Kudirat Abidemi Obisesan

Identificación del origen geográfico de aceite de palma mediante técnicas de clasificación multivariantes

Trabajo de Fin de Grado

Dirigido por la Dra. Itziar Ruisánchez Capelastegui

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2016

ÍNDICE

1. ABSTRACT

Existe un interés creciente en los consumidores por conocer la procedencia de los productos alimentarios, así como los beneficios para la salud que en algunos casos se pueden asociar al origen geográfico de los mismos. El objetivo principal del trabajo es la diferenciación y la clasificación de muestras de aceite de palma de tres orígenes diferentes. Las muestras fueron analizadas en la Universidad de Granada mediante la técnica HPLC, High Performance Liquid Chromatography por huellas dactilares con detector de cargada aerosol (CAD) y ultravioleta-visible (UV). En estudios previos, se han evaluado el contenido en fitoesteroles como posible variable diferenciadora de aceites y se utilizará en este trabajo.

Las técnicas de clasificación fueron evaluadas por SIMCA (Soft Independent Modelling of Class Analogy) y PLS-DA (Partial Least Squares Discriminant Analysis).

Para el establecimiento del modelo de clasificación, se realizaron diferentes estudios entre ellos, las muestras discrepantes, el pretratamiento, la región discriminante y variables latentes para obtener la condición óptima del trabajo. La presencia de las muestras discrepantes han tenido influencia sobre los resultados de CAD. En datos CAD y UV, se observan en el análisis de principal componentes dos agrupaciones de muestras que son Asia y América, en cambio las muestras de África están dispersas por los dos continentes.

Como cualquier técnica de análisis, una vez establecido el modelo, éste se tiene que validar. La validación de modelos de clasificación multivariante están ampliamente aceptados como parámetros de calidad, los valores de porcentajes de: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Estos parámetros dan una idea de la bondad del modelo. PLS-DA-CAD, fue el que presenta los mejores resultados sobre la denominación de origen de las muestras de aceite de palma con porcentajes de verdadera positiva de 96%, 86%, 73% para la clase 1, 2 y 3 respectivamente.

2. OBJETIVO

El objetivo de este trabajo es diferenciar y clasificar muestras de aceite de palma según su origen geográfico. Para ello, muestras de distinta procedencia fueron analizadas por cromatografía líquida (en la Universidad de Granada). En este trabajo, se ha desarrollado modelos multivariantes de clasificación, por lo que el trabajo ha consistido en los siguientes sub-objetivos:

- Estudio del pretratamiento sobre los datos obtenidos del análisis por cromatografía de "fingerprinting" utilizando dos detectores diferentes.
- Estudio previo de la distribución de las muestras mediante la aplicación de técnicas de exploración de datos: análisis de los componentes principales (PCA). Este estudio permite observar el comportamiento de las muestras es decir, la presencia de outliers, valores discrepantes y agrupaciones entre las muestras.
- Desarrollo, implementación y comparación de modelos de clasificación para las distintas clases que se quieren diferenciar (origen geográfico). Se ha trabajado con una técnica de clasificación de tipo modelador (soft Independent Modelling of Class Analogy, SIMCA) y una técnica de tipo discriminante (Partial Least Discriminant Analysis, PLS-DA).
- Los subobjetivos indicados se han realizado en el entorno de programación MATLAB, por lo que se puede decir que un objetivo de este trabajo también ha sido el estudio y utilización del MATLAB.

.

3. INTRODUCCIÓN

El aceite de palma es un aceite de origen vegetal que se obtiene del fruto de la palma (Elais guineensis) y se consume desde hace más de 5000 años [1].

El aceite de palma es originario de África Occidental, del golfo de Guinea y a partir del siglo XV se introdujo en otras regiones de África, el Sudeste Asiático y Latinoamérica a lo largo de la zona ecuatorial. Actualmente, se ha convertido en el aceite con más volumen de producción debido a sus bajos costes de producción y sus múltiples usos. El principal centro de producción de aceite de palma se localiza en países asiáticos principalmente Malasia e Indonesia [2].

El aceite de palma no refinado se caracteriza por tener una coloración rojiza por la aportación de los carotenos. La presencia de los carotenos proporcionan un gran efecto antioxidante, lo que implica una acciona protectora de la piel, los ojos y otras partes del organismo. Uno de los aspectos beneficiosos para la salud es el poder regenerativo y el mantener un correcto estado de la retina ocular y la visión, lo que previene enfermedades como la pérdida de visión por degeneración muscular o la ceguera nocturna [3].

El aceite de palma no refinado está considerado como uno de los alimentos más ricos en vitamina A. Sin embargo, pierde algunas características de su valor nutritivo o calidad de sus ácidos grasos durante el refinado.

Este aceite refinado, también tiene una gran aplicabilidad, entre ellos en las que cabe destacar en la industria de alimentaria (siendo mayoritaria, más de un 50%) para la producción de margarinas, sopas, patatas fritas, helados, bizcochos, galletas, etc. También se utilizan en la industria química, cosmética, alimentación animal y más recientemente, los aceites reciclados son considerados como una alternativa de materia prima para la producción de combustible (biodiesel).

Respecto a su influencia en el medioambiente, las plantaciones de palma pueden tener impacto sobre la biodiversidad, deforestación y sobre la economía. Además, es un cultivo que requiere uso de fertilizantes y las pesticidas, el mal uso de los cuales puede provocar contaminación del ríos y reducir la biodiversidad [4].

El interés en clasificar los aceites de palma en función de su origen geográfico se fundamenta en la demanda de los consumidores sobre la indicación de la procedencia geográfica de los productos y sobre el etiquetado de los alimentos ya que los consumidores lo consideran como un valor añadido al producto.

Por otra parte, el interés de su estudio geográfico es identificar propiedades culinarias específicas (nutrición), cualidades organolépticas (características físico-químicas), o beneficios para la salud asociados con productos regionales.

En cuanto a su composición, el aceite de palma contiene proporciones similares de ácidos grasos saturados e insaturados: alrededor del 40% de ácido oleico (monoinsaturado), un 10% de ácido linoleico (-poliinsaturado), 44% de ácido palmítico (saturado) y 5% de ácido esteárico (saturado) [5].

Un componente que se encuentra en prácticamente todos los alimentos vegetales son los fitoesteroles, esteroles de origen vegetal y se encuentran en mayor concentración en los aceites como aceite de maíz, girasol, soja y colza siendo una excepción el aceite de palma ya que pierde la mayor parte de los esteroles en el proceso de refinado [6].

En algunos estudios [7, 8], se han evaluado el perfil de esterol y contenido total de esteroles como marcadores moleculares para valorar la autenticidad de aceite porque cada especie vegetal tiene un perfil de composición característica de los esteroles.

La técnica instrumental más utilizada para el análisis de aceites de palma es la cromatografía. Si bien, este tipo de análisis no permite una asignación a un determinado origen geográfico. Estudios recientes [9] combinan el análisis mediante técnicas cromatográficas por huella dactilar (fingerprinting) combinada con técnicas quimiométricas (técnicas de clasificación multivariante). De esta forma, se puede verificar el origen geográfico de estos alimentos debido a la cantidad de información proporcionada por la técnica instrumental.

El análisis quimiométrico, mediante la transformación de la señal instrumental en un matriz de datos, permite establecer modelos de clasificación con el fin de extraer eficientemente el máximo de información útil de los datos.

El trabajo consiste en clasificar y diferenciar las muestras del aceite de palma a partir de los datos obtenidos por el análisis cromatográfico del aceite de palma de tres orígenes geográficos (Asia, África y América del sur) aplicando, técnicas de exploración y dos técnicas de clasificación, SIMCA y PLS-DA.

4. DESCRIPCIÓN DE LAS MUESTRAS

Se ha trabajado con un total de 100 muestras de aceite de palma, de diferentes orígenes geográficos, provenientes por la Universidad de Wageinngen (Países Bajos) que fueron analizadas mediante cromatografía líquida por la Universidad de Granada.

La distribución geográfica de las muestras de aceite de palma se indica en la **tabla 1**: 56 muestras de Asía (mayoritariamente de Indonesia y Malasia), 28 muestras de África (mayoritariamente de Ghana) y 16 muestras de América (todas de Brasil). En la tabla 1, muestra los detalles de los orígenes de las muestras analizados**.**

Tabla 1, origen geográfico de las 100 muestras de aceite de palma

Continentes	Países	Muestras	Total
Asia	India	5	56
	Indonesia	24	
	Malasia	19	
	Papúa. N. Guinea	7	
	Salomón	1	
África	Camerún	3	28
	Ghana	18	
	Guinea	2	
	África occidental	5	
América	Brasil	16	16

Como resultado del análisis, los datos analíticos obtenidos (cromatogramas) de los dos detectores fueron dispuestos en dos matrices para establecer el modelo de clasificación. La primera matriz está formada por el conjunto de datos de UV, con dimensiones 100 x 3436. El valor 100 corresponde a los números de muestras estudiados y 3436 a número de variables, en este caso corresponde a la intensidad obtenidos a un tiempo de análisis total aproximadamente 20 min. La segunda matriz corresponde a los datos de CAD de 100 x 1609. La matriz está compuesta de 100 filas (muestras) como en el caso anterior y 1609 columnas (tiempo de retención) como los puntos de datos cromatogramas registradas durante el tiempo de adquisición.

5. FUNDAMENTO TEÓRICO

La Quimiometría es la disciplina química que usa la matemática, estadística, y lógica para analizar matrices de datos que contienen muchas variables. Algunas de las áreas de aplicación más importantes de la quimiometría incluyen la calibración, validación y pruebas de significación, la optimización de las mediciones químicas y procedimientos experimentales y la extracción del máximo de información química [10] a partir de datos analíticos. Actualmente, tienen una gran importancia para garantizar la autenticidad de productos alimenticios y el control de calidad de productos y procesos [11, 12].

5.1. TÉCNICA INSTRUMENTAL

Las técnicas instrumentales más utilizadas en la determinación de la autenticidad de los alimentos suelen ser espectroscópicas como NIR [12] y resonancia magnética (RMN) [13]. Este tipo de técnicas se caracteriza por ser no selectivos, requieren poco pretratamiento de la muestra y el tiempo de análisis es rápido. Al ser no selectivas, implican trabajar con espectro y por lo tanto requieren un tratamiento multivariante (quimiométrico) de los datos. Permiten realizar un análisis cuantitativo y cualitativo, siendo este último el más desarrollado. Por ello, también se refiere a ellas como técnicas de "huellas dactilar"

Las técnicas de análisis cromatográficas, histórica y principalmente se han aplicado en el ámbito cuantitativo (cada pico corresponde a un analito). Recientemente, se están desarrollando cromatografía de huella dactilar (fingerprinting) que no buscan la selectividad en la diferenciación e identificación de los picos (cada pico corresponde a un grupo de analitos). Ello permite, acortar los tiempos de análisis y requiere del análisis quimiométrico de todo el cromatograma (todos los picos) y permite extraer información de tipo cualitativo [14].

La técnica instrumental utilizada para el análisis de esteroles en el aceite vegetal en este trabajo es la técnica cromatográfica de huella dactilar. Los métodos cromatográficos son los métodos más utilizados porque proporciona mucha información sobre la composición de esteroles presentes en la muestra.

El análisis de las composiciones de esteroles en aceite vegetales por cromatografía líquida principalmente se realiza en fase reversa (la fase estacionaria apolar y una fase móvil de polaridad moderada). También se utiliza el método por la fase normal (una fase estacionaria polar y una fase móvil apolar), para la cuantificación absoluta de la cantidad total de fitosteroles pero este método muestra una baja resolución entre los picos cromatográficos. Como consecuencia, no proporcionan información precisa sobre la composición de esteroles [15, 16].

En estudios anteriores [17, 18], se han utilizado el perfil y el contenido de los esteroles

para evaluar la autenticidad de aceite por este motivo, las tres clases de esteroles, dimetilesterol, desmetilesterol y metilesterol, se consideran en este trabajo como herramienta para el desarrollo de análisis para verificar el origen geográfico de aceite de palma.

