Comparación de dos métodos supervisados de reconocimiento de patrones para la clasificación de destilados medios de petróleo mediante Espectroscopia Infrarroja

Yumirka Comesaña García, Ángel Dago-Morales, Isneri Talavera Bustamante,* Reinaldo Fernández Fernández y Diana Porro Muñoz.*

Centro de Investigaciones del Petróleo, Washington No. 169 esquina a Churruca, Cerro, Código Postal 10200, Ciudad de La Habana. *Centro de Aplicaciones de Tecnologías de Avanzada, Calle 7ma entre Calles 218 y 222, No. 21812, Reparto Siboney, Playa, Código Postal 12200, Ciudad de La Habana, Cuba.

Recibido: 19 de noviembre de 2008. Aceptado: 8 de mayo de 2009.

Palabras clave: SIMCA, PLS-DA, kerosinas, quimiometria y Espectroscopia Infrarroja.
Key words: SIMCA, PLS-DA, kerosenes, chemometrics and Infrared Spectroscopy.

RESUMEN. En la industria de refinación del petróleo es frecuente el uso de crudos de diferentes orígenes, lo que origina variaciones apreciables en la composición química de los productos obtenidos durante el proceso de refinación. La aplicación de las técnicas quimiométricas de reconocimiento de patrones ha hecho posible el desarrollo de procedimientos alternativos que permiten, a través de métodos espectroscópicos o cromatográficos, la clasificación y el control de calidad de diversos tipos de combustibles de una forma rápida y con el empleo de pequeña cantidad de muestra. El objetivo de este estudio fue la aplicación de los métodos supervisados de reconocimiento de patrones: modelado suave independiente por analogía de clases y análisis discriminante sobre mínimos cuadrados parciales —(SIMCA) y (PLS-DA), de sus siglas en inglés— para la clasificación de kerosinas mediante la técnica de Espectroscopia Infrarroja; y realizar un estudio comparativo de los resultados obtenidos por ambos métodos. Los modelos desarrollados permitieron diferenciar dos clases de kerosinas con diferente composición química, lo cual fue corroborado por Espectrometría de Masas. El estudio comparativo realizado a través de una validación con muestras externas, demostró que el modelo SIMCA presentó una adecuada sensibilidad y selectividad: no presentó falsos negativos y fue capaz de no clasificar a las muestras de turbocombustibles ajenas al sistema modelado; sin embargo, el modelo PLS-DA presentó problemas de selectividad y no logró diferenciar las muestras de turbocombustibles. Se considera que el problema de selectividad del modelo PLS-DA está dado por la influencia que ejerce el conjunto de entrenamiento en el cálculo de los umbrales de las clases.

ABSTRACT. In the oil refining industry, the use of crudes from several origins is frequent. This leads to considerable variations in the chemical composition of the products obtained during the refining process. The application of the chemometric techniques for pattern recognition has made possible the development of alternative methods that allow the classification and quality control of several types of fuels in a quick way, using small quantities of samples, through spectroscopy and chromatography methods. The objective of this study is the application of the supervised pattern recognition methods: Soft Independent Modeling of Class Analogy (SIMCA) and Partial Least Squares Discriminant Analysis (PLS-DA) in the classification of kerosene through infrared spectroscopy. This study is also intended to carry out a comparative study of the results achieved with the application of both methods. The models developed allowed to differentiate two kerosene groups with different chemical compositions, which were corroborated through mass spectrometry. The comparative study carried out through a validation with external samples showed that the SIMCA model had an adequate sensitivity and selectivity: it didn't show false negatives and was unable to classify the turbo fuel samples alien to the modeled system. Nevertheless, the PLS-DA model showed selectivity problems and was unable to differentiate the turbo fuel samples. It was considered that the selectivity problem presented by the PLS-DA model is due to the influence of the training group in the calculation of the classes' thresholds.

Correspondencia:
Angel Dago-Morales
Centro de Investigaciones del Petróleo, Washington No. 169 esquina a Churruca, Cerro, Código Postal 10200, Ciudad de La Habana, Cuba. Correo electrónico: adago@ceinpet.cupet.cu.

