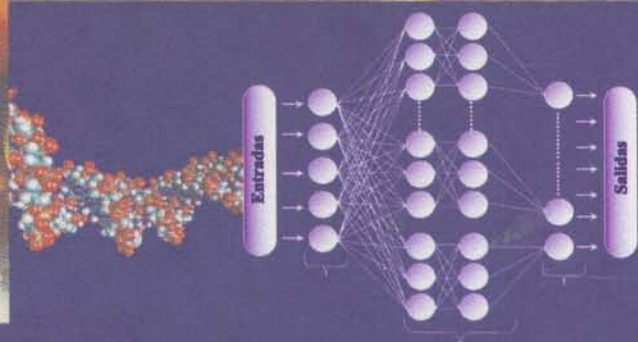
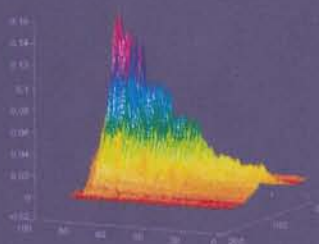




Departamento de Química Analítica
Universidade da Coruña

APLICACIONES AL CONTROL DE CALIDAD INDUSTRIAL DE LA ESPECTROSCOPIA INFRARROJA MEDIA COMBINADA CON MÉTODOS QUIMIOMÉTRICOS MULTIVARIANTES



María Paz Gómez Carracedo
A Coruña, 2005



Departamento de Química Analítica
Universidade da Coruña

APLICACIONES AL CONTROL DE CALIDAD INDUSTRIAL DE
LA ESPECTROSCOPIA INFRARROJA MEDIA COMBINADA CON
MÉTODOS QUIMIOMÉTRICOS MULTIVARIANTES

Memoria presentada por
María Paz Gómez Carracedo
para optar al
Grado de Doctora en Química,
A Coruña, Abril de 2005



UNIVERSIDADE DA CORUÑA
DEPARTAMENTO DE QUÍMICA ANALÍTICA

Campus da Zapateira s/n.
15071 A Coruña. España.
Telf. 981167000.
Fax 981167065

Dra. Dña. SOLEDAD MUNIATEGUI LORENZO, Catedrática y
Directora del Departamento de Química Analítica de la Universidade da
Coruña,

AUTORIZA a Dña. MARÍA PAZ GÓMEZ CARRACEDO a presentar
el trabajo titulado “APLICACIONES AL CONTROL DE
CALIDAD INDUSTRIAL DE LA ESPECTROSCOPIA INFRARROJA
MEDIA COMBINADA CON MÉTODOS QUIMIOMÉTRICOS
MULTIVARIANTES” para optar al grado de Doctora por la
Universidade da Coruña.

A Coruña, Abril de 2005

Dra. Dña. Soledad Muniategui Lorenzo



UNIVERSIDADE DA CORUÑA
DEPARTAMENTO DE QUÍMICA ANALÍTICA
Campus da Zapateira s/n.
15071 A Coruña. España.
Telf. 981167000.
Fax 981167065

Dr. D. DARÍO PRADA RODRÍGUEZ, Catedrático y **Dr. D. JOSÉ M. ANDRADE GARDA**, Profesor Titular del Departamento de Química Analítica de la Universidade da Coruña,


CERTIFICAN

Que la presente Tesis Doctoral titulada “APLICACIONES AL CONTROL DE CALIDAD INDUSTRIAL DE LA ESPECTROSCOPIA INFRARROJA MEDIA COMBINADA CON MÉTODOS QUIMIOMÉTRICOS MULTIVARIANTES”, se ha realizado bajo su dirección en el Departamento de Química Analítica de la Universidade da Coruña.

Y para que así conste, a los efectos oportunos, firman la presente en A Coruña, Abril de 2005.



Dr. D. Darío Prada Rodríguez



Dr. D. José M. Andrade Garda

A mis padres
y a mis hermanas

AGRADECIMIENTOS

Quisiera dar las gracias al Departamento de Química Analítica de la Universidade da Coruña por permitirme realizar este trabajo.

A mis directores el Dr. D. Darío Prada Rodríguez y Dr. D. José Manuel Andrade Garda por haberme introducido en el mundo de la investigación, por su ayuda desinteresada en todo momento y cómo no, por su orientación y apoyo constantes.

Agradezco a la Fundación Mapfre-Universidade da Coruña por la concesión de la beca en la convocatoria 2000/2001.

Gracias al proyecto titulado "Predicción de parámetros de calidad de combustible de aviación mediante espectroscopia FT-IR sometiendo a muestras a distintas condiciones de presión e temperatura en fase vapor" concedido en el bienio 1999-2001 por la Universidade da Coruña.

Gracias a la colaboración prestada por el SCSIE (Servicio Central de Soporte a la Investigación Experimental) de la Universidad de Valencia.

Mi más sincero agradecimiento a todo personal que forma parte del Departamento de Química Analítica de la Universidade da Coruña. Gracias a los profesores que lo componen (Darío, Esther F., Andrade, Sol, Puri, Elisa, Alatzne, Mariago, Isa, Rosa, Esther V., Jorge y Elia) por su trabajo y esfuerzo por la buena convivencia. Gracias a todos ellos. A Esther F., Elisa, Alatzne, Isa y Rosa por esas reuniones a la hora del café. A Sol, Puri y Darío por cuidarnos tanto. A Maribel porque siempre sabe donde están las cosas. A Alatzne, porque aún con el poco tiempo del que dispone, siempre se ha preocupado por todos en todo momento, gracias por ese día de muestreo, que siempre viene bien desconectar un poco de vez en cuando, por tus cuidados cuando nos desborda el trabajo. A Darío por su paciencia y su preocupación. A Andrade, gracias por todo lo que me has ayudado, por el tiempo que has perdido conmigo (y con la quimiometría), por ser siempre accesible en tus explicaciones, tus rápidas correcciones (que son de agradecer), tus explicaciones a distancia (el PLS se me estaba revelando), y un largo etcétera de cosas, por todo eso, muchísimas gracias.

A mis compañeros de laboratorio con los que he compartido todo este tiempo de trabajo en el Departamento, si no hay buen ambiente en el sitio donde pasas gran parte de tu tiempo, no se rinde lo suficiente. A Gloria y a Vero (peque), gracias por contar conmigo para los trabajos con acetona (de la quay), que aunque el trabajo de extracción es muy intenso, también es entretenido todo el tiempo pasado con vosotras y con el gira-gira sin olvidar, cómo no, vuestra constancia en el café. A Vero por esos correos tan entretenidos que me manda y gracias por esos ánimos, por recordarme esas frases célebres que dije en algún momento despistado y que me hacen sonreír en ocasiones. A Gloria por los ánimos y por su compañía en la sala de ordenadores. A Gus, le agradezco el mejor y más original regalo que he tenido en el amigo invisible, además, tu saludo por las mañanas es único. A Chus por ser la veterana y por sus conversaciones de ánimo y consejos; a Carmen y a Fany porque han sido mis compañeras desde mi entrada en el departamento y ¿qué haría yo si no fuera a esas clases de aerobio?, con lo bien que nos lo pasábamos...; a María Piñeiro por esos cuidados que nos proporciona esté donde esté y por esos ánimos; a Merce porque el paint no tiene secretos para ella; a Mónica por su dominio del inglés; a Deborah por sus soluciones en mis dudas informáticas; a Sonia por sus conversaciones; a M^{te} Carmen por esas frases de ánimo durante todo este tiempo. A Pilar y a Diana, por esos cafés entretenidos y por escucharme.

No me puedo olvidar de esa gente que ha compartido conmigo el equipo de FTR, a Miguel, Diana y Raquel, por todo el tiempo pasado peleándonos con el equipo. A Raquel, por ofrecerme su ayuda en cualquier momento, aún no teniendo tiempo para nada.

Durante todo este tiempo ha habido momentos tristes (despedidas) a la vez que alegres (cada vez que se va alguien suele ser para mejorar). A toda la gente que se ha ido del departamento pero con las que he pasado buenos momentos, a Diego (gracias por ese humor), a Mariví (por sus conversaciones), a Miguel (por los conocimientos aportados en mi inicio a la investigación y su buen humor), a Isa G. (por sus charlas), a Manuela (nunca me olvidaré de Rafaela Carrá) a Pi (por ser quien es, claro está), a Diana (por esas comidas y esas conversaciones tan amenas, por hacernos reír siempre), a Vilarriño (por su cultura general), a Vero y a Gerardo (porque aún forman parte del departamento), a María Piñeiro (por sus ánimos y porque siempre nos ha cuidado y se ha preocupado por todos), a Marisa, a Ramón y a María J. (porque en algún momento cada uno de ellos ha formado parte del departamento).

A toda la gente que se ha incorporado recientemente, porque ahora ya forman parte del departamento. En especial a Patricia que va a seguir intentando que el IR nunca se estropee; a Ana por sus conversaciones; a David porque ya lleva 2 años con nosotros y ya está más que adaptado; a Roberto y a Fernando por sus conversaciones a la hora del café; a Luis; a Vero 3 porque aunque ha estado poco tiempo con nosotros, ha sido entretenido dialogar con ella.

Este trabajo no sería posible sin Juanra, Julián y Marcos que nos han introducido en el arduo mundo de las Redes Neuronales y los Algoritmos Genéticos. Gracias por el tiempo que habéis empleado en este trabajo, en las pruebas que se han hecho y en esas reuniones para intercambiar opiniones y deliberar el camino a seguir.

Gracias a todos por vuestro trabajo, por vuestra preocupación, por vuestros consejos, por vuestros ánimos tanto en los buenos como en los malos momentos de esta etapa de mi vida y cómo no, por la amistad que me habéis brindado y que espero seguir manteniendo. Gracias por todo y por hacer de mi lugar de formación y de trabajo mi segunda familia.

Aunque he pasado poco tiempo en el Departamento de Química Analítica de la Universidad de Valencia, agradezco a Miguel de la Guardia y a Boro el aceptarme para hacer una estancia que constituyó una nueva experiencia en esta parte de mi existencia. Agradezco a Guillermo y a Sergio el compartir conmigo sus conocimientos de Raman; a Joseph, a Eva y a Patricia por sus conversaciones y a Xavi y a Paco por ser tal como son. A todos ellos, gracias por esos cafés y las comidas en el tiempo que he estado allí y por seguir manteniendo el contacto, gracias por todo. A Boro, Sergio, Paco y Xavi le agradezco lo bien que me lo pasé con sus charlas.

También es de agradecer durante todo este tiempo, la compañía de buenos amigos. Gracias a Pablo y a Nuria, a Carina, a Ana, a Paula Garrido por saber que se puede contar con vosotros, aunque no estés en tu mejor momento y aunque las "diferencias de opiniones" sean más que frecuentes. A mis compañeras de Facultad (Cris, Sonia, Rosa y Lucía) por seguir manteniendo el contacto y por esos años pasados en Santiago.

Además de todos estos agradecimientos, la parte más importante de la vida de cualquier persona son las personas con las que convive y las que gracias a ellas ha sido posible todo este trabajo desarrollado durante tantos años. Sin el apoyo de mis padres nunca

Llegaría a donde estoy, nada de esto sería posible. Gracias por todo lo que habéis hecho por mi y sobre todo, por aguantarme durante todo este tiempo, incluso en los malos tiempos vividos en ocasiones. Mamá gracias por tu paciencia, aunque a veces me parece excesiva. Papá, gracias por tus consejos. A mis hermanas, Ana y Eva, quiero agradecerles su comprensión y su apoyo durante todo este tiempo, ante todo han sido mis hermanas en todo momento, gracias por escucharme siempre que lo he necesitado, por intentar hacerme entrar en razón (aunque ya sé que soy muy cabezona), gracias por estar ahí cuando os he necesitado, por intentar animarme siempre, o por lo menos, comprenderme, siempre estuvisteis ahí cuando lo he necesitado. Y por último, gracias a mi recién estrenado cuñado, porque ya eres uno más de la familia.

A toda esta gente le tengo que agradecer lo que soy, gracias por todo.

CAPÍTULO 0: OBJETIVOS GENERALES	ix
CAPÍTULO I: INTRODUCCIÓN A LOS MÉTODOS MULTIVARIANTES	1
1. Planteamiento del problema	5
2. Variables predictoras y variables a predecir	8
3. Curvas de potencia	8
4. SIMCA	11
5. Regresión parcial mediante mínimos cuadrados (PLS)	16
5.1. Ventajas de PLS	21
5.2. Inconvenientes	21
5.3. Transformación, escalado y centrado en la media	22
5.4. Selección del número adecuado de VL y validación del modelo	22
5.5. Actualización del modelo	27
5.6. Interpretación y diagnósticos del modelos de PLS	28
5.7. Consideraciones respecto a errores. Exactitud y precisión	31
5.8. Aplicaciones	35
6. Redes de neuronas artificiales	37
6.1. Introducción	37
6.2. Elementos básicos	40
6.3. Redes de propagación hacia atrás (<i>Backpropagation</i>)	46
6.4. Aprendizaje y validación	49
6.5. Aplicaciones	55
7. Diseño de experiencias y optimización	60
7.1. Diseño de experiencias	60
7.2. Método Simplex geométrico	62
8. Bibliografía	65
CAPÍTULO II: MÉTODOS DE SELECCIÓN DE VARIABLES	79
1. Introducción	83
2. Rotación de Procrustes	84
2.1. Concepto	84
2.2. Desarrollo matemático	86
2.3. Necesidades de diagnósticos asociadas a la Rotación de Procrustes	91
2.4. Rotación de Procrustes para comparar subespacios de las variables	95
2.5. Aplicaciones de la Rotación de Procrustes	97

3. Algoritmos Genéticos	101
3.1. Introducción	101
3.2. Fundamentos	103
3.3. Los individuos	104
3.4. Estructura de los AAGG	105
3.5. Los Operadores Genéticos	108
3.6. Ventajas y desventajas del uso de los AAGG	113
3.7. Aplicaciones	113
3.8. Descripción del problema abordado en esta Memoria	115
4. Bibliografía	118
CAPÍTULO III: ESPECTROSCOPIA IR	127
1. Introducción a la espectroscopia IR	131
1.1. Espectroscopia MIR	131
1.2. Espectroscopia Raman	139
2. Aplicaciones de la espectroscopia	140
2.1. Aplicaciones de la espectroscopia IR en el campo petroquímico	140
2.2. Aplicaciones de la espectroscopia IR en el campo alimentario	140
2.3. Aplicaciones de la espectroscopia Raman	142
3. Instrumentación analítica empleada y control de calidad	143
4. Bibliografía	149
CAPÍTULO IV: ANÁLISIS DE QUEROSENOS	159
Parte A.- Aproximación al producto objeto del estudio	163
1. Introducción al combustible de aviación	163
1.1. Composición	165
1.2. Uso, tipos y propiedades del queroseno	166
2. Consideraciones para la salud	170
2.1. Exposición al queroseno	170
2.2. Rutas de exposición	175
Parte B.- Parte experimental	177
1. Optimización de las metodologías alternativas de análisis	177
1.1. Medida de querosenos en fase gas	177
1.2. Medida de querosenos en fase líquida	184

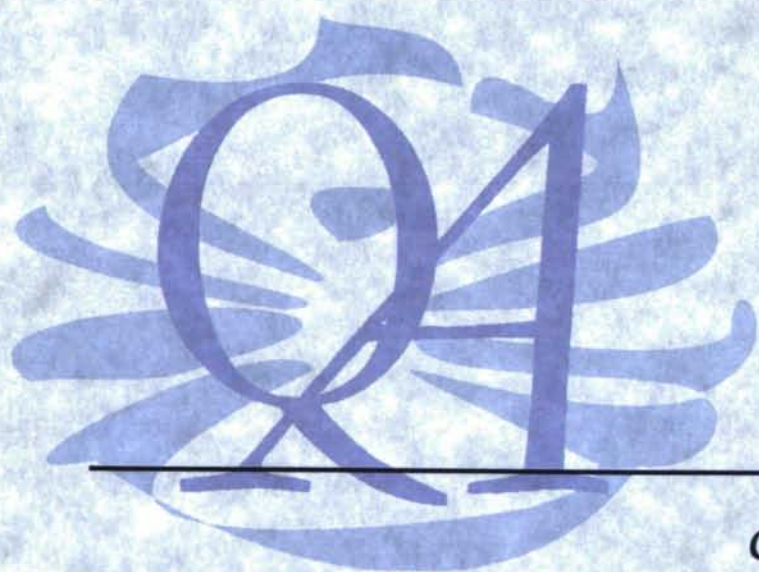
Parte C.- Resultados y discusión	195
1. Modelos multivariantes en fase gas	195
1.1. Modelos multivariantes con el sistema de medida simple	195
1.2. Modelos multivariantes con el sistema de medida complejo	215
1.3. Exactitud y precisión	233
2. Modelos multivariantes en fase líquida	238
2.1. Modelos multivariantes con el sistema de medida FTMIR-ATR	238
2.2. Modelos multivariantes con el sistema de medida Raman	241
Parte D.- Bibliografía	246
 CAPÍTULO V: ANÁLISIS DE ZUMO	 251
Parte A.- Aproximación al producto objeto del estudio	255
1. Definición y clasificación general de los frutos	251
2. Importancia de la fruta	257
3. Las manzanas	258
3.1. Propiedades nutricionales de la manzana	258
3.2. Variedades	259
3.3. Producción mundial	259
3.4. Zumos de manzana	264
3.5. Proceso industrial para obtener zumo	265
4. Composición Química	267
4.1. Compuestos nitrogenados	268
4.2. Carbohidratos	269
4.3. Lípidos	270
4.4. Ácidos orgánicos	271
4.5. Compuestos fenólicos	271
4.6. Compuestos aromáticos	273
4.7. Vitaminas	273
4.8. Minerales	274
5. Cambios en la intensidad respiratoria	274
6. Adulteración	274
7. Legislación	277

Parte B.- Parte experimental	282
1. Aparatos y Software	282
2. Muestras	283
2.1. Zumos/refrescos comerciales	283
2.2. Zumos "100% puros" de manzana	284
2.3. Zumos "puros diluidos" de manzana	287
2.4. Zumos "sintéticos" a partir de azúcares	289
Parte C.- Resultados y discusión	297
1. Clasificación mediante curvas de potencia, rango bajo de zumos (2% al 20%)	297
1.1. Modelado	297
1.2. Validación del modelo	298
1.3. Clasificación de los refrescos comerciales	302
2. Clasificación mediante curvas de potencia, rango alto de zumos (20% - 100%) ...	303
2.1. Modelado	303
2.2. Validación del modelo	304
2.3. Clasificación de los refrescos y zumos comerciales	306
3. Variables que determinan la clasificación de las muestras mediante curvas de potencia	307
4. Clasificación mediante SIMCA, rango bajo de zumos (2% - 20%)	311
4.1. Modelado	311
4.2. Validación del modelo	316
4.3. Clasificación de los refrescos comerciales	317
5. Clasificación mediante SIMCA, rango alto de zumos (20% - 100%)	318
5.1. Modelado	318
5.2. Validación del modelo	319
5.3. Clasificación de los refrescos y zumos comerciales	319
6. Variables que determinan la clasificación de las muestras	321
6.1. Rango bajo	321
6.2. Rango alto	324
7. Clasificación mediante PLS, rango bajo de zumos (2% - 20%)	326
8. Clasificación mediante PLS, rango alto de zumos (20% - 100%)	329
9. Clasificación mediante RRNN, rango bajo de zumos (2% -20%)	331
10. Clasificación mediante RRNN, rango alto de zumos (20% 100%)	332

11. Anexo	333
Parte D.- Bibliografía	334
CAPÍTULO VI: MODELOS CON VARIABLES REDUCIDAS	343
Parte A.- Modelos con variables seleccionadas mediante Rotación de Procustes	349
1. Curvas de Potencia	349
1.1. Rango bajo de zumos (2% al 20%), 1062 y 1064 cm^{-1} (variables 95 y 94)	349
1.2. Rango alto de zumos (20% al 100%), 1060 y 1062 cm^{-1} (variables 96 y 95)	352
2. SIMCA	356
2.1. Rango bajo de zumos (2% al 20%), 1062 y 1064 cm^{-1} (variables 95 y 94)	356
2.2. Rango alto de zumos (20% al 100%), 1060 y 1062 cm^{-1} (variables 96 y 95)	358
3. PLS	360
3.1. Rango bajo de zumos (2% al 20%), 1062 y 1064 cm^{-1} (variables 95 y 94)	360
3.2. Rango alto de zumos (20% al 100%), 1060 y 1062 cm^{-1} (variables 96 y 95)	362
4. Redes de Neuronas Artificiales	364
4.1. Rango bajo de zumos (2% al 20%), 1062 y 1064 cm^{-1} (variables 95 y 94)	364
4.2. Rango alto de zumos (20% al 100%), 1060 y 1062 cm^{-1} (variables 96 y 95)	365
Parte B.- Modelos con variables seleccionadas mediante Rotación de Procustes aplicada sobre los <i>loadings</i>	366
1. Curvas de Potencia	366
1.1. Rango bajo de zumos (2% al 20%), 1084 y 1000 cm^{-1} (variables 84 y 126)	366
1.2. Rango alto de zumos (20% al 100%), 1062 y 1000 cm^{-1} (variables 95 y 126)	369
2. SIMCA	374
2.1. Rango bajo de zumos (2% al 20%), 1084 y 1000 cm^{-1} (variables 84 y 126)	374
2.2. Rango alto de zumos (20% al 100%), 1062 y 1000 cm^{-1} (variables 95 y 126)	376
3. PLS	378
3.1. Rango bajo de zumos (2% al 20%), 1084 y 1000 cm^{-1} (variables 84 y 126)	378
3.2. Rango alto de zumos (20% al 100%), 1062 y 1000 cm^{-1} (variables 95 y 126)	380
4. Redes de Neuronas Artificiales	382
4.1. Rango bajo de zumos (2% al 20%), 1084 y 1000 cm^{-1} (variables 84 y 126)	382
4.2. Rango alto de zumos (20% al 100%), 1062 y 1000 cm^{-1} (variables 95 y 126)	382

Parte C.- Modelos con variables seleccionadas mediante AAGG-búsqueda	
fija	384
1. Curvas de Potencia	384
1.1. Rango bajo de zumos (2% al 20%), 1228 y 934 cm^{-1} (variables 12 y 159)	384
1.2. Rango alto de zumos (20% al 100%), 1228 y 934 cm^{-1} (variables 12 y 159)	387
2. SIMCA	391
2.1. Rango bajo de zumos (2% al 20%), 1228 y 934 cm^{-1} (variables 12 y 159)	391
2.2. Rango alto de zumos (20% al 100%), 1228 y 934 cm^{-1} (variables 12 y 159)	393
3. PLS	395
3.1. Rango bajo de zumos (2% al 20%), 1228 y 934 cm^{-1} (variables 12 y 159)	395
3.2. Rango alto de zumos (20% al 100%), 1228 y 934 cm^{-1} (variables 12 y 159)	397
4. Redes de Neuronas Artificiales	399
4.1. Rango bajo de zumos (2% al 20%), 1228 y 934 cm^{-1} (variables 12 y 159)	399
4.2. Rango alto de zumos (20% al 100%), 1228 y 934 cm^{-1} (variables 12 y 159)	400
Parte D.- Modelos con variables seleccionadas mediante AAGG-búsqueda	
por etapas	401
1. Curvas de Potencia	401
1.1. Rango bajo de zumos (2% al 20%), 1168 y 1098 cm^{-1} (variables 42 y 77)	401
1.2. Rango alto de zumos (20% al 100%), 1168 y 1098 cm^{-1} (variables 42 y 77)	404
2. SIMCA	407
2.1. Rango bajo de zumos (2% al 20%), 1168 y 1098 cm^{-1} (variables 42 y 77)	407
2.2. Rango alto de zumos (20% al 100%), 1168 y 1098 cm^{-1} (variables 42 y 77)	409
3. PLS	411
3.1. Rango bajo de zumos (2% al 20%), 1168 y 1098 cm^{-1} (variables 42 y 77)	411
3.2. Rango alto de zumos (20% al 100%), 1168 y 1098 cm^{-1} (variables 42 y 77)	413
4. Redes de Neuronas Artificiales	415
4.1. Rango bajo de zumos (2% al 20%), 1168 y 1098 cm^{-1} (variables 42 y 77)	415
4.2. Rango alto de zumos (20% al 100%), 1168 y 1098 cm^{-1} (variables 42 y 77)	416
Parte E.- Comparación de los resultados obtenidos mediante los cuatro	
criterios de selección de variables	417
1. Uso de todas las variables espectrales	417
2. Selección de variables por Rotación de Procustes	419
3. Selección de variables por Procustes empleando los <i>loadings</i>	420

4. Selección de variables mediante algoritmos genéticos con búsqueda fija	422
5. Selección de variables mediante algoritmos genéticos con búsqueda por etapas . . .	423
CAPÍTULO VII: CONCLUSIONES	429
ANEXO: PUBLICACIONES Y COMUNICACIONES A CONGRESOS	435



Capítulo 0

Objetivos Generales

Cada día son más complejos los problemas a los que se enfrenta la Química Analítica en los ámbitos de la producción industrial (también llamada Química Analítica de Procesos), especialmente en lo referente a uno de sus campos de actividad primordiales: el Control de Calidad de los Productos manufacturados. Esto hace necesario implementar metodologías analíticas rápidas, sencillas, fiables y que permitan el mínimo consumo de reactivos químicos, con lo que se obtendrá menor cantidad de residuos.

A lo largo de esta Memoria de Tesis Doctoral se empleará como técnica analítica básica de trabajo la espectrofotometría infrarroja con Transformada de Fourier en la zona media, la cual permite cumplir las premisas anteriores y se clasifica como una técnica instrumental de orden uno. Este tipo de técnicas permite trabajar en presencia de concomitantes sin necesidad de conocer su participación en la señal registrada aunque precisan que todos los posibles interferentes estén en las muestras de calibración. El caso ideal sería que, además, su espectro fuese claramente diferente del correspondiente al analito de interés. La situación es incluso más interesante cuando no es un “analito típico” lo que se quiere determinar (p.ej., la concentración de un metal) sino un parámetro (o propiedad) relacionado con el producto pero que “no existe” como tal en la muestra. Esta es, precisamente, la problemática que se aborda en la Memoria: la predicción de parámetros (propiedades) que deben considerarse como el resultado de la composición química del producto (o de su respuesta global ante un estímulo exterior).

Los dos productos comerciales que se han estudiado en la Memoria resultan de gran importancia para las actividades socioeconómicas de la Comunidad Autónoma Gallega. Aunque representan dos sectores productivos claramente diferenciados, algunas de sus problemáticas analíticas pueden abordarse con una misma sistemática de trabajo: los zumos y bebidas de manzana (A Coruña es una de las principales productoras de manzana a nivel nacional) y el combustible de aviación (A Coruña cuenta con la única refinería de crudo de petróleo del Noroeste peninsular).

Tratar de predecir satisfactoriamente parámetros tales como “porcentaje de zumo”, “temperatura de cristalización del combustible” o “temperatura de deflagración” implica no sólo el trabajo con una técnica analítica adecuada, en este caso la espectrometría vibracional, sino también aplicar herramientas quimiométricas que permitan desarrollar modelos adecuados a las necesidades del control de calidad, en particular que aporten predicciones insesgadas. Además de una predicción correcta

interesa detectar la presencia de muestras no consideradas en la calibración y, por tanto, disponer de diagnósticos que permitan detectar características espectrales no habituales (lo cual no deja de ser uno de los objetivos de cualquier sistema de Gestión de la Calidad).

Otro aspecto importante en el control de calidad y el desarrollo de modelos predictivos consiste en evaluar la posibilidad de uso de un pequeño conjunto de variables (en esta Memoria de tipo espectral) que conduzcan a modelos satisfactorios. Algunas ventajas de las técnicas relacionadas con la selección de variables son la mayor rapidez en el proceso de medida y la simplificación de la interpretación química de los modelos quimiométricos y los resultados que éstos ofrecen.

En consecuencia, los principales objetivos que se persiguen en esta Memoria son:

- 1.- Estudiar técnicas quimiométricas orientadas a la selección de un mínimo subconjunto de variables que permita abordar, posteriormente, procesos de clasificación mediante Curvas de Potencia, SIMCA, PLS y Redes de Neuronas Artificiales e implementar modelos de regresión multivariante (PLS lineal). Las técnicas elegidas, rotación de Procustes y Algoritmos Genéticos, tienen principios conceptuales muy diferentes y, por tanto, abordan el problema empleando dos perspectivas opuestas, una determinista (Procustes) y otra basada en procesos estocásticos de evolución (según los principios Darwinianos de la selección y evolución natural (Algoritmos Genéticos).
- 2.- Desarrollar procedimientos analíticos sencillos, rápidos y respetuosos con el Medio Ambiente, basados en la espectroscopia vibracional, que permitan caracterizar las muestras bajo estudio. Además de las técnicas de medida de fase líquida (Raman y Reflectancia Total Atenuada), se trabajará en la puesta a punto de sistemas de medida en fase gas.
- 3.- Comparar las predicciones obtenidas empleando diferentes procedimientos experimentales, con distintas técnicas de clasificación y distintos conjuntos de variables. Analizar sus ventajas e inconvenientes.

A pesar de que al inicio de los diferentes capítulos se irán indicando sus objetivos específicos, la sistemática general de trabajo seguida ha sido: (a) presentación de los conceptos básicos a emplear (espectrales y quimiométricos), (b) revisión bibliográfica comentada, incluyendo una breve discusión acerca de cada uno de los productos considerados, (c) presentación de las hipótesis de trabajo, (d) desarrollo experimental y (e) discusión de los resultados. Con ánimo de evitar reiteraciones, los capítulos 1, 2 y 3 se centran en (a), (b) y (c) y los capítulos 4, 5 y 6 abordan (d) y (e) de forma separada cada producto.



Capítulo I

Introducción a los métodos multivariantes

Objetivo:

En este capítulo se presentan conceptualmente los métodos quimiométricos de los que se hará uso en las aplicaciones experimentales recogidas en esta Memoria. También se recogerán algunos trabajos previos donde se hayan explicado algunas de las ventajas, inconvenientes y problemas asociados a su utilización, además de manifestar los éxitos que permiten alcanzar.

Índice:

1. Planteamiento del problema
2. Variables predictoras y variables a predecir
3. Curvas de potencia
4. SIMCA
5. Regresión parcial mediante mínimos cuadrados (PLS)
 - 5.1. Ventajas de PLS
 - 5.2. Inconvenientes
 - 5.3. Transformación, escalado y centrado en la media
 - 5.4. Selección del número adecuado de VL y validación del modelo
 - 5.5. Actualización del modelo
 - 5.6. Interpretación y diagnósticos del modelos de PLS
 - 5.7. Consideraciones respecto a errores. Exactitud y precisión
 - 5.8. Aplicaciones
6. Redes de neuronas artificiales
 - 6.1. Introducción
 - 6.2. Elementos básicos
 - 6.3. Redes de propagación hacia atrás (Backpropagation)
 - 6.4. Aprendizaje y validación
 - 6.5. Aplicaciones
7. Diseño de experiencias y optimización
 - 7.1. Diseño de experiencias
 - 7.2. Método Simplex geométrico
8. Bibliografía

1. PLANTEAMIENTO DEL PROBLEMA

El control de calidad industrial está basado en la obtención de datos analíticos que describan total o parcialmente el producto que se está obteniendo. Dichos datos (transformados en “resultados” una vez integrados en el contexto adecuado) servirán para decidir acerca de la aceptación, precio, destino y uso que se podrá atribuir al producto manufacturado.

La gran mayoría de métodos instrumentales de análisis son, por su propia naturaleza (a excepción de volumetrías, gravimetrías, algunos métodos electroquímicos y otros isotópicos o basados en la medida de masas moleculares y/o atómicas), métodos relativos de medida en los cuales la propiedad analítica a determinar debe obtenerse mediante una medida que, *a priori*, poco o nada tiene que ver con dicha propiedad. Es necesario, pues, establecer una relación funcional entre la variable de interés y la señal o medida realizada. De esta forma, deberá establecerse antes de realizar cualquier proceso de cuantificación, una etapa de “calibrado” o “modelización”.

Se entenderá por **calibración** o **modelado** el proceso mediante el cual se establece una relación funcional entre una o varias variable/s de interés (p.ej., concentración, densidad, acidez, etc.) y una magnitud (señal) medible (*Beebe et al.*, 1998). En consecuencia, uno de los problemas más frecuentes en la Química Analítica es establecer procedimientos indirectos de medida vía modelos de calibración, o, dicho abreviadamente, “modelos” o “calibraciones”.

El **modelo de calibración** comprende el conjunto de gráficos, ecuaciones matemáticas y datos intermedios que establecen la relación entre la magnitud de la señal medida y la variable de interés (*Martens y Naes*, 1989; *González Dou*, 1991; *Bertran*, 1995). El modelo de calibración puede tener una base físico-química y matemática racional y bien establecida o bien ser puramente empírico. Por lo general, la calibración univariante suele tener un modelo de fondo determinista (ley de Lambert-Bouguer-Beer, ecuación de Ilkovic, etc) pero en el momento en el que se trabaja con más variables (calibración multivariante) lo más habitual es trabajar en un campo mucho más abstracto (*Martens y Naes*, 1989).

Además de las ideas previas que se puedan poseer acerca del problema, lo único cierto en la calibración multivariante es aceptar como punto de partida que la estructura no aleatoria tanto en las variables predictoras como en las predichas es

causada por "algo", aunque ese (esos) fenómeno(s) subyacente(s) sea(n) más o menos desconocido(s) para nosotros. La misión de la calibración multivariante (en su caso la univariante) es, en consecuencia, detectar ese fenómeno en la extensión en la que ello sea posible a partir de los datos de partida y modelizarlo mediante ecuaciones y/o relaciones empíricas que, una vez aplicadas a un nuevo problema experimental, determinen en qué extensión presenta dicho evento una determinada propiedad.

A lo largo del tiempo se ha ido tratando de dar solución a la pregunta ¿qué es un "buen" modelo de calibración?

De la bibliografía y la propia experiencia acumulada, se deduce que la respuesta trivial es la que se mantiene cierta: "*aquel que funcione*". Es decir, aquel que una vez establecido a partir de unas muestras (patrones) "históricas" permita determinar satisfactoriamente (es decir, dentro de unos márgenes de error) la(s) variable(s) de interés en muestras desconocidas y no consideradas hasta ese momento.

De acuerdo con Martens y Naes (Martens y Naes, 1989), todo modelo es "malo" ya que no es real. Al igual que los modelos que se desarrollan en la Química General, Química Cuántica o cualquier otra disciplina (como la Astronomía) los modelos no son "ciertos", aunque pueden ser válidos para unos determinados propósitos. Y es esto, justamente, lo que se busca aquí: modelos de calibración que sean capaces de ofrecer respuestas adecuadas a cada problema que se irá planteando.

Básicamente, nos encontraremos (como se indicará en diversas ocasiones) ante soluciones de compromiso ya que, en general, habrá que buscar un equilibrio entre simplicidad y adecuación. La simplicidad buscará la interpretabilidad y la "parsimonia" matemática; la adecuación (al uso) impondrá el suficiente realismo y descripción de cada problema particular.

¿Cómo crear los modelos de calibración? Si las consideraciones previas pueden considerarse epistemológicas, no es más etérea la respuesta a esta segunda pregunta ya que no hay recetas ni reglas generales. De los diversos métodos matemáticos que pueden aplicarse para alcanzar un modelo se obtiene como conclusión una sistemática de trabajo a seguir en todos los casos (esquemática en la **Figura 1**) y que se ha aplicado a lo largo de esta Memoria. En lo que sigue de este capítulo se hace un breve estudio de cada etapa indicada en este esquema.

En otras ocasiones el problema no consiste tanto en cuantificar el producto

como, más bien, en decidir si la muestra (como representante del lote de producción) tiene características comunes con otro/s conjunto/s de muestras dado/s. En este caso se estaría ante un problema de clasificación. Al respecto deberá recordarse que no sólo será interesante averiguar si la muestra se clasifica en alguno de los conjuntos ya conocidos de antemano sino en investigar si nos enfrentamos a una nueva clase (p.ej., de fármaco, de materia prima, etc.). La problemática de la clasificación también se ha abordado en esta Memoria y se introduce en este capítulo.

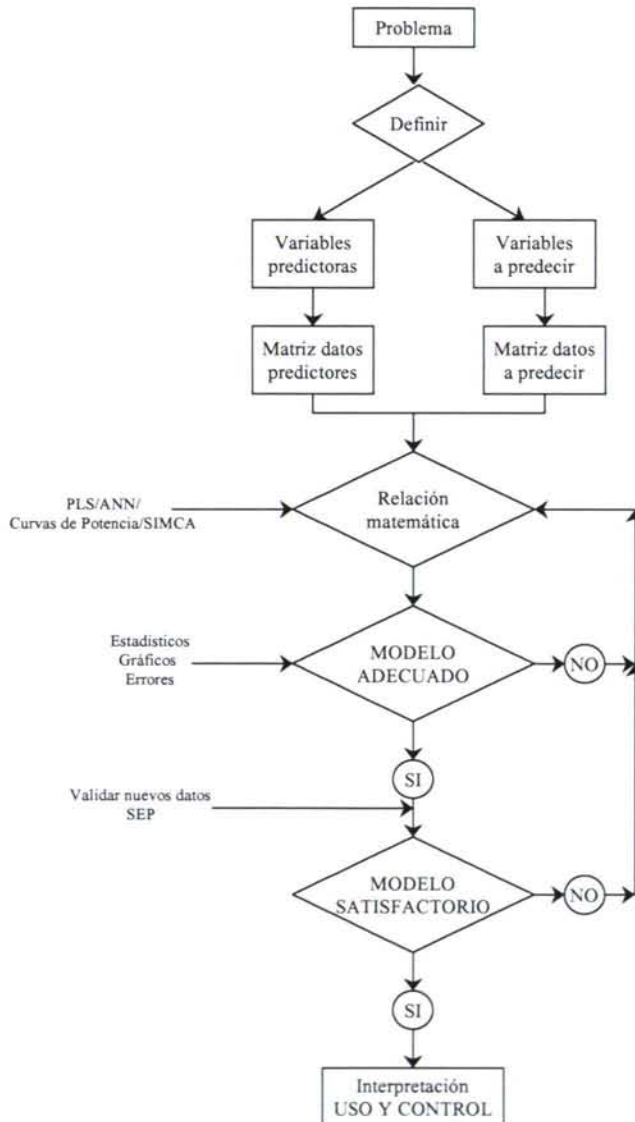


Figura 1: Sistemática de trabajo seguida para desarrollar modelos de calibración.

En cuanto a los modelos matemáticos a utilizar, las posibilidades son numerosas y la decisión ha sido aplicar cuatro técnicas que han demostrado su valía con anterioridad. Se dispone de un cierto conocimiento acerca de sus ventajas e inconvenientes y, además, ofrecen buenos resultados en situaciones complejas. Los métodos elegidos se conocen internacionalmente como curvas de potencia, modelado independiente de clases (*Soft Independent Modelling of Class Analogy* (SIMCA)), Regresión parcial por mínimos cuadrados (*Partial Least Squares* (PLS)) y redes de neuronas artificiales (*Artificial neural networks* (ANN ó RNA)).

Desde luego, ya se admite en este momento que hay muchos otros métodos que no se han aplicado y que también presentan características francamente interesantes. Algunos de ellos se citan en la recopilación bibliográfica.

2. VARIABLES PREDICTORAS Y VARIABLES A PREDECIR

Para abordar soluciones a los casos prácticos presentados en esta Memoria se emplearán modelos multivariantes (de calibración y de clasificación) puesto que la mayor parte de los métodos analíticos actuales pueden considerarse como multicanal y, en consecuencia, ofrecen una gran cantidad de información que, *a priori*, no debería descartarse. Las variables independientes o predictoras serán las absorbancias medidas a los números de onda que constituyan el espectro IR de las muestras; en términos Quimiométricos, se trata de una técnica que origina datos de orden uno (cada muestra un vector) (Boqué y Ferré, 2004).

Dado que los productos que se abordan son muy complejos, no parece factible emplear un sólo número de onda para predecir la(s) propiedad(es) de interés ya que ¿cuál debería elegirse? Está clara, en consecuencia, la necesidad de modelos matemáticos que trabajen con más de una variable. No obstante, se hará una breve incursión en los métodos de selección de variables por la importancia que éstos presentan. Las variables a predecir se irán detallando en cada caso particular pero, en general, serán los parámetros de interés cuyos valores se han determinado previamente en el conjunto de muestras de calibrado.

3. CURVAS DE POTENCIA

Las curvas ó funciones de potencia son métodos supervisados de modelado de clases que fueron definidas en un principio como métodos de clasificación

"estimadores de núcleos de densidad" ("*kernel density estimators*") por *Fix y Hodges (1951)* dentro del contexto del análisis discriminante. Más recientemente, *Coomans y Broeckert (1986)* describieron distintos métodos de curvas de potencia, incluyendo muchas aplicaciones. En su más sencilla aplicación, las curvas de potencia presentan la ventaja de ser más intuitivas que el tradicional análisis lineal discriminante (LDA) y de ofrecer representaciones gráficas que tienden a hacer más comprensible el éxito (o no) de la clasificación (algo que SIMCA no permite). Su fundamento teórico puede resumirse como sigue:

Inicialmente debe disponerse de un conjunto de muestras (llamado conjunto de aprendizaje o modelado), caracterizadas mediante los parámetros que se considere necesarios, sobre el que es posible establecer determinados grupos de acuerdo con algún criterio de interés (origen, calidad, especie, etc.).

En una primera etapa deberá comprobarse, sobre el conjunto de aprendizaje, la existencia de grupos bien definidos, según el criterio seleccionado. Una vez comprobada la existencia de grupos, se asigna una región del espacio como característica de cada grupo en particular. Tras esta definición, una nueva muestra podrá ser asignada a un determinado grupo en función de su proximidad a cada zona característica. Si de alguna forma es posible transformar esta proximidad (distancia) en probabilidad, la decisión queda simplificada y obvia la dicotomía del análisis discriminante. Por otra parte, si una muestra no presenta probabilidades de pertenencia aceptables para ningún grupo puede ser considerada como perteneciente a una nueva clase o bien detectar si se trata de una muestra anómala.

En cuanto al espacio de trabajo, aunque en una primera aproximación podría pensarse en emplear el definido por los parámetros o variables analíticas originales, generalmente dicho espacio no es el más adecuado para describir grupos bien definidos en las muestras, especialmente tratando situaciones multivariantes. Por esta razón es preferible utilizar el espacio de los Componentes Principales (PCs), en el que cada muestra estará descrita por sus coordenadas sobre los primeros componentes (*scores*). Un primer factor determinante del éxito final de la clasificación será, en consecuencia, la cantidad total de información recogida por los primeros PCs (generalmente los dos primeros).

Se han descrito distintas alternativas para realizar la transformación de distancia a probabilidad, superponiendo a cada muestra en el espacio considerado

diferentes funciones de potencial, bien sean distribuciones de probabilidad bien sean funciones de distribución libre, que permitan estimar la densidad de probabilidad (Forina et al., 1991). La función de potencial para cada clase viene entonces definida por la envolvente a las funciones muestrales consideradas y, en consecuencia, es posible determinar los límites de clasificación para diferentes niveles de probabilidad (curvas de isoprobabilidad).

En esta Memoria se ha aplicado un modelo simplificado para el cálculo de las curvas de potencia basado en admitir una curva de potencia para cada clase, es decir, considerar que si cada clase es homogénea, la envolvente puede aproximarse a una única distribución de probabilidad (Tomás y Andrade, 1997).

Bajo esta hipótesis de trabajo (cada grupo de muestras forma un conjunto homogéneamente distribuido) y considerando el espacio definido por los dos primeros PCs, se ha seleccionado como función de potencia la correspondiente distribución de Gauss bivariada que puede definirse mediante la expresión (Cuadras, 1981):

$$[\text{ec. 1}] \quad f(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{A}{2(1-\rho^2)}\right]$$

siendo:

$$A = \left[\frac{X-\mu_x}{\sigma_x}\right]^2 + \left[\frac{Y-\mu_y}{\sigma_y}\right]^2 - 2\rho\left[\frac{X-\mu_x}{\sigma_x}\right]\left[\frac{Y-\mu_y}{\sigma_y}\right]$$

donde X e Y corresponden a las coordenadas sobre el primer y segundo componente principal respectivamente; μ_x y μ_y a los valores medios de las coordenadas de cada grupo; σ_x y σ_y a las desviaciones estándar y ρ al coeficiente de correlación.

La relación incluida en la exponencial define elipses concéntricas de isoprobabilidad en el plano euclídeo:

$$[\text{ec. 2}] \quad a = \frac{1}{(1-\rho^2)} A$$

en la que a es una constante positiva y la probabilidad de que una muestra pertenezca a la clase o grupo para la que se ha definido la distribución puede calcularse como

(coincide con el área de la elipse) (Tomás y Andrade, 1999):

[ec. 3]

$$Prob[Muestra \in Clase] = 1 - \exp(-a/2) = Area$$

Si se van a clasificar nuevas muestras mediante este tipo de modelos deberán ser, primero, proyectadas en el espacio inicial PC1-PC2 (aunque también se podrían usar el PC2-PC3 ó PC1-PC3, según dónde se observen grupos); se recalculan las áreas y se asignan a un determinado grupo según las probabilidades de pertenencia así calculadas. Si una muestra no presenta altas probabilidades para ninguno de los grupos, podrá representar una nueva clase, categoría, etc.

4. SIMCA

La clasificación mediante la técnica SIMCA (*Soft Independent Modelling of Class Analogy*) o "modelado suave de la analogía de clases independientes" también es un ejemplo de una técnica supervisada de reconocimiento de pautas (modelos) usado, por ejemplo, para decidir si una o más muestras nuevas pertenecen a un grupo pre-existente de muestras. El objetivo principal del SIMCA es asignar un nuevo objeto a la clase con la cual muestre mayor similitud (analogía) (Beebe et al., 1998). De esta forma se clasifica el objeto dentro de la clase con la que tiene unas propiedades más comunes.

En SIMCA se lleva a cabo un análisis de componentes principales (PCA) independiente para cada clase (grupo) de muestras con objeto de modelizar su forma y posición. Se construye una "caja multidimensional" para cada clase y para predecir a qué "caja" (o región del espacio) pertenece cada nueva muestra a clasificar es necesario determinar qué región del espacio de medida ocupa.

Para construir las cajas multidimensionales se debe partir de un conjunto de muestras cuyas clases sean conocidas. El conjunto de muestras de entrenamiento se divide en subconjuntos separados, uno para cada clase, y se calculan los componentes principales por separado para cada una de las clases. El número de PCs es determinado para cada clase y el SIMCA se completa definiendo regiones para cada uno de los modelos de PCA. De ahí que este método, también sea conocido como clasificación mediante componentes principales locales. Los datos en cada clase son autoescalados o centrados en la media usando la media y la desviación estándar calculada sobre los

objetos en la clase (Liu et al., 1987):

$$[\text{ec. 4}] \quad \bar{x}_m^{(q)} = (1/N_q) \sum_{n=1}^{N_q} x_{nm}^{(q)}$$

$$[\text{ec. 5}] \quad s_m^{2(q)} = [1/(N_q - 1)] \sum_{n=1}^{N_q} (x_{nm}^{(q)} - \bar{x}_m^{(q)})^2$$

donde:

x_{nm} es el valor de la variable m medida en la muestra n ; N es el número de muestras y M es el número de variables de la matriz de datos $\mathbf{X}_{(N \times M)}$; y q es el número total de clases. Además, $\bar{x}_m^{(q)}$ es la media para la clase q y $s_m^{2(q)}$ es su desviación típica.

De tal forma que los datos autoescalados serán:

$$[\text{ec. 6}] \quad z_{nm}^{(q)} = \frac{(x_{nm}^{(q)} - \bar{x}_m^{(q)})}{s_m^{(q)}}$$

Si se centran en la media serán de la siguiente forma:

$$[\text{ec. 7}] \quad z_{nm}^{(q)} = (x_{nm}^{(q)} - \bar{x}_m^{(q)})$$

Después de escalar, se hace un PCA en el que se calculan los componentes principales (o, autovectores) (\mathbf{Y}) como combinaciones lineales de las variables originales, de tal forma que el primer PC ajusta tanto como le sea posible la varianza presente en las variables originales; el segundo PC ajusta tanto como le sea posible la varianza residual, y así sucesivamente. Los componentes principales formados tienen que ser ortogonales:

$$[\text{ec. 8}] \quad y_{nk}^{(q)} = \sum_{m=1}^M z_{nm}^{(q)} b_{mk}^{(q)} \quad (k = 1, \dots, K^{(q)})$$

donde y_{nk} (normalmente, llamados *scores*) es el valor del K componente principal para el objeto n y representa dónde se sitúa el objeto con respecto a esa nueva dirección; z_{nm} es el vector de datos de la muestra n , escalado usando la media y la desviación

estándar de la clase q , y b_{mk} son los coeficientes m (o *loadings*) del componente principal K de la matriz ($M \times M$) de correlaciones entre las variables. Los *scores* de los componentes principales calculados así tienen media cero y varianza igual al correspondiente autovalor λ_k . Si K_q componentes son retenidos en el modelo de la clase q , los datos se pueden describir en términos del modelo de componentes principales por:

[ec. 9]

$$z_{nm}^{(q)} = \sum y_{nk}^{(q)} b_{km}^{(q)} + e_{nm}^{(q)}$$

porque los componentes son ortogonales; y siendo e_{nm} los residuales de ajuste de los datos al modelo. La varianza residual en la clase calculada en la etapa de modelado será:

[ec. 10]

$$s_0^{2(q)} = [1/(N_q - K_q - 1)(M - K_q)] \sum_{n=1}^{N_q} \sum_{m=1}^M e_{nm}^2$$

que es la medida del "ajuste" o "estrechez" de la clase a las N muestras usadas en el proceso de aprendizaje ó modelado.

Para predecir la clase de una muestra desconocida, es necesario determinar qué región del espacio de medida ocupa. Matemáticamente, esto se hace proyectando el vector de medida de la muestra desconocida en cada uno de los modelos SIMCA. En la predicción puede ocurrir que la muestra sea miembro de una, varias o ninguna de las clases. Si es miembro de más de una clase, esto es debido a que el sistema de medida y los modelos SIMCA no tienen suficiente poder de discriminación para distinguir entre clases. Si la muestra de prueba no pertenece a ninguna clase puede ser debido a un error de medida o el resultado de alguna característica química inusual o desconocida, se considera una muestra "diferente" o, dependiendo de la situación, un anómalo (*outlier*).

Así, una muestra nueva (llámese t) podrá ser clasificada en alguna de las clases previas si se ajusta a alguno de los q modelos. Para ello, deberán calcularse sus *scores* (y_{tk}) a partir de los parámetros anteriores para cada uno de los grupos:

[ec. 11]

$$y_{ik}^{(q)} = \sum_{m=1}^M z_{im}^{(q)} b_{mk}^{(q)} \quad (k = 1, \dots, K_q)$$

Por tanto, la modelización de esta nueva muestra por el modelo previo vendrá dada por:

[ec. 12]

$$\hat{z}_{im}^{(q)} = \sum_{k=1}^{K_q} y_{ik}^{(q)} b_{km}^{(q)} \quad (m = 1, \dots, M)$$

Y, por tanto, la validez del ajuste de la muestra al modelo viene dado por la desviación estándar residual definida por:

[ec. 13]

$$s_t^{2(q)} = [1/(M - K_q)] \sum_{m=1}^M (z_{im}^{(q)} - \hat{z}_{im}^{(q)})^2$$

Dado que se dispone de dos varianzas (la del ajuste del modelo y la de ajuste de la muestra desconocida), se puede aplicar un test F de Fisher-Snedecor como una medida cuantitativa (y objetiva) para la clasificación siendo:

[ec. 14]

$$F_{calc} = \frac{s_t^{2(q)}}{s_0^{2(q)}} \text{ con } (M - K_q) \text{ y } (N_q - K_q - 1)(M - K_q) \text{ grados de libertad}$$

Este valor de F deberá compararse con el tabulado (al grado de confianza deseado) para $(M - K_q)$ y $(N_q - K_q - 1)(M - K_q)$ grados de libertad.

Una gran ventaja de SIMCA (que también comparten las curvas de potencia) es que, en contraste con otras metodologías, no es restringido en el número de variables (M) relativas al número de objetos (N). Los modelos de componentes principales pueden estimarse independientes de la relación entre M y N. La única restricción es que el número de componentes, K_q , tiene que ser más pequeño que ambos M y N. De hecho, la estabilidad de clasificación del SIMCA aumenta con la raíz cuadrada de M (Wold et al., 1981).

Para una mejor clasificación de las muestras mediante SIMCA se ha efectuado el “cierre de las cajas” por los extremos mediante el cálculo de un límite de confianza

obtenido definiendo un valor crítico de la distancia (Euclídea) del modelo. Esto es dado por:

$$[ec. 15] \quad S_{crit} = \sqrt{sF_{crit}}$$

siendo F_{crit} el valor tabulado para $(M-K_q)$ y $(N_q-K_q-1)(M-K_q)$ grados de libertad, generalmente al 95% de confianza. La S_{crit} se usa para determinar el espacio (el cilindro) alrededor de PC1 (ver **Figura 2a**) y los planos alrededor del plano PC1 y PC2 (ver **Figura 2b**). Los objetos con $s < s_{crit}$ pertenecen a la clase, el resto no (Massart et al., 1997).

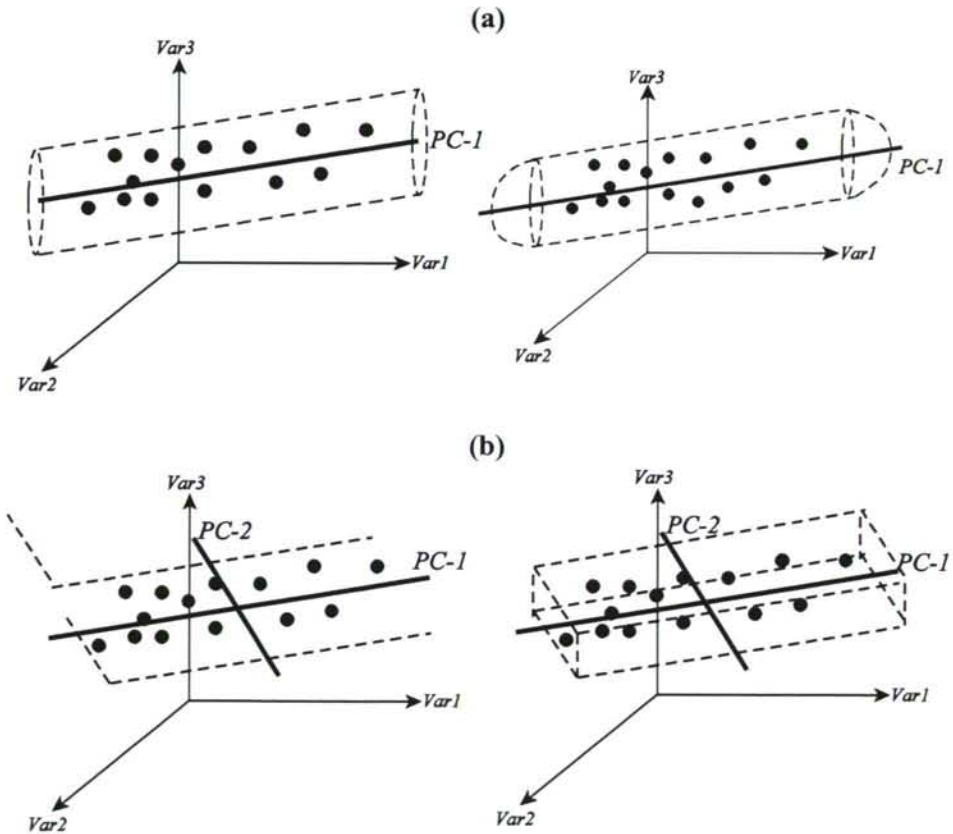


Figura 2: Cierre de la región del espacio en el modelo SIMCA para (a) un PC y (b) dos PCs.

En lo que se refiere a la **calibración** propiamente dicha, y a pesar de que el número de métodos de calibración multivariante es muy amplio, se suelen agrupar en dos grandes bloques. Los métodos "duros", rígidos o deterministas y los modelos "blandos" o flexibles. Existen otras clasificaciones (ver p.ej. *Martens y Naes, 1989; González Dou, 1991; Bertran, 1995; Beebe et al., 1998*) pero ésta es adecuada para los propósitos aquí perseguidos.

Los modelos duros son aquellos en los cuales las variables siguen un modelo pre-establecido (ejemplo típico sería la ley de Lambert-Bouguer-Beer). En ellos se establece una relación directa entre la magnitud de interés (Y) y las variables predictoras (X) del tipo: $y_1 = b_0 + b_{11}x_1 + b_{12}x_2 + \dots + \epsilon$, donde ϵ sería el error.

Los modelos flexibles son aquellos que no parten de una relación funcional y emplean los datos para llegar a una ecuación matemática (empírica) que relaciona las variables medidas con las que se quieren predecir. En este conjunto se encuadran típicamente todos los métodos basados en regresión por componentes principales, regresión parcial por mínimos cuadrados, calibración por redes neuronales, etc.

5. REGRESIÓN PARCIAL MEDIANTE MÍNIMOS CUADRADOS (PLS)

Uno de los métodos más usados actualmente en calibración multivariada es el de regresión parcial mediante mínimos cuadrados (PLS ó PLSR) (*Beebe et al., 1998; Sekulic et al., 1993; Hopke, 2003*), el cual ha ganado importancia en muchos campos de la Química (Analítica, Física, Clínica y control de procesos industriales) (*Geladi y Kowalski, 1986; Adams, 1995; Naes et al., 2002*).

El punto de partida conceptual sería buscar no sólo las direcciones de mayor información en el conjunto de las variables experimentales (o predictoras), matriz **X**, sino también seleccionar aquellas que guarden mayor relación con la(s) variable(s) a predecir, matriz **Y**. Ésta es, de hecho, la idea que subyace en la técnica de PLS, desarrollada por Herman Wold para la econometría a mediados de los años 70 (*Wold, 1975; Martens, 2001*). El propio Wold establece 1977 como el año de nacimiento del método de PLS (*Geladi, 1988*). Más tarde se aplicó en Quimiometría, donde ha conducido a buenos resultados (*Wold et al., 1983; Wold et al., 1984; Wold et al., 2001a; Naes y Risvik, 1996; Kowalski et al., 1982*).

PLS es, pues, una metodología matemática usada para establecer un modelo que relacione la información de dos conjuntos de datos diferentes (pero ligados de

alguna forma desconocida). El modelo de PLS intenta extraer la información importante de ambas fuentes con la única condición de que tales informaciones sean relevantes para establecer la relación entre ambos conjuntos de datos. Esto se consigue desarrollando un modelo de regresión sobre "variables observadas indirectamente". Cada una de estas nuevas "variables" (llamadas Variables Latentes -VL- o factores) se calcula como combinación lineal de las variables originales bajo la condición de que estén correlacionadas con la variable (o variables) a predecir. De esta forma, los modelos de PLS se rigen por un criterio de capacidad predictiva más que por el ajuste del modelo a los datos (Veltkamp y Gentry, 1988). Por esta razón, PLS suele conducir a mejores predicciones con menos factores (variables latentes) que PCR (Naes, 1987; Lorber et al., 1987).

Esto se logra gracias a que los algoritmos empleados relacionan iterativamente la información relevante en la matriz \mathbf{X} con la de \mathbf{Y} , siempre y cuando el modelo mejore su capacidad predictiva. Lo cierto es que la explicación del algoritmo de PLS no es intuitiva y debe hacerse mediante el análisis matemático (aquí, necesariamente, breve) del mismo. A la matriz de las variables predictoras -p.ej., los espectros- se les suele llamar Bloque-X y a la matriz (vector) de variables predichas, Bloque-Y.

Se explican los algoritmos de H.Wold, aplicados por su hijo Svante Wold y basados en el algoritmo NIPALS (Non-Iterative Partial Least Squares). Hay también unas ligeras diferencias entre la escuela de S.Wold y la de H. Martens en las cuales no se entrará (ver Geladi, 1988, para más información) pero que conducen a iguales capacidades predictivas.

La **Figura 3** (tomada de Wold et al., 1987) ilustra las explicaciones y el siguiente resumen se adaptó de Geladi y Veltkamp (Geladi y Kowalski, 1986; Veltkamp y Gentry, 1988) por su carácter pedagógico. Por otro lado, este tipo de cálculo sólo tiene sentido contemplarlo desde el punto de vista de la iteración mediante cálculo con ordenadores.

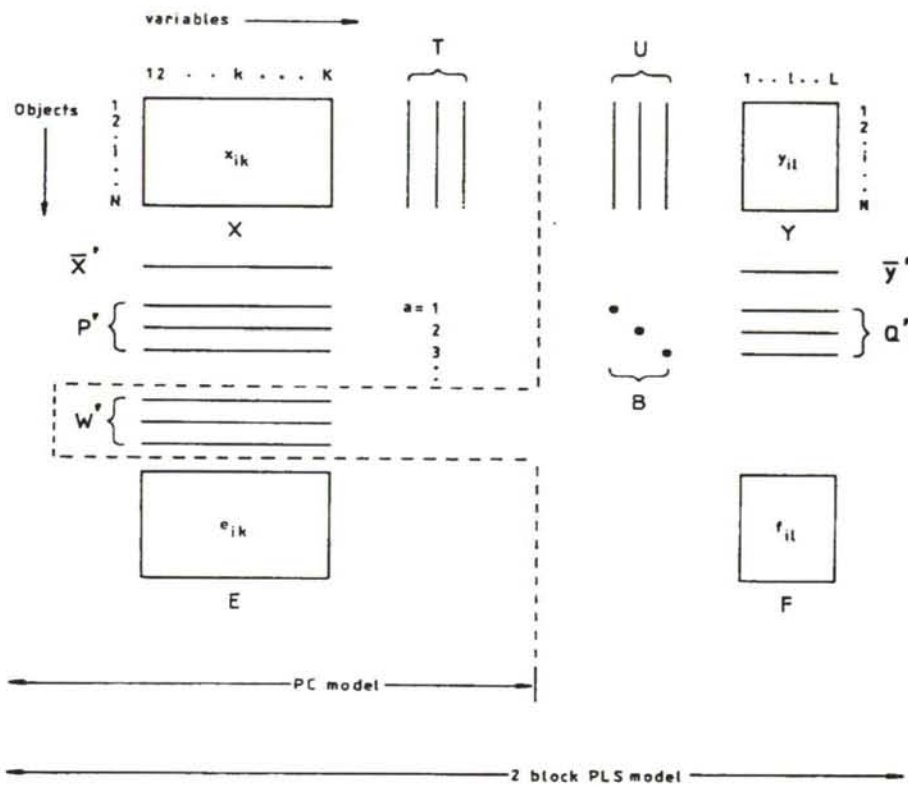


Figura 3: Representación gráfica del algoritmo PLS, empleando NIPALS (Wold et al., 1987).

El modelo de PLS puede considerarse formado a partir de dos relaciones externas (de los dos bloques X e Y , individualmente) y una relación interna que liga ambos bloques (ésta es la clave de PLS). Las relaciones externas no son más que las conocidas descomposiciones en componentes principales (los sumatorios van desde $h=1, 2, \dots, h$ componentes principales). Se asume que X e Y han sido centradas en la media o autoescaladas.

$$[ec. 16] \quad X = TP' + E = \sum t_h p'_h + E$$

$$[ec. 17] \quad Y = UQ' + F = \sum u_h q'_h + F$$

$$[y = Tq' + f, \text{ en el caso de predecir una sólo variable}]$$

donde: X , matriz de variables predictoras originales

Y , matriz (vector, en su caso) de variables dependientes.

T y P , matrices de *scores* y *loadings*, respectivamente, para X .

U y Q , matrices de *scores* y *loadings*, respectivamente, para Y .

E y F , matrices de error asociado.

Dado que la intención es describir Y tan bien como sea posible, eso significa minimizar F y, al mismo tiempo, obtener una relación útil entre X e Y . La relación interna consigue esto mediante la relación de los *scores* t y u . El modelo más sencillo es una relación lineal del tipo

$$[\text{ec. 18}] \quad \hat{u} = b_h t_h$$

donde $b_h = u'_h t_h / t'_h t_h$. De forma que los vectores b juegan el papel de coeficientes de regresión.

De todas formas, éste no es el mejor modelo posible ya que los componentes principales se han calculado de forma independiente para las dos matrices X e Y . Sería mejor darles información relativa de la una a la otra, de forma que se logre una especie de “rotación” que haga que ambos subespacios se acerquen (en la misma idea de la rotación de Procrustes que se abordan en el capítulo siguiente).

Un modelo simplificado es el que sigue, explicado como pseudo-código:

Para el bloque X la descomposición PCA sería:

- (1) tomar $t_{\text{inicio}} = \text{cualquier } x_j$
- (2) hacer $p' = t'X/t't (=u'X/u'u)$
- (3) $p'_{\text{nuevo}} = p'_{\text{viejo}} / \|p'_{\text{viejo}}\|$ donde $\| \|$ es la norma
- (4) $t = Xp/p'p$
- (5) se comparan los vectores t de las etapas 2 y 4; si son iguales, se finaliza, si son diferentes (dentro de un criterio de redondeo), se vuelve a (2)

Para el bloque Y la descomposición PCA es análoga:

- (1) tomar $u_{\text{inicio}} = \text{cualquier } y_j$
- (2) hacer $q' = u'Y/u'u (=t'Y/t't)$
- (3) $q'_{\text{nuevo}} = q'_{\text{viejo}} / \|q'_{\text{viejo}}\|$ donde $\| \|$ es la norma
- (4) $u = Yq/q'q$
- (5) se comparan los vectores u de las etapas 2 y 4; si son iguales, se finaliza, si son diferentes (dentro de un criterio de redondeo), se vuelve a (2).

Mejora de la relación interna

Como se indicó, ésta es una forma de obtener información de cada bloque por separado. La manera de que un bloque "conozca" algo del otro es "cruzar" la información intercambiando los vectores \mathbf{t} y \mathbf{u} , para lo cual se colocaron ya las expresiones entre paréntesis en las etapas (2). De esta forma, el algoritmo quedaría:

- (1) tomar $\mathbf{u}_{\text{inicio}} = \text{cualquier } \mathbf{y}_j$
- (2) hacer $\mathbf{p}' = \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u}$ ($\mathbf{w}' = \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u}$)
- (3) $\mathbf{p}'_{\text{nuevo}} = \mathbf{p}'_{\text{viejo}}/\|\mathbf{p}'_{\text{viejo}}\|$ ($\mathbf{w}'_{\text{nuevo}} = \mathbf{w}'_{\text{viejo}}/\|\mathbf{w}'_{\text{viejo}}\|$)
- (4) $\mathbf{t} = \mathbf{X}\mathbf{p}/\mathbf{p}'\mathbf{p}$ ($\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}'\mathbf{w}$)
- (5) $\mathbf{q}' = \mathbf{t}'\mathbf{Y}/\mathbf{t}'\mathbf{t}$
- (6) $\mathbf{q}'_{\text{nuevo}} = \mathbf{q}'_{\text{viejo}}/\|\mathbf{q}'_{\text{viejo}}\|$
- (7) $\mathbf{u} = \mathbf{Y}\mathbf{q}/\mathbf{q}'\mathbf{q}$

(8) se comparan los vectores \mathbf{t} de la etapa 4 con los de la iteración precedente. Si coinciden, dentro de un error de redondeo, ya hemos finalizado; de lo contrario se repiten iteraciones desde (2).

Si el bloque Y tiene una sólo variable (PLS-1 bloque), se evitan las etapas 5 a 8, colocando $q=1$. La gran ventaja de este algoritmo iterativo es que converge rápidamente.

Matemáticamente hablando hay un problema relacionado con que los vectores \mathbf{t} así obtenidos no son ortogonales, motivo por el cual los vectores \mathbf{p}' no es correcto considerarlos como "loadings" y se les llama pesos (\mathbf{w}' , de ahí las ecuaciones entre paréntesis de los pasos 2, 3 y 4), pero tienen su mismo significado e interpretación (esto es, participación de cada variable original en el componente -que, ahora, se llamará variable latente (VL)-). Aunque no es necesario, si se desea (es lo habitual), se puede lograr la ortogonalidad de los scores, \mathbf{t} , haciendo $\mathbf{p}' = \mathbf{t}'\mathbf{X}/\mathbf{t}'\mathbf{t}$.

Además de la relación lineal entre los vectores \mathbf{t} y \mathbf{u} se pueden establecer relaciones polinomiales (habitualmente no son mayores de orden 2 ó 3) (Wold et al., 1989; Wise y Gallagher, 1998), lo cual aporta variantes interesantes a la técnica de PLS (especialmente para el modelado de relaciones no lineales).

A pesar de ello en esta Memoria no se aplicarán regresiones polinomiales ya que se ha comprobado en numerosas ocasiones que el modelo de PLS lineal es capaz de modelar satisfactoriamente incluso relaciones no lineales, siempre que no sean extremas (Wise y Gallagher, 1998). En algunos casos la capacidad de predicción de

modelos PLS y de redes de neuronas artificiales (RNA) son comparables, lo cual sugeriría que o bien no hay no linealidades o bien que PLS llega a modelarlas tan bien como las redes neuronales (ampliamente sugeridas para problemas con fuertes no-linealidades) (Jacobson y Hagman, 1993; Andrade et al., 1999; Hadjiiski et al., 1999; Yang et al., 2003; Blanco et al., 2000a).

5.1. VENTAJAS DE PLS

Las variables latentes (VL) se extraen de forma que sucesivamente expliquen menos información, por lo cual llegará un momento en que se pueda dejar de introducir variables latentes en el modelo sin perjudicar los resultados finales. Una de sus ventajas es que son ortogonales, lo que puede simplificar la interpretación final del modelo.

El hecho de trabajar con variables latentes permite "visualizar" las muestras en este nuevo subespacio. Por similitud con PCA, a las proyecciones de las muestras en las VL se les llamará *scores*. También las variables se pueden analizar con una cierta facilidad ya que cada VL recoge las relaciones existentes entre ellas y cada variable (original) tiene un *loading* o peso que describe su contribución a la VL en cuestión. Tanto los *scores* como los pesos son muy útiles para estudiar los modelos de PLS.

Una característica esencial de PLS estriba en que puede reducir la influencia de factores dominantes (p. ej. agua en una muestra) pero irrelevantes para el modelo (p.ej. porcentaje de proteína en leche) y, en algunos casos, se reduce la dimensionalidad, lo que hace que la interpretación sea más fácil (MacLaurin et al., 1993). Una de las mayores ventajas de PLS es que puede desarrollar modelos que predigan simultáneamente varios parámetros (a diferencia de los no basados en este método que sólo pueden predecir un parámetro por modelo). Mediante PLS sóloamente hay que establecer también las VL en el bloque de las Y (como se indicó en el algoritmo). Al procedimiento se le denomina PLS de 2 bloques o, simplemente, PLS2 (MacLaurin et al., 1993; Brereton, 2000). Si se predice una sóloa Y, se denomina PLS1 (Brereton, 2000).

5.2. INCONVENIENTES

También existen inconvenientes al aplicar la técnica de PLS. El primero es que la interpretación del modelo puede ser difícil, aunque es cierto que las características de ruido se encuadran en las últimas variables latentes y no deben

usarse en las modelizaciones.

Otro inconveniente es que, a pesar de la sofisticación de los algoritmos, la base está en ajustes por mínimos cuadrados. Por lo que en casos donde existe una excesiva colinealidad entre las variables, los resultados pueden no ser buenos (Todeschini, 1995). Análogamente cabe esperar problemas si hay muestras con comportamiento anómalo (ya sea en las X o las Y).

5.3. TRANSFORMACIÓN, ESCALADO Y CENTRADO EN LA MEDIA

Frecuentemente, antes del análisis, las variables X e Y son transformadas para hacer sus distribuciones equitativamente simétricas y evitar problemas con las escalas de las variables en la matriz X , especialmente si se usan diferentes tipos de propiedades. Con un escalado apropiado, uno puede enfocar el modelo y usar la experiencia para dar más peso a algunas variables X . A pesar de que el proceso de escalado puede distorsionar la situación relativa de las muestras en el espacio de los componentes principales (Ortiz y Sarabia, 1994) el escalado previo tiende a facilitar la interpretación, simplificar el modelo y aportar estabilidad numérica (Wold et al., 2001a). Como regla general se recomienda centrar los datos en la media si todas las variables están expresadas en las mismas unidades y no hay diferencias de órdenes de magnitud en sus valores. Caso contrario se recomienda autoescalar. No obstante, no se puede descartar otro tipo de escalados que, en situaciones complejas, pueden resultar más beneficiosa. Por ejemplo, en espectroscopia IR-Raman en esta Memoria se han empleado combinaciones de dos escalados diferentes, normalización a la unidad y posterior centrado en la media (Andrade et al., 2003), mientras que en espectroscopia IR se han empleado tanto centrado en la media como autoescalado según la propiedad a estudiar como se verá posteriormente (Gómez-Carracedo et al., 2003a; Gómez-Carracedo et al., 2003b). El problema del escalado previo/pretratamiento de datos es muy complejo como lo demuestra el que en la actualidad algunos métodos novedosos para el cálculo de parámetros analíticos multivariantes (sensibilidad y selectividad) dependen de ellos (Faber, 1999).

5.4. SELECCIÓN DEL NÚMERO ADECUADO DE VL Y VALIDACIÓN DEL MODELO

Existen dos tipos de riesgo: el “infraajuste” (*underfitting*) y el “sobreaajuste” (*overfitting*). En el primer caso el modelo no dispondría de la suficiente información

como para generalizar; es decir, predecir muestras que antes no había recibido. En el segundo caso (más frecuente que el primero al trabajar con variables espectrales), el modelo dispone de una excesiva cantidad de información que no representa la generalidad del problema sino que se asocia, preferentemente, a peculiaridades de las muestras empleadas en el calibrado. En consecuencia, en cualquier modelo empírico con fines predictivos es esencial determinar la correcta complejidad (dimensionalidad) del modelo y controlar estrictamente la capacidad predictiva de cada VL (o cada conjunto) y, por tanto, cesar de incrementar su número cuando los modelos empiezan a no mostrar mejorías.

La Cross-Validación o “validación cruzada” (CV, *cross-validation*) es un camino práctico para validar la capacidad predictiva (Höskuldsson, 1988; Tenenhaus, 1998). Es, en general, un método de trabajo que valora objetivamente la magnitud de los errores de predicción (Thomas, 1994) aunque se ha demostrado que puede llevar a modelos de calibración sobreajustados y presenta algunos problemas (Helland, 2001).

Básicamente, la CV se realiza dividiendo el grupo de calibración en dos subgrupos, uno de entrenamiento para construir el modelo de regresión y uno de predicción. Para la CV *leave-one-out* (CV-LOO) usada en esta Memoria, una muestra i es extraída y el resto de las muestras se emplean para desarrollar un modelo que se utilizará para determinar el parámetro de interés de la muestra extraída. Por repetición de este proceso para todas las muestras es posible computar el estadístico “raíz cuadrada del error promedio de CV (RMSECV)” (Lorber y Kowalski, 1990). El error promedio es la media de las diferencias entre los valores predichos y los reales en cada caso, al cuadrado. En algunas ocasiones se habla de la “suma de cuadrados de los residuales predichos (PRESS)” (Wise y Gallagher, 1998) la cual, esencialmente, es lo mismo que el RMSECV sin promediar. Además, mientras el RMSECV aporta información acerca de la capacidad predictiva futura, el PRESS se orienta más a estimar el error de ajuste en función del número de VL consideradas (Veltkamp y Gentry, 1988). A pesar de este matiz PRESS y RMSECV se emplean indistintamente en numerosas ocasiones. En cualquiera de sus denominaciones, el test se utiliza para estimar la capacidad predictiva del modelo (Davies, 1998a; Centner et al., 2000; Esvensen et al., 1994; Estienne et al., 2001).

[ec. 19]

$$RMSECV = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

[ec. 20]

$$PRESS_{VL} = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

En los modelos PLS asumimos que el sistema o proceso investigado es influenciado por tan sólo unas pocas VL (Wold *et al.*, 2001a; Wold *et al.*, 2001b) o, dicho de otra forma, por unas características generales que constituyen la información principal. Obviamente, el parámetro fundamental a la hora de trabajar con PLS es el número de VL (variables latentes o factores) a incluir en el modelo. Es un tema aún abierto en la Quimiometría teórica pero que, a nivel práctico, se soluciona de forma pragmática (según el proceso seguido en esta Memoria).

Cuando no se dispone de información previa lo más sensato es probar diferentes modelos en los cuales la única diferencia sea el número de VL consideradas. Al representar PRESS o RMSECV frente al número de VL en el modelo se obtiene un mínimo en el modelo con capacidades predictivas más adecuadas (ver **Figura 4**). La práctica habitual es emplear el PRESS y/o RMSECV cuando se trabaja con muestras de modelado. Cuando se predice el conjunto de validación independiente el estadístico recibe el nombre de RMSEP (raíz cuadrada del error de predicción) o, abreviadamente, SEP, que es una estimación de la incertidumbre de todas las predicciones futuras hechas por el modelo (Semells *et al.*, 2004).

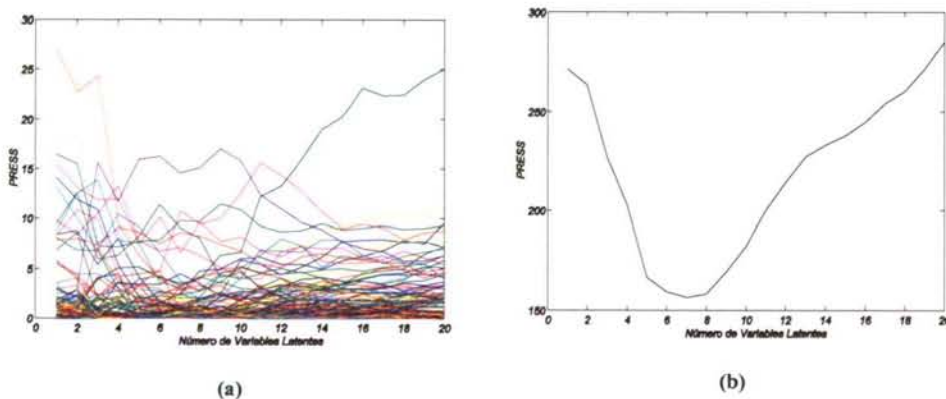


Figura 4: Curva de: (a) Error cometido en la predicción de cada muestra en el proceso de CV-LOO vs VL, (b) PRESS acumulativo vs VL. El mínimo indica el modelo de elección.