En este proyecto para el análisis de las muestras, se emplearon dos sistemas de cromatografía líquida diferentes de fase normal con las siguientes condiciones:

El primero es el sistema de HPLC konik 560 (Konik-Tech, Sant Cugat del Valle, Barcelona, Spain) con una bomba cuaternaria y inyector es automática de 20µl con detector ultravioleta-visible (UV) trabajando a longitud de onda de 202nm.

El segundo sistema es HPLC de Agilent 1100 (Agilent Technologies, Santa Clasra, CA, USA) con bomba cuaternaria, desgasificador, automostrador y detector aerosol cargado (CAD). El funcionamiento del CAD consiste tres etapas que son la nebulización, la evaporación y detección.

En la primera etapa, el efluente (la fase móvil) procedente del sistema cromatográfico se nebuliza, convirtiendo los analitos en partículas con el gas Nitrógeno. El tamaño de las partículas aumenta con la cantidad de analitos. En la etapa de evaporación, la corriente de partículas de analito incide sobre el flujo de gas de nitrógeno cargada positivamente. A continuación, estas cargas se transfieren a un colector en el que las cargas eléctricas se miden mediante un electrómetro de alta sensibilidad. Esto genera una señal cuya intensidad es directamente proporcional a la cantidad de analito [19]. La presión del gas utilizado en el trabajo fue ajustada a 35psi.

En ambos sistemas, la columna es 250 mm de largo, 4mm de diámetro interno y 5µm de tamaño de partícula. La composición de la fase móvil está formado por n-hexano (99%)/2-propanol (1%) sin gradiente y con un flujo de 1.2mlmin^{-1}. El tiempo de análisis en los dos HPLC es aproximadamente de 20min.

El detector CAD es característico por tener una alta sensibilidad, proporcionando una respuesta coherente, y tiene un amplio rango dinámico, particularmente cuando el análisis de compuestos que carecen de cromóforos UV. Además, la respuesta CAD no depende de las características químicas de los compuestos de interés, sino en la concentración inicial del analito proporcionando una respuesta mucho más uniforme en comparación con el detector UV, donde las respuestas puede variar con la longitud de onda utilizada y el coeficiente de extinción [20, 21].

5.2. PREPROCESAMIENTO DE DATOS

El preprocesado de los datos es un paso fundamental que consiste en la manipulación o modificación de los datos realizada antes de aplicar cualquier técnica multivariante.

El objetivo es eliminar o al menos reducir las fuentes de variación que no son de interés. Son fuentes de variabilidad en la señal ya sean de carácter aleatorio o de carácter sistemático que no están relacionadas con el analito o la propiedad de interés. De no ser eliminada, el modelo requiere un trabajo más complejo para obtener la información de interés. Al mismo tiempo, es importante su aplicación de manera correcta ya que lo contrario puede generar resultados erróneos [22].
En estudios anteriores [15, 23, 24] basada en técnicas por huellas dactilares, algunos de los pretratamientos utilizados son autoescalado, corrección de línea de base, alineamiento y centrado.
De este modo, algunos de los preprocesados que se han utilizado a lo largo de trabajo son los siguientes: alineación, Centrado, autoescalado, normalizado y suavizado.

5.2.1. Alineación

La alineación es un pretratamiento necesario para la corrección del desplazamiento del tiempo de retención. Estos desplazamientos característicos en la técnica de HPLC, puede ser causada por variaciones en la composición de la fase móvil, temperatura, flujo, etc. Es importante una correcta alineación de los picos cromatográficos para que al aplicar técnicas quimiométricas sobre las señales de cromatografía, poder interpretar correctamente las relaciones entre muestras y variables [25, 26].

5.2.2. Centrado de datos

El centrado calcula el valor medio de cada variable (columna) de la matriz de datos y se resta a cada valor de la columna según la **ecuación 1**. Donde el $X_{i,centrado}$ es el dato centrado, $X_{i,m}$ es el dato de la fila i (o muestra i) y la columna m (o la variable m) antes del centrado y X_m es la media de la columna m [27].

$$X_{i,centrado} = X_{i,m} - X_m \qquad \text{Ecuación 1}$$

La **figura 1**, muestra el comportamiento de los resultados después de aplicar el centrado. Este tipo del pretratamiento permite mantener las informaciones originales sin modificar la varianza de los datos. La propiedad fundamental de los datos es que el valor medio de cada una de las variables es igual a cero.

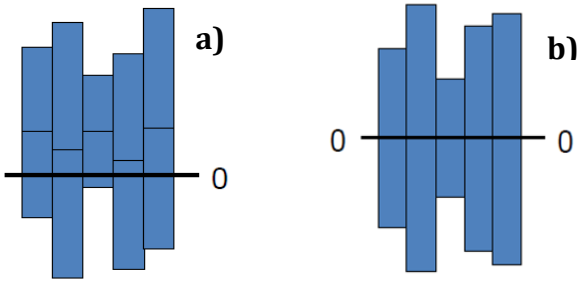

Figura 1, a) datos originales **b)** datos centrados[1]

5.2.3. Autoescalado

El autoescalado se lleva a cabo después de centrado y se divide los valores de cada columna por la desviación estándar de cada columna (**ecuación 2**).

$$\frac{X_{ik}-X_k}{S_k}$$ Ecuación 2

Como muestra en la ecuación 2, la X_{ik} corresponde a los valores de la fila i y de la columna k. El X_k corresponde al valor de la media de la columna k y S_k es la desviación estándar de los valores de la columna k. El resultado obtenido después del autoescalado es que todas las columnas tienen de media cero y varianza unitaria **(Figura 2).**

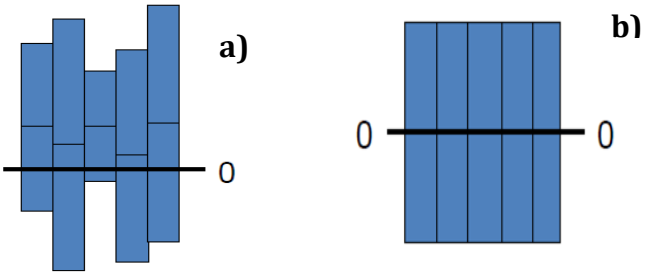

Figura 2, a) datos originales **b)** datos autoescalado[1]

Este tipo de pretratamiento se utiliza cuando las variables originales están expresadas en unidades distintas o cuando hay diferentes variabilidades entre los datos. El autoescalado permite que cada variable tenga la misma influencia en el cálculo [28]. La figura 2, muestra el comportamiento de los resultados después de aplicar el autoescalado a los datos.

5.2.4. Normalización

El pretratamiento por normalización se utiliza para la corrección de diferencia en la intensidad global**.** En cromatografía, la normalización de todo el cromatograma de unidad de área es para eliminar el efecto de volumen de inyección variable.

La **figura 3** es un ejemplo genérico de una aplicación donde se puede ver que el

[1] Figuras obtenidos en: Asignatura técnicas analíticas: quimiometría

cromatograma normalizado permite la eliminación de la variación de volumen de inyección [29].

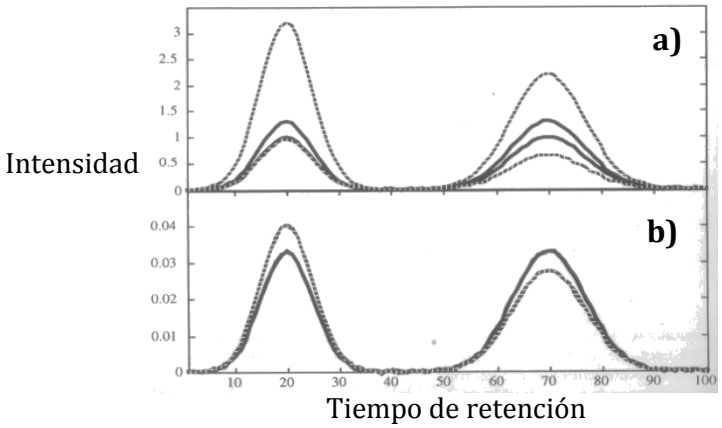

Figura 3, **a)** Cromatograma original **b)** datos normalizado

5.2.5. Suavizado

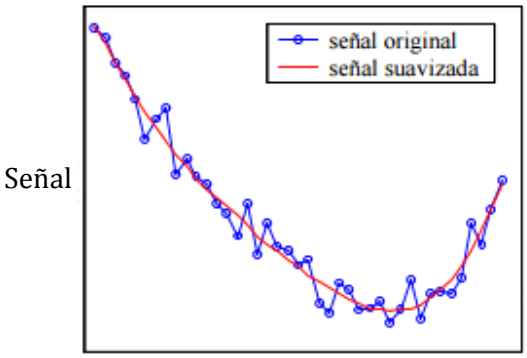

Figura 4, datos suavizados

El objetivo de suavizado es reducir el ruido de la señal analítica instrumental. El suavizado consiste en sustituir cada punto de la señal cromatograma para la media de los valores correspondientes a una serie de puntos adyacentes (se define una ventana alrededor del punto a modificar) [30].

5.3. TÉCNICA DE EXPLORATORIO DE DATOS

El análisis exploratorio de datos son herramientas estadísticas que permite estudiar la estructura de los datos (existencia de agrupaciones) así como detectar si alguna muestra tiene un comportamiento muy diferente al resto (posible outlier).
Una de las técnicas de exploración más utilizadas es el análisis de componentes principales así como las técnicas de agrupación.

5.3.1. Análisis de componentes principales (PCA)

El análisis de componentes principales, PCA, es una técnica de representación en el que las muestras son expresadas en unas nuevas variables (PC's) que son combinación lineal de las variables originales (matriz de datos X). Estas nuevas variables se calculan de manera que retienen el máximo de información presentada en la matriz de datos original. De esta manera, se consigue maximizar la información de la varianza presente en un conjunto de datos y representarla en un menor número de dimensiones [31].
Matemáticamente, la matriz de datos X (I x J), (en caso objeto de estudio, I corresponde a las muestras y J a los valores del cromatograma), se descompone en dos matrices según la **ecuación 3**:

$$X = T P^T + E \qquad \text{Ecuación 3}$$

Donde T es la matriz de los "scores" que tiene tantas files como la matriz original (igual al número de muestras), P es la matriz de los "loadings" que contiene tantas columnas como la matriz original (variables, en el caso objeto de estudio valores del cromatograma) i E es la matriz del error o residuales. La matriz de scores proporciona información del valor de cada muestra y los loadings proporcionan información de las variables. Aunque se pueden calcular tantos componentes principales como el mínimo entre el número de muestras y variables (en nuestro caso 100, ya que se dispone de 100 muestras y un número muy superior de variables), la mayor parte de la varianza esta explicada en los primeros componentes, de manera que el PCA puede utilizarse como un método de reducción de variables.
La expresión de los componentes principales vienen dados según la **ecuación 4** el número de PCs es idéntico al número de las muestras presente en la análisis.

$$PC_1 = A_{11}X_1 + A_{12}X_2 + \ldots$$

$$PC_2 = A_{21}X_1 + A_{22}X_2 + \ldots \qquad \text{Ecuación 4}$$

Donde "X" son las variables originales y los coeficientes "A" son los valores de los loadings, o contribución de cada variable a cada componente. Cuando en la ecuación de cada componente se sustituye el valor que cada variable toma para cada muestra (**figura 5**), se obtienen las coordenadas de cada muestra (componentes principales) es decir, los scores [32].

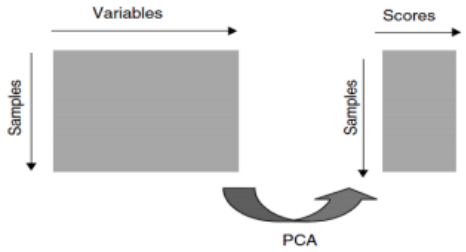

Figura 5, representación del PCA

No todos los componentes principales contienen la misma información. Los primeros componentes son los que describen la mayor variación en los datos, que se asocia a la información más relevante, mientras que los últimos describen variaciones en los datos que pueden ser debidas a ruido o error experimental, y pueden ser descartados, con lo que se consigue una importante reducción del número de variables.