INTRODUCCIÓN

En el control de calidad de los destilados medios del petróleo se emplea un grupo de ensayos físico químicos oficialmente establecidos como procedimientos de referencia por diferentes organizaciones internacionales, como por ejemplo: la ASTM (American Society for Testing Materials), la ISO (Internacional Standard Organization) y el IP (Institute of Petroleum). Se debe destacar que las propiedades físico químicas determinadas a través de los métodos de referencia emplean un tiempo de medición considerable y requieren de apreciable cantidad de muestra; además, cuando los productos presentan pequeñas diferencias o anomalías en su composición química, estas no son fácilmente detectables a través de ellas. Una de las tendencias actuales en la industria petroquímica es la aplicación de la quimiometría sobre la base de los resultados de las técnicas de Cromatografía Gaseosa, o de las espectroscópicas de infrarrojo cercano, medio y ultravioleta visible. Estos procedimientos alternativos son empleados en el control de calidad de diferentes productos, en el control de procesos productivos[1-11] y en la detección de adulteraciones de combustibles a través del empleo de los métodos quimiométricos de reconocimiento de patrones.[12-15]

En este trabajo se empleó la Espectroscopia Infrarroja con transformada de Fourier para la caracterización de kerosinas de variada composición. La definición inicial de las categorías o clases se realizó mediante el método de análisis por componentes principales (PCA, por sus siglas en inglés). Se utilizó la Cromatografía Gaseosa acoplada a la Espectrometría de Masas (CG-EM) para definir el porqué de la separación de las muestras de kerosinas en dos clases diferentes.

Los métodos de reconocimiento de patrones utilizados para la clasificación fueron el análisis discriminante sobre mínimos cuadrados parciales (PLS-DA, por sus siglas en inglés) y el modelado blando independiente de analogías de clase (SIMCA, por sus siglas en inglés). Ambos se catalogan como modelos de clasificación blandos; o sea, son capaces de discernir si las muestras u objetos pertenecen a una clase, a más de una, o a ninguna, a diferencia de los modelos fuertes de clasificación; como por ejemplo, el de los k-ésimos vecinos más cercanos, que siempre clasifican a un objeto acorde a su mayor similitud con alguna de las clases presentes en el modelo, independientemente de que pertenezca o no a la citada clase. Los modelos se calcularon y refinaron sobre la base de un conjunto de muestras de entrenamiento y posteriormente, se sometieron a un proceso de validación con un grupo de muestras independientes (conjunto de validación).

El método SIMCA de reconocimiento de patrones desde su introducción por Svante Wold en 1976[16] ha sido uno de los métodos de clasificación más utilizados en el control de calidad en las industrias alimentiaria[17-18] y farmacéutica.[19-20] En la del petróleo, se ha utilizado para la detección de adulteraciones en gasolinas.[14-15] Es un método supervisado de reconocimiento de patrones que se basa en el principio de analogía entre las muestras que pertenecen a una misma clase y emplea las puntuaciones determinadas mediante PCA para el cálculo de las distancias entre los objetos. El método SIMCA calcula un modelo PCA para cada clase o categoría presente en el sistema objeto de estudio, posteriormente, integra cada una de las clases y calcula sus límites o fronteras con una probabilidad dada, comúnmente del 95 %.

El método PLS-DA es una combinación de técnicas de regresión y clasificación; en general, al igual que el método SIMCA, reduce la dimensionalidad de las variables del sistema, pero en este caso, a través de mínimos cuadrados parciales. Una vez calculadas las nuevas variables latentes se realiza el análisis discriminante y se establecen las fronteras entre las clases.[21] La clasificación de objetos en análisis discriminante se hace en función de su probabilidad de pertenencia a una u otra clase: un objeto clasifica dentro de la clase para la cual se obtiene una mayor probabilidad. La región crítica o frontera que separa dos clases puede ser una recta, un plano o hiperplano, en esta región crítica, se igualan las probabilidades de pertenecer a ambas clases. A diferencia del método SIMCA, este método permite detectar directamente las direcciones de mayor variabilidad, hacer mínima la influencia de aquellas variables que no están directamente asociadas a la respuesta; en resumen, permite lograr una rotación de la proyección de las variables latentes para obtener una separación máxima entre las clases. El algoritmo calcula un umbral de clasificación para cada clase utilizando la estadística bayesiana y la distribución observada de los valores predichos; una muestra u objeto pertenece a la clase cuya probabilidad calculada es mayor.

El objetivo del trabajo consistió en aplicar los métodos de reconocimiento de patrones PLS-DA y SIMCA en la clasificación de kerosinas mediante el empleo de los datos experimentales que brinda la Espectroscopia Infrarroja y realizar un estudio comparativo sobre la capacidad de predicción de ambos modelos.

