



UNIVERSITAT
ROVIRA I VIRGILI



Facultat de
Química
Tarragona

**IDENTIFICACIÓ DE L'ORIGEN GEOGRÀFIC
D'OLI D'OLIVA MITJANÇANT TÈCNiques DE
CLASSIFICACIÓ MULTIVARIANTS I
ESPECTROSCÒPIA FT-RAMAN**

TREBALL DE FI DE GRAU

GRAU DE QUÍMICA

TARRAGONA

JUNY 2018

Alumne: Oriol Salvadó Ruiz

Directora: Itziar Ruisánchez Capelástegui

AGRAÏMENTS

En primer lloc, agrair a la directora del treball, la Dra. Itziar Ruisánchez, per haver-me deixat realitzar el treball en el seu departament i per la seva dedicació i ajuda en tot moment. Al Santiago Macho, que em va ajudar al començament a familiaritzar-me amb el programari Matlab.

També agrair a la Universidad de Granada, concretament al grup de recerca de Análisis en Alimentos y Medio Ambiente (AnAMA) del Departamento de Química Analítica i al Panell de Tast Oficial d'Olis Verges d'Oliva de Catalunya del Departament d'Agricultura, Ramaderia, Pesca i Alimentació de Reus per haver facilitat les mostres per a la realització del treball.

Donar gràcies a l'Institut Català d'Investigació Química (ICIQ) per permetre fer ús de les instal·lacions i l'equip per a realitzar les mesures. Concretament al Fernando Bozoglian, la Mariona Urtasun i al grup SKU en general.

Finalment vull mencionar en aquest apartat als meus amics, família, parella i a tots aquells que durant la realització del treball s'han preocupat i han donat suport moral.

Moltes gràcies.

ÍNDEx

1.	RESUM	5
2.	OBJECTIUS	7
3.	INTRODUCCIÓ.....	8
4.	DESCRIPCIÓ DE LES MOSTRES	10
5.	FONAMENT TEÒRIC	11
5.1.	Tècnica instrumental.....	11
5.2.	Pretractament de dades	13
5.2.1	Centrat (Mean centring).....	13
5.2.2.	Autoescalat.....	14
5.2.3.	Suavitzat (Smoothing) de Savitzky-Golay	14
5.3.	Tècniques d'anàlisi multivariant.....	15
5.3.1.	Representació de dades.....	15
5.3.1.1.	Anàlisi de Components Principals (PCA)	16
5.3.2.	Tècniques de classificació	17
5.3.2.1.	Modelat suau independent d'analogies de classes (SIMCA)	17
5.3.2.2.	Anàlisi discriminant per mínims quadrats parcials (PLS-DA)	19
5.4.	Tècniques de reducció del nombre de variables	21
5.4.1.	Reducció de variables.....	21
5.4.2.	Selecció de variables (VIP scores)	21
5.5.	Paràmetres de qualitat.....	22
6.	PART EXPERIMENTAL	24
7.	RESULTATS I DISCUSSIÓ	25
8.	CONCLUSIONS	38
9.	BIBLIOGRAFIA	40
	ANNEXOS	41

1. RESUM

Existeix un interès en els consumidors per a conèixer la procedència dels productes d'alimentació i els beneficis per a la salut que en alguns casos es poden associar a l'origen geogràfic. El principal objectiu del treball és aconseguir diferenciar i classificar mostres d'oli d'oliva verge extra de la varietat arbequina de dos denominacions d'origen catalanes (Garrigues i Siurana) i un tercer grup de mostres del Sud d'Espanya (Andalusia i Múrcia).

Dins el marc de Treball de Fi de Grau del Grau de Química de la Universitat Rovira i Virgili (URV), les mostres han sigut analitzades a l'Institut Català d'Investigació Química (ICIQ) mitjançant l'equip de l'espectroscòpia Raman amb transformada de Fourier (FT-Raman). Les tècniques de classificació usades han sigut *Soft Independent Modelling of Class Analogy* (SIMCA) i *Partial Least Square-Discriminant Analysis* (PLS-DA).

S'han establert un model SIMCA i un PLS-DA i s'han validat amb 3 paràmetres de qualitat, la sensibilitat, l'especificitat i el percentatge d'inconclusives. Els millors resultats s'han obtingut fent una selecció de variables i els valors obtinguts han sigut >95% (sensibilitat), >85% (especificitat) i <20% (inconclusives) tant per SIMCA com per PLS-DA.

There is an interest for the consumers to know the origin of the food products and the health benefits that in some cases can be associated to the geographic origin. The main objective in this work is to differentiate and classify samples of extra virgin olive oil from the Arbequina variety of two Catalan certificates of origin (Garrigues and Siurana) and a third group of samples from the south of Spain (Andalucía and Murcia), called Sud Espanya. This work has been done during the period of the Final Degree Project of the Chemistry Degree of the Rovira i Virgili University (URV) and the samples have been analysed in the Catalan Institute of Chemistry Investigation (ICIQ) using Raman spectroscopy with Fourier transform as experimental technique. The classification methods used have been Soft Independent Modelling of Class Analogy (SIMCA) and Partial Least Square-Discriminant Analysis (PLS-DA).

A SIMCA model and a PLS-DA model have been established and validated with 3 quality parameters, the sensibility, the specificity and the inconclusive percentage. The best results have been obtained doing a variable selection and the obtained values have been >95% (sensibility), >85% (specificity) and <20% (inconclusive) for the both SIMCA and PLS-DA models.

2. OBJECTIUS

L'objectiu d'aquest treball és diferenciar i classificar mostres d'oli d'oliva verge extra d'arbequina segons el seu origen geogràfic. Les mostres han sigut analitzades per FT-Raman i les dades tractades amb el programari Matlab. Es poden desglossar els següents sub-objectius:

- Aprendre una tècnica nova no explicada ni utilitzada durant el grau de Química com ho és FT-Raman.
- Desenvolupament i comparació de models de classificació capaços de diferenciar les mostres de cada una de les 3 classes.
- Aprendre a utilitzar i familiaritzar-se amb Matlab. Donat que el treball s'ha realitzat amb el programari Matlab i aquest no s'utilitza amb freqüència durant el grau.

3. INTRODUCCIÓ

L'oli d'oliva verge extra és un oli comestible amb un paper molt important en la dieta Mediterrània. Al capdavant dels majors productors hi trobem Espanya, on hi destaquen Andalusia i Catalunya. Aquest producte és molt apreciat i valuós i és per això que hi ha una forta lluita contra les activitats fraudulentas que afecten tan a consumidors com a productors¹.

Actualment a la Unió Europea (UE) existeixen segells de qualitat per als aliments com ho són: Denominació d'Origen Protegit (PDO per les sigles en anglès), Indicació Geogràfica Protegida (PGI per les sigles en anglès) i Especialitat Tradicional Garantida (TSG per les sigles en anglès). Els dos primers tenen una vinculació amb la regió geogràfica on es produeixen els aliments i l'últim amb el procés per el qual s'elabora. PDO fa referència a aquells productes que son produïts, processats i preparats en una zona geogràfica específica. PGI és semblant però només cal que una de les etapes anteriors estigui feta en la regió concreta, les altres es poden dur a terme en altres zones inclús la matèria primera pot provenir d'altres llocs. Per últim STG s'usa per a aquells productes alimentaris que utilitzen matèries primeres tradicionals, mètodes de producció tradicionals o que tenen una composició tradicional, sense cap restricció geogràfica². La Figura 1 mostra els segells PDO, PGI i STG³.



Figura 1. D'esquerra a dreta: PDO, PGI i STG³.

A través d'aquests segells els consumidors poden reconèixer fàcilment la qualitat dels productes en termes d'origen regional o de producció tradicional. Aquests segells també s'utilitzen però com a eina de màrqueting en la UE i altres mercats i es per això que hi ha una protecció legal i requisits per poder usar aquestes distincions i així evitar imitacions o falsificacions⁴.

Els productes amb PDO tenen una reputació i característiques de qualitat que són resultat essencialment de la regió on s'han produït. Catalunya té 11 PDOs de les quals 5 són

referenciades a olis d'oliva verge extra i aquestes són Siurana, Les Garrigues, Oli de la Terra Alta, Oli del Baix Ebre-Montsià i Oli de l'Empordà⁵.

Si es parla d'Andalusia, es poden trobar 25 PDOs i 12 són d'olis d'oliva verge extra. Ho són: Aceite de Lucena, Antequera, Baena, Estepa, Montes de Granada, Montoro-Adamuz, Poniente de Granada, Priego de Córdoba, Sierra de Cádiz, Sierra de Cazorla, Sierra de Segura y Sierra de Mágina⁶.

És important per al productor aconseguir una distinció de qualitat ja que suposa un valor afegit al producte que es pretén vendre i també per al consumidor que sap que els producte amb denominació d'origen tenen un criteri de qualitat. Els olis amb denominació d'origen diferents poden tenir diferent color, gust o aroma depenent de la regió en la que es produeixen⁵. Així doncs, si tenen característiques diferents, es poden diferenciar mitjançant una mesura química?

En aquest estudi es provarà si és possible determinar la denominació d'origen d'olis d'oliva verge extra de la varietat arbequina usant l'espectroscòpia Raman amb transformada de Fourier (FT-Raman d'ara en endavant) com a tècnica instrumental i la quimiometria com a eina de tractament de dades.

4. DESCRIPCIÓ DE LES MOSTRES

S'ha treballat amb un total de 65 mostres d'oli d'oliva verge extra d'oliva arbequina de diferents orígens. Donada la baixa quantitat de mostres que s'han obtingut de denominacions d'origen andaluses i murcianes, finalment la classificació es durà a terme agrupant les mostres provinents del sud d'Espanya i donant-li aquest nom al grup. La distribució geogràfica de les mostres es troba a la taula 1.

Taula 1. Distribució geogràfica i agrupació de les mostres.

Grup	PDO/Província	Mostres	Total
Siurana	Siurana	16	16
Garrigues	Les Garrigues	27	27
Sud Espanya	Córdoba	6	22
	Granada	5	
	Sevilla	2	
	Jaén	5	
	Huelva	2	
	Murcia	2	

Les mostres de Siurana i Garrigues han sigut facilitades pel Panell de Tast Oficial d'Olis Verges d'Oliva de Catalunya del Departament d'Agricultura, Ramaderia, Pesca i Alimentació de Reus. Pel que fa a les mostres agrupades com a Sud Espanya les ha facilitat el grup de recerca de Análisis en Alimentos y Medio Ambiente (AnAMA) del Departamento de Química Analítica de la Universidad de Granada. El llistat de les mostres i la seva nomenclatura es troba en l'Annex 1.

5. FONAMENT TEÒRIC

5.1. Tècnica instrumental

L'espectroscòpia de Raman és una tècnica que permet observar els modes de vibració, rotació i altres modes de baixa freqüència en un sistema.

Es basa en la mesura de la dispersió inelàstica o de Raman d'un feix de llum monocromàtic provinent d'un làser. Aquest feix pertany a l'espectre visible o infraroig proper. En la Figura 2 es mostra un esquema de la dispersió inelàstica. Una molècula en el nivell vibracional fonamental ($v = 0$) pot absorbir un fotó d'energia i tornar-lo a emetre amb menor energia (fins a $v = 1$). Quan la freqüència d'emissió és més baixa que la d'excitació es parla de dispersió d'Stokes. Si per al contrari, una molècula en el primer estat vibracional excitat ($v = 1$) absorbeix un foto i l'emet amb major energia (fins a $v = 0$) es denomina dispersió anti-stokes. Per últim també es pot donar una dispersió elàstica, és a dir que el fotó s'absorbeixi i emeti amb la mateixa energia, llavors s'anomena dispersió de Rayleigh. Aquesta predomina sobre les altres dues. Per regla general, el nivell vibracional fonamental és el més altament poblat i és per aquest motiu que la dispersió d'Stokes és més intensa que la anti-stokes⁷.

