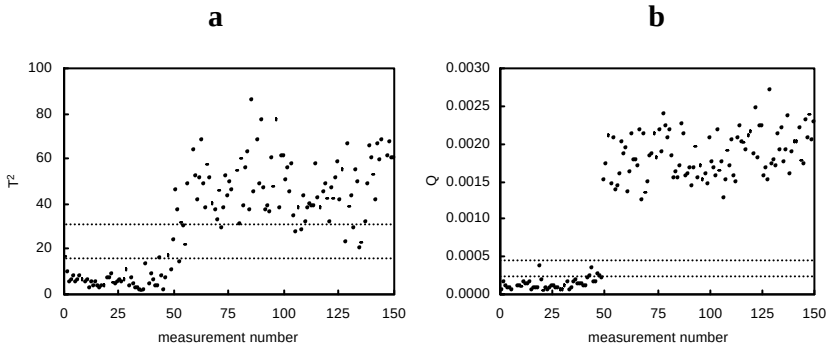


Fig. 3. (a) T^2 graph and (b) Q graph, for the sample.

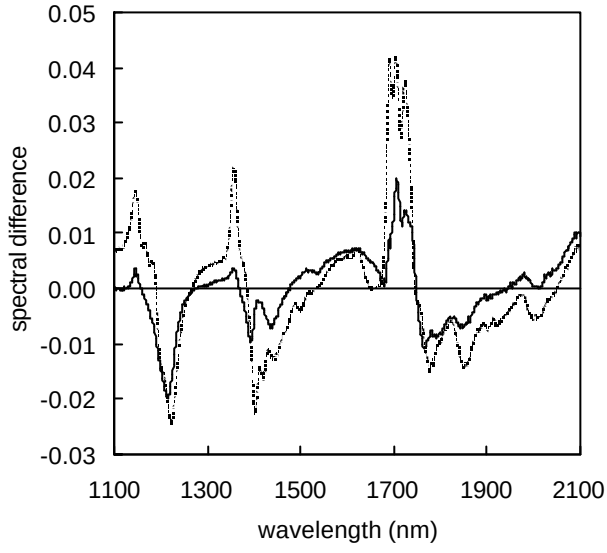


Fig 4. Spectral differences for two consecutive spectra in the first condition (——) and for two spectra in different conditions, first and 57th measurements (· · ·).

Once we had verified that the increase was maintained over time and that we could not resolve the problem by finding a cause that could be corrected, we standardised the model by working with the nine selected samples.

To standardise, we took the concentration and signal values recorded at the beginning of the monitoring process (order 1, first conditions) as the in-control

reference situation. We took the values at order 57, the point up to which we considered that the change was maintained, as the out-of-control reference situation (second conditions).

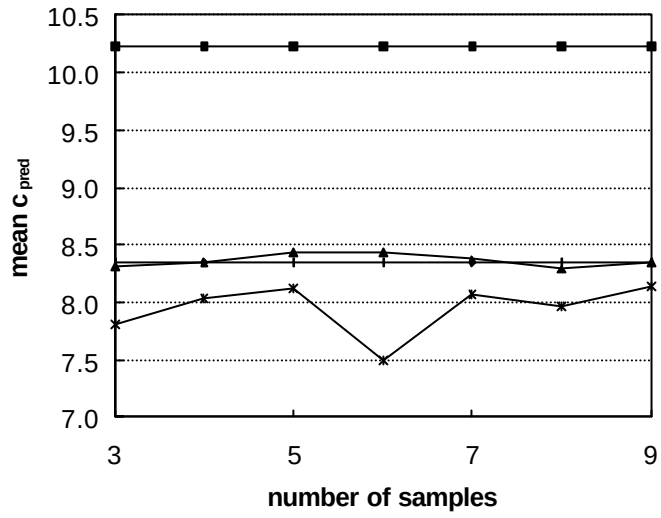


Fig. 5. Ethylene concentration predicted by the model versus number of samples used for the standardisation: first conditions (♦); after the change or second condition (■); after standardisation by SBC (△); and after standardisation by PDS (*).

Fig. 5 shows the results of applying the standardisation using SBC and PDS and different numbers of standardisation samples. For purposes of comparison, we have indicated the predicted concentration at the beginning (first conditions) and after the change (second conditions). For PDS, the number of standardisation samples and the width of the window were the variables taken into account. Various window widths were tried and we observed that, in this case, it was not an important variable. Finally, we decided to use a window width of 5.

The number of samples used in the standardisation is not a parameter of great importance. This is consistent with what was observed when all the samples were

monitored because, as mentioned above, the change was similar in all of them. Fig. 5 shows that the corrections obtained fit better with SBC than with PDS. Nevertheless, we should point out that the definition of the first and second conditions affects the results and should not, therefore, be strictly interpreted.

One further aspect that should be mentioned is that the results are more sensitive to the number of standardisation samples for PDS than for SBC, which means that, in this respect, the latter technique would be best. No general recommendation in favour of one or other of these techniques can be made because the results using the SBC technique are good when the response variation between the first and the second conditions is not complex and affects all the samples in the same direction. When this is not the case, results could be better with a more complex technique such as PDS standardisation, which can remove horizontal and vertical signal shifts.

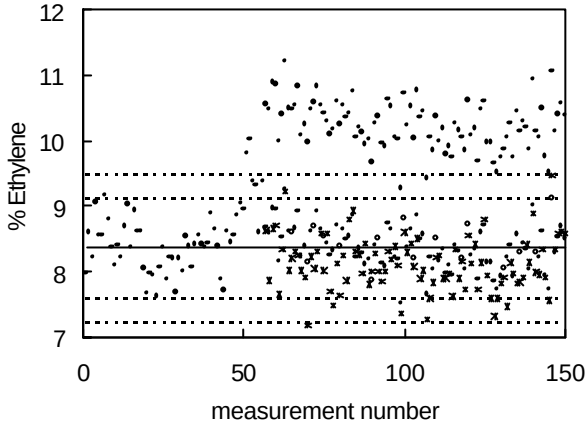


Fig. 6. Shewhart graph of the ethylene concentration in the sample. Uncorrected (●), standardising with SBC (○) and standardising with PDS (*).

Fig. 6 shows the monitoring of the concentration before and after applying both the standardisation techniques with six samples. Most of the standardised values are within the warning limits, though some specific data obtained after

standardisation by PDS are between the warning and the control limits. Therefore, the standardisation process was effective. From the T^2 and Q control charts after PDS standardisation, we can see that most of the T^2 values are below the limits, whereas the Q values are not corrected. This is seen by the presence of artifacts generated by the PDS standardisation, which mainly affects the information that is not contained in the PCA model (i.e. the residual spectra).

5. Conclusions

Monitoring techniques enable an important quality parameter to be established: the validity of the model over time. This is particularly important in multivariate calibration methods in which the experimental cost of establishing a model is high. These techniques are not complicated to implement.

In the case studied, the univariate graphs give us an accurate idea of the evolution of the system. Multivariate graphs confirm the change that has taken place in the system, shown by the univariate graphs.

The application of monitoring techniques has shown the NIR spectroscopy is a technique that shows no drift in the signal in the period studied. The signal may undergo certain specific changes, which will produce an error in the predicted value, but these changes are due to a cause that, once identified, can be avoided. If the cause cannot be identified and corrected there are tools that enable us to keep using the calibration model.

Careful planning of experimental work enables all the tools for carrying out a standardisation process to be available if the model is seen to be no longer valid. These techniques have proved to be an effective tool for correcting variations in the model over time. In the case studied, a technique as simple as SBC was sufficient for correcting them. Few samples are required, although given the low

experimental cost of measuring by NIR, it is thought to be better to analyse as many samples as practically possible.

Acknowledgements

The authors thank the economic support from the Spanish Ministry of Education, Culture and Sports (Project BQU 2000-1256) and the Transformadora de Propileno S.A. (Tarragona) who provide the samples and reference analysis. Santiago Macho would like to thank the Fundación Repsol and the University, Research and Information Society Department of the Generalitat of Catalonia for providing a doctoral fellowship.

References.

1 H. Martens and T. Naes, Multivariate calibration, Wiley, New York, 1991.

2 H. Lammers, M.P.B. van Uum and J.P. de Kleijn, Macromol. Chem. Phys., 196 (1995) 2029.

3 M.P.B. van Uum, H. Lammers and J.P. de Kleijn, Macromol. Chem. Phys., 196 (1995) 2023.

4 K.Aaljoki, H. Hukkanen and P. Jokinen, Process Control Quality 6 (1994) 125.

5 J. Lee and H. Chung, Vibrat. Spectrosc., 17 (1998) 193.

6 M. Defernez and R.H. Wilson, Anal. Chem., 69 (1997) 1288.

7 T. Kourti and J.F. McGregor, Journal of Quality Technology, 28 (1996) 409.

8 S. Macho, F. Sales, M.P. Callao, M.S. Larrechi and F.X. Rius. Appl. Spectrosc. 55 (2000) 1532.

9 E Bouveresse, D.L. Massart, Vibrat. Spectrosc. 11 (1996) 3.

10 E. Dreassi, G. Ceramelli, P.L. Perruccio and P. Corty, Analyst, 123 (1998) 1259.

11 E. Bouveresse, Ch. Casolino and D.L. Massart, Appl. Spectrosc., 52 (1998) 604.

12 R.W. Kennard and L.A. Stone, Technometrics, 11 (1969) 137.

13 W.A. Shewhart, Economic Control of Quality, Van Nostrand, Princeton, NJ, 1931.

14 M. Forina and C. Casolino, Química Analítica, 18 (1999) 49.

15 E. Bouveresse, C. Hartmann, D.L. Massart, I.R. Last, K.A. Prebble, Anal. Chem., 68 (1996) 982.

16 F. Sales, A. Rius and M.P. Callao, Talanta, 52 (2000) 329.

17 Y. Wang, D.J. Veltkamp and B.R. Kowalski, Anal. Chem., 63 (1991) 2750.

18 F. Sales, M.P. Callao and F.X. Rius, Analyst, 124 (1999) 1045.

19 F. Sales, M.P. Callao and F.X. Rius, Analyst, 125 (2000) 883.

20 B.M. Wise, N.L. Ricker, D.F. Weltkamp and B.R. Kowalski, Process Control Quality, 1 (1990) 41.

21 J.E. Jackson, G.S. Mudholkar, Technometrics, 21 (1979) 341.

22 McLennan F., Kowalski, B.R., Process Analytical Chemistry, Chapman & Hall, London, 1995

23 MATLAB. The Mathworks, Sout Natick, MA, USA.

5.3 AMPLIACIÓN DE RESULTADOS

Causa de la perdida de validez del modelo

En el artículo que se acaba de presentar se propuso que un desplazamiento en la posición de las longitudes de onda del instrumento podía ser la responsable de la pérdida de validez del modelo de calibrado. La estabilidad de las longitudes de onda del espectrofotómetro es reconocido frecuentemente como un factor crítico de la calidad de la calibración multivariante [1].

En el caso del instrumento que se utilizó en este trabajo (espectrofotómetro IA500, Bran-Luebbe), diariamente ser realiza un test que comprueba ya ajusta la posición de las longitudes de onda a partir de un análisis de un estándar de poliestireno (6.3 pag. 165). Este ajuste es capaz de realizar correcciones de ± 0.125 nm.

Para comprobar si un desplazamiento en la posición de las longitudes de onda podía ser la causa de la pérdida de validez del modelo, se realizó la siguiente comprobación. Se definió como espectro control un espectro tomado al inicio de experimento, cuando todo el sistema de medida funcionaba correctamente. Este espectro de control se comparará con otros espectros medidos posteriormente, tanto antes como después de detectarse el cambio instrumental. A estos espectros medidos posteriormente se les aplica un desplazamiento deliberado de las longitudes de onda, aplicando un ajuste polinómico por partes (utilizando una estrategia parecida al suavizado de Savitzki-Golay [2]. La comparación entre los espectros con el desplazamiento deliberado y el espectro de control se puede realizar por ejemplo a partir de la suma al cuadrado media de las diferencias (MSE) entre ambos espectros. En caso de no haberse producido ningún desplazamiento accidental, el MSE aumentará tanto si se desplaza el espectro a izquierda o derecha (Fig. 41a), en cambio, si el espectro había sufrido un desplazaminto accidental en

las longitudes de onda, el MSE disminuirá cuando el desplazamiento deliberado compense al accidental (Fig. 41b).

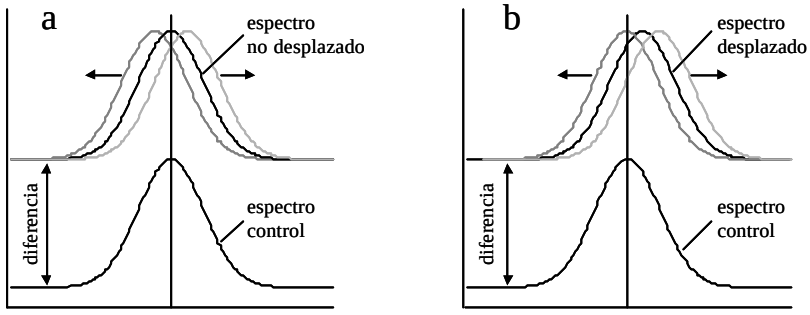


Fig. 41. Esquema de la comparación del espectro de control con otro espectro medido posteriormente, en caso de no haberse producido el desplazamiento (a) o en caso de sí haberse producido (b).

Así, para cualquiera de las nueve muestras analizadas durante los tres meses, se definió el espectro control como la media de los primeros 20 espectros, $\mathbf{r}_{1-20}$. Este espectro control se comparó utilizando la metodología indicada anteriormente con el espectro medio de las medidas 21 a 40, $\mathbf{r}_{21-40}$ (espectro antes del problema instrumental), y con la media de otros 20 espectros tomados después de haberse detectado el cambio en la predicción, por ejemplo $\mathbf{r}_{91-110}$ (espectro después del problema instrumental).

Los espectros $\mathbf{r}_{21-40}$ y $\mathbf{r}_{91-110}$ se desplazaron artificialmente en sus longitudes de onda entre −1 y + 1 nm, utilizando un polinomio de 3er grado, ajustado en una ventana de 6 longitudes de onda. Cada uno de estos espectros se comparó, por diferencia, con el espectro control y se calculó el error medio al cuadrado (MSE) en función del desplazamiento que se había provocado en el espectro. En la Fig. 42 se muestra el resultado de esta comparación para la muestra 15C.

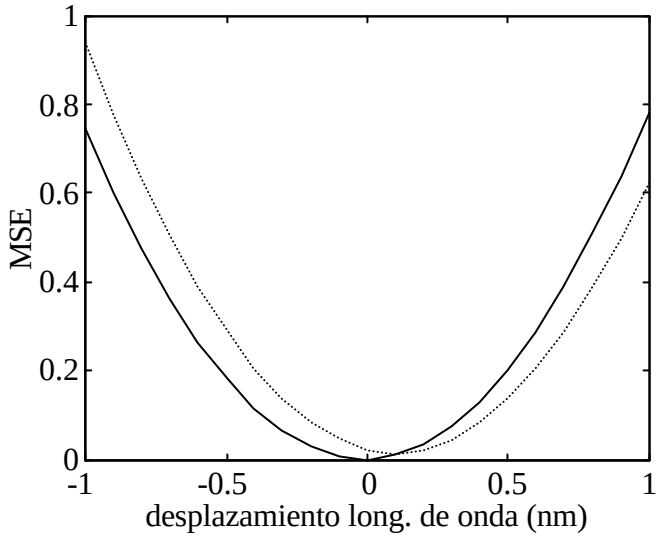


Fig. 42. Evolución del error medio al cuadrado (*mean square error*, MSE) en función del desplazamiento del espectro al comparar un espectro antes del cambio, $\mathbf{r}_{21-40}$ (—) o un espectro después del cambio, $\mathbf{r}_{91-110}$ (----), con el espectro de control, $\mathbf{r}_{1-20}$, para la muestra 15C.

