

PROGETTO FINALE GRUPPO T

Dataset: oliveoil

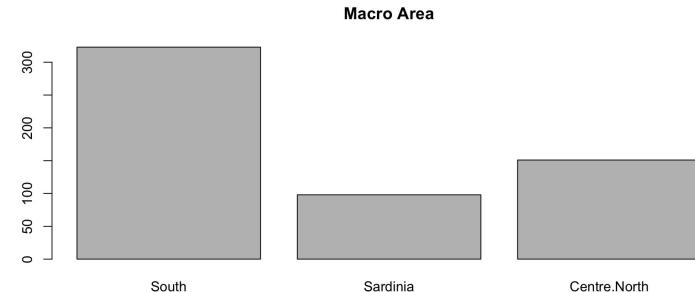
DataSet

oliveoil

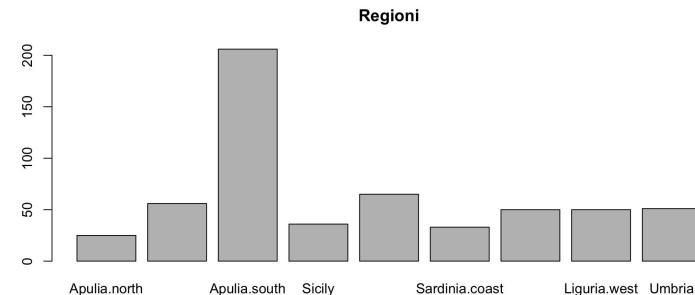
macro.area	region	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
South	Apulia.north	0.10755698	0.007596961	0.02269092	0.7820872	0.06727309	0.003698521	0.006097561	0.002998800
South	Apulia.north	0.10885646	0.007397041	0.02249100	0.7706917	0.07816873	0.003198721	0.006197521	0.002998800
South	Apulia.north	0.09116353	0.005497801	0.02469012	0.8110756	0.05497801	0.003198721	0.006397441	0.002998800
South	Apulia.north	0.09665167	0.005797101	0.02408796	0.7949025	0.06196902	0.005097451	0.007896052	0.003598201
South	Apulia.north	0.10515794	0.006797281	0.02598960	0.7768892	0.06727309	0.005097961	0.008096761	0.004698121
South	Apulia.north	0.09117265	0.004998500	0.02689193	0.7922623	0.06787964	0.005198440	0.007097871	0.004498650
South	Apulia.north	0.09228154	0.006698660	0.02649470	0.7989402	0.06188762	0.004999000	0.005698860	0.002999400
South	Apulia.north	0.11005598	0.006197521	0.02359056	0.7725910	0.07347061	0.003998401	0.006497401	0.003598561
South	Apulia.north	0.10824588	0.006096952	0.02398801	0.7742129	0.07096452	0.004697651	0.008395802	0.003398301
South	Apulia.north	0.10382076	0.005601120	0.02140428	0.7946589	0.06341268	0.002700540	0.005301060	0.003100620
South	Apulia.north	0.10513692	0.003597841	0.02198681	0.7974215	0.06056366	0.002198681	0.006596042	0.002498501
South	Apulia.north	0.10366890	0.005998201	0.02359292	0.7866640	0.06618015	0.003099070	0.006298111	0.004498650
South	Apulia.north	0.10746776	0.007097871	0.02149355	0.7726682	0.07477757	0.005098470	0.007997601	0.003398980
South	Apulia.north	0.08757373	0.005298410	0.02439268	0.8016595	0.06558033	0.004198740	0.007997601	0.003299010
South	Apulia.north	0.09526190	0.004998001	0.02548980	0.7792883	0.07806877	0.005097961	0.007596961	0.004198321

Analisi dati

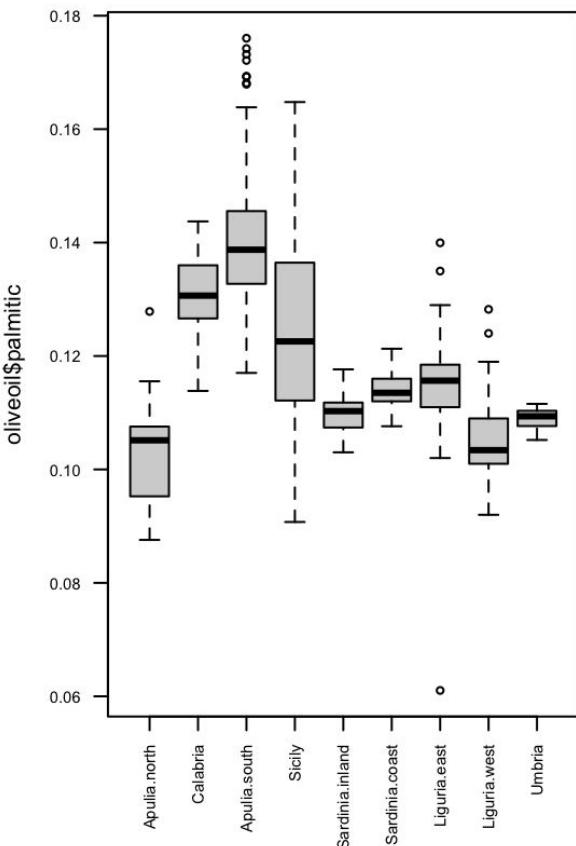
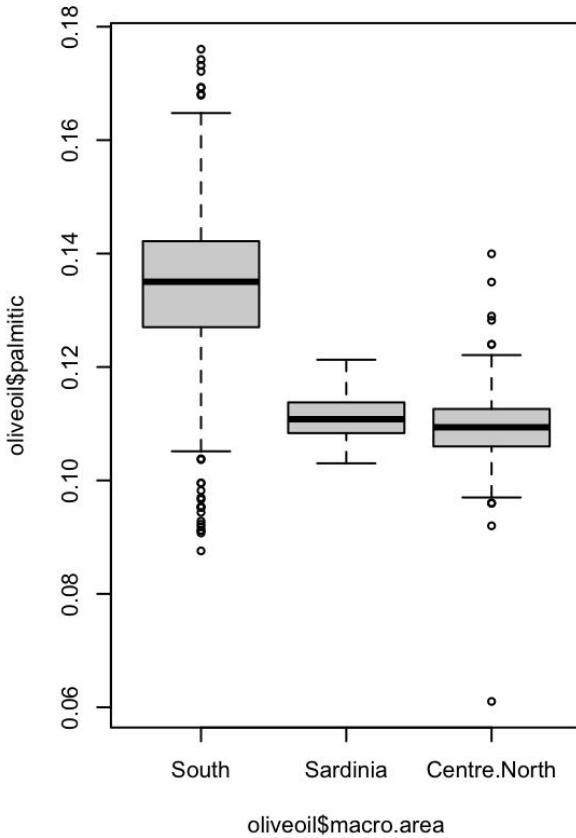
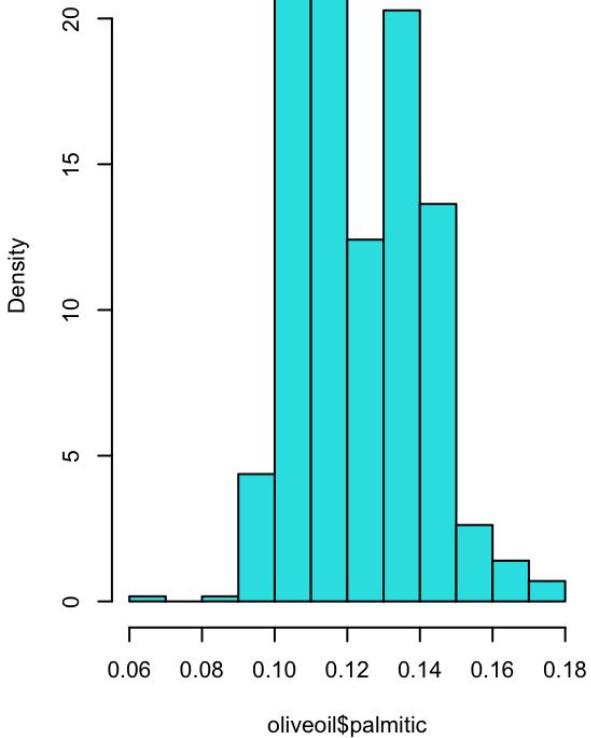
Analisi Univariata: Istogrammi per visualizzare la distribuzione delle concentrazioni di ciascun acido grasso.



