UNIVERSIDAD DE SEVILLA

FACULTAD DE QUIMICA

DEPARTAMENTO DE QUIMICA ANALITICA

DISCRIMINACION DE LAS VARIEDADES DE CAFE VERDE MEDIANTE TECNICAS DE ANALISIS MULTIVARIANTE

TESIS DOCTORAL

Mª JESUS MARTIN VALERO

Sevilla, 1997

D. Alfonso Guiraúm Pérez, Catedrático Director del Departamento de Química Analítica de la Universidad de Sevilla.

CERTIFICA: Que el presente trabajo de investigación ha sido realizado íntegramente en los laboratorios del Departamento de Química Analítica de la Universidad de Sevilla, y reúne las condiciones exigidas a los trabajos de tesis doctoral.

Y para que conste, expido y firmo el presente certificado en Sevilla a 14 de febrero de 1997.

Fdo. D. Alfonso Guiraúm Pérez

DIRECTORES

Fdo. D. Fernando de Pablos Pons

Fdo. D. A. Gustavo González González

Profesores Titulares del Departamento de Química Analítica de la Universidad de Sevilla.

Memoria presentada para aspirar al grado de Doctor de Ciencias Químicas.

Fdo. Mª Jesús Martín Valero

Lcda. en Ciencias Químicas

Deseo expresar mi más sincero agradecimiento:

Al doctor D. Alfonso Guiraúm Pérez por haberme ofrecido la oportunidad de llevar a cabo esta memoria.

A los doctores D. Fernando de Pablos Pons y D. A. Gustavo González González por su dirección, apoyo y paciencia en la realización del presente trabajo y por su amistad.

Al doctor D. Domingo González Arjona por su ayuda, amistad y por haberme guiado en el mundo de las redes neuronales.

Al doctor D. Miguel Angel Bello López por su colaboración.

A la empresa Kraft Jacobs Suchard (Saimaza), en especial a D. José Antonio Sainz de la Maza y D. Joaquín Sainz de la Maza porque sin su colaboración no habría sido posible la realización de esta memoria.

A mis padres y mi familia por su incondicional apoyo.

A Germán por su desinteresada ayuda en la realización de las figuras que ilustran la presente memoria, por su paciencia y por estar siempre ahí.

A mis compañeros y personal del departamento de Química Analítica, en especial, Coral, Flora, Javi y Juan Luis por todos los momentos compartidos y por su sincera amistad.

A Elena, Sofía, Ana, Almudena, Bernardett, Manoli, Pablo, Pepe, Fernando, David y Mariano por su amistad.

A mis padres

INDICE

CAPITULO I:

INTRODUCCION

I. INTRODUCCION

El café es un producto de gran interés tanto comercial como alimenticio, ya que su comercio es el más extendido por todo el mundo después del aceite. Según la leyenda[1-2], el café fue descubierto por un pastor etíope que observó cómo sus cabras permanecían toda la noche despiertas después de haber comido unas bayas del arbusto del café. El pastor comentó a unos monjes el extraño efecto que habían producido las bayas sobre los animales y los monjes empezaron a consumir infusiones preparadas con hojas de dicho arbusto para mantenerse despiertos en las largas sesiones de oración. Cierta o no la historia, se sabe que los orígenes de la *Coffea arabica* se remontan al año 575 a.c. en el sur de Etiopía y fue introducido

en Arabia a través de Yemen por comerciantes de las rutas del golfo de Adén, quienes en sus largos viajes ingerían granos de café machacados y mezclados con grasas. El café empezó a utilizarse habitualmente como bebida en las prácticas religiosas de las ciudades sagradas de La Meca y Medina y a partir de ahí se extendió su uso como bebida por todo el mundo musulmán a través de los numerosos peregrinos.

Hacia el siglo XV, países y regiones como Egipto, el Magreb, Turquía y Persia ya importaban grandes cantidades de café procedentes de Yemen. Como bebida, el café fue introducido en Europa por los turcos en el año 1600 y su consumo creció en todos los países europeos muy rápidamente hasta el punto que la práctica de beber café se hizo muy popular entre la aristocracia europea, de hecho, en 1675 ya había cerca de 3000 cafés públicos en Inglaterra, las cuales se convirtieron en verdaderos centros sociales.

Durante muchos años, la producción de café fue un monopolio de los árabes pero debido a la gran demanda, los europeos intentaron extender el cultivo del café. En los jardines botánicos de Holanda, se sembraron semillas de *Coffea arabica* procedentes de Java, una de las cuales a instancias del rey Luis XIV de Francia[3] se exportó a la Guayana Francesa en 1714 y llegó a ser la progenitora de los millones de arbustos de café que crecen en la actualidad en América del sur y Centroamérica.

También los franceses a principios del siglo XVIII, establecieron una plantación de café en la isla Reunión del océano Indico, entonces conocida como Bourbon, con semillas procedentes directamente de Arabia. Este café se extendió por los trópicos y constituyó una variedad nueva conocida como *Coffea arabica var.*

bourbon[4]. La otra especie importante desde el punto de vista comercial es la *Coffea canephora*, conocida comercialmente como robusta, es originaria del oeste de Africa.

I.1. LA PLANTA DEL CAFE

La planta del café es un arbusto de hoja perenne denominado "cafeto", que puede medir hasta 8 metros, aunque normalmente su altura suele mantenerse sólo hasta los 3 metros para facilitar la recogida de los frutos.

Botánicamente, fue clasificado en la familia de los evónimos y, posteriormente, en la familia de los jazmines; Linneo incorporó el cafeto a la familia de las rubiáceas (*Rubiaceae*) y Jussieu creó el género *Coffea*[5] exclusivamente para él. Este género incluye gran número de especies, pero sólo cuatro de ellas son de importancia económica: *Coffea arabica*, *Coffea canephora*, *Coffea excelsa* y *Coffea liberica*. De estas cuatro especies, *Coffea arabica* y *Coffea canephora* son las más empleadas comercialmente y sus denominaciones comunes son **arábica** y **robusta**.

La variedad arábica se cultiva en regiones situadas entre 500 y 1700 metros e incluso hasta 2000 metros, requiere una temperatura entre 13°C y 21°C y alrededor de 1500 l/m^2 de precipitación anual, mientras que la variedad robusta se encuentra en estado silvestre en casi todos los bosques de la zona tropical africana y asiática, creciendo a altitudes relativamente bajas: la variedad robusta tolera temperaturas más altas (18°C a 27°C) y más lluvia, en torno a 1800 l/m^2 de precipitación anual (humedad relativa del 80 a 90%), que la variedad arábica, la cual requiere un clima más seco y frío. La variedad robusta por el contrario demanda mayor cantidad de humus en el suelo, el cual debe ser permeable y de textura

abierta[6]. En general, la variedad robusta es más resistente que la arábica.

En cuanto al aspecto del grano de café, ambas variedades pueden distinguirse ya que el grano de café arábica tiene una forma ovalada y es de color verde claro, mientras que los granos robusta tienden a ser redondeados y de color marrón.

I.2. PRINCIPALES PAISES PRODUCTORES

Las áreas del mundo donde se cultiva el café están limitadas principalmente por la temperatura, ya que la planta se daña con las heladas, por lo que es en la zona tropical donde se localizan las plantaciones de café, en una franja de 25° norte a 25° sur a ambos lados del ecuador.

Los países productores pueden dividirse según la variedad de café que producen. Así:

La variedad arábica se cultiva, principalmente, en Brasil con una producción de aproximadamente 30 millones de sacos de café (1 saco ≅ 60 Kg) y es el primer productor de arábica. Colombia ocupa el segundo lugar en exportaciones produciendo un café de muy buena calidad, así como Bolivia, Ecuador, Paraguay, Perú, Venezuela y Costa Rica cuyo café es también de excelente calidad y constituye la principal riqueza del país. Cuba y El Salvador, donde la exportación de café constituye el 60% de los ingresos del país, Guatemala (40% de las exportaciones), Haití, Honduras, Jamaica (con poca producción pero de gran calidad), Méjico, Nicaragua y Panamá también son países productores de café.

Como puede observarse, la variedad arábica se cultiva en la zona de Centroamérica y América del Sur, aunque también La India, Etiopía y Kenia

producen café arábica.

Los principales países productores de café de variedad robusta son: Uganda que es el primer productor africano y el café es su fuente principal de ingresos, Angola, Camerún, Tanzania, Madagascar, República Centroafricana que exporta fundamentalmente a Francia e Italia al igual que lo hacen Guinea y Costa de Marfil, cuyo café es de calidad muy uniforme y es su principal exportación.

También producen café de variedad robusta Nigeria, Zaire e Indonesia, que es el tercer país productor del mundo siendo Estados Unidos su principal importador; Tailandia cuyas exportaciones están creciendo en los últimos años y Vietnam, que proporciona café principalmente a Japón, Singapur y países del este de Europa. La principal zona de cultivo de la variedad robusta está localizada en Africa y en el sureste asiático.

A finales del siglo XVIII, la producción mundial de café estaba en algo más de diez mil toneladas; a comienzos del siglo XX era de novecientas mil toneladas y en la actualidad asciende a prácticamente seis millones y medio de toneladas anuales[7], lo cual representa un valor aproximado de 800 billones de pesetas. Hay que mencionar que la producción de robusta, la cual era prácticamente nula en el siglo pasado, ha progresado de manera gigantesca sobre todo en las antiguas colonias francesas como Costa de Marfil. De hecho, en la actualidad se demanda más cantidad de café robusta que arábica y la producción de robusta constituye un 70% de la producción mundial. Los países productores de café tan sólo consumen un millón y medio de toneladas anuales, mientras que sólo Europa demanda anualmente unos dos millones y medio de toneladas.

En la siguiente figura, se muestran los distintos países productores de café.

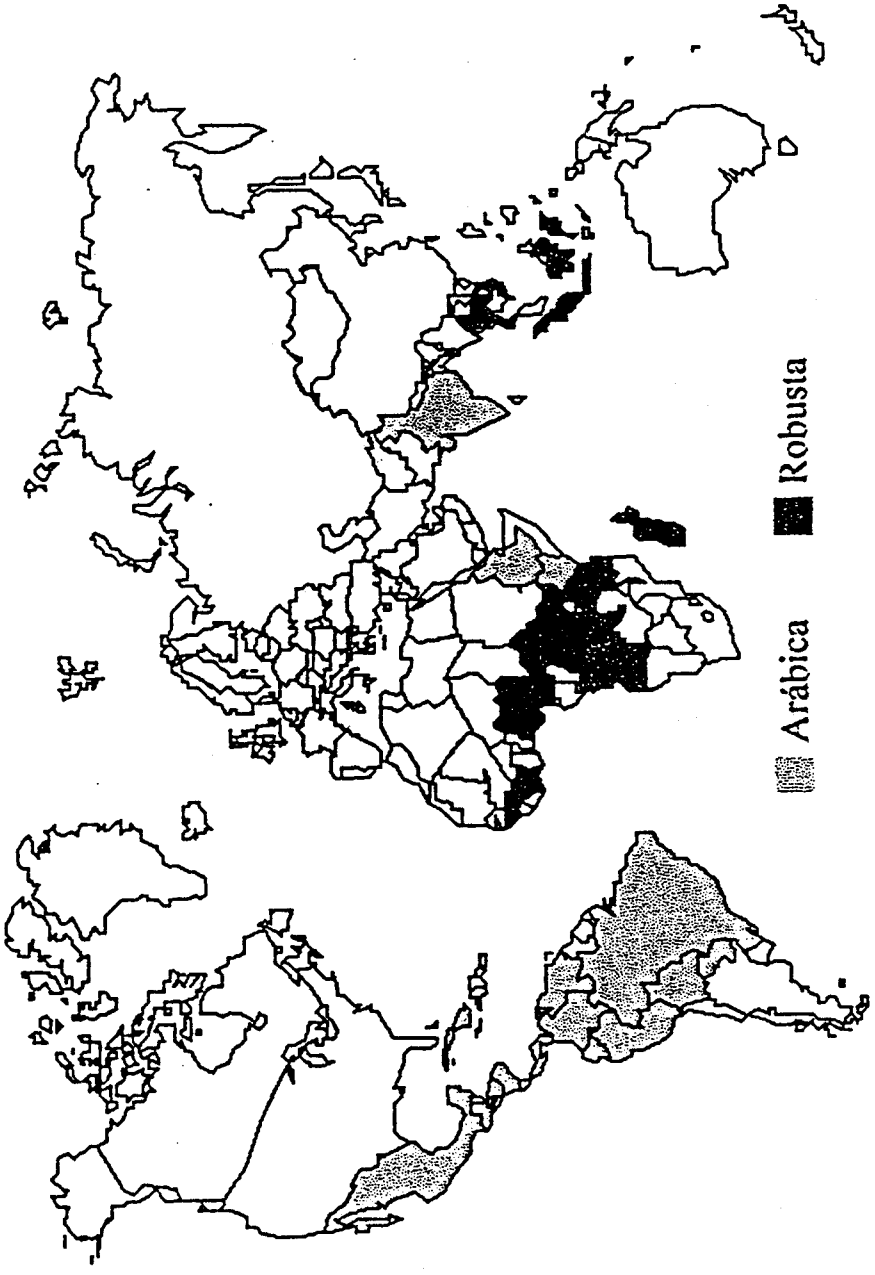

Figura 1. Principales países productores de café.

I.3. OBTENCION DE LOS GRANOS DE CAFE

En una plantación de café, las semillas previamente seleccionadas cuidadosamente se siembran en surcos a veces cubiertos con una capa de arena y protegidos de la luz solar. Cuando alcanzan los 20 ó 30 cm de alto, son trasplantados al campo, donde generalmente hay una densidad de 2500 a 3000 plantas por hectárea; en algunos países, se suelen plantar otros árboles junto a los cafetos para protegerlos del viento y del sol. En materia de abono, es necesario el uso de fertilizantes minerales apropiados, sobre todo el cafeto reclama esencialmente nitrógeno y potasio, mientras que los fosfatos sólo son necesarios en el momento de la formación de los frutos; sus necesidades en calcio son igualmente elevadas y, como hemos mencionado la variedad robusta requiere un alto aporte de humus. La poda sistemática de los arbustos es una operación importante, dejándolos generalmente a una altura no superior a los 3 m; por último, en determinadas áreas se practica la irrigación.

El cafeto no comienza a producir flores hasta los 3 años pero su producción no se hace rentable hasta los 5 años. A lo largo del tronco principal crecen ramas primarias opuestas unas a otras y en un mismo plano. Las hojas están opuestas y de forma lanceolada y son perennes. Las flores aparecen en la axila de las hojas agrupadas formando verticilos de entre 8 y 15 flores.

Una vez realizada la fecundación, hay que esperar entre 6 y 12 meses para que el fruto llegue a la madurez, por eso no es raro ver en una misma planta los frutos del año anterior junto a las flores de la próxima cosecha, tal y como se observa en la siguiente figura.

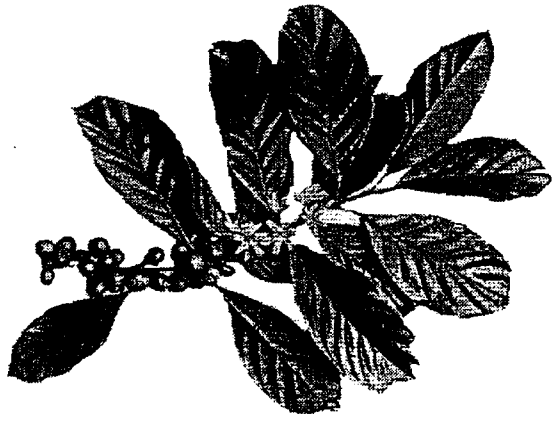

Figura 2. Rama del cafeto con frutos y flores.

El fruto del cafeto es una drupa, también denominada "cereza" debido a su color rojo, con una pulpa blanca y mucilaginosa, dulce y con dos semillas por lo general las cuales constituyen los granos de café. Estos granos están recubiertos por dos envueltas, una primera celulósica de color amarillo pálido llamada "pergamino" y una segunda envuelta consistente en una membrana ligera que recibe el nombre de "película plateada". Seguidamente, en la figura 3, se detallan cada una de las partes del grano.

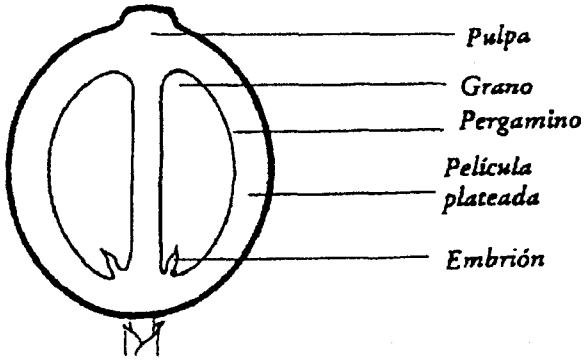

Figura 3. Partes del grano de café.

Las cerezas son recolectadas de forma manual para, posteriormente, separar la pulpa de los granos y dejarlos solamente con la envuelta más interna, este es el llamado *café pergamino*. Una vez obtenido el *café pergamino*, se separa también esta envuelta amarilla y correosa.

El **café verde** es el resultado de dejar a las semillas sin esta última envuelta y de acondicionar su humedad. Este proceso de separación de pulpa y envolturas de las semillas del fruto puede realizarse de dos formas distintas: por *vía seca* o por *vía húmeda*[8]. Antes de proceder al tratamiento propiamente dicho, deben eliminarse cuerpos extraños tales como hojas, restos de madera, piedras, arena y los granos podridos o no maduros. Esta selección se suele realizar con ayuda de seleccionadoras mecánicas que separan las impurezas por corrientes de aire; en algunos países se realiza una selección por densidad sumergiendo la cosecha en un recipiente con circulación de agua, de esta forma, las impurezas ligeras y los granos

más ligeros flotan en la superficie y son arrastrados por la corriente mientras que las piedras e impurezas grandes caen al fondo de la cubeta, quedando solamente los granos sanos de densidad media. Una vez realizada la selección, se procede al tratamiento propiamente dicho de la cereza.

- *Vía seca*:

Se realiza en Brasil y en los países africanos, exceptuando los de Africa oriental, por ser una zona más lluviosa. El método consiste en dejar secar al sol las cerezas extendiéndolas en amplias eras y revolviéndolas de forma regular con palas; así se dejan durante cuatro semanas hasta que su humedad final es menor del 12%, aunque un mejor control se realiza utilizando secaderos de aire caliente. Una vez seco el fruto, se procede al descerezado, con el que el café pasa a un cilindro cuya rotación proyecta los frutos contra las láminas, que rompen las envolturas y dejan libres los granos, eliminándose así los trozos de cáscara y de pergamino.

Seguidamente, el café se calibra y se envasa en sacos de yute de 60 ó 70 Kg de capacidad.

- *Vía húmeda*:

Este método se lleva a cabo en Centroamérica, países sudamericanos (excepto Brasil) y en Africa oriental, Kenia y Tanzania. Los frutos se colocan en balsas con agua, donde se seleccionan, ya que flotan los de peor calidad. Los granos todavía húmedos se pasan por un despulpador en el que la pulpa desmenuzada es arrastrada por una corriente de agua. Los granos que salen del aparato son café pergamino

sobre los que quedan adheridas capas de mucílago (producto gelatinoso y no soluble en agua) que hay que separar. Así pues, los granos se remojan en agua por espacio de 6 a 36 horas para provocar por fermentación la disgregación del mucílago, el cual se arrastra con una corriente de agua muy fuerte. El café con pergamino se seca entonces al sol o con corrientes de aire caliente[9] hasta una humedad inferior al 12%, a continuación se separa el pergamino mecánicamente.

Debido al remojado de los granos, se solubilizan sustancias que contribuyen al amargor del café, por lo que el café obtenido por vía húmeda es más suave que el obtenido por vía seca, recibiendo por ello la denominación de "mild" (suave).

I.4. COMPOSICION QUIMICA DE LOS GRANOS DE CAFE

La composición química de los granos de café verde depende de la variedad en cuestión y, en menor extensión, de otros factores como pueden ser grado de maduración, prácticas agrícolas, condiciones de almacenamiento, tipo de suelo, etc. Debido a la variabilidad que presentan los granos de café en su composición y a los distintos métodos analíticos empleados para la determinación de un componente dado, se hace difícil establecer tablas de composición en las que figuren rangos para los contenidos de los distintos componentes químicos del café verde según su variedad. Según Clifford y Vitzthum[10-11], los principales constituyentes del grano de café pueden resumirse en la siguiente tabla.

COMPONENTE	ARABICA	ROBUSTA
Contenido mineral	3.0 - 4.2	4.0 - 4.5
Cafeína	0.9 - 1.2	1.6 - 2.4
Trigonelina	1.0 - 1.2	0.6 - 0.75
Lípidos	12.0 - 18.0	9.0 - 13.0
Acidos Clorogénicos totales	5.5 - 8.0	7.0 - 10.0
Acidos alifáticos	1.5 - 2.0	1.5 - 2.0
Oligosacáridos	6.0 - 8.0	5.0 - 7.0
Polisacáridos totales	50.0 - 55.0	37.0 - 47.0
Aminoácidos	2.0	2.0
Proteínas	11.0 - 13.0	11.0 - 13.0

Tabla 1. Composición (% base seca) de los granos de café verde según su variedad.

A continuación, vemos cada uno de estos componentes presentes en el café verde:

I.4.1. Contenido mineral

Las sustancias minerales presentes en los granos de café verde suponen, aproximadamente, un 4% en base seca y comprende diferentes elementos cuyos contenidos presentan una gran variabilidad según factores[12] como variedad a la que pertenece la planta, origen geográfico, proceso al que se someten los granos de café (húmedo o seco), empleo de fertilizantes, etc. De hecho, hay evidencias claras de

que los cafés que sufren un proceso por vía seca (variedad robusta) presentan mayor contenido mineral que los granos tratados por vía húmeda (arábicas "mild")[13-14]. Existen diversos estudios sobre los contenidos medios de los distintos elementos metálicos presentes en el café verde[15-17] de los que se puede concluir,
que el elemento mayoritario en el contenido mineral es el potasio, el cual supone alrededor de un 40% del total (1.2%-1.8%) y es un factor dominante en el proceso nutricional de la planta[6]. Después, magnesio (0.16%-0.18%) y calcio (0.07%-1.2%), siendo estos dos elementos de gran importancia para la planta, que los toma del suelo o bien son proporcionados en forma de fertilizantes; fósforo cuya cantidad está directamente influenciada por la utilización de abonos fosfatados. En cuanto a componentes minoritarios[18-19], cabe destacar la presencia de hierro, manganeso, zinc, cobre, bario y estroncio al nivel de µg/g, y escandio, bromo, cobalto, cromo, cesio y lantano a concentraciones del orden de ng/g[20].
No se puede afirmar que exista una correlación definida entre el contenido mineral y la calidad del café, aunque parece que la calidad es mejor cuando hay mayor contenido de zinc y manganeso en los granos de café verde[16].

I.4.2. Carbohidratos

El café verde contiene un amplio rango de diferentes carbohidratos normalmente divididos en dos grupos: polisacáridos y azúcares de bajo peso molecular que incluyen tri-, di- y monosacáridos, los cuales suponen un 40-50%. Sin embargo, existen algunas dudas sobre la naturaleza y cantidad de estos compuestos, sobre todo, para la fracción de polisacáridos. También están presentes sustancias

derivadas de carbohidratos como es el caso de las pectinas.

Dentro de los azúcares de bajo peso molecular, la sacarosa es el compuesto presente en mayor cantidad en los granos de café verde. Por lo general, la variedad arábica contiene más sacarosa que la robusta. Existen numerosos estudios en los que se han llevado a cabo análisis de sacarosa en café verde. Tressl y colaboradores[21] emplearon cromatografía de gases, Trugo y Macrae[22] utilizaron un método cromatográfico de HPLC. Wolfrom y colaboradores[23] realizaron el aislamiento de la sacarosa del café verde a partir del extracto en una disolución de etanol acuoso.

En todos los casos, el paso previo de la extracción es determinante en los resultados finales y otro factor de influencia en el contenido de sacarosa es el tipo de cultivo, estado de maduración del grano, procesado y condiciones de almacenamiento.

En los extractos de café verde, también se ha observado la presencia de pequeñas cantidades (trazas) de otros azúcares simples tales como rafinosa, manosa, arabinosa, galactosa, ribosa y ramnosa. Se han detectado, asimismo, altos niveles de glucosa y fructosa analizadas en muestras de café verde mediante cromatografía en capa fina[21].

La presencia de algunos de estos azúcares puede deberse a un proceso de hidrólisis de los granos de café durante su almacenamiento. Estudios realizados por Porkorny y colaboradores[24] muestran un contenido total de azúcares reductores (expresado como glucosa) de 0.5% en una muestra de café de Colombia almacenada a temperatura ambiente durante un año. Un almacenamiento a altas temperaturas (60°C) y alta humedad provoca una disminución en el contenido de estos azúcares

debido a su reacción de Maillard con los aminoácidos libres también presentes en el grano.

En cuanto al contenido en polisacáridos, suponen un 40-50% del grano en base seca. Dentro de estos polisacáridos, se incluyen tanto glicanos, es decir, polisacáridos compuestos por unidades del mismo monosacárido (principalmente glucosa y arabinosa) como heteroglicanos, en el caso de estar compuestos por varios tipos de monosacáridos. Por otra parte, cuando el compuesto está formado por cadenas de monosacáridos como la manosa recibe el nombre de holocelulosa o hemicelulosa si son cadenas más cortas. Normalmente, en las semillas como es el caso del café, estos compuestos pueden complejarse con otros compuestos como proteínas, o bien estar recubiertos por otras sustancias como lignina y pectinas.

Tras varios estudios fundamentalmente de hidrólisis y reacciones de metilación[25-27], se descubrió que los polisacáridos presentes en los granos de café verde están constituidos por manosa principalmente y, en menor cantidad, por arabinosa, galactosa y glucosa. Investigaciones posteriores realizadas por Thaler y Arneth[28-30] confirmaron estos resultados. La presencia de manosana como principal polisacárido del café confiere características de dureza y resistencia a la planta.

En la tabla 2, se ofrecen datos obtenidos por estos últimos autores acerca de los porcentajes de los principales monosacáridos constituyentes de los polisacáridos presentes en cafés de variedad arábica.

Monosacárido	% base seca
arabinosa	1.8
galactosa	9.3
manosa	20.8
glucosa	6.8
Total	38.7

Tabla 2. Contenido de monosacáridos en polisacáridos de café arábica.

I.4.3. Compuestos nitrogenados

El término compuestos nitrogenados se aplica a todos aquellos compuestos que contengan en su molécula nitrógeno tanto orgánico como inorgánico. En este apartado, vamos a considerar tres grupos fundamentales: alcaloides, bases nitrogenadas y, por último, aminoácidos y proteínas. Hemos de mencionar que existen otros compuestos nitrogenados que no consideraremos aquí, sino en subsiguientes apartados como componentes volátiles y lípidos.

a) Alcaloides: los alcaloides que contiene el café son compuestos derivados de la purina, están ampliamente presentes en el reino vegetal, aunque el anillo de purina como tal no se encuentra en la naturaleza.

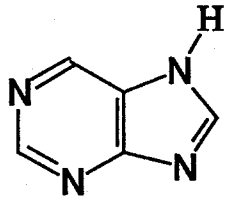

Figura 4. Estructura química de la purina.

Las xantinas, dioxoderivados de la purina, en concreto, la 1,3,7-trimetilxantina es la cafeína, alcaloide principal del café, aunque también pueden encontrarse trazas de teofilina (1,3-dimetilxantina) y teobromina (3,7-dimetilxantina).

La estructura de la cafeína se muestra en la siguiente figura:

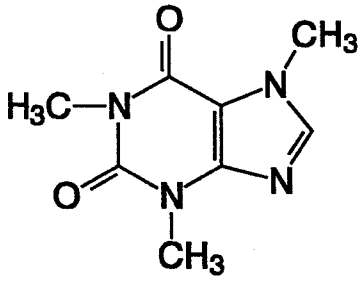

Figura 5. Estructura de la cafeína.

La cafeína es una base anfótera de color blanco, que funde a 236°C y sublima a 178°C. Es soluble en agua aunque su solubilidad aumenta a altas

temperaturas, así 1 gramo de cafeína se disuelve en 46 ml de agua, en 5.5 ml de agua a 80°C y en 1.5 ml de agua hirviendo. Es, asimismo, moderadamente soluble en disolventes orgánicos como etanol, metanol, benceno, cloroformo, éter y CO_2 supercrítico. Muchos de estos disolventes se utilizan para extraer la cafeína de los granos de café y obtener así el café descafeinado. La cafeína presenta un espectro de absorción con un máximo próximo a 270 nm[31].

Al cristalizar la cafeína a partir de soluciones acuosas, se obtiene en forma hidratada con un 6.95% de agua[32]. Por otra parte, la cafeína es relativamente estable en ácidos diluidos y en alcalis, pero puede formar complejos con otros componentes del café como son los ácidos clorogénicos o compuestos aromáticos polinucleares. De hecho, esta propiedad se emplea para realizar extracciones selectivas de dichos compuestos aromáticos en diversos alimentos.

En el café verde, la cafeína se encuentra como sal doble: clorogenato de cafeína y de potasio. El contenido de cafeína en los granos depende de la variedad de que se trate; así, los cafés robusta presentan mayores contenidos en cafeína, del orden de 2.2% en base seca, que los granos de café arábica (1.2%). Esta es una de las principales características por las que se distinguen las dos variedades de café verde. La diferencia en contenido de cafeína se aprovecha a la hora de realizar las mezclas para obtener el café comercial.

b) Bases nitrogenadas: las bases nitrogenadas presentes en el café se pueden dividir en dos grupos, aquellas estables a la temperatura de tueste y aquellas que se descomponen originando compuestos volátiles de importancia organoléptica.

Pertenecientes al primer grupo, son betaína (N,N,N-trimetilglicina) y colina los cuales se encuentran en cantidades traza (por debajo de 0.1%) en los granos de café verde. En el segundo grupo de compuestos, destacan trigonelina y amidas serotonínicas. Estas últimas se encuentran formando parte de las ceras que recubren la superficie de los granos y se extraen del café verde mediante disolventes hidrofóbicos como dietiléter y éter de petróleo.

En cuanto a la trigonelina, su estructura química se muestra en la figura 6.

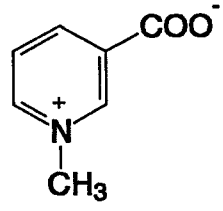

Figura 6. Estructura de la trigonelina.

Se obtiene en forma de cristales incoloros monohidratados extrayendo con etanol acuoso. También se puede obtener en forma anhidra. Se descompone en su fusión a 218°C. Es muy soluble en agua como puede deducirse de su fórmula zwitteriónica, sin embargo es poco soluble en disolventes orgánicos como cloroformo y diclorometano.

El contenido de trigonelina en los granos de café verde depende fundamentalmente de las variedades, encontrándose que la variedad arábica presenta una mayor cantidad de trigonelina que la variedad robusta. A pesar de este hecho

generalizado, los datos disponibles acerca del contenido de trigonelina en los granos presentan una amplia variabilidad debido, presumiblemente, a los distintos métodos analíticos empleados para la determinación[33-35].

La trigonelina ha sido muy estudiada debido a que, durante el tostado se descompone originando como producto mayoritario ácido nicotínico, importante tanto desde el punto de vista sensorial como nutricional.

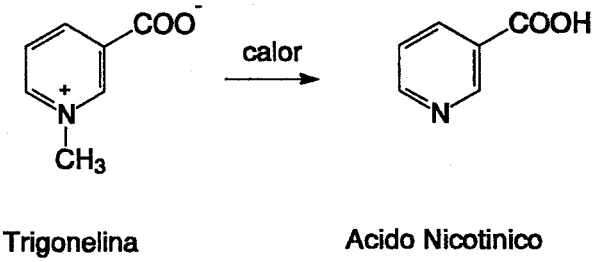

Figura 7. Degradación de trigonelina a ácido nicotínico.

El grado de conversión depende directamente de la temperatura de tueste, aunque en términos generales puede decirse que se descompone del orden del 50-80% de la trigonelina presente en el café verde[36]. Además de ácido nicotínico, se originan otros compuestos tanto no volátiles (N-metilnicotinamida) como volátiles y presentes en el aroma del café, como son piridinas, pirroles y compuestos bicíclicos.

La trigonelina que no se descompone, se extrae en la preparación de las bebidas de café contribuyendo a su amargor.

c) Aminoácidos y proteínas: en el café verde, las proteínas están presentes fundamentalmente en el citoplasma, de forma no enlazada, y en las paredes celulares enlazadas a polisacáridos. El contenido medio de proteínas en el café verde no difiere de una variedad a otra y gira en torno al 8.7-9.7% de base seca[37].

La fracción de proteínas del grano de café se puede dividir en dos grandes grupos: proteínas solubles en agua, que son albúminas y otra fracción insoluble, que es la mayoritaria. Existen diversos estudios en los que se han intentado caracterizar los aminoácidos componentes de las proteínas mediante distintas técnicas analíticas como son filtración sobre gel, diálisis y electroforesis entre otras[38-39] siendo dichos aminoácidos: alanina, arginina, ácido aspártico, ácido glutámico, glicina, histidina, isoleucina, leucina, lisina, prolina, valina, serina, metionina, valina y fenilalanina.

Debido a las altas temperaturas del tueste, esencialmente todas las proteínas se desnaturalizan y liberan aminoácidos.

El contenido en aminoácidos libres del café verde juega un importante papel en el sabor final del café tostado, ya que son precursores del aroma. Su análisis en los granos de café verde no ha recibido demasiada atención, Walter y colaboradores[40] establecieron un rango de 0.15-0.25% para el total de aminoácidos libres. La variedad robusta presenta un mayor contenido de aminoácidos libres que la variedad arábica. En cuanto a aminoácidos en concreto, todos están presentes en mayores cantidades en los cafés robusta que en los arábicas, a excepción del ácido glutámico que se encuentra en más cantidad (50%) en la variedad arábica. También el ácido pipecólico se ha determinado solamente en cafés arábicas y no en muestras de variedad robusta[41-42].

El aislamiento de los aminoácidos libres de los granos de café verde es una operación que entraña cierta dificultad, ya que en la mayoría de los casos se requiere una separación previa de los lípidos seguida de extracción de los aminoácidos libres con alcohol acuoso. En caso utilizar disolventes más eficientes deben separarse del extracto las proteínas junto con otras macromoléculas mediante precipitación.

I.4.4. Acidos clorogénicos

El grano de café contiene una serie de ácidos orgánicos, tales como acético, pirúvico, oxálico, málico y cítrico. Sin embargo, el café verde contiene en mayor proporción otros ácidos, que presentan el carácter de taninos, entre los que se encuentran el ácido caféico y el ácido quínico, cuyas estructuras químicas se muestran a continuación.

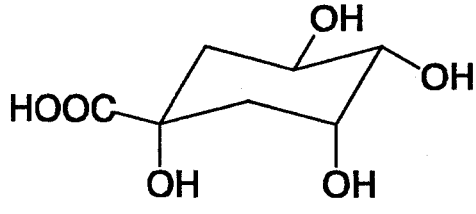

Figura 8. Estructura del ácido quínico.

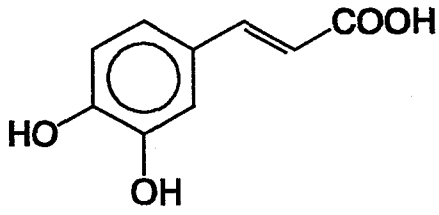

Figura 9. Estructura del ácido caféico.

El ácido quínico como tal ácido libre se encuentra en pequeñas cantidades en el grano de café, ya que un alto porcentaje de ácido quínico se esterifica originando diversos compuestos denominados ácidos clorogénicos. Posiblemente, el primer trabajo donde se describen estos compuestos es el realizado por Robiquet y Boutron[43] en 1837, quienes estudiando sustancias de actividad fisiológica en el café aislaron de los granos de café verde una sustancia de carácter ácido que producía un pigmento verde al ser tratada con cloruro férrico.

En 1903, Griebel proporcionó un punto de fusión de 202-3°C para los cristales de color verdoso de ácido clorogénico y Gorter estableció un punto de fusión de 206-7°C para cristales de color blanco[44]. Mediante hidrólisis alcalina a bajas temperaturas se obtenían cantidades equimoleculares de ácido quínico y ácido cafeico. Años más tarde, en 1920, Freudenberg[45] descubrió que la enzima tanasa hidrolizaba el ácido clorogénico obteniéndose cantidades equimoleculares de ácido quínico y ácido cafeico. En 1932, Fischer y Dangschat[46] mediante sus estudios realizados en la universidad de Berlín dedujeron que el ácido clorogénico era el

ácido 3-cafeoilquínico, ya que:

- el ácido clorogénico no formaba una lactona al calentarse con anhídrido acético, por tanto el -OH del carbono 3 está bloqueado.
- el ácido clorogénico formaba un derivado diacetónico, por lo que los grupos hidroxilos de los carbonos 1, 4 y 5 debían estar libres.
- tras metilación y saponificación, el ácido clorogénico daba ácido 3,4-dimetilcafeico y ácido 1,4,5-trimetilquínico (aislado en forma de lactona), por tanto los -OH de los carbonos 1, 4 y 5 debían estar libres.

Sin embargo, la IUPAC recomienda[47] que el ácido 3-cafeoilquínico se denomine ácido 5-cafeoilquínico.

La estructura química del ácido clorogénico se muestra en la siguiente figura.

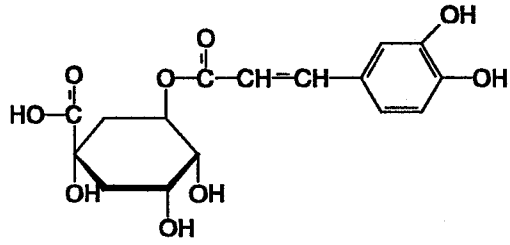

Figura 10. Estructura del ácido clorogénico (5-cafeoilquínico).

En 1950, Barnes y colaboradores[48] descubrieron que este ácido clorogénico no era el único presente en el café verde, por lo que introdujeron el término ácido

isoclorogénico para el ácido que ellos decían era el ácido 3-cafeoilquínico. Como puede observarse, existe toda una controversia creada en torno a la nomenclatura y caracterización de los ácidos clorogénicos del café. Durante los siguientes 15 años y gracias al desarrollo de técnicas analíticas como la cromatografía y la espectroscopía, se llegó al establecimiento de diversos ácidos clorogénicos, incluyendo tres compuestos derivados del anteriormente mencionado ácido isoclorogénico que contienen dos moléculas de ácido cafeico.

Resumiendo, se puede concluir que los tres principales ácidos clorogénicos del café son el ácido clorogénico (5-cafeoilquínico), neoclorogénico (3-cafeoilquínico) e isoclorogénico (ácido dicafeoilquínico). Estos ácidos son solubles en agua y en mezclas alcohol-agua, siendo insolubles en benceno y cloroformo. Los cristales de ácido clorogénico obtenidos en disoluciones de metanol-agua funden a 146°C, mientras que si se obtienen a partir de disoluciones de etilacetato o éter de petróleo, los cristales tienen forma de prisma y su punto de fusión es de 166°C. En cuanto al espectro de absorción ultravioleta en etanol, el ácido clorogénico presenta un máximo a 330 nm[49].

Los compuestos fenólicos, incluyendo los ácidos clorogénicos suelen estar presentes en las plantas como productos secundarios; sin embargo, en el grano de café existe un alto contenido en estos ácidos, lo cual puede ser debido a que posean determinada actividad bioquímica, como protección contra la invasión microbiana, ser precursores de la biosíntesis de la lignina, etc[50].

En cuanto al contenido de ácidos clorogénicos en el café verde, se sabe que la variedad robusta presenta mayores cantidades que la variedad arábica, alrededor

de 7-10% y 5-7.5%, respectivamente. Estos contenidos se refieren al total de estos ácidos, siendo el ácido clorogénico el que se encuentra en mayor cantidad con respecto a los otros tres. También se ha observado un aumento en el contenido de ácidos clorogénicos totales con el grado de maduración de la planta, sobre todo, se observa una mayor presencia del ácido isoclorogénico[51-53]. Por el contrario, Meissner y colaboradores observaron una pérdida de ácido clorogénico en los granos de café verde que habían sido almacenados a una humedad relativa del 80% y a una temperatura entre 30°C y 50°C.

I.4.5. Lípidos

Los lípidos presentes en el café verde son los componentes de un aceite que se encuentra fundamentalmente en el endocarpio del grano, así como de una pequeña cantidad de "cera" localizada en las capas exteriores del grano.

Existen diversos trabajos en los que se proporcionan contenidos de lípidos en el café verde[54-58]. En ellos, queda reflejado una mayor presencia de compuestos lipídicos para la variedad arábica, en torno a un 15% en base seca con respecto a la variedad robusta, la cual presenta un contenido medio de 10%.

Podemos dividir los compuestos lipídicos en dos fracciones: saponificable e insaponificable.

Dentro de los lípidos saponificables, la mayoría son triglicéridos los cuales constituyen el 70-80% de esta fracción[59-60]; en menor cantidad se encuentran ácidos grasos libres, del orden de 0.5-3.0% en granos de buena calidad, aunque pueden llegar a constituir un 20% en el caso de granos de baja calidad[61-62]. Entre estos

ácidos grasos, se encuentran el ácido linoleico, palmítico, oleico, etc.

Por otra parte dentro de la fracción insaponificable, el café verde contiene alcoholes diterpénicos tanto libres como esterificados y esteroles, principalmente estigmasterol, sitosterol y campesterol, aunque también se pueden encontrar trazas de campestanol, colesterol, clerosterol, sitostanal, Δ^5-avenasterol y colestanol[63-65].

I.4.6. Componentes volátiles

Los compuestos volátiles del café se encuentran, en su mayoría, formando parte del aroma del café. Sin embargo, en el grano de café existen una serie de compuestos volátiles aunque la mayoría de ellos aumentan su concentración durante el tostado.

Merrit y colaboradores[66] detectaron la presencia de hidrocarburos alifáticos, procedentes de la oxidación de los lípidos del grano durante su almacenamiento antes del tueste. También encontraron furanos, tiofenos, aldehidos, sulfuros, cetonas y ésteres.

Poisson[67] detectó diversos compuestos volátiles en el grano, tales como piridinas, quinolinas, pirazinas, pirroles, arilaminas y poliaminas. Gutmann y colaboradores[68] compararon los contenidos de volátiles en cafés arábicas y robustas concluyendo que la variedad robusta presenta mayores contenidos de estos compuestos que la variedad arábica y detectando hasta 79 compuestos.

Debido a un almacenamiento de los granos en altas condiciones de humedad relativa (50%) y altas temperaturas, pueden formarse más componentes volátiles[69].

I.4.7. Acidos alifáticos

Se tienen pocos datos acerca de los ácidos alifáticos presentes en el café verde. Deatherage y colaboradores[70-71] y Nakabayashi[72] proporcionan cantidades del orden de 0.5% de ácido cítrico, 0.5% de málico, 0.2% de oxálico y 0.4% de tartárico, en café verde de variedad arábica. En cuanto a la variedad robusta, Northmore[73] encontró grandes cantidades de ácido acético.

Por otra parte, se sabe que existen mayores contenidos de ácidos en granos que hayan permanecido almacenados durante largos períodos de tiempo con respecto a aquellos granos procedentes de cosechas recientes.

I.5. ACCION FISIOLOGICA

La acción fisiológica del café es debida a la presencia de algunos de sus componentes químicos. Entre estos compuestos, la cafeína es uno de los constituyentes más importante.

Así, el metabolismo de la cafeína[74] puede describirse como sigue:

Una vez consumida, la cafeína es rápidamente absorbida, en su totalidad, por el tracto gastrointestinal y, en una hora, es distribuida por todo el cuerpo. Los riñones no la eliminan totalmente de la corriente sanguínea, metabolizándose en paraxantina, teofilina, teobromina y derivados del ácido úrico.