Como se esperaba, en los espectro repetidos antes del cambio, $\mathbf{r}_{21-40}$, cualquier desplazamiento en las longitudes de onda hace aumentar el MSE, por lo que el mínimo de la curva se encuentra muy cerca de 0 (-0.01 nm). En cambio, para los espectros repetidos después del cambio este mínimo aparece mucho más alejado de 0, concretamente a 0.10 nm. Similares resultados se obtienen si se realiza el mismo cálculo con otras repeticiones del espectro o con cualquier otra de la nueve muestras de control.

Efecto de la estandarización en los gráficos de control

Como se discute en el artículo presentado, una estandarización tan simple como la corrección de la pendiente y el sesgo (*slope bias correcction*) tiene la ventaja de ser muy sencilla. Esta sencillez hace que sea muy fácil de aplicar en gran cantidad de situaciones, utilizando sistemas y *software* existente, aspecto muy conveniente en el campo industrial. Sin embargo, presenta el inconveniente de que al corregirse las predicciones del modelo y no los espectros, las técnicas de control multivariante presentadas dejan de ser útiles después de la estandarización. Esta limitación no se produce en el caso del segundo método de estandarización presentado, la estandarización directa por partes (*piecewise direct standardisation*, PDS). En este caso se corrige directamente el espectro de la muestra, por lo que en principio se podría seguir utilizando el control estadístico de procesos multivariante (MSPC).

En la Fig. 43 se muestra el resultado de los estadísticos T^2 y Q, para la muestra 15C, una vez que se había aplicado la estandarización PDS. Se puede observar que en el gráfico de T^2, las muestras después de la corrección sí que vuelven a valores de este estadístico bajo los límites de control. En cambio, en el caso del estadístico Q, las muestras, a pesar de ser correctamente predichas por el modelo, tienen en todos los casos unos valores de Q muy superiores a los que tenía la misma muestra antes de producirse el problema instrumental.

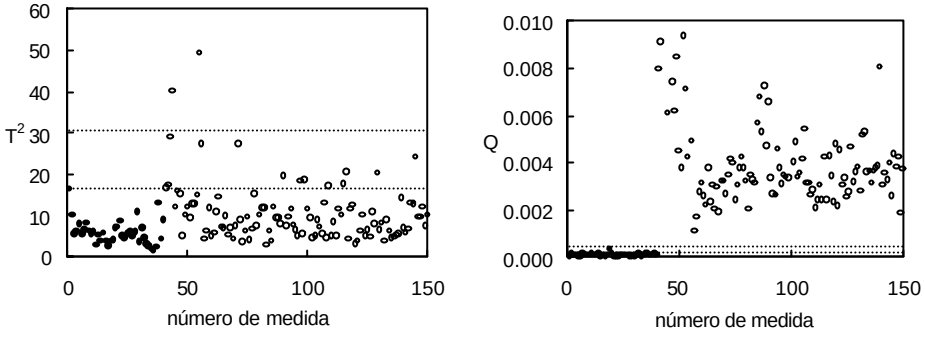


Fig. 43. Representación de los gráficos T^2 y Q de las muestras una vez que los espectros fueron estandarizados por la técnica PDS.

A pesar de que los modelos involucrados en la predicción del % de etileno (modelo PLS) y del control estadístico multivariante (PCA de la 40 primeras repeticiones) no son exactamente el mismo, se puede concluir que la corrección PDS afecta sobre todo a la variación incluida en el modelo (T^2) y corrige en menor medida la variación no incluida en el modelo (Q). Un resultado similar se hubiera obtenido de haber hecho el control de *outliers* en base al *leverage* y al residual espectral del modelo PLS: los valores de *leverage* de las muestras estarían dentro de los límites de calibración una vez estandarizados los espectros pero no así los de los residuales espectrales.

5.4 CONCLUSIONES

En este capítulo se ha puesto de manifiesto la importancia de establecer la validez del modelo con el tiempo, ya que es una característica muy importante en aplicaciones industriales. En determinados casos toda la cadena de producción puede depender de los resultados procedentes de la calibración multivariante, por lo que es de vital importancia detectar rápidamente los cambios que se producen, determinar si estos cambios afectan a la validez del modelo y si es así, es muy importante poder corregir estos cambios de la forma más rápida y sencilla posible.

Se ha propuesto una estrategia de control basada, por un lado en gráficos que representan una respuesta univariante, como sería el gráfico Shewhart de la predicción por el modelo de una muestra control, y gráficos basados en la respuesta multivariante de la muestra (el espectro), como sería la representación de los estadísticos T^2 y Q calculados a partir del espectro y un modelo PCA, que servirían para detectar cambios en las condiciones ambientales o en el sistema de medida, por ejemplo una variación en el espectrofotómetro como ha sido este caso.

El hecho de aplicar métodos de control basados en la medida de una muestra de control, y métodos de estandarización que requieren del análisis de una misma muestra antes y después de producirse la perdida de validez del modelo, hace que la estrategia presentada sea más adecuada en el caso de muestras estables, como sería el caso del polipropileno y su aplicación es menos viable en el caso de una menor estabilidad de las muestras, como sería por ejemplo el caso de las muestras de nafta, en la que la evaporación de los componentes más volátiles puede hacer que la composición varíe con el tiempo.

Se ha demostrado que si la experimentación se planifica correctamente, es posible aprovechar las medidas procedentes del control del modelo para realizar su corrección (estandarización), con lo que el modelo de calibración puede seguir siendo utilizado casi inmediatamente después de detectarse que ha dejado de ser válido. En una aplicación industrial, la eliminación de tiempos muertos (sin modelo útil) es imprescindible.

Se ha mostrado la utilidad de técnicas de estandarización muy sencillas, como la corrección de la pendiente y el sesgo (*slope bias correction*) y de otras más elaboradas como la estandarización directa por partes (PDS). En este segundo caso, como se corrige el espectro de las muestras, los gráficos de control pueden seguir siendo utilizados, aunque se ha mostrado que los espectros de las muestras después de la corrección sólo siguen bajo control estadístico en el estadístico T^2, mientras están fuera de control en el estadístico Q.

BIBLIOGRAFÍA

1 The American Society for Testing and Materials (ASTM), Practice E1655-00. *ASTM Annual Book of Standards*, vol. 03.06, West Conshohocken, PA, USA, **2001,** pag. 17.

2 A. Savitzky, M. J. E. Golay, *Anal. Chem. 36*, **1964**, 1267.

6. Outliers en predicción

6.1 INTRODUCCIÓN

Una vez establecido el modelo de calibrado multivariante, en su uso cotidiano para el análisis de nuevas muestras, que sería paralelo al proceso de control mostrado en el capítulo anterior, es inevitable que ser produzcan errores y fenómenos inesperados. Por ejemplo en el ámbito industrial son comunes los cambios en las materias primas, en las condiciones de operación de las unidades o en la actividad del catalizador. Por este motivo, la detección de observaciones anormales (*outliers*) es una etapa muy importante en la calibración multivariante [1].

La calibración multivariante a diferencia de los modelos univariantes ofrece herramientas para la detección de muestras discrepantes. Hay herramientas que se basan en medidas de la influencia de las muestras en el conjunto de calibración (*leverage*), residuales en la respuesta instrumental (o residuales espectrales) y residuales en la concentración, que han sido presentados en el apartado 2.3.4.2 de esta tesis y utilizados en el desarrollo de los modelos de calibración. Otras estrategias presentan el problema de la detección de *outliers* como un problema de clasificación [2]. También es posible el uso de los estadísticos T^2 y Q, presentados anteriormente en el control del modelo, ya que proporcionan una información similar a la medida del *leverage* y del residual en la respuesta instrumental y su uso en el ámbito industrial está muy extendido.

En el presente capítulo (apartado 6.2) se presenta el artículo *"Outlier detection in the ethylene content determination in propylene copolymer by near-infrared spectroscopy and multivariate calibration"* publicado en la revista *Applied Spectroscopy,* en el que se hace hincapié en la etapa de detección de *outliers,* en el modelo para la predicción del % de etileno en polímeros de EPR/PP (o polipropileno de impacto iPP).

A continuación se presenta una ampliación de los resultados obtenidos (apartado 6.3) así como las conclusiones (apartado 6.4).

6.2 OUTLIER DETECTION IN THE ETHYLENE CONTENT DETERMINATION IN PROPYLENE COPOLYMER BY NEAR-INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

Applied Spectroscopy, 2001, **55**, 1532-1536

S. Macho, F. Sales, M. P. Callao, M. S. Larrechi and F. X. Rius

Departament de Química Analítica i Química Orgànica.
Universitat Rovira i Virgili, Pl. Imperial Tàrraco, 1, 43005-Tarragona, Spain

Abstract

In this study, we employed multivariate control techniques to detect outliers in the determination of ethylene in impact polypropylene samples by near-infrared (NIR) spectroscopy and multivariate calibration partial least-squares (PLS). We also applied an algorithm which identifies those spectral variables responsible for the outlier behavior and that can indicate the source of this behavior.

The outliers in the prediction step may be due to three possible situations: errors associated with the prediction of analyte concentrations in samples that have the same characteristics as the calibration set, but that are beyond the concentration range; changes in the matrix composition; and instrumental errors.

We show that the proposed techniques make it possible to detect whether or not an analyte belongs to the reference set. In addition, we apply an algorithm that identifies the variables that cause outlier behavior and assigns them to a class.

Index Headings: NIR spectroscopy; Multivariate calibration; Outlier detection.
Received 8th February 2001, Accepted 8th June 2001

Introduction

The aim of this study is to apply multivariate control techniques to the detection sample outliers. We applied these techniques in the routine determination of ethylene in polypropylene samples, using near-infrared spectroscopy (NIR) and multivariate calibration.

Several studies in the literature have reported the use of NIR and multivariate calibration for determining the properties of polymers,[1] but we have found no references about its use as a routine laboratory technique. This may be because of the many sources of variability in the industrial process (i.e., changes in the process or in the source of the samples), which make it necessary to check whether the calibration model developed is suitable in new situations that were not taken into account during the development step. These new abnormal situations lead to unrepresentative measurements called outliers.

The most routine type of outlier samples are due to human or instrumentation experimental errors and changes in the sample matrix composition (quite common in industrial products) or in the concentration of the analyte of interest (i.e., out of the calibration range).

The problem of detecting outliers in prediction has been studied for some time. The parameters habitually used for outlier detection in prediction are based on leverage and X-residuals.[2,3] The leverage of an observation refers to the position of the observation in relation to others. The X-residuals reflect the lack-of-fit between the instrumentally measured data X and the resulting prediction ($\hat{\mathbf{X}}$) from the model, due to random noise in X, unmodelled interferences, unmodelled nonlinearities, etc.

One other strategy for detecting outliers is to consider them to be a problem of classification. The calibration samples are usually considered to be a class, and the

new samples are tested to determine whether they belong to this class. An example of this strategy is the use of soft independent modeling of class analogies (SIMCA)[4] for outlier detection.[5,6] It should be pointed out, however, that other strategies have been applied: the concept of centroid, the evaluation of the length of the data vector[7], the use of potential functions,[8] and genetic algorithms.[9] An extensive comparison of different approaches is presented by Walczak and Massart.[10]

In this study, we propose using the Hotelling T^2 and Q residual plots: the former is based on the Mahalanobis distance (MD) and is closely related to the leverage of the observations,[11] and the latter is based on the sum of squares of the residuals in X. We will establish the PCA model and its corresponding limits with the calibration samples and we will determine the T^2 and Q statistics for the unknown samples in such a way that they will be considered outliers if they are beyond the limits.

This approach has some advantages. First, the critical value above which a sample is considered an outlier is based on well-established statistical criteria. In the classical approach, which is based on residuals and leverage, the critical value is sometimes subjective (e.g., for the leverage, a value of 2 or 3 times the mean value of the calibration leverage is proposed as the limit). A second advantage is that the data are graphically presented in a sequential and user-friendly way as they are obtained, which enables the behavior of the samples over time to be observed. These kinds of graphs can be used with the classical parameters (leverage and sum-of-square residuals) but this is not a common industrial practice.

Finally, T^2 and Q graphics have been widely used as tools to monitor and control the quality of processes that generate multivariate data.[12,13] Therefore, the

potential industrial users of multivariate calibration models may be more familiar with these kinds of tools and results.

Likewise, we incorporated an algorithm to detect the variables responsible for the outliers[14] and to help account for why the sample behaves like an outlier.

It should be pointed out that these techniques are useful for solving outlier detection problems such as the ones proposed. A well-established outlier detection step with *user*-friendly tools, which make the tasks of interpretation and automatization easier and highlight abnormal situations, will encourage the use of instrumental multivariate techniques in industrial laboratories.

The proposed application begins by using a validated multivariate calibration model, which has been developed to determine ethylene in impact polypropylene samples (copolymers). The same polymer process also provided two extra data sets of polymer types that were produced during the manufacture of impact PP: the polypropylene homopolymer (H) and the ethylene-propylene random copolymer (R). The laboratory also has other models available for determining ethylene in polypropylene (PP) homopolymers and in ethylene-propylene random copolymers (EPR). The different samples have similar physical characteristics and cannot be distinguished by observing the recorded signal. Therefore, using the copolymer model to predict the ethylene content of an PP homopolymer or EPR sample may be a source of human error.

Theoretical basis

Hotelling T^2 and Q charts[15]

The results of a multivariate measurement of the training set samples can be condensed in a matrix **X** with dimensions *m* (number of samples in the training set) by *n* (variables). This matrix can then be reduced by principal components analysis

to a new matrix **T**, with *m* rows and *k* columns (number of factors), which contains the maximum information on the original data.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad [1]$$

where **T** is the matrix of scores, **P** is the matrix of loadings and **E** is the matrix of the residuals of the model.

For a sample *i*, the statistic Q_i is calculated from the residual $\mathbf{e}_i$ (row *i* of matrix **E**) and its transposed vector $\mathbf{e}_i^{\mathrm{T}}$ by:

$$Q_i = \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} \qquad [2]$$

The Q parameter indicates the lack of fit of the data to the PCA model, i.e., the causes of variability that are not included in the model.

The limit value Q_{lim} is calculated as follows:

$$Q_{lim} = \Theta_1 \left[\frac{c_a \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}} \qquad [3]$$

where h_0 and Θ_i are different combinations of the eigenvalues of the principal components not contained in the model, and c_a is the standard normal deviation for the level of significance, α, chosen[16].