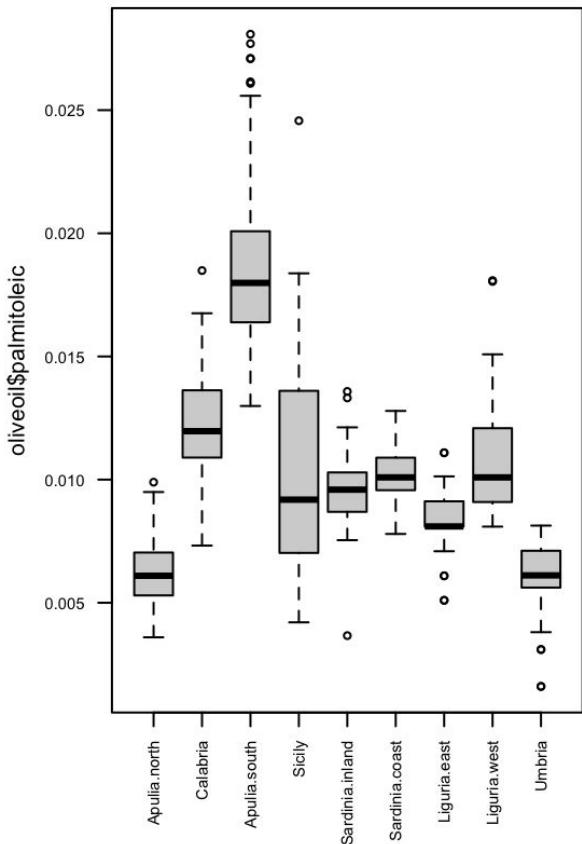
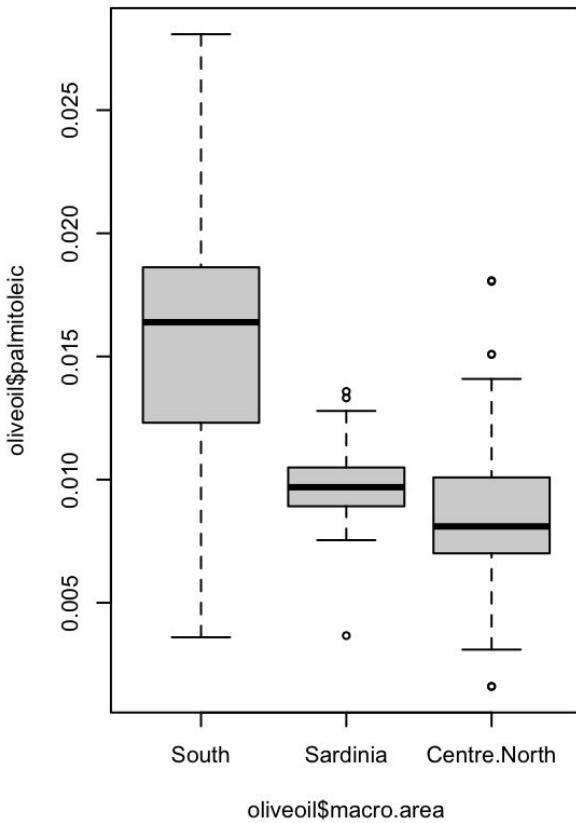
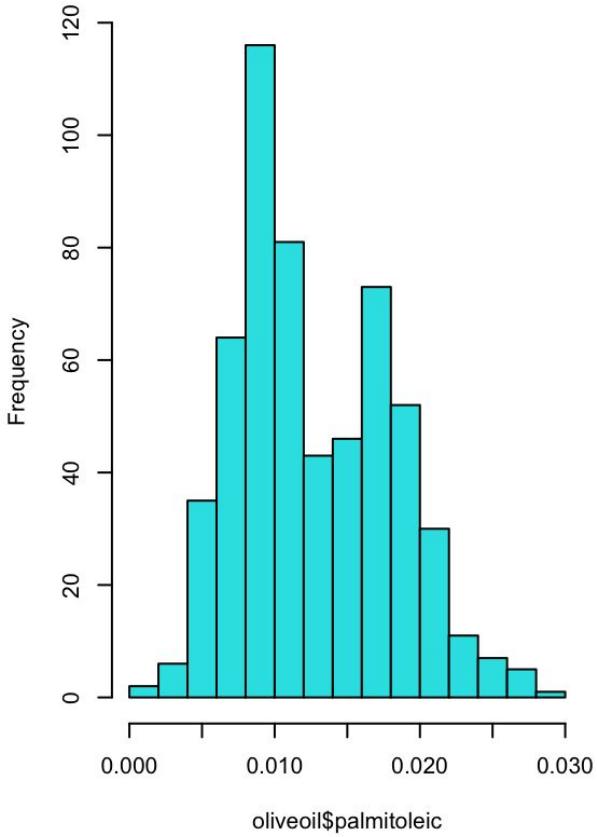
Analisi Bivariata: I boxplot per confrontare le concentrazioni degli acidi grassi tra diverse macro aree.