Hay dos tipos de figuras utilizadas para representar los resultados de PCA que son los gráficos de scores y de loading. En el trabajo estudiado se ha utilizado principalmente el gráfico de scores donde se representan las muestras, normalmente los scores del PC1 frente a los PC2. Este gráfico permite observar las agrupaciones o la dispersión y tendencia de las muestras. La representación también nos permite detectar posibles outlier de las muestras que se alejan a otras [33].

5.4. TÉCNICA DE CLASIFICACIÓN

La técnica clasificatoria consiste en la construcción de modelos capaces de identificar la pertenecía de una muestra a una clase basado en las características de la muestra. Para construir el modelo es necesario disponer de una muestra cuya clase es conocida.
En el trabajo se ha utilizado dos tipos de técnicas de clasificación tales como el SIMCA, Soft Independent Modelling of Class Analogy y PLS-DA, Partial Least Squares Discriminant Analysis.

5.4.1. Soft Independent Modelling of Class Analogy (SIMCA)

SIMCA, se basa en construir un modelo de componentes principales (PCA) independiente para cada una de las clases. El número de componentes principales utilizados para cada clase puede ser diferente dependiendo de la varianza de los datos.
A continuación se muestra un ejemplo de la técnica SIMCA en **la figura 6** donde se presenten dos agrupaciones de muestras que son de clases A y B.
Estas clases se modelan obteniendo unas estructuras lineales (una recta, un plano etc.) dependiendo del número de componentes necesario para establecer el modelo de la clase. Al aplicar el modelo de SIMCA sobre la clase A se obtiene un modelo de

dos PCs que corresponde a un plano, mientras que la clase B se ajusta a un modelo de una única variables, una recta.

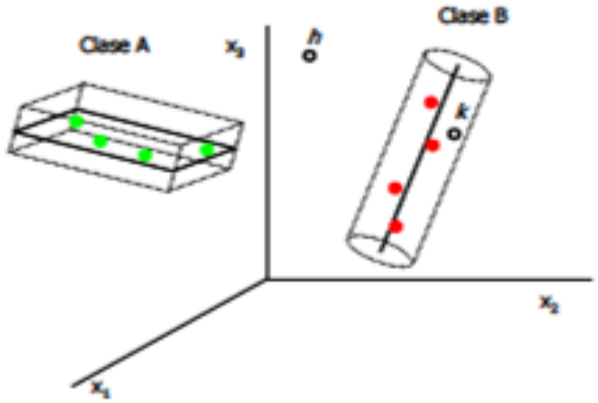

Figura 6, Modelo de SIMCA, clase A con 2 PC y clase A con 1 PC

Para asignar nuevos objetos, se hace comprobando si las muestras están localizadas dentro del modelo establecido. En este caso, la muestra "k" de la figura se puede asignar a la clase B mientras que "h" no pertenece a ninguna clase es un outlier [34, 35].

5.4.2. Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA consiste en establecer una regresión lineal entre una matriz de variables independientes (matriz X, en el trabajo corresponde a los datos de cromatograma de las muestras) y una matriz de variables dependientes (matriz Y). Y es una variable binaria que indica a que clase pertenece las muestras, donde 1 indica pertenencia a la clase y 0 no pertenencia. Dado que en este trabajo se quiere diferenciar entre 3 clases, las muestras de la clase 1 se indican (1, 0, 0), las de la clase 2 (0, 1, 0) y las de la clase 3(0, 0, 1).

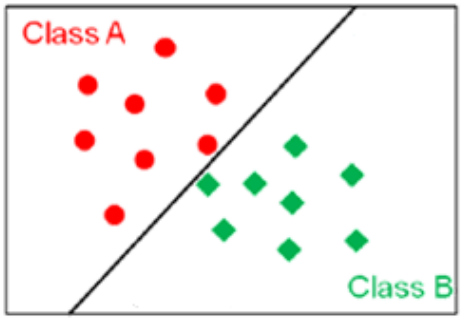

Figura 7, Modelo de PLS-DA para 2 clases A y B. El espacio queda dividido por 1 línea

PLS-DA, como un método discriminante marcan una separación entre clases, dividen el espacio en regiones separadas de las cuales corresponde a una sección de clase, como se puede ver en la **figura 7** [36, 37].

5.4.3. Validación

El método de validación utilizada en el trabajo es la validación cruzada (CV) y concretamente el leave-one-out validación cruzada (LOOCV). Consiste en la división en dos subconjuntos de muestras, uno contiene n-1 muestras y el otro subconjunto 1 muestra, donde "n" es el número total de las muestras. El primero se utiliza para construir el modelo con un determinado PCs o variables latentes en el caso de PLS-DA mientras que la muestra excluida es usada para la predicción. Este método es repetido hasta que todas las muestras sean dejadas fuera y el error calculado [38].

6. PARTE EXPERIMENTAL

El análisis quimiométrico se realizó sobre las cuatro regiones identificadas a diferentes tiempos de retenciones para observar aquella región que permite obtener mejor clasificación de las muestras según su origen geográfico.
Una vez definidas las regiones, con el fin de diferenciar y clasificar las 100 muestras de aceite de palma de orígenes geográficos diferentes, la metodología utilizada consistió:

- **Estudios**

- Análisis exploratorio

En este análisis se realizó una representación de los valores de scores con el conjunto de datos autoescalado, para los dos matrices, con el fin de ver posibles agrupaciones y muestras outliers entre las muestras. En está análisis se espera ver tres agrupaciones de muestras diferentes, ya que según los datos corresponde a tres orígenes geográficos diferentes: Asia, América y África.

- Estudio del Pretratamiento

El estudio del pretratamiento se realizó sobre el conjunto de datos obtenido en el cromatograma. La alineación de los cromatogramas, se realizó previamente por la Universidad de Granada. Los pretratamientos realizados son autoescalado, normalizado, centrado y suavizado. El estudio comparativo entre el pretratamiento se hizo mediante la técnica de clasificación PLS-DA y con seis variables latentes.

- Estudio de región del cromatograma

Una vez obtenido el pretratamiento óptimo, se establecieron modelos de clasificación seleccionando combinaciones de las cuatros regiones con objeto de identificar si mejoraban los valores de clasificación de las muestras. Para la región óptima, se realizó el estudio de las variables latentes.

- **Condiciones óptimas**

Finalmente, una vez elegidas las condiciones óptimas, se desarrollaron los modelos de clasificación SIMCA y PLS-DA. Los dos modelos se validaron mediante el LOOCV. El estadístico evaluado para la selección del número óptimo de PCs (SIMCA) y variables latentes, LVs, (PLS-DA) es la varianza explicada.

La comparación de los resultados obtenidos con SIMCA y PLS-DA se hizo en términos de: muestras verdaderas positivas (TP), falsas positivas (FP), falsas negativas (FN) y verdaderas negativas (TN) [39] y el último las muestras

inconclusive. Estos valores se calculan para cada una de las clases, según las siguientes expresiones (se han indicado para la clase 1):

Verdaderas positivas (TP): Son muestras de clase 1 asignadas correctamente en clase 1 (**ecuación 5)**.

$$\text{TP} = \frac{\text{nº muestras de la clase 1 SI asignadas a la clase 1 (TP)}}{\text{nº total de muestras de la clase 1 (total clase 1)}} \qquad \text{Ecuación 5}$$

Falsas positivas (FP): FP son muestras que no son de la clase 1, pero asignada erróneamente en la clase 1 (**ecuación 6)**.

$$\text{FP} = \frac{\text{nº muestras de la clase 2 y 3 SI asignadas a la clase 1(FP)}}{\text{total clase 2 y 3}} \qquad \text{Ecuación 6}$$

Falsas negativas (FN): Es la proporción de las muestras de la clase 1 que no son clasificadas como clase 1 **(ecuación 7)**.

$$\text{FN} = \frac{\text{nº muestras de la clase 1 SI asignadas a la clase 2 y 3 (FN)}}{\text{total clase 1}} \qquad \text{Ecuación 7}$$

Verdaderas negativas (TN): Es la proporción de las muestras que no son de la clase 1 y no asignada en la clase 1 **(ecuación 8)**.

$$\text{TN} = \frac{\text{nº muestras de la clase 2 y 3 NO asignadas a la clase 1(TN)}}{\text{total clase 2 y 3}} \qquad \text{Ecuación 8}$$

Inconclusive: Es una propuesta realizada en el trabajo, es la proporción de las muestras de clase 1 que son múltiples asignadas en la clase 1 y otras clases **(ecuación 9)**

$$\text{Inconclusive} = \frac{\text{nº muestras de clase 1, múltiples asignadas}}{\text{total clase 1}} \qquad \text{Ecuación 9}$$

En el modelo de clasificación SIMCA, como se ha comentado, se establece un modelo de componentes principales para cada clase. Una vez establecido el modelo, se define un valor umbral distancia al modelo (una frontera o límite para de cada clase) que habitualmente es un valor de 0.5. Para asignar una muestra, ésta se proyecta sobre el modelo de componentes principales de cada clase y si tiene un valor superior al del umbral, quedará asignada a la clase. Si la muestra no queda asignada ninguna clase, quiere decir que tiene una probabilidad menor que el umbral, o tiene una probabilidad superior a él para más de una clase [40].

En el modelo de clasificación de PLS-DA, se define un límite entre las clases. Como en este trabajo se han definido 3 clases, se han de definir tres límites: 1) uno entre la clase 1 y las clases 2 y 3 que corresponde a la matriz Y=(1,0,0); 2) Otro límite para la

clase 2 frente a la clase 1 y 3 matriz Y=(0,1,0) y finalmente 3) un límite para la clase 3 frente a la clase 1 y 2 matriz Y=(0,0,1). La muestra es asignada en la clase para la que cumple el límite definido (por ejemplo, a la clase 1, si cumple el límite de la clase 1). Dado que se definen 3 clases, las posibles asignaciones cuando se predice una muestra son: la muestra pertenece a la clase 1, a la clase 2, a la clase 3, no pertenece a ninguna de las 3 clases o pertenece a más de una clase (puede quedar simultáneamente asignada a la clase 1 y 2, a la clase 2 y 3, ó la clase 1 y 3, ó simultáneamente a las 3 clases).

7. RESULTADOS Y DISCUSSIONES

En primer lugar, se realizó una representación del cromatograma para visualizar todas las regiones posibles.

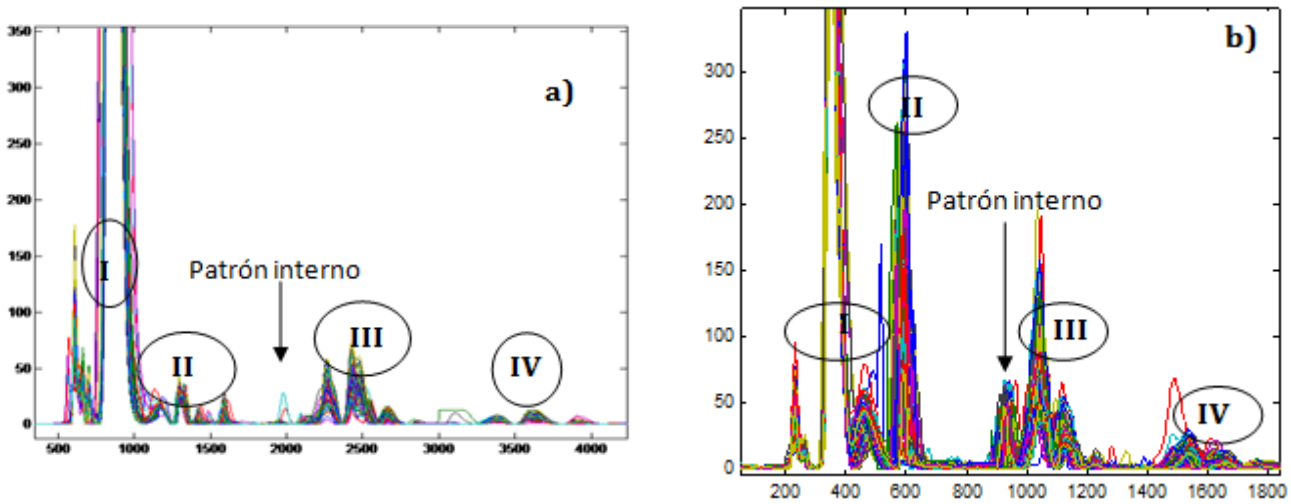

Figura 8, cromatogramas de todas las regiones analizado por a) HPLC-UV b) HPLC-CAD de las muestras del aceite de palma diferenciado por cuatro regiones

La **figura 8,** corresponde a la representación de los cromatogramas de las 100 muestras analizado por cromatografía líquida con los dos detectores (UV y CAD). El tiempo de análisis en ambos casos es aproximadamente 20 min.