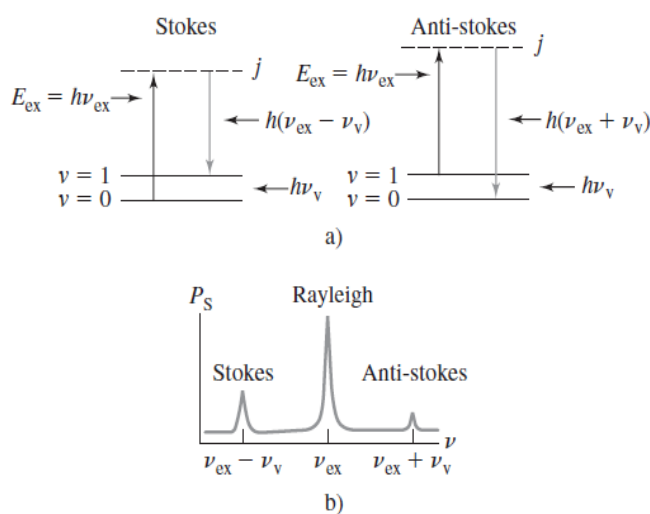


Figura 2. Esquema de la dispersió de Raman. En la a) es mostren les dispersions d'Stokes i anti-stokes. En b) s'observa la relació d'intensitats entre les dispersions.

Així doncs, a l'irradiar una mostra amb un feix de llum monocromàtic aquesta presenta dispersions energètiques elàstiques i inelàstiques i són aquestes últimes les que es separen i es registren.

Concretament en aquest estudi s'ha utilitzat FT-Raman. Tant un espectròmetre de Raman de dispersió com un FT-Raman consten de quatre elements principals:

1. Un làser, font d'un feix de llum monocromàtica (usualment de Nd-YAG, que emet a 1064 nm i no presenta problemes de fluorescència)
2. Lents òptiques per a conduir el feix a la mostra, recollir la dispersió i enfocar-la
3. Monocromador (Raman de dispersió) o interferòmetre (FT-Raman)
4. Detector

Es troben dues diferències principals entre els dos equips. En primer lloc, l'interferòmetre no conté cap reixeta ni element de dispersió, per tant al detector li arriba una senyal més intensa. La segona és que en els instruments dispersius que utilitzen un únic detector, cada element de l'espectre es mesura seqüencialment, mentre que en l'interferòmetre cada un es fa de forma simultània. D'aquesta manera en el temps (T) per a obtenir un espectre en un instrument dispersiu, l'interferòmetre ha registrat cada element de l'espectre per un temps T. Això provoca un augment significatiu de la relació senyal/soroll. Per a aquells instruments amb múltiples detectors com ara díodes en sèrie o de càrrega acoblada la millora en la relació senyal/soroll no serà tan bona, però usualment aquest són ineficients per a treballar a 1064 nm⁸. En la Figura 3 es mostra esquemàticament els components d'un FT-Raman. El feix de llum del làser traspasa la mostra i passa a l'interferòmetre, que consta d'un divisor de feix i dos espills, un fix i un mòbil. El senyal es filtra per a eliminar la radiació làser restant i la dispersió de Rayleigh. Finalment s'enfoca la dispersió al detector⁷.

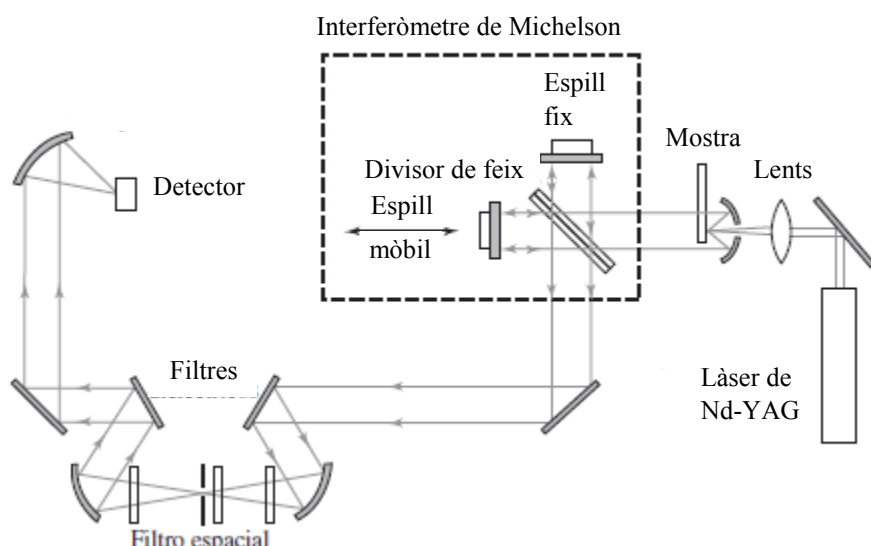


Figura 3. Esquema d'un instrument de FT-Raman.

5.2. Pretractament de dades

De forma general, el pretractament és la modificació de les dades realitzada abans d'aplicar alguna tècnica d'anàlisi les dades. Té com a finalitat eliminar fonts estranyes de variació que no són d'interès, ja sigui fonts de variació aleatòries o sistemàtiques que no estan associades amb l'anàlisi o la propietat d'interès. Si no s'utilitzen aquestes eines, el resultat de l'anàlisi de les dades pot ser erroni o requerir un treball més complex per a obtenir la informació d'interès⁹. A continuació es comenten els pretractaments usats en aquest estudi: centrat, autoescalat i suavitzat.

5.2.1 Centrat (*Mean centring*)

Es tracta de calcular la mitjana de cada columna de la matriu de dades i restar aquest a cada valor de la columna. D'aquesta manera es traslladen els eixos del sistema de coordenades cap al centroide de dades fent que cada mostra només exhibeixi les diferències que té respecte la mostra mitjana de les dades originals¹⁰. L'equació 1 mostra l'expressió matemàtica que es duu a terme, on x_i^{centrada} és el valor de la variable per a una mostra un cop aplicat el centrat, x_i^{original} és el valor original de la variable per a una mostra i $\bar{x}_i$ és la mitjana dels valors de la variable

$$x_i^{\text{centrada}} = x_i^{\text{original}} - \bar{x}_i \quad \text{Equació 1}$$

Aquest pretractament permet mantenir la informació original sense modificar la variància i igualar la mitjana de cada variable a zero. La Figura 4 mostra de forma esquemàtica el comportament d'una sèrie de dades un cop aplicat el centrat.

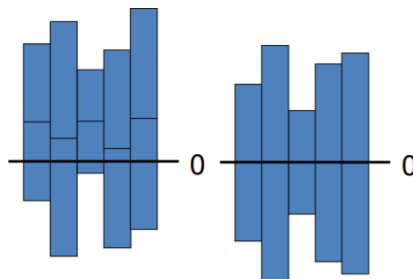


Figura 4. Transformació de les dades originals (esquerra) quan se li aplica el centrat (dreta).

5.2.2. Autoescalat

Es basa en restar la mitjana de cada variable als valors d'aquella variable i posteriorment dividir aquest valor entre la desviació estàndard. L'equació 2 mostra l'expressió matemàtica que es du a terme, on els termes fan referència al mateix que en l'apartat 3.2.2. i la S_i és la desviació estàndard de la variable en qüestió.

$$x_i^{\text{autoescalada}} = \frac{x_i^{\text{original}} - \bar{x}_i}{S_i} \quad \text{Equació 2}$$

D'aquesta manera s'aconsegueix que cada variable tingui la mateixa influència en el càlcul, és a dir, que no es vegi afectat per la mida de les variables. La Figura 5 mostra de forma esquemàtica el comportament d'una sèrie de dades un cop aplicat l'autoescalat.

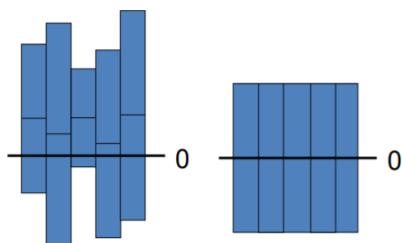


Figura 5. Transformació de les dades originals (esquerra) quan se li aplica un autoescalat (dreta).

5.2.3. Suavitzat (Smoothing) de Savitzky-Golay

Aquest pretractament substitueix cada punt per un nou valor tenint en compte les variables properes. Es pot dur a terme usant la mitjana mòbil dels n valors propers o amb un ajust polinòmic fent servir els n valors propers. En aquest cas s'utilitza la mitjana mòbil. Es basa en utilitzar els valors més propers a la variable i fer una mitjana d'aquests valors. S'han de triar quants valors propers es volen agafar per a fer el suavitzat i a aquest paràmetre se l'anomena amplada de finestra. L'equació 3 mostra l'expressió matemàtica usada en el suavitzat mencionat. En aquest cas x_i és el valor a substituir, n el nombre valors propers escollits i x_{i-n} cada valor proper.

$$x_i = \frac{\sum (x_{i-n} + \dots x_{i-1} + x_i + x_{i+1} + \dots + x_{i+n})}{2n + 1} \quad \text{Equació 3}$$

Té com a finalitat rebaixar el nivell de soroll usualment en matrius de dades d'espectres, però cal anar amb compte ja que també suavitza les senyals de l'espectre desitjades. La Figura 6 mostra el comportament d'una sèrie de dades quan se li aplica un suavitzat de Savitzky-Golay.

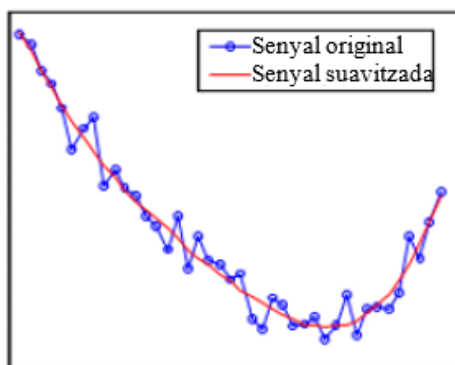


Figura 6. Transformació de les dades originals (blau) quan se li aplica un suavitzat (vermell).

5.3. Tècniques d'anàlisi multivariant

Segons la Societat de Quimiometria Internacional (ICS per les sigles en anglès) la quimiometria és la disciplina química que utilitza la matemàtica, estadística i la lògica formal per a dissenyar o seleccionar els procediments experimentals òptims, proveir el màxim d'informació química útil mitjançant l'anàlisi de les dades i obtenir coneixement sobre el sistema químic¹¹. Hi ha un nombre elevat de tècniques quimiomètriques que es poden utilitzar per a dur a terme els objectius descrits per la ICS, en aquesta secció es descriuran aquelles que han sigut usades en aquest estudi.

5.3.1. Representació de dades

Les tècniques de representació de dades s'utilitzen per a estudiar el comportament de les dades (les mostres). Amb aquestes eines es poden establir relacions entre mostres i possibles agrupacions de les mateixes així com comportaments molt diferenciats de la resta (possibles *outliers*)⁹. A continuació es comentarà la tècnica de representació de dades usada en aquest estudi, l'Anàlisi de Components Principals (PCA per les sigles en anglès).