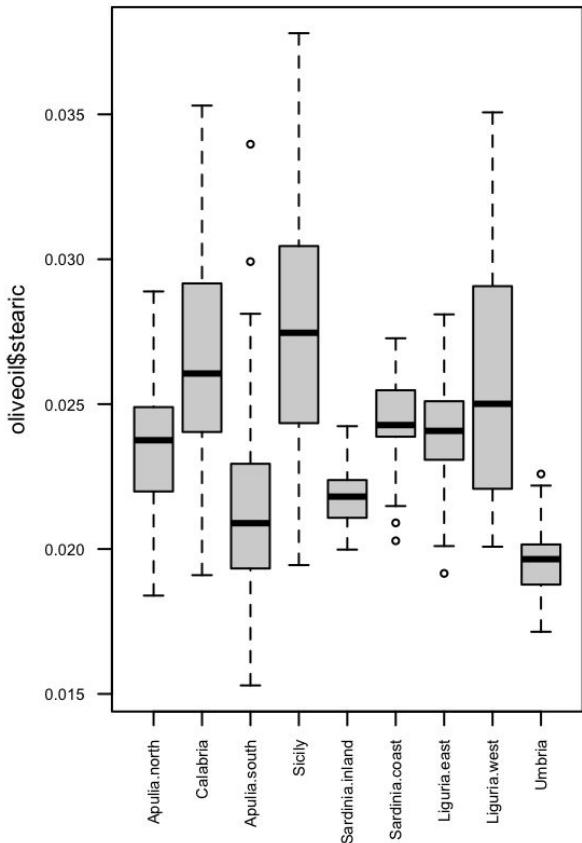
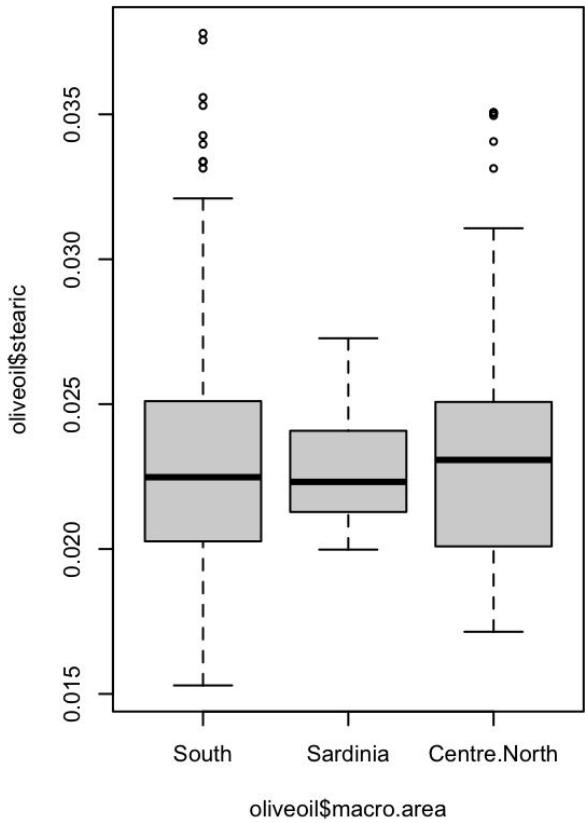
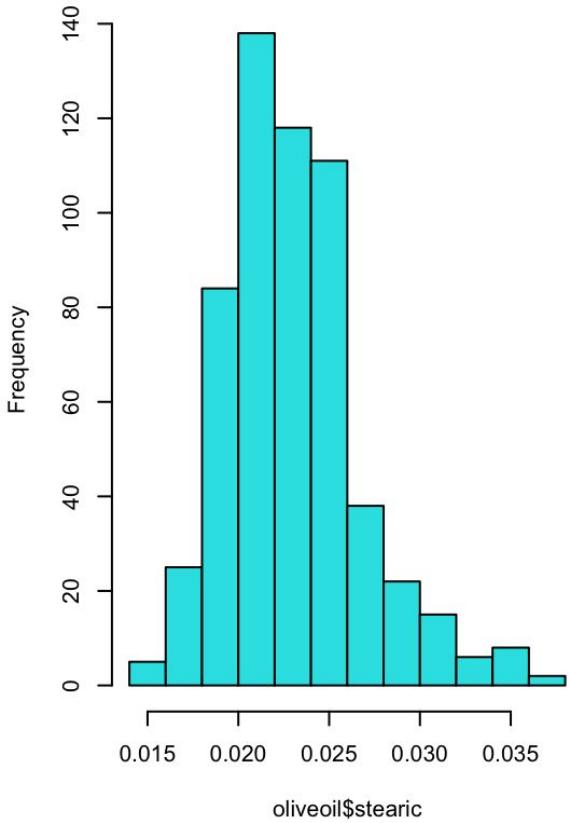