Números de VL inferiores al mínimo producen falta de ajuste (*underfitting*) y valores mayores, un exceso de ajuste (*overfitting*).

Además de la opción de la CV-LOO, existen otras posibilidades muy parecidas pero en las cuales se elimina un determinado número de muestras del conjunto de calibración; el número de muestras dejado fuera es lo que se llama segmento de validación ó cross-validación segmentada (Bertran, 1995; Naes et al., 2002). El caso extremo es, precisamente, la CV-LOO. El principal inconveniente de CV-LOO es el gran volumen de cálculo asociado y el alto tiempo de computación cuando el número de muestras y variables es elevado.

A pesar de esta "validación interna" se ha observado que la "mejor" validación debe consistir en tomar un conjunto de muestras independientes del aprendizaje (teniendo presente que también es conveniente que haya una demora en el tiempo de producción) y predecirlas empleando los modelos (el modelo) que hayan llegado a esta "fase final" (Andrade et al., 1999; Yang et al., 2003; Dieterle et al., 2003; Ruckebusch et al., 2002).

La preparación de un conjunto de validación en el campo del control de calidad industrial no suele presentar grandes problemas ya que tanto las muestras como los valores de referencia se obtienen sin grandes dificultades por el simple hecho de que la producción y el control continúan; en otros casos más relacionados con la investigación pura, la situación puede no ser tan sencilla. Esto último "justificaría" que en una gran cantidad de trabajos no se presenta grupo de validación o se obtiene entresacando las muestras del que ya se disponía en un principio. Se ha visto que esta última opción produce resultados excesivamente optimistas y poco útiles (Davies, 2004). Además, debe tenerse en cuenta que la existencia de un conjunto nuevo de validación resulta indispensable para el modelado con Redes Neuronales Artificiales (como se indicará más adelante).

Una vez validado el modelo, conviene estudiar cómo se comporta en el tiempo antes de disponerlo para el uso rutinario. A esta etapa se le ha denominado "Seguimiento de la operatividad del modelo" y también ha sido destacada por algunos autores (Martens y Naes, 1989; Bertran, 1995), Kvalheim y Karstang (1989) hablan de realizar incluso una "validación química".

Relacionado con este tema está el hecho estudiado en esta Memoria de que las calibraciones que precisan modelos muy complejos (alto número de VL) suelen

tener algún tipo de problemas en los datos de partida de calibración (ya sea en el bloque **X**, ya en el **Y**). También se ha observado que aunque una alta complejidad (alto número de VL) permita una buena predicción de un pequeño conjunto de muestras (del conjunto de aprendizaje o uno de prueba) no necesariamente es ese modelo el más adecuado ante las nuevas muestras (p.ej., espaciadas en el tiempo). En este sentido, los modelos muy complejos han tenido un comportamiento de tipo *overfitting*. Así, «es mejor perder precisión en la predicción de las muestras presentes para ganar capacidad de predicción sobre un rango más amplio de muestras similares» (Martens et al., 1987).

Ian Wakeling (Wakeling y Morris, 1993) ha estudiado el índice de correlación al cuadrado (r^2) obtenido por cross-validación mediante estudios de Monte Carlo para determinar hasta qué punto una determinada correlación es debida a fenómenos aleatorios (en función del número de variables y muestras). Los valores tabulados sirven para decidir si el valor r^2 se debe al azar o no.

Sea n , un número de variables latentes y r_{cv}^2 el coeficiente de regresión (al cuadrado) obtenido por cross-validación. Para tratar de ajustar más la selección del número de variables latentes, empíricamente, se ha observado que una representación de $r_{cv,n}^2 - r_{cv,n-1}^2$ alcanza un máximo en el entorno del número óptimo de variables latentes. El fenómeno es normal dado que r_{cv}^2 está inversamente relacionado con el PRESS (Wakeling, I., *Comunicación privada*); de esta forma, si las gráficas de PRESS presentan un mínimo, la diferencia $r_{cv,n}^2 - r_{cv,n-1}^2$ alcanza un máximo.

Una vez realizadas pruebas previas y teniendo esta consideración en cuenta, se descartó el uso de este estadístico por ser demasiado conservativo (todos los modelos eran altamente significativos) y ser parcialmente redundante con el de PRESS. Si bien en otras aplicaciones podría resultar interesante.

5.4.1. ESTRATEGIA DE TRABAJO EN PLS

De acuerdo con lo apuntado hasta ahora, la **Figura 5** recoge la sistemática de trabajo seguida en esta Memoria para el desarrollo de los modelos de calibración mediante la técnica de PLS. Se ha empleado el software base aportado por PLS-Toolbox (Eigenvector Technologies, versión 4.2.c., Washington, USA) y se ha complementado con programas propios desarrollados en Matlab (The Mathworks, v. 4.2c.1., Natick, MA, USA).

Como se ve, el trabajo sobre validación es intenso ya que no sólo se realiza una primera validación interna sino que se establece una primera validación con un grupo de muestras diferentes y un segundo grupo de validación que, en realidad, es el "seguimiento de la operatividad del modelo", usándolo de forma paralela a los métodos normalizados.

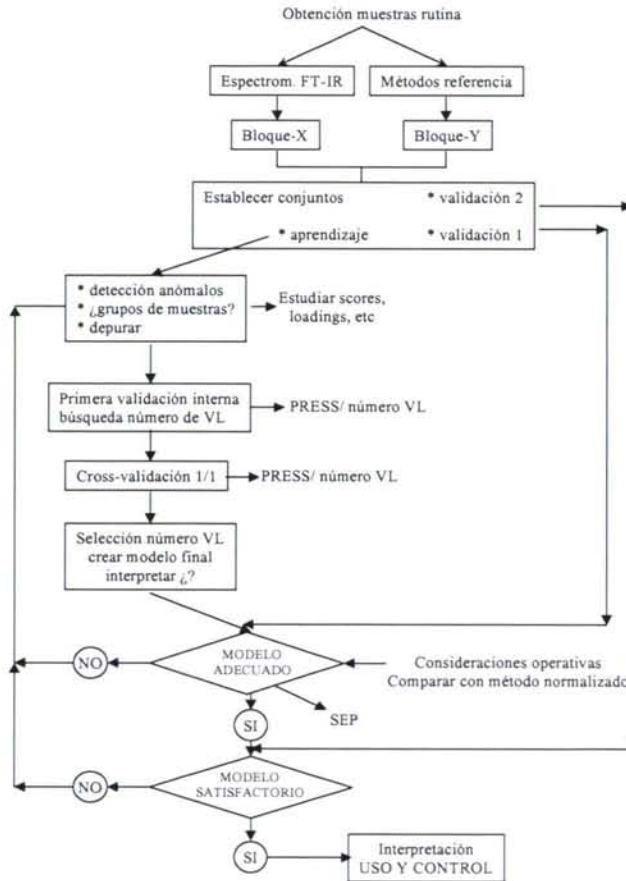


Figura 5: Estrategia de trabajo en PLS.

5.5. ACTUALIZACIÓN DEL MODELO

Una vez seleccionado el modelo "definitivo", se van prediciendo muestras nuevas que se van obteniendo de las distintas Unidades de Producción implicadas. Como las unidades no siempre trabajan bajo las mismas condiciones operativas, las muestras estudiadas presentarán variaciones (más o menos importantes) con el tiempo

que pueden crear discrepancias con el modelo que se tiene establecido. Cuando ocurre esto, las predicciones empiezan a desviarse y se hace necesario introducir en la matriz de calibración muestras nuevas de forma que el modelo las pueda tener en cuenta y las predicciones sean otra vez óptimas. En esta Memoria a este proceso se le ha denominado actualización.

En este sentido es importante que la matriz de calibración incluya la máxima variabilidad posible lo cual implicaría que a nivel industrial se necesita recoger muestras para la matriz de calibración durante un período de tiempo relativamente extenso para intentar incluir todas las posibles combinaciones de crudos, mezclas utilizadas y condiciones de operación. Por ejemplo, Lang (Lang, 1994) indica que no hay un modelo de calibración universal que elimine la necesidad de actualizar. Es escasa la bibliografía que recoge este aspecto de la actualización del modelo de regresión (García-Mencía et al., 2000), ya que, generalmente, se utiliza un grupo de muestras para calibrar y validar pero el estudio se queda ahí, sin que haya un seguimiento de la capacidad del modelo para predecir nuevas muestras en el uso rutinario de los modelos implantados.

5.6. INTERPRETACIÓN Y DIAGNÓSTICOS DEL MODELO DE PLS

El conjunto de las VL que se utilizan en los modelos PLS puede ser interpretado geoméricamente como una hiper-superficie en los espacios donde las coordenadas de las observaciones vienen dadas por las VL (Wold et al., 2001b), de forma análoga a la interpretación de las componentes principales. Como la regresión PLS es un proceso complejo, necesitamos un grupo de herramientas (gráficos) para ayudarnos en la búsqueda del mejor modelo (Davies, 1998b; Davies, 1998c).

Tal como se citó al definir la técnica, los “loadings” (“pesos”) para cada factor indican la contribución de cada variable a la calibración (Wold et al., 2001a; Davies, 1998b), aunque el estudio de los coeficientes de regresión también es interesante para evaluar las variables más importantes en el modelo final. Los *loading* informan sobre cómo se combinan las variables para formar la relación cuantitativa entre X e Y .

Los *scores* contienen información sobre las muestras y sus similitudes/diferencias con respecto al problema dado y al modelo (Wold et al., 2001a).

Hay varios test y representaciones gráficas que diagnostican la existencia o no de muestras anómalas y la bondad del modelo (Beebe et al., 1998; Wise y Gallagher, 1998;

Fernández Pierna et al., 2002; Fernández Pierna et al., 2003). En el caso particular de esta Memoria los test que se han empleado son los que se muestran a continuación:

a) Test T^2 : es la extensión multivariante del test t-student, muestras con un valor T^2 superior al límite tabulado son sospechosas de comportamiento anómalo. El T^2 de Hotelling es una medida de la variación de cada muestra dentro del modelo PCA.

$$[ec. 21] \quad T_i^2 = t_i \lambda^{-1} t_i' = x_i P \lambda^{-1} P' x_i'$$

donde t_i es la fila i-ésima de T_k , la matriz de k scores del modelo de PCA y λ^{-1} es la inversa de la matriz diagonal de los autovalores asociados a los k autovectores (PC) retenidos en el modelo (Wise y Gallagher, 1998).

b) El estadístico Q es una medida de la diferencia, o residual, entre una muestra y su proyección en el k componente principal retenido en el modelo:

$$[ec. 22] \quad Q_i = e_i e_i' = x_i (I - P_k P_k') x_i'$$

donde e_i es la muestra i-ésima de E , siendo E la matriz residual (ver ec. 16), P_k es la matriz de los loadings retenidos en el modelo PCA (donde cada vector es una columna de P_k) e I es la matriz identidad de tamaño apropiado (n x n) (Wise y Gallagher, 1998).

c) Leverage: es una medida de la diferencia que hay entre la muestra i-ésima con respecto a otros grupos de datos (Beebe et al., 1998; Faber, 1999), normalmente se emplea para evaluar la similitud de las muestras desconocidas (conjuntos de validación) a las de calibración y con ello evaluar si hay muestras a predecir que sean muy diferentes (anómalas). Los valores críticos se sitúan en torno a ± 3 . Los “anómalos” en el espacio X pueden ajustarse bien al modelo (anómalo “bueno”) o no (anómalo “malo”). Cuando los datos predichos no son anormales para el objeto, pero ésta se ajusta poco al modelo, se puede hablar de alta observación residual (anómalos en la dirección Y) (Massart et al., 1997). Otro tipo de anómalos serían las observaciones “influenciales”, éstas serían espectros que tienen un alto impacto en la estimación del modelo (cuando se descartan del grupo de calibración se observa un modelo diferente con predicciones distintas).

d) Error estudentizado: La representación del error estudentizado frente al

leverage es una gráfica muy útil para determinar si la muestra es un anómalo en el bloque de las X y en el bloque de las Y. Si una muestra tiene *leverage* pequeño y error estudentizado alto indica que el espectro es muy similar al bloque de calibración pero debe existir algún problema en el bloque de las Y (predicción). La situación contraria también es cierta. Habitualmente los valores críticos se establecen en 3 para el *leverage* y 2.5 para el error estudentizado. En consecuencia las muestras que sobrepasan los límites deberían ser consideradas como anómalas.

El error estudentizado para cada muestra se puede definir de forma sencilla como el cociente entre el error residual (valor real- valor predicho) y una estimación de la varianza asociada a todos los residuales. Esto se lleva a cabo a través de la matriz “hat” (la matriz pseudoinversa de Moore-Penrose) (Veltkamp y Gentry, 1988; <http://su.wikipedia.org>, 2004; Sarabia y Cruz Ortiz, 2004). e) Distancia de Mahalanobis (Fernández Pierna, 2002): se deriva de la matriz de covarianza, se utiliza para evaluar la similitud (distancia) de una muestra desconocida al promedio de muestras (en general considerando este promedio a partir de las muestras de calibración). Un valor alto de la distancia de Mahalanobis indica que el punto se aleja del centro de la nube y, por tanto, es una posible observación influyente a priori.

f) t vs t: *scores* de las muestras en el bloque X. Esta gráfica se utiliza para estudiar la distribución de las muestras de calibrado y detectar grupos ó anómalos.

g) t vs u: *scores* de las muestras (bloque X) frente a *scores* de las mismas muestras en el bloque Y. Se utiliza para estudiar cómo se distribuyen las muestras en lo que se refiere a la relación que existe entre los bloques X e Y. Si, por ejemplo, se está calibrando la disposición de las muestras debería mostrar una tendencia lineal en aquellas VL que contribuyen al modelo. También se usa para identificar muestras anómalas.

h) y_{real} vs y_{pred}: valor real frente al valor predicho. Permite evaluar la bondad del modelo en cuanto a su capacidad de ajuste a las muestras de calibración y, en parte, evaluar su capacidad predictiva. Permite determinar si alguna/s muestra/s puede tener una influencia excesiva en el modelo. Como la recta teórica es la bisectriz, se puede llevar a cabo una evaluación visual de la bondad global; es decir, falta de sesgo.

i) Los residuales de la predicción también son un diagnóstico de interés ya que

son las distancias entre los valores experimentales observados y los predichos mediante el modelo (Wold *et al.*, 2001b). Grandes residuales Y indican que el modelo es pobre, y al poder estudiar si la distribución de los residuales es normal para una variable Y determinada, se pueden identificar outliers (Miller, 1993) en la relación entre T e Y (Wold *et al.*, 2001a).

j) El test F conjunto para pendiente y ordenada en el origen también se emplea como verificación de que el modelo no presenta sesgo, es decir, que es veraz (*trueness*) (test de hipótesis con la hipótesis nula de que la pendiente y ordenada en el origen son estadísticamente iguales a 1 y 0, simultáneamente). La F_{exp} se calcula según la siguiente ecuación:

$$[ec. 23] \quad F_{exp} = \frac{(\beta_0 - b_0)^2 + 2\bar{x}(\beta_0 - b_0)(\beta_1 - b_1) + (\sum x_i^2 / n)(\beta_1 - b_1)^2}{2(S_e^2/n)}$$

donde β_0 es la ordenada en el origen teórica (en nuestro caso es 0), b_0 es la ordenada en el origen real, β_1 es la pendiente teórica (en nuestro caso es 1), b_1 es la pendiente real, x_i son los valores de x , n es el número de puntos de calibrado de la regresión y s_e es el error del ajuste. Dicha F_{exp} debe ser menor que la F_{tab} para el 95% de confianza ($F_{95,2,n-2}$) (Massart *et al.*, 1997).

5.7. CONSIDERACIONES RESPECTO A ERRORES. EXACTITUD Y PRECISIÓN

Se hará una breve referencia a los distintos errores que pueden afectar a la regresión multivariante. Es un tema que actualmente está en estudio y no es de solución sencilla (ver p. ej., Faber, 1999; Ferré y Faber, 2003). De forma general, los principales errores que se pueden cometer al desarrollar un modelo de calibración son:

a) Muestras no representativas

PLS no puede "ajustarlo todo". Deberá seleccionarse racionalmente y de acuerdo con cada problema particular qué muestras se emplearán en el conjunto de calibración. Ya que normalmente en el control de calidad industrial no es posible establecer una calibración mediante diseño experimental, deberá tomarse el mayor número posible de muestras para abarcar todo el rango de interés. Esto obliga a hacer especiales esfuerzos en las etapas de diagnósticos y validación de los modelos.

b) Error aleatorio en los espectros

Este es un foco de incertidumbre que debe controlarse al máximo. La solución es, pues, emplear espectros FT-IR que sean el promedio de un determinado número de barridos y vigilar la estabilidad del equipo. En principio, se debería exigir equipos con una precisión mejor a 0.005 u.a. (Martens y Naes, 1989) (medidas realizadas en el mismo día/sesión de trabajo), lo cual para equipos FT-IR es factible. Frente a lo que cabría esperarse, no se recomienda aportar replicados al conjunto de calibración ya que eso conduciría fácilmente a modelos sobreajustados, especialmente si se aplica CV-LOO (Naes et al., 2002; Faber et al., 1998).

c) Error aleatorio en los valores de Y

Como se ha indicado, los valores Y para la regresión multivariante deben obtenerse a partir de los métodos clásicos. Obviamente dichas metodologías no están exentas de error. En un cálculo somero se ha visto que cuando se añade un 0.5% de error aleatorio a los valores de y, se obtienen errores de predicción del orden de 0.1% (Martens y Naes, 1989) por lo que, a priori, parecería que este fenómeno no es muy influyente. Ciertamente se observa en el trabajo rutinario que PLS tiene unas características operativas tan buenas que logra "reducir" el error promedio del modelo por debajo del propio aportado por los datos de partida en las variables a predecir (ver p.ej. DiFoggio, 1995).

Ahora bien, debe reconocerse que, frecuentemente, el error en los métodos de referencia no es tan bajo como un 0.5%. Este problema, que se acentúa a la hora de realizar algunas medidas en el campo petroquímico, ha sido comentado por algunos autores (p.ej. Kelly y Callis, 1990; Andrade et al., 1995). ¿Hasta qué punto este tipo de errores pueden afectar la capacidad predictiva de los modelos de PLS?

Es claro, que la solución dependerá de muchos factores pero lo que sí se puede dar fácilmente es un límite superior de propagación del error (Martens y Naes, 1989) (si se admite que no hay sesgo importante, el RMSEP de Martens se puede sustituir por el SEP de uso más común).

La incertidumbre en el "valor de referencia" (\bar{y}) debido al error aleatorio en el método de referencia ($\sigma_{j,rel}$) en las n muestras empleadas en el conjunto de calibración es $\sigma(\bar{y}) = \sigma_{j,rel}/\sqrt{n}$. Esta incertidumbre contaminará cualquier predicción

\hat{y}_{ij} . Por tanto, el nivel de ruido en el método de referencia debe ser, al menos, tan pequeño que permita que el error promedio de predicción (SEP) tenga el valor adecuado (necesario) en las predicciones.

$$\sigma_{j,ref} \ll SEP_j/\sqrt{n}$$

Por ejemplo la desviación típica ASTM para la determinación del porcentaje de aromáticos en querosenos (deducida de la reproducibilidad) es 1.35. Si se suponen 17 muestras de calibrado, esto llevaría un error mínimo en torno al 0.3.

Adicionalmente no se puede asegurar que en los datos de las **Y** empleadas en el conjunto de validación no hay errores groseros. La única "solución" es analizar qué ha podido suceder en el caso de las discrepancias más groseras (estudio de los residuales, test de Mahalanobis (*Massart et al., 1997*), etc). De esta forma, si se observa un error grosero pero el intervalo de confianza asociado es bajo y no hay otras causas asignables, puede pensarse en un error en el dato del método de referencia.

Como profundización de estos estudios, *Faber y Kowalski (1997)* y *Faber et al. (1998)* han presentado dos estudios en los cuales se evalúa con mayor detalle la influencia del error de las **Y** en el modelo multivariante (PLS).

Una de las reglas más importantes en regresión PLS es no desarrollar modelos de calibración en los que haya sobreajuste (*overfitting*) ó defecto de ajuste (*underfitting*) de las muestras de calibración. En este sentido, es importante evaluar la contribución del error de medida en los valores de referencia al SEP, el cual puede ser denominado "aparente". Como resultado, el SEP (aparente) sobreestimaré sistemáticamente el verdadero SEP puesto que hasta los valores de referencia contienen un componente aleatorio que no puede ser predicho por ningún modelo (*Faber y Kowalski, 1997; Faber et al., 1998*). Siendo el SEP aparente (*Faber y Kowalski, 1997; Faber et al., 2004*):

[ec. 24]

$$SEP_{aparente} = \left[(1/n_t) \sum_{i=1}^{n_t} (\hat{y}_i - y_{ref,i})^2 \right]^{1/2}$$

donde n_t es el número de muestras del grupo de validación, \hat{y}_i es la propiedad predicha para la muestra i ($i= 1, \dots, n_t$) e $y_{ref,i}$ es el valor de referencia asociado.

De hecho, se ha demostrado que el SEP "verdadero" aportado por el modelo multivariante puede ser más pequeño que el propio error inherente al método de

referencia usado para construir el modelo (como ya adelantara DiFoggio de forma empírica) (Fodor et al., 1999). Una valoración cuantitativa y real del “verdadero” SEP se obtiene mediante la siguiente ecuación (Faber y Kowalski, 1997; Faber et al., 2004):

[ec. 25]

$$SEP_{\text{corregido}} = \left[SEP_{\text{aparente}}^2 - s_{\text{ref}}^2 \right]^{1/2}$$

Normalmente la varianza verdadera es desconocida, pero se emplea una estimación de la varianza del método de referencia corregido por un factor que recoge los grados de libertad (ν) y la distribución de probabilidades chi (χ) para el SEP aparente (Faber y Kowalski, 1997):

[ec. 26]

$$SEP_{\text{corregido}}(\nu, \alpha) = \left[SEP_{\text{aparente}}^2 - \hat{s}^2 \frac{\nu}{\chi_{\nu, \alpha}^2} \right]^{1/2}$$

siendo ν el número de grados de libertad asociado con la varianza estimada, $\chi_{\nu, \alpha}^2$ es el valor de la distribución χ con ν grados de libertad y α es el nivel predeterminado de confianza (5%) (Faber et al., 1997).

d) Precisión: repetibilidad y reproducibilidad

Si el SEP mide la diferencia promedio que existe entre el método de referencia y el multivariante, los valores de precisión se establecen de la forma tradicional; es decir, como la desviación típica de una serie de medidas. En los casos aquí abordados, se ha optado por hablar de “precisión global”. Con este término se quiere indicar que los valores que se presentan son precisiones de metodología, incluyéndose en ellos precisión del equipo de FT-IR, de los cálculos estadísticos y otros errores típicos de redondeo, errores aleatorios, etc.

De todas formas, tampoco aquí se está exento de una cierta polémica ya que, como se ha hecho notar (Garrigues et al., 1995) existe un desajuste entre las definiciones de la IUPAC y los de ASTM. La IUPAC toma como valores de precisión las desviaciones típicas (tanto en repetibilidad (r) como reproducibilidad (R)). Por su lado, ASTM habla de rangos o intervalos de confianza. No es difícil la interconversión, pero debe recordarse a la hora de establecer comparaciones.

Cabe esperar que los valores de precisión (r y R) sean mejores para la metodología FT-IR-PLS que los métodos de referencia gracias al buen funcionamiento

de los equipos actuales. No obstante, se debe hacer una evaluación de hasta qué punto se ve afectada la precisión en la predicción ante pequeñas diferencias espectrales producidas en la misma muestra ya sea a corto o largo plazo (problema abordado en los párrafos anteriores, según los trabajos de Faber). La repetibilidad (r) y la reproducibilidad (R) se estiman al 95% de confianza como $2\sqrt{2} * S_{\text{corto}}$ y $2\sqrt{2} * S_{\text{largo}}$, respectivamente. Donde S_{corto} y S_{largo} son las desviaciones típicas de varias medidas realizadas sobre una misma muestra en un corto espacio de tiempo (en principio, consecutivas) y a “largo plazo” (varias sesiones de trabajo) (Taylor, 1987).

Por tanto, además de los valores de SEP que se muestran rutinariamente, es conveniente aportar también los de r y R ya que son parámetros que permiten ampliar el conocimiento de la bondad del método multivariante empleado.

5.8. APLICACIONES

No se pretende hacer una recopilación exhaustiva de los diferentes trabajos encontrados en donde se aplica la técnica de regresión parcial por mínimos cuadrados, aunque sí presentar algunas aplicaciones interesantes para el desarrollo de esta Memoria:

- Davies (1996) explica cómo funcionan y van surgiendo MLR (regresión lineal múltiple), PCR (regresión de componentes principales) y PLS (mínimos cuadrados parciales).
- Forina et al. (1994) comparan tres métodos de validación (SES (single evaluation set), CV (cross-validation) y RES (repeated evaluation set)) para evaluar la desviación estándar predictiva y la complejidad del modelo de regresión basado en PLS.
- Havel et al. (1993) usan PLS para la determinación simultánea de analitos en mezclas mediante un método cinético.
- Ortiz et al. (1993) aplican la técnica PLS y el análisis Cluster para la clasificación de bebidas alcohólicas.
- Durán-Merás et al. (1993) usan la técnica de regresión multivariante para la determinación de tres potenciadores de sabor en mezclas de productos alimentarios.
- Fodor et al. (1993) y Fodor (1994) analizan los destilados medios del fuel empleando MIR y la técnica de clasificación multivariante PLS.

- *Andrade et al. (1999)* emplean el método multivariante en el control de calidad industrial de gasolinas reformadas.
- *Macho et al. (1999)* lo usan en la determinación de parámetros composicionales relacionados con el contenido de hidrocarburos en nafta.
- *Chung et al. (1999)* comparan NIR y MIR para la determinación de propiedades de destilación de keroseno mediante PLS. *Chung et al. (2000)* comparaban NIR, IR y Raman para el análisis de productos de petróleo pesados.
- *Blanco et al. (2000b)* determinan el valor de penetración de bitúmenes mediante NIR.
- *Tjomsland et al. (1996)* aplican la técnica para comparar espectro de IR e impedancia de fracciones de petróleo.
- *Hamett et al. (1996)* modelan procesos de planta productiva mediante PLS.
- *Le Thanh et al. (2000)*, *Ayora-Cañada et al. (2000)*, *Sivakesava y Irudayaraj (2000)*, *Rodríguez-Saona et al. (2001)* y *Duarte et al. (2002)* aplican la técnica de regresión PLS para la determinación de ácidos y/o azúcares en refrescos y/o zumos de frutas.
- *Sivakesava e Irudayaraj (2001)* usan PLS en la determinación de la adulteración de mermelada con azúcar de caña invertido.
- *López-Anreus et al. (1998)* emplean PLS para determinar butil acetato, tolueno y metil-etil-cetona en pinturas.

En cuanto al desarrollo de la idea PLS, además del algoritmo lineal, se trabaja en PLS de tipo no lineal (*Wold et al., 1989; Hassel et al., 2002*); extendiendo PLS a matrices tridimensionales (en general matrices n-dimensionales) (*Wold et al., 1987*), combinando la estadística de tipo no paramétrico (algoritmos kernel) con PLS, en lo que se ha dado en llamar PLS-Kernel (*Lindgren et al., 1993; Rännar et al., 1994; Gao y Ren, 1999; Wold et al., 2001a*) o estudiando nuevos algoritmos de PLS que permitan selección iterativa de variables (*Lindgren et al., 1994; Abrahamsson et al., 2003; Koshoubu et al., 2000; Koshoubu et al., 2001; Estienne et al., 2001; Estienne et al., 2004; Centner et al., 2000*). *Barker y Rayens (2003)* comentan la eficacia de PLS para la discriminación de grupos (aplicación usada rutinariamente).

Además, se están abordando problemas tan importantes como la deducción de los parámetros analíticos de un modelo multivariante (por semejanza con los

univariantes, se habla de la sensibilidad, límites de detección y cuantificación, especificidad y selectividad, etc.) (Ferré y Faber, 2003).

6. REDES DE NEURONAS ARTIFICIALES

6.1. INTRODUCCIÓN

Como consecuencia del importante desarrollo de la informática, especialmente en las últimas décadas, los Químicos Analíticos están aplicando métodos cada vez más complejos para explorar las correlaciones multivariantes entre un conjunto de variables predictoras y otra(s) a predecir. A pesar del aumento de la exactitud y precisión de los métodos de medida, se observa que no todos los efectos de interés no pueden ser descritos por correlación simple univariante ni por correlaciones sencillas multivariantes. Debe acudir a técnicas cada vez más sofisticadas (Smits *et al.*, 1994; Zupan, 1994).

Éstas permiten implementar soluciones para resolver problemas que antes resultaban difíciles o imposibles de abordar. Sin embargo, se observa una limitación importante: ¿qué ocurre cuando el problema que se quiere resolver no admite un tratamiento algorítmico, como es el caso, por ejemplo, de la clasificación de objetos cuyas características son muy similares y donde el análisis cluster ha fallado? ¿y si la información de partida tiene un alto margen de error? Este ejemplo demuestra que las nuevas técnicas (más versátiles) requieren un enfoque distinto del problema. La inteligencia artificial es un campo de la informática que intenta descubrir y describir aspectos de la inteligencia humana que pueden ser simulados mediante sistemas informáticos (incluyendo software y hardware). Esta disciplina, desarrollada fuertemente en los últimos años, ha demostrado su eficacia en diversos campos científicos (Pazos Sierra, 1996).

Las redes de neuronas artificiales o redes neuronales (ANN, RNA o RN) son una forma de emular ciertas características propias de los humanos, como la capacidad de memorizar y de asociar hechos. Si se examinan con atención aquellos problemas que no pueden expresarse a través de un algoritmo, se observará que todos ellos tienen una característica en común: la experiencia. No en vano, el hombre es capaz de resolver situaciones acudiendo a la experiencia acumulada. Así, parece claro que una forma de aproximarse al problema consista en la construcción de sistemas que sean capaces de reproducir esta característica humana. Las redes neuronales tratan de ser

un modelo artificial y simplificado de un cerebro, que es el ejemplo más perfecto del que disponemos para un sistema que es capaz de adquirir conocimiento a través de la experiencia. Una red neuronal es “un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona” (ver **Figura 6**) (www.modeloingenieria.edu.ar, 2003).

En las **Figuras 7 y 8** se compara la estructura y funcionamiento (conceptual) de una neurona biológica con una artificial. Una neurona biológica (y su análoga artificial) precisa una señal (función) de entrada (*input function*), una señal (función) de activación (*activation function*) y una señal (función) de salida (*output function*).

De la observación detallada del proceso biológico se han hallado las siguientes analogías con el sistema artificial (**Figura 7**):

- Las entradas, x_i , representan las señales que provienen de otras neuronas y que son capturadas por las dendritas.
- Los pesos, W_i , son la intensidad de la sinapsis que conecta dos neuronas; tanto x_i como W_i son valores reales.
- θ es la función umbral que la neurona debe sobrepasar para activarse; este proceso ocurre biológicamente en el cuerpo de la célula. Esta “ordenada en el origen” es la encargada de ajustar el valor de la red para cada elemento de procesado, de forma que caiga en un margen adecuado para ser procesado por la función de salida. θ también se denominará *bias* o término de tendencia.
- J son las salidas de las neuronas.

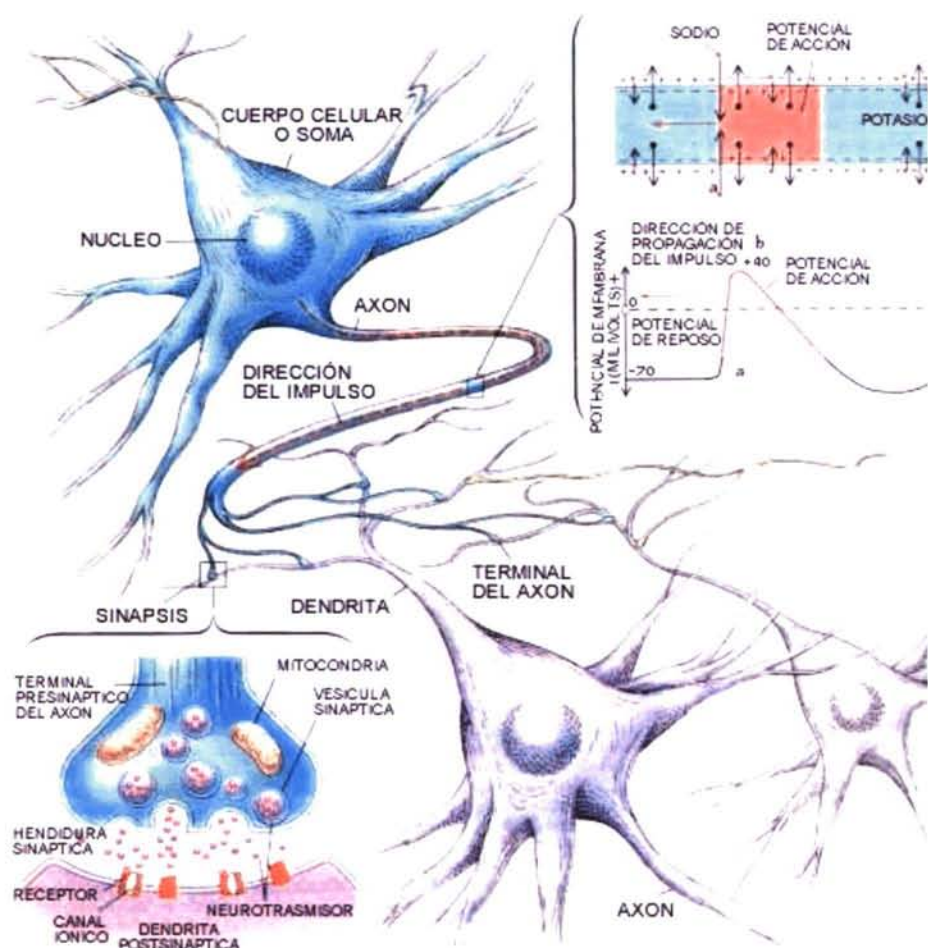


Figura 6: Comunicación entre neuronas biológicas (<http://ohm.utp.edu.co/neuronales>, 2003).

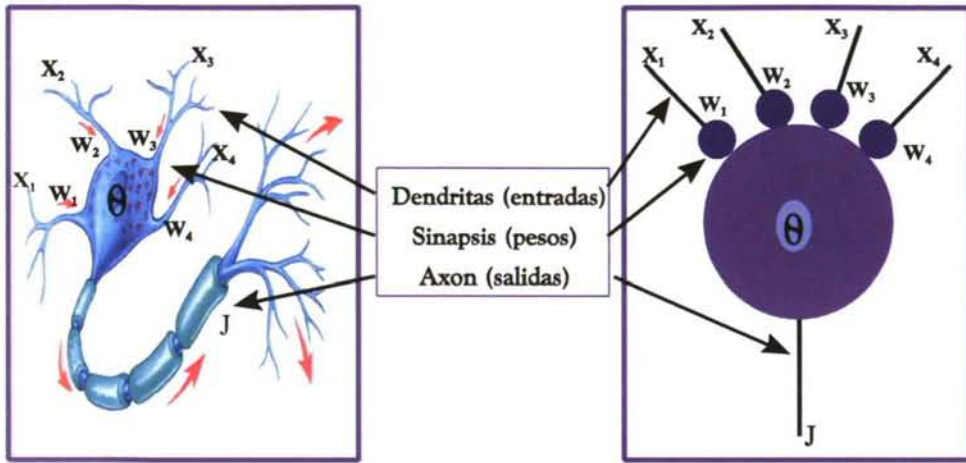


Figura 7: Comparación entre una neurona biológica (izquierda) y una artificial (derecha).

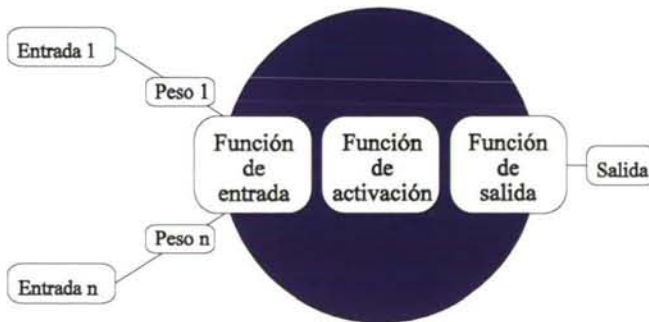


Figura 8: Ejemplo de funcionamiento conceptual de una neurona.

6.2. ELEMENTOS BÁSICOS

En consecuencia, la RN está constituida por neuronas (nodos, neuronodos, celdas, unidades o elementos de procesamiento) interconectadas y organizadas en capas (esto último puede variar). Los datos ingresan por medio de la “capa de entrada” (que recibe directamente la información proveniente de las fuentes externas de la red), pasan a través de una o varias “capa/s oculta/s” (que son internas a la red y no tienen contacto directo con el entorno exterior) y salen por la “capa de salida”

(transfiriendo la información de la red hacia el exterior) (Figura 9). El número teórico de niveles o capas ocultas está entre cero y un número elevado, pudiendo estar interconectadas sus neuronas de distintas maneras, lo que determina (junto con su número) las distintas “topologías” de las RRNN. Todas las unidades de una capa están conectadas con todas las unidades de la siguiente capa. El número de unidades de entrada y salida depende de las representaciones de los objetos de entrada y salida, respectivamente (Smits *et al.*, 1994; Zorriassatine y Tannock, 1998; Adams, 1995). La terminología usada para describir la topología de una RN puede variar dependiendo del autor. En esta Memoria se considera el número de capas como el número de capas ocultas (o capas intermedias).

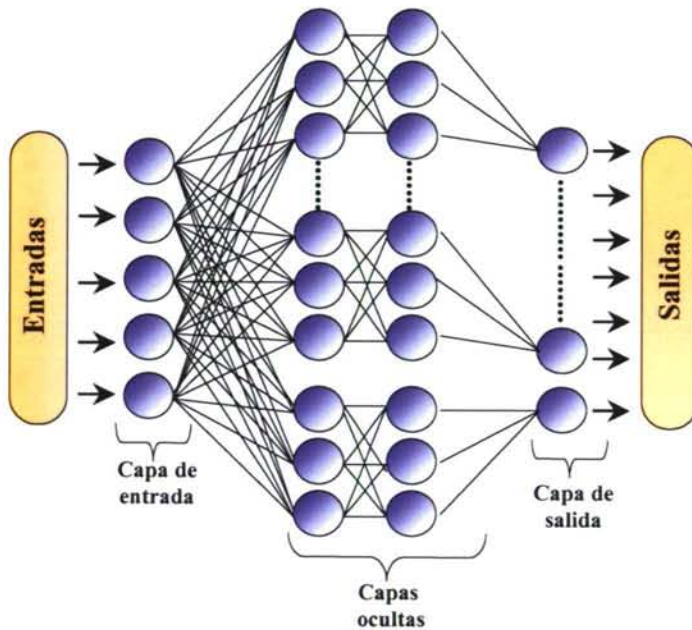


Figura 9: Ejemplo de una red neuronal totalmente conectada.

La topología o arquitectura de una RN consiste en la organización y disposición física de las neuronas en la misma. En este sentido, los parámetros fundamentales de la red son: el número de “capas” o agrupaciones de neuronas más o menos alejadas de la entrada y salida de dicha red, el número de neuronas por capa, el grado de conectividad y el tipo de conexiones entre neuronas (<http://cursos.itam.mx>, 2003). Para conseguir una RN que sirva para solucionar el problema abordado deberá estudiarse la topología más adecuada a cada caso. En consecuencia, deberán abordarse

estudios empíricos que permitan determinar el número de capas y el número de neuronas por cada capa que serán más adecuadas. No obstante, siempre se parte de una capa de entrada (en general una neurona por cada variable experimental) y una de salida (en donde pueden darse varias posibilidades). Es importante recalcar que no existe una técnica para determinar el número de capas ocultas, ni el número de neuronas que debe contener cada una de ellas para un problema específico, esta elección es determinada por la experiencia del químico. La única regla que existe es que según el teorema de Kolmogorov cualquier función de n variables se puede representar por la superposición de $2n+1$ funciones. A continuación se podría ir disminuyendo el número de capas ocultas hasta encontrar la red más pequeña que sea adecuada (Erb, 1993). Recientemente, se ha propuesto un algoritmo que permitiría simplificar esta etapa por ser casi innecesaria gracias al proceso de cálculo seguido (según Boger, 2003; Boger y Weber, 2000).

6.2.1. FUNCIÓN DE ENTRADA (*Input Function*)

Una neurona normalmente recibe muchas entradas simultáneas. Las señales de entrada a una neurona artificial x_1, x_2, \dots, x_n son variables continuas. Cada señal de entrada pasa a través de una “ganancia” o peso, llamado peso sináptico o fortaleza de la conexión cuya función es análoga a la de la función sináptica de la neurona biológica. Los pesos pueden ser positivos (excitatorios) reforzando la transmisión de la señal, o negativos (inhibitorios) debilitándola, el nodo sumatorio acumula todas las señales de entradas multiplicadas por los pesos (ponderada). De esta forma, la entrada neta a cada unidad puede escribirse de la siguiente manera:

[ec. 27]

$$Net_j = \sum_{i=1}^q w_{ji} x_i + \theta_j$$

siendo w_{ji} el peso o la intensidad con que la señal x_i afecta al cómputo global de Net_j (entrada a la unidad) y θ_j el *bias*. Los pesos pueden proceder de un proceso previo de cálculo o inicializarse aleatoriamente. Los pesos tienen, por tanto, una misión de ponderación de la influencia de los valores de entrada en la RN final.

6.2.2. FUNCIÓN DE ACTIVACIÓN (*Activation Function*)

Una neurona biológica puede estar activa (excitada) o inactiva (no excitada); es decir, que tiene un “estado de activación”. Las neuronas artificiales también tienen diferentes estados de activación; algunas de ellas solamente dos, al igual que las

biológicas, pero otras pueden tomar cualquier valor dentro de un conjunto determinado. En la **Figura 10** se visualiza el modelo más académico que facilita el estudio de las neuronas.

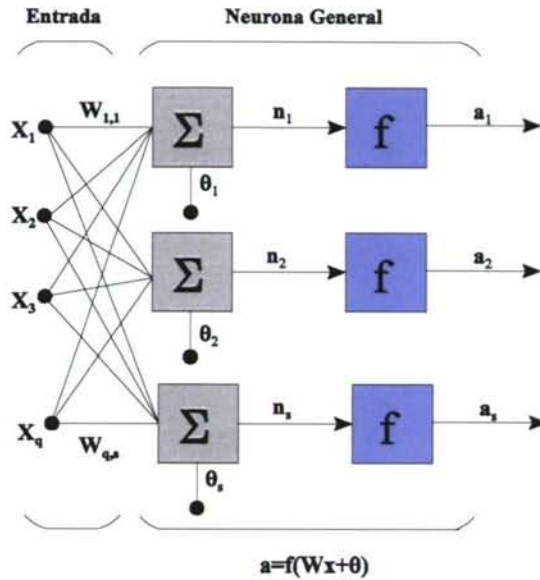


Figura 10: Capa de S neuronas.

La salida total o respuesta, a , para cada neurona está determinada por la función de transferencia, f , la cual puede ser una función lineal o no lineal de n , y que es escogida dependiendo de las especificaciones del problema que la neurona tenga que resolver. Es decir:

[ec. 28]
$$a_j = f_j(Net_j)$$

$$a_j = f_j\left(\sum_{i=1}^q w_{ji}x_i + q_j\right)$$

La actividad de la unidad determina la señal transmitida, n , de la unidad vía la función de transferencia. El *bias* se puede tratar como el peso de la unidad de entrada a la unidad i . Este término se ha fijado al comienzo y θ_j se puede ir modificando como si fuese un peso más (Smits et al., 1994).

La función de activación calcula el estado de actividad de una neurona; transformando la entrada global (n) en un valor (estado) de activación, cuyo rango

normalmente va de (0 a 1) o de (-1 a 1). Esto es así porque una neurona puede estar totalmente inactiva (0 o -1) o activa (1). Existen diferentes funciones de activación, las más comúnmente utilizadas son: la función lineal, la función umbral (o escalón), la función gaussiana, las funciones sigmoidales (entre las que se encuentran la función sigmoidea (con valores de 0 a 1) y la tangente hiperbólica (con valores de -1 a 1)) (ver **Tabla I**).

Cada una de ellas tiene ventajas e inconvenientes y no existen reglas generales para elegir una u otra de cara a su aplicación en un problema determinado. En cada uno de ellos deberán hacerse pruebas para seleccionar la más adecuada a cada caso (Despaigne y Massart, 1998). A pesar de ello, la función sigmoidea de la tangente hiperbólica se usa mucho cuando las salidas de los elementos de procesado deben ser valores continuos. Esta función es parecida a la función sigmoidal. Sus rangos de salida están comprendidos entre -1 y +1, frente a los rangos de la sigmoidal, comprendidos entre 0 y 1. La salida de la función de transferencia es utilizada como multiplicador en la ecuación de actualización de pesos, un rango entre 0 y 1 significa un multiplicador pequeño cuando la suma es baja, y un multiplicador alto cuando la suma es alta (es aconsejable cuando predominan las salidas "1"). La tangencial hiperbólica da pesos iguales a valores bajos y altos (Pazos Sierra, 1996).

Nombre	Relación Entrada/Salida	Icono	Función
Limitador Fuerte (Heaviside)	$a=0 \quad n<0$ $a=1 \quad n\geq 0$		<i>hardlim</i>
Limitador Fuerte Simétrico	$a=-1 \quad n<0$ $a=+1 \quad n\geq 0$		<i>hardlims</i>
Lineal Positiva	$a=0 \quad n<0$ $a=n \quad 0\leq n$		<i>poslin</i>
Lineal	$a=n$		<i>purelin</i>
Lineal Saturado	$a=0 \quad n<0$ $a=n \quad 0\leq n\leq 1$ $a=1 \quad n>1$		<i>satlin</i>
Lineal Saturado Simétrico	$a=-1 \quad n<-1$ $a=n \quad -1\leq n\leq 1$ $a=+1 \quad n>1$		<i>satlins</i>
Sigmoidal Logarítmico	$a = \frac{1}{1 + e^{-n}}$		<i>logsig</i>
Tangente Sigmoidal Hiperbólica	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$		<i>tansig</i>
Competitiva	$a=1$ Neurona con n max $a=0$ El resto de neuronas		<i>compet</i>

Tabla I: Funciones de activación más comúnmente empleadas (<http://ohm.utp.edu.co/neuronales>, 2003).

6.2.3. FUNCIÓN DE SALIDA (*Output Function*)

El valor resultante de esta función es la salida de la neurona, por lo que la función de salida determina el valor que se transfiere a las neuronas vinculadas. Esto es lo que ocurre con la neurona biológica en la que hay muchas entradas y sólo una salida. Si la función de activación está por debajo de un umbral determinado, ninguna salida se pasa a la neurona subsiguiente. Habitualmente, los valores de salida están comprendidos en el rango $[0, 1]$ o $[-1, 1]$. Dos de las funciones de salida más comunes son:

- Ninguna: este es el tipo de función más sencillo, tal que la salida es la misma que la entrada. Es también llamada función identidad.

- Binaria: siendo 1 si es mayor que el umbral y 0 si es menor.

6.3. REDES DE PROPAGACIÓN HACIA ATRÁS (*Backpropagation*)

De todas las topologías posibles, las redes de propagación hacia atrás, retropropagación o *backpropagation* (BPN) constituyen, sin duda, el paradigma de las RRNN de mayor aplicabilidad (<http://ttt.alc.upv.es>, 2002; López Fandiño, 1997; Andersson y Kauffmann, 2000; www.dacs.dtic.mil, 2003; Kalogirou, 2003). Esto es debido a su gran versatilidad para resolver una amplia gama de problemas, centrados principalmente en aplicaciones de regresión y clasificación supervisada. Si bien su desarrollo data de la década de los setenta (Werbos, 1974), las BPN no comenzaron a utilizarse de forma generalizada hasta finales de los años ochenta. Esto era debido a la falta de algoritmos de aprendizaje efectivos que permitieran procesar problemas de dimensión elevada (Jansson, 1991).

Uno de los grandes avances de la *backpropagation* es que aprovecha la naturaleza paralela de las redes neuronales para reducir el tiempo requerido por un procesador secuencial para determinar la correspondencia entre unos patrones dados (Luera Peña y Minim, 2001). Además, el tiempo de desarrollo de cualquier sistema que se esté tratando de analizar se puede reducir como consecuencia de que la red puede aprender el algoritmo correcto sin que alguien tenga que deducir por anticipado el algoritmo en cuestión.

La *backpropagation* es un tipo de RN que emplea un ciclo propagación-adaptación de dos fases, de acuerdo con un aprendizaje supervisado (ver **Figura 11**). Una vez que se ha introducido un patrón a la entrada de la red como estímulo, éste

se propaga desde la primera capa a las siguientes hasta generar una salida, la cual se compara con la salida deseada y se calcula un error en el aprendizaje para cada una de las salidas (<http://ohm.utp.edu.co/neuronales>, 2003; Burke y Ignizio, 1997; Kateman, 1993; Gasteiger y Zupan, 1993; www.dd.chalmers.se/~f96jost/superresolution, 2002).

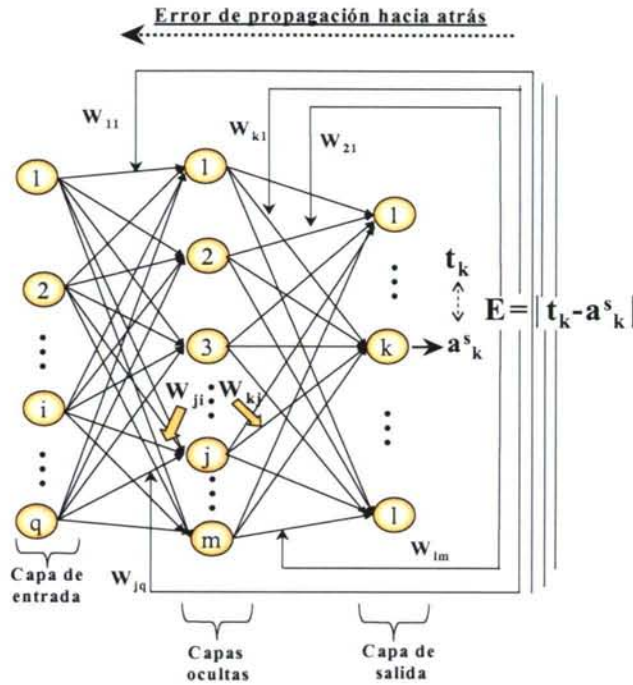


Figura 11: Red de propagación hacia atrás.

El error calculado se propaga hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida. Sin embargo las neuronas de la capa oculta sólo reciben una fracción del total del error, basándose aproximadamente en la contribución relativa que haya aportado cada neurona de salida original. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido una señal de error que describa su contribución relativa al error total. Basándose en la "señal" de error percibida, se actualizan los pesos de conexión de cada neurona, para hacer que la red converja hacia un estado que permita clasificar correctamente todos los patrones de entrenamiento (Erb, 1993). Normalmente se admite un margen de error predefinido, o bien se da por finalizado el proceso cuando se alcanza un número predeterminado de presentaciones de los patrones de entrenamiento (iteraciones). En la Figura 12 se observa el esquema de entrenamiento de la BPN.

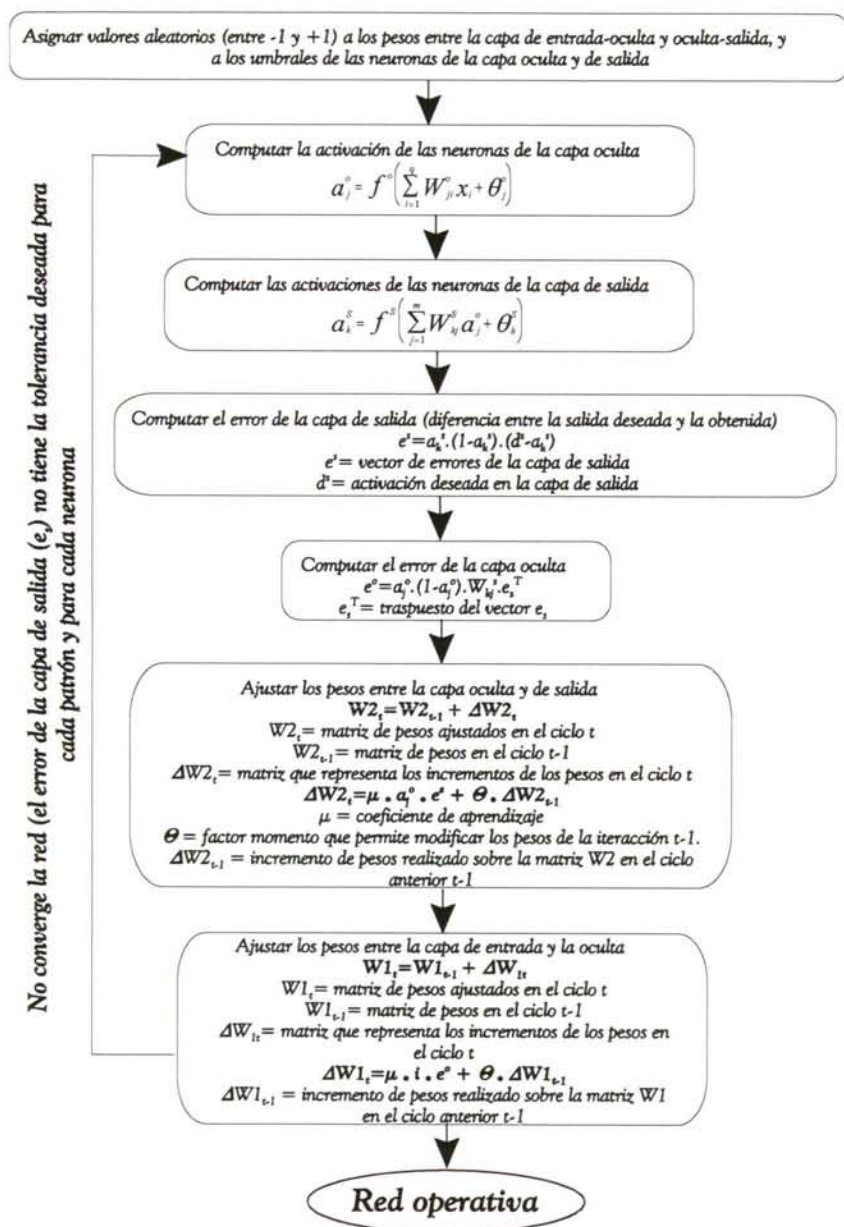


Figura 12: Esquema de entrenamiento de una RN de retropropagación con una capa oculta (Pazos Sierra, 1996).

A medida que se va entrenando la red, las neuronas de las capas intermedias se organizan a sí mismas de tal modo que las distintas neuronas aprenden a reconocer distintas características del conjunto total de entrada. Después del entrenamiento cuando se introduzca una muestra arbitraria de entrada, la RN responderá con una salida si la nueva muestra contiene características que las neuronas individuales hayan aprendido a reconocer durante su entrenamiento. Y a la inversa, las unidades de las capas ocultas tienen una tendencia a inhibir su salida si el patrón de entrada no contiene las características en las cuales basan su “reconocimiento”, para las cuales han sido entrenadas.

Varias investigaciones han demostrado que durante el proceso de entrenamiento la red de propagación hacia atrás tiende a desarrollar relaciones internas entre neuronas con el fin de organizar los datos de entrenamiento en clases (<http://ohm.utp.edu.co/neuronaes>, 2003). Eso se puede extrapolar para llegar a la hipótesis de que las unidades de la capa oculta de una *backpropagation* se asocian de alguna manera a características específicas de los patrones de entrada como consecuencia del entrenamiento (como sucede en el aprendizaje humano). En general, esta asociación no es interpretable para el observador humano (aunque se ha hecho algún intento (*Ruckebusch et al.*, 2002)), lo importante es que la red ha encontrado una representación interna que le permite generar las salidas deseadas cuando se le dan las entradas.

6.4. APRENDIZAJE Y VALIDACIÓN

En este epígrafe se va a presentar de forma resumida el proceso por el cual una RN se dice que aprende. También se verá cómo validar dicha RN de cara a su uso práctico. Las explicaciones se harán para una red con una capa de entrada, una capa oculta y una capa de salida y luego se generalizará para redes que tengan más de una capa oculta. En todos los casos:

i = número de variables del vector de entrada (1, 2, ..., i).

j = número de neuronas de la capa oculta (1, 2, ..., j).

k = número de neuronas de la capa de salida (1, 2, ..., k).

Cuando se presenta a la red un patrón de entrenamiento, éste se propaga a través de las conexiones produciendo una entrada neta (n) en cada una de las neuronas de la siguiente capa. La capa de neuronas de entrada recibe, simplemente, el dato correspondiente a las variables predictoras (p.ej. los espectros), una vez escaladas. La entrada neta a la neurona j de la siguiente capa (la primera oculta) está

dada por la ecuación suma del apartado 6.2.1:

[ec. 29]

$$n_j^0 = \sum_{i=1}^i W_{ji}^0 x_i + \theta_j^0$$

siendo W_{ji}^0 el peso que une la neurona (o nodo) i de la entrada con la neurona j de la primera capa oculta; x_i la variable i del vector de entrenamiento; θ_j^0 la ganancia (o *bias*) de la neurona j de la capa oculta y el superíndice (0) representa la capa a la que pertenece cada parámetro, en este caso, la capa oculta.

Cada una de las neuronas de la capa oculta tiene como salida o respuesta a_j^0 , calculada por la ecuación:

[ec. 30]

$$a_j^0 = f^0(n_j^0)$$

siendo f^0 la función de transferencia de las neuronas de la capa oculta. Ahora, las j salidas, a_j^0 , de las neuronas de la capa oculta son las entradas ponderadas con otros pesos de conexión de las neuronas de la capa de salida. Como sucedía antes, la señal neta de entrada para una neurona de la capa de salida está descrita por la ecuación suma (el superíndice s indica la capa de salida):

[ec. 31]

$$n_k^s = \sum_{j=1}^j W_{kj}^s a_j^0 + \theta_k^s$$

siendo W_{kj}^s el peso que une la neurona j de la capa oculta con la neurona k de la capa de salida, θ_k^s la ganancia (o *bias*) de la neurona k de la capa de salida y n_k^s la entrada neta a la neurona k de la capa de salida.

La red produce una salida o respuesta final (para cada neurona de salida), descrita por la ecuación:

[ec. 32]

$$respuesta = a_k^s = f^s(n_k^s)$$

La salida de la red de cada neurona a_k^s se compara con la salida deseada t_k para calcular el error en cada unidad de salida:

[ec. 33]

$$\delta_k = (t_k - a_k^s)$$

El error total cometido al predecir la/s respuesta/s para cada patrón propagado se evalúa mediante:

[ec. 34]

$$ep^2 = \frac{1}{2} \sum_{k=1}^k (\delta_k)^2$$

siendo ep^2 el error medio cuadrático para cada patrón de entrada y δ_k el error en la neurona k de la capa de salida.

Este proceso se repite para el número total de patrones de entrenamiento (r). Para un proceso de aprendizaje exitoso el objetivo del algoritmo es actualizar todos los pesos y ganancias (*bias*) de la red minimizando el error medio cuadrático total descrito en:

[ec. 35]

$$e^2 = \sum_{r=1}^r ep^2$$

donde e^2 es el error total en el proceso de aprendizaje en una iteración después de haber presentado a la red los r patrones de entrenamiento.

El error que genera una red neuronal en función de sus pesos genera un espacio de n dimensiones, donde n es el número de pesos de conexión de la red, al evaluar el gradiente del error en un punto de esta superficie se obtendrá la dirección en la cual la función del error tendrá un mayor crecimiento. Como el objetivo del proceso de aprendizaje es minimizar el error debe tomarse la dirección negativa del gradiente para obtener el mayor decrecimiento del error y de esta forma su minimización, condición requerida para realizar la actualización de la matriz de pesos en el algoritmo *backpropagation*.

6.4.1. MECANISMOS DE APRENDIZAJE

Dado que los datos de entrada se procesan a través de la red neuronal con el propósito de lograr una salida, la RN debe aprender a calcular la salida correcta para cada vector de entrada (muestra/patrón) en el conjunto de ejemplos mediante los algoritmos arriba descritos. Este proceso de aprendizaje se denomina: *proceso de entrenamiento o acondicionamiento*. El conjunto de datos (o conjunto de ejemplos) sobre el cual se basa este proceso se llama *conjunto de datos de entrenamiento*.

Como, en general, ni la topología de la red ni las funciones que operan en

cada neurona (entrada, activación y salida) cambiarán durante el aprendizaje, el aprendizaje de una RN se limita, pues, a la *adaptación de los pesos*. En otras palabras, el aprendizaje es el proceso por el cual una RN modifica sus pesos en respuesta a la información de entrada. Los cambios que se producen se reducen a la “destrucción”, modificación y “creación” de conexiones entre las neuronas (al igual que los sistemas biológicos). En los modelos de RRNN la creación de una nueva conexión implica que el peso de la misma pasa a tener un valor distinto de cero. De la misma manera, una conexión se destruye cuando su peso pasa a ser cero. El proceso habrá terminado (la red ha aprendido) cuando los valores de los pesos permanecen estables.

Un aspecto importante respecto al aprendizaje de las redes es determinar cómo se modifican los valores de los pesos, es decir, los criterios para cambiar los pesos durante el aprendizaje. Existen diversos métodos de aprendizaje (www.modeloingenieria.edu.ar, 2003; <http://ttt.alc.upv.es>, 2002):

a) Aprendizaje supervisado: el proceso de aprendizaje o entrenamiento se controla por un agente externo (supervisor) que analiza la salida de la red y, si no coincide con la deseada, procederá a modificar los pesos de las conexiones con el fin de conseguir que la salida obtenida se aproxime a la deseada (www.ulbra.tche.br/~danielnm, 2003). En Química Analítica la mayor parte de los aprendizajes son de este tipo. Hay varias posibilidades:

1- Aprendizaje por corrección de error (p.ej. retropropagación o *backpropagation*): consiste en ajustar los pesos de las conexiones de la red en función de la diferencia entre los valores deseados y los obtenidos a la salida de la red, es decir, en función del error cometido en la salida (habitualmente ésta es la opción elegida).

2- Aprendizaje por refuerzo: durante el entrenamiento no se indica exactamente la salida que se desea que proporcione la red ante una determinada entrada.

3- Aprendizaje estocástico: consiste en realizar cambios aleatorios en los valores de los pesos de las conexiones de la red y se evalúa su efecto a partir del objetivo deseado y de distribuciones de probabilidad.

b) Aprendizaje no supervisado: la red no recibe ninguna información del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta (www.ulbra.tche.br/~danielnm, 2003).

1- Aprendizaje hebbiano: pretende medir la similitud entre las

muestras o extraer características comunes de los datos de entrada.
2- Aprendizaje competitivo y comparativo: si un nuevo patrón (o muestra) se determina que pertenece a una clase reconocida previamente, entonces la inclusión de este nuevo patrón a esta clase matizará (corregirá, ampliará, etc.) la representación de la misma.

6.4.2. EVOLUCIÓN DE LOS PESOS

Antes de comenzar el entrenamiento se debe escoger un conjunto inicial de pesos para las diferentes conexiones entre las neuronas de la red neuronal. Esto puede realizarse por varios criterios aunque el más habitual consiste en otorgar un valor aleatorio del peso a cada conexión, encontrándose los mismos dentro de un cierto intervalo. A la hora de propagar hacia atrás los errores, existen dos posibilidades en cuanto a la actualización de los pesos:

- a) Actualización continua, los cambios en los pesos se realizan inmediatamente después de aplicar un patrón de entrada; de esta forma cuando ha transcurrido una época (iteración de la RN con todos los patrones), todos los pesos de la red se habrán modificado r veces.
- b) Actualización periódica, en donde las modificaciones de los pesos no se realizan hasta que se ha completado una época, se suman todos los cambios de pesos para todas las observaciones, aplicándose al final.

La aproximación periódica, análoga al método de la regresión parcial por mínimos cuadrados (PLS), tiene como principal desventaja la necesidad de almacenar una gran cantidad de información durante el entrenamiento. Por contra, la carga computacional es menor y, además, permite emplear parámetros de aprendizaje más elevados, con lo que la convergencia se alcanza más rápidamente.

Para cada iteración diferente de la primera, los nuevos pesos se calculan como: $\text{nuevo peso} = (\text{coeficiente o velocidad de aprendizaje}) * (\text{error}) + (\text{momento}) * (\text{peso anterior})$, donde el "error" es la diferencia entre el valor de salida de la RN y el valor verdadero, el coeficiente o velocidad de aprendizaje (abreviadamente, lr) es un factor ajustable que controla la velocidad del proceso de aprendizaje (a valores altos, el proceso de aprendizaje es más rápido, pero si la relación es demasiado grande, las oscilaciones en el cambio de pesos puede impedir la convergencia para obtener una solución óptima). El factor momento es una constante que tiene el efecto de suavizar las oscilaciones de cambio de pesos durante el proceso de optimización (valores pequeños del momento ayuda a prevenir nuevas fluctuaciones en el cambio de pesos que pueden impedir que

la red llegue a la solución óptima). En las aplicaciones de esta Memoria sólo se podrá ajustar el coeficiente ó velocidad de aprendizaje (lr), ya que suele tomar valores entre 0.3 y 0.6 y su influencia es menor que la del coeficiente de aprendizaje (Smits *et al.*, 1994). La situación de esta ecuación en el proceso general de aprendizaje de la RN se muestra en la **Figura 12** de este capítulo.

6.4.3. DETENCIÓN DEL PROCESO DE APRENDIZAJE

Para determinar cuándo se detendrá el proceso de aprendizaje, es necesario establecer una *condición de detención*. El entrenamiento se detiene (convergencia de la RN) cuando el error observado está por debajo de un determinado umbral. Se acostumbra a emplear el error estándar normalizado (NSE) (o error promedio) como medida del error global de la red (López Fandiño, 1997; <http://zhanshou.hypermart.net.htm>, 2003):

[ec. 37]

$$NSE = \frac{1}{rh} \sum_{k=1}^k \sum_{r=1}^r (y_{kr} - \bar{y}_{kr})^2$$

en el que r sería el número de patrones en el conjunto de datos y k es el número de neuronas de salida; y_{kr} y \bar{y}_{kr} representan el valor de salida deseado y obtenido, respectivamente, de la unidad k en el patrón r (Smits *et al.*, 1994).

6.4.4. CODIFICACIÓN DE LOS DATOS DE ENTRADA

Para facilitar el trabajo de la RN los datos de entrada y salida suelen codificarse, es decir, se lleva a cabo un proceso de escalado (centrado en la media, autoescalado, normalización 0-1, etc.). De forma que la RN (como otros métodos multivariantes de regresión) no se enfrente a los problemas de escala en los datos de partida. Análogamente, también es frecuente codificar los datos de salida, especialmente si se trata de clasificación en grupos, discriminación de las salidas, etc. No existe un proceso de escalado universalmente bueno y en cada problema deberán probarse algunos, el escalado 0-1 (división por el valor máximo) es muy habitual.

6.4.5. VALIDACIÓN DE LA RED NEURONAL

Después del entrenamiento la red tiene que ser validada. La experiencia demuestra que las RRNN convergen con cierta facilidad y que, por tanto, aprenden

muy bien el conjunto de los patrones ó muestras de entrenamiento. Esto supone un problema operativo porque llegan a “memorizar” en lugar de generalizar, en cuyo caso el sobreajuste impide la utilidad real de la RN (Richards et al., 2002). En consecuencia, se debe comprobar si la red neuronal puede resolver nuevos problemas que no había visto hasta ese momento. Por lo tanto, la validación de la RN requiere otro conjunto de datos, independiente del usado anteriormente, denominado conjunto de validación o *test*. La fase de test es un camino para determinar la bondad del aprendizaje de la red y su capacidad de generalización. Durante esta fase los casos de test son presentados a la red y ésta entrega sus resultados. Si se conoce la respuesta correcta, se puede calcular el error de la red. Si los casos de test son representativos del problema real que se desea abordar, uno se puede hacer una idea del funcionamiento de la red en dicho problema.

La mayor diferencia entre entrenamiento y test radica en la no modificación de los pesos durante los test. Cada ejemplo del conjunto de validación contiene los valores de las variables de entrada, con su correspondiente solución. Pero esta solución no será conocida para la RN por lo que el Químico Analítico deberá comparar la solución calculada para cada ejemplo de validación con la solución conocida y tomar las decisiones pertinentes.

6.5. APLICACIONES

En el campo de la calibración multivariable, la RN se definiría como un estimador de regresión no lineal y no paramétrica (Despaigne y Massart, 1998), siendo los métodos no paramétricos los no basados en una asunción *a priori* de una forma de modelo específica. Las RRNN estiman las relaciones entre una o varias variables de entrada (llamadas variables independientes o descriptores) y una o varias variables de salida (llamadas variables dependientes o respuestas).

Debido a su constitución y a sus fundamentos, las RRNN presentan un gran número de características semejantes a las del cerebro. Esto hace que ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando en múltiples áreas. Entre las ventajas se incluyen (www.modeloingenieria.edu.ar, 2003; www.uta.cl, 2003; <http://lisisu02.usal.es/~airene>, 2003; www.doc.ic.ac.uk/~nd/surprise.html, 2003):

a) Adquirir el “conocimiento” a través de la experiencia, el cual es almacenado, al igual que en el cerebro, en el peso relativo de las conexiones interneuronales (aprendizaje adaptativo).

- b) Gran adaptabilidad, son capaces de cambiar dinámicamente junto con el medio (auto-organización).
- c) Alto nivel de tolerancia a fallos, es decir, pueden sufrir un daño considerable y continuar teniendo un buen comportamiento, al igual que ocurre en los sistemas biológicos.
- d) Tener un comportamiento altamente no-lineal, lo que les permite procesar información procedente de fenómenos no lineales o con elevado "ruido" (información no relacionada con el problema de interés).

La idea básica de las RRNN no es un desarrollo reciente. Este campo fue establecido antes del advenimiento de los computadores, pero su verdadero desarrollo tuvo lugar cuando las simulaciones por computador fueron factibles por capacidad de procesamiento y bajo costo. Tras un período inicial de entusiasmo, las RRNN cayeron en un período de frustración y desprestigio tras las críticas de *Minsky y Papert (1969)*. Durante este período, cuando el soporte computacional era mínimo y se disponía de pocos estudios, sólo unos pocos investigadores consiguieron logros importantes que sembraron frustración general en la comunidad científica contra las RRNN, que aceptó las conclusiones de estos investigadores sin un mayor análisis. Actualmente, las redes neuronales constituyen un campo en el cual resurgió el interés. La historia de las RRNN puede resumirse a través de varios hitos:

Alan Turing (1936) fue el primero en estudiar el cerebro como una forma de ver el mundo de la computación aunque quienes concibieron los fundamentos de la computación neuronal fueron *Warren McCulloch y Walter Pitts (1943)*. Por su parte, *Donald Hebb (1949)* fue el primero en explicar los procesos del aprendizaje desde un punto de vista psicológico y estudió las semejanzas entre el aprendizaje y la actividad nerviosa. Los trabajos de Hebb formaron las bases de la Teoría de las Redes Neuronales. *Karl Lashley (1950)* encontró que la información era distribuida a lo largo de él. *Frank Rosenblatt (1957)* comenzó el desarrollo del Perceptrón (la RN más antigua que sigue utilizándose aún hoy), el cual es capaz de generalizar. Dos años más tarde, confirmó que, bajo ciertas condiciones, el aprendizaje del Perceptrón convergía hacia un estado finito (Teorema de Convergencia del Perceptrón). *Bernard Widrow y Marcian Hoff (1960)* desarrollaron el modelo Adaline (ADAPTative LINear Elements), la primera RN aplicada a un problema real (filtros adaptativos para eliminar ecos en las líneas telefónicas). En 1969 se produjo un fuerte retroceso en el uso de las RRNN, ya que *Marvin Minsky y Seymour Papert (1969)* probaron (matemáticamente) que el

Perceptrón no era capaz de resolver problemas relativamente fáciles, tales como el aprendizaje de una función no-lineal. Esto demostró que el Perceptrón era muy débil, dado que las funciones no-lineales son extensamente empleadas en computación y en los problemas del mundo real.

Paul Werbos (1974) desarrolló la idea básica del algoritmo de aprendizaje de propagación hacia atrás (backpropagation) cuyo significado quedó definitivamente aclarado en 1985. En este mismo año, *Hopfield y Tank (1985)* consiguieron el “renacimiento” de las redes neuronales gracias a su libro “Computación neuronal de decisiones en problemas de optimización”. *David Rumelhart et al. (1986)* redescubrieron el algoritmo de aprendizaje de propagación hacia atrás (backpropagation).

La tipología y el número de aplicaciones de las RRNN se ha incrementado sustancialmente en los últimos años (ver **Tabla II**) ya que pueden ofrecer una solución en aquellos casos en los que otras técnicas quimiométricas fallan o no aportan resultados correctos. A continuación se hará un brevísimo recorrido por estas aplicaciones.

Las RRNN deberían ser usadas principalmente cuando se sospecha que un grupo de datos no es lineal (desde un punto de vista matemático, un modelo realmente es no lineal cuando la no linealidad ocurre en la relación entre la respuesta y los descriptores, p.ej. $y=x^2$). En Química Analítica, la distinción entre no linealidad verdadera y aparente no siempre puede ser detectada visualmente en líneas de calibración o en modelos de los residuales. Algunos tipos de no linealidad pueden observarse al usar sensores o datos espectroscópicos. Por ejemplo, la ley de Lambert-Bouger-Beer que relaciona linealmente la absorbancia de una especie en una mezcla con su concentración es una aproximación que sólo es válida para sistemas diluidos y no saturados. Desviaciones de la linealidad se pueden observar si una muestra es altamente absorbente o no homogénea, si el tamaño de partícula no es constante en todas las muestras o si algunas señales solapan. Una respuesta no lineal del detector o la presencia de radiación difusa (debido a imperfecciones ópticas de un espectrómetro) introduce curvatura en la función concentración-respuesta.

NOMBRE	AÑO	AUTOR	APLICACIONES
Perceptron	1957	Rosenblatt	Reconocimiento de caracteres impresos
ADALINE	1960	Widrow	Filtros adaptativo de señales. Modems
Avalancha	1967	Grossberg	Reconocimietno del habla. Control de robots
Cerebellatron	1969	Marr, Albus	Control de robots
BPN	1974-1985	Werbos, Rumelhart	Reconocimietno de caracteres manuscritos
Brain-State-in-a-box	1977	Anderson	Extracción del conocimiento
Neocognitron	1978-1984	Fukushima	Reconocimiento de caracteres manuscritos
SOM	1980-1984	Kohonen	Optimización, clustering, codificación
Hopfield	1982	Hopfield	Reconstrucción de patrones, optimización
Boltzman y Cauchy	1985-1986	Hinton, Sejnowski	Reconocimiento de patrones, optimización
BAM	1985	Kosko	Memoria heteroasociativa
CPN	1986	Hecht-Nielsen	Compresión de imágenes
ART	1986	Carpenter, Grossberg	Reconocimiento de patrones complejos

Tabla II: Cronología de los principales tipos de RRNN (Hilera y Martínez, 1995).

Las RRNN se pueden usar cuando no se tiene una indicación *a priori* referente a la naturaleza matemática de la relación y se necesita rápidamente un modelo, aunque no deben aplicarse las RRNN cuando el modelo sea lineal. Finalmente, se recomienda la RN para monitorizar procesos on-line, donde las variables medidas poseen ruido y perturbaciones (p.ej. efectos de temperatura que introducen una no linealidad en el modelo) (Despaigne y Massart, 1998).

A continuación se presenta un brevísimo resumen de algunos ejemplos de aplicaciones de las RRNN en el área de la Química Analítica. En esta revisión no se ha pretendido ser exhaustivo sino, únicamente, indicar diversos tipos de problemas en los cuales las RRNN han resultado ventajosas. El número de aplicaciones ha aumentado mucho, tal como ilustraban en Zupan y Gasteiger (1991) en su revisión. En la revisión fundamental de quimiometría de 1996 (Brown *et al.*, 1996), se reportaron las aplicaciones de las RRNN en campos tan diferentes como el procesado de señales, la resolución de curvas, calibración, parámetros de estimación, técnicas de

reconocimiento de pautas y, por supuesto, en inteligencia artificial. Tutoriales de redes neuronales en Química fueron propuestos por *Smits et al. (1994)* y *Svozil et al. (1997)* y diferentes tipos de aplicaciones de RRNN para espectroscopía fueron revisadas por *Cirovic (1997)*. Las aplicaciones quimiométricas son muy extensas y, aunque en muchos casos son de naturaleza multidisciplinar, merecen una consideración aparte. En general ofrecen una alternativa clara a los métodos estadísticos tradicionales (*Lippmann, 1987; Sánchez y Sarabia, 1995; Schulz et al., 1995; Wienke y Kateman, 1994*), mejorando en un gran número de casos los procedimientos de reconocimiento de pautas y análisis de regresión.

La automatización del proceso de interpretación de espectros de IR mediante ordenador puede realizarse con ayuda de un sistema modular basado en las RRNN (*van Est et al., 1993; Affolter y Clerc, 1993; Jacobsson y Hagman, 1993*).

