REGRESIÓN DEL ÍNDICE DE CETANO DEL DIÉSEL A PARTIR DE PROPIEDADES MACROSCÓPICAS Y ESPECTROS INFRARROJOS UTILIZANDO MLR Y PLS

JORGE MARIO MORALES MONTAÑO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOQUÍMICAS
ESCUELA DE INGENIERÍA QUÍMICA
BUCARAMANGA
2015

REGRESIÓN DEL ÍNDICE DE CETANO DEL DIÉSEL A PARTIR DE PROPIEDADES MACROSCÓPICAS Y ESPECTROS INFRARROJOS UTILIZANDO MLR Y PLS

JORGE MARIO MORALES MONTAÑO

Trabajo de grado presentado para optar al título de Ingeniero Químico

Director
Prof. Giovanni Morales Medina, Ing. Qco. Dr.

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOQUÍMICAS
ESCUELA DE INGENIERÍA QUÍMICA
BUCARAMANGA
2015

TABLA DE CONTENIDO

Pág.

INTRODUCCIÓN 12

1. MARCO TEÓRICO 13
1.1. EL DIÉSEL 13
1.2. PROPIEDADES DEL DIÉSEL 14
1.2.1. Número de Cetano. 14
1.2.2. Otras propiedades. 16
1.3 QUIMIOMETRIA 16
1.4 REGRESIÓN LINEAL MÚLTIPLE 17
1.5 ANÁLISIS POR COMPONENTES PRINCIPALES (PCA) 18
1.6 REGRESIÓN POR MÍNIMOS CUADRADOS PARCIALES (PLS) 20
1.7 ESPECTROSCOPIA INFRARROJA 21
1.8 REGIONES ESPECTRALES 22

2. METODOLOGÍA 23
2.1 REGRESIÓN LINEAL MÚLTIPLE 24
2.2 REGRESIÓN MULTIVARIADA POR PCR y PLS 26

3. ANÁLISIS DE RESULTADOS 28
3.1 RLM DE PROPIEDADES MACROSCÓPICAS 28
3.2 PLS Y PCA DE PROPIEDADES MACROSCÓPICAS 38
3.2.1 Análisis por componentes Principales (PCA). 38
3.2.2 Desarrollo del Modelo PLS. 40
3.3 PCA Y PLS A PARTIR DE ESPECTROS INFRARROJOS 41

4. CONCLUSIONES 47

5. RECOMENDACIONES.....48

REFERENCIAS BIBLIOGRAFICA.....49

BIBLIOGRAFÍA.....53

ANEXOS.....56

LISTA DE TABLAS

Pág.

Tabla 1. Estadísticos univariados representativos de la base de datos macroscópica. 28
Tabla 2. Análisis de la varianza (ANOVA) del modelo inicial. 30
Tabla 3. Características modelo inicial. 31
Tabla 4. Puntos atípicos identificados 31
Tabla 5.Estadísticos del Modelo sin puntos atípicos e influyentes 32
Tabla 6. ANOVA Modelo Inicial sin Outliers 33
Tabla 7. ANOVA Modelo con 7 Variables 34
Tabla 8. % Peso de las Variables 36
Tabla 9. ANOVA modelo reducido 36
Tabla 10.Estadísticos de desempeño para los modelos de regresión. 37
Tabla 11. Intervalos de Confianza Modelo reducido 38
Tabla 12. Rangos Índice de cetano en el espectro Infrarrojo 42
Tabla 13.Contribución de los PCs para la base de espectros NIR 42
Tabla 14. Estadísticos de Desempeño de los Modelos 45

LISTA DE FIGURAS

Pág.

Figura 1. Productos de la destilación del Petróleo 13
Figura 2. Descomposición en componentes principales 19
Figura 3. Descripción gráfica del Modelo PLS 20
Figura 4. Densidad vs Índice de Refracción. 29
Figura 5.Temp 50% vs Peso Molecular. 29
Figura 6.Leverage para los datos del Modelo 8 Var. 31
Figura 7. Distancia de Cook para los datos del Modelo 8 Var. 31
Figura 8. Modelo 7 Variables MLR.....34
Figura 9. Modelo RLM reducido a 4 variables.....37
Figura 10. Varianza vs PCs para el PCA de propiedades macroscópicas.....39
Figura 11.Scores PCA Propiedades 39
Figura 12. Posibles puntos atípicos (región de influencia). 40
Figura 13.Modelo PLS Propiedades 41
Figura 14.Score Espectro Infrarrojos 43
Figura 15. Gráfico de influencia del PCA de espectros NIR. 43
Figura 16. Ajuste PLS basada en NIR. Calibración en azul, validación cruzada en rojo. 44

LISTA DE ANEXOS

Pág.

Anexo A. Datos de las propiedades Macroscópicas 56
Anexo B.. Matriz de Correlación 68
Anexo C. Codigos 69
Anexo D. Técnicas de pre-tratamiento de datos espectrales.....71
Anexo. E. Leverage y distancia COOK 73
Anexo F. CI 976 vs CI 4737 77
Anexo G. Escalamiento de Datos 81
Anexo I. Espectros Normalizados 85

RESUMEN

TITULO: REGRESIÓN DEL ÍNDICE DE CETANO DEL DIÉSEL A PARTIR DE PROPIEDADES MACROSCÓPICAS Y ESPECTROS INFRARROJOS UTILIZANDO MLR Y PLS.[1]

AUTOR: JORGE MARIO MORALES MONTAÑO[2]

PALABRAS CLAVES: MLR, PCA, PLS, ESPECTROS INFRARROJOS, MACRO PROPIEDADES, DIÉSEL, ÍNDICE DE CETANO.

RESUMEN:

Una base de datos de 140 muestras de propiedades macroscópicas del diésel fue analizada mediante las técnicas de regresión mediantes modelos lineales y modelos no-lineales: Regresión Lineal Múltiple (MLR) y Regresión por mínimos cuadrados (PLS) para obtener modelos matemáticos que me permitan cuantificar el índice de cetano del diésel.

Para el análisis de estos modelos se tuvieron en cuenta diferentes criterios como el coeficiente de correlación el error estándar y el criterio de selección de Akaike's. Además se analizaron 62 espectros infrarrojos de muestras de diésel proporcionadas por la sección de *blending* de la gerencia refinería de Barrancabermeja de ECOPETROL S.A. donde se encontró una relación entre las intensidades espectrales y el incide de cetano, los cuales fueron utilizados para la creación del modelo no-lineal.

Entre las principales ventajas de utilizar los modelos de regresión se encuentran: comprensión de las variables influyentes en los procesos, determinación de los estadísticos que presentan los datos, análisis de datos atípicos e influyentes y predicción de nuevos valores de propiedades.

Los resultados de este trabajo demuestran que los modelos obtenidos mostraron una gran capacidad de predicción del Índice de cetano, sin embargo por la facilidad de disposición y obtención de los datos se afirma que el mejor modelo de predicción corresponde a la regresión parcial de mínimos cuadrados a partir de espectroscopia infrarroja cercana. Donde se recomienda ampliar la base de datos con la finalidad de validar estos modelos para obtener una mayor exactitud a la hora de aplicar estos modelos.

[1] Trabajo de Grado

[2] Facultad de Ingenierías Fisicoquímicas, Escuela de Ingeniería Química. Director: Giovanni Morales

ABSTRACT

TÍTLE: REGRESSION OF CETANE NUMBER OF DIESEL FROM MACROSCOPIC PROPERTIES AND INFRARED SPECTRA USING MLR AND PLS[3]

AUTHOR: JORGE MARIO MORALES MONTAÑO.[**]

KEY WORDS: Multiple linear regression; Partial least squares; Diesel; Near infrared spectroscopy; Cetane index

DESCRIPTIÓN:

A database of 140 samples of macroscopic properties of diesel was analyzed using regression techniques using linear models and nonlinear models: Multiple Linear Regression (MLR) and Regression least squares (PLS) for mathematical models that allow me to quantify the cetane number of diesel.

For the analysis of these models different criteria such as the correlation coefficient and the standard error of the selection criteria Akaike's were taken into account. Besides infrared spectra of 62 samples provided by diesel blending section management Barrancabermeja refinery ECOPETROL SA analyzed where a relationship between the spectral intensities and incident cetane, which were used to create non-linear model was found.

Among the main advantages of using regression models include: understanding the influential variables in the process, determination of the statistical data presented, analysis of outliers and influential data and prediction of new property values.

The results of this work show that the obtained models showed a great ability to predict cetane index, however for ease of layout and data collection is stated that the best prediction model corresponding to the partial least squares regression from near-infrared spectroscopy. Where it is recommended to expand the database in order to validate these models for greater accuracy when applying these models.

[3] Degree Work

[**] Faculty of Phisico-Chemical Engineering. School of Chemical Engineerign. Director: Giovanni Morales Medina

INTRODUCCIÓN

En la industria petroquímica existe un gran interés por el continuo desarrollo de tecnologías y metodologías que permitan obtener información de calidad, que requieran menor costo y menor tiempo de trabajo. Es así como los modelos de regresión han surgido como alternativa para la obtención de datos para la toma de decisiones sobre la marcha.

Los modelos de regresión se han desarrollado utilizando diferentes datos, usualmente provenientes de ensayos estandarizados de laboratorio; para la industria del petróleo, los ensayos estándar se toman mayoritariamente de las normas ASTM. En las últimas décadas, los modelos se han desarrollado utilizando los resultados de técnicas espectroscópicas. Los resultados de estas técnicas han reemplazado a costosos ensayos de laboratorio para la obtención de propiedades macroscópicas; la principal ventaja de estas técnicas es que no requieren pretratamientos complejos para las muestras, la toma de espectros toma poco tiempo y en ocasiones presenta un menor costo [1].

La información, ya sea proveniente de pruebas estándar o por técnicas espectroscópicas, puede ser analizada por medio de técnicas estadísticas multivariadas con el objetivo de interpretar y proponer regresiones de propiedades de difícil o costosa determinación. El objetivo de este proyecto es el desarrollo de modelos de regresión lineal múltiple y modelos de mínimos cuadrados parciales (PLS: *partial least square*) para la predicción del índice de cetano del diésel a partir de datos de propiedades macroscópicas y espectros de infrarrojo cercano (NIR).

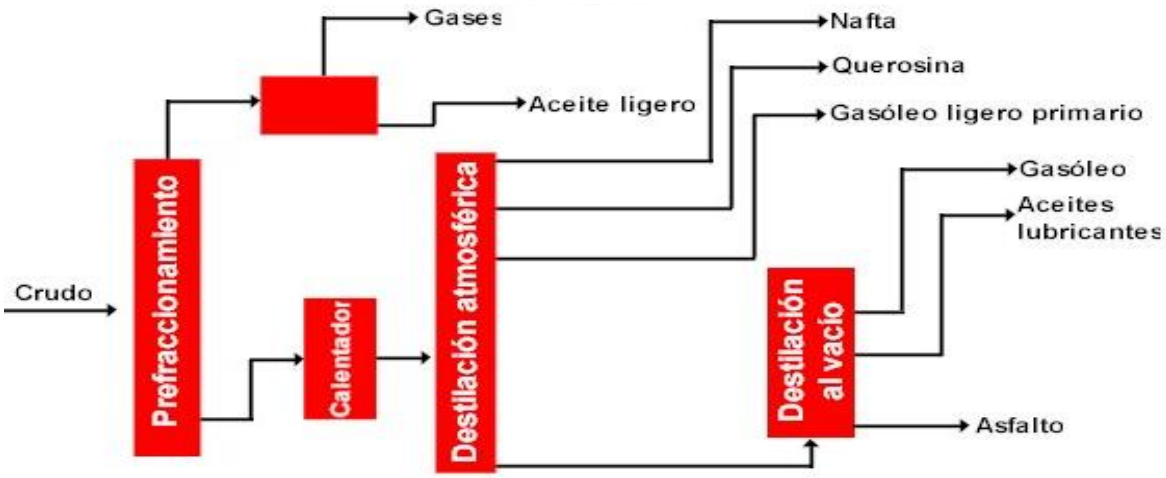
1. MARCO TEÓRICO

1.1. EL DIÉSEL

En una refinería, el petróleo es convertido a una variedad de productos mediante procesos físicos y químicos. El primer proceso al que se somete el petróleo en la refinería, es la destilación para ser separado en diferentes fracciones. Dentro de las torres de destilación, los líquidos y los vapores se separan en fracciones de acuerdo a su punto de ebullición (Figura 1). Las fracciones más ligeras, incluyendo gasolinas y gas LP, vaporizan y suben hasta la parte superior de la torre donde se condensan. Los líquidos medianamente pesados, como el queroseno y la fracción diésel, se quedan en la parte media, mientras que los más pesados efluyen desde el fondo. Las gasolinas contienen fracciones con punto de ebullición por debajo de 200°C, mientras que para el diésel, el límite superior de ebullición corresponde a 350°C. Lo anterior se debe a que el diésel está constituido por moléculas de entre 10 y 20 carbonos, mientras que los componentes de la gasolina se ubican en el orden de menos de 12 carbonos [2].

El combustible diésel, también se procesa, en muchos casos a partir de mezclas de gasóleos con queroseno, y aceite de cíclico ligero, el cual es producto del proceso de craqueo catalítico fluidizado. Con estas corrientes de diésel disponibles, el proceso de producción de diésel comercial en una refinería comprende escoger y mezclar las diferentes fracciones para cumplir con las especificaciones del mercado.

Figura 1. Productos de la destilación del Petróleo



1.2. PROPIEDADES DEL DIÉSEL

1.2.1. Número de Cetano. Es una medida de la calidad de ignición de un combustible e influye en las respectivas emisiones de humo y en la eficiencia de los motores; su determinación está especificada en la norma ASTM D 613. La escala de medición se basa en las características de ignición de dos hidrocarburos, el *n*-hexadecano (cetano) que tiene un periodo corto de retardo durante la ignición se le asigna un número de cetano de 100, mientras que para el heptametilnonano que reporta un periodo largo de retardo se le asigna un número de cetano de 15.[3]

En la práctica la determinación del número de cetano es costosa y laboriosa, por lo cual el instituto americano del petróleo propone una ecuación de dos parámetros (ASTM D 976) para la estimación del número de cetano por medio del índice de cetano, según la Ecuación 1 [4]:

$$Indice\ de\ cetano\ (ASTM\ D-976) = 454.74 - 1641.416\,D + 774.74\,D^2 - 0.554B + 97.803(logB)^2$$

(Ecuación 1)

Donde, D: Densidad a 15 °C, g/mL, determinada mediante D1298 o D4052 y B:

Temperatura media de ebullición (T50%), °C, determinada mediante ASTM D-86 y la presión barométrica estándar.

El número de cetano o su índice (IC) corresponden al porcentaje de cetano en una mezcla cetano- heptametilnonano con las mismas características de ignición que la muestra de diésel analizada. El IC depende del diseño y tamaño del motor, de las variaciones de la carga y velocidad y condiciones de arranque y atmosférica, típicamente los motores se diseñan para utilizar índices de cetano entre 40 y 50, debajo de 38 se incrementa rápidamente el retardo de la ignición, los ruidos en el motor y el peso molecular de las emisiones. Asimismo, el IC se incrementa a medida que aumenta la longitud de la cadena. En general, los aromáticos y los alcoholes tienen un índice de cetano bajo. Por ello el porcentaje de gasóleo utilizado en la preparación de diésel se ve limitado por su contenido de aromáticos [5]. También, existen algunos aditivos que pueden elevar el IC.

La norma ASTM D 4737 reporta una regresión basada en cuatro parámetros para calcular el IC, a este se le llama índice de cetano calculado y se puede utilizar cuando no está disponible un motor de prueba para determinar esta propiedad por la norma ASTM D-613 [6]. El IC calculado se puede determinar mediante la siguiente ecuación (considerando diferentes limitaciones comentadas en la ASTM D-4737). [7]

$$CCI = 45.2 + (0.0892)(T_{10N}) + [0.131 + (0.901)B][T_{50N}] + [0.0523 - (0.420)B][T_{90N}] + [0.00049][(T_{10N})^2 - (T_{90N})^2] + 107B + 60B^2$$

(Ecuación 2)

Donde:

$$B = \exp[-3.5(DN)] - 1; DN = d_{15} - 0.85; d_{15} = Densidad\ a\ 15°C; T_{10N} = T_{10} - 215; T_{50N} = T_{50} - 260; T_{90N} = T_{90} - 310; T_{10} = Temperatura\ de\ ebullicion\ al\ 10\%; T_{50} = Temperatura\ de\ ebullicion\ al\ 50\%; T_{90} = Temperatura\ de\ ebullicion\ al\ 90\%; Norma\ ASTM\ D86$$

1.2.2. Otras propiedades. El azufre contribuye al desgaste del motor y a la aparición de depósitos que varían considerablemente en importancia dependiendo en gran medida de las condiciones de funcionamiento del motor.