From the measurement of a sample *i*, $\mathbf{x}_i$ (row *i* of matrix **X**), T_i^2 is calculated as:

$$T_i^2 = \mathbf{x}_i \mathbf{P}_k \, l^{-1} \mathbf{P}_k^{\mathrm{T}} \mathbf{x}_i^{\mathrm{T}} \qquad [4]$$

where $\mathbf{P}_k$ is the loading matrix of the model for *k* factors, and l is the diagonal matrix of the eigenvalues, which is given by:

$$\mathbf{X}^{\mathrm{T}} \mathbf{X} = \mathbf{P}_k \, l^{-1} \mathbf{P}_k^{\mathrm{T}} \qquad [5]$$

Hotelling's T^2 statistic includes the variation of the samples in the PCA model. The T^2 charts have the control limit T_{lim}^2 :

$$T_{lim}^2 = \frac{k(m-1)}{m-k} F_{k,m-k,a} \qquad [6]$$

where $F_{k,m-k,a}$ is the *F*-distribution value for k and m-k degrees of freedom and a given α. The samples with values beyond the limits indicate changes which are out of the ordinary. High Q values must be interpreted as increases in the data which is not explained by the PCA model, whereas the points which are beyond the T^2 control limit represent abnormal deviations of the samples inside the PCA model.

BESI algorithm.

Once the charts reveal an error in the system, the source of this fault should be detected; but this is not obvious in a multivariate context. To overcome this problem, backward elimination sensor identification (BESI) attempts to diagnose the unreliable sensors that produce variations in the measurement. This algorithm has been applied to the unmodeled disturbances by means of a Q-ratio.[14]

The procedure is applied to the outlying Q samples, in which each sensor is sequentially removed from the model and the anomalous sample. The Q_{-j} values of the detected sample without sensor j are compared to the corresponding upper limit value Q_{lim-j} as follows:

$$Q_{ratio-j} = Q_{-j} / Q_{lim-j} \qquad [7]$$

If the minimum Q-ratio is less than 1, the procedure stops and the sensor is identified. If not, the sensor with the smallest Q-ratio is removed and the procedure iterates until a sensor with a minimum Q-ratio less than 1 is obtained. This last sensor and all the previous ones with the smallest values are detected as the disturbed sensors for the sample.

When the most influential sensors are removed, the value of the numerator Q decreases until the Q-ratio is less than 1, i.e., the samples are in control. This method combines the identification capability of the univariate charts with the potential sensibility of the multivariate approach.

Experimental

Samples

177 samples of impact-resistant poly(propylene-ethylene) copolymer (impact PP) from TPD, Tarragona, Spain, were obtained over a 9-month period of production. This data set was split into two different sets: 80 samples for calibration purposes and 97 samples for validation. The samples are labeled with a C (copolymer).

A second set of 88 impact PP samples with the same characteristics is available. These samples behave abnormally in the calibration and validation step. An

analysis of the process showed that their spectra were acquired before the instrument was properly stabilized. These samples are labeled with a C'.

The third set of samples comprises 69 samples of polypropylene homopolymer labeled as H and the fourth set of 70 samples of ethylene-propylene random copolymer is labeled as R. Impact PP is produced by in situ copolymerization of propylene and ethylene in the presence of PP homopolymer.

The last three extra data sets are included in the study to show how the detection techniques behaved when applied to samples not considered in the set up step.

Instrumentation and software

Spectroscopic data were collected by diffuse reflectance on a Bran+Luebbe IA500 spectrometer, equipped with a rotating cup drawer. This system can scan a larger area of the sample, which improves the ability to reproduce the spectrum. The results were processed with calculation routines programmed with MATLAB (MathWorks, Version 4.0, 1993).

Results and discussion

To establish the T^2 and Q plots (and their limits), a PCA model was calculated with 80 impact PP samples (training set). Table I shows the retained and cumulated variance for each principal component (PC). The first four principal components retain 98.3% of the information, that is to say, practically all of it. The optimum number of PCs in a PCA is a critical parameter. There are several statistical tests that enable the optimum number of factors to be chosen objectively[17] but here we consider that the choice of the number of PCs is not such a critical step, because the T^2 and Q information complement each other

Table I. Percent variance captured of the PCA model

PC number	1	2	3	4	5	6
This PC (%)	74.6	15.7	5.9	2.1	1.1	0.1
Total (%)	74.6	90.3	96.2	**98.3**	99.4	99.6

Figures 1a and 2a show the T^2 and Q statistics of the training set samples. Two control limits were established for both statistics –95% and 99.8%. These values were chosen because they are similar to the control and warning lines used in univariate control. As an example, in Fig. 1c the leverage calibration of the samples is shown. The outlier limit is set at twice the calibration mean leverage value [2 (k+1)/m].

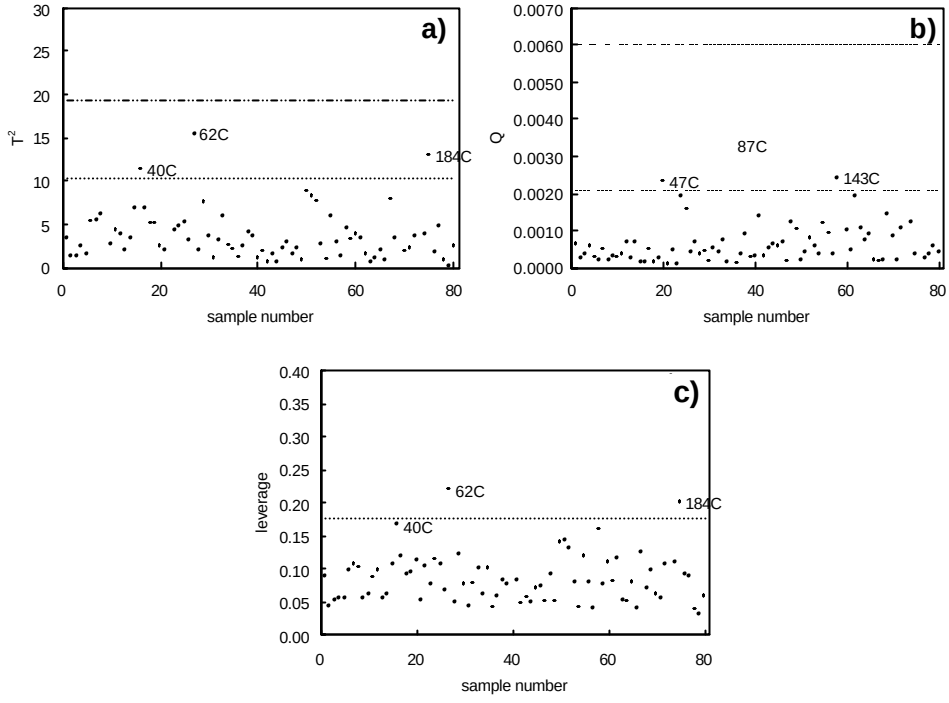


Fig.1. (**a**) T^2, (**b**) Q, and (**c**) leverage plots for the impact PP training samples.

The following out-of-control samples can be observed in Figs. 1a and 1b: no sample is above the control line (99.8%), but three samples (62C, 184C and 40C) are above the control line for the 95% limit for T^2, but in control for the Q statistic. Finally, three samples (47C, 143C and 87C) are in control for T^2 and are beyond the 95% limit for the Q statistic. As was stated in the Introduction, the leverage (Fig. 1c) gives information that is very similar to T^2. The only difference is sample 40C, which is slightly below the outlier limit.

The T^2 and Q values are not much above the limits, and these samples had not been detected as outliers when the calibration model was established and validated using standard outlier detection techniques of leverage, residual standard deviation in X, and the residual in y. Therefore we decided to use the samples to construct the control limits.

Once the T^2 and Q limits had been established, the values of these parameters were calculated for the four data sets described in the Experimental section. Figure 2 shows the T^2 and Q statistics for the impact PP test set samples (○).

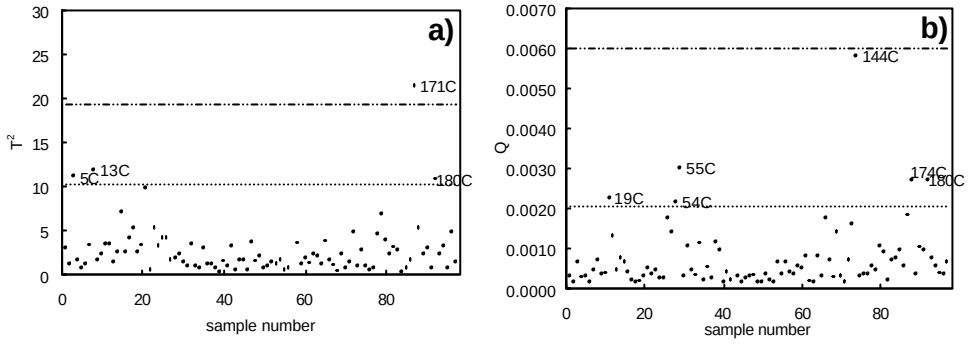


Fig. 2. (**a**) T^2 and (**b**) Q plots for the impact PP test set samples.

These samples behave in a very similar way to those of the training set. If a sample in one of the plots falls beyond the established limits, then the possibility of an anomalous situation must be considered. It should be pointed out that none of

the test samples gave a simultaneous warning in both of the plots studied. There are a variety of points which may be of interest to the user. A high value in the T^2 plot but not in the Q plot may be a warning that the calibration model is being extrapolated, while a high value in Q and not in T^2 may be because a variability not considered in the PCA model —but which has no great effect on it— is being incorporated into the signal recorded.

Figures 3a and 3b show the results for the remaining samples: PP homopolymer, EPR, and impact PP samples when the instrument was not correctly stabilized.

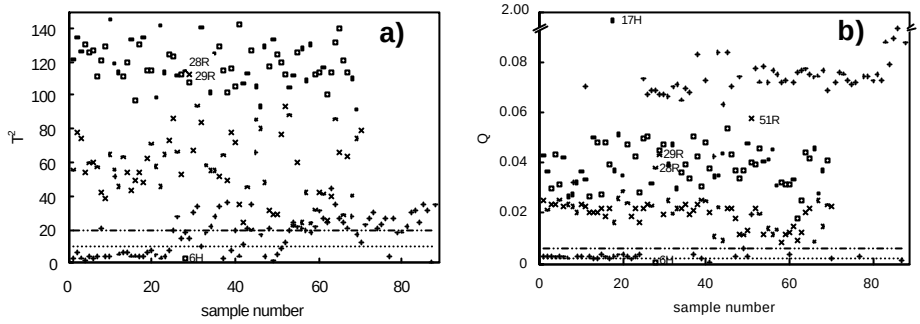


Fig. 3. (**a**) T^2 and (**b**) Q plots for the three extra data sets: PP homopolymer (□), EPR (x), and impact PP with nonstabilized instrument (+) samples.

Overall it can be seen that all the samples are some considerable distance from the limits; the lower their ethylene content, the greater the distance. Therefore, the samples of homopolymers are more different than the random polymers with higher amounts of ethylene, some of which can even be seen to be near the established limits. There are two different groups of nonstabilized samples, one within T^2 limits but outside Q limits, and the other outside both limits. This is interpreted as meaning that the instrumental variability that affects the response for

that set of samples is not constant and depends on when the spectrum was measured.

A detailed analysis of the plots provides additional information. For example, there is one homopolymer sample (17H) which is clearly different from all the other samples. It has exceptionally high Q values and normal T^2 values. The spectrum of this sample shows that there was a severe instrumental failure, which caused an abnormally low value of log (1/R) for one wavelength. The rest of the spectrum fits very well with the mean spectrum of the samples analyzed. In this extreme case, a brief visual analysis of the spectrum is sufficient to reject the analysis carried out

Likewise, the homopolymer sample (6H) is within the limits of the impact PP samples in both plots. A study was made of the sample and it was revealed that there had been a labeling error: the sample was not a homopolymer, but impact polypropylene.

Two of the random copolymer samples (28R and 29R) are mixed with the homopolymer samples. These samples are those that have the lowest ethylene content (0.5 and 0.9% ethylene, when the mean of the EPR samples have 4%). Therefore, it is normal that they behave in a way similar to the PP homopolymer (the ethylene content of which is not significant).

The user may also be interested in knowing the reason why a sample is an outlier. Therefore, the BESI algorithm was applied to the sets that had out-of-control samples in the Q statistic. The results in Table II show three error zones, which depend on the order in which the algorithm detects the variables.

Table II. Wavelengths detected as responsible for outlier behaviour

	1st error zone (nm)	2nd error zone (nm)	3rd error zone (nm)
PP homopolymer	1728	1844	1710
EPR copolymer	1728	1174	1846
Impact PP nonstabilized 1st type	1688	1742	-
Impact PP nonstabilized 2nd type	1688	1742	1356

The zones detected correspond to the frequencies associated with the vibrations of the first (1700-1800 nm) and second overtones (1150-1250 nm) of the methyl and methylene groups and the combination bands of the methylene groups (around 1400 nm). The homopolymer differs from the random copolymer in the second zone detected (1174 nm). This zone is associated with the second CH_2 overtone and will therefore enable the sample detected to be assigned as an outlier in its source group.

The most important sensors in the samples analyzed with the nonstabilized instrument are around 1688 nm, which is the zone with the steepest slope in the spectrum (side peaks). The following zones (1742 nm) correspond to similar points in other bands (side peaks). This may be because these samples have a slight shift in the x-axis. Small variations in the x-axis will be more critical in the spectrum in which there are considerable changes from one wavelength to the next (maximum slope points, side peaks). The difference between the two types of nonstabilized samples is that the second type has more error zones than the first (e.g. 1356 nm).

This information is important because the spectra will not give the users any indication that the conditions in which they are operating are below standard and that their results will therefore not be valid.

Figure 4a shows a characteristic spectrum for each of the groups studied. The zones detected by the BESI algorithm, which have been discussed above, have been indicated. Due to the similitude between samples and in order to help the sample identification, Fig. 4b shows the difference between the spectra of homopolymer PP and impact PP and EPR and impact PP.

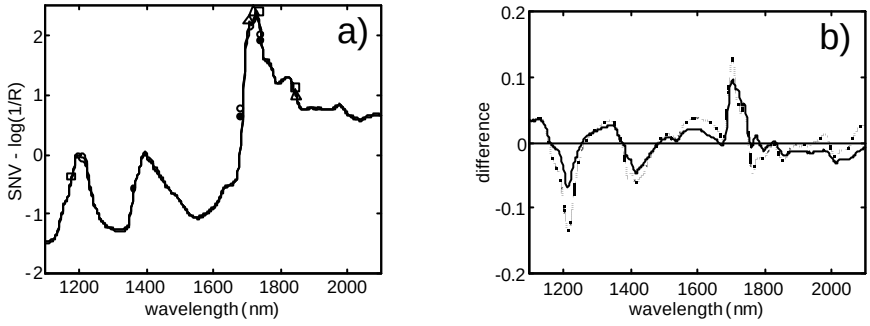


Fig. 4. (**a**) NIR reflectance spectra of the three polymer products: polypropylene, ethylene-propylene rubber, and impact polypropylene. The error zones for PP homopolymer (△), EPR (□), and impact PP nonstabilized 1st and 2nd conditions (○,●) are shown. (**b**) Difference spectrum between homopolymer PP and impact PP (⋯⋯) and between EPR (—) and impact PP.