Ingestas moderadas de cafeína, del orden de 0.25 gramos, proporcionan efectos beneficiosos para el cuerpo humano; de hecho, la cafeína es diurética, en el sistema nervioso central facilita la percepción de las excitaciones sensoriales, el ejercicio de las funciones cerebrales y estimula la excitabilidad refleja de la médula espinal;

favorece asimismo la función pulmonar aumentando la frecuencia y la amplitud de los movimientos respiratorios. Acrecienta, también, el trabajo de los músculos estriados; la cafeína, por otra parte, es tonicardíaca, aumenta la capacidad cardíaca y asegura una dilatación de los vasos. La dosis máxima recomendable es de 1.50 gramos, siendo 10 gramos la dosis letal.

A pesar de los mencionados efectos beneficiosos, algunos estudios epidemiológicos afirman que la cafeína puede producir cáncer, úlceras, defectos en fetos y enfermedades coronarias. A este respecto, el *American Council on Science and Health*[75] desmiente estas hipótesis aunque recomienda que las mujeres embarazadas y aquellas que intenten quedarse embarazadas limiten el consumo diario de cafeína.

Por otro lado, la presencia de ácidos clorogénicos ejerce una acción estimulante sobre el aparato digestivo, circulatorio y el sistema nervioso central, a la vez que facilita la fijación de ciertas proteínas.

Otros ácidos presentes en el café como acético, málico, cítrico, pero, sobre todo, ácidos clorogénicos, junto con los ácidos quínico y cafeico, los cuales presentan el carácter de taninos, pueden tener una acción astringente sobre mucosas y tejidos.

En cuanto al contenido vitamínico, es importante la presencia de trigonelina en el café verde, ya que ésta en la etapa de tueste se descompone en bastante proporción formando como producto principal de la degradación ácido nicotínico, el cual constituye un aporte de vitamina PP (factor antipelagra) considerable, en torno a 1-2 mg.

Por último, el café es una buena fuente de minerales como potasio, calcio, magnesio, hierro y manganeso con gran valor nutritivo[18].

I.6. EVALUACION DE LA CALIDAD DEL CAFE

Para evaluar la calidad del café se realizan varios ensayos en los que se determinan diversos parámetros, tales como humedad, contenido mineral, extracto acuoso, ácidos clorogénicos, carbohidratos, sustancias nitrogenadas y componentes volátiles[50].

Con fines clasificatorios, son de gran importancia aquellos parámetros que caracterizan los diferentes tipos de café y evalúan los niveles de las sustancias responsables del aroma, color y acción fisiológica. Dentro de este grupo, podemos considerar compuestos como son los polifenoles, alcaloides como la cafeína, metales como potasio, calcio, hierro, manganeso, magnesio, bario, estroncio, cobre, entre otros. El contenido en ácidos clorogénicos también es una medida directa de la calidad del café[76]. El extracto acuoso y los aminoácidos son también parámetros de interés, ya que contribuyen al cuerpo y aroma, respectivamente[20,77] y están recomendados por la asociación de Químicos Analíticos Oficiales de los Estados Unidos (A.O.A.C.) para evaluar la calidad del café.

En este trabajo se han considerado, como clasificatorios, los siguientes parámetros químicos: extracto acuoso, polifenoles totales, aminoácidos libres totales, cafeína, ácido clorogénico y metales, todos ellos referidos a base seca.

I.6.1. EXTRACTO ACUOSO

El extracto acuoso es un parámetro que indica la cantidad total de sólidos solubles en una muestra. Täufel y colaboradores[78] emplearon un método para la determinación del extracto acuoso junto con el valor de pH y la cantidad de ácidos libres en café.

En la presente memoria, la determinación del extracto acuoso se lleva a cabo mediante un método gravimétrico recomendado por la A.O.A.C.[79-80], disolviendo un peso conocido de muestra en agua caliente, filtrando el resto insoluble para, una vez evaporado la totalidad del disolvente, pesar el residuo.

I.6.2. POLIFENOLES

Los polifenoles naturales son compuestos caracterizados porque precipitan las proteínas y alcaloides, dan precipitados de color azul con sales de hierro, sirven para curtir las pieles y son astringentes. Reciben el nombre genérico de materias tánicas o taninos.

Existen numerosos trabajos en los que se ha llevado a cabo la determinación de compuestos fenólicos, concluyendo que el empleo de los reactivos Folin (Folin-Denis, 1912 y Folin-Ciocalteu, 1927) es muy adecuado para estas determinaciones.

Así, Sharama y Krishnan[81] determinaron sustancias húmicas en suelos, de Hann[82] determinó también sustancias húmicas en aguas, Polvoledo y colaboradores[83] analizaron compuestos polifenólicos en sedimentos, Berk y Schroeder[84] estudiaron una amplia gama de sustancias fenólicas y, en general, los reactivos Folin han sido empleados como método estándar para la determinación de taninos y ligninas[85-86].

Singleton y colaboradores[87] realizaron un estudio comparativo de estos reactivos, llegando a la conclusión que el reactivo Folin-Ciocalteu es más apropiado para la determinación de polifenoles totales que el reactivo Folin-Denis, debido a la obtención de un compuesto coloreado más intenso y a la no reactividad de posibles sustancias interferentes.

Por tanto, la determinación de los polifenoles se realiza mediante un método espectrofotométrico, utilizando el reactivo Folin-Ciocalteu. Tiene lugar una reacción entre los polifenoles y el ácido fosfowolfrámico presente en el reactivo, el cual se reduce originando W_2O_5, que es la especie coloreada. Dicha reacción de oxido-reducción puede describirse de forma general como sigue:

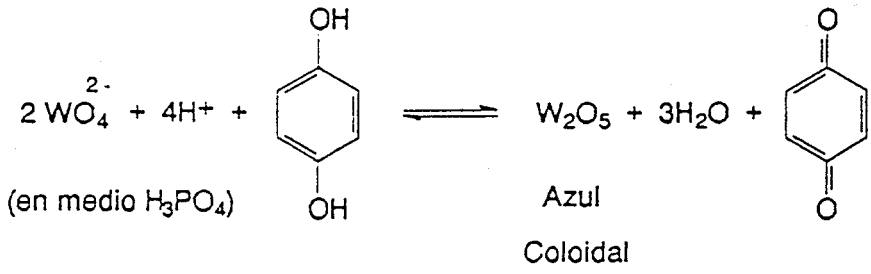

Figura 11. Reacción para la determinación de los polifenoles totales.

I.6.3. AMINOACIDOS LIBRES

Un método muy utilizado para la determinación cuantitativa de los aminoácidos es el procedimiento espectrofotométrico propuesto por Moore y Stein[88], basado en la formación de un compuesto coloreado entre los aminoácidos y la ninhidrina. Este método se ha empleado para la determinación de aminoácidos libres totales en muestras de té[89] y en muestras de café tostado[90].

La reacción que tiene lugar se muestra en la siguiente figura:

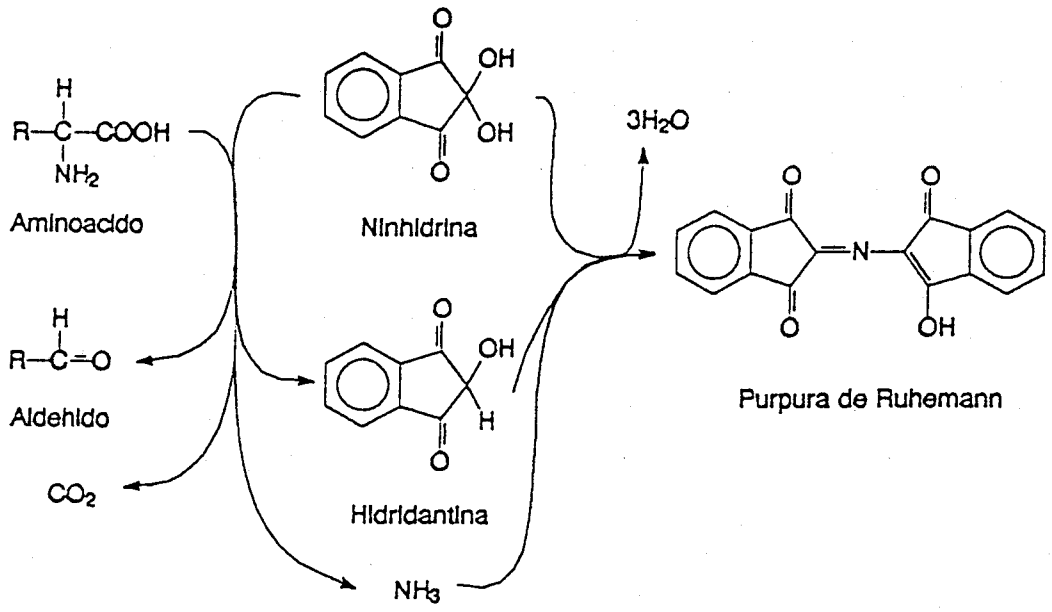

Figura 12. Reacción para la determinación de los aminoácidos libres totales.

Hay que señalar que la ninhidrina reacciona con numerosos compuestos que contienen nitrógeno en su molécula, por lo que es importante encontrar las

condiciones de reacción adecuadas para conseguir mejorar la selectividad.

I.6.4. CAFEINA

La cafeína, metil derivado de la xantina, es el alcaloide presente en una mayor proporción en el café verde y el compuesto más frecuentemente determinado en muestras de café.

Los métodos de análisis empleados para la determinación de cafeína han sufrido grandes cambios durante los últimos años. Incialmente, se emplearon métodos gravimétricos, así como el método de Kjeldahl para la determinación del nitrógeno del alcaloide. Estos fueron los métodos oficiales de análisis considerados durante unos años[91-93]. Debido a que la cafeína absorbe en la región ultravioleta presentando un máximo sobre 270 nm, también se ha empleado la espectrofotometría ultravioleta-visible[94-96] para su determinación, asimismo, se ha empleado la cromatografía de capa fina[97-98], cromatografía de gases[99-100], cromatografía de gel filtración[101] y valoración potenciométrica[102].

Una de las técnicas más interesante para la determinación de cafeína es la cromatografía líquida (LC). Maeda y colaboradores[103] determinaron cafeína, vitaminas solubles y conservantes en bebidas mediante LC en fase reversa con detección espectrofotométrica a 254 nm. Campiglia y colaboradores[104] determinaron cafeína, teofilina y teobromina por LC con detección fosforimétrica.

Sin embargo, estos métodos cromatográficos se han visto desbancados por el desarrollo de la cromatografía líquida de alta resolución (HPLC), que ha resultado ser la técnica más rápida y resolutiva para el análisis de cafeína; además, como este

compuesto contiene varios grupos cromóforos en su molécula es fácilmente detectable mediante absorción UV. A continuación, se citan diversas publicaciones en las que se emplea esta técnica. Tan y colaboradores[105] determinaron cafeína en bebidas mediante HPLC con detección espectrofotométrica a 254 nm; Moia y col.[106] analizaron cafeína y teofilina en plasma por HPLC con detección espectrofotométrica a 280 nm, Kuhr y Engelhardt[107] determinaron cafeína, flavonoles y ácido gálico en té usando HPLC en gradiente. Muthladi y colaboradores[108] analizaron cafeína en diferentes productos alimenticios con detección espectrofotométrica a 274 nm comparando esta técnica con la cromatografía de gases. Miceli y Chapman[109] analizaron cafeína en orina mediante HPLC con detector de fila de diodos, barriendo en la zona comprendida entre 255 nm y 290 nm. Otros trabajos[110-111] encontrados en la bibliografía también emplean HPLC con detección espectrofotométrica para la determinación de cafeína en alimentos. Por otra parte, el Centro de Investigación y Control de la Calidad (C.I.C.C.) recomienda la cromatografía líquida de alta resolución en fase reversa como método oficial para análisis de cafeína en alimentos[112].

I.6.5. ACIDO CLOROGENICO

Junto con la cafeína, el ácido clorogénico (ácido 5-cafeoilquínico) es uno de los compuestos característicos del café verde. Muy frecuentemente y junto con el resto de ácidos clorogénicos pertenecientes a la misma familia, se determina como contenido en ácidos clorogénicos totales.

Clifford[113] realizó un amplio estudio acerca de las técnicas analíticas

empleadas para la determinación de ácidos clorogénicos, entre ellas mencionaremos métodos espectrofotométricos (λ=320-330 nm), originariamente recomendados por la A.O.A.C. para el contenido de ácidos clorogénicos totales en granos de café verde, aunque para el caso de cafés tostados pueden existir interferencias de compuestos no fenólicos que absorben en esa región del espectro.

Se han utilizado diversas técnicas espectroscópicas para la identificación de ácidos clorogénicos en el café. Así, podemos citar la Resonancia Magnética Nuclear[114-115], Espectrometría de Masas[114,116] y la espectroscopía de Infra-Rojo[117]. Entre los métodos cromatográficos, se ha empleado la cromatografía de gases[118] pero la técnica más comúnmente aplicada para el análisis de estos compuestos ha sido la cromatografía líquida. Hanson y Zucker[119] emplearon este método en régimen de gradiente (cliclohexano-cloroformo 10:90 y alcohol-*t*-butílico-cloroformo 10:30) para la separación de ácidos clorogénicos.

Sin embargo, la mayoría de los trabajos encontrados en la bibliografía utilizan HPLC para el análisis, ya que se evitan etapas previas de derivatización. Posiblemente, los primeros trabajos publicados en los que se determinan ácidos clorogénicos por HPLC en fase reversa se atribuyen a Court y Rees[120-121]. Van der Stegen y Van Dujin[122] lograron separar hasta doce ácidos clorogénicos y dos ácidos cinámicos mediante HPLC en fase reversa empleando metanol-agua a pH 2.5 como fase móvil en gradiente. Clifford y colaboradores[52] pusieron a punto un método de HPLC en régimen isocrático para la determinación de nueve ácidos clorogénicos, tres ácidos cinámicos y cafeína. En todos estos trabajos, se ha empleado detección espectrofotométrica a 310-320 nm. Trugo y Macrae[123] determinaron nueve ácidos

clorogénicos en cafés instantáneos mediante HPLC en gradiente, estos mismos autores estudiaron el efecto del tueste en la composición de ácidos clorogénicos del café[124]. Ramírez[125] determinó diversos compuestos fenólicos en pulpa de café mediante HPLC. Morishita y colaboradores[114] realizaron la separación de siete ácidos clorogénicos en café mediante HPLC llevando a cabo su posterior identificación por espectrometría de masas y resonancia magnética nuclear. De Maria y colaboradores[35] analizaron simultáneamente ácidos clorogénicos totales, trigonelina y cafeína en muestras de café verde mediante cromatografía de filtración sobre gel de alta resolución.

En la presente memoria, se realiza la determinación de cafeína y ácido clorogénico mediante HPLC en fase reversa y en régimen isocrático.

I.6.6. TRIGONELINA

La trigonelina es un componente del grano de café verde y su importancia se debe, principalmente, a que es el precursor de diversos compuestos presentes en el aroma del café tostado, además de tener importancia desde el punto de vista nutricional.

Los primeros métodos de análisis de trigonelina se basaban en su precipitación y posterior determinación como un complejo de yodo[126]. También se realizaron determinaciones colorimétricas y se utilizaron otros métodos espectroscópicos[127-128].

Sin embargo, todos estos métodos han sido reemplazados por los métodos cromatográficos. En primeros trabajos, se utilizó cromatografía en capa fina para la

determinación de trigonelina y cafeína en café con posterior detección ultravioleta a 265 nm[129], aunque, posteriormente, se ha preferido la cromatografía líquida de alta resolución debido a su precisión y sensibilidad.

Los métodos de HPLC, ya sean de intercambio iónico (cromatografía iónica) o de fase reversa, requieren etapas previas de "clean-up" de las muestras. Stennert y colaboradores[130] determinaron trigonelina en café mediante HPLC y cromatografía en capa fina, concluyendo que HPLC era una técnica de análisis más adecuada y rápida. Mazzafera[34] también llevó a cabo el análisis de trigonelina en café mediante HPLC en régimen isocrático con detección espectrofotométrica a 272 nm; Trugo, Macrae y Dick[131] analizaron alcaloides y trigonelina en muestras de café soluble por HPLC en régimen de gradiente con detección, también, a 272 nm.

En lo referente a los métodos de cromatografía iónica, Van Dujin y colaboradores[132] determinaron cafeína y trigonelina en café instantáneo mediante esta técnica analítica empleando detección espectrofotométrica a 254 nm y 280 nm.

En el presente trabajo, se ha puesto a punto un método de análisis de trigonelina mediante cromatografía iónica con detección espectrofotométrica.

I.7. METALES

El contenido mineral total del café verde supone un 4% en base seca. Generalmente, este contenido mineral total se ha expresado como cenizas tal y como establece el Centro de Investigación y Control de Calidad (C.I.C.C.)[112]; en la bibliografía se encuentran diversos trabajos en los que se determina el contenido en cenizas en café[13,133-134].

Para llevar a cabo el análisis de los elementos metálicos, la técnica analítica más empleada ha sido la espectrometría de absorción atómica. Con este método, diversos autores establecieron que el metal mayoritario presente en el café verde es el potasio; así, Clarke y Walter[14] determinaron este elemento en 18 muestras de café y Tserevitnov[15] confirmó dichos resultados. También por absorción atómica[13,17,135], se han determinado diversos elementos metálicos presentes en pequeñas cantidades en café, tales como manganeso, cobre, zinc, sodio, rubidio, hierro y calcio. Krivan y colaboradores[20] realizaron la determinación de hasta 20 elementos en muestras de café verde de distinto origen geográfico mediante absorción atómica con cámara de grafito y absorción atómica de llama.

En los últimos años, y debido a la rapidez en el análisis, se tiende a reemplazar la espectrofotometría de absorción atómica por la espectroscopía de emisión atómica con plasma inducido acoplado (ICP-AES). Koch y colaboradores[136] determinaron aluminio en muestras de té y café tanto por absorción atómica con cámara de grafito como por ICP.

En esta memoria, se lleva a cabo la determinación de elementos metálicos mediante ICP-AES. Previo al análisis, se ha de preparar la muestra de café verde realizando una destrucción de la materia orgánica por vía húmeda[137] utilizando una mezcla de ácidos sulfúrico y nítrico.

FUNDAMENTO DE LAS TECNICAS DE RECONOCIMIENTO DE PATRONES APLICADAS EN ESTA MEMORIA

II.1. INTRODUCCION A LOS METODOS DE RECONOCIMIENTO DE PATRONES

Antes de comenzar, conviene precisar algunos términos fundamentales como *Quimiometría*, *Reconocimiento de Patrones (Pattern Recognition)* o *Análisis Multivariante*, que constituyen la base de toda interpretación racional a partir de datos químicos multivariantes.

La *Quimiometría*, tal como se define en el *Journal of Chemometrics and Intelligent Laboratory System*, es la disciplina química que emplea los métodos matemáticos y estadísticos para diseñar o seleccionar procedimientos óptimos y experimentos, así como para proporcionar un máximo de información química a

partir del análisis de los datos químicos.

El *Reconocimiento de Patrones* es una rama de la inteligencia artificial, desarrollada a partir del final de la década de los sesenta, que proporciona una aproximación general a la resolución de problemas de análisis de grandes conjuntos de datos, que pueden aglomerarse en diversas clases. En pocas palabras, el planteamiento general del problema sería: "Dado un conjunto de objetos y una serie de medidas realizadas sobre esos objetos, ¿es posible encontrar y/o predecir una propiedad de los objetos, que no puede medirse directamente, pero que sabemos que está relacionada con las medidas mediante una relación desconocida?[138]. Su campo de aplicación no sólo atañe a las ciencias experimentales, también incluye otras disciplinas como son psicología, política, pedagogía, lingüística, medicina... Debido a sus numerosas aplicaciones, ha sufrido un espectacular desarrollo en los últimos años el cual se ha visto influido muy positivamente por las grandes posibilidades que actualmente ofrece la informática.

En química el término "objetos" puede cubrir desde elementos puros o compuestos hasta complicados productos industriales o naturales, y cada uno de ellos viene caracterizado por un conjunto de medidas. Por lo tanto entramos de lleno en el la interpretación de datos químicos multivariantes (cada objeto viene caracterizado no por una sino por un conjunto de medidas) y ello entraña el empleo, entre otras herramientas de trabajo, del *Análisis Multivariante.*

Se entiende por *Análisis Multivariante* a la rama de la estadística y el análisis de datos que estudia, interpreta y elabora el material estadístico sobre la base de un conjunto de n>1 variables que pueden ser de tipo cuantitativo, cualitativo o una

mezcla de ambos. La información obtenida en Análisis Multivariante es, por lo tanto, de carácter multidimensional, por lo que utiliza extensamente los métodos del álgebra lineal, cálculo numérico, geometría lineal y otras clases de geometrías[139]. No obstante, debemos precisar que el Análisis Multivariante trabaja con variables que siguen una determinada función de distribución, ya que pertenecen a la estadística.

El Reconocimiento de Patrones (RP) puede usar métodos de Análisis Multivariante o bien otros métodos que no se basen en la estadística, como por ejemplo, los algoritmos neuronales, para conseguir sus objetivos.

Dentro de esta disciplina, el término patrón se refiere a acontecimientos, objetos o entes que presentan características establecidas y definidas, las cuales pueden ser propiedades físicas o químicas. A un conjunto de patrones caracterizados por una relación común se le conoce como *clase.*

Desde el punto de vista quimiométrico, con estos métodos pueden predecirse características de muestras que no se observan directamente a partir de un conjunto de medidas químicas pero que están relacionadas con alguna propiedad química. En RP se trabaja siempre con patrones o "casos" los cuales, químicamente no son más que muestras descritas por un conjunto de variables o "descriptores", es decir, parámetros físicos o químicos que caracterizan dichas muestras.

El primer nivel de RP consiste en el establecimiento de fronteras entre clases de patrones y de reglas de clasificación para ubicar un patrón desconocido dentro de alguna de las clases previamente conocida. Es una técnica de "modelado duro"(*hard modelling*), ya que el patrón va a ser asignado a una determinada clase.

Si consideramos el segundo nivel, las fronteras se establecen no entre las categorías existentes, sino entre cada clase y el resto de los datos. Es decir, que un patrón no tiene necesariamente que pertenecer a alguna de las categorías. Las técnicas que se aplican en este segundo nivel son de "modelado suave"(*soft modelling*) y consideran la posibilidad de datos aberrantes (*outliers*) que no pueden clasificarse en ninguna de las clases conocidas de antemano. No consideraremos en esta memoria niveles superiores de RP[140].

II.1.1. Conceptos básicos en el Reconocimiento de Patrones

*** Espacio de modelos (Pattern Space):**

Cualquier patrón u objeto i que se estudia, es decir, cada muestra, va a venir descrito por un conjunto de c variables y se va a representar por un vector fila

$$\boldsymbol{x}_{ij} \quad (j=1 \ a \ c)$$

cuyas componentes serán $\{ x_{i1}, x_{i2}, ..., x_{ic} \}$, estos vectores son los que forman las filas de la matriz de datos **X** en el espacio de las variables y reciben el nombre de vectores patrón (*pattern vector*).

Por el contrario, si cambiamos de perspectiva y trabajamos en el espacio de los patrones tendremos r vectores columna

$$\boldsymbol{x}_{ij} \quad (i=1 \ a \ r)$$

cuyas componentes son los valores de una misma variable en cada una de las muestras $\{ x_{1j}, x_{2j}, ..., x_{rj} \}$ y constituyen las columnas de la matriz de datos. Esto

conduce a dos principales categorías de técnicas analíticas llamadas modo R y modo Q. Las técnicas en modo R tratan las relaciones entre las variables del experimento y examinan las dependencias entre las columnas de la matriz de datos, mientras que las técnicas en modo Q tratan las relaciones o agrupamientos entre los casos examinando las dependencias entre las filas de la matriz **X**.

Si se trabaja en modo R, el punto de partida es calcular la matriz de covarianzas **C**, que es una matriz simétrica obtenida según $\mathbf{C} = \mathbf{X}^T\mathbf{X}$. Normalmente, para evitar el predominio de descriptores que presenten valores elevados sobre otros de valores más pequeños, se suele suelen homogeneizar las escalas de las variables realizando el denominado autoescalado de las variables, según la expresión:

$$x'_{ij} = \frac{x_{ij} - \bar{x}}{s_j} \qquad \text{(II.1)}$$

donde:

$$\bar{x} = \frac{\sum_{i=1}^{r} x_{ij}}{r} \qquad s_j = \sqrt{\frac{\sum_{i=1}^{r} x_{ij} - \bar{x}}{r-1}} \qquad \text{(II.2)}$$

Así, la matriz de covarianza para los datos autoescalados se transforma en la matriz de correlación **R**, donde los elementos de la diagonal r_{ii} son siempre la unidad y el resto de elementos son los coeficientes de correlación $r_{ij} = r_{ji} < 1$.

Si se trabaja en modo Q, se comienza con una matriz de distancias en el espacio de los modelos.

*** Distancias en el espacio patrón:**[141]

Vamos a considerar dos vectores patrón $\mathbf{x_a}$ y $\mathbf{x_b}$ que pertenecen al espacio c-dimensional. Así: $\mathbf{x_a} = \{x_{a1}, x_{a2}, ..., x_{ac}\}$ $\mathbf{x_b} = \{x_{b1}, x_{b2}, ..., x_{bc}\}$

Las distancias que más suelen emplearse son las siguientes:

- *Distancia euclídea*:

$$d_{ab} = \sqrt{\sum_{i=1}^{c} (x_{ai} - x_{bi})^2} \qquad \text{(II.3)}$$

esta distancia es una de las más utilizadas en quimiometría.

- *Distancia de Minkowski*:

$$d_{ab} = \left(\sum_{i=1}^{c} |x_{ai} - x_{bi}|^k\right)^{\frac{1}{k}} \qquad \text{(II.4)}$$

donde k es un entero.

- *Distancia City Block o Manhattan*:

$$d_{ab} = \sum_{i=1}^{c} |x_{ai} - x_{bi}| \qquad \text{(II.5)}$$

es equivalente a la Minkowski para k=1

- *Distancia Hamming*:

$$d_{ab} = \sum_{i=1}^{c} XOR(x_{ai}, x_{bi}) \qquad \text{(II.6)}$$

esta distancia corresponde a la City Block cuando las variables se codifican

de forma binaria, escalando con valores 0 ó 1.

- *Distancia de Mahalanobis*:

$$d_{ab}^2 = (x_a - x_b)^T C^{-1} (x_a - x_b) \qquad \text{(II.7)}$$

donde **C** es la matriz de covarianzas, es decir, esta distancia corresponde a la distancia euclídea eliminando la posible correlación de las variables.

II.1.2. Preprocesado de los datos

El preprocesado de los datos sobre los que se va a trabajar consiste en manipulaciones algebraicas sobre la matriz de datos inicial, para una mejor realización de las técnicas de Reconocimiento de Patrones. Fundamentalmente, encontramos dos variantes: *escalado* y *ponderación*; en la primera, se equiparan los valores de las variables y se emplea para evitar los distintos rangos de magnitud de las medidas, la segunda diferencia entre los valores de descriptores que pertenecen a distintas clases y su uso permite ver las variables más importantes a la hora de distinguir entre categorías.

Dentro de las *técnicas de escalado* se distinguen[142]:

- *Centrado*: cuya representación de los nuevos valores viene dada por la expresión

$$x'_{ij} = x_{ij} - \bar{x}_j \qquad \textit{donde} \qquad \bar{x}_j = \frac{1}{r}\sum_i x_{ij} \tag{II.8}$$

- *Normalización de columnas*: basta con dividir los valores de las variables por la desviación estándar

$$x'_{ij} = \frac{x_{ij}}{s_j} \tag{II.9}$$

-*Autoescalado*: las variables originales sufren una transformación tipo Student. Así

$$x'_{ij} = \frac{x_{ij} - \bar{x}}{s_j} \tag{II.10}$$

- *Escalado del rango*: transformación del intervalo de valores entre un máximo de 1 y un mínimo de 0.

$$x_{ij} = \frac{x_{ij} - \min_j(x_{ij})}{\max_j(x_{ij}) - \min_j(x_{ij})} \tag{II.11}$$

- *Perfiles de fila*: consigue una frecuencia de aparición de cada variable.

$$x_{ij} = \frac{x_{ij}}{\sum_j x_{ij}} \qquad \text{(II.12)}$$

Criterios de ponderación y Selección de variables:

Antes de pasar a la enumeración de los distintos métodos, consideraremos algunas expresiones que se van a utilizar. Así,

$$\boldsymbol{m} = \frac{1}{r}\sum_i^r \boldsymbol{x}_i \qquad \text{(II.13)}$$

es el valor promedio de todos los patrones considerados. Expresión análoga se tiene para el patrón promedio de una determinada clase L.

$$\boldsymbol{m}^{(l)} = \frac{1}{r_l}\sum_i^{r_l} \boldsymbol{x}_i^{(l)} \qquad \text{(II.14)}$$

La matriz de varianza total es

$$\boldsymbol{S_T} = \sum_{i=1}^{r} (\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T \qquad \text{(II.15)}$$

para una variable, mientras que la covarianza entre las variables j y k es

$$(S_T)_{jk} = \sum_{i=1}^{r} (x_{ij} - m_j)(x_{ik} - m_k) \qquad \text{(II.16)}$$

La matriz intraclase, que mide la dispersión dentro de cada clase se define como

$$S_w = \sum_{l=1}^{L} \sum_{i=1}^{r_l} (x_i^{(l)} - m^{(l)})(x_i^{(l)} - m^{(l)})^T \qquad \text{(II.17)}$$

cuando es entre dos variables j y k se tiene:

$$(S_w)_{jk} = \sum_{l=1}^{L} \sum_{i=1}^{r_l} (x_{ij} - m_j)(x_{ik} - m_k) \qquad \text{(II.18)}$$

De forma análoga, la matriz de varianza entre clases para una variable y para dos variables j y k es:

$$S_B = \sum_{l=1}^{L} r_l (m^{(l)} - m)(m^{(l)} - m)^T \qquad \text{(II.19)}$$

$$(S_B)_{jk} = \sum_{l=1}^{L} r_l (m_j^{(l)} - m_j)(m_k^{(l)} - m_k) \qquad \text{(II.20)}$$

A continuación, citaremos algunos de los criterios más habituales de ponderación considerando las clases 1 y 2:

- *Ponderación de la varianza*[143]:

$$WV_j = \frac{r_1(m_j^{(1)} - m_j)^2 + r_2(m_j^{(2)} - m_j)^2}{S_j^{(1)^2} + S_j^{(2)^2}} \qquad \text{(II.21)}$$

mientras mayor sea WV más discriminante será la variable j.

- *Pesos de Fisher*[144]:

$$FW_j = \frac{(m_j^{(1)} - m_j^{(2)})^2}{S_j^{(1)^2} + S_j^{(2)^2}} \qquad \text{(II.22)}$$

es un criterio más utilizado que el anterior.

- *Ponderación de Coomans*[145]:

Cuando se consideran más de dos clases es ventajoso el empleo del criterio siguiente:

$$g_j = \frac{|m_j^{(1)} - m_j^{(2)}|}{S_j^{(1)} + S_j^{(2)}} \qquad \text{(II.23)}$$

- *Método* Λ *de Wilks*[146-147]:

$$\Lambda = \frac{\det(S_w)}{\det(S_T)} \qquad \text{(II.24)}$$

los valores del parámetro van desde 1.0 (ningún poder discriminatorio) a 0.0 (poder discriminatorio perfecto).

II.2. METODOS DE VISUALIZACION DE DATOS

II.2.1. Análisis en Componentes Principales

El Análisis en Componentes Principales (PCA), también llamado Autoanálisis o transformación de Karhunen Loewe, comenzó en la primera década del presente siglo; fue en 1931 cuando se desarrolló para perfiles de comportamiento. Hasta 1950, no se aplicó a problemas de tipo químico. Es un procedimiento que puede aplicarse a cualquier conjunto de datos como técnica exploratoria con excelentes resultados[148-149]. PCA transforma las c variables originales, posiblemente correlacionadas entre sí, en otros nuevos c ejes llamados *componentes principales* o *PC's* mediante un giro en el espacio de r dimensiones. Los nuevos ejes son ortogonales entre sí y son combinaciones lineales de las variables originales. En este tipo de análisis, no se tiene en cuenta ningún modelo de distribución para las variables.

Estos PC's se van a generar sucesivamente, de forma que el primero explica la mayor parte de la varianza entre los datos y los siguientes explican cantidades decrecientes de la varianza residual. Por tanto, al convertir las variables originales en componentes principales las correlaciones entre los ejes quedan eliminadas y la mayor parte de la varianza, es decir, de la información contenida en los datos originales queda explicada por los primeros PC's.

En el sentido matemático, se parte de la matriz de datos $\mathbf{X}_{r,c}$; el primer paso es realizar un preprocesado (centrado, autoescalado de los datos originales, etc). La matriz de covarianzas viene dada por $\mathbf{C} = \mathbf{X}^T\mathbf{X}$. El análisis en componentes principales busca una matriz de similaridad (transformación ortogonal) $\mathbf{U}_{c,c}$ que

actúe sobre la matriz de datos $\mathbf{X_{r,c}}$ para originar otra matriz $\mathbf{Y_{r,c}}$, llamada matriz de scores, en la que los datos están referidos a los nuevos ejes o componentes principales:

$$\boldsymbol{Y_{r,c} = X_{r,c} U_{c,c}} \qquad \text{(II.25)}$$

La matriz $\mathbf{Y_{r,c}}$ debe cumplir:

$$\boldsymbol{Y_{r,c} Y_{r,c}^{T}} = \Lambda_{c,c} \qquad \text{(II.26)}$$

donde $\Lambda_{c,c}$ es la nueva matriz de covarianzas, que es una matriz diagonal y no debe confundirse con el parámetro Λ de *Wilks*.
Por tanto:

$$\boldsymbol{Y^{T}Y} = \boldsymbol{(XU)^{T}(XU)} = \boldsymbol{U^{T}CU} = \Lambda \qquad \text{(II.27)}$$

luego la expresión final que obtenemos es:

$$\Lambda = \boldsymbol{U^{T}CU} \qquad \text{(II.28)}$$

esta ecuación es la llamada *Transformación de Karhunen-Loewe.*

Se pretende encontrar los vectores columna de la matriz de transformación U, como ésta es ortogonal se cumple que $\mathbf{U^{T} = U^{-1}}$. De forma que multiplicando por U por la izquierda la expresión de Karhunen, se tiene: $\mathbf{U\,\Lambda = C\,U}$. Si reescribimos la misma expresión pero en lugar de en forma matricial lo hacemos vector a vector, podemos poner: $\mathbf{u_j\,\lambda_j\,I = C\,u_j}$ es una clásica ecuación de autovalores o

autovectores; es decir $\mathbf{u}_j$ ($\mathbf{C}$-λ_j $\mathbf{I}$) = 0 luego debe cumplirse que det ($\mathbf{C}$-$\lambda\mathbf{I}$) = 0 donde $\mathbf{I}$ es la matriz unidad. Lo cual conduce a un polinomio de grado c:

$$\lambda^c + a_1\lambda^{c-1} + a_2\lambda^{c-2} + \ldots + a_c = 0 \qquad \text{(II.29)}$$

debido a que la matriz $\mathbf{C}$ es simétrica las c soluciones del polinomio son reales y positivas. Los valores de λ_j son los elementos de la diagonal principal de la matriz Λ y son las varianzas de los datos referidas a los nuevos ejes. Se ordenan en sentido decreciente de los valores de λ_j, de forma que:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_c$$

la suma de todos los λ_j debe ser igual a la suma de las varianzas de las variables originales. A partir de estos autovalores calculamos los vectores columna $\mathbf{u}_j$ y con ellos queda ya calculada la matriz de transformación $\mathbf{U}$.

Las coordenadas de los datos en los nuevos ejes , Y, reciben el nombre de *scores*. Como $\mathbf{U}$ es una transformación ortogonal es posible expresar linealmente las variables en función de los PC's y viceversa. Así:

$$x_{ij} = a_{j1}PC_{i1} + a_{j2}PC_{i2} + \ldots + a_{jc}PC_{ic} \qquad \text{(II.30)}$$

Los coeficientes a_{jk} se denominan *loadings* e indican la contribución del componente principal PC_k a la variable original. Por tanto, los PC's calculados son combinaciones lineales de las variables originales y pueden no tener sentido físico.

Debido a que los primeros componentes principales contienen la mayor información sobre la variabilidad de los datos, siempre es posible descartar aquellos PC's que no contengan información significativa sobre los mismos, mediante ensayos adecuados. La contribución a la varianza total de cada PC viene dada por la expresión:

$$\%var = \frac{\lambda_j}{\sum_{j=1}^{c} \lambda_j} \qquad \text{(II.31)}$$

de forma que pueden escogerse los primeros PC's que expliquen la mayor parte de la varianza conduciendo a una disminución de la dimensionalidad.

Algunos criterios utilizados para seleccionar el número adecuado de componentes principales se enumeran a continuación:

- *Criterio de Kaiser*[150]:

Este es uno de los más sencillos y más empleados. Admite que los datos están autoescalados y considera componentes principales explicativos aquellos cuyos autovalores correspondientes son mayores que la unidad. Es decir $\lambda>1$.

- *Criterio de la proporción de la varianza explicada*[151]:

Se deja en cierto modo al criterio de cada uno, normalmente se escogen los PC's que expliquen hasta por lo menos un 70% de la varianza.

- *Criterio de las comunalidades*[151]:

Según este método, nos quedamos con aquellos componentes principales

cuyas *comunalidades* sean ≈0.8 nunca se escogen PC's de comunalidad menor de 0.6. Si se combina este criterio con el de Kaiser se obtienen muy buenos resultados.

- *Criterio de la función indicador (IND) de Malinowski*[152]:

La expresión de dicha función es la siguiente:

$$IND = \frac{RSD}{(c-f)^2} \qquad \text{(II.32)}$$

donde RSD es la desviación estándar relativa, cuya expresión es:

$$RSD = \sqrt{\frac{\sum_{i=f-1}^{c} \lambda_i}{r(c-f)}} \qquad \text{(II.33)}$$

en la que f representa los componentes principales seleccionados. El numerador de la expresión representa la varianza del error y el numerador es una medida de los grados de libertad. Esta desviación estándar relativa es una medida de cómo se reproduce la matriz original cuando en lugar de los c componentes principales se usan sólo f PC's.

El método consiste en ir variando f hasta que la representación gráfica de la función IND frente a f presente un mínimo, entonces ese valor de f es el número óptimo de PC's a seleccionar.

- *Criterio de la función F*[153]:

También es debido a Malinowski, la función F viene representada por

$$F = \frac{RSD_f^2}{RSD_{f+1}^2} \qquad \text{(II.34)}$$

según este criterio se van cogiendo componentes principales hasta que la varianza explicada por f PC's no sea significativamente mayor a la explicada por f+1 PC's.

- *Método de la validación cruzada*[154]:

Fue desarrollado por el químico sueco Wold a mediados de los años 70 y está basado en el algoritmo NIPALS (Nonlinear Iterative PArtial Least Squares), que es un método iterativo para calcular PCs y lo hace ideal para su aplicación en una computadora. La validación cruzada consiste en eliminar un determinado porcentaje de los datos en la matriz **X** y calcular los PC's. El criterio de bondad del ajuste consiste en calcular el valor PRESS (Predicted Residual Error Sum of Squares) que viene dado por

$$PRESS_f = \sum_{i=1}^{r} \sum_{j=1}^{c} (x_{ij} - x_{ij}^*)^2 \qquad \text{(II.35)}$$

donde x_{ij} son los datos eliminados de la matriz y x_{ij}^* los predichos empleando f PC's. Si por ejemplo eliminamos la cuarta parte de los datos los cálculos serían los siguientes:

1.- Eliminar el 25% de los datos de **X**.

2.- Calcular el primer PC empleando NIPALS en ausencia del 25% de los datos.

3.- Predecir los valores eliminados (x_{ij}^{*}).

4.- Restituir los datos eliminados y quitar ahora el 25% de otros diferentes y volver al paso 1 hasta un total de 4 veces (cualquier dato ha sido eliminado de la matriz alguna vez). Entonces calcular $PRESS_1$.

Repetir la secuencia calculando en el paso 2, dos, tres, etc componentes principales y calcular $PRESS_f$. Un $\mathbf{PC_f}$ se considerará significativo cuando $PRESS_f/PRESS_{f-1}$ sea menor que la unidad.

Una vez que han sido seleccionados los f componentes principales significativos, mediante alguno de los criterios indicados, podemos escribir las variables originales como:

$$x_j = a_{j1}PC_1 + a_{j2}PC_2 + \ldots + a_{jf}PC_f + e_j \qquad \text{(II.36)}$$

el término e_j describe el error de ajuste para la variable $\mathbf{x_j}$.

Como los f PC's son independientes, si aplicamos la ley de propagación de la varianza al modelo anterior, debido a esta independencia y a que $var(PC_k)=1$, se cumple que:

$$var(x_j) = a_{j1}^2 + a_{j2}^2 + ... + a_{jf}^2 + var(e_j) \qquad \text{(II.37)}$$

Como las variables originales suelen estar autoescaladas, var(x_j)=1. La fracción de varianza de la variable explicada por los f componentes principales se denomina *comunalidad* y su expresión es:

$$\sum_{k=1}^{f} a_{jk}^2 \qquad \text{(II.38)}$$

mientras más se aproxime a 1 la comunalidad, mejor vendrá descrito el sistema por los PC's escogidos.

II.2.2. Biplots

Una vez realizado el análisis en componentes principales, una forma muy efectiva de visualizar las relaciones entre casos y descriptores es el uso de los BIPLOTS[155], en los que se representan tanto los *scores* de los objetos como las contribuciones de las variables (*loadings*) con respecto a los dos primeros PC's, que son los que explican mayor porcentaje de varianza.

Las variables son vectores en el BIPLOT y los objetos son puntos. Asimismo, la distancia euclídea entre dos puntos del BIPLOT sería la distancia existente entre dos casos; por tanto, esta representación permite visualizar posibles agrupamientos de los casos así como la mayor o menor separación de las variables en el plano de los dos primeros componentes principales.

Así pues, la representación BIPLOT es una buena herramienta para establecer las tendencias de los objetos e intuir. En lo que respecta a los descriptores,

observando esta representación es posible establecer cuales son las variables que mejor discriminan entre las clases, ya que las que presenten mayores valores de *loadings* (tanto positivos como negativos) para el primer PC serán los más discriminantes; de igual forma, variables con valores de *loadings* próximos van a proporcionar el mismo tipo de información acerca de los objetos y descriptores cuyos *loadings* estén próximos al valor cero no aportarán mucha información.

II.3. RECONOCIMIENTO DE PATRONES NO SUPERVISADO

II.3.1. Análisis Cluster

Cuando queremos establecer relaciones de pertenencia de los casos o variables estudiadas a determinadas categorías, los métodos difieren según conozcamos *a priori* o no la existencia de tales clases. En el primer caso, se aplican técnicas de Reconocimiento de Patrones Supervisadas (las cuales explicaremos más adelante en este capítulo), y en el segundo, se emplean técnicas de Reconocimiento de Patrones No Supervisadas, como es el Análisis Cluster (CA).

Un conjunto de objetos o patrones puede ser normalmente agrupado en *clusters* o grupos; llegar a describir y localizar estos clusters ayuda a una mejor descripción de la estructura de los objetos, lo cual, aparte de ser útil en sí mismo, simplifica y resuelve muchos problemas de clasificación de patrones o casos. Se denomina Análisis Cluster al conjunto de métodos y técnicas que describen y localizan estas agrupaciones de acuerdo con su similaridad en el espacio patrón[156-157].