MATERIALES Y MÉTODOS

Se utilizaron 60 muestras de kerosinas y 10 muestras de turbocombustibles provenientes de una refinería de la Unión Cuba Petróleo. Los espectros infrarrojos se obtuvieron en un espectrómetro FT-MIR Avatar 360Esp (Nicolet Instrument Corp.). El funcionamiento del espectrómetro se chequeó de forma periódica con el patrón de poliestireno. Se utilizó celda fija con ventanas de cloruro de sodio de 0,03 mm . La corrección del fondo se realizó con la celda vacía y seca. Previo al análisis, la celda se endulzó dos veces con alícuotas de la propia muestra. Los espectros se midieron en absorbancia mediante la técnica de transmisión en el intervalo entre 4 000 y 400 cm^{-1} con una resolución de 4 cm^{-1}, un paso de 1,929 cm^{-1} y 32 barridos por punto. El análisis PCA realizado a partir de los resultados espectrales evidenció que las muestras de kerosinas se dividen en dos grupos diferentes, los cuales se identificaron como C y V. Los espectros de masas de una muestra de cada grupo se obtuvieron en un cromatógrafo de gases con detector másico HRGC-MS (Konik-Tech), en un intervalo de temperatura desde 50 hasta 320 °C, rampa de calentamiento de 3 °C /min y columna DB-5 de 30 m . El equipo emplea fuente de impacto electrónico a 70 eV, con intervalo de masas de 29 a 600 Dalton.

Para el desarrollo de los modelos de clasificación se emplearon los métodos supervisados de reconocimiento de patrones PLS-DA y SIMCA. El cálculo de los modelos se realizó sobre la base de las dos clases de kerosinas definidas en el análisis PCA. Se utilizaron 40 muestras de kerosinas para conformar el conjunto de entrenamiento, 26 de ellas de la clase C y 14 de la clase V. Para validar los modelos se utilizaron 30 muestras: 13 de la clase C, 7 de la V y se agregaron 10 de turbocombustibles para verificar la selectividad y robustez de los modelos. Para el cálculo de los modelos, se utilizó la zona espectral comprendida entre 1 750 y 690 cm^{-1}, (553 variables por muestra) que es la zona de mayor informa-

ción o huella espectral del sistema objeto de estudio; por lo cual las dimensiones de las matrices X de trabajo para el conjunto de calibración y validación fueron 40 x 553 y 30 x 553, respectivamente. Para minimizar el efecto del corrimiento de la línea base, se aplicó la corrección multiplicativa de la señal. Se utilizó como preprocesamiento el centrado en la media. El número óptimo de componentes principales se seleccionó mediante validación cruzada. Para la evaluación del modelo SIMCA se utilizaron los diagramas Coomans y los estadígrafos Q residual y T^2 de Hotelling. Se fijó en un 5 % el nivel de significación de los límites de las fronteras que definen cada clase. En el procesamiento matemático y estadístico se emplearon los programas PLS-Toolbox 3,5[22] y Pirouette versión 3,11.[23]

RESULTADOS Y DISCUSIÓN

Por simple inspección visual no es posible detectar diferencias apreciables entre los espectros infrarrojos de las 60 muestras de kerosinas en el intervalo espectral (690 a 1 750 cm^{-1}) en el cual se calcularon los diferentes modelos de clasificación (Fig. 1). Sin embargo, después de aplicar el método de análisis por componentes principales a estos datos espectrales (Fig. 2), sí se observaron diferencias entre las muestras de kerosinas del conjunto global: a lo largo del primer componente principal las muestras se separan en dos grupos; las muestras del grupo identificado como C se concentran en la parte negativa del primer componente principal, las muestras del grupo V se ubican en su parte positiva.

Para confirmar la causa de la separación en dos clases diferentes se seleccionó una muestra de cada una de clases y se analizaron por CG-EM. Los resultados demostraron que la separación en dos clases está dada por apreciables diferencias en su composición química. Los cromatogramas de iones totales (CIT) de las muestras analizadas (Fig. 3) revelaron que la kerosina clase V presenta componentes más pesados que la de la clase C, y está menos enriquecida en compuestos ligeros. De igual forma, los fragmentogramas obtenidos en el espectrómetro de masas corroboraron que la muestra clasificada como clase V presenta los iones de mayor peso molecular. El fragmentograma del ión 57 tipifica la riqueza en hidrocarburos saturados alifáticos de la muestra de la clase V en comparación con la muestra de la clase C.