Se pueden diferenciar cuatro regiones. La región I corresponde a diferentes compuestos de ácidos grasos presente en el aceite de palma, siendo el compuesto mayoritariamente del aceite eso se ve reflejado en el pico con mayor intensidad. La región II es característica de los esteroles que corresponde a una clase de fitoesteroles que es dimetilesterol. En cuanto a la región III, cuenta con dos clases de fitoesteroles que son desmetilesterol, metilesterol, y con otros compuestos como alcoholes grasos. Finalmente, la región IV está asociada a la presencia de alcoholes terpénico presenta en el aceite de palma. También se observa un pico que corresponde al pico de patrón interno que se encuentra entre región III Y IV en ambos cromatogramas [16]. Este pico se ha eliminado antes del establecimiento del modelo de clasificación ya que no aporta información sobre la muestra objeto de estudio.

Se han desarrollado los modelos de clasificación considerando las cuatro regiones y combinaciones de éstas: I, II, III y IV; II y III; III y IV; II, III y IV, donde el patrón interno no está incluido en todas las regiones.

7.1. Datos ultravioleta-visible (UV)

7.1.1. Estudio de condición óptima

- Análisis de principal componente(PCA)

A continuación, se realizó un estudio de PCA considerando las cuatro regiones I, II, III y IV. **La figura 9**, muestra la representación del PC1 vs PC2 para cada una de las 3 clases estudiadas (datos autoescalados).
Donde la **figura 9a** corresponde a las 56 muestras de continente Asia. En esta representación se expresa el 48.99% de la información original, en el cual el PC1 explica la mayor parte de las informaciones con el 31.21%.

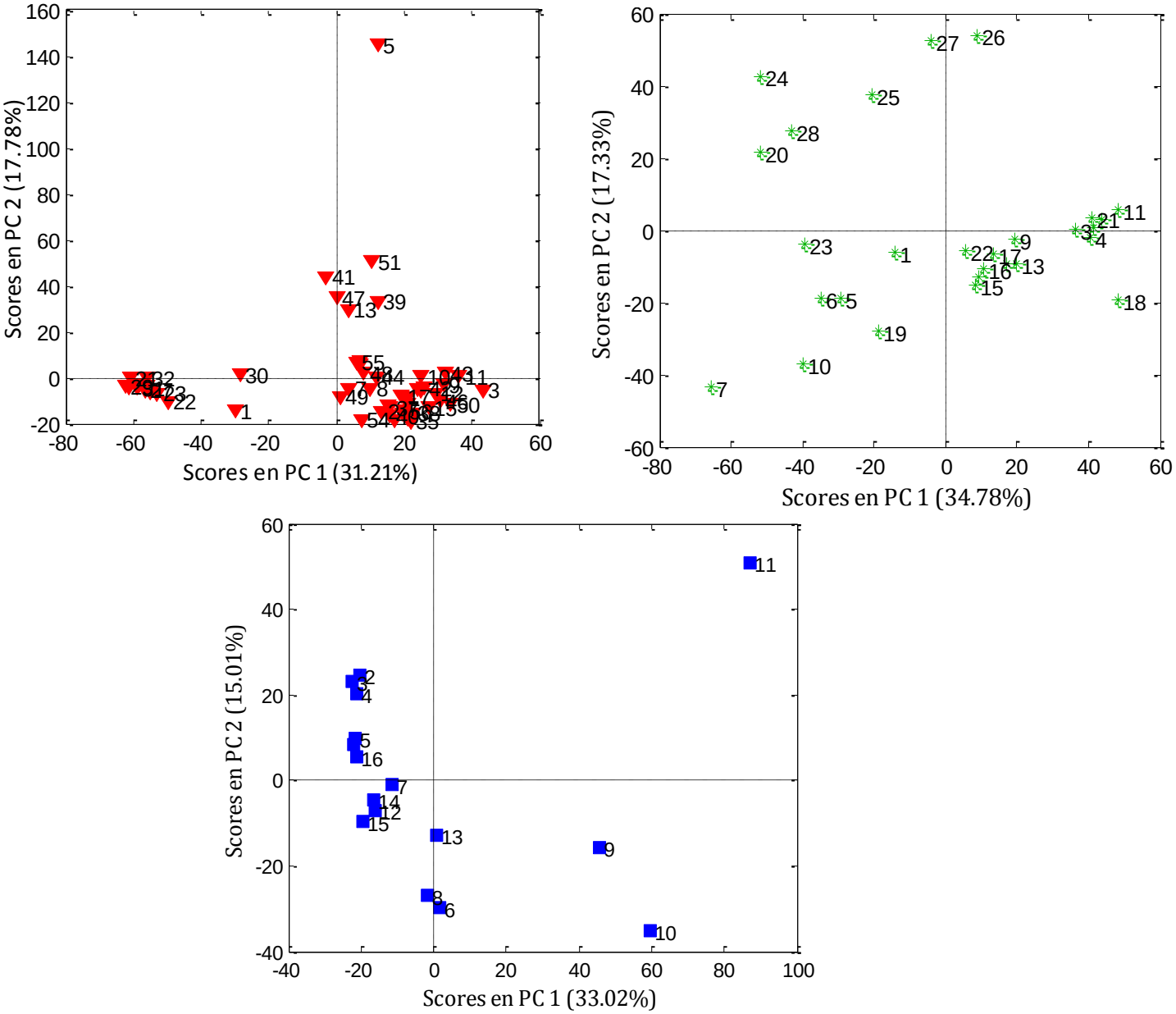

Figura 9, Gráfico de scores de PC1 vs PC2 **a)** Asia **b)** África **c)** América

En la figura 9a, se observan agrupaciones entre las muestras sobre el PC1 excepto una muestra (muestra 5) que tiene un comportamiento diferente al resto. Por lo tanto, esta muestra podría ser considerada sospechosa/outlier. También se puede observar la presencia de tres subgrupos de muestras, uno de ellos se encuentra a valores positivos de PC1, y las demás a valores negativos de PC1 y positivos de PC2. Estos subgrupos pueden ser debidos a que dentro de la categoría Asia, existen diferentes procedencias (tabla 1), como India, Indonesia y Malasia finalmente Papúa. N. Guinea y Salomón, pero al no disponer de esta información no podemos establecer los grupos.

La **figura 9b)** representa las 28 muestras del origen geográfico de África, se puede observar que las muestras están dispersas sobre los dos PCs, donde principalmente el PC1, explica la mayor variabilidad de las informaciones (34,78%). En cuanto a la **figura 9c)** corresponde a las 16 muestras de América con una varianza explicada, 33.02% de la información, en esta representación también se puede observar una muestra (muestra 11) que se aleja de los grupos, esta muestra se puede asociar como muestra discrepante u outlier.

Tanto la muestra 5 de Asia y la muestra 11 de América, se consideraban como dos muestras sospechosas en los datos. La presencia de estas muestras puede ser causada tanto por error experimental como por la química. De están manera en el análisis posterior se realizó un estudio de la influencia de estas muestras discrepantes.

- Estudio de muestras discrepantes

Todo el estudio realizado en el trabajo se hizo mediante el modelo de PLS-DA, con el método de LOOCV. Para la elección del variables latentes (LVs), las opciones razonables incluirían 5, 6 ó 7 que son los LV's que permite obtener menos error posible para la clasificación de las muestras. De esta manera, para los estudios comparativos, se ha fijado seis como el LV.
En este apartado lo que se quiere conseguir es ver la influencia tanto con o sin la presencia de las muestras considerado anteriormente discrepantes o sospechosas. La **tabla 2** muestra un resumen de los resultados de los porcentajes de clasificación obtenido de las muestras discrepantes con los datos de la región I, II, III, IV después de aplicar el pretratamiento autoescalado.

Tabla 2. Porcentajes de asignación correcta en cada clase

Muestras	Clase 1	Clase 2	Clase 3
Todas las muestras	88	71	81
Muestras sin nº 5 (Asía)	91	75	81
Muestras sin nº 11 (América)	89	71	73
Muestras sin nº 5 y 11	89	71	80

En base a los resultados, las dos muestras (muestras de nº 5 y 11) fueron conservadas ya que no se observa una variación significante en los porcentajes de clasificación. La presencia o ausencia de estás muestras no modifican la clasificación de las muestras. De esta manera, el análisis posterior del estudio de los pretratamientos se realizó con todas las muestras.

- Elección de pretratamiento

En este apartado, consiste en evaluar diferentes pretratamientos sobre la región I, II, III y IV, con el fin de elegir el pretratamiento con clasificación máxima. Se aplicaron el centrado, autoescalado, normalizado y suavizado + autoescalado.

Tabla 3. Porcentajes de análisis de pretratamiento sobre la región I, II, III, y IV

Pretratamiento	Clase 1	Clase 2	Clase 3
Autoescalado	88	71	81
Centrado	78	75	50
Normalizado	75	57	50
Suavizado + autoescalado	88	71	75

El autoescalado (**tabla 3**) es el pretratamiento que da mejores porcentajes de clasificación para las tres clases. También se puede ver que con el suavizado+autoescalado, permite obtener una buena clasificación. No obstante, sólo el autoescalado es suficiente y es el que se ha elegido como óptimo.

- Elección de región y variables latentes

Una vez definido el pretratamiento óptimo el autoescalado, en esta etapa del trabajo, se pasa al estudio de la región óptima. Para ello se aplicó el pretratamiento autoescalado para cada región (región I, II, III y IV; II y III; III y IV; II, III y IV) y mediante la técnica de clasificación de PLS-DA permite obtener una clasificación como se puede ver en la **tabla 4.**

En base de los resultados, en todas las regiones, se puede diferenciar el origen geográfico de las muestras para la clase 1 (Asia) y la clase 2 (África) en cambio para la clase 3 (América), la región I, II, III y IV es la única región que permite obtener una clara diferenciación. Este comportamiento de las muestras de la clase 3 (América), con menor porcentaje de clasificación puede ser debido a que las muestras no son representativas (pocas muestras).

Teniendo en cuenta esto, la región óptima es la región I, II, III y IV. Esto indica que no solo los esteroles son marcadores moleculares para diferenciar claramente el origen geográfico de aceite palma sino también son necesarios otros componentes que se encuentran entre las regiones I y IV.

Tabla 4. Porcentajes de clasificación de las regiones

Regiones	Clase 1	Clase 2	Clase 3
I, II, III y IV	88	71	81
II, III y IV	91	61	56
III y IV	89	71	56
II y III	86	71	50

Un parámetro importante a ser definido es el número de variables latentes (LV's) para establecer el modelo. Para saber cuántos son necesarios, se realizó un estudio de la varianza explicada en función de los PC's (**figura 10**). El objetivo de este análisis consiste en identificar el menor PCs que explique la mayor parte de la varianza. A partir del gráfico se puede observar que hasta siete PCs se considera como los valores de PCs significativas, pero a partir de ocho PCs no se puede observar caída significante y diferenciar las variancias. La variabilidad a partir del ocho PCs puede ser debido al error experimental, error instrumental por el ruido que presente la señal.