Conclusions

As the T^2 and Q plots are complementary, a sample will be considered as, or suspected to be, an outlier if it appears in either of the plots because this means that the sample has information which is not the same as in the model, either in the part contained in the model (plot T^2) or in the part not contained (plot Q).

The graphs used in this study detect whether a sample is of the type used to set them up; therefore, reliable predictions are expected. In a few cases, when the value is close to the limits, a warning may be given, but this does not cause the prediction of analyte content in the samples to be abnormal. We believe that these

conclusions can be easily extrapolated to similar situations in which there are several calibration models set up to analyze samples of similar physical characteristics.

T^2 and Q statistics make it possible to use well-established statistical limits that are not available in other outlier detection strategies and that depend more on the analyst's skills.

The BESI algorithm can determine the variables that cause a sample to be detected as anomalous and, if there are sample sets of outlier samples available, as may be the case in industry, will guide users to the kind of problems that cause the abnormal behavior.

Acknowledgments

The authors are grateful for the economic support of the Spanish Ministry of Education, Culture, and Sports (project BQU 2000-1256) and the Transformadora de Propileno S.A. (Tarragona), which provided the samples and reference analysis. Santiago Macho would like to thank the Fundación Repsol and the University Research and Information Society Department of the Generalitat of Catalonia for providing a doctoral fellowship.

References

1 J. S. Lee and H. Chung, Vib. Spectrosc., **17**, 193 (1998).

2 H. Martens and T. Næs, *Multivariate Calibration* (John Wiley and Sons, New York, 1989), Chap. 5, pp. 291-293.

3 K. R. Beebe, R. J. Pell and M.B. Seasholtz, *Chemometrics a Practical Guide*, (John Wiley and Sons, New York, 1998), Chap. 5, pp. 292-304.

4 D. L. Massart, B. M. G. Vandegiste, S. N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: A textbook* (Elsevier, Amsterdam, 1988) Chap. 23, pp. 403-406.

5 B. Mertens, M. Thompson and T, Fearn, Analyst, **119**, 2777 (1994).

6 R. De Maesschalk, A. Candolfi, D. L. Massart and S. Heuerding, Chemom. Intell. Lab. Syst., **47**, 65 (1999).

7 W. J. Egan and S. L. Morgan, Anal. Chem., **70**, 2372 (1998).

8 D. Jouan-Rimbaud, E. Bouveresse, D.L. Massart and O.E. de Noord, Anal. Chim. Acta, **388**, 283 (1999).

9 B. Walczak, Chemom. Intell. Lab. Syst., **28**, 259 (1995).

10 B. Walczak and D. L. Massart, Chemom. Intell. Lab. Syst., **41**, 1 (1998).

11 R. De Maesschalck, D. Jouan-Rimbaud and D. L. Massart, Chemom .Intell. Lab. Syst., **50**, 1 (2000).

12 A. Nijhmis, S. De Jong and B.G.M. Vandengiste, Chemom. Intell. Lab. Syst., **47**, 107 (1999).

13 C. Wikström, C. Albano, L. Eriksson, A. Fridén, E. Johansson, A. Nordahl, S. Rännar, M. Sandberg, N. Kettanen-Wol and S. Wold. Chemom. Intell. Lab. Syst., **42**, 221 (1998).

14 C. L. Stork, D.J. Veltkamp and B. R. Kowalski, Anal. Chem., **69,** 5031 (1997),.

15 T. Kourti and J. F. MacGregor, J. Qual. Tech., **28**, 409 (1996).

16 J. E. Jackson and G. S. Mudholkar, Technometrics, **21**, 341 (1979),.

17 B. R. Kowalski and M. B. Seasholtz, J. Chemom., **5**, 129 (1991).

6.3 AMPLIACIÓN DE ALGUNOS ASPECTOS EXPERIMENTALES

Pretratamiento de los espectros

En los espectros originales (Fig. 44a, c y e) de los diferentes tipos de polímeros se observan variaciones en la línea base como consecuencia de la dispersión no específica de la radiación en la superficie de la muestra, que se produce debido a diferencias en el tamaño de las partículas entre muestras. Este efecto es más importante en el caso de los polímeros de EPR/PP y EPR y es menos importante en el caso del PP homopolímero.

En todos los casos el pretratamiento *standard normal variate* (SNV), aplicado entre 1100 y 2100 nm, corrige el problema (Fig. 44b, d y f). En estas figuras se puede observar como las diferencias en la línea base de los espectros, consecuencia de los efectos multiplicativos, se minimiza.

En el caso de los espectros del homopolímero de PP se observa como uno de los espectros sufrió un problema instrumental muy evidente, que consiste en una la lectura de log 1/R anormalmente baja a una única longitud de onda (un pico negativo). Un analista puede detectar este problema a partir de la simple inspección visual del espectro en el momento de su medida y corregirlo repitiendo la medida. A pesar de este hecho, se utilizó el espectro incorrecto para observar el comportamiento de las herramientas de detección de *outliers* con este tipo de muestras y de cara a la utilización del método de forma automática, sin la intervención directa del analista, como sería el caso de implementar este método en una aplicación en línea.

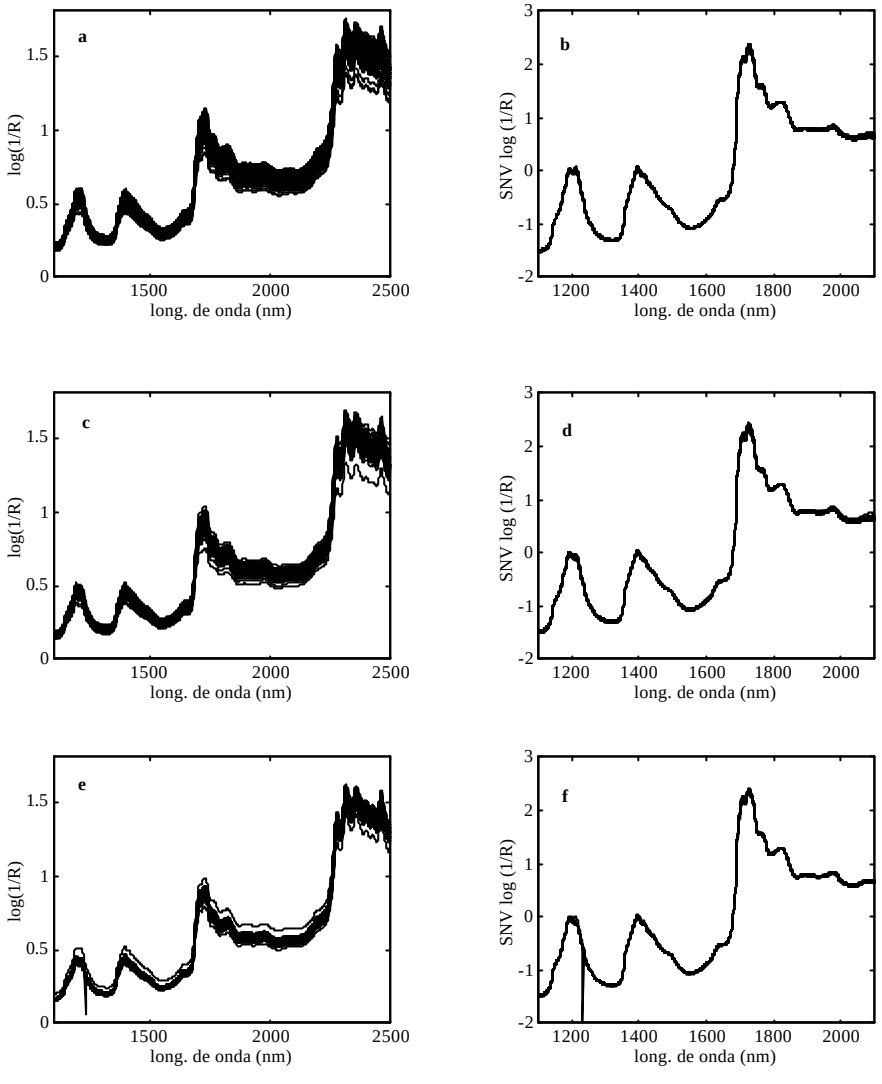


Fig. 44. Muestras de los diferentes polímeros de polipropileno. Espectros originales de EPR/PP (**a**), EPR (**c**) y PP homopolímero (**e**) y después de aplicarles la corrección standard normal variate (SNV) (**b**,**d** y **f**).

Exactitud de las longitudes de onda

El fabricante del espectrofómetro BL IA500 recomienda realizar dos test de control de calidad del funcionamiento del espectrofotómetro. Un test mensual que registra el nivel de ruido a partir de 10 medidas repetidas de una muestra estándar de un material basado en el teflón, con un nivel de absorbancia media, y un test diario que asegura la exactitud de la posición de las longitudes de onda a partir de la medida de una muestra estándar de poliestireno y por comparación de los picos encontrados con su posición teórica.

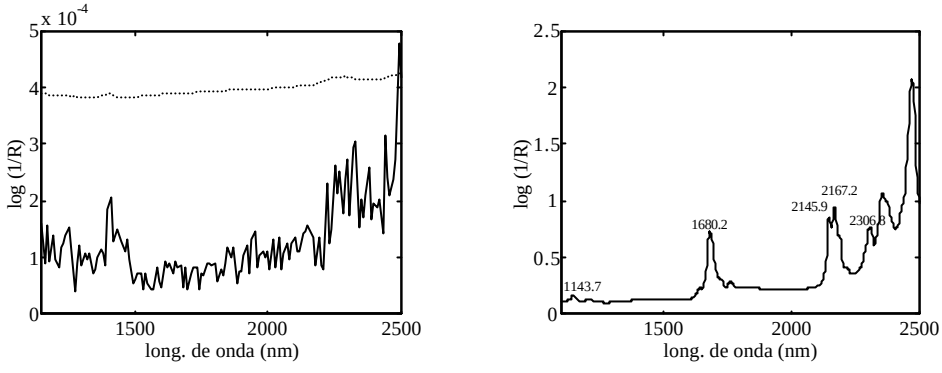


Fig. 45. Izq. Test de ruido: desviación estándar de la medida repetida (n=10) de un material estándar de teflón, así como el nivel de ruido permitido por el fabricante (---). Der. Espectro del material estándar de poliestireno para la medida la exactitud de la longitud de onda.

Al realizarse el test diario de exactitud de las longitudes de onda, en caso de detectarse diferencias en la posición de los picos superiores a 1 nm, el aparato puede realizar un ajuste automático de su escala de longitudes de onda. El fabricante recomienda que para realizar este test el instrumento debe haber sido encendido un mínimo de 10 minutos antes. Se detectó que, en algunas ocasiones, este tiempo debe ser sensiblemente mayor (por ejemplo en caso de largos periodos de inactividad del aparato) para asegurar una estabilidad de las longitudes de onda.

En estos casos, si se realiza el test de exactitud de las longitudes de onda antes de que el instrumento se haya estabilizado correctamente, se produce un error ya que el instrumento corrige su escala de longitudes de onda antes de que ésta sea estable. Este fue el problema instrumental que se produjo en el caso del conjunto de 88 muestras de copolímero EPR/PP medido en condiciones instrumentales no correctas.

6.4 CONCLUSIONES

En este capítulo se ha puesto de manifiesto la importancia de la detección de las observaciones discrepantes en la etapa de predicción de muestras desconocidas. El proceso de detección de las observaciones discrepantes no debe limitarse únicamente a señalar las muestras con un comportamiento diferente, sino que se debe extraer la máxima cantidad de información del proceso y siempre que sea posible se debe justificar el comportamiento anormal de la muestra, identificando la causa del problema.

Para facilitar la tarea de identificar las causas de los comportamientos discrepantes, hay que conocer los principios en los que se basan los estadísticos utilizados en la detección de *outliers*. En este caso han sido los estadísticos T^2 y Q, aunque podrían haber sido otros equivalentes (como el *leverage* y el residual espectral, por ejemplo). Cada uno de ellos señala errores de diferente origen y que pueden servir como una primera indicación del origen del comportamiento anómalo de la muestra.

Otra herramienta que ayuda a la interpretación del origen de las muestras *outlier* es el algoritmo BESI, que identifica las variables más importantes que han generado el comportamiento anómalo. En este caso este algoritmo ha sido especialmente útil en el caso del grupo de espectros de EPR/PP (o iPP) medido en

condiciones inadecuadas, ya que en base a los resultados del algoritmo BESI se pudo interpretar la causa del comportamiento discrepante de estos espectros.

BIBLIOGRAFÍA

1 H. Martens and T. Næs, *Multivariate Calibration*,Wiley, New York, **1989**, pag. 267.

2 B. Walczak and D.L. Massart, *Chemom. Intell. Lab. Syst*, *41*, **1998**, 1.

7. Evaluación de las posibilidades de esta metodología

7.1 INTRODUCIÓN

Llegados a este punto de la tesis doctoral, en base a la experiencia adquirida en su desarrollo, así como en los contactos que en este periodo ha habido entre el Grupo de Quimiometría y empresas del polígono químico de Tarragona, se constata que los métodos basados en la combinación de la espectroscopia NIR con las técnicas de calibración multivariante son muy atractivos en el ámbito industrial, debido a su rapidez de análisis, la posibilidad de determinar diferentes propiedades simultáneamente y a su versatilidad para aplicarse en análisis en línea.

Así, en el ámbito de la industria petroquímica ha tenido un fuerte impacto el éxito de la determinación en línea del índice de octano en gasolina mediante NIR frente al lento método tradicional que utiliza un motor de explosión. El éxito de algunas aplicaciones ha producido que en ocasiones se olviden algunos conceptos básicos de la metodología y que se considere la calibración multivariante una "caja negra" matemática, que extrae la información contenida en el espectro, y proporciona unos resultados sin que el usuario controle y comprenda muy bien cómo lo hace [1].

Por este motivo, en este apartado de la tesis doctoral aborda algunos planteamientos generales que se deben considerar al desarrollar aplicaciones multivariantes, tales como la existencia de una relación entre la propiedad de interés y el espectro NIR de la muestra, si se está trabajando en un intervalo de concentraciones amplio o restringido, qué factores afectan al error en predicción o la capacidad de controlar la validez en el tiempo del modelo. Sobre el aspecto de la validez en el tiempo se presentan técnicas de estandarización como las presentadas en el apartado 2.3.4.5, así como otras técnicas que intentan reducir el coste y esfuerzo del mantenimiento de los modelos, como son los modelos robustos o la actualización de modelos.

En el presente capítulo se presenta el artículo "*Near infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry*", publicado en la revista *Trends in Analytical Chemistry*, en el que se desarrollan las reflexiones que se han presentado en esta introducción, utilizando como ejemplos las aplicaciones desarrolladas para el análisis de nafta y polímeros, así como otras aplicaciones publicadas, en el campo de la industria petroquímica.

En el apartado siguiente (7.2) se presenta la publicación realizada y a continuación, en el apartado 7.3, se presentan las conclusiones de este trabajo.

7.2 NEAR INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION FOR THE QUANTITATIVE DETERMINATION OF CERTAIN PROPERTIES IN THE PETROCHEMICAL INDUSTRY

Trends in Analytical Chemistry, in press.