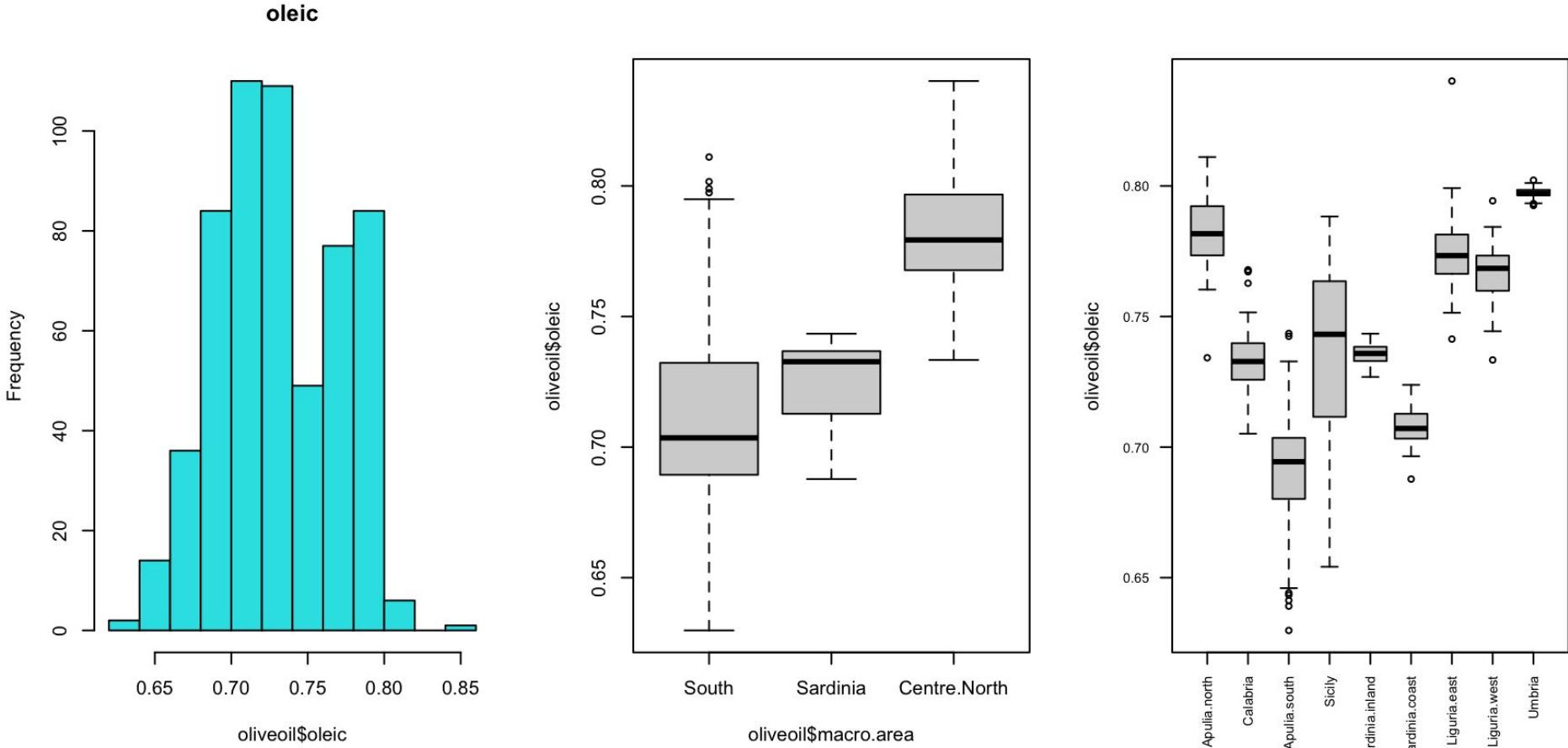
palmitic



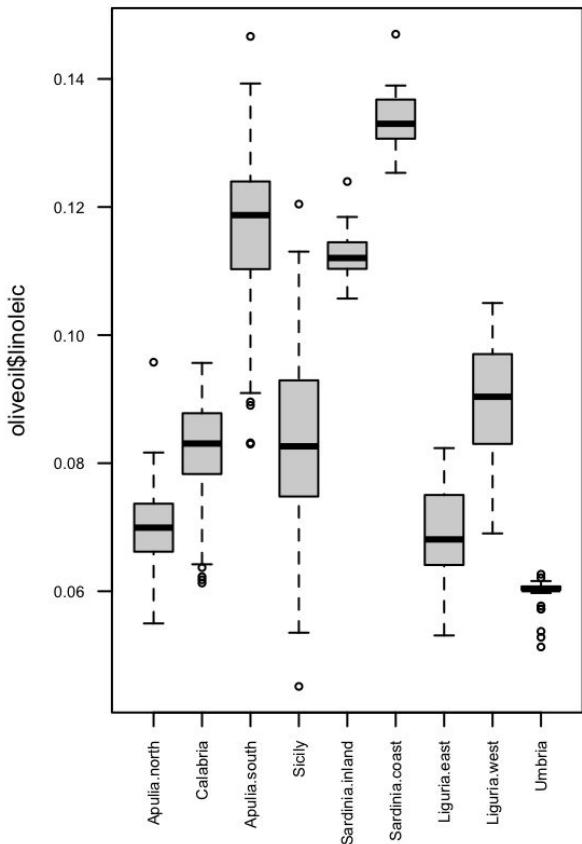
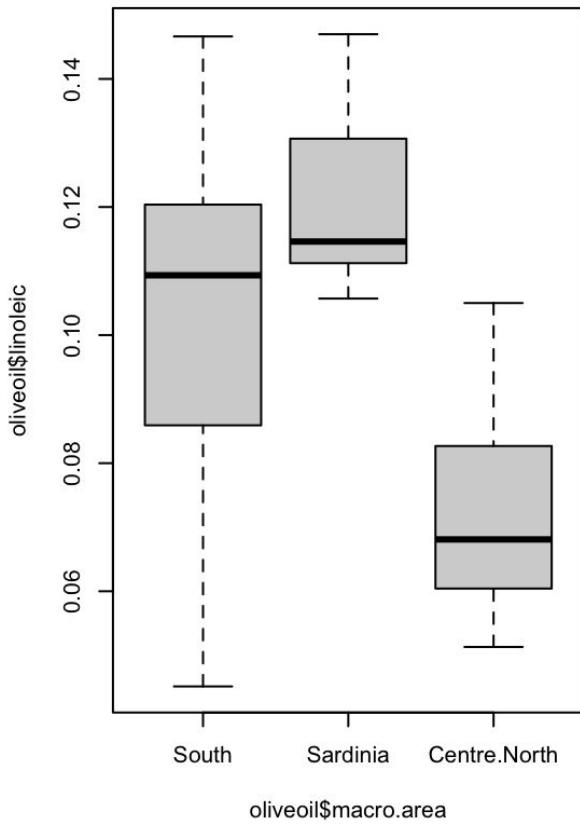
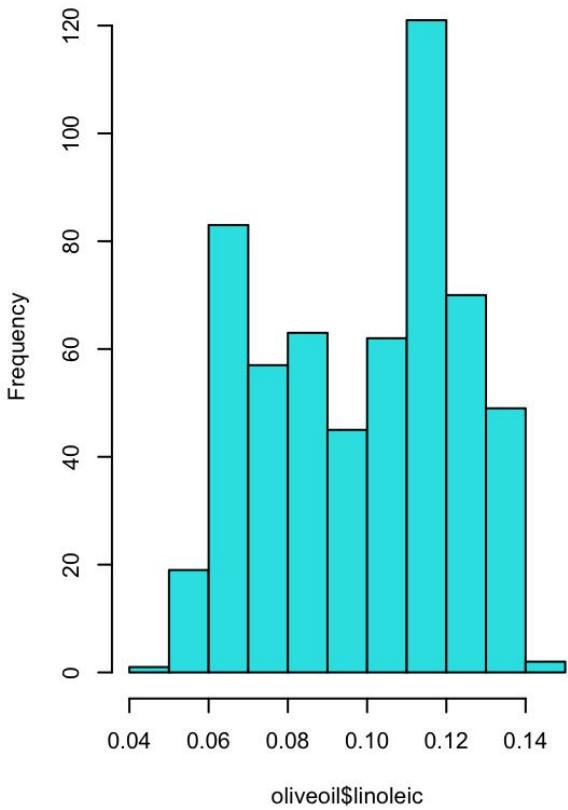
palmitoleic