Las variaciones en la densidad y viscosidad de los combustibles resultan en variaciones en la potencia del motor y consecuentemente en las emisiones y el consumo. La influencia del contenido de poliaromáticos en el combustible afecta la formación de PM y las emisiones de este tipo de hidrocarburos en el tubo de escape. [8]. Los valores para las propiedades macroscópicas del diésel corresponden a una consecuencia de su composición y varían en un determinado rango para los diferentes diésel del petróleo. De acuerdo a los modelos predictivos propuestos para la predicción de las propiedades de las fracciones del petróleo [9], el índice de cetano puede ser predicho a través de un modelo de regresión utilizando otras propiedades del diésel como variables independientes.

1.3 QUIMIOMETRIA

El termino quimiometria puede ser descrito como una herramienta matemática que permite separar la información más relevante contenida en los datos obtenidos de un experimento de análisis químico por medio del análisis estadístico multivariado

[10]. Una muestra puede contener una o más variables independientes que se requieran determinar una o más propiedades de interés. En la mayoría de los métodos de análisis instrumental cientos o miles de datos que referencian propiedades microscópicas son generados para analizar el comportamiento macroscópico de una determinada muestra. Los métodos estadísticos permiten el análisis metodológico de estos conjuntos de datos y han sido aplicados en el análisis de sistemas complejos, incluyendo aguas residuales, alimentos, aceites y el petróleo y sus derivados [11].

De acuerdo a la cantidad de variables independientes y dependientes, el análisis de los datos de propiedades químicas puede desarrollarse por métodos estadísticos tradicionales o métodos estadísticos multivariados [12].En particular, para la información proveniente de la espectroscopia IR, los métodos estadísticos corresponden a los multivariados. En el análisis estadístico multivariado la información química de un conjunto de mezclas patrón (representada en longitudes de onda, absorbancias, tiempos de retención, intensidad de señal, etc.), es transformada utilizando diferentes metodologías para extraer valores que pueden ser traducidos en agrupaciones, tendencias y regresiones [13].

1.4 REGRESIÓN LINEAL MÚLTIPLE

En la regresión lineal múltiple (RLM) se utilizan más de una variable explicativa o independiente x_k para la predicción de una variable dependiente y. En la RLM se considera que los valores de la variable dependiente y han sido generados por una combinación lineal de las variables explicativas, según:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \cdots + b_k * x_k + u$$

Donde:

y: Variable de respuesta o variable dependiente.

$b_0, b_1, b_2, \ldots, b_k$: Son los coeficientes de la regresión.

$x_1, x_2, x_3, \ldots, x_k$: Son las variables predictoras o independientes.

Los coeficientes son obtenidos de forma que la suma de cuadrados entre los valores observados y los pronosticados sea mínima (sumatoria del cuadrado de los residuos). Para el análisis del modelo de regresión múltiple se realizan diferentes pruebas y diagnósticos tales como la prueba F, el análisis de los residuos y el análisis de normalidad [14], con el objetivo de verificar los supuestos de la RLM.

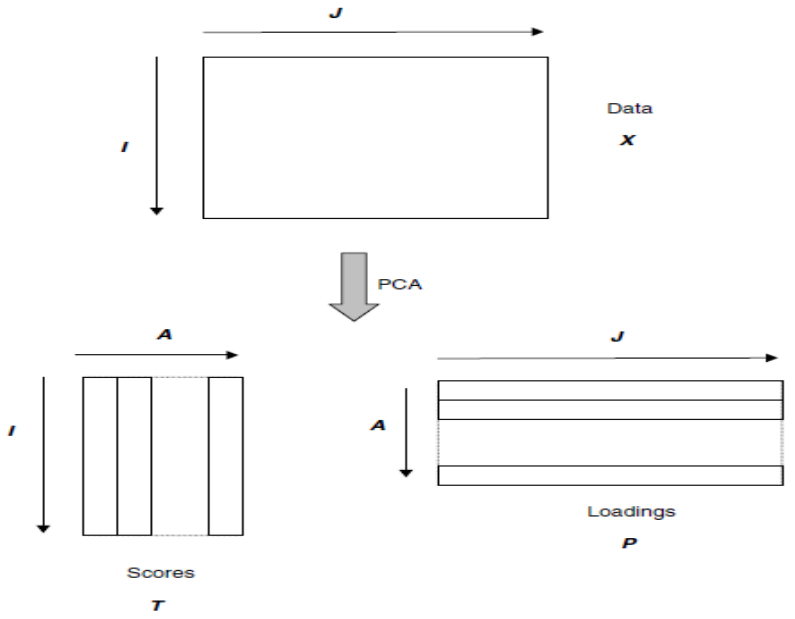
1.5 ANÁLISIS POR COMPONENTES PRINCIPALES (PCA)

El análisis componentes principales, PCA (principal component analysis), reduce el número de dimensiones del sistema, concentrando una gran proporción de la varianza en un conjunto reducido de variables no correlacionadas denominadas componentes principales (CP); el análisis de los resultados de la obtención de los CP puede ser utilizado para el reconocimiento de tendencias, agrupaciones y regresión y predicción de propiedades [15].

Los CP son ejes ortogonales sobre los cuales son proyectados los datos iniciales, anulando de esta forma la colinealidad entre las variables del problema inicial. Para esto se parte de la matriz de datos iniciales X, donde cada fila es una muestra o caso y cada columna es una variable. La matriz de covarianza de X es

sometida a un proceso de denominado diagonalización para obtener los PC, del cual se derivan unos valores propios (*scores*) y vectores propios (*loadings*) que contienen información sobre las interrelaciones entre los casos y las variables del problema inicial (Figura 2) [10].

Figura 2. Descomposición en componentes principales

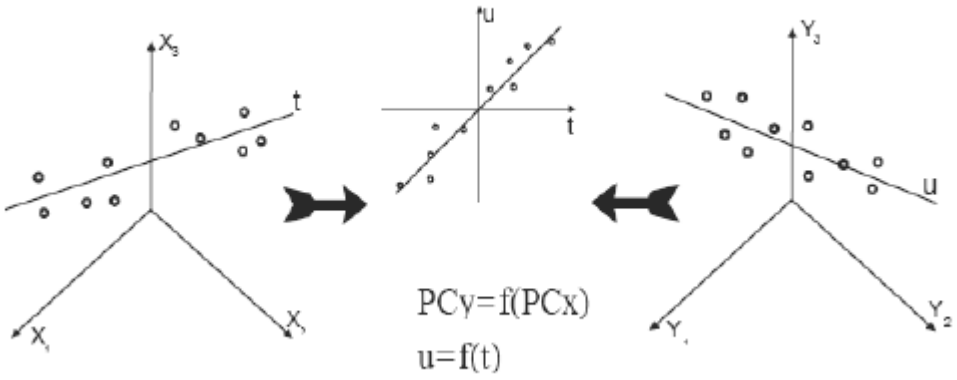


El primer componente principal obtenido en el PCA corresponde a la dirección que explica la máxima variabilidad de la matriz de datos iniciales X. El segundo PC explica la máxima variabilidad de la matriz X no contenida en el primer componente; por consiguiente, los subsecuentes PC explican las varianzas no contenidas en los anteriores [10]. De esta forma, los PC pueden ser utilizados como variables independientes no colineales en un algoritmo de regresión por mínimos cuadrados para obtener una ecuación de predicción de una propiedad del problema inicial [10]; este algoritmo se denomina comúnmente como regresión por componentes principales (PCR, *principal component regression*).

1.6 REGRESIÓN POR MÍNIMOS CUADRADOS PARCIALES (PLS)

Uno de los métodos más usados en quimiometria es el método de mínimos cuadrados parciales, PLS (*partial least square*). Este método, relacionado con la regresión de componentes principales, PCR posee ventajas teóricas y computacionales que han llevado a innumerables aplicaciones [15]. Durante la etapa de calibración, el PLS utiliza tanto la información contenida en la matriz de datos como la información contenida en la matriz de la propiedad a determinar, para encontrar un conjunto de variables auxiliares llamadas variables latentes **VL** que son utilizadas para la predicción. La siguiente figura ilustra el proceso de calibración en el PLS. [16].

Figura 3. Descripción gráfica del Modelo PLS



Las nuevas variables se pueden representar como un producto de matrices, según,

$$X = TP^T + E = \sum t_a p_a^T + E$$

$$Y = UQ^T + F = \sum u_a q_a^T + F$$

Donde:

T y **U** son las matrices de puntuaciones (score) de **X** y respectivamente.
P y **Q** son las matrices de pesos (loadings) de **X** y respectivamente.
E y **F** son los residuos durante el proceso de calibración.
En el caso de calibrar una sola propiedad (Y es un vector), el algoritmo recibe el nombre de PLS1, mientras que si se calibran simultáneamente varia propiedades, el algoritmo es denominado PLS2. [16][17]

1.7 ESPECTROSCOPIA INFRARROJA

La espectroscopia NIR, es una clase de espectroscopia vibracional que emplea fotones de energía en el rango de 2.65x10-19 a 7.96x10-20 Julios (J) correspondientes al rango de frecuencias entre 780 y 2500 nm. Esta energía es mayor que la necesaria para promover las moléculas a su primer estado vibracional excitado y menor que la requerida para llevar a cabo una excitación electrónica. Solo las frecuencias que puedan suplir la diferencia energética entre dos estados vibracionales en una molécula serán completamente absorbidas, mientras otras frecuencias serán parcialmente o no serán absorbidas. [18]

La espectrometría es un método que permite medir la cantidad de luz que absorbe una sustancia química por medio de la intensidad de luz, como un haz de luz pasa por la solución de la muestra. El equipo, está compuesto por una fuente de luz, un colimador, un monocromador, un selector de longitud de onda, una cubeta de solución de las muestras, un detector fotoeléctrico y un monitor [19]

1.8 REGIONES ESPECTRALES

La región infrarroja del espectro electromagnético se extiende entre la zona del visible y la de las microondas. La región de mayor utilidad práctica de la región IR es la que se extiende entre 7000 y 4000 cm^{-1} denominada región infrarroja media (MIR), donde tienen lugar las vibraciones fundamentales; es una de las técnicas analíticas disponibles más importantes para conseguir información sobre los aspectos cualitativos u cuantitativos en tiempo real. En la región IR lejana se encuentra ubicada entre 700 y 200 cm^{-1}; en esta región se producen absorciones energéticas debidas a cambios rotacionales, las cuales pueden ser utilizadas para el análisis de compuesto órgano-metálicos o inorgánicos (átomos pesados, enlaces débiles) [20].

De otro lado, la región IR cercana entre 12500 y 4000 cm^{-1} presenta absorciones energéticas debidas a sobre tonos y combinaciones de las bandas vibracionales de tensión fundamental que se producen en la región de 3000 a 1700 cm^{-1}; los enlaces implicados son por lo general C–H, N–H y O–H. El IR en la región cercana ha sido utilizado para determinaciones cuantitativas de especies tales como agua, proteínas, hidrocarburos de peso molecular bajo, grasas en productos agrícolas y productos petroquímicos, entre otros. [21].

2. METODOLOGÍA

En este proyecto se utilizó una base de datos de 140 muestras para la regresión del índice de cetano del diésel por medio de propiedades macroscópicas; la base de datos contiene mediciones de propiedades fisicoquímicas de diferentes muestras de diésel tales como: densidad, temperatura al 10% de destilado (T10), temperatura al 50% de destilado (T50), temperatura al 90% de destilado (T90), viscosidad, índice de refracción, contenido de hidrogeno, punto de anilina y peso molecular. Esta base de datos fue tomada del manuscrito publicado por Strative D. y colaboradores [22] y puede ser consultada en el Anexo A.

La base de datos fue complementada utilizando la temperatura de ebullición promedio (T_M) y el flash point (T_f), los cuales fueron calculados por medio de las ecuaciones (1) y (2), respectivamente [23]. Estas propiedades fueron adicionadas ya que influyen en el retardo de la ignición.

$$T_M = \sum_{i=1}^{n} \frac{T_i}{n} \qquad \text{(Ecuación 3)}$$

Donde:

Ti: Temperatura de Ebullición

$$\frac{1}{\left(1.8 * (T_f[K] - 0.33\right)} = -0.014568 + \frac{2.84947}{1.8 * (T_M[K] - 0.33} + 1.093 * 10^{-3} \ln(T_M[K]) - 0.33$$

(Ecuación 4)

2.1 REGRESIÓN LINEAL MÚLTIPLE

Para el ajuste de un modelo lineal se seleccionó el 70% de los datos iniciales para la calibración y el 30% restante para la validación; la selección se efectúo de manera aleatoria. Las actividades seguidas para el ajuste RLM fueron las siguientes (Figura 2.1) [23]:

- Aplicación de estadística univariada y bivariada para describir y analizar la base de datos.
- Autoescalado de datos: Después de centrar cada columna, se divide el resultado por la desviación estándar de la misma de forma que la varianza de cada variable obtenga el valor de la unidad. Con el autoescalado se cambian las unidades originales de cada variable a unidades de desviación estándar. De esta forma todos los ejes tienen la misma longitud y cada variable tiene la misma influencia de cálculo.
- Aplicación del método de búsqueda *stepwise* para proponer el mejor modelo de calibración: Se realizó la mayor cantidad de combinaciones entre las variables independientes para partir de un modelo depurado para la regresión del índice de cetano con base en el coeficiente de correlación.
- Análisis de varianza el modelo inicial de calibración para determinar la significancia estadística del modelo y de sus coeficientes ($\alpha \leq 5\%$).
- Generación de modelos adicionales de calibración omitiendo variables sin significancia estadística ($\alpha > 5\%$).
- Análisis de los residuos, puntos atípicos y puntos influyentes de los modelos de calibración:
  - Análisis de los *Leverage*: permite identificar posibles puntos atípicos que alteren el modelo.
  - Prueba de Shapiro-Wilk: permite comprobar el supuesto de distribución normal de los residuos.

- Distancia de Cook: Mide la influencia de cada caso en el modelo. Una distancia de Cook grande indica que ese caso tiene un peso considerable en la estimación de los coeficientes de regresión (puntos influyentes).

➢ Obtención de estadísticos de desempeño, validación de los modelos y proposición del mejor modelo de regresión:

- Cálculo del coeficiente de determinación (R^2) y la raíz del error cuadrático medio (RMSE). Para calibración R^2 = 1 – RSS/SST, donde RSS corresponde a la suma de los cuadrados de los residuos de calibración y SST representa la varianza total del conjunto. Para validación R^2 corresponde a Q^2 = 1 – PRESS/SST, donde PRESS es la suma de los cuadrados de los residuos de predicción según el conjunto externo (30%) y SST representa la varianza total del conjunto.
- Criterio de Información de Akaike (AIC): la comparación entre los valores de este criterio para los diferentes modelos conduce a la proposición del mejor modelo de regresión; a menor valor de AIC mejor modelo de regresión. El AIC se define según [25].

$$AIC = n.log\left(\frac{RSS}{n}\right) + 2p$$

$$RSS = \sum (y_{ti} - y_{pi})^2 = \sum e^2$$

Donde yti y ypi corresponden al valor de referencia de la variable independiente y al valor predicho de la variable independiente, n al número de muestras y p al número de variables.

2.2 REGRESIÓN MULTIVARIADA POR PCR y PLS

Las regresiones multivariadas PCR y PLS fueron aplicadas a la base de datos de propiedades macroscópicas de muestras de diésel y a una nueva base de datos conformada por 61 espectros NIR de diésel y sus respectivas propiedades. Esta nueva base de datos fue facilitada por la sección de *blending* de la gerencia refinería de Barrancabermeja de ECOPETROL S.A.

Las actividades desarrolladas para la aplicación de las regresiones multivariadas en cada base de datos se resumen en las siguientes [26]:

- Pretratamiento por normalización de los datos mediante los métodos de Área bajo la curva y suavización por el procedimiento de Savitzky-Golay para eliminar ruido y otras interferencias.
- Selección del rango del espectro según reportes de literatura [27].

- Análisis por PCA para generar los componentes principales y determinar posibles puntos atípicos en las base de datos por medio del test estadístico de Hotelling (elipse T^2).
- Calibración de las regresiones PCR y PLS a la base de datos con y sin los posibles puntos atípicos, con lo cual se puede verificar el sesgo de estos puntos en los respectivos modelos.
- Obtención de estadísticos de desempeño, validación de los modelos y proposición del mejor modelo de regresión. Esta etapa sigue los pasos de la última actividad en la metodología de regresión RLM (sin aplicar AIC); para la base de datos de espectros NIR la calibración se realiza con todos los casos y se valida por medio del procedimiento de validación cruzada.

Los modelos considerados en el presente documento fueron ajustados y evaluados utilizando los paquetes *R* y The Unscrambel.X.