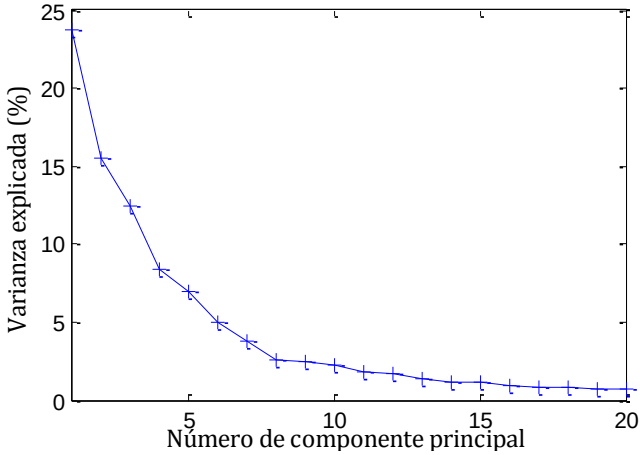

Figura 10. Gráfico de PCs vs varianza explicada

En la **tabla 5**, se muestra el resultado del porcentaje de clasificación variando el número de LV's (entre cinco y siete). Un valor mayor de LV's causa una posibilidad de error de sobreajuste de los modelos. Por lo contrario, un valor menor LV's, puede dejar fuera aquellas informaciones relevantes para establecer el modelo.

En este estudio, se puede ver que con seis LV's permite diferenciar las tres clases correctamente con una varianza sobre X de 62.62% y varianza sobre la Y de 36.07%.

Tabla 5. Porcentajes de clasificación en LVs

Variables latentes	Clase 1	Clase 2	Clase 3
5	88	64	63
6	88	71	81
7	88	71	75

7.1.2. Establecimiento del modelo de clasificación óptimo

Las condiciones óptimas son región I, II, III y IV con datos autoescalado y seis LV's. Dicho esto, el estudio de clasificación se realizó sobre esta condición.

Previo al establecimiento del modelo, se ha realizado un gráfico de scores con las 3 clases (**La figura 11**), donde los datos "1" corresponde a las muestras de continente Asía, los datos "2" corresponde a África y finalmente los datos "3" a América. El PC1 tiene mayor importancia con un 23.79 % respecto el PC2 (15.47%).

Aunque se observa un gran solapamiento de las 3 clases, sobre el PC1 se pueden diferenciar las muestras de América (color azul) que presentan mayoritariamente valores de scores negativos de PC1.

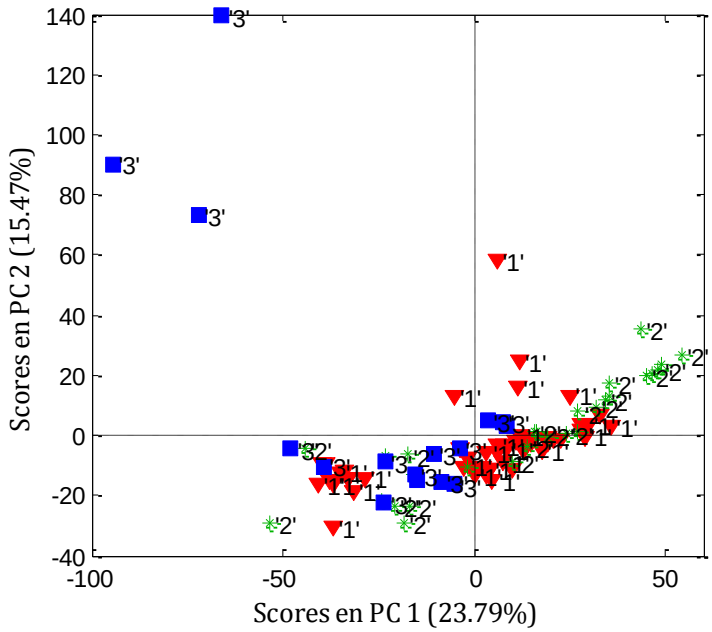

Figura 11. Gráfico de scores PC1 vs PC2

El otro grupo es Asia (color rojo) en el cual, las muestras se encuentran mayoritariamente en valores positivos de PC1 y un subgrupo compacto con valores negativos. Las muestras que provienen de África (color verde) tienen un comportamiento diferente en comparación con los otros dos continentes. Las muestras de África están más dispersas a lo largo del PC1 (con valores positivos y negativos) y no se ve ninguna agrupación entre ellos, y se ven muy solapados con las muestras de Asia (color rojo) y América (color azul).

En cuanto PC2, no se puede observar ninguna tendencia entre las muestras sobre el origen geográfico, pero se puede identificar agrupaciones de tres muestras de América (azul), donde tiene mayor influencia sobre valores positivos de este PC. Una de esta muestra (muestra 94) corresponde a la muestra considerado como sospechosa en el estudio realizado anterior de manera individual de las clases sobre los PCs. Dicho esto para análisis comparativa entre las técnicas, se conserva las tres muestras

- Comparación y evaluación entre las técnicas: PLSDA y SIMCA

A continuación, se presentan los resultados de clasificación usando las dos técnicas de clasificación que son SIMCA y PLS-DA utilizadas en el trabajo (**tabla 6**). Se presentan las muestras asignadas y las no asignadas para cada una de las tres clases. En el modelo SIMCA, se han seleccionado como número de PCs 6, 6, y 7 para la clase 1, clase 2 y clase 3 respectivamente. La mayor parte de los errores son por múltiple asignación en todas las clases, siendo mayoritarios en la clase 1 y de la 2. Los porcentajes de asignación correcta son muy bajos (inferiores al 50%) para las clases 1 y 2, y aceptable para la clase 3 (75%). Esto indica que SIMCA no es una técnica de clasificación válida para la autentificación del origen geográfico de las muestras de aceite de palma analizadas por la técnica cromatográfica indicada.

Con el modelo PLS-DA, en general se puede decir que se obtienen buenos porcentajes de clasificación para las 3 clases estudiadas ya que las muestras están mayoritariamente asignadas en su propia clase. Hay un porcentaje pequeño (7%) de muestras de la clase 1 y de la 2 que presentan doble asignación, y hay una muestra de la clase 2, que no es asignada en ninguna clase. De la clase 3, hay 3 muestras mal asignadas lo que representa un 18%, pero no hay dobles asignaciones o no asignaciones.

Tabla 6. Número de muestras asignadas y no asignadas en las tres clases

	Clase Real	PLSDA		Clase Real	SIMCA[a]	
Predicho	**Clase 1**	**Clase 2**	**Clase 3**	**Clase 1**	**Clase 2**	**Clase 3**
Clase 1	49	3	1	11	0	0
Clase 2	3	20	2	0	12	0
Clase 3	0	2	13	0	0	12
Múltiples	4	2	0	40	16	4
No asignadas	0	1	0	5	0	0
Total	56	28	16	56	28	16

[a]: Número de PCs para establecer el modelo SIMCA: clase 1(6), clase 2(6) clase (7)

De esta manera, el orden de clasificación correcta sigue está tendencia en base de modelo de PLS-DA, clase 1 (88%) > clase 3 (81%) > clase 2 (71%). Sigue está tendencia ya que hay muestras de las clases 2 que son múltiples y mal asignadas. Por otra parte, debido a que las muestras del aceite de palma tienen su originaria en África por la cual las muestras están dispersas por el espacio sobre el PCs, lo que impide obtener una buena clasificación respecto otras clases.
De los resultados mostrados en la **tabla 7** se destacan las siguientes observaciones por clase. En el caso de clase 1 (Asia), las muestras consideradas verdaderas positivas y falsas positivas son más altas en comparación con las otras clases.
No obstante, las verdaderas y falsas negativas, para dicha clase presentan menos muestras. En cuanto a la clase 2 (África), presenta mayor muestras verdaderas negativas en comparación con las muestras verdaderas positivas.

Tabla 7. Porcentaje de muestras clasificadas

	Clase 1	Clase 2	Clase 3
TP	88	71	81
TN	84	88	89
FP	9	7	2
FN	7	21	19
Inconclusive	5	7	0

Similarmente como ocurre en muestras de la clase 3 (América). El interés de estos parámetros es obtener valores elevados de verdaderos positivos y verdaderos negativos y valores bajos de falso positivo y falso negativo.

Como queda reflejado en la tabla 7, también se ha definido un parámetro denominado inconclusive. Corresponde a las muestras que presentan múltiple asignación y/o que no han sido asignadas. Tal y como se observa en los resultados de la tabla 6, la clase 1 y 2, son las únicas muestras inconclusives con un porcentaje de 7%. Mediante este parámetro también se puede afirmar que PLS-DA es una técnica válida ya que no presenta muchas muestras inconclusives.

Para estudiar las muestras dobles asignadas, se realizó un estudio más detallado de la clasificación de las muestras en cada una de las tres clases.

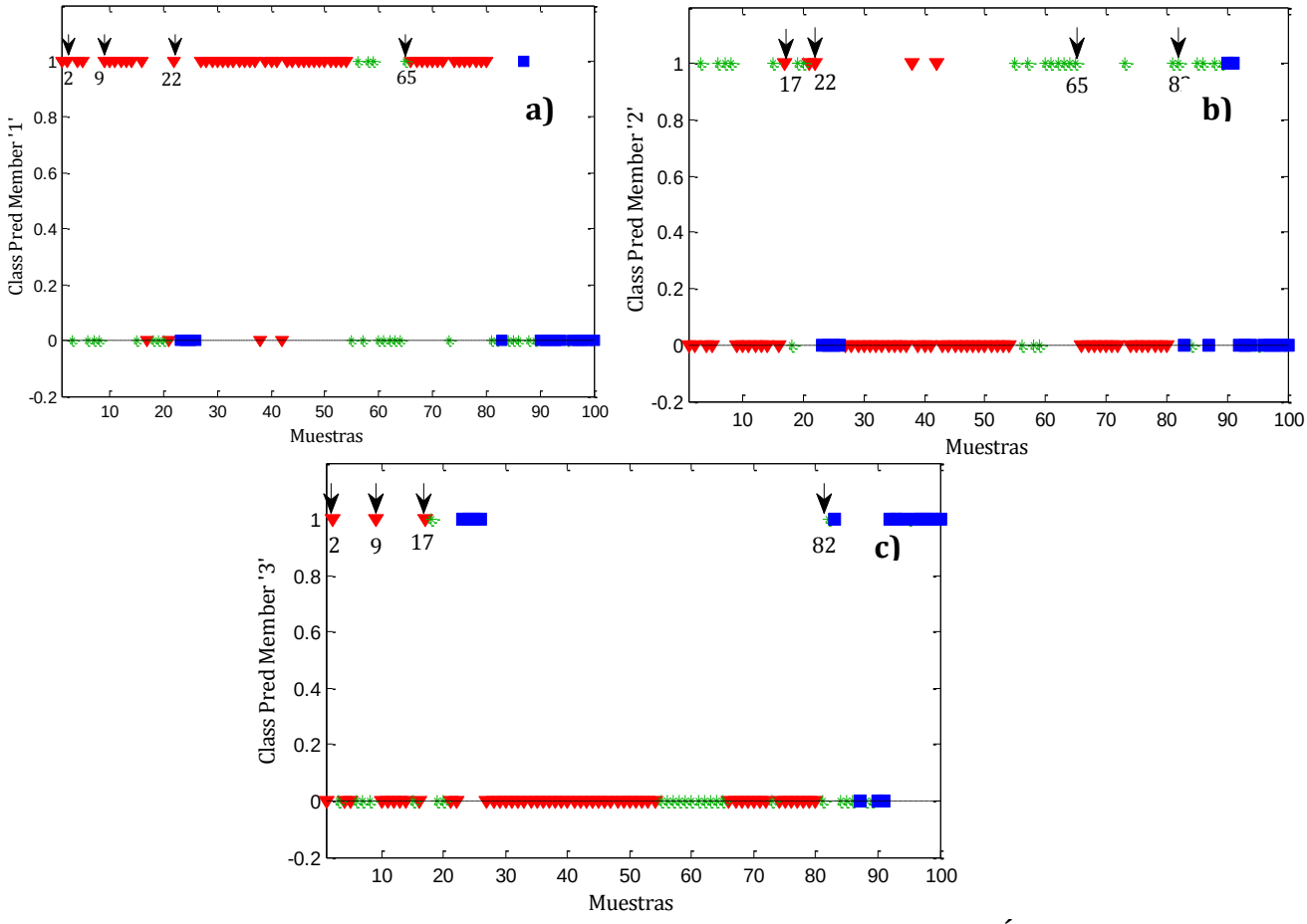

Figura 12. PLS-DA asignaciones **a)** Clase 1(Asía, rojo) **b)** Clase 2(África, verde) **c)** Clase 3 (América, azul)

La **figura 12** es una representación de la predicción de las muestras en el intervalo de clasificación entre 0 y 1. Donde las muestras que se tienen en valor "0", corresponden a las que no se ajustan al modelo de la clase considerada, y las muestras con el valor "1" indica que si cumplen el modelo. Con esta representación se puede identificar las muestras que son doble asignadas y las de su propia clase. Como indica la tabla 6, hay seis muestras que son múltiples asignadas, cuatro de ellas corresponden a la de clase 1 y dos de la clase 2. Las cuatro muestras de clase 1, son la 2, 9, 17 y 22. Estas muestras coincidan con las indicadas en la figura 12. En el cual las muestras 2 y 9, están doble asignadas en clase uno y tres. En cambio, la muestra 17 están asignadas en la clase 2 y clase 3 y la muestra 22, en la clase 1 y clase 2. Finalmente, las dos muestras múltiples asignadas de clase 2 son la 65 y 82. En el caso de muestra 65 está asignada tanto en clase 1 y 2, y muestra 82, en clase 2 y 3.