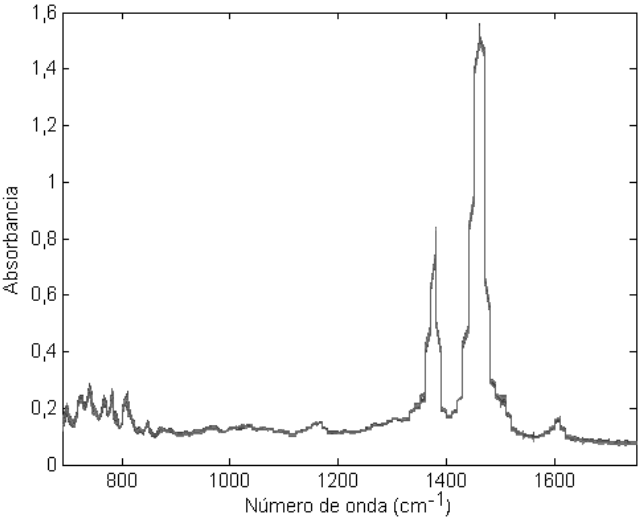


Fig. 1. *Espectros infrarrojos de las muestras de kerosinas (conjunto global).*

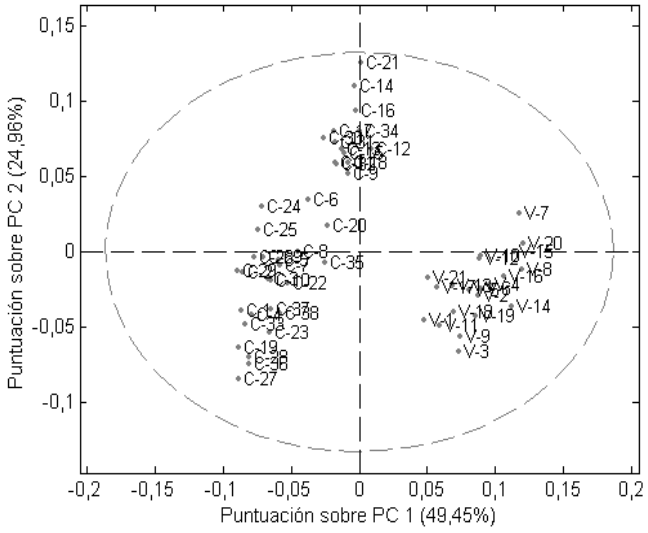


Fig. 2. *Resultados del modelo PCA obtenido a partir de los espectros infrarrojos de las muestras de kerosinas: puntuaciones sobre los dos primeros componentes principales.*

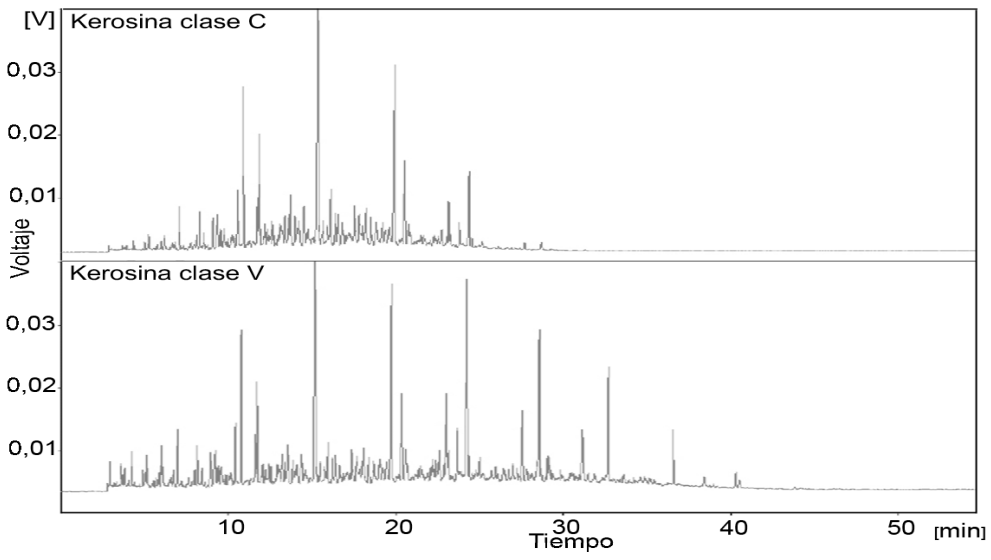


Fig. 3. *Cromatogramas de iones totales de muestras de kerosinas de las clases C y V.*