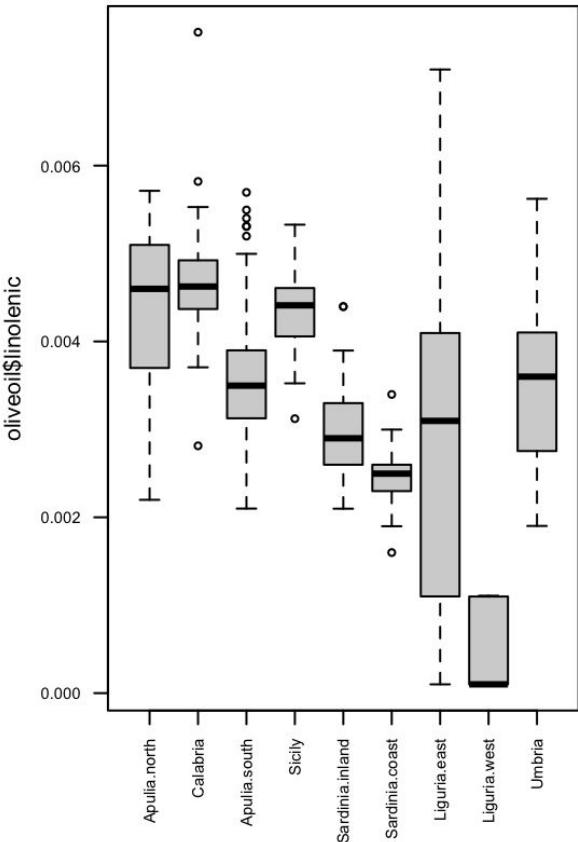
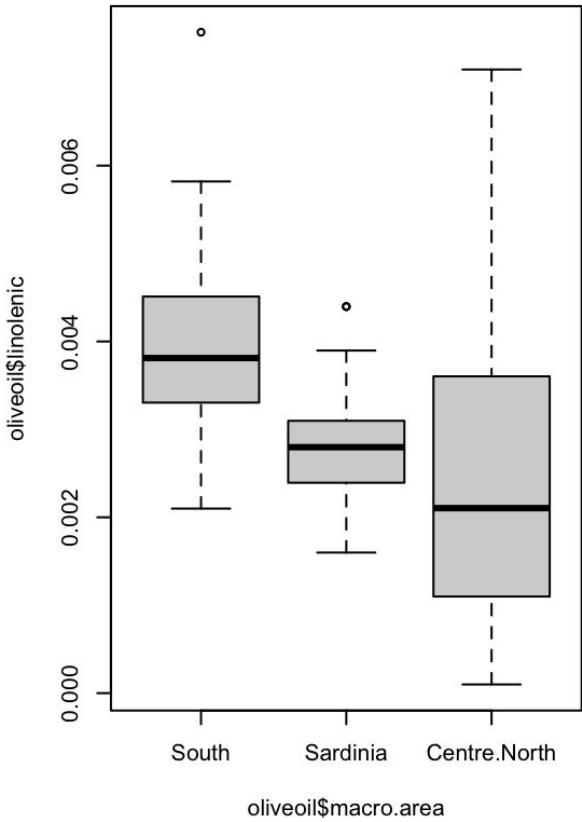
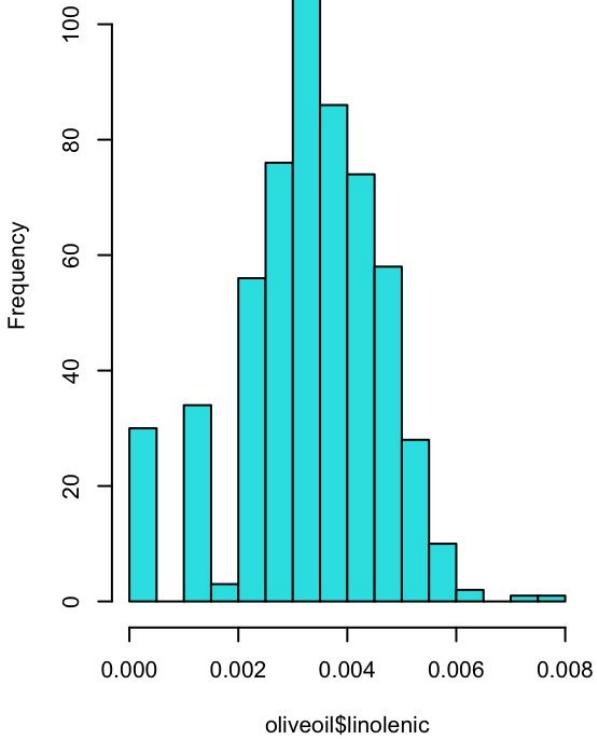
stearic



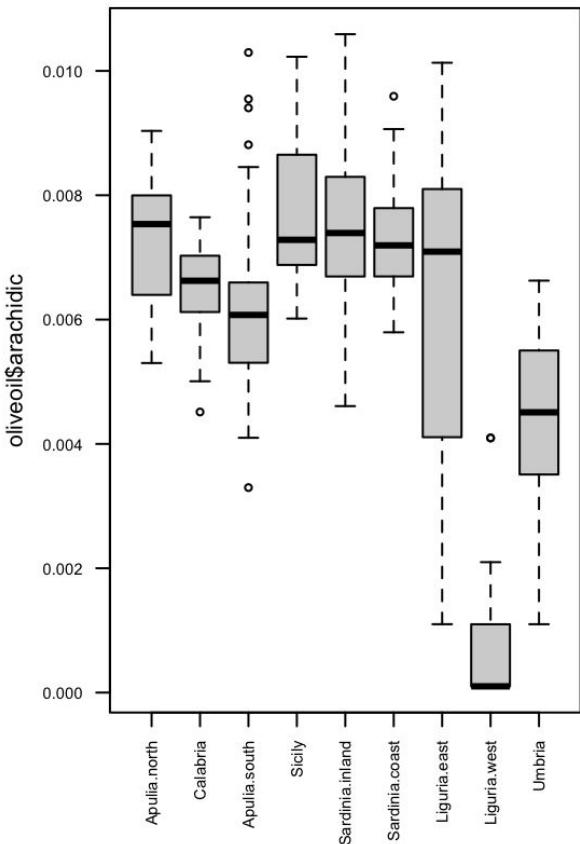
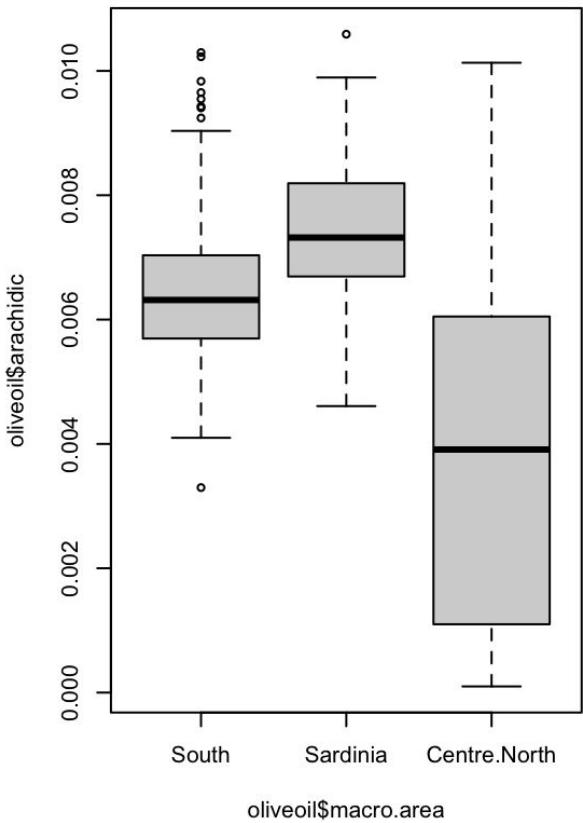
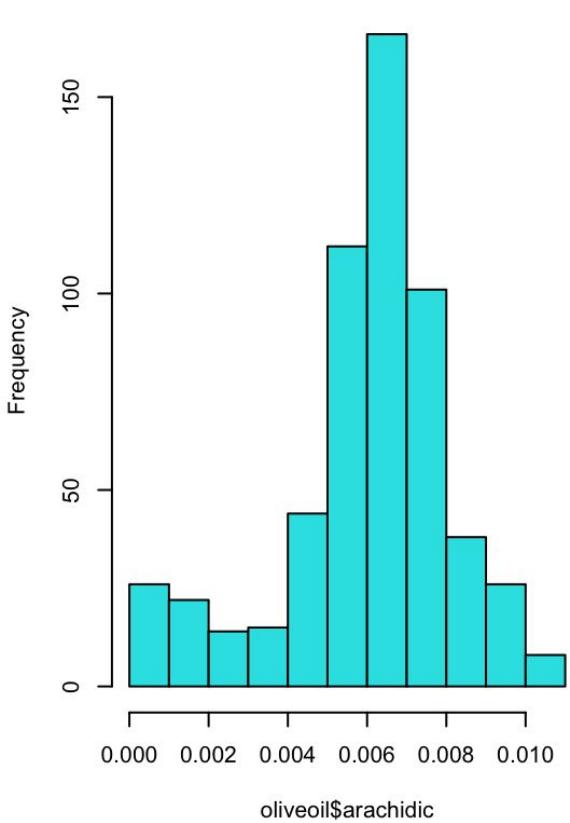
linoleic