Así, la utilidad fundamental del CA puede ser:

a) Un mejor estudio de los casos analizando las causas intrínsecas de la agrupación de los mismos.

b) La muestra puede corresponder únicamente a una clase y los posibles grupos a formar pueden ser subclases de la misma clase.

c) Pueden desconocerse las categorías a las cuales pertenecen los objetos pero, en cambio, puede que de acuerdo con las variables seleccionadas, si se encuentran clusters naturales, éstos correspondan a clases naturales de tal forma que, en base a ellas, se puedan diseñar reglas de clasificación de futuros casos cuya categoría de pertenencia es desconocida.

Para llegar a agrupar los objetos en clases naturales, el análisis cluster utiliza el criterio de minimizar la desviación interna de los patrones de un mismo grupo, maximizando, por tanto, la distancia entre los diversos grupos.

Una vez considerado que el objetivo del análisis cluster consiste en encontrar agrupaciones naturales del conjunto de objetos, es preciso definir qué se entiende por agrupaciones naturales y, por tanto, con arreglo a qué criterio se puede afirmar que dos grupos son más o menos similares; así, si se forma un cluster éste no tiene por qué representar una clase y viceversa, una clase o categoría no siempre va a aparecer como un cluster aislado.

Todo esto nos lleva a establecer una medida de la similitud entre dos muestras, la forma más obvia de medir la similitud o divergencia entre dos casos es la distancia entre ambos. Por tanto, la manera más idónea de comenzar un CA es definir una apropiada distancia métrica, (diversas definiciones de distancias se dieron en el apartado II.1.1.).

Las cualidades que debe tener una adecuada distancia son:

- Dos patrones distintos deben tener distancia positiva.
- La distancia de un patrón consigo mismo ha de ser nula.
- Debe cumplir la propiedad conmutativa.
- La distancia debe ser invariante a rotaciones y transformaciones. Normalmente, el cumplimiento de esta condición se consigue autoescalando las variables.
- La distancia debe tener en cuenta la posible correlación de las variables.

Dicho esto, la distancia más idónea a escoger es la distancia de Mahalanobis, o bien la distancia euclídea si los datos han sido previamente autoescalados.

*** Tipos de técnicas de Análisis Cluster**

En los últimos años, ha crecido considerablemente el número de métodos diferentes para agrupar objetos y se han realizado estudios exhaustivos de recopilación de métodos cluster[158-162].

- *Métodos de reagrupamiento*:

Se considera que un método de análisis cluster es de *reagrupamiento* cuando habiendo determinado el número de clusters a formar, se van distribuyendo los objetos entre los diversos grupos de tal forma que se establecen iterativamente los centrotipos de los grupos minimizando la distancia de los patrones con respecto a su centrotipo (centroide ó un patrón que defina el centro del grupo), maximizando la distancia entre los distintos centrotipos determinados. Entre estas técnicas están el método de las K-medias y Fuzzy clustering.

Si se desconoce el número de clusters o clases esperable, primero se realiza un análisis cluster *jerárquico.*

- *Métodos jerárquicos*:

Son los métodos más utilizados a la hora de realizar un CA y su expresión visual más común es el *dendrograma.* Estos métodos tienen por objetivo agrupar clusters para formar uno nuevo hasta llegar a un solo grupo o bien separar clusters formando nuevos subconjuntos que salen del anterior.

Por tanto, los métodos jerárquicos se pueden subdividir a su vez en:

1) *Métodos aglomerativos*: se parte de n grupos, tantos como puntos había en el espacio patrón y se van uniendo hasta llegar a un cluster común que engloba todas las muestras, procediendo en cada nivel a fusionar aquellos dos grupos que sean más similares. Este concepto se puede apreciar en la siguiente figura.

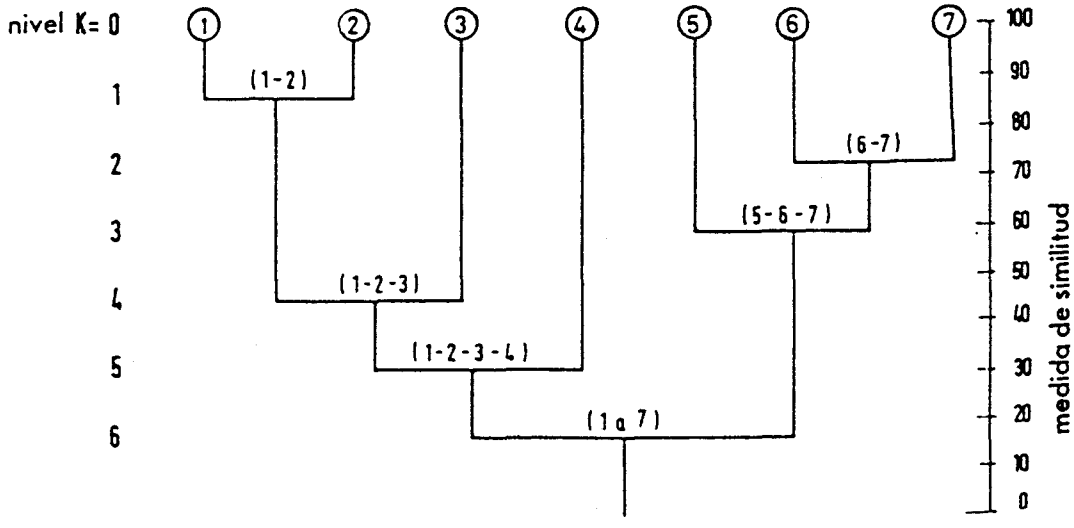

Figura 13. Aglomeración jerárquica de siete patrones en seis niveles.

2) *Métodos divisivos*: se parte de un grupo formado por todos los casos en el nivel K=0, en el siguiente nivel K=1 se obtienen dos grupos repartiendo los objetos en base a maximizar sus divergencias. Se va procediendo de esta forma sucesivamente hasta conseguir n clusters correspondientes a todos los puntos del espacio. En la figura 14, se ofrece un ejemplo de este método.

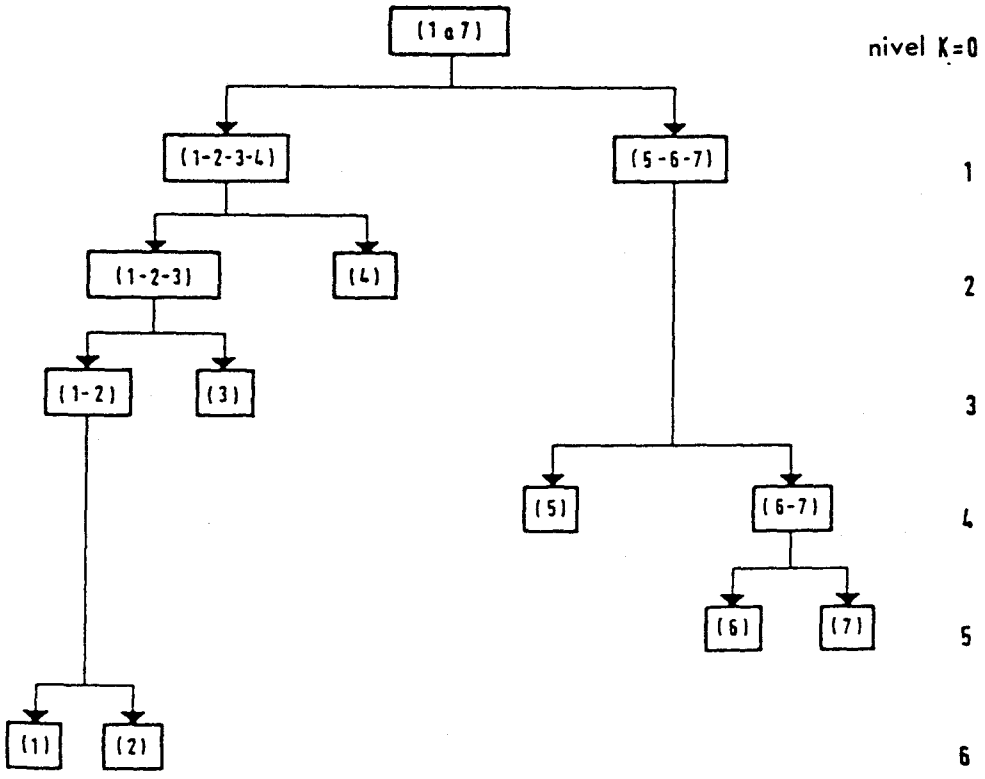

Figura 14. División jerárquica para siete patrones en seis niveles.

Vamos a centrarnos en el análisis cluster jerárquico, puesto que es el más empleado. Una vez seleccionada la distancia, debe establecerse cuidadosamente una distancia umbral d_0, ya que de ella va a depender la formación de clusters. Si d_0 es muy grande se formará un solo grupo, mientras que si por el contrario esta distancia umbral es muy pequeña, puede darse la posibilidad de que cada muestra forma un cluster. Así, un valor intermedio de d_0 proporcionará un número determinado de clusters. Este hecho queda ilustrado en la figura 15.

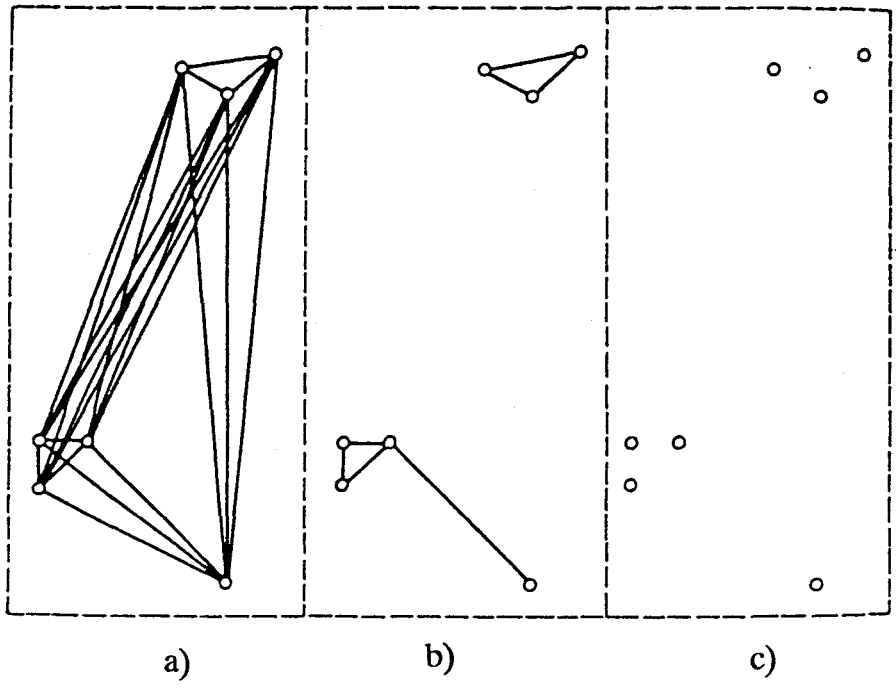

Figura 15. Efecto del umbral en la formación de clusters. Las líneas recogen los patrones cuya distancia es menor que d_0.
a) d_0 grande b) d_0 intermedio c) d_0 pequeño

El procedimiento para realizar un análisis cluster jerárquico aglomerativo comienza por calcular la matriz de distancias **D** de dimensión rxr que es simétrica, sus elementos diagonales son todos nulos y el resto son números positivos. Una vez que se ha calculado la matriz **D** cada ciclo del agrupamiento iterativo tiene lugar en los siguientes pasos:

1.- Buscar los dos objetos más próximos, cuyo elemento d_{ij} de la matriz **D** sea menor.

2.- Agrupar los objetos i y j, en un cluster (ij).

3.- Actualizar la matriz **D**, eliminando las distancias con respecto a i y a j y calculando las distancias nuevas de los demás objetos con respecto al centroide del cluster (ij).

4.- Con la nueva matriz (r-1)x(r-1) repetir el procedimiento entero hasta haber realizado r-1 agrupamientos.

Otra cuestión importante a la hora de llevar a cabo un CA jerárquico es la elección de la *Regla de amalgamación* para la formación de clusters. La distancia entre dos clusters H y K puede escribirse mediante una ecuación formalmente idéntica aunque con diferentes parámetros, dependiendo éstos de la regla de amalgamación seleccionada, llamada ecuación de Lance-Williams[163]:

$$d(H,K) = a_I d(H,I) + a_J d(H,J) + b\, d(I,J) + g(d(H,I) - d(H,J)) \qquad \text{(II.39)}$$

La ecuación de Lance-Williams permite el cálculo de la distancia entre dos clusters H y K, d(H,K), donde K representa el cluster formado más recientemente (I,J). De manera más conveniente la distancia d(H,K) debería escribirse d(H,(I,J)). Los coeficientes a_I, a_J, b y g son diferentes dependiendo del método de agrupar los clusters, como se indica en la tabla siguiente:

Regla de amalgamación	Coeficientes de la ecuación de Lance-Williams			
	a_I	a_J	b	g
Distancia mínima[164]	0.5	0.5	0	-0.5
Distancia máxima[165]	0.5	0.5	0	0.5
Distancia promedio ponderada[166]	n_I/n_K	n_J/n_K	0	0
Distancia promedio no ponderada[167]	0.5	0.5	0	0
Centroide[168]	n_I/n_K	n_J/n_K	$-n_In_J/n^2_K$	0
Mediana[168]	0.5	0.5	-0.25	0
Método de Ward[169]	n_{HI}/n_{HK}	n_{HJ}/n_{HK}	$-n_H/n_{HK}$	0

Tabla 3. Coeficientes de la ecuación de Lance-Williams según la regla de amalgamación para siete modalidades distintas. El valor n_L representa el número de objetos en el cluster L (L=I, J, H, K) y el valor n_{LM} la suma $n_L + n_M$.

Los resultados obtenidos por el método de Ward son óptimos porque conduce a formación de los clusters más homogéneos posibles, aunque presenta la tendencia

de formar un número mínimo de éstos.

Métodos no jerárquicos

Vamos a introducir de manera breve dos procedimientos de reagrupamiento muy utilizados en la bibliografía que ya han sido mencionados en los comienzos de este epígrafe, el método de las K medias y el agrupamiento borroso (Fuzzy clustering)

Método de las K medias[170]

El método de las K medias, también llamado de la partición óptima sigue el siguiente algoritmo: Inicialmente, se selecciona el número de clusters a formar. Los datos se dividen al azar en el número apropiado de grupos y cada uno de ellos se representa por su centrotipo (centroide o miembro típico del grupo). Entonces se calcula la distancia media dentro de los grupos como el promedio de las distancias entre cada elemento del grupo y su centrotipo. Los objetos se transfieren y reagrupan de un grupo a otro de modo que se minimice esta distancia media entre grupos. El resultado conduce a una serie de clusters muy homogéneos.

Método del agrupamiento borroso[171]

En este método la idea fundamental es el concepto de Zadeh de pertenencia parcial a un conjunto[172]. Así, se considera que:

a) Una muestra (vector patrón) puede pertenecer simultáneamente a más de un grupo, con un grado de pertenencia en cada cluster

particular representado por un número en el intervalo (0,1).

b) La pertenencia total de una muestra determinada a todos los clusters es la unidad. El Algoritmo de agrupamiento borroso usa la notación $u_{ik} = u_i(\mathbf{x_k})$ para representar el grado de pertenencia del vector patrón $\mathbf{x_k}$ en el cluster i. Las dos condiciones anteriores pueden entonces expresarse como:

$$\begin{aligned} &a)\ 0 \le u_{ik} \le 1 \quad \forall i, k \\ &b)\ \sum_{i=1}^{n} u_{ik} = 1 \quad \forall k \end{aligned} \tag{II.40}$$

donde n es el número de clusters, que debe conocerse *a priori.*

El algoritmo es un proceso iterativo. En la primera iteración uno divide los objetos en los correspondientes grupos y calcula los centrotipos (centroides). Después, se calcula la distancia de cada uno de los objetos a cada centrotipo. El grado de pertenencia de cada objeto a cada grupo viene asignado por la ecuación:

$$u_{ik} = \frac{1}{\sum_{j=1}^{n} \left(\frac{d_{ik}}{d_{jk}} \right)^2} \tag{II.41}$$

donde d_{jk} denota la distancia entre el patrón $\mathbf{x_k}$ y el centrotipo del cluster j.

El paso siguiente es calcular nuevos centrotipos v_i según:

$$v_i = \frac{\sum_{k=1}^{r} u_{ik}^2 x_k}{\sum_{k=1}^{r} u_{ik}^2} \qquad \text{(II.42)}$$

Estos nuevos centrotipos se usan para calcular de nuevo el grado de pertenencia de cada muestra a cada grupo, que a su vez se emplea en el cálculo de nuevos centrotipos. Las iteraciones continuan hasta que se consigue algún criterio de finalización. Uno muy usado consiste en parar las iteraciones cuando la separación máxima en la posición de los centrotipos en dos iteraciones consecutivas no supera un umbral previamente establecido.

Para concluir, recordemos que la hipótesis de cualquier método de RP es que objetos similares con respecto a una propiedad determinada se encontrarán próximos entre sí en el espacio patrón y formarán clusters. Si los clusters están bien definidos, y dichas agrupaciones no contravienen nuestra intuición y experiencia química, podemos adscribirlos a las diversas clases que constituyen el conjunto de datos.

II.4. RECONOCIMIENTO DE PATRONES SUPERVISADO.

II.4.1. Conceptos generales.

Las técnicas de Reconocimiento de Patrones Supervisadas suponen que se conoce *a priori* el número de clases así como la pertenencia a las mismas de cada uno de los miembros del conjunto de datos. Este recibe el nombre de conjunto conocido (*known set*). La finalidad es diseñar y aplicar reglas de clasificación para predecir las clases a las que pertenecen un conjunto de muestras desconocidas (*unknown set*). El diseño del clasificador recibe el nombre de entrenamiento (*training*). Durante el entrenamiento, se utilizan dos subconjuntos del conjunto total, el conjunto de entrenamiento (*training set*) y el conjunto de evaluación o predicción (*evaluation, prediction or test set*)[173]. Las reglas de clasificación se desarrollan empleando los objetos contenidos en el conjunto de entrenamiento y estas se comprueban utilizando el conjunto de predicción. El porcentaje de objetos del conjunto de entrenamiento clasificados correctamente recibe el nombre de tasa de reconocimiento (*recalling or recognition rate*). El clasificador se comprueba con los objetos del conjunto de predicción y el porcentaje de casos correctamente clasificados recibe el nombre de facultad predictora (*prediction ability*). Así pues, el conjunto total de objetos se divide en el conjunto de entrenamiento, el cual suele estar constituido por un 75% de los casos del conjunto de datos conocido inicial, y el conjunto de ensayo, formado por el 25% restante[174]. La selección de estos porcentajes se lleva a cabo de manera aleatoria pero aplicándose a los objetos de cada clase (es decir tomar estocásticamente el 75% de objetos en la clase I, la II, etc para formar el conjunto de entrenamiento). Los resultados de la eficacia en el

reconocimiento y la facultad predictiva se dan como valores promedio de diez repeticiones del procedimiento para diferentes constituciones aleatoriamente seleccionadas de los conjuntos de entrenamiento y evaluación.

Otra modalidad para operar a la hora de determinar la facultad predictiva y recognitiva del clasificador es el llamado "método de dejar uno fuera" (*leave-one-out method*)[175]. En esta técnica, se parte del conjunto conocido completo y a continuación, se coge aleatoriamente uno de los objetos y se la "deja fuera", es decir, él solo va a constituir el conjunto de ensayo y el resto de casos formarán el conjunto de entrenamiento; se desarrolla la regla de clasificación y se contabilizan los casos que hayan quedado clasificados correctamente. Seguidamente, se repite el proceso de forma iterativa cada vez dejando un objeto distinto fuera (en el conjunto de ensayo); así se sigue, finalizando el proceso cuando se llega al último caso. La eficacia predictiva se calcula contabilizando los éxitos y fallos en clasificar cada patrón.

Los métodos basados en el aprendizaje supervisado pueden dividirse en paramétricos y no paramétricos. Los primeros suponen que las funciones de densidad de probabilidad de las variables son conocidas o pueden estimarse, y se aplican ensayos estadísticos en la discriminación. Si pueden hacerse suposiciones razonables acerca de estas funciones de distribución, es posible estimar la probabilidad de una clasificación correcta o al menos el riesgo de fallo en la aplicación del clasificador. Desafortunadamente, la mayor parte de las veces no podemos hacer suposiciones dado el limitado conjunto de datos para el análisis. Por

tanto, los métodos no paramétricos se emplean con mayor profusión. Estos últimos ignoran las funciones de distribución de las variables y realizan la clasificación empleando en general medidas de proximidad entre patrones. Seguidamente presentaremos una exposición de los métodos utilizados en esta memoria que pertenecen a una de estas dos categorías.

II.4.2. Métodos paramétricos.

Antes de comentar los distintos métodos utilizados, indicaremos los fundamentos estadísticos de la clasificación supervisada, que serían los métodos puramente bayesianos.

Consideramos el caso particular de dos clases I y II, cada caso x_{ij} está representado por un vector fila cuyas componentes son los contenidos en cada variable, para mayor simplicidad vamos a denominarlo **x**. Se denomina probabilidad *a priori* de la clase I, P(I), a la probabilidad de que sin conocer el valor de las variables x, un caso pertenezca a la clase I. Como solo consideramos la existencia de dos clases, se cumple que P(I) + P(II) = 1. Se define *probabilidad condicional*, P(**x**/I), a la probabilidad de que los valores de las variables de un caso que pertenece a la clase I sean los componentes de x. Según los axiomas de la probabilidad, se tiene que:

$$P(x) = P(x/I)P(I) + P(x/II)P(II) \qquad \text{(II.43)}$$

Por otro lado, se define probabilidad *a posteriori* a la probabilidad de que el caso

pertenezca a la clase I conocido el valor de sus variables x. Según el teorema de Bayes:

$$P(I/x) = P(x/I)\ P(I)/P(x) \qquad \text{(II.44)}$$

de forma análoga:

$$P(II/x) = P(x/II)\ P(II)/P(x) \qquad \text{(II.45)}$$

de acuerdo con esta aproximación estadística[176], se puede considerar que la regla óptima de decisión, en el sentido que minimiza la probabilidad de error, es aquella que asigna el caso a la clase para la cual la probabilidad *a posteriori* es mayor, tal que si $P(I/x) \geq P(II/x)$, el caso se asigna a la clase I.

En definitiva, el problema fundamental para un método bayesiano consiste en la determinación de los parámetros de la función de probabilidad P(x/I) y P(x/II) en base al conjunto de evaluación. En la mayoría de los casos, tal función no se conoce y es preciso efectuar su estimación. Esta función de probabilidad en muchos casos es multimodal, por ello, generalmente se recurre a otros sistemas no bayesianos como los métodos no paramétricos y los métodos paramétricos basados en el cálculo de las denominadas funciones discriminantes, como veremos a continuación.

*** Análisis Discriminante Lineal (LDA)**

Este es uno de los métodos supervisados más aplicado, el cual estima la probabilidad *a posteriori* de que un objeto pertenezca a una determinada categoría, creando funciones de decisión que separan las clases a las que pertenecen los objetos a partir del conjunto de entrenamiento.

A continuación, vamos a describir el análisis discriminante lineal que, como su propio nombre indica, consiste en encontrar las mencionadas funciones discriminantes *lineales* admitiendo que la separación entre categorías pueda realizarse mediante hiperplanos.

Esta técnica se basa en los trabajos desarrollados por Fisher a finales de los años treinta[177-178]. El procedimiento es encontrar combinaciones lineales de las variables, que van a ser las denominadas funciones discriminantes, tal que sea máximo el cociente:

$$F = \frac{S_B}{S_w} \qquad \text{(II.46)}$$

Para realizar el cálculo de las funciones discriminantes, previamente, se seleccionan las variables con mayor poder discriminatorio utilizando para ello criterios de selección sucesiva (*Stepwise Criteria*) de forma que se incluyen solo aquellas R variables que son verdaderamente relevantes para la discriminación[179].

Así, definíamos Λ *de Wilks*[146] como:

los valores de Λ van desde 1 (ningún poder discriminante) hasta 0 (máximo poder discriminatorio).

$$\Lambda = \frac{|S_w|}{|S_B|} \tag{II.47}$$

Análogamente, se define Λ *parcial de Wilks* como la medida de Λ para la contribución de una variable determinada en el modelo:

$$\Lambda_{parcial} = \frac{\Lambda(\textit{después de añadir la variable})}{\Lambda(\textit{antes de añadir la variable})} \tag{II.48}$$

este valor puede convertirse en un valor de F de la forma:

$$F = \frac{N-k-R}{k-1} \cdot \frac{1-\Lambda_{parcial}}{\Lambda_{parcial}} \tag{II.49}$$

donde N es el número total de objetos, k es el número de clases y R el número de variables consideradas.

Si se van seleccionando las variables más discriminantes por un método "hacia delante" (*forward stepwise*), al comienzo, en el paso 0, no hay variables en el modelo y, por definición, Λ=1. Después de añadir la primera variable, se cumple que $\Lambda_{parcial}=\Lambda$.

De todas las variables, se selecciona como primera aquella cuyo valor de F tiene un menor nivel de significación (*p-level*). Se continua añadiendo variables hasta que F presente un *p-level* por encima de un valor especificado de antemano (por ejemplo 0.05). Existe un método análogo pero esta vez se parte de todas las variables y se van eliminando una, atendiendo a su menor poder discriminante, este es el *backward selection.*

Una vez seleccionadas las R variables más significativas, se procede al cálculo de las funciones discriminantes, también denominadas variables canónicas[180] como combinación lineal de las R variables discriminantes de modo que las clases se observen lo más separadas posibles y se reduzca la dimensionalidad sin pérdida de diferenciación entre dichas clases. Estas funciones constituirán los nuevos ejes donde mejor se contemplarían las diferencias entre categorías.
Así, la expresión analítica de tales funciones es la siguiente:

$$f(x_1,x_2,\ldots,x_R) = v_1x_1 + v_2x_2 + \ldots + v_Rx_R = v^Tx \qquad \text{(II.50)}$$

En definitiva, estas funciones discriminantes son rotaciones en el espacio de las variables reducidas que generan combinaciones lineales que deben caracterizar bien una clase.

El objetivo de la técnica es calcular los vectores $\mathbf{v}_T$ para los que sea máximo el cociente:

$$F = \frac{v^TS_Bv}{v^TS_wv} \qquad \text{(II.51)}$$

es decir, que los promedios entre cada categoría sean lo más diferentes entre sí y, por otra parte, que cada clase tenga la menor desviación interna.
Un ejemplo claro del problema a resolver se visualiza en la figura 16, para un caso bidimensional en el que ni la variable $\mathbf{x}_{1j}$ ni $\mathbf{x}_{2j}$ discriminan aisladamente las clases 1 y 2, ya que en ambos supuestos habría patrones mal clasificados. En cambio, la

nueva variable $R_j = v_1x_{1j} + v_2x_{2j}$ permite la máxima separación de ambas clases, minimizando al mismo tiempo la desviación intraclase.

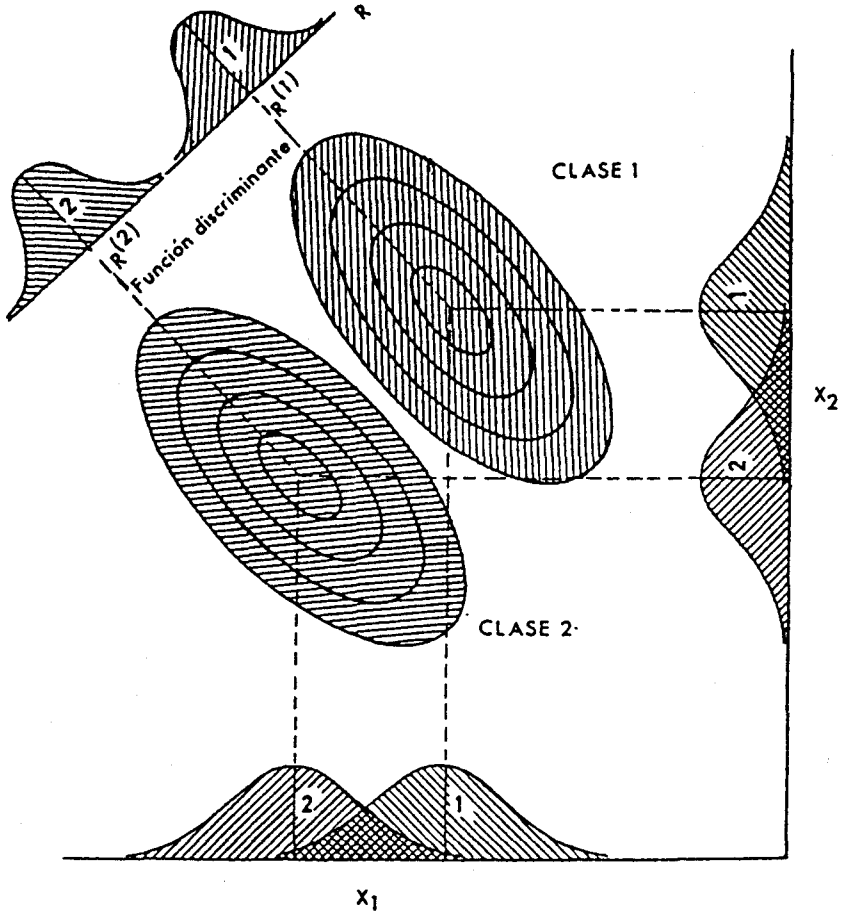

Figura 16. Desolapamiento de las funciones de distribución de probabilidad *a posteriori* para dos clases, empleando la función discriminante como nuevo eje.

El vector v para el cual se maximiza la F será aquel para el cual la derivada primera sea nula:

$$\frac{\partial F}{\partial v} = \frac{\partial (v^T S_B v / v^T S_w v)}{\partial v} = 0 \qquad \text{(II.52)}$$

Realizando la derivada, se tiene:

$$\frac{\partial F}{\partial v} = \frac{S_B v (v^T S_w v) - (v^T S_B v) S_w v}{(v^T S_w v)^2} = 0 \qquad \text{(II.53)}$$

luego debe cumplirse que:

$$S_B v (v^T S_w v) - (v^T S_B v) S_w v = 0 \qquad \text{(II.54)}$$

es decir, que:

$$S_B v - F S_w v = 0 \qquad \text{(II.55)}$$

llegando a la expresión:

$$S_w^{-1} S_B v = F v \qquad \text{(II.56)}$$

Es la típica ecuación de autovalores y autovectores. Así pues, habrá un número determinado de autovalores distintos de cero que corresponderán al número de funciones discriminantes para separar las categorías. Ese número, normalmente, es el mínimo entre el número de clases menos uno y el número de variables es decir: min (k-1,R).

Por otra parte, se denomina *discriminant score* de un objeto $\mathbf{x_i}$ que pertenece a la clase k a la expresión:

$$DS_i^{(k)} = v^T x_i^{(k)} \qquad \text{(II.57)}$$

de esta forma, el objeto $\mathbf{x_i}$ se puede clasificar según su posición con respecto al nuevo eje f, comparando con los centroides de las clases existentes.

Supongamos dos clases 1 y 2 cuyos centroides son respectivamente:

$$\overline{DS}^{(1)} \quad y \quad \overline{DS}^{(2)}$$

entonces, el objeto $\mathbf{x_i}$ pertenecerá a la clase 1 si se cumple que:

$$|DS_i - \overline{DS}^{(1)}| < |DS_i - \overline{DS}^{(2)}| \qquad \text{(II.58)}$$

y en caso contrario, pertenecerá a la clase 2.

Esta es una manera de clasificar objetos pero, generalmente en el análisis discriminante lineal, se suelen calcular reglas de clasificación basadas en la estimación de las probabilidades *a posteriori* de la pertenencia de un objeto a una clase empleando estimación Bayesiana.

De acuerdo con las definiciones de probabilidad *a priori*, probabilidad condicional y probabilidad *a posteriori*, expresadas en el apartado II.4.2. de este capítulo y según el teorema de Bayes, se tiene que la probabilidad *a posteriori*, para un objeto $\mathbf{x_i}$, de pertenecer a una clase c_k es:

$$P(c_k/x_i) = \frac{P(x_i/c_k)P(c_k)}{\sum_{j=1}^{k} P(x_i/c_j)P(c_j)} = \frac{P(x_i/c_k)P(c_k)}{P(x_i)} \qquad \text{(II.59)}$$

Normalmente, si se desea establecer la regla de clasificación por el método bayesiano, se suelen hacer tres suposiciones:

1) Las características x_{ij} son estadísticamente independientes, lo cual significa que la matriz de covarianzas es diagonal con todos los elementos de la diagonal iguales.

2) Las distribuciones $P(x_{ij}/c_k)$ siguen la ley de Gauss.

3) Estas distribuciones tienen la misma varianza y solo se diferencian en sus valores promedio $\mu_j(I)$ y $\mu_j(II)$.

Así:

$$P(x_i/c_1) = \frac{1}{(2\pi)^{R/2}\sigma_1\sigma_2\ldots\sigma_R} e^{-\frac{1}{2}\left[\left(\frac{x_{i1}-\mu^{(1)}}{\sigma_1^{(1)}}\right)^2+\ldots+\left(\frac{x_{iR}-\mu_R^{(1)}}{\sigma_R^{(1)}}\right)^2\right]} \qquad \text{(II.60)}$$

luego se puede reescribir la expresión anterior como:

$$P(x_i/c_1) = \frac{1}{(2\pi)^{R/2}(\det\Sigma)^{1/2}} e^{-\frac{1}{2}[(x_i-\mu^{(1)})^T\Sigma^{-1}(x_i-\mu^{(1)})]} \qquad \text{(II.61)}$$

donde:

$$\Sigma^{-1} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ . & . & . & . \\ 0 & 0 & 0 & \sigma_R^2 \end{bmatrix} \qquad \text{(II.62)}$$

en general:

$$P(x_i/c_1) = \frac{1}{(2\pi)^{R/2}(\det\Sigma)^{1/2}} e^{-\frac{1}{2}[(x_i-\mu^{(k)})^T\Sigma^{-1}(x_i-\mu^{(k)})]} \qquad \text{(II.63)}$$

Tomando como función discriminante:

$$g(x_i)^{(k)} = \ln P(c_k) + \ln P(x_i/c_k) \qquad \text{(II.64)}$$

sustituyendo $P(x_i/c_k)$ por su expresión, queda:

$$g(x_i)^{(k)} = \ln P(c_k) - \frac{1}{2}(R\ln 2\pi + \ln(\det\Sigma)) - \\ -\frac{1}{2}(x_i^T-\mu^{(k)})^T\Sigma^{-1}(x_i-\mu^{(k)}) \qquad \text{(II.65)}$$

los dos primeros términos de la expresión anterior son independientes de la clase k, ya que se admite que las probabilidades *a priori* son iguales $P(c_j) = P(c_k)$ y Σ es común. Por tanto:

$$g(x_i)^{(k)} = -\frac{1}{2}(x_i-\mu^{(k)})^T\Sigma^{-1}(x_i-\mu^{(k)}) = -\frac{1}{2}D_i^{(k)^2} \qquad \text{(II.66)}$$

donde $D_i^{(k)}$ es la distancia de Mahalanobis del patrón x_i al centroide de la clase k. operando, se tiene:

$$g(x_i)^{(k)} = -\frac{1}{2}x_i^T\Sigma^{-1}x_i + x_i^T\Sigma^{-1}\mu^{(k)} - \frac{1}{2}\mu^{(k)T}\Sigma^{-1}\mu^{(k)} \qquad \text{(II.67)}$$

el primer término de la expresión carece de poder discriminante, por lo que ésta se puede escribir como:

$$g(x_i)^{(k)} = x_i^T\Sigma^{-1}\mu^{(k)} - \frac{1}{2}\mu^{(k)T}\Sigma^{-1}\mu^{(k)} \qquad \text{(II.68)}$$

a esta expresión también se le denomina *classification score*, así:

$$CS_i^{(k)} = x_i^T\Sigma^{-1}\mu^{(k)} - \frac{1}{2}\mu^{(k)T}\Sigma^{-1}\mu^{(k)} \qquad \text{(II.69)}$$

En la práctica, se emplea S_w es decir la varianza intraclase como una mejor estimación de Σ y, además, si llamamos $\mathbf{m}^{(k)}$ a la expresión:

$$m^{(k)} = \frac{1}{r_k}\sum_{i=1}^{r_k} x_i^{(k)} \qquad \text{(II.70)}$$

el *classification score* queda:

$$CS_i^{(k)} = x_i^T S_w^{-1} m^{(k)} - \frac{1}{2}m^{(k)T} S_w^{-1} m^{(k)} \qquad \text{(II.71)}$$

esta es la *función de clasificación de Fisher* por la que cada objeto x_i se asigna a

la clase k para la cual la función de clasificación $SC_i^{(k)}$ presente el mayor valor.

*** SIMCA (*Soft Independent Modelling of Class Analogy*)**

Esta técnica supervisada de reconocimiento de patrones es un procedimiento blando de modelización, ya que puede que algún no se clasifique dentro de alguna de las categorías existentes, es decir, lleva a cabo el nivel 2 del reconocimiento de patrones. Así, el método SIMCA[140,154,181-190] asocia a cada clase conocida un modelo matemático individual de tal forma que cada patrón desconocido se compara con cada uno de los modelos establecidos para comprobar si se ajusta a alguno de ellos, perteneciendo entonces a dicha categoría, o bien pertenece a otro tipo de distribución constituyendo en ese caso un *outlier* con respecto a las clases conocidas en el conjunto de entrenamiento.

SIMCA divide la matriz de datos original en submatrices correspondientes a cada una de las k categorías y realiza un análisis en componentes principales por separado para cada una de las clases del conjunto de entrenamiento[191]:

$$\boldsymbol{X}^{(k)} = \boldsymbol{Y}^{(k)}\boldsymbol{U}^{T(k)} = \boldsymbol{y}_1^{(k)}\boldsymbol{u}_1(k) + \boldsymbol{y}_2^{(k)}\boldsymbol{u}_2^{(k)} + \ldots \qquad \text{(II.72)}$$

mediante el método de la validación cruzada y empleando el algoritmo NIPALS, explicado en el apartado II.2.1., se seleccionan f PC's explicativos de la mayor parte de la varianza de los datos para cada clase; así, se tiene:

$$X^{(k)} = \sum_{i=1}^{f(k)} y_i^{(k)} u_i^{T(k)} + E \qquad \text{(II.73)}$$

donde **E** es la matriz error que contabiliza los residuales de cada caso, es decir, la diferencia entre los datos originales y la estimación del modelo.
La varianza residual para la clase K con r_k objetos es:

$$s_k^2 = \frac{SS}{(r_k - f^{(k)} - 1)(v - f^{(k)})} \qquad \text{(II.74)}$$

donde v es el número de variables originales y ss viene dada por la expresión:

$$SS = \sum_{i=1}^{r_k} \sum_{j=1}^{f} e_{ij}^2 \qquad \text{(II.75)}$$

Para un elemento cualquiera, se tiene:

$$x_{kl}^{(k)} = \sum_{a=1}^{f(k)} y_{ka}^{(k)} u_{al}^{(k)} + \varepsilon_{kl} \qquad \text{(II.76)}$$

se calcula el valor estimado para x_{kl} según el modelo PCA realizado para la clase k y se evalúa el error de ajuste. Esto se repite para cada uno de los elementos en cada una de las clases.

Una vez evaluados los errores, la varianza para el elemento q será s_q^2.
Si s_q^2 es de la misma magnitud que s_k^2, el objeto q se considera miembro de la clase k. El criterio F de Fisher proporciona una medida cuantitativa para dicha clasificación:

$$s_q^2 = \frac{\sum_{j=1}^{v} e_{qj}^2}{v - f^{(k)}} \qquad \text{(II.77)}$$

$$F = \frac{s_q^2}{s_k^2} \qquad \text{(II.78)}$$

Al emplear un criterio F para establecer la pertenencia de un objeto a la clase considerada, el método SIMCA es evidentemente paramétrico.

II.4.3. Métodos no paramétricos

* Método de los K vecinos más próximos (KNN)

Este es un método empírico en el que un caso desconocido se clasifica de acuerdo con el "voto" de clase de sus K vecinos más próximos en el conjunto de aprendizaje de un espacio multidimensional[192].

Así, tenemos un objeto x_i, un número de muestras totales n y un número K de vecinos que van a utilizarse para clasificar. La elección del número K óptimo puede realizarse según la recomendación de Duda y Hart[193], quienes afirman que K debe ser igual a $n^{1/2}$, otro criterio es seleccionar K por un procedimiento de "dejar uno fuera", generalmente, K es un número impar. Para un conjunto de datos de tamaño limitado K=1 puede ser una buena elección, sin embargo en el caso de tener clases solapadas K=3 ó K=5 mejora considerablemente la habilidad de predicción.

En definitiva una vez establecido el número de vecinos que van a decidir, el método consiste en ir calculando las distancias euclídeas de cada caso $\mathbf{x_i}$ a los K vecinos más próximos de cada clase. El objeto $\mathbf{x_i}$ queda clasificado dentro de aquella categoría que presenta una mayor proximidad con respecto a las K muestras que han dado su voto.

El KNN es un "método duro" ya que los objetos quedarán clasificados dentro de alguna de las categorías existentes, es decir, se aplica el nivel 1 del reconocimiento de patrones.

*** Máquina de Aprendizaje Lineal (LLM)**

Este es un método no paramétrico que viene aplicándose desde prácticamente el comienzo de las técnicas de reconocimiento de patrones[194]. Está basado en la clasificación de los objetos en sus respectivas categorías estableciendo las llamadas "superficies de decisión" en el espacio multidimensional, en particular, superficies de decisión lineales en el espacio patrón. Así, supongamos dos clases de patrones c- dimensionales a separar por un hiperplano, cuya dimensión es c-1. Dicho hiperplano puede venir representado por un vector **w** normal a él, que denominaremos vector de pesos. Para cualquier dirección del vector **w**, la posición de un objeto dado $\mathbf{x_i}$ con respecto al hiperplano de separación puede calcularse a partir del producto escalar de **w** y $\mathbf{x_i}$:

$$\boldsymbol{w}\boldsymbol{x}_i = |\boldsymbol{w}||\boldsymbol{x}_i|\cos\alpha \qquad \text{(II.79)}$$

donde α es el ángulo formado por **w** y $\mathbf{x_i}$. El signo del producto escalar determina

en qué lado del hiperplano se sitúa el objeto $\mathbf{x_i}$.

El problema de la clasificación, por tanto, está en establecer el vector de pesos **w**, representativo del hiperplano de separación, que sea capaz de separar bien las clases a las cuales pertenecen los casos.

La Máquina de Aprendizaje Lineal consiste en un proceso iterativo[195-197] para la clasificación de los casos cuya entrada está constituida por el vector de pesos **w** y el objeto $\mathbf{x_i}$, proporcionando una salida binaria (+1 ó -1) que da cuenta de la clase a la cual pertenece el objeto. Al principio, se toma un valor inicial para el vector **w** de forma aleatoria y todos los objetos del conjunto de entrenamiento se van introduciendo en el algoritmo de cálculo donde se realiza el producto escalar y se van clasificando los casos de acuerdo con el signo de la operación. Si la clasificación ha sido correcta, se da entrada al siguiente objeto y se procede de igual modo; si, por el contrario el objeto se ha clasificado mal, el vector de pesos **w** se modifica en cierta extensión (factor de aprendizaje) de forma que el nuevo vector de pesos **w'** es:

$$w' = w + Cx_i \qquad donde \qquad C = -\frac{2wx_i}{x_i x_i} \qquad \text{(II.80)}$$

así, se va repitiendo el algoritmo hasta obtener una buena clasificación del objeto x_i y análogamente con el total de casos. Este procedimiento encontrará, si existe, el hiperplano de separación de las clases, las cuales obviamente deben ser linealmente separables.

El significado físico-químico del plano discriminante depende en gran medida del cociente entre el número de objetos en el conjunto de entrenamiento y la dimensión de los datos[198-200]; un cociente bajo y variables poco adecuadas pueden conducir a una perfecta separación aunque carente de sentido químico.

