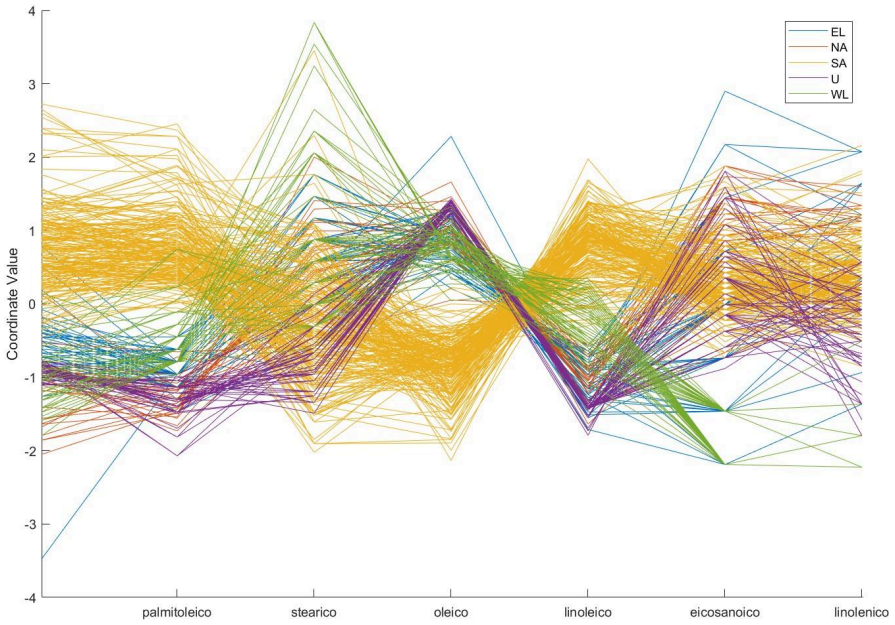
Primo Assignment

1) *È possibile individuare una o più variabili in grado di distinguere alcune delle categorie? (specificare quali variabili e quali categorie).*