3. ANÁLISIS DE RESULTADOS

3.1 RLM DE PROPIEDADES MACROSCÓPICAS

La siguiente tabla resume los valores de las propiedades que conforman la base de datos de propiedades microscópicas (Anexo A). Las propiedades temperatura de ebullición media y punto de flama fueron calculadas según las ecuaciones reportadas en la sección de la metodología. Según los estadísticos univariados, la base de datos presenta un posible punto atípico, el cual corresponde a la muestra con índice de cetano de 13 (columna Mínimo, Tabla 1). Este caso no se retira de la base de datos ya que sólo con el soporte univariado no es posible decidir si esta muestra no es representativa para la regresión en turno. Los demás casos muestran valores típicos para el diésel que puede encontrarse en la destilación del petróleo.

Tabla 1. Estadísticos univariados representativos de la base de datos macroscópica.

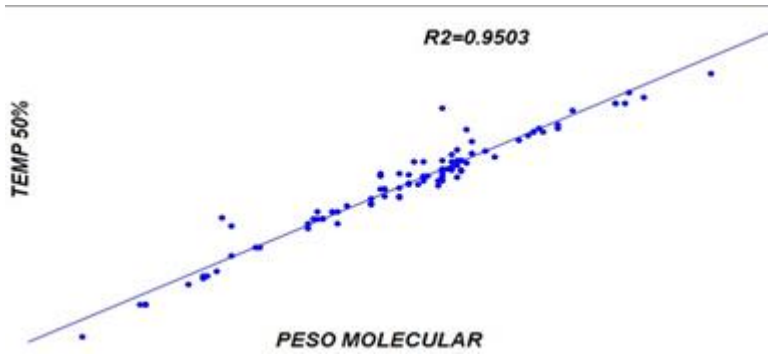
Propiedades	Unidades	Máximo	Promedio	Mínimo	Desviación Estándar
CI ASTM D-4737	-	74,800	48,613	13,000	10,173
Densidad	[kg/m3]	1013,000	848,072	790,200	34,001
T10%	[K]	615,150	505,333	401,150	40,856
T50%	[K]	618,650	541,492	437,150	36,282
T90%	[K]	643,150	584,663	471,150	43,235
Viscosidad	[mm2/s]	1.079	0.348	0.095	0.186
Peso Molecular	[kg/kmol]	259,000	194,564	128,000	25,947
Índice de	[K]	274,736	274,624	274,591	0,022

Propiedades	Unidades	Máximo	Promedio	Mínimo	Desviación Estándar
Refracción					
Punto de Anilina	[K]	368,350	339,605	294,150	11,603
Contenido de Hidrogeno	%	14,750	12,935	8,390	0,958
Temperatura Ebullición media	[K]	802,300	710,901	562,300	57,033
Flash Point	[K]	444,635	433,636	392,522	12,527

Asimismo, las gráficas de dispersión (Figura 4 y 5) muestran que la densidad y el índice de refracción, así como el peso molecular y la T50% presentan una buena relación lineal, con la cual se decide descartar el peso molecular y el índice de refracción del modelo de regresión. La aplicación del método de regresión *stepwise* codificado en el paquete libre *R* [28] reporta que las restantes variables (i.e. densidad, T10, T50, T90, viscosidad, punto de anilina, contenido de hidrógeno y *flash point*) excepto la temperatura de ebullición media (debido a que esta resulta de la combinación lineal de T10, T50 y T90), son significativas para el modelo desde el punto de vista de la explicación de la varianza. Las propiedades fueron previamente autoescaladas para trabajar con variables adimensionales.

Figura 4. Densidad vs Índice de Refracción.

Figura 5.Temp 50% vs Peso Molecular.



El modelo inicial obtenido con la metodología *stepwise* fue analizado para determinar su significancia estadística y la hipótesis de normalidad. Lo anterior debido a que este modelo puede no tener significancia estadística para explicar la variación que presenta el IC; asimismo, no todas las variables pueden presentar significancia estadística para el modelo. Según la tabla ANOVA (Tabla 2, columna p-value), el contenido de H2 no aporta significativamente a la explicación de la varianza obtenida con la regresión; la tabla ANOVA considera el aporte de cada variable cuando se adicionan de forma secuencial a la regresión.

Tabla 2. Análisis de la varianza (ANOVA) del modelo inicial.

	SS	df	MS	F ratio	p-value	Coeficientes
Densidad	0,141	1,000	0,141	10,764	1,47E-03	0,813
T10	0,296	1,000	0,296	22,624	7,46E-06	1,143
T50	0,205	1,000	0,205	15,688	1,49E-04	-2,000
T90	0,495	1,000	0,495	37,848	2,07E-08	0,255
Viscosidad	0,275	1,000	0,275	21,034	1,45E-05	-0,262
Punto de anilina	0,205	1,000	0,205	15,625	1,54E-04	2,718
Contenido H2	0,034	1,000	0,034	2,620	1,09E-01	-1,081
Punto de Flama	0,082	1,000	0,082	6,259	1,42E-02	-0,579

De otro lado, los datos utilizados para calibrar el modelo inicial de RLM fueron analizados para determinar puntos atípicos e influyentes que puedan afectar el

desempeño de las regresiones. Estos puntos son identificados por su alto valor de *Leverage* y de su distancia de Cook; la Tabla 3 presenta los valores de máximos tolerados para los anteriores estadísticos, mientras que la Tabla 4 presenta los puntos detectados que exhiben valores de *Leverage* y distancia de Cook superiores a los máximos (Figuras 6 y 7).

Tabla 3. Características modelo inicial.

	R^2	R^2 Ajustado	*Leverage* Max	Distancia Cook Max	Shapiro-Wilk
Modelo Inicial	0,986	0,985	0,184	0,059	4,21E-02

Tabla 4. Puntos atípicos identificados

	Leverage	Distancia Cook
Modelo Inicial	57, 58	58, 60

Figura 6.Leverage para los datos del Modelo 8 Var.

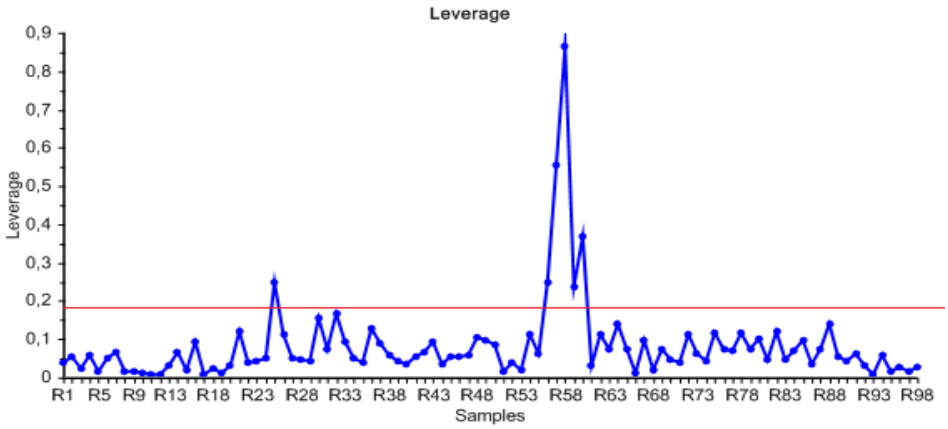
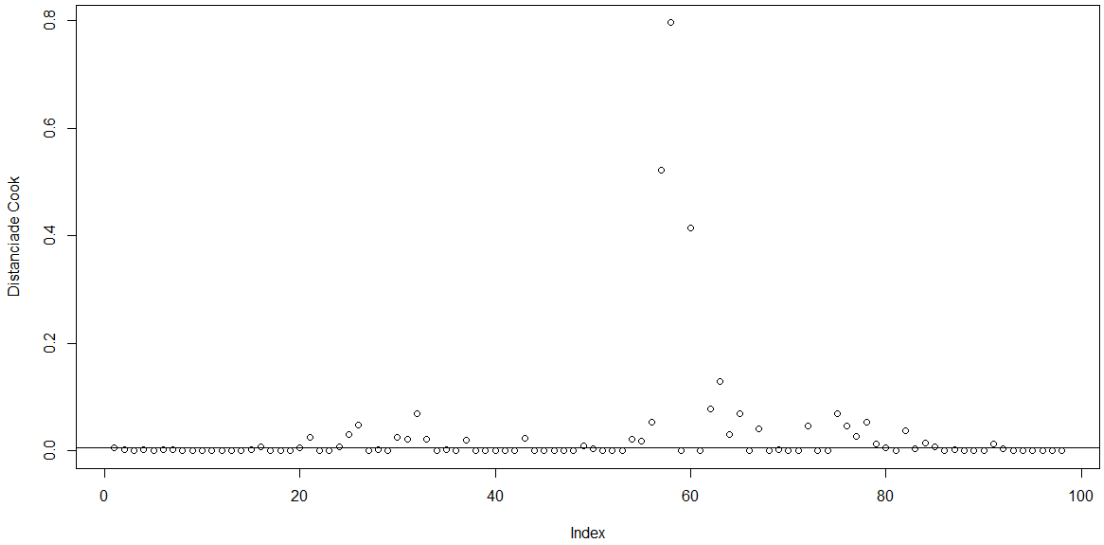


Figura 7. Distancia de Cook para los datos del Modelo 8 Var.



Los puntos de la Tabla 4 pueden sesgar el desempeño del modelo, con lo cual el mismo fue reajustado sin considerar el punto 58 (punto común); lo anterior generó un nuevo modelo (Tabla 5). Al comparar el nuevo modelo (Tabla 5) con el anterior (Tabla 4) es posible observar que el punto influyente no modifica el desempeño ni la tendencia en la normalidad del modelo (columna Shapiro-Wilk); con esto, el punto 58 no fue removido en los análisis posteriores. Asimismo, es posible observar que el nuevo modelo también presenta un valor-p del contenido de hidrogeno mayor al establecido (>0.05) superando la significancia estadística del 5% (Tabla 6). Lo anterior reafirma la necesidad de plantear un nuevo modelo sin incluir el contenido de hidrógeno en la regresión.

Tabla 5.Estadísticos del Modelo sin puntos atípicos e influyentes

	R^2	R^2 Ajustado	Shapiro -

			Wilk
Modelo sin caso 58	0,984	0,983	1,09E-02

Tabla 6. ANOVA Modelo Inicial sin Outliers

	SS	df	MS	F ratio	p value	Coeficientes
Densidad	0,084	1	0,084	6,414	1,31E-02	1,234
T10	0,299	1	0,299	22,833	6,93E-06	1,148
T50	0,213	1	0,213	16,288	1,15E-04	-2,046
T90	0,497	1	0,497	37,943	2,07E-08	0,255
Viscosidad	0,205	1	0,205	15,669	1,52E-04	-0,241
Punto de anilina	0,172	1	0,172	13,163	4,76E-04	2,560
Contenido H2	0,007	1	0,007	0,498	4,82E-01	-0,586
Punto de Flama	0,083	1	0,083	6,366	1,34E-02	-0,584

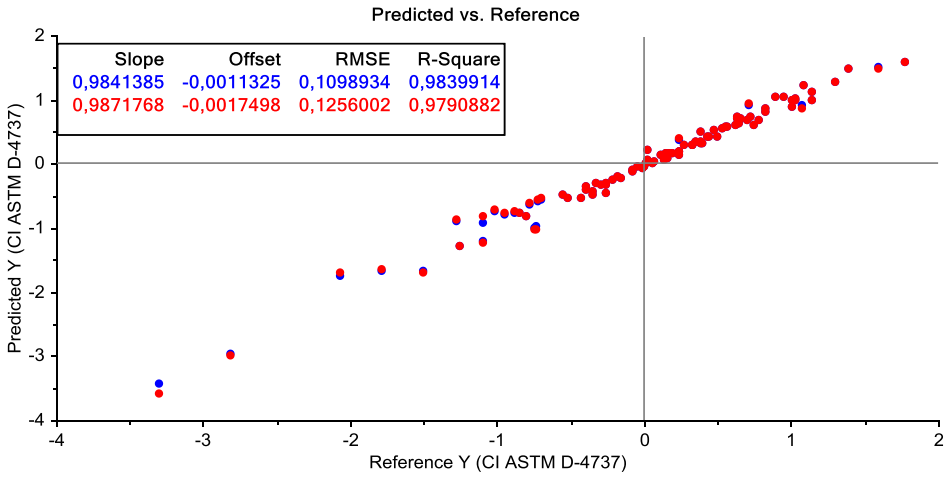
En la Tabla 7 se presenta el análisis de varianza para el modelo de regresión sin considerar el contenido de hidrógeno. En esta tabla es posible verificar la significancia estadística de cada variable en el nuevo modelo de regresión. Este modelo con 7 variables reporta un coeficiente de determinación de 0.984 (0.979 para validación cruzada), y un RMSE de 0.1099 (0.1256 para validación cruzada) (Figura 10). La ecuación de regresión para la estimación del índice de cetano a partir de 7 variables toma la siguiente estructura (Ecuación 5):

$$CI = 1{,}425 * \rho + 1{,}197 * T10 - 1{,}806 * T50 + 0{,}255 * T90 - 2{,}236 * \mu$$
$$+2{,}122 * Punto.Anilina - 0{,}631 * Punto.Flama \quad \text{(Ecuación 5)}$$

Tabla 7. ANOVA Modelo con 7 Variables

	SS	df	MS	F ratio	p-value	Coeficientes
Densidad (ρ)	0,162	1,000	0,162	12,480	6,51E-04	1,425
T10	0,355	1,000	0,355	27,273	1,13E-06	1,197
T50	0,303	1,000	0,303	23,264	5,72E-06	-1,806
T90	0,494	1,000	0,494	37,958	1,99E-08	0,255
Viscosidad (μ)	0,200	1,000	0,200	15,354	1,73E-04	-0,236
Punto de anilina	0,522	1,000	0,522	40,112	9,24E-09	2,122
Punto de Flama	0,106	1,000	0,106	8,122	5,42E-03	-0,631

Figura 8. Modelo 7 Variables MLR



La regresión propuesta en el presente trabajo (Ecuación 5) se acerca con alto grado de exactitud a la predicción del índice de cetano que se realiza a través de

la norma ASTM D-4737 mediante cuatro parámetros (Densidad, T10%, T50% y T90%). Stratiev y colaboradores reportan que el índice de cetano obtenido mediante la norma ASTM D-4737 presenta una explicación del 91.5% de la varianza obtenida con los 140 datos experimentales (Anexo F) [29]; El modelo 7 es de utilidad cuando se tienen los diferentes datos experimentales ya que reduce la complejidad de la ecuación en la norma ASTM D-4737. Asimismo, Stratiev y colaboradores proponen una regresión que alcanza una explicación del 95.6% de la varianza de los datos experimentales (número de cetano) utilizando los cuatro parámetros que requiere la norma ASTM D-4737.

Con el objetivo de proponer una regresión que dependa de una menor cantidad de variables, la Ecuación 5 fue reducida a cuatro variables independientes, las cuales poseen las mayores contribuciones al modelos según el valor absoluto de sus coeficientes; estas variables poseen la mayor contribución al modelo y corresponden a Densidad, T10, T50 y Punto de anilina (Tabla 8); en total estas variables poseen un peso del 85% en el modelo de la ecuación 1. Una nueva regresión RLM fue calibrada considerando el modelo reducido de cuatro variables. La Tabla 9 presenta los coeficientes y la significancia estadística obtenida con el modelo reducido; asimismo, la Figura 9 ilustra los estadísticos y los errores obtenidos para este modelo. La Ecuación 6 muestra la regresión reducida ajustada para el índice de cetano.

La explicación de la varianza del índice de cetano (97.2%) obtenida con el modelo reducido desarrollado en el presente trabajo (Figura 9, Ecuación 2) mejora la predicción del índice de cetano a partir de la norma ASTM D-976 (basada en los 140 datos experimentales descritos en el Anexo 1); la ecuación propuesta en la norma ASTM D-976 alcanza una explicación del 88.8% del índice de cetano calculado a partir de la norma ASTM D-4737 (Figura 36, Anexo F). La ventaja de la norma ASTM D-976 referente al modelo reducido (ecuación 6) es que requiere sólo dos parámetros, la densidad y el T50. Sin embargo, es posible reducir esta

ventaja utilizando ecuaciones propuestas en la literatura para la predicción de T10 y el punto de anilina [30]. Las posibles ecuaciones de predicción de estas propiedades pueden ser analizadas en el Anexo H.