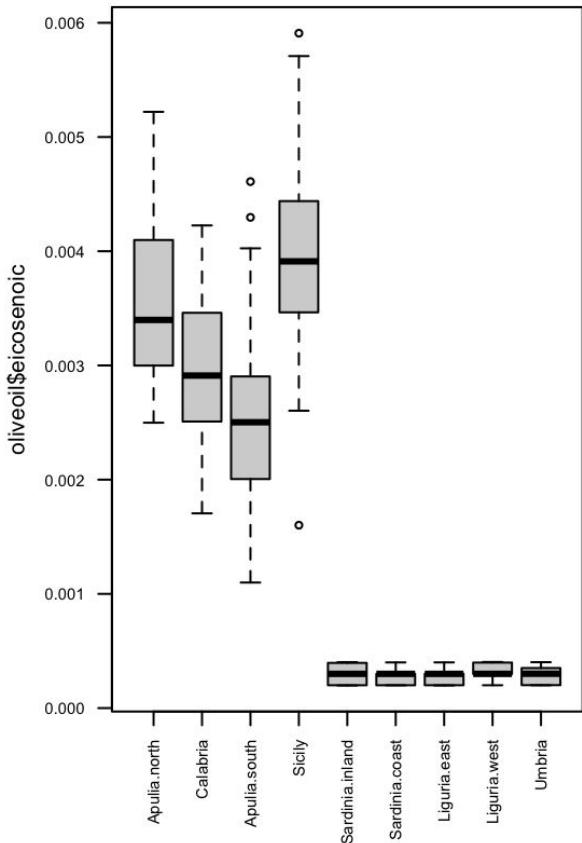
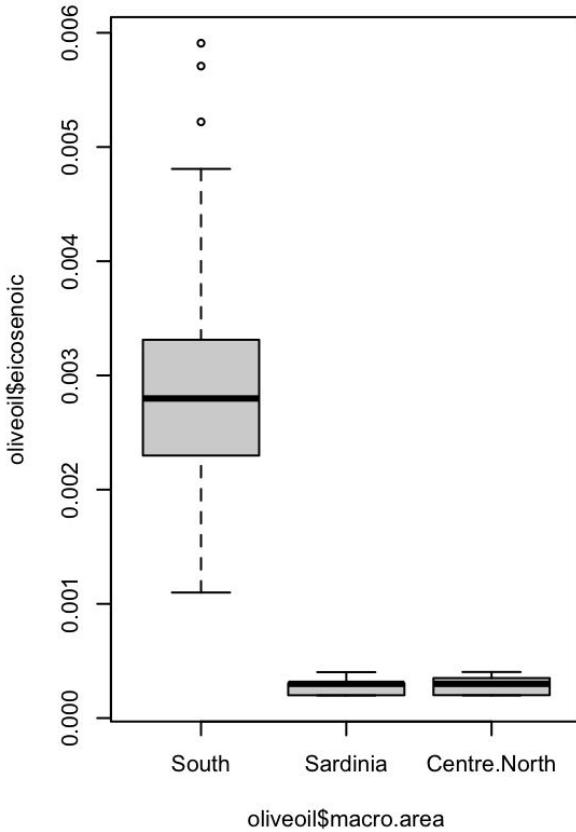
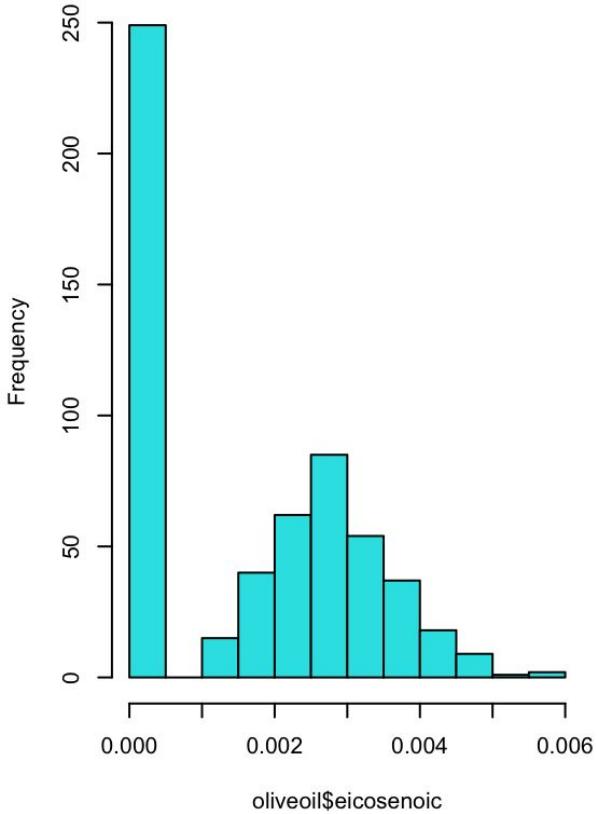
linolenic



arachidic



eicosenoic



Operazioni sui dati

Presenza di zeri dovuti alle misurazioni al di sotto del livello di sensibilità degli strumenti con i quali è stata effettuata l'analisi.