Los paquetes estadísticos empleados para la realización de las técnicas de reconocimiento de patrones son: CSS STATISTICA, SIRIUS y programas diseñados en el Departamento de Química Analítica y en el Departamento de Química Física de la Facultad de Química de Sevilla.

Otro método no paramétrico empleado en la memoria es el que se basa en Algoritmos de Redes Neuronales Artificiales, pero debido a su diferente filosofía y por su enorme ámbito de aplicación vamos a considerarlos en un capítulo separado.

RECONOCIMIENTO DE PATRONES MEDIANTE ALGORITMOS BASADOS EN REDES NEURONALES ARTIFICIALES

III.1. INTRODUCCION

A lo largo de la historia, el principal problema de los investigadores ha sido buscar la forma de obtener los datos para que produzcan información. En la actualidad debido al desarrollo sufrido por las técnicas instrumentales, la problemática dentro del campo de la quimiometría, no es ya la forma de obtener los datos, sino cómo rechazar toda la información superflua y seleccionar solamente aquellos datos verdaderamente relevantes de un experimento específico. En este marco se incluyen las denominadas *redes neuronales*, consistentes en algoritmos puramente matemáticos que trabajan, tal y como indica su nombre, mimificando el funcionamiento del cerebro humano.

Como veremos, el funcionamiento de estos métodos no radica tanto en el término "neuronal" sino en cómo están interconectadas esas neuronas artificiales, es decir, en la arquitectura de la "red". Precisamente, una de las razones por las que los algoritmos neuronales gozan de una amplia adaptabilidad para el tratamiento de datos es la mencionada arquitectura de las redes, junto con el número de neuronas que las componen y la posibilidad de llevar a cabo tanto un aprendizaje supervisado como no supervisado, por lo cual tienen múltiples campos de aplicación.

Las redes neuronales artificiales han sido el objeto de múltiples trabajos centrados básicamente en reconocimiento de imágenes. Dentro del campo de la química, existen numerosas aplicaciones con más de 500 artículos publicados en el Chemical Abstract[201], referentes a control de procesos, estudio de la estructura secundaria de las proteínas, elucidación de estructuras mediante métodos espectroscópicos, clasificación de objetos en distintas categorías, reactividad de enlaces, optimización de métodos cromatográficos, etc.

En principio, podemos considerar la red neuronal como una caja negra que acepta una serie de datos de entrada y proporciona uno o más datos de salida, tal y como se representa a continuación en la figura 17.

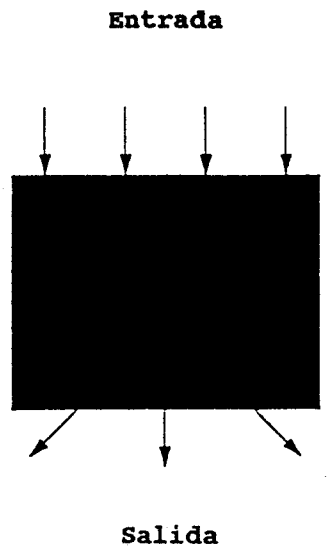

Figura 17. Neurona en forma de caja negra.

En el interior de dicha caja, existen unidades básicas conectadas unas con otras llamadas *neuronas* tal y como se aprecia en la figura 18.

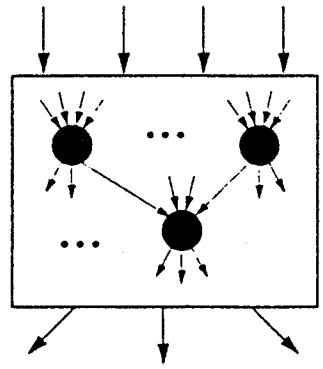

Figura 18. Unidades básicas dentro de la caja.

Las entradas pasan a través de estas conexiones, que son las líneas de la red, transformándose y distribuyéndose en cada una de estas neuronas de forma que producen una o varias salidas. Los datos de entrada son m-dimensionales, siendo la salida n-dimensional. Normalmente, se utilizan redes en las que los datos de salida tienen una dimensión menor que los de entrada.

Las aplicaciones más comunes de las redes neuronales son:

- *Autoasociación*: el sistema es capaz de reconstruir un dato de salida correcto a partir de un dato de entrada incompleto, es decir, identifica patrones corruptos.
- *Heteroasociación*: el sistema realiza asociación uno a uno entre los miembros de dos conjuntos de patrones.
- *Clasificación*: asigna a los objetos una clase en base a una o varias propiedades que caracterizan a una determinada clase. El proceso de clasificación puede realizarse tanto de forma supervisada como no supervisada. El entrenamiento se realiza con un subconjunto de los datos. Posteriormente, la red es capaz de predecir la clase o el cluster al cual pertenecen objetos desconocidos.
- *Transformación*: la red lleva a cabo la transformación de datos del espacio multivariante a otro tipo de datos, también multivariantes de igual o menor dimensionalidad manteniendo la topología de los datos. Es decir, manteniendo la relación entre los datos.

- *Modelización*: consiste en buscar una función analítica o "modelo" que proporcione unos datos de salida n-dimensionales para una entrada m-dimensional. En la modelización clásica, es necesario conocer de antemano la función de modelización, en el caso de los algoritmos neuronales sólo se necesita un número suficiente de datos de entrada, lo suficientemente espaciados como para que la red se adapte a cualquier relación no lineal entre los datos de entrad y los de salida.

III.2. NEURONAS Y REDES

Las redes neuronales artificiales son modelos matemáticos con una similitud muy superficial a las redes neuronales biológicas. Seguidamente, vamos a describir de forma breve y simplificada el funcionamiento de una neurona biológica en el proceso de transmisión de información.

El sistema nervioso humano consta de aproximadamente 10000 millones de neuronas. Una neurona biológica consta de unas ramificaciones llamadas dendritas, un cuerpo o *soma* y una ramificación más alargada denominada axón. Las dendritas reciben la información en forma de impulsos nerviosos y la transmiten al axón y de ahí a la siguiente neurona.

Las *sinapsis* son posiciones situadas en las dendritas y en el axón donde se realiza la transmisión de información. La señal generada por la neurona y transportada por el axón es una señal eléctrica, mientras que la señal que se transmite entre los terminales axónicos de una neurona y las dendritas de las neuronas siguientes es de origen químico. Esta última se realiza mediante sustancias

químicas llamadas neurotransmisores, los cuales modulan la información y fluyen a través de las sinapsis transmitiendo así el impulso nervioso hacia las dendritas de otras neuronas[202]. El paso de información de una neurona a otra se realiza de forma unidireccional en el sentido dendrita-axón. El tipo de señal producida por las neuronas es muy similar independientemente de la especie sin embargo, la intensidad de señal o la frecuencia de transmisión es variable aunque no influye en el tipo de información producida.

Una conclusión que se deduce de la similitud existente entre las señales es que la información memorizada en el cerebro no depende tanto del papel realizado por las neuronas en sí mismas, sino que está más relacionado con los valores sinápticos de las conexiones entre las neuronas y su estructura.

La sinapsis implica la acción de los neurotransmisores modificando la señal que se transmite. En neurobiología, a esta modulación se le denomina *fuerza sináptica.* En el caso de las neuronas artificiales, la fuerza sináptica se denomina *peso (w).* La fuerza sináptica determina la cantidad de señal que entra en el cuerpo de una neurona a través de las dendritas. Los cambios en la fuerza sináptica, incluso entre dos impulsos consecutivos, constituyen un mecanismo vital para el correcto funcionamiento del proceso de aprendizaje. Este proceso de aprendizaje en la red no es más que un cierto número de cambios en los valores de los pesos. Por tanto, si designamos la señal de entrada como s_i (i=1, ..., m), que actúa sobre los pesos w_i, la señal de entrada global será una función de todas las s_i que llegan en cada momento y de todos los pesos w_i, es decir, la señal global o neta (Net) va a ser una suma de los productos $w_i s_i$.

Esta idea se representa en el siguiente esquema:

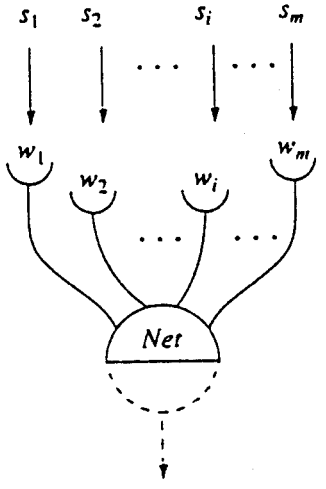

Figura 19. Cálculo de la señal de entrada global a una neurona artificial.

La expresión analítica de la señal Net es:

$$Net = w_1 s_1 + w_2 s_2 + \ldots + w_m s_m \qquad \text{(III.1)}$$

Es conveniente combinar todas las señales s_i en un vector de entrada **X**, cuyas componentes van a ser las s_i. Así:

$$\{s_1, s_2, \ldots s_m\} = X(x_1, x_2, \ldots, x_m) \qquad \text{(III.2)}$$

análogamente, las fuerzas sinápticas se escriben en forma de un vector de pesos **W**:

$$\boldsymbol{W} = (w_1, w_2, \ldots, w_m) \tag{III.3}$$

de manera que la función Net será el producto escalar de los dos vectores **W** y **X**:

$$Net = \boldsymbol{WX} \tag{III.4}$$

generalmente, se añade un *off-set* adicional llamado *bias*:

$$Net = \boldsymbol{WX} + \theta \tag{III.5}$$

la adición de este término permite una mayor adaptabilidad para el aprendizaje.

Por tanto, la expresión final de la función Net será:

$$Net = \sum_{i=1}^{m} W_i X_i + \theta \tag{III.6}$$

Todo proceso de aprendizaje implica, como se indicó anteriormente, el cálculo del vector de pesos óptimo que proporcione decisiones correctas. Para evaluar este vector **W** se utiliza la denominada *Regla Delta.*

Inicialmente, el vector de pesos se escoge al azar y se calcula Net para un objeto determinado **X**, si no se ha obtenido un resultado correcto, deberá modificarse **W** en cierta extensión:

$$\Delta \boldsymbol{W} = \boldsymbol{W}^{(nuevo)} - \boldsymbol{W}^{(viejo)} \tag{III.7}$$

esta modificación es proporcional a una determinada cantidad δ del vector de pesos **W**:

$$\Delta W \sim \delta X \tag{III.8}$$

por tanto, δ se puede expresar como:

$$\delta = \eta\left(\frac{\Delta W}{X}\right) \tag{III.9}$$

donde η es una constante de proporcionalidad llamada factor de aprendizaje.
La ecuación estándar para la regla delta es:

$$\Delta W = \eta \delta X \tag{III.10}$$

Al principio del proceso de aprendizaje, es conveniente mantener η=1 y conforme va aumentando el tiempo de cálculo, mejorándose el aprendizaje, es aconsejable ir reduciendo el valor de η.

Una vez obtenido **W** y calculada Net, hay que transformar de forma no lineal esta función Net de forma análoga al comportamiento biológico. Esta modificación de Net para calcular la salida de la neurona se denomina *Función de transferencia (Sal)*:

$$Sal = f(Net) \tag{III.11}$$

esta función de transferencia debe cumplir tres requisitos:

- no ser negativa.
- debe ser continua y estar confinada dentro de un determinado rango.
- debe ser derivable.

Dentro de las funciones matemáticas que cumplen dichas características, una de las más empleadas en numerosas aplicaciones de redes neuronales es la siguiente:

* *Función sigmoidea (sf)*:

la expresión analítica de esta función es:

$$sf(Net) = \frac{1}{1+\exp(-Net)} \qquad \text{(III.12)}$$

Una de las ventajas que presenta esta función de transferencia es que el cambio de pendiente es mucho más suave que para otras funciones, además la zona de cambio es la de máximo aprendizaje. Una ventaja adicional es que la derivada de la función contiene a la propia función.

Con lo explicado hasta ahora, ya tenemos definida la unidad básica o neurona artificial, cuya representación gráfica se puede observar en la siguiente figura.

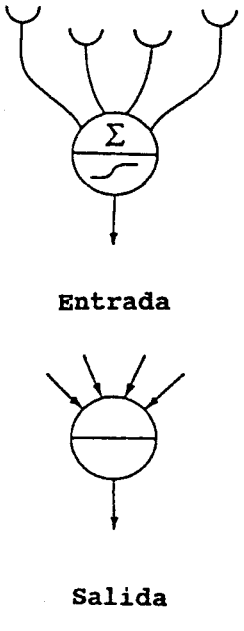

Figura 20. Representación gráfica de una neurona artificial.

El modelo de red debe procesar la información de forma paralela. La señal de entrada debe adecuarse al conjunto de neuronas conectadas en la red, por lo tanto, la señal de entrada debe atacar a todas las neuronas a la vez, cada una de ellas con sus propios vectores de pesos, **W**, y señales de salida (Sal).

Al grupo de neuronas que reciben el mismo conjunto de señales de entrada **X** y producen otras tantas salidas simultáneamente se les denomina *capa.* Todas las neuronas de la misma capa deben tener la misma dimensión. Como cada neurona j produce su propia función Net_j y su correspondiente señal de salida Sal_j, cada señal individual de una misma capa se puede englobar en los respectivos vectores **X, Net** y **Sal**. El vector de salida **Sal** de una capa será el vector de entrada **X** de la

siguiente capa de neuronas.

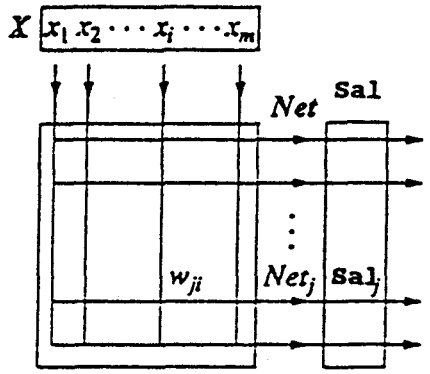

Figura 21. Representación matricial de una red neuronal

A partir de su situación dentro de la red, se pueden distinguir tres tipos de capas:

- *Capa de entrada*: es la capa que recibe directamente la información proveniente de las fuentes externas de la red, los datos de entrada. Es una capa no activa.
- *Capas ocultas*: son internas a la red y no tienen contacto directo con el entorno exterior. El número de niveles ocultos puede estar entre cero y un número elevado. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina, junto con su número, las distintas topologías de redes neuronales.
- *Capa de salida*: transfiere información de la red al exterior.

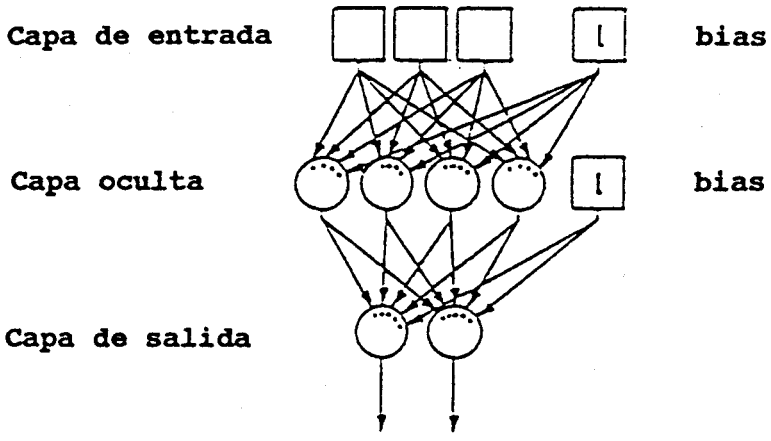

Figura 22. Representación gráfica de las distintas capas de una red neuronal. En este caso, la red de la figura es 3x4x2.

Sea cual fuere la arquitectura de una red neuronal, se debe cumplir que a todas las neuronas de una misma capa les ha de llegar igual número de entradas, incluyendo una entrada adicional (*bias*). Además, el número de pesos en cada neurona es viene fijado por el número de señales procedentes de la capa anterior.

Hasta el momento, hemos representado gráficamente la red neuronal como capas de círculos (neuronas) interconectados mediante flechas de una capa a la siguiente, indicativas de la dirección de flujo de las señales. Desde el punto de vista matemático, una representación matricial es mucho más precisa y explícita.
Vamos a considerar una capa de n neuronas, cada una de ellas con un vector de pesos de m-dimensional, por lo que tendremos una matriz de pesos **W**. para una red multicapa, cada elemento de esta matriz de pesos **W** tendrá un superíndice indicativo

de la capa a la cual pertenece y un subíndice j representativo de su neurona correspondiente (w^l_{ji}).

Para aplicar una señal de entrada m-dimensional a una red neuronal con una capa de n neuronas, cada una de ellas con m pesos, tendremos que multiplicar un vector **X** (x_1, x_2, ..., x_{m-1}, 1) por una matriz de pesos **W** de dimensión nxm. El resultado será un vector **Net** (Net_1, Net_2, ..., Net_n).

Empleando una notación extendida, cada componente Net_j para una capa l se calcula como:

$$Net_j^l = \sum_{i=1}^{m} w_{ji}^l x_i^l \qquad \text{(III.13)}$$

para j=1, 2, ..., n.

En forma matricial:

$$\boldsymbol{Net} = (Net_1, Net_2, ..., Net_j, ..., Net_n) = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \cdot & \cdot & w_{ji} & \cdot \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_i \\ \cdot \\ \cdot \\ x_{m-1} \\ 1 \end{bmatrix} \qquad \text{(III.14)}$$

Para una red multicapa, se cumple que la señal de entrada de la capa l es la salida de la capa l-1:

$$X^l = Sal^{l-1} \tag{III.15}$$

por tanto, puede escribirse:

$$Net_j^l = \sum_{i=1}^{m} w_{ji}^l sal_i^{l-1} \tag{III.16}$$

para j=1, 2, ..., n.

La señal de salida de la capa l se obtiene a partir de Net^l mediante la señal de transferencia, en el caso de utilizar la función sigmoidea:

$$Sal^l = sf(Net^l) \tag{III.17}$$

III.3. APRENDIZAJE NO SUPERVISADO

Las redes neuronales biológicas están dotadas de la facultad de aprender a partir de la experiencia y adaptarse a entornos nuevos con extrema facilidad. Esta capacidad de aprendizaje se debe a la aptitud del organismo de modificar la permeabilidad sináptica entre neuronas, lo cual incide en el poder que una neurona tiene de excitar a otra con la que está conectada y por tanto, en la facilidad de propagación de los impulsos electroquímicos por el entramado de la red.

La capacidad de modificación de la permeabilidad sináptica puede estar regida por dos mecanismos diferentes, que dan lugar a los grandes bloques en que pueden dividirse las redes neuronales artificiales: sistemas supervisados y no supervisados.

Los algoritmos neuronales con aprendizaje no supervisado son capaces de modificar sus parámetros internamente, adaptándose al entorno de la mejor manera posible. Se trata básicamente de que la red debe descubrir por sí sola características, correlaciones o categorías de los datos de entrada y obtenerlas de forma codificada a la salida. Se puede afirmar, por tanto, que estas unidades y conexiones muestran cierto grado de autoorganización.

Uno de los modelos más característicos de redes no supervisadas es el de mapas autoorganizados de Kohonen[203].

III.3.1. Mapas autoorganizativos (SOM)

Una de las características del cerebro hace que unidades estructuralmente idénticas tengan diferente funcionalidad debida a parámetros internos que evolucionan de forma distinta según la ordenación de las células. Esta propiedad *topológica* del cerebro parece ser de fundamental importancia para la representación de cierto tipo de información. Estos mapas topológicos se encuentran presentes en la corteza cerebral y se encargan de diversas tareas de tipo sensorial y motor.

Según Kohonen[204], ciertas redes neuronales pueden adaptar sus respuestas de tal forma que la posición de la célula que produce la respuesta pasa a ser específica de una determinada característica de la señal de entrada. Esta especificidad se da en

el mismo orden topológico para la red que el que existe entre las características de las señales de entrada.

Estudios realizados sobre el neocortex cerebral[205] explican que está compuesto por capas de células bidimensionales interconectadas en cada capa por conexiones laterales, de manera que cada neurona está conectada con otras de su entorno de tal forma que produce una excitación en las más próximas y una inhibición en las más alejadas, como se muestra, a continuación, en la figura 23.

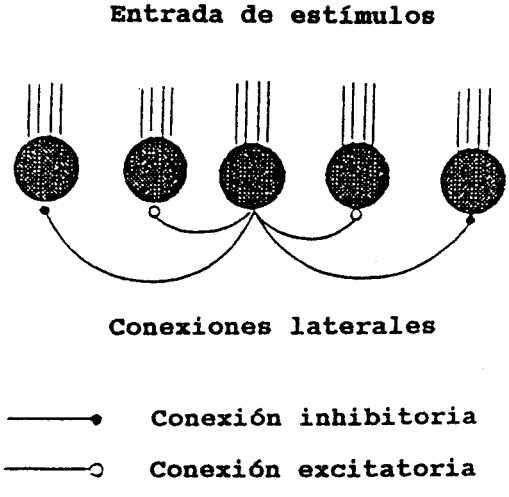

Figura 23. Conexiones laterales de neuronas biológicas.

Tanto la excitación como la inhibición laterales son gradualmente más débiles a medida que nos alejamos de la neurona en cuestión. Así, la posición de las neuronas influye directamente en la forma en que los estímulos van a ser propagados a través de la red y en su respuesta. Esta interacción lateral viene

definida por una función matemática en forma de "sombrero mejicano", cuya representación gráfica se ofrece en la figura 24.

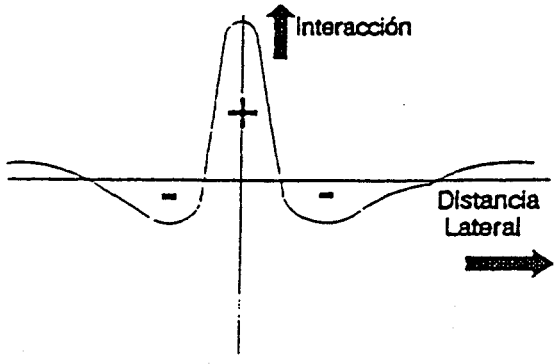

Figura 24. Representación gráfica de la función denominada "sombrero mejicano".

A partir de estos estudios, Teuvo Kohonen diseñó un modelo de red neuronal denominada *mapa autoorganizativo*, basada en una sola capa de neuronas fijadas en un plano bidimensional con una topología muy bien definida: cada neurona está rodeada de un número determinado de neuronas vecinas, las cuales pueden estar dispuestas de dos formas distintas, en cuadrados o en hexágonos. Es decir, cada neurona va a estar rodeada por ocho o seis neuronas en la primera capa de vecindad, tal y como se observa en la figura 25.

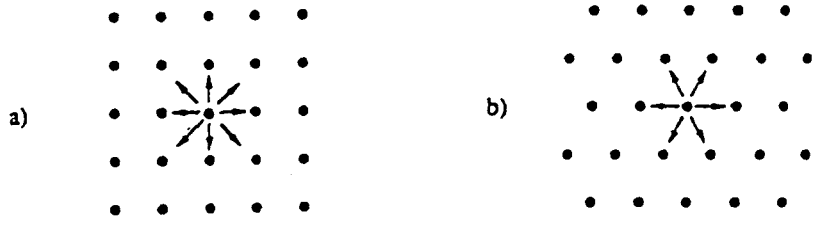

Figura 25. Disposición de la primera capa de vecindad.
a) cuadrada b) hexagonal.

La red SOM hace que señales de entrada similares activen neuronas vecinas (en términos de distancia espacial). En definitiva, relaciona la similitud con la distancia topológica. Es importante que todas las neuronas tengan el mismo número de vecinas, pero debido a que la red es finita se da la circunstancia que las neuronas situadas en las esquinas no tengan el mismo número de vecinas. Una forma de solventar el problema, es plegar la red para dar origen a un toroide, tal como se muestra en la siguiente figura.

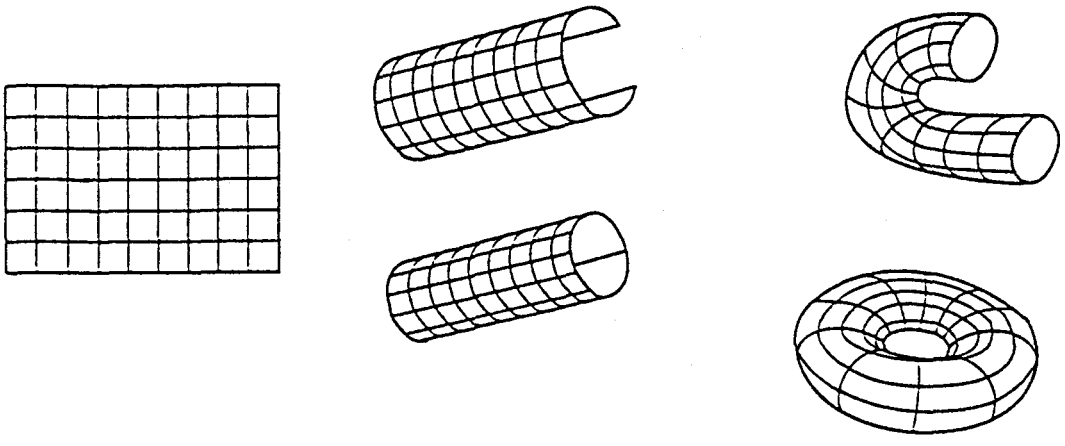

Figura 26. Proceso de conversión de la red bidimensional a un toroide.

En esta red se puede hablar de dos capas: la de entrada y la de salida, siendo ésta última la capa activa. La red aprende por un sistema de *aprendizaje competitivo*, en el que solo una neurona de la capa activa, denominada neurona central, se selecciona de acuerdo a dos criterios posibles:

a) la neurona elegida como central es la que produce el mayor valor de salida:

$$Sal_c = \max(sal)_j = \max\left(\sum_{i=1}^{m} w_{ji} x_{si}\right) \qquad \text{(III.18)}$$

para j=1, 2, ..., n; subíndice que hace referencia a una neurona en particular, n es el número de neuronas, m es el número de pesos por neurona y s identifica una señal de entrada.

b) el vector de pesos de la neurona elegida como central es el más similar al vector de entrada (igual dimensión para **W** y $\mathbf{X_s}$):

$$Sal_c = \min\left(\sum_{i=1}^{m}(x_{si}-w_{ji})^2\right) \qquad \text{(III.19)}$$

para j=1, 2, ..., n.

Este segundo criterio es el más utilizado.

Una vez elegida la neurona central, han de corregirse los pesos w_{ji} de las neuronas vecinas de forma proporcional a sus distancias con respecto a la neurona central. La función matemática que mejor se ajusta al procedimiento de aprendizaje competitivo es la función denominada "sombrero mejicano", cuya expresión analítica es:

$$f = \eta(t)a(d_c-d_j) \qquad \text{(III.20)}$$

donde $a(d_c-d_j)$ es una función de dependencia topológica, puesto de representa la distancia topológica entre la neurona central c y una neurona j; por otro lado, $\eta(t)$ es una función monotónicamente decreciente donde t es el número de objetos que entran en el proceso de aprendizaje.

De esta forma, la corrección de los pesos disminuye al aumentar el número de objetos que se entrenan, al igual que disminuye el número de anillos de neuronas vecinas a los cuales se les hace dicha corrección:

$$w_{ji}^{nuevo} = w_{ji}^{viejo} + \eta(t)a(d_c - d_j)(x_i - w_{ji}^{viejo}) \qquad \text{(III.21)}$$

Así pues, resumiendo el algoritmo empleado por la red de Kohonen consta de los siguientes pasos:

- Introducción del objeto m-dimensional en la red.
- Cálculo de la salida de todas las neuronas.
- Encontrar la neurona central según el criterio seleccionado.
- Corrección de los pesos de la neurona central.
- Corrección de los pesos de los distintos anillos de vecindad.
- Normalización de pesos, sólo en el caso de haber empleado el criterio a) para elegir la neurona central.
- Introducción del siguiente objeto en la red.

III.4. APRENDIZAJE SUPERVISADO

En los sistemas de aprendizaje supervisado, se dispone de algún tipo de información que permite decidir cuándo dejar de aprender, con qué intensidad aprender o cada cuánto tiempo hacerlo. Es posible también manejar cierto tipo de información de error que permite ponderar la rectificación que debemos introducir en la red.

En 1986, Rumelhart y colaboradores[206] basándose en los trabajos de otros investigadores[207-208] idearon un método para que una red neuronal "aprendiera" la asociación existente entre los patrones de entrada a la misma y las clases correspondientes empleando varias capas de neuronas. Este método es conocido

como *backpropagation* o propagación del error hacia atrás (retropropagación) y está basado en la regla delta. A pesar de sus limitaciones, el rango de aplicaciones de este tipo de red neuronal se ha ampliado tanto que casi el 90% de las publicaciones en las que se emplean redes neuronales en el campo de la química se utiliza el aprendizaje por retropropagación[209].

III.4.1. Aprendizaje por Retropropagación

La principal ventaja de un aprendizaje por retropropagación es que la corrección de los pesos está muy bien definida, dicha corrección empieza a ser aplicada a los pesos de la última capa (capa de salida) continuando hacia atrás hasta llegar a la capa de entrada. Seguidamente, en la figura 27, se muestra un esquema de dicha corrección.

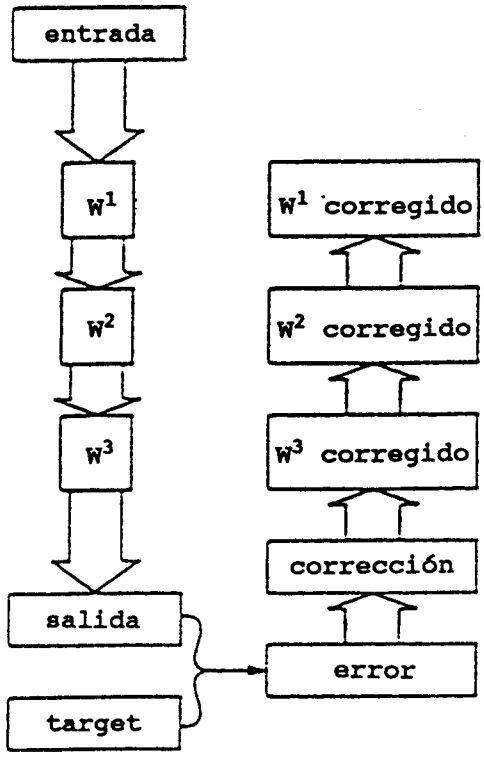

Figura 27. Corrección de pesos en el aprendizaje por retropropagación.

La arquitectura de la red de retropropagación, es decir, el número de capas, número de neuronas en cada capa y la forma en la que se conectan estas neuronas es muy flexible. Generalmente, las capas de neuronas están conectadas de forma total. El número de capas así como el número de neuronas en cada capa depende de la aplicación para la cual se esté desarrollando la red y, normalmente, se determina por ensayo-error. No obstante, en lo referente a esta cuestión existen

diversas recomendaciones en la bibliografía, entre ellas que el número de pesos total no supere el número de objetos introducidos en la red, o bien que el número de pesos sea aproximadamente igual al número de objetos en el conjunto de entrenamiento; por otra parte, también suele ser aconsejable que, en la capa oculta, haya un número menor de neuronas que en la capa de entrada.

En la mayoría de los casos, una red de retropropagación consiste en una capa de entrada con tantas neuronas como componentes tenga la señal de entrada (**X**) y dos capas activas, la capa oculta y la capa de salida.

El método de aprendizaje es supervisado debido a que necesita de un conjunto predefinido de parejas de entradas consistentes en objetos (**X**) junto con unos objetivos o *targets* asociados (**T**). Durante el proceso de entrenamiento, se ajustan los pesos de la red con el fin de que las entradas introducidas produzcan las salidas deseadas.

Antes de comenzar el entrenamiento, es conveniente inicializar los pesos a pequeños valores elegidos al azar con lo cual se asegura, por una parte, que la red no se sature con valores grandes de los pesos y por otra, que se comience en un punto aleatorio de la superficie de error.

Los pasos seguidos en el entrenamiento son:

- Elegir dos conjuntos de objetos de entrada y de objetivos e introducir el primer vector de entrada (**X**) en la red.
- Calcular la salida de la red.
- Calcular el error entre la salida de la red y la salida deseada (objetivo).

- Modificar los pesos de la red con el fin de minimizar el error, empleando la Regla Delta.
- Se repiten los pasos anteriores para cada vector del conjunto de entrenamiento hasta disminuir convenientemente el error.

Las neuronas de la capa de entrada actúan como *buffers* de entrada de los datos, los cuales son ponderados por cada una de las neuronas de la capa. La siguiente capa, capa oculta, recibe las salidas ya ponderadas de la capa de entrada, suma todas las entradas en cada neurona y pasa dicha suma a través de la función de transferencia. El resultado obtenido se pondera y pasa hacia la capa de salida donde se repite el proceso de suma y transformación. El valor de la función de transferencia de la capa de salida constituye la respuesta de la red del objeto introducido.

El proceso de entrenamiento se realiza de forma iterativa introduciendo en la red todos los miembros del conjunto de parejas conocido objeto-objetivo (conjunto de entrenamiento). Cada paso del conjunto completo se denomina época (*epoch*). El aprendizaje termina cuando no se mejora la diferencia entre la salida de la red y el objetivo correspondiente al objeto, o bien, cuando se llega a un número determinado de épocas.

Cuando la función de transferencia empleada es la sigmoidea, el error en la capa de salida viene dado por:

$$\delta_{kl} = (t_{kl} - sal_{kl}) sal_{kl} (1 - sal_{kl}) \qquad \text{(III.22)}$$

donde δ_{kl} es el error para el objeto k en la salida de la neurona l, t_{kl} es el valor de *target* u objetivo correspondiente al objeto k y sal_{kl} es la salida del objeto k en la neurona l.

El error en la neurona j de la capa oculta cuando se utiliza la función sigmoidea es:

$$\delta_{kj} = sal_{kj} (1 - sal_{kj}) \sum_{l=1}^{L} \delta_{kl} w_{lj} \qquad \text{(III.23)}$$

Los errores de la capa oculta se propagan hacia atrás por la red corrigiendo los pesos según la regla delta, así:

$$\Delta w_{ji}(n) = \eta \, \delta_{kj} sal_{ki} + \mu \, \Delta w_{ji}(n-1) \qquad \text{(III.24)}$$

donde Δw_{ji} es la corrección de pesos entre la neurona j de la capa oculta y la neurona i de la capa de entrada, η es la velocidad de aprendizaje y μ se denomina momento y es un parámetro utilizado para salir de los mínimos locales. El término n se refiere a la iteración actual y el término n-1 a la iteración anterior.

Durante el entrenamiento de la red, los objetos pueden introducirse al azar para, de esta forma, evitar posibles tendencias o desviaciones de la red, debido a que ésta puede memorizar el conjunto de aprendizaje extrayendo dependencias.

Otra cuestión importante es la normalización de los objetos de entrada para que no se produzca desbordamiento de la red neuronal.

Este tipo de aprendizaje con esta red es muy adecuado para clasificación de objetos. Para estos casos, el número de neuronas de la capa de salida se hace coincidir con el número de clases existentes y los objetivos asociados a los objetos de entrada serán las clases a las cuales pertenecen dichos objetos. En la mayoría de los casos, es aconsejable escalar los objetivos entre valores 0 y 1; incluso debido al carácter no lineal que presenta la función de transferencia, se puede escalar la salida de la red a valores entre 0.1 y 0.9 ó 0.2 y 0.8. Esta codificación conlleva tres ventajas importantes:

- La comparación de la salida de la red y los objetivos se hace más fácil.
- Se realiza un mejor cálculo del error.
- Se facilidad el posterior cálculo de la respuesta correcta de la red.

Los paquetes estadísticos utilizados para llevar a cabo el análisis mediante redes neuronales son: WINNN 0.97 y programas realizados en el Departamento de Química Analítica y en el Departamento de Química Física de la Facultad de Química de Sevilla.

PARTE EXPERIMENTAL

IV.1. MATERIAL Y REACTIVOS EMPLEADOS

El trabajo experimental llevado a cabo en el presente trabajo, se ha realizado utilizando los siguientes reactivos de calidad "para análisis":

A) REACTIVOS

- Acido tánico (Fluka).
- Carbonato sódico anhidro (Panreac).
- Ninhidrina (Sigma).
- 2-Metoxietanol (Metilcellosolve) (Fluka).
- Acido ascórbico (Panreac).

- Glicina (Merck).
- Acido cítrico (Merck).
- Etanol absoluto (Romil).
- Cafeína (Merck).
- Acido clorogénico (Fluka).
- Acido fórmico (85%) (Panreac).
- Hidróxido potásico en lentejas (85%) (Panreac).
- Metanol (calidad HPLC) (Romil).
- Trigonelina (Sigma).
- Acido clorhídrico (32%) (Merck).
- Acido nítrico (65%) (Merck).
- Acido sulfúrico concentrado (Panreac).

B) DISOLUCIONES

- Carbonato sódico saturado.
- Acido tánico 0.2 g/l.
- Reactivo Folin-Ciocalteu (Fluka).
- Tampón de ácido cítrico/citrato sódico 0.2M (pH 4.6).
- Ninhidrina 10 g/l en metilcellosolve con 0.03% de ácido ascórbico.
- Glicina 0.3 g/l.
- Etanol 60%.
- Metanol-agua (20:80) (pH 4.5).
- Acido clorhídrico 2mM (pH 3).

- Patrón de bario 1 g/l.
- Patrón de calcio 0.5 g/l.
- Patrón de cobre 1 g/l.
- Patrón de estroncio 1 g/l.
- Patrón de fósforo 0.1 g/l.
- Patrón de hierro 1 g/l.
- Patrón de magnesio 1 g/l.
- Patrón de manganeso 1 g/l.
- Patrón de potasio 1 g/l.
- Patrón de sodio 1 g/l.
- Patrón de zinc 0.5 g/l.

A menos que se indique lo contrario, todas las disoluciones empleadas son acuosas. En todos los casos, se ha utilizado agua milli-Q.

En las determinaciones cromatográficas, los disolventes utilizados como fases móviles son de calidad HPLC; además dichas fases móviles fueron filtradas con un filtro de tamaño de poro de 0.45 µm.

IV.2. APARATOS

Las medidas de humedad y extracto acuoso han sido realizadas en una estufa P-Selecta con controlador de temperatura.

Las pesadas se han realizado en balanza de precisión Mettler AE 200 y en granatario, para el caso de cantidades que no requerían demasiada precisión.

Las medidas fotométricas se llevaron a cabo en un espectrofotómetro UV-V

Phillips PU 8720 provisto de cubetas de cuarzo de 1 cm de paso de luz.

Las determinaciones de pH se realizaron con un medidor de pH Crison 2002 equipado con un electrodo combinado de vidrio / Ag/AgCl.

Las determinaciones cromatográficas han sido realizadas empleando un cromatógrafo líquido de alta resolución, equipado con los siguientes módulos:

- Controlador de gradiente Waters AGC-680.
- Bomba Waters 510.
- Inyector tipo Rheodyne con bucle de 20 µl.
- Detector UV Waters 440.
- Integrador CE Instruments DP700.

Para las medidas realizadas mediante Cromatografía Iónica, se utilizó un equipo consistente en:

- Bomba Waters 501.
- Inyector tipo Rheodyne con bucle de 100 µl.
- Detector UV Waters 440.
- Integrador Hewlett-Packard 3395.

Los metales se determinaron con un espectrómetro de emisión atómica de plasma inducido acoplado Fisons-ARL 3410.

IV.3. PROCEDIMIENTOS

IV.3.1. Toma y tratamiento de muestras

En la presente memoria, se han analizado 41 muestras de café verde de distinto origen geográfico, de las que 28 pertenecen a la variedad arábica y 13 son

de la variedad robusta.

Seguidamente, en la tabla 4, se ofrece una lista de todas las muestras estudiadas, junto con un código de identificación, que se utilizará en lo sucesivo.

VARIEDAD	ORIGEN	CODIGO	VARIEDAD	ORIGEN	CODIGO
Arábica	Brasil	1A	Arábica	Salvador	22A
Arábica	Brasil	2A	Arábica	Nicaragua	23A
Robusta	Tailandia	3R	Arábica	Brasil	24A
Arábica	Brasil	4A	Arábica	Brasil	25A
Robusta	Indonesia	5R	Arábica	Colombia	26A
Arábica	Salvador	6A	Robusta	Uganda	27R
Robusta	Costa Marfil	7R	Arábica	Brasil	28A
Arábica	Brasil	8A	Arábica	Brasil	29A
Arábica	Costa Rica	9A	Arábica	Nicaragua	30A
Robusta	Uganda	10R	Arábica	Brasil	31A
Arábica	Colombia	11A	Arábica	Brasil	32A
Robusta	Costa Marfil	12R	Arábica	Brasil	33A
Arábica	Honduras	13A	Arábica	Salvador	34A
Arábica	Nicaragua	14A	Arábica	Honduras	35A
Robusta	Camerún	15R	Robusta	Uganda	36R
Arábica	Guatemala	16A	Robusta	Indonesia	37R
Arábica	Colombia	17A	Arábica	Colombia	38A
Robusta	Costa Marfil	18R	Arábica	Nicaragua	39A
Robusta	Uganda	19R	Robusta	Camerún	40R
Arábica	Brasil	20A	Robusta	Vietnam	41R
Arábica	Honduras	21A			

Tabla 4. Muestras de café verde analizadas.

Todos los parámetros analizados están referidos a una determinada cantidad de muestra seca, por lo que cada una de las muestras fueron secadas en estufa a 103°C durante 2 horas.

El tanto por ciento de humedad, indicado en el apartado correspondiente, se ha obtenido a partir del peso inicial y del peso de muestra seca.

Los parámetros extracto acuoso, polifenoles totales, aminoácidos libres totales, cafeína, ácido clorogénico y trigonelina se determinaron a partir de disoluciones acuosas obtenidas por lixiviación de las muestras de café verde, previa molienda de las mismas.

En cuanto al análisis de metales, se ha realizado una mineralización de todas las muestras.

IV.3.2. Determinación de la humedad

Como se ha comentado anteriormente, todos los resultados se han expresado en relación a base seca, por ello, se determinó previamente la humedad de las muestras, según el procedimiento establecido en la norma ISO 11294[210]:

5g de cada muestra de café verde fueron secados en estufa a 103°C durante 2 horas, se deja enfriar y finalmente se pesa. La pérdida de peso resultante es expresada como agua en tanto por ciento. Los resultados obtenidos se ofrecen a continuación en la tabla 5.

MUESTRA	% (m/m)	MUESTRA	% (m/m)	MUESTRA	% (m/m)
1A	8.6	15R	9.4	29A	8.4
2A	6.2	16A	8.6	30A	9.4
3R	9.2	17A	9.8	31A	11.8
4A	8.4	18R	9.0	32A	8.8
5R	9.2	19R	11.2	33A	8.6
6A	7.8	20A	8.0	34A	9.8
7R	9.2	21A	10.0	35A	9.0
8A	8.0	22A	10.2	36R	9.8
9A	9.0	23A	10.0	37R	7.8
10R	8.8	24A	7.6	38A	8.6
11A	10.2	25A	9.4	39A	9.2
12R	9.4	26A	7.0	40R	10.8
13A	9.0	27R	10.2	41R	9.3
14A	9.4	28A	8.4		

Tabla 5. Humedad de las muestras de café.

Como puede observarse, en ningún caso el contenido en humedad de las muestras supera un 12%, límite máximo establecido por la Reglamentación Técnico-Sanitaria para la elaboración, almacenamiento, transporte y comercialización del café, en su apartado correspondiente a materias primas[211].

IV.3.3. Determinación del extracto acuoso

El análisis del extracto acuoso nos proporciona un parámetro indicativo de la cantidad existente de sólidos solubles en agua.