Grafico



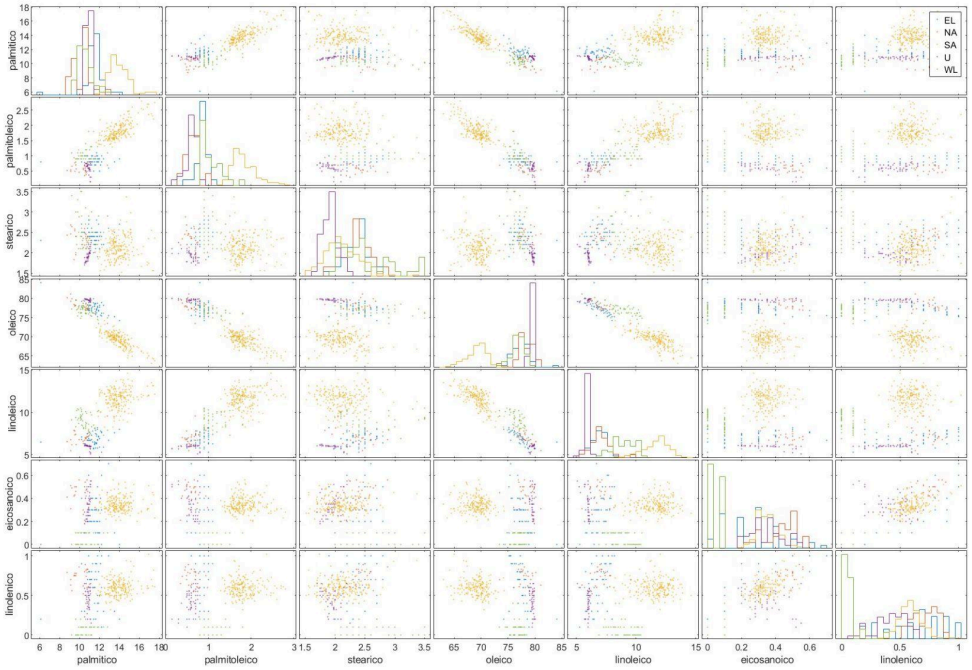
Risposta

Sì, è possibile individuare delle variabili in grado di distinguere alcune categorie di olio.
L'olio proveniente dalla Liguria Sud si separa nettamente dagli altri campioni, caratterizzandosi per concentrazioni superiori di acido Palmitoleico e Linoleico, a fronte di livelli inferiori di acido Oleico.
Anche l'olio proveniente dalla Liguria Ovest mostra un profilo che si distingue nettamente dalle altre categorie. In questo caso, le variabili che ne determinano la separazione sono l'acido Stearico, che presenta valori molto alti, e l'acido Eicosanoico, che si presenta invece valori bassi.

---

2) *Quali variabili da sole (diagonale) mostrano una distribuzione non normale che possa indicare la presenza di gruppi ? quali categorie sono distinguibili? Quali scatter plots (extradiagonali) mettono meglio in luce le diverse categorie o alcune di esse rispetto alle altre? Individuati un paio di scatter plot significativi fate (solo per questi) uno scatter hist. Ci sono variabili correlate (quali)?*

Grafico

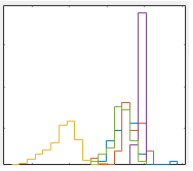


Risposta

> *Quali variabili da sole (diagonale) mostrano una distribuzione non normale che possa indicare la presenza di gruppi? Quali categorie sono distinguibili?*

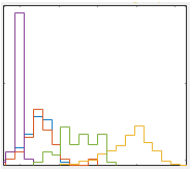
Una distribuzione non normale si può notare in diversi acidi (più evidenti):

1. Acido Oleico



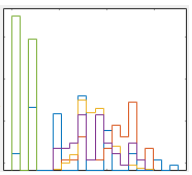
Distribuzione bimodale, con un primo picco della Puglia del Sud nei valori bassi e un secondo picco nei valori alti dei restanti gruppi. È ben distinguibile la Puglia del Sud che, da sola, ha valori molto più bassi di questo acido rispetto alle altre categorie di olio.

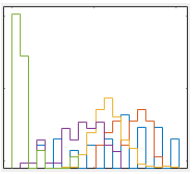
2. Acido Linoleico



Descrizione praticamente identica alla oleica con l'unica differenza che il grafico è specchiato.

3. Acido Eicosanoico e Linolenico



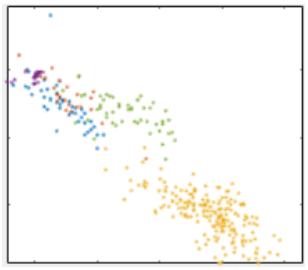


Distribuzione bimodale, con un primo picco della Liguria Ovest nei valori bassi e un secondo picco nei valori alti dei restanti gruppi. È ben distinguibile la Liguria Ovest che, da sola, ha valori molto più bassi di questo acido rispetto alle altre categorie di olio.

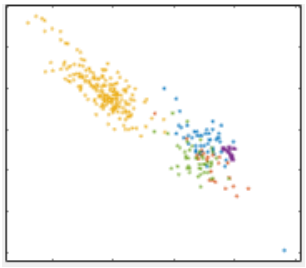
> *Quali scatter plots (extradiagonali) mettono meglio in luce le diverse categorie o alcune di esse rispetto alle altre? Individuati un paio di scatter plot significativi fate (solo per questi) uno scatter hist. Ci sono variabili correlate (quali)?*

Alcuni scatter plots che mettono meglio in luce le diverse categorie sono le seguenti combinazioni:

- Acido oleico e linoleico

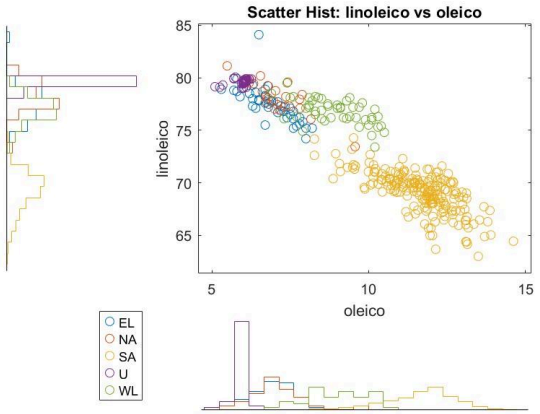


- Acido palmitico e oleico



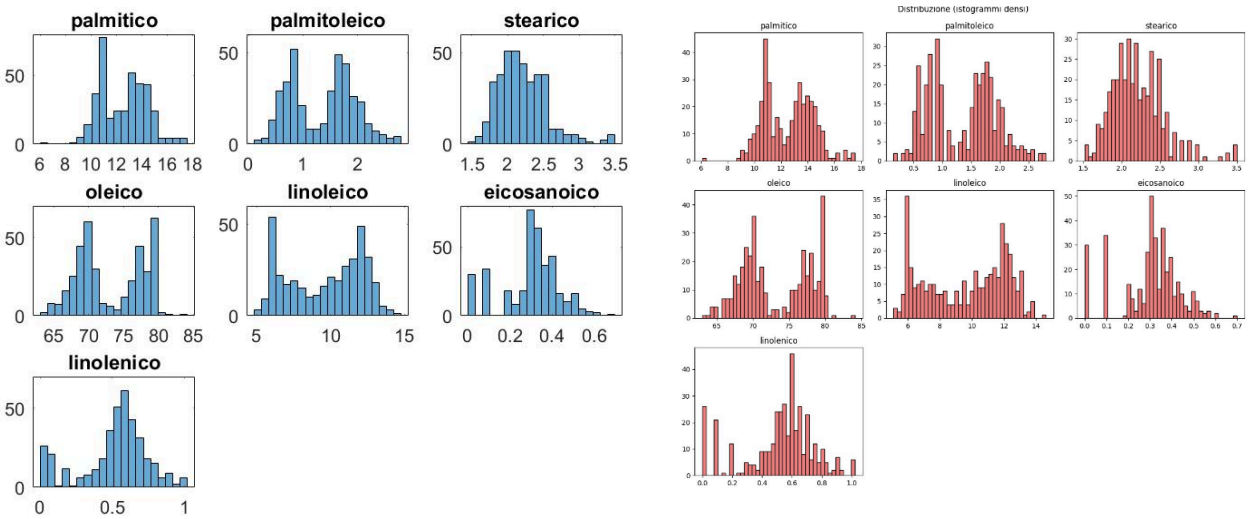
I restanti scatter plots, per la maggior parte, mettono in luce in modo significativo la categoria del Sud della Puglia, rendendola ben separata ed evidente rispetto alle altre (un'eccezione si ha ad esempio nei grafici della riga dello stearico, nelle ultime due colonne).

Isolando un caso (acido Oleico e Linoleico) e creando uno scatter hist, si evidenzia una netta correlazione negativa tra le due: all'aumentare dell'Oleico, Linoleico diminuisce. Le regioni si separano chiaramente in cluster omogenei e distinti : la Puglia del sud forma un gruppo nettamente isolato in basso a destra, mentre tutte le altre si distribuiscono separatamente con valori bassi di Oleico e alti di Linoleico.



3) *Commentate le distribuzioni delle variabili e l'eventuale presenza di valori estremi.*