Se han realizado aplicaciones para el reconocimiento de isómeros en bencenos sustituidos mediante técnicas espectrofotométricas (*Weigel y Herges, 1992*). Aunque los sistemas expertos también se han utilizado con éxito para la interpretación de este tipo de espectros (*Andreev et al., 1993*).

La predicción de propiedades físico-químicas de compuestos orgánicos se ha realizado empleando un sistema basado en descriptores gráficos junto con la técnica de RRNN (*Artemenko et al., 2003*). Los análisis por RMN también cuentan con aplicaciones de RRNN de cara a la predicción de propiedades moleculares (*Kvasnicka et al., 1992*).

Se han realizado comparaciones de diferentes métodos quimiométricos de análisis para la clasificación de tres tipos de agentes quimioterapéuticos (*Tominaga, Y, 1999*). *Livingstone et al. (1997)* emplean la RN en el modelado de datos en el diseño de drogas y presentan alguna estrategia para realizar selección de variables en función de la magnitud de los pesos y del error obtenido al ir eliminando variables.

Se han llevado a cabo estudios de la adulteración alimentaria aplicando la técnica de redes neuronales (*Downey, 1998; Goodacre et al., 1997; Goodacre et al., 1995, Penza y Cassano, 2004; Pérez-Magariño et al., 2004*), de caracterización de variedades (*Marini et al., 2004a; Marini et al., 2004b*), de modelización del tratamiento térmico de alimentos (*Luera Peña y Minim, 2001*) y de predicción de viscosidad de zumos (*Rai et al., 2005*).

Esta técnica quimiométrica se ha empleado también para determinar metales

en diferentes tipos de muestras (Kompany-Zareh *et al.*, 1999; Hernández-Caraballo *et al.*, 2003; Hernández-Caraballo *et al.*, 2004), composición y propiedades de gomas (Borosy, 1999), análisis de drogas (Tong y Cheng, 1999), predicción de la temperatura de fusión de cenizas de carbón (Yin *et al.*, 1998), determinación de la humedad relativa y la composición de CO₂ en mezclas de gases (Henkel y Schmeißer, 2002), estimación de propiedades de suelos (Ramadan *et al.*, 2005), clasificación de aceites (Kapur *et al.*, 2004), *screening* de hidrocarburos aromáticos policíclicos (Fernández-Sánchez *et al.*, 2004), predicción de los índices de retención de alquilbencenos (Zhang *et al.*, 1999) y de los grupos hidroxilo y número de ácidos presentes en las resinas de poliéster (Marengo *et al.*, 2004).

Li *et al.* (2000) consideraban la red neuronal recurrente para la compresión de espectros UV-Visibles, encontrándola una buena aplicación comparable a WT (transformación de wavelet) o incluso mejor.

Uno de los aspectos más relevantes y que prometen un gran desarrollo es el del análisis sensorial (Naes y Risvik, 1996), en concreto la modelización del sentido del gusto (Chen *et al.*, 2001), del olfato (Nakamoto *et al.*, 1991; Nakamoto *et al.*, 1992; Nakamoto *et al.*, 1993; Lu *et al.*, 2000) y del oído (Ritter *et al.*, 1992).

7. DISEÑO DE EXPERIENCIAS Y OPTIMIZACIÓN

En la puesta a punto de un procedimiento analítico es muy frecuente que deba decidirse, en primer lugar, qué variables analíticas son aquellas que influyen de forma más acusada en el resultado final (de una lista de posibles parámetros, que decide el analista). Esta selección deberá ser objetiva y una vía de hacerlo es aplicando diseños de experiencias.

Posteriormente, las variables seleccionadas habrán de optimizarse; es decir, deducir qué valor numérico debería fijarse para cada una de ellas, una vía objetiva es emplear el método simplex.

7.1. DISEÑO DE EXPERIENCIAS

Son numerosas las variables que influyen en el proceso de medida, por lo que el estudio univariante implicaría un gran número de experimentos y, además, la información obtenida podría ser incompleta ya que no se tendrían en cuenta las posibles interacciones entre las variables consideradas. Entre los métodos disponibles

para llevar a cabo esta tarea de forma sistemática destacan las técnicas de *Diseño de Experiencias* cuyo objetivo es obtener la máxima información con el mínimo esfuerzo y coste (lo que suele equivaler a menor número de experimentos).

Para ello, definimos:

- a) Factores: son las variables con las que se trabajarán y que pueden afectar al resultado.
- b) Espacio de los factores: es el espacio n dimensional asociado con los n factores.
- c) Superficie de respuesta: es la función de respuesta representada en el espacio de los factores.
- d) Niveles: cada uno de los valores experimentales considerados para cada factor (el número de niveles es igual para todos los factores).

Un tipo particular de diseños, muy aplicados en la Química Analítica, son los atribuidos a *Plackett y Burman (1946)*, los cuales son diseños factoriales parciales saturados que permiten seleccionar, de entre todos los ensayos que podrían ser realizados, solamente aquellos que permitan establecer qué variables influyen sobre el sistema y cómo lo hacen. Con este tipo de diseños es posible estudiar $k = N - 1$ variables en N ensayos, en donde N tiene que ser múltiplo de 4.

Ahora bien, conviene recordar que si bien los diseños de Plackett-Burman reducen el número de experiencias a realizar, incrementan la complejidad de la interpretación de los resultados debido al efecto conocido como “confusión”. Éste supone que el efecto atribuido a una variable en realidad puede incluir también el efecto debido a combinaciones de otras variables.

Como la aplicación que se va a realizar en esta Memoria de Diseño de experiencias es muy sencilla y su objetivo tan sólo será definir la distribución espacial de patrones acuosos de tres azúcares para llevar a cabo un calibrado multivariante, se ha elegido un diseño completo a tres niveles, siendo el número de combinaciones de valores de los n factores $N=3^n$. Se emplea este diseño para poder estimar los coeficientes de un polinomio de segundo grado, conteniendo 3 niveles para cada una de las variables independientes (factores). De hecho, *Massart et al. (1997)* establecen que uno de los dos grandes objetivos de los diseños de experiencias consiste en modelar la relación entre x e y con un mínimo de experimentos. Esto requiere un esquema eficiente y ordenado del espacio experimental, obteniendo resultados con el menor coste posible.

El diseño factorial completo de tres niveles se lleva a cabo para determinar la magnitud del efecto de las interacciones entre dos o más factores que pueden llevar a un efecto en la respuesta. Este tipo de diseño de 3^n , puede ser usado para obtener modelos cuadráticos (Massart *et al.*, 1997).

7.2. MÉTODO SIMPLEX GEOMÉTRICO

Es un método de optimización muy empleado tanto por su simplicidad como los buenos resultados que ofrece sin provocar un trabajo analítico excesivo.

Desde el punto de vista matemático, el método simplex geométrico es un algoritmo secuencial de búsqueda directa que, como todos los de este grupo, se caracteriza por la no utilización explícita de derivadas de la función objetivo y por basarse en un examen secuencial de soluciones. Cada solución (respuesta) ensayada se compara con la mejor obtenida hasta el momento y existe una estrategia para determinar (en función de los resultados previos) qué nuevo ensayo debe realizarse.

Un simplex es una figura geométrica definida por un número de puntos igual al número de dimensiones del espacio (k variables) más uno, de forma que para 2 dimensiones tenemos un triángulo equilátero, para $k=3$ un tetraedro y así, sucesivamente. Estas figuras geométricas tienen la interesante propiedad de que se puede formar un nuevo simplex regular sobre cualquier cara de un simplex dado por la adición, tan sólo, de un nuevo punto equidistante de los de la cara fija.

El método Simplex, introducido en su forma original por Spendley *et al* (1962), no se basa en planeamientos factoriales y por eso requiere pocos experimentos para moverse, desplazándose en la dirección del óptimo. La aplicación del método Simplex en Química Analítica fue efectuada por primera vez en 1969 (www.chemkeys.com, 2004; Cela, 1984). El método Simplex original, a lo largo de estos años, ha sufrido modificaciones que obligaron a la distinción del mismo dentro de las estrategias de optimización, así el método Simplex original pasó a ser llamado el Método Simplex Básico (MSB).

El procedimiento a seguir para alcanzar el óptimo consiste en calcular el vértice en el que la función objetivo toma su menor valor (supongamos que queremos maximizar la función) y buscar un simétrico con respecto al centro de gravedad (centroide) del simplex. A continuación, evaluamos la función objetivo en este nuevo

vértice y continuamos como en el caso anterior. El avance del Simplex se efectúa siguiendo las reglas que se explican a continuación:

1. Se realiza un movimiento tras cada observación de respuesta (experimento).
2. El movimiento conduce hasta un nuevo simplex adyacente, que se obtiene despreciando el punto del simplex anterior que corresponde a la respuesta menos deseable y sustituirlo por su imagen especular respecto a la cara (o hipercara) definida por los puntos restantes.
3. Si el punto reflejado tiene una respuesta menos satisfactoria en el nuevo simplex, no aplicar la regla 2, sino, en su lugar, eliminar el segundo punto menos satisfactorio del simplex original y continuar la obtención de otro nuevo simplex.
4. Si se ha retenido un vértice en un simplex $k + 1$ dimensional $2k$ veces, antes de aplicar la regla 2, reobservar la respuesta en el vértice persistente.
5. Si el nuevo vértice cae fuera de los límites de las variables, asignar una respuesta arbitraria muy indeseable a dicho vértice.

La mayor desventaja del método simplex puede ser el gran número de experimentos que se deban realizar para alcanzar el óptimo. Este número será tanto mayor cuanto más alejado del óptimo hayamos iniciado el proceso. La razón está en la forma regular de la figura geométrica que utilizamos. Lo lógico es que cuando iniciamos el proceso los avances sean mayores en la dirección del óptimo y cuando estemos muy próximos al óptimo el simplex fuera capaz de realizar avances más cortos, evitando la necesidad de iniciar uno nuevo en esa región. Todo esto se puede lograr mediante el simplex modificado, propuesto por *Nelder y Mead (1965)*, un método especialmente eficiente si el número de variables no excede de 5 ó 6.

En este caso existen tres operaciones básicas para el movimiento del simplex:

- 1) Reflexión
- 2) Expansión
- 3) Contracción

Por otra parte, si el concepto del simplex modificado se lleva hasta sus últimas consecuencias se llega al método de contracción masiva, que consiste en iniciar la búsqueda con un simplex de gran tamaño, tanto que consideremos que el óptimo

estará en su interior y, por tanto, a partir del simplex inicial todos los movimientos serán contracciones sucesivas a simplex de menor tamaño hasta localizar el óptimo.

El método simplex ha alcanzado una considerable popularidad en los laboratorios analíticos en los últimos 15 años. Ciertamente, cuando el número de variables no es demasiado elevado y, sobre todo, cuando la superficie de respuesta es unimodal, los resultados son excelentes. Además, resulta especialmente indicado en procesos de desarrollo de metodología, en los que, normalmente, no se tienen indicaciones precisas de cuales serán los valores en las inmediaciones del óptimo. Es en estos casos donde resulta claramente recomendable. Por el contrario, si el proceso es relativamente conocido de antemano, es decir, si disponemos de información acerca de la localización aproximada del óptimo, o si se trata de efectuar mejoras en la calidad del mismo por ajuste fino de las variables los procedimientos de diseño discutidos anteriormente resultan generalmente preferibles.

8. BIBLIOGRAFÍA

Abrahamsson, Ch.; Johansson, J.; Sparén, A.; Lindgren, F. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. *Chemom. Intell. Lab. Syst.* 69, 3-12 (2003).

Adams, M.J. Chemometrics in analytical spectroscopy. RSC Analytical spectroscopy monographs edited by N.W. Barnett. The Royal Society of Chemistry (1995).

Affolter, Ch; Clerc, J.T. Prediction of infrared spectra from chemical structures of organic compounds using neural networks. *Chemometrics and Intelligent Laboratory Systems Information Management* 21, 151-157 (1993).

Andersson, G.G.; Kaufmann, P. Development of a generalized neural network. *Chemom. Intell. Lab. Syst.* 50, 101-105 (2000).

Andrade, J.M.; Muniategui, S.; López, P.; Prada, D. Costs, laboratory safety, productivity and faster research octane number and motor octane number determinations in industrial chemistry laboratories. *Analyst* 120, 249-253 (1995).

Andrade, J.M.; Sánchez, M.S.; Sarabia, L.A. Applicability of high-absorbance MIR spectroscopy in industrial quality control of reformed gasolines. *Chemom. Intell. Lab. Syst.* 46, 41-55 (1999).

Andrade, J.M.; Garrigues, S.; de la Guardia, M.; Gómez-Carracedo, M.; Prada, D. Non-destructive and clean prediction of aviation fuel characteristics through Fourier transform-Raman spectroscopy and multivariate calibration. *Anal. Chim. Acta* 482, 115-128 (2003).

Andreev, G.N.; Argirov, O.K.; Penchev, P.N. Expert system for the interpretation of infrared spectra. *Anal. Chim. Acta* 284, 131-136 (1993).

Artemenko, N.V.; Baskin, I.I.; Palyulin, V.A.; Zefirov, N.S. Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds. *Russian Chemical Bulletin, International Edition* 52(1), 20-29 (2003).

Ayora-Cañada, M.J.; Lendl, B. Sheath-flow Fourier Transform infrared spectrometry for the simultaneous determination of citric, malic and tartaric acids in soft drinks. *Anal. Chim. Acta* 417, 41-50 (2000).

Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* 17, 166-173 (2003).

Beebe, K.R.; Pell, R.J.; Seasholtz, M.B. Chemometrics: A Practical Guide. John Wiley & Sons (1998).

Bertran i Gimferrer, Q. Aplicación de la calibración multivariada al análisis industrial de control de calidad por espectrofotometría en UV-VIS, IR Medio e IR Cercano, Memoria de Tesis Doctoral, Departamento de Química Analítica, Universitat

Autónoma de Barcelona (1995).

Blanco, M.; Coello, J.; Iturriaga, H.; MasPOCH, S.; Pagés, J. NIR calibration in non-linear systems different PLS approaches and artificial neural networks. *Chemom. Intell. Lab. Syst.* 50, 75-82 (2000a).

Blanco, M.; MasPOCH, S.; Villarroja, I.; Peralta, X.; González, J.M.; Torres, J. Determination of the penetration value of bitumens by near infrared spectroscopy. *Analyst* 125, 1823-1828 (2000b).

Boger, Z.; Weber, R. Finding an optimal artificial neural network topology in Real-Life Modeling. In: The ICSC Symposium on Neural Computation, article nº 1, 1403/109 (2000).

Boger, Z. Selection of the quasi-optimal inputs in chemometric modeling by artificial neural networks analysis. *Anal. Chim. Acta* 490 (1-2), 31-40 (2003).

Boqué, R.; Ferré, J. Using second-order data in chromatographic analysis. *LC-GC Europe* 17(7), 402-407 (2004).

Borosy, A.P. Quantitative composition-property modelling of rubber mixtures by utilising artificial neural networks. *Chemom. Intell. Lab. Syst.* 47, 227-238 (1999).

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

Brown, S.D.; Sum, S.T.; Despaigne, F.; Lavine, B.K. Chemometrics. *Analytical Chemistry* 68(12), 21R-62R (1996).

Burke, L.; Ignizio, J.P. A practical overview of neural networks. *Journal of Intelligent Manufacturing* 8, 157-165 (1997).

Cela, R.; Pérez-Bustamante, J.A. El método simplex y sus aplicaciones en química analítica. *Química Analítica* 3(2), 87-128 (1984).

Centner, V.; Verdú-Andrés, J.; Walczak, B.; Jouan-Rimbaud, D.; Despaigne, F.; Pasti, L.; Poppi, R.; Massart, D-L.; de Noord, O.E. Comparison of Multivariate Calibration Techniques Applied to Experimental NIR Data Sets. *Applied Spectroscopy* 54(4), 608-623 (2000).

Cirovic, D.A. Feed-forward artificial neural networks: applications to spectroscopy. *Trends in Analytical Chemistry* 16(3), 148-155 (1997).

Coomans, D.; Broeckaert, I. Potential pattern recognition in chemical and medical decision making. Chemometric Series D.; D. Bawden (Ed.). Research Studies 3 Press, Letchworth (1986).

Cuadras, M. Métodos de análisis multivariante, EUNIBAR, Barcelona (1981).

Chen, H; Yu, Z.Y.; Zhu, G.B. Recognition of the bouquet of Chinese spirits by

- artificial neural network analysis. *Journal of AOAC International* 84(5), 1579-1585 (2001).
- Chung, H.; Ku, M-S; Lee, J-S. Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene. *Vibrational Spectroscopy* 20, 155-163 (1999).
- Chung, H.; Ku, M-S. Comparison of near-infrared, infrared, and Raman spectroscopy for the analysis of heavy petroleum products. *Applied Spectroscopy* 54(2), 239-245 (2000).
- Davies, A.M.C. Kissing cousins-the relationship between three popular regression methods. *Spectroscopy Europe* 8(6), 26-28 (1996).
- Davies, A.M.C. Cross-validation: do we love it too much? *Spectroscopy Europe* 10(2), 24-25 (1998a).
- Davies, A.M.C. The value of pictures. *Spectroscopy Europe* 10(4), 28-31 (1998b).
- Davies, A.M.C. More pictures from PLS regression. *Spectroscopy Europe* 10(6), 20-22 (1998c).
- Davies, A.M.C. Independence rules (or Rules for independence). *Spectroscopy Europe* 16(4), 27-28 (2004).
- Despagne, F.; Massart, L. Neural networks in multivariate calibration. *The Analyst* 123, 157R-178R (1998).
- Dieterle, F.; Busche, S.; Gauglitz, G. Growing neural networks for a multivariate calibration and variable selection of time-resolved measurements. *Anal. Chim. Acta* 490, 71-83 (2003).
- DiFoggio, R. Examination of Some Misconceptions about Near-Infrared Analysis. *Applied Spectroscopy* 49(1), 67-75 (1995).
- Downey, G. Food and food ingredient authentication by mid-infra-red spectroscopy and chemometrics. *Trends in Analytical Chemistry* 17(7), 418-424 (1998).
- Duarte, I.F.; Barros, A.; Delgadillo, I.; Almeida, C.; Gil, A.M. Application of FTIR spectroscopy for the quantification of sugars in mango juice as a function of ripening. *Journal of Agricultural and Food Chemistry* 50, 3104-3111 (2002).
- Durán-Merás, I.; Muñoz de la Peña, A.; Espinosa-Mansilla, A.; Salinas, F. Multicomponent Determination of flavour enhancers in food preparations by partial least squares and principal component regression modelling of spectrophotometric data. *Analyst* 118, 807-813 (1993).
- Erb, R.J. Introduction to Backpropagation Neural Network Computation. *Pharmaceutical Research* 10(2), 165-170 (1993).

Estienne, F.; Pasti, L.; Centner, V.; Walczak, B.; Despagne, F.; Rimbaud, D.J.; De Noord, O.E.; Massart, D.L. A comparison of multivariate calibration techniques applied to experimental NIR data sets. Part II: Predictive ability under extrapolation conditions. *Chemom. Intell. Lab. Syst.* 58, 195-211 (2001).

Estienne, F.; Despagne, F.; Walczak, B.; de Noord, O.E.; Massart, D.L. A comparison of multivariate calibration techniques applied to experimental NIR data sets. Part III: Robustness against instrumental perturbation conditions. *Chemom. Intell. Lab. Syst.* 73, 207-218 (2004).

Esvensen, K.; Midtgaard, T.; Schönkopf, S. *Multivariate Analysis in Practice. Computer-Aided Modelling As (CAMO)*, Norway (1994).

Faber, K.; Kowalski, B.R. Improved Prediction Error Estimates for Multivariate Calibration by Correcting for the Measurement Error in the Reference Values. *Applied Spectroscopy* 51(5), 660-665 (1997).

Faber, N.K.M.; Duewer, D.L.; Choquette, S.J.; Green, T.L.; Chesler, S.N. Characterizing the Uncertainty in Near-Infrared Spectroscopic Prediction of Mixed-Oxygenate Concentrations in Gasoline: Sample-Specific Prediction Intervals. *Anal. Chem.* 70, 2972-2982 (1998).

Faber, N.K.M. Multivariate Sensitivity for the Interpretation of the Effect of Spectral Pretreatment Methods on Near-Infrared Calibration Model Predictions. *Anal. Chem.* 71(3), 557-565 (1999).

Faber, N.M.; Schreutelkamp, F.H.; Vedder, H.W. Estimation of prediction uncertainty for a multivariate calibration model. *Spectroscopy Europe* 16/1, 17-20 (2004).

Fernández Pierna, J.A.; Wahl, F.; de Noord, O.E.; Massart, D.L. Methods for outlier detection in prediction. *Chemom. Intell. Lab. Syst.* 63, 27-39 (2002).

Fernández Pierna, J.A.; Jin, L.; Daszykowski, M.; Wahl, F.; Massart, D.L. A methodology to detect outliers/inliers in prediction with PLS. *Chemom. Intell. Lab. Syst.* 68, 17-28 (2003).

Fernández-Sánchez, J.F.; Segura Carretero, A.; Benítez-Sánchez, J.M.; Cruces-Blanco, C.; Fernández-Gutiérrez, A. Fluorescence optosensor using an artificial neural network for screening of polycyclic aromatic hydrocarbons. *Anal. Chim. Acta* 510, 183-187 (2004).

Ferré, J.; Faber, N.M. Net analyte signal calculation for multivariate calibration. *Chem. Intell. Lab. Syst.* 69, 123-136 (2003).

Fix, E.; Hodges, J.L. Discriminatory analysis, nonparametric estimation: consistency properties. Report 4, Project 21-49-004. USAF School of Aviation Medicine. Randolph Field (TX) (1951).

- Fodor, G.E.; Kohl, K.B. Analysis of Middle Distillate Fuels by midband infrared spectroscopy. *Energy and Fuels* 7, 598-601 (1993).
- Fodor, G.E. Analysis of Petroleum Fuels by Midband Infrared Spectroscopy. International Congress & Exposition; Detroit, Michigan, February 28-March 3 (1994).
- Fodor, G.E., Mason, R.A.; Hutzler, S.A. Estimation of Middle Distillate Fuel Properties by FT-IR. *Applied Spectroscopy* 53(10), 1292-1298 (1999).
- Forina, M.; Armanino, C.; Learde, R.; Drava, G. A class-Modelling technique based on potential functions. *J. Chemometrics* 5, 435-453 (1991).
- Forina, M.; Drava, G.; Boggia, R.; Lanteri, S.; Conti, P. Validation procedures in near-infrared spectrometry. *Anal. Chim. Acta* 295, 109-118 (1994).
- Gao, L.; Ren, S.X. Simultaneous spectrophotometric determination of four metals by the Kernel partial least squares method. *Chemom. Intell. Lab. Syst.* 45 (1-2), 87-93 (1999).
- García-Mencía, M.V.; Andrade, J.M.; López-Mahía, P.; Prada, D. An empirical approach to update multivariate regression models intended for routine industrial use. *Fuel* 79, 1823-1832 (2000).
- Garrigues, S.; Andrade, J.M.; De la Guardia, M.; Prada, D. Multivariate calibrations in Fourier transform infrared spectrometry for prediction kerosene properties. *Anal. Chim. Acta* 317 (1-3), 95-105 (1995).
- Gasteiger, J.; Zupan, J. Neural networks in Chemistry. *Angew. Chem. Int. Ed. Engl.* 32, 503-527 (1993).
- Geladi, P.; Kowalski, B.R. Partial Least Squares: A Tutorial. *Analytica Chimica Acta* 185, 1-17 (1986).
- Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemometrics* 2, 231-246 (1988).
- Gómez-Carracedo, M.P.; Andrade, J.M.; Calviño, M.; Fernández, E.; Prada, D.; Muniategui, S. Multivariate prediction of eight kerosene properties employing vapour-phase mid-infrared spectrometry. *Fuel* 82, 1211-1218 (2003a).
- Gómez-Carracedo, M.P.; Andrade, J.M.; Calviño, M.A.; Prada, D.; Fernández, E.; Muniategui, S. Generation and mid-IR measurement of a gas-phase to predict security parameters of aviation jet fuel. *Talanta* 60, 1051-1062 (2003b).
- González Dou, F. Modelos de calibración en la determinación simultánea de mezclas, Memoria de Tesis Doctoral, Departamento de Química Analítica, Universitat Autònoma de Barcelona (1991).

- Goodacre, R.; Kell, D.B.; Bianchi, G. Food adulteration exposed by neural networks. *Analysis Europa* 5, 35-37 (1995).
- Goodacre, R.; Hammond, D.; Kell, D.B. Quantitative analysis of the adulteration of orange juice with sucrose using pyrolysis-mass spectrometry and chemometrics. *Journal of Analytical and Applied Pyrolysis* 40-41, 135-158 (1997).
- Hadjiiski, L.; Geladi, P.; Hopke, Ph. A comparison of modeling nonlinear systems with artificial neural networks and partial least squares. *Chemom. Intell. Lab. Syst.* 49, 91-103 (1999).
- Hartnett, M.; Lightbody, G.; Irwin, G.W. Chemometric techniques in multivariate statistical modelling of process plant. *Analyst* 121, 749-754 (1996).
- Havel, J.; Cuesta, F.; Jancar, L. PLS self-calibration method for the evaluation of rate constants and or initial concentrations in multicomponent kinetic-analysis. *Reaction Kinetics and Catalysis Letters* 49(1), 189-195 (1993).
- Hassel, P.A.; Martin, E.B.; Morris, J. Non-linear partial least squares. Estimation of the weight vector. *J. of Chemometrics* 16, 419-426 (2002).
- Hebb, D.O. *The Organization of Behavior*, Wiley, New York (1949).
- Helland, I.S. Some theoretical aspects of partial least squares regression. *Chemom. Intell. Lab. Syst.* 58, 97-107 (2001).
- Henkel, K.; Schmeißer, D. Back-propagation-based neural network with a two sensor system for monitoring carbon dioxide and relative humidity. *Anal. Bioanal. Chem.* 374, 329-337 (2002).
- Hernández-Caraballo, E.A.; Avila-Gómez, R.M.; Capote, T.; Rivas, F.; Pérez, A.G. Classification of Venezuelan spirituous beverages by means of discriminant analysis and artificial neural networks based on their Zn, Cu and Fe concentrations. *Talanta* 60, 1259-1267 (2003).
- Hernández-Caraballo, E.A.; Avila-Gómez, R.M.; Rivas, F.; Burguera, M.; Burguera, J.L. Increasing the working calibration range by means of artificial neural networks for the determination of cadmium by graphite furnace atomic absorption spectrometry. *Talanta* 63(2), 419-424 (2004).
- Hilera, J.R.; Martínez, V.J. *Redes Neuronales Artificiales. Fundamentos, Modelos y Aplicaciones*. RA-MA Editorial. Madrid (1995).
- Hopfield, J.J.; Tank, D.W. "Neural" computation of decisions in optimization problems. *Biological Cybernetics* 52, 141-152 (1985).
- Hopke, Ph. The evolution of chemometrics. *Anal. Chim. Acta* 500, 365-377 (2003).
- Höskuldsson, A. PLS regression methods. *J. Chemom.* 2, 211-218 (1988).

<http://cursos.itam.mx>, 2003.

<http://lisisu02.usal.es/~airene>, 2003.

<http://ohm.utp.edu.co/neuronales>, 2003.

<http://su.wikipedia.org>, 2004.

<http://ttt.alc.upv.es>, 2002.

<http://zhanshou.hypermart.net.htm>, 2003.

Jacobson, S.P.; Hagman, A. Chemical composition analysis of carrageenans by infrared spectroscopy using partial least squares and neural networks. *Anal. Chim. Acta* 284, 137-147 (1993).

Jansson, P.A. Neural Networks: An Overview. *Analytical Chemistry* 63(6), 357A-362A (1991).

Kalogirou, S.A. Artificial intelligence for the modeling and control of combustion processes: a review. *Progress in Energy and Combustion Science* 29, 515-566 (2003).

Kapur, G.S.; Sastry, M.I.S.; Jaiswal, A.K.; Sarpal, A.S. Establishing structure-property correlations and classification of base oils using statistical techniques and artificial neural networks. *Anal. Chim. Acta* 506, 57-69 (2004).

Kateman, G. Neural networks in analytical chemistry? *Chem. Intell. Lab. Syst.* 19 (2), 135-142 (1993).

Kelly, J.J.; Callis, J.B. Nondestructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. *Anal. Chem.* 62, 1444-1451 (1990).

Kompany-Zareh, M.; Massoumi, A.; Pezeshk-Zadeh, Sh. Simultaneous spectrophotometric determination of Fe and Ni with xylenol orange using principal component analysis and artificial neural networks in some industrial samples. *Talanta* 48, 283-292 (1999).

Koshoubu, J.; Iwata, T.; Minami, S. Application of the Modified UVE-PLS Method for a Mid-Infrared Absorption Spectral Data Set of Water-Ethanol Mixtures. *Applied Spectroscopy* 54(1), 148-152 (2000).

Koshoubu, J.; Iwata, T.; Minami, S. Elimination of the Uninformative Calibration Sample Subset in the Modified UVE (Uninformative Variable Elimination)-PLS (Partial Least Squares) Method. *Analytical Sciences* 17, 319-322 (2001).

Kowalski, B.R.; Gerlach, R.; Wold, H. Chemical Systems Under Indirect Observation. In: Jöreskog, K.G. and Wold, H. (Eds.). *Systems Under Indirect Observation, Part II*. Amsterdam: North-Holland (1982).

- Kvalheim, O. ; Karstang, T.V. Interpretation of latent-variable regression models. *Chemom. Intell. Lab. Syst.* 7 (1-2), 39-51 (1989).
- Kvasnicka, V.; Skenák, S.; Pospíchal, J. Application of recurrent neural networks in chemistry. Prediction of C^{13} NMR chemical shifts in a series of monosubstituted benzenes. *Journal of Chemical Information and Computer Science* 32, 742-747 (1992).
- Lang, G.A. NIRs monitor critical gasoline parameters. *Hydrocarbon Processing* 73 (2), 69-71 (1994).
- Lashley, K. In search of the engram. Society of Experimental Biology, Symposium 4, 454-482 (1950).
- Le Thanh, H.; Lendl, B. Sequential injection Fourier Transform infrared spectroscopy for the simultaneous determination of organic acids and sugars in soft drinks employing automated solid phase extraction. *Anal. Chim. Acta* 422, 63-69 (2000).
- Li, L-K.; Chau, F-T.; Leung, A. K-M. Compression of ultraviolet-visible spectrum with recurrent neural network. *Chemom. Intell. Lab. Syst.* 52, 135-143 (2000).
- Lindgren, F.; Geladi, P.; Wold, S. The Kernel algorithm for PLS. *J. Chemometrics* 7 (1), 45-59 (1993).
- Lindgren, F.; Geladi, P.; Rännar, S.; Wold, S. Interactive Variable Selection (IVS) for PLS. Part I. Theory and Algorithms. *J. Chemometrics* 8, 349-363 (1994).
- Lippmann, R.P. An Introduction to Computing with Neural Networks. *IEEE ASSP Magazine* 3 (4), 4-22 (1987).
- Liu, X.; Van Espen, P.; Adams, F.; Shou He Yan; Vanbelle, M. Classification of Chinese tea samples according to origin and quality by principal component techniques. *Anal. Chim. Acta* 200, 421-430 (1987).
- Livingstone, D.J.; Manallack, D.T.; Tetko, I.V. Data modelling with neural networks: Advantages and limitations. *Journal of Computer-Aided Molecular Design* 11, 135-142 (1997).
- López-Anreus, E.; Garrigues, S; de la Guardia, M. Simultaneous vapour phase Fourier Transform infrared spectrometric determination of butyl acetate, toluene and methyl ethyl ketone in paint solvents. *Analyst* 123, 1247-1252 (1998).
- López Fandiño, V.M. Análisis de componentes principales no lineales mediante redes neuronales artificiales de propagación hacia atrás: Aplicaciones del modelo de Kramer. Tesis realizada en el Instituto Químico de Sarria en el Departamento de Quimiometría, dirigida por Dr. Vicente García Espeso. Universitat Ramon Llull (1997).
- Lorber, A.; Wangen, D.E.; Kowalski, B.R. A theoretical Foundation for the PLS algorithm. *J. Chemometrics* 1, 19-31 (1987).

- Lorber, A.; Kowalski, B.R. Alternatives to Cross-Validatory Estimation of the Number of Factors in Multivariate Calibration. *Applied Spectroscopy* 44(9), 1464-1470 (1990).
- Luera Peña, W.E.; Minim, L.A. Aplicación de redes neuronales artificiales en la modelización del tratamiento térmico de alimentos. *Cienc. Tecnol. Aliment.* 3(2), 81-88 (2001).
- Lu, Y.; Bian, L.; Yang, P. Quantitative artificial neural network for electronic noses. *Anal. Chim. Acta* 417, 101-110 (2000).
- Macho, S.; Boqué, R.; Larrechi, M.S.; Rius, F.X. Multivariate determination of several compositional parameters related to the content of hydrocarbon in naphtha by MIR spectroscopy. *Analyst* 124, 1827-1831 (1999).
- MacLaurin, P.; Worsfold, P.J.; Norman, Ph.; Crane, M. Partial Least Squares Resolution of Multianalyte Flow Injection Data. *Analyst* 118, 617-622 (1993).
- Marengo, E.; Bobba, M.; Robotti, E.; Lenti, M. Hydroxyl and acid number prediction in polyester resins by near infrared spectroscopy and artificial neural networks. *Anal. Chim. Acta* 511(2), 313-322(2004).
- Marini, F.; Zupan, J.; Magrí, A.L. On the use of counterpropagation artificial neural networks to characterize Italian rice varieties. *Anal. Chim. Acta* 510, 231-240 (2004a).
- Marini, F.; Balestrieri, F.; Bucci, R.; Magrí, A.D.; Magrí, A.L.; Marini, D. Supervised pattern recognition to authenticate Italian extra virgin olive oil varieties. *Chemom. Intell. Lab. Syst.* 73, 85-93 (2004b).
- Martens, H.; Karstang, T.; Naes, T. Improved selectivity in spectroscopy by multivariate calibration. *J. Chemometrics*, 1, 201-219 (1987).
- Martens, H.; Naes, T. *Multivariate Calibration*, John Willey & Sons (1989).
- Martens, H. Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression. *Chemom. Intell. Lab. Syst.* 58, 85-95 (2001).
- Massart, D.L.; Vandeginste, B.G.M.; Buydens, L.M.C.; De Jong, S.; Lewi, P.J.; Smeyers-Verbeke, J. *Hanbook of Chemometrics and qualimetrics*, Elsevier (1997).
- McCulloch, W.S.; Pitts, W.H. A Logical Calculus of the Ideas Immanent in Neural Nets. *Bulletin of Mathematical Biophysics* 5, 115-133 (1943).
- Miller, J.N. Outliers in Experimental Data and Their Treatment. *Analyst* 118, 455-461 (1993).
- Minsky, M.L.; Papert, S.S. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, M.A. (1969).
- Naes, T. The design of calibration in near infrared reflectance. *J. Chemometrics* 1 (2), 121-134 (1987).

- Naes, T.; Risvik. Multivariate Analysis of data in sensory science. Data Handling in Science and Technology, volume 16. Elsevier Science B.V. (1996).
- Naes, T.; Isaksson, T.; Fearn, T.; Davies, T. Multivariate Calibration and Classification. NIR Publications (Chichester, UK) (2002).
- Nakamoto, T.; Fukuda, A; Moriizumi, T. Improvement of Identification Capability in an Odorsensing System. *Sensors and Actuators B* 3, 221-226 (1991).
- Nakamoto, T.; Fukuda, A; Moriizumi, T. Gas/Odor Identification by Semiconductor Gas-sensor Array and an Analog Artificial Neural Network Circuit. *Sensors and Actuators B* 8, 181-186 (1992).
- Nakamoto, T.; Fukuda, A; Moriizumi, T. Perfume and Flavour Identification by Odour-sensing System Using Quartz-resonator Sensor Array and Neural Network Pattern Recognition. *Sensors and Actuators B* 10, 85-90 (1993).
- Nelder, J. A. and Mead, R. A Simplex Method for Function Minimization. *Comput. J.* 7, 308-313 (1965).
- Ortiz, M.; Saez, J.A.; López Palacios, J. Typification of Alcoholic distillates by Multivariate Techniques using data from chromatographic analyses. *Analyst* 118, 801-805 (1993).
- Ortiz, M.C.; Sarabia, L.A. Componentes principales y correspondencias en Avances en Quimiometría Práctica, R. Cela. Editorial Universidad de Santiago de Compostela (1994).
- Pazos Sierra, A. Redes de neuronas artificiales y algoritmos genéticos. Ed. Univ. da Coruña (1996).
- Penza, M.; Cassano, G. Recognition of adulteration of Italian wines by thin-film multisensor array and artificial neural networks. *Anal. Chim. Acta* 509, 159-177 (2004).
- Pérez-Magariño, S.; Ortega, Heras, M.; González-San José, M.L.; Boger, Z. Comparative study of artificial neural network and multivariate methods to classify Spanish DO rose wines. *Talanta* 62(5), 983-990 (2004).
- Plackett, R.L.; Burman, J.P. The Design of Optimal Multifactorial Experiments. *Biometrika* 33, 305-325 (1946).
- Rai, P.; Majumdar, G.C.; DasGupta, S.; De, S. Prediction of the viscosity of clarified fruit juice using artificial neural network: a combined effect of concentration and temperature. *Journal of Food Engineering* 68 (4), 527-533 (2005).
- Ramadan, Z.; Hopke, Ph.K.; Johnson, M.J.; Scow, K.M. Application of PLS and Back-Propagation Neural Networks for the estimation of soil properties. *Chemom. Intell. Lab. Syst.* 75(1), 23-30 (2005).

- Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. *J. Chemometrics* 8, 111-125 (1994).
- Richards, E.; Bessant, C.; Saini, S. Optimisation of a neural network model for calibration of voltametric data. *Chemom. Intell. Lab. Syst.* 61, 35-49 (2002).
- Ritter, H.; Martinetz, T.; Schulten, K. Neural Computation and Self-Organizing Maps. Addison-Wesley Publishing Company, New York (1992).
- Rodriguez-Saona, L.E.; Fry, F.S.; McLaughlin, M.A.; Calvey, E.M. Rapid analysis of sugars in fruit juices by FT-NIR spectroscopy. *Carbohydrate Research* 336, 63-74 (2001).
- Rosenblatt, F. The Perceptron: a Perceiving and Recognizing Automation (project PARA). Technical Report 85-460-1, Cornell Aeronautical Laboratory (1957).
- Ruckebusch, C.; Duponchel, L.; Huvenne, J.-P. Interpretation and improvement of an artificial neural network MIR calibration. *Chemom. Intell. Lab. Syst.* 62, 139-198 (2002).
- Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by backpropagating errors. *Nature* 323, 533-536 (1986).
- Sánchez, M.S.; Sarabia, L.A. Efficiency of multi-layered feed-forward neural network on classification in relation to linear discriminant analysis, quadratic discriminant analysis and regularized discriminant analysis. *Chemom. Intell. Lab. Syst.* 28, 287-303 (1995).
- Sarabia, L.A.; Cruz Ortiz, M. El orden de la señal y el calibrado. Escuela de Quimiometría, Universidad de Burgos, Junio 2004.
- Schulz, H.; Derrick, M.; Stulik, D. Simple encoding of infrared spectra for pattern recognition. Part 2. Neural network approach using back-propagation and associative Hopfield memory. *Anal. Chim. Acta* 316, 145-159 (1995).
- Sekulic, S.; Seasholtz, M.B.; Wang, Z.; Kowalski, B.R.; Lee, S.E.; Holt, B.R. Nonlinear Multivariate Calibration Methods in Analytical Chemistry. *Analytical Chemistry* 65(19), 835A-845A (1993).
- Serneels, S.; Croux, Ch.; Van Espen, P.J. Influence properties of partial least squares regression. *Chemom. Intell. Lab. Syst.* 71, 13-20 (2004).
- Sivakesava, S.; Irudayaraj, J. Determination of sugars in aqueous mixtures using mid-infrared spectroscopy. *Applied Engineering in Agriculture* 16(5), 541-550 (2000).
- Sivakesava, S.; Irudayaraj, J. Prediction of Inverted cane sugar adulteration of honey by Fourier Transform Infrared Spectroscopy. *Journal of Food Science* 66(7), 972-978 (2001).

- Smits, J.R.M.; Melssen, W.J.; Buydens, L.M.C.; Kateman, G. Using artificial neural networks for solving chemical problems. Part I: Multi-layer feed-forward networks. *Chemom. Intell. Lab. Syst.* 22(2), 165-189 (1994).
- Spendley, W. ; Hext, G. R.; Himsworth, F. R. Sequential application of simplex designs in optimization and evolutionary operation. *Technometrics* 4, 441-461 (1962).
- Svozil, D.; Kvasnicka, V.; Pospichal, J. Introduction to multi-layer feed-forward neural networks. *Chemom. Intell. Lab. Syst.* 39(1), 43-62 (1997).
- Taylor, J.K. Quality Assurance of Chemical Measurements. Lewis Publishers, Michigan, USA (1987).
- Tenenhaus, M. La Regression PLS: Theorie et Pratique, Technip, Paris (1998).
- Thomas, E.V. A Primer on Multivariate Calibration. *Analytical Chemistry* 66(15), 795A-804A (1994).
- Tjomsland, T.; Hilland, J.; Christy, A.A.; Sjöblom, J.; Riis, M.; Friisø, T.; Folgerø, K. Comparison of infrared and impedance spectra of petroleum fractions. *Fuel* 75(3), 322-332 (1996).
- Todeschini, R. Modeling and prediction of molecular properties. Theory of Grid-Weighted Holistic Invariant Molecular (G-WHIM) Descriptors, Chemometrics '95, Oral presentation, 3-7 July (1995).
- Tong, C.S.; Cheng, K.C. Mass spectral search method using the neural network approach. *Chemom. Intell. Lab. Syst.* 49, 135-150 (1999).
- Tomás, X.; Andrade, J.M. Factores determinantes para la clasificación multivariada mediante curvas de potencia. *Afinidad LIV*, 468, 103-108 (1997).
- Tomás i Morer, X.; Andrade Garda, J.M. Application of simplified potential curves to classification problems. *Química Analítica* 18, 117-120 (1999).
- Tominaga, Y. Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. *Chemom. Intell. Lab. Syst.* 49, 105-115 (1999).
- Turing, A.M. On Computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 42, 230-265 (1936).
- Van Est, Q.C.; Schoenmakers, P.J.; Smits, J.R.; Nijssen, W.P. Practical implementation of neural networks for the interpretation of infrared spectra. *Vibrational Spectroscopy* 4, 263-272 (1993).
- Veltkamp, D.; Gentry, D. PLS-Modeling: User's Manual, Center for Process Analytical Chemistry, University of Washington (Seattle, Washington) (1988).

- Wakeling, I.N.; Morris, J.J. A test of significance for partial least squares regression. *J. Chemometrics*, 7 (4), 291-304 (1993).
- Weigel, U.M.; Herges, R. Automatic interpretation of infrared spectra: recognition of aromatic substitution patterns using neural networks. *Journal of Chemical Information and Computer Science* 32, 723-731 (1992).
- Werbos, P.J. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Cambridge M.A.: Harvard University dissertation (1974).
- Widrow, B.; Hoff, M. Adaptive switching Circuits, 1960 IRE WESCON Convention Record, Part 4, August 1960.
- Wienke, D.; Kateman, G. Adaptive resonance theory based artificial neural networks for treatment of open-category problems in chemical pattern recognition-application to UV-Vis and IR spectroscopy. *Chemom. Intell. Lab. Syst.* 23, 309-329 (1994).
- Wise, B.M.; Gallagher, N.B. Pls-Toolbox for use with MATLAB™. Eigenvector Technologies (1998).
- Wold, H. Soft Modeling by Latent Variables; the Non-linear Iterative Partial Least Squares Approach, en, GANI, J., Papers in Honour of M.S. Bartlett: Perspectives in Probability and Statistics, Academic Press (London)(1975).
- Wold, S.; Johansson, E.; Jellum, E.; Bjørnson, I.; Nesbakken, R. Application of SIMCA multivariate data analysis to the classification of gas chromatographic profiles of human brain tissues. *Anal. Chim. Acta* 133, 251-159 (1981).
- Wold, S.; Albano, C.; Dunn III, W.J.; Esbensen, K.; Hellberg, S.; Johansson, E.; Sjöström, M. Pattern Recognition: Finding and Using Irregularities in Multivariate Data. In: Martens, H.; Russwurm, H.Jr. (Eds.). Food research and Data Analysis. London: Applied Science Publishers (1983).
- Wold, S.; Albano, C.; Dunn III, W.J.; Esbensen, K.; Hellberg, S.; Johansson, E.; Sjöström, M. Multivariate Data Analysis in Chemistry. In: Kowalski, B.R. (Ed.). Chemometrics: Mathematics and Statistics in Chemistry. The Netherlands: D. Reidel (1984).
- Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. Multi-way principal components and PLS-analysis. *J. Chemometrics* 1, 41-56 (1987).
- Wold, S.; Kettaneh-Wold, N.; Skagerberg, B. Nonlinear PLS modeling. *Chemom. Intell. Lab. Syst.* 7(1-2), 53-65 (1989).
- Wold, S.; Sjöström, M.; Eriksson. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58, 109-130 (2001a).

Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* 58, 131-150 (2001b).

www.chemkeys.com/esp/md/peyo_5/mdoeq_1/metsim_3/metsim_3.htm, 2004.

www.dacs.dtic.mil, 2003.

www.dd.chalmers.se/~f96jost/superresolution, 2002.

www.doc.ic.ac.uk/~nd/surprise.html, 2003

www.modeloingenieria.edu.ar, 2003.

www.ulbra.tche.br/~danielnm, 2003.

www.uta.cl, 2003.

Yang, H.; Griffiths, P.R.; Tate, J.D. Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra. *Anal. Chim. Acta* 489, 125-136 (2003).

Yin, Ch.; Luo, Z.; Ni, M.; Cen, K. Prediction coal ash fusion temperature with a back-propagation neural network model. *Fuel* 77(15), 1777-1782 (1998).

Zhang, R.; Yan, A.; Liu, M.; Liu, H.; Hu, Z. Application of artificial neural networks for prediction of the retention indices of alkylbenzenes. *Chemom. Intell. Lab. Syst.* 45, 113-120 (1999).

Zorriassatine, F.; Tannock, J.D.T. A review of neural networks for statistical process control. *Journal of Intelligent Manufacturing* 9, 209-224 (1998).

Zupan, J.; Gasteiger, J. Neural networks: A new method for solving chemical problems or just a passing phase?. *Anal. Chim. Acta* 248(1), 1-30 (1991).

Zupan, J. Introduction to Artificial Neural Network (ANN) Methods: What They Are and How to Use Them. *Acta Chimica Slovenica* 41(3), 327-352 (1994).



Capítulo II

Métodos de selección de variables

Objetivo:

Se aborda en este capítulo el estudio metodológico de dos técnicas quimiométricas orientadas a la selección del mínimo subconjunto de variables que permita abordar procesos de clasificación y de regresión multivariante. Las dos técnicas elegidas, rotación de Procustes y algoritmos genéticos, tienen principios conceptuales muy diferentes y, por tanto, abordan el problema empleando dos perspectivas diferentes. Su separación del capítulo anterior se justifica por la entidad que posee esta temática y el momento de aplicación. No en vano, las técnicas aquí estudiadas podrían constituir (así se presenta en capítulos posteriores) un "pretratamiento" de los datos antes del modelado.

Índice:

1. Introducción
2. Rotación de Procustes
 - 2.1. Concepto
 - 2.2. Desarrollo matemático
 - 2.3. Necesidades de diagnósticos asociadas a la Rotación de Procustes
 - 2.4. Rotación de Procustes para comparar subespacios de las variables
 - 2.5. Aplicaciones de la Rotación de Procustes
3. Algoritmos Genéticos
 - 3.1. Introducción
 - 3.2. Fundamentos
 - 3.3. Los individuos
 - 3.4. Estructura de los AAGG
 - 3.5. Los Operadores Genéticos
 - 3.6. Ventajas y desventajas del uso de los AAGG
 - 3.7. Aplicaciones
 - 3.8. Descripción del problema abordado en esta Memoria
4. Bibliografía

1. INTRODUCCIÓN

Mientras que en las décadas previas a los años 80 los conjuntos de datos con “muchas variables” presentaban alrededor de la centena de variables, hoy en día la situación habitual consiste en tener un número de variables mucho mayor, incluso claramente mayor que el número de muestras (Höskuldsson, 2001; Wold et al., 2001). Este aumento puede causar problemas para el modelado de los datos, principalmente en la interpretación. Aunque los diferentes métodos multivariantes (PLS, SIMCA,...) han sido diseñados para manipular matemáticamente muchas variables, sus resultados (*loadings*, residuales, etc) buscan simplificar la visualización del problema y, muchas veces, no permiten su interpretación química de manera obvia. Una tendencia habitual para tratar de “solucionar” el problema es seleccionar unas pocas variables (supuestamente) descriptoras del problema e ignorar el resto. Como dicha selección es altamente subjetiva en muchos casos, esto supone un riesgo ya que puede llegar a disminuir la eficiencia del diagnóstico y la validez de los métodos multivariantes.

Ahora bien, la correcta selección de variables que representen el problema puede suponer un gran ahorro en:

- ▶ Tiempo de análisis de cada muestra
- ▶ Tiempo de trabajo del analista
- ▶ Tiempo para la emisión de resultados
- ▶ Los costos generales del análisis y, por ende, del laboratorio
- ▶ Además de una mayor simplificación de los estudios, mayor atención y control en aquellas variables consideradas como “fundamentales” y la no necesidad de prestar atención (medir, tratar, vigilar, etc) a las variables “redundantes”.

En esta Memoria se trata de abordar una selección rigurosa y objetiva de un pequeño conjunto de variables descriptoras del problema que se considere. Para ello se han implementado y comparado dos métodos objetivos de selección de variables: la denominada “rotación de Procustes” y los algoritmos genéticos. No en vano se trata de dos metodologías totalmente diferentes en sus principios conceptuales y modo de abordar el problema. La primera, rotación de Procustes, es paramétrica y determinista; la segunda, algoritmos genéticos, se basa en el actualmente denominado cálculo

“natural” y sigue un patrón de búsqueda de soluciones guiado por el principio de “supervivencia del más apto” o, abreviadamente, de la mejor solución.

2. ROTACIÓN DE PROCUSTES

2.1. CONCEPTO

Si se estudia la etimología de la palabra "Procustes" (o Procusto), además de la curiosidad cultural, en este caso, se pueden encontrar las ideas básicas que definirán la técnica matemática.

En la mitología griega, Procustes (o Procusto) (Προχρούστης) era el apodo de un gigante y anciano ladrón llamado Polifemón (Πουπήμων) o Damastes (Δαμάστης) que vivía en las riberas del río Cefisos (Κηφισός) en Ática. Polifemón tenía como costumbre invitar a los viajeros a su posada para que pasaran la noche y, una vez éstos dormían, los torturaba acostándolos en una cama de hierro y cortándoles o estirándoles las piernas hasta que encajasen exactamente en la longitud de su cama. De hecho, el apelativo Procustes está formado por dos palabras griegas. Un prefijo-conjunción, *πρὸ* (que significa antes, en frente) y un verbo, *χρούστης* (que significa pegar, golpear, empujar, dejar fuera de combate). Consecuentemente, Procustes es "aquel que golpea y estira" o "el que golpea para elongar". Raíces etimológicas similares se encuentran en el eslavo antiguo (*kruchu*=separar a puñetazos, golpear); en el lituano (*kruszu*=magullar, aporrear) y en el letón (*krauset*=dar patadas).

La técnica matemática de Procustes, en esencia, lo que hace es comparar dos (en general n) espacios k -dimensionales de forma que se busque su máxima similitud mediante giros, traslaciones, elongaciones, etc. De alguna forma, pues, lo que se pretende es adaptar un espacio dimensional a otro (u otros). Por cuestiones prácticas, uno de los espacios se mantiene fijo y es el otro el que se ajusta (se "tortura"). Si esto se aplica a la idea de reducción de variables, se podrá evaluar la diferencia entre el espacio original y el "reducido".

En consecuencia, en el momento en el que un subespacio definido por algunas de las **variables originales** se diferencie poco o nada del espacio definido por el conjunto total, se habrá logrado el objetivo de reducir variables sin perder información (o perdiendo la mínima posible).

Además de su base matemática, la rotación Procustes es un "camino natural" bastante intuitivo de comparación de objetos similares, pudiendo ser algo que nuestro

cerebro empleara en la vida diaria. Como ejemplo, piénsese en lo que ocurre cuando tenemos una visión difusa de la cara de una persona mientras corremos por una calle, seguramente nuestro cerebro hará alguna clase de "comparación Procustea" similar a la planteada en la **Figura 1**.

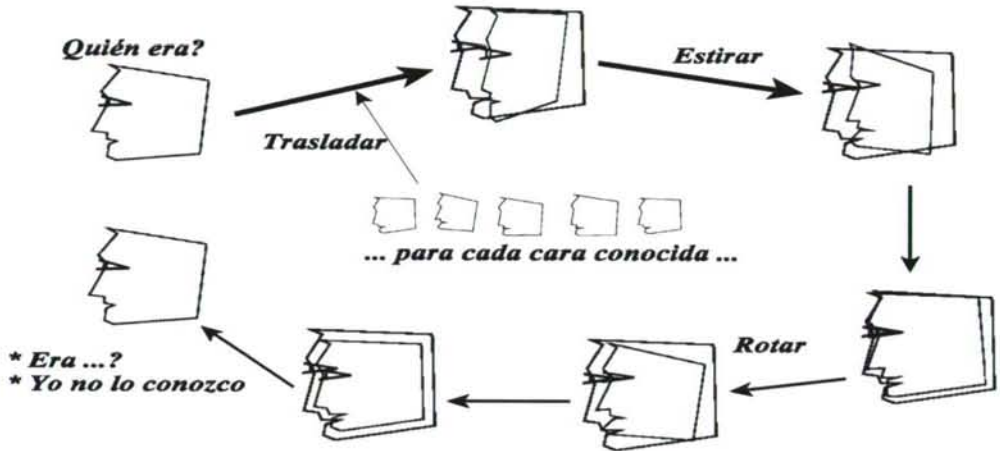


Figura 1: Ejemplo de rotación de Procustes.

Con esta idea general en mente, la forma inmediata de trabajar a la hora de reducir el número de las variables que describen un problema es evidente y consistiría en comparar sistemáticamente todos los posibles conjuntos de variables con los datos originales. De hecho, tan sólo habría que definir una función objetivo a minimizar y tomar como selección más adecuada la que conduzca a dicho mínimo. Ahora bien, existen dos grandes problemas:

- 1.- El volumen de cálculo es enorme
- 2.- En qué orden introducir las variables en el modelo

A semejanza de lo que ocurre en regresión lineal múltiple, podría pensarse en procedimientos de inclusión "hacia adelante" (*forward*), "hacia atrás" (*backward*) o "paso a paso" (*step by step*). Sin embargo, la experiencia conduce a resultados divergentes por los tres métodos, por lo que la vía no parece aconsejable.

2.2. DESARROLLO MATEMÁTICO

La técnica de Procrustes es bastante conocida en el mundo matemático pero no ha sido hasta muy recientemente cuando se logró su desarrollo (especialmente, la generalización de los teoremas) y aplicación práctica.

Tres de los grandes investigadores en este área son Gower (Gower, 1975), quien generalizó la teoría, Krzanowski (Krzanowski, 1979; Krzanowski, 1987a; Krzanowski, 1987b; Krzanowski, 2001) y Arnold (Arnold y Williams, 1986; Arnold, 1992; Arnold y Collins, 1993). Arnold hace una revisión hasta 1986 (Arnold y Williams, 1986) donde explica el uso de la técnica de Procrustes generalizada. Recientemente el método ha recibido mayor atención, por la disponibilidad de software y por algunas críticas del método (Dijksterhuis, 1996).

Krzanowski se basó en los desarrollos de las rotaciones de matrices para permitir el desarrollo de la "rotación de Procrustes" (Eckart y Young, 1936; Schönemann y Carroll, 1970; Gower, 1975) y en la descomposición de valores singulares de Bunch (Bunch et al., 1978; Bunch y Nielsen, 1978). Conceptualmente, hay dos grandes vías de aplicación de la técnica de Procrustes (Krzanowski, 2001):

- 1.- Dados n -puntos sobre los cuales se han medido dos conjuntos de características diferentes, se desea analizar si las pautas de distribución en los conjuntos de variables son similares. A pesar de que éste no es un problema típico de la Química Analítica, su análisis matemático es, precisamente, la base para establecer la técnica de la Rotación de Procrustes (o rotaciones Procrusteanas).
- 2.- Mucho más frecuente en la Química Analítica es disponer de un grupo de p -variables a medir en distintos conjuntos de muestras y desear saber el grado de coincidencia de un conjunto de subespacios en relación a esos ejes. Dicho de otra forma, determinar si las fuentes de variabilidad son iguales o similares a lo largo de los conjuntos de muestras. Típicamente, éste puede ser un método de validación de estudios previos, y que también hace uso de la Rotación de Procrustes.

Pártase de la idea central de que se desean eliminar las variables redundantes sin perder información esencial. Esto debe hacerse mediante la minimización de una función objetivo. Las aproximaciones de MacCabe y Jolliffe no eran satisfactorias,

básicamente porque sólo consideraban la reproducción de la matriz de varianza-covarianza (Krzanowski, 1987a).

Para saber si se pierde o no información al eliminar una variable, la vía más lógica parece comparar el espacio de objetos antes y después de su eliminación, estableciendo una medida de discrepancia. A menor discrepancia observada, más probable es que la variable sea redundante. Dado que en ambos espacios dimensionales el número de objetos es el mismo puede derivarse el criterio de las Rotación de Procrustes (Gower, 1971; Sibson, 1978; Krzanowski, 1987a; Krzanowski, 1987b).

Surge un problema que debe solucionarse ya en este momento. ¿Qué espacios dimensionales van a compararse? Aunque podría pensarse en emplear las variables originales directamente hay, al menos, dos motivos que lo desaconsejan:

- 1.- El "ruido" de los datos, que puede enmascarar la información relevante
- 2.- Al eliminar una variable puede haber diferencias importantes si la comparación es directa debido a problemas de escala, correlaciones, etc.

Por este motivo, se compararán los subespacios dimensionales obtenidos mediante PCA. La comparación (igualamiento) deberá estudiarse en cuatro operaciones:

1. **Traslación:** desplazamiento fijo de todos los puntos a un origen de coordenadas común.
2. **Rotación rígida:** desplazamiento de los puntos en un ángulo constante.
3. **Dilatación:** (contracción), del espacio por un movimiento constante de todos los puntos.
4. **Reflexión:** la reflexión puede considerarse formada por una combinación de rotaciones y se puede tomar como englobada en el punto 2.

A partir de todo ello se puede definir un estadístico sencillo que mida el grado de coincidencia de dos configuraciones como la suma de las diferencias al cuadrado entre las coordenadas de los puntos al realizar secuencialmente las cuatro operaciones anteriores. A este estadístico se le llama **Rotación de Procrustes (M^2)**.

Sea $X_{(n \times p)}$ la matriz de datos originales (n objetos y p variables medidas). Supóngase que la dimensionalidad esencial de los datos (número óptimo de

componentes principales) es A . Sea $Y_{(n \times A)}$ la matriz de scores de las muestras en los A componentes principales y que representan la mejor aproximación A dimensional a la configuración inicial de X . Sea $\tilde{X}_{(n \times q)}$ la matriz de datos que retiene solamente q -variables seleccionadas y sea $Z_{(n \times A)}$ la matriz de los scores de las muestras en los datos reducidos (el número de componentes ha de ser el mismo en ambos espacios). Esta última es la mejor aproximación A dimensional a la configuración q -dimensional. Para medir las discrepancias entre las configuraciones Y y Z se realiza un análisis procusteano; es decir, se calculan las diferencias al cuadrado entre los puntos correspondientes de las configuraciones después de su comparación por traslación, rotación y dilatación:

[ec. 1]

$$M^2 = \sum_{i=1}^n \left[\sum_{j=1}^A (y_{ij} - z_{ij})^2 \right]$$

El esquema de trabajo se muestra en la **Figura 2**.

2.2.1. IGUALAMIENTO POR TRASLACIÓN

La existencia de un desplazamiento constante de todos los puntos no es un motivo de diferenciación de pautas entre objetos. La comparación se hace entre los centroides de ambas configuraciones. Los centroides G_y y G_z tendrán coordenadas $(\bar{y}_1, \dots, \bar{y}_A)$ y $(\bar{z}_1, \dots, \bar{z}_A)$ donde la barra superior indica valores medios.

Sumando y restando los centroides a la ec.1, se obtiene (Krzanowski, 1990):

[ec. 2]

$$M^2 = \sum_{i=1}^n \left[\sum_{j=1}^A [(y_{ij} - \bar{y}_j) - (z_{ij} - \bar{z}_j) + (\bar{y}_j - \bar{z}_j)]^2 \right]$$

Tomando los dos primeros términos entre paréntesis como una unidad, se tiene un binomio $(a+b)^2$ que se desarrolla. Además por propiedad de valor medio,

$$\sum_{i=1}^n (y_{ij} - \bar{y}_j) = \sum_{i=1}^n (z_{ij} - \bar{z}_j) = 0, \text{ con lo que:}$$

[ec. 3]

$$M^2 = \sum_{i=1}^n \sum_{j=1}^A [(y_{ij} - \bar{y}_j) - (z_{ij} - \bar{z}_j)]^2 + n \sum_{j=1}^A (\bar{y}_j - \bar{z}_j)^2$$

pero el primer sumando está formado por elementos de las matrices Y y Z después de

centrar en la media y , por tanto, lograr la coincidencia de centroides. El segundo sumando sería la distancia entre los centroides, la cual evidentemente es cero si los datos de Y y Z se centran en la media ya de partida. Así, este término se ha simplificado totalmente teniendo en cuenta la práctica de centrado habitual en Química Analítica. Con ello también se simplifica el proceso de ajuste por dilatación /contracción (Krzanowski, 2001).

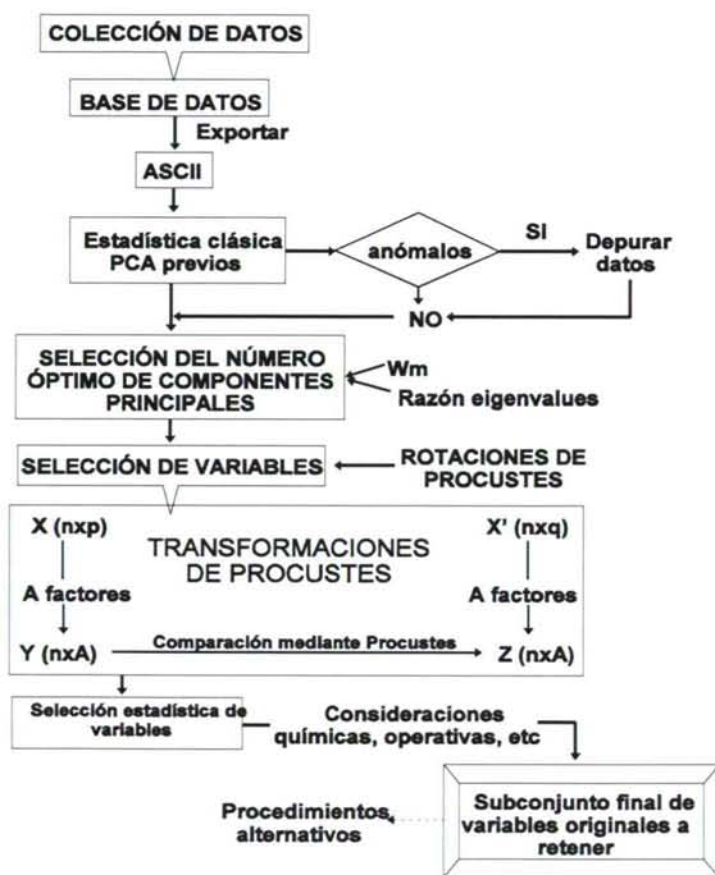


Figura 2: Esquema de trabajo para seleccionar variables mediante la técnica de Procrustes.

2.2.2. IGUALAMIENTO POR ROTACIÓN

Cualquier rotación de **Z** sobre **Y** se puede expresar mediante una matriz ortogonal **Q**. Las coordenadas posteriores a la rotación son las filas de **ZQ** (Krzanowski, 1987a).

La ec.1 puede escribirse como (el apóstrofe indica traspuesta; Tr indica traza de la matriz):

[ec. 4]
$$M^2 = \text{Tr}[(Y-Z)(Y-Z)']$$

[ec. 5]
$$M^2 = \text{Tr}(YY' + ZZ' - 2YZ')$$

Después de la rotación, **Z** se convierte en **ZQ** y se obtiene:

[ec. 6]
$$M^2 = \text{Tr}(YY' + ZQQ'Z' - 2YQ'Z')$$

Como **Q** es una matriz ortogonal, **QQ'**=**I** (matriz identidad). Así:

[ec. 7]
$$M^2 = \text{Tr}(YY' + ZZ' - 2YQ'Z')$$

Para lograr un sumatorio de residuales (M^2) pequeño, debe elegirse una matriz de rotación (**Q**) que aproxime al máximo los dos subespacios. Una vía para lograrlo es mediante la descomposición en valores singulares de la matriz **Y'Z**. Así, **Q=VU'**, donde **UΣV'** es la descomposición de **Y'Z** y donde **U'U=I_A**, **V'V=VV'=I_A** y **Σ=diag(σ₁, ..., σ_A)**. Como se cumple que **Tr(AB)=Tr(BA)**, la **Tr(YQ'Z')** = **Tr(Q'Z'Y)** y teniendo en cuenta la definición de **Q** y la descomposición en valores singulares:

[ec. 8]
$$\text{Tr}(Q'Z'Y) = \text{Tr}(UV'VΣU') = \text{Tr}(V'VΣU'U) = \text{Tr}(Σ)$$

Con lo cual (Krzanowski, 1987a; Deane y Macfie, 1989; Andrade et al., 1993):

[ec. 9]
$$M^2 = \text{Tr}(YY' + ZZ' - 2Σ)$$

A la vista de 2.2.1 y 2.2.2 la ecuación 9 es la que define el Criterio de Rotación de Procustes. El proceso de trabajo es secuencial y se estructura en bucles:

- 1.- Elegir el tamaño del subespacio (número de componentes principales, A)
- 2.- Obtener una matriz **Z** para cada variable eliminada sucesivamente e identificar la que conduce a una suma de cuadrados de Procustes (M^2) mínima
- 3.- Eliminar la variable recién identificada

4.- Continuar el ciclo hasta que sólo queden A variables

2.3. NECESIDADES DE DIAGNÓSTICOS ASOCIADAS A LA ROTACIÓN DE PROCUSTES

2.3.1. SELECCIÓN DEL NÚMERO ÓPTIMO DE COMPONENTES PRINCIPALES

La selección del número más adecuado de componentes principales para describir un sistema es un tema abierto en Quimiometría. Hay que aceptar que no hay un único criterio válido y que lo más adecuado y que conduce a mejores resultados es el uso conjunto de varios criterios en cada problema para alcanzar una conclusión final.

Sarabia (Sarabia y Ortiz, 1994) y Brereton (Brereton, 1992) recopilan algunos criterios tales como el del valor propio, bastón roto, criterio de Malinowski, el test F, relación de *eigenvalues*, criterio de la F de Krzanowski, etc.

Wold (Wold et al., 1987) ya establece una primera discusión que luego es seguida, entre otros, por Malinowski (Malinowski, 1987), el cual estudia la distribución de errores en los *eigenvalues* y su influencia en la determinación del número adecuado de factores en datos espectroscópicos. Sutter (Sutter et al., 1992), discute este problema en el entorno de la regresión por componentes principales. Scarponi (Scarponi et al., 1990) realiza un estudio aplicado al análisis de vinos. Fay (Fay et al., 1991), analiza el poder discriminante de los componentes principales cuando los datos (espectrales) tienen una fuerte carga de ruido. Por su parte, Osten (Osten, 1988) también hace estudios basados en diversas técnicas de selección del número de componentes más adecuado en función de distintos modos de evaluar el PRESS. Del estudio deduce que la selección del mínimo absoluto en función de la crossvalidación puede conducir a problemas y demuestra que la selección basada en el test-F es más robusta.

De la bibliografía y experiencia aplicada, se deduce que el criterio de Malinowski conduce a resultados francamente buenos (y es sencillo) (Carlosena et al., 1995); lo mismo cabe decir para el criterio de la razón de *eigenvalues*.

La tendencia general es aceptar como más fiable y "robusta" la selección del número de componentes mediante técnicas de validación cruzada o *cross-validation* a pesar de que se sabe que en los casos de modelización para regresión multivariante puede conducir a un sobreajuste (Dieterle et al., 2003). Precisamente, Eastment y Krzanowski (Eastment y Krzanowski, 1982; Krzanowski, 1987a) proponen un estadístico

que se basa en la técnica de *cross-validation* (modificada) empleada por Wold y que conduce a resultados muy interesantes. Tal estadístico se ha revelado muy útil en diversas aplicaciones realizadas (Krzanowski, 1987b; Deane y Macfie, 1989; Scarponi et al., 1990; Andrade et al., 1993; Andrade et al., 1994; Carlosena et al., 1995; Andrade et al., 2003).

2.3.1.1. CRITERIO W DE KRZANOWSKI

El uso de este criterio en el presente trabajo se debe a tres razones fundamentales:

- 1.- Está basado en la técnica de *cross-validation*.
- 2.- Krzanowski consigue mejorar el criterio de Wold ya que hace real la expresión "dejar cada dato fuera".
- 3.- Además de ser el criterio empleado (y recomendado) por Krzanowski, se ha comprobado que conduce a resultados satisfactorios a pesar de que tiene el problema de que no es fácil la presentación gráfica de resultados (Osten, 1988; Deane y Macfie, 1989; Scarponi et al., 1990; Andrade et al., 1993).

El desarrollo completo del test se presenta en Eastment y Krzanowski (1982) y aquí se resume. Si se emplea la descomposición de valores singulares para calcular los componentes principales, el problema es elegir el mínimo número de ellos (sea, A) que permite representar el espacio original con el menor error posible.

$$[\text{ec. 10}] \quad x_{ij} = \sum_{t=1}^A (u_{it} s_t v_{tj}) + \epsilon_{ij}$$

Asociado a cada valor de A , habrá un predictor

$$[\text{ec. 11}] \quad \hat{x}_{ij}(A) = \sum_{t=1}^A (u_{it} s_t v_{tj})$$

El criterio de *cross-validation* implica ir eliminando grupos de datos de la matriz original y predecirlos una vez recalculado el modelo. La función objetivo es una relación entre los valores reales y los predichos. El valor de A óptimo será aquel para el cual esa función objetivo sea mínima a lo largo de los subgrupos predichos. De acuerdo con Wold, la discrepancia entre los valores reales y predichos se denomina "Predicted Residual Error Sum of Squares" (Sumatorio de los Errores Residuales al cuadrado en la Predicción) (**PRESS**) y se expresa como:

[ec. 12]

$$\text{PRESS}(A) = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (\hat{x}_{ij}(A) - x_{ij})^2$$

El problema principal es que un mismo dato no se debe emplear a la vez para modelizar y predecir. La solución de Eastment y Krzanowski (1982) es que para predecir el dato x_{ij} , no se deben usar ni la columna j ni la fila i . Es decir, en realidad se eliminan variables y elementos, lo que permite una verdadera *cross-validation* "uno a uno". Sea $\mathbf{X}_{(-j)}$ el resultado de eliminar la columna j de \mathbf{X} y $\mathbf{X}_{(-i)}$ el de eliminar la fila i . La descomposición de valores singulares conduce a:

[ec. 13]

$$\begin{aligned} \mathbf{X}_{(-j)} &= \tilde{\mathbf{U}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}' \\ \mathbf{X}_{(-i)} &= \bar{\mathbf{U}} \bar{\mathbf{D}} \bar{\mathbf{V}}' \\ \tilde{\mathbf{U}} &= (\tilde{u}_{st}); \tilde{\mathbf{V}} = (\tilde{v}_{st}); \tilde{\mathbf{D}} = \text{diag}(\tilde{d}_1, \dots, \tilde{d}_{p-1}) \\ \bar{\mathbf{U}} &= (\bar{u}_{st}); \bar{\mathbf{V}} = (\bar{v}_{st}); \bar{\mathbf{D}} = \text{diag}(\bar{d}_1, \dots, \bar{d}_n) \end{aligned}$$

Con lo cual, el predictor de x_{ij} que se busca es:

[ec. 14]

$$\hat{x}_{ij}(A) = \sum_{t=1}^A (\tilde{u}_{it} \sqrt{\tilde{d}_t}) (\bar{v}_{tj} \sqrt{\bar{d}_t})$$

ya que el factor de la izquierda proviene de la descomposición sin considerar la columna j y el factor de la derecha de no considerar la fila i . Esta operación hay que hacerla para cada valor de la matriz de datos y a lo largo de todos los A valores. Así, para cada valor de A se tiene el estadístico:

[ec. 15]

$$W_A = \frac{\frac{\text{PRESS}_{(A-1)} - \text{PRESS}_{(A)}}{D_A}}{\frac{\text{PRESS}_{(A)}}{D_R}}$$

Donde D_A son los grados de libertad para ajustar el A componente ($D_A = n + p - 2A$) y D_R son los grados de libertad restantes ($D_R = (n-1)p - D_A$). Así, W_A representa el incremento en la información predictiva aportada por el A componente dividida por la información promedio en el resto de los componentes. En la práctica, un factor "importante" será aquel con un $W_A \geq 0.9$, si bien a veces la situación es más compleja.

2.3.1.2. OBJETOS ANÓMALOS

Como toda técnica multivariante basada en los componentes principales, la existencia de datos anómalos distorsiona totalmente los modelos. Aprovechando los estudios que se acaban de desarrollar, es fácil estudiar la influencia de cada objeto en el modelo.

Dado que \bar{V} contiene los coeficientes de los componentes principales cuando se ha eliminado la fila i , una comparación entre las primeras A columnas de \bar{V} y V indicará la influencia del objeto i . Krzanowski ha demostrado (Krzanowski, 1979) que el ángulo que forman ambos subespacios se puede calcular como:

$$[\text{ec. 16}] \quad \theta = \cos^{-1}(d)$$

donde d es el valor propio (*eigenvalue*) más pequeño de la descomposición de $V'_{(A)} \bar{V}_{(A)}$. A mayor ángulo, mayor influencia del objeto. Los estudios de Monte Carlo realizados no permitieron establecer criterios cuantitativos (Krzanowski, 1987b), por ello, lo que se hace es representar gráficamente los ángulos obtenidos para las muestras y observar la pauta de distribución, de acuerdo con ello se pueden tomar decisiones.

2.3.1.3. INFLUENCIA DE CADA VARIABLE

Análogamente al epígrafe anterior, \tilde{U} y \tilde{D} contienen la información en ausencia de la variable j ; ($\tilde{Z} = \tilde{U} \tilde{D}$, son los *scores*) y su influencia se analiza comparando dos configuraciones dimensionales de los mismos puntos. Esto es, de nuevo, una aplicación de la Rotación de Procrustes en donde se comparan las primeras A columnas de \tilde{Z} con Z para cada valor de A .

Deane y Macfie (1989) han sugerido una comparación entre la estructura perdida (M^2) para cada variable (conjunto de variables) y la información total del sistema ($\text{Tr}(\mathbf{Y}\mathbf{Y}')$). La gráfica obtenida es muy interesante ya que permite ver si, a pesar de que se haya identificado una variable como la que produce una menor pérdida, dicha variable es crítica y conviene no eliminarla (según el conocimiento que se tenga del sistema). La información perdida se define mediante la ecuación 17:

[ec. 17]

$$\%Infor.perdida = \left[\frac{Tr(YY' + ZZ' - 2\Sigma) \cdot 100}{Tr(YY')} \right]$$

donde $Tr(YY')$ es la información total que existe originalmente en los datos.

2.4. ROTACIÓN DE PROCUSTES PARA COMPARAR SUBESPACIOS DE LAS VARIABLES

Ésta es la segunda vía de aplicación de esta técnica. Resulta de gran utilidad cuando se pretende averiguar si dos (en general n) conjuntos de objetos presentan pautas similares de distribución. El punto de partida es que sobre ellas se han analizado las mismas p variables. En esta Memoria no se hará uso de esta aplicación por lo que tan sólo se dan algunos detalles.

Sean X e Y dos muestras multivariantes de n_1 y n_2 objetos respectivamente. Supóngase que un estudio de componentes principales ha indicado que A es el número adecuado de componentes (restricción que no es esencial, pero sí conveniente por su comodidad para la explicación). Sean x_1, \dots, x_p las variables originales medidas en los objetos, l_{ij} y m_{ij} son los coeficientes de los *loadings*. Se trata de comparar líneas, superficies o hiperespacios empleando como medida de similitud los ángulos que forman entre ellas. La visión clásica es que dos líneas (ejes, componentes principales) se parecen tanto más cuanto menor es su ángulo. La misma idea conceptual aplica en p -dimensiones.

Sean L y M las matrices ($A \times p$) que contienen los *loadings* respectivos. Defínase $N = LM'ML'$ ($=TT'$, donde $T = LM'$). Krzanowski ha demostrado los dos resultados siguientes (Krzanowski, 1979):

Teorema 1:

El mínimo ángulo entre un vector arbitrario en el espacio de los A primeros componentes principales de X y el vector paralelo más próximo a éste en el espacio de los primeros k componentes principales de Y viene dado por $\cos^{-1}\sqrt{\lambda_1}$, donde λ_1 es el valor propio (*eigenvalue*) mayor de N .

Teorema 2:

Sea λ_i el i -*eigenvalue* de N , a_i su *eigenvector* asociado y $b_i = L'a_i$ ($i = 1, \dots, A$).