Tabla 8. % Peso de las Variables

Variable	Coeficiente	Valor Abs. Coef	Peso Variable	%Peso Variable	
Punto de anilina	2,122	2,122	0,277	27,664	1
T50	-1,806	1,806	0,235	23,536	2
Densidad	1,425	1,425	0,186	18,574	3
T10	1,197	1,197	0,156	15,608	4
Punto de Flama	-0,631	0,631	0,082	8,222	5
T90	0,255	0,255	0,033	3,318	6
Viscosidad	-0,236	0,236	0,031	3,079	7

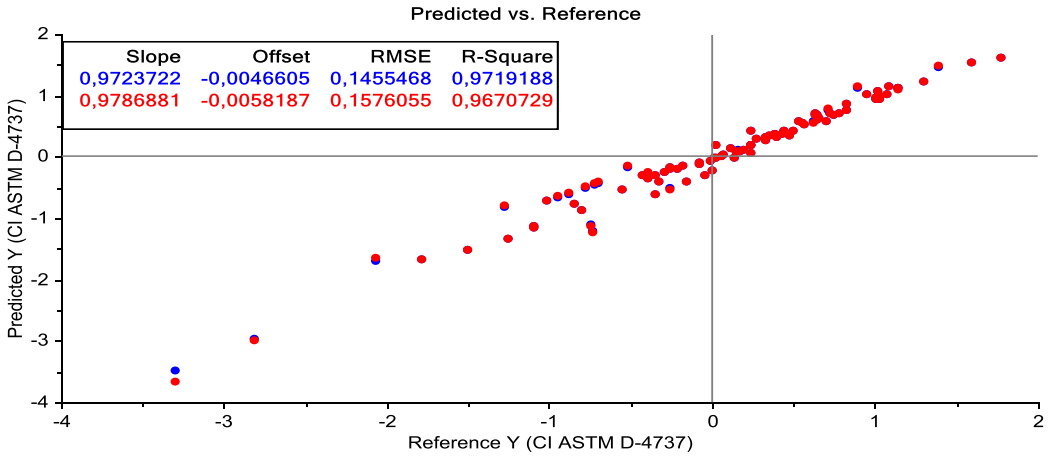
Tabla 9. ANOVA modelo reducido

	SS	df	MS	F ratio	p valué	Coeficientes
Densidad	0,451	1	0,451	20,401	1,84E-05	2,246
T10	2,634	1	2,634	119,195	2,40E-18	0,284
T50	0,579	1	0,579	26,183	1,66E-06	-2,352
Punto de anilina	1,078	1	1,078	48,768	4,23E-10	2,867

Una comparación entre el modelo de 7 variables y el modelo reducido puede ser observada en la Tabla 10. Según los estadísticos consignados en la Tabla 10, el modelo reducido presenta un desempeño cercano al modelo de 7 variables; i.e.

sólo se deja de explicar el 1% de la varianza del índice de cetano con la disminución en el número de variables; el criterio de información de Akaike para el modelo de 7 variables y el reducido corresponde a -137,205 y -88,664 respectivamente. Desde luego, los desempeños de estas regresiones deben ser comprobados en un conjunto mayor al utilizado en el presente trabajo. De otro lado, la Tabla 11 muestra los intervalos de confianza para las variables del modelo reducido.

Figura 9. Modelo RLM reducido a 4 variables



$$CI = 2{,}246 * \rho + 0{,}284 * T10 - 2{,}352 * T50 + 2{,}867 * Punto.Anilina$$

(Ecuación 6)

Tabla 10.Estadísticos de desempeño para los modelos de regresión.

Modelos		R^2	R^2 Ajustado	**RMSE**
Modelo	Calibración	0,984	0,983	0,110

Ecuación 5	Validación cruzada	0,979	0,975	0,126
	Validación Externa	0,993	0,992	0,177
Modelo Reducido	Calibración	0,972	0,971	0,146
	Validación cruzada	0,967	0,964	0,158
	Validación Externa	0,963	0,959	0,218

Tabla 11. Intervalos de Confianza Modelo reducido

	2,50%	97,50%
Densidad	1,271	3,220
T10	0,233	0,335
T50	-3,253	-1,451
Punto Anilina	2,063	3,672

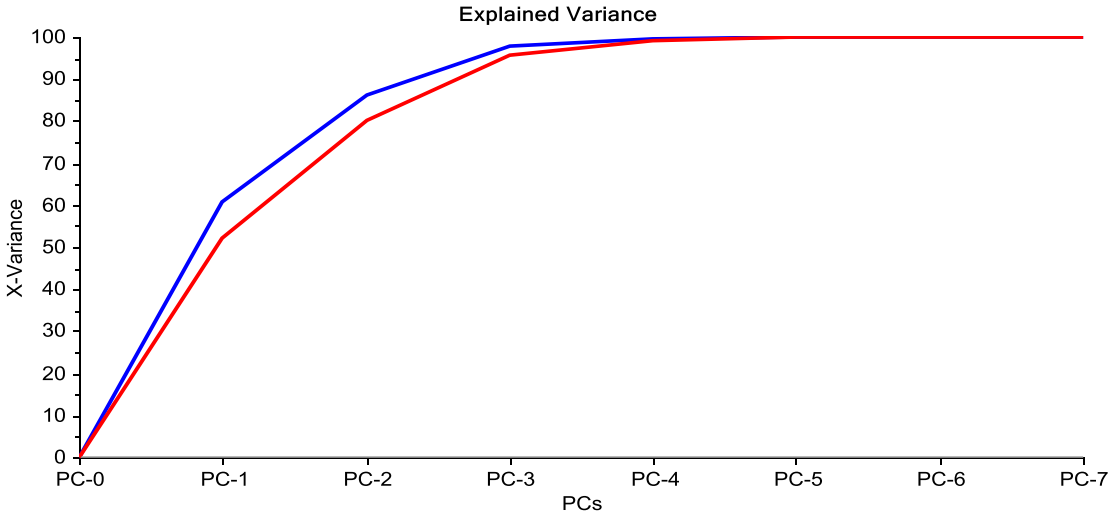
3.2 PLS Y PCA DE PROPIEDADES MACROSCÓPICAS

Para desarrollar estos modelos se utilizó la misma base de datos que para el ajuste por RLM [22].

3.2.1 Análisis por componentes Principales (PCA). Se aplicó PCA a toda la matriz de propiedades con el fin de identificar agrupaciones naturales y el número de componentes principales que determinan la mayor variación de la información contenida en las variables originales. Según la Figura 10, los 3 primeros PCs

explican aproximadamente el 99% de la varianza de la base de datos de propiedades macroscópicas de diésel del petróleo. En la Figura 10, la línea azul corresponde a calibración, mientras que la línea roja representa validación.

Figura 10. Varianza vs PCs para el PCA de propiedades macroscópicas



De otro lado, según la Figura 11 que reporta los *score* para los primeros componentes principales, al parecer las muestras que componen la base de datos de propiedades macroscópicas de diésel pertenecen a un mismo conjunto; i.e. en la gráfica de los *score* no se aprecian agrupaciones que puedan indicar la aplicación de diferentes regresiones en diferentes regiones del rango total de propiedades de diésel; sin embargo, una prueba cuantitativa como la distancia de Mahalanobis es requerida para validar esta conclusión. Asimismo, la Figura 12 muestra que la base de datos puede presentar puntos atípicos que pueden influir negativamente en un modelo de regresión; lo anterior basado en el estadístico de Hotelling.

Figura 11.Scores PCA Propiedades

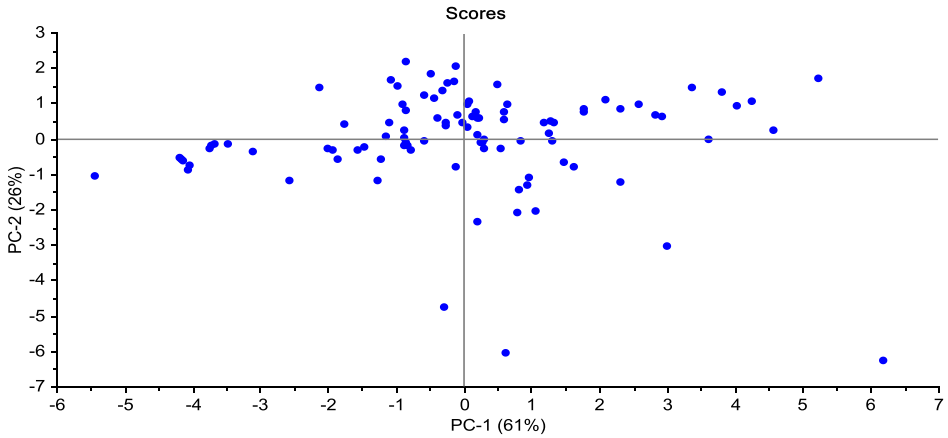
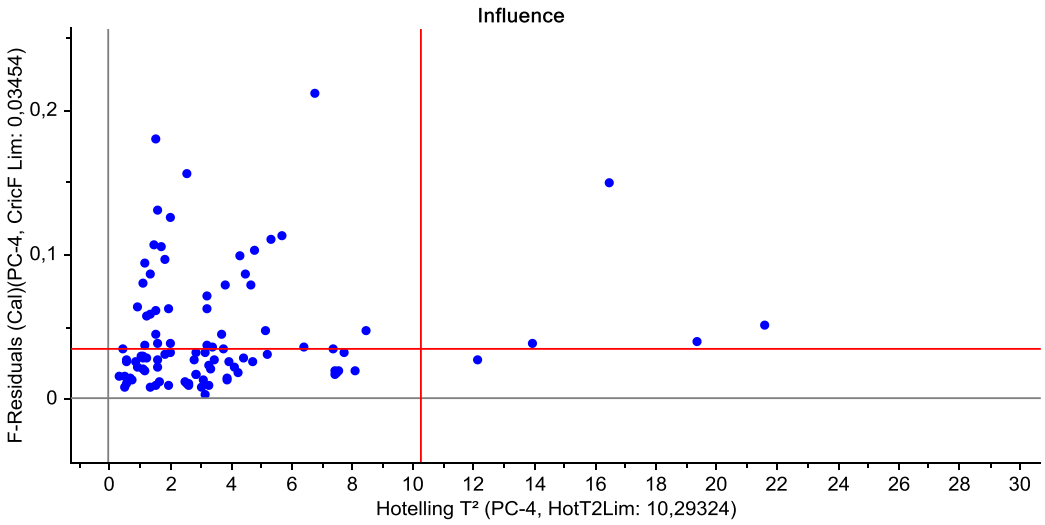


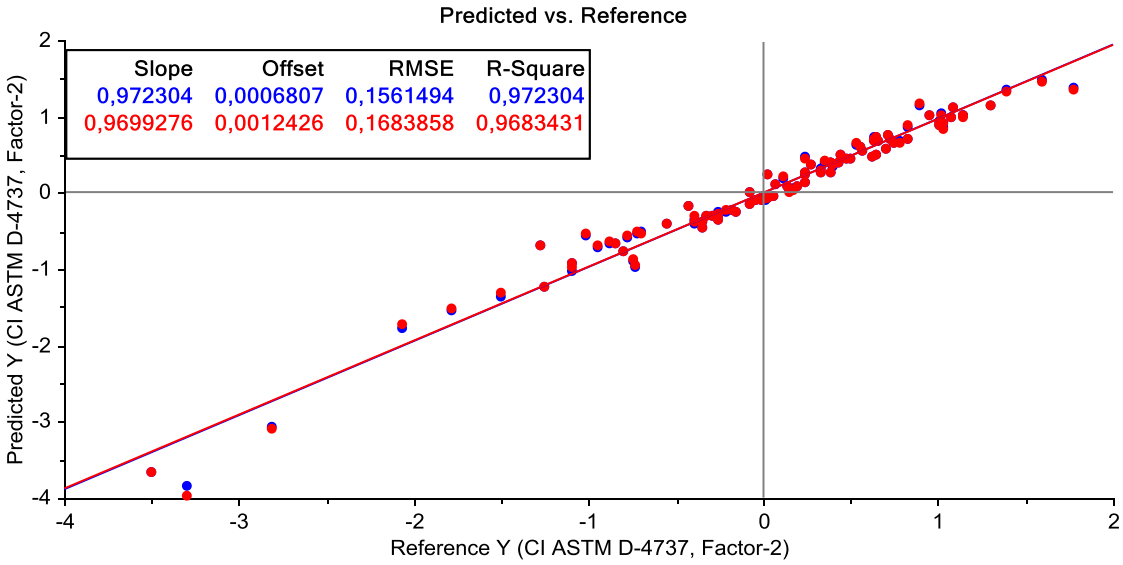
Figura 12. Posibles puntos atípicos (región de influencia).



3.2.2 Desarrollo del Modelo PLS. El modelo PLS fue calibrado sin considerar los puntos atípicos obtenidos por el análisis PCA (Figura 12). Estos puntos fueron eliminados de forma secuencial para verificar su influencia en los respectivos modelos por medio de los estadísticos R^2 y $RMSE$. Según esta tabla, el modelo PLS de regresión con mejor desempeño reportó un $R^2 = 0.9723$ y un $RMSE = 0.1561$ (Figura 13). El modelo referido puede el índice de cetano a partir de propiedades macroscópicas con un error promedio del 16%. Asimismo, el desempeño del modelo calibrado PLS resultó consistente con el desempeño

mostrado en la validación cruzada mostrando una buena capacidad predictiva para los rangos que describen las variables de la base de datos.

Figura 13.Modelo PLS Propiedades



3.3 PCA Y PLS A PARTIR DE ESPECTROS INFRARROJOS

La base de datos de 61 espectros NIR de diésel fue suministrada por la coordinación de *blending* de la refinería de Barrancabermeja. El parámetro de Índice de cetano fue calculado a partir de la ecuación consignada en la norma ASTM D-4737. Los espectros NIR (Figura 26 del Anexo I) fueron sometidos a pretratamiento (normalización con el área bajo la curva y suavización por Savitzky-Golay según se presenta en el Anexo A) con el objetivo de disminuir las fuentes de variabilidad entre los mismos. Después del pretratamiento, un análisis por componentes principales (PCA) a la zona del espectro recomendada para la propiedad índice de cetano [19]; los espectros reportaron aproximadamente 812 datos en la zona definida por:

Tabla 12. Rangos Índice de cetano en el espectro Infrarrojo

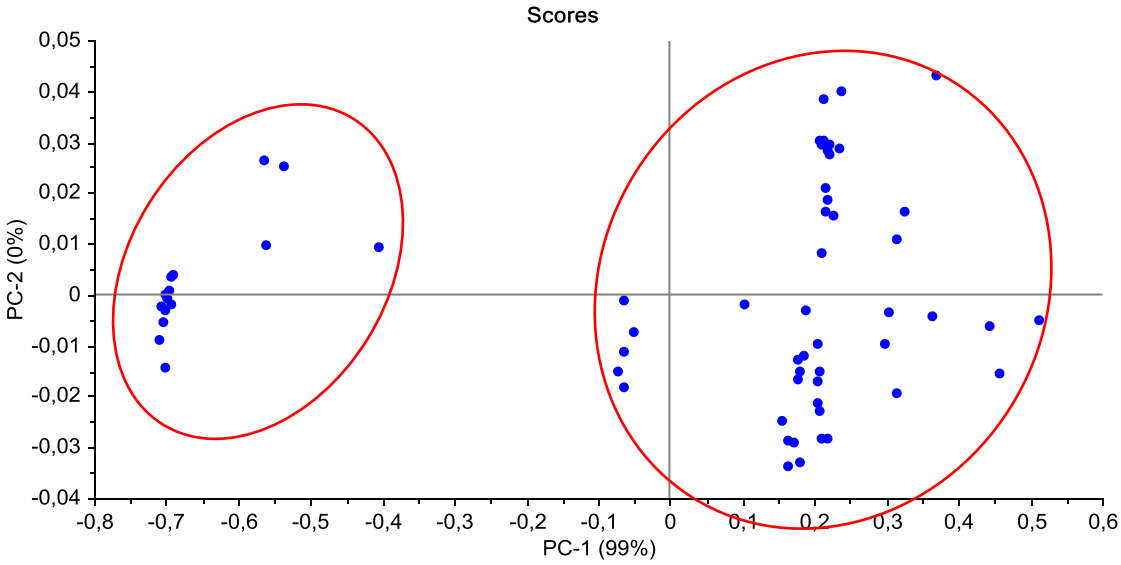
	Rango	cm^{-1}
Índice de Cetano	4440 - 4760	
	5820 - 6220	
	8130 - 8430	

La Tabla 13 presenta la varianza explicada por cada PC resultante del análisis PCA. De esta tabla se puede concluir que con sólo el primer componente se explica la gran mayoría de la varianza del conjunto (95.7%); los otros PCs presentan una contribución no significativa a la explicación de la varianza de los datos. Asimismo, el gráfico de los *score* (Figura 14) sugiere que los espectros pueden corresponder a muestras estructuralmente diferentes. La Figura 14 muestra espectros con valores de score opuestos formando, desde una vista cualitativa, dos conjuntos señalados por círculos de color rojo; desde luego, una comprobación cuantitativa es sugerida para validar los mencionados resultados. Estas agrupaciones pueden deberse a que las muestras fueron recolectadas en diferentes tiempos, para los cuales las cargas de crudos a la refinería presentaron amplias variaciones en sus propiedades.