Su determinación se realiza a partir de una disolución acuosa de café verde, la cual se deja enfriar y seguidamente se filtra. Después, la disolución se lleva a un vaso de precipitado, que ha sido tarado con antelación, y se deja evaporar todo el disolvente. Finalmente, pesamos el residuo seco obtenido.

El procedimiento seguido se detalla a continuación:

5g de una muestra de café verde son llevados a un matraz de fondo redondo de 500 ml y se adicionan 200 ml de agua. A continuación, se calienta a reflujo durante una hora a 80°C. Una vez fría la disolución, se filtra y se trasvasa a un matraz aforado de 250 ml y se enrasa con agua destilada. Seguidamente, 50 ml de la disolución obtenida se llevan a un vaso de precipitado tarado, el cual se pone en una estufa a 105°C hasta total evaporación del disolvente. Se deja enfriar y se pesa, expresando el extracto acuoso obtenido como porcentaje en peso.

En la tabla 6, pueden observarse los datos para las 41 muestras analizadas.

MUESTRA	% (m/m)	MUESTRA	% (m/m)	MUESTRA	% (m/m)
1A	27.35	15R	23.18	29A	25.11
2A	27.53	16A	24.18	30A	23.18
3R	23.13	17A	28.83	31A	27.71
4A	27.29	18R	24.78	32A	24.13
5R	25.33	19R	24.78	33A	29.54
6A	23.29	20A	28.26	34A	26.61
7R	26.43	21A	28.89	35A	23.08
8A	27.17	22A	24.50	36R	25.50
9A	26.37	23A	25.97	37R	23.86
10R	27.41	24A	24.89	38A	22.03
11A	25.61	25A	22.08	39A	20.93
12A	27.41	26A	23.66	40R	22.42
13A	24.18	27R	24.50	41R	27.47
14A	22.08	28A	25.11		

Tabla 6. Extracto acuoso en el café verde.

Los valores obtenidos oscilan entre un 22% y un 29.5%, datos que concuerdan con los encontrados en la bibliografía[7].

IV.3.4. Determinación de polifenoles totales

La cantidad de polifenoles presentes en el café verde se determina mediante un método espectrofotométrico, midiendo el color azul debido a la especie W_2O_5, originada en la reacción que tiene lugar entre los polifenoles del café y el reactivo

Folin-Ciocalteu. Como se observa en la figura 28, el espectro de absorción del azul de wolframio presenta un máximo a una longitud de onda de 748 nm, por lo que en dicha longitud de onda se llevaron a cabo todas las medidas.

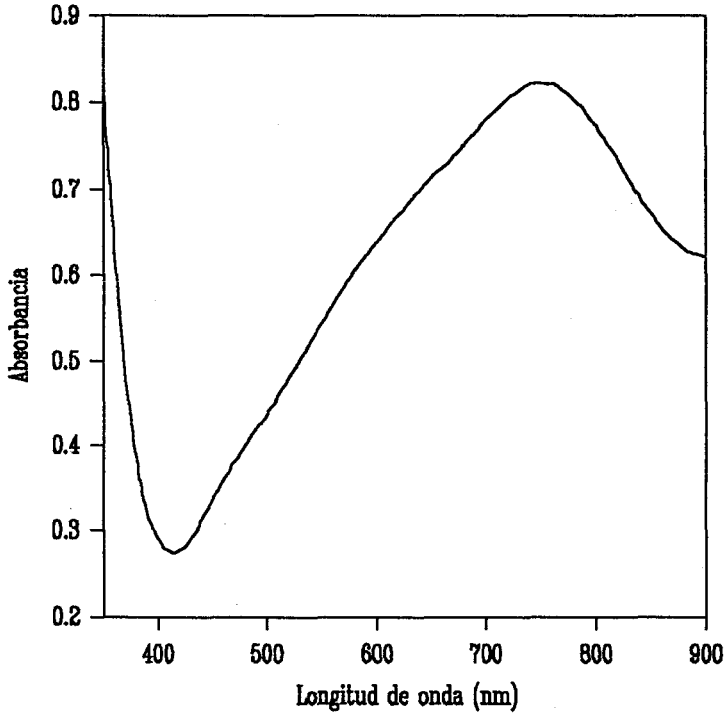

Figura 28. Espectro de absorción de azul de wolframio. Concentración de ácido tánico 10 $\mu g \cdot ml^{-1}$.

El procedimiento seguido se describe a continuación:

5g de café se llevan a un matraz aforado de 500 ml, se añaden 200 ml de agua y se calienta a reflujo por espacio de una hora, a una temperatura de 80°C. A

continuación, se filtra y lava con varias porciones de agua caliente; con ello, se obtiene una disolución que se lleva a un matraz aforado de 250 ml y se enrasa con agua. De esta disolución se toman 10 ml y se llevan a un matraz aforado de 50 ml que se enrasa con agua. 0.25 ml de esta nueva disolución son llevados a un matraz aforado de 25 ml, al cual se añaden 1.30 ml de reactivo Folin-Ciocalteu y 2.5 ml de carbonato sódico saturado. Una vez enrasada con agua, se homogeneiza la disolución y se espera una hora antes de medir, con el propósito de que se estabilice la especie coloreada. Seguidamente, la absorbancia se mide a 748 nm frente a un blanco de reactivo.

Los polifenoles totales contenidos en la muestra de café verde son determinados mediante la recta de calibrado mostrada en la figura 29, en la que se ha utilizado ácido tánico como patrón.

Dicha recta de calibrado tiene como ecuación:

$$A = -9.33E{-}4 + 0.104[ac.tánico] \qquad \text{(IV.1)}$$

El coeficiente de correlación es 0.9945 y el límite de detección es 1.243E-4 %[212].

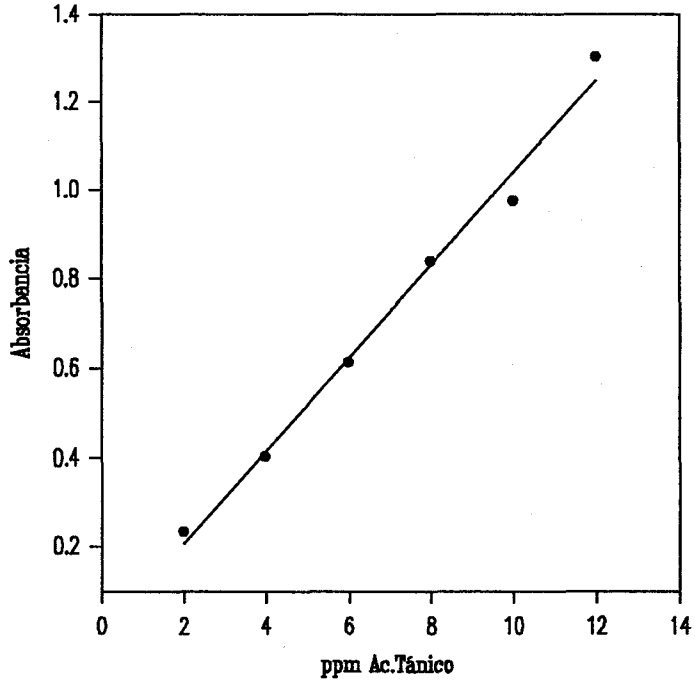

Figura 29. Recta de calibrado para la determinación de polifenoles.

Los valores de las absorbancias así como las correspondientes concentraciones de las muestras empleadas para la recta de calibrado se indican en la tabla 7.

AC.TANICO*	Abs_{748}
2	0.234
4	0.403
6	0.614
8	0.839
10	0.975
12	1.304

Tabla 7. Recta de calibrado para la determinación de polifenoles.
*Cantidades expresadas en $\mu g \cdot ml^{-1}$

En la siguiente tabla, se ofrecen los valores de contenido en polifenoles encontrados en las 41 muestras analizadas, expresados como porcentaje de ácido tánico.

MUESTRA	% (m/m)	MUESTRA	% (m/m)	MUESTRA	% (m/m)
1A	5.0	15R	7.4	29A	6.4
2A	5.2	16A	7.2	30A	6.8
3R	7.5	17A	8.2	31A	7.5
4A	5.0	18R	7.8	32A	5.5
5R	9.5	19R	7.0	33A	5.5
6A	5.6	20A	5.8	34A	4.4
7A	7.8	21A	5.7	35A	5.1
8A	5.9	22A	7.0	36R	6.1
9A	5.9	23A	7.5	37R	8.2
10R	8.2	24A	5.0	38A	6.2
11A	5.2	25A	5.0	39A	6.6
12R	8.4	26A	4.9	40R	8.1
13A	6.1	27R	6.8	41R	8.1
14A	6.4	28A	4.6		

Tabla 8. Contenido en polifenoles totales en el café verde.

Los valores encontrados oscilan entre un máximo de 9.5% para la muestra 5R y un valor mínimo de 4.4% correspondiente a la muestra 34A. Se puede, asimismo, observar que para las muestras de café robusta el contenido en polifenoles encontrado es más alto que para las muestras de café pertenecientes a la variedad arábica. Los valores medios de las dos variedades son 5.9% y 7.8% para arábica y robusta, respectivamente. Aplicando el test t^{212} se comprueba que sí hay diferencia significativa entre ambos valores.

Entre los polifenoles existentes, el compuesto mayoritario es el ácido clorogénico. Esto se comprueba en el apartado IV.3.6. y concuerda con los datos consultados en la bibliografía[213].

IV.3.5. Determinación de aminoácidos libres totales

Los aminoácidos se han determinado midiendo la absorbancia del compuesto coloreado, Púrpura de Ruheman, que se forma. Su espectro de absorción se muestra en la figura 30.

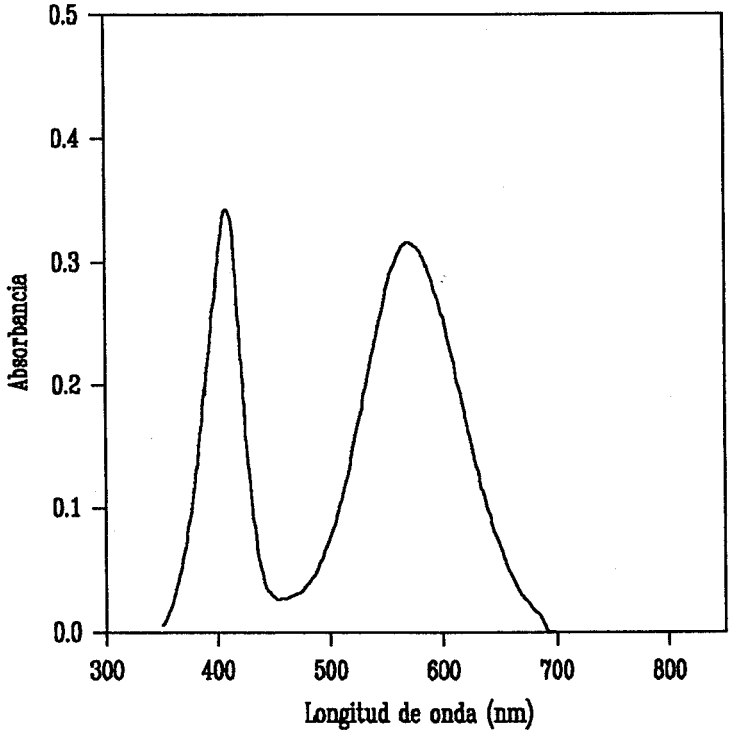

Figura 30. Espectro de absorción de la Púrpura de Ruheman.

La concentración de glicina empleada fue de 1.8 $\mu g \cdot ml^{-1}$. Como puede observarse, el espectro presenta dos máximos a 405 nm y 570 nm. A esta última longitud de onda se han realizado todas las medidas de absorbancia.

La preparación de las disoluciones de café se realizó de forma análoga a lo explicado en el apartado IV.3.4.; excepto que, en este caso, las alícuotas se tomaron directamente del matraz de 250 ml y se siguió el siguiente protocolo:

Se toman 5 ml de la disolución madre de café y se llevan a un matraz aforado de 25 ml, a los cuales se les añade 2.3 ml de tampón cítrico/citrato, 3.6 ml de agua destilada y 4.6 ml de disolución de ninhidrina. Se homogeneiza y se calienta al baño maría durante 15 minutos a una temperatura controlada de 80°C. Se enfría y se enrasa con etanol al 60% (v/v) para estabilizar el color. Se espera una hora y se mide la absorbancia a la longitud de onda indicada anteriormente frente a un blanco de reactivo, previa filtración de las muestras con un filtro de tamaño de poro de 0.45 µm.

El contenido en aminoácidos libres totales en el café verde se determina a partir de la recta de calibrado mostrada en la figura 31, empleando el aminoácido glicina como patrón. Dicha recta de calibrado responde a la siguiente ecuación:

$$A = -4.17E{-}3 + 0.0287[glicina] \qquad \text{(IV.2)}$$

El coeficiente de correlación es de 0.9985. y el límite de detección es de 7.03 E-5%[212].

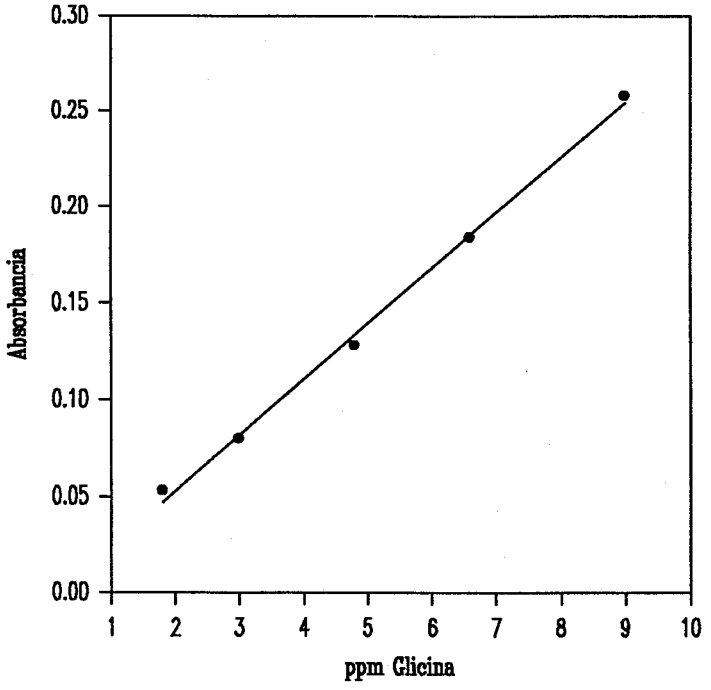

Figura 31. Recta de calibrado para la determinación de aminoácidos libres totales.

Los datos de concentraciones y absorbancias utilizadas para realizar la recta de calibrado se muestran, a continuación en la tabla 9.

GLICINA*	Abs_{570}
1.8	0.053
3.0	0.080
4.8	0.128
6.6	0.184
9.0	0.258

Tabla 9. Recta de calibrado para la determinación de aminoácidos libres totales.

*Cantidades expresadas en $\mu g \cdot ml^{-1}$

Seguidamente, se indican los contenidos de aminoácidos libres totales encontrados en las muestras de café verde analizadas, expresadas todas ellas en porcentaje en peso de glicina.

MUESTRA	% (m/m)	MUESTRA	% (m/m)	MUESTRA	% (m/m)
1A	0.19	15R	0.21	29A	0.23
2A	0.26	16A	0.15	30A	0.26
3R	0.26	17A	0.18	31A	0.24
4A	0.27	18R	0.22	32A	0.23
5R	0.29	19R	0.25	33A	0.24
6A	0.19	20A	0.26	34A	0.19
7R	0.16	21A	0.25	35A	0.20
8A	0.19	22A	0.23	36R	0.18
9A	0.22	23A	0.26	37R	0.26
10R	0.22	24A	0.18	38A	0.20
11A	0.17	25A	0.15	39A	0.24
12R	0.24	26A	0.23	40R	0.16
13A	0.19	27R	0.25	41R	0.17
14A	0.26	28A	0.25		

Tabla 10. Contenido de aminoácidos libres totales en café verde.

Los valores obtenidos se encuentran entre un máximo de 0.29% y un mínimo

de 0.15%, datos que concuerdan con lo encontrado en la bibliografía[40,214] en la que se establece un rango entre 0.15% y 0.25%.

IV.3.6. Determinación de cafeína y ácido clorogénico

Para el análisis de cafeína y ácido clorogénico, se ha empleado la Cromatografía Líquida de Alta Resolución en fase reversa, con fase estacionaria enlazada C-18. En cuanto a la fase móvil, se realizó un estudio para establecer las condiciones óptimas de composición y pH en las que se obtiene una adecuada separación de nuestros analitos. Aunque en la bibliografía[114,123,125] hace referencia, fundamentalmente, a determinaciones en gradiente, se intentó optimizar un método en modo isocrático, que es más rápido ya que no hay que esperar ningún tiempo entre dos análisis sucesivos.

Debido a la presencia de los ácidos isoclorogénico y neoclorogénico, es necesario optimizar las condiciones cromatográficas, para poder resolver los picos correspondientes a estos compuestos además de los de la cafeína y el ácido clorogénico.

Con este fin, se prepararon mezclas de 40, 30 y 20% (v/v) en metanol-disolución acuosa tamponada a pH 3. Al emplear las fases móviles de composición 40% y 30%, se observa que los picos de ácido clorogénico y cafeína no se resuelven bien; sin embargo, en el caso de la mezcla de 20% en metanol, los picos se separan un poco más, aunque sin llegar a resolverse completamente.

Por ello, se optó por mantener la composición de la mezcla binaria en un 20% de metanol y estudiar el efecto del pH.

Se ensayaron varias fases móviles 20% (v/v) metanol-agua con distintos valores de pH y se comprobó que con un valor de pH de 4.5 se obtiene una buena resolución de los picos, tal y como puede apreciarse en la figura 32.

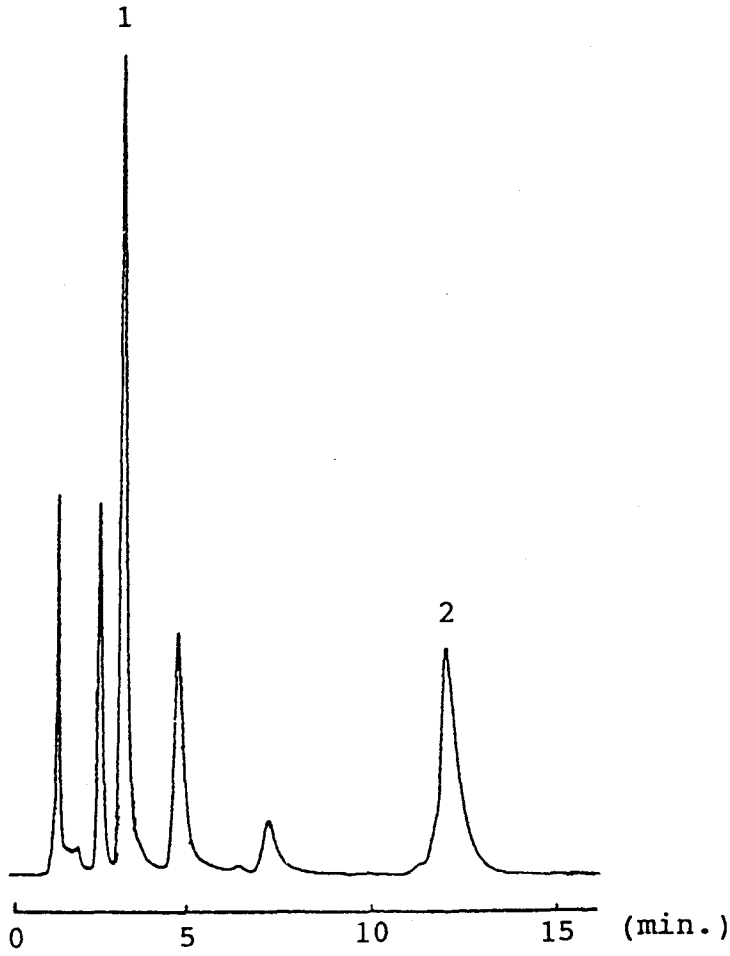

Figura 32. Cromatograma de una muestra de café empleando como fase móvil metanol-agua (20:80) a pH 4.5. 1) Acido clorogénico 2) Cafeína

En lo referente al sistema de detección, se utilizó un detector ultravioleta-visible. Los espectros de absorción de la cafeína y del ácido clorogénico se muestran a continuación en las figuras 33 y 34.

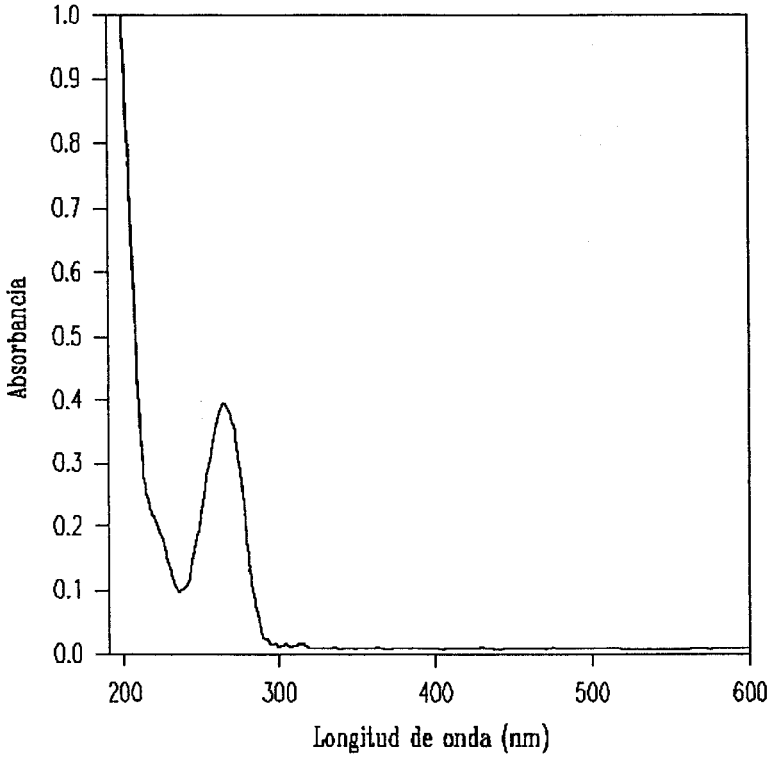

Figura 33. Espectro de absorción de cafeína ($4 \cdot 10^{-5}$ M).

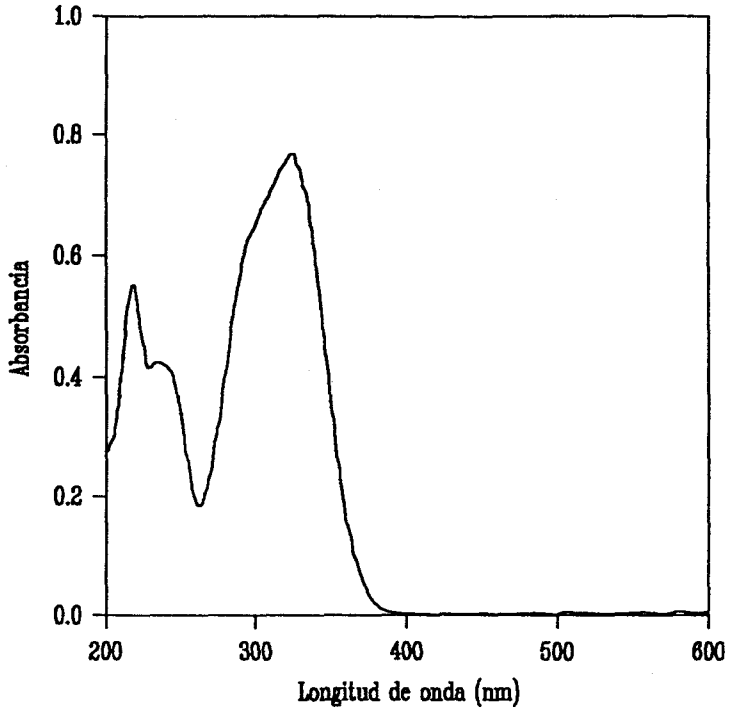

Figura 34. Espectro de absorción de ácido clorogénico ($4 \cdot 10^{-5}$ M).

Las condiciones cromatográficas óptimas fueron las siguientes:

- Fase móvil: metanol-agua (20 : 80) pH 4.5.
- Flujo de fase móvil: 1.5 ml/min.
- Columna: 150 x 4 mm Lichrosorb RP-18 de 5 µm.
- Detector: Ultravioleta de longitud de onda fija a 254 nm.

- Preparación de la fase móvil:

Se disuelven 2.5 g de KOH en agua milli-Q y se añaden 4 ml de ácido fórmico. A continuación, se adicionan 200 ml de metanol, se enrasa con agua milli-Q hasta 1 litro y se microfiltra con un filtro de tamaño de poro de 0.45 µm. El pH final de la fase móvil es de 4.5.

- Preparación de las muestras:

Las muestras de café verde analizadas se preparan a partir de 3g de café molido, los cuales se llevan a un matraz de fondo redondo, se añaden 200 ml de agua y se calienta a reflujo durante una hora. Una vez fría la disolución, se filtra a vacío y se lleva a un matraz de 250 ml y se enrasa con agua. 10 ml de la disolución obtenida se llevan a un matraz de 25 ml y se enrasa con fase móvil. Previamente a la inyección en el cromatógrafo, una porción de esta última disolución se microfiltra (0.45 µm) y se lleva a un vial. Seguidamente, se inyectan alícuotas de 20 µl.

El contenido de ácido clorogénico y cafeína se determina a partir de las rectas de calibrado que se muestran en la figura 35. En ellas, se representan las alturas de los picos cromatográficos frente a las respectivas concentraciones.

La ecuación correspondiente a la recta de calibrado del ácido clorogénico es:

$$Señal = -2.511E+4 + 1.074E+4[Ac.clorogénico] \qquad (IV.3)$$

Mientras que la recta de calibrado correspondiente a la cafeína tiene de ecuación:

$$Señal = -1.803E+3 + 3.241E+3[Cafeína] \qquad \text{(IV.4)}$$

Los coeficientes de correlación son 0.9995 y 0.9977 y los límites de detección son 1.844E-4% y 8.84E-4% respectivamente.

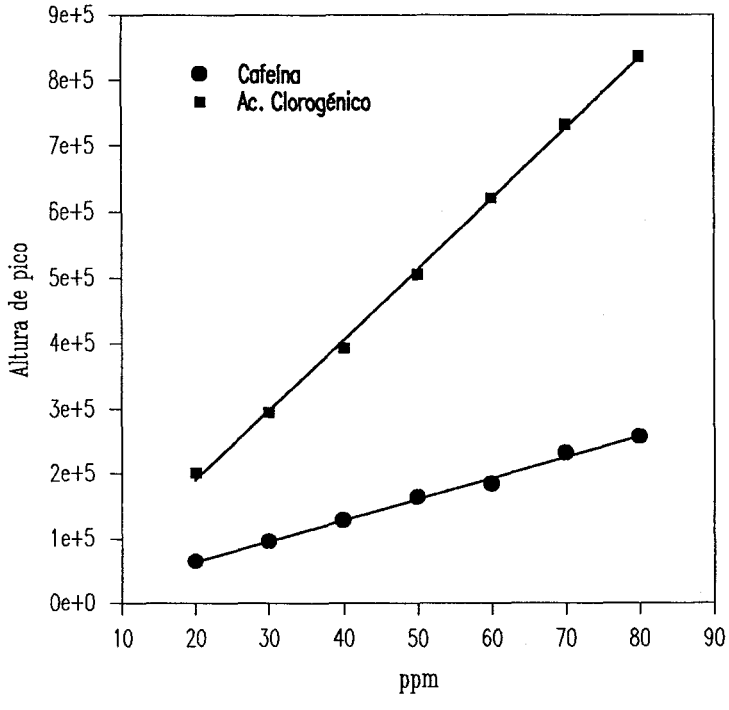

Figura 35. Rectas de calibrado para la determinación de cafeína y ácido clorogénico.

Los valores de concentraciones de las muestras empleadas para las rectas de calibrado, así como las alturas de pico se indican en las tablas 11 y 12.

CAFEINA*	ALTURA DE PICO
20	63637
30	94873
40	128185
50	163363
60	183628
70	231778
80	256421

Tabla 11. Recta de calibrado para la determinación de cafeína.
*Cantidades expresadas en $\mu g \cdot ml^{-1}$

Ac.CLOROGENICO*	ALTURA DE PICO
20	200754
30	295161
40	393800
50	505100
60	620827
70	732071
80	836447

Tabla 12. Recta de calibrado para la determinación de Ac. Clorogénico.
*Cantidades expresadas en $\mu g \cdot ml^{-1}$

A continuación, en la figura 36, se muestra un cromatograma de una disolución patrón de ácido clorogénico y cafeína, obtenido bajo las condiciones cromatográficas optimizadas.

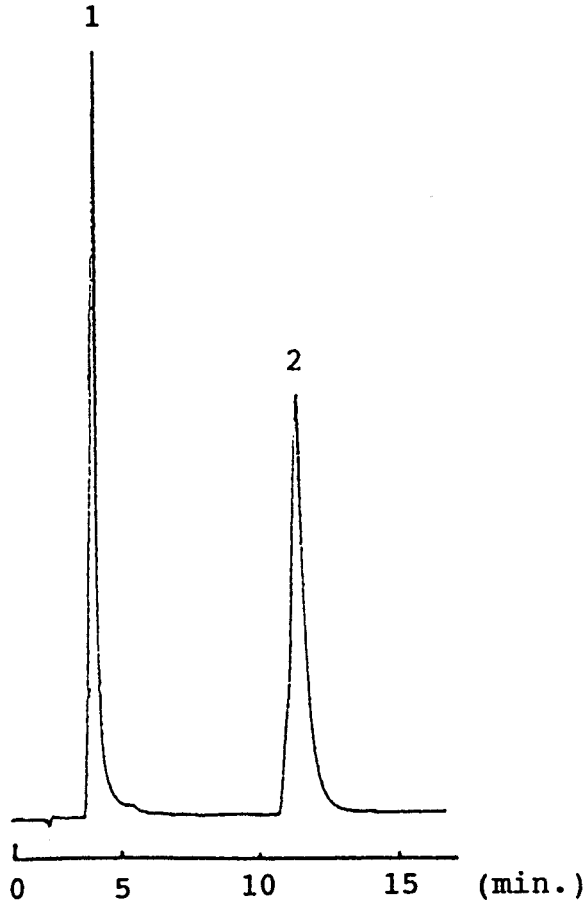

Figura 36. Cromatograma de una disolución patrón de cafeína y ácido clorogénico. 1) Acido clorogénico 2) Cafeína. Concentración 80 $\mu g \cdot ml^{-1}$.

Las siguientes figuras corresponden a cromatogramas de muestras de café verde analizadas de variedad arábica y robusta. Todos los cromatogramas de las muestras se realizaron por triplicado.

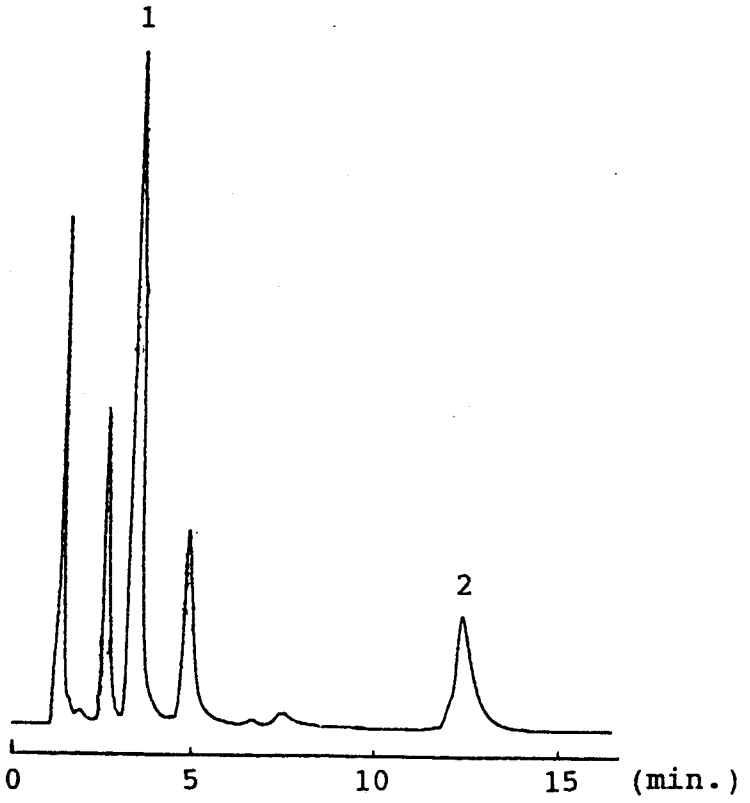

Figura 37. Cromatograma de una muestra de café arábica.

1) Acido clorogénico 2) Cafeína.

Figura 38. Cromatograma de una muestra de café robusta.

1) Acido clorogénico 2) Cafeína.

Los valores obtenidos para cada una de las muestras, expresados en porcentaje de muestra seca, se indican en la tabla 13.

MUESTRA	% (m/m)		MUESTRA	% (m/m)	
	Clorogénico	Cafeína		Clorogénico	Cafeína
1A	3.2	0.9	22A	3.4	1.3
2A	3.4	1.1	23A	4.0	1.2
3R	4.9	2.2	24A	2.8	1.3
4A	3.9	1.3	25A	2.7	1.3
5R	4.2	2.3	26A	3.3	1.3
6A	3.7	1.2	27R	4.2	2.2
7R	5.6	2.7	28A	4.3	1.3
8A	3.9	1.3	29A	4.8	1.4
9A	3.5	1.3	30A	5.0	1.4
10R	4.6	2.4	31A	3.4	1.2
11A	4.1	1.2	32A	3.3	1.1
12R	4.7	2.8	33A	3.1	1.1
13A	3.7	1.1	34A	3.2	1.3
14A	3.4	1.0	35A	3.6	1.4
15R	3.8	2.4	36R	3.8	2.3
16A	3.7	1.1	37R	3.8	2.2
17A	3.7	1.1	38A	3.3	1.1
18R	4.6	2.7	39A	3.4	1.0
19R	4.7	3.2	40R	3.3	1.7
20A	3.9	1.8	41R	3.6	1.6
21A	3.1	1.2			

Tabla 13. Contenido en ácido clorogénico y cafeína en café verde.

Los valores encontrados para el ácido clorogénico, en el caso de muestras de café robusta, son más altos que los correspondientes a la variedad arábica, hecho que concuerda con lo encontrado en la bibliografía[122,213,215]. En cuanto al contenido en cafeína, según los datos de la tabla anterior se aprecian valores más altos de cafeína para la especie robusta (valor medio de 2.4%), que en el caso de muestras arábicas (1.2%). Esta distinción es muy característica de las variedades existentes de café verde y ha sido objeto de numerosos estudios, confirmando todos ellos los resultados mencionados[54,134,216,217]. Destaca la muestra 1A por su bajo contenido en cafeína (0.9%, variedad arábica) y la muestra 19R por su valor extremadamente elevado (3.2%, variedad robusta).

IV.3.7. Determinación de trigonelina

Se ha llevado a cabo la puesta a punto de un nuevo método, sencillo y rápido, para la determinación de trigonelina en las muestra de café verde.

La trigonelina, a pH ácido, se encuentra completamente protonada, por lo que actúa como un catión. Esto permite la determinación de trigonelina mediante Cromatografía Iónica, empleando una columna catiónica, cuya fase estacionaria son partículas de sílice recubiertas de ácido polibutadien-maleico. Este tipo de empaquetamiento es muy adecuado para la retención de compuestos catiónicos monovalentes nitrogenados[218,219].

La fase móvil utilizada fue una disolución acuosa de ácido clorhídrico 2mM (pH 3) en la que la trigonelina se encuentra como catión monovalente. Debido a las propiedades absorbentes de la trigonelina, las cuales se ponen de manifiesto en el

espectro de la figura 39, el detector ultravioleta-visible es muy adecuado, ya que proporciona una sensibilidad mucho mayor que el detector de conductividad.

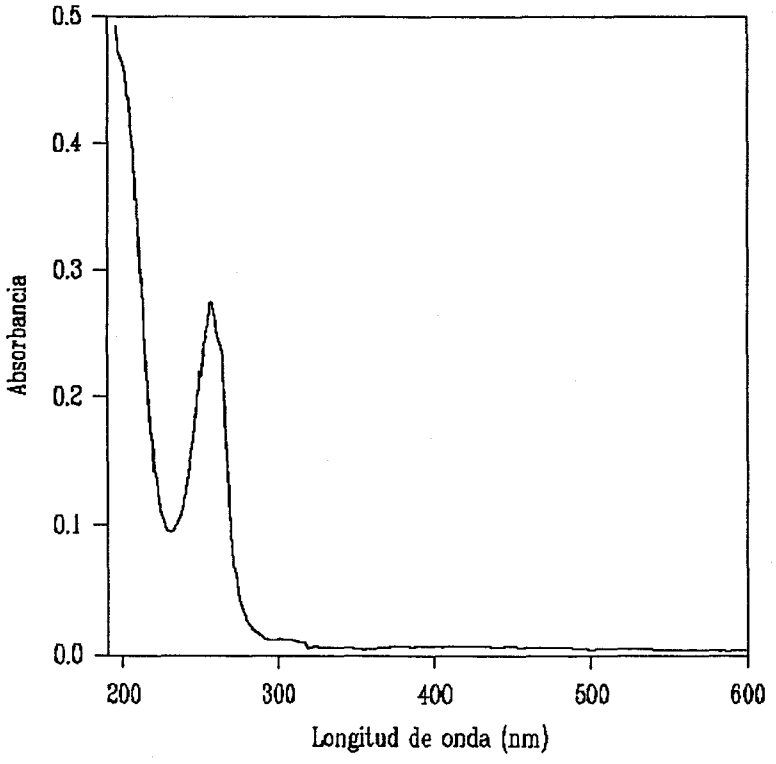

Figura 39. Espectro de absorción de trigonelina. Concentración $4 \cdot 10^{-5}$ M.

El espectro obtenido muestra un máximo de absorción a una longitud de onda de 265 nm y, además, corrobora la elección del ácido clorhídrico para la fase móvil por no absorber a la longitud de onda usada para la detección.

Las condiciones cromatográficas optimizadas se enuncian a continuación:

- Fase móvil: disolución acuosa de ácido clorhídrico de pH 3.
- Flujo de la fase móvil: 1 ml/min.
- Columna: 3.9 x 150 mm Waters IC PacK C M/D de 5 µm.
- Detector: Ultravioleta de longitud de onda fija a 254 nm.

- Preparación de la fase móvil:

Disolución de ácido clorhídrico de pH 3, que se filtra con un filtro de tamaño de poro de 0.45 µm.

- Preparación y clean-up de las muestras:

La disolución de las muestras de café verde previamente molido se prepara pesando 3g de café, los cuales se llevan a un matraz de fondo redondo, se añaden 200 ml de agua milli-Q y se calienta a reflujo durante una hora. Una vez fría la disolución, se filtra y se lleva a un matraz aforado de 250 ml que se enrasa con agua. 5 ml de dicha disolución se llevan a un matraz aforado de 25 ml y se vuelve a enrasar con agua.

Debido a la presencia de compuestos orgánicos neutros del café, los cuales podrían quedar retenidos en la columna por interacciones hidrofóbicas con la fase estacionaria, provocando interferencias en la determinación de la trigonelina, se realiza un clean-up de las muestras antes de inyectarlas en el cromatógrafo.

Los pasos seguidos para el clean-up de dichos compuestos son los siguientes: 3 ml de la última disolución obtenida se pasan a través de un cartucho de C-18 (300 mg), previamente acondicionado con fase móvil; a continuación, se eluye con fase móvil hasta completar un volumen de 10 ml. Con ello, los compuestos iónicos, incluida la trigonelina, son los únicos que pasan a través del cartucho, quedando retenidos los compuestos interferentes.

Seguidamente, una porción de la disolución de 10 ml se microfiltra (0.45 µm) y se lleva a un vial, inyectando alícuotas de 100 µl.

Para determinar la cantidad de trigonelina en las muestras de café verde analizadas, se ha utilizado la recta de calibrado que se muestra en la figura 40, en la que se representa la altura de los picos cromatográficos frente a la concentración de trigonelina.

La ecuación de la recta es:

$$Señal = 5.968E{+}4 + 4.595E{+}4[Trigonelina] \qquad \text{(IV.5)}$$

El coeficiente de correlación es 0.9999. El límite de detección calculado como 3s/b es 0.9E-4%.

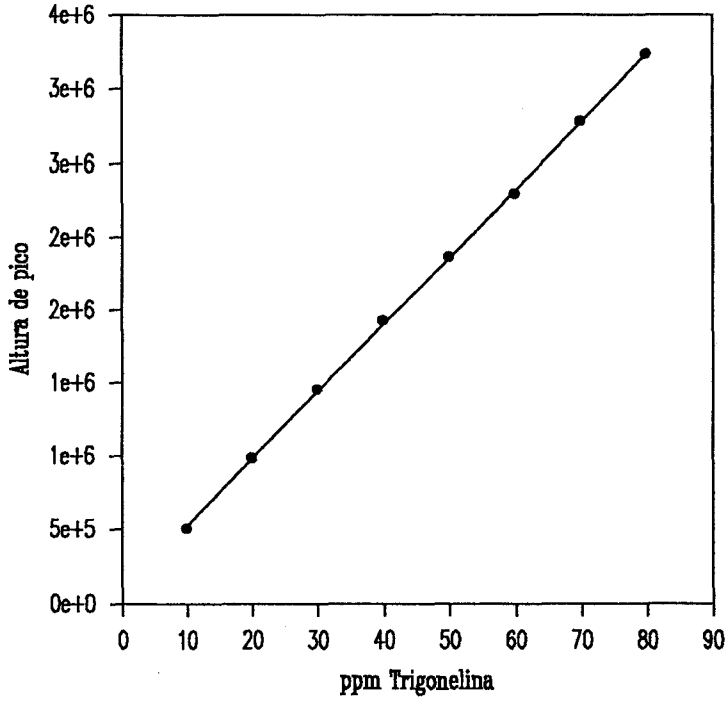

Figura 40. Recta de calibrado para la determinación de trigonelina.

Las concentraciones y alturas de pico correspondientes a las muestras utilizadas para la recta de calibrado se ofrecen en la tabla 14.

TRIGONELINA *	ALTURA DE PICO
10	500060
20	979828
30	1448155
40	1922060
50	2364606
60	2792741
70	3279310
80	3731810

Tabla 14. Recta de calibrado para la determinación de trigonelina.
*Cantidades expresadas en $\mu g \cdot ml^{-1}$

Debido a que no se dispone de muestras certificadas de referencia, la exactitud del método propuesto para la determinación de trigonelina, se ha comprobado mediante un procedimiento de adición estándar[220].

Así, se han establecido tres niveles de fortificación: 10, 20 y 30 $\mu g \cdot ml^{-1}$, para cada uno de los cuales se ha realizado el cálculo de las recuperaciones obtenidas; a continuación se obtiene la media de dichas recuperaciones, encontrándose este dato entre un 0.98 y un 1.02. Según el test de Student[220], como la recuperación es próxima a la unidad, la exactitud del método queda probada.

Cromatogramas correspondientes a una disolución patrón y a muestras de café verde, de variedad arábica y robusta, se ofrecen en las figuras 41, 42 y 43.

Los cromatogramas de todas las muestras se realizaron por triplicado.

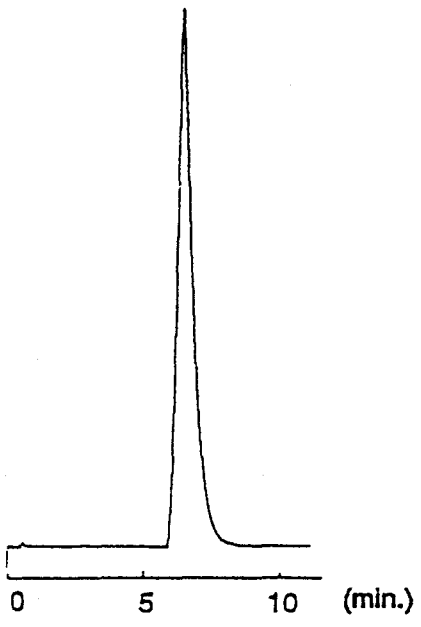

Figura 41. Cromatograma de una disolución patrón de trigonelina Concentración 80 $\mu g \cdot ml^{-1}$.