Entonces, $\mathbf{b}_1, \dots, \mathbf{b}_A$ forman un conjunto de vectores ortogonales embebidos en el subespacio \mathbf{A} y $\mathbf{M}'\mathbf{M}\mathbf{b}_1, \dots, \mathbf{M}'\mathbf{M}\mathbf{b}_A$ el correspondiente conjunto de vectores en el subespacio \mathbf{B} en el cual se reparten las diferencias entre los subespacios. El ángulo entre la pareja $\mathbf{b}_i, \mathbf{M}'\mathbf{M}\mathbf{b}_i$ viene dado por $\cos^{-1}\sqrt{\lambda_i}$, ($i= 1, \dots, A$).

Por tanto, la cuantificación de hasta cuánto difieren dos espacios A -dimensionales viene dado por los ángulos $\cos^{-1}\sqrt{\lambda_1}, \dots, \cos^{-1}\sqrt{\lambda_A}$.

Para interpretar químicamente las similitudes o diferencias entre \mathbf{X} e \mathbf{Y} , se considera la pareja de autovectores \mathbf{b}_i y $\mathbf{M}'\mathbf{M}\mathbf{b}_i$ asociados al *eigenvalue* λ_i . Estos vectores están evidentemente definidos a partir de los p -ejes originales \mathbf{y} , por tanto, tienen interpretación.

Para ello, se define un nuevo vector en el espacio original de p -dimensiones que está próximo tanto a \mathbf{b}_i como a $\mathbf{M}'\mathbf{M}\mathbf{b}_i$ (de hecho es su bisector) y que se llama "**vector consenso**" (ver Figura 3), ec.18.

[ec. 18]

$$\mathbf{c}_i = [2(1+\sqrt{\lambda_i})]^{-1/2} * (\mathbf{I} + \frac{1}{\sqrt{\lambda_i}} \mathbf{M}'\mathbf{M}) \mathbf{b}_i$$

donde \mathbf{I} es la matriz identidad ($p \times p$).

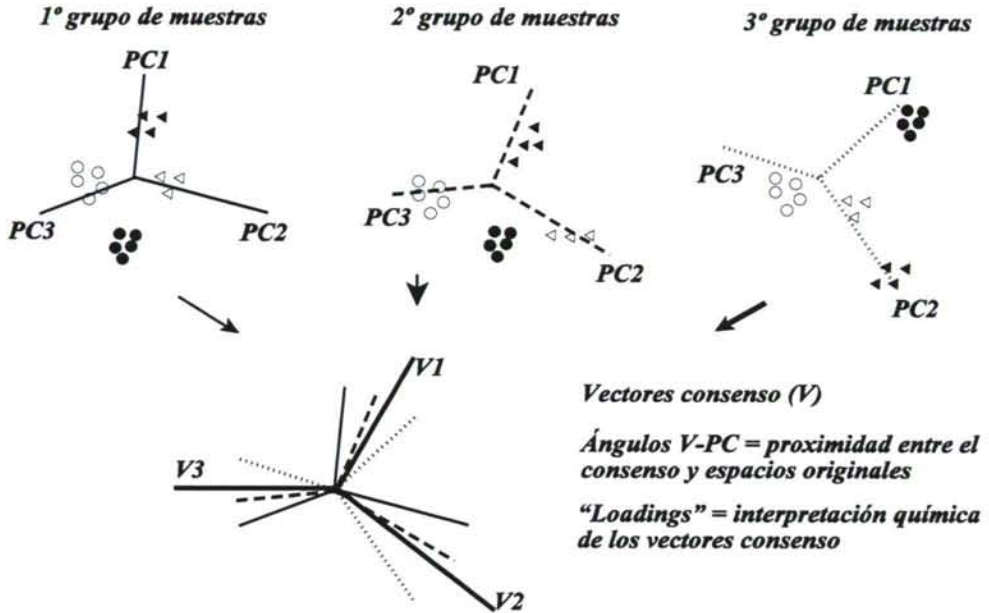


Figura 3: Interpretación gráfica de los vectores consenso de la Rotación de Procustes.

2.5. APLICACIONES DE LA ROTACIÓN DE PROCUSTES

La gran utilidad de esta técnica es seleccionar tan sólo unas cuantas variables analíticas reales (estadísticamente el menor subconjunto posible) que deben ser medidas. Esto constituye una ventaja comparativa con otras opciones donde el objetivo es localizar rangos o intervalos. A modo de ejemplo de esta última opción cabe citar los siguientes trabajos:

-Grimalt y Olive (1993) emplean un análisis de factores a partir de cuyos *loading* seleccionan las variables aparentemente más importantes.

- Centner et al. (1996) emplean un método en el que se hace una regresión preliminar con todas las variables; añaden ruido aleatorio artificial y seleccionan las variables experimentales que muestran más importancia que las artificiales (de acuerdo con un criterio basado en los coeficientes de regresión).

- Garrido-Frenich et al. (1999) seleccionan las variables con más altos coeficientes tras hacer un PLS.

-Goicoechea y Oliveri (1999) emplean un error indicador de análisis de regresión lineal híbrido.

- Ferré y Rius (1997) emplean un método basado en considerar la región de confianza de las concentraciones estimadas y optimizar cinco criterios diferentes, seleccionando el grupo óptimo valorando todas las posibles combinaciones de variables.

Hasta hace relativamente poco, la mayor parte de las aplicaciones de rotación de Procustes se realizaban en el campo del análisis sensorial (Dijksterhuis, 1996; *www.mapp.asb.dk*, 2003) y alimentario para estudiar relaciones entre paneles de sensores o la inter-relación de catadores/degustadores. Arnold (Arnold y Williams, 1986) recopila el uso de esta técnica al análisis sensorial y/o de cata (Arnold y Collins, 1993).

Scarponi (Scarponi et al., 1990) realiza una aplicación centrada esencialmente en la selección de un subconjunto de variables que le permita diferenciar entre distintos tipos de vinos italianos (Chianti, Brunello y Mazemino el Trentino). Boschelle et al. (1994) estudiaron la respuesta de diferentes cultivos de oliva en el área climática del Golfo de Trieste.

El control de calidad industrial constituye un campo de aplicación de esta técnica. Deane y Macfie (1989) realizan una selección de variables de queroseno con fines de control de calidad aunque el principal inconveniente es que no emplean todos los test de las especificaciones. También en queroseno, aunque empleando los 26 parámetros de calidad medidos rutinariamente en las especificaciones de la OTAN, Andrade (Andrade et al., 1993, Andrade et al., 1997) presenta trabajos en los que realizan estudios iniciales de selección de variables y donde se plantea el problema de la validación y uso final de este tipo de estudios.

En el campo medioambiental, Andrade et al. (1994) presentaron un estudio piloto en el cual se seleccionaban parámetros de aguas de pozos y acuíferos de la Comunidad Valenciana y se estudiaba si la importancia de las variables seleccionadas se mantenía a lo largo de diversos muestreos. Carlosena et al. (1995) realizan un estudio en el cual se trata de determinar las variables que definen las pautas de distribución de suelos de la provincia de A Coruña sometidos a tráfico intenso. También se realiza un seguimiento a lo largo de un año que revela que dos variables son las fundamentales para definir y diferenciar hasta cuatro subgrupos de suelos.

Kubista ha aplicado el concepto de rotación de Procustes a la espectroscopia

con objeto de resolver mezclas monómero-dímero o monómero-dímero-trímero empleando relaciones físicoquímicas y espectroscopia de fluorescencia. El gran avance en este campo es que no precisa de una calibración directa (Kubista, 1990; Kubista et al., 1992; Kubista et al., 1993; Scarminio y Kubista, 1993; Nygren et al., 1995; Kubista et al., 1999).

Vigneau et al. (1995) emplean la idea de Rotación de Procustes para comparar-correlacionar-reconstruir espectros en el infrarrojo medio mediante datos de infrarrojo cercano. La idea es francamente interesante y muy prometedora y tiene grandes posibilidades en cuanto a su aplicación. Anderson y Kalivas (1999) emplearon también esta técnica para realizar comparaciones entre espectros.

Se han encontrado estudios relacionados con transporte aéreo de contaminantes. Se observó que hay dos posibilidades: (i) seleccionar variables para describir los grupos de muestras sin perder demasiada información (King y Jackson, 1999) y/o (ii) comparar el espacio de datos "experimental" con algún espacio objetivo (por ejemplo, una lista de fuentes de polución (Richman y Vermette, 1993).

La descripción de moléculas orgánicas no es una aplicación analítica habitual en la rotación Procustes pero se han encontrado dos aplicaciones interesantes (Robert y Carbo-Dorca, 1998; Tomas et al., 1999).

Se ha aplicado también esta técnica como apoyo de estudios instrumentales complejos. Por ejemplo, para determinar el número de masas significativas en GC-MS (Demir et al., 1996); comparar la información obtenida por dos detectores en un diodo array acoplado triplemente con un sistema de espectrometría de masas electroespray cromatografía de líquidos (Bessant et al., 1999) reducir el espectro de masas de 2- y 3-hidroxipiridinas a 20 masas significativas con los dos detectores (Dunkerley et al., 1998) y operar con los datos cuando se trabaja con muestras multicomponente (Schulze y Stilbs, 1993).

Antonov et al. (1999) comparaban el comportamiento de esta técnica con otra de resolución simultánea de solapamiento de bandas y encontraban un funcionamiento similar cuando se estudiaba el equilibrio monómero-dímero.

Poco a poco se han llevado a cabo estudios más profundos empleando la técnica de Procustes. A título de ejemplo se recogen algunos de ellos.

Books y Kowalski (1994) demostraban que la rotación de Procustes eran equivalentes a aplicar una técnica denominada "Rank Anihilation Factor Analysis"

para el análisis de espectros correlacionados, y presentaban otros progresos de Lorber, Sanchez y Kowalski, donde se sugirió que más que dos subespacios relacionados pueden ser comparados aplicando descomposición trilineal.

Aunque en los estudios presentados en esta Memoria se mantienen los ejes ortogonales, se ha demostrado que también se pueden llevar a cabo rotaciones oblicuas donde los ejes no se mantienen perpendiculares (Reyment y Jöreskog, 1993).

González-Arjona et al. (1999) desarrollaban un método de análisis discriminante basado en Procustes. Se comparaban tres técnicas: PLS discriminante, modelo de clase discriminante lineal y redes de neuronas artificiales (*backpropagation*).

Vinzi (2001) ha presentado un estudio donde comparaba Procustes con diferentes métodos para comparar matrices múltiples estructuradas.

Guo et al. (2002) presentaba un método basado en rotación de Procustes para seleccionar un subgrupo de variables en PCA. Un sistema similar fue aplicado por Guo et al. (2001) en combinación con *projection pursuit*.

Heberger y Andrade (2004) comparaban el método de selección de variables por Rotación de Procustes con un método no paramétrico basado en el estudio de variables apareadas (*generalized Pair-Wise correlation procedure*, GPCM) observando que se alcanzaban resultados similares.

3. ALGORITMOS GENÉTICOS

3.1. INTRODUCCIÓN

Para abordar el estudio de esta herramienta es conveniente hacer una breve disgresión inicial acerca de la idea de “evolución”. Para entender el modelo evolutivo natural es preciso reconocer que la naturaleza es un ingeniero consumado. Los seres vivos representan la solución al problema de la supervivencia de un organismo que se enfrenta a condiciones determinadas por el entorno en el que se desarrolla. El más conocido pionero y defensor de este punto de vista fue Darwin quien observó que las especies evolucionan poco a poco a lo largo de intervalos de tiempo que, en la escala humana, resultan enormes, y que las especies actuales son el fruto de innumerables cambios paulatinos de otras que las antecedieron. Además, lo que dirige los cambios es la pertinencia que estos tienen para favorecer la subsistencia de los individuos de la especie en función de su medio ambiente. A mediados de este siglo se empezaron a comprender las razones por las que así ocurre. Watson y Crick explicaron el mecanismo de la herencia identificando la molécula de ADN como la portadora de la información (a nivel celular) que determina las características del individuo. El ADN codifica al individuo (el genotipo) usando un alfabeto químico muy simple y, al ser interpretado por el organismo que lo gesta, da origen al individuo mismo (el fenotipo). El ADN, pues, contiene la información de la mejor solución (el individuo) al problema fundamental del ser vivo (la supervivencia). La teoría de la evolución que considera esta forma de transmisión de la información de una generación a otra se denomina neodarwinismo (<http://cursos.itam.mx/akuri/Semestre2/AGS/19RNSyAGS.pdf>, 2003).

La información contenida en el ADN de los seres vivos es información intrínseca que se deriva de un proceso de aprendizaje generacional. La idea de la Computación Evolutiva, y de los Algoritmos Genéticos (AAGG) en particular, es simular parcialmente el proceso anterior para lograr que un programa de software “aprenda” y sea capaz de “dar respuesta” a un problema planteado por el Químico Analítico. Es decir, se trata de buscar la mejor solución posible (el mejor ADN) ante el problema a resolver (las condiciones de entorno).

Los Algoritmos Genéticos son una herramienta matemática que se utiliza para resolver problemas asociados a la optimización y a la búsqueda de soluciones óptimas en problemas complejos. Estos algoritmos basan su idea conceptual en la Teoría de la Evolución de los seres vivos y han alcanzado gran importancia dentro de la comunidad científica a lo largo de los últimos años.

El algoritmo genético (AG) se basa en los mecanismos de selección que utiliza la naturaleza. Según estos, los individuos más aptos de una población son los que sobreviven al adaptarse más fácilmente a los cambios que se producen en su entorno. Hoy en día se sabe que estos cambios se efectúan en los genes del individuo (unidad básica de codificación de cada uno de los atributos de un ser vivo) y que los atributos más deseables del mismo (i.e., los que permiten a un individuo adaptarse mejor a su entorno) se transmiten a sus descendientes cuando éste se reproduce sexualmente.

Un investigador de la Universidad de Michigan, llamado John Holland, era un estudioso de la selección natural y, a fines de los años 60, desarrolló una técnica que permitió incorporar esa idea en un programa de software (Galaviz-Casas, 1998; www.redcientifica.com, 2003; www.devdept.com, 2003; www.ub.rug.nl, 2003). El propósito original de Holland no era diseñar algoritmos para resolver problemas concretos sino estudiar formalmente el fenómeno de adaptación y desarrollar vías para extrapolar esos mecanismos de adaptación natural a los sistemas computacionales, es decir, lograr que las computadoras aprendieran por sí mismas (www.uv.es, 2003). A la técnica que inventó Holland se le llamó originalmente “planes reproductivos”, pero se hizo popular bajo el nombre “algoritmo genético” tras su publicación en 1975 (Forrest y Mitchell, 1993; Holland, 1975; www.cs.us.es, 2003; <http://members.tripod.com>, 2003; www.galeon.com, 2004).

Una definición bastante completa de un algoritmo genético es la propuesta por John Koza (1992): “Es un algoritmo matemático (...) que transforma un conjunto de objetos matemáticos individuales a lo largo del tiempo usando operaciones modeladas de acuerdo al principio Darwiniano de reproducción y supervivencia del más apto y tras haber pasado una serie de operaciones genéticas naturales de entre las que destaca la recombinación sexual. Cada uno de estos objetos matemáticos suele ser una cadena de caracteres (letras o números) de longitud fija que se ajusta al modelo de las cadenas de cromosomas y se les asocia con una cierta función matemática que refleja su aptitud” (www.cs.us.es, 2003; <http://members.tripod.com>, 2003; www.galeon.com, 2004).

El enorme desarrollo en la capacidad de cálculo de las computadoras hace plausible abordar ciertos problemas usando técnicas que hace pocos años se hubiesen desechado por ineficientes ya que requieren muchos recursos de cómputo o mucho tiempo. Dentro de estas técnicas se encuentran primordialmente aquellas que, de alguna manera, pretenden emular fenómenos naturales de gran complejidad (por ejemplo, AAGG, Redes de Neuronas Artificiales, etc) (Leardi, 2001; <http://cursos.itam.mx>,

2003).

3.2. FUNDAMENTOS

En los Algoritmos Genéticos se parte de un conjunto de posibles soluciones para un problema dado (denominado “población inicial”). A cada una de las posibles soluciones se les llama “individuos” (Pazos Sierra, 1996; www.orcero.org, 2003). Por medio de una serie de operaciones matemáticas (llamadas Operadores Genéticos) los individuos de la población inicial van evolucionando en cada generación hacia poblaciones mejores. De esta forma, en un instante dado de la evolución, podrá existir un individuo de la población que sea la solución “óptima” al problema dado (www.mundo-electronico.com, 2003; Pazos Sierra, 1996). En la **Figura 4** se muestra el ciclo típico de funcionamiento de un Algoritmo Genético.

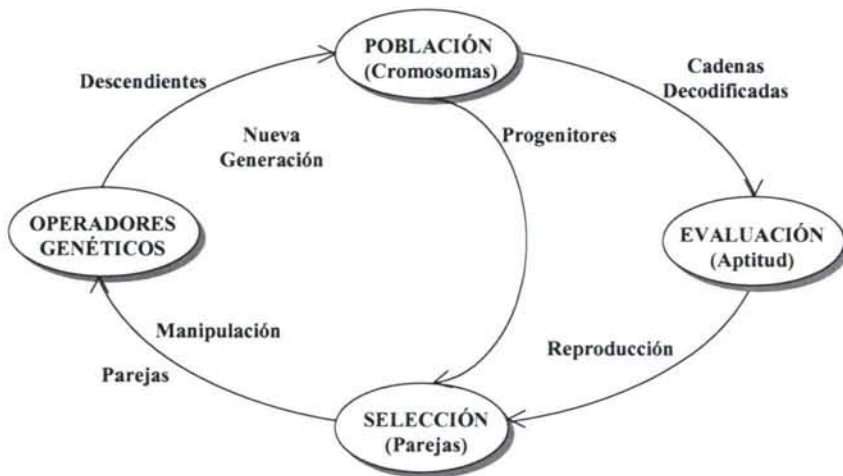


Figura 4: Ciclo de un AG (www.mundonomades.com, 2003).

Los Operadores Genéticos que se aplican sobre las poblaciones para que éstas vayan evolucionando a lo largo del tiempo son variados y reciben muchas denominaciones (no siempre coincidentes): reproducción, mutación, vecindad, creación de nichos y especies, diferenciación sexual, dominación, etc. Para simplificar la explicación de los Algoritmos Genéticos se van a abordar solamente la reproducción, el cruce y la mutación por ser los tres operadores genéticos básicos (www.mundo-electronico.com, 2003).

3.3. LOS INDIVIDUOS

La representación de cada individuo dentro de una población se realiza por medio de una cadena de bits (=genes) de longitud fija. Los individuos pueden estar formados por uno o varios genes. A continuación se explican algunos conceptos:

- 1.- Población: es un conjunto de individuos.
- 2.- Individuo: es un “ser” caracterizado por un cromosoma, el cual es el código de información sobre el cual opera el algoritmo.
- 3.- Genoma: son todos los parámetros, o genes diferentes, que caracterizan a todos lo individuos de una población.
- 4.- Genotipo: es la información genética particular que describe a un individuo específico, es decir, los genes que posee.
- 5.- Gen: es cada unidad estructural que forma el cromosoma. Por ello los genes que formen un individuo establecerán su comportamiento. El número de genes dentro de un individuo es variable y depende de la aplicación pero dentro de cada aplicación todos los individuos tienen el mismo número de genes. Aunque cada gen puede constar de uno o más bits consecutivos (**Figura 5**), lo habitual es que sólo esté formado por un bit (*Pazos Sierra, 1996*).
- 6.- Alelo: es cada uno de los bits que forman un gen (**Figura 5**); generalmente se trabaja con un alelo/bit y toma valores de 0 ó 1.

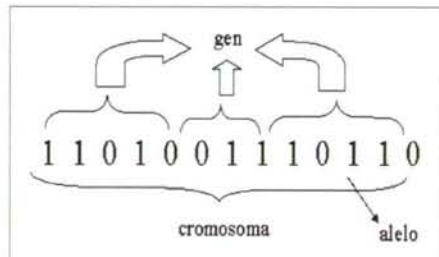


Figura 5: Esquema de definición de cromosoma, gen y alelo.

- 7.- Fenotipo: es la expresión del genotipo, es decir, una cadena de números binarios.
- 8.- Función de idoneidad (*Fitness*, función de adaptación, función de ajuste

o función de evaluación): la única restricción para usar un programa basado en AAGG es que exista una función que le informe de cuán bueno es un individuo en la solución de un problema. Esta función *fitness* o de evaluación (idoneidad) es el principal enlace entre el AG y el problema real. En un caso ideal, la función *fitness* debería ser similar, sino igual, a la función objetivo que se quiere optimizar. El resultado de aplicar la función *fitness* (p.ej. una función como $y=a+bx$) determina la supervivencia del individuo y se utiliza como parámetro de los operadores y guía la obtención de nuevas poblaciones (<http://members.tripod.com>, 2003).

3.4. ESTRUCTURA DE LOS AAGG

En un AG se parte de una población de m posibles soluciones a un problema dado (m individuos). Sobre esta población inicial (P_0) se realizan diferentes procesos. Podemos caracterizar los AAGG a través de las siguientes etapas (www.ica.ele.puc-rio.br, 2003; Pazos Sierra, 1996):

- 1- Problema a ser optimizado.
- 2- Representación de Soluciones del Problema (codificación).
- 3- Inicialización.
- 4- Evaluación.
- 5- Operadores Genéticos.
- 6.- Parámetros y criterios de parada.

1.- Problema

Los AAGG son particularmente útiles en problemas complejos de optimización: problemas con muchos parámetros o características que precisan ser combinadas en busca de la mejor solución; problemas con muchas restricciones o condiciones que no pueden ser representadas matemáticamente y problemas con grandes espacios de búsqueda. En el apartado 3.7. de este capítulo se recogen brevemente algunos problemas que han sido abordados mediante AAGG. El problema deberá plantearse, pues, como una secuencia de operaciones en la que -al menos- una etapa consistirá en un proceso de optimización (o búsqueda) mediante los AAGG. De

hecho, los algoritmos genéticos constituyen una herramienta muy potente para combinar explotación con exploración de posibles soluciones.

2.- Representación o codificación

La representación de las posibles soluciones dentro del espacio de búsqueda de un problema define la estructura del cromosoma que va a ser manipulado por el algoritmo. Normalmente se emplean la representación binaria por ser más fácil de manipular a través de los operadores genéticos y ser fácil de transformar en un número entero o real.

3.- Inicialización de la Población

La inicialización de la población consiste en la creación de los individuos que constituyen la población inicial (P_0). Normalmente, la población inicial se forma a partir de m individuos creados aleatoriamente. Las poblaciones iniciales creadas aleatoriamente pueden ser “sembradas” con cromosomas altamente adecuados para conseguir una evolución más rápida, si se dispone de información *a priori*.

4.- Evaluación (idoneidad, ajuste o *fitness*)

La idoneidad establece la relación entre el AG y el mundo externo. La evaluación se realiza a través de una función que representa de forma adecuada el problema y tiene como objetivo suministrar una medida de la aptitud de cada individuo de la población actual.

5.- Operadores genéticos (Parsons et al., 1995; Hong et al., 2000; Hong et al., 2002):

a) Selección: En este proceso se eligen n individuos de la población inicial P_0 que van a pasar a la siguiente generación, generalmente $n \leq m$. Estos n individuos forman una población intermedia P_i . De estos n individuos se seleccionan p individuos ($p \leq n$) que serán utilizados como progenitores y utilizarán la reproducción para generar descendientes.

b) Cruce o reproducción (*crossover*): En este proceso los p individuos de la población intermedia P_i se juntan en parejas y generan q descendientes. Ahora la población intermedia P_i está formado por $n+q$ individuos.

En este proceso, de los $n+q$ individuos de la población intermedia P_i , se eligen n individuos, generalmente los más aptos una vez evaluados por la función

fitness, que son los que formarán la siguiente población.

c) Mutación: en este proceso, se altera el contenido de una posición del cromosoma, según una determinada probabilidad (en general baja); esta posibilidad es lo que confiere al algoritmo su capacidad de exploración de nuevas regiones de eventuales soluciones.

6.- Parámetros y Criterios de Parada

Varios parámetros controlan la evolución (Kalogirou, 2003):

a.- Tamaño de la Población: número de individuos del espacio de búsqueda que son considerados en paralelo.

b.- Tasa de *Crossover* (cruce): probabilidad de un individuo de ser recombinado con otro.

c.- Tasa de Mutación: probabilidad de que el contenido de cada posición/gen del cromosoma sea alterado.

d.- Número de Generaciones: número total de ciclos de evolución de un AG.

e.- Total de Individuos: número total de tentativas (tamaño de la población x número de generaciones).

Los dos últimos parámetros son empleados generalmente como criterio de parada de un AG.

Por tanto, un AG puede ser descrito, de forma sencilla, como un proceso continuo que repite ciclos de evolución controlados por un criterio de parada, como se muestra en la **Figura 6**. Algunos detalles recogidos en la misma se explican a continuación.

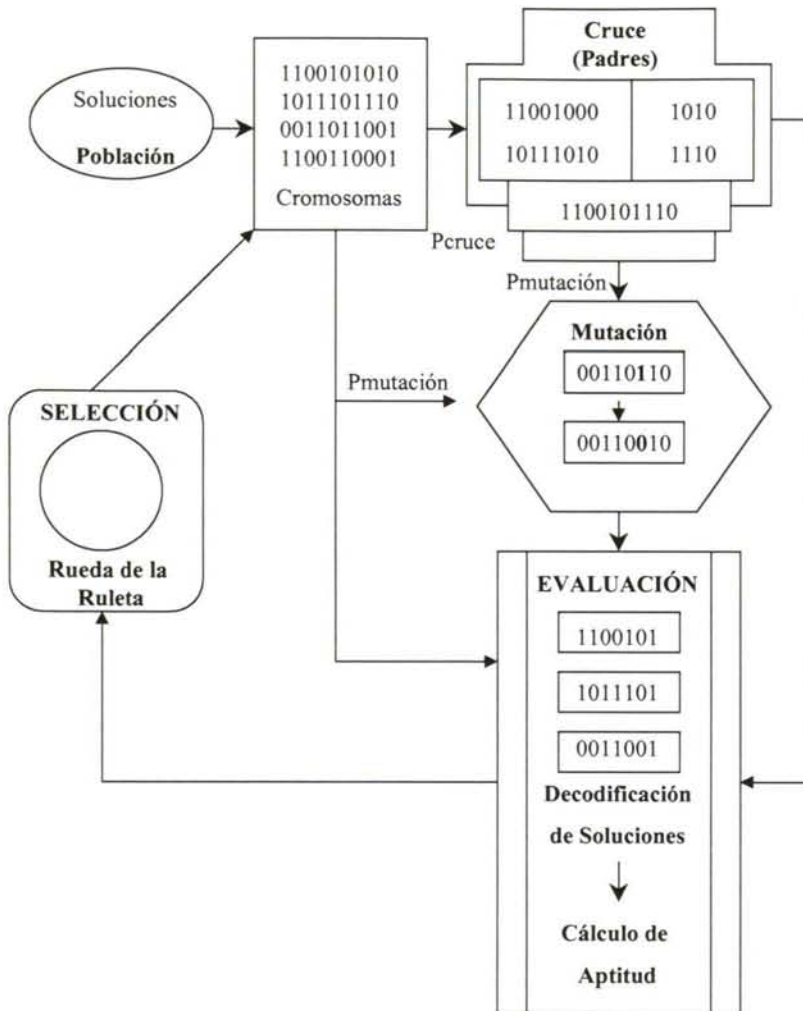


Figura 6: Estructura general de un algoritmo genético, utilizando los operadores genéticos más comunes (www.asee.org, 2003). Siendo P_{cruce} : probabilidad de cruce y $P_{\text{mutación}}$: probabilidad de mutación.

3.5. LOS OPERADORES GENÉTICOS

En esta Memoria se van a describir los operadores Selección, Reproducción (Cruce), Mutación y reinicialización de la población ya que son los operadores básicos de todo algoritmo genético (Pazos Sierra, 1996; Libelli et Alba, 2000; Shapiro, 2001). Estos operadores se utilizan para hacer evolucionar a los individuos de tal forma que vayan sobreviviendo los mejores. En las referencias indicadas se encuentran más detalles (www.redcientifica.com, 2003; www.cs.us.es, 2004; Estévez Valencia, 1997; <http://members.tripod.com>, 2003; <http://cursos.itam.mx>, 2003; www.des.ime.br, 2003).

3.5.1. SELECCIÓN

Para que un AG progrese debe producirse una evolución en las poblaciones de individuos. Si se piensa en la evolución de las especies, es necesario elegir unos progenitores. El operador de selección se encarga de escoger los individuos genéticos de la población sobre los que se aplicarán las operaciones de reproducción y mutación. Existen varios métodos de selección:

a) El **método de la ruleta** (De Jong, 1985; www.cs.us.es, 2004; www.ub.rug.nl, 2003; <http://benli.bcc.bilkent.edu.tr>, 2003): selecciona un individuo frente a otros en función de una probabilidad. Equivale a asignar a un individuo un determinado número de posiciones de una ruleta de acuerdo a su idoneidad. Cuanto mejor sea la solución proporcionada por ese individuo más posiciones se le asignarán (**Figura 7**) y, por lo tanto, más posibilidades tendrá de ser seleccionado por el azar (sorteo). Después se hace girar la ruleta (lo que a nivel práctico se lleva a cabo con un generador de números aleatorio) y se selecciona el individuo sobre el que se detiene. De esta manera se establece una presión de selección mayor sobre los mejores individuos; a los peores se les asigna un menor número de posiciones de la ruleta por lo que tienen menos posibilidades de ser escogidos .

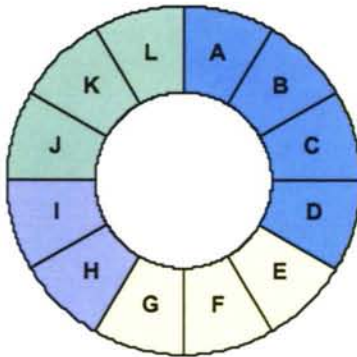


Figura 7: Método de la ruleta para seleccionar un progenitor en los AAGG.

b) El **método del Torneo** (Shapiro, 2001; Glibovets y Medvid, 2003; www.cs.us.es, 2004): selecciona al azar una serie de individuos de la población, generalmente 2. Si el torneo es determinístico se selecciona siempre el mejor de los individuos y si es probabilístico unas veces se selecciona el mejor individuo y otras uno peor. De esta manera se disminuye la presión de selección y los individuos con peor ajuste tienen posibilidades de ser

seleccionados.

c) **Selección aleatoria:** todos los individuos de la población tienen las mismas posibilidades de ser escogidos, independientemente de si representan una solución buena o mala.

3.5.2. CRUCE O REPRODUCCIÓN

El operador de cruce se encarga de obtener la descendencia combinando los genes de los padres. De esta manera si uno de los descendientes recibe los genes causantes de la bondad de los dos padres este nuevo individuo representará una mejor solución que cada uno de los padres por separado. Hay varias opciones de trabajo (Belfiore y Esposito, 1998; Glibovets y Medvid, 2003; Tian, 2001):

a) En el “cruce de un punto” se selecciona un punto al azar del cromosoma, de manera que éste se divide en dos segmentos (Figura 8). Para obtener la descendencia se intercambian los segmentos más pequeños de los padres. De esta manera los dos hijos generados heredan información genética de ambos padres (Richards et al., 2002).

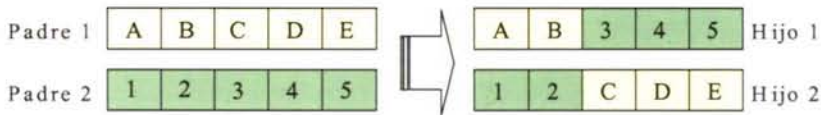


Figura 8: Uso de un sólo punto de cruce entre 2 individuos. Cada pareja de cromosomas da origen a 2 descendientes para la siguiente generación. Si el punto de cruce es cualquiera de los 2 extremos de la cadena, no hay cruce.

b) En el “cruce de dos puntos” se corta el cromosoma por 2 puntos, por lo que se crean 3 segmentos (Figura 9). Para obtener los hijos se intercambian los segmentos centrales de los padres. Uniendo el segmento final del cromosoma con el inicial puede observarse la ventaja de este tipo de cruce con respecto al anterior (Richards et al., 2002).

Utilizar un único punto de cruce equivaldría a poder variar únicamente el punto final del segmento que se intercambia, es decir, podría escogerse (en función del azar) un segmento máximo hasta el gen D (Figura 8). Sin embargo el punto de inicio del primer segmento permanecería siempre fijo en el primer gen (llamado A en la Figura 8). Trabajando con dos puntos de

cruce también se puede variar el inicio de este segmento y hacer que empiece, por ejemplo, a partir del gen etiquetado con la letra B (Figura 9). En principio se pueden añadir tantos puntos de cruce como se desee pero existen estudios que demuestran que trabajar con más de dos puntos de cruce reduce el rendimiento del Algoritmo Genético.

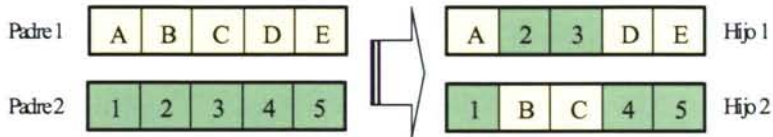


Figura 9: Uso de dos puntos de cruce entre dos individuos. En este caso se mantienen los genes de los extremos, y se intercambian los del centro. Si uno o ambos puntos de cruce se encuentran en los extremos de la cadena, se hará un cruce usando un sólo punto, o ningún cruce, según corresponda.

c) En el “cruce uniforme”, cada uno de los cromosomas de los hijos tiene las mismas posibilidades de proceder de uno u otro padre. Se genera una plantilla con tantas posiciones como genes tengan los individuos y se completa con 0 y 1 de manera aleatoria. Si en una posición hay un 1 se pasa al hijo el situado en esa posición en uno de los padres y si tiene un 0 el del otro. Para generar el segundo hijo se intercambia la posición de los padres.

3.5.3. MUTACIÓN

Con la mutación se introduce en alguno de los genes de la población valores nuevos que no se podrían obtener aplicando únicamente operaciones de cruce. De esta manera se asegura que no queda ninguna región del espacio de búsqueda sin explorar.

Las mutaciones más usuales consisten en variar de manera aleatoria el valor de alguno de los genes o bien de intercambiar los valores de un par de genes (Feam y Davies, 1997).

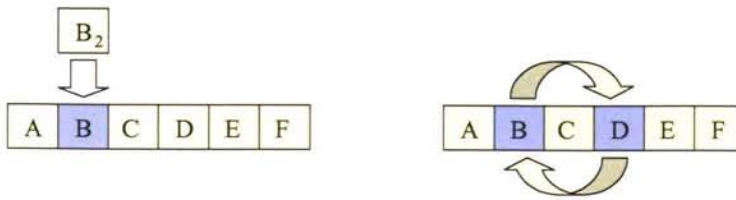


Figura 10: Mutaciones posibles en AAGG.

3.5.4. REINICIALIZACIÓN DE LA POBLACIÓN

Una vez aplicados los operadores genéticos debe insertarse la descendencia en la población. Si el Algoritmo Genético trabaja con una población temporal se inserta directamente en ésta y cuando ésta se llena se pasan todos sus individuos a la población del Algoritmo Genético, reemplazando a los existentes hasta ese momento. A continuación la población temporal se vacía y comienza de nuevo el proceso. Por el contrario, si se prescinde de esta población temporal las inserciones deben hacerse directamente sobre la población. Pero como debe mantenerse siempre el mismo número de individuos, antes de insertar la descendencia se ha de hacer sitio.

Para escoger los individuos que se eliminarán existen varias opciones (Pazos Sierra, 1996):

- a) La más sencilla es escogerlos al azar pero se corre el riesgo de estar eliminando, precisamente, aquellos individuos que representan las mejores soluciones.
- b) Otra opción sería eliminar de la población a los padres para dejar sitio a los hijos.
- c) Como, por lo general, la población está ordenada con respecto a la bondad de los individuos (*fitness*) se pueden eliminar algunas de las peores soluciones de la población.
- d) También se puede insertar la descendencia reemplazando para ello alguno de los individuos que presenten un ajuste similar al suyo.

3.6. VENTAJAS Y DESVENTAJAS DEL USO DE LOS AAGG

A continuación se resumen algunas de las diferentes ventajas y desventajas más mencionadas en lo que se refiere al uso de los AAGG con respecto a otras técnicas de búsqueda (*www.cs.us.es, 2003; www.redcientifica.com, 2003*). Como ventajas se cita que:

- No necesitan conocimientos específicos sobre el problema que intentan resolver.
- Operan de forma simultánea con varias soluciones, en vez de trabajar de forma secuencial, como las técnicas tradicionales.
- Cuando se usan para problemas de optimización (maximizar una función objetivo) resultan menos afectados por los máximos locales (falsas soluciones) que las técnicas tradicionales. Aportan múltiples soluciones para el mismo problema y podrá elegirse la más ventajosa en función de diversos requisitos que puedan establecerse.
- Resulta sumamente fácil ejecutarlos en las modernas arquitecturas de computadoras paralelas.
- Usan operadores probabilísticos, en vez de los típicos operadores determinísticos de las otras técnicas.

Los problemas habituales son que:

- Pueden tardar mucho en converger, o no converger en absoluto, dependiendo en cierta medida de los parámetros que se utilicen (tamaño de la población, número de generaciones, etc...).
- Pueden converger prematuramente y no llegar a la zona del óptimo (*www.ub.rug.nl, 2003*).

3.7. APLICACIONES

El uso de AAGG ha crecido exponencialmente desde que Holland (1975) publicó el primer artículo. Gracias al extraordinario aumento en el poder de cálculo, hoy en día es posible aplicarlos a problemas muy complejos. Ha sido publicado un considerable número de artículos en los que se aplicaron los AAGG. En particular, a continuación se recogen algunos trabajos en Química.

Guo *et al.* (2001, 2002) comparaban la ejecución de diferentes métodos, incluyendo una variante de rotación Procrustes acoplado a AAGG para seleccionar variables.

Ramadan *et al.* (2001) usaban los AAGG como método de selección de variables para la clasificación de muestras ambientales de suelo.

Bangalore *et al.* (1996) empleaban PLS en la determinación de diferentes compuestos orgánicos haciendo una previa selección de variables aplicando AAGG.

Kemsley (1998) empleaba los AAGG para el cálculo de variables canónicas.

Meusinger y Moros (1999) determinan la relación cuantitativa de estructuras moleculares de las gasolinas con el número de octano por AAGG.

Shao *et al.* (2000) emplean AAGG y un algoritmo “inmune” para la resolución del solapamiento de cromatogramas.

Vivó-Truyols *et al.* (2001a, 2001b) usaban un AG híbrido para un problema de optimización combinatoria.

Tsukimoto y Hatano (2003) emplean AAGG para la focalización funcional de una red neuronal.

Arifovic y Gençay (2001) usan un AG para seleccionar la arquitectura de una RNA feedforward. Algo que Sánchez también realizó al incluir un AG como motor de búsqueda de la mejor RN (Sánchez Pastor, 1997).

Lavine *et al.* (2001) desarrollan un AG para el análisis de reconocimiento de pautas de datos químicos multivariados.

Fatemi *et al.* (2003) predice el factor de bioconcentración combinando AAGG con una RN.

Da Costa Filho y Poppi (2002) determinaban azúcares mediante espectroscopia MIR (fructosa, glucosa y maltosa), haciendo una selección previa de variables empleando AAGG.

Pavan *et al.* (2003) presentaban una combinación novedosa de AAGG y diagramas Hasse para seleccionar variables.

3.8. DESCRIPCIÓN DEL PROBLEMA ABORDADO EN ESTA MEMORIA

A continuación se resume cómo se abordó en esta Memoria el problema de la selección de variables espectrales en la región infrarroja media mediante el uso de AAGG. El objetivo principal es determinar las variables (números de onda) que contienen la máxima información posible para desarrollar modelos de clasificación y predicción de nuevas muestras.

Se investigaron dos AAGG (búsqueda por etapas y búsqueda fija) para reducir el número de variables espectrales que, posteriormente, serán empleadas bien para entrenar una RN bien para obtener modelos de clasificación. La idea clave es, pues, encontrar un subconjunto pequeño de variables capaz de clasificar muestras con resultados comparables al uso de todas las variables. También se estudia la utilidad de las variables seleccionadas para desarrollar modelos de calibración con fines predictivos. En ambos casos, la aplicación práctica se llevará a cabo empleando espectros MIR de bebidas basadas en zumo de manzana.

En cualquiera de las dos aproximaciones (búsqueda fija o por etapas), el AG debe disponer de una función de idoneidad adecuada. La configuración básica del AG se puede resumir como un AG de “estado estable” (*steady state*) con un tamaño total de población de 100 individuos. *Steady state* denota un AG con sólo una población donde 2 padres son seleccionados y las dos descendencias son incluidas en la población (Goldberg, 1989) (dos individuos son rechazados para mantener la población total constante). Esto contrasta con el AG con dos poblaciones (la población de padres y una nueva población donde la descendencia está siendo almacenada); después se completa el proceso aplicando el operador de mutación.

Para llevar a cabo la selección de los dos padres, uno es seleccionado de acuerdo con su función *fitness* por medio del método de la ruleta, el otro se elige aleatoriamente.

Las probabilidades de cruce y mutación fueron de 95% y 5%, respectivamente. Esto significa que, para cada nueva generación, el 95% de los individuos provienen del cruce sexual y el 5% de las mutaciones (sólo un gen puede mutar por cada individuo seleccionado para mutación). Se utiliza un cruce uniforme, lo que significa que cada gen de la descendencia es seleccionado aleatoriamente a partir de los genes correspondientes de los dos padres, como se indicó en el apartado 3.5.2. Adicionalmente, también se han considerado algoritmos de cruce de uno y dos puntos

pero fueron descartados porque requerían un tiempo de convergencia demasiado alto. Los individuos generados a partir del cruce reemplazan a dos miembros de la población genética que tengan *fitness* similar, en un intento de mantener la diversidad de la población.

3.8.1. ESTRATEGIA DE BÚSQUEDA POR ETAPAS

La primera modalidad de AAGG empleado (denominada “**búsqueda por etapas**”) considera inicialmente todas las variables y en cada etapa (iteración) va descartando, gradualmente, grupos de ellas. El proceso continúa mientras que los resultados de clasificación (predicción) sean iguales (o similares) a los obtenidos empleando todas las variables. Así, el AG determina cuántos y qué números de onda se considerarán para la clasificación (predicción).

3.8.1.1. CODIFICACIÓN DEL PROBLEMA

El primer paso para usar el AG es codificar el problema en una serie de valores (o individuos). Por lo tanto, cada individuo en la población inicial tiene 176 genes, uno para representar cada número de onda. Cada gen puede llevar a un valor, 1 ó 0, el cual indica si la absorbancia asociada a esta variable es considerada (1) o no (0). Cada individuo en la población inicial es creado dando valores aleatorios (1 o 0) para cada gen. De esta forma ya se crean candidatos a distintas soluciones, algunas de ellas mostrando menos variables que otras.

3.8.1.2. EVALUACIÓN DE LOS INDIVIDUOS GENÉTICOS

Cada uno de los individuos (posibles soluciones) se evalúa mediante el uso de una RN clasificatoria donde cada clase se había codificado de forma binaria; p.ej., 1000000 para la clase del 2%, 0100000 para la del 4%, etc. Los resultados de clasificación se comparan con el modelo original (considerando las 176 variables). Mientras los resultados de clasificación sean iguales, o al menos comparables, el AG es aceptado. Esta opción requiere mucho tiempo de entrenamiento por lo que se fijó el número de iteraciones del entrenamiento. Es decir, las RRNN no se entrenan hasta sus óptimos absolutos sino, únicamente, en la misma extensión. Esto resulta suficiente para evaluar su bondad (Gestal et al., 2004).

Además del error (o éxito) que proporciona la RN en su entrenamiento reducido, y que podría pensarse que es un criterio suficiente para elegir el mejor individuo, lo cierto es que éste determina sólo una parte de la calidad de dicho individuo (solución). La otra parte debería reflejar la cantidad de números de onda

considerados para la clasificación, por ejemplo, el número de genes con valor "1" en el cromosoma. Por tanto, el error de evaluación de la red de neuronas artificiales deberá "combinarse" con el número de "unos" contenido en el genotipo de un individuo, y, por tanto, favorecer los individuos con una cantidad más pequeña de números de onda. De esta manera:

$$\text{fitness}(k) = \text{Error cuadrático medio de la RN} \times \text{Unos en genotipo de } k / n^0 \text{ de variables}$$

Esta función de evaluación automáticamente favorece los individuos que emplean menos información (menos variables) para obtener un determinado error durante la clasificación de la muestra. Así, el AG se desarrolla hacia individuos que reducen gradualmente el número de variables requerido para determinar el porcentaje de zumo de manzana en la muestra.

3.8.2. ESTRATEGIA DE BÚSQUEDA FIJA

La segunda modalidad de AAGG (denominada "búsqueda fija") no sigue un procedimiento secuencial. El número final de variables se fija previamente (de acuerdo con algunos criterios externos, previo cálculo, criterios químicos u otros) y, así, el AG tendrá que buscar el subgrupo de variables que se considerará finalmente. Para simplificar la comparación de resultados, el número de variables requeridas en esta Memoria será en todos los casos igual al mínimo número de componentes principales que pueden describir nuestro grupo de datos, según el criterio W_m de Krzanowski.

3.8.2.1. CODIFICACIÓN DEL PROBLEMA

A diferencia de la aproximación previa, se pueden emplear aquí números reales ordinarios para caracterizar los individuos. Cada gen contiene un número que denota el número de onda a ser considerado. La población genética individual consistirá en 100 individuos con "n" genes, siendo "n" la cantidad de números de onda que son considerados a priori para la solución.

3.8.2.2. EVALUACIÓN DE LOS INDIVIDUOS GENÉTICOS

Como la cantidad total de números de onda a considerar permanece constante durante todo el proceso de selección, este sistema usa el error cuadrático medio de la evaluación de la RN como la función de evaluación.

4. BIBLIOGRAFIA

Andrade, J.M.; Prada, D.; Muniategui, S.; Gómez, B.; Pan, M. Multivariate selection of variables in industrial quality control: optimizing aviation fuel final control. *J. Chemometrics* 7, 427-438 (1993).

Andrade, J.M.; Prada, D.; Muniategui, S.; Alonso, E.; López, P.; De la Fuente, P.; Quijano, M.A. Selection of analytical variables to optimize laboratory efforts in future groundwater studies. *Anal. Chim. Acta* 292, 253-261 (1994).

Andrade, J.M.; Muniategui, S.; Lopez-Mahia, P.; Prada, D. Use of multivariate techniques in quality control of kerosene productions. *Fuel* 76(1), 51-59 (1997).

Andrade, J.M.; Gómez-Carracedo, M.P.; Fernández, E.; Elbergali, A.; Kubista, M.; Prada, D. Classification of commercial apple beverages using a minimum set of mid-IR wavenumbers selected by Procrustes rotation. *Analyst* 128, 1193-1199 (2003).

Anderson, C.E.; Kalivas, J.H. Fundamentals of Calibration Transfer through Procrustes Analysis. *Appl. Spectrosc.* 53(10), 1268-1276 (1999).

Antonov, L.; Gergov, G.; Petrov, V.; Kubista, M.; Nygren, J. UV-Vis spectroscopic and chemometric study on the aggregation of ionic dyes in water. *Talanta* 49, 99-106 (1999).

Arifovic, J.; Gençay, R. Using genetic algorithms to select architecture of a feedforward artificial neural network. *Physica A* 289, 574-594 (2001).

Arnold, G.M.; Williams, A.A., The use of Generalised Procrustes Techniques in Sensory Analysis, en, *Statistical Procedures in Food Research*. Ed. J.R. Piggott, Elsevier (1986).

Arnold, G.M., Scaling Factors in Generalised Procrustes Analysis, en, *Computational Statistics (Volume 1)*, Dodge, Y. y Whittaker, J., editores, Springer-Verlag, Proceedings of the 10th Symposium on Computational Statistics, Neuchâtel, Switzerland, August (1992).

Arnold, G.M.; Collins, A.J. Interpretation of Transformed Axes in Multivariate Analysis. *Appl. Statist.* 42, 381-400 (1993).

Bangalore, A.S.; Shaffer, R.E.; Small, G.W. Genetic Algorithm-Based Method for Selecting Wavelengths and Model Size for Use with Partial Least-Squares Regression: Application to Near-Infrared Spectroscopy. *Anal. Chem.* 68, 4200-4212 (1996).

Belfiore, N.P.; Esposito, A. Theoretical and Experimental Study of Crossover Operators of Genetic Algorithms. *Journal of Optimization theory and applications* 99 (2), 271-302 (1998).

Bessant, C.; Breerton, R.G.; Dunkerley, S. Integrated processing of triply coupled diode array liquid chromatography electrospray mass spectrometric signals by

chemometric methods. *Analyst* 124, 1733-1744 (1999).

Books, K.S.; Kowalski, B.R. Comments on the DATa Analysis (DATAN) Algorithm and Rank Annihilation Factor Analysis for the Analysis of Correlated Spectral Data. *J. Chemom* 8, 287-292 (1994).

Boschelle, O.; Giomo, A.; Conte, L.; Lercker, G. Caratterizzazione delle cultivar di olivo del Golfo di Trieste mediante metodi chemiometrici applicati ai dati chimico-fisici. *Riv. Ital. Sostanze Grasse* 71(2), 57-65 (1994).

Breteron, R.G., *Multivariate Pattern Recognition in Chemometrics*. Edit. Elsevier (1992).

Bunch, J.R.; Nielsen, C.P. Updating the singular value decomposition. *Numer. Math.* 31, 111-129 (1978).

Bunch, J.R.; Nielsen, C.P.; Sorensen, D.C. Rank-one modification of the symmetric eigenproblem. *Numer. Math.* 31, 31-48 (1978).

Carlosona, A.; Andrade, J.M.; Kubista, M.; Prada, D. Procrustes Rotation as a way to Compare Different Sampling Seasons in Soils. *Anal. Chem.* 67, 2373-2378 (1995).

Centner, V.; Massart, D.L.; De Noord, O.E.; de Jong, S.; Vadegeinste, B.M.; Sterna, C. Elimination of Uninformative Variables for Multivariate Calibration. *Anal. Chem.* 68, 3851-3858 (1996).

Da Costa Filho, P.A.; Poppi, R.J. Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio, determinação simultânea de glicose, maltose e fructose. *Quim. Nova* 25(1), 46-52 (2002).

Deane, J.M.; Macfie, J.H. Testing for Redundancy in Product Quality Control Test Criteria: An Application to Aviation Turbine Fuel. *J. Chemometrics* 3, 477-491 (1989).

De Jong, K.A. Genetic Algorithms: a 10 years perspective, in: J.J. Grefenstette (Ed.), *Proceedings of the First International Conference on Genetic Algorithms*, Lawrence Erlbaum Associates (1985).

Demir, C.; Hindmarch, P.; Breteron, R.G. Procrustes Analysis for Determination of Number of Significant Masses in GCMS. *Analyst* 121(10), 1443-1449 (1996).

Dieterle, F.; Busche, S.; Gauglitz, G. Growing neural networks for a multivariate calibration and variable selection of time-resolved measurements. *Anal. Chim. Acta* 490, 71-83 (2003).

Dijksterhuis, G. Procrustes analysis in sensory research. In T. Naes & E. Risvik, *Multivariate analysis of data in sensory science* (p.185-217). *Data handling in science and technology*, volume 16. Elsevier Science B.V. (1996).

- Dunkerley, S.; Crosby, J.; Brereton, R.g.; Zissis, K.D.; Escott, R.E. Chemometric analysis of high performance liquid chromatography diode array detection electrospray mass spectrometry of 2- and 3- hydroxypyridine. *Analyst* 123(10), 2021-2033 (1998).
- Eastment, H.T.; Krzanowski, W.J. Cross-Validatory Choice of the Number of Principal Components from a Principal Component Analysis. *Technometrics* 24 (1), 73-77 (1982).
- Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* 1 (3), 211-218 (1936).
- Estévez Valencia, P. Optimización mediante Algoritmos Genéticos. *Anales del Instituto de Ingenieros de Chile*, 83-92 (1997).
- Fatemi, M.H.; Jalahi-Heravi, M.; Konuze, E. Prediction of bioconcentration factor using genetic algorithm and artificial neural network. *Anal. Chim. Acta* 486, 101-108 (2003).
- Fay, M.J.; Proctor, A.; Hoffmann, D.P.; Hercules, D.M. Improved principal component analysis of noisy data. *Anal. Chem.* 63 (11), 1058-1063 (1991).
- Fearn, T.; Davies, T. Genetic algorithms: the evolutionary solution to an old problem. *Spectroscopy Europe* 9(6), 25-27 (1997).
- Ferré, J.; Rius, F.X. A graphical criterion to examine the quality of multicomponent analysis : Implications for wavelength selection. *Trends Anal. Chem.* 16, 155-162 (1997).
- Forrest, S.; Mitchell, M. What Makes a Problem Hard for a Genetic Algorithm?. Some Anomalous Results and Their Explanation. *Machine Learning* 13, 285-319 (1993).
- Galaviz-Casas, J. Selection Analysis in Genetic Algorithms. *Lecture Notes in Computer Science* 1484, 283-292 (1998).
- Garrido-Frenich, A.; Gil-García, M.D.; Martínez-Vidal, Martínez-Galera. PLS and MLR methods in wavelength selection for multicomponent spectrophotometric data: a comparative study. *Quím. Anal.* 18 (4), 319-327 (1999).
- Gestal, M.; Gómez-Carracedo, M.P.; Andrade, J.M.; Dorado, J.; Fernández, E.; Prada, D.; Pazos, A. Classification of apple beverages using artificial neural networks with previous variable selection. *Anal. Chim. Acta* 524(1-2), 225-234 (2004).
- Glibovets, N.N.; Medvid, S.A. Genetic algorithms used to solve scheduling problems. *Cybernetics and Systems Analysis* 39 (1), 81-90 (2003).
- Goicoechea, H.C.; Olivieri, A.C. Determination of bromhexine in cough-cold syrups by absorption spectrophotometry and multivariate calibration using partial least-squares and hybrid linear analyses. Application of a novel method of wavelength selection. *Talanta* 49, 793-800 (1999).

Goldberg, D.E. Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.

González-Arjona, D.; López-Pérez, G.; González, A.G. Performing procrustes discriminant analysis with HOLMES. *Talanta* 49 (1), 189-197 (1999).

Gower, J.C. Statistical methods of comparing different multivariate analyses of the same data, in: F.R. Hodson, D.G. Kendall, P. Tautu (Eds). *Mathematics in the Archaeological and Historical Sciences*, Edinburgh Univ. Press, Edinburgh, Scotland, 138-149 (1971).

Gower, J.C. Generalized Procrustes Analysis. *Psychometrika* 40 (1), 33-51 (1975).

Grimalt, J.; Olive, J. Source input elucidation in aquatic systems by factor and principal component analysis of molecular marker data. *Anal. Chim. Acta* 278 (1), 159-176 (1993).

Guo, Q.; Wu, W.; Massart, D.L.; Boucon, C.; de Jong, S. Feature selection in sequential projection pursuit. *Anal. Chim. Acta* 446, 85-96 (2001).

Guo, Q.; Wu, W.; Massart, D.L.; Boucon, C.; de Jong, S. Feature selection in principal component analysis of analytical data. *Chemom. Intell. Lab. Syst.* 61, 123-132 (2002).

Heberger, K.; Andrade, J.M. Procrustes Rotation and Pair-wise Correlation: a Parametric and a Non-Parametric Method for Variable Selection. *Croat. Chem. Acta* 77, (2004).

Holland, J.H. *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press (1975).

Hong, T-P.; Wang, H-Sh.; Chen, W-Ch. Simultaneously Applying Multiple Mutation Operators in Genetic Algorithms. *Journal of Heuristics* 6, 439-455 (2000).

Hong, T-P.; Wang, H-Sh.; Lin, W-Y.; Lee, W-Y. Evolution of Appropriate Crossover and Mutation Operators in a Genetic Process. *Applied Intelligence* 16, 7-17 (2002).

Höskuldsson, A. Variable and subset selection in PLS regression. *Chemom. Intell. Lab. Syst.* 55, 23-38 (2001).

<http://benli.bcc.bilkent.edu.tr/~lors/ie572/burhaneddin.pdf>, 2003.

<http://cursos.itam.mx/akuri/Semestre2/AGS/19RNSyAGS.pdf>, 2003.

<http://members.tripod.com/jesus-alfonso-lopez/AgIntro.html>, 2003.

Kalogirou, S.A. Artificial intelligence for the modeling and control of combustion processes: a review. *Progress in Energy and Combustion Science* 29, 515-566 (2003).

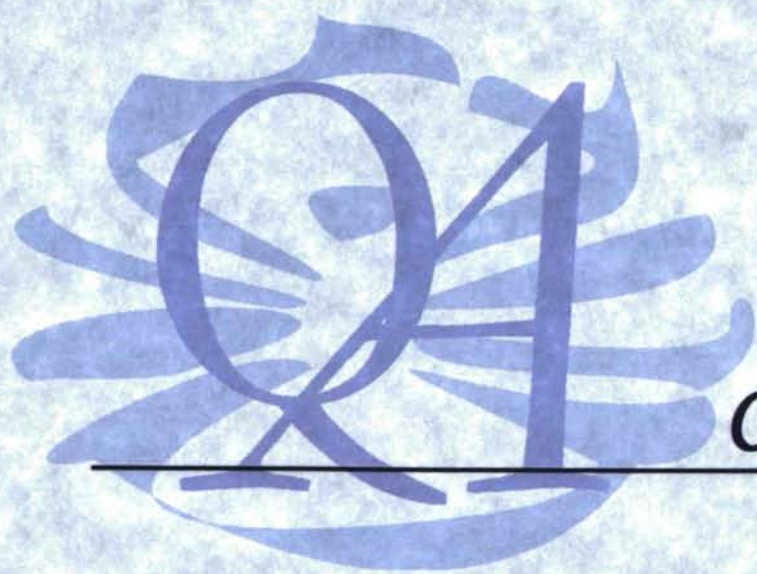
Kemsley, E.K. A genetic algorithm (GA) approach to the calculation of canonical

- variates (CVs). *Trends in Analytical Chemistry* 17 (1), 24-34 (1998).
- King, J.R.; Jackson, D.A. Variable selection with principal components analysis of environmental data. *Environmetrics* 10(1), 67-77 (1999).
- Koza, J.R. Genetic Programming. On the Programming of Computers by Means of Natural Selection. The MIT Press (1992).
- Krzanowski, W.J. *J. American Statistical Association* 74, 703-707 Corrección en 76, 1022 (1979).
- Krzanowski, W.J. Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components. *Appl. Statist.* 36 (1), 22-33 (1987a).
- Krzanowski, W.J. Cross-validation in principal component analysis. *Biometrics* 43, 575-584 (1987b).
- Krzanowski, W.J., Principles of Multivariate Analysis: A User's Perspective. Edit. Clarendon Press (Oxford, London) (1990).
- Krzanowski, W.J. Principles of multivariate analysis: a user's perspective. Oxford University Press, Inc., New York, NY, USA (2001).
- Kubista, M., A new method for the analysis of correlated data using procrustes rotation which is suitable for spectral analysis. *Chemom. Intell. Lab. Syst.* 7 (3), 273-279 (1990).
- Kubista, M.; Eriksson, S.; Albinsson, B. Quantitative Spectral Analysis Without Reference Samples. *Spectroscopy Europe* 4 (6), 28-29 (1992).
- Kubista, M.; Sjöback, R.; Albinsson, B. Determination of equilibrium constants by chemometric analysis of spectroscopic data. *Anal. Chem.* 65, 994-998 (1993).
- Kubista, M.; Nygren, J.; Elbergali, A.; Sjöback, R. Making reference samples redundant. *Crit. Rev. Anal. Chem.* 29(1), 1-28 (1999).
- Lavine, B.K.; Davidson, C.E.; Moores, A.J. Innovative genetic algorithms for chemoinformatics. *Chemom. Intell. Lab. Syst.* 60, 161-171 (2002).
- Leardi, R. Genetic Algorithms in chemometrics and chemistry: a review. *J. Chemom.* 15, 559-569 (2001).
- Libelli, S.M.; Alba, P. Adaptive mutation in genetic algorithms. *Soft Computing* 4, 76-80 (2000).
- Malinowski, E.R. Theory of the distribution of error eigenvalues resulting from principal component analysis with applications to spectroscopic data. *J. Chemometrics* 1, 33-40 (1987).
- Meusinger, R.; Moros, R. Determination of quantitative structure-octane rating

- relationships of hydrocarbons by genetic algorithms. *Chemom. Intell. Lab. Syst.* 46, 67-78 (1999).
- Nygren, J.; Andrade, J.M.; Kubista, M., Combining Thermodynamic and spectroscopic information in spectral analysis; characterization of a single sample; Conference on Chemometrics, Pardubice (Czech Republic), 3 al 7 de Julio de 1995. Presentación tipo Poster (1995).
- Osten, D.W. Selection of optimal regression models via cross-validation. *J. Chemometrics* 2, 39-48 (1988).
- Parsons, R.J.; Forrest, S.; Burks, Ch. Genetic Algorithms, Operators, and DNA Fragment Assembly. *Machine Learning* 21, 11-33 (1995).
- Pavan, M.; Consomni, V.; Todeschini, R. Development of Order Ranking Models by Genetic Algorithm Variable Subset Selection (GA-VSS), Conferentia Chemometrica 2003, Budapest, 27-29 October 2003.
- Pazos Sierra, A. Redes de neuronas artificiales y algoritmos genéticos. Ed. Univ. da Coruña (1996).
- Ramadan, Z.; Song, X.H.; Hopke, P.K.; Johnson, M.J.; Scow, K.M. Variable selection in classification of environmental soil samples for partial least square and neural network models. *Anal. Chim. Acta* 446, 231-242 (2001).
- Reyment, R.; Jöreskog, K.G. Applied Factor Analysis in the Natural Sciences. Cambridge University Press (1993).
- Richards, E.; Bessant, C.; Saini, S. Optimisation of a neural network model for calibration of voltametric data. *Chemom. Intell. Lab. Syst.* 61, 35-49 (2002).
- Richman, M.B.; Vermette, S.J. The use of Procrustes Target Analysis to discriminate dominant source regions of fine sulfur in the western United States. *Atmospheric Environment* 27A, 475-481 (1993).
- Robert, D.; Carbo-Dorca, R. A Formal Comparison between Molecular Quantum Similarity Measures and Indices. *Journal of Chemical Information and Computer Sciences* 38 (3), 469-475 (1998).
- Sánchez Pastor, M. S. Tesis doctoral. Redes Neuronales en Clasificación. Universidad de Valladolid (Departamento de Matemática Aplicada y Computación) (1997).
- Sarabia, L.; Ortiz, M.C., Componentes principales y correspondencias, en, Avances en Quimiometría Práctica, editor Cela, R., Edit. Universidade de Santiago de Compostela (Santiago, España) (1994).
- Scarmínio, I.; Kubista, M. Analysis of correlated spectral data. *Anal. Chem.* 65(4), 409-416 (1993).

- Scarponi, G.; Moret, I.; Capodaglio, G.; Romanazzi, M. Cross-validation, influential observations and selection of variables in chemometric studies of wines by principal component analysis. *J. Chemometrics* 4, 217-240 (1990).
- Schönemann, P.H.; Carroll, R.M., Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika* 35 (2), 245-255 (1970).
- Schulze, D.; Stilbs, P. Analysis of multicomponent FT-PGSE experiments by multivariate statistical methods applied to the complete bandshapes. *Journal of Magnetic Resonance* 105(1), 54-58 (1993).
- Shao, X.; Chen, Z.; Lin, X. Resolution of multicomponent overlapping chromatogram using an immune algorithm and genetic algorithm. *Chemom. Intell. Lab. Syst.* 50, 91-99 (2000).
- Shapiro, J. Genetic Algorithms in Machine Learning. *Lecture Notes in Computer Science* 2049, 146-168 (2001).
- Sibson, R. Studies in the robustness of multidimensional scaling: Procrustes statistics. *J. Royal Statistics Society, Series B*, 40, 234-238 (1978).
- Sutter, J.M.; Kalivas, J.H.; Lang, P.M. Which principal components to utilize for principal component regression. *J. Chemometrics* 6, 217-225 (1992).
- Tian, L. The Nature of Crossover Operator in Genetic Algorithms. *Lecture Notes in Computer Science* 2005, 619, 623 (2001).
- Tomas, X.; Andrade, J.M.; Alvarez-Larena, A. Chemometric analysis of skeletal data from non-fused and non- π -complexed pentafulvenes. *Talanta* 48 (4), 781-794 (1999).
- Tsukimoto, H.; Hatano, H. The functional focalization of neural networks using genetic algorithms. *Neural Networks* 16, 55-67 (2003).
- Vigneau, E.; Devaux, M.F.; Safar, M., J. Application of Procrustean methods to mid- and near-infrared spectral data. *Chemometrics* 9(2), 125-135 (1995).
- Vinzi, V.E. Explanatory methods for comparative analyses. *Chemom. Intell. Lab. Syst.* 58 (2), 275-286 (2001).
- Vivó-Truyols, G.; Torres-Lapasió, J.R.; García-Alvarez-Coque, M.C. A hybrid genetic algorithm with local search: I. Discrete variables: optimisation of complementary mobile phases. *Chemom. Intell. Lab. Syst.* 59, 89-106 (2001a).
- Vivó-Truyols, G.; Torres-Lapasió, J.R.; Garrido-Frenich, A.; García-Alvarez-Coque, M.C. A hybrid genetic algorithm with local search: II. Continuous variables: multibatch peak deconvolution. *Chemom. Intell. Lab. Syst.* 59, 107-120 (2001b).
- Wold, S., Svante Wold, Kim Esbensen and Paul Geladi. Principal component analysis. *Chemom. Intell. Lab. Syst.* 2, 37-52 (1987).

- Wold S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* 58, 131-150 (2001).
- www.asee.org/international/INTERTECH2002/833.pdf, 2003.
- www.cs.us.es/~delia/sia/html98-99/pag-alumnos/web11/indice.html, 2004.
- www.des.ime.eb.br/~sneto/paa/Genetico.pdf, 2003.
- www.devdept.com/tod/docs/sadhana.pdf, 2003.
- www.galeon.com/addyalfaro/geneticos.html, 2004.
- www.ica.ele.puc-rio.br/pesquisa/download/paper.pdf, 2003.
- www.mapp.asb.dk/WPpdf/wp20.pdf, 2003.
- www.mundo-electronico.com/PDF/Any2000/313_octubre/Algoritmos.pdf, 2003.
- www.mundonomades.com/von/aGEsoullier.pdf, 2003.
- www.orcero.org/irbis/disertacion/node189.html, 2003.
- www.redcientifica.com/doc/doc199904260011.html, 2003.
- www.ub.rug.nl/eldoc/dis/science/m.m.lankhorst/c2.pdf, 2003.
- www.uv.es/asepuma/jornadas/madrid/J24C.pdf, 2003.



Capítulo III

Espectroscopia IR

Objetivo:

Se presentan aquí las técnicas analíticas instrumentales empleadas en esta Memoria, haciendo una pequeña discusión de sus ventajas e inconvenientes. Se recopilan algunos trabajos previos que constituyen la base de los estudios que se abordan en Capítulos posteriores.

Índice:

1. *Introducción a la espectroscopia IR*
 - 1.1. *Espectroscopia MIR*
 - 1.2. *Espectroscopia Raman*
2. *Aplicaciones de la espectroscopia*
 - 2.1. *Aplicaciones de la espectroscopia IR en el campo petroquímico*
 - 2.2. *Aplicaciones de la espectroscopia IR en el campo alimentario*
 - 2.3. *Aplicaciones de la espectroscopia Raman*
3. *Instrumentación analítica empleada y control de calidad*
4. *Bibliografía*

1. INTRODUCCIÓN A LA ESPECTROSCOPIA IR

La espectroscopia vibracional es una poderosa herramienta que se emplea para el estudio de la estructura y/o perfil composicional de productos de naturaleza orgánica e inorgánica (Schmidt *et al.*, 2001), tanto sustancias puras como mezclas complejas. Al compararla con otras técnicas usadas en problemas complejos, en general, presenta grandes ventajas en cuanto a simplicidad de uso, manipulación de muestra, mantenimiento y poco tiempo de análisis, lo cual es importante en ambientes industriales (Chalmers and Griffiths, 2002).

A lo largo de dos décadas se han empleado abundantemente tres técnicas vibracionales en control de calidad: la espectrometría de infrarrojo cercano (NIR), muy usada en control de calidad de alimentación (Simnaeve *et al.*, 1997; Osborne *et al.*, 1993; Esteban-Díez *et al.*, 2004) y en el campo petroquímico (Bohacs *et al.*, 1998; Parisi *et al.*, 1990; Macho y Larrechi, 2002; Breitzkreitz *et al.*, 2003); la espectrometría de infrarrojo medio (MIR) y la espectroscopia Raman, bastante menos utilizada por su mayor complejidad y carestía.

La espectroscopia vibracional, además, ha cobrado nuevo vigor gracias a los desarrollos de la espectroscopia infrarroja cercana (NIR) y las diversas técnicas de trabajo disponibles para la zona media (MIR).

1.1. ESPECTROSCOPIA MIR

La zona MIR presenta ventajas importantes respecto a la de tipo NIR (actualmente muy aplicada) tales como estabilidad de los equipos, estandarización sencilla y ampliamente establecida, técnicas de trabajo sencillas y fácil correlación entre el espectro obtenido y la composición genérica del producto bajo análisis. Esta zona será la que se aborde aquí.

1.1.1. TÉCNICAS DE REFLEXIÓN

Las técnicas de reflexión difieren de las de transmisión (típicamente las más aplicadas) en que el haz IR es "despedido" fuera de la muestra en lugar de pasar a través de ella. Una ventaja de las técnicas de reflexión frente a las de transmisión es que no son destructivas (ya que no suelen exigir disponer de la muestra en disolución) y, además, la preparación de la muestra es más rápida y más fácil que para las técnicas de transmisión (frecuentemente, innecesaria). Ahora bien, el espectro de reflexión

puede ser más ruidoso que el de transmisión para un determinado número de barridos y una resolución fija debido a que es difícil capturar toda la luz reflejada fuera de la superficie de la muestra.

Hay tres métodos principales de reflexión (Smith, 1996):

1. **Reflectancia especular:** Cuando el ángulo de incidencia es igual al ángulo de reflexión. Una muestra transparente es situada en frente de una superficie de reflexión. El haz pasa a través de la muestra y se refleja de vuelta también a través de la muestra para producir un espectro de transmisión equivalente al doble de transmisión (doble transmisión).

Un diagrama óptico de un accesorio de reflectancia especular simple consiste en dos espejos planos y una plataforma con un agujero. La muestra se sitúa en la plataforma. El primer espejo dirige la radiación a la muestra, mientras el segundo espejo captura la radiación reflejada fuera de la muestra y la dirige al detector. Se usa frecuentemente para examinar películas de polímeros en metales.

2. **Reflectancia difusa (DRIFTS):** La muestra se centra debajo de un gran espejo elipsoide. El haz es concentrado sobre la muestra sólida (en general, finamente dividida) causando que ésta refleje de vuelta la luz en un patrón difuso el cual puede ser recogido y analizado. La técnica se puede aplicar a polvos, espumas, fibras, carbón, plásticos y hojas, donde las técnicas convencionales son difíciles o imposibles de usar.

El diagrama óptico de un accesorio DRIFTS consiste en espejos planos que dirigen la radiación al espejo elipsoidal o esférico enfocado. La radiación es enfocada en el compartimento de la muestra y es difusivamente reflejada en un círculo de 360°. La radiación difusivamente reflejada se compone de radiación dispersa, absorbida, transmitida y reflejada por la muestra. Un segundo espejo elipsoidal o esférico recoge la radiación reflejada difusivamente que, a continuación, es enfocada al detector.

3. **Reflexión interna:** Se trata con detenimiento a continuación ya que será la que se use en los análisis presentados en esta Memoria.

1.1.1.1. ESPECTROSCOPIA DE REFLEXIÓN INTERNA

La espectroscopia de reflexión interna, o de reflectancia total atenuada (ATR), fue introducida en la espectroscopia IR en 1959 a través de 2 publicaciones que aparecieron en la misma época. *Fahrenfort* (1961) y *Harrick* (1960) desarrollaban el

método de reflexión total atenuada (ATR) para análisis IR en 1959. De hecho, el fenómeno físico de espectroscopia de reflexión interna fue observado por Newton, pero no fue hasta las publicaciones de *Fahrenfort (1961)* y *Harrick (1960)* cuando se encontraron posibles aplicaciones analíticas. *Fahrenfort* se refería a la técnica como reflectancia total atenuada, ATR, y la trataba desde el punto de vista de la química analítica, mientras que *Harrick* prefería la terminología física de espectroscopia de reflexión interna y la trataba desde un punto de vista físico. Más tarde (en 1967) las normas ASTM definían los términos (*Jemison et al., 1992*).

El sistema de ATR es una técnica útil y ampliamente aplicada para obtener espectros de muestras difíciles de manejar en la zona MIR tales como gomas, muestras acuosas, alimentos o resinas solidificadas.

1.1.1.2. LA FÍSICA DE LA REFLEXIÓN INTERNA

Los sistemas de ATR están basados en las propiedades de la refracción de la luz y sus leyes básicas. La refracción de la radiación se define como el cambio de velocidad, dirección y sentido que experimenta un haz de radiación al pasar de un medio transparente a otro de diferente índice de refracción.

De acuerdo con la segunda ley de Snell, la relación entre los senos de los ángulos de incidencia y refracción es igual a la relación entre los respectivos valores de la velocidad de la radiación en ambos medios y, en consecuencia, inversamente proporcional a sus respectivos índices absolutos de refracción. De acuerdo con ello, se formula la siguiente regla nemotécnica: "*si un rayo pasa de un medio de mayor índice de refracción a otro medio de menor índice de refracción, se aleja de la normal; y a la inversa*". De esta forma, al ir variando (aumentando) el ángulo de incidencia se llegará a uno (ángulo crítico) para el cual el rayo refractado será rasante con la interfase. Para un ángulo de incidencia mayor, el ángulo de refracción será mayor de 90° y se pasa a un fenómeno de reflexión total.

En primer lugar considérese lo que ocurre cuando un haz de radiación pasa a través de una interfase entre dos materiales transparentes de distintos índices de refracción. Para un haz perpendicular a la interfase, la radiación será parcialmente transmitida y parcialmente reflejada, con una reflectancia, R , (la fracción de radiación que ha sido reflejada) dada por:

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

Donde n_1 y n_2 son los índices de refracción de los dos medios en la interfase.

Si la radiación pasa de un medio de un determinado índice de refracción, n_1 , a través de una interfase a un medio de más bajo índice de refracción, n_2 , a ángulos próximos a 90° , entonces "los objetos podrán ser vistos desde detrás de la interfase" (fenómeno de refracción).

El ángulo crítico α_c para el que ocurre la reflexión total será calculado como (tégase en cuenta que el ángulo de refracción es 90°):

$$\text{sen } \alpha_c = \frac{n_2}{n_1}$$

La reflexión a ángulos mayores que el ángulo crítico es del 100%. La reflexión interna que se produce confinando la radiación en un medio es virtualmente "perfecta" en contraste con la reflexión externa (donde algún porcentaje de la radiación incidente es perdida en cada reflexión) por tanto, un haz reflejado internamente puede tener miles de reflexiones sin pérdida de energía, excepto las absorciones del medio.

El uso de ATR en espectroscopia se basa en que a pesar de que la reflexión interna completa ocurre en la interfase, la radiación penetra una corta distancia en el medio de índice de refracción más bajo y, precisamente, esta radiación evanescente se usa en análisis de IR para obtener el espectro.

Como ya se indicó, la Reflectancia Total Atenuada es una técnica de IR conocida y aplicada desde hace unos años, especialmente en el análisis industrial de polímeros. La característica esencial del ATR es que permite el muestreo superficial de películas, líquidos, sólidos y resinas depositadas en un material transparente a IR sin prácticamente preparación previa de dichas muestras (Piccolo, 1994; Smith, 1979).

El sistema ATR horizontal es un accesorio colocado de manera que la radiación IR enfocada sobre él viaja de forma horizontal y cuyo "núcleo" básico es un material transparente al IR de alto índice de refracción (KRS-5, ioduro de

talio/bromuro de talio; Germanio ó ZnSe, para el análisis de sustancias viscosas, disoluciones orgánicas y acuosas) (Piccolo, 1994). La radiación incidente es dirigida mediante espejos paraboloides al cristal seleccionado, que refleja internamente la onda a lo largo de la longitud del cristal, actuando éste como guía de la radiación. Cada vez que se produce una reflexión total de radiación se crea una onda, llamada onda evanescente (ver Figura 1). La propiedad más importante de la onda evanescente es que es ligeramente más grande que el cristal y, por tanto, se extiende una pequeña distancia por encima de la superficie del cristal. La amplitud (o energía) de la onda evanescente disminuye exponencialmente con la distancia en el material de índice de refracción menor y su intensidad, que varía con el cuadrado de la amplitud, decae muy rápidamente (Smith, 1979). Una muestra en contacto con el cristal interactúa con la onda evanescente, absorbe la radiación IR y esto permite registrar su espectro de IR. La onda evanescente es atenuada por la absorbancia de la muestra, de ahí el nombre de ATR. La intensidad de absorción resultante es proporcional al número de reflexiones del haz de IR en el cristal y a la profundidad de penetración de la onda evanescente (Piccolo, 1994) aunque en la práctica la señal obtenida a la salida del cristal es normalmente del 20-50% de la radiación incidente.

Se necesita buen contacto entre la muestra y el cristal para asegurar que la onda evanescente penetra en la muestra, por eso el cristal debe estar bien limpio y sin ralladuras. También, a veces, es necesaria presión para asegurar el buen contacto entre el cristal y la muestra (Smith, 1996).