$$y_j = \frac{x_j + 1}{\sum_{j=1}^d (x_j + 1)}$$

Compositional data

I dati compositivi sono dati quantitativi che descrivono una parti di un insieme unitario.

$$S^d = \{x \in \mathbb{R}^d | x_i > 0, \sum_{i=1}^d x_i = k\}$$

ALR Transform

$$alr : S^d \subset \mathbb{R}^d \rightarrow \mathbb{R}^{d-1}$$

$$\text{alr}(x) = \left[\log \frac{x_1}{x_D} \dots \log \frac{x_{D-1}}{x_D} \right]$$

Cluster analysis

Analisi dei cluster

1. K-means
2. PAM
3. DBSCAN
4. PdfCluster

Indice ARI

1. Macro aree
2. Regioni

K-MEANS

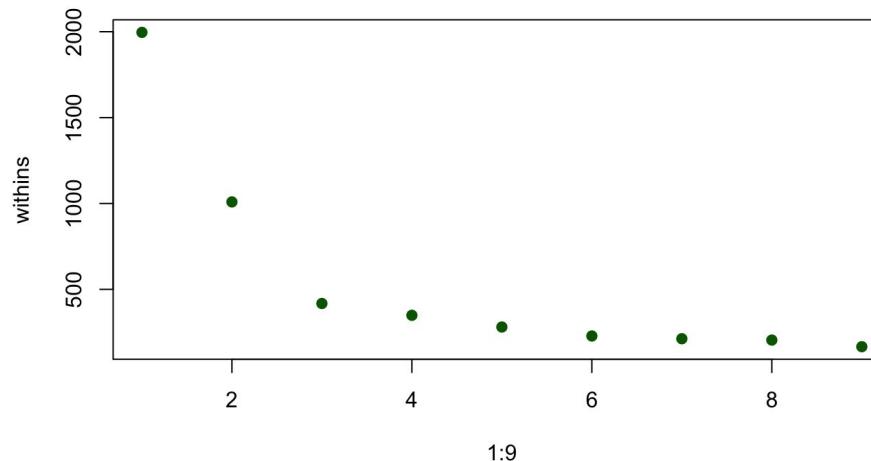
K-MEANS

1. Si scelgono K punti casuali diversi dai punti del dataset che sono i centroidi dei cluster
2. Si associa ogni dato al centroide più vicino
3. Per ogni gruppo si trova il punto medio che diventerà il nuovo centroide di quel gruppo
4. Itero dal punto 2 fino a quando nessun dato cambia gruppo tra un'iterazione e l'altra

Scelta di K

tot.withinss : somma delle distanze tra ogni punto e il centroide del cluster a cui è associato

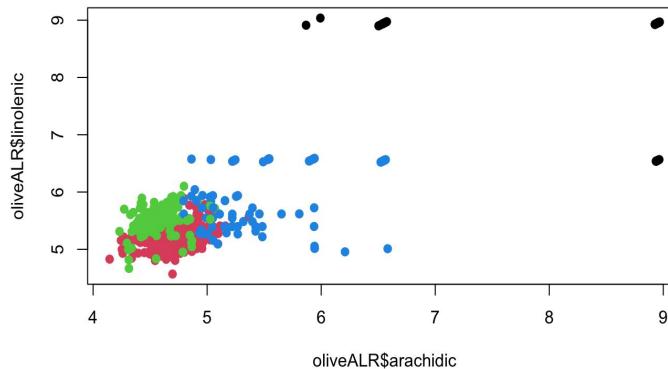
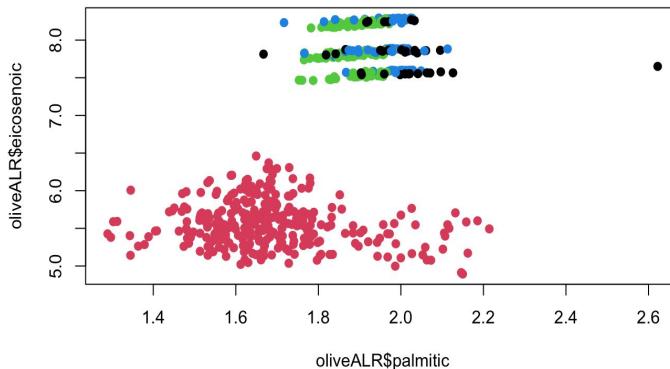
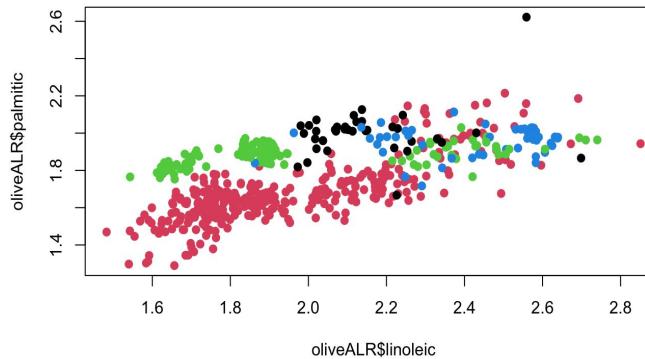
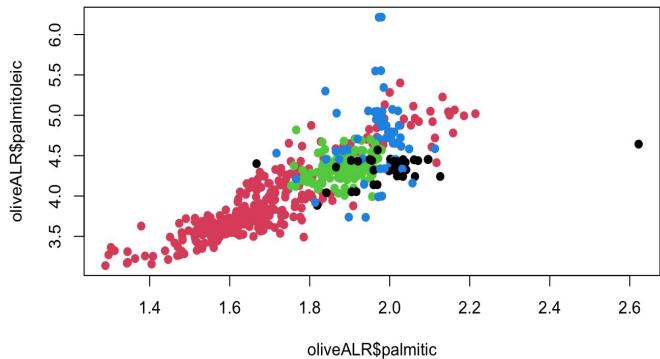
Elbow method :
4 cluster



```
km.out <- kmeans(oliveALR[,3:9], centers = 4, nstart = 15)
```