Tabla 13.Contribución de los PCs para la base de espectros NIR

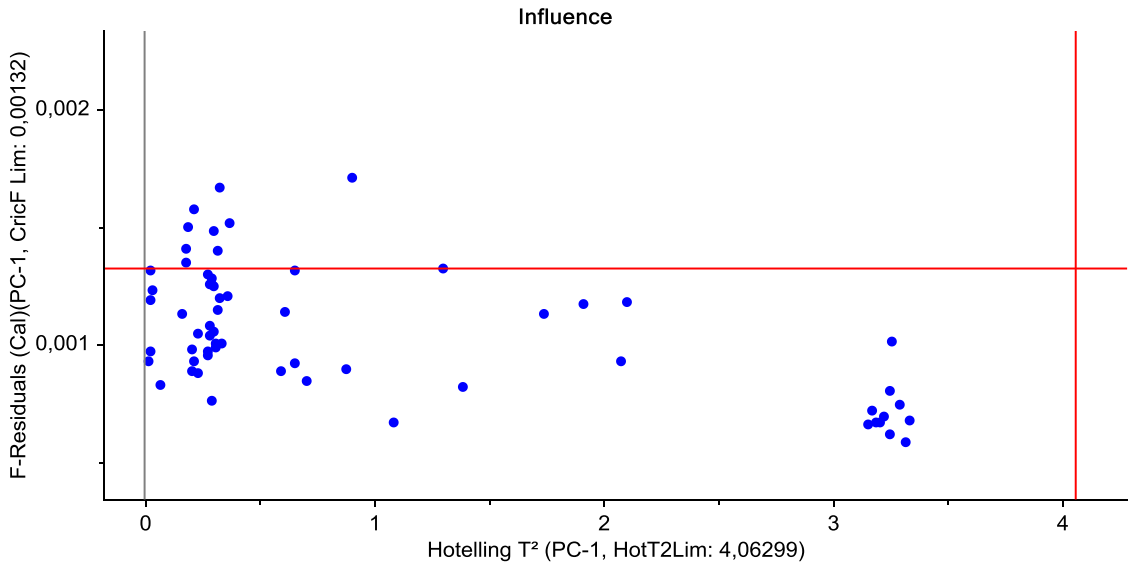
Componente Principal	Varianza Explicada (%)	Varianza Acumulada (%)
PC 1	95.7	95.7
PC 2	2.66	98.36
PC 3	1.04	99.4

Figura 14.Score Espectro Infrarrojos



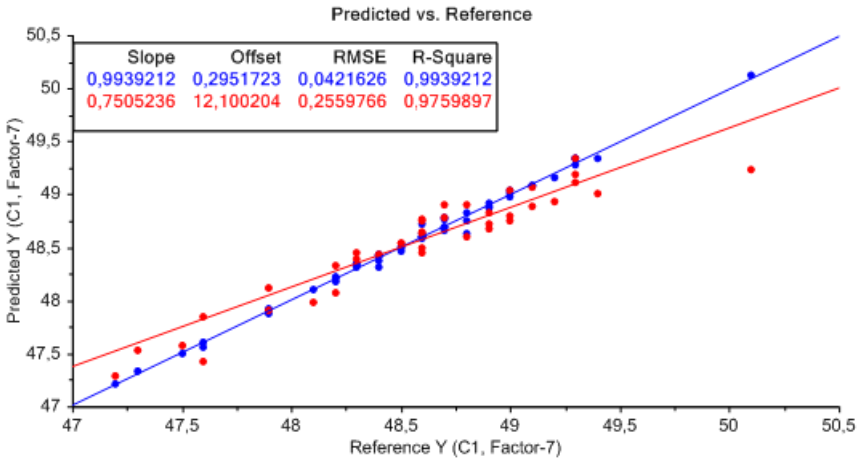
Asimismo, la Figura 15 presenta los resultados del estadístico de Hotelling sobre las muestras de espectros NIR. Según la región del estadístico T2, ninguna muestra aparece como atípica para el conjunto de espectros NIR (límite horizontal derecho); esto puede considerarse como una consecuencia de la agrupación propuesta en la Figura 14.

Figura 15. Gráfico de influencia del PCA de espectros NIR.



Aunque la Figura 14 sugiere agrupar las muestras según su valor de score en el primer componente principal, el modelo PLS fue aplicado al conjunto completo (61 espectros) debido a la reducción en el número de casos para la calibración. El modelo PLS calibrado reportó los siguientes estadísticos: $R^2 = 0.994$ y RMSE = 0.042 (Figura 15). A pesar del buen desempeño del modelo calibrado, los estadísticos $R^2 = 0.976$ y RMSE = 0.26 obtenidos con el procedimiento de validación interna o cruzada, reflejan la no homogeneidad de la base de datos de espectros NIR; lo anterior sugiere como válida la agrupación presentada en la Figura 14. Aunque la calibración del modelo PLS muestra un error bajo en la reproducción del índice de cetano (4.2% en promedio), la validación cruzada reporta una desmejora en la regresión cuando se utiliza en predicciones (error promedio del 26%, Figura 15). Lo anterior sugiere ampliar la base de datos de espectros NIR para desarrollar un modelo PLS validado para la predicción del índice de cetano.

Figura 16. Ajuste PLS basada en NIR. Calibración en azul, validación cruzada en rojo.



La Tabla 14 presenta los modelos desarrollados en el presente documento con sus respectivos estadísticos de desempeño. De acuerdo a esta tabla, el modelo RLM de 7 variables basado en propiedades macroscópicas reporta un mejor desempeño predictivo para el índice de cetano; en segundo lugar se encuentra el modelo PLS basado en propiedades macroscópicas y en tercero puesto el modelo PLS basado en espectros NIR. Aunque los modelos PLS resultaron con un desempeño por debajo de los modelos RLM, se recomienda verificar con pretratamientos adicionales y con un mayor número de muestras respectivas para validar los resultados obtenidos. Asimismo, la aplicación de los resultados mostrados en este trabajo es recomendada como una aproximación preliminar al índice de cetano considerando los límites de las variables que conforman las bases de datos respectivas.

Tabla 14. Estadísticos de Desempeño de los Modelos

	Modelo	Modelo	PLS Variables	PLS Espectros

	Final	Reducido	Macroscópicas	Infrarrojos
$R^2 Calibracion$	0,9840	0,9719	0,9723	0,9939
$R^2 Validacion$	0,9790	0,9670	0,9683	0,9760
RMSE Calibración	0,1098	0,1455	0,1561	0,0422
RMSE Validación	0,1256	0,1576	0,1684	0,2560

4. CONCLUSIONES

- Los modelos desarrollados en el presente proyecto de grado basados en propiedades macroscópicas y en espectros NIR muestran un desempeño importante (R2 superior al 95%) en la cuantificación del índice de cetano del diésel.

- Según los resultados obtenidos, la regresión RLM basada en propiedades macroscópicas mostró un mejor desempeño que los modelos PLS. Sin embargo, se debe ampliar las bases de datos y los pretratamientos para verificar los desempeños de los modelos.

- Con base en las 140 muestras que conforman la base de datos de propiedades macroscópicas, la regresión RLM permitió la generación de una ecuación dependiente de dos parámetros con desempeño superior a la reportada por la norma ASTM D-976.

5. RECOMENDACIONES

- Los modelos desarrollados en este proyecto deben ser validados en un conjunto mayor de muestras para determinar su aplicación a nivel predictivo.

- Utilizar las bases de datos descritas en el presente proyecto para continuar con el desarrollo de modelos predictivos de otras propiedades del diésel.

- Ampliar el análisis PCA para verificar agrupaciones y tendencias mostradas por las muestras de las bases de datos.

REFERENCIAS BIBLIOGRAFICA

1. SASTRY, S; CHOPRA, A; SARPAL, S; JAIN, K; SRIVASTAVA, P and BHATNAGAR, K. Determination of physicochemical properties and carbóntype analysis of base oils using Mid-IR spectroscopy and partial least squares regression analysis. Energy & Fuels 1998. Vol 12. 304-11 p.

2. http://www.ref.pemex.com/octanaje/24DIESEL.htm (en Línea)

3. SALVATORE J. Significance of Tests for Petroleum Products. Seventh Edition. ASTM Manual Series Pag. 70.

4. Riazi, 2005. Characterization and properties of petroleum fractions

5. Ayala M. et al., 1998. Biocatalytic oxidation of fuel as an alternative to bio desulfurization. Fuel Processing Technology 57. 101-111.

6. ASTM D-976. (s.f.). Standard Terminology Relating to Petroleum, Petroleum Products, and Lubricants.

7. Totten G, Fuel and lubricants handbook: Tenchnology, properties, performance and testing. ASTM Manual Series: MNL37WCD.

8. CASTELLANOS C, ORTIZ Y. Desulfuración de Diésel por oxidación y extracción con solvente. Bucaramanga 2004. Tesis de Grado (ingeniería Química) Universidad Industrial de Santander. Facultad de Ingenierías Fisicoquímicas. Escuela Ingeniería Química

9. Xiao-dong, G., et al, 2002. SSHT process: A low cost solution for low sulfur and low aromatic diesel. Paper presented at the 17th World Petroleum Congress, Rio de Janeiro, and September 1–5

10. Smilde, A., Bro, R., Geladi, P. (2004). Multi-way Analysis with Applications in the Chemical Sciences. John Wiley & Sons, Ltd. England.

11. Brereton, R.G. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. 1st edition. USA: Wiley, 2002.

12. AHAMD AL-GHOUTI, M., SALIM AL-DEGS, Y., and AMER, M. Application of chemometrics and FTIR for determination of viscosity index and base number of motor oils. En: Talanta 81 (2010) 1096–110.

13. GEMPERLINE P. J. Introduction to Chemometrics.Practical Guide to Chemometric, Second Edition. 2006

14. Mendenhall, W., Beaver, R., & Beaver, B. (2006). Introducción a la probabilidad y la estadística.

15. MILLER, J.N, MILLER, J.C, Estadística y Quimiometria para la química analítica. Cuarta edición. España: Prentice Hall, 2002. P.595

16. MACHO-APARICIO, Santiago. Metodología analítica basada en espectroscopia de infrarrojo y calibración multivariante. Aplicación industria petroquímica. Tarragona. Tesis doctoral (Doctor en Química) Univertar Rovira i Virgili. Área de Química analítica. Departament de Química Analítica i Química orgánica. 2002.

17. Quimiometria aplicada al control. Lopez, pedro. Nuevo Leon, Mexico: ciencia UNAL, 1996, Vol. 6.

18. CIURCZAK, E. W. Handbook of near-infrared analysis. New York. Taylor & Francis Group, 2001.

19. SANTOS VO, OLIVEIRA FCC, LIMA DG, PETRY AC, GARCIA E, SUAREZ P a. Z, 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.

20. VARGAS, G. Predicción del análisis SARA de fondos de vacío colombianos utilizando foto acústica en el infrarrojo medio y métodos quimiométrico. Bucaramanga 2011. Tesis de grado (Química). Universidad Industrial de Santander. Facultad de Ciencias. Escuela de Química.

21. APARICIO M, S. Metodologías analíticas basadas en espectroscopia de infrarrojo y calibración multivariante, aplicación a la industria petroquímica. Tarragona. Tesis Doctoral, Universidad y Virgili. Departamento de Química analítica y química orgánica. 2002. 20-29p.

22. Stratiev, D., Marino, I., Dinkov, R., Shishkova, I., Velkov, I., Sharafutdinov, I., Rudnev‖, N. (2015). Opportunity to Improve Diesel-Fuel Cetane-Number Prediction from Easily Available Physical Properties and Application of the Least Squares Method and Artificial Neural Networks.Energy Fuels 29, 1520−1533.

23. Xinshuai L. Zhenyhi L. Research Progress on Flash Point Prediction. Journal of Chemical Engineering Data, Vol. 55, No. 9, 2010, 2943-2950

24. Hair, J.f.; Black, W.C.; Babin, B.J.; Anderson, R.E. 2010. Multivariate Data Analysis. Prentice Hall, 7th edition, USA. Draper, N.R. and Smith, H. 1998. Applied Regression Analysis. John Wiley & Sons. Third edition. USA.

25. Garza J. Morales B. Gonzales B. Análisis estadístico Multivariante: Un enfoque teorico y practico. MACGRAW-HILL, 2013

26. Brereton, R.G. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. 1st edition. USA: Wiley, 2002; Hair, J.f.; Black, W.C.; Babin, B.J.; Anderson, R.E. 2010. Multivariate Data Analysis. Prentice Hall, 7th edition, USA.

27. Vianney O. Santos Jr., Flavia C.C. Oliveira, Daniella G. Lima, Andrea C. Petry,Edgardo Garcia, Paulo A.Z. Suarez, Joel C. Rubim. A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis.

28. https://www.r-project.org/ (En Linea)

29. ASTM D-4737. (s.f.). Standard Terminology Relating to Petroleum, Petroleum Products, and Lubricants.

30. Riazi, 2005. Characterization and properties of petroleum fractions. Professor of Chemical Engineering, Kuwait University. Equation 2.51

31. Riazi, 2005. Characterization and properties of petroleum fractions. Professor of Chemical Engineering, Kuwait University. Equation 2.111

BIBLIOGRAFÍA

APARICIO R., C. (2013). *Predicción del número de Bromo de naftas colombianas utilizando espectroscopia infrarrojo en la región cercana y metodos quimiométricos.* Bucarmanga: Universidad Industrial de Santander.

ASTM D4175-15a. (s.f.). *Standard Terminology Relating to Petroleum, Petroleum Products, and Lubricants.*

BALLESTEROS, S. Prediccion de las fracciones SARA de fondos de vacio de crudos colombianos, por Metodos Quimiometricos utilizando Espectroscopia de fluorecencia inducida por laser (LIF). Bucaramanga Tesis de grado (Quimica) Universidad Industrial de Santander. Facultad de ciencias. Escuela de Quimica, 2010.

BARBIERI GONZAGA, F., & Pasquini, C. (2010). A low cost short wave near infrared spectrophotometer: Application for determination of quality parameters of diesel fuel. *Analytica Chimica*, 92-97.

FALLA, F., LARINI, C., LE ROUX, G., QUINA, F., MORO, L., & Nascimento, C. (2006). Characterization of crude petroleum by NIR. *Journal of Petroleum Science and Engineering*, 127-137.

FENG, F., WU, Q., & ZENG, L. (2015). Rapid analysis of diesel fuel properties by near infrared reflectance spectra. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 149*, 271-278.

L

ARAMIE: 2006. Chemical Industries Series. 916 p.

MELÉNDEZ, L. V., LACHE, A., ORREGO-RUIZ, J. A., PACHÓN, Z., & MEJÍA-OSPINO, E. (2012). Prediction of the SARA analysis of Colombian crude oils using ATR–FTIR spectroscopy and chemometric methods. *Journal of Petroleum Science and Engineering*, 56-60.

MENDENHALL, W., BEAVER, R., & BEAVER, B. (2006). *Introduccion a la probabilidad y la estadistica.*

MENDOZA, L. *, predicción de propiedades fisicoquímicas de productos destilados del delayed coking de fondos de vacío a partir de parámetros estructurales determinados por espectroscopia infrarroja FTIR-ATR y métodos quimiométricos. Universidad industrial de Santander.* bucaramanga : s.n., 2014.

MONROY L. Predicción de la densidad en Asfáltenos Colombianos utilizando espectroscopia fotoacústica en la región del infrarrojo medio y métodos quimiométricos. Bucaramanga 2010. Tesis de grado (químico). Universidad Industrial de Santander. Facultad de ciencias. Escuela de Química.

PINILLA TORRES, A. M., & ARDILA ANTOLINES, J. V. (2014). *Optimización de un proceso limpio para la desulfurización de una muestra de diésel colombiano enriquecida utilizando el líquido iónico tetrafluoroborato de 1-butil-3-metilimidazolio.* Bucaramanga: Tesis de Química, UNIVERSIDAD INDUSTRIAL DE SANTANDER.

QINQIN, C., et al. nvironmental risk source management system for the petrochemical industry. J Process Safety and Environmental Protection. Acta [online] 2014. Vol. 92, p. 251–260. Available from: ELSEVIER-SCIENCE DIRECT. [Biblioteca Universidad Industrial de Santander].

SANTOS JR, V., OLIVEIRA, F., LIMA, D., PETRY, A., GARCIA, E., SUAREZ, P., & RUBIM, J. (2005). A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. *Analytica Chimica*, 188-196.

SPEIGHT, J. G. The chemistry and technology of petroleum. Cuarta edición:

STRATIEV, D., MARINOV, I., DINKOV, R., SHISHKOVA, I., VELKOV, I., SHARAFUTDINOV, I., RUDNEV‖, N. (2015). Opportunity to Improve Diesel-Fuel Cetane-Number Prediction from Easily Available Physical Properties and Application of the Least-Squares Method and Artificial Neural Networks. Energy Fuels 29, 1520−1533

THE UNSCRAMBLER X METHODS. Sofware para el diseño de experimentos y analisis multivariado (En Linea) http://www.camo.com [citado Mayo 2015]

VARGAS, Gloria. Predicción del análisis SARA de fondos de vacio colombianos utilizando espectroscopia fotoacústica en la región del infrarrojo medio y métodos quimiométricos. Bucaramanga 2011. Tesis de grado (Química). Universidad industrial de Santander. Facultad de ciencias. Escuela de Química.