S. Macho*, M. S. Larrechi

Departament de Quimica Analítica i Q. Orgánica.
Universitat Rovira i Virgili.Pl. Imperial Tarraco, nº1. 43005 Tarragona

Abstract

Near infrared spectroscopy in conjunction with chemometric techniques allows on-line monitoring in real time, which can be of considerable use in industry. If it is to be correctly used in industrial applications some basic considerations should be taken into account. This is not always the case. This study discusses some of the considerations that would help evaluate the possibilities of applying multivariate calibration in combination with NIR in properties of industrial interest. Examples of these considerations are whether there is a relation between the NIR spectrum and the property of interest, what the calibration constraints are or how a sample-specific error of prediction can be quantified. Various strategies for maintaining a multivariate model after it has been installed are also presented and discussed.

Keywords: Multivariate calibration, NIR spectroscopy, Petrochemical industry, Partial least squares (PLS), Principal component regression (PCR).

Received 15 November 2001, Accepted 22th January 2002

1. Introduction

Near infrared spectroscopy is particularly popular for industrial applications because it can rapidly and non-destructively analyse samples with minimum preparation [1]. In recent years, several studies have described how it can be used with multivariate calibration techniques, principal components regression (PCR) or partial least squares (PLS) as an alternative to classic methods —which are generally complex and time consuming— for the quantitative determination of properties that are of interest to the petrochemical industry [2, 3, 4, 5]. The undeniable success of these techniques in such important industrial applications as the determination of the octane index of gasoline [6, 7, 8], together with the fact that sometimes the applications published supply limited information, has meant that the technique has been somewhat idealised. For example, many studies assess the goodness of the technique solely in terms of the overall error of the model (root mean square error of crossvalidation, RMSCV, root mean square error of prediction, RMSEP, etc.) obtained with a small sample set. A small —and, therefore, hardly representative— calibration/validation set may lead to over-optimistic results [9]. Along these lines and in an application of these techniques in the field of oil derivatives, Swarin [10] shows that a large calibration set (359 gasoline samples from 23 different cities) or a restricted set (26 gasoline samples all from the same city) can lead to considerable differences in the complexity of models for determining properties of commercial gasoline. Although other performance parameters or figures of merit such as selectivity, net analyte signal or sensitivity can be evaluated [11, 12, 13] to provide a better assessment of the model, in practice this is not often done.

All manufacturers of instruments also include software with their products. This enables users to handle the most common chemometric techniques (MLR, PCR and PLS) even though they may not be very familiar with the underlying mathematics.

On occasions this means that users are unaware of basic theoretical considerations [14, 15] that are of the utmost importance.

This study reviews some basic concepts about the problem that we are to deal with and their implications on the multivariate calibration model. We believe that these concepts should be considered when an application is to be set up. We ask the following questions: Can the property of interest be determined with NIR spectroscopy? Will the sensitivity of the technique be sufficient for us to work in a narrow interval of the property of interest? What sources of error will affect the result? How can these errors be quantified? On what aspects does the validity of the model over time depend? The answers to these questions will allow us to make an a priori assessment of the chances that the technique will be successful.

For the sake of example, we shall present two kinds of applications: the determination of hydrocarbons and density in naphtha samples, and the determination of ethylene and viscosity in polypropylene copolymers These applications are examples of two different problems: naphtha is the feedstock of the cracking process and naphtha analysis enables the process to be monitored whereas the physical and chemical properties of polypropylene must be analysed for the sake of quality control.

2. Relation between the property and the NIR spectrum

In the near infrared zone, intensity values are associated to the combination and harmonic bands of the C-H, N-H, O-H and S-H bonds. This suggests that it can be used to determine parameters that refer to the chemical composition of the sample such as hydrocarbons in naphtha [16] or the percentage of ethylene in samples of polypropylene [17]. In both cases it is quite possible that two samples with the same value of the property (concentration) will have the same spectrum. However, it is much more unlikely that this is so for physical and chemical properties such as the viscosity of a polymer, which is directly related to the length and orientation of

the polymer chain. Two samples of polypropylene with the same number of C-H bonds may have different viscosities.

Some studies [18,19] describe near infrared spectroscopy as a method for monitoring the polymerisation process, showing changes in the intensities of the harmonic and combination bands of ethylene and propylene. Viscosity is related to the length of the polymer chain, which can be monitored with NIR. This suggests that it may be possible to determine viscosity using spectral data recorded for the samples. Several studies have been published that describe quantitative applications for determining viscosity both in synthetic polymers [20, 21] and other types of samples such as cellulose [22]. In the latter case, however, a previous transformation of the signal (orthogonal signal correction, OSC) is required if a valid model is to be achieved. This correction removes the part of the spectra orthogonal (i.e. uncorrelated) to the property of interest.

An in-depth analysis of the property-spectrum relation will enable us to make an initial assessment of the difficulty of the application. The relation between the spectrum and the property is simple so the application will be easier to set up than if there were no direct relation.

3. What constraints does my problem have?

To set up an application based on modelling NIR spectra, a sample set that is representative of the problem must be collected. For these samples, the value of the property of interest is known because they have already been analysed by the reference technique. The analysed samples must cover both the concentration interval and all the other sources of variability in the system [23]. To achieve this aim systematically, some researchers have proposed experimental design strategies [23]. Usually, these strategies are only applicable when the samples can be prepared in the laboratory. However, laboratory samples may poorly approximate the real industrial process sample matrix, and current practice favours using real

process samples. In this way, the complex sample matrix is accounted for in the model. In industrial applications, it is too expensive or not possible to control the composition or other sources of variation (e.g. temperature) of real samples. Therefore if the sample composition variation is very restricted (or other sources of variation have not been taken into account), the calibrations developed may have poor predictive performance outside a narrow working range. In practice, the sampling stage should be extended when determining hydrocarbon content so that usual refinery events are taken into account: changes in the type of crude, importation of intermediate products, different operation modes of the units, different recipes depending on the product to be obtained (e.g. summer and winter gasoline), halts in production units, changes in the activity of the catalyst, etc. [2] To ensure that the calibration sample set is appropriate, sampling may have to take place over a period of several months of normal refinery operations. In other simpler synthetic processes, such as the determination of ethylene content or the viscosity of polypropylene copolymers, there are considerably fewer sources of variation that need to be monitored.

Fig 1a shows a problem of this type: the reference values versus the NIR predicted values of a model for predicting density in naphtha samples. The main group of samples has a density value between 0.65-0.73 and a few other samples (15) have density values above 0.75. Some small differences are appreciable in the samples' spectra, and a principal component analysis (Fig. 1b) shows that these samples (marked as + in the figure) are outside the main group of samples. After this had been noticed, a study of the samples revealed that they had been produced by a different refinery unit and that they were, therefore, different from the main sample group.

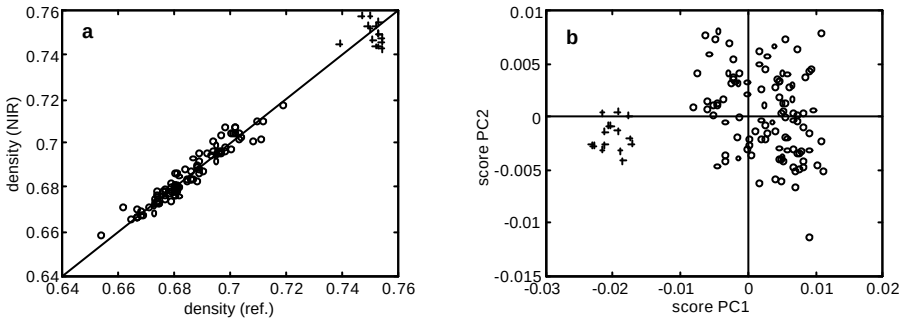


Fig. 1. a) PLS model for the determination of density in naphtha samples, (o) main naphtha samples and (+) secondary unit naphtha samples. With all samples the RMSECV=0.0038; using only main samples, RMSCV=0.0028. b) Principal component scores of the same naphtha samples.

Although the high density sample group seems to follow the general model tendency, its prediction errors are greater than those of the main sample group. The recommendation, therefore, is that the model should be used to predict only samples that are similar to the main sample group (o) and that a new model (with an appropriate number of samples) should be constructed to predict the density of the second type of samples (+). The samples that behave differently can be revealed in the calibration and prediction step with such outlier detection techniques as PCA representation, spectral residuals or leverage values.

Assuming that there is a relation between the NIR spectrum and the property of interest, another common problem arises when variations in the property of the samples produce spectral changes that are smaller than the reproducibility of the NIR measurement. A problem of this type may occur, for example, if one attempts to predict a property in a very narrow interval of values. Fig. 3 compares two models used to predict the composition (C_6 and C_{10} aromatic compounds) in 132 naphtha samples.

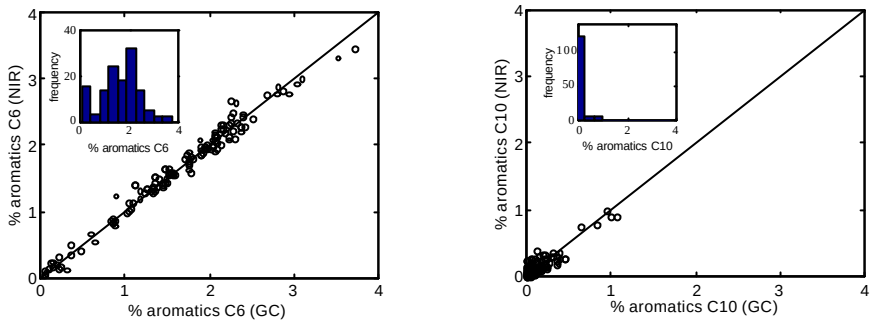


Fig. 2. a) PLS model for the determination of % C6 aromatics and histogram plot of the spread of reference values. b) PLS model for the determination of % C10 aromatics and histogram plot of the spread of reference values.

In refineries, multi-dimensional gas chromatography is used to determine aromatic compounds according to the number of carbon atoms [24] for 132 naphtha samples. It can be seen that the interval of the content of C_6 aromatics (benzene) is four times greater than the C_{10} aromatics. Moreover, most values for C_{10} aromatics are below 0.2% and only five samples have values near 1%. In the same conditions of instrumental reproducibility requirements, it will be more difficult to calibrate for the C_{10} aromatics because the interval of contents is very narrow, the samples are not evenly spread throughout the working range of contents (0-1%) and there is also the problem of their low concentration. In general, near infrared spectroscopy is not the best technique for quantifying analyte concentrations below 1% unless the analyte provides a strong signal that can be distinguished from the matrix spectrum [25]. It is possible to find good calibration models in the literature for determining the total content of aromatics in naphthas or similar products [26, 16] and even for some individual components such as benzene or toluene. However, it is not clear that NIR is an alternative to gas chromatography for determining low level aromatic compounds in naphtha.

4. What sources of error will affect the result?

In industry the maximum tolerable error is often known a priori, because (1) it is controlled by a standard or by legislation (e.g. content of aromatics and oxygenates in gasoline), (2) there is an economic interest (e.g. producers do not want their gasoline to have a MON that is higher than the minimum requirement) or (3) there is an operative interest (on-line process control, eg. naphtha steam cracking). For this reason it is important to consider all error sources involved in the multivariate calibration in order to evaluate the precision of the final result.

In several published applications, a single referenced mean value represents the predictive ability of the model. However, in practice, what people are often interested in is the individual assessment of error committed in each analysis. It seems logical to assume that the model developed will not always predict the property of interest with the same precision. Although there is not a generally accepted approach, several equations have been put forward for determining these individual confidence intervals [27, 28, 29] and they enable the individual uncertainty associated to each sample to be estimated. The following expression to calculate the prediction error variance was proposed by Faber and Kowalski [30]:

$$s^2_{\hat{y}_u} = \left(\frac{1}{N} + h_u\right)\left(s^2_e + s^2_{\Delta y} + \|\mathbf{b}_A\|^2 s^2_{\Delta X}\right) + s^2_{e_u} + \|\mathbf{b}_A\|^2 s^2_{\Delta X_u}$$

Where N is the number of calibration samples, and h_u is the leverage of the unknown sample. The leverage is a measure of how close the sample is to the multivariate model center and can easily be calculated from the calibration model. The variance of the reference measurement ($s^2_{\Delta y}$), the variance of the instrumental signal for calibration ($s^2_{\Delta X}$) and the new prediction object ($s^2_{\Delta X_u}$) can be obtained by repetitions. The square Euclidean norm of the regression vector obtained for A-

dimensional PCR or PLS is symbolized by $\|\mathbf{b}_A\|$. The variance of the unmodeled part of y (the model 'residuals') for calibration and prediction (new) objects are denoted by s_e^2 and $s_{e_u}^2$, which are determined by mean-square error of prediction (MSEC) in the absence of bias. For a detailed application of this formula in a practical case, see ref. [31], where this expression is used to provide a sample-specific uncertainty quantifying single oxygenate compounds in gasoline. Ether oxygenates have been used instead of organometallic additives to boost the octane number. In this application it is important to provide a sample-specific prediction error because the samples analysed are Standard Reference Materials (SRMs); therefore, the content oxygenate value and its uncertainty should be given. The use of the non-destructive NIR method reduces the uncertainty value of SRMs and this may help to improve the accuracy of oxygenate analysis in gasoline. The Faber and Kowalski expression generates good results even when the measurement error in the reference method is relatively large [32], a common situation in chemometrical applications. This reference method variance can be minimised if multiple measurements are made but, unfortunately, this is not habitual practice [14].

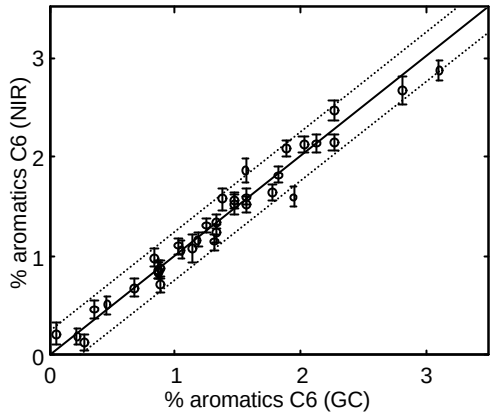


Fig. 3. Reference versus predicted NIR value of C6 aromatics model. Sample-specific error predictions have been added as error bars. Tolerable error based on reference method reproducibility is displayed as dotted lines.

Fig. 3 shows the reference versus NIR predicted values of a model for predicting benzene. It includes the confidence intervals of individual prediction, based on the Faber-Kowalski expression, and tolerable user error (0.25%) ,based on reference method reproducibility. These values can help to decide whether the model performs sufficiently well and to search for outlier samples.

5. Validity over time

Setting up an optimum NIR multivariate model is costly and, therefore, to make it worthwhile, it must be useful for as long as possible. However, the model may lose its validity and make erroneous predictions for several reasons: changes in instrument response (e.g. baseline offsets, changes in cell path length, cell positioning error, etc), ambient changes such as temperature or pressure, chemical-based variations due to process upsets, changes in raw materials, etc. and changes in physical sample conditions (e.g. viscosity). Failures in the model or instrument performance should be detected using outlier detection or statistical process control techniques [33] (for example, monitoring a stable sample or periodically comparing NIR predictions with a reference method value).