Grafico



Risposta

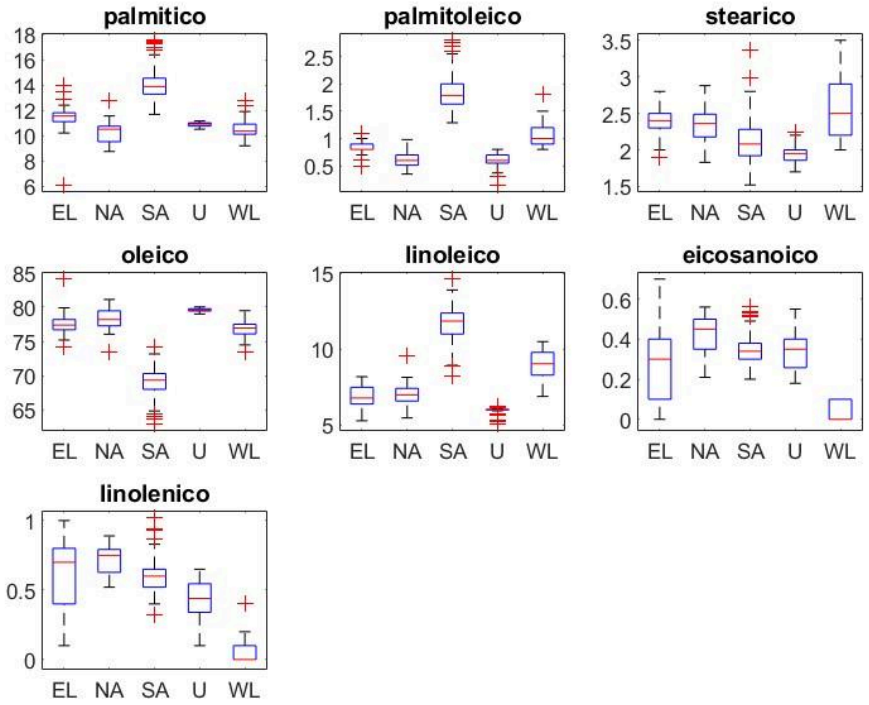
- Acido Palmitico: distribuzione bimodale con due picchi distinti (il primo da 8 a 12 e il secondo da 12 a 18). È presente un valore estremo a 6.
- Acido Palmitoleico: distribuzione bimodale con due picchi distinti (il primo da 0 a 1 e il secondo da 1 a 3) suggerendo la presenza di due gruppi distinti. Non ci sono valori estremi.
- Acido Stearico: distribuzione unimodale (picco principale tra 2 e 2.25) con un asimmetria a destra. I valori finali verso 3.5, mostrano la presenza di valori estremi.
- Acido Oleico: distribuzione bimodale con due picchi distinti (il primo da 65 a 75 e il secondo da 75 a 85) che forma quindi due gruppi distinti. Non ci sono valori estremi.
- Acido Linoleico: distribuzione bimodale con due picchi distinti (il primo da 0 a 8 e il secondo da 8 a 15) che forma quindi due gruppi distinti. Non ci sono valori estremi.
- Acido Eicosanoico: distribuzione unimodale asimmetrica a sinistra. Non ci sono valori estremi.
- Acido Linolenico: distribuzione bimodale con due picchi distinti (il primo da 0 a 0.25 e il secondo da 0.25 a 1). Non ci sono estremi.

  In generale, le distribuzioni di Palmitoleico, Oleico e Linoleico sono quelle che suggeriscono maggiormente la presenza di categorie distinte. In particolare, queste tre variabili si confermano le più utili per la distinzione tra i gruppi grazie alla presenza di più mode ben riconoscibili (ad esempio, il doppio picco in Linoleico e Palmitoleico).

---

4) *Quali variabili mostrano di poter distinguere le diverse categorie o alcune di esse?*