Mediante están representación se puede ver claramente las muestras que son múltiples asignadas y en que clase están asignadas. Estas muestras doble asignada indica que una muestra de un continente, puede tener unos características de otras continentes. Como por ejemplo en el caso de muestras 82, que está doble asignada en clase dos y tres Eso quiere decir, que esta muestra tiene alguna característica común de África y América.

7.2. Análisis con datos de detector de aerosol cargado (CAD)

7.2.1. Análisis de principal componentes

La metodología utilizado con los datos obtenido del detector de carga aerosol, también fue la misma como en el caso de datos con ultravioleta-visible.

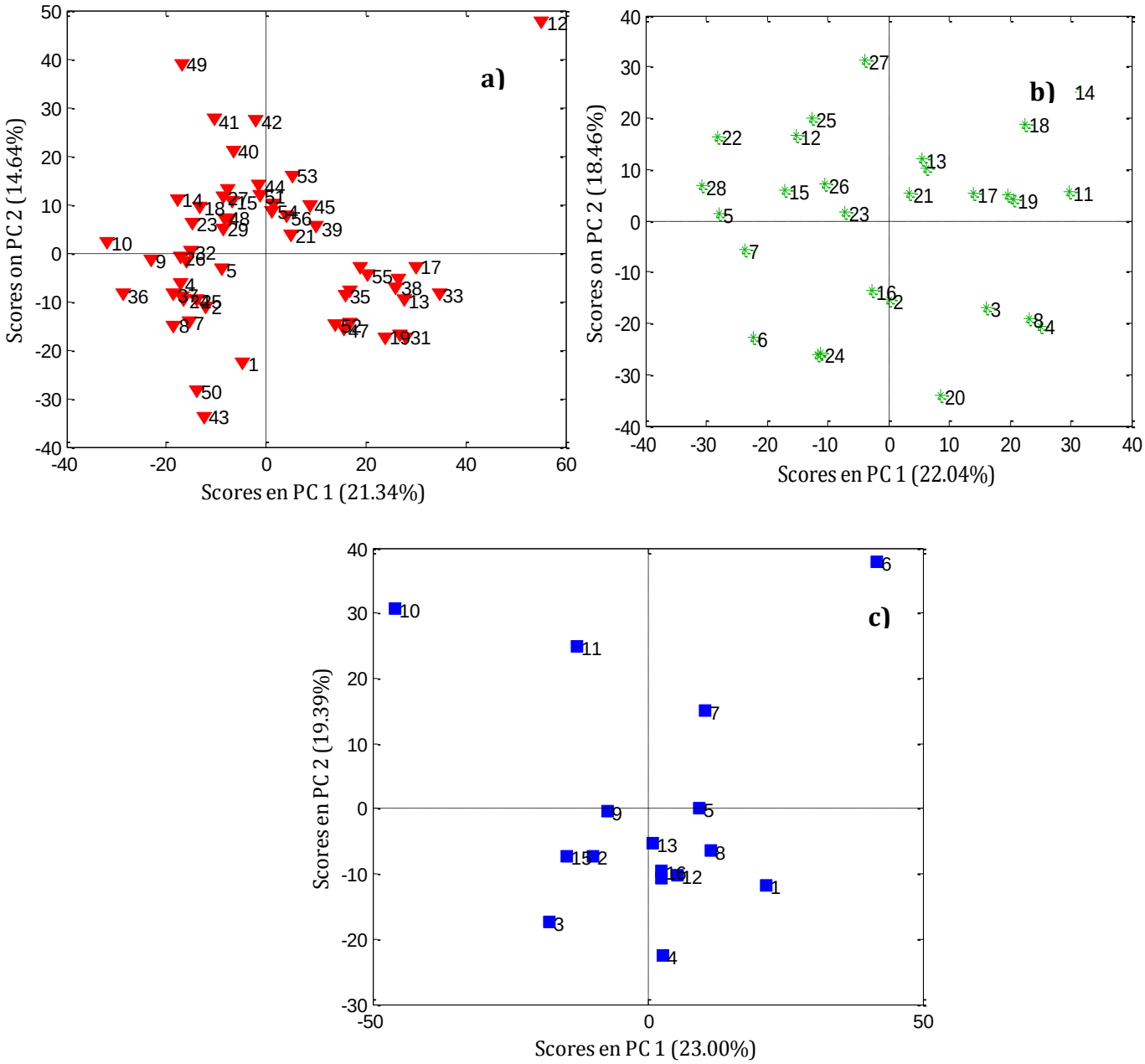

Figura 13, Gráfico de scores de PC1 vs PC2 **a)** Asia **b)** África **c)** América

En primer lugar se estudio el comportamiento de las muestras de las diferentes clases de manera individual sobre los gráficos de scores. Este estudio se realizó con 56 muestras de Asia, 28 de África y 16 de América. Como se puede observar en la **figura 13**, las muestras muestran una tendencia similar como en el caso de los datos de UV.
En todos casos, la varianza explicada es aproximadamente 40% de la información original, donde el PC1 explica la información mayoritaria. En los gráficos de scores, se pueden observar agrupaciones entre las muestras en los diferentes clases excepto la clase 1 y clase 3 donde tienen dos muestras, muestras 12 y 6 respectivamente muy separadas del resto. Como en datos UV, en cuanto al gráfico de **13a)** Asía, también se observa tres subgrupos. Estos tres subgrupos no se pueden asociar a ningún país del continente ya que en el trabajo no se dispone de esa información.
Posteriormente se analizó la presencia de las muestras discrepantes mediante el modelo de PLS-DA, para observar su influencia sobre la clasificación de las muestras en función de su origen geográfico.

- Estudio de muestras discrepantes

Tabla 8. Porcentajes de análisis de muestras discrepantes sobre la región I, II, III, y

Muestras	Clase 1	Clase 2	Clase 3
Muestras con todas	95	71	63
Muestras sin nº 12 (Asía)	96	82	69
Muestras sin nº 6 (América)	95	75	67
Muestras sin nº 12 y 6	96	86	73

Después de haber estudiando las muestras, tal como se puede ver en la **tabla 8** tanto la presencia y ausencia de las dos muestras discrepantes, en el resultado se ve que la ausencia de las dos muestras permite obtener una clasificación mayores respectos las demás clases. Principalmente se observar una clara influencia de la muestra 6 en conjunto de las muestras discrepantes en la clasificación. Lo que indica que la presencia de la muestra 6 no permite ver correctamente la clasificación de su propia clase (clase 3) y otra clase (Clase 2). Por este motivo, el análisis posterior se realizó sin las dos muestras discrepantes.

- Elección de Pretratamiento

De la misma manera como en los datos de UV, se realizó un estudio de pretratamiento óptimo sobre la región I, II, III, y IV sin muestras discrepantes (muestras nº 12 y 6). Para ello se realizó los siguientes pretratamiento: Centrado, autoescalado, normalizado y suavizado + autoescalado.

Tabla 9. Porcentajes de análisis de diferentes pretratamiento de región I, II, III, y IV

Pretratamientos	Clase 1	Clase 2	Clase 3
Autoescalado	96	86	73
Centrado	87	68	60
Normalizado	85	68	67
Suavizado+ autoescalado	96	82	67

La **tabla 9** muestra los resultados obtenido con la técnica de PLS-DA. Se puede ver que tras realizar los diferentes, pretratamientos en el modelo de PLS-DA, la clasificación entre las muestras es mejor con el autoescalado. Por este motivo, para el estudio de la elección de la región óptima se utilizó el autoescalado con los datos sin outlier ya que nos permite obtener mayor clasificación de las muestras por el origen geográfico elevado.

- Elección de la región y variables latentes

Una vez definida la región con el pretratamiento con mayor clasificación. Se procede al estudio sobre las regiónes. La elección de la región, se realizó sobre los datos autoescalado en los diferentes cuatros regiones para observar aquella región que permite obtener una habilidad de clasificación ideal.

Tabla 10. Porcentajes de clasificación de las regiones

Regiones	Clase 1	Clase 2	Clase 3
I, II, III y IV	96	86	73
II, III y IV	93	75	73
III y IV	91	71	53
II y III	85	71	67

Como se puede ver en la **tabla 10**, la región I, II, III y IV muestra una buena clasificación de igual manera como en datos de UV.
Finalmente, para elegir el número de componentes principales se representa la varianza explicada en función del número de PC´s (**Figura 14**) siendo el número óptimo aquel a partir del cual no se observa una variabilidad significativa.

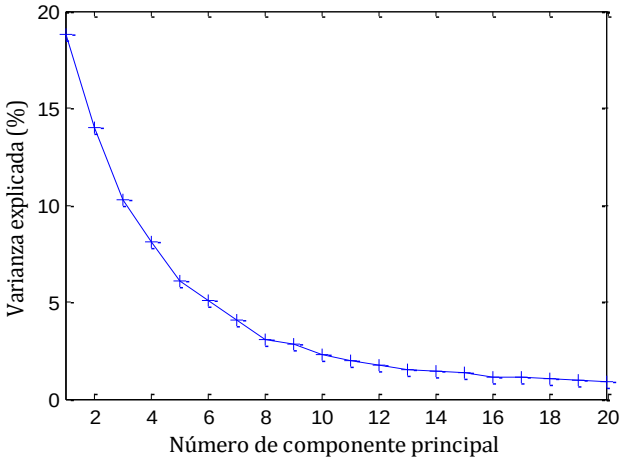

Figura 14. Gráfica de PCs vs varianza explicada

Como no se observa un punto de inflexión claro, se han calculado los porcentajes de clasificación con 5, 6 y 7 PC´s considerado como valores de 5, 6 y 7 LV´s.
Los resultados se muestran en la **tabla 11**. Como se puede observar, los mejores valores se obtienen con 6LVs

Tabla 11. Porcentajes de clasificación

Variables latentes	Clase 1	Clase 2	Clase 3
5	93	64	67
6	96	86	72
7	98	89	67

7.2.2. Análisis quimiométrica para la condición óptima

Una vez estudiado las condiciones óptimas que son región I, II, II y IV sin las muestras discrepantes con datos autoescalado y seis variables latentes. Con esta condición, se realizó la clasificación del aceite de palma.