To overcome the loss of validity of the model without a complete recalibration procedure, several strategies of correction have been proposed. One group of techniques uses the calibration model to transform the measured spectra, the model's regression parameters or the predictions (e.g. bias/slope correction [34] or piece wise standardisation [35]). These techniques are useful when there is instrumental drift but requires measures of a subset of samples before and after the change. When these data are not available, e.g. in the case of non-stable samples, this approach cannot be applied. The model for controlling the percentage of ethylene in polypropylene polymers, which is one example of a stable sample, was corrected after an instrumental drift due to a change in the wavelength position (Fig. 4). The model predictions were corrected using a simple slope/bias correction

that enabled the same calibration model to be used without further adjustment for three months.

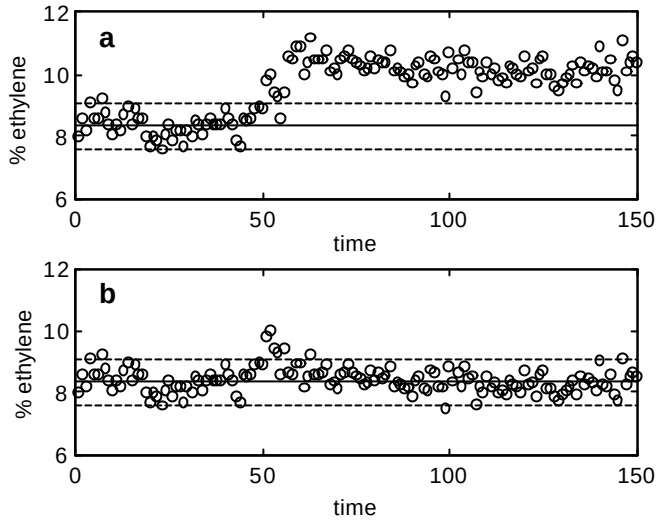


Fig. 4. Three-month polypropylene control sample prediction: a) without model correction and b) with simple slope bias correction after time 55.

In industrial applications, these kinds of techniques are useful for transferring a model developed on a laboratory instrument to an instrument in the field or for transferring models between different industrial sites. However, in the industrial environment, some ambient conditions or physical samples can change continuously and have a strong influence on model predictions, e.g. temperature, pressure or viscosity. As these changes can be constant and random, strategies of the first group cannot be applied, e.g. a model built to work at 50ºC can be corrected to work at 65ºC, but if the sample temperature increases again (e.g. 80ºC) the corrected model will also fail. To overcome this problem, a second group of techniques aims to make robust multivariate models insensitive to instrumental changes, ambient variations, etc. Robust multivariate models are easy to maintain and require infrequent recalibration [36]. In practice, model selectivity can be improved by data preprocessing, including measurement conditions in the

calibration model or using robust multivariate calibration techniques [37]. Thus, Swierenga *et al.* [37] studied the effect of temperature on determining the density of heavy oil products. Robust models are achieved by using global calibrations (i.e. the effect of temperature is implicitly included in the model because the samples are measured at different temperatures) and techniques based on variable selection (i.e. the model uses only a subset of spectral values, which are not sensitive to the changing conditions). The results of the variable selection technique were slightly better but the disadvantage is that expertise and specific software is needed. As far as variations in temperature are concerned, Wülfert et al. [38] found that various strategies for explicitly including temperature in the model (e.g. a new X variable) did not make the calibration models better than the global models, which included temperature implicitly. In the industrial situation, if temperature fluctuations cannot be avoided, the construction of a global model covering different temperatures may be the most practical solution.

In some situations a third group of techniques can be applied. If new samples contain an unmodeled chemical component that invalidates the existing model, full recalibration can be avoided if the model is updated with additional calibration samples. This situation is very common in the petrochemistry industry where changes in the type of crude, the importation of new intermediate products, halts in production units, etc. [2] can introduce chemical variability that is not accounted for in the original model. Moreover, in practice, it is usual to have only a few calibration samples that contain the new chemical component and their inclusion in the calibration model does not perceptibly improve the predictions. This problem can be overcome by assigning higher weights to samples with the unmodeled chemical. A very simple weighting method is to include two or more copies of each sample when constructing the updated model. Stork et al. [39] present a more formal weighting criterion based on leverage and apply it to predict a physical property in gasoline samples collected over the course of a year's refinery production. The author uses 1 to 11 new samples to update the model and states

that the weighting procedure is particularly useful when noisy instrumental responses are used. One empirical approach to model updating [40] is reported in a nine-month application to predict the octane number of naphtha in a petrochemical refinery. This updating scheme consists of a simple process of periodically introducing new samples into the calibration set and removing old spectra samples if the software boundaries are reached.

6. Conclusions

Near infrared spectroscopy with multivariate calibration is a very attractive technique for industry. Whether it is successful or not will depend on careful planning and evaluating all the aspects involved in practical situations. A relation between the spectrum and the property of interest should exist, and an *a priori* study of this relation can determine the difficulties of setting up a calibration model to predict the property. When a strong non-linearity relation is expected, various non-linear modelling techniques, such as non-linear PLS, artificial neural networks or local regression may give better results.

Calibration samples must be designed appropriately. All usual sources of variation should be included in the model and samples should be monitored for cluster presence. If some groups of samples behave differently, then some sources of variation may not have been included in the model. A very narrow and unevenly spread composition range is another example of inadequate calibration design.

The results and interpretation of NIR calibration can be improved if the error prediction of the model is expressed using sample-specific error expressions and not only an overall model error of prediction. Thus the multivariate NIR model acts more like univariate models which have well known prediction intervals.

One very important point when applying NIR in industry is that it must be regularly maintained. In this paper, we have reviewed techniques such as

transformations, robust models and model updating which can reduce the cost and time of maintenance.

Acknowledgements

The authors thank the Spanish Ministry of Education, Culture and Sports for economic support (project BQU 2000-1256) and Repsol Petróleo S.A. (Tarragona), who provided the samples and reference analyses. Santiago Macho would like to thank the Fundación Repsol and the University, Research and Information Society Department of the Autonomous Government of Catalonia for providing a doctoral fellowship.

References

1 C.M. Henry, Anal. Chem. News & Features (1999) 625A.

2 A. Espinosa, D. Lambert, M. Valleur, Hydrocarbon Processing 74 (1995) 86.

3 J.P. Coates, Spectroscopy 9 (1994) 36.

4 C.E. Miller, B.E. Eichinger, Appl. Spectrosc. 44 (1990) 887.

5 G. Lachenal, Vibrat. Spectrosc. 9 (1995) 93.

6 D. Lambert, A. Martens, E.P. 0 285 251 A1 (1988).

7 J.J. Kelly, C.H. Barlow, M. Jinguji, J.B. Callis, Anal. Chem. 61 (1989) 313.

8 M.S. Zetter, B.A. Politzer, Hydrocarbon Processing 72 (1993) 1.

9 H.A. Martens, P. Dardenne, Chemom. Intell. Lab. System 44 (1998) 99.

10 S.J. Swarin, Ch. A. Drumm, Spectroscopy 7 (1992) 42.

11 J.H. Kalivas, P.M. Lang, Chemom. Intell. Lab. System 32 (1996) 135.

12 N.M. Faber, A. Lorber, B.R. Kowalski, J. Chemom. 11 (1997) 419.

13 N.M. Faber, B.R. Kowalski, J. Chemom. 11 (1997) 181.

14 R. DiFoggio, Appl. Spectrosc. 54 (2000) 94A.

15 K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics a Practical Guide, Wiley, New York, 1998.

16 S.M. Maggard, U.S. Pantent 5 349 188 (1994).

17 J.S. Lee, H. Chung, Vibrat. Spectrosc. 17 (1998) 193.

18 M. Buback, B. Huchestein, V. Leinos, Makromol. Chem. Rapid. Commun. 8 (1987) 473.

19 C. Tosi, A. Pinto, Spetrochim. Acta, 28 (1972) 585.

20 C. Zhu, M. Heiftje, Appl. Spectrosc. 46 (1992) 69.

21 M. P.B. van Uum, H. Lammers, J.P. de Kleijn, Macromol. Chem. Phys. 196 (1995) 2023.

22 S. Wold, H. Antti, F. Lindgren, J. Öhman, Chemom. Intel. Lab. Syst. 44 (1998) 175.

23 T. Næs, Isaksson, Appl. Spectrosc. 43 (1989) 328.

24 J. Beens, U.A.Th. Brinkman, Trends Anal. Chem. 19 (2000) 260.

25 T. Hirschfeld, Anal. Chem. 56 (1984) 933.

26 J.J. Kelly, J.B. Callis, Anal. Chem. 62 (1990) 1444.

27 A.J. Hardy, W. Wegscheider, S.J. Haswell, P.A. Hailey, Analyst 120 (1995) 1875.

28 S.De Vries, C.J.F. Ter Braak, Chemom. Intell. Lab. Syst. 30 (1995) 239.

29 M. Hoy, K. Steen, H. Martens, Chemom. Intell. Lab. Syst. 44 (1998) 123.

30 N.M. Faber, B. R. Kowalski, Chemom. Intell. Lab. Syst., 34 (1996) 283.

31 N.M. Faber, D.L. Duewer, S.J. Choquette, T.L. Gree, S.N. Chesler. Anal. Chem. 70 (1998) 2972.

32 N.M. Faber, Chemom. Intell. Lab. Syst., 52 (2000) 123.

33 E.L. Grandt, R.S. Leavenworth, Statistical Process Control, McGraw-Hill, 1988.

34 E. Bouveresse, C. Sterna, J.L. Linossier, D.L. Massart, Analusis 24 (1994) 394.

35 Y. Wang, D.J. Veltkamp, B.R. Kowalski, Anal. Chem., 63 (1991) 2750.

36 P.J. Gemperline, Chemom. Intell. Lab. Syst., 39 (1997) 29.

37 H. Swierenga, F. Wülfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens, Anal. Chim. Acta, 411 (2000) 121.

38 F. Wülfert, W.T. Kok, O.E. de Noord, A.K. Smilde, Chemom. Intell. Lab. Syst., 51 (2000) 189.

39 C.L. Stork, B.R. Kowalski, Chemom. Intell. Lab. Syst., 48 (1999) 151.

40 M.V. García-Mencía, J.M. Andrade, P. López-Mahía, D. Prada, Fuel, 79 (2000) 1823.

7.3 CONCLUSIONES

En este trabajo se ha querido dejar constancia de la innegable utilidad que tiene la espectroscopia de infrarrojo cercano en combinación con la calibración multivariante. Sin dejar de reconocer este hecho, es importante que el método se aplique correctamente, para lo que hay que conocer algunos aspectos básicos de la técnica.

Algunos de estos aspectos básicos son muy simples, como la necesidad de la existencia de una cierta relación entre el espectro y la propiedad de interés, otros son más complejos y difíciles de controlar, como un buen diseño del conjunto de muestras de calibrado, o los factores que afectan al error de predicción. En cualquier caso, sean o no sencillos de controlar, es importante la reflexión previa sobre los diferentes aspectos del problema, lo que permitirá obtener resultados más satisfactorios.

Otros aspectos muy importantes en las aplicaciones industriales de la calibración multivariante son: el tiempo durante el cuál el modelo resulta útil, cómo se puede extender este tiempo de vida útil o cómo se puede corregir en caso de presentarse fuentes de variación inesperada. En el artículo se han presentando algunas de las estrategias generales que se pueden aplicar.

BIBLIOGRAFÍA

1 R. DiFoggio, *Appl. Spectrosc. 54*, **2000**, 94A

8. Análisis espectral para la selección de variables

8.1 INTRODUCCIÓN

Como ya se ha se discutido en el capítulo 5, las metodologías analíticas basadas en métodos multivariantes están sometidas a variaciones instrumentales que es necesario controlar y corregir para asegurar la validez de las predicciones. Los modelos multivariantes que utilizan datos espectroscópicos habitualmente están construidos utilizando todo el intervalo de longitudes de onda disponible. En esta situación, con un número de longitudes de onda grande, se puede mejorar y simplificar el modelo reduciendo el número de longitudes de onda utilizadas. La simplificación del modelo también puede aportar la ventaja adicional, de hacerlo más robusto, y por tanto menos sensible a variaciones instrumentales.

En este último capítulo, se ha abordado la selección de variables en calibración multivariante con el objetivo de incidir en la facilidad de interpretación de los modelos y de su robustez. Por otra parte, no se espera que la selección de variables altere las conclusiones generales sobre la aplicación de estas metodologías al análisis de las muestras estudiadas.

Con este objetivo, en el presente capítulo se discute una estrategia de selección de longitudes de onda basada en métodos de interpretación espectral, como la identificación de bandas de absorción de grupos funcionales en el espectro original o en la segunda derivada, o la interpretación de los coeficientes asociados a los *loadings* de una descomposición en componentes principales.

Esta estrategia de selección se discute para el caso de la determinación de parafinas, aromáticos y naftenos en muestras de nafta. Los resultados se describen en el apartado 8.2, donde se presenta el artículo "*Wavelength selection of naphtha near infrared spectra using conventional spectral analysis and PCA*" enviado para publicación en la revista *Talanta*. A continuación se presentan las conclusiones de este trabajo (apartado 8.3).

8.2 WAVELENGTH SELECTION OF NAPHTHA NEAR INFRARED SPECTRA USING CONVENTIONAL SPECTRAL ANALYSIS AND PCA.

Talanta, submited

S. Macho and M.S. Larrechi

Departament de Química Analítica i Química Orgánica.
Universitat Rovira i Virgili.Pl. Imperial Tàrraco, 1, 43005-Tarragona, Spain

Abstract

Near infrared spectroscopy (NIR) and multivariate calibration have been progressively used in several industrial applications because it provides fast analysis and is suitable for on-line use. Multivariate calibration methods such as principal components regression (PCR) or partial least squares (PLS) regression can use the whole spectrum. These methods have several advantages (e.g. error-averaging on signal, outlier detection, etc), but wavelength selection can improve the performance of the calibration model by excluding the spectral regions that have no useful information and only introduce noise to the model. This study uses classical spectral analysis methods such as second derivative and chemometrics methods such as principal components analysis (PCA) to select the best wavelength combination using chemical knowledge of samples. This wavelength selection is used to construct PLS models to determine the percentages of aromatics, naphthenes and paraffines in naphthta samples. The errors in prediction are very similar to those with the full PLS models: aromatics, full PLS model RMSCV=0.42%, selected wavelengths PLS, RMSCV=0.38%; naphthenes, full PLS model RMSCV=1.12%, selected wavelengths PLS, RMSCV=1.09%; paraffins, full PLS model RMSCV=0.93%, selected wavelengths PLS, RMSCV=0.99%.

Index Headings: Near infrared spectroscopy; Partial Least Squares (PLS); Wavelength selection; Principal Components.

Introduction

Near infrared spectroscopy (NIR) and multivariate calibration are very important in industrial applications [1-3], food analysis [4] and pharmaceutical analysis [5]. They can rapidly and non-destructively analyse samples with minimal preparation and are an alternative to traditional methods, which are generally complex and time consuming.

The simplest way to build a multivariate calibration model is to use the entire spectrum, which is easily collected by modern analytical instruments, and calibration methods such as partial least squares (PLS) regression, which can deal with over-determined systems [6,7]. Using a large number of wavelengths has an error-averaging effect that is beneficial for the accuracy and precision of the model. However, a wavelength selection procedure improves the performance of the calibration model by excluding the spectral regions that do not contain information that is correlated with the constituent of interest.