5.3.1.1. Anàlisis de Components Principals (PCA)

El PCA és una tècnica exploratòria de dades en el que les dades originals es transformen i s'expressen en funció d'unes noves variables anomenades components principals (PCs d'ara en endavant) que són combinacions lineals de les variables originals¹¹. Aquests PCs es calculen de forma que continguin la major variància de la matriu de dades original, però no tots els PCs contenen la mateixa informació, els primers que es calculen descriuen la major variació de les dades i per tant contenen la informació més rellevant mentre que els últims descriuen variacions molt petites que poden estar associades a soroll o error experimental. Es poden resumir aquestes noves variables de forma que:

- **PC₁**: conté la màxima variació possible de les dades → reté la màxima variància
- **PC₂**: conté la màxima variació de les dades no explicada en el PC₁ → menor variància que en el PC₁. Ha de ser ortogonal a PC₁
- **PC₃**: conté la màxima variació de les dades no explicada en el PC₁ i PC₂ → menor variància que en el PC₂. Ha de ser ortogonal a PC₁ i PC₂.
- **PC_a**: conté la màxima variació de les dades no explicada en el PC₁, PC₂, PC₃, ... i PC_{a-1} → menor variància que en el PC_{a-1}. Ha de ser ortogonal a PC₁, PC₂, PC₃, ... i PC_{a-1}.

Hi ha dos tipus de gràfics usats per a representar els resultats del PCA que són el gràfic d'scores i el de loadings. El gràfic d'scores usualment té el PC₁ i el PC₂ com a eixos de coordenades i s'hi representen els valors de les mostres en cada un d'aquests PCs, conté informació sobre les mostres. El gràfic de loadings és la representació gràfica de les variables originals i la seva contribució en cada PC, conté informació sobre les variables⁹.

L'equació 4 mostra l'expressió matemàtica que desglossa la matriu de dades originals en altres matrius de mida inferior.

$$X = TP^T + E \quad \text{Equació 4}$$

En aquest cas X és la matriu de dades originals; T és la matriu de *scores* (degut a la falta d'una traducció literal s'usarà aquest terme durant tot l'estudi), que són els valors de cada mostra en les noves variables; P^T és la matriu de *loadings* (mateix comentari que als *scores*), que conté el pes de cada variable original en les noves variables, transposada; i finalment E és la matriu residual, que conté la variació de les dades no continguda en els PCs escollits. La Figura 7 mostra visualment la base de l'equació (4) on P' substitueix

l'anterior P^T , n és el nombre de mostres, m el nombre de variables originals i a el nombre de PCs calculats.

$$\begin{array}{|c|} \hline \mathbf{X} \\ \hline (n \times m) \\ \hline \end{array} = \begin{array}{|c|} \hline \mathbf{T} \\ \hline (n \times a) \\ \hline \end{array} \begin{array}{|c|} \hline \mathbf{P}' \\ \hline (a \times m) \\ \hline \end{array} + \begin{array}{|c|} \hline \mathbf{E} \\ \hline (n \times m) \\ \hline \end{array}$$

Figura 7. Esquema del fonament matemàtic del PCA.

5.3.2. Tècniques de classificació

Aquestes eines tracten de construir models capaços d'identificar la pertinença (o no) d'una mostra a una classe tenint en compte les característiques d'aquestes. Per a poder fer el model es necessita disposar d'un grup de mostres de classe coneguda. A continuació es descriuen de forma breus les tècniques de classificació usades en aquest estudi: el *Soft Independent Modeling of Class Analogy (SIMCA)* i el *Partial Least Squares-Discriminant Analysis (PLS-DA)*¹¹.

5.3.2.1. Modelat suau independent d'analogies de classes (SIMCA)

El modelat suau independent d'analogies de classes (SIMCA d'ara en endavant), es tracta d'un mètode de classificació on cada classe es modela utilitzant un PCA de forma independent¹¹. D'aquesta manera cada classe té el seu propi model amb un nombre òptim de PCs que pot ser igual o diferents entre classes. Té una gran importància escollir el nombre de PCs a l'hora de fer la representació ja que triar-ne un nombre molt gran podria provocar un sobre ajust del model.

Per a dur a terme el SIMCA és necessari:

1. Fer un model independent per a cada classe, triant els PCs òptims en cada cas. La Figura 8 mostra un exemple de SIMCA amb mostres d'una classe 1 i unes de classe 2.

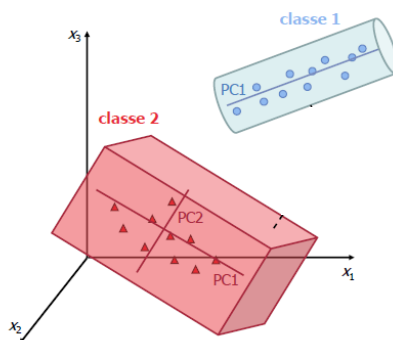


Figura 8. Exemple de modelat SIMCA amb dades d'una classe 1 (blaves) i una classe 2 (vermelles).

Com es pot observar, la classe 1 s'ha definit a partir del PC1 com un cilindre i la classe 2 a partir dels 2 primers PCs com un ortoedre. Per a saber el nombre de PC cal mirar en el gràfic de la variància explicada vs PCs a partir de quin PC la variància no canvia significativament i es seleccionen el nombre de PCs anterior a aquest.

2. Definir els límits de cada model. Aquests límits es defineixen a partir de dos paràmetres estadístics, la Q_{lim} i T^2_{lim} de Hotelling (T^2_{lim} d'ara en endavant). La T^2_{lim} fa referència a la variància retinguda pel model i la Q_{lim} a la variància residual no retinguda pel model. En aquest estudi s'utilitza un interval de confiança del 95% per a establir aquests límits¹².
3. Realitzar una validació interna. En aquest estudi s'utilitza la validació creuada (Cross-Validation) amb el mètode *Leave-One-Out* (s'utilitzarà aquest terme provinent de l'anglès per falta d'una traducció literal establerta). Aquesta validació deixa una mostra fora de la matriu de dades i aquest es prediu amb els models construïts amb la resta de dades. Aquest procés es repeteix successivament deixant cada cop una mostra diferent fora de la matriu de dades fins que tots els objectes han sigut usats per a predir¹¹.

4. Predir mostres de pertinença desconeguda. Cada mostra té uns valors estadístics T_i^2 i Q_i i si ambdós són inferiors als valors límits d'una classe la mostra queda assignada en ella. Per a simplificar s'utilitzen els paràmetres de $Q_{i,red}$ i $T_{i,red}^2$ (reduïts) per a fer l'assignació. En aquest cas es divideixen els paràmetres de cada mostra entre els límits del model respectivament i si els dos són inferiors a 1 queden assignades a la classe. Les equacions 5 i 6 mostren les expressions matemàtiques usades per al càlcul dels paràmetres reduïts.

$$T_{i,red}^2 = \frac{T_i^2}{T_{lim}^2} \quad \text{Equació 5}$$

$$Q_{i,red} = \frac{Q_i}{Q_{lim}} \quad \text{Equació 6}$$

S'obtenen valors de $T_{i,red}^2$ i $Q_{i,red}$ de cada mostra per a cada model diferent i cada mostra podrà ser classificada a un model, diversos models o a cap model. Si seguim l'exemple de la Figura 8 un objecte nou podria tenir les següents assignacions: pertany a la Classe 1, pertany a la Classe 2, pertany a la Classe 1 i a la Classe 2 o no pertany a cap classe.

5.3.2.2. Anàlisi discriminant per mínims quadrats parcials (PLS-DA)

Es tracta d'un anàlisi basat en la regressió per mínims quadrats parcials¹² adaptat per a classificar. Aquest mètode requereix d'una matriu X que contingui les variables independents (en aquest estudi les dades espectrals recollides) i una matriu Y que contingui les variables dependents (en aquest estudi fa referència a quina classe pertany cada mostra). La matriu Y utilitza un codi binari (0 i 1) per a especificar la classe de cada mostra sent (1,0,0) una mostra que pertany a la classe 1, (0,1,0) a la classe 2 i (0,0,1) a la classe 3. Posteriorment s'estableix un model PLS, que consisteix en calcular un model de components principals conjunt per a tota la matriu X de dades tenint en compte la informació que conté la matriu Y¹⁰. D'aquesta manera s'obté un únic model independentment del nombre de classes assignades en la matriu Y.

Per a fer un PLS-DA és necessari:

1. Triar el nombre de variables latents (LVs d'ara en endavant), que vindria a ser l'equivalent dels PCs. En aquest cas es pot mirar el gràfic de RMSECV (*Root Mean Square Error of Cross-Validation*) vs LVs i seleccionar el nombre de variables latents amb un mínim error de validació.
2. Establir un llindar entre les classes. Aquest separa les classes entre elles i es basa en el teorema de Bayes, que troba la millor discriminació entre les classes amb la menor probabilitat d'errors de classificació en futures prediccions¹³. La Figura 9 mostra el funcionament esquemàtic d'un PLS-DA amb 3 classes.

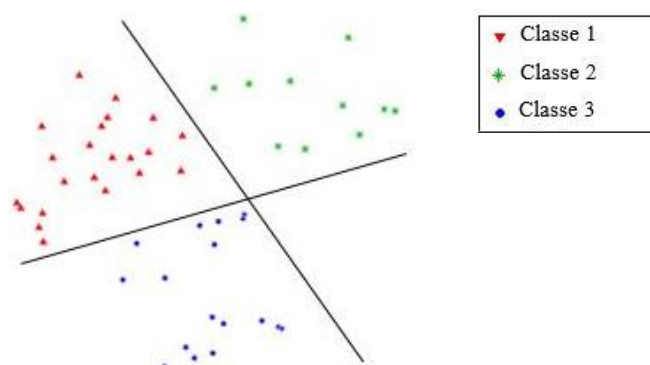


Figura 9. Esquema de funcionament del PLS-DA.

3. Realitzar una validació interna. Com en l'apartat de SIMCA, s'ha utilitzat una validació creuada amb el mètode Leave-One-Out.
4. Predir mostres de pertinença desconeguda. Per a assignar mostres a les classes s'utilitza el valor de Y predit. Existeix un valor de Y predit de cada mostra per a cada classe i aquest valor es troba al voltant de 0 si la mostra no pertany a una classe i al voltant de 1 si hi pertany¹⁴. En aquest estudi les mostres poden quedar per sobre i per sota del llindar de diferents classe i per tant les possibles assignacions són: assignació a una classe, a múltiples classes o no classificada a cap classe.

5.4. Tècniques de reducció del nombre de variables

Aquestes tècniques tenen com a objectiu eliminar les variables de la matriu de dades que siguin menys significatives per al desenvolupament del model. Són de gran ajuda ja que com indica el seu propi nom permeten reduir el nombre de variables de treball i utilitzar aquelles que tinguin una importància dins del modelatge corresponent. A continuació es descriuen breument les dues tècniques de reducció de variables usades en aquest estudi.

5.4.1. Reducció de variables

Aquest mètode consisteix en reduir el nombre de variables fent combinacions de les variables originals. Com que exactament això és el que fa un PCA s'utilitza la matriu d'scores com a nova matriu X a les tècniques de classificació. Per a establir aquesta nova matriu X s'ha de fer una elecció del nombre de PCs que s'utilitzaran. Aquest nombre ha de contenir un nivell de variància elevat ja que volem concentrar la màxima informació possible en el mínim nombre de variables. En aquest estudi s'ha agafat com a criteri el nombre de PCs que inclogués un 98% de variància.