Grafico



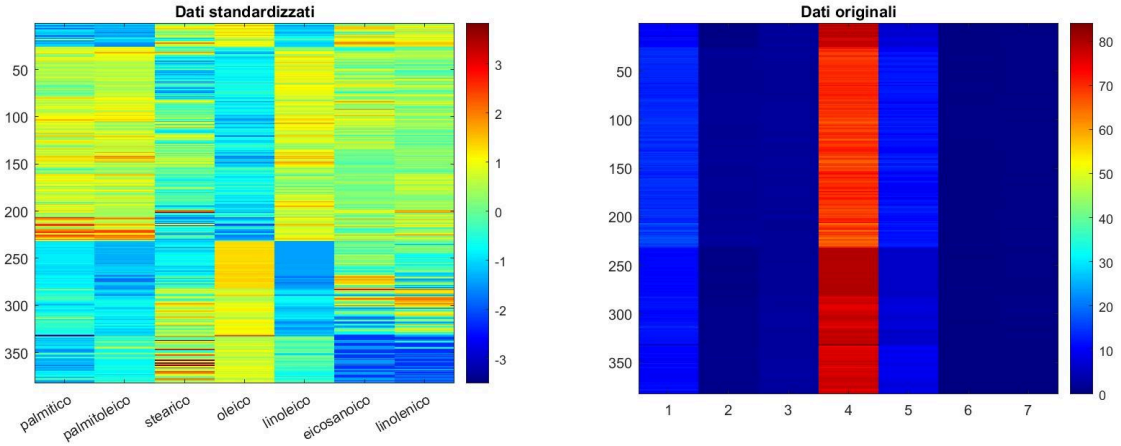
Risposta

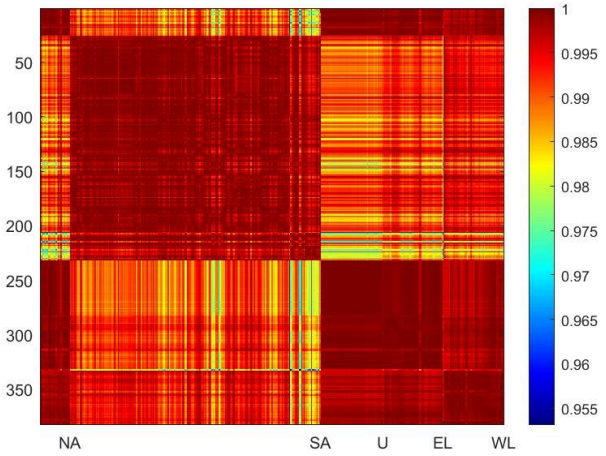
Le variabili più efficaci nel distinguere le categorie principali sono Palmitoleico, Oleico e Linoleico.
L'acido Palmitoleico e Linoleico isolano in modo netto la Puglia del sud: questa presenta valori molto alti in entrambi i casi, creando un'asimmetria rispetto alle restanti categorie, che si raggruppano su valori inferiori.
Per quanto riguarda l'acido Oleico, mostra una correlazione negativa con il caso precedente. Per questa variabile, infatti, tutte le categorie si posizionano su valori alti, ad eccezione della Puglia del sud (che presenta valori bassi). Importante notare che per l'Umbria si ha una distribuzione molto compatta, segno di oli simili.
Infine, l'acido Eicosanoico separa nettamente la Liguria ovest (caratterizzata da valori bassi) dagli altri gruppi.

---

5) *Sono utili queste rappresentazioni per avere un'idea della somiglianza tra campioni (righe della matrice)? Quale di più? Commentando le somiglianze che riuscite ad individuare.*

Grafico





Risposta

La Heatmap sulle variabili autoscalate è molto più efficace per la valutazione rispetto ai dati originali, in quanto la normalizzazione evidenzia meglio le differenze relative tra i campioni. Il plot è estremamente utile per giudicare la somiglianza tra campioni e individuare gruppi affini: righe di colore uniforme indicano oli dalle proprietà chimiche simili. Facendo un confronto per l'interezza dei campioni, si nota che le variabili Palmitico, Palmitoleico e Linoleico hanno una forte somiglianza. Tramite questa rappresentazione è possibile anche distinguere i vari gruppi: Oleico, Palmitoleico e Linolenico mostrano la maggiore variabilità relativa (fasce di colori accesi come giallo/rosso o blu intenso). Infine, è facile cogliere anche le correlazioni osservando i pattern di colori opposti (ad esempio Oleico alto associato a Linoleico basso).
La Matrice di Correlazione al Quadrato è una delle rappresentazioni più chiare per visualizzare la somiglianza e la distinzione tra tutti i campioni. Nonostante non mostri le singole variabili più discriminanti, è evidente che i campioni di una stessa categoria hanno una fortissima similarità reciproca, formando blocchi di colore rosso brillante lungo la diagonale. Le categorie degli oli Umbri e della Liguria Ovest risultano particolarmente omogenee internamente. La separazione netta tra i blocchi indica che l'insieme delle variabili utilizzate è altamente efficace nel distinguere le categorie di olio.