To select the wavelength several criteria have been used, most of which were related to selectivity [8-12]. These criteria use quantitative measures of concepts that are very simple in univariate calibration but which are not so straightforward when the model is multivariate. In this study we used classical spectral analysis methods like second derivative [13,14] and chemometrics methods like principal components analysis (PCA) [15-17] to select an optimal subset using the chemical knowledge of the sample.

We considered the NIR spectra of one hundred naphtha samples to determine the hydrocarbon family content i.e : (P) paraffins (or normal alkanes), (N) naphthenes (cycloalkanes) and (A) aromatics. To evaluate our selection, we compared the predictions from the PLS model with different criteria with those from a full spectrum PLS model.

Experimental

Samples and Analysis

One hundred samples were received in aluminium bottles from a naphtha feedstock over a period of several months (REPSOL Petróleo, Tarragona, Spain) and stored at 5 ºC. The samples were characterised by a multi-column gas chromatography system. The range of concentrations of the samples in each hydrocarbon family is shown in Table 1.

Table 1. PNA composition intervals of samples.

	Composition interval (%)
naphthenes	10.7 – 32.0
paraffins	55.8 – 86.9
aromatics	3.5 - 11.8

Instrumentation and software

Spectra between 1100 nm and 2230 nm were collected by diffuse reflectance on a Bran+Luebbe IA500 spectrometer using a temperature-controlled cell at 15º C. The results were processed with calculation routines programmed with MATLAB (MathWorks, Version 4.0, 1993).

Results and discussion

Spectral analysis

The naphtha samples under study were characterised by their total paraffin content, which ranged from 86.9% to 55.8%. The samples that were low in paraffinic content were therefore high in naphthene and aromatic content. Fig. 1 shows the spread of the samples in terms of the total paraffin content and the aromatic and naphthenic content.

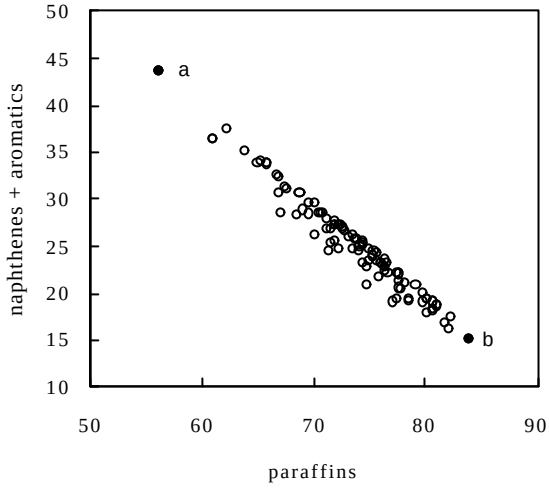


Fig. 1. Spread of naphtha samples according to hydrocarbon type. Sample b corresponds to a naphtha with a high paraffin content. Sample a corresponds to a naphtha with high naphthene and aromatic content.

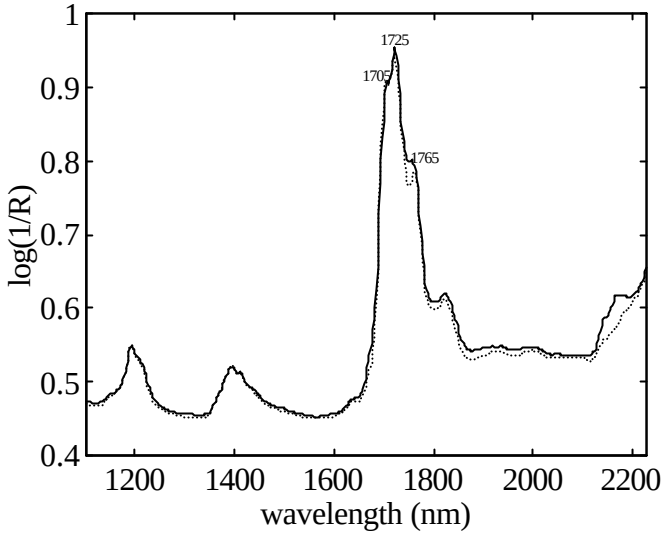


Fig. 2. NIR spectrum between 1100 and 2228 nm of the two extreme naphtha samples. a) high naphthene and aromatic content (—) and b) high paraffin content (---).

Fig. 2 shows the NIR spectrum (between 1100 and 2228 nm of the two extreme samples), the sample with the highest paraffin content (b), and the sample with the lowest paraffin content and therefore the highest naphthene and aromatic content (a).

There are two relatively weak bands, one around 1200 nm which is due to the second overtone of the C-H bond of CH_3 and CH_2 groups [4,18], and another band around 1400 nm which is due to combination bands of both groups and C-H aromatic. In this spectrum zone, the most paraffinic sample and the aromatic sample are not significantly different. Between 1600 and 1850 nm there is a strong band with a maximum at 1725 nm, and two shoulders at 1705 nm and 1765 nm. This spectrum zone contains referenced [4,18] bands of CH_3, CH_2 and CH groups. The difference between the peaks at 1765 nm (second overtone of CH_2) is greater than the difference between the peaks in all other zones in the spectrum. At 1824 nm there is a weak broad band that is not assigned. Beyond 2100 nm the two samples behave very differently.

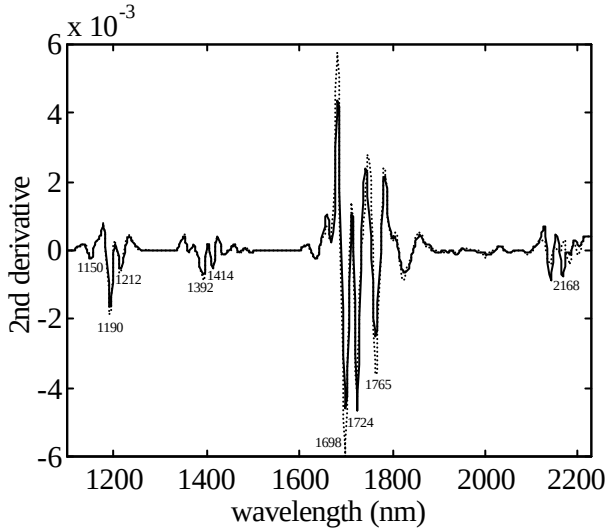


Fig. 3. Second derivative spectra of the two samples of naphtha: (---) high paraffin content and (—) high aromatic content.

Fig. 3 represents the second derivative of the spectra in Fig. 2. The quadratic background in the original spectra has been removed to reveal peaks that were not clearly identified in the original spectra. There are three very similar weak downward peaks in both samples. The first is at 1150 nm and, despite the contribution in this zone of the second overtone of aromatic C-H bond at 1145 nm [18], this corresponds to the second overtone of bands of CH_3 [4,18]. The second is at 1190 nm. This is also referenced as the second overtone of CH_3 [18], and is stronger in the paraffinic sample. The third is at 1212 nm (the second overtone of CH_2). Other low intensity downward peaks appear at 1392 nm and 1414 nm and are assigned to combination bands of the CH_2 [4,18].

Bigger downward peaks appear at 1698 nm, 1724 nm and 1765 nm. At 1698 nm (the first overtone of CH_3), the peak is stronger in the paraffinic sample. At 1724 nm, (the first overtone of CH_2 [18]) the peak is stronger in the naphthenic-aromatic sample, possibly due to overlapped CH group absorptions. At 1765 nm (the first overtone of CH_2), the peak is stronger in the paraffinic sample.

There is a small downward peak at 1824 nm that is slightly more intense in the paraffinic sample than in the aromatic one. There are three sharp peaks at 2140 nm, 2168 nm and 2182 nm. The peak at 2168 nm, which is clearly bigger in the aromatic sample, may be due to the CH aromatic [19].

After this analysis, the most significant bands for determining aromatics are 1150 nm, 2168 nm and 1724 nm, and for determining paraffin content the most significant bands are 1190 nm, 1824 nm, 1698 nm and 1765 nm.

PCA Analysis

The NIR second derivative spectra of the samples of naphtha, were mean-centered and subjected to PCA. This procedure detects independent factors of variation or principal components (PC) that include the information in the overall

sample set. Each factor may be related to one variability source and not necessarily to a chemical constituent of the sample (i.e. two chemical components that vary simultaneously in the sample are covered by only one PC). The first three principal components account for 92 % of the total variance (the first principal component accounts for 59.1%, the second for 24.2% and the third for 8,7 %). Loading explains the relationship between the original variable and the PCs. Variables with a high degree of systematic variation have a large absolute loading. The first PC lies along the direction of maximum variance in the data set, so if we plot loading for PC 1 (Fig.4a), we can see the spectral variables that contribute to the most important variability.

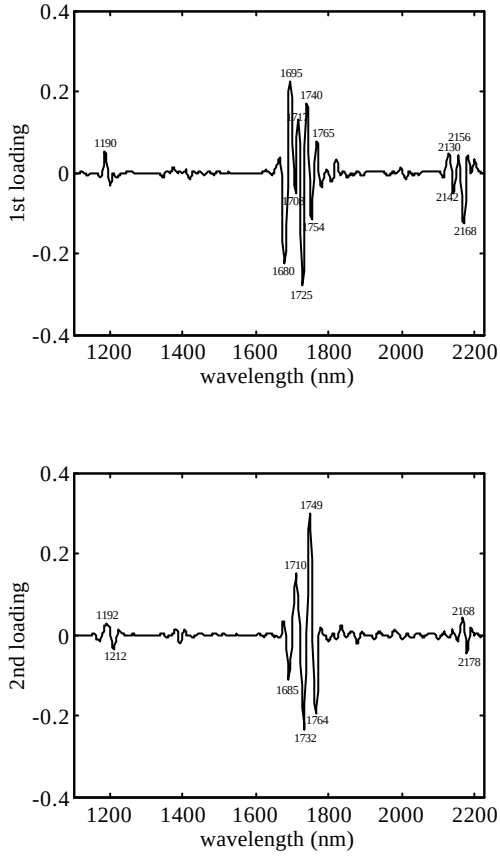


Fig. 4. a) First loading vector from PCA analysis (59.1% variance) and b) second loading vector (24.2% of variance).

There is a weak positive peak near 1190 nm (second overtone of CH_3) and a stronger one at 1695 nm (first overtone of CH_3), which is negatively correlated with a peak at 1680 nm (first overtone CH aromatic [18]). This negative correlation is interpreted as the decrease in the content of paraffin and the increase in the content of naphthenes and aromatic. There are other negative peaks at 1708 nm, 1725 nm and 1754 nm, negatively correlated with positive peaks at 1717 nm, 1740 nm and 1765 nm. Some of these peaks have been previously assigned as CH_2 group absorptions, but it is difficult to make a more precise assignment to justify the negative correlation between bands assigned to the same functional group.

Other negative contributions are found beyond 2100 nm, at 2142 nm and 2168 nm, and can be attributed to CH aromatic [19].

In this zone there are positive features at 2130 nm and 2156 nm. Around this spectral zone there are olefinic groups in the literature but these are minor groups (less than 1.0 % olefin) in the samples studied.

From a general analysis of the information provided by the first PC, which accounts for 59.1% of the total variance, it appears that the positive features provide information about the paraffinic components (linear and branched), while the negative peaks are related to the aromatic, and possibly naphthenic, chemical components.

This can be verified by Fig. 5, which shows the sample projection onto the space of the first two principal components. The samples are labelled with their aromatic content in percentages. The samples have highest aromatic content, and then lowest paraffins content, along the first principal component axe.

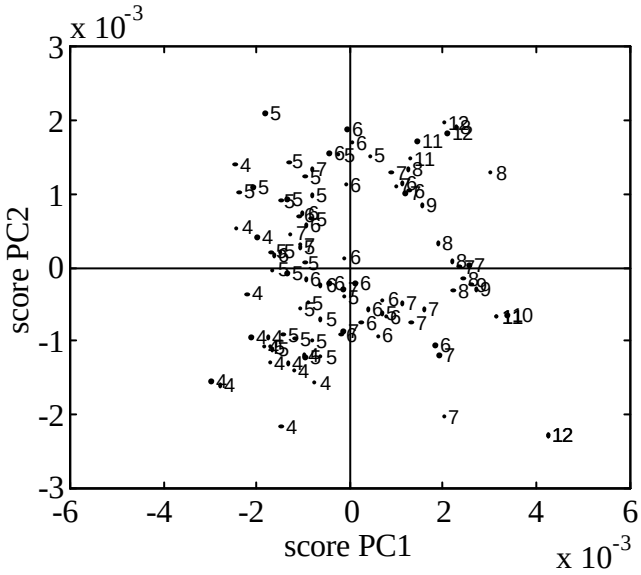


Fig. 5. Scores (PC1 and PC2) for the PCA model. The number near the points is its % aromatic content.

After this first analysis, and in accordance with the correlation between aromatics and paraffins, the wavelengths selected for building the model for aromatics or paraffins are those with positive coefficients in the first principal component loading (1190 nm, 1690 nm, 1717 nm, 1740 nm, 1766 nm, 1824 nm, 2130 nm and 2156 nm), or those with negative coefficients in the first principal component loading (1200 nm, 1680 nm, 1708 nm, 1754 nm, 2142 nm and 2170 nm).

The second PC (Fig 4b) lies along the direction that is orthogonal to the first PC and in the direction of the second largest variance. A peak at 1210 nm can be assigned to the second overtone of CH_2. This peak is negatively correlated with the peak at 1192 nm, which arises from the second overtone of CH_3. The peak at 1685 nm (CH aromatic) is weaker than it is in the first PC. This peak, and peaks at 1710 nm and 1749 nm, are negatively correlated with negative peaks at 1690 nm, 1732 nm and 1764 nm, which are characteristic of the CH_2 group [18]. A non-

referenced peak at 2178 nm is negatively correlated with the 2168 nm aromatic peak.

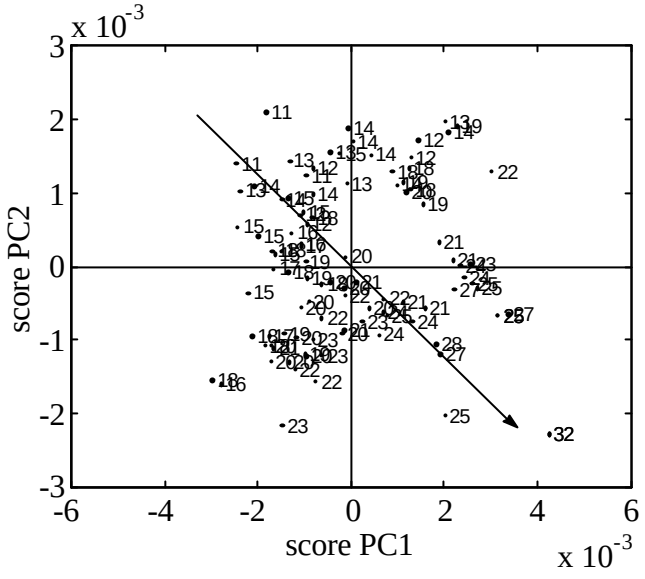


Fig. 6. Scores (PC1 and PC2) for the PCA model. The number near the points is its % naphthenes content.

The combination of the first and second principal components can also explain the percentage of naphthenes (Fig. 6). Samples with a high percentage of naphthenes have high positive values in PC1 and high negative values in PC2, while the samples with the lowest percentage of naphthenes have negative values of PC1 and positive values of PC2. This indicates that the variables with high coefficients in PC2 loading are important and should be incorporated to construct the PLS model for naphthenes. These variables are 1150 nm, 1210 nm, 1690 nm, 1732 nm, 1764 nm and 2178 nm or their correlated bands with positive coefficients in PC2 loading.