5.4.2. Selecció de variables (VIP scores)

Hi ha diferents mètodes per a reduir el nombre de variables seleccionant quines es volen utilitzar i en aquest estudi s'ha utilitzat l'anomenada selecció de variable per la importància de la variable en la projecció (VIP per les seves sigles en anglès), per tant es descriu únicament aquest mètode. El nom més popular i el que s'utilitzarà per a fer referència a aquesta selecció de variables en aquest estudi és el de *VIP scores*^{12,15}. Es basa en seleccionar les variables originals que tenen més importància per al model. Per a dur a terme aquesta selecció es necessari:

1. Aplicar un PLS-DA al conjunt de dades
2. Fer el càlcul del VIP score de cada variable per a cada classe
3. Seleccionar aquelles variables amb un VIP score superior a 1 en alguna de les 3 classes.

D'aquesta manera s'aconsegueix reduir la matriu de treball original seleccionant les variables originals per a seguir amb l'estudi.

5.5. Paràmetres de qualitat

Els paràmetres de qualitat són l'eina que ens permet validar i comparar la qualitat dels models creats. En aquest estudi s'han utilitzat la sensibilitat, l'especificitat i les inconclusives com a paràmetres per a comparar els diferents models¹⁶. A continuació es descriuen breument aquests tres termes.

- Sensibilitat: indica la capacitat de model de reconèixer mostres realment positives, és a dir, la probabilitat d'assignar correctament mostres a la seva classe. Aquest paràmetre s'expressarà en % i s'avalua individualment per a cada classe. L'equació 7 mostra l'expressió matemàtica usada per al seu càlcul.

$$Sensibilitat_j = \frac{TP_j}{n^oS_j} \quad \text{Equació 7}$$

On TP_j són els positius vertaders (mostres de la classe j correctament assignades a la classe j i n^oS_j el nombre total de mostres que realment pertanyen a la classe j).

- Especificitat: indica la capacitat del model de reconèixer mostres que realment són diferents de la classe d'estudi, és a dir, la probabilitat de no assignar mostres a una classe a la que no pertanyen¹⁶. Aquest paràmetre s'expressarà en % i s'avalua individualment per a cada classe. L'equació 8 mostra l'expressió matemàtica usada per al seu càlcul.

$$Especificitat_j = \frac{TN_j}{n^oS_{no\ j}} \quad \text{Equació 8}$$

On TN_j són els negatius vertaders (mostres que no són de la classe j i no han estat predites en la classe j) i $n^\circ S_{no j}$ el nombre total de mostres que no pertanyen a la classe j .

- Inconclusives: indica la proporció de mostres que no poden ser assignades a una classe de forma concloent perquè tenen una assignació múltiple o una no assignació. Aquest paràmetre s'expressarà en % i s'avalua individualment per a cada classe. L'equació 9 mostra l'expressió matemàtica usada per al seu càlcul.

$$Inconclusives_j = \frac{NA_j + MA}{n^\circ S_j} \quad \text{Equació 9}$$

On NA_j són les mostres de la classe j no assignades a cap classe, MA les mostres de la classe j múltiples assignades a més d'una classe i $n^\circ S_j$ el nombre total de mostres que pertanyen a la classe j .

6. **PART EXPERIMENTAL**

Les mostres no han requerit cap preparació ni tractament previ a les mesures espectrals.

Les mesures de les diferents mostres d'oli d'oliva verge extra s'han realitzat emprant la tècnica del FT-Raman. Concretament s'ha utilitzat l'equip Thermo Nicolet 5700 FT-IR equipat amb un mòdul NXR de FT-Raman, un detector de InGaAs, un divisor de feix de CaF_2 i un làser de Nd-YAG. Les condicions de mesura òptimes trobades han estat:

- Radiació de 1064 nm del làser
- Poder del làser de 0,25W
- Resolució de 4 cm^{-1}
- 124 scans per cada espectres
- Rang de mesura de $100\text{-}3700 \text{ cm}^{-1}$

Les dades espectrals han sigut exportades i posteriorment tractades al programari Matlab usant la versió Matlab R2012b i l'eina suplementària PLS Toolbox.

7. RESULTATS I DISCUSSIÓ

En primer lloc s'ha realitzat una representació del conjunt de dades espectrals per observar la forma dels espectres, possibles diferències entre aquests i pretractaments a realitzar. La Figura 10 mostra la representació del total de mostres mesurades (65 mostres).

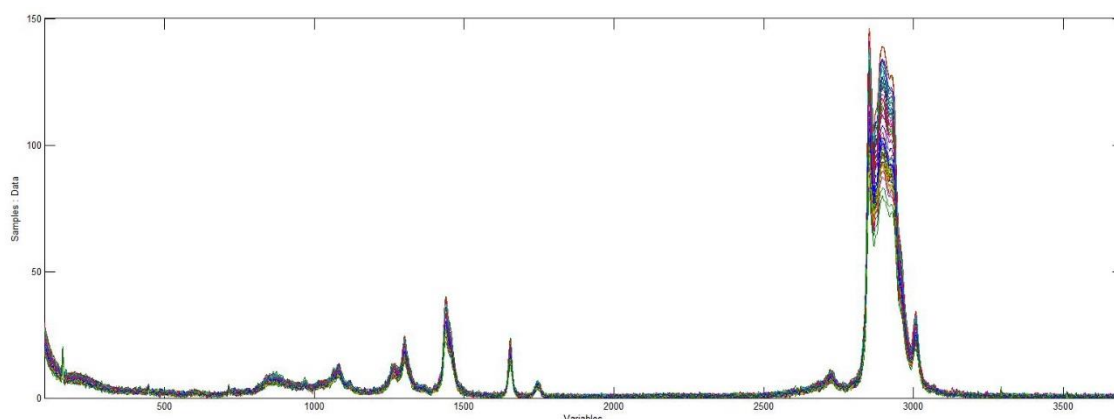


Figura 10. Representació de les dades obtingudes en la mesura dels espectres de les 65 mostres

Es pot observar que no s'aprecien diferències entre els olis estudiats en funció del seu origen geogràfic i que en la part final de l'espectre no hi ha cap tipus de senyal així que s'ha decidit tallar la matriu de dades i utilitzar les dades que van de l'100 a 3200 cm^{-1} . Tanmateix es veu que l'espectre presenta soroll, per tant s'ha considerat oportú aplicar un suavitzat a les dades per a disminuir la presència de soroll. S'ha fet amb una amplada de finestra petita (5) ja que es desconeix la rellevància dels pics més petits de l'espectre i no es volen menysprear. Així doncs es decideix tallar la matriu de dades, realitzar un suavitzat i també un centrat de les dades.

El següent pas ha sigut realitzar un PCA de cada una de les 3 classes (Garrigues, Siurana i Sud Espanya) per a observar el comportament de les mostres. A més també s'ha fet un PCA del conjunt total de mostres per a veure si hi ha tendències associatives entre les mostres d'una mateixa classe. La Figura 11 mostra el gràfic d'scores PC1 vs PC2 dels 4 PCAs anteriors.

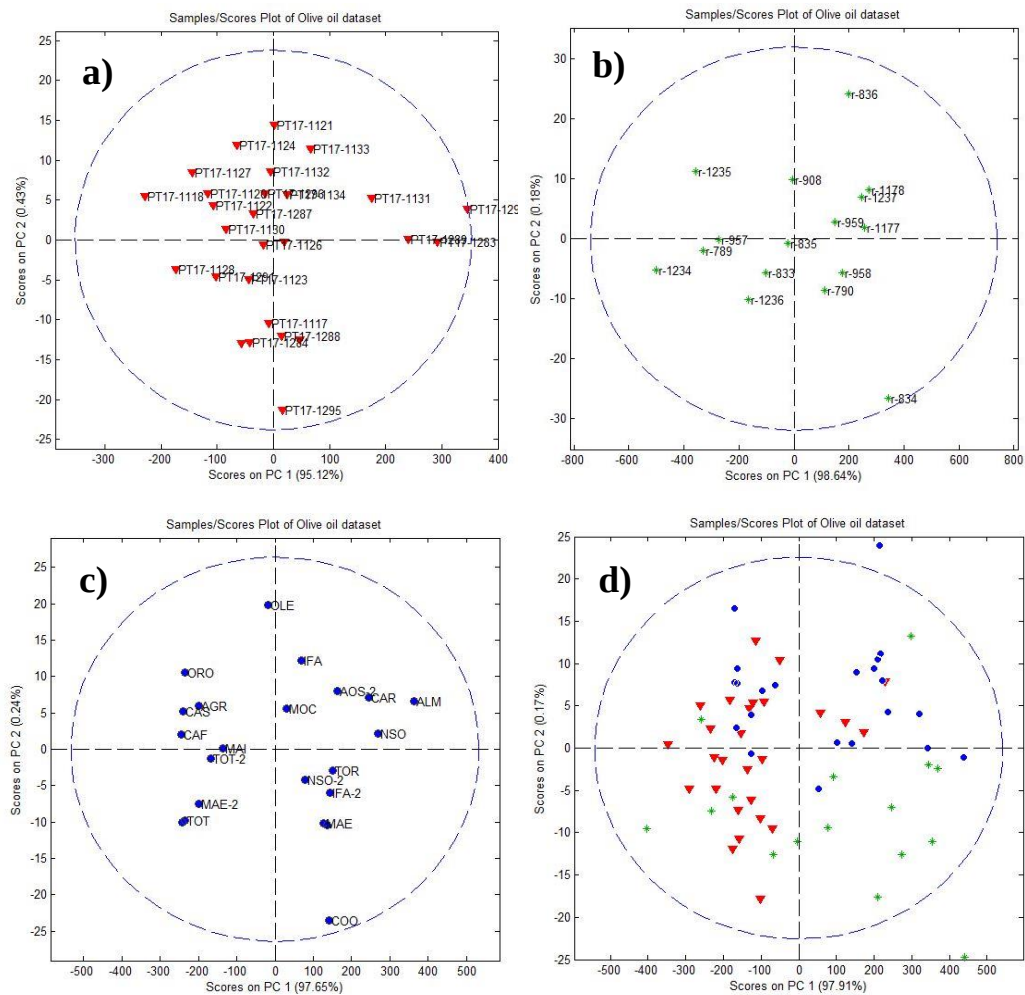


Figura 11. Gràfics de PC1 vs PC2 dels PCAs de: a) Garrigues, b) Siurana, c) Sud Espanya i d) totes les mostres (en aquest cas es mantenen les marques de les representacions individuals, vermell-Garrigues, verd-Siurana i blau-Sud Espanya). En tots els casos les circumferències amb línies discontinües marquen el límit de confiança del 95%.

Es pot observar com en els PCAs de cada classe es troba alguna mostra separada de les altres però dins del límit de confiança així que no es tracten en cap cas com a outlier. Si es centra l'atenció en la Figura 11 a) es pot observar com, en general, les mostres tenen una lleugera tendència a trobar-se al llarg del PC2. En la Figura 11 b) es pot veure com les mostres es troben al voltant del centre de coordenades. Si es mira la Figura 11 c) no sembla haver-hi cap tendència associativa. Tot i això les tendències observades són mínimes.

- Classificació amb les variables originals

Un cop s'han representat les dades s'ha procedit a fer l'anàlisi SIMCA. Per a fer-lo s'han utilitzat 2 PCs per a les 3 classes amb una informació retinguda del 95,56% per a Garrigues, 98.82% per a Siurana i 97,89% per a Sud Espanya. La Figura 12 mostra el gràfic de variància explicada a partir del qual s'han seleccionat els PCs òptims de la classe Garrigues.