XINSHUAI, L., and ZHENYI, L.. Research progress on flash point prediction. *Journal of Chemical and Engineering*. Acta [online] 2010. Vol. 55 N°9, p. 2943–2950. Available from: ELSEVIER-SCIENCE DIRECT. [Biblioteca Universidad Industrial de Santander].

ANEXOS

Anexo A. Datos de las propiedades Macroscópicas

No	CI ASTM D-4737	Densidad [Kg/m3]	10%[K]	50%[K]	90%[K]	Viscosidad[mm2/s]
1	45	840,9	473,15	518,15	584,15	0,232
2	51	820,9	469,15	512,15	588,15	0,203
3	51	842,1	482,15	548,15	610,15	0,322
4	47	833,2	466,15	515,15	597,15	0,219
5	50	836,1	475,15	532,15	605,15	0,264
6	43	877,3	525,15	563,15	609,15	0,441
7	45,6	801,1	452,15	473,15	502,15	0,132
8	54,4	844,1	516,15	555,15	608,15	0,352
9	56	841,6	523,15	557,15	597,15	0,356
10	52,5	843,7	501,15	552,15	606,15	0,339
11	53,7	837,9	502,15	545,15	596,15	0,307
12	52	840,8	507,15	543,15	589,15	0,308
13	57	855,9	545,15	581,15	620,15	0,5
14	48,6	834,1	469,15	523,15	610,15	0,239
15	55	855,4	541,15	573,15	614,15	0,456
16	45	804,2	447,15	477,15	503,15	0,129
17	51	844,8	504,15	547,15	593,15	0,306
18	47,8	859,4	508,15	558,15	611,15	0,388
19	49,3	849,2	503,15	548,15	597,15	0,332
20	55	861,9	548,15	583,15	623,15	0,513

No	CI ASTM D-4737	Densidad [Kg/m3]	10%[K]	50%[K]	90%[K]	Viscosidad[mm2/s]
21	39,7	800,3	455,15	458,65	471,15	0,141
22	47,8	836,5	515,15	518,65	531,15	0,258
23	50,6	846,4	535,15	538,65	551,15	0,315
24	56,6	863,3	575,15	578,65	591,15	0,65
25	64,8	872,8	615,15	618,65	631,15	1,001
26	39	801,8	455,15	458,65	471,15	0,121
27	44,6	826,5	495,15	498,65	511,15	0,199
28	48,5	834,8	515,15	518,65	531,15	0,253
29	50,3	847,2	535,15	538,65	551,15	0,308
30	52,5	856,6	555,15	558,65	565,15	0,353
31	56,2	864	575,15	578,65	591,15	0,689
32	59,5	870,8	595,15	598,65	611,15	0,895
33	41,3	796,8	455,15	458,65	471,15	0,117
34	46,4	822,4	495,15	498,65	511,15	0,195
35	50,2	831,2	515,15	518,65	531,15	0,248
36	61,8	865,9	595,15	598,65	611,15	0,743
37	41,5	796,4	455,15	458,65	471,15	0,112
38	44,6	808,6	475,15	478,65	491,15	0,139
39	46,8	821,5	495,15	498,65	511,15	0,176
40	48,7	834,3	515,15	518,65	531,15	0,227
41	48,9	850,2	535,15	538,65	551,15	0,293
42	55,2	850,9	555,15	558,65	565,15	0,405
43	40,7	798,2	455,15	458,65	471,15	0,117
44	49,2	833,3	515,15	518,65	531,15	0,251

No	CI ASTM D-4737	Densidad [Kg/m3]	10%[K]	50%[K]	90%[K]	Viscosidad[mm2/s]
45	50,1	847,7	535,15	538,65	551,15	0,29
46	54,9	851,6	555,15	558,65	565,15	0,416
47	59,1	857,9	575,15	578,65	591,15	0,578
48	58,8	872,3	595,15	598,65	611,15	0,793
49	46	803,6	448,15	479,15	505,15	0,131
50	46	805,6	453,15	482,15	507,15	0,137
51	53	847	518,15	553,15	602,15	0,348
52	52	839	513,15	540,15	564,15	0,267
53	54	857,7	534,15	576,15	618,15	0,476
54	51	890,3	587,15	606,15	633,15	0,848
55	48,8	885,9	570,15	593,15	624,15	0,68
56	20	911,3	481,15	513,15	538,15	0,245
57	15	942,7	493,15	519,15	546,15	0,255
58	13	1013	573,15	595,15	621,15	0,932
59	43,3	871,5	483,15	581,15	632,15	0,542
60	37,5	790,2	401,15	437,15	519,15	0,095
61	49,8	832,9	469,15	534,35	601,15	0,246
62	27,6	930,5	531,25	580,65	615,25	0,515
63	35,6	881,1	493,85	566,05	611,35	0,38
64	30,4	891,2	469,85	549,45	605,25	0,292
65	38,3	863,9	476,55	549,05	604,65	0,298
66	52,5	843,8	503,15	552,15	602,15	0,34
67	66,7	817	526,15	547,15	581,15	0,287
68	51,4	855,7	518,15	565,15	619,15	0,415

No	CI ASTM D-4737	Densidad [Kg/m3]	10%[K]	50%[K]	90%[K]	Viscosidad[mm2/s]
69	40	891,7	530,15	572,15	618,15	0,523
70	45,3	872,2	530,15	559,15	603,15	0,412
71	55,7	823	492,15	528,15	595,15	0,241
72	57,7	864,7	561,65	602,15	642,15	0,687
73	35,9	888,4	512,15	548,15	588,15	0,377
74	40,4	876,4	514,65	549,65	590,15	0,371
75	41,1	833,1	439,15	493,15	576,15	0,175
76	41	853,2	451,15	523,15	601,15	0,254
77	46	838	450,15	523,15	601,15	0,242
78	33,3	893,1	474,15	558,15	613,15	0,435
79	37,5	882,5	484,15	558,15	615,15	0,421
80	44,2	864,1	482,15	564,15	614,15	0,425
81	55,2	829	473,15	547,15	629,15	0,302
82	55,8	806	452,15	515,15	618,15	0,198
83	58,3	827	490,15	552,15	633,15	0,316
84	59,6	820	487,15	545,15	632,15	0,285
85	62,7	811	484,15	541,15	624,15	0,262
86	55,3	829	486,15	542,15	624,15	0,286
87	53,4	827	483,15	529,15	616,15	0,247
88	48,1	841	483,15	527,15	619,15	0,255
89	60,2	814	486,15	533,15	601,15	0,245
90	59,1	817	486,15	534,15	603,15	0,25
91	60,2	826	501,15	551,15	596,15	0,312
92	58,8	826	500,15	545,15	598,15	0,308

No	CI ASTM D-4737	Densidad [Kg/m3]	10%[K]	50%[K]	90%[K]	Viscosidad[mm2/s]
93	52,6	837	492,15	542,15	599,15	0,282
94	59	834	508,15	561,15	618,15	0,385
95	52,2	844	497,15	553,15	612,15	0,311
96	54,2	838	493,15	552,15	623,15	0,326
97	53,1	842	496,15	554,15	617,15	0,335
98	57	834,9	513,35	549,45	597,75	0,284
99	37,8	834,1	458,15	487,15	527,15	0,166
100	46	804,2	455,15	478,15	504,15	0,132
101	43,8	828,3	495,15	498,65	511,15	0,223
102	53,1	855,3	555,15	558,65	565,15	0,333
103	64,8	872,8	615,15	618,65	631,15	1,079
104	43	812	475,15	478,65	491,15	0,16
105	51,5	844,4	535,15	538,65	551,15	0,307
106	54,1	853,2	555,15	558,65	565,15	0,352
107	58,1	860,1	575,15	578,65	591,15	0,598
108	59,6	857	575,15	578,65	591,15	0,554
109	57,7	874,5	595,15	598,65	611,15	0,798
110	46,8	821,5	495,15	498,65	511,15	0,189
111	46	809,2	445,15	483,15	525,15	0,137
112	58	861,7	562,15	590,15	625,15	0,606
113	54	859,4	533,15	584,15	643,15	0,491
114	19	966,7	501,15	541,15	601,15	0,338
115	39	869,7	479,65	559,55	609,55	0,334
116	36	865	468,05	545,45	602,65	0,28

No	CI ASTM D-4737	Densidad [Kg/m3]	10%[K]	50%[K]	90%[K]	Viscosidad[mm2/s]
117	34,1	877,3	480,75	554,65	606,15	0,31
118	41,7	854	469,85	542,45	602,45	0,274
119	31	900,2	494,45	570,45	612,25	0,47
120	45,6	872,7	521,15	570,15	619,15	0,474
121	34,5	913,7	539,15	576,15	622,15	0,595
122	24,4	947,1	553,15	583,15	624,15	0,724
123	58,9	839,2	516,65	568,65	622,15	0,402
124	57,6	857	543,15	593,65	639,65	0,592
125	31,9	899,6	509,15	546,15	585,65	0,376
126	45,6	864	518,15	552,65	591,15	0,369
127	36,9	848,6	438,15	494,15	577,15	0,184
128	45,9	818,1	439,15	493,15	575,15	0,167
129	36,6	868,3	449,15	524,15	601,15	0,267
130	45,4	852,2	494,15	536,15	585,15	0,293
131	55,6	836	494,15	556,15	638,15	0,344
132	57,7	821	485,15	538,15	622,15	0,265
133	57,9	814	481,15	525,15	597,15	0,226
134	54,4	818	479,15	518,15	592,15	0,214
135	58,9	826	501,15	545,15	597,15	0,309
136	60,9	826	502,15	554,15	599,15	0,29
137	58,5	826	501,15	543,15	599,15	0,288
138	74,8	807	507,15	569,15	619,15	0,39
139	52,5	814	486,15	504,15	542,15	0,193
140	53,2	839	487,15	552,15	615,15	0,257

No	Peso Molecular[kg/kmol]	Índice de Refracción[K]	Punto de anilina[K]	Contenido H2[%m]	T ebull prom[K]	Flash Point[K]
1	176	274,621	333,850	12,910	695,550	433,812
2	175	274,607	340,450	13,580	695,750	433,848
3	200	274,620	344,350	13,210	734,850	439,698
4	175	274,616	336,150	13,150	705,050	435,463
5	188	274,617	341,150	13,250	721,650	437,994
6	205	274,643	334,750	12,140	745,750	440,910
7	150	274,596	334,750	13,830	597,150	406,864
8	206	274,621	346,050	13,220	739,950	440,287
9	208	274,619	347,850	13,330	731,750	439,322
10	203	274,621	345,150	13,200	735,150	439,734
11	199	274,617	345,050	13,330	722,750	438,146
12	196	274,619	343,050	13,200	715,950	437,174
13	227	274,628	350,250	13,060	766,650	442,766
14	181	274,616	338,650	13,220	721,050	437,910
15	219	274,628	347,650	13,000	756,850	441,970
16	153	274,598	334,850	13,760	599,550	407,754
17	199	274,622	342,750	13,100	721,250	437,938
18	205	274,631	340,450	12,700	743,350	440,658
19	199	274,625	341,250	12,960	725,250	438,485
20	227	274,631	348,450	12,870	770,650	443,054
21	141	274,597	329,550	13,640	562,300	392,522
22	177	274,618	335,950	13,070	652,300	424,284
23	191	274,623	338,950	12,950	682,300	431,251

No	Peso Molecular[kg/kmol]	Índice de Refracción[K]	Punto de anilina[K]	Contenido H2[%m]	T ebull prom[K]	Flash Point[K]
24	222	274,633	346,150	12,780	742,300	440,545
25	259	274,637	356,350	12,830	802,300	444,635
26	140	274,598	328,950	13,580	562,300	392,522
27	164	274,612	332,950	13,190	622,300	415,583
28	178	274,617	336,650	13,140	652,300	424,284
29	191	274,624	338,650	12,920	682,300	431,251
30	206	274,629	341,850	12,810	706,900	435,767
31	222	274,633	345,850	12,760	742,300	440,545
32	239	274,637	350,150	12,710	772,300	443,167
33	141	274,594	331,150	13,780	562,300	392,522
34	165	274,609	334,750	13,350	622,300	415,583
35	178	274,614	338,250	13,270	652,300	424,284
36	241	274,633	352,250	12,880	772,300	443,167
37	141	274,594	331,350	13,800	562,300	392,522
38	153	274,601	333,450	13,610	592,300	405,027
39	165	274,609	335,150	13,380	622,300	415,583
40	178	274,616	336,950	13,150	652,300	424,284
41	190	274,626	337,350	12,810	682,300	431,251
42	207	274,625	344,350	13,010	706,900	435,767
43	141	274,595	330,550	13,730	562,300	392,522
44	178	274,616	337,350	13,190	652,300	424,284
45	191	274,624	338,450	12,900	682,300	431,251
46	207	274,626	344,050	12,980	706,900	435,767
47	224	274,629	348,550	12,970	742,300	440,545

No	Peso Molecular[kg/kmol]	Índice de Refracción[K]	Punto de anilina[K]	Contenido H2[%m]	T ebull prom[K]	Flash Point[K]
48	239	274,638	349,550	12,660	772,300	443,167
49	154	274,598	335,950	13,810	602,450	408,813
50	156	274,599	336,150	13,780	606,250	410,173
51	203	274,623	344,050	13,090	733,750	439,566
52	194	274,618	342,750	13,240	692,550	433,259
53	221	274,629	347,650	12,950	761,250	442,343
54	242	274,650	344,650	12,130	795,050	444,378
55	230	274,647	341,950	12,170	778,750	443,575
56	159	274,670	347,650	10,430	652,450	424,323
57	157	274,691	294,750	9,510	663,850	427,164
58	203	274,736	297,050	8,390	777,350	443,491
59	223	274,638	343,650	12,530	771,250	443,096
60	128	274,591	325,950	13,710	589,350	403,885
61	191	274,614	343,350	13,390	718,550	437,554
62	208	274,679	320,250	10,610	760,600	442,290
63	206	274,646	334,250	12,040	746,050	440,941
64	190	274,654	324,350	11,520	729,860	439,085
65	196	274,635	335,250	12,450	729,790	439,076
66	203	274,621	345,050	13,190	731,750	439,322
67	206	274,603	355,350	14,140	712,650	436,677
68	212	274,628	344,550	12,910	755,050	441,810
69	209	274,653	332,150	11,750	758,850	442,143
70	203	274,640	335,450	12,270	738,850	440,163
71	188	274,608	345,550	13,690	712,350	436,630

No	Peso Molecular[kg/kmol]	Índice de Refracción[K]	Punto de anilina[K]	Contenido H2[%m]	T ebull prom[K]	Flash Point[K]
72	245	274,632	354,050	12,960	798,600	444,512
73	190	274,652	324,950	11,600	718,050	437,482
74	194	274,643	330,350	12,020	720,850	437,882
75	159	274,617	327,950	12,870	672,450	429,146
76	177	274,629	330,450	12,520	711,150	436,445
77	180	274,619	336,950	13,070	711,050	436,429
78	197	274,654	326,650	11,560	741,750	440,485
79	199	274,647	330,850	11,910	744,550	440,785
80	209	274,634	340,650	12,600	746,450	440,982
81	203	274,611	349,850	13,680	750,550	441,389
82	181	274,597	348,550	14,190	722,550	438,118
83	207	274,609	352,550	13,810	758,350	442,100
84	203	274,605	353,250	14,000	753,650	441,682
85	202	274,599	355,950	14,300	744,150	440,743
86	198	274,611	347,950	13,630	744,850	440,816
87	188	274,611	344,050	13,550	730,850	439,210
88	183	274,620	337,150	13,010	732,550	439,420
89	194	274,602	351,550	14,100	719,650	437,712
90	194	274,604	350,550	13,990	721,950	438,035
91	207	274,609	352,650	13,840	725,650	438,538
92	202	274,609	350,450	13,770	724,350	438,364
93	196	274,617	344,350	13,330	722,950	438,173
94	214	274,614	352,750	13,650	751,150	441,447
95	204	274,621	345,350	13,200	740,650	440,365

No	Peso Molecular[kg/kmol]	Índice de Refracción[K]	Punto de anilina[K]	Contenido H2[%m]	T ebull prom[K]	Flash Point[K]
96	205	274,617	347,650	13,400	749,650	441,302
97	205	274,619	346,550	13,280	745,550	440,889
98	203	274,615	348,050	13,490	727,460	438,777
99	154	274,618	325,350	12,750	627,250	417,143
100	153	274,598	335,250	13,780	601,750	408,559
101	164	274,613	332,150	13,120	622,300	415,583
102	206	274,628	342,450	12,850	706,900	435,767
103	259	274,637	356,350	12,830	802,300	444,635
104	152	274,603	331,850	13,470	592,300	405,027
105	192	274,622	339,850	13,020	682,300	431,251
106	207	274,627	343,250	12,930	706,900	435,767
107	223	274,631	347,550	12,890	742,300	440,545
108	224	274,628	348,950	13,000	742,300	440,545
109	238	274,639	348,550	12,590	772,300	443,167
110	165	274,609	335,150	13,380	622,300	415,583
111	156	274,601	334,850	13,650	622,150	415,535
112	234	274,631	351,050	12,950	777,350	443,491
113	229	274,630	349,850	12,970	787,650	444,053
114	169	274,707	294,150	9,070	725,150	438,472
115	204	274,638	336,650	12,360	739,760	440,265
116	193	274,636	333,550	12,370	725,340	438,497
117	198	274,644	331,750	12,050	734,360	439,640
118	193	274,628	337,050	12,720	723,840	438,295
119	206	274,659	328,250	11,460	749,120	441,250