Los prismas (cristales) empleados en ATR pueden ser de una gran variedad de geometrías y de ángulos de entrada (normalmente son de 45° ó 60°). En cualquier caso, el ángulo de incidencia tiene que ser más grande que el ángulo crítico para que ocurra la reflexión total interna. El hecho de que el ángulo crítico varía con las longitudes de onda que forman una banda de absorción sugiere que α (ángulo de incidencia) debería ser apreciablemente más grande que α_c (ángulo crítico).

Otros factores que influyen en las medidas con ATR son el número de reflexiones (el cual será mayor para ángulos de incidencia altos) y el área de contacto de la muestra. El número de reflexiones puede ser fácilmente aumentado usando un elemento multireflexión más delgado o más largo (Smith, 1979), de acuerdo con la expresión:

$$N = \frac{L}{t} \cot \alpha$$

donde L es la longitud del cristal, t el espesor, α es el ángulo de incidencia y N es el número total de reflexiones.

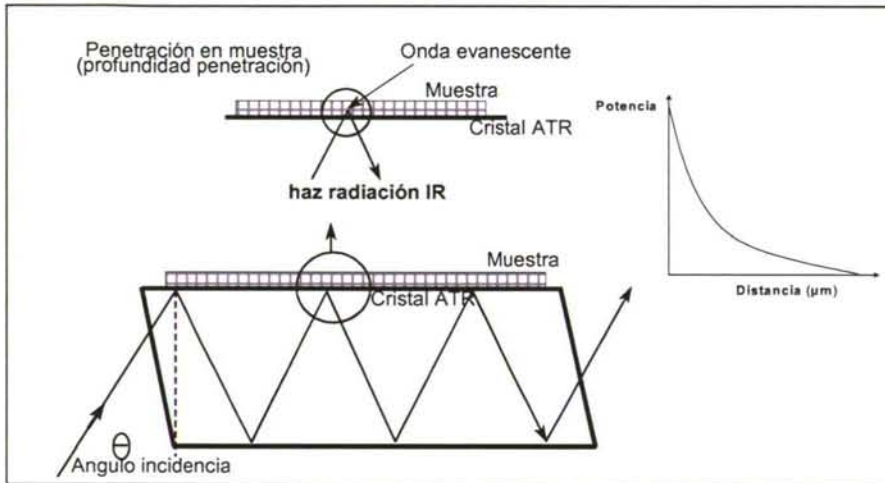


Figura 1: Esquema simplificado del sistema de Reflectancia Total Atenuada Horizontal. Detalle del concepto de onda evanescente y profundidad de penetración.

En el caso de esta Memoria, se ha trabajado con un cristal trapezoidal de ZnSe con un ángulo de incidencia de 45° y $N=12$.

La distancia que penetra la radiación en la muestra es conocida como profundidad de penetración (d_p) (Grdadolnik, 2002), la cual es definida como la profundidad a la cual la onda evanescente es atenuada a un 36.8% ($1/e$) de su intensidad total. La d_p se calcula por:

$$d_p = \frac{\lambda}{2\pi n_c \sqrt{\sin^2 \alpha - (n_s / n_c)^2}}$$

Donde la λ es la longitud de onda expresada en mm; n_c es el índice de refracción del cristal, que en el caso del ZnSe es de 2.4; n_s es el índice de refracción de la muestra y α es el ángulo del cristal expresado en grados, que en el caso de esta Memoria es de 45° (como se indicó arriba).

La profundidad de penetración es dependiente del número de onda. La d_p disminuye cuando el número de onda aumenta. Así, la radiación de número de onda bajo penetra más lejos en la muestra que la de onda alto. Como consecuencia, el espectro de ATR muestra picos que son más intensos a números de onda bajos que a números de onda altos. La dependencia de la profundidad de penetración del número de onda causa intensidades de pico relativas distintas para el espectro de ATR y el espectro de transmisión de la misma muestra (Smith, 1996). Debido a esto se usa un factor de corrección al que llamamos MIR (Reflexión Interna Múltiple), que incluye además un factor de corrección de contacto (C). Este algoritmo de corrección compensa el área de muestreo (profundidad de penetración), análogamente a la longitud de paso óptico en espectroscopia de transmisión. Cuando la compensación es completa, el espectro resultante corresponde a un espectro de transmisión para el material muestreado. La corrección MIR se calcula de la siguiente forma:

$$\text{corrección MIR} = \frac{\text{valor } (\%T \text{ ó } A)}{\lambda - C}$$

El valor de C depende de la pérdida del área de muestreo debido al pobre contacto ente la muestra y el cristal. En nuestro caso, al ser la muestra un líquido, se considera que hay buen contacto entre la muestra y el cristal, que el contacto es uniforme y, por tanto, que $C=0$.

La d_p disminuye al aumentar el índice de refracción del cristal ATR por lo que cambiando el material del cristal se puede obtener el espectro desde diferentes profundidades en la muestra, lo que se conoce como perfil de profundidad. Análogamente, se podría obtener el cambio de la composición de la muestra con la profundidad o, en el caso de películas laminadas que poseen capas de composición, su composición con la profundidad.

La d_p también disminuye al aumentar el ángulo de incidencia de la radiación de IR. El ángulo de incidencia se puede variar cambiando el ángulo de biselado del

cristal de ATR o cambiando el ángulo de la radiación. Algunos sistemas del ATR poseen espejos móviles que cambian el ángulo de incidencia y, entonces, se varía la d_p de forma controlada.

Afortunadamente, el índice de refracción de la mayoría de los compuestos orgánicos es muy parecido, así que la d_p varía poco de muestra a muestra (Smith, 1996). La longitud de campo efectivo nominal (EPL) será entonces:

$$EPL = N * d_p$$

donde N es el número de reflexiones y d_p es la profundidad de penetración. La profundidad de penetración que se calcula en el ATR es un valor nominal; en realidad la profundidad real es más grande que la profundidad de penetración (normalmente entre 2.3 a 2.5 veces mayor) (Jakusch et al., 1997). La longitud de campo efectivo es equivalente a la longitud de campo de la celda o espesor de la muestra en el experimento de transmisión tradicional.

1.1.1.3. VENTAJAS DEL ATR

1. La preparación de la muestra es fácil ó innecesaria. Además, la limpieza y manipulación del soporte de la muestra es sencillo, aunque no debe descuidarse.
2. Los espectros son independientes del espesor de la muestra.
3. Aumentando el número de reflexiones múltiples, por cambio del cristal del ATR, puede aumentar la longitud de campo y, por tanto, la absorbancia medida.
4. Las intensidades de banda del espectro de ATR varían con la concentración de acuerdo con la ley de Beer.
5. Con el ATR el análisis cuantitativo de disoluciones acuosas es posible, algo no viable usando la espectroscopia IR clásica y que resulta imprescindible para parte del trabajo presentado en esta Memoria.
6. Es una técnica que implica poca manipulación de las muestras y, en consecuencia, facilita el trabajo rutinario del laboratorio y acelera el proceso analítico.

1.2. ESPECTROSCOPIA RAMAN

La teoría de la dispersión Raman demuestra que el fenómeno está relacionado con el mismo tipo de cambios vibracionales cuánticos que se producen en la absorción infrarroja clásica y, por tanto, la diferencia de longitud de onda entre la radiación incidente y la dispersada corresponde a las longitudes de onda de la región del infrarrojo medio. En efecto, el espectro de dispersión Raman y el espectro de absorción infrarrojo para una especie determinada a menudo suelen parecerse mucho entre sí. Sin embargo, hay suficientes diferencias entre los tipos de grupos funcionales que son activos en el infrarrojo y los que lo son en Raman como para que ambas técnicas no resulten competitivas sino complementarias (Skoog y Leary, 1994). Para resolver algunos problemas el infrarrojo clásico es una técnica superior, sin embargo, para otros el Raman proporciona unos espectros más útiles.

La espectroscopia Raman ha sido menos aplicada en la industria debido a su alto coste, baja relación señal/ruido cuando se compara con NIR o MIR, problemas de fluorescencia de la muestra cuando se usan bajas longitudes de onda de excitación del laser y muy pequeñas diferencias espectrales entre muestras. Aunque la espectroscopia Raman continúa siendo bastante cara; la mayoría de las desventajas han sido solucionadas debido al acoplamiento rutinario de la espectroscopia Raman con la Transformada de Fourier, láseres más potentes y las técnicas quimiométricas.

Una de las propiedades más interesantes de la espectroscopia Raman (desde el punto de vista industrial) es la capacidad de medir las muestras directamente en viales de vidrio corriente ya que esto simplifica y acelera la manipulación de la muestra. Así, la espectroscopia Raman puede ser considerada como una técnica "limpia", con bajo consumo de muestra líquida (usando viales de vidrio), no necesita reactivos previamente a la medida, no hay residuos y no se generan vapores durante la manipulación de la muestra (Andrade et al., 2003). Todas estas ventajas hacen de la espectroscopia Raman una técnica bastante interesante y prometedora para el control de calidad industrial de rutina.

2. APLICACIONES DE LA ESPECTROSCOPIA

2.1. APLICACIONES DE LA ESPECTROSCOPIA IR EN EL CAMPO PETROQUÍMICO

Aunque esta metodología no es nueva en diferentes campos de la petroquímica, sí que resulta de cierta novedad en cuanto a su aplicación al combustible de aviación. La espectroscopia se ha aplicado para analizar diversos combustibles (Fodor et al., 1993; Fodor et al., 1994; Fodor et al., 1996; Iob et al., 1996; Macho et al., 1999; Bernabei et al., 2003), determinar propiedades físicas (Garrigues et al., 1995; Chung et al., 1999), determinar benceno (Gallignani et al., 1993; Ródenas-Torralba et al., 2004), tolueno (Gallignani et al., 1994) y metil ter-butyl éter (De la Guardia et al., 1993). Dube et al. (2004) comparan ATR-FTIR y GPC para la monitorización de la producción de biodiesel.

También se han encontrado diversos artículos que estudian las bandas de absorción de IR de los diferentes grupos funcionales que pueden ayudar en la identificación de los diferentes constituyentes del queroseno (Bertie et al., 1993; Bertie et al., 1994a; Bertie et al., 1994b; Bertie et al., 1995; Bertie y Zhang, 1994; Haluki et al., 1995; Etzkorn, 1999), además de los textos clásicos (Morcillo y Madroñero, 1962; Conley, 1979; Robinson, 1991; Roeges, 1994; Pretsch et al., 1998; Günzler y Gremlich, 2002).

2.2. APLICACIONES DE LA ESPECTROSCOPIA IR EN EL CAMPO ALIMENTARIO

La espectroscopia IR también se ha empleado en diversos estudios de productos alimenticios (Wilson y Tapp, 1999); en la determinación de azúcares (Le Thanh et al., 2000; Rodríguez-Saona et al., 2001; Tewari et al., 1999; Cadet, 1999; Duarte et al., 2002; Garrigues et al., 1998; Rambla et al., 1998; Irudayaraj y Tewari, 2003; Sivakesava y Irudayaraj, 2000, Da Costa Filho y Poppi, 2002); diferentes ácidos orgánicos en zumos y refrescos (Ayora-Cañada et al., 2000; Le Thanh et al., 2000; Tewari et al., 1999; Irudayaraj y Tewari, 2003); determinación de sacarosa en raíz de remolacha (Garrigues et al., 2000), determinación de vitaminas (Yang y Irudayaraj, 2002; Wojciechowski et al., 1998); estudio de aceite de oliva (Iñón et al., 2003; Tapp et al., 2003; Innawong y Zulick, 1996; Tay et al., 2002); estudio de vino (Patz et al., 2004; Urtubria et al., 2004; Moreira y Santos, 2004; Kupina y Shrikhande, 2003; Endelmann et al., 2003; Schindler et al., 1998); control de calidad de cerveza (Duarte et al., 2004); estudio de mermelada (Kelly et al., 2004; Tewari y Irudayaraj, 2004; Sivakesava y Irudayaraj, 2001); estudio de jarabe de arce (Paradkar et al., 2002a;

Paradkar, 2002b; Paradkar, 2002c); determinación del origen geográfico del queso Emmental (Pillonel *et al.*, 2003); determinación de parámetros nutricionales en muestras de leche (Iñón *et al.*, 2004) y detección de adulteración de zumos (Sivakesava *et al.*, 2001).

A continuación se presentan algunos trabajos recientes en los cuales se emplea la técnica de ATR:

- *Baucells et al. (1991)* comparan distintas técnicas (CIRCLE, ATR, DRIFT y transmisión) para el análisis cuantitativo de la cafeína.
- *Defernex et al. (1995)* aplican la técnica ATR para clasificar purés de frutas en tres tipos: fresas, frambuesas y manzanas, también trata de distinguir entre purés frescos y descongelados.
- *McQueen et al. (1995)* comparan la espectroscopia de infrarrojo optotérmica con la FTIR-ATR aplicadas a 24 muestras de quesos para obtener el contenido de grasas, proteínas y mezclas de contenidos.
- *Bellon-Maurel y Gilles (1994)* y *Bellon-Maurel et al. (1995)* usan FTIR-ATR para cuantificar azúcares individuales en mezclas reales extraídas durante la hidrólisis de azúcares y almidón a nivel industrial.
- *Kemsley et al. (1996)* aplican FTIR-ATR para detectar la adulteración de puré de frambuesa con sacarosa, manzana o ciruela.
- *Cadet y Offmann (1997)* determinan sacarosa en azúcar de caña mediante FTIR-ATR y procesan los datos mediante PCA y PCR.
- *Rambla et al. (1998)* utilizan la técnica FTIR-ATR para la determinación directa de azúcares en zumos (manzana, naranja, uva) y refrescos.
- *Garrigues et al. (1998)* comparan la técnica de HATR con la CIRCLE para la determinación de azúcares en refrescos y zumos de fruta empleando PLS.
- *Cadet (1999)* mide el contenido de azúcar mediante FTIR-ATR, identifica las bandas características y efectúa un análisis de componentes principales para medir sus concentraciones.
- *Garrigues et al. (2000)* desarrollaron un procedimiento simple y rápido para la determinación directa de sacarosa en muestras de raíz de remolacha mediante ATR-FTIR.

- Inawong y Zulick (1996) e Inawong et al. (2004) determinan la calidad del aceite frito.
- Duarte et al. (2004) emplea FTIR y RMN para el control de calidad de la cerveza.
- Kelly et al. (2004) y Sivakesava e Irudayaraj (2001) estudian la adulteración de la mermelada empleando FTIR-ATR y métodos quimiométricos.
- Irudayaraj y Tewari (2003) emplean FTIR-ATR para la monitorización de ácidos orgánicos y azúcares en zumos de manzana.
- Sivakesava et al. (2001) estudian la adulteración de zumo de manzana.
- Iñón et al. (2003) determinan la acidez del aceite de oliva.
- Yang e Irudayaraj (2002) determinan rápidamente la vitamina C empleando diversas técnicas vibracionales (NIR, MIR y FT-Raman).

2.3. APLICACIONES DE LA ESPECTROSCOPIA RAMAN

La espectroscopia Raman se aplica para resolver problemas en diferentes campos de la analítica: análisis ambiental (Skoulika et al., 2000; Jager et al., 2000; Mosier-Boss y Lieberman, 2000); estudios de arte e históricos (Burgio y Clark, 2000; Burgio et al., 2000; Burgio et al., 2003; Zuo et al., 1999a; Zuo et al., 1999b; Brown y Clark, 2004a; Brown y Clark, 2004b); bioanálisis (Pappas et al., 2000); alimentación (Marigheto et al., 1998; Aparicio y Baeten, 1998; Davies et al., 2000; Marquardt y Wold, 2004; Baeten et al., 1998); polímeros (Chalmers y Everall, 1996); briquetas para encendido del hogar (Alia et al., 1999); clasificación de maderas (Yang et al., 1999), etc.

En cuanto al área de la petroquímica, de interés en esta Memoria, la espectroscopia Raman ha sido menos empleada, aunque una de las primeras aplicaciones consistía en predecir porcentajes de masa de mezclas líquidas de fuel de gasolina sin plomo, gasolina súper sin plomo y diesel (Seasholtz et al., 1989). Dos artículos realizaban aplicaciones “cualitativas” en diesel (Zhang et al., 1996) y combustibles de turbinas de aviación. Otro trabajo más reciente interpretaba el espectro y diferenciaba algunos de los diferentes tipos de fueles de aviación (Chung et al., 1991).

La espectrometría Raman y las calibraciones multivariadas se han empleado para evaluar muestras de residuos atmosféricos de productos pesados del petróleo (Chung y Ku, 2000) y para la determinación de benceno, tolueno y etilbenceno en

petróleo tipo "mock" (Flecher et al. 1996). Se empleó FT-Raman y PLS para cuantificar el peso del porcentaje de oxígeno en gasolinas (Cooper et al., 1996 y Cooper et al., 1997) y para comparar su perfil con otros modelos multivariados desarrollados usando los rangos espectrales de IR cercano y medio. Se predijo el número de octano y la presión de vapor Reid de gasolinas empleando espectroscopia Raman dispersiva con fibra óptica y detección CCD (Flecher et al., 1997). No se han encontrado aplicaciones de espectroscopia Raman para predecir las propiedades físico-químicas del combustible de aviación (keroseno) en test diarios ni en control de calidad industrial de muestras comerciales, objetivos incluidos en esta Memoria.

3. INSTRUMENTACIÓN ANALÍTICA EMPLEADA Y CONTROL DE CALIDAD

El equipo instrumental empleado ha sido un espectrofotómetro de IR con Transformada de Fourier, Perkin-Elmer modelo 16PC, operando en la región media del IR ($4000-400\text{cm}^{-1}$) con un detector DTGS (sulfato de triglicina deuterado), un beamsplitter de KBr y Ge, acoplado con un accesorio de reflexión interna (Reflectancia Total Atenuada, ATR) horizontal con cristal trapezoidal de ZnSe con un ángulo de incidencia de 45° y doce reflexiones nominales. La profundidad de penetración se calcula según la ecuación indicada con anterioridad. Dado que ésta depende del número de onda, se presenta una figura que recoge el "perfil de profundidades" obtenido en la región de trabajo (Figura 2).

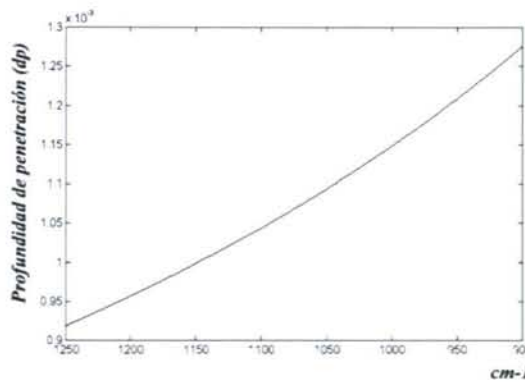


Figura 2: Variación en la profundidad de penetración nominal del haz en la muestra en función de los números de onda considerados (antes de efectuar la corrección).

El comportamiento del espectrofotómetro FTIR tiene que ser verificado (validado) regularmente. Hay diversos procedimientos relativamente sencillos para controlar el buen funcionamiento del equipo IR empleado. Siguiendo a Smith (1996) y la norma ASTM E 1421 (2004), semanalmente se efectuaron algunos chequeos allí propuestos: aspecto del interferograma, porcentaje del beamsplitter, línea del 100%, relación S/R, energía del láser y perfil del fondo.

Para monitorizar la exactitud del número de onda del equipo se midió semanalmente el espectro de una película de poliestireno y se compararon las longitudes de onda de algunas de sus bandas de absorción con los valores de referencia y sus intervalos de confianza (Garfield, 2000; Smith, 1991; Barnes y Dent, 1994; NIST, 2004). Los resultados encontrados están siempre en los rangos aceptables. Aunque el poliestireno es generalmente la opción elegida por su simplicidad, hay otras posibilidades como el uso de indeno o mezclas de indeno-campfor-ciclohexanona que se usan, principalmente, a nivel de fabricantes de los equipos (Smith, 1991).

Mensualmente, se purga con N₂ seco, se revisan las ventanas de KBr y se cambia el desecante (semestralmente). Si un parámetro está fuera de control, se repite el chequeo completo y si el problema se confirma, se realizan operaciones de mantenimiento. Los chequeos efectuados se resumen a continuación:

a- **Energía del láser:** se mide y registra regularmente la energía con el ATR colocado.

b- **Aspecto del interferograma:** Se debe tener un pico agudo (llamado *centerburst*) y sus lóbulos laterales (*wings*) tienen que decaer poco a poco, no deben visualizarse inflexiones bruscas.

c- **Perfil del fondo:** Se registra el espectro del fondo, con apodización fuerte, un único barrido, resolución de 4 cm⁻¹ en el rango de 4400- 450 cm⁻¹ y se hallan las relaciones de intensidades mostradas a continuación. El perfil debería ser reproducible de un día a otro y se chequea si los picos del agua y del CO₂ son extremadamente intensos.

$$c1) f(4400) / f(2600) > 0.25$$

$$c3) f(2600) * 100 > 65$$

$$c2) f(450) / f(2600) > 0.025$$

$$c4) \frac{f(1657.9) - f(1651.9)}{f(1657.9)} * 100 > 10\%$$

A modo de ejemplo, se presenta la carta de control para la primera de las relaciones, c1) (Figura 2). El comportamiento de las otras tres es análogo.

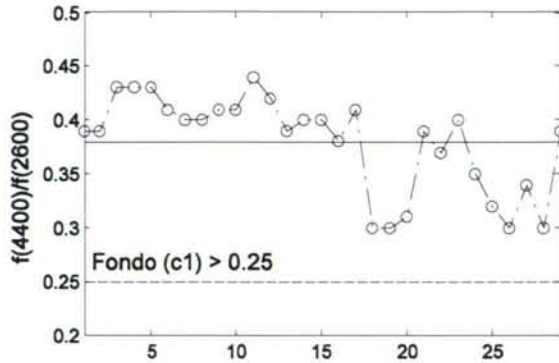


Figura 2: Relación Perfil de fondo para $f(4400)/f(2600) > 0.25$

d- **Porcentaje del beamsplitter:** Se hace un espectro del fondo con un barrido, apodización fuerte, resolución 4cm^{-1} y rango $4400\text{-}450\text{ cm}^{-1}$ y se calcula la relación de intensidades entre la energía a 4000 cm^{-1} y a la banda del CO_2 :

$$\%B = f(4000)/f(\text{CO}_2) \geq 0.25\%$$

e- **Relación S/N (pico/pico):** Se hace el espectro del fondo (16 barridos) y, a continuación, un espectro del aire con un barrido, apodización fuerte, resolución de 4cm^{-1} y en el rango $4400\text{-}450\text{ cm}^{-1}$. Se determinan las intensidades máxima y mínima en el rango $2250\text{-}2150\text{cm}^{-1}$ y se calcula:

$$S/N (\text{pico/pico}) = e(\text{máx}) - e(\text{mín}) < 0.1\%$$

En la **Figura 3** se presenta la carta de control asociada a este chequeo. Aunque se aprecia que, ocasionalmente, la relación S/N excede el límite de especificación del equipo para una medida particular, la repetición de las medidas permitía comprobar que la relación S/N se situaba en los niveles correctos (8 de cada 10 medidas eran correctas, como indica la especificación del equipo).

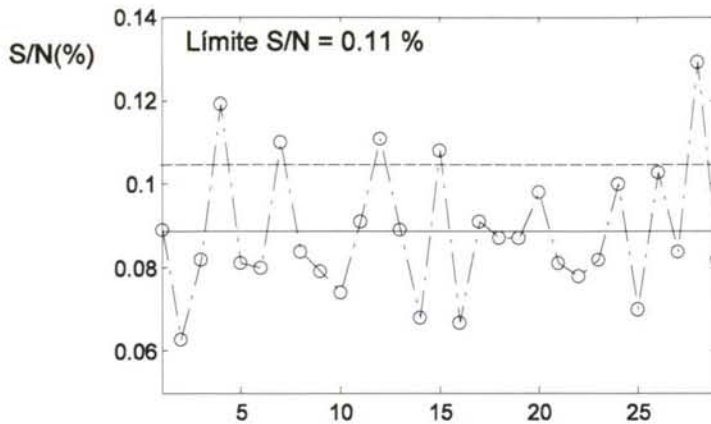


Figura 3: Relación señal/ruido

- La **exactitud del número de onda** se evalúa empleando una película de poliestireno, midiéndola con apodización fuerte en el rango $4000-600\text{ cm}^{-1}$, con 16 barridos y resolución de 2 cm^{-1} con apertura de Jackinot, a los siguientes números de onda: $2851.5 \pm 3\text{ cm}^{-1}$ (Garfield, 2000); $1601.8 \pm 1.5\text{ cm}^{-1}$ (Garfield, 2000) ($1601.4 \pm 0.5\text{ cm}^{-1}$ (NIST, 2004)); $1028.3 \pm 1.5\text{ cm}^{-1}$ (Garfield, 2000) ($1028.4 \pm 0.5\text{ cm}^{-1}$ (NIST, 2004)); $3027.1 \pm 0.3\text{ cm}^{-1}$ ($3028.2 \pm 0.5\text{ cm}^{-1}$ (NIST, 2004)); $2924 \pm 2\text{ cm}^{-1}$; $1944 \pm 1\text{ cm}^{-1}$; $1871 \pm 0.3\text{ cm}^{-1}$; $906.7 \pm 0.3\text{ cm}^{-1}$ (Smith, 1991), verificando la posición de los picos espectrales obtenidos experimentalmente a esos números de onda.

g- La **exactitud en transmitancia** se evalúa determinando dos relaciones de picos (Barnes y Dent, 1994) leyendo las intensidades a los números de onda de 3060, 3001 y 906 cm^{-1} . Se hallaron las dos relaciones de pico siguientes:

$$A(3060)/A(906)$$

$$A(3001)/A(906)$$

En la **Figura 4** se resumen los datos obtenidos en el control de la exactitud en el número de onda y en la absorbancia.

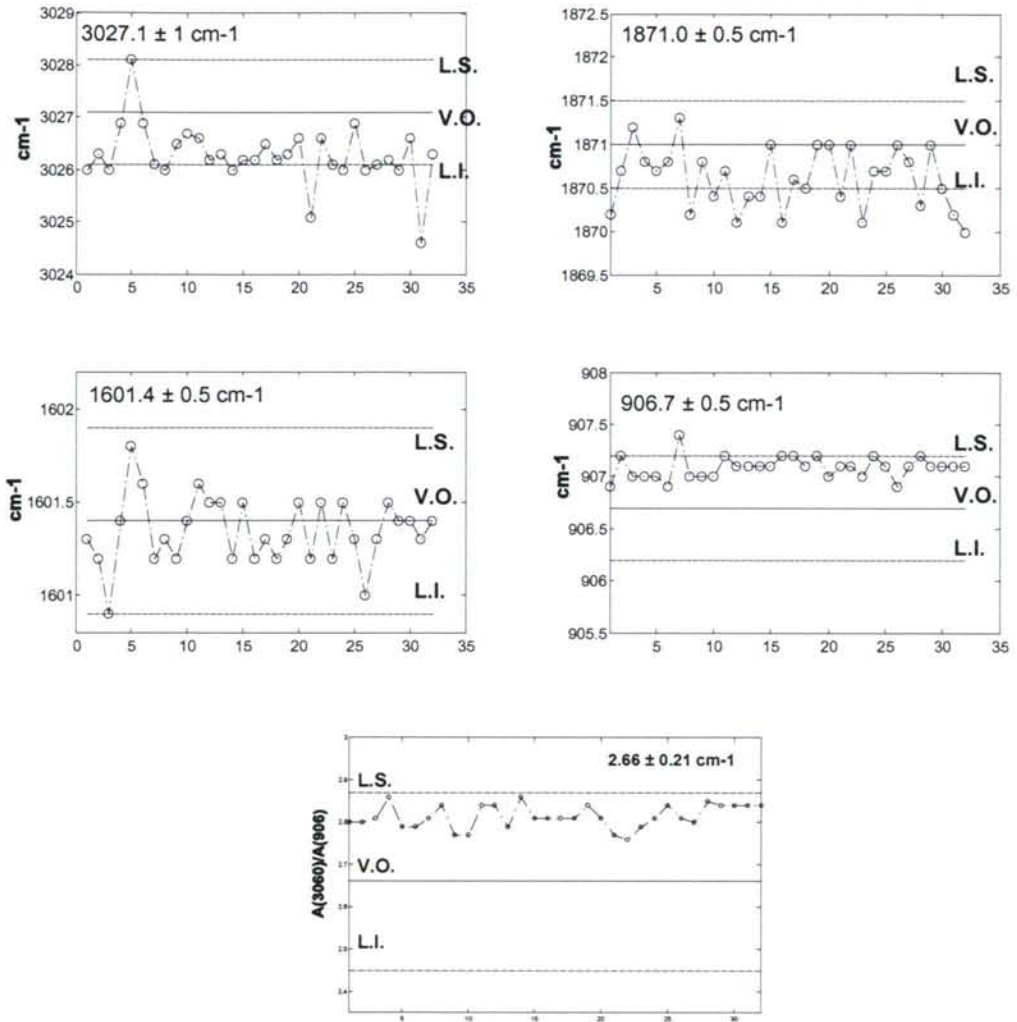


Figura 4: Medida de la exactitud y relación de la absorbancia de los picos $A(3060)/A(906)$. L.S.= límite superior, V.O.= valor objetivo, L.I.= límite inferior (LS y LI corresponden a los intervalos de confianza).

h- Precisión:

a) Repetibilidad (r): es una medida de la precisión de un sistema cuando se utiliza consecutivamente en una misma sesión de trabajo, por el mismo operador y empleando la misma muestra. Se mide como la diferencia máxima que puede existir entre dos medidas realizadas consecutivamente en las condiciones anteriores. Para estimarla se parte de la desviación típica "a corto

plazo" y se transforma según:

$$r = 2\sqrt{2} * SD_{\text{corto}} \quad \text{donde } SD_{\text{corto}} \text{ es la desviación típica determinada a partir de varias medidas realizadas en una sesión de trabajo.}$$

En esta Memoria la repetibilidad del equipo FTIR empleado se ha establecido en $6.4 * 10^{-4}$ u.A., a 1061 cm^{-1} , $A=0.12$ u.A.; lo que da lugar a una $r=0.53\%$.

b) Reproducibilidad (R): es una medida de la precisión de un sistema cuando se utiliza en diferentes sesiones de trabajo, por operadores diferentes y empleando la misma muestra. Se mide como la diferencia máxima que puede existir entre dos medidas realizadas en días diferentes en las condiciones anteriores. En función de estas definiciones, es evidente que la reproducibilidad debe ser mayor que la repetibilidad. Para estimarla se parte de la desviación típica "a largo plazo" y se transforma según:

$$R = 2\sqrt{2} * SD_{\text{largo}} \quad \text{donde } SD_{\text{largo}} \text{ es la desviación típica determinada a partir de varias medidas realizadas en diferentes sesiones de trabajo.}$$

En esta Memoria la reproducibilidad del equipo FTIR empleado se ha establecido en $7.4 * 10^{-4}$ u.A., a 1061 cm^{-1} , $A=0.12$; lo que da lugar a $R=0.62\%$, la cual, como se comentaba anteriormente, es mayor que la repetibilidad.

Nota: el factor 2 asume una distribución del 95% de confianza con un número muy elevado de medidas y puede sustituirse por el estadístico t de Student (en función del grado de confianza y grados de libertad experimentales).

El equipo de espectrometría Raman empleado fue el espectrómetro FT-Raman Bruker RFS 100, equipado con un láser Nd:YAG a 1064 nm y un detector de Ge, en modelo back-scattering y acumulando 25 barridos (tiempo de adquisición de 150 s) con una potencia del láser de 300 mW. En lo que se refiere a su control, antes de cada sesión de trabajo se llenaba el depósito del detector con nitrógeno líquido (operación que se repetía cada 4 horas), se verificaba la energía del láser y la situación de máxima energía para el posicionamiento de las muestras. El resto de comprobaciones se integraban en la rutina de mantenimiento seguida por el SCSIE (Servicio Central de Soporte a la Investigación Experimental) de la Universidad de Valencia.

4. BIBLIOGRAFÍA

Alia, J.M.; Edwards, H.G.M.; Garcia-Navarro, F.J.; Parras-Armenteros, J.; Sanchez-Jimenez, C.J. Application of FT-Raman spectroscopy to quality control in brick clays firing process. *Talanta* 50(2), 291-298 (1999).

Andrade, J.M.; Garrigues, S.; de la Guardia, M.; Gómez-Carracedo, M.; Prada, D. Non-destructive and clean prediction of aviation fuel characteristics through Fourier transform-Raman spectroscopy and multivariate calibration. *Anal. Chim. Acta* 482, 115-128 (2003).

Aparicio, R.; Baeten, V. Fats and oils authentication by FT-Raman. *Ocl.-oleagineux corps gras lipides* 5(4), 293-295 (1998).

ASTM E1421, Standard Practice for Describing and Measuring Performance of Fourier Transform Mid-Infrared (FT-MIR) Spectrometers Level Zero and Level One Tests, *Annual Book of ASTM Standards*, Vol. 11.02. (2004).

Ayora-Cañada, M.J.; Lendl, B. Sheath-flow Fourier transform infrared spectrometry for the simultaneous determination of citric, malic and tartaric acids in soft drinks. *Anal. Chim. Acta* 417, 41-50 (2000).

Baeten, V.; Hourant, P.; Morales, M.T.; Aparicio, R. Oil and fat classification by FT-Raman spectroscopy. *J. Agric. Food Chem.* 46(7), 2638-2646 (1998).

Barnes, D.; Dent, G. "Polystyrene films as a performance check for FT-IR spectrometers". *Spectroscopy Europe* 6 (2), 3-14, (1994).

Baucells, M.; Ferrer, N.; Lacort, G.; Roura, M. Comparación de diferentes técnicas (CIRCLE, ATR, DRIFT, transmisión) en el análisis cuantitativo por IRTF. *Química Analítica*, Vol. 10, nº 3, p. 211-219 (1991).

Bellon-Maurel, V.; Gilles, T. Process control in sugar industry by FT-IR spectrometer coupled with an ATR accessory. *Dev. Food Sci*, vol. 36, p. 11-18 (1994).

Bellon-Maurel, V.; Vallat, C.; Goffinet, D. Quantitative Analysis of Individual Sugars during Starch Hydrolysis by FT-IR/ATR Spectrometry. Part I: Multivariate Calibration Study-Repeatibility and Reproducibility. *Applied Spectroscopy*, vol. 49, nº 5, p. 556-562 (1995).

Bernabei, M.; Reda, R.; Galiero, R.; Bocchinfuso, G. Determination of total and polycyclic aromatic hydrocarbons in aviation jet fuel. *J. of Chromatography A* 985(1-2), 197-203 (2003).

Bertie, J.E.; Jones, R.N.; Keefe, C.D. Infrared Intensities of Liquids XII: Accurate Optical Constants and Molar Absorption Coefficients Between 6225 and 500 cm^{-1} of Benzene at 25°C, from Spectra Recorded in Several Laboratories. *Applied Spectroscopy* 47 (7), 891-911 (1993).

- Bertie, J.E.; Jones, R.N.; Apelblat, Y.; Keefe, C.D. Infrared Intensities of Liquids XII: Accurate Optical Constants and Molar Absorption Coefficients Between 6500 and 435 cm^{-1} of Toluene at 25°C, from Spectra Recorded in Several Laboratories. *Applied Spectroscopy* 48 (1), 127-143 (1994a).
- Bertie, J.E.; Jones, N.; Apelblat, Y. Infrared Intensities of Liquids XIII: Accurate Optical Constants and Molar Absorption Coefficients Between 4800 and 450 cm^{-1} of Chlorobenzene at 25°C from Spectra Recorded in Several Laboratories. *Applied Spectroscopy* 48 (1), 144-159 (1994b).
- Bertie, J.E.; Zhang, S.L. Infrared intensities of liquids XVII. Infrared refractive indices from 8000 to 350 cm^{-1} , absolute integrated absorption intensities, transition moments, and dipole moment derivatives of methan- d_3 -ol and methanol- d_4 at 25°C. *J. Chem. Phys.* 101(10), 8364-8379 (1994).
- Bertie, J.E.; Lan, Z.; Jones, R.N.; Apelblat, Y. Infrared Intensities of Liquids XVIII: Accurate Optical Constants and Molar Absorption Coefficients between 6500 and 800 cm^{-1} of Dichloromethane at 25°C, from Spectra Recorded in Several Laboratories. *Applied Spectroscopy* 49 (6), 840-851 (1995).
- Bohacs, G.; Ovadi, Z.; Salgo, A. Prediction of gasoline properties with near infrared spectroscopy. *J. Near Infrared Spectroscopy* 6, 341-348 (1998).
- Breitkreitz, M.C.; Raimundo, I.M.; Rohwedder, J.J.R.; Pasquini, C.; Dantas Filho, H.A.; José, G.E.; Araújo, M.C.U. Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. *Analyst* 128, 1204-1207 (2003).
- Brown, K.L.; Clark, R.J.H. Analysis of Key Anglo-Saxon manuscripts (8-11th centuries) in the British Library: pigment identification by Raman microscopy. *J. Raman Spectrosc.* 35(3), 181-189 (2004a).
- Brown, K.L.; Clark, R.J.H. Three English manuscripts post-1066 AD: pigment identification and palette comparisons by Raman microscopy. *J. Raman Spectrosc.* 35(3), 217-223 (2004b).
- Burgio, L.; Clark, R.J.H. Comparative pigment analysis of six modern Egyptian papyri and an authentic one of the 13th century BC by Raman microscopy and other techniques. *J. Raman Spectrosc.* 31(5), 395-401 (2000).
- Burgio, L.; Clark, R.J.H.; Stratoudaki, T.; Doulgeridis, M.; Anglos, D. Pigment identification in painted artworks: A dual analytical approach employing laser-induced breakdown spectroscopy and Raman microscopy. *Applied Spectroscopy* 54(4), 463-469 (2000).
- Burgio, L.; Clark, R.J.H.; Theodoraki, K. Raman microscopy of Greek icons: identification of unusual pigments. *Spectrochimica Acta A* 59(10), 2371-2389 (2003).
- Cadet, F.; Offmann, B. Direct spectroscopic sucrose determination of raw sugar cane

- juices. *J. Agric. Food Chem.* 45(1), 166-171 (1997).
- Cadet, F. Mesasurement of sugar content by multidimensional analysis and mid-infrared spectroscopy. *Talanta* 48, 867-875 (1999).
- Conley, R.T. *Espectroscopia infrarroja*. Edit. Alhambra (1979).
- Cooper, J.B.; Wise, K.L.; Welch, W.T.; Bledsoe, R.R.; Sumner, M.B. Determination of Weight Percent Oxygen in Commercial Gasoline: Comparison between FT-Raman, FT-IR, and Dispersive Near-IR Spectroscopies. *Applied Spectroscopy* 50 (7), 917 (1996).
- Cooper, J.B.; Wise, K.L.; Welch, W.T.; Sumner, M.B.; Wilt, B.K.; Bledsoe, R.R. Comparison of near-IR, Raman, and Mid-IR Spectroscopies for the Determination of BTEX in Petroleum Fuels. *Applied Spectroscopy* 51 (11), 1613 (1997).
- Chalmers, J.M.; Griffiths, P.R. (Eds.), *Handbook of Vibrational Spectroscopy*, Wiley, Chichester (2002).
- Chalmers, J.M.; Everall, N.J. FTIR, FT-Raman and chemometrics: applications to the analysis and characterisation of polymers. *Trends in Analytical Chemistry* 15, 18-25 (1996).
- Chung, W.M.; Wang, Q.; Sezerman, U.; Clarke, R.H. Analysis of Aviation Turbine Fuel Composition by Laser Raman Spectroscopy. *Applied Spectroscopy* 45(9), 1527-1532 (1991).
- Chung, H.; Ku, M-S; Lee, J-S. Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene. *Vibrational Spectroscopy* 20, 155-163 (1999).
- Chung, H.; Ku, M.S. Comparison of near-infrared, infrared, and Raman spectroscopy for the analysis of heavy petroleum products. *Applied Spectroscopy* 54(2), 239-245 (2000).
- Da Costa Filho, P.A.; Poppi, R.J. Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio. Determinação simultânea de glicose, maltose e frutose. *Quim. Nova* 25(1), 46-52 (2002).
- Davies, A.N.; McIntyre, P.; Morgan, E. Study of the use of molecular spectroscopy for the authentication of Raman spectroscopy. *Appl. Spectrosc.* 54(12), 1864-1867 (2000).
- Defernez, M.; Kemsley, E.K.; Wilson, R.H. Use of infrared spectroscopy and chemometrics for the authentication of fruit purees. *J. Agric. Food Chem.* 43(1), 109-113 (1995).
- De la Guardia, M.; Gallignani, M.; Garrigues, S. Flow-injection derivative Fourier transform infrared determination of methyl tert-butyl ether in gasolines. *Anal. Chim. Acta* 282, 543-550 (1993).

- Duarte, I.F.; Barros, A.; Delgadillo, I.; Almeida, C.; Gil, A.M. Application of FTIR Spectroscopy for the Quantification of Sugars in Mango Juice as a Function of Ripening. *J. Agric. Food Chem.* 50, 3104-3111 (2002).
- Duarte, I.F.; Barros, A.; Almeida, C.; Spraul, M.; Gil, A.M. Multivariate analysis of NMR and FTIR data as a potential tool for the quality control of beer. *J. Agric. Food Chem.* 52(5), 1031-1038 (2004).
- Dube, MA; Zheng, S; McLean, DD, Kates, M. A comparison of attenuated total reflectance-FTIR spectroscopy and GPC for monitoring biodiesel production. *J Am. Oil Chem Soc* 81 (6), 599-603 (2004).
- Endelmann, A.; Diewok, J.; Baena, J.R.; Lendl, B. High-performance liquid chromatography with diamond ATR-FTIR detection for the determination of carbohydrates, alcohols and organic acids in red wine. *Analytical and bioanalytical chemistry* 376(1), 92-97 (2003).
- Esteban-Díez, I.; González-Sáiz, J.M.; Pizarro, C. Prediction of sensory properties of espresso from roasted coffee samples by near-infrared spectroscopy. *Anal. Chim. Acta* 525, 171-182 (2004).
- Etzkorn, T.; Klotz, B.; Sørensen, S.; Patroescu, I.V.; Barnes, I.; Becker, K.H.; Platt, U. Gas-phase absorption cross sections of 24 monocyclic aromatic hydrocarbons in the UV and IR spectral ranges. *Atmospheric Environment* 33, 525-540 (1999).
- Fahrenfort, J. Attenuated Total Reflection: A New Principle for The Production of Useful Infrared Reflection Spectra of Organic Compound. *Spectrochimica Acta* 17, 698-709 (1961).
- Flecher, P.E.; Cooper, J.B.; Vess, T.M.; Welch, W.T. Remote fiber optic Raman analysis of benzene, toluene, and ethylbenzene in mock petroleum fuels using partial least squares regression analysis. *Spectrochimica Acta* 52 (10), 1235-1244 (1996).
- Flecher, P.E.; Welch, W.T.; Albin, S.; Cooper, J.B. Determination of octane numbers and Reid vapor pressure in commercial gasoline using dispersive fiber-optic Raman spectroscopy. *Spectrochimica Acta* 53(2), 199-206 (1997).
- Fodor, G.E.; Kohl, K.B. Analysis of Middle Distillate Fuels by Midband Infrared Spectroscopy. *Energy & Fuels* 7(5), 598-601 (1993).
- Fodor, G.E. Analysis of Petroleum Fuels by Midband Infrared Spectroscopy. *International Congress & Exposition, Detroit, Michigan. February 28-March 3* (1994).
- Fodor, G.E.; Kohl, K.B.; Mason, R.L. Analysis of Gasolines by FT-IR Spectroscopy. *Analytical Chemistry* 68, 23-30 (1996).
- Galignani, M.; Garrigues, S.; De la Guardia, M. Direct determination of benzene in gasoline by flow-injection Fourier transform infrared spectrometry. *Anal. Chim. Acta* 274, 267-274 (1993).

- Galignani, M.; Garrigues, S.; De la Guardia, M.; Burquera, J.L.; Burquera, M. Comparative study of different approaches for the flow-injection-Fourier transform infrared determination of toluene in gasolines. *Talanta* 41 (5), 739-745 (1994).
- Garfield, F.M. "Quality Assurance Principles for Analytical Laboratories". Edit. AOAC, 3ª Edición (2000).
- Garrigues, S.; Andrade, J.M.; de la Guardia, M.; Prada, D. Multivariate calibration in Fourier Transform infrared spectrometry for prediction of kerosene properties. *Anal. Chim. Acta* 317, 95-105 (1995).
- Garrigues, S.; Rambla, F.J.; De La Guardia, M. Comparative study of reflectance cells for PLS-FTIR determination of sugars in soft drinks. *Fresenius J. Anal. Chem.* 362, 137-140 (1998).
- Garrigues, J.M.; Akssira, M.; Rambla, F.J.; Garrigues, S.; De La Guardia, M. Direct ATR-FTIR determination of sucrose in beet root. *Talanta* 51, 247-255 (2000).
- Grdadolnik, J. ATR-FTIR spectroscopy: its advantages and limitations. *Acta Chim. Slov.* 49, 631-642 (2002).
- Günzler, H.; Gremlich, H.-U. IR Spectroscopy. An Introduction. Wiley-VCH (2002).
- Hakuli, A.; Kytökiivi, A.; Lakomaa, E.-L.; Krause, O. FT-IR in the Quantitative Analysis of Gaseous Hydrocarbon Mixtures. *Anal. Chem.* 67, 1881-1886 (1995).
- Harrick, N.J. Surface Chemistry from Spectral Analysis of Totally Internally Reflected Radiation. *Journal of Physical Chemistry* 64, 1110-1114 (1960).
- Innawong, B.; Zulick, D.L. The determination of frying oil quality using Fourier transform infrared attenuated total reflectance. *Acta Chromatographica* 6, 7-13 (1996).
- Innawong, B; Mallikarjunan, P; Irudayaraj, J, Marcy, JE. The determination of frying oil quality using Fourier transform infrared attenuated total reflectance. *Lebensm-Wiss Technol* 37(1), 23-28 (2004).
- Iñón, F.A.; Garrigues, J.M.; Garrigues, S.; Molina, A.; de la Guardia, M. Selection of calibration set samples in determination of olive oil acidity by partial least squares-attenuated total reflectance-Fourier transform infrared spectroscopy. *Anal. Chim. Acta* 489, 59-75 (2003).
- Iñón, F.A.; Garrigues, S.; de la Guardia, M. Nutritional parameters of commercially available milk samples by FTIR and chemometric techniques. *Anal. Chim. Acta* 513, 401-412 (2004).
- Iob, A.; Ali, M.A.; Tawabini, B.S.; Abbas, N.M. Hydrocarbon group (PONA) analysis of reformat by FT-IR spectroscopy. *Fuel* 75 (9), 1060-1064 (1996).
- Irudayaraj, J.; Tewari, J. Simultaneous monitoring of organic acids and sugars in fresh and processed apple juice by Fourier transform infrared-attenuated total reflectation spectroscopy. *Applied Spectroscopy* 57(12), 1599-1604 (2003).

- Jager, M.J.; McClintic, D.P.; Tilotta, D.C. Measurement of petroleum fuel contamination in water by solid-phase microextraction with direct Raman spectroscopic detection. *Applied Spectroscopy* 54(11), 1617-1623 (2000).
- Jakusch, M.; Mizaiakoff, B.; Kellner, R.; Katzir, A. Towards a remote IR fiber-optic sensor system for the determination of chlorinated hydrocarbons in water. *Sensors and Actuators B* 38-39, 83-87 (1997).
- Jemison, H.B.; Burr, B.L.; Davison, R.R.; Bullin, J.A.; Glover, C.J. Application and use of the ATR, FT-IR Method to asphalt aging Studies. *Fuel Science and Technology Int'l* 10 (4-6), 795-808 (1992).
- Kelly, J.F.D.; Downey, G.; Fouratier, V. Initial study of honey adulteration by sugar solutions using midinfrared (MIR) spectroscopy and chemometrics. *J. Agric. Food Chem.* 52(1), 33-39 (2004).
- Kemsley, E.K.; Holland, J.K.; Defernez, M.; Wilson, R.H. Detection of adulteration of raspberry purees using infrared spectroscopy and chemometrics. *J. Agric. Food Chem.* 44(12), 3864-3870 (1996).
- Kupina, S.A.; Shrikhande, A.J. Evaluation of a Fourier transform infrared instrument for rapid quality-control wine analyses. *American Journal of Enology and Viticulture* 54(2), 131-134 (2003).
- Le Thanh, H.; Lendl, B. Sequential injection Fourier transform infrared spectroscopy for the simultaneous determination of organic acids and sugars in soft drinks employing automated solid phase extraction. *Anal. Chim. Acta* 422, 63-69 (2000).
- Macho, S.; Boqué, R.; Larrechi, M.S.; Rius, X. Multivariate determination of several compositional parameters related to the content of hydrocarbon in naphtha by MIR spectroscopy. *Analyst* 124, 1827-1831 (1999).
- Macho, S.; Larrechi, M.S. Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. *Trends in Analytical Chemistry* 21(12), 799-806 (2002).
- Marigheto, N.A.; Kemsley, E.K.; Defernez, M.; Wilson, R.H. A Comparison of mid-infrared and Raman spectroscopies for the authentication of edible oils. *J. Of the American oil Chemists' society* 75(8), 987-992 (1998).
- Marquardt, B.J.; Wold, J.P. Raman analysis of fish: a potential method for rapid quality screening. *Lebensm-Wiss Technol.* 37(1), 1-8 (2004).
- McQueen, D.H.; Wilson, R.; Kinnunen, A.; Jensen, E.P. Comparison of two infrared spectroscopic methods for cheese analysis. *Talanta* 42, 2007-2015 (1995).
- Morcillo, J.; Madroñero, R. Aplicaciones prácticas de la espectroscopia infrarroja. Edit. Universidad de Madrid (Facultad de Ciencias) (1962).
- Moreira, J.L.; Santos, L. Spectroscopic interferences in fourier transform infrared wine analysis. *Anal. Chim. Acta* 513 (1), 263-268 (2004).

- Mosier-Boss, P.A.; Lieberman, S.H. Detection of nitrate and sulfate anions by normal Raman spectroscopy and SERS of cationic-coated, silver substrates. *Applied Spectroscopy* 54(8), 1126-1135 (2000).
- NIST, Certified Polystyrene (2004).
- Osborne, B.G.; Fearn, T.; Hindle, P.H., Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, Longman, Harlow, Essex, UK (1993).
- Pappas, D.; Smith, B.W.; Winefordner, J.D. Raman spectroscopy in bioanalysis. *Talanta* 51(1), 131-144 (2000).
- Paradkar, M.M.; Sivakesava, S.; Irudayaraj, J. Discrimination and classification of adulterants in maple syrup with the use of infrared spectroscopic techniques. *Journal of the Science of Food and Agriculture* 83(7), 714-721 (2002a).
- Paradkar, M.M.; Sakhamuri, S.; Irudayaraj, J. Comparison of FTIR, FT-Raman, and NIR spectroscopy in a maple syrup adulteration study. *Journal of Food Science* 67(6), 2009-2015 (2002b).
- Paradkar, M.M.; Sivakesava, S.; Irudayaraj, J. Discrimination and classification of adulterants in maple syrup with the use of infrared spectroscopic techniques. *Journal of the Science of Food and Agriculture* 82(5), 497-504 (2002c).
- Parisi, A.F.; Nogueiras, L.; Prieto, H. On-line determination of fuel quality parameters using near-infrared spectrometry with fibre optics and multivariate calibration. *Anal. Chim. Acta* 238, 95-100 (1990).
- Patz, C.-D.; Blicke, A.; Ristow, R.; Dietrich, H. Application of FT-MIR spectrometry in wine analysis. *Anal. Chim. Acta* 513, 81-89 (2004).
- Piccolo, A. Advanced Infrared Techniques (FT-IR, DRIFT, and ATR) Applied to Organic and Inorganic Soil Materials. *Congr. Soil Sci 15th, vol 3a, p 3-22. International Society of Soil Science* (1994).
- Pillonel, L.; Luginbühl, W.; Picque, D.; Schaller, E.; Tabacchi, R.; Bosset, J. Analytical methods for the determination of the geographic origin of Emmental cheese: mid- and near-infrared spectroscopy. *European Food Research and Technology* 216(2), 174-178 (2003).
- Pretsch, E.; Clerc, T.; Seibl, J.; Simon, W. Tablas para la determinación estructural por métodos espectroscópicos. Springer-Verlag Ibérica, Barcelona (1998).
- Rambla, F.J.; Garrigues, S.; Ferrer, N.; De La Guardia, M. Simple partial least-squares-attenuated total reflectance Fourier transform infrared spectrometric method for the determination of sugars in fruit juices and soft drinks using aqueous standards. *Analyst* 123, 277-281 (1998).
- Robinson, J.W. Practical Handbook of Spectroscopy. CRC Press Inc. (1991).
- Ródenas-Torralba, E.; Ventura-Gayete, J.; Morales-Rubio, A.; Garrigues, S.; De la

- Guardia, M. Multicommutation Fourier transform infrared determination of benzene in gasoline. *Anal. Chim. Acta* 512, 215-221 (2004).
- Rodriguez-Saona, L.E.; Fry, F.S.; McLaughlin, M.A. Rapid analysis of sugars in fruit juices by FT-NIR spectroscopy. *Carbohydrate research* 336, 63-74 (2001).
- Roeges, N.P.G. A Guide to the complete interpretation of infrared spectra of organic structures. John Wiley & Sons Ltd. (1994).
- Schindler, R.; Vonach, R.; Lendl, B.; Kellner, R. A rapid automated method for wine analysis based upon sequential injection (SI)-FTIR spectrometry. *Fresenius Journal of Analytical Chemistry* 362(1), 130-136 (1998).
- Schmidt, T.C.; Duong, H-A; Berg, M.; Haderlein, S.B. Analysis of fuel oxygenates in the environment. *Analyst* 126, 405-413 (2001).
- Seasholtz, M.B.; Archibald, D.D.; Lorber, A.; Kowalski, B.R. Quantitative Analysis of Liquid Fuel Mixtures with the use of Fourier Transform Near-IR Raman Spectroscopy. *Applied Spectroscopy* 43(6), 1067-1072 (1989).
- Sinnaeve, G.; Dardenne, P.; Agneessens, R.; Lateur, M., Hallet, A. Quantitative analysis of raw apple juices using near infrared, Fourier-transform near infrared and Fourier-transform infrared instruments: a comparison of their analytical performances. *J. Near Infrared Spectrosc.* 5 (1), 1-17 (1997).
- Sivakesava, S.; Irudayaraj, J. Determination of sugars in aqueous mixtures using mid-infrared spectroscopy. *Applied Engineering in Agriculture* 16(5), 543-550 (2000).
- Sivakesava, S.; Irudayaraj, J. Prediction of Inverted Cane Sugar Adulteration of Honey by Fourier Transform Infrared Spectroscopy. *Journal of Food Science* 66(7), 972-978 (2001).
- Sivakesava, S.; Irudayaraj, J.M.K.; Korach, R.L. Detection of adulteration in apple juice using mid infrared spectroscopy. *Applied Engineering in Agriculture* 17(6), 815-820 (2001).
- Skoog, D.A.; Leary, J.J. Análisis instrumental (4ª Edición). McGraw-Hill (1994).
- Skoulíka, S.G.; Georgiou, C.A.; Polissiou, M.G. FT-Raman spectroscopy-analytical tool for routine analysis of diazinon pesticide formulations. *Talanta* 51(3), 599-604 (2000).
- Smith, A.L. Applied Infrared Spectroscopy: Fundamentals, Techniques, and Analytical Problem-solving. *Chemical Analysis, vol. 54. Ed. John Wiley & Sons* (1979).
- Smith, A. L. "Practical Handbook of Spectroscopy". Edit by J.W. Robinson, CRC Press (1991).
- Smith, B.C. Fundamentals of Fourier Transform Infrared Spectroscopy. Ed. CRC Press (1996).
- Tapp, H.S.; Defernez, M.; Kemsley, E.K. FTIR spectroscopy and multivariate analysis

- can distinguish the geographic origin of extra virgin olive oils. *J. Agric. Food Chem.* 51(21), 6110-6115 (2003).
- Tay, A.; Singh, R.K.; Krishnan, S.S.; Gore, J.P. Authentication of olive oil adulterated with vegetable oils using Fourier transform infrared spectroscopy. *Lebensmittel-Wissenschaft und-Technologie-Food Science and Technology* 35(1), 99-103 (2002).
- Tewari, J.; Joshi, M.; Gupta, A.; Mehrotra, R.; Chandra, S. Determination of Sugars and Organic Acid Concentration in Apple Juices Using Infrared Spectroscopy. *Journal of Scientific & Industrial Research* 58, 19-24 (1999).
- Tewari, J.; Irudayaraj, J. Quantification of saccharides in multiple floral honeys using Fourier transform infrared microattenuated total reflectance spectroscopy. *J. Agric. Food Chem.* 52(11), 3237-3243 (2004).
- Urtubria, A.; Pérez-Correa, J.R.; Meurens, M.; Agosin, E. Monitoring large scale wine fermentations with infrared spectroscopy. *Talanta* 64(3), 778-784 (2004).
- Wilson, R.H.; Tapp, H.S. Mid-infrared spectroscopy for food analysis: recent new applications and relevant developments in sample presentation methods. *Trends in Analytical Chemistry* 18(2), 85-93 (1999).
- Wojciechowski, C; Dupuy, N; Ta, CD, Huvenne, JP; Legrand, P. Quantitative analysis of water-soluble vitamins by ATR-FTIR spectroscopy. *Food Chem.* 63(1), 133-140 (1998).
- Yang, H.; Irudayaraj, J. Rapid determination of vitamin c by NIR, MIR and FT-Raman techniques. *J. Pharm. Pharmacol.* 54(9), 1247-1255 (2002).
- Yang, H.S.; Lewis, I.R.; Griffiths, P.R. Raman spectrometry and neural networks for the classification of wood types. 2. Kohonen self-organizing map. *Spectrochimica Acta A* 55(14), 2783-2791 (1999).
- Zhang, S.L.; Michaelian, K.H.; Bulmer, J.T.; Hall, R.H.; Hellman, J.L. Fourier transform Raman spectroscopy of fuels: curve-fitting of C-H stretching bands. *Spectrochimica Acta A* 52(11), 1529-1540 (1996).
- Zuo, J.; Xu, C.Y.; Wang, C.S.; Yushi, Z. Identification of the pigment in painted pottery from the Xishan site by Raman microscopy. *Journal of Raman Spectroscopy* 30(12), 1053-1055 (1999a).
- Zuo, J.A.; Wnag, C.S.; Xu, C.Y.; Qin, P.; Xu, G.J.; Zhao, H.B. Raman microscopy study of the pigments on the ancient wall painting from the large grave in Wanzhang (Hebel, China). *Spectroscopy Letters* 32(5), 841-850 (1999b).



Capítulo IV

Análisis de queroseno

Objetivo:

En este capítulo se desarrollan procesos analíticos rápidos, sencillos de aplicar, fiables y con el mínimo consumo de reactivos químicos, que permitan llevar a cabo la predicción de múltiples propiedades del queroseno empleando diferentes sistemas de medida en fase gas y líquida. Para ello se combinará la espectroscopia FTIR-ATR con técnicas quimiométricas, en concreto la técnica de mínimos cuadrados parciales (PLS).

Índice:

Parte A.- Aproximación al producto objeto del estudio

- 1. Introducción al combustible de aviación*
 - 1.1. Composición*
 - 1.2. Uso, tipos y propiedades del queroseno*
- 2. Consideraciones para la salud*
 - 2.1. Exposición al queroseno*
 - 2.2. Rutas de exposición*

Parte B.- Parte experimental

- 1. Optimización de las metodologías alternativas de análisis*
 - 1.1. Medida de querosenos en fase gas*
 - 1.2. Medida de querosenos en fase líquida*

Parte C.- Resultados y discusión

- 1. Modelos multivariantes en fase gas*
 - 1.1. Modelos multivariantes con el sistema de medida simple*
 - 1.2. Modelos multivariantes con el sistema de medida complejo*
 - 1.3. Exactitud y precisión*
- 2. Modelos multivariantes en fase líquida*
 - 2.1. Modelos multivariantes con el sistema de medida FTMIR-ATR*
 - 2.2. Modelos multivariantes con el sistema de medida Raman*

Parte D.- Bibliografía

PARTE A.- APROXIMACIÓN AL PRODUCTO OBJETO DEL ESTUDIO

1. INTRODUCCIÓN AL COMBUSTIBLE DE AVIACIÓN

Uno de los productos obtenidos del refino del crudo de petróleo que más destacan por su aplicación es el queroseno ó combustible para aviación a reacción. Físicamente es un líquido amarillo transparente, con olor penetrante y que se obtiene como una fracción intermedia entre la destilación de la gasolina y la del gasóleo. Queroseno (keroseno) es el nombre genérico que se aplica a un grupo de destilados medios derivados del petróleo con punto de ebullición entre 145º y 300º C (*www.concawe.be, 2002a*).

El queroseno se obtiene mediante refino del crudo de petróleo en las refinerías. Una refinería se puede definir como un complejo industrial donde el crudo de petróleo se somete a un proceso de destilación o separación física y, luego, a procesos químicos que permiten extraerle buena parte de la gran variedad de componentes que contiene. Las refinerías son muy distintas unas de otras según las tecnologías y los esquemas de proceso que se utilicen, así como según su capacidad. Unas, procesan petróleos ligeros, otras petróleos pesados o mezclas de ambos y, por consiguiente, los productos que se obtienen varían de una a otra (*www.monografias.com, 2002*).

El proceso de refino se lleva a cabo en varias etapas, por lo que una refinería dispone de numerosas torres de destilación, unidades de tratamiento, equipos auxiliares y tuberías. En términos simplificados, el funcionamiento de una refinería se indica a continuación.

En primer lugar se lleva a cabo el refino del crudo de petróleo en las torres de "destilación primaria" o "destilación atmosférica". En su interior, las torres operan a una presión cercana a la atmosférica y están divididas en numerosos compartimentos que se denominan "bandejas" o "platos". Cada bandeja tiene una temperatura diferente y cumple la función de recoger los hidrocarburos líquidos aprovechando sus diferentes puntos de ebullición/condensación.

El crudo llega a estas torres después de pasar por un horno, donde se calienta a temperaturas de hasta 400 °C que lo convierten en vapor. Los vapores entran por la parte inferior de la torre de destilación y ascienden por entre las bandejas. A medida que suben pierden calor y se enfrían. Cuando cada componente vaporizado alcanza

una temperatura inferior a la característica de ebullición, condensa y deposita en su respectiva bandeja, a la cual están conectados conductos por los que se recogen las distintas corrientes que se separaron en esta etapa. Al fondo de la torre cae el "crudo reducido", es decir, aquel que no alcanzó a evaporarse en esta primera etapa. Se cumple así el primer paso del refino. De arriba hacia abajo se obtienen, en orden: gas de refinería, gas licuado de petróleo (GLP), nafta, queroseno, diesel, gasoil pesado y un residuo que corresponde a los compuestos más pesados que no llegaron a evaporarse. Algunos de los productos, como jet-A, queroseno y diesel, son productos ya finales.

Las demás corrientes se envían a otras torres y unidades para someterlas a nuevos procesos, al final de los cuales se obtendrán los demás derivados del petróleo. En dichos procesos, el "petróleo" se separa en fracciones o destilados que requieren un tratamiento posterior para la obtención de los productos demandados por el mercado. Algunos de los tratamientos secundarios o de depuración son los siguientes: procesos de separación (por ejemplo la destilación a vacío utilizada en la mayoría de las refinerías, en donde se separan nuevamente los productos como consecuencia de los diferentes punto de ebullición pero a una presión muy reducida), procesos de purificación (por ejemplo el proceso dulcificante o hidrodesulfuración, en donde se emplean reacciones químicas para eliminar los compuestos presentes en cantidades traza que dan al material una mala calidad (mercaptanos y otros derivados del azufre) y los procesos de conversión (cambian la estructura larga de las moléculas, normalmente mediante craqueo, en estructuras pequeñas).

Brevemente, algunos de los procesos de conversión son:

La Unidad de Craqueo Catalítico o Cracking; recibe gasóleos y crudos reducidos para producir (fundamentalmente) gasolina y gas propano. Las unidades de Recuperación de Vapores reciben los gases de las demás plantas y obtienen gas combustible, gas propano, propileno y butanos. La planta de mezclas (*blending*) es la que recibe las distintas corrientes de naftas para obtener la gasolina de automoción. La unidad de aromáticos produce a partir de las naftas: tolueno, xilenos, benceno, ciclohexano y otros productos con usos petroquímicos. La unidad de Parafinas recibe destilados parafínicos y nafténicos para sacar parafinas y bases lubricantes.

De todo este proceso también se obtienen azufre, asfaltos, fuel-oil y coque de petróleo, que son los últimos productos que se pueden obtener del petróleo.

En resumen, el principal producto (económicamente hablando) que sale del refino del petróleo es la gasolina. El volumen de gasolina que cada refinería obtiene es el resultado del esquema que utilice. En promedio, cada barril de petróleo que entra a una refinería se convierte en un 40 - 50 por ciento de gasolina.

En la **Figura 1** se muestra el proceso de obtención de los principales productos de petróleo usados hoy en día.

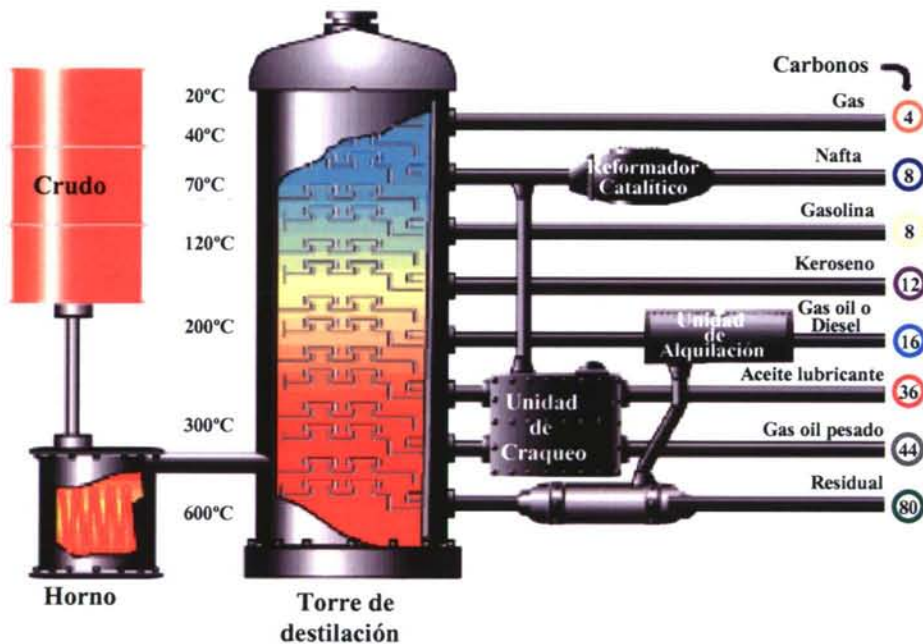


Figura 1: Principales productos obtenidos por destilación del crudo de petróleo, entre los que se encuentra el queroseno (<http://auto.howstuffworks.com/oil-refining5.htm>, 2003).

1.1. COMPOSICIÓN

La composición química detallada del queroseno depende de la naturaleza de los crudos de petróleo empleados en los procesos de refinería, ya que se obtiene a partir de la destilación atmosférica. Está compuesto de mezclas complejas de hidrocarburos alifáticos (parafinas, C_nH_{2n+2}) y naftenos (o ciclo parafinas, C_nH_{2n}). Al menos el 70% en volumen de queroseno está formado por alcanos, cicloalcanos, alquenos (los cuales le dan su típico olor), aromáticos y aromáticos policíclicos (principalmente

alquilbenzenos y alquilnaftalenos que no exceden del 25% en volumen) con 9 a 16 átomos de carbono predominantemente (www.chevron.com, 2002a; www.concawe.be, 2002a; McKay, 2001; <http://193.51.164.11/htdocs/>, 2002; Mattie et al., 1991; www.metrokc.gov, 2001; ATSDR, 1993; Smith et al. 1997; Liu et al, 1999).

Las olefinas (que no deben exceder del 5% en volumen) son constituyentes indeseables de los querosenos ya que son relativamente inestables y pueden causar formación de goma cuando se queman. Las concentraciones de benceno (punto de ebullición de 80°C) y n-hexano (punto de ebullición de 69°C) son siempre inferiores al 0.01% de masa debido a que el rango de destilación típico de los querosenos es de 145 a 300°C (www.concawe.be, 2002a).

Además de mezclas de los componentes indicados, los querosenos algunas veces reciben naftas y bajas concentraciones de aditivos (generalmente menos del 0.1% (p/p)) para obtener una mejor estabilidad y un mejor resultado en su utilización final. Los aditivos típicos incluyen antioxidantes, desactivadores de metal, inhibidores de corrosión, aditivos antihielo, disipadores estáticos y biocidas (www.metrokc.gov, 2001; ATSDR, 1995; ATSDR, 1998a; ATSDR, 1998b; www.chevron.com, 2002a).

Todos los querosenos tienen que estar libres de sólidos en suspensión y, particularmente, agua. Adicionalmente, tienen que ser bombeables a muy bajas temperaturas y deben ser estables a temperaturas muy altas, así como actuar como medio de transferencia de calor para lubricantes y fluidos hidráulicos.

El agua puede estar de tres formas diferentes en los querosenos: disuelta en el combustible, como una fase líquida que se puede separar (agua libre), y como una emulsión de agua y combustible. El agua disuelta no es un problema, pero el agua libre y las emulsiones de agua pueden ser potencialmente peligrosas, ya que desestabilizan el combustible y tienen que ser eliminadas (www.chevron.com, 2002a).

1.2. USO, TIPOS Y PROPIEDADES DEL QUEROSENO

El uso más habitual del queroseno es como combustible de la aviación a reacción, como combustible de calentamiento industrial o doméstico y como componente de insecticidas, productos de limpieza, pesticidas, etc. Se utilizaba mucho como combustible en los quinqués y en otras lámparas. Por ello era conocido como "petróleo de quemar" (www.concawe.be, 2002a; www.concawe.be, 2002b).

En términos generales, hay tres tipos de querosenos (www.concawe.be, 2002a;

www.concawe.be, 2002b):

- Tipo queroseno, normalmente combinando los querosenos.
- Tipo “corte ancho”, en el que los querosenos son mezclados con naftas de bajo punto de deflagración.
- Tipo queroseno de alto punto de deflagración, normalmente mezclados con querosenos que tienen un mínimo punto de deflagración de 60°C.

Los nombres comunes usados para querosenos (o “*jet fuels*”) se resumen como se indica a continuación:

- ✓ Jet A-1: se usa en aviación civil, con un máximo de 25% (v/v) de contenido en hidrocarburos aromáticos. Punto de congelación máximo de 47°C “bajo cero”.
- ✓ Jet A: igual que el Jet A-1, pero con punto de congelación máximo de 40°C “bajo cero”.
- ✓ Jet B: tipo “corte ancho” usado en aviación civil, con máximo del 25% (v/v) de contenido de hidrocarburos aromáticos.
- ✓ JP- 4: tipo “corte ancho” usado en aviación militar, con contenido máximo de hidrocarburos aromáticos del 25% (v/v).
- ✓ JP-5: tipo queroseno de alto punto de deflagración usado en aviación naval, con contenido máximo de hidrocarburos aromáticos del 25%(v/v).
- ✓ JP-8: tipo queroseno usado en aviación militar, con contenido máximo de hidrocarburos aromáticos del 25% (v/v).

El queroseno puede ser peligroso si no es manipulado apropiadamente. Aunque no es un líquido inflamable sí puede explosionar (especialmente en presencia de chispas producidas por la acumulación de carga electrostática) y quemarse rápidamente. Las normas de la N.F.P.A. (Asociación Nacional de Protección contra el Fuego, EE.UU) clasifican a los líquidos en **inflamables** (punto de inflamación menor a 38°C, pueden inflamarse a temperatura ambiente) y **combustibles** (punto de inflamación superior a 38°C, requieren calentamiento previo para incendiarse) (*www.concawe.be, 2002b; www.chevron.com, 2002a; http://library.cbest.chevron.com, 2001*). La deflagración se produce tras una evaporación más o menos rápida en la cual se quema un vapor, que puede incendiarse y quemarse en presencia de aire produciendo una

explosión violenta (<http://library.cbest.chevron.com>, 2001; www.chevron.com, 2002a).

Dados el amplísimo uso comercial del producto, la peligrosidad de manipulación y los riesgos para la salud durante la exposición, es evidente la importancia de evaluar su calidad, no sólo mediante procedimientos “clásicos” sino con otros más actuales que mejoren la productividad del laboratorio (Caswell et al., 1989; Fodor and Kohl, 1993; Fodor, 1994; Garrigues et al., 1995; Litani-Barzilai et al., 1997; Fodor et al., 1996; Job et al., 1996; Macho et al., 1999; Collette, 1997).

Las principales características de calidad vienen definidas en las especificaciones del *British Ministry of Defence* (U.K.) (DERD 2494, 1999) y las guías ASTM (ASTM D 86, ASTM D 445, ASTM D 1319, ASTM D 2386). Ambas, plantean un mínimo de 26 tests analíticos para llevar a cabo en los laboratorios. Los “más importantes” en la valoración de la calidad de los querosenos son: punto de deflagración, densidad, punto inicial de destilación, punto final de destilación, pérdida de destilación, punto de congelación, conductividad, acidez total, contenido en olefinas, punto de humo e índice de separación de agua.

En esta Memoria se estudiarán 8 variables por su importancia en la manipulación del queroseno:

- Punto de deflagración (*flash point*, según el método Abel (IP-170)): Es la temperatura mínima a la que el producto líquido produce vapores en concentraciones tales que pueden inflamarse en contacto con una fuente de ignición (llama, chispas, etc.). Cuanto menor sea el punto de inflamación de un producto, más inflamable y por esto, más peligroso será éste.

- Punto de congelación (*freezing point*) (ASTM D 2386): Es la temperatura por debajo de la cual, el líquido pasa a sólido.

- Punto inicial de destilación, 10% de destilación, 90% de destilación y punto final de destilación (ASTM D 86): Salvo productos que tienen cierto grado de pureza (hexano, tolueno o xileno) los distintos derivados del petróleo son mezclas de hidrocarburos, por lo que no presentan un punto de ebullición definido sino un rango de destilación más o menos amplio, de acuerdo a la especificación de cada producto. A grandes rasgos, a menor punto inicial de destilación corresponde una mayor volatilidad del producto (producto ligero) y, por lo tanto, una mayor peligrosidad, tanto desde el punto de vista del riesgo de incendio como de la generación de vapores tóxicos. Otra medida de la volatilidad es la presión de vapor, siendo el producto tanto

más peligroso cuanto mayor sea ésta.

- Porcentaje de aromáticos (ASTM D 1319). Se suele emplear este parámetro para evaluar la existencia de fracciones “pesadas” ó hidrocarburos insaturados en el queroseno. Es una medida de “pureza”.

- Viscosidad (ASTM D 445): es la medida de la resistencia del líquido a fluir sobre presión generada por gravedad o una fuente mecánica. También se relaciona con la composición del producto: a menor viscosidad, mayor contenido de fracciones ligeras.

En la **Tabla 1** se presentan las variables a estudiar y los valores de precisión (repetibilidad y reproducibilidad) asociados a su determinación así como las guías de referencia.

Según ASTM, la repetibilidad (r) se define como la diferencia máxima que puede existir entre dos resultados de ensayo obtenidos consecutivamente por el mismo operador con el mismo aparato, con las mismas condiciones de operación e idéntico material de ensayo (95% de confianza). Análogamente, la reproducibilidad (R) se define como la diferencia máxima que puede existir entre dos resultados sencillos e independientes obtenidos por diferentes operadores trabajando en distintos laboratorios y con idéntico material de ensayo (95% de confianza).

Parámetro	Guía	r	SD_r	R	SD_R
Flash Point (°C)	IP 170	1.0	0.36	1.5	0.54
Freezing (°C)	ASTM D 2386	1.0	0.36	2.5	0.89
PI (°C)	ASTM D 86	3.5	1.25	8.5	3.04
PD10 (°C)	ASTM D 86	4.3	1.54	8.8	3.14
PD90 (°C)	ASTM D 86	4.1	1.46	8.8	3.14
FBP (°C)	ASTM D 86	3.5	1.25	10.5	3.75
% aromáticos (% v/v)	ASTM D 1319	1.3	0.46	2.7	0.96
Viscosidad (cSt)	ASTM D445	0.02	0.007	0.04	0.01

Tabla I: Repetibilidad (r) y reproducibilidad (R) de las propiedades a estudiar según se indica en las guías oficiales. SD_r y SD_R son las respectivas desviaciones típicas estimadas a partir de ellas (a través del factor de 2.8).

2. CONSIDERACIONES PARA LA SALUD

En condiciones industriales normales de almacenamiento, manipulación y uso, los querosenos no presentan un riesgo para la salud. En uso doméstico, hay un riesgo significativo de daño pulmonar cuando se aspira el líquido a los pulmones.

2.1. EXPOSICIÓN AL QUEROSENO

En el caso concreto de la exposición ocupacional a los combustibles de aviación, ésta se produce en los trabajos indicados de forma resumida en la **Tabla II** (www.concawe.be, 2002b). También se describen los grupos de trabajo que tienen riesgo de exposición ocupacional a queroseno, las tareas de exposición y sus medidas de control. En todos los casos se da por supuesto que se toman todas las medidas posibles para evitar accidentes, explosión, etc.

- Fabricación y distribución de queroseno, lo que comprende una variedad de trabajos: desde producción y operaciones relacionadas con la refinería a la distribución de los productos (*Kalabokas et al.*, 2001).

- Fabricación de productos formulados que contienen queroseno, como agentes de limpieza, desengrasado, pinturas, barnices, herbicidas, insecticidas y pesticidas.