Cálculo de los modelos PLS-DA y SIMCA

Los modelos se calcularon en base a las dos clases definidas con anterioridad por PCA y CG-EM. Tanto para el modelo SIMCA como el PLS-DA los números de muestras para las clase C y V fueron 26 y 14, respectivamente. Para el modelo SIMCA el número de factores resultó cinco para ambas clases y para el modelo PLS-DA, seis. La varianza explicada (en tanto por ciento) por los modelos resultaron: modelo SIMCA, clase C (93,4), clase V (88,4) y modelo PLS-DA, para ambas clases (88,9).

Ambos modelos clasifican correctamente a todas las muestras del conjunto de entrenamiento. En el diagrama de Coomans calculado para el modelo SIMCA se reportan las distancias ortogonales de las muestras del conjunto de entrenamiento con respecto a las dos clases: se destaca la clasificación correcta de las muestras de las clases C y V (Fig. 4). Se obtuvo una distancia entre las clases de 3,8: la separación entre las clases se considera adecuada cuando este parámetro es mayor que 3,0.[23]

Los resultados del modelo PLS-DA, en particular, los valores de puntuación de las muestras del conjunto de entrenamiento en la predicción de la clase C (Fig. 5), revelan una evidente separación de ambas clases en torno al umbral de probabilidad calculado y que no se presentan errores de clasificación.

Validación de los modelos PLS-DA y SIMCA

El grupo de validación se conformó con 13 muestras de la clase C, 7 de la clase V y 10 muestras de turbocombustibles (T1-T10). Estas últimas eran ajenas a los modelos

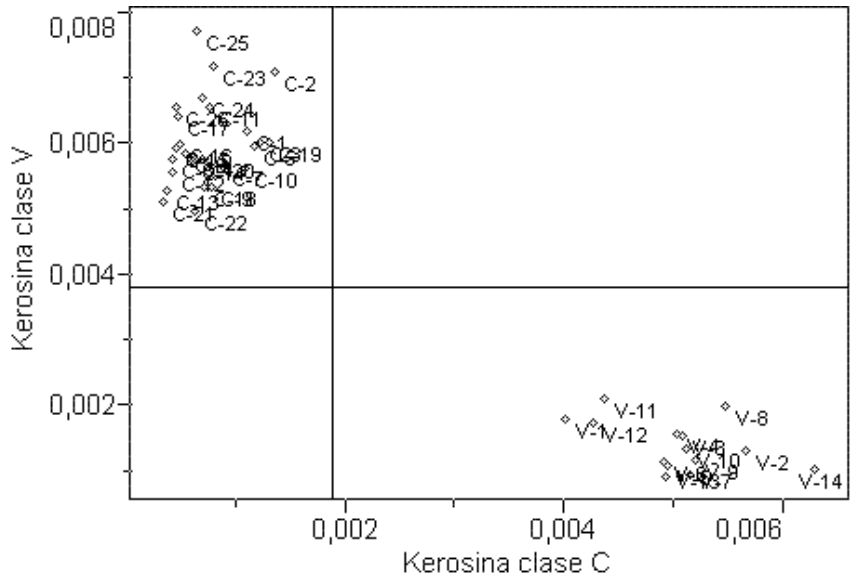


Fig. 4. *Modelo SIMCA: Diagrama de Coomans: Distancias de las muestras del conjunto de calibración con respecto a las clases C y V (con líneas continuas se destacan los límites de las clases para un nivel de significación del 5 %).*