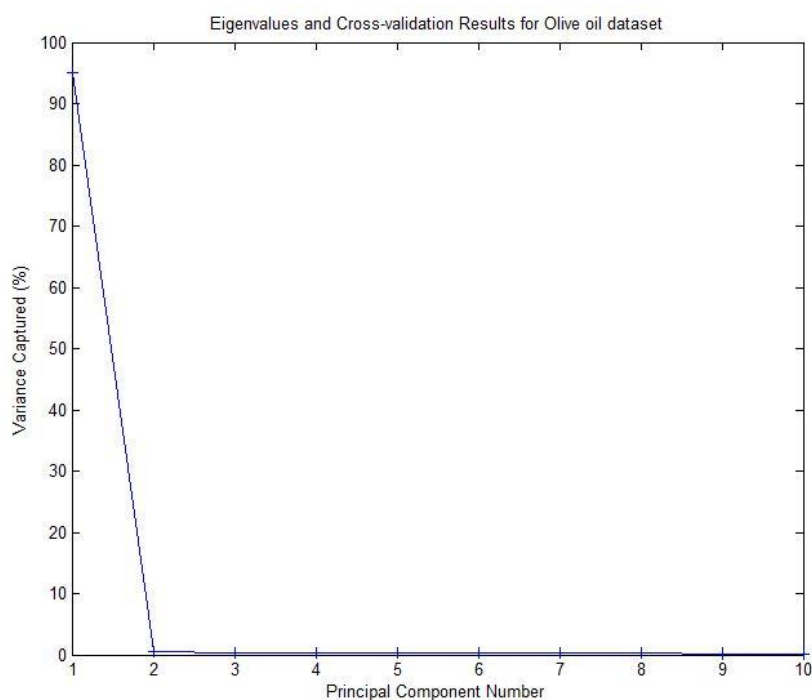


Figura 12. Gràfic de variància explicada vs PCs de la classe Garrigues.

Com es pot veure a partir del PC2 la variància explicada per a cada PC pràcticament no canvia. La variància del PC2 és del 0,19%; la del PC3 del 0,12% i la del PC4 del 0.10% així que es considera que no canvia significativament a partir del PC3 i es seleccionen 2 PCs. Es segueix el mateix criteri per a totes les classes i tots els models SIMCA d'aquest estudi.

La Figura 13 mostra els gràfics de T^2_{red} vs Q_{red} de les tres classes de forma que les mostres que queden enquadrades per sota de 1 en abscisses i ordenades són les que s'assignen a aquella classe.

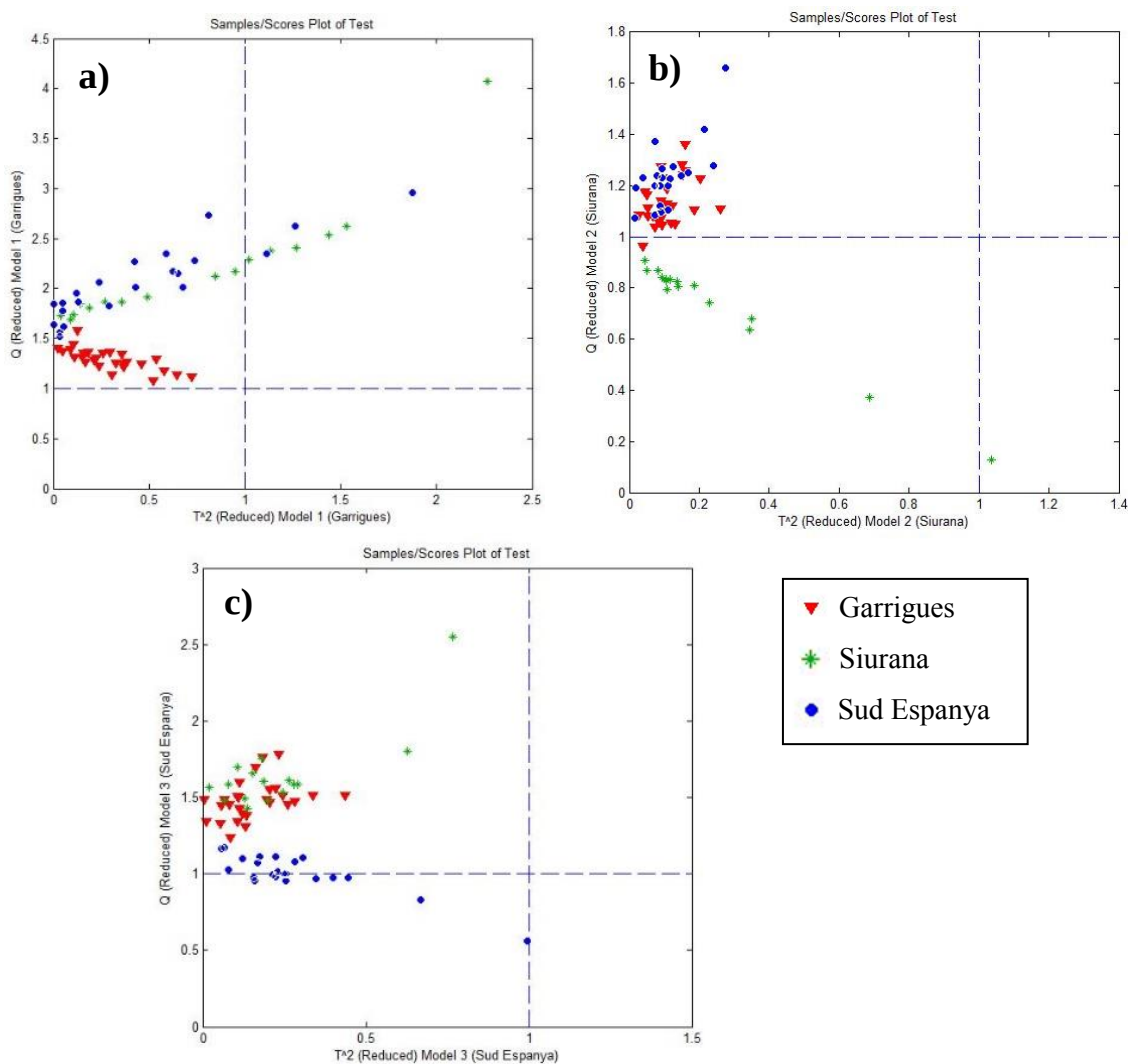


Figura 13. Gràfics de les assignacions de les 3 classes: a) Garrigues, b) Siurana i c) Sud Espanya per a un model SIMCA usant els pretractaments centrat i suavitzat.

Es pot observar com el model no classifica cap mostra com a Garrigues, assigna totes les de Siurana menys 1 correctament i aproximadament la meitat de les de Sud Espanya també correctament. Tot i no assignar totes les mostres, només hi ha una mostra que ha assignat incorrectament, les altres han quedat sense classificar. La Taula 2 mostra els valors dels paràmetres de qualitat.

Taula 2. Paràmetres de qualitat del model SIMCA amb els pretractaments centrat i suavitzat.

	% Sensibilitat	% Especificitat	% d'inconclusives
Garrigues	0	100	96,30
Siurana	93,75	97,96	6,25
Sud Espanya	45,45	100	54,55

La taula mostra valors elevats d'especificitat però valors baixos de sensibilitat per a les classes Garrigues i Sud Espanya i un % d'inconclusives elevat per a les mateixes classes.

Es prova a fer el model mitjançant un PLS-DA el qual s'ha fet usant 3 LVs, amb una variància de 98,11% sobre el bloc X i 76,74% sobre el bloc Y. La Figura 14 mostra el gràfic de RMSECV vs LVs de les tres classes.

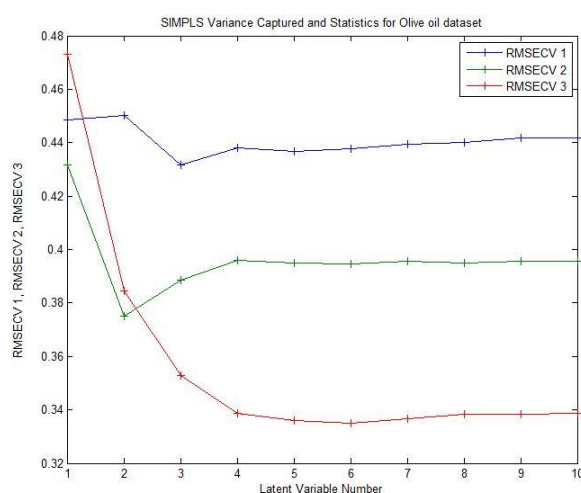


Figura 14. Gràfic de RMSECV vs LVs de cada classe. Blau-Garrigues, Verd-Siurana, Vermell-Sud espanya.

Es pot veure que tant el RMSECV 1 com el 2 tenen un valor mínim en 2 i 3 variables latents respectivament. En canvi el RMSECV 3 disminueix i es manté pràcticament constant a partir de 4 variables latents. S'ha decidit usar 3 LVs ja que és el valor mínim per a Garrigues i el segon valor més petit per a Siurana, i el RMSECV 3 es més petit amb 3 LVs que amb 2 LVs.

El resultat ha sigut una classificació amb els paràmetres mostrats a la Taula 3.

Taula 3. Paràmetres de qualitat del model PLS-DA amb els pretractaments centrat i suavitzat.

	% Sensibilitat	% Especificitat	% d'inconclusives
Garrigues	85,19	57,89	44,44
Siurana	81,25	85,71	31,25
Sud Espanya	86,36	82,76	40,91

S'observa un augment de la sensibilitat respecte el model SIMCA però també una baixada de l'especificitat i uns valors d'inconclusives massa elevats.

- Classificació amb selecció de variables

Per a intentar millorar els resultats es decideix reduir del nombre de variables pel mètode de reducció de variables. Es realitza un PCA del conjunt de totes les mostres amb els mateixos pretractaments i s'agafa els valor dels scores dels primers 5PCs (98,28% de variància) de totes les mostres. Amb aquests valors es crea una matriu amb totes les mostres i únicament 5 variables. A partir d'aquí es torna a crear un model SIMCA amb aquesta matriu. Per a les tres classes es trien 3 PCs amb una informació retinguda del 99,91% per a Garrigues, 99,96% per a Siurana i 99,89% per a Sud Espanya. La Figura 15 mostra els gràfics de T^2_{red} vs Q_{red} de les 3 classes.