- Uso industrial de productos formulados.

REFINERÍA	
Operarios in-situ	Operaciones en las plantas de producción de queroseno, por ejemplo, válvula de operación, colección de muestras, drenaje de contenedores y conductos para el mantenimiento.
Operarios en el exterior	Operaciones auxiliares llevadas a cabo por trabajadores de la refinería, por ejemplo, técnicos de laboratorio (control de calidad/investigación), actividades del tanque de almacenamiento exterior (lavado/muestreo/descarga del agua del fondo del tanque), limpieza de recipientes de muestra, etc
Trabajador de mantenimiento	Lleva a cabo una variedad de tareas que pueden suponer una exposición a queroseno líquido y vapor cuando se vacía, limpia, abre y se trabaja con equipos en un sitio cerrado.
Limpieza de tanque	Limpieza de los tanques de almacenamiento.
DISTRIBUCIÓN	
Conductor de camión cisterna	Carga y descarga del camión cisterna. Test para mirar el contenido de agua. La carga de asfalto de viscosidad reducida se hace con adición de un diluyente volátil (potencial exposición a vapor de queroseno).

Operadores de carga de los vagones del ferrocarril	Operadores que llevan a cabo la carga de los vagones.
Operarios de los vagones del ferrocarril (no de carga)	Operarios que hacen la conexión/desconexión de manguera y muestreo. Nota: esta operación puede formar parte de las tareas de los Operadores de Terminal.
Personal de muelle	Supervisión de las operaciones de carga, muestreo, limpieza de tanque, manipulación de mangueras.
Colla de descarga (buques)	Carga de petroleros usando mangueras flexibles y ventilación por medio de válvulas de seguridad del tanque de carga. Tareas que incluyen conexión/desconexión de tubería de carga, chequeo de los niveles de llenado del tanque, limpieza del tanque.
Operarios de Terminal	Llevar a cabo tareas como los operarios del exterior de la Refinería, como en la estación de distribución de productos petroleros.
Calibradores de contadores	Prueban los contadores y reparan el equipo de medida. Conexión/desconexión de mangueras de los vehículos.
Mecánico del vehículo	Reparación de mangueras y test de presión. Uso de queroseno como desengrasante para limpiar los componentes del vehículo. Inspección del compartimento del tanque.
Llenado de envases pequeños	Llenado de, por ejemplo, contenedores de 25L, bidones de 200L, etc. Se puede llevar a cabo esta tarea con una salida a pequeña escala.
AEROPUERTOS	
Operarios	Carga de vehículos, como el camión repostador, sistema de repostaje por manguera, análisis para el agua y densidad específica.
Personal de taller	Análisis para agua y densidad específica, test de eficiencia en la separación de agua, cambio de filtros, pruebas de medida y servicio de medida.
Mecánico del vehículo	Reparación de mangueras y test de presión. Uso de queroseno como desengrasante para limpiar los componentes del vehículo. Inspección del compartimento del tanque.
Mecánico de taller	Lleva a cabo una variedad de tareas que pueden suponer una exposición a queroseno líquido y vapor cuando se vacía, limpia, abre y se trabaja en un sitio cerrado con equipos.
Limpieza de tanque	Limpieza de los tanques de almacenamiento.
RESIDENCIA DOMÉSTICA	
Suministrador de calefacción doméstica	Reemplazo de calderas, tanques de almacenamiento que pueden comprender vaciado de calderas.
Mantenimiento de calefacción doméstica	Servicio que incluye inspección visual y mecánica del sistema entero incluyendo los tanques de almacenaje, calderas y radiadores. Ocurre poco contacto o ninguno con queroseno.

Tabla II: Tareas que comprenden exposición potencial para trabajadores y medidas de control (www.concawe.be, 2002b).

El personal de pista de los aeropuertos está expuesto a inhalación de vapores y humos tanto de restos de combustible como de productos de combustión, algo que incide directamente sobre su Seguridad y Salud. Además de la seguridad del pasaje durante el transcurso del vuelo, se debe tener en cuenta la seguridad en las etapas de trasiego de los camiones cisterna al tanque de almacenaje del aeropuerto y, por supuesto, en el trasiego interno de las refinerías.

El queroseno producido en la refinería pasa por un control de calidad estricto para que cumpla todos los parámetros analíticos recogidos en la legislación. El combustible puede ser distribuido directamente al tanque de almacenamiento de un aeropuerto aunque habitualmente la distribución no es directa, ya que incluye una o más etapas intermedias de almacenamiento (terminales), como se observa en la **Figura 2**. Algunos modos de transporte pueden ser: por medio de oleoductos, barcos, vagones cisterna, petroleros o camiones cisterna, pero no todos ellos son posibles para cualquier destino.

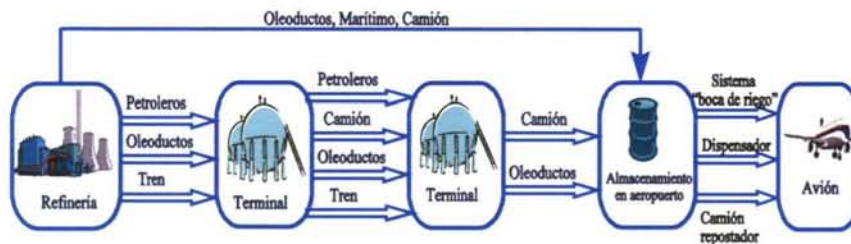


Figura 2: Sistema de distribución del combustible para turbinas de combustión.

En la **Figura 3** se muestra el mapa de oleoductos existentes en España, las refinerías y los aeropuertos existentes.

Los oleoductos son el mejor medio de distribución para el transporte de grandes cantidades de combustible. Por esta razón, los querosenos son normalmente distribuidos mediante oleoductos. La gasolina de aviación (*avgas*) normalmente es distribuida con camiones, ferrocarril o petroleros debido a la menor cantidad necesitada en los aeropuertos.

Pocas son las refinerías que están directamente conectadas a los aeropuertos mediante oleoductos específicos para combustible para turbinas de combustión. La mayoría del queroseno es distribuido por oleoductos multiproductos. Por ello, cuando el queroseno se distribuye por medio de oleoductos, corre el riesgo de contaminarse

Después del almacenamiento en los tanques del aeropuerto hay varias formas de dispensación al avión: sistema “boca de riego” o sistema de repostaje por manguera, camión repostador y equipo distribuidor de gasolinas (gasolina o queroseno).

El sistema “boca de riego” se usa en la mayoría de los grandes aeropuertos comerciales. En este sistema una red de conductos conectan los tanques de almacenamiento a cada puerta de embarque. La unidad “boca de riego” usa pequeños camiones equipados con equipo de filtración y de medida de volumen para llenar de combustible el avión. Los camiones con “boca de riego” normalmente tienen filtros/separadores de agua para introducir la mínima contaminación en los tanques de combustible del avión (**Figura 5**).

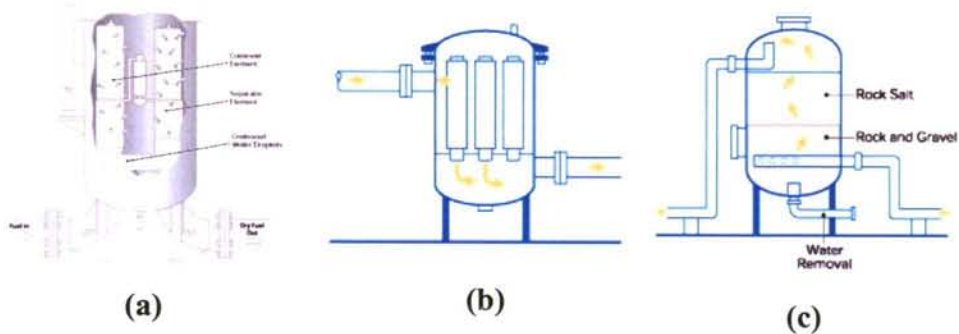


Figura 5: Purificación del queroseno mediante: (a) un filtro separador, (b) filtro de partículas, (c) desecador con sales.

El equipo distribuidor de combustible (dispensador) y el camión de repostaje se usan en aeropuertos pequeños. Los dos tienen bombeo, filtración y equipo de medida de volumen. Los camiones de repostaje llevan el combustible a los aviones en tanques propios (**Figura 6**). El dispensador es una bomba de localización fija diseñada específicamente para combustibles de aviación.



Figura 6: Trabajador llenando el tanque de un avión comercial mediante un camión repostador.

2.2. RUTAS DE EXPOSICIÓN

Las tres rutas de exposición al queroseno son: contacto, ingestión e inhalación.

2.2.1. EXPOSICIÓN POR CONTACTO

El contacto repetido o prolongado con el queroseno puede causar irritación de la piel, eritema, ampollas, enrojecimiento, hinchazón y quemazón (www.chevron.com, 2002b) e incluso dermatitis (ver **Figura 7**) (www.concawe.be, 2002a).



Figura 7: Dermatitis crónica de exposición a queroseno. La piel está muy gruesa, hiperpigmentada, seca y con fisuras (www.cdc.gov/niosh/ocderm4.html, 2002).

2.2.2. EXPOSICIÓN POR INGESTIÓN

La ingestión de querosenos no se suele encontrar en el uso industrial de los productos aunque hay muchos casos de ingestión accidental de queroseno por niños en accidentes domésticos.

El sabor y olor del queroseno normalmente limita la ingestión a pequeñas cantidades. Aunque el queroseno es de baja toxicidad oral, el vómito es una reacción común a su ingestión y esto produce un riesgo de aspiración del líquido a los pulmones. La ingestión puede también causar irritación de los labios, del tracto gastrointestinal, diarrea, tos, neumonía, somnolencia, intranquilidad, irritabilidad, convulsiones, coma e incluso la muerte (www.atsdr.cdc.gov, 2002; www.concawe.be, 2002a).

2.2.3. EXPOSICIÓN POR INHALACIÓN

En el rango normal de temperaturas ambientales, la presión de vapor del queroseno es demasiado baja para acumular concentraciones significativas de vapor. Sin embargo, la combinación de espacios cerrados y elevada temperatura puede dar lugar a concentraciones elevadas de vapor de queroseno. Además, las aplicaciones en

spray de productos que contienen queroseno, al ser una mezcla de vapor y aerosol, puede causar irritación del tracto respiratorio (www.concawe.be, 2002a).

La exposición crónica a queroseno puede causar disfunción renal, hepática, neurológica, pulmonar (Harris et al., 2000a; http://ocf.nps.navy.mil/jetfuel/docs/liv_kid_jp8.doc, 2001), disfunción emocional, electroencefalogramas anormales, pérdida de atención y disminución de la velocidad sensitivo-motora (Harris et al., 2000a; Harris et al., 2000b; Harris et al., 1997a; Harris et al., 1997b; www.thermo.com, 2002; Liu et al., 2001; Harris et al., 2000c), dolores de cabeza, náuseas, confusión, ataxia, mala articulación en el habla (Josefson, 1997), vértigo, dificultades en el equilibrio, cansancio general, anorexia y dificultades en la concentración (www.atsdr.cdc.gov, 2002), visión borrosa, somnolencia, confusión y desorientación. Una exposición excesiva puede causar daños en el sistema nervioso central produciendo depresión respiratoria, temblores o convulsiones, pérdida de consciencia, coma o muerte (www.chevron.com, 2002b; Ribak et al., 1995). Algunos estudios hechos en ratones han demostrado que puede causar fibrosis pulmonar intersticial, pero en humanos no se ha demostrado (McKay et al., 2001).

En Estados Unidos, la NAVOSH (Naval Occupational Safety and Health Department) recomendaba exposiciones menores de 350 mg/m^3 para un periodo de trabajo de 8 horas y no más que 1000 mg/m^3 para 15 min (Harris et al., 2000a; Harris et al., 2000b; Harris et al., 1997a; Harris et al., 1997b), aunque se han informado de exposiciones por encima de 10295 mg/m^3 para entradas a tanques (www.aircareintl.org, 2002).

PARTE B.- PARTE EXPERIMENTAL

1. OPTIMIZACIÓN DE LAS METODOLOGÍAS ALTERNATIVAS DE ANÁLISIS

La cuestión que se trata de abordar en este capítulo es, por tanto, desarrollar una metodología rápida, sencilla de aplicar y fiable para determinar algunas propiedades físico-químicas del queroseno. Esta(s) metodología(s) sería(n) aplicable(s) de forma alternativa a los métodos oficiales, algunos de los cuales son lentos, subjetivos y con algunos problemas operacionales. La aplicabilidad de las metodologías que aquí se estudian no quedaría restringida sólo al control de calidad industrial sino también a la realización de estudios en el campo de la seguridad y salud de los trabajadores.

1.1. MEDIDA DE QUEROSENO EN FASE GAS

Se han preparado dos dispositivos experimentales diferentes para caracterizar el queroseno en fase gaseosa mediante espectroscopia infraroja en la zona media (FT-MIR). Uno de ellos, instrumentalmente más sencillo, emplea una celda de cuarzo típica de la espectroscopia UV e IR (Smith, 1996) y otro, más complejo, hace uso de una celda de gases (López-Anreus *et al.*, 1998; Pérez-Ponce *et al.*, 1998; Pérez-Ponce *et al.* 2000; Hren *et al.*, 2000).

1.1.1. SISTEMA DE MEDIDA EN LA CELDA DE CUARZO

El instrumental que se precisa es:

- a) Celda de cuarzo con paso óptico fijo de 1 cm.
- b) Cámara termostatzada (Spectroscopy Central Ltd., U.K.) ó estufa de laboratorio.

En la **Figura 8** se representa un esquema del dispositivo experimental necesario para realizar la medida y un ejemplo de espectros de diferentes muestras obtenidos mediante el mismo.

Las medidas se efectuaron haciendo uso de una celda de cuarzo grado espectroscopia UV e IR con paso óptico fijo de 1 cm. Se introdujeron en la celda 50 μL de queroseno y se obtuvo la fase vapor calentando a 37°C durante 2 min en una cámara termostatzada. Una vez obtenida de esta manera la fase vapor, se introdujo rápidamente la celda de cuarzo en el compartimento de espectrómetro para obtener el espectro.

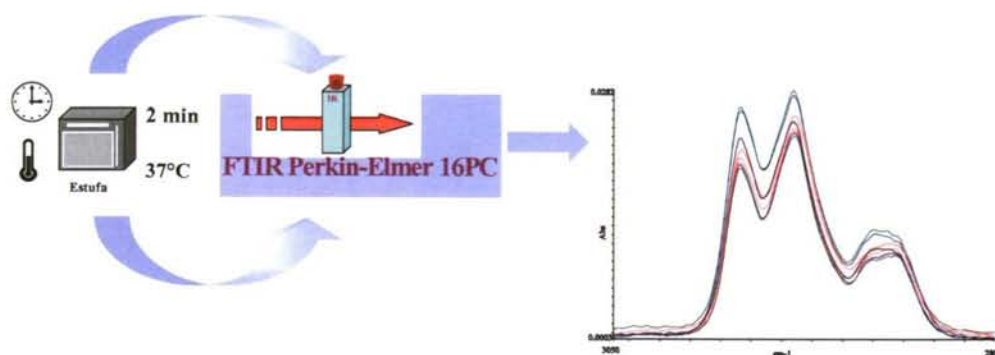


Figura 8: Esquema de trabajo y espectros de querosenos en fase vapor mediante el dispositivo experimental con celda de cuarzo de 1 cm ("simple").

1.1.1.1. OPTIMIZACIÓN

Como se ha indicado, este sistema es muy simple ya que se toma una gota de la muestra de queroseno líquido ($50\mu\text{L}$ en todos los casos), se coloca en el interior de la celda de cuarzo (tapada) y se calienta ligeramente hasta lograr la evaporación. Las variables a optimizar son pues, la temperatura mínima a la que se conseguirá la evaporación sin que exista condensación en la celda durante el proceso posterior de medida y el tiempo de calentamiento de la celda.

Para optimizar estas dos variables, se llevó a cabo un simplex modificado de Nelder y Mead (método de la "contracción masiva", ver Figura 9). El máximo de absorbancia (con bajo ruido de fondo) se obtuvo a 37°C y 2.5 min.

El espectro final de la muestra resulta de una media de 25 interferogramas, corrigiendo la línea de base en el rango espectral de $3500\text{-}2500\text{ cm}^{-1}$ (zona espectral para la que es transparente la celda de cuarzo) y usando la apodización de Beer-Norton fuerte. El espectro del fondo se realizó calentando la celda vacía a 37°C durante 2.5 minutos.

La interpretación química de esta pequeña ventana espectral es muy sencilla ya que las tres bandas muestran las señales correspondientes a las tensiones de los enlaces C-H de las unidades CH_3 y CH_2 en hidrocarburos alifáticos: aproximadamente a 2930 cm^{-1} se observa la vibración de tensión asimétrica del enlace CH en grupos CH_3 solapada con tensión asimétrica CH en CH_2 . A $2850\text{-}2870\text{ cm}^{-1}$ se observa la tensión simétrica de los enlaces CH en CH_3 solapada con tensión simétrica de los CH y CH_2 . La banda situada aproximadamente en 2970 cm^{-1} se interpreta como la tensión

del enlace C-H en las unidades aromáticas y olefínicas. Nótese que la posición de las bandas está ligeramente desplazada de los valores “típicos” por tratarse de un espectro en fase gas.

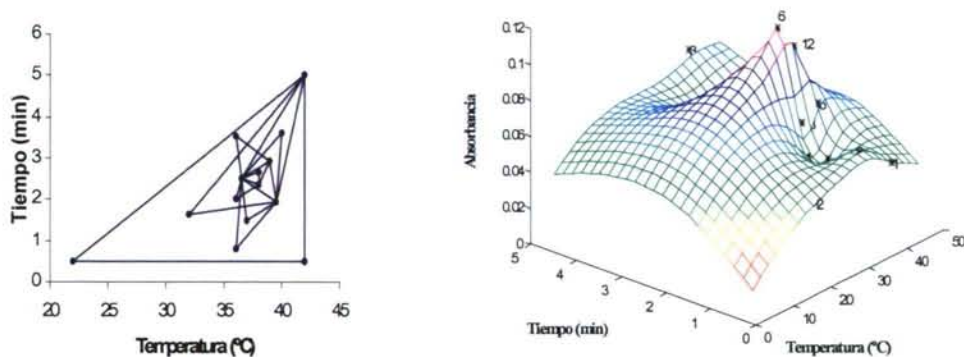


Figura 9: Simplex modificado de Nelder y Mead empleado para la optimización de la temperatura y el tiempo de calentamiento de la muestra para la generación de vapor. Óptimo a 37°C durante 2.5 min.

1.1.2. SISTEMA DE MEDIDA CON LA CELDA PARA GASES

El instrumental empleado para el sistema “complejo” de generación y medida vapor de queroseno mediante celda de gases requirió el siguiente material:

- Celda de gases de paso óptico 10 cm (Wilma Glass Company, USA).
- Ventanas de ZnSe.
- Gas N₂ como fluido impulsor.
- Recipiente de calentamiento de vidrio Pyrex.
- Cinta calefactora (Selecta, Barcelona, España).
- Baño termostático (Selecta Precistern 2L, Barcelona, España).

El dispositivo para realizar la medida de queroseno en fase gas se muestra en la **Figura 10**, teniendo en cuenta que entre las distintas muestras se debe realizar un blanco de N₂ siguiendo el mismo procedimiento que para las muestras.

En resumen, el proceso de análisis consiste en introducir el queroseno en un recipiente de vidrio Pyrex y calentarlo empleando un baño termostático para producir vapor de queroseno. A continuación, para poder arrastrar el vapor hacia la celda de gases, se hace pasar una corriente de N₂, se cierra la corriente de gas, se deja termostatar y se obtiene el espectro FTIR.

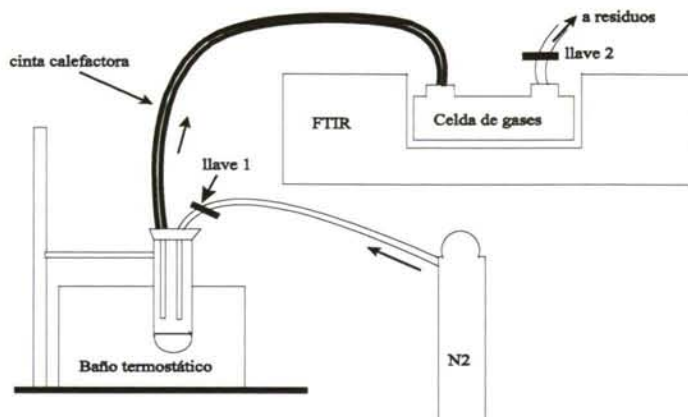


Figura 10: Dispositivo experimental para la celda para gases. El conducto por el que pasan los vapores de queroseno y la celda de gases se rodean con una cinta calefactora para que no haya condensación.

1.1.2.1. OPTIMIZACIÓN

Como se señaló, el sistema “complejo” (cuyo uso parece más general) implica tener en cuenta un número relativamente elevado de variables (y su consiguiente optimización) para generar el vapor y mantenerlo como tal durante el tiempo necesario para llevar a cabo la medida IR. En este trabajo se ha intentado realizar el proceso de optimización de forma multivariante siempre que ha sido posible. Las variables a considerar fueron:

- Flujo de gas portador
- Tiempo de arrastre del vapor.
- Volumen de muestra (0.2-3 mL).
- Temperatura de la línea de transferencia del vapor y celda de medida (170-190°C).
- Temperatura del baño termostático (130-190°C).
- Tiempo de generación del vapor en el baño termostático (120-360s).

Las variables a) y b) no se pudieron incluir en el método multivariante de optimización ya que cuando se fijaban niveles elevados de presión o caudal existían problemas de sobrepresión en la celda y conexiones. Tras algunas pruebas, sus valores se fijan en 2 L/min y 4 s.

Con respecto al volumen de muestra líquida que debe usarse, esta variable debe fijarse ya en los primeros pasos del estudio. Por ello se hizo un estudio univariante sencillo en el cual se estudiaba si el volumen de muestra introducido en el recipiente

sencillo en el cual se estudiaba si el volumen de muestra introducido en el recipiente de calentamiento influía en la absorbancia del espectro (caracterizado por el pico espectral a 1464 cm^{-1}). En la **Figura 11** se aprecia que el volumen de muestra no influye siempre y cuando sea $\geq 1.5\text{ mL}$. Por seguridad se fija en 2 mL (la precisión es satisfactoria a partir de 1.5 mL ; del orden de 3.23 de %RSD).

Aunque la variable d) inicialmente se había incluido en la optimización multivariante, hubo de extraerse de ella porque siempre que se alcanzaban temperaturas en torno a 190°C , las cintas calefactoras se fundían (a pesar de que las especificaciones del fabricante indicaban temperaturas de trabajo de hasta 200°C). Por consiguiente se estableció un límite de seguridad de 180°C . El rango operativo de la variable queda así entre 170 - 180°C . La temperatura inferior se fijó un poco más elevada que la temperatura inicial de destilación de los querosenos (150 - 165°C) puesto que la misión de la cinta es evitar la condensación de los vapores. La variable d) se ha optimizado entre 170 - 180°C , de forma univariante encontrándose un valor ligeramente más elevado para la absorbancia medida a 1464 cm^{-1} a 173°C (se fijó este valor como adecuado aunque se permitía un rango de ± 3 (ver **Figura 12**).

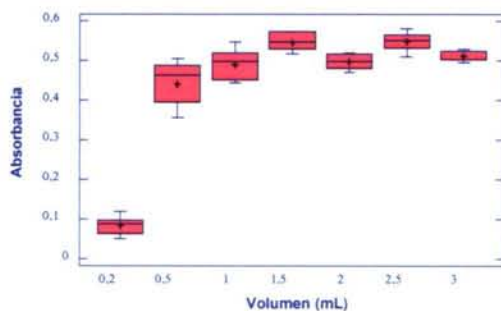


Figura 11: Optimización del volumen del líquido. Se establece un mínimo de 1.5 mL .

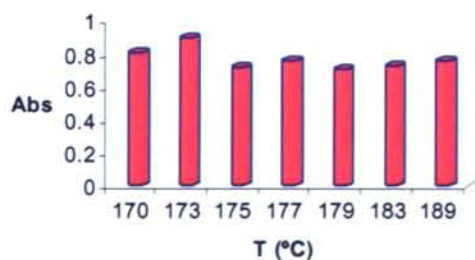


Figura 12: Optimización de la temperatura de la línea de transferencia (absorbancia a 1464 cm^{-1}). Se establece la temperatura de 173°C ($\pm 3^{\circ}\text{C}$).

La temperatura del baño termostático y el tiempo de permanencia del queroseno dentro de él (variables e y f) se han optimizado mediante un método multivariante de optimización (un SIMPLEX modificado de Nelder y Mead) debido a la fuerte relación existente entre ellas y su gran influencia en los resultados (ver Figura 13). El óptimo se estableció en 186°C y 5 min.

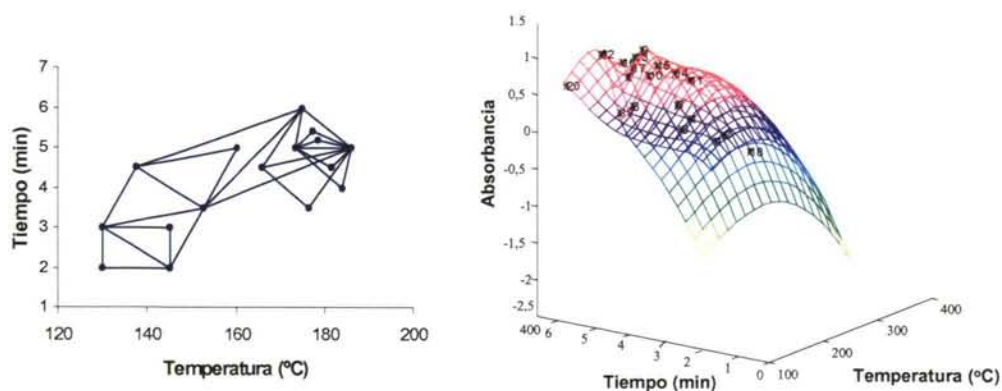


Figura 13: Simplex modificado de Nelder y Mead empleado para la optimización de la temperatura del baño termostático y el tiempo de generación de vapor. Óptimo a 5 min (300 s) y 186 °C, después de 12 iteraciones.

Una vez optimizadas las condiciones experimentales se procede a medir las muestras con el espectrómetro FTIR entre 3500-600 cm^{-1} , realizando 50 barridos, apodización fuerte de Beer-Norton y corrección de línea de base entre 1700 y 1100 cm^{-1} . A cada espectro se le resta el blanco realizado previamente en las mismas condiciones experimentales.

Se introducen 2 mL de queroseno en un recipiente de vidrio Pyrex de 50 mL de capacidad en las condiciones optimizadas. El vapor se arrastra hacia la celda de gases con una corriente de N_2 . Se cierra la corriente de gas, se esperan 35 s y se obtiene el espectro FTIR.

La limpieza de la celda se realiza pasando una corriente de N_2 durante 2 min. Cada 5 muestras se limpia totalmente el sistema utilizando 5 mL de hexano en las mismas condiciones que las muestras. El hexano se elimina totalmente haciendo pasar N_2 durante 10 min.

Una vez realizada la medida experimental se procede a su digitalización en formato LOTUS 123® para abordar el desarrollo de los modelos estadísticos predictivos de las propiedades de los querosenos.

En cuanto a la interpretación química de los espectros (ver Figuras 14 y 15), además de las asignaciones ya realizadas en el apartado 1.1.1.1. se deben incluir las bandas espectrales a 1464 cm^{-1} (flexión asimétrica de los enlaces C-H en CH_3 y/o flexiones simétricas en CH_2); 1377 cm^{-1} (flexión simétrica de CH en CH_3 y asimétrica en CH_2) y 1600 cm^{-1} (tensión de los enlaces C=C en olefinas y anillos bencénicos).

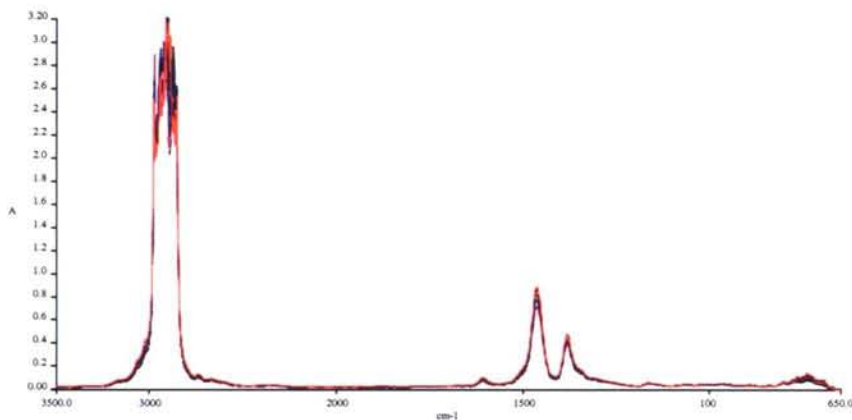


Figura 14: Espectros IR de muestras de queroseno medidas mediante la celda de gases en el rango $3500\text{-}650\text{ cm}^{-1}$.

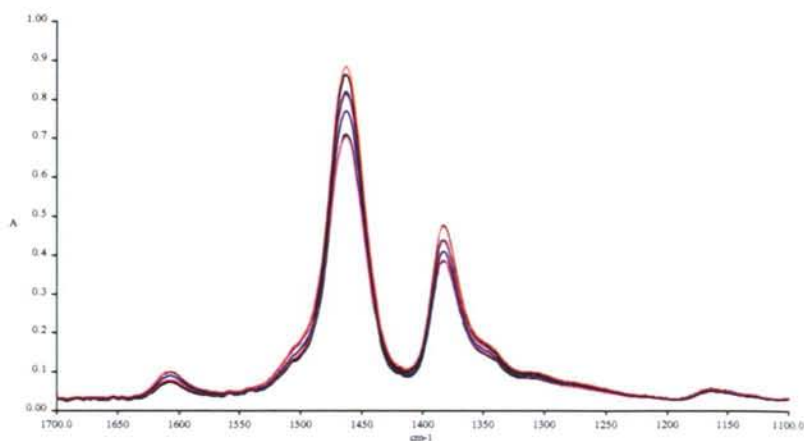


Figura 15: Espectros IR de seis muestras de queroseno medidas mediante la celda de gases en el rango $1700\text{-}1100\text{ cm}^{-1}$.

1.2. MEDIDA DE QUEROSENO EN FASE LÍQUIDA

Se ha trabajado con dos metodologías experimentales diferentes para la caracterización del queroseno en fase líquida: (a) mediante espectroscopia infrarroja en la zona media (FT-MIR) empleando reflectancia total atenuada y (b) la espectroscopia Raman.

1.2.1. SISTEMA DE MEDIDA MEDIANTE REFLECTANCIA TOTAL ATENUADA (ATR)

Las medidas se efectuaron haciendo uso de un accesorio de ATR como se detalla en el Capítulo III. Se introdujo 1 mL de muestra y se obtuvo el espectro de queroseno en fase líquida con una media de 25 barridos en el rango de $4000\text{-}600\text{ cm}^{-1}$, y corrigiendo línea de base entre $1650\text{-}1200$ y $870\text{-}650\text{ cm}^{-1}$.

En la **Figura 16** se superponen algunos espectros obtenidos mediante este sistema de medida.

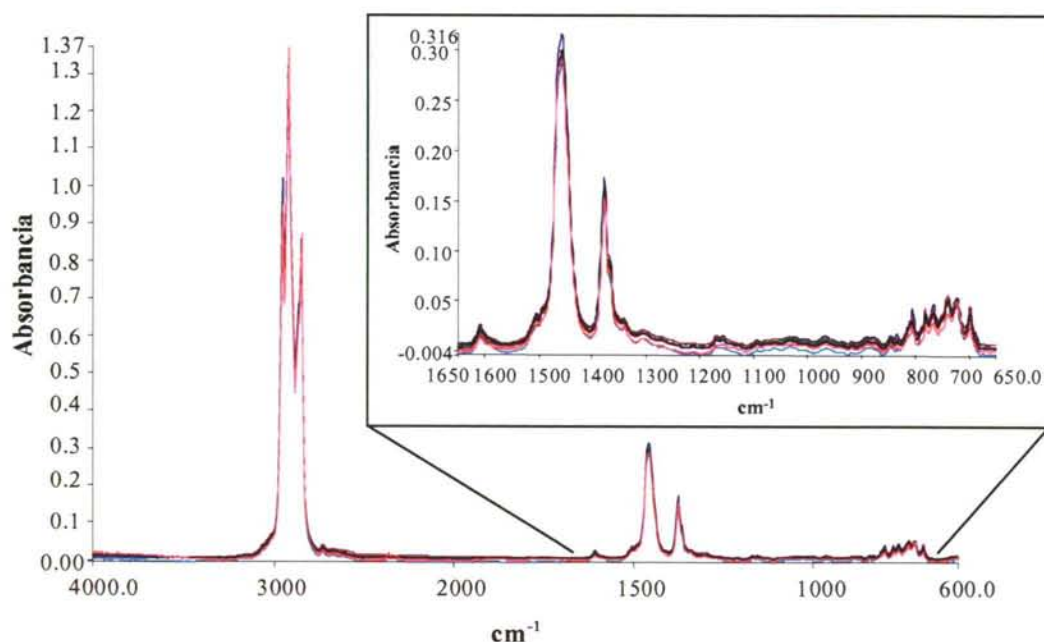


Figura 16: Espectros IR de muestras de queroseno medidas mediante FT-MIR-ATR en el rango $4000\text{-}600\text{ cm}^{-1}$, escogiendo la zona de $1650\text{ a }650\text{ cm}^{-1}$ para la obtención de modelos de PLS.

En cuanto a la interpretación química de los espectros en el rango espectral seleccionado ($1650\text{-}650\text{ cm}^{-1}$) (ver **Figura 16**) podría ser la que se indica a continuación: 1600 cm^{-1} sería la vibración de tensión del enlace C=C de las olefinas y/o aromáticos; 1500 cm^{-1} correspondería con la vibración asimétrica del anillo en el plano; 1460 cm^{-1} sería la flexión de balanceo antisimétrica de grupos CH_3 ; 1455 cm^{-1} se asigna a la vibración de tijera de CH_2 ; 1420 cm^{-1} sería la flexión $\text{CH}_2=\text{C}$ en el plano; 1375 cm^{-1} la flexión de balanceo simétrica de grupos CH_3 ; las bandas a $1170\text{-}1100\text{ cm}^{-1}$ indicarían la presencia de alcoholes terciarios, aunque ese mismo rango ($1175\text{-}1140\text{ cm}^{-1}$) también podrían asociarse a bandas de esqueleto de grupos $\text{C}(\text{CH}_3)_2$, lo que explica el desdoblamiento de la banda de 1375 , también podrían ser bandas de esqueleto debidas al $\text{C}(\text{CH}_3)_3$; la estructura a $840\text{-}790\text{ cm}^{-1}$ correspondería a bandas de esqueleto de los grupos $\text{C}(\text{CH}_3)_2$, el pico a 740 cm^{-1} sería el movimiento del esqueleto $(\text{CH}_2)_n$ y la banda a 699 cm^{-1} corresponde a los anillos aromáticos bencénicos monosustituidos (flexión del C=C-H fuera del plano, junto con la de cerca de 750 cm^{-1}).

1.2.2. SISTEMA DE MEDIDA MEDIANTE ESPECTROSCOPIA RAMAN

Los espectros Raman de las muestras de queroseno se obtuvieron usando un espectrómetro Raman con Transformada de Fourier Bruker RFS 100 equipado con un sistema láser Nd:YAG a 1064 nm con una potencia de salida máxima de 2 W , un beamsplitter de cuarzo y un detector de Ge enfriado criogénicamente con nitrógeno líquido. La excitación del láser se fijó en 300 mW para minimizar el calentamiento de la muestra y la fluorescencia. La potencia del láser se mantiene tan estable como sea posible. El espectro Raman se registró en modo de retrodispersión, haciendo una media de 25 barridos (tiempo de adquisición de 150 s) a una resolución de 4 cm^{-1} y apodizado usando el algoritmo de Blackman-Harris 4-Term. No se observó en el espectro ninguna señal de fluorescencia. Además se necesitó:

- Viales de vidrio estándar de 2 mL ($12 \times 32\text{ mm}$).
- N_2 líquido para enfriar el detector criogénicamente.

Las muestras se introdujeron en viales estándar de vidrio de 2 mL ($12 \times 32\text{ mm}$) para automuestreadores de HPLC (Agilent Technologies) y se taparon con septums de teflón (hechos de láminas comerciales de teflón), para prevenir la volatilización. Los viales de vidrio se dejaron a temperatura ambiente ($18^\circ\text{C} \pm 2^\circ\text{C}$) antes de medir.

Se obtuvieron todos los espectros para desplazamientos Raman entre 0 y 3500 cm^{-1} . Los desplazamientos entre 0 y 70 cm^{-1} se asocian a la influencia del láser y se descartan. Los espectros fueron digitalizados y exportados mediante el software propio de Bruker. Apenas se observaron diferencias entre los picos de los distintos espectros recogidos para las diferentes muestras (valoración visual solapando los espectros). La **Figura 17** muestra algunos espectros típicos en el rango entre 3500 y 70 cm^{-1} , donde se pueden observar pocas diferencias entre muestras (como es normal en producción de rutina).

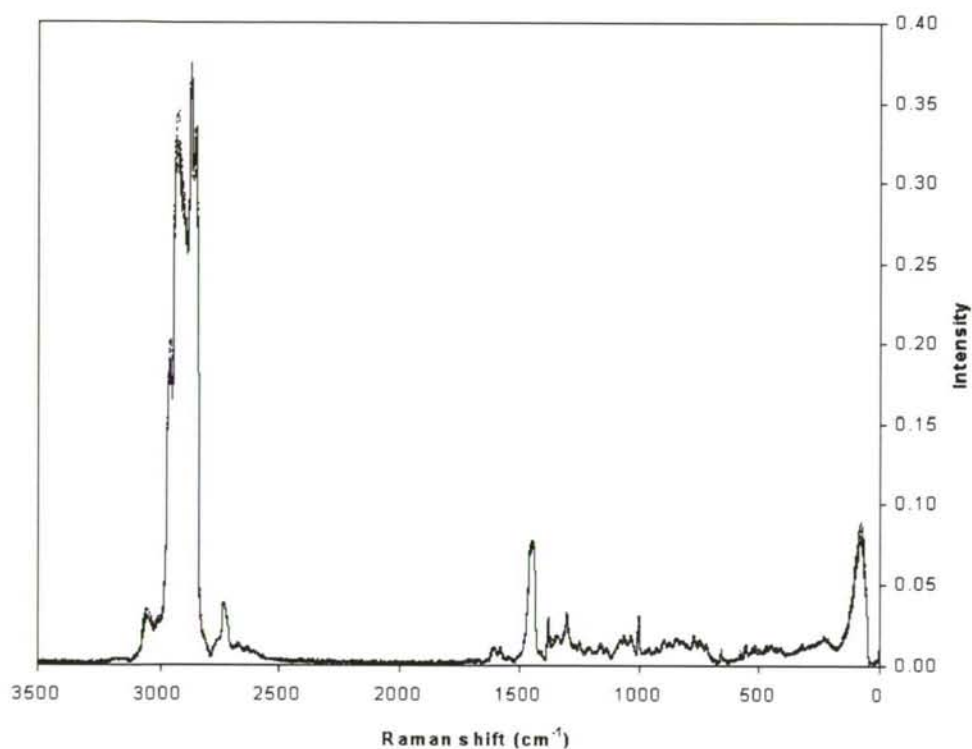


Figura 17: Espectros Raman típicos de queroseno.

1.2.2.1. OPTIMIZACIÓN

a) Efecto de la potencia del láser y de la posición de la muestra.

Teniendo en cuenta que la estrategia de medida debe ser tan robusta como sea posible se decidió determinar la influencia de pequeños cambios en la posición de la muestra y en la potencia del láser sobre las predicciones.

La dependencia de la intensidad de la altura de pico con la energía del láser se estudió considerando algunos picos espectrales (de diferente magnitud) y algunas energías. Las intensidades de los picos a 1303 cm^{-1} , 1443 cm^{-1} , 1457 cm^{-1} , 2853 cm^{-1} , 2873 cm^{-1} y 2933 cm^{-1} se determinaron registrando el espectro a 100, 200, 300, 400, 500, 600 y 700 mW (ver **Tabla III**). La regresión lineal obtenida usando el criterio de mínimos cuadrados ordinario conducía a ajustes lineales con coeficientes de regresión desde 0.9997 (para el pico a 2873 cm^{-1} , muy intenso) a 0.9951 (para el pico débil a 1303 cm^{-1}) (ver **Figura 18**). Estas altas correlaciones indican que sería posible relacionar los valores espectrales obtenidos a una energía de láser con los obtenidos usando otras energías. Esta propiedad interesante se usará más tarde cuando se discuten las opciones de escalado. Estudiando el efecto de la potencia del láser en la relación señal/ruido, se seleccionó la potencia de 300 mW para tener una buena relación señal/ruido (87 a 2931 cm^{-1} , 20 a 1445 cm^{-1} y 10 a 2730 cm^{-1}) y evitar problemas relacionados con el calentamiento de la muestra o fluorescencia.

(a)		Intensidad de picos en u.A. (corregida línea de base)					
picos (mW)	$1457,69\text{cm}^{-1}$	$1443,28\text{ cm}^{-1}$	1303 cm^{-1}	$2932,87\text{ cm}^{-1}$	$2873,15\text{ cm}^{-1}$	$2853,24\text{ cm}^{-1}$	
100	0,0248	0,0239	0,0077	0,102	0,117	0,113	
200	0,0416	0,0413	0,0119	0,187	0,217	0,204	
250	0,0487	0,0487	0,01557	0,227	0,26	0,246	
300	0,057	0,0578	0,0178	0,268	0,306	0,289	
400	0,0729	0,0759	0,0237	0,357	0,405	0,379	
500	0,0894	0,0958	0,0281	0,45	0,511	0,476	
600	0,108	0,111	0,0302	0,53	0,602	0,548	
700	0,12	0,123	0,0367	0,599	0,687	625	

(b)		Intensidad de picos en u.A. (sin corregir línea de base)					
picos (mW)	$1457,69\text{ cm}^{-1}$	$1443,28\text{ cm}^{-1}$	1303 cm^{-1}	$2932,87\text{ cm}^{-1}$	$2873,15\text{ cm}^{-1}$	$2853,24\text{ cm}^{-1}$	
100	0,0253	245	1063	103	12	115	
200	0,0429	0,0446	0,018	0,19	0,221	0,208	
250	0,0504	0,0523	0,0226	0,232	0,266	0,251	
300	0,0583	0,0596	0,0273	0,273	0,313	0,296	
400	0,0758	0,0793	0,0358	0,363	0,414	0,387	
500	0,0948	0,1002	0,0439	0,458	0,522	0,486	
600	0,115	0,116	0,0494	0,54	0,614	0,559	
700	0,127	0,131	0,0574	0,61	0,7	64	

Tabla III: Proporcionalidad de la señal de los picos espectrales con la potencia del láser: (a) corregida la línea de base y (b) sin corregir línea de base.

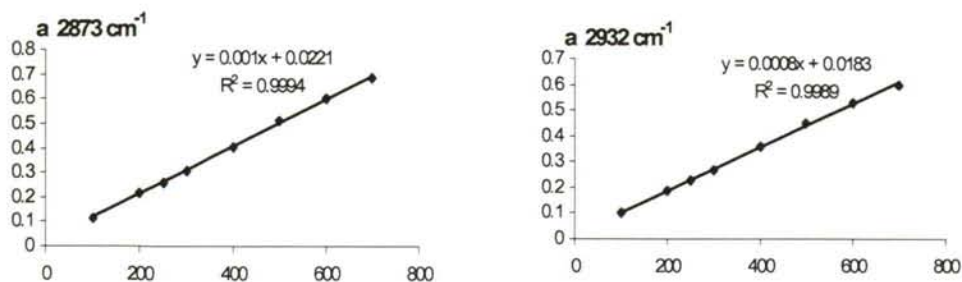


Figura 18: Proporcionalidad de la señal de los picos con la potencia del láser a diferentes longitudes de onda.

Con objeto de estudiar la influencia de la posición de la muestra se empleó un sistema automático de posicionamiento para seleccionar distintas localizaciones de la muestra en el equipo de Raman. La “mejor” situación vendrá determinada previamente por la máxima intensidad observada en el detector. El movimiento de este dispositivo se efectúa en pequeños pasos y nosotros consideraremos cada paso como una “unidad”.

Seis alícuotas de una muestra típica de queroseno se colocaron en seis viales de vidrio diferentes y sujetos a diferentes potencias nominales del láser; 200, 225, 250, 275, 300 (potencia nominal al cual se hizo este trabajo), 325, 350, 375 y 400 mW. A continuación se midieron otras seis porciones de una muestra a una potencia nominal del láser de 300 mW pero moviendo la muestra hacia delante y hacia atrás de la posición de referencia inicial (considerada como la posición 0); denominadas, +12, +6, +3, 0, -3, -6 y -12.

De la Figura 19 y la Tabla IV se puede deducir que la posición de la muestra y la potencia del láser no afectan mucho a los valores medios predichos excepto para situaciones extremas. Para valorar tales efectos se hizo un análisis de la varianza de una vía (ANOVA) para cada potencia del láser y cada desplazamiento y considerando las diferentes propiedades estudiadas. Se puede observar fácilmente que las potencias de 200 mW y 225 mW conducían a valores medios más altos, lo cual se puede atribuir a la baja relación señal/ruido observada en el espectro correspondiente y, por tanto, a un proceso *backscattering* no eficiente. La situación cambiaba para el punto final de destilación y no encontramos una razón clara excepto que se produjese algún problema experimental en estas medidas en particular. Es importante destacar que pequeñas fluctuaciones de la potencia del láser alrededor del óptimo, seleccionado mediante observación visual del máximo del detector, no afectaban a las predicciones.

Adicionalmente, los dos desplazamientos más alejados (por defecto) de la mejor posición de la muestra (-12 y -6) conducían a predicciones mucho menores. Seguramente, estos puntos están afectados por una baja energía del láser en la muestra (una baja eficiencia en el haz del láser). Situaciones leves de “sobreaproximación” no conducían a valores medios significativamente diferentes, ni tampoco una leve “infraaproximación” de -3 unidades. Como sucedía en el estudio de la energía del láser, el punto final de destilación presenta un comportamiento ligeramente diferente, lo que se puede atribuir al mismo problema antes comentado.

En conclusión, cambios mínimos en la potencia del láser o posición de la muestra no afectan a las predicciones medias de los modelos multivariados, esto puede ser debido al efecto positivo que provoca el pretratamiento de normalización unidad (normalización 0→1) que se comentará más adelante.

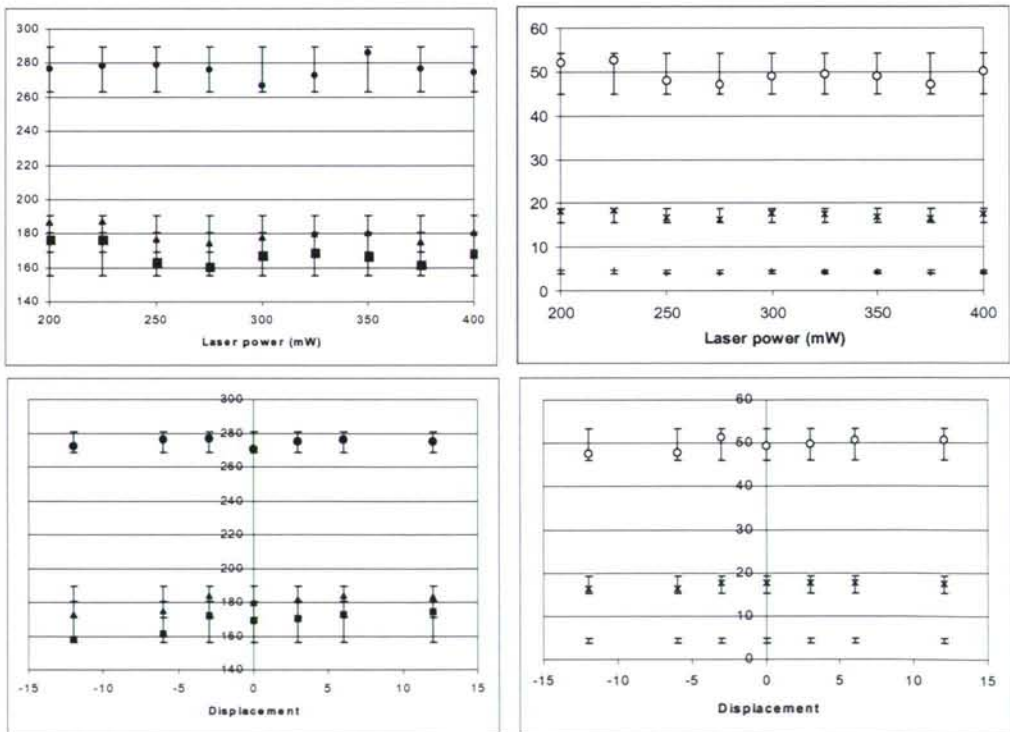


Figura 19: Influencia de la potencia del láser y la posición de la muestra en las propiedades predichas (ver texto y Tabla IV). (○)=Punto de deflagración, (■)=IBP, (▲)=10% destilación, (●)=FBP, (+)= % de aromáticos, (x)=viscosidad. La gráfica muestra el valor medio (símbolo central) y las predicciones máximas y mínimas.

Propiedad (unidad)	Potencia del Laser	Posición de la muestra
Abel (°C)	$F_{exp} = 8.68$ Las medias diferentes son: Altas: 200 & 225 mW	$F_{exp} = 5.66$ Las medias diferentes son: Bajas: -12 & -6 mW
IBP (°C)	$F_{exp} = 11.58$ Las medias diferentes son: Altas: 200 & 225 mW	$F_{exp} = 29.08$ Las medias diferentes son: Bajas: -12 & -6 mW
10% destilación(°C)	$F_{exp} = 7.75$ Las medias diferentes son: Altas: 200 & 225 mW	$F_{exp} = 20.32$ Las medias diferentes son: Bajas: -12 & -6 mW
FBP (°C)	$F_{exp} = 5.46$ Las medias diferentes son: Altas: 350 mW	$F_{exp} = 4.66$ Las medias diferentes son: Bajas: -12 & 0 mW
Viscosidad (cSt)	$F_{exp} = 14.56$ Las medias diferentes son: Altas: 200 & 225 mW	$F_{exp} = 12.57$ Las medias diferentes son: Bajas: -12 & -6 mW
% Aromáticos	$F_{exp} = 26.18$ Las medias diferentes son: Altas: 200 & 225 mW	$F_{exp} = 31.65$ Las medias diferentes son: Bajas: -12 & -6 mW

Tabla IV: Influencia de la potencia del láser y la posición de la muestra (se llevó a cabo mediante ANOVA de una vía, al 95% de confianza). Se indica la potencia del láser y la posición de la muestra que conduce a los valores medios más altos y más bajos (como responsable para las diferencias del ANOVA).

b) Escalado y región espectral a utilizar en los modelos

En general se aconseja disponer de un sistema automático de control de la potencia del láser (Flecher *et al.*, 1996; Cooper *et al.*, 1996) ó, es su defecto, llevar a cabo una vigilancia permanente de la misma.

A pesar de que el equipo utilizado disponía de control automático, fluctuaciones menores de láser no pueden ser despreciadas y así, deberían ser estudiados sus efectos en los modelos de predicción. Este tema se abordaba aplicando simultáneamente varios escalados de datos de la región espectral antes de desarrollar las regresiones multivariantes (Flecher *et al.*, 1996; Flecher *et al.*, 1997). De acuerdo con esto, aquí se aplicó una aproximación análoga. Aunque el centrado en la media (para cada variable) es el que considera frecuentemente para el pretratamiento espectral, se estudiaron dos opciones diferentes: normalización a la unidad (0 a 1, mediante la cual para cada espectro, la altura del pico más alto del espectro, a 2871 cm^{-1} , era escalado a 1 y todas las demás señales escaladas relativas a él) y la combinación de los dos escalados (primero, normalización a la unidad y, después, centrado en la media).

Como se ha indicado en el apartado anterior, existe una buena correlación entre altura de pico y potencia del láser y, por tanto, sería posible “predecir” una altura de pico dada la potencia del láser. La idea clave es que si se pueden normalizar todos los espectros minimizando las diferencias espectrales causadas por la energía del láser, podríamos llevar a cabo modelos predictivos relativamente independientes de la energía del láser empleada para obtener el espectro Raman. De ahí que cada espectro sea normalizado por su altura de pico máximo, lo que puede hacerse usando cualquier software de control.

Un estudio adicional consiste en seleccionar la región espectral más apropiada para implementar los modelos multivariantes. A pesar de que PLS es una técnica de “espectro completo”, es interesante valorar en qué región espectral se presentan (si es que sucede) algunas características (“artifacts” o fuertes no linealidades) que podrían deteriorar los modelos.

Se plantearon tres opciones de trabajo: (i) espectro completo (193.5 cm^{-1} a 3500.9 cm^{-1}), (ii) las dos regiones espectrales donde hay picos más intensos (1158.5 cm^{-1} a 1541 cm^{-1} y $2698.6\text{-}3180.8 \text{ cm}^{-1}$) y (iii) la región de la “huella digital” (193.5 cm^{-1} a 1688.1 cm^{-1}). La región de 0 a 193 cm^{-1} se rechazó en todos los casos porque corresponde a la señal de la banda Rayleigh desde el láser.

Para evaluar la influencia de ambos, el escalado y la región espectral, en los modelos PLS se seleccionó el punto de deflagración como parámetro representativo por su importancia en el control diario de producción de queroseno. La **Tabla V** resume las propiedades principales de los modelos predictivos desarrollados para cada situación. Se puede deducir que la mayoría de los modelos ventajosos eran obtenidos normalizando cada espectro (0 a 1) y seleccionando la región “huella digital”. Nótese que sólo se consideraron 46 muestras para calcular el RMSEP porque las cuatro restantes conducían a grandes errores en cualquier modelo, independientemente del número de VL. El estudio de estas muestras revelaba que sus valores de referencia habían sido redondeados a una expresión de “ $>65^{\circ}\text{C}$ ” y no podían usarse como referencias. A pesar de ello, los errores de predicción son bastante similares para la mayoría de los modelos, la opción seleccionada de región huella y normalización unidad conducía a mejores curvas de regresión (real frente a predicho), menos errores de predicción extremos y modelos más parsimoniosos y fue, en consecuencia, la opción elegida.

Región (cm ⁻¹)	Escalado		
	Centrado en la media	Normalización unidad	Ambos
	7 VL (+)	7 VL	7 VL
193.5 – 3500.9	CV-RMSEC = 2.45	CV-RMSEC = 2.44	CV-RMSEC = 2.44
	RMSEP(*)=2.13	RMSEP=2.13	RMSEP=2.12
	4 VL	5 VL	4 VL
193.5 – 1688.1	CV-RMSEC = 2.56	CV-RMSEC = 2.57	CV-RMSEC = 2.47
	RMSEP=2.17	RMSEP=2.08	RMSEP=2.10
1115.8 – 1541.5	5 VL	5 VL	4 VL
&	CV-RMSEC = 2.44	CV-RMSEC = 2.56	CV-RMSEC = 2.74
2698.6 – 3180.8	RMSEP=2.62	RMSEP=2.54	RMSEP=2.61

(*) n= 46 muestras de validación (ver texto para detalles)

Tabla V: Selección de la región espectral (longitud de onda) y opción de escalado, modelos PLS1 para punto de deflagración.

c) Influencia de los viales de vidrio

Para evaluar el efecto de los viales de vidrio en las predicciones se analizaron dos muestras seis veces, empleando diferentes viales de vidrio (de distintas partidas de producción) y se emplearon sus espectros para predecir sus propiedades. Parece evidente que, *a priori*, los viales de vidrio no deberían aumentar la varianza asociada a las predicciones debido a efectos espectrales ni otros efectos indeseables. De acuerdo con esto, se calculó la desviación estándar relativa (RSD) para cada propiedad y comparó con otros estudios previos donde los viales de vidrio no habían sido cambiados. Los valores de RSD cambiando los viales fueron 2.25%, 0.99%, 0.87%, 0.90%, 1.24% y 2.42% para el punto de deflagración, punto inicial de destilación, 10% de destilación, punto final de destilación, porcentaje de aromáticos y viscosidad, respectivamente. Todos ellos se consideran dentro de la precisión típica de la espectroscopia Raman. Los valores no superaban las RSD de muestras medidas sin cambiar el vial.

d) Interpretación espectral

En las figuras siguientes puede apreciarse que las diferencias espectrales entre las muestras (codificadas como k) son mínimas. Además, los picos espectrales principales observados en la figura que se presenta a continuación se pueden relacionar con los grupos funcionales presentes en este tipo de combustibles (Chung *et al.*, 1991; Lin-Vien *et al.*, 1991; Robinson, 1991). La banda a 3055 cm⁻¹ es típica de la tensión CH en aromáticos mientras que las bandas alrededor de 2960 cm⁻¹, 2934 cm⁻¹,

2926 cm^{-1} , 2872 cm^{-1} y 2852 cm^{-1} corresponden a los modos de tensión simétricos y antisimétricos de CH_2 y/o CH_3 en n-alcanos (ver **Figura 20**). El pico pequeño a 2730 cm^{-1} no ha sido identificado. Nótese que desde 2700 cm^{-1} a 1600 cm^{-1} no aparece señal excepto ruido aleatorio y, por ello, esta región no era útil cuando se desarrollaban los modelos de predicción antes citados.

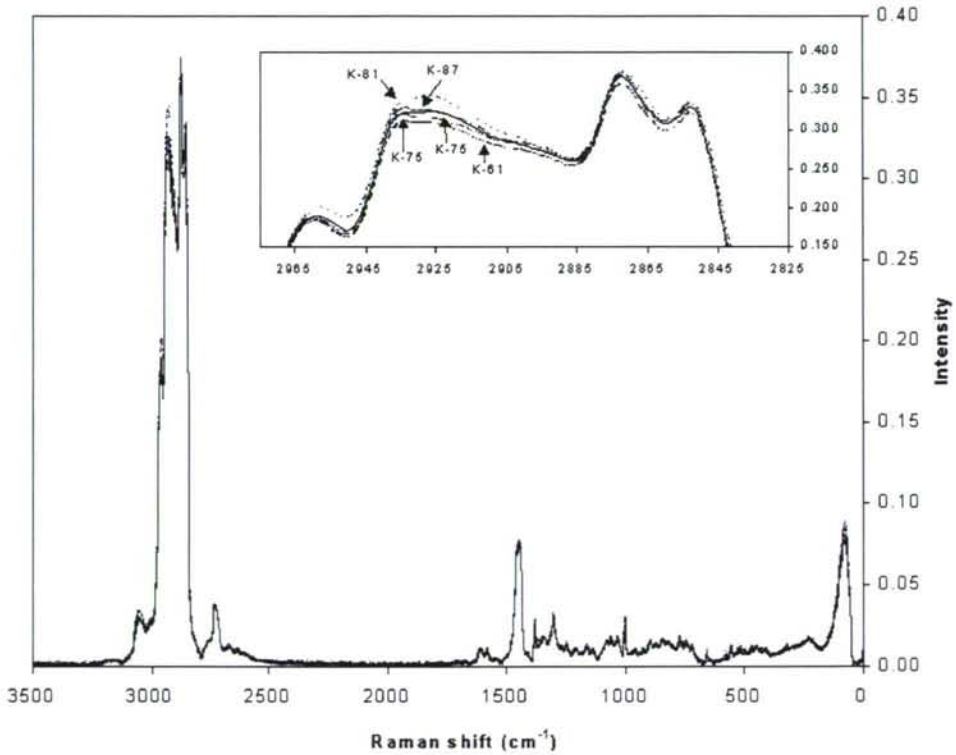


Figura 20: Espectro Raman en “todo” el rango, en el inserto se muestra la zona de 2975-2825 cm^{-1} .

El pico desdoblado centrado en 1600 cm^{-1} (1612 y 1582 cm^{-1}) se puede asignar a la tensión de los anillos de derivados bencénicos (la tensión $\text{C}=\text{C}$ de alquenos lineales puede también estar solapada) y la banda fuerte de 1459 cm^{-1} corresponde a la deformación CH_2 y/o CH_3 de n-alcanos. El pico a 1304 cm^{-1} se puede atribuir a la torsión del n-alcano y a los modos de vibración de la oscilación.

El pico estrecho y bien definido a 1381 cm^{-1} es representativo de la tensión de los anillos bicíclicos de las fracciones aromáticas mientras el pico a 1003 cm^{-1} se asigna al modo de deformación "breathing" en anillos para los componentes aromáticos monocíclicos en la mezcla de los querosenos (Chung et al., 1991).

En la **Figura 21** se observa la banda a 1063 cm^{-1} , que puede ser causada por la tensión C=S (componentes de azufre presentes en los querosenos) y/o vibraciones del anillo para bencenos orto-disustituídos. La banda no resuelta entre 700 y 900 cm^{-1} puede asociarse a un número grande de vibraciones de esqueleto ("modos de breathing") de diferentes estructuras de anillos (Lin-Vien et al., 1991; Robinson, 1991).

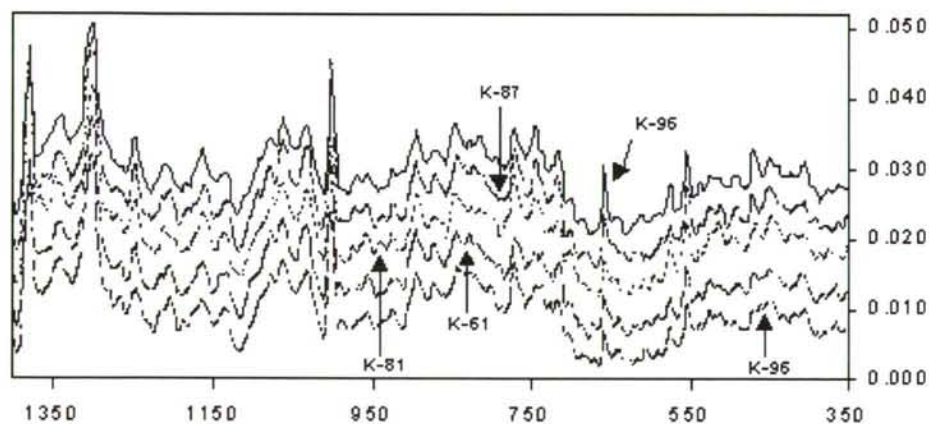


Figura 21: Espectro Raman en el rango $1350\text{-}350\text{ cm}^{-1}$ en donde los espectros se han desplazado para evitar el solapamiento de muestras.

PARTE C.- RESULTADOS Y DISCUSIÓN

1. MODELOS MULTIVARIANTES EN FASE GAS

1.1. MODELOS MULTIVARIANTES CON EL SISTEMA DE MEDIDA SIMPLE

1.1.1. PUNTO DE DEFLAGRACIÓN

El modelo elegido incluye 7 VL y explica el 99.79% de la información existente en las variables predictoras y el 95.87% de la varianza en la variable a predecir. En la **Figura 22** se presentan dos índices para la detección de muestras anómalas: los gráficos de T^2 y Q . Los posibles anómalos estarán claramente alejados de los valores límite (líneas horizontales). En la **Figura 22** se aprecian posibles anómalos: muestra 32 (**Figura 22 (a)**) y muestras 34, 39 y 44 (**Figura 22 (b)**), aunque como al observar la **Figura 23** no se ve una clara desviación de estas muestras del valor esperado no se consideran anómalas (excepto la 32 que continúa siendo sospechosa). La **Figura 23 (b)** permite llegar a conclusiones similares. A mayor peso (*leverage*) y más error normalizado (*error estudentizado*), más posibilidad de que las muestras sean anómalas.

La eliminación de la muestra 32 no provocó mejorías apreciables en el modelo como podía preverse puesto que aunque su *loading* es alto, el error *estudentizado* es pequeño.

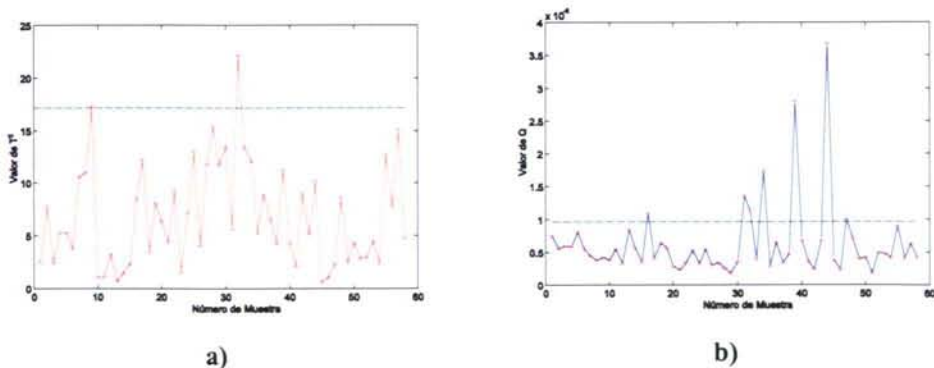


Figura 22: (a) Gráfico de T^2 y (b) gráfico de Q para el punto de deflagración.

En la **Figura 23 (a)** se representa el valor “real” (obtenido por el método clásico IP-170-Abel) frente al valor predicho (obtenido mediante el modelo) durante

la calibración. Se observa que los valores predichos se ajustan bastante bien al valor real (el comportamiento teórico perfecto se indica con una línea recta azul). Aunque hay algún error destacable, se puede aceptar que las predicciones obtenidas son buenas (con un SEC-CV-LOO ($n_c = 58$) = 1.74 °C; SEP ($n_v = 34$) = 1.03 °C) (n_c = número de muestras de calibración, n_v = número de muestras de validación). A pesar de que el SEC ha resultado un poco elevado, el SEP encontrado para las muestras de validación es satisfactorio (al comparar con los valores oficiales recogidos en la sección 1.2. del Capítulo, **Tabla I**) y se encuentra dentro del margen de reproducibilidad permitido para la propia norma. Las muestras 7, 8 y 9 (que se encuentran alejadas del rango habitual) hay que indicar que no se eliminaron porque constituyen un tipo de querosenos de baja volatilidad que se producen con una baja frecuencia (queroseno de alto punto de deflagración, tipo JP5) y deben formar parte del modelo. A pesar de que la muestra 9 sesga ligeramente el modelo, al eliminarla no se conseguían mejoras apreciables ni en el SEC ni en el SEP, por lo que se decidió mantenerla. Cuando se disponga de otras muestras similares deberá cambiarse por ellas (actualización del modelo).

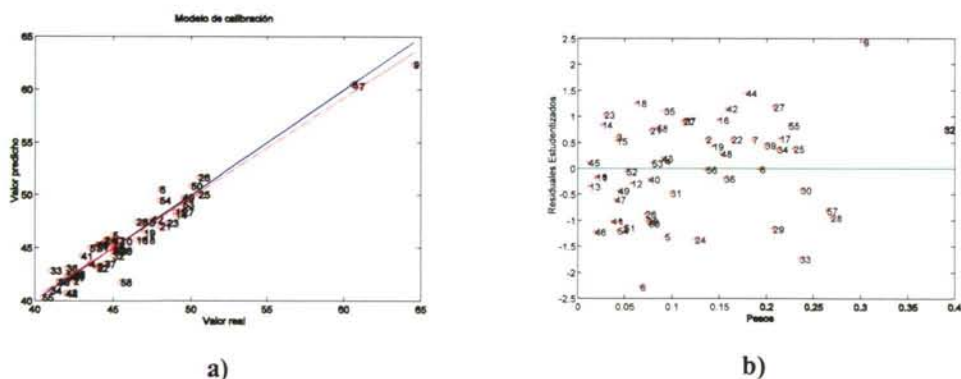


Figura 23: (a) Valor real frente al valor predicho del punto de deflagración para muestras de calibración empleando 7 VL. Ajuste ideal, línea azul; ajuste experimental, línea roja (b) gráfica de los “pesos frente a residuales estudentizados”.

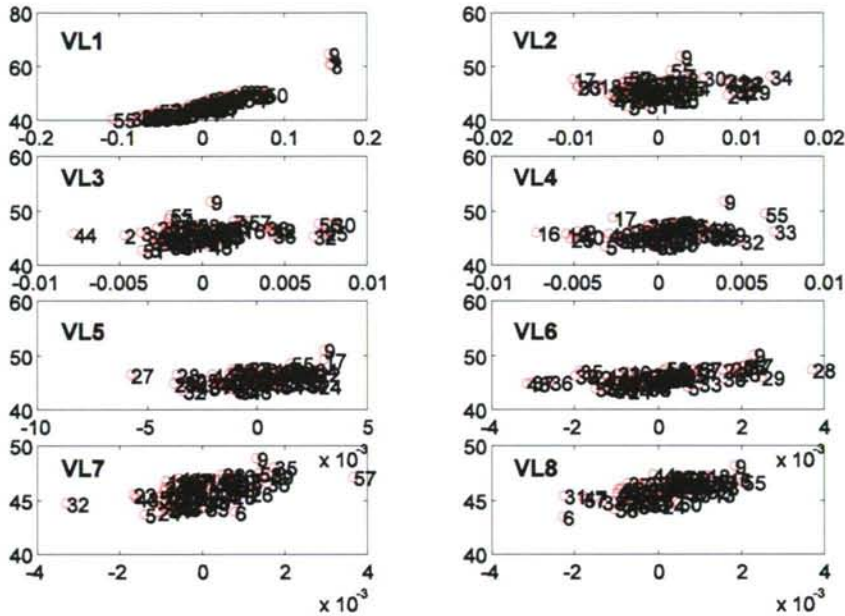


Figura 24: Relaciones entre los scores X (t, en abscisas) e Y (u, en ordenadas) para el modelo PLS (7 VL), punto de deflagración.

En las Figuras 24 y 25 se muestran las relaciones entre los scores de los datos X e Y para el modelo PLS elegido para el *flash point*-Abel. Se muestran las gráficas para las 7 variables latentes. La Figura 24 permite analizar la relación entre los scores de las muestras en el espacio de las variables predictoras, X, (siempre en abscisas) y los scores en el espacio de parámetro a predecir, Y, (siempre en ordenadas). Para todas las VL del modelo se aprecia una buena linealidad entre ambos subespacios. La característica más relevante es el subgrupo diferenciado que forman las muestras 7, 8 y 9 (JP5), como cabía esperar. Se observa que alguna muestra esporádicamente presenta características ligeramente diferenciadas del resto lo que podría ser una consecuencia de alguna diferencia en la composición (el “corte” del queroseno). Esto último se deduce de que el alejamiento del bloque principal de muestras se da en el eje de abscisas, ligado a los scores en X. En la Figura 25 sólo aparecen diferenciadas las ya citadas muestras 7, 8 y 9.

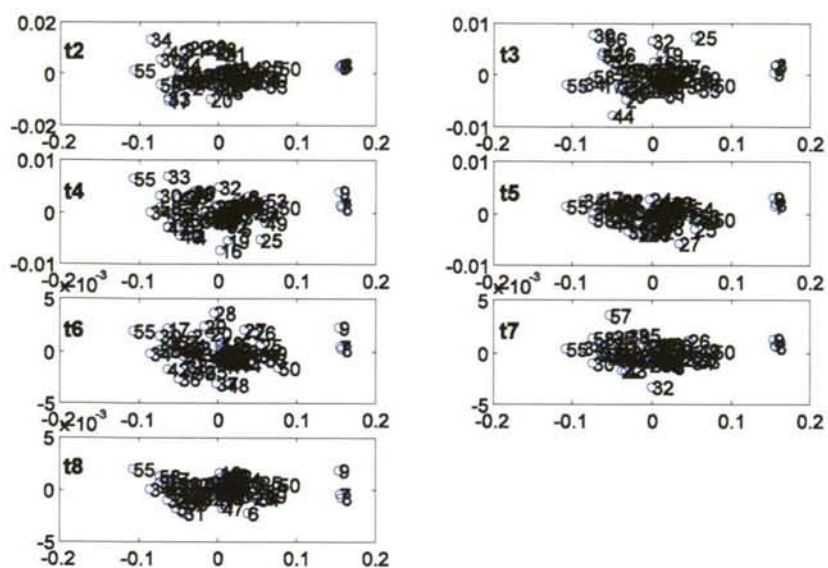


Figura 25: Distribución de las muestras en el espacio de las variables predictoras (scores t_1 vs t_k) para el modelo PLS (7 VL), punto de deflagración (abscisas, t_1).

1.1.2. PUNTO DE CONGELACIÓN

Para la predicción del punto de congelación se han seleccionado 9VL, las cuales conducen al 99.91% de varianza explicada para las predictoras y el 92.40 % para la variable predicha.

En la Figura 26 se muestran los gráficos de T^2 y Q , en donde se revela la posible anomalía de 6 muestras (la 45 en la figura (a) y las 3, 6, 16, 30 y 40 en la figura (b)). Pese a ello, tras estudiar las Figuras 27 (a) y (b) y las Figuras 28 y 29 se descartaron los comportamientos anómalos.

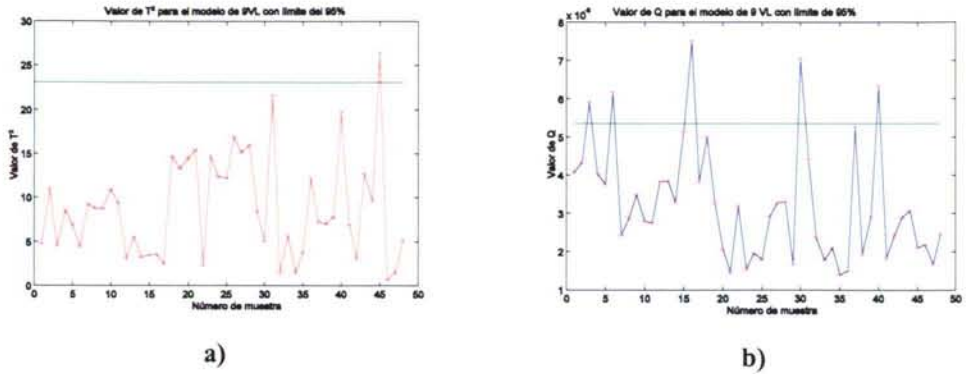


Figura 26: (a) Gráfico de T^2 y (b) Gráfico de Q , para el punto de congelación.

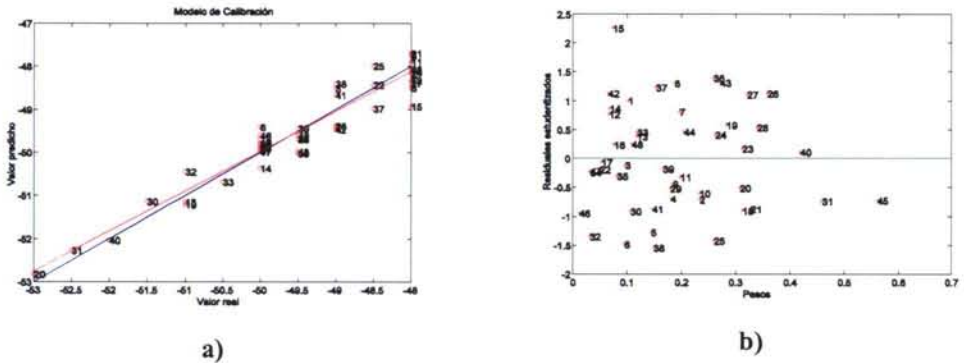


Figura 27: (a) Valor real frente al valor predicho del punto de congelación para muestras de calibración predichas empleando 9 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

En las Figuras 28 y 29 se observan las relaciones entre los *scores* (X vs Y) y las gráficas de *scores* en el espacio X , respectivamente para el modelo elegido de PLS. En lo que se refiere a la Figura 28 se ve que el modelo es complejo porque las primeras VL no explican la relación lineal X - Y . Más bien parecen modelar varianza no útil para la predicción. Esto podría tener relación con alguna de las dificultades que se encuentran al aplicar el método oficial (el vaso Dewar se empaña y dificulta la visualización de los primeros cristales de queroseno y, por tanto, la decisión). No en vano, el valor de reproducibilidad del método oficial es 2.5°C . El pequeño grupo de muestras que se aprecia a la derecha en la Figura 29 no parece afectar al modelo.

La predicción obtenida es aceptablemente buena, con un SEC-CV-LOO ($n_c =$

48) = 1.24°C y un SEP ($n_v = 44$) = 1.26°C. Se aprecia que el error promedio en calibración y predicción es inferior a la reproducibilidad oficial.

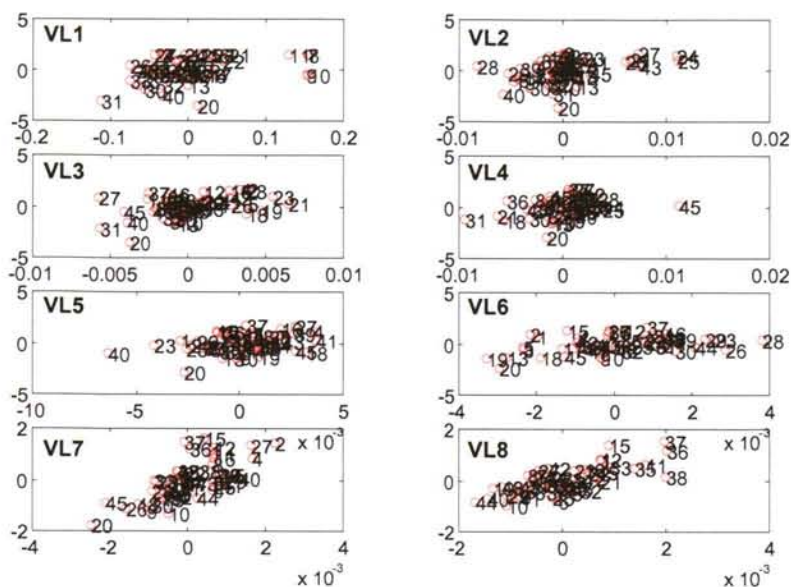


Figura 28: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para cada VL del modelo PLS (9 VL), punto de congelación.