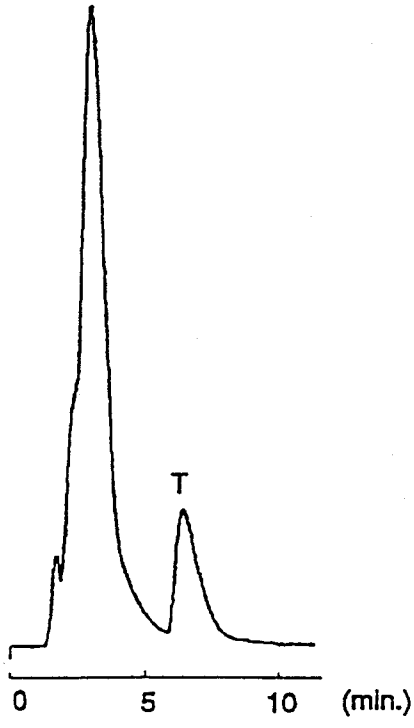

Figura 42. Cromatograma de una muestra de café arábica.

T) Pico correspondiente a trigonelina.

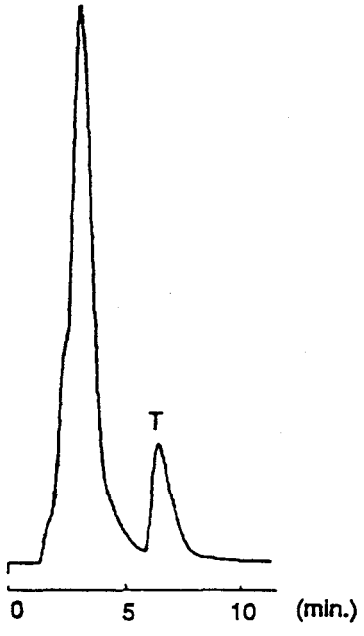

Figura 43. Cromatograma de una muestra de café robusta.
T) Pico correspondiente a trigonelina.

Puede observarse la presencia de un pico al inicio del cromatograma, el cual aparece en todas las muestras que han sido analizadas y que, posiblemente corresponda a especies aniónicas del café, como nitratos o fosfatos, que no son retenidas por la columna, saliendo con el frente de disolvente.

Los datos obtenidos en las determinaciones, expresados en porcentaje de muestra seca, se indican en la tabla 15.

MUESTRA	% (m/m)	MUESTRA	% (m/m)	MUESTRA	% (m/m)
1A	1.78	15R	1.61	29A	1.45
2A	1.79	16A	1.66	30A	1.00
3R	1.35	17A	1.92	31A	1.39
4A	1.68	18R	1.94	32A	1.51
5R	0.94	19R	1.83	33A	1.42
6A	1.10	20A	1.21	34A	1.26
7R	0.91	21A	1.14	35A	1.18
8A	1.10	22A	1.14	36R	1.11
9A	1.27	23A	1.10	37R	1.24
10R	0.96	24A	1.31	38A	1.28
11A	1.04	25A	1.41	39A	1.24
12R	1.72	26A	1.20	40R	1.10
13A	1.12	27R	1.14	41R	1.30
14A	1.19	28A	1.45		

Tabla 15. Contenido de trigonelina en muestras de café verde.

El contenido de trigonelina para las muestras arábicas se encuentra en un rango de 1.00-1.92%, mientras que los cafés robusta presentan un rango de 0.91%-1.94%. También se han calculado los valores medios de trigonelina para ambas variedades, siendo respectivamente 1.334%$\pm$0.247 (arábica) y 1.318%$\pm$0.351 (robusta). Según el test t [212], se comprueba que no hay diferencia significativa entre ambos valores.

IV.3.8. Determinación de metales

Se ha llevado a cabo la determinación del contenido en magnesio, calcio, potasio, fósforo, zinc, manganeso, hierro, sodio, cobre, estroncio y bario en las muestras de café verde. La técnica analítica utilizada para realizar dicha determinación ha sido la Espectroscopía de Emisión Atómica de Plasma Inducido Acoplado (ICP-AES).

Previamente, todas las muestras se mineralizaron según un procedimiento por vía húmeda que se describe a continuación:

Se pesa 1 gramo de café verde molido y se lleva a un vaso de precipitado, se añaden 3 ml de ácido sulfúrico y 10 ml de ácido nítrico y se calienta a la llama hasta obtención de humos blancos. En los casos en que la disolución adquiera un color oscuro, debido a una carbonización de la muestra, se añaden gotas de ácido nítrico tantas veces como sea necesario hasta obtener una disolución prácticamente incolora. Dicha disolución se deja enfriar, se filtra, se lleva a un matraz de 100 ml y, finalmente, se enrasa con agua milli-Q.

Las condiciones instrumentales del espectrómetro de emisión se indican a continuación:

- Radiofrecuencia:

  Potencia incidente: 650 w

  Potencia reflejada: 0 w

  Frecuencia: 27.12 MHz

- Flujos de Argon:

  Coolant: 7.5 l/min

  Plasma: 0.8 l/min

  Carrier: 0.8 l/min

- Nebulizador:

  Flujo de aspiración: 2.3 ml/min

Las longitudes de onda de emisión empleadas en las distintas determinaciones se muestran en la siguiente tabla:

METAL	λ (nm)
Magnesio	279.553
Calcio	393.366
Potasio	766.490
Fósforo	214.914
Zinc	213.856
Manganeso	257.610
Hierro	259.940
Sodio	589.592
Cobre	324.754
Estroncio	407.771
Bario	455.403

Tabla 16. Longitud de onda de emisión de los metales analizados.

La longitud de onda elegida para medir el fósforo es distinta a la de mayor sensibilidad (λ=213.618 nm) para evitar la interferencia debido a la presencia del cobre.

Los cálculos realizados para la determinación cuantitativa de los metales en las muestras de café se hicieron a partir de rectas de calibrado, cuyos parámetros de regresión se indican, junto con las concentraciones de los patrones utilizados (expresadas en $\mu g \cdot ml^{-1}$) y sus correspondientes señales, en las siguientes tablas.

COBRE	SEÑAL
0.01	0.794
0.05	1.032
0.1	1.332
0.2	1.988
0.5	3.886
Ordenada Origen: 0.720	
Pendiente: 0.629E+1	
Coef.Correlación: 0.9997	
L.O.D. 1.338E-6%	

Tabla 17. Recta de calibrado para el cobre.

ESTRONCIO	SEÑAL
0.01	2.324
0.05	8.365
0.1	15.718
0.2	31.378
0.5	78.023
Ordenada Origen: 0.53939	
Pendiente: 0.15478E+3	
Coef.Correlación: 0.9999	
L.O.D. 5.521E-8%	

Tabla 18. Recta de calibrado para el estroncio.

BARIO	SEÑAL
0.01	1.178
0.05	2.835
0.1	5.039
0.2	9.456
0.5	22.824
Ordenada Origen: 0.64769	
Pendiente: 0.44295E+2	
Coef.Correlación: 0.9999	
L.O.D. 1.181E-6%	

Tabla 19. Recta de calibrado para el bario.

ZINC	SEÑAL
0.01	0.353
0.05	2.010
0.1	3.611
0.2	6.989
0.5	17.213
Ordenada Origen: 0.16492	
Pendiente: 0.34130E+2	
Coef.Correlación: 0.9998	
L.O.D. 1.80E-6%	

Tabla 20. Recta de calibrado para el zinc.

CALCIO	SEÑAL
5	60.466
10	110.083
15	170.731
20	215.462
Ordenada Origen: 0.77765E+1	
Pendiente: 0.10516E+2	
Coef.Correlación: 0.9985	
L.O.D. 3.902E-5%	

Tabla 21. Recta de calibrado para el calcio.

HIERRO	SEÑAL
0.5	4.199
1.0	7.402
1.5	11.512
2.0	14.267
Ordenada Origen: 0.76650	
Pendiente: 0.76650	
Coef.Correlación: 0.68642E+1	
L.O.D. 8.09E-6%	

Tabla 22. Recta de calibrado para el hierro.

POTASIO	SEÑAL
100	25.195
150	35.943
200	45.305
250	57.918
Ordenada Origen: 0.34544E+1	
Pendiente: 0.21506	
Coef.Correlación: 0.9988	
L.O.D. 1.367E-3%	

Tabla 23. Recta de calibrado para el potasio.

MAGNESIO	SEÑAL
5	121.041
10	238.025
20	469.566
30	703.727
Ordenada Origen: 0.46465E+1	
Pendiente: 0.23289E+2	
Coef.Correlación: 0.9999	
L.O.D. 1.83E-5%	

Tabla 24. Recta de calibrado para el magnesio.

MANGANESO	SEÑAL
0.5	27.524
1.0	50.614
1.5	79.325
2.0	98.756
Ordenada Origen: 0.34530E+1	
Pendiente: 0.48481E+2	
Coef.Correlación: 0.9976	
L.O.D. 6.0E-6%	

Tabla 25. Recta de calibrado para el manganeso.

SODIO	SEÑAL
0.5	2.484
1.0	3.821
1.5	4.888
2.0	6.131
Ordenada Origen: 0.13290E+1	
Pendiente: 0.24016E+1	
Coef.Correlación: 0.9991	
L.O.D. 5.36E-5%	

Tabla 26. Recta de calibrado para el sodio.

FOSFORO	SEÑAL
5	1.431
10	2.760
15	4.348
20	5.557
Ordenada Origen: 0.325E-1	
Pendiente: 0.27573	
Coef.Correlación: 0.9999	
L.O.D. 3.96E-5%	

Tabla 27. Recta de calibrado para el fósforo.

Los datos sobre el contenido metálico en las muestras analizadas se ofrecen, a continuación, en la siguiente tabla expresados en tanto por ciento de muestra seca.

MUESTRA	ZN	P	MN	FE	MG	CA	NA	K	CU	SR	BA
1A	6.13E-3	0.142	3.54E-3	5.47E-3	0.172	0.110	6.00E-3	1.532	7.69E-3	3.83E-4	3.83E-4
2A	7.04E-4	0.141	2.86E-3	3.91E-3	0.175	9.86E-2	4.22E-3	1.589	1.81E-3	4.58E-4	4.16E-4
3R	1.50E-3	0.188	1.97E-3	3.95E-3	0.177	0.125	5.13E-3	1.606	1.52E-3	3.30E-4	1.76E-4
4A	4.69E-4	0.153	3.36E-3	3.23E-3	0.183	0.093	2.84E-3	1.555	1.75E-3	4.80E-4	5.13E-4
5R	1.34E-3	0.187	1.52E-3	3.48E-3	0.168	0.123	7.50E-3	1.735	2.03E-3	6.17E-4	1.87E-4
6A	6.29E-4	0.148	3.23E-3	2.80E-3	0.184	0.110	5.85E-3	1.381	1.50E-3	5.10E-4	3.47E-4
7R	8.59E-4	0.172	1.53E-3	4.88E-3	0.179	0.127	8.39E-3	1.545	2.25E-3	6.06E-4	2.09E-4
8A	5.33E-4	0.158	3.41E-3	3.51E--3	0.195	0.107	3.91E-3	1.538	1.76E-3	5.43E-4	4.89E-4
9A	3.18E-3	0.150	3.75E-3	3.25E-3	0.201	0.107	5.75E-3	1.431	1.67E-3	1.30E-4	5.49E-4
10R	8.44E-4	0.208	1.50E-3	3.91E-3	0.171	0.123	1.82E-3	1.616	2.04E-3	6.69E-4	5.81E-4
11A	5.90E-4	0.170	3.10E-3	2.71E-3	0.188	0.103	3.50E-3	1.492	1.48E-3	4.90E-4	4.68E-4
12R	1.63E-3	0.180	1.47E-3	3.47E-3	0.181	0.110	5.07E-3	1.693	2.22E-3	5.63E-4	1.77E-4
13A	1.77E-3	0.148	3.30E-3	2.99E-3	0.188	0.110	3.81E-3	1.498	1.43E-3	4.18E-4	6.70E-4
14A	1.43E-3	0.155	2.11E-3	2.86E-3	0.187	0.101	7.56E-3	1.662	1.59E-3	4.30E-4	5.41E-4
15R	1.99E-3	0.201	1.85E-3	2.88E-3	0.166	0.152	6.13E-3	1.664	2.12E-3	8.72E-4	4.30E-4
16A	7.77E-4	0.148	3.46E-3	2.76E-3	0.183	0.103	6.90E-3	1.482	1.62E-3	5.25E-4	5.25E-4
17A	6.98E-4	0.152	4.72E-3	2.48E-3	0.179	0.106	4.25E-3	1.432	1.52E-3	9.53E-4	7.32E-4
18R	9.23E-4	0.176	1.60E-3	4.75E-3	0.186	0.133	9.35E-3	1.619	2.58E-3	6.04E-4	1.65E-4
19R	5.41E-4	0.219	1.61E-3	4.83E-3	0.183	0.137	3.12E-3	1.614	2.21E-3	8.78E-4	5.18E-4
20A	7.50E-4	0.163	3.71E-3	4.42E-3	0.197	0.103	6.88E-3	1.370	1.84E-3	4.57E-4	3.70E-4
21A	1.23E-3	0.161	3.32E-3	3.14E-3	0.197	0.096	5.51E-3	1.211	1.62E-3	4.67E-4	6.00E-4

Tabla 28. Contenido en metales, en %(m/m), de las muestras de café verde.

Tabla 28 (cont.). Contenido en metales, en %(m/m), de las muestras de café verde.

MUESTRA	ZN	P	MN	FE	MG	CA	NA	K	CU	SR	BA
22A	1.60E-3	0.147	3.05E-3	4.02E-3	0.183	0.103	4.76E-3	1.362	1.65E-3	7.24E-4	4.90E-4
23A	1.39E-3	0.145	1.81E-3	3.02E-3	0.179	0.106	4.74E-3	1.308	1.70E-3	5.56E-4	4.78E-4
24A	5.52E-4	0.154	2.66E-3	2.79E-3	0.186	0.133	3.38E-3	1.238	1.80E-3	3.14E-4	2.92E-4
25A	3.64E-4	0.150	3.01E-3	2.86E-3	0.183	0.137	3.08E-3	1.824	1.57E-3	3.86E-4	3.75E-4
26A	5.91E-4	0.156	5.00E-3	2.54E-3	0.197	0.103	6.81E-3	1.677	1.84E-3	1.16E-3	7.85E-4
27R	9.69E-4	0.220	1.51E-3	4.98E-3	0.197	0.096	4.55E-3	1.863	2.31E-3	9.69E-4	4.57E-4
28A	4.91E-4	0.148	3.32E-3	3.08E-3	0.185	0.132	4.50E-3	1.701	1.72E-3	4.80E-4	3.93E-4
29A	4.26E-4	0.152	3.55E-3	3.67E-3	0.187	0.106	3.80E-3	1.652	1.64E-3	4.37E-4	3.71E-4
30A	7.95E-4	0.147	1.95E-3	5.51E-3	0.182	0.099	6.67E-3	1.618	1.58E-3	4.64E-4	5.96E-4
31A	8.39E-4	0.152	2.64E-3	4.30E-3	0.206	0.103	6.80E-3	1.669	1.92E-3	3.40E-4	2.49E-4
32A	3.62E-4	0.150	3.17E-3	2.64E-2	0.198	0.109	4.34E-3	1.828	1.58E-3	4.17E-4	3.51E-4
33A	5.14E-4	0.154	3.87E-3	3.61E-3	0.184	0.110	5.31E-3	1.719	1.54E-3	4.49E-4	3.17E-4
34A	4.88E-4	0.145	2.72E-3	3.54E-3	0.177	0.125	4.48E-3	1.664	1.50E-3	6.10E-4	4.55E-4
35A	5.49E-4	0.157	3.36E-3	2.79E-3	0.187	0.093	5.46E-3	1.645	1.52E-3	4.73E-4	6.81E-4
36R	1.09E-3	0.212	1.45E-3	7.43E-3	0.176	0.145	0.010	1.869	2.49E-3	8.43E-4	6.43E-4
37R	6.29E-4	0.198	1.69E-3	9.33E-3	0.183	0.137	5.48E-3	1.896	2.61E-3	8.13E-4	3.58E-4
38A	5.36E-4	0.159	4.79E-3	3.04E-3	0.179	0.105	2.84E-3	1.517	1.68E-3	5.47E-4	5.91E-4
39A	7.82E-4	0.150	1.62E-3	3.94E-3	0.185	0.115	0.012	1.598	1.64E-3	5.40E-4	4.96E-4
40R	5.38E-4	0.195	1.75E-3	4.66E-3	0.175	0.162	3.08E-3	1.681	2.24E-3	1.00E-3	4.93E-4
41R	1.14E-3	0.186	1.91E-3	7.31E-3	0.160	0.095	3.55E-3	1.809	1.87E-3	4.96E-4	4.85E-4

Según los datos obtenidos, se puede observar la existencia de un grupo de metales mayoritarios, cuyas concentraciones oscilan entre los valores 1.896% y 0.093%. Cabe destacar el contenido en potasio (1.896%-1.211%), el cual ha sido ampliamente estudiado por diversos autores empleando distintas técnicas de medida[12-15], así como magnesio (0.206%-0.160%), fósforo (0.220%-0.141%) y, en menor cantidad, calcio (0.162%-0.093%).

En cuanto al fósforo, hay que mencionar una clara diferencia de contenido para muestras de variedad arábica ($0.152\% \pm 0.00644$) y las pertenecientes a la variedad robusta ($0.196\% \pm 0.0158$).

En relación a zinc, manganeso, hierro, sodio, cobre, estroncio y bario, cuyos contenidos son menores, del orden de 10^{-3}-10^{-4}%, cabe indicar que estos datos concuerdan con los publicados en la bibliografía[36].

En el caso del manganeso, se aprecia una clara diferencia entre las dos variedades, ya que el valor medio para la arábica es 3.22E-3% y para robusta 1.64E-3%. Esto concuerda con lo indicado por Wilbaux[221]. Para el cobre, también hay diferencias entre ambas variedades, pero en este caso el contenido es mayor en los robusta (2.04E-3%), que en los arábica (1.65E-3%).

DISCUSION DE RESULTADOS

V. DISCUSION DE RESULTADOS

En la presente memoria, hemos analizado las variables extracto acuoso (EXT), polifenoles totales (POLI), aminoácidos libres totales (AA), ácido clorogénico (CLOROG), cafeína (CAF), trigonelina (TRIG), zinc (ZN), fósforo (P), manganeso (MN), hierro (FE), magnesio (MG), calcio (CA), sodio (NA), potasio (K), cobre (CU), estroncio (SR) y bario (BA) en un total de 41 muestras de café verde, de las que 28 pertenecen a la variedad arábica y 13 a la variedad robusta.

Las posibles fuentes de variación que influyen en la dispersión o aglomeración de las muestras son múltiples: lugar de origen, variedad de la planta, condiciones de almacenamiento, cosecha, clima, abonos; además de todos los

tratamientos a que es sometido el café desde su recolección hasta su tueste. La mayoría de estas fuentes no pueden determinarse con exactitud, sólo nos consta con certeza la variedad a la que pertenecen las muestras y el país de procedencia.
En primer lugar deberemos establecer cual será la principal fuente de variación atendiendo a los descriptores seleccionados para, posteriormente, poder construir reglas de clasificación adecuadas, utilizando los procedimientos correspondientes de reconocimiento de patrones ya explicados en el capítulo anterior.

V.1. TECNICAS DE PREPROCESADO

Para realizar los cálculos, los 17 descriptores químicos analizados se han dividido en dos grupos: variables no metálicas EXT, POLI, AA, TRIG, CLOROG y CAF y, por otro lado, variables de carácter metálico ZN, P, MN, FE, MG, CA, NA, K, CU, SR Y BA.

V.1.1. Análisis en Componentes Principales (PCA)

Como se ha explicado en el apartado II.2.1. comenzamos el estudio quimiométrico con esta técnica de preprocesado con la cual, analizaremos el poder discriminante de los descriptores elegidos a la hora de separar las muestras de café verde. Además, a partir del "scores plot" nos haremos una primera idea acerca de la distribución, en el espacio de las variables, de los casos que nos ocupan.

Tal y como se ha mencionado al inicio de este capítulo, vamos a tratar por separado el perfil metálico y el no metálico de nuestras muestras.

- *Variables de carácter no metálico:*

La matriz de datos está ahora compuesta por 6 columnas y 41 filas.

	EXT	POLI	AA	TRIG	CLORG	CAF		EXT	POLI	AA	TRIG	CLORG	CAF
1A	27.35	5.0	0.19	1.78	3.2	0.9	22A	24.50	7.0	0.23	1.14	3.4	1.3
2A	27.53	5.2	0.26	1.79	3.4	1.1	23A	25.97	7.5	0.26	1.10	4.0	1.2
3R	23.13	7.5	0.26	1.35	4.9	2.2	24A	24.89	5.0	0.18	1.31	2.8	1.3
4A	27.29	5.0	0.27	1.68	3.9	1.3	25A	22.08	5.0	0.15	1.41	2.7	1.3
5R	25.33	9.5	0.29	0.94	4.2	2.3	26A	23.66	4.9	0.23	1.20	3.3	1.3
6A	23.29	5.6	0.19	1.10	3.7	1.2	27R	24.50	6.8	0.25	1.14	4.2	2.2
7R	26.43	7.8	0.16	0.91	5.6	2.7	28A	25.11	4.6	0.25	1.45	4.3	1.3
8A	27.17	5.9	0.19	1.10	3.9	1.3	29A	25.11	6.4	0.23	1.45	4.8	1.4
9A	26.37	5.9	0.22	1.27	3.5	1.3	30A	23.18	6.8	0.26	1.00	5.0	1.4
10R	27.41	8.2	0.22	0.96	4.6	2.4	31A	27.21	7.5	0.24	1.39	3.4	1.2
11A	25.61	5.2	0.17	1.04	4.1	1.2	32A	24.13	5.5	0.23	1.51	3.3	1.1
12R	27.41	8.4	0.24	1.72	4.7	2.8	33A	29.54	5.5	0.24	1.42	3.1	1.1
13A	24.18	6.1	0.19	1.12	3.7	1.1	34A	26.61	4.4	0.19	1.26	3.2	1.3
14A	22.08	6.4	0.26	1.19	3.4	1.0	35A	23.08	5.1	0.20	1.18	3.6	1.4
15R	23.18	7.4	0.21	1.61	3.8	2.4	36R	25.50	6.0	0.18	1.11	3.8	2.3
16A	24.18	7.2	0.15	1.66	3.7	1.1	37R	23.86	8.2	0.26	1.24	3.8	2.2
17A	28.83	8.2	0.18	1.92	3.7	1.1	38A	22.03	6.2	0.20	1.28	3.3	1.1
18R	24.78	7.8	0.22	1.94	4.6	2.7	39A	20.93	6.6	0.24	1.24	3.4	1.0
19R	24.78	7.0	0.25	1.83	4.7	3.2	40R	22.42	8.1	0.16	1.10	3.3	1.7
20A	28.26	5.8	0.26	1.21	3.9	1.8	41R	27.47	8.1	0.17	1.30	3.6	1.6
21A	28.90	5.7	0.25	1.14	3.1	1.2							

Figura 44. Matriz de datos con las variables de carácter no metálico.

Después de realizar el PCA, obtenemos tres Componentes Principales que explican el 72.77% de la varianza, como se muestra en la tabla 29.

COMPONENTE	AUTOVALOR	% VARIANZA ACUMULADA
PC1	2.19	36.5
PC2	1.23	57.1
PC3	0.94	72.8

Tabla 29. Varianzas acumuladas para los tres primeros Componentes Principales.

Los valores de comunalidades de estos factores nunca fueron menores de 0.6, tal y como se muestra en la tabla 30.

	PC1	PC2	PC3	COMUNALIDAD
EXT	.079	-.766	.085	.60
POLI	.773	.127	-.184	.65
AA	.362	-.322	.787	.85
TRIG	-.017	-.725	-.475	.75
CLOROG	.839	.056	.051	.71
CAF	.867	.022	-.231	.80

Tabla 30. Contribuciones de los PC's a las variables y comunalidades.

Para una mejor visualización de la información obtenida con esta técnica de preprocesado, realizamos la representación BIPLOT tanto de los *scores* como de los *loadings* para los dos primeros PC's. Observando la figura 45, podemos hacernos una primera idea de la distribución de las variables.

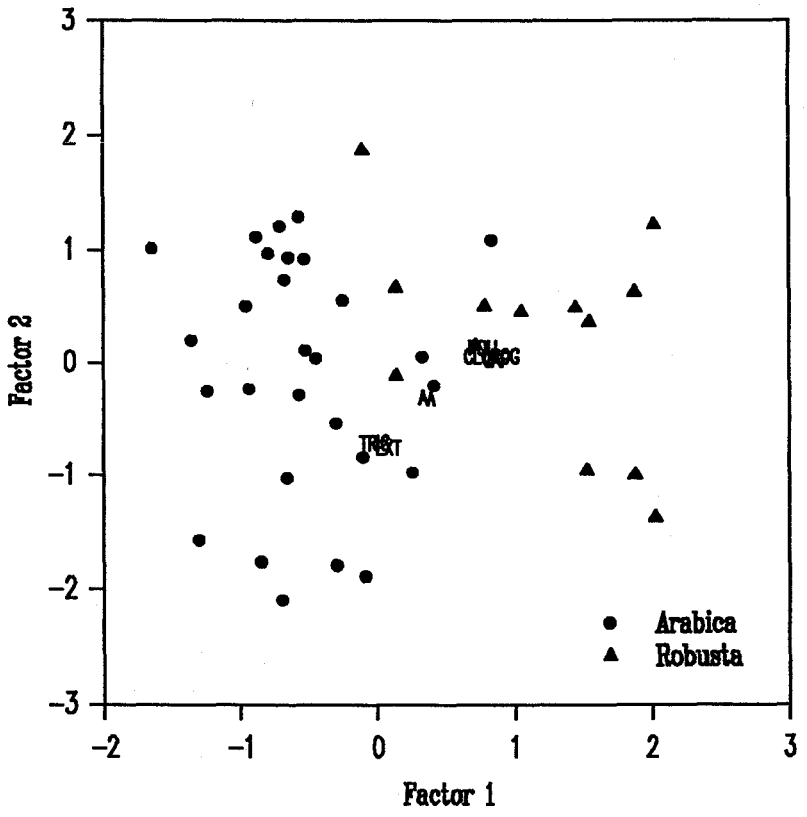

Figura 45. BIPLOT para el perfil no metálico de muestras y descriptores.

Así, vemos cómo los descriptores POLI, CAF y CLOROG son los que más contribuyen al primer factor, lo que nos lleva a pensar que su poder discriminatorio será elevado; ofreciendo, asimismo, una información análoga puesto que sus *loadings* son prácticamente coincidentes. Por el contrario, debido a que los *loadings* de TRIG y EXT están muy próximos a cero para el PC 1, podemos deducir que no van a contribuir a la diferenciación de las muestras. En cuanto a los AA, poseen un valor de *loadings* intermedio y, en principio, su contribución no parece que vaya a ser muy significativa.

Si pasamos a estudiar la distribución de los casos en el nuevo espacio de factores, se observa una buena separación de las muestras *arábicas*, las cuales se sitúan en la parte negativa del primer componente principal (excepto los casos 23A, 29A y 30A), y los cafés robusta ubicados en la parte positiva del eje (exceptuando la muestra 40R).

- Variables de carácter metálico:

En este caso, la matriz de datos se compone de 11 columnas y 41 filas tal y como se muestra en la figura 46.

	ZN	P	MN	FE	MG	CA	NA	K	CU	SR	BA
1A	6.13E-3	0.142	3.54E-3	5.47E-3	0.172	0.110	6.00E-3	1.532	7.69E-3	3.83E-4	3.83E-4
2A	7.04E-4	0.141	2.86E-3	3.91E-3	0.175	9.86E-2	4.22E-3	1.589	1.81E-3	4.58E-4	4.16E-4
3R	1.50E-3	0.188	1.97E-3	3.95E-3	0.177	0.125	5.13E-3	1.606	1.52E-3	3.30E-4	1.76E-4
4A	4.69E-4	0.153	3.36E-3	3.23E-3	0.183	0.093	2.84E-3	1.555	1.75E-3	4.80E-4	5.13E-4
5R	1.34E-3	0.187	1.52E-3	3.48E-3	0.168	0.123	7.50E-3	1.735	2.03E-3	6.17E-4	1.87E-4
6A	6.29E-4	0.148	3.23E-3	2.80E-3	0.184	0.110	5.85E-3	1.381	1.50E-3	5.10E-4	3.47E-4
7R	8.59E-4	0.172	1.53E-3	4.88E-3	0.179	0.127	8.39E-3	1.545	2.25E-3	6.06E-4	2.09E-4
8A	5.33E-4	0.158	3.41E-3	3.51E--3	0.195	0.107	3.91E-3	1.538	1.76E-3	5.43E-4	4.89E-4
9A	3.18E-3	0.150	3.75E-3	3.25E-3	0.201	0.107	5.75E-3	1.431	1.67E-3	1.30E-4	5.49E-4
10R	8.44E-4	0.208	1.50E-3	3.91E-3	0.171	0.123	1.82E-3	1.616	2.04E-3	6.69E-4	5.81E-4
11A	5.90E-4	0.170	3.10E-3	2.71E-3	0.188	0.103	3.50E-3	1.492	1.48E-3	4.90E-4	4.68E-4
12R	1.63E-3	0.180	1.47E-3	3.47E-3	0.181	0.110	5.07E-3	1.693	2.22E-3	5.63E-4	1.77E-4
13A	1.77E-3	0.148	3.30E-3	2.99E-3	0.188	0.110	3.81E-3	1.498	1.43E-3	4.18E-4	6.70E-4
14A	1.43E-3	0.155	2.11E-3	2.86E-3	0.187	0.101	7.56E-3	1.662	1.59E-3	4.30E-4	5.41E-4
15R	1.99E-3	0.201	1.85E-3	2.88E-3	0.166	0.152	6.13E-3	1.664	2.12E-3	8.72E-4	4.30E-4
16A	7.77E-4	0.148	3.46E-3	2.76E-3	0.183	0.103	6.90E-3	1.482	1.62E-3	5.25E-4	5.25E-4
17A	6.98E-4	0.152	4.72E-3	2.48E-3	0.179	0.106	4.25E-3	1.432	1.52E-3	9.53E-4	7.32E-4
18R	9.23E-4	0.176	1.60E-3	4.75E-3	0.186	0.133	9.35E-3	1.619	2.58E-3	6.04E-4	1.65E-4
19R	5.41E-4	0.219	1.61E-3	4.83E-3	0.183	0.137	3.12E-3	1.614	2.21E-3	8.78E-4	5.18E-4
20A	7.50E-4	0.163	3.71E-3	4.42E-3	0.197	0.103	6.88E-3	1.370	1.84E-3	4.57E-4	3.70E-4
21A	1.23E-3	0.161	3.32E-3	3.14E-3	0.197	0.096	5.51E-3	1.211	1.62E-3	4.67E-4	6.00E-4

Figura 46. Matriz de datos con las variables de carácter metálico.

	ZN	P	MN	FE	MG	CA	NA	K	CU	SR	BA
22A	1.60E-3	0.147	3.05E-3	4.02E-3	0.183	0.103	4.76E-3	1.362	1.65E-3	7.24E-4	4.90E-4
23A	1.39E-3	0.145	1.81E-3	3.02E-3	0.179	0.106	4.74E-3	1.308	1.70E-3	5.56E-4	4.78E-4
24A	5.52E-4	0.154	2.66E-3	2.79E-3	0.186	0.133	3.38E-3	1.238	1.80E-3	3.14E-4	2.92E-4
25A	3.64E-4	0.150	3.01E-3	2.86E-3	0.183	0.137	3.08E-3	1.824	1.57E-3	3.86E-4	3.75E-4
26A	5.91E-4	0.156	5.00E-3	2.54E-3	0.197	0.103	6.81E-3	1.677	1.84E-3	1.16E-3	7.85E-4
27R	9.69E-4	0.220	1.51E-3	4.98E-3	0.197	0.096	4.55E-3	1.863	2.31E-3	9.69E-4	4.57E-4
28A	4.91E-4	0.148	3.32E-3	3.08E-3	0.185	0.132	4.50E-3	1.701	1.72E-3	4.80E-4	3.93E-4
29A	4.26E-4	0.152	3.55E-3	3.67E-3	0.187	0.106	3.80E-3	1.652	1.64E-3	4.37E-4	3.71E-4
30A	7.95E-4	0.147	1.95E-3	5.51E-3	0.182	0.099	6.67E-3	1.618	1.58E-3	4.64E-4	5.96E-4
31A	8.39E-4	0.152	2.64E-3	4.30E-3	0.206	0.103	6.80E-3	1.669	1.92E-3	3.40E-4	2.49E-4
32A	3.62E-4	0.150	3.17E-3	2.64E-2	0.198	0.109	4.34E-3	1.828	1.58E-3	4.17E-4	3.51E-4
33A	5.14E-4	0.154	3.87E-3	3.61E-3	0.184	0.110	5.31E-3	1.719	1.54E-3	4.49E-4	3.17E-4
34A	4.88E-4	0.145	2.72E-3	3.54E-3	0.177	0.125	4.48E-3	1.664	1.50E-3	6.10E-4	4.55E-4
35A	5.49E-4	0.157	3.36E-3	2.79E-3	0.187	0.093	5.46E-3	1.645	1.52E-3	4.73E-4	6.81E-4
36R	1.09E-3	0.212	1.45E-3	7.43E-3	0.176	0.145	0.010	1.869	2.49E-3	8.43E-4	6.43E-4
37R	6.29E-4	0.198	1.69E-3	9.33E-3	0.183	0.137	5.48E-3	1.896	2.61E-3	8.13E-4	3.58E-4
38A	5.36E-4	0.159	4.79E-3	3.04E-3	0.179	0.105	2.84E-3	1.517	1.68E-3	5.47E-4	5.91E-4
39A	7.82E-4	0.150	1.62E-3	3.94E-3	0.185	0.115	0.012	1.598	1.64E-3	5.40E-4	4.96E-4
40R	5.38E-4	0.195	1.75E-3	4.66E-3	0.175	0.162	3.08E-3	1.681	2.24E-3	1.00E-3	4.93E-4
41R	1.14E-3	0.186	1.91E-3	7.31E-3	0.160	0.095	3.55E-3	1.809	1.87E-3	4.96E-4	4.85E-4

Figura 46 (cont.). Matriz de datos con las variables de carácter metálico.

Al llevar a cabo un PCA, obtenemos cinco PC's que explican hasta un 79.9% de la varianza, asimismo sus comunalidades en ningún caso fueron menores de 0.7, como puede verse en las tablas 31 y 32.

COMPONENTE	AUTOVALOR	% VARIANZA ACUMULADA
PC1	3.69	33.6
PC2	1.76	49.7
PC3	1.41	62.5
PC4	1.11	72.6
PC5	0.80	79.9

Tabla 31. Varianzas acumuladas para los cinco primeros componentes principales.

	PC1	PC2	PC3	PC4	PC5	COMUNALIDAD
ZN	.119	-.908	.579	.196	.291	.73
P	.844	.241	-.154	-.151	.153	.84
MN	-.671	-.026	-.389	.233	.094	.75
FE	.194	.093	-.360	.545	.110	.74
MG	-.584	.159	.141	.581	.313	.78
CA	.841	.113	-.059	.080	.145	.75
NA	.221	-.252	.616	.417	-.579	.96
K	.611	.238	-.405	.188	-.235	.72
CU	.308	-.861	.047	.073	.088	.76
SR	.449	.203	-.165	.351	.055	.90
BA	-.213	.227	-.418	.128	-.388	.85

Tabla 32. Contribuciones de los PC's a las variables y comunalidades.

La representación BIPLOT se muestra en la figura 47.

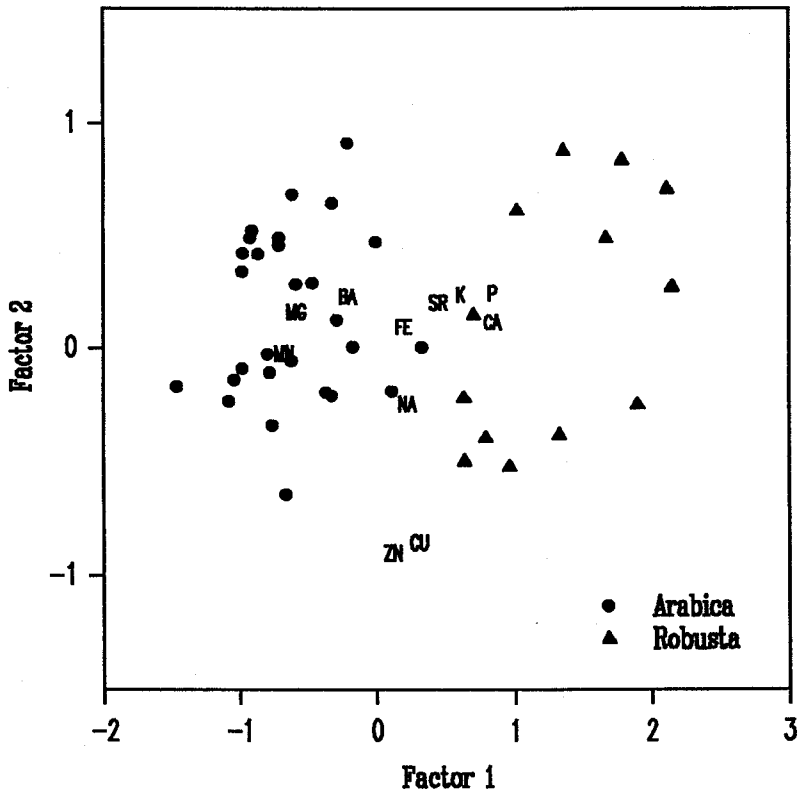

Figura 47. BIPLOT para el perfil metálico de muestras y descriptores.

Se aprecia una mayor importancia significativa para los descriptores P y CA, situados en los valores positivos del PC1, y también para MG y MN, situados en los valores negativos. Las variables FE, NA, ZN y CU seguidas de SR y K, con valores de *loadings* intermedios, no parecen aportar mucha información.

Por lo que respecta a los casos, de nuevo se observa una buena separación entre variedades, incluso mejor que la obtenida con los descriptores de carácter no metálico, siendo la distribución de las muestras arábicas más homogénea que la

correspondiente a los casos de café robusta.

Por tanto, una primera conclusión que se puede deducir de la aplicación del PCA es que, en cuanto a la tendencias de nuestras muestras, se observa una clara separación de las muestras de café verde atendiendo a la variedad de la planta: arábica y robusta. Por otra parte, en lo referente a los descriptores, las variables de carácter no metálico más significativas a la hora de diferenciar entre variedades parecen ser CAF, POLI y CLOROG, así como las de carácter metálico P (o bien CA) y MG (o bien MN).

El contenido en CAF es una de las principales diferencias que caracterizan la variedad a la cual pertenece el grano de café, tal y como queda reflejado en el BIPLOT, siendo además muy utilizada a la hora de realizar las mezclas para la elaboración del café comercial.

En lo referente al contenido metálico, es claro el poder discriminatorio de P, cuyo contenido es mucho mayor en los cafés robusta. El origen del fósforo en el grano de café es debido a fuentes naturales de este elemento en el suelo, aunque puede influir en gran medida la adición a la tierra de cultivo de abonos fosforados.

V.2. RECONOCIMIENTO DE PATRONES NO SUPERVISADO

V.2.1. Análisis Cluster

A pesar de que el Análisis Cluster es una técnica de reconocimiento de patrones no supervisada y, por tanto, se aplica en el caso de que no se conozcan a priori la pertenencia de las muestras a las distintas categorías, vamos a aplicarla para corroborar los resultados del PCA y ver cómo se agrupan naturalmente las muestras de café verde.

- Variables de carácter no metálico:

Como hemos visto en el BIPLOT que las variables más significativas son POLI, CLOROG y CAF, realizamos un cluster jerárquico de los casos empleando la distancia euclídea como medida de similaridad y el Método de Ward como regla de amalgamación. En el dendrograma obtenido se observan clusters en los que hay una mayoría de muestras pertenecientes a una misma clase, aunque se aprecian mezclas de categorías. Este hecho nos hace pensar que debiéramos emplear sólo una de las variables (POLI, CLOROG o CAF) ya que proporcionan el mismo tipo de información; así pues, llevamos a cabo de nuevo el análisis cluster en las mismas condiciones pero, en este caso, utilizando las variables CAF y AA (por ser estos últimos los que poseen un valor de *loading* intermedio). En la figura 48, se ofrece el dendrograma obtenido en el que se observa una clara separación de los casos en dos grandes clusters, conteniendo cada uno muestras de una variedad distinta; esto ocurre incluso a distancias casi inferiores al 10% de la máxima.

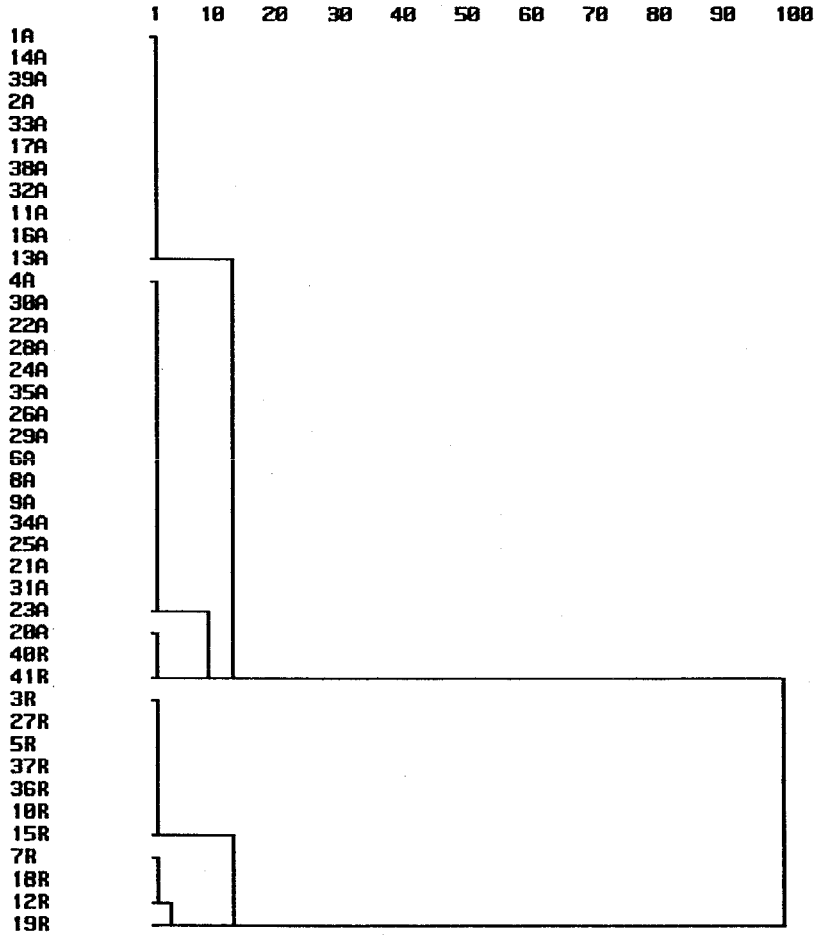

Figura 48. Dendrograma de las muestras de café usando las variables CAF y AA.

- Variables de carácter metálico:

Al realizar el preprocesado, habíamos visto que las variables más significativas eran P y MG, por lo que llevamos a cabo un análisis cluster jerarquizado de casos empleando la distancia euclídea y el método de Ward.