No	Peso Molecular[kg/kmol]	Índice de Refracción[K]	Punto de anilina[K]	Contenido H2[%m]	T ebull prom[K]	Flash Point[K]
120	212	274,640	339,150	12,370	757,850	442,057
121	208	274,668	324,950	11,080	765,350	442,668
122	206	274,690	315,050	10,130	772,050	443,150
123	219	274,617	353,150	13,530	759,350	442,185
124	238	274,628	354,350	13,140	790,250	444,175
125	186	274,660	319,950	11,210	714,500	436,958
126	199	274,635	336,550	12,480	723,600	438,262
127	156	274,628	321,750	12,310	673,750	429,434
128	162	274,607	334,650	13,440	671,550	428,945
129	175	274,639	324,450	12,000	711,450	436,491
130	188	274,628	335,550	12,710	707,550	435,873
131	209	274,615	349,950	13,520	765,250	442,660
132	197	274,606	350,150	13,880	740,950	440,398
133	187	274,602	348,550	14,000	711,550	436,507
134	181	274,605	344,050	13,760	703,350	435,179
135	202	274,609	350,450	13,770	723,550	438,255
136	210	274,609	353,750	13,870	729,950	439,096
137	200	274,609	349,750	13,750	724,350	438,364
138	229	274,595	368,350	14,750	755,950	441,890
139	171	274,603	340,650	13,740	652,050	424,218
140	204	274,618	347,150	13,370	741,850	440,496

Anexo B.. Matriz de Correlación

Matriz Correlación	CI	Densidad	T10	T50	T90	Viscosidad	Peso Molecular	Índice de Refracción	Punto de Anilina	Contenido de H2	Temp Ebull	Flash Point
CI	1											
Densidad	0,2936	1										
T10	0,1196	0,2323	1									
T50	0,0846	0,379	0,6437	1								
T90	0,0669	0,2055	0,1382	0,6703	1							
Viscosidad	0,0278	0,4014	0,7288	0,7547	0,3388	1						
Peso Molecular	0,2455	0,173	0,6348	0,9425	0,6345	0,7037	1					
Índice de Refracción	0,362	0,9937	0,1855	0,3043	0,1582	0,342	0,1179	1				
Punto de Anilina	0,9287	0,251	0,0967	0,1382	0,1373	0,342	0,3302	0,3217	1			
Contenido de H2	0,5563	0,8948	0,0763	0,1284	0,0526	0,0424	0,0184	0,9317	0,5221	1		
Temp Ebull	0,0863	0,2912	0,3369	0,8726	0,9395	0,1887	0,8269	0,2282	0,1547	0,0843	1	
Flash Point	0,0775	0,269	0,2611	0,7704	0,9063	0,3693	0,7013	0,2119	0,1265	0,0797	0,9318	1

Anexo C. Codigos

Código R-proyec.

```
read.excel<-
function(header=TRUE,...){read.table("clipboard",sep="\t",header=header,dec=",",..
.)}
cetano=read.excel()
fix(cetano)
Regresion1=lm(CI~0+Den+t10+t50+t90+V+Pan+CH2+PF, data=cetano)
summary(Regresion1)
layout(matrix(c(1,2,3,4),2,2)) # 4 figuras por página
plot(Regresion1)
residuos=residuals(Regresion1)
shapiro.test(residuos)
lev=hat(model.matrix(Regresion1))
levmax=2*sum(lev)/98 # máximo leverage.
plot(lev)
abline(levmax,0,col="red",lwd=1,lty=3)
abline(levmax,0)
cetano[lev>levmax,]
levmax
cook= cooks.distance(Regresion1)
plot(cook,ylab="Distanciade Cook")
abline(cook,0)
cookmax=2*sum(cook)/98 # máximo cook.
cetano[cook>cookmax,]
cookmax
restu=rstudent(Regresion1)
plot(restu)
abline(restu,0)
```

```
cetano[abs(restu)>1.95,] # Puntos fuera del rango
abline(1.9553,0,lty=3)
abline(-1.9553,0,lty=3)
cetano[abs(restu)>1.9553,]
fix(restu)
plot(restu)
rstan=rstandard(Regresion1)
plot(rstan)
fix(rstan)
cook[cook>cookmax]
anova(Regresion1)
```

Anexo D. Técnicas de pre-tratamiento de datos espectrales

Como su nombre lo indica, estos métodos son aplicados antes de analizar y correlacionar los datos espectrales para eliminar los efectos (ruido, baja resolución, desviación línea base etc.) que distorsionan o desmejoran la información ya sea por el tipo de muestra que se esté analizando o por la herramienta espectroscópica usada. Los pre-tratamientos más usados en espectroscopia son (Hopke & Philip, 2003):

Promediado de espectros: Se realiza para eliminar ruido debido a que este es de naturaleza aleatoria. Consiste en realizar un promedio de las señales adquiridas bajo las mismas condiciones y de la misma muestra aumentando así la relación señal/ruido.

Suavizado: Es usado para eliminar el ruido espectral, cuando el promediado no es suficiente. Este es realizado a través de algoritmos matemáticos o filtros que evalúan y discrimina las señales del ruido. Los algoritmos más usados son el de Savitsky-Golay (Savitsky & Golay, 1964), transformada de Fourier (Oppenheim, 2007) y transformada Wavelet (Chau & Liang, 2004).

Derivación: Se usa para resolver bandas y corregir la línea base, pero presenta como desventaja el aumento del ruido. Las más comúnmente usadas son la primera y segunda derivada. La primera mide la pendiente de la curva espectral en cada punto, con lo cual se hace cero aquellos puntos donde se encuentra el máximo de una señal (resuelve señales) y la segunda mide el cambio de la pendiente de la curva, con lo cual se realiza corrección de línea Base y se resuelve picos cercanos. El algoritmo más usado es el de Savitsky- Golay.

Corrección de línea Base: La corrección de la línea se utiliza para ajustar el cero espectral y eliminar así efectos no químicos del levantamiento del espectro. Los métodos más usados es el método *Offset* que consiste en encontrar como punto de referencia cero la señal más baja y a partir de ella realizar el ajuste siguiendo la ecuación 8, y la corrección *lineal,* que consiste en tomar dos puntos de referencia como cero, trazándose una recta inclinada, que al aplicarse se vuelve horizontal.

$$f_x = X - minx$$

Normalización: Los datos se normalizan cuando se quiere analizar un conjunto de espectros de diferentes muestras para lograr que los datos estén aproximadamente a la misma escala. Esta puede realizarse por varios métodos:

Normalización por área:

$$X_i = x_i / \sum_j x_{ij}$$

Normalización por vector unitario:

$$X_i = x_i / \sqrt{\sum_j x_i^2}$$

Normalización con la media:

$$X = \frac{xi}{x}$$

Normalización con máximo:

$$X_i = x_i / x_{max}$$

Normalización por rangos:

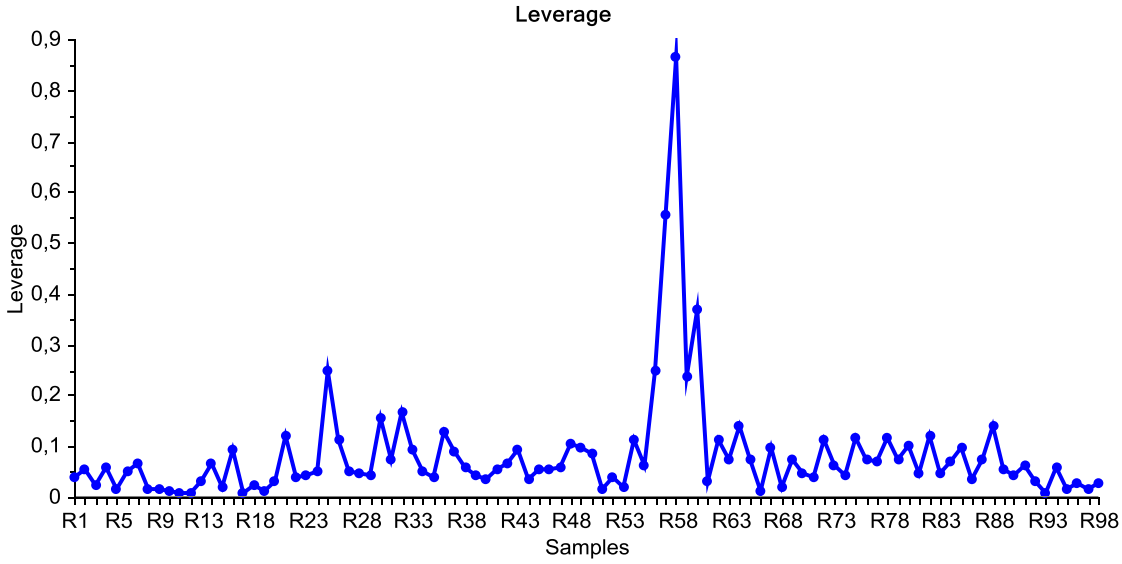
$$X_i = x_i / (x_{max} - x_{min})$$

Normalización por pico:

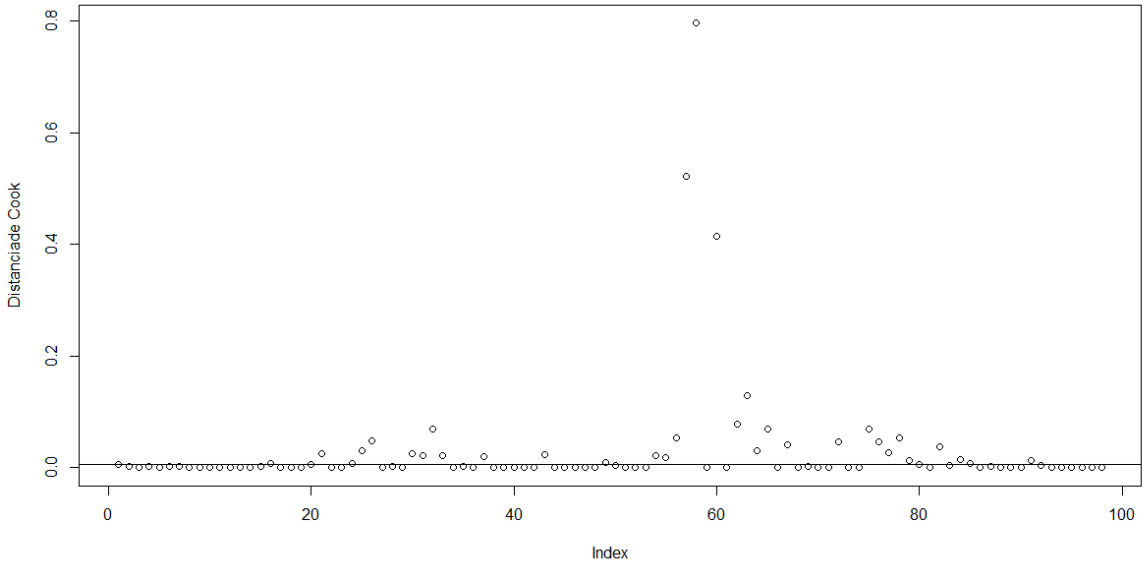
$$X_i = \frac{x_i}{x_{ik}}$$

Anexo. E. Leverage y distancia COOK

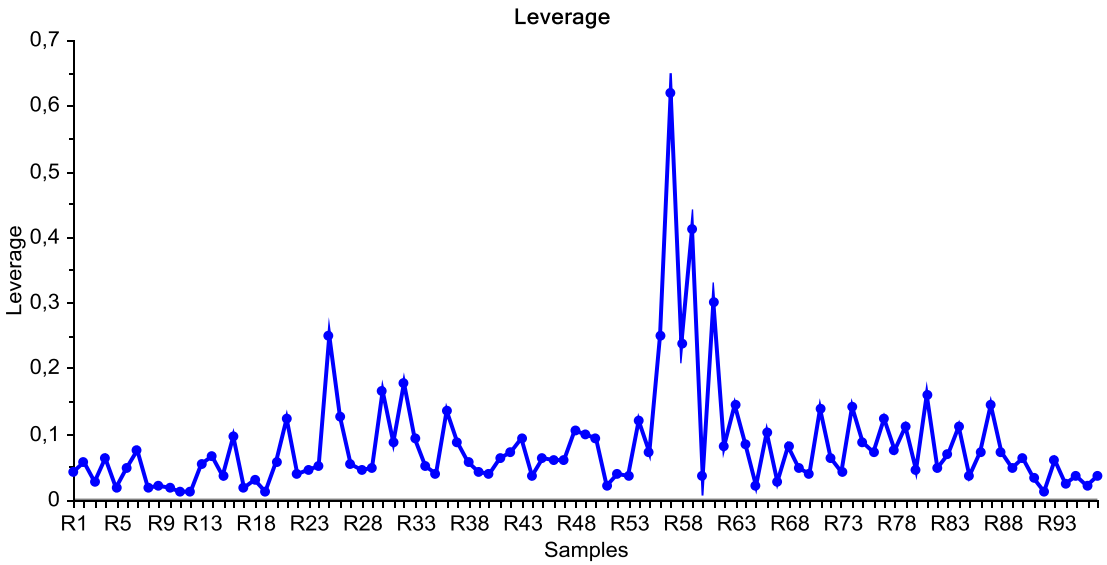
Leverage Modelo Inicial 8 Variables



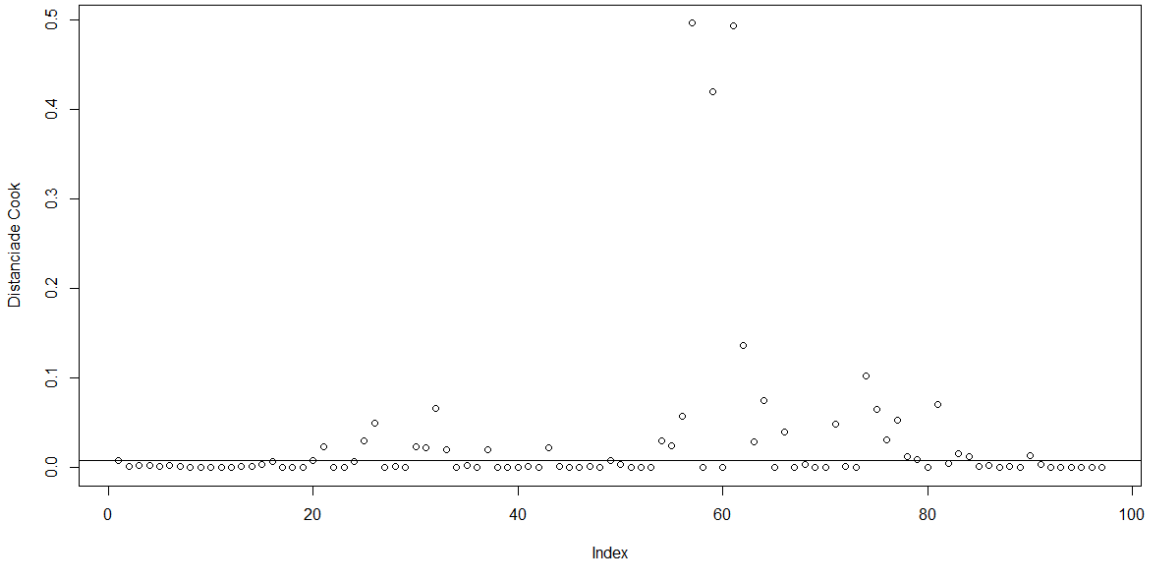
Distancia de Cook modelo Inicial 8 Variables