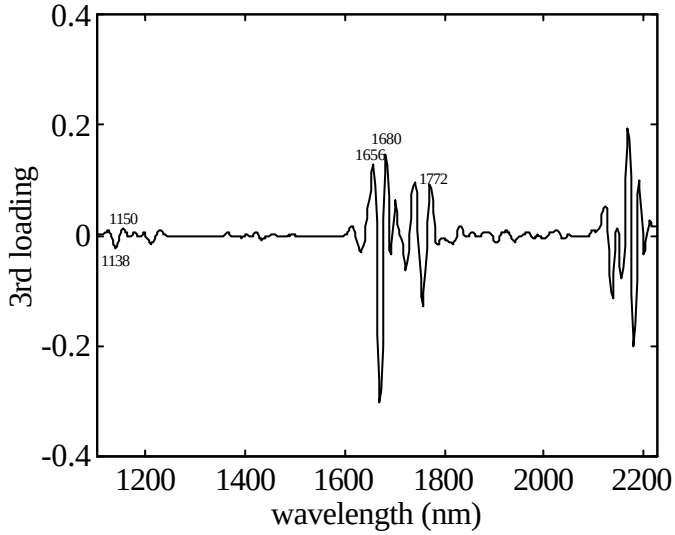


Fig. 7. Third loading vector from PCA analysis (8.7% variance).

The third principal component (Fig. 7) explains 8.7% of variance and provides information that is different from that provided by the first two. This PC is interpreted in the same way as the others. There is a peak at 1138 nm that is not assigned and is negatively correlated with the characteristic CH_3 absorption at 1150 nm. Between 1600 nm and 1780 nm, where there are referenced absorptions bands of CH_2 and CH_3, there are peaks at 1656 nm, 1680 nm and 1772 nm. Below 2100 nm there are several peaks but these are difficult to assign.

It is very difficult to conclude the variation explained by the third PC by evaluating the graph of the first three scores. This third PC seems to help distinguish between naphthenes and paraffines. Samples with a high naphthene content and a low paraffine content are close to the PC1-PC2 plane. High PC3 loading coefficient variables can therefore help determine naphthene content and so were included in the model.

PLS models and wavelength selection

Using the information from the second derivative and principal components analysis, we selected the wavelengths in Table 2 to construct PLS models to predict aromatics, naphthenes and paraffines. We compared the predictions of these models (RMSEP) with those of the full PLS models. The results are also given in the table. There were no significant differences.

Table 2. RMSEP of full PLS and selected wavelengths PLS models showing the number of latent values (LV) and wavelengths.

	LV	RMSEP full PLS	Wave-lengths	RMSEP PLS	wavelengths
aromatics	5	0.42	13	0.38	1150, 1200, 1656, 1680, 1708, 1728, 1740, 1754, 1772, 2124, 2142, 2170, 2192
naphthenes	4	1.12	11	1.09	1150, 1212, 1631, 1670, 1690, 1722, 1732, 1756, 2138, 2157, 2178
paraffins	6	0.93	16	0.99	1138, 1190, 1210, 1432, 1631, 1670, 1694, 1717, 1740, 1756, 1766, 1824, 2130, 2156, 2182, 2202

Conclusions

From spectral analysis of original and second derivative spectra, we can interpret the chemical source of bands. Principal component analysis is a chemometric method that provides valuable information to relate non-specific signals to analytes. By successively interpreting principal components we can identify suitable wavelengths for distinguishing samples by their composition. We have shown that with wavelength selection, we can work with fewer wavelengths without increasing error prediction.

Acknowledgments

The authors would like to thank the Spanish Ministry of Education, Culture and Sports for their economic support (project BQU 2000-1256) and Repsol Petróleo (Tarragona) for providing the samples and reference analysis. Santiago Macho would like to thank the *Fundación Repsol* and the University, Research and

Information Society Department of the Generalitat of Catalonia for providing a doctoral fellowship.

References

[1] J.J. Kelly, C.H. Barlow, M. Jinguji and J.B. Callis, Anal. Chem. 61 (1989) 313.

[2] S.J. Swarin and C.A. Drumm, Spectroscopy, 7 (1992) 42.

[3] M.S. Zetter and B.A. Politzer, Hydrocarbon Processing, 72 (1993) 1.

[4] B.G. Osborne, T. Fearn and P.H. Hindle, Practical NIR spectroscopy with applications in food and beverage analysis, 2nd edn., Longman Scientific & Technical, Harlow, 1993, pp 29-32.

[5] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch and C. de la Pezuela, Analyst, 123, (1998) R135.

[6] H. Martens and T. Naes, Multivariate Calibration, Wiley, New York, 1989.

[7] D.M. Haaland, Multivariate Calibration Methods Applied to Quantitative FT-IR Analyses, in J. R. Ferraro and K. Krishnan (Ed.), Practical Fourier Transform Infrared Spectroscopy, Academic Press, New York, 1989, pp. 396-468.

[8] M. Otto and W. Wegscheider, Anal. Chim. Acta, 180 (1986) 445.

[9] A. Lorber, Anal. Chem. 58 (1986) 1167.

[10] C.B. Lucasius, M.L.M. Beckers and G. Kateman, Anal. Chim. Acta, 286 (1994) 135.

[11] J.H. Kalivas, Anal. Chem. 58 (1986) 989.

[12] L.L. Juhl and J.H. Kalivas, Anal. Chim. Acta, 207 (1988) 125.

[13] Y. Liu, R. K. Cho, K. Sakuri, T. Miura and Y. Ozaki, Appl. Spectrosc. 48 (1994) 1249.

[14] J. Wang, M.G. Soma, M.K. Ahmed and H.H. Mantsch, J. Phys. Chem. 98 (1994) 4748.

[15] P. Robert, M.F. Devaux, N. Mouhous and E. Dufour, Appl. Spectrosc. 53 (1999) 226.

[16] G. Domjan, K.J. Kaffka, J.M. Jako, I.T. Valyi-Nagy, J. Near Infrared Spectrosc. 2 (1995) 67.

[17] K. Murayama, B. Czarnik-Matusewicz, Y. Wu, R. Tsenkova and Y. Ozaki, Appl. Spectrosc. 54 (2000) 978.

[18] J.J. Kelly and J.B. Calllis, Anal. Chem. 62 (1990) 1444.

[19] N. Asker, S. Kokot, Appl. Spectrosc. 45 (1991) 1153.

8.3 CONCLUSIONES

En este trabajo se han aplicado diferentes herramientas para el análisis espectral, como la segunda derivada o el análisis en componentes principales (PCA), que han permitido identificar las longitudes de onda más características de cada compuesto de la muestra.

Esta selección de longitudes de onda ha permitido el uso de modelos de calibración PLS más simples, con errores de predicción comparables a los modelos PLS utilizando el espectro completo.

El criterio de selección de variables utilizado en este trabajo ha sido interpretativo. Si se concluyese que estos modelos mejoran en gran medida su robustez, se podrían aplicar otros métodos de selección de variables, basados en criterios matemáticos, que no necesitan la interpretación de los espectros.

9. Conclusiones generales

9.1 CONCLUSIONES

El éxito del objetivo general enunciado al principio de esta tesis: "establecimiento de una metodología analítica basada en medidas de infrarrojo cercano y métodos de calibrado multivariante, para la determinación de propiedades de interés en el ámbito de la industria petroquímica", está directamente ligado a un ejercicio previo que analice: consideraciones propias del problema a abordar, aspectos básicos de la medida experimental a utilizar, en este caso la espectroscopia de infrarrojo, y otras consideraciones esenciales para el correcto desarrollo de los métodos quimiométricos. Estos aspectos fueron revisados en el capítulo 7, en el que se aportan algunas pautas que facilitan la incorporación de estas metodologías a la industria.

A continuación se vuelve a enunciar los objetivos detallados al inicio de esta tesis para comentar las conclusiones unidas a cada uno de ellos.

1 Desarrollo de modelos de calibrado multivariantes basados en medidas espectroscópicas.

La espectroscopia de infrarrojo cercano en combinación con técnicas de calibrado multivariante como el PLS se ha mostrado como una técnica alternativa, en la mayoría de los casos, a la cromatografía de gases para la determinación del contenido en parafinas, isoparafinas, naftenos y aromáticos en muestras de nafta. Su rapidez de análisis así como su versatilidad para análisis en línea la convierten en una técnica muy atractiva, aunque el proceso de obtención y validación del modelo de calibración serían uno de sus principales inconvenientes.

La espectroscopia de infrarrojo medio produce resultados similares a la espectroscopia de infrarrojo cercano, aunque las características de la

instrumentación NIR la hacen más adecuada para su aplicación en línea. Ha sido en esta última técnica en la que se ha profundizado más a lo largo de esta tesis.

Los modelos de calibrado desarrollados son satisfactorios, en la mayoría de los casos, para la determinación tanto de los porcentajes desglosados por número de carbono de las diferentes familias de hidrocarburos, como para los porcentajes globales. La determinación del contenido de parámetros de composición presenta ciertos problemas cuando el contenido es inferior al 1%, así como cuando los intervalos de contenidos son estrechos (1-2% de variación), como por ejemplo en el caso de los compuestos C_3, C_4, C_{10} o C_{11}.

La estimación de incertidumbres asociadas a las predicciones individuales proporciona información relevante al usuario, acostumbrado muchas veces a disponer de unos límites de tolerancia para el parámetro cuantificado.

La aplicación de esta metodología para la determinación de etileno en muestras de copolímeros de EPR/PP proporciona resultados comparables a la técnica de referencia, con la ventaja de evitar el pretratamiento de la muestra e incrementar la reproducibilidad de los resultados. En el caso del copolímero EPR, el intervalo de composición es estrecho (diferencia del 2%) en comparación con la reproducibilidad de la técnica de referencia. Por lo que se concluye que la determinación del contenido en etileno es posible, siempre y cuando el intervalo de composición de las muestras sea más amplio.

Los modelos de calibrado desarrollados a partir de las medidas NIR para la determinación de parámetros mecánicos, como la viscosidad o el índice de fluidez, no han proporcionado resultados comparables a la técnica de referencia. La existencia de una relación directa entre estas propiedades y el espectro NIR de las muestras no se ha podido determinar a través del uso de una técnica de calibración

multivariante lineal como la utilizada. Es posible que utilizando métodos de calibrado multivariante no lineales los resultados mejorasen.

El estudio de agrupaciones naturales de los espectros de las muestras de polímeros, realizado mediante el análisis de componentes principales, demostró la existencia de grupos que se corresponden con los diferentes tipos de polímeros estudiados: copolímero EPR/PP, copolímero EPR y homopolímero de PP.

La selección de variables basadas en métodos interpretativos, tales como la segunda derivada de los espectros, o el análisis de componentes principales, permite una selección efectiva de las variables más adecuadas para la determinación del contenido en parafinas, naftenos y aromáticos. Estos métodos facilitan una interpretación química de los modelos desarrollados.

El desarrollo de los modelos de calibrado utilizando las variables seleccionadas producen valores de error de predicción comparables a los encontrados utilizando la totalidad del espectro para calibrar.

2 Estudio de la estabilidad en el tiempo de los modelos desarrollados y transferencia de los modelos a nuevas situaciones experimentales.

La viabilidad de la aplicación de la calibración multivariante a partir de medidas espectroscópicas en el ámbito industrial depende en gran medida de la estabilidad de los modelos desarrollados, así como en su adaptación a nuevas situaciones experimentales.

En este sentido, se ha demostrado que tanto las técnicas de control estadístico de procesos (SPC) univariantes consideradas, gráficos de Shewhart, como las técnicas de control estadístico multivariante (MSPC), en concreto los estadísticos

T^2 y Q, son útiles para monitorizar la estabilidad del modelo que determina el contenido en etileno en copolímeros EPR/PP.

Los gráficos T^2 y Q proporcionan información útil que ayuda a interpretar la causa de la pérdida de validez del modelo de calibrado.

La estrategia planteada para la selección de las muestras de control, ha permitido adaptar el modelo de calibrado existente a las nuevas condiciones experimentales. Esta estrategia reduce significativamente el esfuerzo requerido en la etapa de estandarización de los modelos de calibrado.

Las técnicas de estandarización, corrección del sesgo y la pendiente (SBC) y la estandarización directa por partes (PDS), permiten adecuar el modelo establecido a la nueva situación experimental producida por variaciones en el proceso de medida.

3 Detección de situaciones anómalas, no contempladas en el proceso de calibrado (detección de muestras discrepantes).

Así como es importante el control de la estabilidad del modelo, es imprescindible la adopción de técnicas de detección de muestras *outlier* que sean fácilmente interpretables por el usuario, si se desea una amplia difusión de este tipo de técnicas en el ámbito industrial.

Los gráficos de control basados en los estadísticos T^2 y Q, han demostrado ser eficaces en la detección de muestras discrepantes cuando se aplicaron a la determinación del porcentaje de etileno en copolímeros de EPR/PP. Estas herramientas son fáciles de utilizar e interpretar por un usuario no especializado en quimiometría.

El empleo de una herramienta adicional como el algoritmo BESI, ayuda a identificar las causas que han originado el comportamiento discrepante de la muestra. Esta información es muy valiosa, ya que orienta al usuario sobre la acción más adecuada ante cada situación.

9.2 PERSPECTIVAS FUTURAS

Después de presentar los resultados y comentar las conclusiones obtenidas, se esbozan muy brevemente algunos de los temas tratados en la presente tesis que serían susceptibles de ser objeto de un trabajo posterior:

- El uso de nuevos métodos de calibración que mejoren los resultados en aquellos modelos que no han tenido una solución óptima. Entre los posibles métodos se encuentran los modelos basados en agrupaciones (k vecinos más próximos), que podrían solventar el problema derivado de la poca variabilidad de la propiedad de interés en las muestras, o los métodos no lineales (PLS no lineal, redes neuronales, etc) que pueden ser más eficaces cuando la propiedad de interés no guarda una relación lineal y sencilla con el espectro, como podría ser el caso de la viscosidad o el índice de fluidez.

- El desarrollo de estrategias de control en el caso de muestras cuya constitución no hiciera viable el uso de una muestra estable de control. Tales estrategias serían necesarias por ejemplo en el caso de los modelos de las muestras de nafta e implicarían encontrar una sustancia pura y estable lo suficientemente similar a la muestra. Por ejemplo en el campo del análisis de hidrocarburos esta muestra estable podría ser el tolueno.

- Mejora de las herramientas de estandarización, de forma que una vez aplicadas se pudieran seguir aplicando las herramientas de control y detección de *outliers* iniciales.

- Desarrollar nuevas estrategias de calibración que aumentasen la robustez de los modelos, alargando el tiempo entre estandarizaciones o incluso haciendo éstas innecesarias.

Esta Tesis se ha realizado con el soporte de la Fundación Repsol y del Departament d'Universitats Recerca i Societat de la Informació de la Generalitat de Catalunya.

Aquesta Tesis ha estat realitzada amb el suport de la Fundación Repsol i del Departament d'Universitats Recerca i Societat de la Informació de la Generalitat de Catalunya.