En esta ocasión, se obtiene un dendrograma (figura 49) donde a partir de distancias incluso inferiores al 20% de la máxima, aparecen dos clusters uno que corresponde a la categoría arábica y el otro correspondiente a muestras de variedad robusta, exceptuando el caso número 11A que como indica su código es arábica y se ubica en el racimo de la clase robusta.

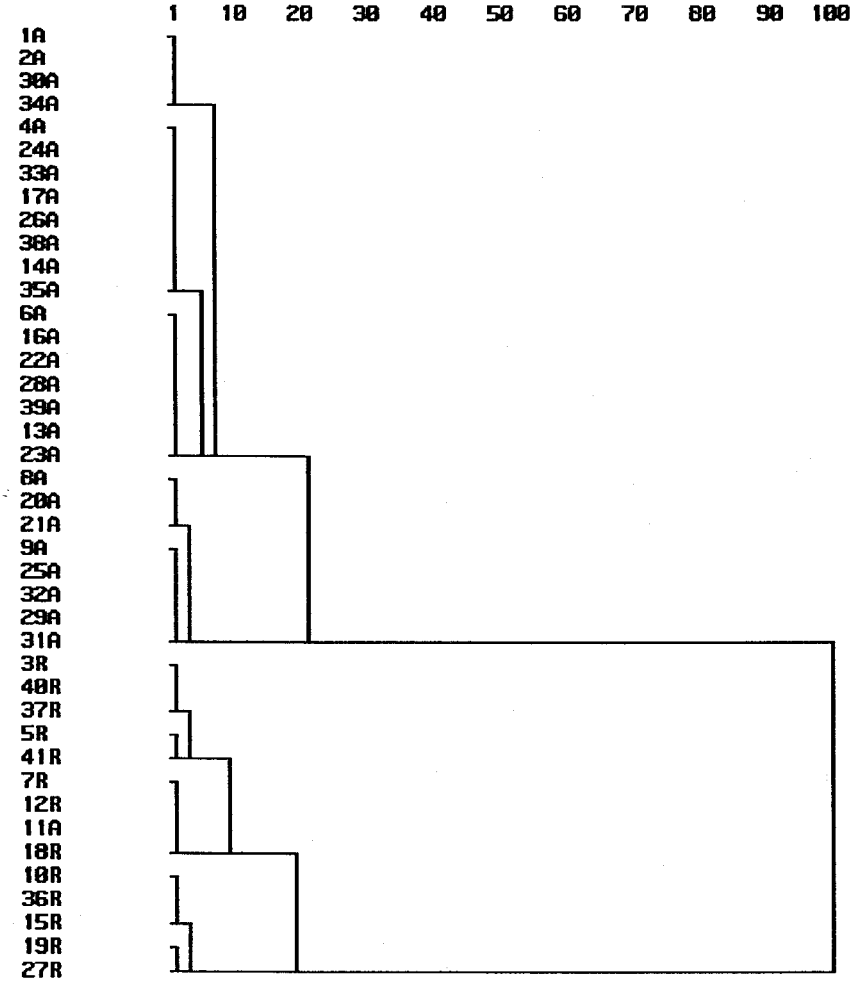

Figura 49. Dendrograma de las muestras de café usando las variables P y MG.

Esto confirma los resultados del PCA para los descriptores metálicos, es decir, que P y MG eran las variables más significativas a la hora de discriminar entre las dos variedades.

El mismo dendrograma se obtiene cuando se emplean, en las mismas condiciones, las variables P, MG y BA.

Con toda la información obtenida hasta ahora, después de llevar a cabo los métodos de preprocesado y análisis cluster y ya que, en nuestro caso, conocemos a priori la categoría de las muestras estudiadas vamos a aplicar métodos de reconocimiento de patrones supervisados que nos permitan establecer y definir reglas de clasificación para asignar categorías a las muestras.

V.3. RECONOCIMIENTO DE PATRONES SUPERVISADO

Antes de llevar a cabo el análisis, es conveniente realizar una selección de variables en cuanto a su poder discriminante; para ello se calculan los siguientes parámetros estadísticos para los distintos descriptores: valor medio ($\bar{x}$) y varianza (var) de los datos para las dos clases arábica (A) y robusta (R). Todo ello se muestra en las tablas 33 y 34.

	$\bar{x}$ (A)	$\bar{x}$ (R)	var(A)	var(R)
EXT	25.323	25.092	5.459	2.929
POLI	5.904	7.776	0.959	0.733
AA	0.218	0.220	1.24E-3	1.789
TRIG	1.334	1.318	0.061	0.123
CLOROG	3.613	4.291	0.269	0.421
CAF	1.231	2.349	0.026	0.174

Tabla 33. Valores medios y varianza de las variables no metálicas para cada clase.

	$\bar{x}$ (A)	$\bar{x}$ (R)	var(A)	var(R)
ZN	1.04E-3	1.08E-3	1.34E-6	1.92E-7
P	0.152	0.196	4.14E-5	2.51E-4
MN	3.12E-3	1.64E-3	9.49E-7	3.09E-8
FE	4.24E-3	5.07E-3	1.95E-5	3.47E-6
MG	0.187	0.176	7.13E-5	5.68E-5
CA	0.107	0.132	7.29E-5	3.17E-4
NA	5.21E-3	5.63E-3	3.62E-6	6.48E-6
K	1.543	1.708	0.026	0.013
CU	1.86E-3	2.19E-3	1.32E-6	8.70E-8
SR	5.47E-4	7.12E-4	5.24E-8	4.0E-8
BA	4.83E-4	3.75E-4	1.82E-8	2.96E-8

Tabla 34. Valores medios y varianza de los descriptores para cada clase.

Con estos valores, se pueden calcular los Pesos de Fisher para cada una de las variables de acuerdo a la expresión:

$$FW(A,R) = \frac{|\bar{x}(A) - \bar{x}(R)|}{var(A) + var(R)} \qquad (V.1)$$

Los valores de pesos de Fisher obtenidos para cada grupo de variables se ofrecen en las tablas 35 y 36.

	EXT	POLI	AA	TRIG	CLOROG	CAF
FW	0.03	1.11	0.61	0.09	0.98	5.58

Tabla 35. Pesos de Fisher para las variables de carácter no metálico.

	FW
ZN	22.15
P	149.11
MN	1504.62
FE	36.14
MG	88.88
CA	64.22

	FW
NA	41.10
K	4.24
CU	233.79
SR	1784.92
BA	2258.99

Tabla 36. Pesos de Fisher para las variables de carácter metálico.

Observando estos valores parece ser que las variables de carácter no metálico con mayor poder discriminante son CAF y POLI y, dentro del grupo de descriptores metálicos, los de mayor poder discriminante son SR seguido de MN.

En primer lugar, vamos a aplicar al conjunto de muestras uno de los primeros métodos establecidos para aprendizaje supervisado.

V.3.1. Máquina de Aprendizaje Lineal (LLM)

Este es un método iterativo y completamente empírico en el cual se pretende separar clases linealmente separables mediante un plano el cual, viene caracterizado por un vector unitario. Una vez realizado un autoescalado de los datos, se elige al azar un vector representativo del plano de separación el cual se multiplica por cada uno de los vectores representativos de cada una de las muestras; según el signo del producto escalar, la muestra pertenecerá a una u otra categoría. A continuación, se va comprobando si los casos han sido bien clasificados modificando el plano de separación en caso de clasificación errónea. El proceso se va repitiendo de forma iterativa hasta obtener una separación óptima de las muestras, creando fronteras lineales entre las clases existentes.

- Variables de carácter no metálico:

En nuestro caso, cada muestra de café va estar definida por un vector de tres componentes: contenido en polifenoles totales, contenido en ácido clorogénico y contenido en cafeína. La velocidad de mejora η se va optimizando durante el proceso de aprendizaje, partiendo del valor 1 y disminuyendo en 0.1 en cada

iteración. El conjunto completo formado por las 41 muestras, se ha dividido aleatoriamente en un conjunto de entrenamiento y otro de ensayo.

Este proceso se ha repetido cinco veces y el resultado obtenido fue, en todos los casos, de un 100% de eficacia en la clasificación tanto en el reconocimiento como en la predicción. La velocidad de mejora fue η=0.98, η=0.97, η=0.97, η=0.98 y η=0.98.

- Variables de carácter metálico:

Para el perfil metálico, cada muestra está representada por un vector de tres componentes: contenido en fósforo, contenido en magnesio y contenido en bario.

Se ha seguido un proceso análogo al explicado anteriormente para el perfil no metálico, obteniéndose también en este caso un 100% en la capacidad de reconocimiento y en la capacidad de predicción. En este caso, los valores finales para las distintas velocidades de mejora son: η=0.98, η=0.97, η=0.97, η=0.97 y η=0.98.

Seguidamente, vamos a aplicar métodos tanto paramétricos como no paramétricos para construir las funciones discriminantes que distingan entre las dos categorías a las que pertenecen las muestras de café estudiadas.

V.3.2. Métodos paramétricos

* Análisis Discriminante Lineal (LDA)

Para llevar a cabo esta técnica, generalmente se divide el total de las muestras en dos grupos: un conjunto de aprendizaje (training set) y un conjunto de prueba (test set). Con el primero, que estará constituido por un 75% de los casos, se construye la regla de clasificación, cuya eficacia se comprueba aplicando la regla al conjunto de ensayo (25% restante de las muestras).

En nuestro caso, hemos llevado a cabo el llamado Método de Dejar Uno Fuera (*Leave One Out*), proceso iterativo en el que cada vez se toman todos los casos formando parte del conjunto de aprendizaje menos uno que se deja fuera al azar (test set), se construye la regla de clasificación y ésta se aplica al test set; a continuación, se repite el proceso dejando fuera un caso distinto; así, se sigue hasta completar todo el conjunto de muestras.

El criterio de ponderación de las variables es distinto, en esta técnica, al explicado de los Pesos de Fisher; en este caso, se emplea el Criterio λ de Wilks que minimiza el cociente entre la varianza intraclase y la varianza interclase (o maximiza el cociente inverso).

- Variables de carácter no metálico:

Se lleva a cabo un LDA sobre la matriz de datos correspondiente, aplicando el método *Leave One Out* y empleando el método de Backward Stepwise. La clasificación se basa en las probabilidades *a posteriori.*

El resultado del análisis así realizado es que los descriptores con mayor poder diferenciador entre clases son POLI y CAF. Se construyeron funciones de clasificación con estas dos variables, cuyas expresiones se muestran a continuación:

CLASE	FUNCION DE CLASIFICACION
ARABICA	-35.0328 + 20.6718 * CAF + 7.4286 * POLI
ROBUSTA	-84.8515 + 37.5622 * CAF + 10.1793 * POLI

Tabla 37. Funciones discriminantes para las clases arábica y robusta.

No se obtiene ningún fallo en la asignación de las clases por lo que la efectividad de la clasificación es del 100%.

- Variables de carácter metálico:

Se realiza el mismo tipo de análisis sobre el perfil metálico de las muestras, empleando también el método backward stepwise y se obtienen funciones de clasificación solamente con los descriptores P, MG y BA, que son las variables más significativas. Igualmente, el resultado es un 100% de efectividad en la clasificación.

Las funciones discriminantes se muestran en la siguiente tabla.

CLASE	FUNCIONES DE CLASIFICACION
ARABICA	-371.69 + 794.70 * P + 3183.94 * MG + 50068.95 * BA
ROBUSTA	-384.45 + 1528.62 * P + 2640.35 * MG + 10816.36 * BA

Tabla 38. Funciones discriminantes para las clases arábica y robusta.

*** SIMCA (Soft Independent Modelling of Class Analogy)**

Es un procedimiento blando de modelización separada de clases vía PCA, ya que en esta técnica de reconocimiento de patrones supervisado se realiza un análisis en componentes principales por separado para cada una de las categorías existentes según el algoritmo NIPALS. A continuación, se va ajustando cada objeto para cada una de las clases para obtener así el porcentaje de eficacia en la clasificación tanto para el conjunto de aprendizaje como para el conjunto de ensayo (valores de F a un nivel de confianza del 95%).

Para la aplicación de esta técnica, la matriz de datos está compuesta por 30 casos, de los cuales 17 son de la variedad arábica y 13 robusta. La razón por la que sólo fueron empleados 30 de los 41 casos disponibles es la limitación impuesta por el paquete estadístico (SIRIUS) utilizado para aplicar esta técnica. Debemos mencionar que se eligieron 30 muestras representativas del conjunto total; así dichos casos son: 2A, 3R, 4A, 5R, 6A, 7R, 9A, 10R, 12R, 13A, 14A, 15R, 16A, 17A, 18R, 19R, 21A, 22A, 23A, 25A, 26A, 27R, 30A, 34A, 35A, 36R, 37R, 38A, 40R y 41R.

- *Variables de carácter no metálico:*

En este caso, hemos utilizado los tres descriptores más significativos: POLI, CAF y CLOROG.

El conjunto de aprendizaje para la clase arábica está constituido por los casos: 4A, 6A, 9A, 14A, 16A, 17A, 21A, 22A, 23A, 25A, 30A, 35A y 38A, elegidos aleatoriamente. Al realizar el PCA sobre la clase, se obtuvo un primer factor que dio cuenta del 81% de la varianza total de los datos y un segundo factor que explicó hasta el 99.5% de la varianza. Ninguno de estos dos PC's resultó ser significativo, por lo que no nos quedamos con ningún factor para realizar el análisis.

Análogamente, procedemos con la categoría de cafés robusta, donde el *training set* está constituido por las diez muestras siguientes: 3R, 5R, 7R, 10R 12R, 15R, 18R, 19R, 27R y 36R, también elegidas al azar. Se aplica un PCA sobre la clase, obteniendo un primer PC que da cuenta solo de un 55.2% de la varianza y con el segundo se explica hasta un 95.1%. De nuevo, ninguno de los dos PC's resultó ser significativo por lo que realizamos el análisis sin tener en cuenta ningún factor.

A continuación, realizamos el SIMCA sobre el total de los 30 casos de que disponemos; así, podremos ver la "capacidad de reconocimiento" y la "capacidad de predicción" al mismo tiempo. El resultado obtenido no fue satisfactorio en ninguno de los casos, por lo que decidimos forzar el ajuste teniendo en cuenta los dos PC's calculados anteriormente para cada una de las clases a pesar de no ser significativos. De esta forma, al llevar a cabo el análisis obtuvimos una capacidad de reconocimiento del 100% y un 93% de efectividad en la capacidad de predicción.

- *Variables de carácter metálico:*

En esta ocasión, las variables empleadas para el análisis son P, MG y BA que eran las de mayor poder discriminante según el LDA realizado previamente. Se utilizaron los mismos conjuntos de aprendizaje empleados para las variables de carácter no metálico y se procedió a extraer los componentes principales de cada clase.

Para la clase arábica, el primer PC explica un 69.4% de la varianza y el segundo factor llega al 100%. Para la variedad robusta, el primer PC da cuenta de un 79.4% y con los dos primeros factores se llega a explicar el 100% de la varianza.

De nuevo, ninguno de los PC's fue significativo, por tanto, procedimos de forma análoga a lo realizado para las variables de carácter no metálico. Se llevó a cabo el SIMCA sobre el total de las 30 muestras sin coger ningún factor y se obtuvo un 100% de eficacia tanto para la capacidad de reconocimiento como para la de predicción. Idéntico resultado se obtuvo al forzar el análisis utilizando los dos factores calculados para cada clase.

Podemos concluir que el perfil metálico de las muestras de café es más homogéneo que el perfil no metálico puesto que las muestras se ajustan perfectamente al modelo sin tener en cuenta ningún PC.

V.3.3. Métodos no paramétricos

*** Método de los K vecinos más próximos (KNN)**

Seguidamente, aplicamos un método duro de modelización.

- Variables de carácter no metálico:

Análogamente a las técnicas aplicadas anteriormente, de las seis variables iniciales vamos a utilizar sólo las tres más significativas: POLI, CLOROG y CAF.

Se realiza un Leave One Out y se selecciona así el número de vecinos que "van a dar su voto". Al llevar a cabo un 3NN se obtiene un 95.2% de eficacia en la asignación de clases, puesto que las muestras 40R y 41R las clasifica como pertenecientes a la variedad arábica cuando son de variedad robusta.

Cuando se realiza un 3NN con las variables POLI y CAF, solamente se obtiene un fallo en la clasificación (muestra 41R), lo que supone un 97.5% de efectividad en la asignación de clases.

- Variables de carácter metálico:

Los descriptores con mayor poder discriminante una vez aplicadas las técnicas anteriores, hemos visto que parecen ser P, MG y BA. Con ellas vamos a realizar un 3NN en las mismas condiciones explicadas para las variables de carácter no metálico. El resultado es un 100% de efectividad en la asignación de clases, puesto que todos los casos quedan clasificados correctamente.

V.4. REDES NEURONALES ARTIFICIALES

Una vez realizado el análisis multivariante con los Métodos convencionales de Reconocimiento de Patrones, vamos a llevar a cabo un análisis de la información obtenida con los datos químicos de que disponemos pero, esta vez, empleando las llamadas *Redes Neuronales Artificiales.*

V.4.1. Métodos de aprendizaje no supervisado

* Mapa Autoorganizativo (SOM)

Este tipo de red neuronal no requiere un conocimiento previo de la clase a la que pertenecen las muestras. Es una red monocapa que va a trabajar atendiendo a la relación topológica de los datos. Así, señales de entrada similares, es decir, muestras de café con contenidos parecidos de las variables seleccionadas, van a activar neuronas vecinas. Dentro de la capa activa, se elige una neurona central cuya salida es la más parecida a la señal de entrada. Una vez corregidos los pesos de esta neurona central y los pesos del número de anillos vecinos establecido, la señal de salida final va a ser bidimensional; en el problema que nos ocupa, clase arábica o clase robusta.

- Variables de carácter no metálico:

Las señales de entrada van a ser los contenidos en polifenoles, ácido clorogénico y cafeína, junto con una etiqueta para denotar la clase a la que pertenece cada muestra. El aprendizaje de la red se va a realizar con todas las muestras de café, empleando un factor de aprendizaje de η=0.2.

Se van probando distintos tamaños de red: 40 x 40, 50 x 50 y 60 x 60.

1) Cuando se utiliza una red de dimensión 40 x 40, siendo el número de vecinos que deciden la clase igual a tres y corrigiendo los 20 anillos más próximos a la neurona central, el error de cuantificación es de un 1.847.

2) Probamos, seguidamente con una matriz de 50 x 50. Número de anillos que se modifican igual a 25, número de vecinos que deciden la clase igual a tres y el factor de aprendizaje sigue siendo η=0.5. El error es ahora de 1.686. Si se mantiene el mismo tamaño de red pero sólo se corrigen los 20 anillos más próximos, el error aumenta a 2.403.

3) Al emplear una red mayor, de 60 x 60, se corrigen 30 anillos y el número de vecinos que deciden la clase es también tres, pero en esta ocasión utilizamos un factor de aprendizaje de η=0.8, el error vuelve a aumentar siendo su valor de 1.969.

Se siguieron realizando distintas modificaciones pero no se mejoró el error cometido, por lo que nos quedamos con la red SOM de dimensión 50 x 50, corregimos 25 anillos vecinos con un factor de aprendizaje η=0.5 y deciden la clase tres vecinos.

El mapa autoorganizativo que resulta se ofrece en la figura 50.

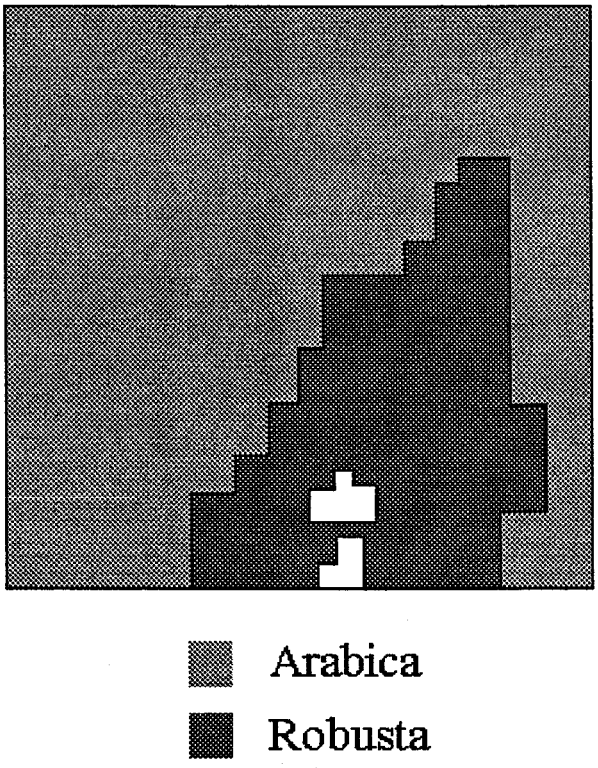

Figura 50. Mapa autoorganizativo de las muestras de café según el perfil no metálico.

Se aprecia cómo las muestras de café pertenecientes a la variedad robusta aparecen localizadas en la misma zona del mapa, rodeadas por los casos de la variedad arábica.

Por tanto, se confirma una relación entre la similitud en los contenidos de POLI, CLOROG y CAF y la topología que presentan las muestras, puesto que casos de la misma variedad poseen una ubicación muy próxima. También pueden

observarse huecos dentro del grupo de muestras robusta, lo cual puede ser indicativo de una menor homogeneidad en esta clase con respecto a la clase de cafés arábicas; hecho que se sospechó al examinar la representación BIPLOT correspondiente.

- *Variables de carácter metálico:*

En esta ocasión, las señales de entrada a cada neurona van a ser los contenidos en fósforo, magnesio y bario, asignando una etiqueta relativa a la variedad de cada muestra.

Se han ensayado distintos tamaños de matrices, pero hasta llegar a una matriz 70 x 70 no se consigue una aparición en el mapa de las dos clases existentes.

En la figura 51, se muestra la representación SOM correspondiente.

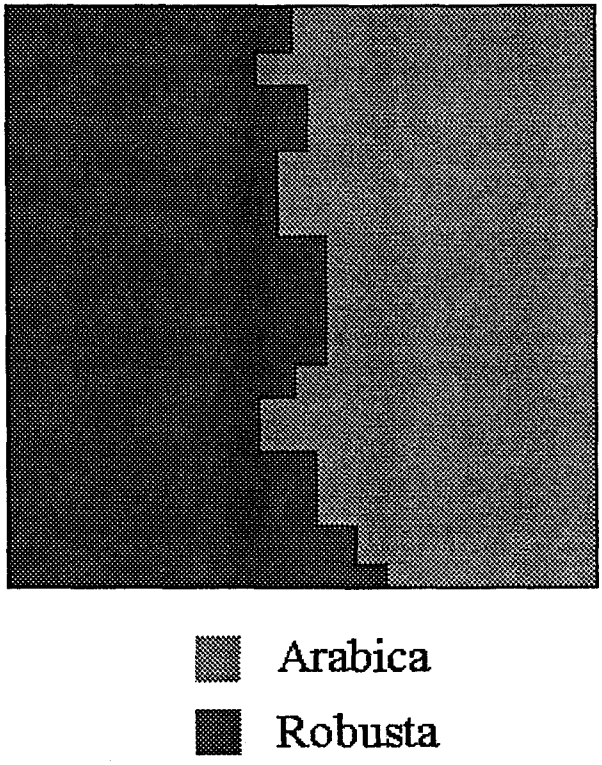

Figura 51. Mapa autoorganizativo de las muestras de café según el perfil metálico.

La dimensión de la matriz es 70 x 70, el factor de aprendizaje es η=0.5, el número de vecinos que deciden la clase es igual a tres y los anillos a modificar son 35. El error de cuantificación que se obtiene es 6.34E-3.

V.4.2. Métodos de aprendizaje supervisado

* Aprendizaje por Retropropagación

En esta técnica, la red está compuesta por varias capas de neuronas interconectadas de forma total. En este caso, cada muestra lleva asociado un"target" u objetivo, que indica la clase a la que pertenece.

Generalmente, la red consta de tres capas: la capa de entrada contiene tantas unidades básicas como descriptores y es inactiva; la siguiente se denomina capa oculta y recibe la salida ponderada de cada neurona de la capa de entrada; sobre esta entrada actúa una función de transferencia, la cual suele ser de forma sigmoidea. El resultado de esta operación se pondera de nuevo y constituye la entrada de la última en la que se repite el mismo proceso anterior, obteniéndose una salida alusiva a la clase a la cual pertenece la muestra introducida en la red neuronal.

Una vez obtenida esta salida, se compara con el target asociado a la muestra; si no se ha realizado una clasificación correcta, se corrigen los pesos. Esta corrección es función de dos parámetros η=velocidad de aprendizaje y μ=momento, siguiendo también la Regla Delta. Una vez llevada a cabo dicha corrección, los datos de entrada vuelven a pasar por la red y se sigue el mismo proceso iterativamente hasta que no se obtiene ninguna mejora en la salida, o bien se alcanza un número determinado de iteraciones, también llamadas "epochs".

- Variables de carácter no metálico:

Las muestras de café van a estar descritas por tres variables: POLI, CLOROG y CAF, por lo que la capa de entrada de la red va a tener tres neuronas.

La capa de salida de la red consta de dos neuronas, puesto que dicha salida corresponde a cada una de las clases existentes.

El conjunto total de muestras se dividió al azar en dos: conjunto de aprendizaje y conjunto de ensayo, repitiendo siempre cada ensayo con diez conjuntos diferentes. Previamente a la aplicación de la técnica, los datos se normalizaron entre 1 y -1. Los valores de pesos iniciales se tomaron de forma aleatoria entre -0.1 y 0.1.

Se han probado distintas arquitecturas de la red neuronal y distintas funciones de transferencia, empleando siempre tres capas de neuronas.

* Función sigmoidea:

1) En un primer ensayo, se ha utilizado una red 3x3x2 (más bias), donde la velocidad de aprendizaje y el momento permanecieron fijos durante todo el proceso η=0.2 y μ=0.5. Una vez realizadas 150 iteraciones, se obtuvo una eficacia del 100% en la capacidad de reconocimiento y de un 91% en la capacidad de predicción cuando el error permitido era de 0.1. Si el error se aumenta a 0.2, se obtiene un 100% de eficacia en el reconocimiento y 93% en la predicción. Se alcanzaron 400 iteraciones.

2) Seguidamente, se probó una arquitectura de red 3x2x2 (más bias), donde la velocidad de aprendizaje y el momento también permanecieron fijos durante todo el proceso con los mismos valores anteriores. Cuando el error máximo permitido era de 0.1, se obtuvo

una eficacia de 100% en la capacidad de reconocimiento y un 95% en la capacidad de predicción. Se alcanzaron 400 iteraciones. Sin embargo, cuando se aumentó el error a 0.2, se obtuvo un 100% en la capacidad de reconocimiento pero sólo un 88% en la predicción. El número de iteraciones fue de 160.

3) Por último, se cambió la arquitectura de la red a 3x1x2, manteniendo los mismos valores de velocidad de aprendizaje y momento. Si el error permitido es de 0.1, al llegar a 350 épocas se alcanza un 100% de eficacia en el reconocimiento y un 95% en la predicción. Si el error es de 0.2, en 100 iteraciones se llega a un 100% de eficacia en la capacidad de reconocimiento y un 94% en la predicción.

A continuación, vamos a emplear otro tipo de función de transferencia.

* Función sigmoidea + 0.1:

Se ha probado con la arquitectura de red 3x1x2 con un error permitido de 0.1 y al llegar a 200 épocas se consigue un 100% de eficacia en la capacidad de reconocimiento y un 96% en la predicción. Si disminuimos el error a 0.05, se necesitan 400 iteraciones para alcanzar un 100% de eficacia en el reconocimiento y pero la capacidad de predicción disminuye a un 94%.

Se siguió modificando la arquitectura de la red neuronal, pero no se consiguió mejorar los resultados.

En consecuencia, la red neuronal óptima para establecer la separación de nuestras muestras de café atendiendo a la variedad a la que pertenecen, es:

3x1x2 η=0.2 μ=0.5 Función sigmoidea + 0.1

A continuación, se muestra una representación gráfica de dicha red.

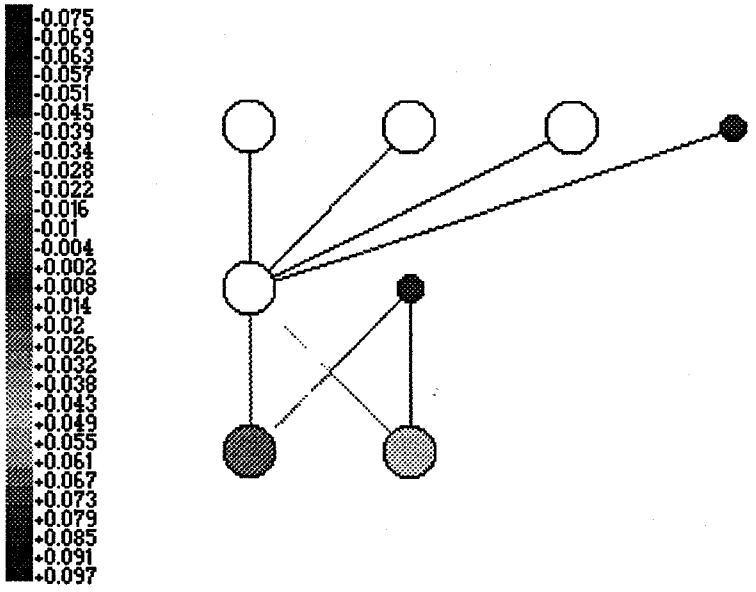

Figura 52. Arquitectura óptima de la red neuronal.

- *Variables de carácter metálico:*

Para el perfil metálico, la señal de entrada a la red será el contenido en fósforo, magnesio y bario de las muestras de café.

Se van a probar las mismas estructuras de la red neuronal que las estudiadas para el perfil no metálico. Así:

* Función sigmoidea:

1) 3x3x2 η=0.2 μ=0.5. El error máximo permitido es 0.1. Al realizar 200 iteraciones, se consigue un 100% en la capacidad de reconocimiento y un 94% en la predicción. Si el error permitido es 0.2, seguimos obteniendo un 100% en el reconocimiento y llegamos a un 96% en la predicción, sólo con 100 epochs.

2) 3x2x2 η=0.2 μ=0.5. Error=0.1. Una vez alcanzadas las 250 iteraciones, se obtiene un 100% de eficacia en el reconocimiento y un 97% en la predicción. Con un error de 0.2, al llegar a 100 iteraciones, se vuelve a obtener un 100% en el reconocimiento y un 95% en la predicción.

3) 3x1x2 η=0.2 μ=0.5. Error=0.1. Epochs=200. 100% de eficacia en la capacidad de reconocimiento y 97% en la capacidad de predicción. Al aumentar el valor del error permitido a 0.2, al alcanzar 80 iteraciones, se consigue un 100% en el reconocimiento y se baja a un 96% de eficacia en la predicción.

Análogamente a lo realizado para el caso de las variables de carácter no metálico, vamos a probar con otra función de transferencia para las arquitecturas de red que dieron un mayor porcentaje de eficacia.

* Función sigmoidea + 0.1:

1) 3x2x2 η=0.2 μ=0.5. Error=0.1. Con 250 iteraciones, se consigue un 100% de eficacia tanto para la capacidad de reconocimiento como para la capacidad de predicción. Los mismos resultados se obtienen para un error máximo permitido de 0.2; en este caso, hay que realizar 300 iteraciones.

2) 3x1x2 η=0.2 μ=0.5. Error=0.1. Hacen falta tan sólo 100 epochs para conseguir un 100% de eficacia tanto en el reconocimiento como en la predicción. Al disminuir el error permitido a 0.05, también se obtiene un 100% de eficacia con el conjunto de aprendizaje y con el conjunto de ensayo, pero aumenta un ligeramente el número de iteraciones a 150.

Se puede concluir que tan sólo utilizando una neurona en la capa oculta de la red (arquitectura 3x1x2), ésta es capaz de clasificar bien la totalidad de las muestras de café, empleando como datos de entrada el valor de los descriptores metálicos más significativos. Dicha red se muestra en la siguiente figura.

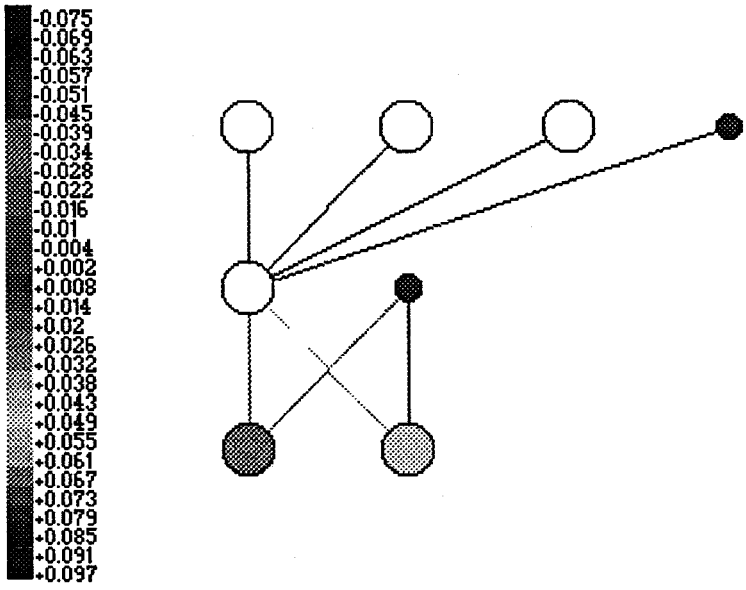

Figura 53. Arquitectura óptima de la red neuronal.

* LVQ

En esta técnica, las neuronas de la red se tratan como vectores, cuya dimensión corresponde al número de descriptores con mayor poder discriminante entre las muestras, cada uno de estos vectores lleva asociado un determinado peso.

El objetivo del método es encontrar los "*codebook vectors*" especializados en reconocer la pertenencia de los casos a una clase determinada. Los pesos asociados a dichos vectores se van a ir modificando siguiendo un aprendizaje competitivo a medida que se van introduciendo cada uno de los casos que componen el training

set.

Así, se va optimizando cada *codebook vector* para cada una de las categorías existentes y, una vez obtenidos dichos vectores, la efectividad del método a la hora de predecir pertenencia a clases de las muestras del conjunto de ensayo, se lleva a cabo empleando la técnica de los k vecinos más próximos.

- *Variables de carácter no metálico:*

Cada muestra va a ser un vector de tres componentes: polifenoles, ácido clorogénico y cafeína a los cuales, se les va a asignar una clase conforme va progresando el aprendizaje realizando un 5NN. El factor de aprendizaje inicial fue η=1 y el número de "*codebook vectors*" es dos, uno por cada una de las categorías a las que pertenecen las muestras de café.

El conjunto de las 41 muestras de café se ha dividido de forma aleatoria en dos grupos, para formar el conjunto de entrenamiento y el conjunto de ensayo. La técnica se ha aplicado cinco veces obteniendo un 100% de eficacia en la capacidad de reconocimiento y un 98% en la capacidad de predicción.

- *Variables de carácter metálico:*

En este caso, los casos se tratan como vectores también de tres componentes: contenido en fósforo, contenido en magnesio y contenido en bario.

El total de muestras de café se dividió al azar en un conjunto de entrenamiento y uno de ensayo. El factor de aprendizaje η inicial fue igual a 1 y se tomaron dos "*codebook vectors*". El método se ha aplicado cinco veces, siendo

ambos conjuntos de entrenamiento y ensayo distintos cada vez. El resultado obtenido fue excelente, ya que se consiguió un 100% de eficacia tanto en la capacidad de reconocimiento como en la capacidad de predicción.

A continuación, en la tabla 39, mostramos un resumen de los resultados obtenidos al llevar a cabo cada uno de los métodos, tanto quimiométricos como algoritmos neuronales, sobre el conjunto de las muestras de café.

METODO	PERFIL NO METALICO (POLI, CLOROG, CAF)		PERFIL METALICO (P, MG, BA)	
	% Recon.	% Predic.	% Recon.	% Predic.
LDA	100	100	100	100
SIMCA	100	93	100	100
KNN	95		100	
LLM	100	100	100	100
BPL	100	96	100	100
LVQ	100	98	100	100

Tabla 39. Eficacia de reconocimiento y predicción de las técnicas aplicadas.

V.5. TRATAMIENTO CON LA MATRIZ REDUCIDA

Hasta ahora, hemos aplicado los distintos métodos pertenecientes a la quimiometría clásica por un lado a los descriptores metálicos y por otro a las variables de carácter no metálico; de forma análoga se ha procedido con las técnicas basadas en redes neuronales artificiales.

Así, para cada uno de los perfiles: metálico y no metálico, se han estudiado cuales de los descriptores analizados en las muestras eran los que poseían un mayor poder discriminante. Una vez realizados llevados a cabo todos los métodos, podemos concluir que las variables más significativas y que mejor diferencian los casos, atendiendo a la clase a la cual pertenecen (arábica o robusta) son POLI, CLOROG, CAF, P, MG y BA.

Por tanto, seguidamente vamos a volver a emplear las técnicas de reconocimiento de patrones pero, esta vez, la matriz de datos va a estar constituida por 41 filas (total de muestras de café) y 6 columnas (las variables que acabamos de enumerar). Esta nueva matriz de datos se muestra en la figura 54.

	POLI	CLOROG	CAF	P	MG	BA
1A	5.0	3.2	0.9	0.142	0.172	3.83E-4
2A	5.2	3.4	1.1	0.141	0.175	4.16E-4
3R	7.5	4.9	2.2	0.188	0.177	1.76E-4
4A	5.0	3.9	1.3	0.153	0.183	5.13E-4
5R	9.5	4.2	2.3	0.187	0.168	1.87E-4
6A	5.6	3.7	1.2	0.148	0.184	3.47E-4
7R	7.8	5.6	2.7	0.172	0.179	2.09E-4
8A	5.9	3.9	1.3	0.158	0.195	4.89E-4
9A	5.9	3.5	1.3	0.150	0.201	5.49E-4
10R	8.2	4.6	2.4	0.208	0.171	5.81E-4
11A	5.2	4.1	1.2	0.170	0.188	4.68E-4
12R	8.4	4.7	2.8	0.180	0.181	1.77E-4
13A	6.1	3.7	1.1	0.148	0.188	6.70E-4
14A	6.4	3.4	1.0	0.155	0.187	5.41E-4
15R	7.4	3.8	2.4	0.201	0.166	4.30E-4
16A	7.2	3.7	1.1	0.148	0.183	5.25E-4
17A	8.2	3.7	1.1	0.152	0.179	7.32E-4
18R	7.8	4.6	2.7	0.176	0.186	1.65E-4
19R	7.0	4.7	3.2	0.219	0.183	5.18E-4
20A	5.8	3.9	1.8	0.163	0.197	3.70E-4
21A	5.7	3.1	1.2	0.161	0.197	6.00E-4

Figura 54. Matriz de datos con las variables más significativas.

	POLI	CLOROG	CAF	P	MG	BA
22A	7.0	3.4	1.3	0.147	0.183	4.90E-4
23A	7.5	4.0	1.2	0.145	0.179	4.78E-4
24A	5.0	2.8	1.3	0.154	0.186	2.92E-4
25A	5.0	2.7	1.3	0.150	0.183	3.75E-4
26A	4.9	3.3	1.3	0.156	0.197	7.85E-4
27R	6.8	4.2	2.2	0.220	0.197	4.57E-4
28A	4.6	4.3	1.3	0.148	0.185	3.93E-4
29A	6.4	4.8	1.4	0.152	0.187	3.71E-4
30A	6.8	5.0	1.4	0.147	0.182	5.96E-4
31A	7.5	3.4	1.2	0.152	0.206	2.49E-4
32A	5.5	3.3	1.1	0.150	0.198	3.51E-4
33A	5.5	3.1	1.1	0.154	0.184	3.17E-4
34A	4.4	3.2	1.3	0.145	0.177	4.55E-4
35A	5.1	3.6	1.4	0.157	0.187	6.81E-4
36R	6.0	3.8	2.3	0.212	0.176	6.43E-4
37R	8.2	3.8	2.2	0.198	0.183	3.58E-4
38A	6.2	3.3	1.1	0.159	0.179	5.91E-4
39A	6.6	3.4	1.0	0.150	0.185	4.96E-4
40R	8.1	3.3	1.7	0.195	0.175	4.93E-4
41R	8.1	3.6	1.6	0.186	0.160	4.85E-4

Figura 54 (cont.). Matriz de datos con las variables más significativas.

V.5.1. Análisis en Componentes Principales (PCA)

Una vez que hemos obtenido la matriz de datos definitiva de la que hemos eliminado las columnas correspondientes a las variables menos significativas, vamos a empezar el tratamiento quimiométrico aplicando la técnica de preprocesado; de esta forma, obtendremos una idea intuitiva de cuales son los descriptores que verdaderamente discriminan entre las muestras. Además, observando la distribución de los casos en el nuevo espacio de factores veremos la mayor o menor homogeneidad de éstas.

Así pues, hemos realizado un análisis en componentes principales sobre nuestra matriz de datos. Nos quedamos con tres componentes principales que conllevan valores de comunalidades en ningún caso inferiores a 0.6, tal y como se muestra junto a los valores de loadings para cada una de las variables iniciales, en la tabla 40.

	PC1	PC2	PC3	COMUNALIDAD
P	.823	.299	.164	.793
MG	-.519	-.613	.489	.884
BA	-.393	.764	.476	.964
POLI	.764	8.40E-4	-.129	.601
CLOROG	.713	-.284	.360	.718
CAF	.923	-.067	.161	.883

Tabla 40. Contribuciones de los PC's a las variables y comunalidades.

Para visualizar mejor los resultados obtenidos al llevar a cabo este preprocesado, realizamos la representación BIPLOT en donde podemos observar la distribución de las muestras en el espacio de los dos primeros componentes principales, al igual que la disposición de los descriptores. Dicha representación se ofrece, a continuación, en la figura 55.

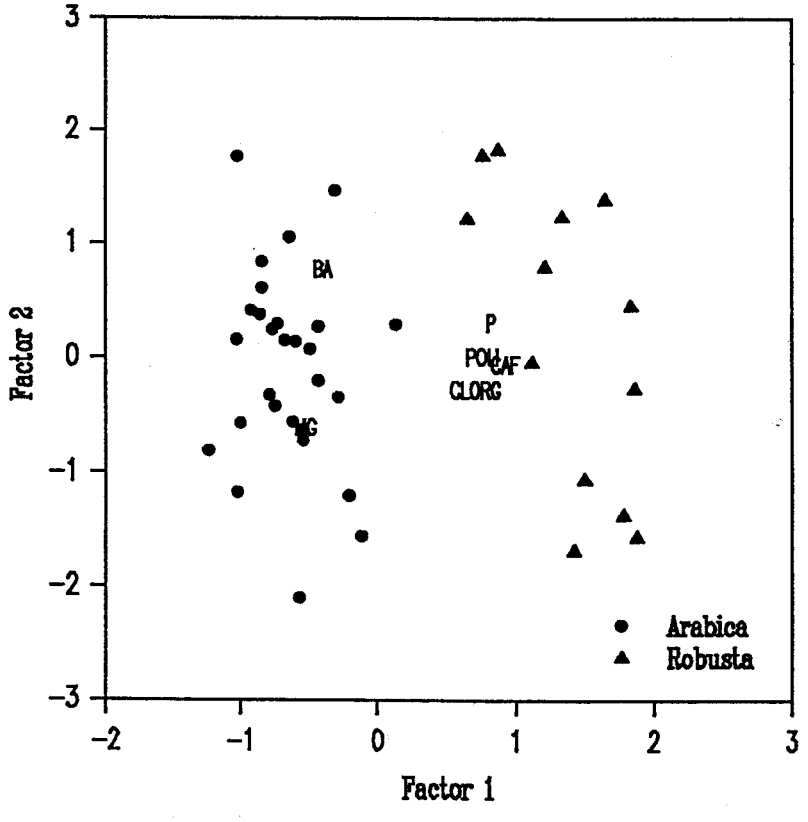

Figura 55. BIPLOT de las muestras de café y los descriptores más significativos.