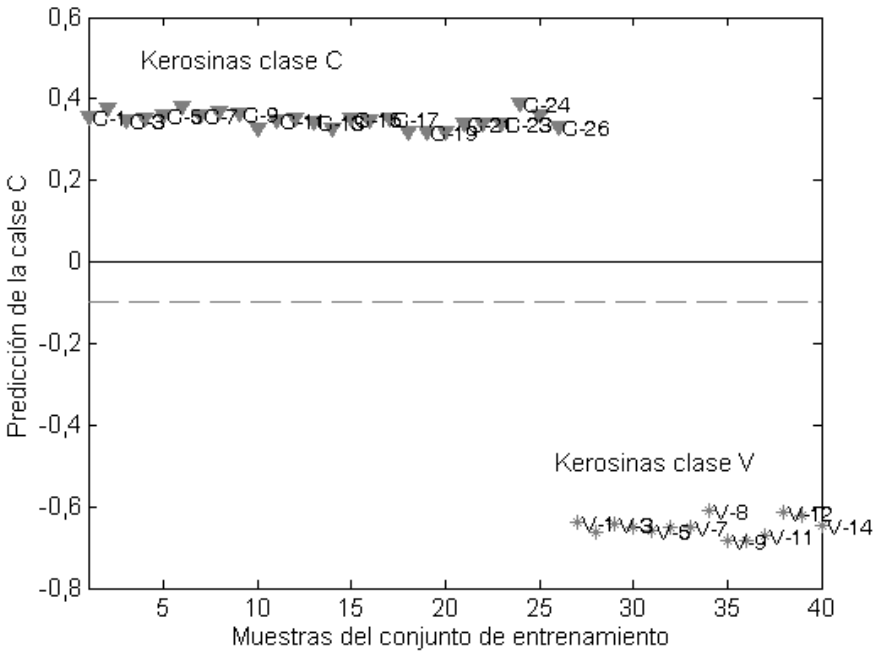


Fig. 5. *Modelo PLS-DA. Resultados de la predicción de la clase C (con líneas discontinuas se presenta el umbral de clasificación calculado para la clase C).*

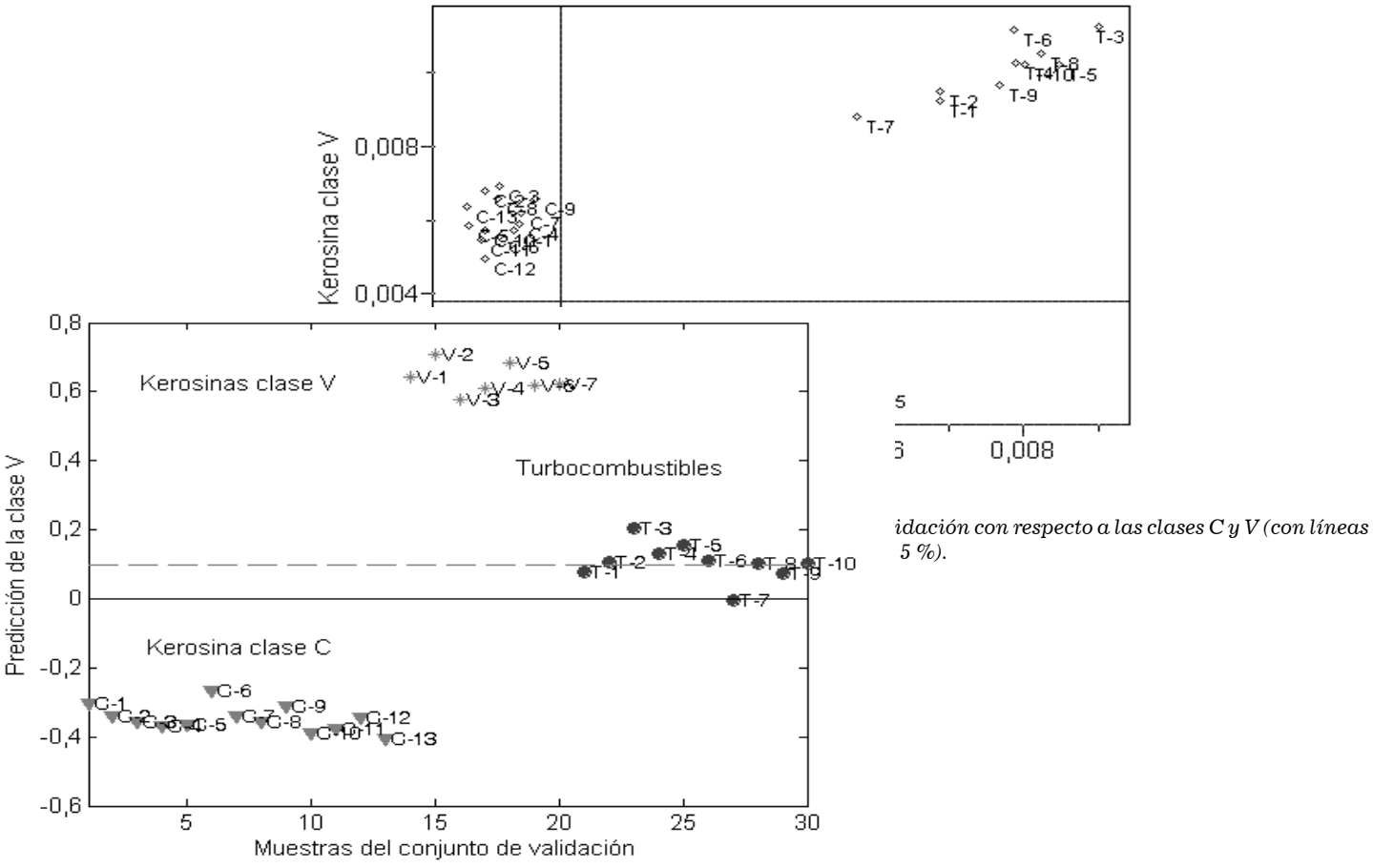
desarrollados y fueron incluidas en el grupo de validación para verificar la selectividad de estos; es decir, para comprobar si los modelos eran capaces de reconocerlas como que no pertenecen a ninguna de las clases.