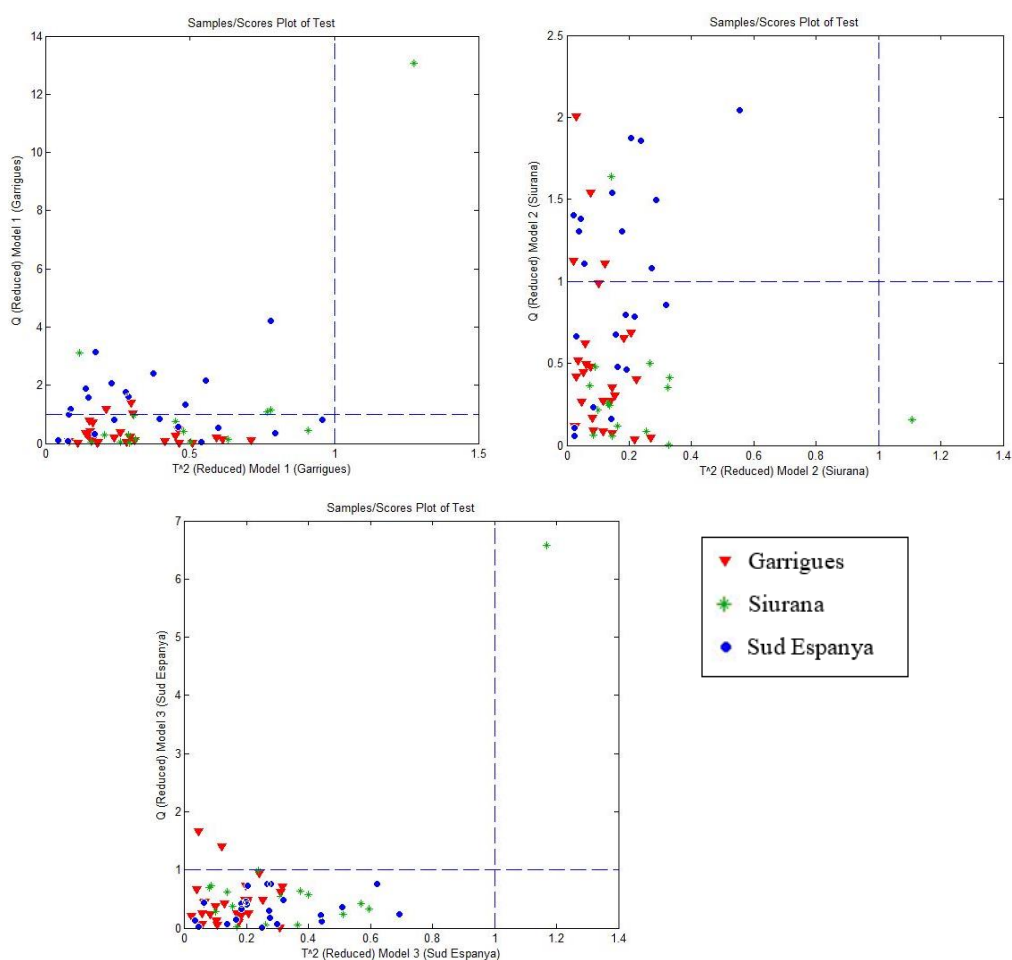


Figura 15. Gràfics de les assignacions de les 3 classes. Superior esquerra Garrigues, superior dreta Siurana i inferior Sud Espanya per a un model SIMCA usant els scores dels 5 primers PCs.

Com es pot observar per a les tres classes hi ha múltiples assignacions de mostres que pertanyen a classes diferents. Es pot dir que hi ha una certa aleatorietat a l'hora de classificar les mostres així que com a model és dolent. Així ho confirmen les dades dels paràmetres de qualitat mostrats en la Taula 4. En ella es mostren valors molt baixos

d'especificitat i molt alts d'inconclusives el que deix entreveure un elevat nombre de mostres amb assignació múltiple.

Taula 4. Paràmetres de qualitat del model SIMCA creat amb els scores dels 5 primers PCs.

	% Sensibilitat	% Especificitat	% d'inconclusives
Garrigues	88,89	39,47	96,30
Siurana	87,5	30,61	93,75
Sud Espanya	100	13,04	63,64

Els resultats, molt lluny de millorar, empitjoren així que es decideix reduir el nombre de variables pel mètode dels VIP scores. S'obtenen els valors de totes les variables per a cada classe en la matriu *output* (de resultats) del PLS-DA. Les variables que tenen un VIP scores superior a 1 en qualsevol de les 3 classes es seleccionen en la matriu de dades original i la resta no s'utilitzaran per a la creació dels models. A partir d'aquest moment es treballa amb aquesta matriu en la que s'han seleccionat les variables. Han quedat seleccionades un total del 1240 variables de les 6431 que es tenien abans. La Figura 16 mostra la representació de l'espectre (sense pretractaments) amb només les variables seleccionades.

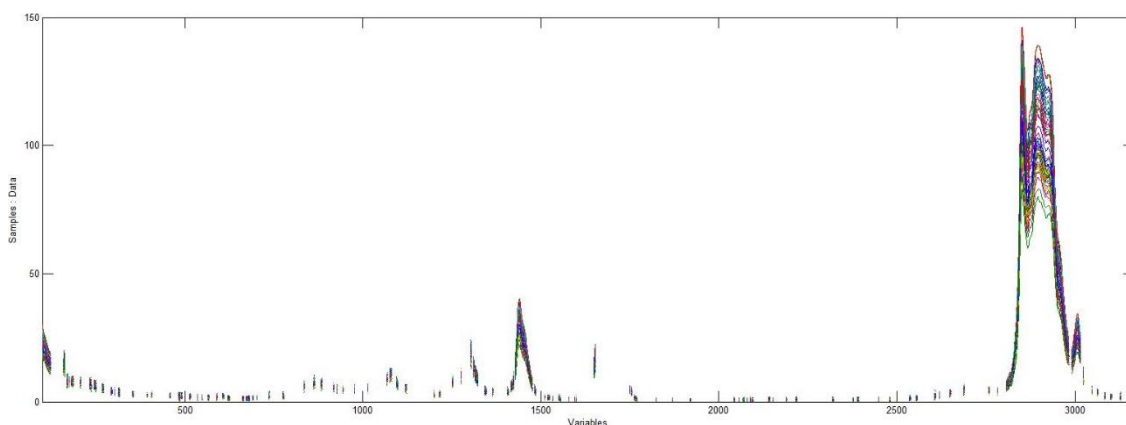


Figura 16. Representació de l'espectre amb només les variables amb VIP score major que 1 seleccionades.

Un cop es té la matriu amb les variables desitjades seleccionades es procedeix a fer un model SIMCA amb aquestes dades. Aquest modelatge s'ha fet utilitzant els mateixos pretractaments previs (centrat i suavitzat d'amplada de finestra 5) amb 3 PCs per a Garrigues (99,02% de variància explicada), 2 PCs per a Siurana (99,73% de variància explicada) i 4 PCs per a Sud Espanya (99,59% de variància explicada). El temps de creació del model per part del software disminueix notablement respecte si es fa servir la

matriu de dades original. La Figura 17 mostra les assignacions que s'han donat en aquest model i la Taula 5 els paràmetres de qualitat.

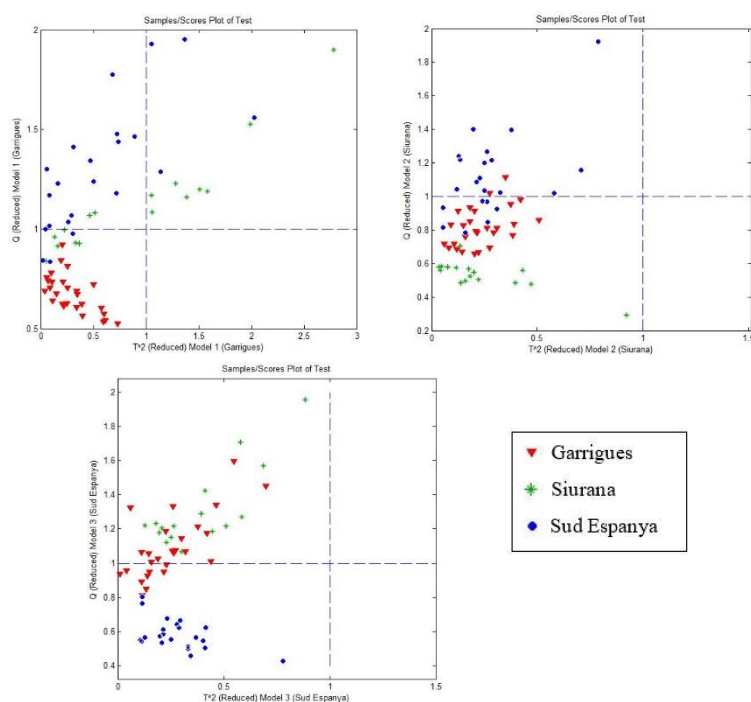


Figura 17. Gràfics de les assignacions de les 3 classes. Superior esquerra Garrigues, superior dreta Siurana i inferior Sud Espanya per a un model SIMCA les variables VIP scores i els pretractaments centrat i suavitzat.

Taula 5. Paràmetres de qualitat del model SIMCA creat la matriu de variables seleccionades i els pretractaments centrat i suavitzat

	% Sensibilitat	% Especificitat	% d'inconclusives
Garrigues	100	73,68	96,30
Siurana	100	34,69	37,5
Sud Espanya	100	18,18	31,82

S'obtenen unes sensibilitats perfecte però unes especificitats molt baixes i uns valors d'inconclusives massa elevats.

S'ha fet també un PLS-DA usant 4 LVs (99,57% de variància sobre el bloc X i 83,66 sobre el bloc Y) i paràmetres obtinguts es mostren en la Taula 6.

Taula 6. Paràmetres de qualitat del model PLS-DA amb la matriu de variables seleccionades i els pretractaments centrat i suavitzat.

	% Sensibilitat	% Especificitat	% d'inconclusives
Garrigues	96,30	97,37	3,70
Siurana	100	100	0
Sud Espanya	95,45	100	0

S'observa fàcilment que els resultats obtinguts són molt bons per als 3 paràmetres amb valors de sensibilitat i especificitat molt elevats i d'inconclusives molt baixos.

S'ha pensat que potser la banda que apareix sobre 2900 cm^{-1} pot estar eclipsant informació dels altres pics més petits ja que els triplica en intensitat així que s'ha decidit fer un PCA conjunt de totes les dades usant els pretractaments de centrat i autoescalat amb la matriu original i posteriorment amb la matriu de variables seleccionades. Els resultats d'aquests PCAs es mostren en la Figura 18.

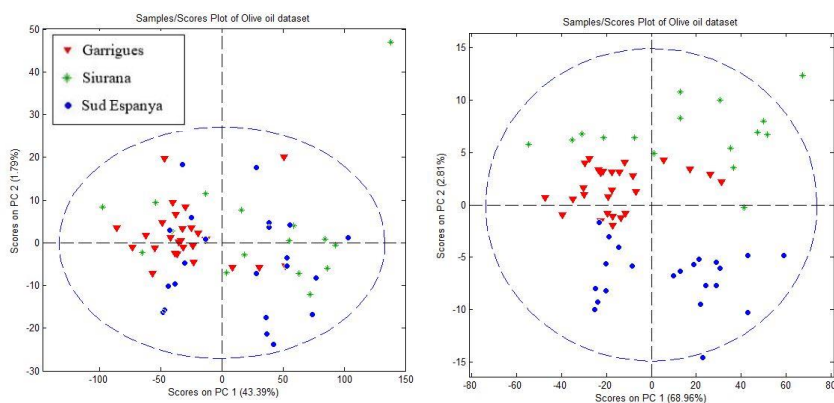


Figura 18. PCAs realitzats amb la matriu de dades original (esquerra) i amb la matriu de variables seleccionades per VIP scores (dreta).

Ràpidament es veu que en el PCA realitzat amb la matriu que conté les variables seleccionades per VIP scores hi ha tendències associatives de les diferents classes. Veient aquestes agrupacions es descarta utilitzar l'autoescalat amb la matriu de dades original i es torna a fer un model SIMCA i PLS-DA utilitzant la matriu de dades de variables seleccionades i els pretractaments de centrat i autoescalat.