En la figura, se puede ver cómo las variables CAF y P aparecen muy próximas, en el extremo positivo del factor 1, lo que nos lleva a pensar que van a proporcionar el mismo tipo de información. Por el contrario, situado en el extremo negativo del primer componente principal aparece el descriptor MG. Una primera conclusión que podemos extraer es que las variables que van a definir mejor la separación de muestras según la variedad arábica o robusta son CAF (o bien P) y MG. Precisamente, estos descriptores eran los que aparecían como más significativos cuando realizamos el estudio por separado del perfil metálico y el no metálico.

En cuanto a la distribución de los casos, es obvia la clara separación existente entre las muestras arábicas, situadas en la parte negativa del factor 1 y con una distribución muy homogénea, y las muestras de variedad robusta, las cuales aparecen en la zona positiva de PC1 y presentan una distribución un poco menos homogénea que las anteriores.

Una vez extraídas estas primeras conclusiones, vamos a aplicar los métodos quimiométricos tanto supervisados como no supervisados.

V.5.2. Análisis Cluster

En un primer ensayo, realizamos el análisis cluster jerárquico utilizando como variables solamente MG y CAF, las cuales tiene un mayor poder discriminante según lo estudiado en el PCA. La regla de amalgamación empleada es el método de Ward y la distancia es la euclídea.

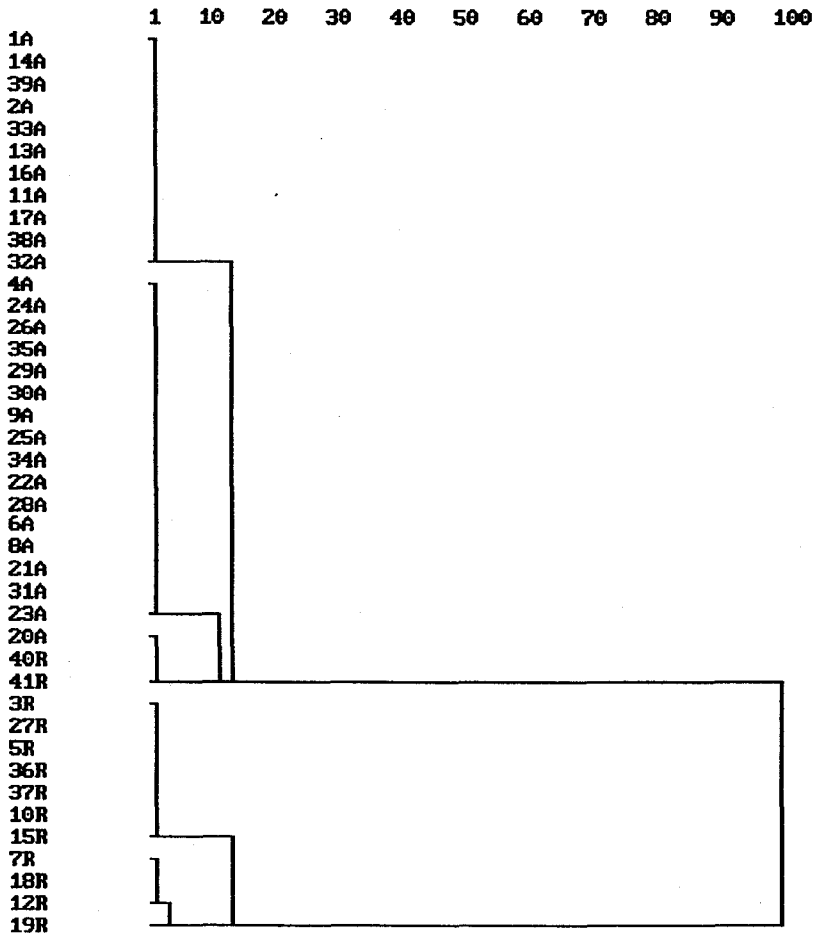

Figura 56. Dendrograma de las muestras de café empleando las variables MG y CAF.

En el dendrograma obtenido, mostrado en la figura 56, se observa que incluso a distancias menores del 20% de la máxima (casi del 10%), aparecen dos clusters conteniendo cada uno de ellos muestras correspondientes a cada una de las

variedades de café. El caso 40R aparece en el racimo de las muestras arábicas siendo robusta, pero está ubicada justo en la frontera entre ambos clusters. Esta figura nos confirma nuestras suposiciones anteriores, es decir, que las variables MG y CAF bastan para diferenciar entre las muestras de café atendiendo a su variedad.

Por otra parte, según el PCA realizado anteriormente y a la vista del BIPLOT podemos deducir que las variables CAF y P proporcionan el mismo tipo de información (poseen *loadings* muy próximos). Por consiguiente, vamos a volver a realizar el análisis cluster, bajo idénticas condiciones, únicamente cambiando el descriptor CAF por P.

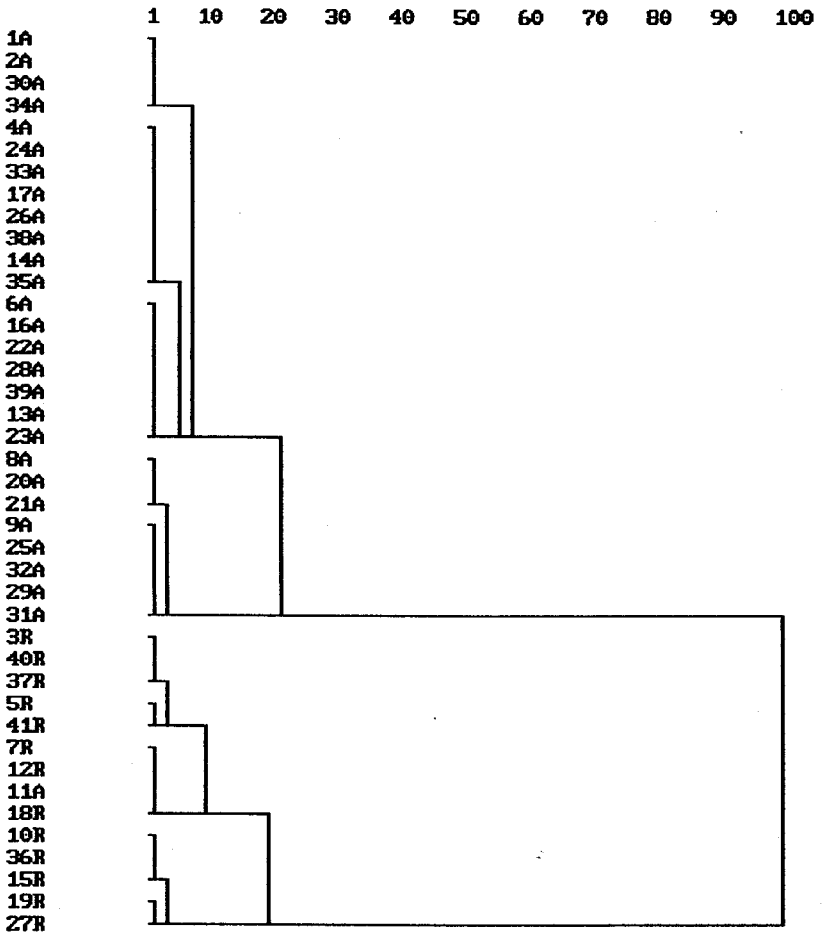

Figura 57. Dendrograma de las muestras utilizando las variables MG y P.

En la figura 57, puede apreciarse que este último dendrograma y el mostrado en la figura 56 son prácticamente idénticos. Por tanto, para discriminar entre los cafés arábica y los robusta basta con realizar un análisis del contenido en magnesio y fósforo.

V.5.3. Análisis Discriminante Lineal (LDA)

Una vez hemos llevado a cabo el preprocesado de los datos y, seguidamente, hemos aplicado un método de reconocimiento de patrones no supervisado, vamos a realizar el análisis de discriminante lineal ya que conocemos *a priori* a qué variedad pertenece cada uno de nuestros casos, es decir, vamos a aplicar una técnica supervisada.

En primer lugar, llevamos a cabo el LDA empleando el backward stepwise el cual parte de todas las variables de la matriz de datos y va eliminando una a una hasta que ya no se discrimina entre las distintas categorías. Así, el programa solamente elimina la variable CLOROG; al estudiar las probabilidades *a posteriori*, observamos que todas las muestras están correctamente clasificadas.

A continuación, realizamos el análisis discriminante utilizando el método forward stepwise, el cual comienza considerando la variable más discriminativa y va añadiendo descriptores parándose cuando no se produce una mejora en la discriminación entre clases. El resultado obtenido es que son necesarias todas las variables excepto el bario para discriminar entre las muestras; en las probabilidades *a posteriori* vuelve a mostrarse un 100% de eficacia en la clasificación.

Seguidamente, vamos a emplear el método stándar para el que sólo vamos a considerar los descriptores fósforo y cafeína. Al estudiar las probabilidades *a posteriori*, se observa una clasificación correcta en todos los casos.

Las funciones de clasificación que se obtienen para cada una de las categorías con estas dos variables se muestran en la tabla 41.

CLASE	FUNCIONES DE CLASIFICACION
ARABICA	-115.4 + 1409.28 * P + 12.90 * CAF
ROBUSTA	-208.41 + 1791.44 * P + 27.34 * CAF

Tabla 41. Funciones discriminantes para las clases arábica y robusta.

Lo mismo ocurre cuando empleamos las variables MG y CAF. En este caso, las funciones discriminantes se muestran en la siguiente tabla.

CLASE	FUNCIONES DE CLASIFICACION
ARABICA	-274.26 + 3062.26 * MG - 20.26 * CAF
ROBUSTA	-231.85 + 2616.51 * MG + 0.81 * CAF

Tabla 42. Funciones discriminantes para las clases arábica y robusta.

Tal y como hicimos en el análisis cluster, empleamos las variables P y MG para crear funciones de clasificación que se muestran en la siguiente tabla.

CLASE	FUNCIONES DE CLASIFICACION
ARABICA	-325.45 + 1109.62 * P + 2574.92 * MG
ROBUSTA	-356.41 + 1554.12 * P + 2315.50 * MG

Tabla 43. Funciones discriminantes para las clases arábica y robusta.

Observando las probabilidades *a posteriori* vemos que se clasifican bien todos los casos exceptuando las muestras 7R y 18R.

Por último, hemos repetido el método stándar pero, esta vez, intentamos emplear un sólo descriptor. Para ello, se ha aplicado de nuevo el análisis discriminante lineal utilizando el método stándar y empleando solamente la variable P; al estudiar las probabilidades *a posteriori*, se observa que quedan bien clasificados todos los casos excepto la muestra 7R que es robusta y se le asigna la variedad arábica. Cuando se realiza el mismo análisis utilizando sólo la variable CAF, las probabilidades *a posteriori* muestran una clasificación errónea solo en dos casos: las muestras 40R y 41R.

Si se emplea solamente el descriptor MG, se obtienen más fallos en la clasificación: 1A, 2A, 3R, 7R, 12R, 18R, 19R, 27R Y 37R.

Una vez aplicados todos los métodos del Análisis Discriminante Lineal, en lugar de dividir en conjunto total de muestras en dos: conjunto de aprendizaje y de ensayo, se realiza un Leave One Out empleando tanto el backward stepwise como el método standard.

Cuando se realiza un Leave One Out utilizando el método backward stepwise, se consigue un 100% de eficacia en la asignación de clases. El programa sólo elimina la variables ácido clorogénico.

Al llevar a cabo un Leave One Out con el método standard y empleando tan sólo las variables fósforo y cafeína, vuelve a conseguirse un 100% de acierto en la

clasificación de las muestras. Idéntico resultado se obtiene al emplear los descriptores P y MG.

Por tanto, tras el amplio estudio quimiométrico realizado, puede concluirse que de todas los parámetros químicos estudiados, bastaría con analizar el contenido en fósforo y magnesio de las muestras de café para diferenciar entre las variedades comerciales arábica y robusta. En lugar de estudiar el contenido en magnesio, también puede analizarse el contenido en cafeína obteniéndose idénticos resultados.

V.6. CORRELACION ENTRE VARIABLES

Una vez tratadas todas las muestras de café desde el punto de vista quimiométrico, estudiar la correlación existente entre las variables analizadas es asunto sencillo puesto que puede deducirse a partir de la matriz de correlación, la cual no es más que la matriz de covarianzas para los datos autoescalados. Sin embargo, las correlaciones así deducidas son lineales y éstas conllevan una distribución normal o gaussiana de las variables, es decir, correlaciones en el sentido de Pearson; no obstante, como en el caso que nos ocupa esta condición no se cumple para la totalidad de los descriptores empleados, es más adecuado aplicar un test no paramétrico como el de Spearman para tal fin.

En primer lugar, hemos estudiado las correlaciones existentes entre los pares de variables de carácter no metálico, considerando correlación significativa para aquellas parejas cuyo nivel de significación p-level es menor de 0.005 tal y como se muestra en la tabla 44.

VARIABLES	R	p-level
POLI & CLOROG	0.464	0.002
POLI & CAF	0.428	0.005
CLOROG & CAF	0.603	0.000

Tabla 44. Correlaciones más significativas para las variables no metálicas.

Como puede apreciarse, existe correlación significativa entre las variables polifenoles, ácido clorogénico y cafeína hecho que ya podía observarse cuando realizamos la representación BIPLOT en el preprocesado de los datos, puesto que estos tres descriptores presentaban loadings prácticamente coincidentes para los dos primeros componentes principales.

Otro hecho que cabe destacar es que la correlación existente es positiva, lo cual concuerda con lo indicado en la parte experimental en la que se afirma que los cafés de variedad robusta presentan valores más altos de estas tres variables que los pertenecientes a la variedad arábica.

A continuación, hemos estudiado las posibles correlaciones existentes entre las variables de carácter metálico, considerando igualmente correlación significativa aquellas cuyo p-level es menor de 0.005 y son las que se muestran en la siguiente tabla.

VARIABLES	R	p-level
ZN & NA	0.467	0.002
P & CU	0.621	0.000
MN & CU	-0.533	0.000
FE & CU	0.512	0.000
CA & SR	0.543	0.000
CU & SR	0.490	0.001

Tabla 45. Correlaciones más significativas entre las variables metálicas.

Se aprecia correlación positiva en todos los casos exceptuando la pareja manganeso y cobre, que presenta correlación negativa. Este hecho está en concordancia con lo indicado en el apartado IV.3.8. de la parte experimental donde se indica que las muestras de café arábica presentan un mayor contenido en manganeso comparado con las robustas y, por el contrario, en lo que respecta al cobre esta relación es en sentido inverso, es decir, los cafés robusta son los que tienen un mayor contenido en este metal.

Si nos fijamos en la relación entre metales mayoritarios y minoritarios, se aprecia que existe siempre una correlación positiva: ZN & Na, P & CU y CA & SR, mientras que para el caso de metales minoritarios entre sí ésta es tanto negativa como positiva. Por otra parte, no se aprecia correlación importante entre metales mayoritarios.

Por último, estudiamos la correlación existente entre las tres variables metálicas más significativas, fósforo, magnesio y bario, y las tres con mayor poder discriminante del bloque de descriptores de carácter no metálico, es decir, polifenoles, ácido clorogénico y cafeína. En la siguiente tabla, se muestran sólo las parejas de variables que presentan correlación significativa (p-level < 0.005).

VARIABLES	R	p-level
P & POLI	0.474	0.001
P & CAF	0.694	0.000

Tabla 46. Correlaciones significativas entre las variables más discriminantes.

Como puede observarse, los descriptores más significativos de cada grupo de variables, los cuales son fósforo y cafeína, son los que presentan una más alta correlación siendo ésta positiva. Este resultado está en perfecto acuerdo con lo expresado en la parte experimental, ya que las muestras de café con mayor contenido en fósforo son las robusta, variedad ésta que análogamente presenta mayores contenidos en cafeína.

De igual modo, se aprecia correlación también positiva entre fósforo y polifenoles, lo cual es lógico ya que al estudiar las correlaciones entre las variables de carácter no metálico vimos la relación existente entre cafeína y polifenoles, luego si la cafeína está correlacionada con el fósforo y con polifenoles, también presentarán correlación y también será positiva fósforo y polifenoles. En todos los casos, es la variedad robusta la que presenta contenidos mayores de estos tres descriptores.

RESUMEN Y CONCLUSIONES

1. Se ha realizado un estudio quimiométrico sobre un conjunto de 41 muestras de café verde para diferenciar las variedades arábica y robusta.
2. Los parámetros químicos utilizados como descriptores han sido: extracto acuoso, polifenoles totales, aminoácidos libres totales, cafeína, ácido clorogénico, trigonelina, zinc, fósforo, manganeso, hierro, magnesio, calcio, sodio, potasio, cobre, estroncio y bario.
3. La determinación de estos parámetros se ha realizado como sigue:

   * El extracto acuoso se ha determinado mediante un método gravimétrico.

   * Los polifenoles totales se han determinado mediante un método espectrofotométrico, usando el reactivo Folin-Ciocalteu.

   * Los aminoácidos libres totales se han determinado mediante un método espectrofotométrico, basado en su reacción con ninhidrina.

   * La cafeína y el ácido clorogénico se han determinado por cromatografía líquida de alta resolución en fase reversa. Se utiliza una columna de octadecilsilano y una fase móvil metanol-agua (20:80) de pH 4.5.

   * La trigonelina se ha determinado mediante un nuevo método de cromatografía iónica. Se emplea una columna

catiónica Waters IC Pack C M/D, como fase móvil una disolución acuosa de ácido clorhídrico $2 \cdot 10^{-3}$ M de pH 3 y detección ultravioleta a 254 nm.

* Los metales se han determinado mediante espectroscopía de emisión atómica de plasma inducido acoplado.

4. Una vez realizado el estudio quimiométrico, ambas variedades de café pueden distinguirse fidedignamente, tanto a partir de su perfil metálico como no metálico.

   Del conjunto total de descriptores, los más discriminativos fueron magnesio, fósforo y cafeína. Basta seleccionar una pareja de descriptores como magnesio-fósforo o cafeína-fósforo para diferenciar perfectamente entre las variedades arábica y robusta; si bien, la determinación de magnesio-fósforo puede realizarse simultáneamente en un sólo experimento mediante espectroscopía de emisión atómica de plasma inducido acoplado.

BIBLIOGRAFIA

1. N. Kolpas, Coffee, J. Murray (Ed.), Londres, 1979.
2. W.H. Ukers, All about coffee, 2ª Ed., Gale Research Co., Detroit, 1976.
3. J.M. Restrepo, Memorias sobre el cultivo del café, Publicaciones del Banco de la República, Archivo de la Economía Nacional, Bogotá, 1952.
4. A.E. Haarer, Modern Coffee Production, L. Hill, Londres, 1962.
5. M. Vanier, Libro del amante del café, J.J. Olañeta, Barcelona, 1983.
6. F.J.E. Van Dierendonk (Ed.), The manuring of Coffee, Cocoa, Tea and Tobacco, Cantre D'etude de L'azote, Génova, 1959.
7. Comunicación personal.
8. B. Rothfos, Coffee Production, Gordian-Marx-Rieck, Hamburgo, 1980.
9. M. Silvetz y N.W. Desrosier, Coffee Technology, AVI Publishing Co., Westport Conn., 1979.
10. M.N. Clifford, Proc. Biochem., 19, 1975, 13-16.
11. O.G. Vitzthum, In Kaffee und Coffein, O. Eichler (Ed.), Springer-Verlang, Berlín, 1975.
12. R.J. Clarke y L.J. Walker, J. Sci. Fd. Agric., 25, 1974, 1389.
13. L.A.B. Ferreira, M.A.C. Fragoso, M.F. Peratta, M.C.C. Sirvo y M.C. Rebeto, Proc. 5º Coll. ASIC, 1973, 51-62.
14. R.J. Clarke y L.J. Walker, Proc. 7º Coll. ASIC, 1975, 159-163.
15. O.B. Tserevitinov y col., Voprosy Pitaniya, 31, 1972, 85 citado en Food Sci. Technol. Abstr., 4, 1972, 6H824.

16. M.A. Peratta y M.F. Silvo, Proc. 5º Coll. ASIC, 1973, 51-62.
17. C. Horowitz y col., South African Medical Journal, 48, 1974, 230, citado en Food Sci. Technol. Abstr., 6, 1974, 8H1334.
18. R.J. Clarke y R. Macrae, Coffee, Vol.3 (Physiology), Elsevier, Londres, 1985.
19. H.G. Maier, Kaffee, P. Parey (Ed.), Berlín, 1981.
20. V. Krivan, P. Barth y A. Feria Morales, Mikrochim. Acta, 110 (4-6), 1993, 217-236.
21. R. Tessl, M. Holzer y H. Kamperschroer, Proc. 10º Coll. ASIC, 1982, 279-292.
22. L.C. Trugo y R. Macrae, Proc. 10º Coll. ASIC, 1982, 187-192.
23. M.L. Wolfrom y col., J. Agric. Fd. Chem., 8. 1960, 58-65.
24. J. Pokorny y col., Nahrung, 18, 1974, 799-805.
25. M.L. Wolfrom y col., J. Org. Chem., 26, 1961, 4533-4535.
26. M.L. Wolfrom y D.L. Patin, J. Agric. Fd. Chem., 12, 1964, 376-377.
27. M.L. Wolfrom y D.L. Patin, J. Org. Chem., 30, 1965, 4060-4063.
28. H. Thaler y W. Arneth, Z. Lebensm. Unters. Forsch., 138, 1968, 137-145.
29. H. Thaler y W. Arneth, Z. Lebensm. Unters. Forsch., 140, 1969, 101-109.
30. H. Thaler, Z. Lebensm. Unters. Forsch., 143, 1970, 342-348.
31. Hartley, J. Chem. Soc., 87, 1905, 1802.
32. H. Bothe y N.K. Cammenga, Proc. 9º Coll. ASIC, 1980, 135-144.
33. A. Sternnert y H.G. Maier, Z. Lebensm. Unters. Forsch., 196, 1993, 430.
34. P. Mazzafera, Phytochemistry, 30, 1991, 2309.
35. C.A.B. De Maria y col., Food Chem., 52, 1995, 447.

36. M.N. Clifford, Proc. Biochem., 1975, 20-29.
37. H. Thaler y R. Gaigl, Z. Lebensm. Unters. Forsch., 119, 1963, 10-25.
38. H.V. Amorim y R.V. Josephson, J. Fd. Sci., 40, 1975, 1179-1184.
39. H.V. Amorim y R.V. Josephson, Proc. 7º Coll. ASIC, 1975, 109-114.
40. W. Walter, H.G. Grigal y J. Heukeshoven, Naturwiss, 57, 1970, 246-247.
41. L.S. Campos y J.M.L. Rodrigues, Proc. 5º Coll. ASIC, 1971, 91-96.
42. A. Pereira y M.M. Pereira, Proc. 5º Coll. ASIC, 1971, 85-90.
43. Robiquet y Boutron, Annalen der Pharmacie, 23, 1837, 93-95.
44. K. Gorter, Annalen, 358, 1907, 327-348.
45. K. Freudenberg, Ber., 53, 1920, 232-239.
46. H.O.L. Fisher y G. Dangschat, Ber., 65, 1932, 1037-1040.
47. IUPAC, Biochem. J., 153, 1976, 23-31.
48. H.M. Barnes, J.R. Feblman y W.V. White, J. Amer. Chem. Soc., 72, 1950, 4178-4182.
49. K. Rubach, disertación, Technische Universität, Berlín, 1969.
50. R.J. Clarke y R. Macrae, Coffee, Vol.1 (Chemistry), Elsevier, Londres, 1985.
51. O. Ohiokpehai, tesis doctoral, Universidad de Surrey, 1982.
52. M.N. Clifford, O. Ohiokpehai y H. de Menezes, Proc. 11º Coll., ASIC, 1985.
53. O. Ohiokpehai, G. Brumen y M.N. Clifford, Proc. 10º Coll., ASIC, 1982, 177-186.
54. H.G. Maier, Kaffee, Paul Parey, Hamburgo, 1981, 21.

55. J. Poisson, Proc. 8º Coll., ASIC, 1977, 33-58.

56. M.N. Clifford, Proc. Biochem., 1975, 3-8.

57. M.N. Clifford y K.C. Willson (Ed.), Coffee: Botany, Biochemistry and Production of Beans and Beverage, Croom-Helm, Londres, 1985.

58. H. Strevli, Proc. 6º Coll. ASIC, 1973, 61-72.

59. L. Hartmann y col., J. Amer. Oil Chem. Soc., 45, 1968, 577.

60. H.P. Kaufman y col., Fette, Seifen, Anstrichmittel, 64, 1962, 206.

61. A. Carisano y col., J. Sci. Fd. Agric., 15, 1964, 619..

62. J. Pokorny y col., Sb vys Sk Chem-technol. Praze, 28, 1970, 73 citado en Food Sci. Technol. Abstr., 6, 1974, 1H147.

63. T. Itoh, T. Matsumoto y T. Tamura, J. Amer. Oil Chem. Soc., 50, 1973, 122-125.

64. T. Itoh, T. Matsumoto y T. Tamura, J. Amer. Oil Chem. Soc., 50, 1973, 300-303.

65. X. Tomas y J.J. Molins, Afinidad XLVII, 1990 Mayo-Junio, 427.

66. C. Merrit, D.H. Robertson y D.J. McAdoo, Proc. 4º Coll. ASIC, 1969, 144-148.

67. J. Poisson, Proc. 8º Coll. ASIC, 1977, 33-57.

68. W. Gutmann, P. Werkhoff, M. Barthels y O.G. Vitzthum, Proc. 8º Coll. ASIC, 1977, 153-161.

69. J. Pokorny, C.Nguyênhuy, E. Smidrkalova y G. Janicek, Z. Lebensm. Unters. Forsch., 158, 1975, 87-92.

70. A.F. Mabrouk y F.E. Deatherage, Fd. Technol., 10, 1956, 194-197.

71. C. Lentner y F.E. Deatherage, Fd. Res., 24, 1959, 483-492.

72. T. Nakabayashi, Jap. Soc. Fd. Sci. Technol., 25, 1978, 142-146.

73. J.M. Northmore, Proc. 4º Coll. ASIC, 1970, 47-54.

74. R.W. Von Borstel, Fd. Technol., 37(9), 1983, 40-43, 46-47.

75. American Council on Science an Health, Health Effects of Caffeine, ACSH, Nueva York, Marzo 1981.

76. H.V. Amorim, E. Malavolta, A.A. Teixeira, V.F. Cruz, M. Meb, M.A. Guercio, E. Fossa, O. Breviglieri, S.E. Ferrari y D.M. Silva, Proc. 6º Coll. ASIC, 1973, 113-127.

77. M.J. Martín, F. Pablos y A.G. González, Discrimination of green coffee varieties by means of pattern recognition techniques, Euroanalysis IX, Bolonia (Italia), 1996.

78. K. Täufel, Handbuch der Lebensmittelchemie, Vol.VI, 32.

79. F. Tateo, Analisi dei Prodotti Alimentari, 2ª Ed., Vol.2, Chiriotti, Pinerolo, 1978.

80. Official Methods of Analysis AOAC, 4ª Ed., Arlington, Virginia, 1984,275.

81. O.K. Sharma y P.S. Krishman, Analyt. Biochem., 14, 1966, 11-16.

82. H. Haan, On the determination of soluble humic substances in freshwaters. In humic substances: their structure and function in the biosphere, D. Polvoledo & H.L. Goltermann (Ed.), Produc. Wageningen, Holanda, 1975, 53-62.

83. D. Polvoledo y M. Gerletti, Mitt. int. Verein. Theor. angew. Limmol,

14, 1968, 145-154.

84. A.A. Berk y W.C. Schroeder, Ind. Engng. Chem., 14, 1942, 456-459.

85. Standard Methods for the Examination of water and wastewater (13ª Ed.), American Public Health Association (APHA), American Water Works Association and Water Pollution Control Federation, Washington D.C., 1971.

86. Standard Methods for the Examination of water and wastewater (13ª Ed.), American Public Health Association (APHA), American Water Works Association and Water Pollution Control Federation, Washington D.C., 1976.

87. V.L. Singleton y P. Esau, Phenolic Substances in Grapes and Wines and their significance, Adv. Food Res., Suppl.1, Academic Press, Nueva York, 1969.

88. S. Moore y W.H. Stein, J. Biol. Chem., 176, 1955, 367.

89. P. Valera, F. Pablos y A.G. González, Talanta, 43, 1996, 415-419.

90. M.J. Martín, F. Pablos y A.G. González, Anal. Chim. Acta, 320, 1996, 191-197.

91. Official Methods of Analysis (8ª Ed.), Washington D.C., 1955, 238.

92. E. Borker, J. Assoc. Offic. Anal. Chem., 43, 1960, 620-622.

93. W. Horwitz, Methods of Analysis AOAC, (13ª Ed.), AOAC, Washington D.C. 1980, 234.

94. F. Chassevent y col., Café, cacao, thé, 18, 1974, 49.

95. J.M. Newton, J. Assoc. Offic. Anal. Chem., 62, 1979, 705.

96. Oi-Wah Lau, Shiv-Fai Luk, Oi-Ming Cheng y Teresa P. Y. Chiv, Analyst, 117, 1992, 777.
97. U.M. Senanayke y R.O.B. Wijeskera, J. Sci. Food Agr., 22, 1971, 262.
98. H. Harlos, Lebensm. Ind., 20, 1973, 412.
99. N.R. Strahl, H. Lewis y R. Fargen, J. Agr. Food Chem., 25, 1977, 233.
100. G. Lehmann y col., Z. Physiol. Chem., 1965 citado en Biol. Abstr., 40, 1967, 40496.
101. L.C. Trugo, C.A.B. De Maria y C.C. Werneck, Food Chem., 42, 1991, 81-87.
102. E.C. Greenhoward y L.E. Spencer, Analyst, 98, Londres, 1973.
103. Y. Maeda, M. Yamamoto, K. Owada y S. Satt, Shizuoka Prefectural Institute of Public Health and Environ. Sci., Shizuoka 420, 1988.
104. A.D. Campiglia, J.J. Laserna, A. Berthod y J.D. Winefordner, Anal. Chim. Acta, 244, 1991, 215-222.
105. Z. Tang, C. Hey y S. Fan, Fenxi Huaxue, 19, 1991, 1402-1404.
106. F. Moia, U. Pellegatta, R. Rosso, G. Vignati y E. Svigo, G. Ital. Chim. Clin., 15, 1990, 411-415.
107. S. Kuhr y U.H. Engelhardt, Lebensm. Unters. Forsh., 192, 1991, 526-529.
108. F.J. Muthladi, S.S. El-Hawary y M.S. Hilnawy, J. Liq. Chromatogr., 13, 1990, 1013-1028.
109. J.N. Miceli y W. Chapman, J. Liq. Chromatogr., 13, 1990, 2239-2251.
110. J.L. Blauch y S.M. Tarka Jr., J. Fd. Sci., Vol.48, 1983, 745-750.
111. S.H. Ashoor, G.J. Sperich, W.C. Monte y J. Welty, J. Assoc. Off.

Anal. Chem., Vol.66 nº3, 1983.

112. Análisis de Alimentos, Métodos Oficiales y recomendados por el Centro de Investigación y Control de la Calidad; Ministerio de Sanidad y Consumo Secretaría General para el Consumo. Dirección General de Control y Análisis de la Calidad, Servicio de Publicaciones, 1985.

113. M.N. Clifford, In Coffee: Botany, Biochemistry and Production of beans and beverage, M.N. Clifford y K.C. Willson (Ed.), Croom Helm, Londres, 1985, 305-374.

114. H. Morshita, H. Iwahashi, N. Osaka y R. Kido, J. Chromatography, 315, 1984, 253-260.

115. J. Corse y R.E. Lundin, J. Org. Chem., 35, 1970, 1904-1909.

116. E. Bombardelli, B. Gabetta y E.M. Martinelli, Fitoterapia, 48, 1977, 143-152.

117. E. Haslam, G.K. Makinson, M. Naumann y J. Cunningham, J. Chem. Soc. 1964, 2137-2146.

118. B. Möller y K. Herrmann, J. Chromatogr., 241, 1982, 371-379.

119. K.R. Hanson y M. Zucker, J. Biol. Chem., 238, 1963, 1105-1115.

120. W.A. Court, J. Chromatogr., 130, 1977, 287-291.

121. D.I. Rees y P.D. Theaker, Proc. 8ª Coll. ASIC, 1977, 79-84.

122. G.H.D. Van der Stegen y J. Van Dujin, Proc. 9º Coll. ASIC, 1980, 107-112.

123. L.C. Trugo y R. Macrae, Analyst, Vol.109, 1984, 263-266.

124. L.C. Trugo y R. Macrae, Food Chemistry, 15, 1984, 219-227.

125. J.R. Ramirez-Martinez, J. Sci. Food Agric., 43, 1988, 135-144.

126. F.E. Nottbohm y F. Mayer, Z. Lebensm. Unters. Forsch, 61, 1931, 429-435.

127. W.A. Perlzweig, E.D. Levy y H.P. Scarett, J. Biol. Chem., 136, 1940, 729-745.

128. R.G. Moores y D.M. Greninger, Anal. Chem., 23, 1951, 327-331.

129. L. Kogan, F.J. DiCarlo y W.E. Mayard, Anal. Chem., 25, 1953, 1118-1120.

130. A. Stennert y H.G. Maier, Z. Lebensm. Unters. Forsch, 196, 1993, 430-434.

131. L.C. Trugo, R. Macrae y J. Dick, J. Sci. Food Agric., 34, 1983, 300-306.

132. J. Van Dujin y G.H.D. Van der Stegen, J. Chromatography, 179, 1979, 199-204.

133. P. Navellier, Proc. 2º Coll. ASIC, 1965, 49-54.

134. U. Kroplien, Green and Roasted Coffee Tests, Gordian, Hamburgo, 1961.

135. J.A. Roffi y A. Corte dos Santos, Proc. 5º Coll. ASIC, 1971, 179-200.

136. K.R. Koch, B. Pougnet y S. De Villiers, Analyst, 114, 1989, 911-913.

137. D.C. Harris, Quantitative Chemical Analysis (3ª Ed.), International Student Edition, Nueva York, 1982.

138. B.R. Kowalski y C.F. Bender, J. Am. Chem. Soc., 94, 1972, 5632-5639.

139. C.M. Cuadras, Métodos de Análisis Multivariante, PPU, Barcelona, 1991.

140. C. Albano, W.J. Dunn III, E. Edlund, E. Johanson, B. Norden, M. Sjöström y S. Wold, Anal. Chim. Acta, 103, 1978, 429.

141. K. Varmuza, Pattern Recognition in Chemistry, Springer-Verlag, Heidelberg, 1980, 25-29.

142. M. Meloun, J. Militky y M. Forina, Chemometrics for analytical chemistry, Vol.I:PC-aided statistical data analysis, Ellis Horwood Limited, Chimester, 1992, cap.5.

143. B.R. Kowalski y C.F. Bender, J. Am. Chem. Soc., 94, 1972, 5632.

144. R.A. Fisher, Ann. Eugenics, 7, 1936, 179.

145. D. Coomans, I. Broeckaert, M. Jonckheer, D.L. Massart y P. Blockx, Anal. Chim. Acta, 103, 1978, 409.

146. S. Wilks, Multidimensional Scatter in Olkin (Ed.), Contributions to Probability and Statistics, Stanford Press, 1960, 597-614.

147. G.P. McCabe, Technometrics, 17, 1975, 103.

148. C. Chatfield y A.J. Collins, Introduction to Multivariate Analysis, Chapman and Hall, Londres, 1980.

149. P.E.T. Auf der Heyde, J. Chem. Educ., 67, 1990, 461.

150. H.F. Kaiser, Educ. Psych. Means, 20, 1966, 141.

151. J.R. Piggott (Ed.), Statistical Procedures in Food Research, Elsevier Applied Science, Londres, 1986.

152. E.R. Malinowski, Anal. Chem. 49, 1977, 612.

153. E.R. Malinowski, J. Chemom., 3, 1988, 49.

154. S. Wold, Technometrics, 20, 1978, 397.

155. K.R. Gabriel, Biometrika, 58, 1971, 453.

156. H.C. Romesburg, Cluster Analysis for Researchers, Lifetime Learning Publications, Belmont, California, 1984.

157. D.L. Massart y L. Kaufmann, The interpretation of analytical chemical

data by the use of Cluster Analysis, J. Wiley, Londres, 1983.

158. G.H. Ball, Fall Joint Computer Conference, AFIPS Conf. Proc., 27, 1, Spartan Books, Washington, 553-559.
159. G.N. Lance y W.T. Williams, Comp. J., 10, 1976, 271-277.
160. W.T. Williams y M.B. Dale, Advenced Bot. Res., 2, 1969, 35-68.
161. R.M. Cormack, J. of the Royal Statistical Society, 1971, 321-367.
162. J.N. Kennedy, AIEE Transactions, 6, 3, 1974, 216-227.
163. G.H. Lance y W.T. Williams, Comp. J., 9, 1966, 373.
164. K.J. Florek, Colloquium, Math., 2, 1951, 282.
165. P.H.A. Sneath y R.R. Sokal, Numerical Taxonomy, Freeman, San Francisco, 1973, 222.
166. R.R. Sokal y C.D. Michener, Kansas University Science Bulletin, 38, 1958, 1409.
167. P.H.A. Sneath y R.R. Sokal, Numerical Taxonomy, Freeman, San Francisco, 1973, 228.
168. P.H.A. Sneath y R.R. Sokal, Numerical Taxonomy, Freeman, San Francisco, 1973, 235.
169. J.H. Ward, J. Am. Statist. Ass., 58, 1963, 236.
170. R. Henrion y G. Henrion, Multivariate Datenanalyse, Springer, Heidelberg, 1995, 56-57.
171. E.H. Ruspini, Inf. Sci., 2, 1970, 319.
172. L.A. Zadeh, Inf. and Conf., 8, 1965, 338.
173. K. Varmuza, Anal. Chim. Acta, 122, 1980, 227.

174. L. Kryger, Talanta, 28, 1981, 871-887.

175. A.J. Stuper y P.C. Jurs, J. Pharm. Sci., 67, 1978, 745.

176. L.F. Escudero, Reconocimiento de Patrones, Paraninfo S.A., madrid, 1977.

177. R.A. Fisher, Ann. Eugenics, 8, 1938, 376-378.

178. M.M. Tatsouka y D.V. Tiedman, Rev. Educ. Res., 24, 1954, 402-420.

179. J.J. Powers y E.S. Keith, J. Food Sci., 33, 1968, 207.

180. D. Coomans, D,L. Massart y L. Kaufman, Anal. Chim. Acta, 112, 1979, 97-122.

181. S. Wold, Pattern Recognition, 8, 1976, 127.

182. O. Strouf y S. Wold, Acta Chem. Scand, A31, 1977, 391.

183. S. Wold y O. Strouf, ibid., A33, 1979, 521.

184. S. Wold y O. Strouf, ibid., A33, 1979, 463.

185. S. Wold y M. Sjöström, SIMCA: A method for analyzing chemical data in terms of similarity and analogy, in reference 1, 243.

186. D.L. Duewer, B.R. Kowalski y T.F. Schatzki, ibid, 47, 1975, 1573.

187. B.E.H. Saxberg, D.L. Duewer, J.L. Booker y B.R. Kowalski, Anal. Chim. Acta, 103, 1978, 201.

188. M. Sjöström y U. Edlund, J. Magn. Resonance, 25, 1977, 285.

189. S. Wold y K. Anderson, J. Chromatrog., 80, 1973, 43.

190. S. Wold y M. Sjöström, Linear Free Energy Relationships as Tools for Investigating Chemical Similarity-Theory and Practice, in Correlation Analysis in Chemistry, N.B. Chapman y J. Shorter (Ed.), Plenum

Press, Nueva York, 1978.

191. X.L.P. Van Espen, F. Adams, S.H. Yan y M. Vanbelle, Anal. Chim. Acta, 200, 1987, 421-430.

192. E. Fix y J. Jodges, USAF School of Aviation Medicine Report nº4, 1951, Randolph Field, Texas.

193. R.D. Duda y P.E. Hart: Pattern Classification and Scinece Analysis, J. Wiley, cap.4, 1973.

194. P.C. Jurs, B.R. Kowalski y T.L. Isenhour, Anal. Chem., 41, 1969, 21.

195. C.E. Klopfenstein y C.L. Wilkins (Ed.), Computers in Chemical and Biochemical Research, Vol.2, Academic Press, Nueva York, 1974.

196. P.R. Griffiths (Ed.), Transform Techniques in Chemistry, Plenum Press, Nueva York, 1978.

197. N.J. Nilsson, Learning Machines, McGraw-Hill, Nueva York, 1965.

198. G.L. Ritter y H.B. Woodruff, ibid, 49, 1977, 2116.

199. C.P. Weisel y J.L. Fasching, ibid, 49, 1977, 2114.

200. N.A.B. Gray, ibid, 48, 1976, 2265.

201. J. Zupan y J. Gasteiger, Neural Networks for Chemists, VCH Publishers, 1993.

202. J.R. Hilera y V.J. Martinez, Redes Neuronales Artificiales. Fundamentos, modelos y aplicaciones, Serie Paradigma. Ed. ra-ma, Madrid, 1995.

203. I. Olmeda y S. Barba-Romero (Ed.), Redes Neuronales Artificiales. Fundamentos y aplicaciones. Servicio de publicaciones de la

Universidad de Alcalá de Henares, Alcalá de Henares, 1993.

204. T. Kohonen, Biological Cybernetics, 43, 1982, 59-69.

205. D.H. Hubel y T.N. Wiesel, J. Physiol., 166, 1962, 106-154.

206. D. Rumelhart, G. Hinton y R. Williams, Nature, 323, 1986, 533-536.

207. P. Werbos, Beyond Regression: New tools for prediction and analysis in the behavioral sciences, Ph. Tesis, Harvard University, 1974.

208. D. Parker, Learning logic, Invention Report, 581-664, File1, Office of Technology Licensing, Stanford University, 1982.

209. J. Zupan y J. Gasteiger, Anal. Chim. Acta, 248, 1991, 1-30.

210. International Organization of Standarization, ISO 11294, 1994.

211. Reglamentación Tecnico-Sanitaria para la elaboración, almacenamiento, transporte y comercialización del café. Real Decreto 664/1983, B.O.E., 30, 3, 1983.

212. J.C. Miller y J.N. Miller, Statistics for Analytical Chemistry (2ªEd.), Ellis Horwood Limited, Chichester, 1988.

213. H.G. Maier y A. Grimsehl, Kaffee und Tee Markt, 32, 1982, 3-5.

214. H.G. Maier, Kaffee, P. Parey, Berlín, 1981, 28.

215. W.A. Köning y R. Sturm, Proc. 10º Coll. ASIC, 1982, 271-278.

216. A. Charrier y J. Berthaud, Café, cacao, thé, 1975, 19, 251-264.

217. J. Wurziger, Kaffee und Tee Markt, 1977, 27, 3-8.

218. G. Schomburg, P. Kolla y M.W. Läubli, Int. Lab. April, 40, 1989.

219. R.M. Brunet, Waters Chromatography, 1996. Comunicación personal.

220. L. Cuadros, A.M. García, M. García Campaña, F. Alés, C. Jiménez y

M. Román Ceba, J. AOAC Int., 78, 1995, 471.

221. R. Wilbaux, Le Traitment du Café, FAO, Roma, 1961.

MARIA JESUS MARTIN VALERO

DISCRIMINACION DE LAS VARIEDADES DE CAFE VERDE MEDIANTE TECNICAS DE ANALISIS MULTIVARIANTE

Apto "cum laude"

por unanimidad

24 ABRIL 97

Mª Jesús Martín