Los resultados de la validación del modelo SIMCA fueron satisfactorios, las muestras pertenecientes a las clases C y V clasificaron correctamente en sus respectivas clases y las muestras de turbocombustibles cayeron fuera de los límites que definen las fronteras de las clases; o sea, fueron correctamente clasificadas como no pertenecientes a ninguna de las clases del modelo (Fig. 6).

La clasificación de las muestras de las clases C y V en el proceso de validación del modelo PLS-DA fue correcta (Fig. 7) y la separación entre las clases fue adecuada; sin embargo, las muestras de turbocombustibles —ajenas al modelo— no clasificaron bien: T1, T7 y T9 clasificaron como pertenecientes a la clase C y el resto, como pertenecientes a la clase V.

Los resultados del estudio comparativo demuestran que el modelo SIMCA tuvo una adecuada sensibilidad y selectividad: no presentó falsos negativos y fue capaz de no clasificar a las muestras de turbocombustibles ajenas al modelo. Los autores consideran que el problema de selectividad que presentó el modelo PLS-DA estuvo dado por la influencia que ejerce el conjunto de entrenamiento en el cálculo de los umbrales de las clases; o sea, si las muestras problemas presentan pequeñas variaciones en su composición química y estas variaciones no fueron contempladas durante el desarrollo del modelo, entonces el umbral calculado no es representativo y no delimita correctamente a las muestras discrepantes.

Se debe destacar que las propiedades físico químicas que comúnmente se miden en el control de calidad de los destilados medios (viscosidad, densidad, azufre, temperatura de inflamación y destilación) no fueron capaces de detectar las pequeñas diferencias de la composición química del sistema estudiado. El trabajo desarrollado demuestra que la Espectroscopia Infrarroja, en combinación con el método SIMCA de reconocimien-



idación con respecto a las clases C y V (con líneas 5 %).

Fig. 7. *Validación del modelo PLS-DA. Resultados de la predicción de la clase V (con líneas discontinuas se presenta el umbral de clasificación calculado para la clase V).*

to de patrones puede utilizarse en el control de calidad de los destilados medios del petróleo como procedimiento alternativo, lo cual no solo presenta ventajas en cuanto a la rapidez de la respuesta analítica, sino también, en el consumo de muestras y reactivos. Las refinerías y laboratorios de investigación cubanos pueden hacer extensiva esta metodología a otros destilados medios del petróleo (diesel y turbocombustibles).

CONCLUSIONES

Se emplearon los métodos de reconocimiento de patrones supervisados SIMCA y PLSDA para la clasificación de kerosinas a partir de los datos que brinda la técnica de espectroscopía infrarroja. Los modelos SIMCA y PLS-DA calculados se validaron mediante un grupo de muestras que no formaron parte del conjunto de entrenamiento. El estudio comparativo llevado a cabo a partir de los resultados de la validación demostró que ambos modelos poseen una adecuada sensibilidad; sin embargo, solo el modelo SIMCA tiene la selectividad y robustez que exige este tipo de sistema. La metodología desarrollada puede constituir una herramienta útil para el control de calidad de los destilados medios del petróleo.