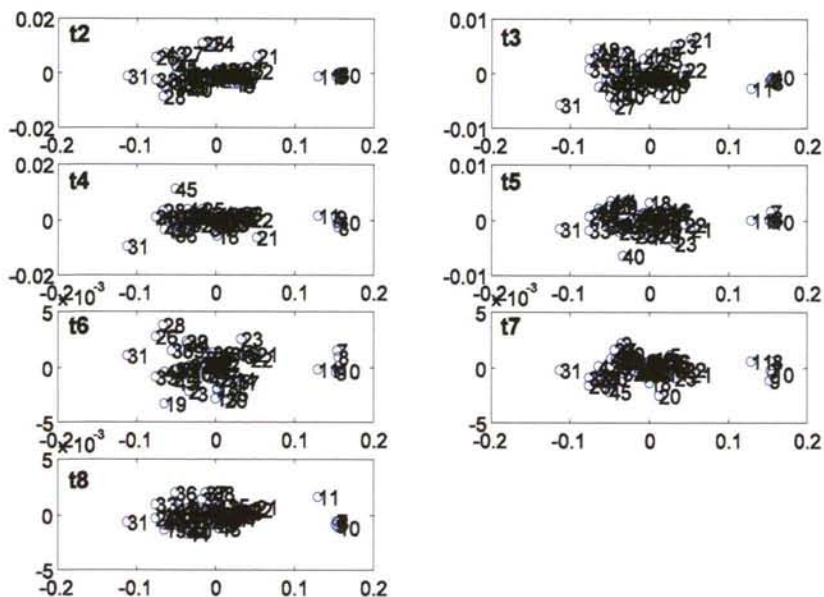


Figura 29: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (9 VL), punto de congelación (abscisas, t1).

1.1.3. PUNTO INICIAL DE DESTILACIÓN

Para el punto inicial de destilación se han seleccionado 2 variables latentes, explicando un 99.38% de la varianza en las X y un 92.83% de varianza en las Y. La **Figura 30** muestra los gráficos de T^2 y Q . Aunque parece que hay varios anómalos (muestras 24, 40 y 45 en (a) y (b) respectivamente), más tarde se observó que no tenían mucha influencia en el modelo (ver **Figura 31**), donde tan sólo la número 24 puede presentar problemas (si bien su *leverage* -peso- numéricamente es pequeño y no debería afectar excesivamente al modelo). Una vez eliminada se apreció que el modelo no mejoraba su comportamiento.

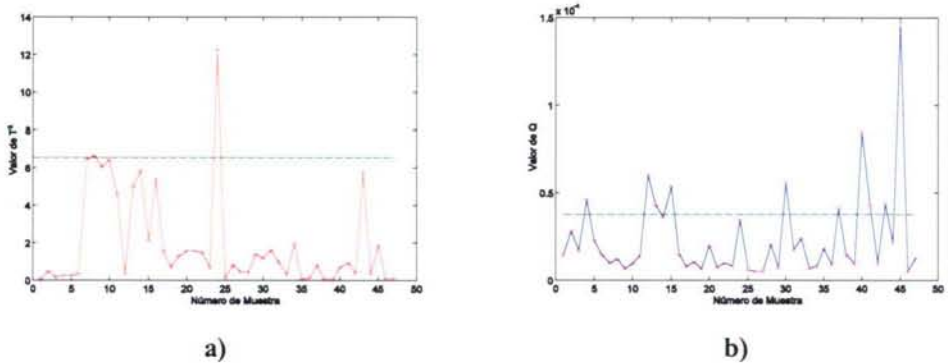


Figura 30: (a) Gráfico de T^2 y (b) Gráfico de Q , para el punto inicial de destilación.

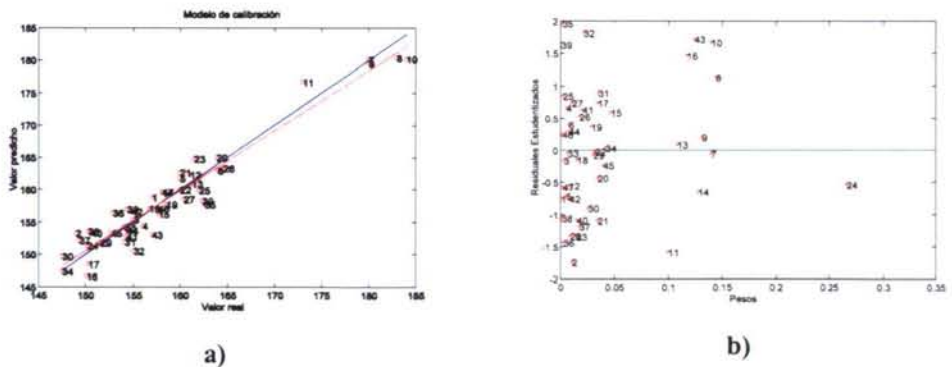


Figura 31: (a) Valor real frente al valor predicho del punto inicial de destilación para muestras de calibración predichas empleando 2 VL (línea azul: comportamiento ideal, línea roja: comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

En las Figuras 32 y 33 se muestran las relaciones entre los scores de los datos X e Y para el modelo PLS elegido para punto inicial de destilación. La complejidad inherente al error intrínseco al método de referencia (volatilización de los componentes más volátiles, dificultad de medida de las primeras gotas en los destiladores y variabilidad en la composición de las muestras) parece revelarse en una relación de scores no lineal en las VL superiores (Figura 32). También se aprecia un pequeño grupo de muestras poco volátiles (números 7, 8, 9, 10, 11, ver Figuras 31 y 33) que es bueno mantener en el modelo para mejorar el rango de aplicación del mismo. Obsérvese que las muestras 7, 8 y 9 ya habían conducido a la misma situación al predecir el punto de deflagración, por la misma razón (corte pesado, JP5).

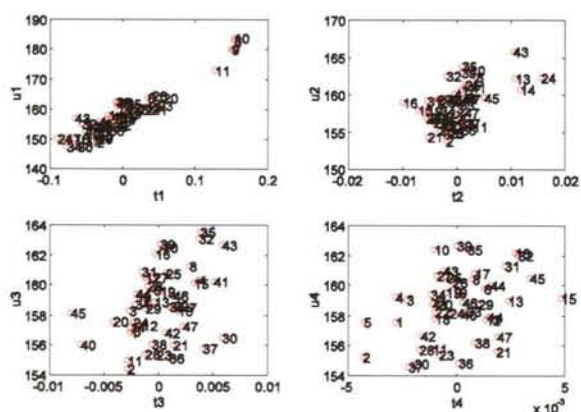


Figura 32: Relaciones entre los scores X (en ordenadas) e Y (en abscisas) para el modelo PLS, punto inicial de destilación (se muestran 4 VL, aunque el modelo escogido sólo considera 2 VL).

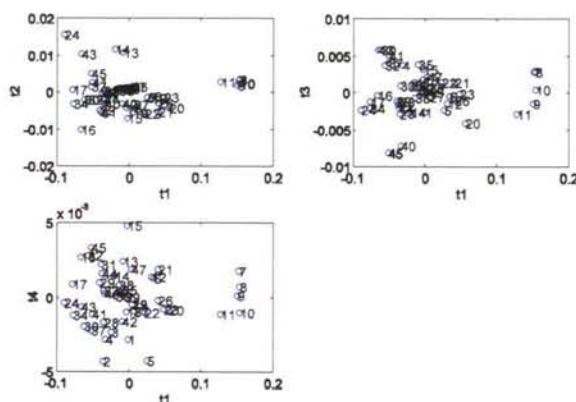


Figura 33: Relaciones entre los scores para el modelo PLS (2 VL), punto inicial de destilación (abscisas, t1).

En general, se puede aceptar que las predicciones obtenidas son buenas (SEC-CV-LOO ($n_c=47$) = 2.56 °C; SEP ($n_v= 43$) = 3.37 °C). Tanto el error de calibración como el de validación son muy buenos comparados con la reproducibilidad del método ASTM (8.5°C), por lo que se considera que el modelo elegido es bueno, incluso a pesar de tener algunas muestras no perfectamente modeladas.

1.1.4. PUNTO 10% DE DESTILACIÓN

Para predecir el 10% de destilación se han seleccionado 8 VL, las cuales conducen al 99.86% de varianza explicada para las X y el 99.03% para la variable a predecir.

En la **Figura 34** se muestran los gráficos de T^2 y Q , en donde se baraja la posibilidad de la posible existencia de 4 anómalos (muestras 19 y 35, gráfica (a) y las 31 y 49 (gráfica (b)). A pesar de ello, al estudiar las **Figuras 35 (a)** y (b) se descarta que tales muestras se comporten como anómalos debido a la buena linealidad observada en la gráfica de “valor real frente a predicho”, y la no existencia de anómalos claros en la gráfica de “pesos frente a residuales estudentizados” de hecho, los valores de pesos son muy bajos. El estudio de las **Figuras 35** y **36** concluye que las muestras 19 y 35 deben presentar alguna característica espectral poco frecuente que las hace ligeramente diferentes en las VL superiores aunque hay otras con comportamientos similares. La observación visual de los espectros no permitió decidir cuál. Al eliminarlas del modelo éste no mejoraba y se decidió mantenerlas en él.

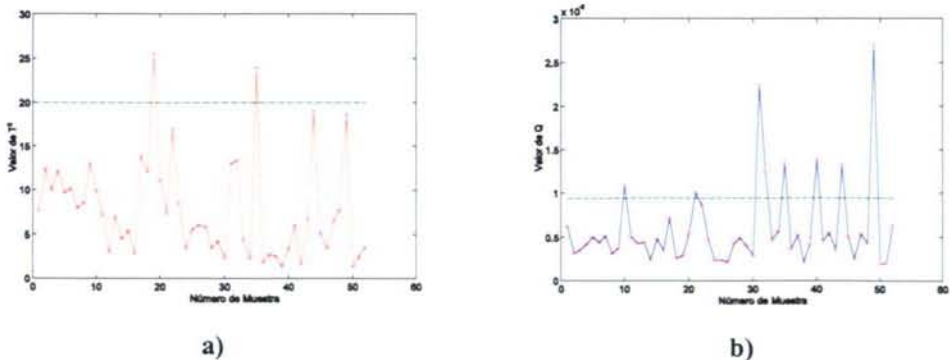


Figura 34: (a) Gráfico de T^2 y (b) Gráfico de Q , para el 10% de destilación.

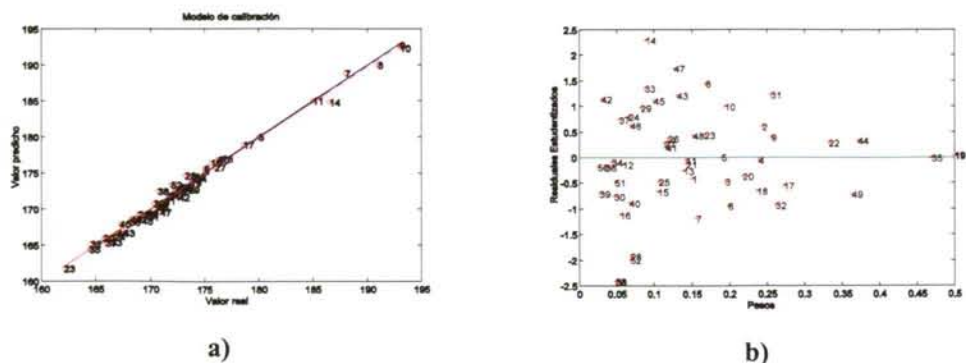


Figura 35: (a) Valor real del 10% de destilación frente al valor predicho para muestras de calibración predichas por el modelo con 8 VL y (b) Gráfica de “pesos frente a residuales estudentizados”.

En las Figuras 36 y 37 observamos que, en general, las relaciones entre los scores de los datos X e Y para el modelo PLS elegido para el 10% de destilación parecen seguir un buen comportamiento lineal para las 8 variables latentes, la excepción serían las variables latentes 4 y 5, más asociadas a las muestras 19 y 35, respectivamente. En la Figura 37 se muestra la gráfica de scores en el espacio de las predictoras para el modelo elegido de PLS para 10% de destilación. Como en los modelos anteriores, las muestras 7, 8 y 9 (junto con alguna otra) muestran sus mayores valores de los puntos de destilación.

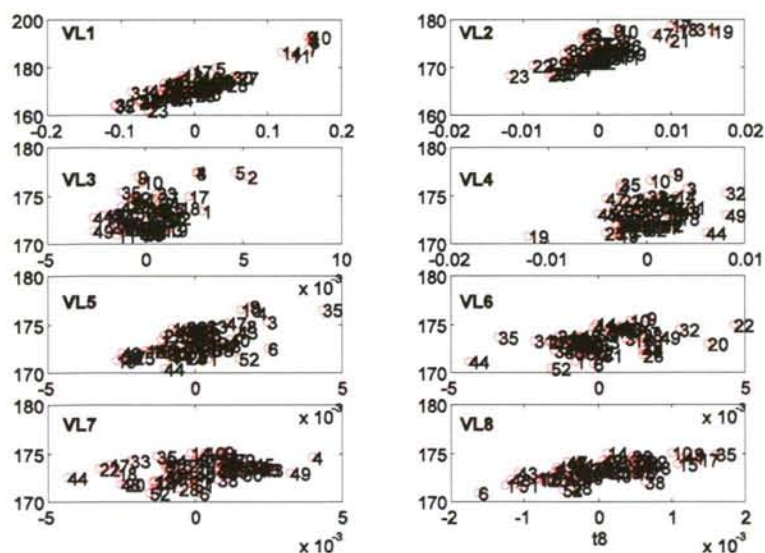


Figura 36: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (8 VL), 10% de destilación.

Por comparación con la reproducibilidad de la **Tabla I**, se puede aceptar que las predicciones obtenidas son muy buenas ($SEC-CV-LOO$ ($n_c = 52$) = 1.64 °C; SEP ($n_v = 46$) = 2.28 °C).

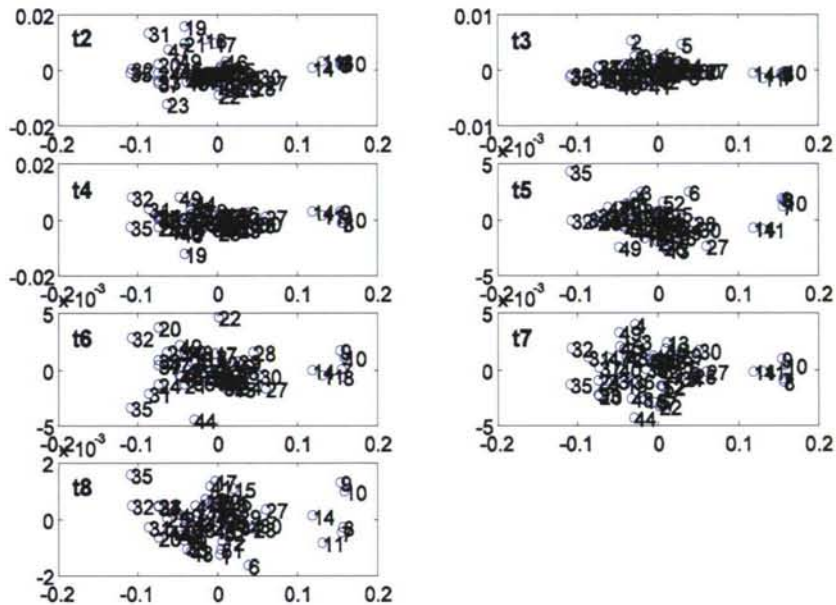


Figura 37: Distribución de las muestras en el espacio de las variables predictoras (scores t_1 vs t_k) para el modelo PLS (8 VL), 10% de destilación (abscisas, t_1).

1.1.5. PUNTO DEL 90% DE DESTILACIÓN

El modelo PLS para la predicción del 90% de destilación precisa un autoescalado de los datos en vez de centrado en la media (los modelos estudiados con el centrado no fueron buenos y presentaban muchos problemas), esto se podría deber a los problemas operativos de medida de los destiladores automáticos al final de la curva de destilación (Dyloff, 1998), lo que se refleja en los valores altos de reproducibilidad que la propia norma ASTM establece. Esto, por tanto, podría ser la razón principal de las grandes dificultades encontradas a la hora de relacionar el espectro con los valores de referencia en un modelo suficientemente satisfactorio.

El modelo con 7VL explica 91.65% de la varianza en el espacio espectral y el 93.93% de la varianza a predecir. En la **Figura 38** se observaron 6 posibles anómalos que se desvían de los valores límite (6, 35, 43 y 48 en la **Figura 38** (a) y 3, 4 en la **Figura 38** (b)). Al observar la **Figura 39** se desmiente la existencia de posibles

anómalos ya que las muestras se ajustan bastante bien a la linealidad “real frente a predicho”.

Obsérvense también las Figuras 40 y 41 pertenecientes a las relaciones entre los scores de los datos X e Y para el modelo elegido de 90% de destilación.

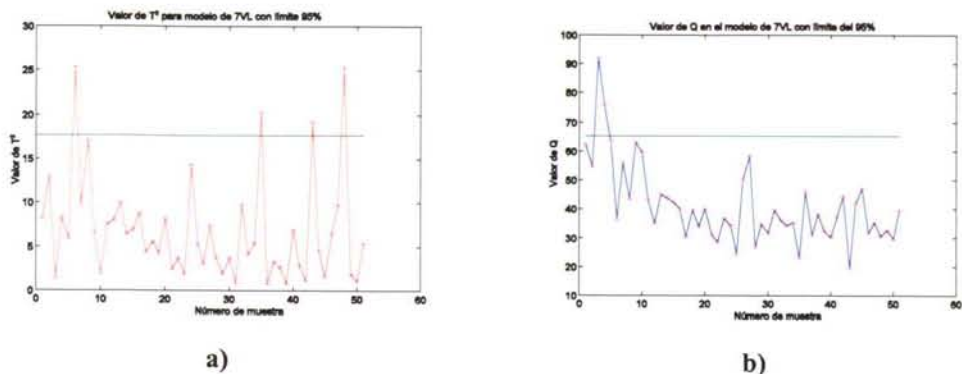


Figura 38: (a) Gráfico de T^2 y (b) Gráfico de Q para el 90% de destilación..

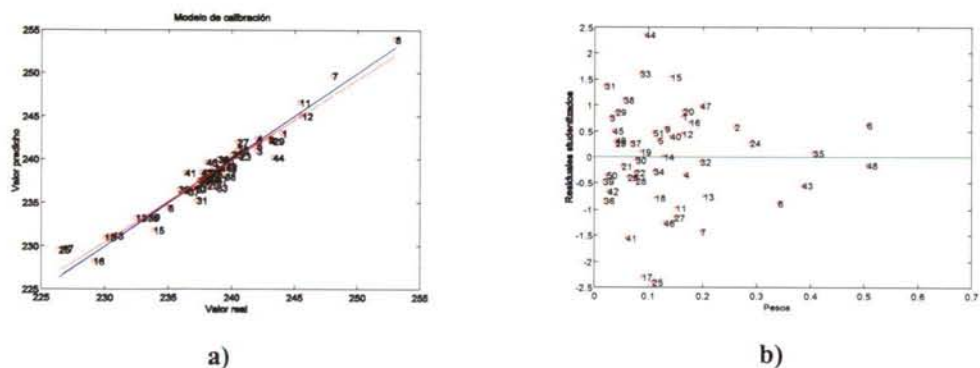


Figura 39: (a) Valor real del 90% de destilación frente al valor predicho para muestras de calibración predichas por el modelo con 7 VL y (b) Gráfica de “pesos frente a residuales estudentizados”.

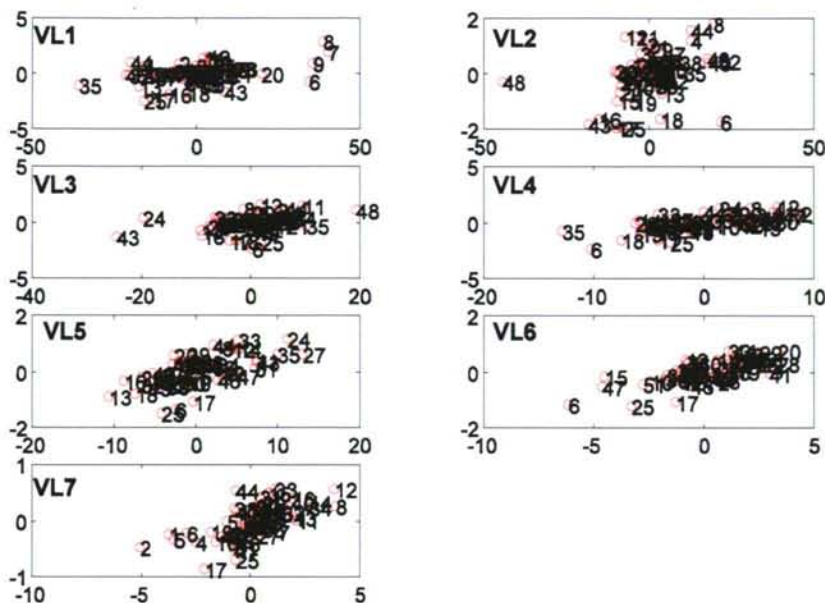


Figura 40: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (7 VL), 90% de destilación.

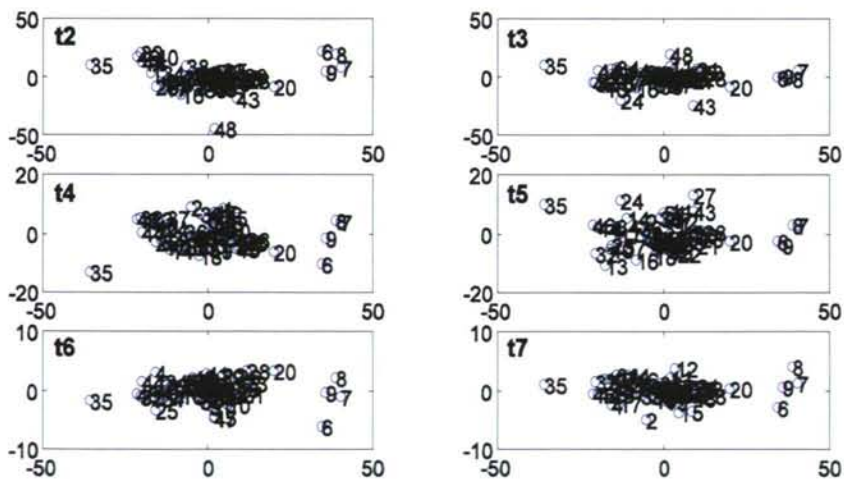


Figura 41: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (7 VL), 90% de destilación (abscisas, t1).

Las muestras 7, 8 y 9 continúan demostrando su diferente composición (productos más pesados, en este caso acompañadas también por la muestra 6).

Las predicciones son satisfactorias (con un SEC-CV-LOO ($n_c = 51$) = 3.89; SEP ($n_v = 37$) = 3.42) si se comparan con la reproducibilidad del método ASTM (8.8°C).

1.1.6. PUNTO FINAL DE DESTILACIÓN

También se aplicó el método PLS para el punto final de destilación autoescalando (por la misma razón que en el caso del 90% de destilación) y se observó que la varianza explicada para las X era del 91.51% mientras que para las Y la varianza explicada fue de un 94.53%. Se han seleccionado 7 VL. En las **Figuras 42** y **43** se observan las diferentes gráficas para el modelo PLS escogido para la propiedad del punto final de destilación. Aunque en la **Figura 42 (a)** se ven cuatro muestras (6, 35, 42 y 47) y en la **Figura 42 (b)** hay otras cuatro (1, 2, 4 y 46) que se desvían ligeramente de los valores límite, cabe esperar que su comportamiento no sea problemático para el modelo. Esto parece confirmarse con la **Figura 43**, donde no se observan muestras sospechosas aunque, ciertamente, la muestra 6 presenta un comportamiento sorprendentemente bajo (especialmente porque en 90% de destilación no se había manifestado su composición ligera -aunque, efectivamente, en el espacio de las X sí había indicado una composición ligeramente diferente al resto-).

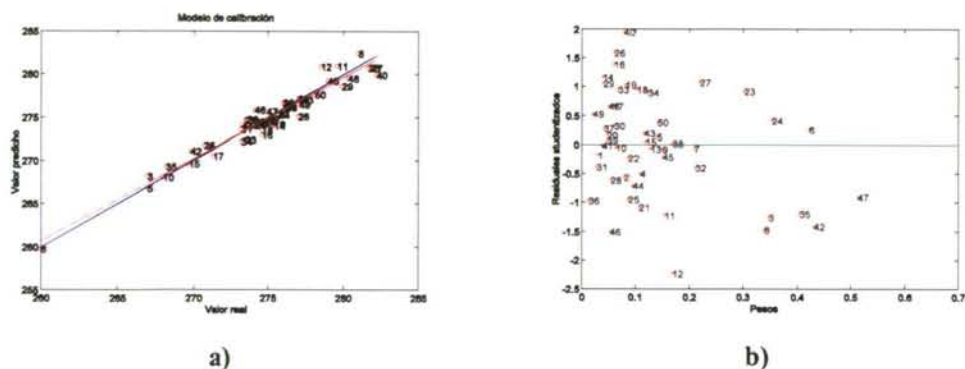


Figura 42: (a) Gráfico de T^2 y (b) Gráfico de Q para el punto final de destilación.

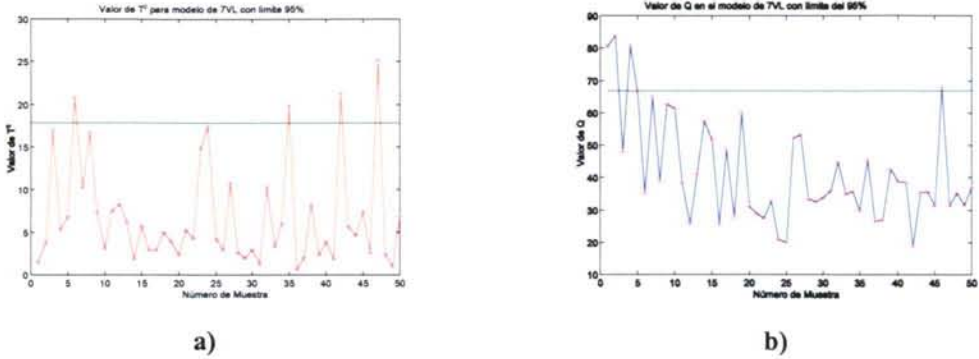


Figura 43: (a) Valor real del punto final de destilación frente al valor predicho para muestras de calibración predichos por el modelo con 7 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

En las **Figuras 44 y 45** se observan las relaciones entre los scores de los datos X e Y y las gráficas de scores, respectivamente, para el modelo elegido de PLS donde, salvo el ya señalado comportamiento de las muestras 7, 8 y 9, y la existencia de alguna muestra conflictiva, no se observan aspectos más relevantes.

A pesar de que la regresión “valor real frente a valor predicho” no es tan buena como en las propiedades anteriores, la predicción obtenida es francamente satisfactoria ($SEC-CV-LOO$ ($n_c = 50$) = 3.99 °C; SEP ($n_v = 35$) = 3.99 °C), especialmente si se recuerda que la reproducibilidad del método ASTM es 10.5 °C.

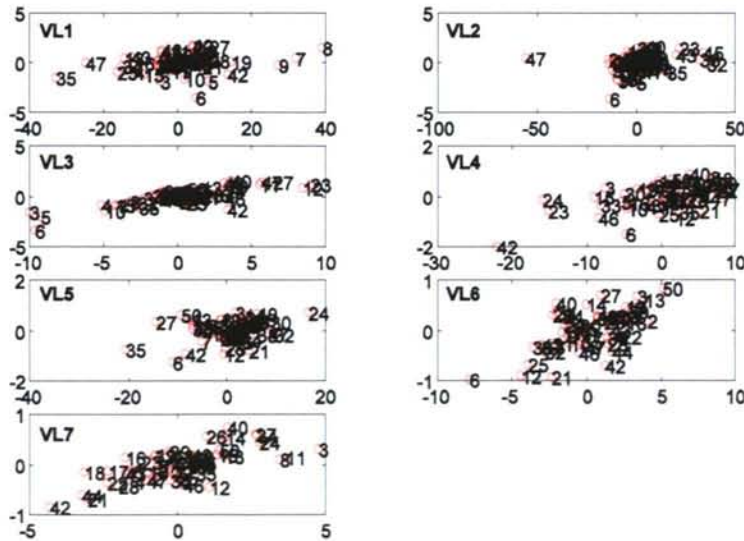


Figura 44: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (7 VL), punto final de destilación.

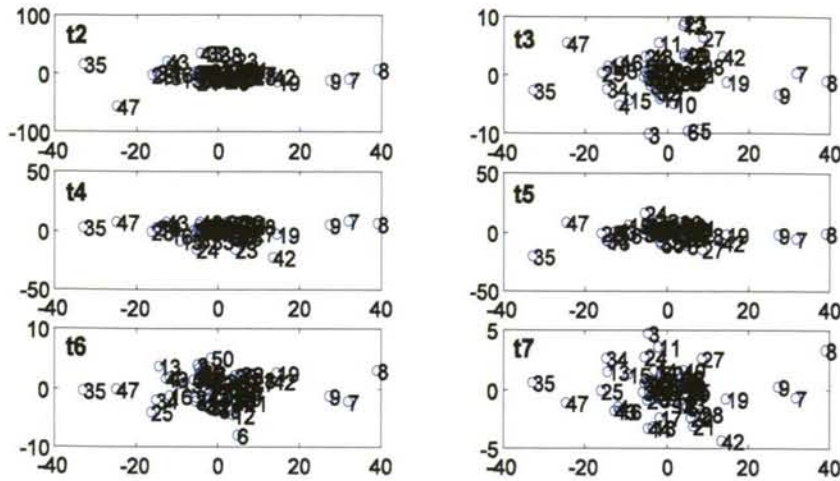


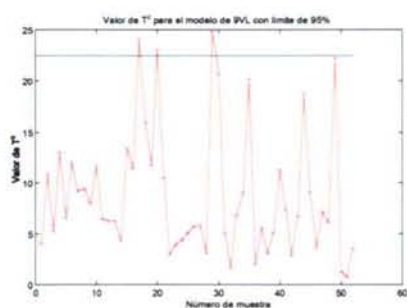
Figura 45: Relaciones entre los scores para el modelo PLS (7 VL), punto final de destilación (abscisas, t1).

1.1.7. PORCENTAJE DE AROMÁTICOS

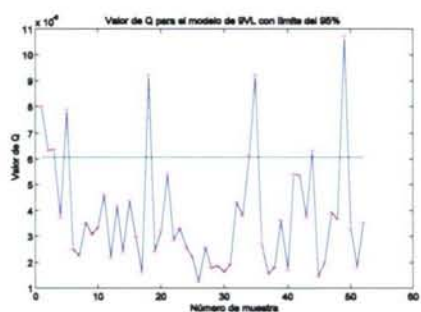
Se ha aplicado la regresión multivariante PLS con objeto de predecir el porcentaje de aromáticos para las muestras de queroseno en fase vapor, obteniéndose un modelo bueno para 9 VL, que explican un 99.91% de la varianza para las X y un 93.95% de la varianza para las Y.

En las Figuras 46 y 47 se muestran las gráficas obtenidas para el modelo PLS. En las Figuras 48 y 49 se observan las relaciones entre los scores de los datos X e Y para el modelo PLS elegido y las gráficas de scores, respectivamente, para el porcentaje de aromáticos. En la Figura 46 se observa que las muestras 17, 20 y 29 (en (a)) y 1, 2, 3, 5, 18, 34, 35, 44 y 49 (en (b)) podrían ser anómalas, pero esto se desmiente al observar la Figura 47, ya que estas muestras no se desvían mucho del valor real y no destacan en cuanto a los errores normalizados, por lo tanto no serán consideradas como anómalas y seguirán siendo constituyentes del modelo de predicción obtenido.

Aunque hay alguna muestra cuyo comportamiento se aleja del promedio, se observa que las Figuras 46 y 47 no revelan ningún anómalo claro, por lo que las predicciones obtenidas se consideran fiables (SEC-CV-LOO ($n_c = 52$) = 1.10%; SEP ($n_v = 46$) = 1.29%). Comparando con la Tabla I (reproducibilidad ASTM = 2.7%), las predicciones son buenas.

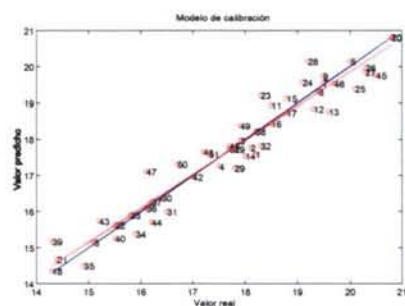


a)

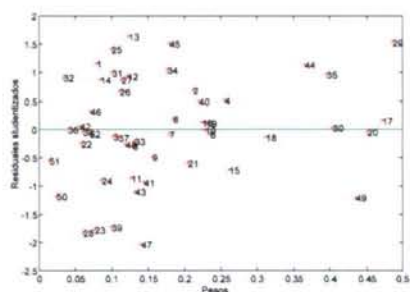


b)

Figuras 46: (a) Gráfico de T^2 para porcentaje de aromáticos y (b) Gráfico de Q .



a)



b)

Figura 47: (a) Valor real del porcentaje de aromáticos frente al valor predicho para muestras de calibración medidas por el modelo obtenido (9 VL) (línea azul= comportamiento ideal, línea roja= comportamiento experimental y (b) Gráfica de los “pesos frente a residuales estudentizados”.

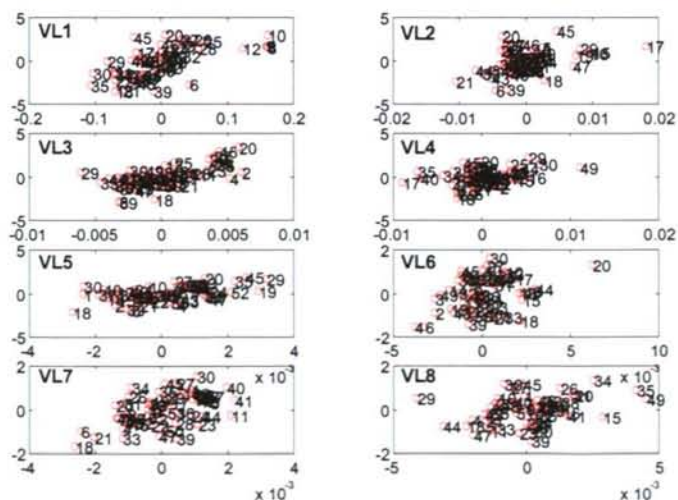


Figura 48: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS elegido para el porcentaje de aromáticos (9 VL).

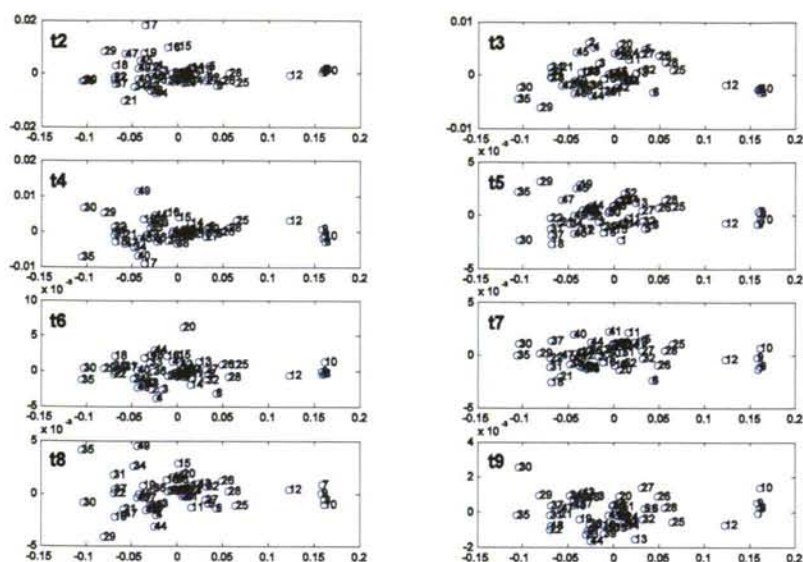


Figura 49: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS elegido para el porcentaje de aromáticos (9 VL) (abscisas, t1).

1.1.8. VISCOSIDAD

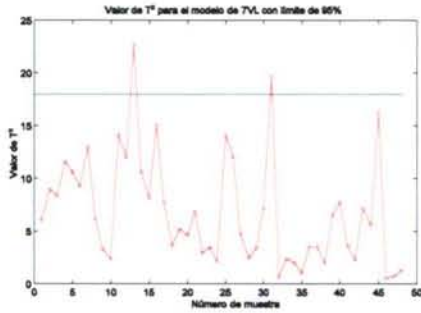
Para predecir la viscosidad de las muestras de queroseno en fase vapor se precisaron 7 VL, que explican un 99.83% de la varianza en X y un 92.99% de la varianza en Y (datos centrados en la media). Las predicciones obtenidas no son demasiado buenas (con un SEC-CV-LOO ($n_c = 48$) = 0.19 cSt; SEP ($n_v = 41$) = 0.13 cSt) (reproducibilidad ASTM = 0.04 cSt).

En la Figura 50 se observan varios posibles anómalos como son las muestras 13 y 31 (en el gráfico (a)) y las 25, 36, 40 y 45 (en el gráfico (b)), aunque no se ven como anómalos claros en la Figura 51, 52 ni 53, por lo tanto no se eliminaron del modelo, aunque éste no es demasiado bueno para la predicción de esta propiedad. Esto puede atribuirse a:

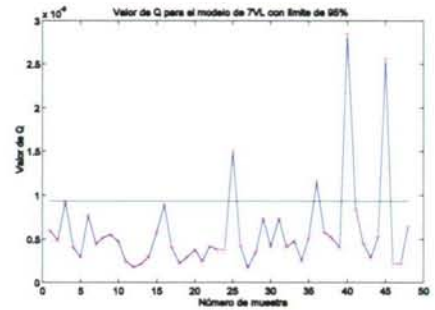
- 1.- La viscosidad se define en función del uso de un viscosímetro de vidrio a una temperatura determinada (-20°C) y ni en las medidas IR (a temperatura ambiente) ni en los modelos se pudieron introducir variables predichas que tuviesen en cuenta que las medidas espectrales se realizan a aproximadamente 20°C (temperatura del laboratorio).
- 2.- El método ASTM tiene una precisión y exactitud excelentes con las cuales es difícil competir. A esta misma situación se había llegado en otros estudios

abordados para querosenos (Garrigues *et al.*, 1995).

En las **Figuras 50** y **51** se muestran las gráficas obtenidas para el modelo PLS con 7 variables latentes. En las **Figuras 52** y **53** se observan las relaciones entre los *scores* de los datos X e Y para el modelo PLS elegido y las gráficas de *scores* para el porcentaje de aromáticos, respectivamente. Se observa que las muestras 5, 6 y 7 tienen valores elevados de viscosidad (las muestras 8 y 9 de modelos anteriores debieron eliminarse de éste por los muchos problemas que causaban).

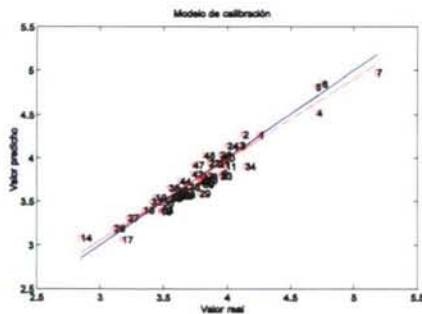


a)

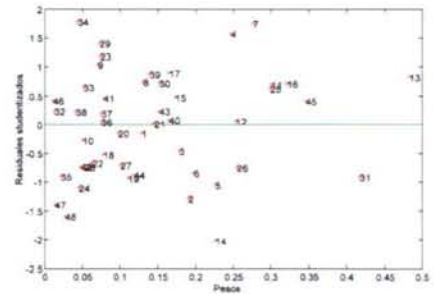


b)

Figuras 50: (a) Gráfico de T^2 que muestra la no existencia de muestras anómalas en el calibrado para la viscosidad y (b) Gráfico de Q que muestra algunas muestras alejadas del promedio.



a)



b)

Figura 51: (a) Valor real de la viscosidad frente al valor predicho para muestras de calibración predichas por el modelo con 7 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

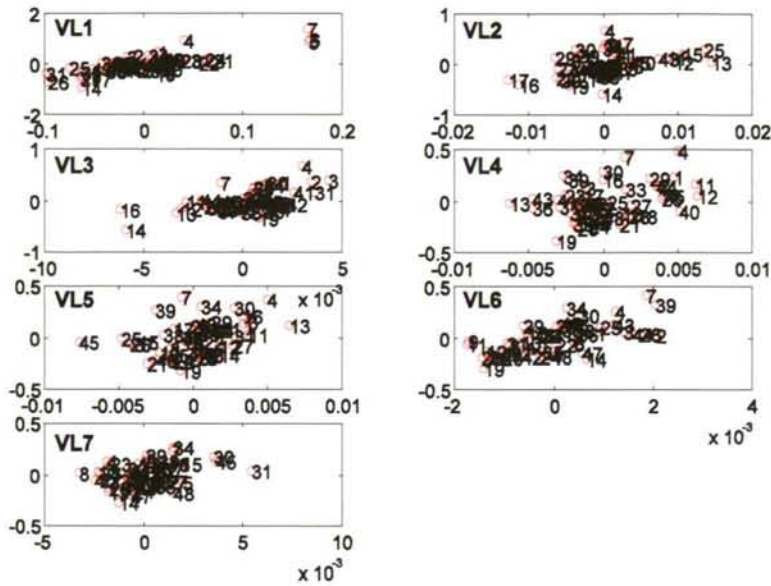


Figura 52: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (7 VL), viscosidad.

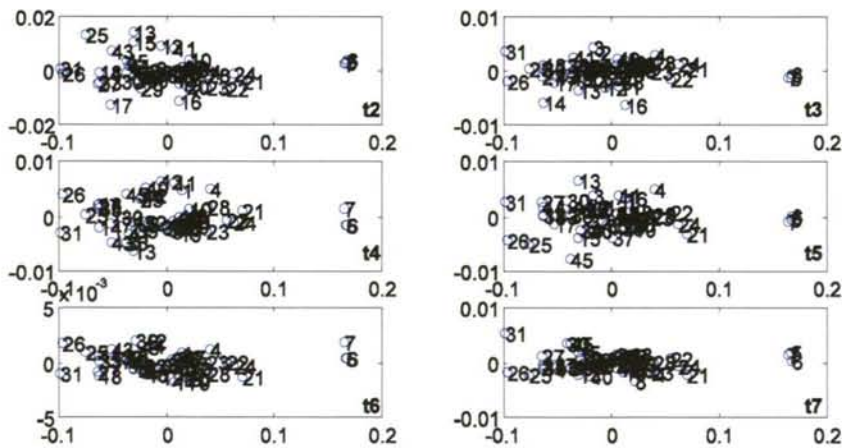


Figura 53: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (7 VL), viscosidad (abscisas, tk).

1.2. MODELOS MULTIVARIANTES CON EL SISTEMA DE MEDIDA COMPLEJO

1.2.1. PUNTO DE DEFLAGRACIÓN

El modelo PLS para la predicción del punto de deflagración de queroseno a partir de espectros IR mediante el sistema de la celda para gases precisa 9 VL que explican en la matriz **X**, el 99.98% de la varianza y en la **Y**, el 93.10%. Se obtuvieron valores de SEC-CV-LOO ($n_c = 56$) de 1.71 °C y el SEP de validación de 1.51 °C con $n_v = 32$, algo mayores de lo que sería deseable (reproducibilidad IP= 1.5 °C).

En las **Figuras 54** y **55** se muestran las gráficas obtenidas para este modelo. Las muestras que parecen anómalas en la **Figura 54** no afectan al comportamiento del modelo (se probó a su eliminación), simplemente se corresponden con valores límite de la propiedad (valores muy altos o muy bajos); por ejemplo, la muestra 3 posee un valor alto de punto de deflagración (64.5 °C), pero es conveniente no quitarla para permitir que el modelo pueda modelizar bien los valores altos de esta propiedad.

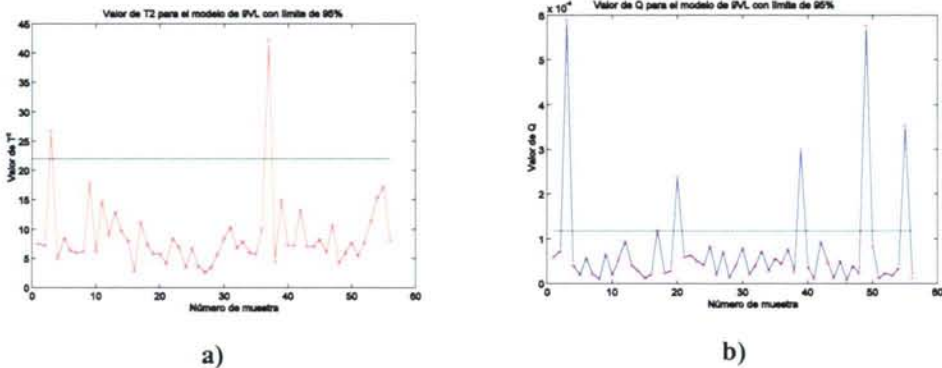


Figura 54: (a) Gráfico de T^2 y (b) Gráfico de Q , punto de deflagración.

En las **Figuras 56** y **57** se muestran las relaciones entre los scores de los datos **X** e **Y** para el modelo PLS elegido para esta propiedad (mostrándose sólo 8 VL de las 9 VL que tiene el modelo). Además de las muestras con valores altos ó bajos de la propiedad, destacan algunas otras que deben tener alguna característica espectral diferente al resto. Tal es el caso de la #37 (la cual marca a la sexta VL) pero no se ha encontrado el motivo. Como su eliminación no mejoraba el modelo y su predicción era buena (error estudentizado es casi cero) se decidió mantenerla (*leverage* alto pero bien modelada).

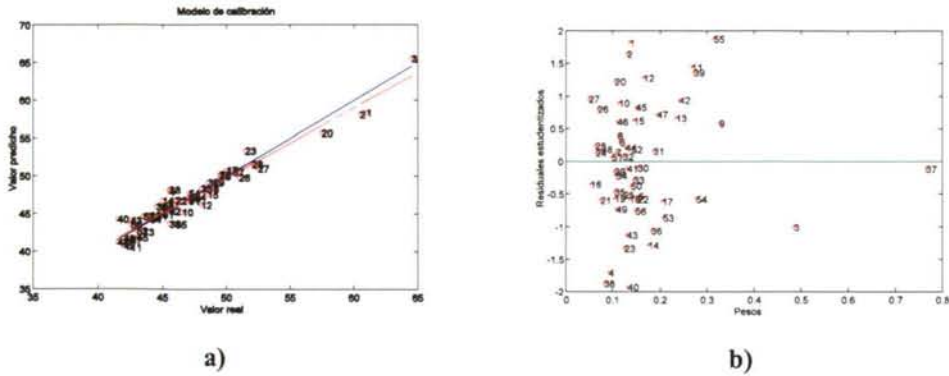


Figura 55: (a) Valor real del punto de deflagración frente al valor predicho para muestras de calibración predichos por el modelo con 9 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estandarizados”.

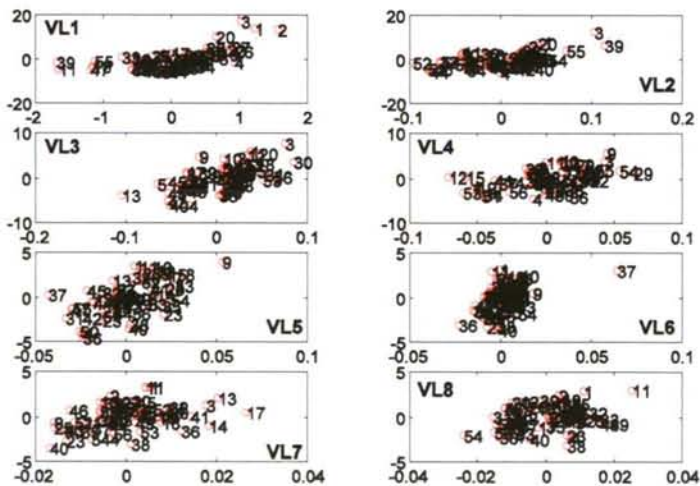


Figura 56: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (9 VL), punto de deflagración.

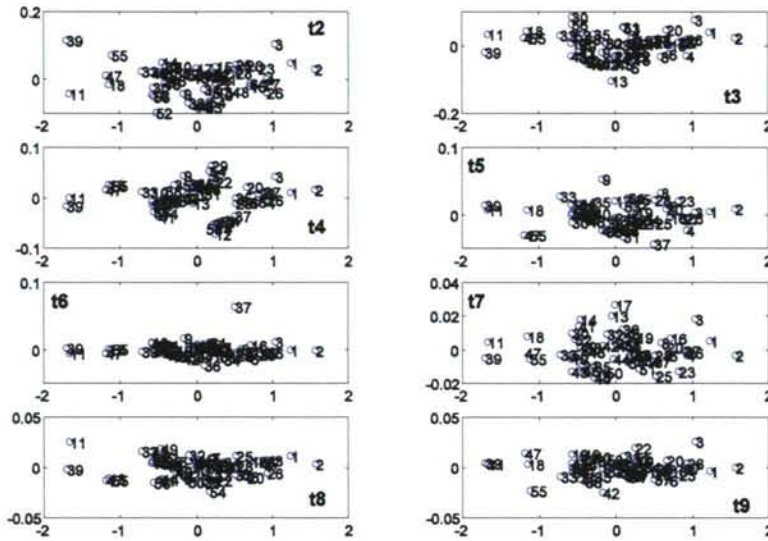


Figura 57: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (9 VL), punto de deflagración (abscisas, t1).

1.2.2. PUNTO DE CONGELACIÓN

Para el punto de congelación fue necesario desarrollar un modelo con 13 VL en el que se explica el 99.99% de la varianza en X y el 68.24 en Y, obteniendo de esta forma un modelo cuyo SEC-CV-LOO ($n_c=58$) = 1.23 y SEP ($n_v=35$) = 1.36 (reproducibilidad ASTM = 2.5 °C).

En la Figura 58 y 59 se observan algunas muestras posibles sospechosas en el modelo escogido. No se encuentra relación entre los posibles anómalos de las Figuras 58 ((a) y (b)) y las de la 59 (a) y (b). La Figura 58 (a) no es buena, pero no se consiguió mejorar; ni tan siquiera al eliminar las muestras sospechosas (3, 13, 27 y 45).

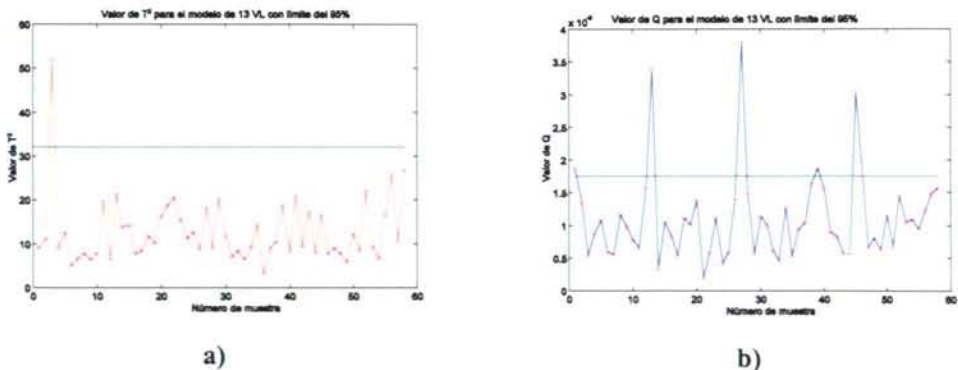


Figura 58: (a) Gráfico de T^2 y (b) Gráfico de Q, punto de congelación.

La muestra 3 ya había presentado un comportamiento especial en la sección anterior, aquí es modelada correctamente (error estudentizado aproximadamente cero, leverage alto).

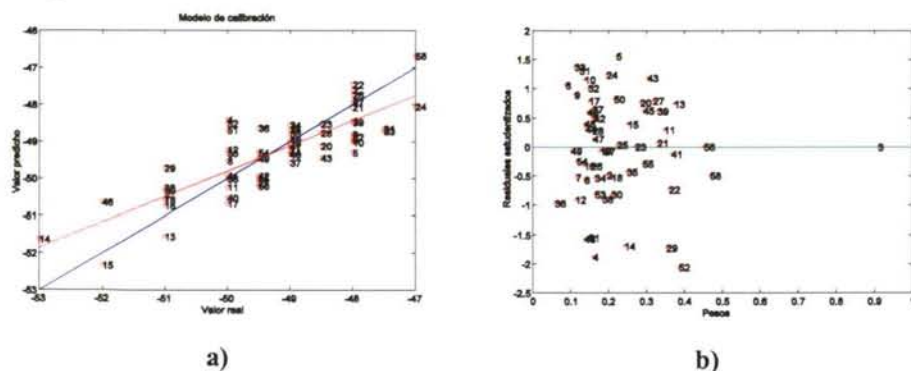


Figura 59: (a) Valor real del punto de congelación frente al valor predicho para muestras de calibración predichas por el modelo con 13 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

En las Figuras 60 y 61 se muestran las relaciones entre los scores de los datos X e Y para el modelo PLS elegido y las gráficas de scores, respectivamente para las primeras VL. En las gráficas (y en las otras no presentadas aquí) no se aprecia ningún comportamiento especial.

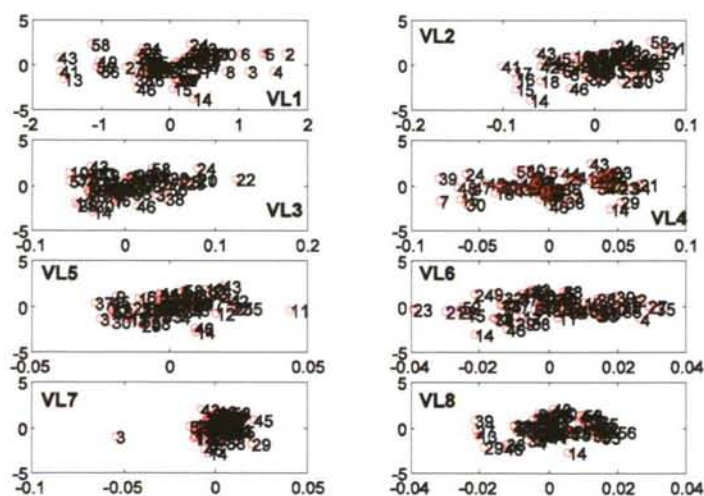


Figura 60: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (13 VL, se muestran sólo las primeras), punto de congelación.

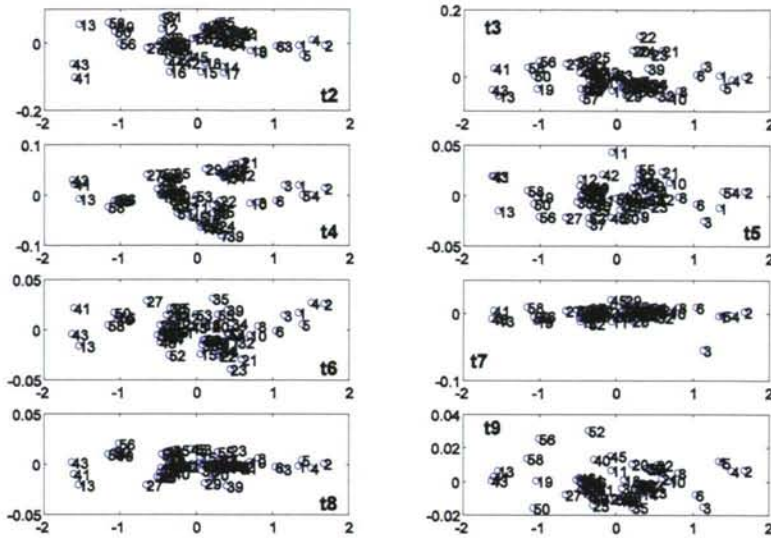


Figura 61: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (13 VL), punto de congelación (abscisas, t1).

1.2.3. PUNTO INICIAL DE DESTILACIÓN

Para el punto inicial de destilación se ha escogido un modelo con 13 VL en donde se explica el 99.99% de varianza en X y el 89.43% en Y. A continuación se evalúa la bondad de este modelo, que tiene un SEC-CV-LOO ($n_c = 54$) de 6.21 °C y un error estandar de validación (SEP) para las 34 muestras de validación de 3.88 °C, valores satisfactorios (reproducibilidad ASTM = 8.5 °C).

En la Figura 62 se presentan los índices de T^2 y Q. La Figura 63 (a) representa el valor real (obtenido por el método clásico ASTM D-86) frente al valor predicho mediante el modelo en el proceso de calibración. Se observa que los valores predichos se ajustan bastante bien al valor real y, por tanto, se puede aceptar que las predicciones obtenidas son buenas. En la Figura 63 (b) se muestra la gráfica de “pesos frente a residuales estudentizados”. Se observa que no hay anómalos evidentes, ya que las muestras que parecen anómalas en la Figura 62, no afectan al comportamiento del modelo, simplemente se corresponden con valores límite de la propiedad (valores muy altos o muy bajos); por ejemplo, la muestra 1, posee un valor alto de punto inicial de destilación (180°C), pero es conveniente no quitarla (aunque no se prediga su valor de forma “perfecta”) para permitir que el modelo pueda modelizar bien los valores altos de esta propiedad.

En las Figuras 64 y 65 se muestran las relaciones entre los scores de los datos X e Y para el modelo PLS elegido y las gráficas de scores, respectivamente.

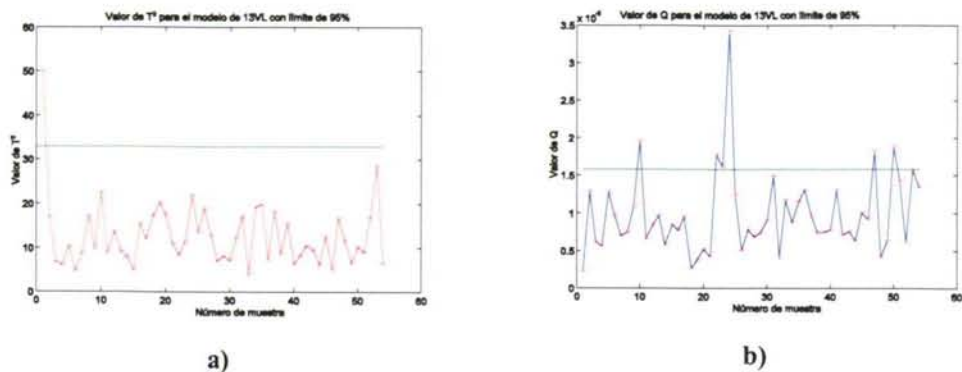


Figura 62: (a) Gráfico de T^2 y (b) Gráfico de Q, punto inicial de destilación.

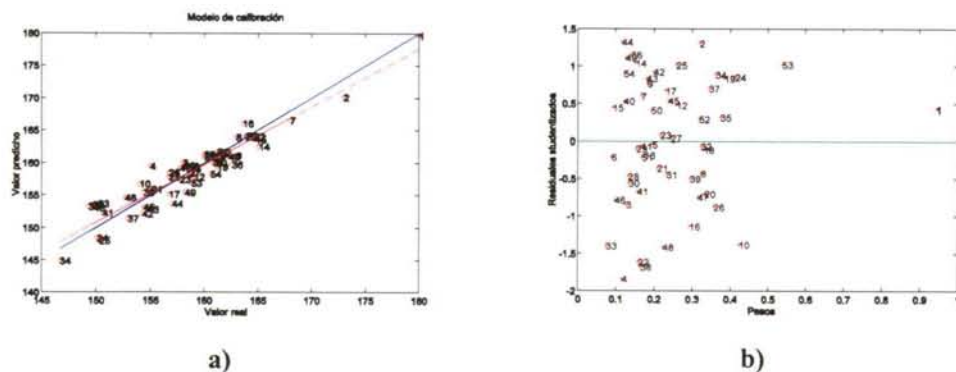


Figura 63: (a) Valor real del punto inicial de destilación frente al valor predicho para muestras de calibración predichas por el modelo con 13 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

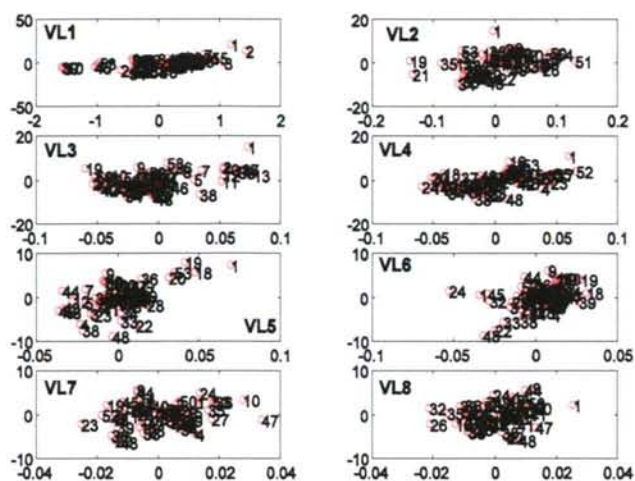


Figura 64: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (13 VL, se muestran sólo las primeras), punto inicial de destilación.

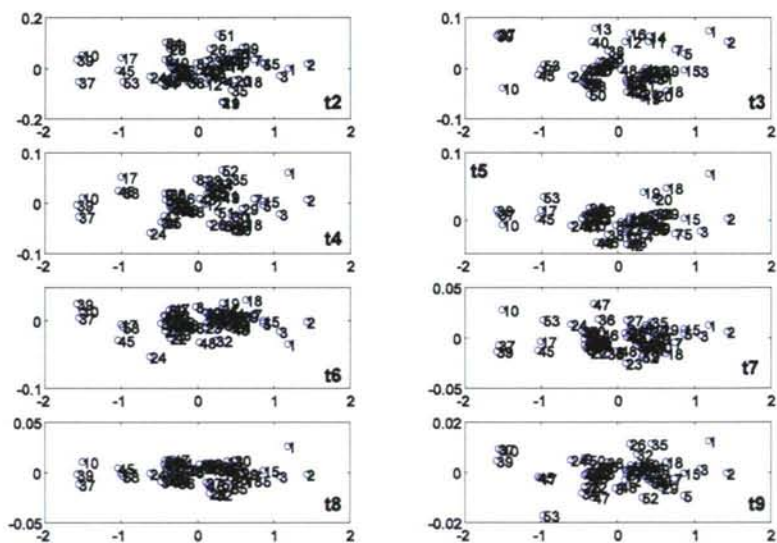


Figura 65: Distribución de las muestras en el espacio de las variables predictoras (scores t_1 vs t_k) para el modelo PLS (13 VL), punto inicial de destilación (sólo se muestran las primeras; abscisas, t_1).

1.2.4. PUNTO DE 10% DE DESTILACIÓN

El modelo seleccionado para el 10% de destilación explica el 99.99% de la varianza en el bloque de las X y el 91.76% en el bloque de las Y haciendo uso de 11 VL (con SEC-CV-LOO ($n_c = 58$) = 2.29 °C y SEP ($n_v = 38$) = 2.07 °C, errores promedio que son satisfactorios ya que la reproducibilidad del procedimiento ASTM es 8.8 °C). En las Figuras 66 y 67 se analiza la posible existencia de anómalos. Aunque la muestra 1 podría parecer un anómalo, se confirma que no lo es aunque sí constituye un punto influyente debido a que tiene el valor experimental más elevado de esta propiedad (193 °C). De hecho, resulta una muestra que permite expandir la modelización a valores un poco más elevados de lo habitual (como sucedía en el epígrafe anterior).

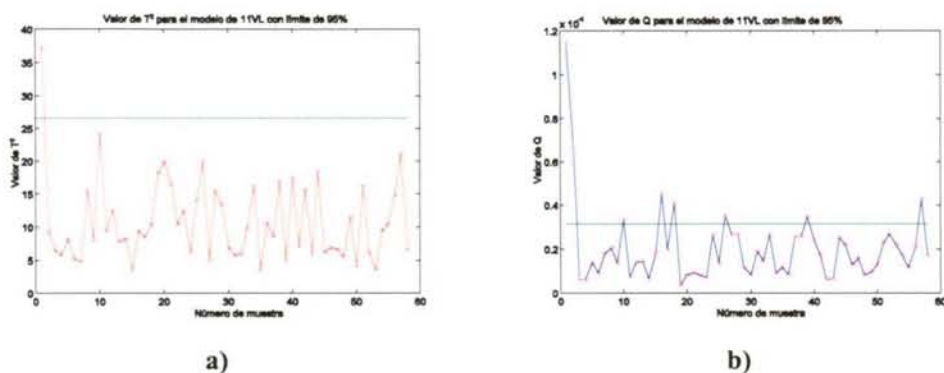


Figura 66: (a) Gráfico de T^2 y (b) Gráfico de Q, 10% de destilación.

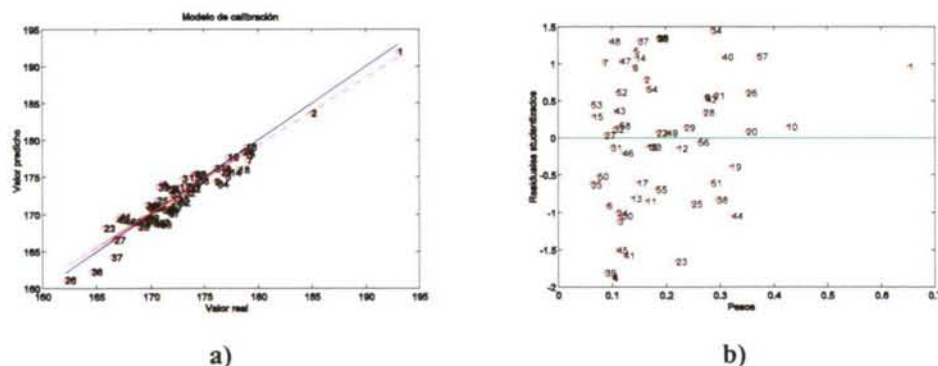


Figura 67: (a) Valor real del 10% de destilación frente al valor predicho para muestras de calibración predichas por el modelo con 11 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

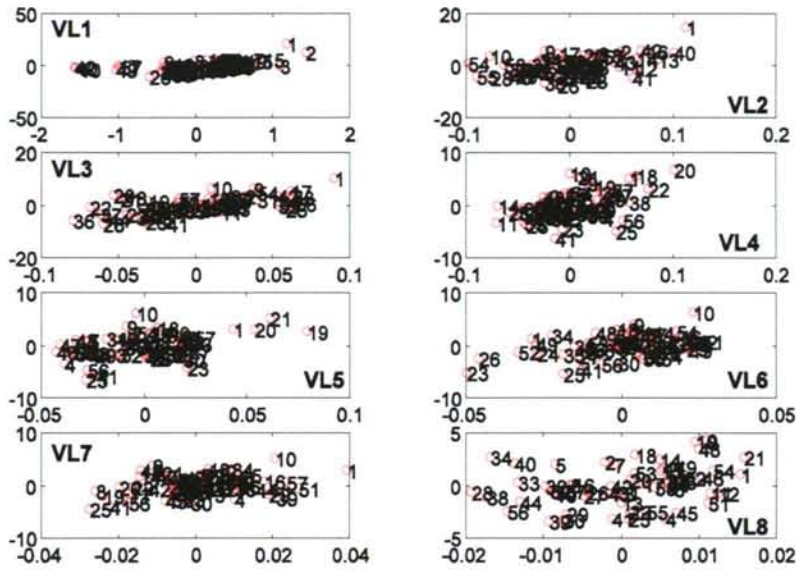


Figura 68: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (11 VL), 10% de destilación (sólo se muestran las primeras).

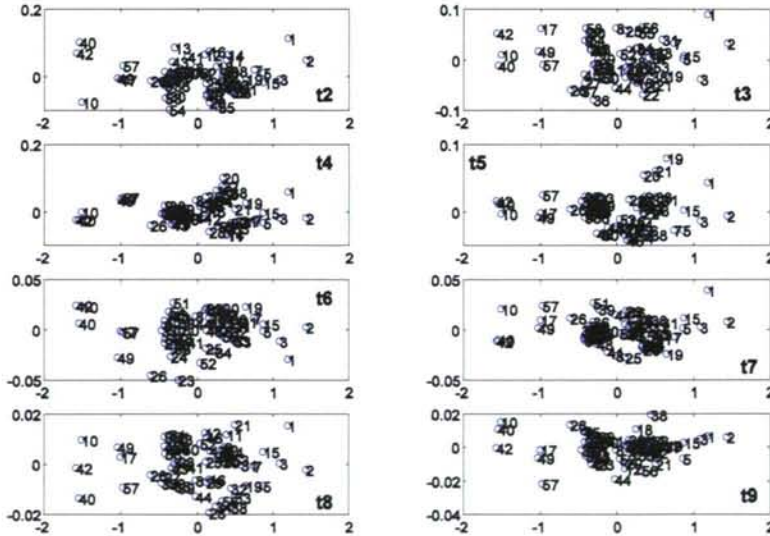


Figura 69: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (11 VL), 10% de destilación (sólo se muestran las primeras; abscisas, t1).

1.2.5. PUNTO DEL 90% DE DESTILACIÓN

El modelo seleccionado para dicha propiedad es de 13 VL (empleando autoescalado de los espectros), que explican un 99.96% y 90.82% de la varianza en el bloque de las X y las Y, respectivamente (con SEC-CV-LOO ($n_c=56$) = 3.04°C y SEP ($n_v=32$) = 2.83°C, errores promedio satisfactorios). En las Figuras 70 y 71 se estudia la posible existencia de anómalos encontrándose un valor de Q elevado para las muestras 44, 45 y 56. Al eliminar esas muestras del modelo, no se ha encontrado una mejora, por lo que no fueron eliminadas del modelo.

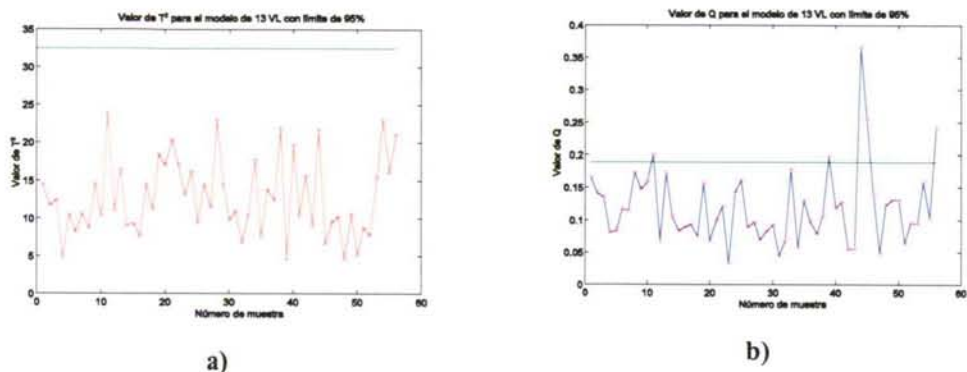


Figura 70: (a) Gráfico de T^2 y (b) Gráfico de Q, 90% de destilación.

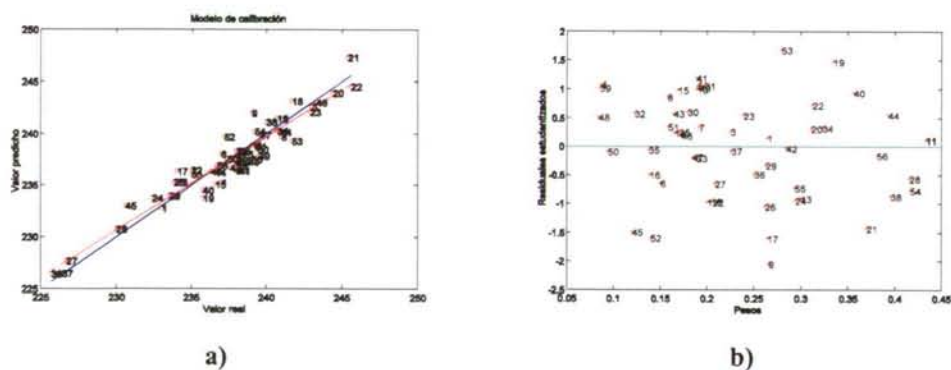


Figura 71: (a) Valor real del 90% de destilación frente al valor predicho para muestras de calibración predichas por el modelo con 13 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

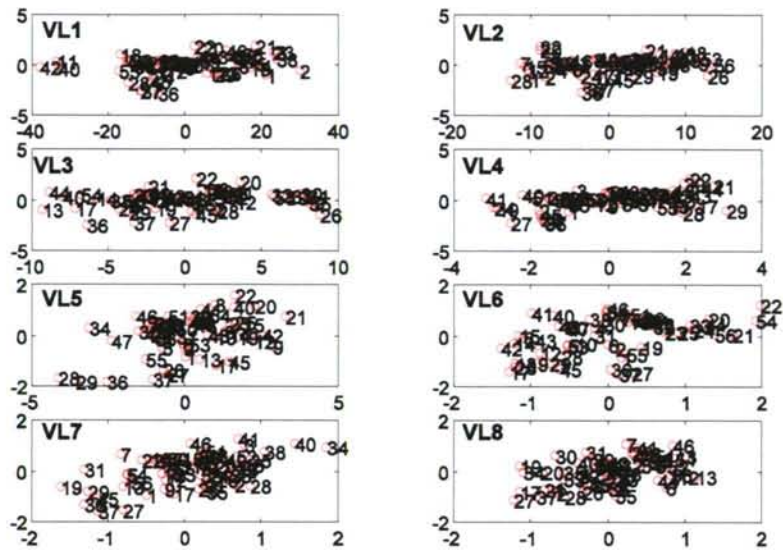


Figura 72: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (13 VL), 90% de destilación (sólo se muestran las primeras).

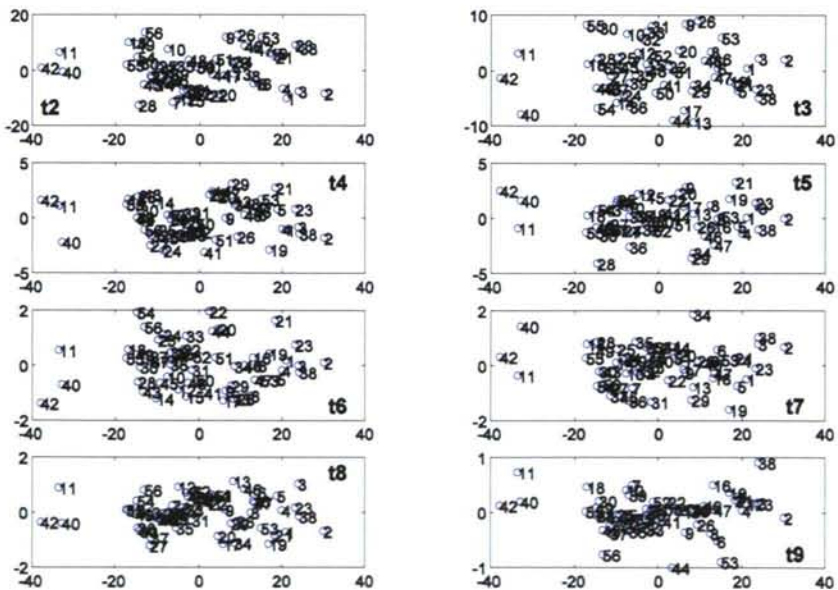


Figura 73: Distribución de las muestras en el espacio de las variables predictoras (scores t_1 vs t_k) para el modelo PLS (13 VL), 90% de destilación (sólo se muestran las primeras; abscisas, t_1).

1.2.6. PUNTO FINAL DE DESTILACIÓN

Se ha aplicado la metodología PLS para predecir el punto final de destilación para muestras de queroseno en fase vapor medidas empleando la celda de gases, obteniéndose un modelo bueno para 14 VL, que logra explicar un 99.96% de la varianza en las X y un 90.17% de la varianza en las Y (autoescalado de los datos).

En las Figuras 74 y 75 se muestran las gráficas obtenidas para evaluar el modelo PLS con 14 variables latentes.

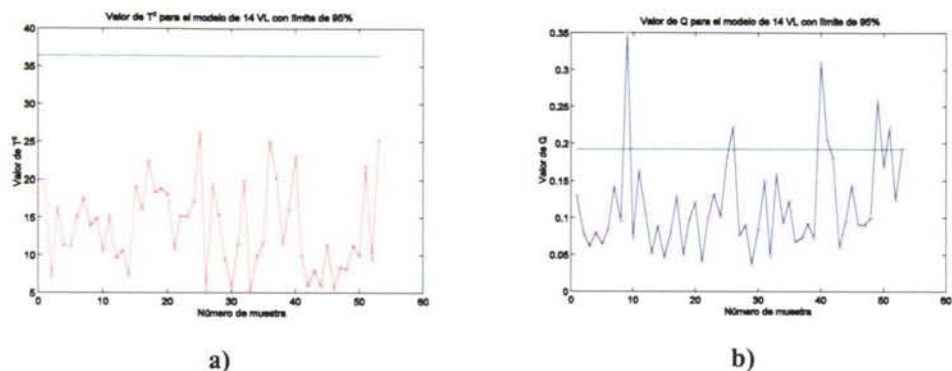


Figura 74: (a) Gráfico de T^2 que muestra la no existencia de muestras anómalas en el calibrado para el punto final de destilación (fbp) y (b) Gráfico de Q que muestra algunas muestras alejadas del promedio.