Per al SIMCA es seleccionen 6 PCs (74,96% de variància) per a Garrigues, 4 PCs (82,37% de variància) per a Siurana i 6 PCs (81,82% de variància) per a Sud Espanya. La Figura 19 mostra les assignacions obtingudes i la Taula 7 els paràmetres de qualitat.

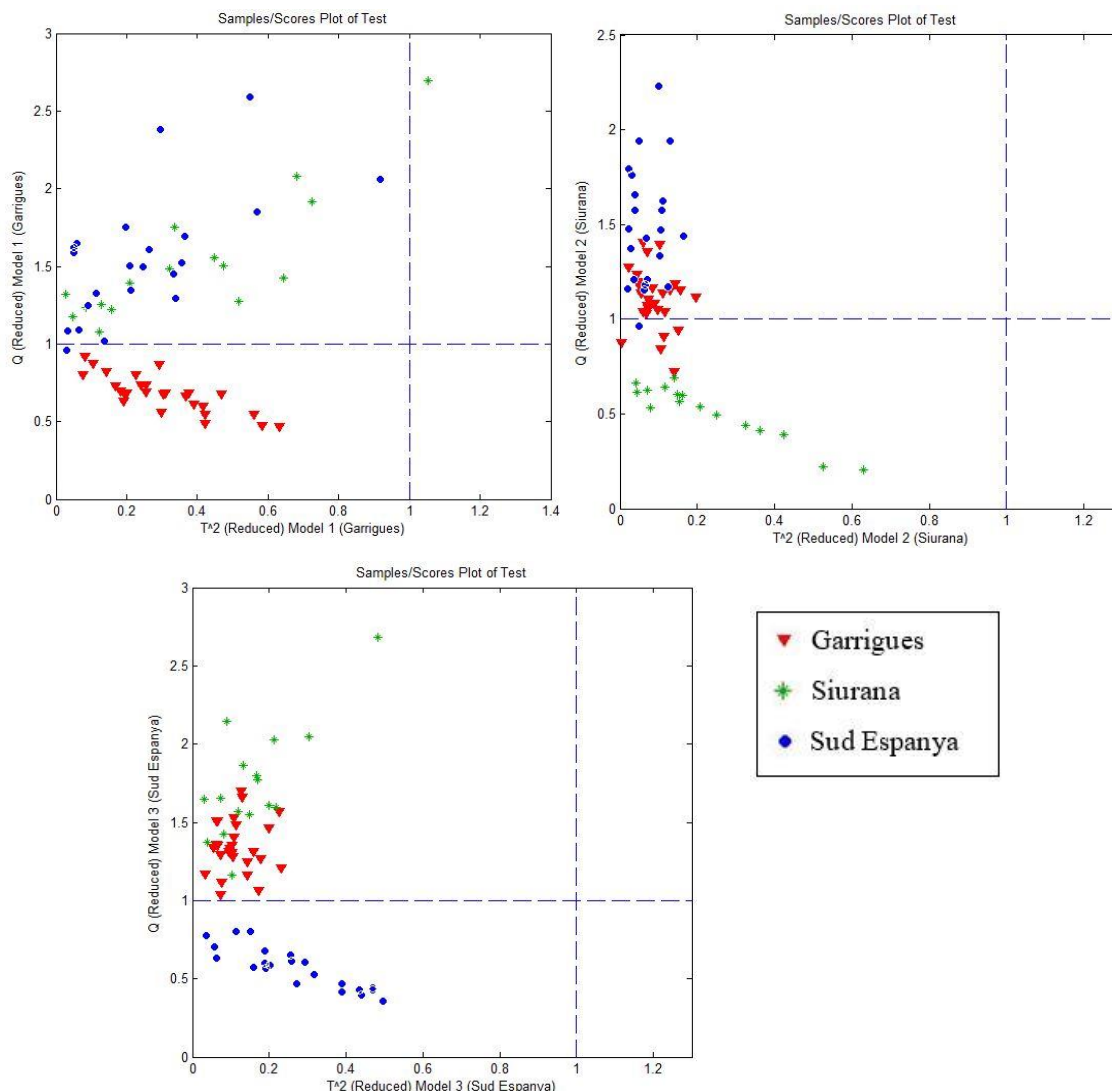


Figura 19. Gràfics de les assignacions de les 3 classes. Superior esquerra Garrigues, superior dreta Siurana i inferior Sud Espanya per a un model SIMCA les variables VIP scores i els pretractaments centrat i autoescalat.

Taula 7. Paràmetres de qualitat del model SIMCA amb la matriu de variables seleccionades per VIP scores i els pretractaments centrat i autoescalat.

	% Sensibilitat	% Especificitat	% d'inconclusives
Garrigues	100	97,9	18,52
Siurana	100	87,76	0
Sud Espanya	100	100	4,55

Aquest model presenta unes sensibilitats perfectes i uns valors d'especificitat molt elevats. El % d'inconclusives és baix per a les tres classes essent la de Garrigues la que el té una mica més elevat. Aquests paràmetres són molt millors que els obtinguts anteriorment en qualsevol SIMCA.

El PLS-DA s'ha modelat amb 3 LVs amb una variància del 72,63% sobre el bloc X i del 80,95% sobre el bloc Y. Els paràmetres obtinguts en aquest PLS-DA es presenten en la Taula 8.

Taula 8. Paràmetres de qualitat del model PLS-DA amb la matriu de variables seleccionades per VIP scores i els pretractaments centrat i autoescalat.

	% Sensibilitat	% Especificitat	% d'inconclusives
Garrigues	92,59	94,74	7,41
Siurana	100	100	6,25
Sud Espanya	100	96,43	9,09

Es pot observar que els valors de sensibilitat són molt bons així com els d'especificitat i els d'inconclusives.

S'ha vist que amb els pretractaments de centrat i autoescalat i utilitzant la matriu de variables seleccionades es poden diferenciar les mostres segons el seu origen. La Figura 20 mostra el PCA conjunt ja vist en la Figura 18. Es pot veure que la diferenciació s'ha donat gràcies als valors dels scores de les mostres en el PC2. Les tres classes s'estenen al llarg del PC1 de forma similar però la classe Siurana es troba a valors del PC2 superiors a 5, Garrigues entre -2 i 5 i Sud Espanya inferiors a -5 aproximadament.

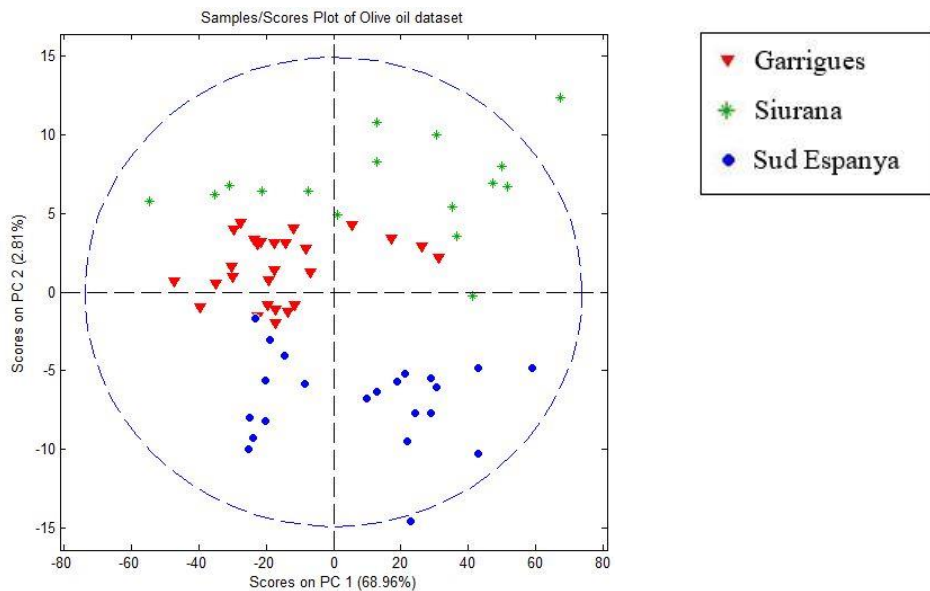


Figura 20. PCA realitzat amb la matriu de variables seleccionades i pretractaments centrat i autoescalat.

Així doncs la variància retinguda en el PC2 ha sigut la que ens ha permès fer la diferenciació de classes. És interessant doncs, mirar el gràfic de loadings per a observar quines variables han tingut més rellevància en el PC2 i saber així quina zona de l'espectre és la que presenta diferenciacions entre les classes. La Figura 21 mostra el gràfic de loadings de les variables en el PC2.

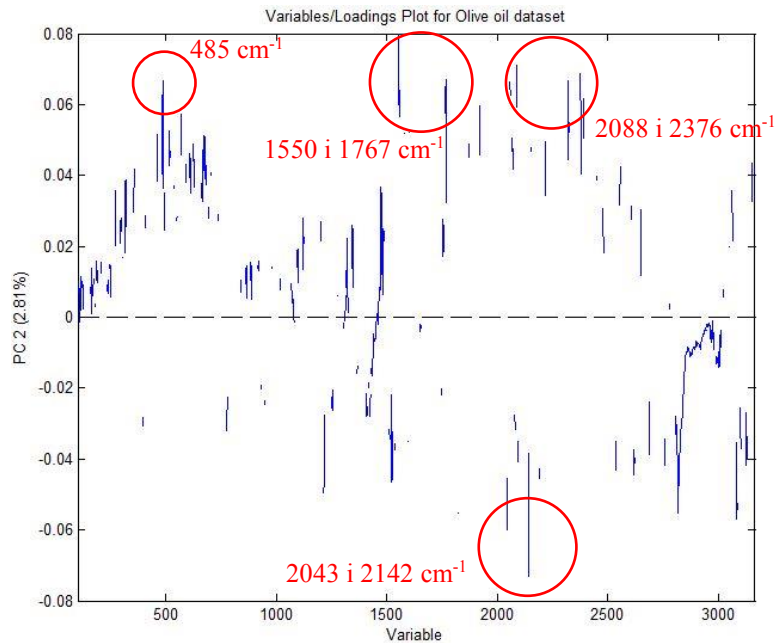


Figura 21. Gràfic dels loadings del PC2.

D'aquest gràfic cal destacar que les variables amb més contribució en el PC2 són aquelles amb major valor de loadings (en valor absolut) en el PC2. Aquestes variables es troben encerclades en la Figura 21 i les etiquetes mostren els nombres d'ona. Si es compara amb l'espectre que es mostra en la Figura 22 (mateix que la Figura 16) es pot veure que aquests nombres d'ona corresponen a zones de baixa intensitat de l'espectre i, per tant, no es poden assignar a cap banda fonamental del FT-Raman. Tot i això en aquestes regions les mostres d'una mateixa classe mostren tendències en la forma de l'espectre i a l'aplicar l'autoescalat prenen importància i ens permeten classificar.

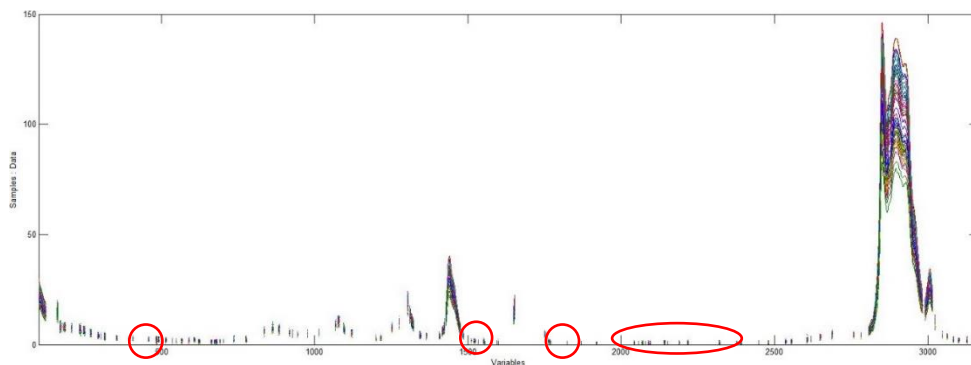


Figura 22. Espectre amb només les variables seleccionades.