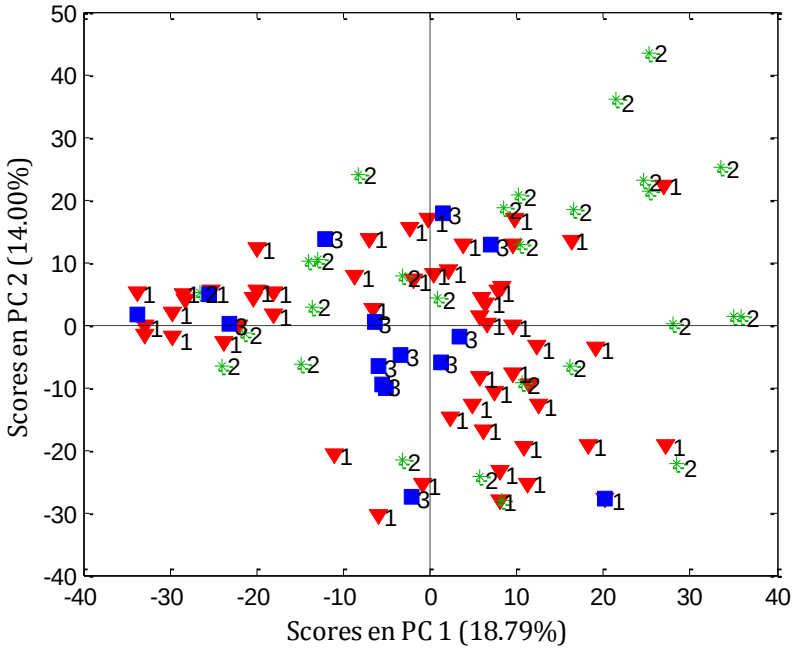

Figura 15. Gráfica de scores de PC1 vs PC2

De la misma manera como en los datos de UV se realizó un análisis más profundo de PCs. La **figura 15,** muestra el gráfico de scores resultante al utilizar los datos que corresponden a la región I, II, III y IV sin muestras discrepantes autoescalado. Con los dos primeros componentes se explicó un 32.79% de la varianza de todas las muestras donde el PC1 tiene el mayor peso, el 18.79% siendo el componente que permite obtener una discriminación entre las muestras de aceite de palma de

diferente origen geográfico. En cambio, por el PC2, no se muestra ninguna tendencia. Cosa que pasa con los datos de UV.

También se obtuvo una agrupación similar a la obtenida utilizando los datos de UV. En el cual se observa una separación de dos grupos de muestras, por una parte América por valores mayoritariamente negativa (clase 3, color azul) de PC1, y por otra parte los de Asia por valores mayoritariamente positiva de PC1 (clase 1, rojo). Sin embargo, las muestras de África (clase 2) se encuentran muy dispersas y no se logra una clara diferenciación de estos grupos.

- Comparación y evaluación entre las técnicas: PLSDA y SIMCA

De la misma manera como en los datos de UV, para estudiar las muestras entre las clases, la **tabla 12** corresponde a los datos autoescalado de la región I, II, III, y IV, para la identificación del origen geográfica de las muestras de aceite de palma para los modelos de PLS-DA y SIMCA. Como observado para datos de UV, el modelo de SIMCA muestra gran clasificación para múltiples asignadas, así que para la clasificación de las muestras de aceite de palma la técnica SIMCA no es adecuado.

Para el modelo de PLS-DA, en la clase 1, la mayoría de las muestras están asignadas en la clase 1 y no presenta muestras múltiples asignadas. El 7% y 20% de las muestras de la clase 2 y clase 3 respectivamente fueron clasificadas como múltiples y mal asignadas. Para la clase 2, presenta tantas muestras mal y múltiples asignadas, pero la mayoría de las muestras de clase 2 están asignadas en su propia clase. La clase 3 sigue la misma tendencia como las muestras de clase 2, pero en este caso no hay muestras mal asignadas.

Tabla 12. Número de muestras asignadas y no asignadas en las tres Clases

	Clase Real	**PLS-DA**		**SIMCA[a]**		
Predicho	**Clase 1**	**Clase 2**	**Clase 3**	**Clase 1**	**Clase 2**	**Clase 3**
Clase 1	53	0	0	15	0	0
Clase 2	0	24	0	0	20	0
Clase 3	1	1	11	1	0	6
Múltiples	0	2	3	37	8	9
No asignadas	1	1	1	2	0	0
Total	55	28	15	55	28	15

[a]: Número de pcs para el modelo SIMCA: clase 1(8), clase 2(8) clase (7)

Comparando las clasificaciones obtenido por la técnica de PLS-DA de los datos de UV con los datos de CAD, se puede decir que en datos de UV la clase 3 no presenta muestras múltiples asignadas y en datos CAD la clase 1. Pero en ambos casos las múltiples asignadas para la clase 2.

Tabla 13. Porcentaje de habilidad de clasificación

	Clase 1	Clase 2	Clase 3
TP	96	86	73
TN	84	93	93
FP	0	0	2
FN	2	4	0
Inconclusive	0	7	20

La siguiente **tabla 13,** muestra otras estadísticas para la clasificación en el cual se destacan las siguientes observaciones por clase: en el caso de clase 1, el porcentaje mayor de las muestras verdaderas positivas corresponden a la clase 1, con un 96%, seguido por la clase 2, con 86%. Sin embargo este último presenta también la mayor clasificación de falsas positivas, con una proporción de 17%.
Comparando los resultados obtenido en datos (CAD y UV), el CAD permite obtener el porcentaje de muestras falsas positivas bajos respectos UV. También se podría decir que en ambos datos presenta el porcentaje de muestras inconclusive.

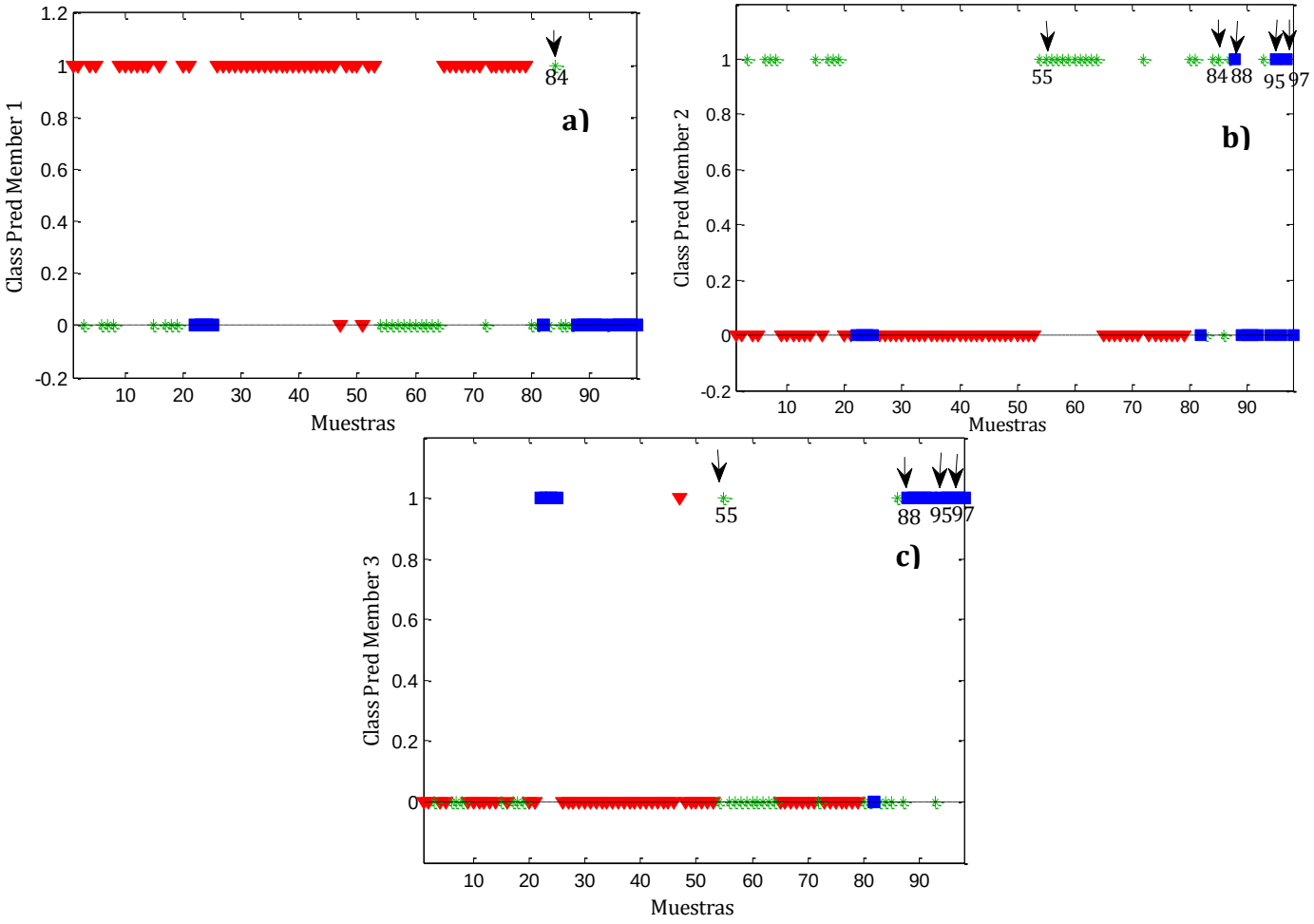

Figura 16. PLS-DA asignaciones **a)** Clase 1(Asía, rojo) **b)** Clase 2(África, verde) **c)** Clase 3 (América, azul)

Como indica la **tabla 12,** las muestras doble asignadas son dos de la clase 2 y tres de la clase 3. En cuanta a la muestra de la clase 2 que corresponde al África, representa la misma desempeña tanto en dato UV y CAD, que es lo esperable por ser el originario del aceite de palma. Por el análisis de estas muestra se hizo una representación de la **figura 16,** en ello se puede ver que las dos muestras doble asignadas de la clase 2 son muestras 84 y 55. Donde la muestra 84, pertenece a clase 1 y 2. En el caso de muestra 55 para la clase 2 y 3.

En cuanto a las tres muestras de clase tres que son 88, 95 y 97. Las tres muestras están doble asignadas en la clase 2 y 3.

8. CONCLUSIÓN

In this project, a multivariate method has been developed for the classification and differentiation of palm oil samples according to their geographic origin. Classification method has been implemented using software matlab.

The chromatograms using two different detectors, CAD and UV, has been used to build the classification models by means of two classification techniques: SIMCA and PLS-DA. PCA has been used to check sample distribution among the three pre-defined classes. In this analysis, we can see that samples from Africa has a different behaviour respect to the other two continents (Asia and America) in which is overlapped with the other continents, this behaviour occurs because the original samples of palm oil is Africa. These two continents, Asia and America, revealed a tendency to separate so therefore we can apply the technique classification. In addition, PCA has allowed outlier detection when working with CAD detector.

The study of the optimal conditions has been carried out where the best preprocessing is auto-scaling and the best results were obtained working with all four chromatographic regions (bands of the chromatogram) region I, II, III and IV.

PLS-DA provides the best result for the classification, while with SIMCA it has not been possible to establish reliable classification models. Comparing the result of model PLS-DA with cross-validation optimized by leave-one-out method, reliable classification values has been obtained for both types of detector (CAD and UV). CAD show better classification results: percentages values of correct assignation of 96%, 86%, 73% from Asia, Africa and America, respectively while the values obtained from the UV were 88%, 71% and 81%, also for Asia, Africa and America, respectively.

For this reason, it can be stated, that this chromatographic method associate with chemiometric fingerprinting provides a rapid tool for palm oil classification according to geographical origin and could serve as a technique to verify the labelling compliance of the oil.

9. BIBLIOGRAFÍA

[1] Gran velada. http://www.granvelada.com/es/donde-comprar-aceites-mantecas-para-jabon-cosmetica/14-aceite-de-palma-natural.html?gclid=CNPo1YiBu8kCFRQTGwodWmAN6w (fecha de consulta 29, octubre 2015). Aceite de palma natural

[2] Killman, W.; Non-forest tree plantations. ftp://ftp.fao.org/docrep/fao/006/ac126e/ac126e00.pdf (fecha de consulta 29, octubre 2015). The African oil palm.