BIBLIOGRAFÍA

1. Kidajat K, Chong SM. Quality characterization of crude oils by partial least squares calibration of NIR spectral profile. J. of Near Infrared Spect. 2000;8:53-9.
2. Reboucas MV, Barros Neto. Near infrared spectroscopic prediction of physical properties of aromatic-rich hydrocarbon mixtures. J. of Near Infrared Spect. 2001;9:263-73.
3. Gómez-Carracedo M, Andrade J, Calviño M, Prada D, Fernández E, Muniategui S. Generation and mid-IR measurement of a gas-phase to predict security parameters of aviation jet fuel. Talanta. 2003;60:1051-62.
4. Felício Candida C, Bras Ligia P, Lopez João A, Cabrita L, Menezes J. Comparison of PLS algorithms in gasoline and gas oil parameter monitoring with MIR and NIR. Chemometrics Intell. Lab. Syst. 2005;78:74-80.
5. Vianney O, Flavia C, Daniella G, Andréa C, Edgardo G, Paulo A, *et al*. A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. Anal Chim Acta. 2005;547:188-96.
6. Brudzewski K, Kesik A, Kolodziejczyk K, Zborowska U, Ulaczyk J. Gasoline quality prediction using gas chromatography and FT-IR spectroscopy: An artificial intelligence approach. Fuel. 2006;85:553-8.
7. Pasadakis N, Sourligas S, Foteinopoulos Ch. Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks. Fuel. 2006;85:1131-7.
8. Hongfu Y, Xiaoli C, Haoran L, Yupeng X. Determination of multi-properties of residual oils using mid-infrared attenuated total reflection spectroscopy. Fuel. 2006; 85:1720-8.
9. Caneca AR, Fernada Pimentel M, Kawakami Harrop RG, Eliane da Matta C, Rodrigues de Carvalho F, Raimundo MIJr, *et al*. Assesment of infrared spectroscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricanting oils. Talanta. 2006;70:344-52.
10. Figueiredo dos Santos R, Kawakami Harrop RG, Ugulino Araujo MC, Cirino da Silva E. Improvement of prediction ability of PLS models employing the wavelet packet transform: A case study concerning FT-IR determination gasoline parameters. Talanta. 2007;71:1136-43.
11. Balabin RM, Safieva RZ, Lomakina EI. Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction. Chem. and Int. Lab. Syst. 2007;88:183-8.
12. Pasadakis N, Kardamakis Andreas A. Identifying constituents in commercial gasoline using Fourier transform-infrared spectroscopy and independent component analysis. Anal. Chim. Acta. 2006;578:250-5.
13. Santos de Oliveira F, Gomes Teixeira LS, Ugulino Araujo MC, Korn M. Screening analysis to detect adulterations in brazilian gasoline samples using distillation curves. Fuel. 2004;83:917-23.
14. Wiedemann LSM, d´Avila LA, Azevedo DA. Adulteration detection of brazilian gasoline samples by statistical analysis. Fuel. 2005;84:467-73.
15. Flumignan DL, Tininis AG, Ferreira FO, de Oliveira JE. Screening brazilian C gasoline quality: Application of the SIMCA chemometric method to gas chromatographic data. Anal Chim Acta. 2007;595:128-35.
16. Wold S. Pattern recognition by means of disjoint principal component models. Pattern Recognition. 1976;8:127-139.
17. Camara José S, Arminda Alves M, Marques José C. Multivariate analysis for the classification and diferentiation of Madeira wines according to main grape varieties. Talanta. 2006;68:1512-21.
18. Guler C. Characterization of Turkish bottled waters using pattern recognition methods. Chemometrics and Intelligent Laboratory System. 2007;86:86-94.
19. Gabrielsson J, Lindberg NO, Lundstedt T. Multivariate methods in pharmaceutical applications. J of Chemometrics. 2002;16:141-60.
20. Nic Daéid N, Waddell Ruth JH. The analytical and chemometric procedures used to profile illicit drug seizures. Talanta. 2005;67:280-5.
21. Barker M, Rayens W. Partial squares for discrimination. J Chemometrics. 2003;17:166-173.
22. PLS-Toolbox 3,5 for use with MATLAB, Eigenvector Research, Inc, version 3,5, 2004.
23. Pirouette, Infometrix Inc, Woodinville WA, version 3,11, 2003.