Tenim doncs, 3 models que en base als seus paràmetres de qualitat ens classifiquen correctament les mostres. Aquests són:

1. Model SIMCA realitzat amb la matriu de variables seleccionades per VIP scores i els pretractaments centrat i autoescalat. Paràmetres de qualitat en la Taula 7.
2. Model PLS-DA realitzat amb la matriu de variables seleccionades per VIP scores i els pretractaments centrat i autoescalat. Paràmetres de qualitat en la Taula 8.
3. Model PLS-DA realitzat amb la matriu de variables seleccionades per VIP scores i els pretractaments centrat i suavitzat amb amplada de finestra de 5. Paràmetres de qualitat en la Taula 6.

Entre els dos models PLS-DA (vàlids els dos), el que s'ha fet amb el suavitzat presenta menor percentatge d'inconclusives per a totes les classes i major especificitat i té sensibilitats molt semblants a l'altre PLS-DA. Per tant s'han obtingut un model SIMCA i un PLS-DA capaços de diferenciar les tres classes de mostra. La Taula 9 mostra una comparació dels paràmetres per al model SIMCA i el PLS-DA. En ambdós models s'obtenen bones sensibilitats i especificitats però destaca el nombre d'inconclusives de la classe Garrigues en el SIMCA ja que és més elevat que en les altres classes i models perquè 5 mostres de Garrigues queden doblement assignades a les classes Garrigues i Siurana.

Taula 9. Comparació dels paràmetres de qualitat del model SIMCA i PLS-DA

		% Sensibilitat	% Especificitat	% d'inconclusives
SIMCA	Garrigues	100	97,9	18,52
	Siurana	100	87,76	0
	Sud Espanya	100	100	4,55
PLS-DA	Garrigues	96,30	97,37	3,70
	Siurana	100	100	0
	Sud Espanya	95,45	100	0

8. CONCLUSIONS

En aquest treball s'ha desenvolupat un model de classificació i diferenciació de mostres d'oli d'oliva verge de la varietat arbequina segons la seva procedència. Aquests models s'han establert analitzant les mostres per FT-Raman i utilitzant el programari Matlab per a l'anàlisi de dades. Els models han sigut validats amb 3 paràmetres de qualitat.

S'ha aconseguit establir i validar un model SIMCA aplicant dos pretractaments a les dades (un centrat i un autoescalat) i seleccionant aquelles variables que tenen un valor de VIP score superior a 1. Aquest model presenta sensibilitats del 100% per a totes les classes; especificitats del 97,9%, 87,76% i 100% per a Garrigues, Siurana i Sud Espanya respectivament; i inconclusives del 18,52%, 0% i 4,55% per a Garrigues, Siurana i Sud Espanya respectivament.

Per altra banda s'ha establert i validat un model PLS-DA aplicant dos pretractaments a les dades (un centrat i un suavitzat amb amplitud de finestra de 5) i seleccionant aquelles variables amb un valor de VIP score superior a 1. Les sensibilitats d'aquest model són del 96,3%, 100% i 95,45% per a Garrigues, Siurana i Sud Espanya respectivament. Les especificitats, del 97,37% per a Garrigues i del 100% per a les altres dues classes. Finalment els valors d'inconclusives són del 3,7% per a Garrigues i del 0% per a Siurana i Sud Espanya.

Per aquestes raons, es pot dir que l'espectroscòpia de Raman amb transformada de Fourier permet classificar i diferenciar mostres d'oli d'oliva verge extra d'arbequina segons el seu origen. També es pot dir que s'ha après el funcionament d'una tècnica (i el seu equip) no explicada durant el grau (FT-Raman). De la mateixa forma, s'ha millorat l'ús del Matlab en l'àmbit d'anàlisi de dades ja que s'havien utilitzat altres programaris per a fer els anàlisis de dades durant el grau.

In this project, it has been developed a differentiation and classification method for arbequina extra virgin olive oil according to their origin. These models have been established analysing the sample through FT-Raman and using the Matlab software to analyse the data. The models have been validated with 3 quality parameters.

A SIMCA model has been established and validated applying two data preprocesses (mean centring and autoscaling) and selecting those variables with a VIP score bigger than 1. This model presents sensibilities of 100% in all of the classes; specificities of 97,9%, 87,76% and 100% for Garrigues, Siurana and Sud Espanya respectively; and inconclusive of 18,52%, 0% and 4,55% for Garrigues, Siurana and Sud Espanya respectively.

On the other hand, a PLS-DA model has been established and validated applying two data preprocesses (mean centring and 5 filter width smoothing) and also selecting those variables with a VIP scores bigger than 1. This model sensibilities are 96,3%, 100% and 95,45% for Garrigues, Siurana and Sud Espanya respectively. The specificities are 97,37% for Garrigues and 100% for the other two classes. Finally, the inconclusive percentage are 3,7% for Garrigues and 0% for the Siurana and Sud Espanya

For this reason, it can be state that the Fourier transform Raman spectroscopy allows us to classify and to differentiate arbequina extra virgin olive oil sample according to their geographical origin. Also, it can be state that the functioning of a new technique and equipment (FT-Raman) has been learned. In the same way, Matlab operating has been improved in particular in the data analysis field so it the degree other software have been used to analyse the data.

9. BIBLIOGRAFIA

- (1) Sánchez-López, E.; Sánchez-Rodríguez, M. I.; Marinas, A.; Marinas, J. M.; Urbano, F. J.; Caridad, J. M.; Moalem, M. *Talanta* **2016**, *156–157*, 180–190.
- (2) Dias, C.; Mendes, L. *Food Res. Int.* **2018**, *103*, 492–508.
- (3) Aliments d'origen – Entitat sense ànim de lucre, que agrupa la majoria dels Consell Reguladors DOP-IGP de Catalunya. <http://www.alimentsdorigen.cat/> (Data de consulta 16 de maig del 2018).
- (4) EU quality logos | Agricultura y Desarrollo Rural https://ec.europa.eu/agriculture/quality/schemes_es (Data de consulta 16 de maig del 2018).
- (5) Extra virgin olive oils. Catalonia Protected Designations of Origin <http://www.olisdopcatalans.cat/index.php?module=denominacio&lang=eng> (Data de consulta 16 de maig del 2018).
- (6) Junta de Andalucía - Aceite de Oliva Virgen Extra <http://www.juntadeandalucia.es/organismos/agriculturapescayderosarrollorural/areas/industrias-agroalimentarias/calidad-promocion/paginas/denominaciones-calidad-aceite-oliva.html> (Data de consulta 16 de maig del 2018).
- (7) Skoog, D. A.; Holler, F. J.; R. Crouch, S. *Principio de análisis instrumental*, Sexta Edic.
- (8) Parker, S. F. *Spectrochim. Acta Part A Mol. Spectrosc.* **1994**, *50*, 1841–1856.
- (9) Ricardo Lucio Gutiérrez, J.; Coello Bonilla, J. Aplicación de Métodos Quimiométricos para la Caracterización y Control de Calidad de Plantas Medicinales, Universitat Autònoma de Barcelona, 2011.
- (10) Brereton, R. G. *Appl. Chemom. Sci.* **2007**, 1–379.
- (11) Massart, D. L.; Vandeginste, B. G. M.; Buydens, J. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verberke, J.; Buydens, L. M. C.; Jong, S. De; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; 1997.
- (12) Wise, B. M.; Gallagher, N. B.; Windig, W. *Chemometrics Tutorial for PLS – Toolbox and Solo*; 2006.
- (13) Moncayo Martin, S.; Cáceres Gianni, J. O. Desarrollo y aplicación de métodos quimiométricos para el estudio de muestras mediante Espectroscopia de Ablación Láser (LIBS), Universidad Complutense de Madrid, 2016.
- (14) Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166–173.
- (15) Farrés, M.; Platikanov, S.; Tsakovski, S.; Tauler, R. *J. Chemom.* **2015**, *29*, 528–536.
- (16) Márquez, C.; López, M. I.; Ruisánchez, I.; Callao, M. P. *Talanta* **2016**, *161*, 80–86.

ANNEXOS

MOSTRA	ORIGEN	DATA D'ENTREGA	MOSTRA	ORIGEN	DATA D'ENTREGA
r-789	S	06/03/2018	AOS	COR	27/04/2018
r-790	S	06/03/2018	AOS-2	COR	27/04/2018
r-833	S	06/03/2018	AGR	COR	27/04/2018
r-834	S	06/03/2018	ALM	COR	27/04/2018
r-835	S	06/03/2018	CAR	COR	27/04/2018
r-836	S	06/03/2018	HOJ	COR	27/04/2018
r-908	S	06/03/2018	CAF	GRA	27/04/2018
r-957	S	06/03/2018	MAE	GRA	27/04/2018
r-958	S	06/03/2018	MAE-2	GRA	27/04/2018
r-959	S	06/03/2018	CAS	GRA	27/04/2018
r-1177	S	06/03/2018	MOC	GRA	27/04/2018
r-1178	S	06/03/2018	NSO	HUE	27/04/2018
r-1234	S	06/03/2018	NSO-2	HUE	27/04/2018
r-1235	S	06/03/2018	IFA	JAE	27/04/2018
r-1236	S	06/03/2018	IFA-2	JAE	27/04/2018
r-1237	S	06/03/2018	COO	JAE	27/04/2018
PT17-1296	G	06/03/2018	TOR	JAE	27/04/2018
PT17-1283	G	06/03/2018	ORO	JAE	27/04/2018
PT17-1295	G	06/03/2018	MAI	SEV	27/04/2018
PT17-1291	G	06/03/2018	OLE	SEV	27/04/2018
PT17-1133	G	06/03/2018	TOT-2	MUR	27/04/2018
PT17-1293	G	06/03/2018	TOT	MUR	27/04/2018
PT17-1126	G	06/03/2018			
PT17-1134	G	06/03/2018			
PT17-1284	G	06/03/2018			
PT17-1132	G	06/03/2018			
PT17-1289	G	06/03/2018			
PT17-1288	G	06/03/2018			
PT17-1129	G	06/03/2018			
PT17-1124	G	06/03/2018			
PT17-1128	G	06/03/2018			
PT17-1287	G	06/03/2018			
PT17-1125	G	06/03/2018			
PT17-1127	G	06/03/2018			
PT17-1117	G	06/03/2018			
PT17-1118	G	06/03/2018			
PT17-1120	G	06/03/2018			
PT17-1123	G	06/03/2018			
PT17-1119	G	06/03/2018			
PT17-1122	G	06/03/2018			
PT17-1130	G	06/03/2018			
PT17-1121	G	06/03/2018			
PT17-1131	G	06/03/2018			

S = SIURANA

G = GARRIGUES

COR = CORDOBA

GRA = GRANADA

SEV = SEVILLA

JAE = JAEN

HUE = HUELVA

MUR = MURCIA

Sud Espanya

Annex 1. Taula de nomenclatura de les mostres.