[3] Botánica online. http://www.botanical-online.com/aceite_de_palma_propiedades.htm (Fecha de consulta 29, octubre 2015), Propiedades de aceite de palma, Beneficios de aceite de palma

[4] Esmiol, S.; Amigos de la Tierra. http://www.tierra.org/spip/IMG/pdf/Aceite_de_Palma.pdf (Fecha de consulta 30, octubre 2015), Aceite de palma: usos, orígenes e impactos.

[5] Unipalma S.A.; http://www.unipalma.com/aceite-de-palma (Fecha de consulta 6, Noviembre 2015), Aceite de palma, composición general

[6] Castaño, P.E.; Aplicabilidad del perfil de esteroles para la cuantificación de aceite de oliva en alimentos. [En línea] **2012,** pp 15, http://hera.ugr.es/tesisugr/21607862.pdf (Fecha de consulta 6, Noviembre 2015)

[7] Homapour, M.; Ghavami, M.; Piravi-Vanak, Z.; Hosseini, S. E.; Chemical properties of virgin olive oil from Iranian cultivars grown in the Fadak and Gilvan regions. *Grasas aceites.* [En línea] **2014**, *65,* doi:10.3989/gya.0351141. http://grasasyaceites.revistas.csic.es/index.php/grasasyaceites/article/viewArticle/1508 (fecha de consulta 9, Noviembre 2015)

[8] Lerma-García, M. J.; Ramis-Ramos, G.; Herrero-Martínez , J.M.; Simó-Alfonso, E.F.; Classification of vegetable oils according to their botanical origin using sterol profiles established by direct infusion mass spectrometry. *Rapid Commun. Mass Spectrom.* [En línea] **2008**, *22,* doi:10.1002/rcm.3459. http://www.ncbi.nlm.nih.gov/pubmed/18320541 (fecha de consulta 12, Noviembre 2015)

[9] Mata-Espinosa, P.; Bosque-Sendra, J.M.; Bro, R.; Cuadros-Rodríguez, L.; Olive oil quantification of edible vegetable oil blends using triacylglycerols chromatographic fingerprints and chemometric tools. [En linea] **2011,** *85*, pp 177-182. http://ac.els-cdn.com/S0039914011002542/1-s2.0-S0039914011002542-main.pdf?_tid=d647ade8-a6ae-11e5-9362-

00000aacb362&acdnat=1450570735_d8fea2a9e0f7fffe8d868f55817a18b5 (fecha de consulta 13, Noviembre 2015)

[10] Gutiérrez, J. Aplicación de Métodos Quimiométricos para la Caracterización y Control de Calidad de Plantas Medicinales. [En línea] **2012,** pp 16. http://grupsderecerca.uab.cat/chemometrics/sites/grupsderecerca.uab.cat.chemometrics/files/Tesis%20JRLG.pdf (fecha de consulta 16, Noviembre 2015)

[11] Ruiz-Samblás, C.; Arrebola-Pascual,C.; Tres, A.; Ruth, S.; Cuadros-Rodríguez, L.; Authentication of geographical origin of palm oil by chromatographic fingerprinting of triacylglycerols and partial least square-discriminant analysis. [En línea] **2013,** *116*, pp788-793. http://www.sciencedirect.com/science/article/pii/S0039914013006279 (fecha de consulta 17, Noviembre 2015)

[12] Oliveri, P.; López, MI.; Casolino, MC.; Ruisánchez, I.; Callao, MP.; Medini, L.; Lanteri,S; Partial least squares density modeling (PLS-DM) – A new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy. [En línea] **2014,** *851*, pp 30-36. http://www.ncbi.nlm.nih.gov/pubmed/25440661 (fecha de consulta 18, Noviembre 2015)

[13] Alonso-Salces, R.M.; Moreno-Rojas, J.M.; Holland, M.; Reniero, F.; Guillou, C.; Hérberger, K.; Virgin Olive Oil Authentication by Multivariate Analyses of ^{1}H NMR Fingerprints and $\delta\ ^{13}C$ and $\delta^{2}H$ Data. *J. Agric. Food. Chem.* [En línea] **2010,** *58*, pp 5586-5596. http://pubs.acs.org/doi/pdf/10.1021/jf903989b (fecha de consulta 19, Noviembre 2015)

[14] Geng,P.; Zhang,M.; Harnly,J.M.; Luthria,D.L.; Chen,P.; Use of fuzzy chromatography mass spectrometric (FCMS) fingerprinting and chemometric analysis for differentiation of whole-grain and refined wheat (T. aestivum) flour. *Anal Bioanal Chem.* [En línea] **2015,** *26*, pp 7875-88. http://www.ncbi.nlm.nih.gov/pubmed/26374564 (fecha de consulta 20, Noviembre 2015)

[15] Ruiz-Samblás, C.; Tres, A.; Koot, A.; Van Ruth, S.; González-Casado, A.; Cuadros-Rodríguez, L.; Proton transfer reaction-mass spectrometry volatile organic compound fingerprinting for monovarietal extra virgin olive oil identification. *Anal. Methods.* [En línea] **2012,** *134*, pp 589-596. http://www.sciencedirect.com/science/article/pii/S0308814612003500 (fecha de consulta 23, Noviembre 2015)

[16] Pérez-Castaño, E.; Ruiz-Samblás, C.; Medina-Rodríguez,S.; Quirós Rodríguez,V.; Jiménez-Carvelo, A. Valverde-Som, L.; González-Casado, A.; Cuadros-Rodríguez, L.

Comparison of different analytical classification scenarios: application for the geographical origin of edible palm oil by sterolic (NP) HPLC fingerprinting. *Anal. Methods.* [En línea] **2015,** *7,* pp 4192-4201, http://pubs.rsc.org/en/content/articlepdf/2015/ay/c5ay00168d (fecha de consulta 24, Noviembre 2015)

[17] Lerma-García, M.; Concha-Herrera, V.; Herrero-Martínez, J.; Simó-Alfonso, E.; Classification of Extra Virgin Olive Oils Produced at *La Comunitat Valenciana* According to Their Genetic Variety Using Sterol Profiles Established by High-Performance Liquid Chromatography with Mass Spectrometry Detection. *Agric. Food Chem.* [En línea] **2009,** *22,* pp 10512-10517, http://pubs.acs.org/doi/abs/10.1021/jf902322c (fecha de consulta 25, Noviembre 2015)

[18] Lerma-García, M.; Lusardi, R.; Chiavaro, E.; Cerretani, L.; Bendini, A.; Ramis-Ramos, G.; Simó-Alfonso, E.; Use of triacylglycerol profiles established by high performance liquid chromatography with ultraviolet-visible detection to predict the botanical origin of vegetable oils. *Chromatogr A.* [En línea] **2011**, *42*, pp 7521-7, http://www.ncbi.nlm.nih.gov/pubmed/21855883 (fecha de consulta 26, Noviembre 2015)

[19] Sillero, I.; nuevas aplicaciones de detectores analíticos no convencionales basados en procesos de ionización, [En línea] **2013,** pp 46, http://helvia.uco.es/xmlui/bitstream/handle/10396/10946/2013000000824.pdf?sequence=1 (fecha de consulta 27, Noviembre 2015)

[20] Thermo scientific, https://www.thermoscientific.es/about-us/promotions/thermo-scientific-dionex-corona-veo-charged-aerosol-detector.html (fecha de consulta 27, Noviembre 2015), los máximos ocultos se detectan gracias una visión más clara

[21] Award, A.; Emanuele, M.; Hartley, D.; Swartz, M.; LC GC, (fecha de consulta 30 Noviembre 2015) http://www.chromatographyonline.com/charged-aerosol-detection-pharmaceutical-analysis-overview. Charged Aerosol Detection in Pharmaceutical Analysis: An Overview, 2009.

[22] Guitiérrez, J.; Aplicación de métodos Quimiométricos para la caracterización y conrol de calidad de plantas medicinales; Univesidad Autónoma de Barcelona, 2012, pp 17

[23] Ruiz-Samblás, C. Marini, F.; Cuadros-Rodríguez, L. González-Casado, A. Quantification of blending of olive oils and edible vegetable oils by triacylglycerol fingerprint gas chromatography and chemometric tools. [En línea] **2012**, *910,* pp 71-77.http://www.sciencedirect.com/science/article/pii/S157002321200058X (fecha de consulta 2, diciembre 2015)

[24] Mata-Espinosa, P.; Bosque-Sendra, J.; Bro, R.; Olive oil quantification of edible vegetable oil blends using triacylglycerols chromatographic fingerprints and chemometric tools. [En linea] **2011,** *85,* pp 177-182. http://www.researchgate.net/publication/51195868_Olive_oil_quantification_of_edible_vegetable_oil_blends_using_triacylglycerols_chromatographic_fingerprints_and_chemometric_tools (fecha de consulta 2, diciembre 2015)

[25] Guitiérrez, J.; Quimiometria. In *Aplicación de métodos Quimiométricos para la caracterización y conrol de calidad de plantas medicinales*; Univesidad Autónoma de Barcelona, 2012, pp 19

[26] Mata, P.; Preprocesamieno de los datos. In *Aplicabilidad de la cromatografía líida y espectromería vibracional ara desarrollar modelos multivariantes para la detección y cantificación de aceite de oliva en mezclas de aceites vegetales;* Granada, 2011, pp 146

[27] Samblás, C.; Preprocesado de los datos. In *Autentificación de aceites vegetales mediane el emleo de cromatografía de gases y espectrometría de masas. Cuanificación de aceit de oliva.* Granada-España, 2012; pp 267

[28] Blanco, M. Regresión lineal por mínimos cuadrados. Calibrado univariable. In *Temas Avanzados de Quimiometría;* Cerdá, V., Ed: IIIes Balears, 2007; pp 251

[29] Beebe, K.; Pell, R.; Seasholz, M.; Preprocessing. In *Chemometrics.* Wiley, J.; Inc, S.; New York, 1998; pp 28

[30] Macho, S.; Técnicas de pretratamiento de datos. In *Metodología análitica basadas en espectroscoia de infrarrojo y calibración multivariane. Aplicación a la industria petroquímica;* Universidad Rovira y Virgili – Tarragona, 2002, pp 37

[31] Guitérrez, J.; Quimiomería. In *Aplicación de métodos Quimioméricos para la Caracterizacion y Control de Calidad de lanas Medicinales;* Univesidada Autónoma de Barcelona, 2012, pp 21

[32] Castaño, E.; Métodos no supervisados. In *Aplicalidad del perfil de esteroles para la cuantificación de aceites de oliva en alimentos;* Universidad de Granada, 2012, pp 86

[33] Romia, M.; Métodos de referencia. In *La espectroscoia NIR en la determinación de propiedades físicas y composición química de intermedios de producción y producos acabados*; Universidad autònoma de Barcelona, 2010, pp 35-36

[34] Samblás, C.; Herramientas Quimioméricas. In *Autentificación de aceites vegetales mediane el emleo de cromatografía de gases y espectrometría de masas. Cuanificación de aceit de oliva.* Granada-España, 2012; pp 331

[35] Hernández, L. Reconocimiento de pautas supervisadas. In *Tipicación y caracterización de café comercial mediante métodos instrumentales y quimiometría;* México, 2011; pp 31

[36] Vilardell, M.; Development and validation of qualitative methods in the food field; Tarragona, 2015; pp 6

[37] Guitiérrez, J.; Aplicación de métodos Quimiométricos para la caracterización y conrol de calidad de plantas medicinales; Univesidad Autónoma de Barcelona, 2012, pp 24

[38] Pantoja, P.; validación del modelo. In *Espectroscopia de infrarrojo cercano como resesa alternaiva en la caracerización y valoración de crudo en la refinería de petróleo*; Lima-Peru: 2013; pp 44

[39] López, M. I.; Callao , M. P.; Ruisánchez, I.; A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. *Anal. Chim. Acta,* [En linea] **2015,** *891,* pp 62-72, http://www.sciencedirect.com/science/article/pii/S00032670150083387 (Fecha de consulta 3, diciembre 2015)

[40] Sample Classification Predictions, http://wiki.eigenvector.com/index.php?title=Sample_Classification_Predictions (fecha de consulta 4, diciembre 2015), Class Pred strict