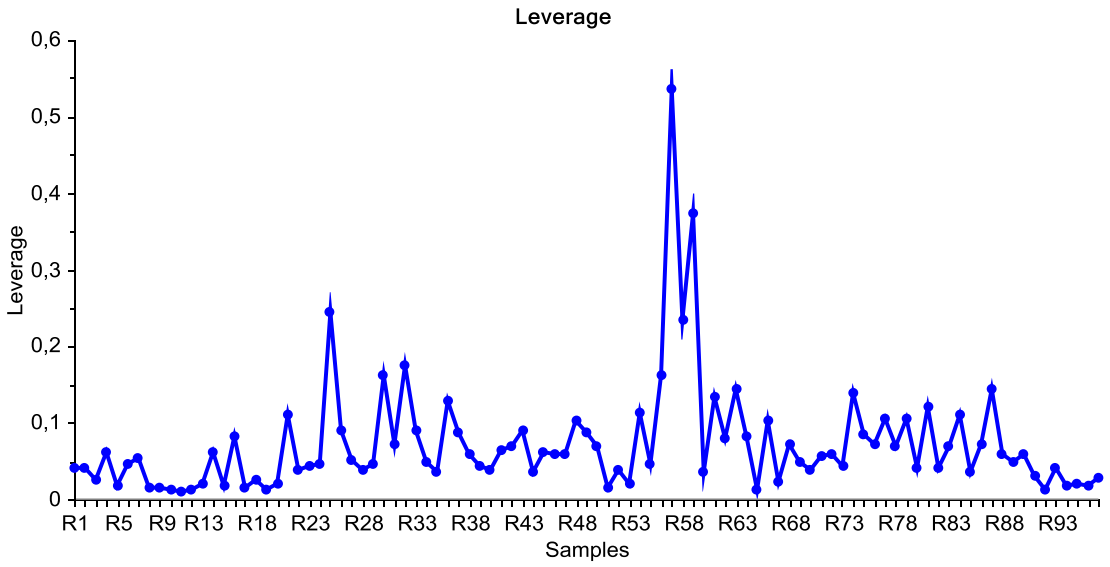
Leverage Modelo Inicial sin Outliers



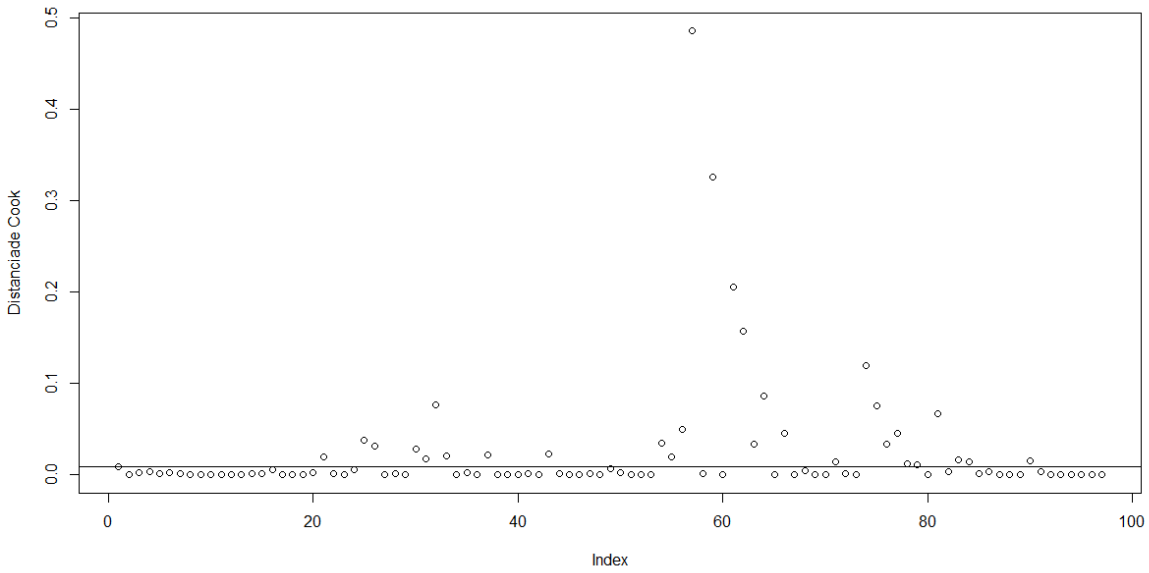
.Distancia de Cook Modelo inicial sin Outliers



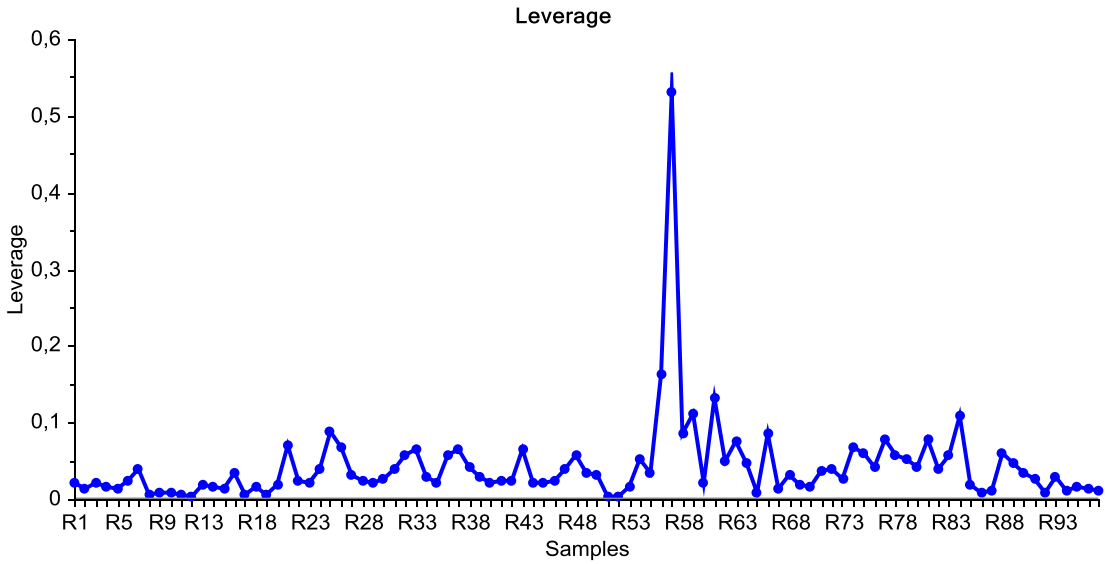
Leverage Modelo 7 sin outliers



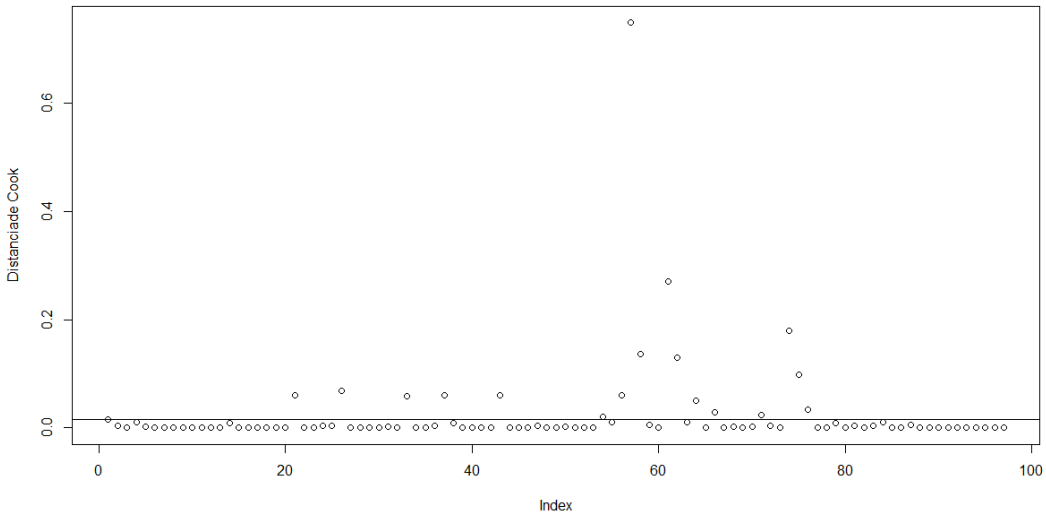
.Distacia de Cook Modelo 7 sin Outliers



Leverage Modelo 4 sin Outliers



Distancia de Cook Modelo 4 sin Outliers



Anexo F. CI 976 vs CI 4737

CI ASTM D-976	CI ASTM D-4737
44,843	45
50,225	51
51,508	51
46,659	47
50,065	50
43,368	43
44,041	45
52,181	54
53,383	56
51,752	52,5
52,327	53,7
50,921	52
52,450	57
48,518	48,6
51,489	55
44,470	45
50,404	51
47,802	47,8
49,158	49,300
50,848	55
37,787	39,7
46,484	47,8
48,076	50,6
49,823	56,6
51,101	64,8

CI ASTM D-976	CI ASTM D-4737
37,187	39
43,974	44,6
47,073	48,5
47,812	50,3
48,764	52,5
49,611	56,2
49,969	59,5
39,202	41,3
45,466	46,4
48,335	50,2
51,419	61,8
39,364	41,5
43,353	44,6
45,797	46,8
47,247	48,7
46,833	48,9
50,580	55,2
38,634	40,7
47,596	49,2
47,648	50,1
50,354	54,9
51,486	59,1
49,533	58,8
45,516	46
45,910	46
50,847	53
50,890	52

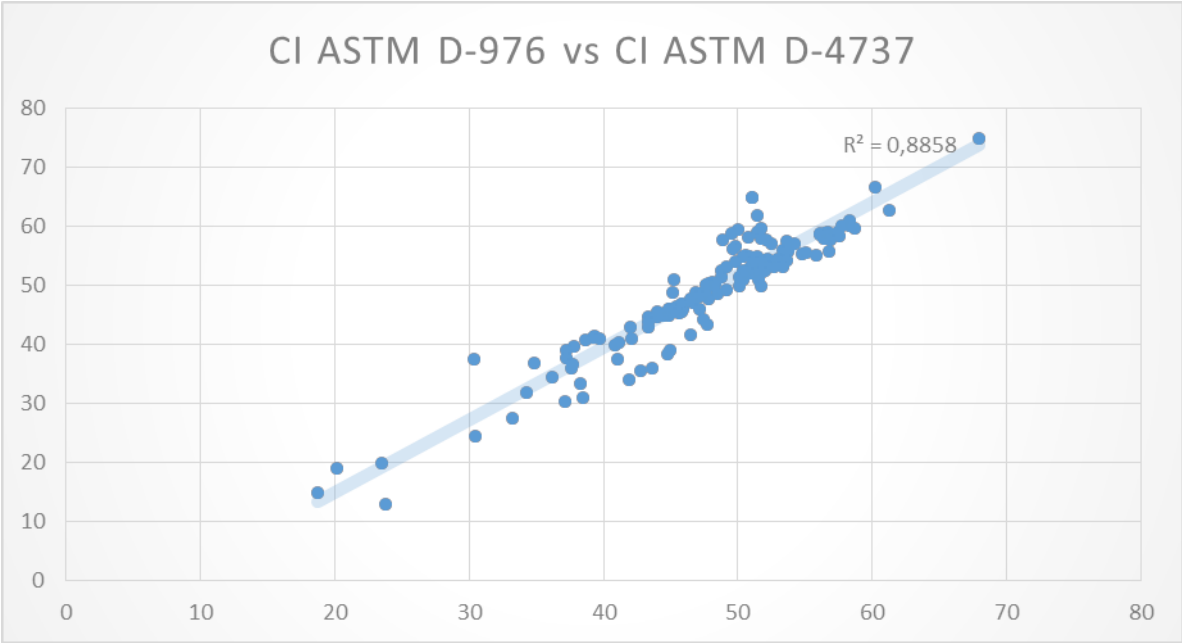
CI ASTM D-976	CI ASTM D-4737
51,200	54
45,284	51
45,150	48,8
23,448	20
18,689	15
23,728	13
47,721	43
30,373	37,5
51,696	49,8
33,180	27,6
42,784	35,6
37,107	30,4
44,717	38,3
51,719	52,5
60,244	66,7
50,152	51,4
40,891	40
44,140	45,3
53,758	55,7
52,126	57,7
37,585	35,9
41,171	40,4
39,749	41,1
42,135	41
47,169	46
38,242	33,3
41,059	37,5

CI ASTM D-976	CI ASTM D-4737
47,392	44,2
55,850	55,2
56,763	55,8
57,548	58,3
58,717	59,6
61,273	62,7
54,812	55,3
52,552	53,4
47,169	48,1
58,324	60,2
57,424	59,1
57,717	60,2
56,520	58,8
52,007	52,6
56,694	59
51,841	52,2
53,682	54,2
52,699	53,1
54,229	57
37,235	37,8
44,876	46
43,327	43,8
49,174	53,1
51,101	64,8
42,041	43
48,739	51,5
49,841	54,1

CI ASTM D-976	CI ASTM D-4737
50,803	58,1
51,767	59,6
48,899	57,7
45,797	46,8
44,892	46
51,766	58
51,745	54
20,143	19
44,940	39
43,659	36
41,870	34,1
46,438	41,7
38,482	31
45,803	45,6
36,134	34,5
30,407	24,4
56,128	58,9
53,619	57,6
34,312	31,9
45,384	45,6
34,850	36,9
45,181	45,9
37,748	36,6
45,627	45,4
55,116	55,6
56,845	57,7
56,362	57,9

CI ASTM D-976	CI ASTM D-4737
53,000	54,4
56,520	58,9
58,286	60,9
56,102	58,5
67,993	74,8
50,385	52,5
53,340	53,2

CI ASTM D976 vs CI ASTM D4737



Anexo G. Escalamiento de Datos

Para el escalamiento de los datos iniciales se utilizó el método de Mead and Center.

Como resultado el modelo que se obtuvo fue:

$$CI = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + \cdots + a_nX_n$$

$$CI = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + \cdots + a_nX_n$$

$$CI = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + \cdots + a_nX_n$$

$$CI = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + \cdots + a_nX_n$$

Donde:

x: Variable

$\bar{x}$: Promedio de la Variable

Y así sucesivamente con todas las variables, para regresar a las variables originales aplicamos la siguiente ecuación:

$$x = (X_1 * Desv.Est\ x) + \bar{x} \qquad \text{(Ecuación 6)}$$

Los valores de Desviación Estándar y Promedios de las propiedades se encuentran en Tabla 17

$$X_1 = \frac{x - \bar{x}}{Desv.Est\ x}$$

$$X_1 = \frac{x - \bar{x}}{Desv.Est\ x_1}$$

$$X_1 = \frac{x - \bar{x}}{Desv.Est\ x}$$

$$X_1 = \frac{x - \bar{x}}{Desv.Est\ x}$$

Desviación .Estándar y Promedio de las Macro propiedades Autoescaladas

Propiedades Macroscópicas	Promedio	Desviación Estándar
Densidad	846,347	33,299
T10	504,916	40,435
T50	540,023	36,736
T90	582,757	44,941
Viscosidad	0,348	0,186
Punto de anilina	340,227	10,247
Contenido Hidrogeno	12,979	0,911
Flash Point	361,208	22,623

Anexo H. Calculo de Flash Point

Para determinar el punto de anilina se usan las siguientes ecuaciones.

Punto de Anilina

$$PA = -183.3 + 0.27API\sqrt[5]{T_{av}} + 0.317T_{av} \quad [22]$$

$$Contenido\ de\ Hidrogeno\ (wt\%) = 30.346 + \frac{82.952 - 65.341RI_{20}}{d_4^{20}} - \frac{306}{MW} \quad [22]$$

Donde: $RI_{20}: Indice\ de\ Refraccion\ \ a\ 20°C\ ; MW: Peso\ Molecular$

$$d_4^{20}: Densida\ \ relativa = \frac{densida\ a\ 20°C\ \left(\frac{g}{cm^3}\right)}{densidad\ \ del\ agua\ a\ 4°C\ \left(\frac{g}{cm^3}\right)},$$

$$densida\ \ agua\ a\ 4°C = 1000\,\frac{g}{cm^3} \quad [22]$$

$$PM = 42.965[\exp(2.097x10^{-4}T_b - 7.78712SG + 2.08476x10^{-3}T_bSG)]T_b^{1.26007}SG^{4.98308}$$

[30]

Donde:

$SG = Gravedad\ Especifica$, $T_b = Temperatura\ de\ Ebullicion$

$d_{20} = SG - 4.5x10^{-3}(2.34 - 1.9SG)$ [31]

Donde: $d_{20}: Densidad\ \ a\ 20°C$

Índice de Refracción a 20°C

$$IR_{20} = 1 + 0.8447SG^{1.2056}T_{av}^{-0.0557}MW^{-0.0044} \quad [22]$$

Donde: $SG: gravedad\ especifica, MW: Peso\ molecular, T_{av}: Temperatura\ 50\%$

Temperatura 50%

$$ASTM\ D\ 86(50\ vol\%) = 255.4 + 0{,}79424[SD(50\ wt\%) - 255.4]^{1.0395}$$

Dónde: SD (50% wt%) y ASTM D 86(50 vol%) son las temperaturas en 50% destiladas en Kelvin. La diferencia entre los puntos de corte adyacentes se calcula de la siguiente ecuación.

$$U_i = ET_i^F$$

Donde: $U_i =$ diferencia en la norma ASTM D86 entre la temperatura en dos puntos de corte [K].

$T_i =$ Diferencia observada en las temperaturas SD entre dos puntos de corte [K].

$E, F =$ Constantes que varían en cada punto de corte.

Para determinar la temperatura de ASTM D86 en cualquier porcentaje destilado, los cálculos deben comenzar con una temperatura 50% ASTM D86 y la adicción o sustracción de la diferencia de la temperatura adecuada Ui.

$$ASTM\ D86\ (10\%) = ASTM\ D86(50\%) - U_4 - U_5$$

Anexo I. Espectros Normalizados