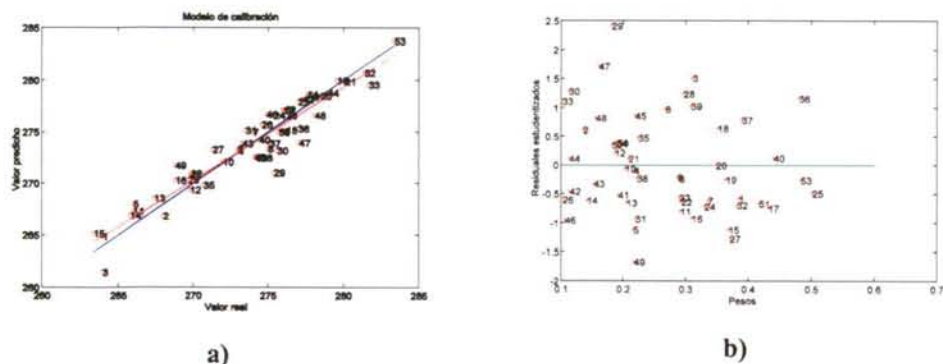


Figura 75: (a) Valor real del punto final de destilación frente al valor predicho para muestras de calibración predichas por el modelo con 14 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de “pesos frente a residuales estudentizados”.

En las Figuras 76 y 77 se observan las relaciones entre los scores de los datos X e Y para el modelo PLS elegido y las gráficas de scores, respectivamente, para el punto final de destilación (se dibujan sólo las primeras VL). El modelo escogido conduce a SEC-CV-LOO ($n_c=53$) = 4.34 °C y SEP ($n_v=32$) = 6.74 °C, mucho mejores que la reproducibilidad del método ASTM (= 10.5 °C).

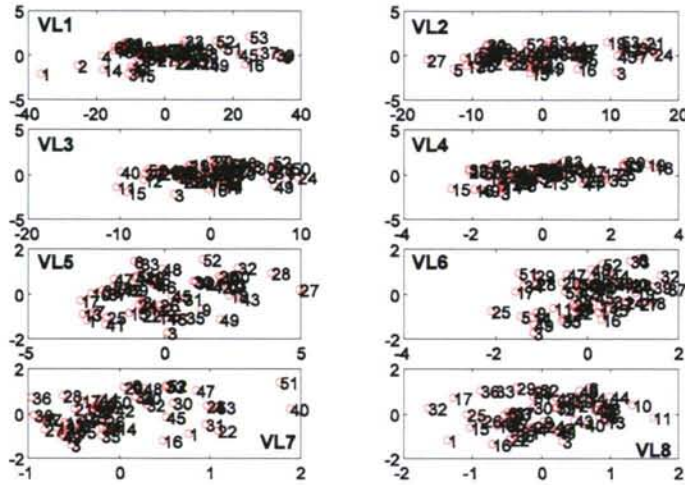


Figura 76: Relaciones entre los scores X e Y para el modelo PLS (14 VL), punto final de destilación (sólo se muestran las primeras).

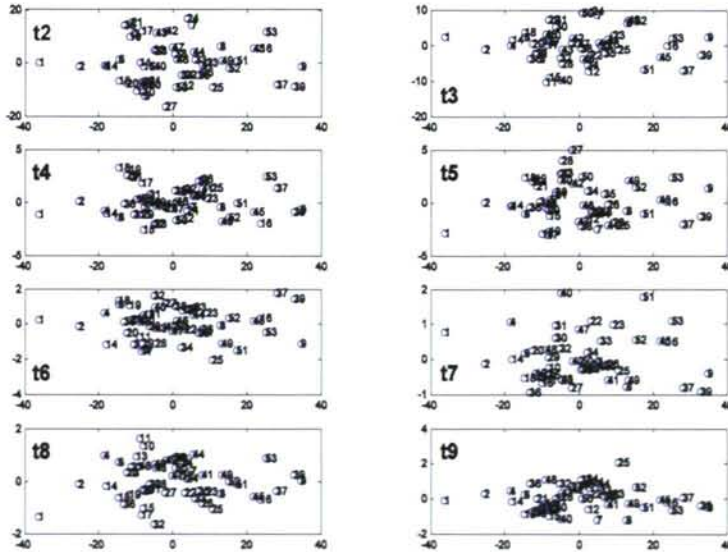


Figura 77: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (14 VL), punto final de destilación (sólo se muestran las primeras; abscisas t1).

1.2.7. PORCENTAJE DE AROMÁTICOS

Se recogen a continuación las gráficas que permiten evaluar la bondad del modelo PLS escogido para predecir el porcentaje de aromáticos empleando 13 VL (Figuras 78, 79, 80 y 81).

Se explica el 99.99% de varianza en el bloque de las X y el 90.71% en el bloque de las Y. Se aprecia que ninguna muestra tiene un comportamiento anómalo y que la relación obtenida real-predicho es buena. El SEC-CV-LOO ($n_c = 58$) es de 1.24% y el SEP en muestras de validación ($n_v = 32$) fue bueno, con valor de 0.77%. Los modelos conducen a predicciones promedio excelentes (reproducibilidad ASTM = 2.7%).

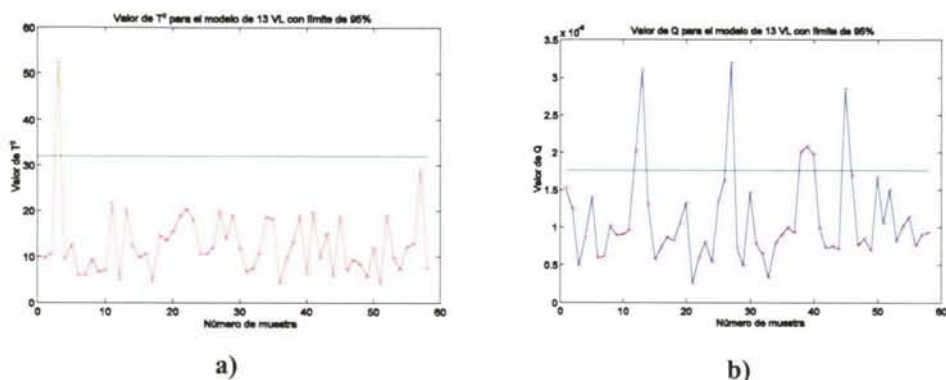


Figura 78: (a) Gráfico de T^2 y (b) Gráfico de Q, porcentaje de aromáticos.

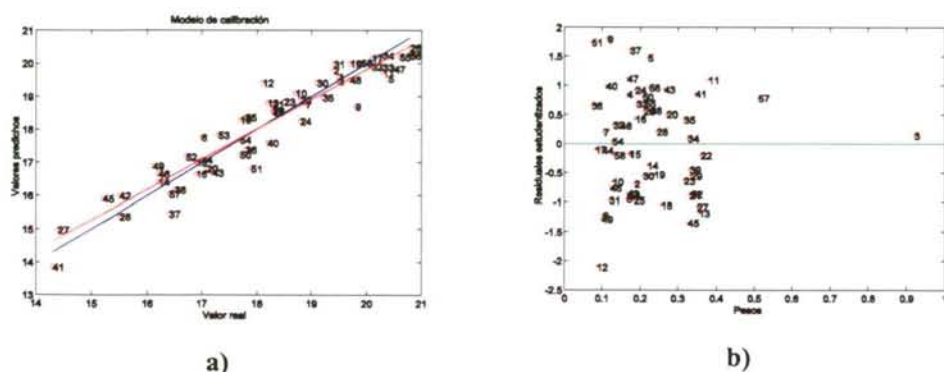


Figura 79: (a) Valor real del porcentaje de aromáticos frente al valor predicho para muestras de calibración predichas por el modelo con 13 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de "pesos frente a residuales estudentizados".

Se observa que no hay anómalos evidentes, ya que las muestras que parecen anómalas en la **Figura 78**, no lo corroboran ni la **Figura 79** ni la **80**. El comportamiento de la muestra 3 en la **Figura 79b** coincide con el observado en secciones previas (*leverage* alto pero buena modelización).

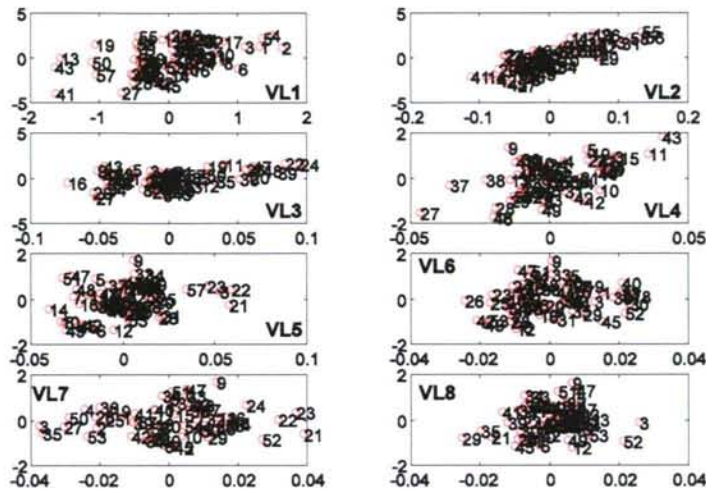


Figura 80: Relaciones entre los *scores* X (en abscisas) e Y (en ordenadas) para el modelo PLS (13VL), porcentaje de aromáticos (sólo se muestran las primeras).

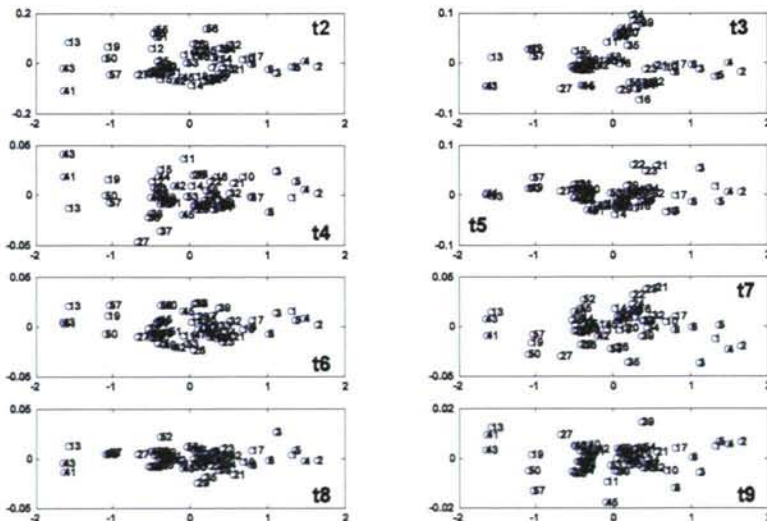


Figura 81: Distribución de las muestras en el espacio de las variables predictoras (*scores* t1 vs tk) para el modelo PLS (13 VL), porcentaje de aromáticos (sólo se muestran las primeras; abscisas, t1).

1.2.8. VISCOSIDAD

La predicción de la viscosidad para las muestras de queroseno en fase vapor medidas empleando la celda de gases, necesita hasta 11 VL que explican un 99.99% de la varianza en las X y un 93.20% de la varianza en las Y. Este número tan elevado de VL que nunca fue necesario en otros casos para modelizar este parámetro casos debe atribuirse a la dificultad de modelización de este parámetro físico y los bajos errores que se permiten en el método ASTM D-445. El SEC-CV-LOO fue de 0.12 cSt ($n_c=58$) y el SEP ($n_v=35$) de 0.12 cSt. A pesar de que estos modelos tienen un rango menor de trabajo que los de la sección 1.1.8. (se eliminaron las muestras con valores más elevados) y, efectivamente, el error es ligeramente menor, los modelos siguen arrojando errores promedio mayores que la reproducibilidad de ASTM (incluso considerando 11 VL), lo que puede atribuirse a las mismas situaciones que las señaladas en el epígrafe correspondiente al uso del sistema más simple (Parte C, 1.1.8.).

En las Figuras 82, 83, 84 y 85 se muestran las gráficas obtenidas para el modelo PLS escogido con 11 variables seleccionadas.

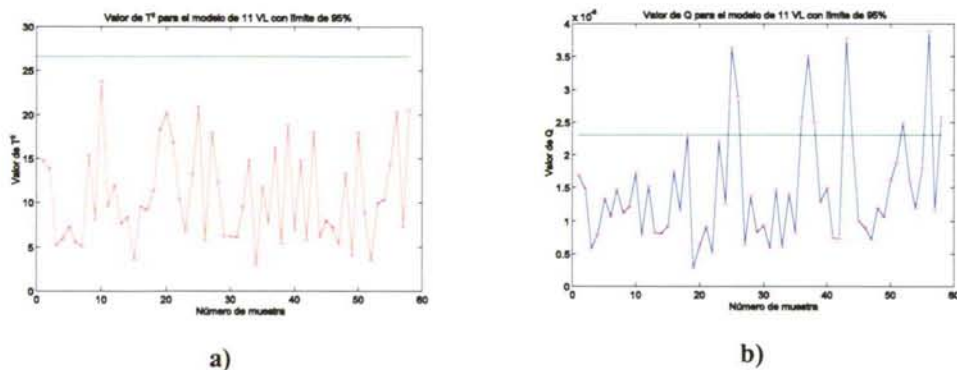


Figura 82: (a) Gráfico de T^2 y (b) Gráfico de Q, viscosidad.

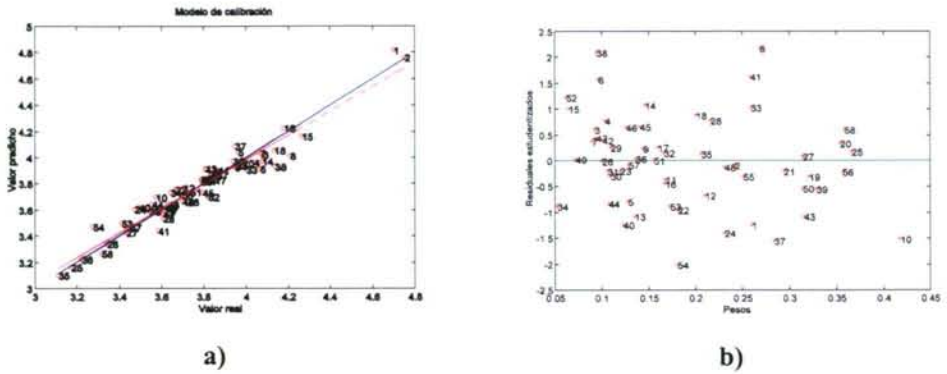


Figura 83: (a) Valor real de la viscosidad frente al valor predicho para muestras de calibración predichas por el modelo con 11 VL (línea azul= comportamiento ideal, línea roja= comportamiento experimental) y (b) Gráfica de "pesos frente a residuales estandarizados".

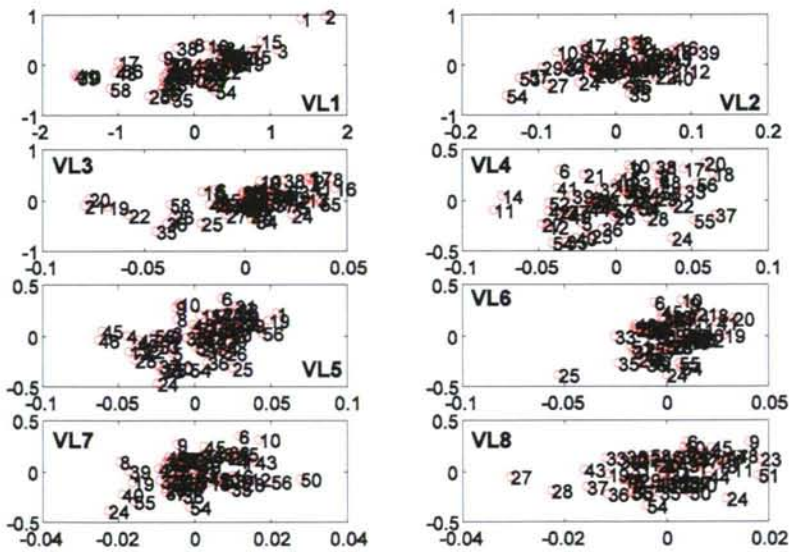


Figura 84: Relaciones entre los scores X (en abscisas) e Y (en ordenadas) para el modelo PLS (11 VL), viscosidad (sólo se muestran las primeras).

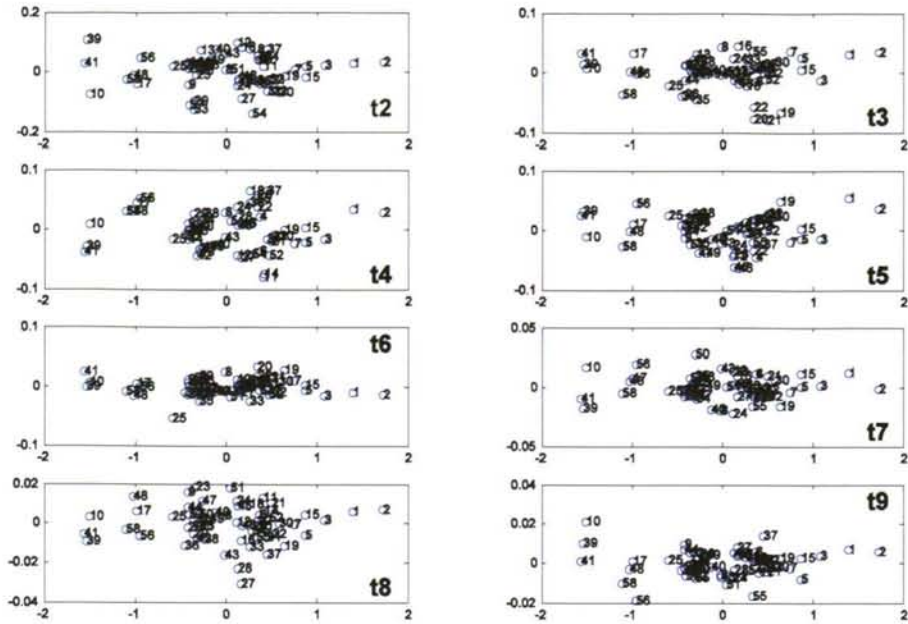


Figura 85: Distribución de las muestras en el espacio de las variables predictoras (scores t1 vs tk) para el modelo PLS (11 VL), viscosidad (sólo se muestran las primeras; abscisas, t1).

1.3. EXACTITUD Y PRECISIÓN

Las metodologías propuestas para la predicción de las ocho propiedades del queroseno empleando espectroscopia FTMIR-ATR y PLS son exactas, lo cual se pone de manifiesto a través del estudio de los estadísticos SEP (ver **Tabla VI**) y los test F para la ordenada en el origen y la pendiente (simultáneamente iguales a 0 y 1). Dicha tabla resume los modelos de predicción de PLS para cada propiedad y cada sistema experimental. En general se observan bajos errores promedio (menores que la reproducibilidad del método oficial) y buenos valores de precisión (r y R), que son casi siempre mejores que los métodos estándar.

De la tabla, se deduce que las metodologías usadas para realizar las medidas espectrales conducen a errores promedio similares. Como se ha ido comentando, en este trabajo se ha tomado como criterio que los modelos que se consideren adecuados para realizar las predicciones deberán exhibir un error promedio (SEP) inferior a la reproducibilidad de los métodos “clásicos”. Es decir, los modelos que se seleccionen deben proporcionar una predicción cuyo error respecto al “valor oficial” sea inferior a la variabilidad máxima permitida para el propio procedimiento oficial. Esta circunstancia se cumplió siempre, excepto en la determinación de la viscosidad y del punto de deflagración Abel (sistema de celda de gases). En el caso de la viscosidad, no alcanzar este criterio se atribuye a que ésta viene definida en función del uso de un viscosímetro de vidrio y, por tanto, éste será el mejor sistema de medida de la misma. A pesar de esto, nótese que el error que se comete es tan sólo del 3.75% (lo cual indica que, en la práctica, aunque los modelos son buenos, la reproducibilidad del método oficial es difícil de superar).

En cuanto al punto de deflagración, el método de la celda de gases iguala la reproducibilidad del método IP-170. Sin embargo, el uso de la celda de cuarzo conduce a mejores resultados. Esto puede deberse a la mayor dificultad de manipulación del sistema basado en el flujo de gas.

PROPIEDAD	Valores de referencia		Medida con célula de cuarzo										Medida con célula de gases									
	r	R	VL	SEP	SEP _{cor}	n _s	r	R	F _{tab} (95%)	F _{tab} (99%)	F _{exp}	LV	SEP	SEP _{cor}	n _s	r	R	F _{tab} (95%)	F _{tab} (99%)	F _{exp}		
Abel	1.0	1.5	7	1.03	0.93	34	0.95	1.4	3.16	4.99	1.21	9	1.51	1.44	32	2.36	3.91	3.17	5.02	2.00		
Freezing	1.0	2.5	9	1.26	1.02	44	0.74	3.11	3.20	5.10	1.89	13	1.36	1.14	35	0.95	1.24	3.16	4.99	13.03		
IBP	3.5	8.5	2	3.37	2.24	43	1.17	2.62	3.20	5.11	1.74	13	3.88	2.95	34	3.18	4.58	3.18	5.03	3.07		
PD10	4.3	8.8	8	2.28	2.28**	46	2.39	3.12	3.18	5.05	0.24	11	2.07	2.07**	38	2.18	4.64	3.16	4.99	2.51		
PD90**	4.1	8.8	7	3.42	2.22	37	5.65	9.45	3.19	5.07	1.58	13	2.83	1.12	33	4.44	6.91	3.17	5.02	2.73		
FBP*	3.5	10.5	7	3.99	2.51	35	3.57	4.88	3.19	5.08	1.39	14	6.74	5.98	32	4.99	19.71	3.18	5.04	2.78		
% Aromáticos	1.3	2.7	9	1.29	1.02	46	1.19	2.15	3.18	5.05	1.61	13	0.77	0.77**	38	0.93	1.67	3.16	4.99	2.87		
Viscosidad	0.02	0.04	7	0.14	0.14	41	0.20	0.50	3.20	5.10	1.73	11	0.12	0.12	35	0.11	0.16	3.16	4.99	2.04		

* el SEP corregido es negativo, el error estándar de redondeo es mayor que el error de predicción

** Datos autocorrelacionados

Tabla VI: Modelos para cada una de las propiedades de los querosenos.

Destacar que las propiedades que dependen (principalmente) de las fracciones pesadas del queroseno (% de aromáticos, punto final de destilación y viscosidad) se predicen mejor (menos error y mejor precisión) haciendo uso del sistema de medida complejo. Esto puede explicarse porque la generación de la fase vapor es más efectiva que en la celda de cuarzo sólo. Por contra, el sistema simple es más efectivo para predecir el resto de propiedades porque éstas se asocian más fuertemente a las fracciones ligeras (las cuales se pueden modelizar sin problema en la celda de cuarzo). También resulta significativo que el número de factores en los modelos desarrollados con el sistema simple es menor que en el caso de emplear la celda de gases, posiblemente porque el sistema complejo implica una mayor variabilidad y, por tanto, más información incorrelada con la propiedad a estudiar.

Otro estadístico útil a la hora de la valoración del modelo es el test F en el que se determina si la ordenada y pendiente de la línea de regresión “real vs predicho” son estadísticamente y de forma simultánea iguales a 0 y 1, respectivamente, y si los modelos multivariantes conducen a predicciones sin desviación. En la **Tabla VI** se presentan las F teóricas (95 y 99% de probabilidad) y se observó que sólo el modelo multivariante desarrollado para predecir el punto de congelación usando la celda de gases (sistema complejo de medida) conducía a desviaciones en la predicción. No se encontró explicación a este comportamiento observado.

En la **Figura 86** se representan las predicciones obtenidas para las muestras de validación en algunos de los parámetros estudiados. Se aprecia que todas se distribuyen en torno al “comportamiento ideal”; es decir, la pendiente de la regresión “real vs. predicho” es la unidad y la ordenada en el origen es cero.

En lo que hace referencia a la precisión, se han evaluado los dos parámetros asociados (repetibilidad y reproducibilidad) como la diferencia máxima entre resultados consecutivos obtenidos en una secuencia de trabajo (r) y en días diferentes de trabajo (R). Las precisiones que presentan las metodologías puestas a punto son en muchos casos del orden de las repetibilidades de los métodos clásicos, lo cual es muy positivo.

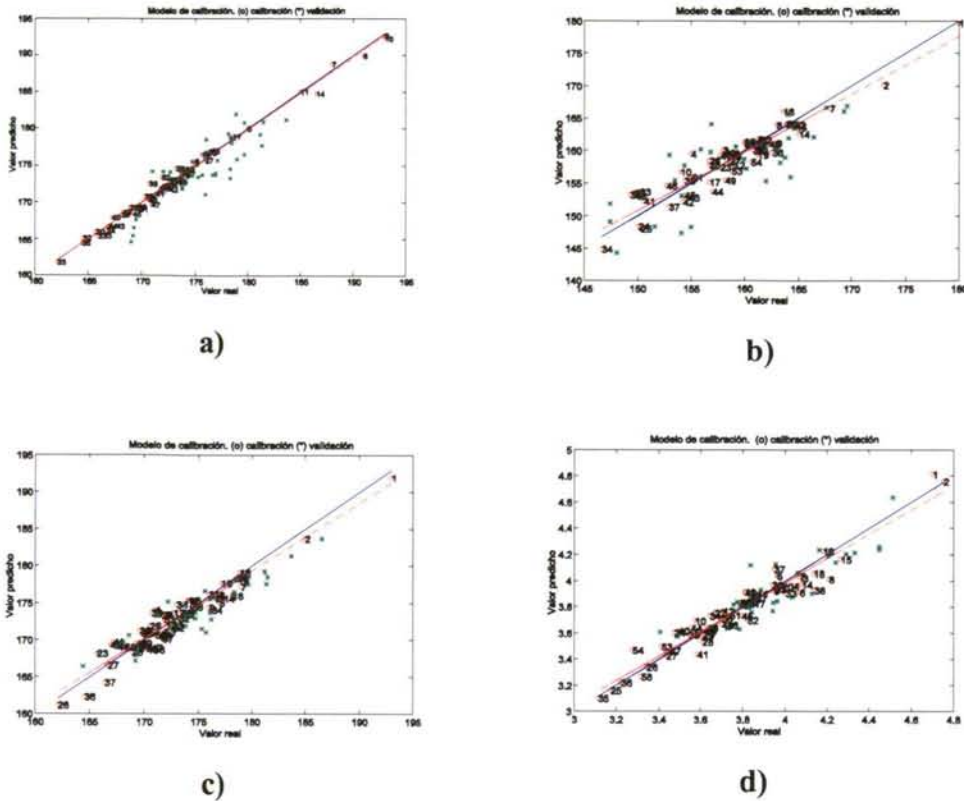


Figura 86: Predicciones obtenidas para las muestras de validación (*) superpuestas al modelo de calibración (o) para (a) 10% de destilación (sistema simple); (b) punto inicial de destilación (sistema complejo); (c) 10% de destilación (sistema complejo) y (d) viscosidad (sistema complejo).

Para obtener predicciones buenas y mejores errores promedio, dos parámetros físico-químicos (90% de destilación y punto final de destilación) necesitaban autoescalado en vez de centrado en la media. La principal razón ya se indicó que estaría asociada a problemas de los destiladores automáticos para medir el punto final de destilación de la curva de destilación (Dyroff, 1998), lo que se refleja en la elevada reproducibilidad del método ASTM y las dificultades de relacionar el espectro con los valores de referencia.

En lo que hace referencia al error promedio corregido (según las ideas de Faber y Kowalski, 1997) con objeto de separar el error de modelado del debido al método oficial) se puede observar que, efectivamente, el error inherente a la regresión PLS (SEPCor) es menor que el SEP global, lo cual mejoraría la validez intrínseca de las metodologías desarrolladas. En algunos casos, la participación del error aleatorio en

el método de referencia es tan alta que el SEP corregido sería negativo, lo cual obviamente no puede ser y se optó por dejar el SEP original (ver Tabla VI).

Loadings

En la Figura 87 se observan los cuatro primeros *loadings* de los modelos multivariantes (datos centrados en la media), los cuales son comunes a todos los parámetros estudiados. La primera VL representa el perfil general de los querosenos. La segunda VL se define principalmente por cuatro números de onda: 1385 y (mucho menos) 1460 cm^{-1} (con *loading* positivo) y 1610 cm^{-1} y el codo a 1500 cm^{-1} (con *loadings* negativos). Este factor parece diferenciar entre las estructuras lineales (*loadings* positivos) y las aromáticas/olefínicas (*loadings* negativos). De forma análoga ocurre con la tercera VL. La cuarta VL representa casi exclusivamente al pico de 1460 cm^{-1} (el pico espectral más intenso y con mayor varianza). El resto de VL se asocian a características espectrales menores. En el caso de VL muy elevadas (8^a, 9^a, 10^a, etc.) no hay una interpretación química sencilla y cabe suponer que son características espectrales menores y/o muy relacionadas con alguna/s muestra/s específica/s.

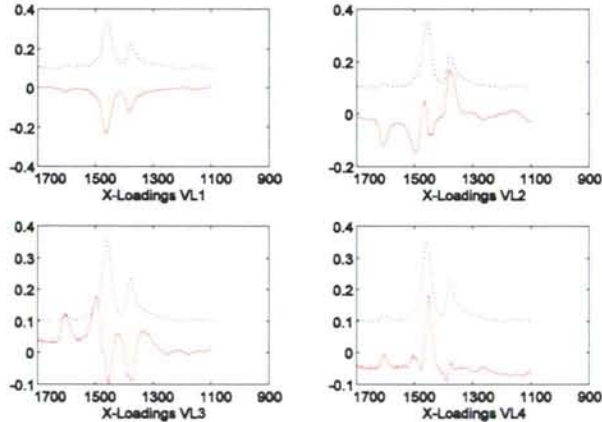


Figura 87: *Loadings* (línea continua roja) y espectro original (línea discontinua azul) para las cuatro primeras VL, 10% de destilación para el sistema de medida “complejo” (en abscisas se indican cm^{-1}).

2. MODELOS MULTIVARIANTES EN FASE LÍQUIDA

2.1. MODELOS MULTIVARIANTES CON EL SISTEMA DE MEDIDA FTMIR-ATR

Siguiendo la misma sistemática de trabajo que para las medidas en fase vapor, se han obtenido los modelos multivariantes de PLS para muestras en fase líquida medidas mediante el sistema FTMIR-ATR, obteniendo los resultados que se resumen en la **Tabla VII**:

Propiedad (unidad)	VL	%X-info	%Y-info	CV	RMSEP	RMSEP corregido	F _{exp}	r		R	
								Max. dif	RSD (%)	Max. dif	RSD (%)
Abel (°C)	11	99.10	94.36	2.15	2.1 (33)	2.05	1.73	2.16	1.91	2.57	2.30
Free (°C)	8	97.78	59.42	1.23	1.16 (31)	0.90	16.73	0.39	0.41	0.58	0.47
IBP (°C)	7	96.23	91.39	3.7	4.4 (32)	3.61	2.31	3.53	1.24	4.29	1.23
Pd10 (°C)	14	99.44	98.32	1.64	1.57 (33)	1.57*	0.48	0.95	0.33	1.9	0.67
Pd90 (°C)	12	99.26	93.38	2.37	1.95 (33)	1.95*	1.67	2.90	0.65	5.36	1.02
FBP (°C)	12	99.35	85.06	3.68	2.93 (29)	2.93*	4.30	4.65	0.94	5.42	0.99
% arom (v/v)	8	97.32	91.83	0.75	0.94 (33)	0.50	2.45	0.55	1.91	0.86	2.40
Visco (cSt)	13	99.29	95.89	0.16	0.14 (35)	0.14	1.14	0.11	1.37	0.11	1.07

%X-info= porcentaje de información explicada en las variables espectrales

%Y-info= porcentaje de información explicada en las propiedades predichas

(*)= el SEP corregido es negativo, el error estándar de redondeo es mayor que el error de predicción.

Tabla VII: Características principales de los modelos de regresión para el sistema de medida de muestras en fase líquida FTMIR-ATR. Region espectral: 1650 cm⁻¹ a 650 cm⁻¹, centrado en la media.

En todos los casos, los espectros se centraron en la media y se buscó el modelo con las mejores capacidades predictivas. En la **Figura 88** se muestran algunos de los ajustes de los modelos a los valores reales, encontrándose buenos resultados (los errores promedio son inferiores a las reproducibilidades oficiales, excepto para la viscosidad) con los modelos escogidos. El estadístico conjunto para la ordenada (=0) y la pendiente (=1) mostró siempre valores de F experimentales inferiores a los tabulados (al 95% y 99% de confianza), excepto en el caso del freezing. Para el FBP se observó que el valor de la F_{exp} presentaba un valor intermedio entre la F_{tab} del 95% y la del 99%.



“Aplicaciones al control de calidad industrial de la espectroscopia infrarroja media combinada con métodos quimiométricos multivariantes”

Relación alfabética de abreviaturas

AG / AAGG:	Algoritmo Genético / Algoritmos Genéticos.
ASTM:	<i>American Society for Testing and Materials.</i>
ATR / HATR:	Reflectancia Total Atenuada / Reflectancia Total Atenuada Horizontal.
BPN:	Backpropagation, Retropropagación ó Propagación hacia atrás.
CV:	Cross-validación o validación cruzada.
CV-LOO:	Cross-validación <i>leave one out</i> , validación cruzada uno-uno.
DTGS:	Detector de Sulfato de Triglicina Deuterado.
FBP:	Punto Final de Destilación.
FT:	Transformada de Fourier.
GC-MS:	Cromatografía de Gases- Espectrometría de Masas.
GLP:	Gas Licuado de Petróleo.
IBP, PI:	Punto Inicial de Destilación.
IR:	Espectroscopia Infrarrojo.
LDA:	Análisis Lineal Discriminante.
lr:	Coefficiente ó Velocidad de Aprendizaje.
LV / VL:	Variable Latente.
MIR:	Espectroscopia de Infrarrojos en la zona Media.
MLR:	Regresión Lineal Múltiple.
N_c:	Número de muestras en la matriz de calibración.
N_v:	Número de muestras en la matriz de validación.
NIPALS:	<i>Non-Iterative Partial Least Squares.</i>
NIR:	Espectroscopia de Infrarrojos en la zona Cercana.
NSE:	Error Estándar Normalizado ó Error Promedio (uso en RRNN).

PC:	Componente Principal.
PCA:	Análisis mediante Componentes Principales.
PCR:	Regresión Mediante Componentes Principales.
PD10:	10% de Destilación.
PD90:	90% de Destilación.
PLS, PLSR:	Regresión por Mínimos Cuadrados Parciales.
PRESS:	Suma de Cuadrados de los Residuales Predichos (<i>predicted Residual Error Sum of Squares</i>).
r:	Repetibilidad.
R:	Reproducibilidad.
RMSECV:	Raíz Cuadrada del Error Promedio de CV.
RMSEC, SEC:	Raíz Cuadrada del Error Promedio de Calibración.
RMSEP, SEP:	Raíz Cuadrada del Error Promedio de Predicción.
RN / RRNN:	Red de Neuronas Artificiales /Redes de Neuronas Artificiales.
RSD:	Desviación estándar Relativa.
SD:	Desviación estándar.
SIMCA:	Técnica de Modelado Suave de Clases Mediante Analogía (<i>Soft Independent Modelling of Class Analogy</i>).
S/N:	Relación señal/ruido.
UV:	Ultravioleta.

En las ecuaciones:

- (^o) indica Traspuesto/a.
- a** indica vector a.
- A** indica matriz A (de dimensiones n x p).
- a_{ij} indica elemento i, j de la matriz.
- a** indica un escalar.

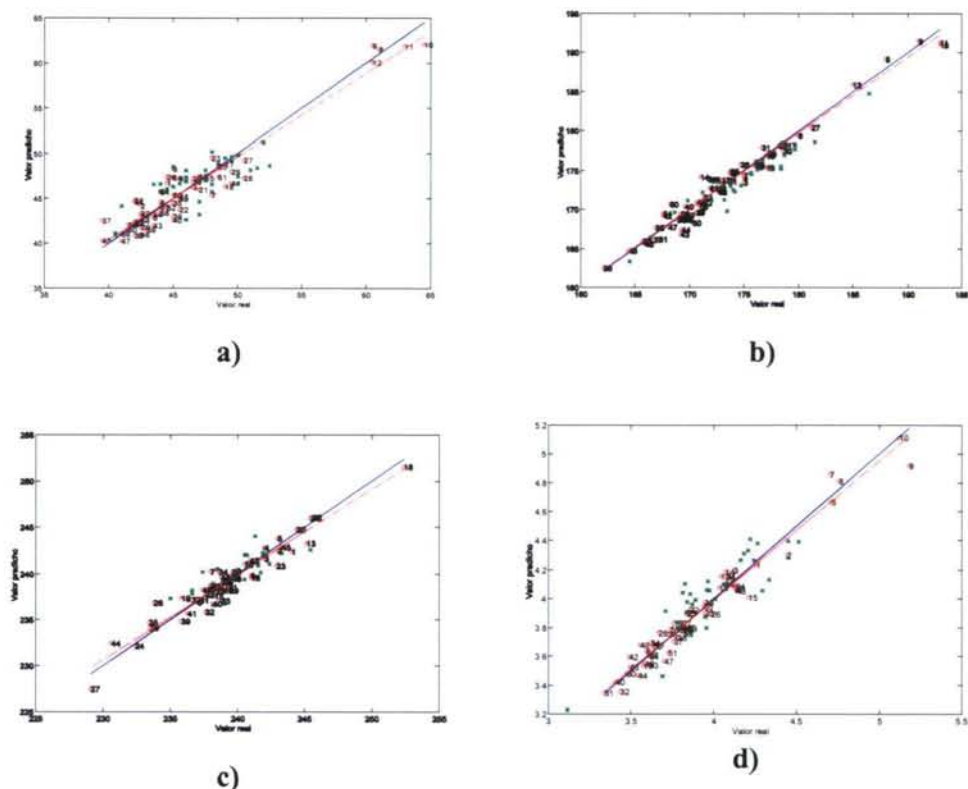


Figura 88: Ejemplos de predicciones obtenidas mediante PLS para las muestras de validación(*) empleando el sistema de medida de líquidos mediante ATR para (a) punto de deflagración; (b) 10% de destilación; (c) 90% de destilación; (d) viscosidad. Los (○) y los números corresponden a las muestras de los modelos de calibración. Las líneas azules representan el comportamiento ideal y las rojas el comportamiento experimental.

En cuanto a la comparación de los modelos con los obtenidos en fase gas, en general, se necesitan más factores para alcanzar el mismo nivel de error promedio empleando la fase líquida. De hecho, los valores SEP no son sustancialmente distintos a los encontrados en fase gas (método simple y método complejo) aunque sí hay mejorías para las predicciones del *freezing point* (punto de cristalización), 10% de destilación, 90% de destilación (aquí la mejora es clara) y FBP. Los modelos de viscosidad siguen presentando problemas.

Loadings

La **Figura 89** muestra cuatro vectores típicos de *loadings* correspondientes a las primeras cuatro variables latentes. Para simplificar su visualización e interpretación, se incluyó el espectro original desplazado verticalmente (en color azul en la parte superior).

Los *loadings* más importantes corresponden a los picos espectrales más intensos de los espectros FTIR-ATR. En la **Figura 89** se observa mayormente la oposición de estructuras lineales frente a estructuras aromáticas (1ª y 2ª variables latentes). En la 3ª VL se observa que la banda espectral de 850-650 cm^{-1} se desdobra en dos, una parte con *loadings* negativos (correspondiente, con probabilidad, a las bandas de esqueleto de los grupos $\text{C}(\text{CH}_3)_2$), y al movimiento del esqueleto (CH_2_n) y la parte con *loadings* positivos (correspondiente a los anillos aromáticos bencénicos monosustituídos). La cuarta VL muestra mayormente la banda de aromáticos. La VL quinta o la sexta (dependiendo de la propiedad) se corresponderían con el espectro promedio del queroseno. Se observó que el FBP y el % de aromáticos daban más importancia a las estructuras aromáticas, como cabría esperar.

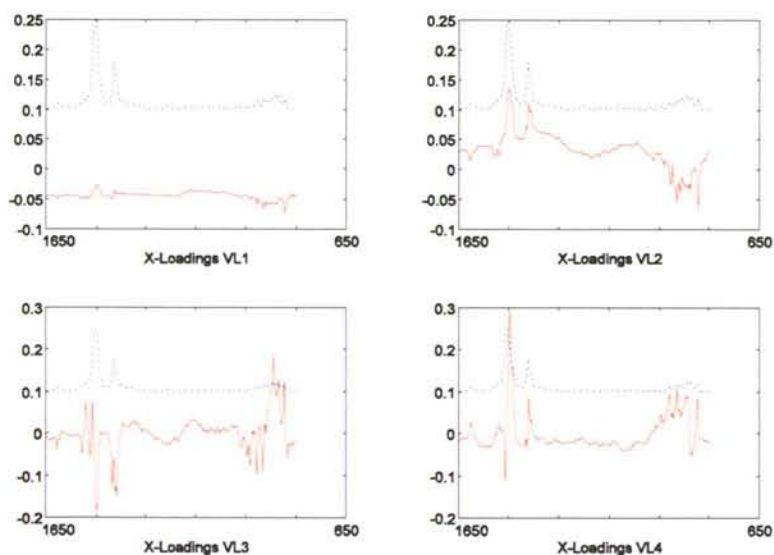


Figura 89: Aspecto típico de los *loadings* considerando el punto de deflagración. En azul se indica el espectro típico mientras que en rojo se observan los diferentes *loadings*.

2.2. MODELOS MULTIVARIANTES CON EL SISTEMA DE MEDIDA RAMAN

De forma análoga a los casos anteriores de medida de queroseno, se han construido los modelos PLS para las propiedades de dicho producto, obteniéndose los resultados mostrados de forma resumida en la **Tabla VIII**. Dado que éste constituía un estudio piloto, se descartaron las propiedades del punto de cristalización y del 90% de destilado.

En la **Figura 90** se muestra el ajuste de los modelos a los valores reales, encontrándose buenos resultados con los modelos escogidos.

Propiedad (unidad)	VL	%X-info	%Y-info	CV	SEP	RMSEP	r		R	
							Max. dif	RSD (%)	Max. dif	RSD (%)
Abel (°C)	5	99.87	99.98	2.6	2.1 (46)	2.05	2.4	2.2	3.5	2.5
IBP (°C)	4	99.85	99.97	5.4	3.7 (45)	2.71	4.4	1.2	5.4	1.1
Pd10 (°C)	5	99.87	100.0	3.1	3.1 (49)	1.69	4.3	1.3	6.6	1.5
FBP (°C)	4	99.85	99.90	5.8	4.0 (48)	2.52	6.6	1.4	7.6	1.4 (*)
% Arom (v/v)	2	99.73	99.70	1.1	1.3 (48)	1.02	0.8	1.7	0.9	1.4
Visco (cSt)	4	99.85	99.97	0.29	0.19 (49)	0.19	0.35	3.06	0.16	2.18

%X-info= porcentaje de información explicada en las variables espectrales

%Y-info= porcentaje de información explicada en las propiedades predichas

* = sólo dos valores se usaron para calcular el valor medio

Tabla VIII: Características principales de los modelos de regresión empleando espectroscopia Raman.
Region espectral: 193.5 cm^{-1} a 1688.1 cm^{-1} , normalización unidad.

Se observaron unos resultados de predicción similares (aunque un poco superiores) a los de la celda de cuarzo (sistema simple para medida de gases), excepto en el caso del punto de deflagración. Si se comparan con los observados para el sistema complejo de medida de gases, se ve una mejor predicción para las propiedades a excepción del punto de deflagración, porcentaje de aromáticos y viscosidad. Además, los modelos obtenidos mediante el sistema de medida de líquidos por Raman son comparables a los que se obtuvieron mediante el sistema de medida de líquidos por ATR.

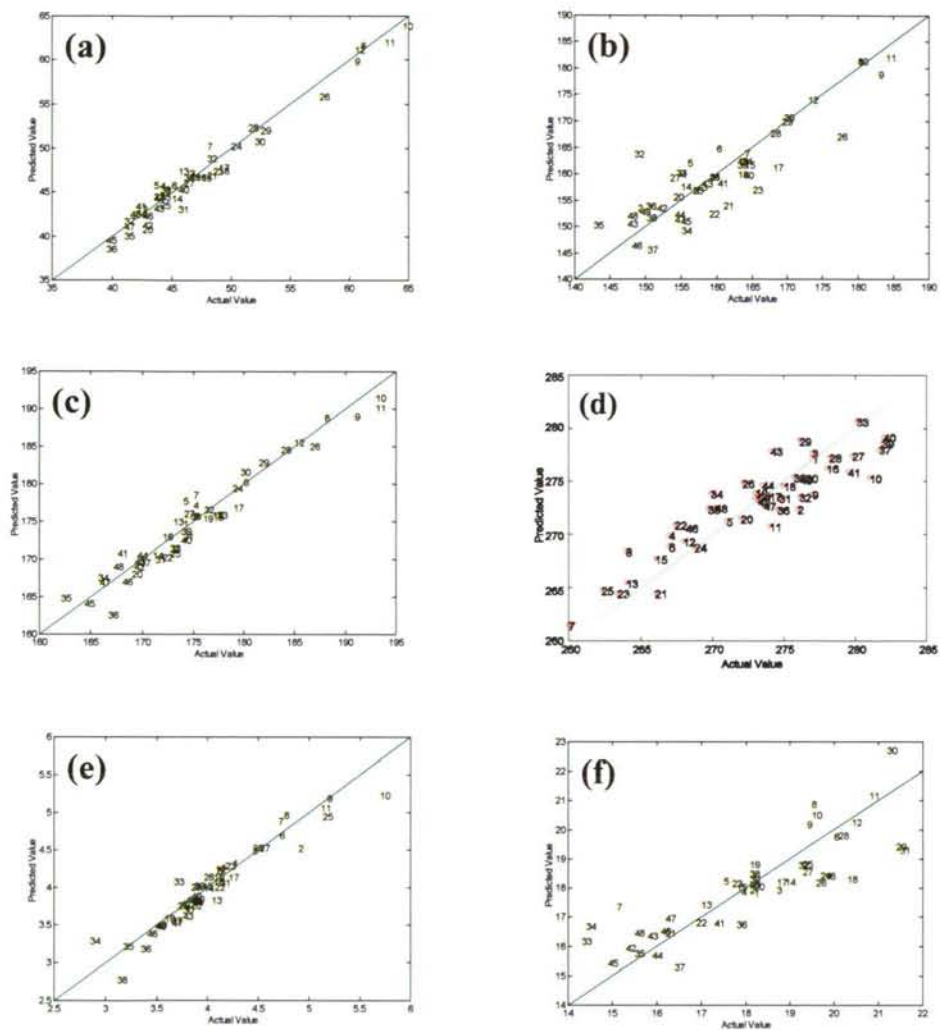


Figura 90: Valor real frente a predicho para cada una de las propiedades del queroseno (la línea recta indica la situación perfecta): (a) flash point; (b) IBP; (c) 10% de destilación; (d) FBP; (e) viscosidad; (f) % aromáticos.

Ahora bien, debe destacarse que los modelos empleando la espectroscopia Raman son, claramente, menos complejos que los necesarios al emplear tanto fase vapor (en ambas modalidades) como en ATR. Esto puede atribuirse, en parte, a la utilidad del tratamiento de normalización unidad llevado a cabo. Resulta, pues, interesante la opción de combinar la medida directa de muestras líquidas mediante

espectrometría Raman con PLS ya que, si bien no condujo a los menores errores promedio en las variables estudiadas, sí conduce a modelos cuyo error promedio es inferior a la reproducibilidad de los métodos oficiales excepto para el punto de deflagración (Abel) y el, ya reseñado, de determinación de la viscosidad.

Adicionalmente, se presenta una breve comparación entre los modelos Raman encontrados y otros modelos encontrados en la bibliografía (no se han hallado otros trabajos que empleasen Raman) (ver **Tabla IX**).

Propiedad	Raman		ATR-MIR (1)		ATR-MIR (2)		NIR (2)		Microcelda MIR (3)	
	VL	RMSEP	VL	RMSEP	VL	RMSEP	VL	RMSEP	VL	RMSEP
Abel	5	2.1	17	6.0	---	---	---	---	3	2.1
IBP	4	3.7	14	10.0	8	3.6	6	3.5	3	2.7
PD10	5	3.1	15	6.0	5	2.2	10	1.3	---	---
FBP	4	4.0	16	11.0	6	3.4	7	2.1	3	2.0
% Aromáticos	2	1.3	18	0.5	---	---	---	---	3	2.3
Viscosidad	4	0.19	19	0.09	---	---	---	---	3	0.2

(1) Fodor, G.E.; Mason, R.A.; Hutzler, S.A. *Applied Spectroscopy* 53, 1292 (1999).

(2) Chung, H.; Ku, M.S.; Lee, J.S. *Vibrational Spectroscopy* 20, 155 (1999).

(3) Garrigues, S.; Andrade, J.M.; De la Guardia, M.; Prada, D. *Analitica Chimica Acta* 317, 95 (1993).

Tabla IX: Comparación de capacidades predictivas de distintos sistemas vibracionales (el número entre paréntesis se corresponde con la referencia).

Loadings

Aunque la interpretación química de los *loadings* de PLS no es siempre sencilla, es interesante relacionar los modelos con la química conocida. Esto daría confianza no sólo en cómo se realizaron los modelos sino también en la aplicación de los modelos a la rutina de trabajo. Además de la interpretación espectral indicada en el epígrafe (d) de la página 192 se trata ahora de relacionar los modelos (a través de sus *loadings*) con aquella interpretación estructural.

Figura 91 presenta cuatro vectores típicos de *loadings* correspondientes a las primeras 4 variables latentes. Se usarán para ver la influencia de cada una de las regiones espectrales (grupos funcionales) en el modelo PLS. Para simplificar su visualización e interpretación, se incluyó el espectro original desplazado verticalmente (dibujado en color azul en la parte superior).

Puede verse fácilmente que los *loadings* más importantes se corresponden con los picos espectrales más intensos presentes en la región espectral estudiada, 193.5 cm^{-1} (variable número 1) a 1688.1 cm^{-1} (variable número 775). La primera variable latente representa un perfil perfecto del espectro, como se esperaba, ya que los datos no fueron centrados ni la varianza escalada y, por tanto, el primer factor es “la media del espectro de queroseno”.

Los *loadings* restantes son más específicos. La segunda variable latente (VL) está esencialmente definida por el pico a 1612 cm^{-1} , correspondiente a una tensión del anillo de derivados del benceno (también puede ser la tensión C=C de alquenos lineales). La tercera VL no es fácil de interpretar pero se puede atribuir a 1459 cm^{-1} (estructuras de n-alcano), 1612 cm^{-1} , (alquenos lineales o tensión del anillo) y (con signo opuesto) 750 cm^{-1} (diferentes modos de “*breathing*” (“respiración”) de estructuras de anillos), de esta manera esta VL parece mostrar el comportamiento opuesto de las estructuras lineales y las del anillo con respecto a las propiedades físico-químicas consideradas aquí (por ejemplo, las estructuras del anillo aumentan los puntos de deflagración y de destilación mientras las estructuras lineales tienen tendencia a disminuirlos). La cuarta VL es mayormente definida por 1003 cm^{-1} (modos de “*breathing*” de compuestos aromáticos).

Todos los modelos de las diferentes propiedades revelaban patrones similares de *loadings* a lo largo de las primeras cuatro variables latentes, mientras la quinta VL se relacionó a variables espectrales “menores” (cuando tenía que ser usado este factor,

se relacionaba con diferencias espectrales en los picos más débiles).

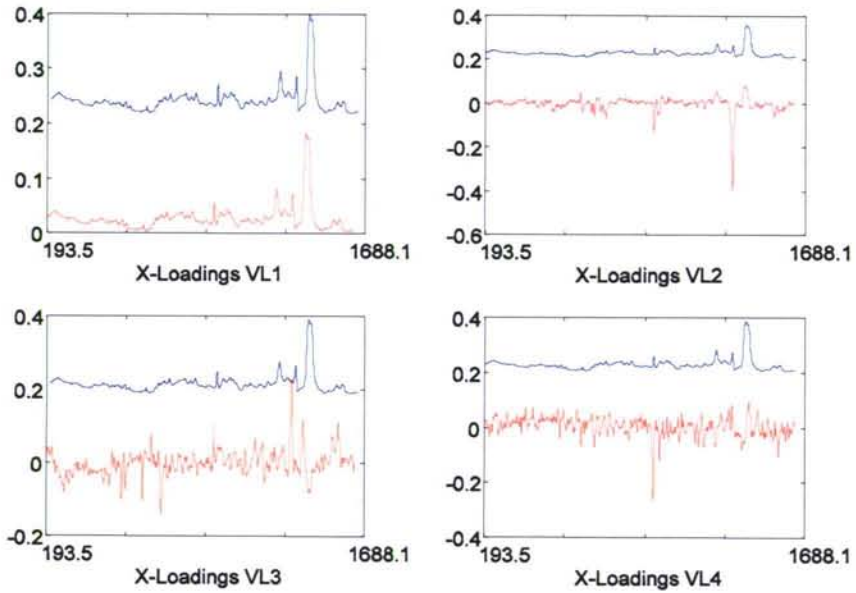


Figura 91: Aspecto típico de los *loadings* considerando el punto de deflagración. En azul se indica el espectro típico mientras que en rojo se observan los diferentes *loadings* (en abscisas se indica el desplazamiento Raman, en cm^{-1}).

PARTE D.- BIBLIOGRAFÍA

ATSDR 1993. Case Studies in Environmental Medicine: Jet Fuel Toxicity. US Department of Human Health Services Public Health Service, Agency for Toxic Substances and Disease Registry. September 1993.

ATSDR 1995. Toxicological Profile for Fuel Oils. US Department of Human Health Services Public Health Service, Agency for Toxic Substances and Disease Registry. June 1995.

ATSDR 1998a. Toxicological Profile for Jet Fuels (JP-5 and JP-8). US Department of Human Health Services Public Health Service, Agency for Toxic Substances and Disease Registry. August 1998.

ATSDR 1998b. Toxicological Profile for Total Petroleum Hydrocarbons (THP). US Department of Human Health Services Public Health Service, Agency for Toxic Substances and Disease Registry. September 1998.

ASTM D86. Test Method for Distillation of Petroleum Products, Annual Book of ASTM Standards, Vol. 05.01 (1995).

ASTM D445. Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (the Calculation of Dynamic Viscosity), Annual Book of ASTM Standards, Vol. 05.01 (1995).

ASTM D1319. Test Method for Hydrocarbon Types in Liquid Petroleum Products by Fluorescent Indicator Adsorption, Annual Book of ASTM Standards, Vol. 05.01 (1995).

ASTM D2386. Test Method for Freezing Point of Aviation Fuels, Annual Book of ASTM Standards, Vol. 05.02 (1995).

Caswell, K.A.; Glass, T.E.; Swann, M.; Dorn, H.C. Rapid Prediction of Various Physical Properties for Middle Distillate Fuels Utilizing Directly Coupled Liquid Chromatography/¹H Nuclear Magnetic Resonance. *Anal. Chem.* 61, 206-211 (1989).

Collette, T.W. Predicting environmental fate parameters with infrared spectroscopy. *Trends in analytical chemistry* 16 (1), 24-36 (1997).

Cooper, J.B.; Wise, K.L.; Welch, W.T.; Bledsoe, R.R.; Sumner, M.B. Determination of Weight Percent Oxygen in Commercial Gasoline: Comparison between FT-Raman, FT-IR, and Dispersive Near-IR Spectroscopies. *Applied Spectroscopy* 50(7), 917-921 (1996).

Chung, W.M.; Wang, Q.; Sezerman, U.; Clarke, R.H. Analysis of aviation turbine-fuel composition by laser Raman spectroscopy. *Applied Spectroscopy* 45(9), 1527-1532 (1991).

- DERD 2494. Turbine Fuel, Aviation Kerosine Type, Jet A-1. Nato Code: F-35. Joint Service Designation: AVTUR. Ministry of Defence. Defence Standard 91-91/Issue 3; 12 November 1999.
- Dyroff, G.V. Manual on significance of tests for petroleum products. ASTM Manual Series, MNL1, Philadelphia, MA, USA: ASTM; 1998.
- Faber, K.; Kowalski, B.R. Improved Prediction Error Estimates for Multivariate Calibration by Correcting for the Measurement Error in the Reference Values. *Applied Spectroscopy* 51 (5), 660-665 (1997).
- Flecher, P.E.; Cooper, J.B.; Vess, T.M.; Welch, W.T. Remote Fiber optic Raman analysis of benzene, toluene, and ethylbenzene in mock petroleum fuels using partial least squares regression analysis. *Spectrochimica Acta* 52(10), 1235-1244 (1996).
- Flecher, P.E.; Welch, W.T.; Albin, S.; Cooper, J.B. Determination of octane numbers and Reid vapor pressure in commercial gasoline using dispersive fiber-optic Raman spectroscopy. *Spectrochimica Acta* 53(2), 199-206 (1997).
- Fodor, G.E.; Kohl, K.B. Analysis of Middle Distillate Fuels by Midband Infrared Spectroscopy. *Energy & Fuels* 7(5), 598-601 (1993).
- Fodor, G.E. Analysis of Petroleum Fuels by Midband Infrared Spectroscopy. *International Congress & Exposition, Detroit, Michigan. February 28-March 3, 1994.*
- Fodor, G.E.; Kohl, K.B.; Mason, R.L. Analysis of Gasolines by FT-IR Spectroscopy. *Analytical Chemistry* 68, 23-30 (1996).
- Garrigues, S.; Andrade, J.M.; de la Guardia, M.; Prada, D. Multivariate calibration in Fourier Transform infrared spectrometry for prediction of kerosene properties. *Analytica Chimica Acta* 317, 95-105 (1995).
- Harris, D.T.; Sakiestewa, D.; Robledo, R.F.; Witten, M. Immunotoxicological effects of JP-8 Jet Fuel Exposure. *Toxicology and Industrial Health* 13 (1), 43- 45 (1997).
- Harris, D.T.; Sakiestewa, D.; Robledo, R.F.; Witten, M. Protection from JP-8 Jet Fuel Induced Immunotoxicity by Administration of Aerosolized Substance P. *Toxicology and Industrial Health* 13 (5), 571- 588 (1997).
- Harris, D.T.; Sakiestewa, D.; Robledo, R.F.; Young, R.S.; Witten, M. Effects of short-term JP-8 jet fuel exposure on cell-mediated immunity. *Toxicology and Industrial Health* 16, 78-84 (2000a).
- Harris, D.T.; Sakiestewa, D.; Titone, D.; Robledo, R.F.; Young, R.S.; Witten, M. Substance P as prophylaxis for JP-8 jet fuel-induced immunotoxicity. *Toxicology and Industrial Health* 16, 253-259 (2000b).
- Harris, D.T.; Sakiestewa, D.; Titone, D.; Robledo, R.F.; Young, R.S.; Witten, M. Jet fuel-induced immunotoxicity. *Toxicology and Industrial Health* 16, 261-265 (2000c).

- Hren, B.; Katona, K.; Mink, J.; Kohán, J.; Isaák, Gy. Long-path FTIR spectroscopic studies of air pollutants in the Danube refinery plant. *Analyst* 125, 1655-1659 (2000).
<http://auto.howstuffworks.com/oil-refining5.htm>, 2003.
- <http://193.51.164.11/htdocs/monographs/Vol45/45-04.htm>, 2002.
- <http://library.cbest.chevron.com>, 2001.
- http://ocl.nps.navy.mil/jetfuel/docs/liv_kid_jp8.doc, 2001.
- Iob, A.; Ali, M.A.; Tawabini, B.S.; Abbas, N.M. Hydrocarbon group (PONA) analysis of reformat by FT-IR spectroscopy. *Fuel* 75 (9), 1060-1064 (1996).
- IP 170. Flash Point by the Abel Closed Cup, British Institute of Petroleum (1995).
- Josefson, D. Jet fuel may cause nerve damage. *British Medical Journal (BMJ)* 315, 269-274 (1997).
- Kalabokas, P.D.; Hatzianestis, J.; Bartzis, J.G.; Papagiannakopoulos. Atmospheric concentrations of saturated and aromatic hydrocarbons around a Greek oil refinery. *Atmospheric Environment* 35, 2545-2555 (2001).
- Lin-Vien, D.; Clothup, N.B.; Fateley, W.G.; Grasselli, J.G. The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules, Academic Press, San Diego, CA, USA, 1991.
- Litani-Barzilai, I.; Sela, I.; Bulatov, V.; Zilberman, I.; Schechter, I. On-line remote prediction of gasoline properties by combined optical methods. *Analytica Chimica Acta* 339, 193-199 (1997).
- Liu, S.; Pleil, J.D. Method for liquid-liquid extraction of blood surrogates for assessing human exposure to jet fuel. *Journal of Chromatography B* 728, 193-207 (1999).
- Liu, S.; Pleil, J.D. Optimized determination of trace jet fuel volatile organic compounds in human blood using in-field liquid-liquid extraction with subsequent laboratory gas chromatographic-mass spectrometric analysis and on-column large-volume injection. *Journal of Chromatography B* 752, 159-171 (2001).
- López-Anreus, E.; Garrigues, S.; de la Guardia, M. Simultaneous vapour phase Fourier transform infrared spectrometric determination of butyl acetate, toluene and methyl ethyl ketone in paint solvents. *Analyst* 123, 1247-1252 (1998).
- Macho, S.; Boqué, R.; Larrechi, M.S.; Rius, X. Multivariate determination of several compositional parameters related to the content of hydrocarbon in naphtha by MIR spectroscopy. *Analyst* 124, 1827-1831 (1999).
- Mattie, D.R.; Alden, C.L.; Newell, T.K.; Gaworski, C.L.; Flemming, C.D. A 90-day continuous vapor inhalation toxicity study of JP-8 jet fuel followed by 20 or 21 months

of recovery in Fischer 344 rats and C56BL/6 mice. *Toxicol. Pathol.* 19 (2), 77-87 (1991).

Mckay MD, Ch. A.; Hart MD, K.; Siddiqi MBBS, M.; McCormick PharmD, M.A.; Bayer MD, M. J. Case Report: Pulmonary Interstitial Fibrosis Associated With Inhalation Exposure to Aviation Fuels. *Int. J. Med. Toxicol.* 4(3), 20 (2001).

Pérez-Ponce, A.; Garrigues, S.; de la Guardia, M. Direct determination of total SO₂ in musts and wines by vapour phase Fourier transform infrared spectrometry. *Química Analítica* 19, 151-158 (2000).

Pérez-Ponce, A.; Garrigues, J.M.; Garrigues, S.; de la Guardia, M. Vapour phase Fourier transform infrared spectrometric determination of carbonate in sediments. *Analyst* 123, 1817-1821 (1998).

Ribak, J.; Rayman, R.B.; Froom, P. Occupational Health in Aviation. Maintenance and Support Personnel. Academic Press, 1995.

Robinson, J.M. Practical Handbook of Spectroscopy, CRC Press, Boca Raton, FL, USA, 1991.

Smith, Brian C. . Fundamentals of Fourier Transform Infrared Spectroscopy. CRC Press, 1996.

Smith, L.B.; Bhattacharya, A.; Lemasters, G.; Succop, P.; Puhala, E.; Medvedovic, M.; Joyce, J. Effect of Chronic Low-level Exposure to Jet Fuel on Postural Balance of US Air Force Personnel. *Journal of Occupational and Environmental Medicine* 39 (7), 623-632 (1997).

www.aircareintl.org, 2002.

www.atsdr.cdc.gov, 2002.

www.cdc.gov/niosh/ocderm4.html, 2002.

www.chevron.com/prodserv/fuels/bulletin/aviationfuel/toc.shtm, 2002a.

www.chevron.com/prodserv/aviation/jeta0513.shtm, 2002b.

www.concawe.be/download/reports/rpt_94-106.pdf, 2002a.

www.concawe.be/download/reports/rpt_99-52.pdf, 2002b.

www.interlink.es/ccoo-clh/CLH.HTM, 2002.

www.metrokc.gov/health/phnr/eapd/reports/cancer/jetfuel.pdf, 2001.

www.monografias.com/trabajos/petroleo2/petroleo2.shtml, 2002.

www.thermo.com/eThermo/CDA/BU_Home/BU_Homepage/0,1285,101,00.html, 2002.



Capítulo V

Análisis de zumo

Objetivo:

En este capítulo se desarrolla un proceso analítico rápido, sencillo de aplicar, fiable y sin consumo de reactivos químicos, que permita llevar a cabo operaciones de screening con objeto de evaluar si un zumo ó refresco comercial de manzana poseen el porcentaje declarado de zumo puro de manzana. Para ello se combinará la espectroscopia FTIR-ATR con técnicas quimiométricas.

Índice:

Parte A.- Aproximación al producto objeto del estudio

- 1. Definición y clasificación general de los frutos*
- 2. Importancia de la fruta*
- 3. Las manzanas*
 - 3.1. Propiedades nutricionales de la manzana*
 - 3.2. Variedades*
 - 3.3. Producción mundial*
 - 3.4. Zumos de manzana*
 - 3.5. Proceso industrial para obtener zumo*
- 4. Composición Química*
 - 4.1. Compuestos nitrogenados*
 - 4.2. Carbohidratos*
 - 4.3. Lípidos*
 - 4.4. Ácidos orgánicos*
 - 4.5. Compuestos fenólicos*
 - 4.6. Compuestos aromáticos*
 - 4.7. Vitaminas*
 - 4.8. Minerales*

5. Cambios en la intensidad respiratoria

6. Adulteración

7. Legislación

Parte B.- Parte experimental

1. Aparatos y Software

2. Muestras

- 2.1. Zumos/refrescos comerciales*
- 2.2. Zumos "100% puros" de manzana*
- 2.3. Zumos "puros diluidos" de manzana*
- 2.4. Zumos "sintéticos" a partir de azúcares*

Parte C.- Resultados y discusión

- 1. Clasificación mediante curvas de potencia, rango bajo de zumos (2% al 20%)*
 - 1.1. Modelado*
 - 1.2. Validación del modelo*
 - 1.3. Clasificación de los refrescos comerciales*
- 2. Clasificación mediante curvas de potencia, rango alto de zumos (20% - 100%)*
 - 2.1. Modelado*
 - 2.2. Validación del modelo*
 - 2.3. Clasificación de los refrescos y zumos comerciales*
- 3. Variables que determinan la clasificación de las muestras mediante curvas de potencia*
- 4. Clasificación mediante SIMCA, rango bajo de zumos (2% - 20%)*
 - 4.1. Modelado*
 - 4.2. Validación del modelo*
 - 4.3. Clasificación de los refrescos comerciales*
- 5. Clasificación mediante SIMCA, rango alto de zumos (20% - 100%)*
 - 5.1. Modelado*
 - 5.2. Validación del modelo*
 - 5.3. Clasificación de los refrescos y zumos comerciales*
- 6. Variables que determinan la clasificación de las muestras*
 - 6.1. Rango bajo*
 - 6.2. Rango alto*
 - 6.3. Clasificación de los refrescos y zumos comerciales*
- 7. Clasificación mediante PLS, rango bajo de zumos (2% - 20%)*
- 8. Clasificación mediante PLS, rango alto de zumos (20% - 100%)*
- 9. Clasificación mediante RRNN, rango bajo de zumos (2% -20%)*
- 10. Clasificación mediante RRNN, rango alto de zumos (20% 100%)*
- 11. Anexo*

Parte D.- Bibliografía

PARTE A.- APROXIMACIÓN AL PRODUCTO OBJETO DEL ESTUDIO

1. DEFINICIÓN Y CLASIFICACIÓN GENERAL DE LOS FRUTOS

La fruta se define como "el fruto, la infrutescencia, la semilla o las partes carnosas de órganos florales, que hayan alcanzado un grado adecuado de madurez y sean propias para el consumo humano" (Código alimentario español y disposiciones complementarias, 1997). Se denominan frutos a los productos del desarrollo de una flor después de la fecundación; en ellos quedan contenidas las semillas (Diccionario de la Real Academia Española, 2001).

En la cubierta del fruto, que recibe el nombre de pericarpio, se pueden distinguir tres zonas o capas: la externa o epicarpio, la intermedia o mesocarpio y la interna o endocarpio (Tabla I). Atendiendo a su origen y otras características, los frutos se clasifican en (Enciclopedia Universal Ilustrada Europeo Americana, 1980) simples (proceden de la transformación de un gineceo constituido por un sólo carpelo o por varios soldados, el carpelo es la hoja transformada para formar un pistilo o parte de un pistilo (Diccionario de la Real Academia Española, 2001)), múltiples (fresa, rosa, granada) y compuestos ó infrutescencias (higo, mora, piña americana). Los primeros, a su vez, se pueden dividir en secos y carnosos. Dentro de estos últimos se diferencian:

- a.- Drupa (melocotón, aceituna, cereza): es el fruto con mesocarpio carnoso y endocarpio (hueso) leñoso. En este caso el mesocarpio suele llamarse sarcocarpio (carne).
- b.- Baya: es el fruto carnoso en todo el espesor del pericarpio (pulpa).
 - Hesperidio (naranja, limón).
 - Pepónide (calabaza, melón).
- c.- Pomo (manzana, pera).

Zona	Denominación	Características	Formado por
Externa	Epicarpio	Tejido epidérmico y sus productos	cera, pelos, alas, espinas, etc
Intermedia	Mesocarpio	Mesocarpio	
		Heterogéneo	Parénquima Esclerenquima Prosénquima Pequeños haces fibrovasculares Conductos secretores Glándulas Vasos lactíferos
Interna	Endocarpio	Homogéneo	Parenquimatoso Esclerenquimatoso Fibroso
		Heterogéneo	Parenquimatoso y fibroso Esclerenquimatoso y fibrosovascular

Tabla I: Estructura anatómica de los frutos.

La manzana, objeto de estudio en esta Memoria, cuyo nombre científico es *Malus sylvestris*, de la familia de las *Rosaceae*, es un pomo y se usa cruda, desecada, en compota, pulpa, jalea, zumo y aguardiente (Belitz y Grosch, 1992).

Existen otras clasificaciones a tener en cuenta además de la presentada, como las recogidas en el *Código alimentario español y disposiciones complementarias (1997)*:

1. Por su naturaleza:

- 1.1. Frutas carnosas: su parte comestible posee en su composición, cuando menos, el 50% de agua.
- 1.2. Frutas secas: son aquellas cuya parte comestible posee en su composición menos del 50% de agua.
- 1.3. Frutas oleaginosas: son aquellas empleadas para la obtención de grasas y para el consumo humano.

2. Por su estado:

- 2.1. Fruta fresca: es la destinada al consumo inmediato sin sufrir tratamiento alguno que afecte a su estado natural.
- 2.2. Fruta desecada: es el producto obtenido a partir de frutas frescas, a las que se ha reducido la proporción de humedad mediante procesos apropiados

y autorizados. El grado de humedad residual será tal que impida toda alteración posterior.

2.3. Fruta congelada.

3. Por su calidad comercial, las que se determinen en cada caso por la reglamentación correspondiente.

2. IMPORTANCIA DE LA FRUTA

El consumo de frutas y vegetales se ha asociado a unas bajas proporciones de incidencia y mortalidad de cáncer en humanos (Doll, 1990; Dragsted *et al.*, 1993; Ames *et al.*, 1993; Willett, 1994a). En experimentos con animales se ha encontrado que los vegetales comunes a las dietas humanas presentan efectos antitumorigénicos (Belman, 1983; Maltzman *et al.*, 1989; Stoewsand *et al.*, 1988; Stoewsand *et al.*, 1989; Bingham, 1990; Bresnick *et al.*, 1990; Wattenberg y Coccia, 1991). Algunos autores también han indicado una correlación negativa altamente significativa entre la toma de frutas y vegetales totalmente frescos y la mortalidad por dolencia de la isquemia en el corazón, tal es el caso de Armstrong *et al.* (1975) en Bretaña y Verlangieri *et al.* (1985) en los Estados Unidos. Datos similares fueron encontrados entre grupos de vegetarianos (Phillips *et al.*, 1978; Burr y Sweetnam, 1982). Tanto personas vegetarianas como no vegetarianas con consumo de frutas y vegetales altos han visto reducida su presión arterial (Sacks and Dass, 1988; Ascherio *et al.*, 1992). También fue hallada una asociación negativa entre el consumo de frutas y vegetales y la mortalidad por dolencia cerebrovascular (Acheson y Williams, 1983; Wang *et al.*, 1996).

La protección que las frutas y vegetales tienen contra algunas enfermedades, incluyendo cáncer, dolencias cardiovasculares y cerebrovasculares, ha sido atribuida a los antioxidantes que contienen (Ames, 1983; Steinberg *et al.*, 1989; Gey, 1990; Steinberg, 1991). Actualmente, hay muchas evidencias que indican que los radicales libres causan perjuicio a lípidos, proteínas y ácidos nucleicos. Los radicales libres pueden ser la raíz de la etiología de un gran número de enfermedades, incluyendo cáncer y arterioesclerosis. Además, los antioxidantes que pueden neutralizar los radicales libres, pueden ser de importancia central en la prevención de estas enfermedades. Se han asociado bajos niveles en plasma de las vitaminas antioxidantes con un riesgo alto de mortalidad de cáncer (Stähelin *et al.*, 1991a,b; Willett, 1994b) y, así mismo, bajos niveles en plasma de vitaminas E y C aumentaban el riesgo de angina en hombres; también se observó una correlación inversa entre niveles de vitamina E en plasma y mortalidad

de isquemia en el corazón (Riemersma, 1989; Gey et al., 1991; Wang et al., 1996).

3. LAS MANZANAS

3.1. PROPIEDADES NUTRICIONALES DE LA MANZANA

Las propiedades nutricionales de la manzana son (www.juver.es, 2004):

- Importante fuente de vitamina C.
- Buena fuente de fibra, ya que la mitad de una manzana de tamaño medio contiene 5 g de fibra, mientras que los cereales tienen entre 1-5 g. La fibra ayuda a nuestro cuerpo facilitando la digestión y añade volumen a la dieta (www.dole5aday.com, 2004). También es buena para el corazón y la circulación, efectiva contra el estreñimiento y la diarrea, limpia los dientes y fortalece las encías, acción antiviral, etc. El consumo ideal sería de 2 manzanas al día.

La información nutricional de una manzana se resume como sigue (datos por 150 g) (www.juver.es, 2004): 80 calorías, grasa: 0 g, colesterol: 0 mg, sodio: 0 mg, carbohidratos: 22 g, fibra: 5 g, azúcares: 16 g, proteínas: 0 g.

Los datos nutricionales para el zumo son, por 240 mL (www.cepri.cl, 2000): 134 calorías, grasa total: 0 g, grasas saturadas: 0 g, colesterol: 0 g, sodio: 5 mg, carbohidratos totales: 33 g (fibra dietaria: 0 g, azúcares: 26 g), proteína: 0.5 g, calcio: 0.5%.

Una manzana tiene el 20% de fibra que necesita nuestro cuerpo cada día para tener buena salud. Hay dos clases de fibra (soluble e insoluble) siendo, aproximadamente, el 80% de la fibra de las manzanas de tipo soluble. La fibra insoluble puede prevenir distintos tipos de cáncer (www.dole5aday.com, 2004). La manzana es rica en pectina, una fibra soluble, que ayuda al cuerpo a eliminar el colesterol y a protegerse contra los efectos de la polución ambiental. Diversos estudios en Francia, Italia e Irlanda han demostrado que con dos manzanas al día se puede reducir hasta un 10% el nivel de colesterol, al mismo tiempo que la pectina ayuda a nuestro cuerpo a eliminar metales nocivos tales como el plomo y el mercurio. Dos manzanas al día pueden ser un tónico para el corazón y la circulación. Las manzanas contienen, además, ácido málico y tartárico, que son especialmente eficaces como ayuda en la digestión de alimentos ricos en grasas. La vitamina C que se encuentra en la manzana ayuda a reforzar el sistema inmunológico. Tradicionalmente, las manzanas se han usado para combatir problemas gastrointestinales y, por ejemplo, (debido a sus

contenidos en fibra soluble) se usa para combatir el estreñimiento. La manzana se usa, así mismo, para problemas de artritis, reumatismo, gota, diarrea, gastroenteritis y colitis. También se ha visto que el simple olor a manzanas tiene un efecto relajante y ayuda a bajar la tensión.

El azúcar de las manzanas es mayormente fructosa, un azúcar simple que se descompone lentamente en el cuerpo y ayuda a mantener un nivel equilibrado de azúcar en sangre.

Las manzanas pueden ser almacenadas unos cinco meses si son enfriadas a aproximadamente 0°C. En atmósferas especiales controladas para el almacenamiento, las manzanas pueden ser almacenadas casi 12 meses porque la temperatura, humedad, oxígeno y dióxido de carbono son constantemente monitorizados y controlados con objeto de prevenir la maduración rápida. Las manzanas frescas flotan porque el 25% de su volumen es aire (*www.dole5aday.com, 2004*).

3.2. VARIEDADES

Hay unas 8000 variedades distintas de manzanas; algunas de las más importantes pueden ser: Granny Smith, Cox Orange, Boskoop, Elstar, Fuji, Braeburn, Gloster, Vista Bella, Jersey Mac, Paulared, McIntosh, Gala, Cortland, Jonathan, JonaGold, Red Delicious, Empire, Idared, Rome, Golden Delicious, Melrose, Crispin/Mutsu, Criterion, Winter Banana, Stayman Winesap, Dorset Golden, Anna, Primicia, Slor, Topred, Bravo Esmolfe, Reineta, Esperiega, Earligold, Gingergold, Pink Lady, etc. (*www.philipsfruit.com, 2004*).

3.3. PRODUCCIÓN MUNDIAL

En los últimos años se ha producido un aumento del cultivo de manzanos en distintas zonas del mundo, pero no siempre ha habido un aumento en su consumo. Por su parte, el mercado europeo de manzanas frescas es el mayor del mundo. La producción anual es de alrededor de 8 millones de Tm. Esto representa un valor anual (a precios de consumidor) de 7 billones de dólares. Con el nuevo mercado desarrollado con los países del Este y Rusia el consumo total ha aumentado a 15 millones de Tm (*www.idfta.org, 2004*). En la **Tabla II** se muestran las cinco variedades más consumidas en Europa y en EEUU (*http://postharvest.tfrec.wsu.edu, 2005*). Como se puede observar en la tabla, las variedades más consumidas son parecidas. España representa el 12% del total de la producción total de la Unión Europea (*www.fas.usda.gov, 2004*).