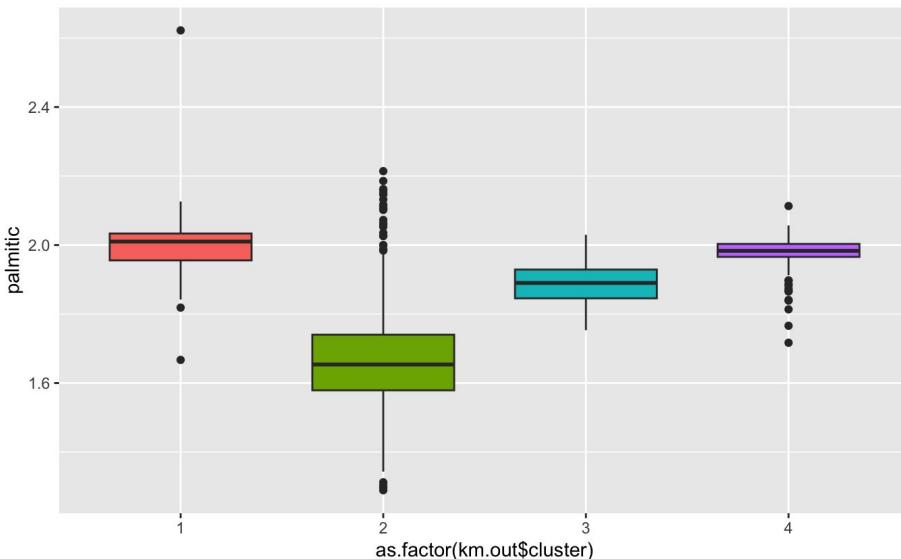
Cluster	1	2	3	4
Size	37	323	132	80

Cluster e correlazione

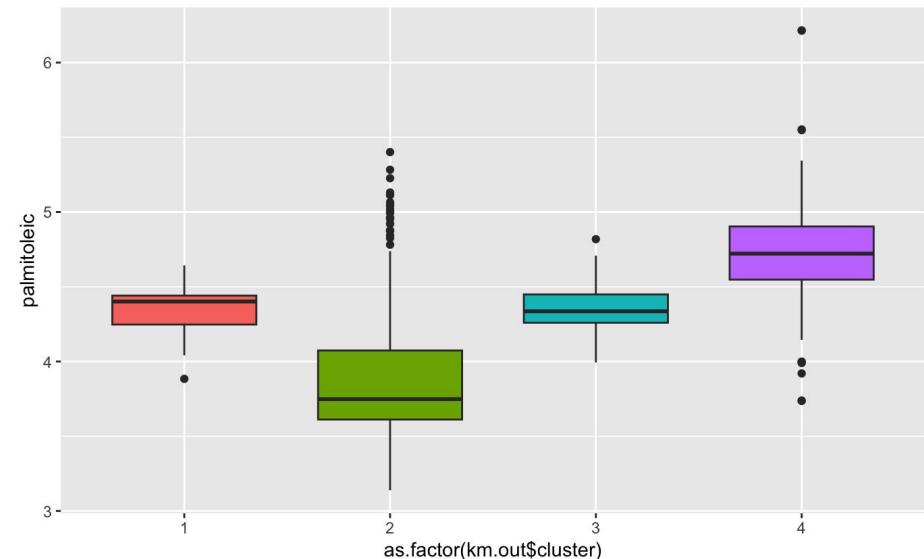


- Acidi fortemente Correlati
- Cluster 1 e 3 simili,
- Numero di outlier e variabilità maggiore nel cluster 2

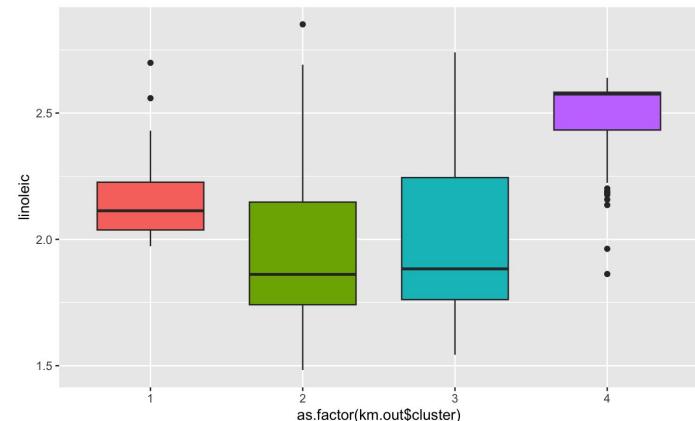
Acido Palmitico



Acido Palmitoleico

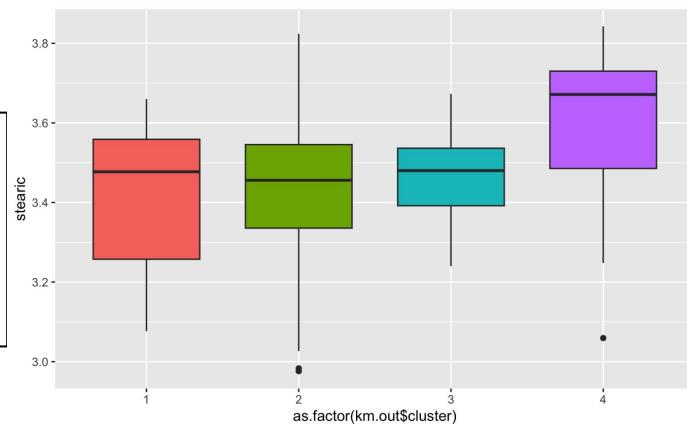


Acido Linoleico



- Cluster 2 e 3 presentano valori simili e una variabilità più elevata
- Cluster 4 contiene outlier

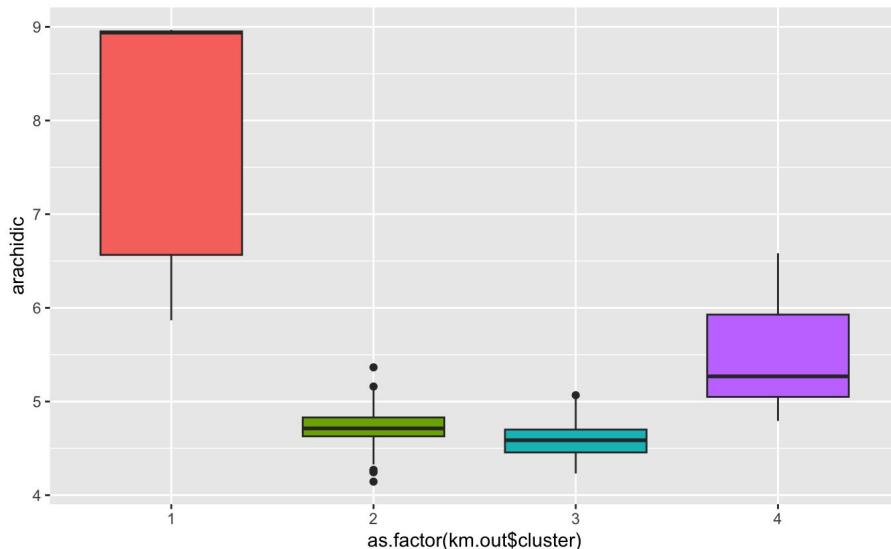
Acido Stearico



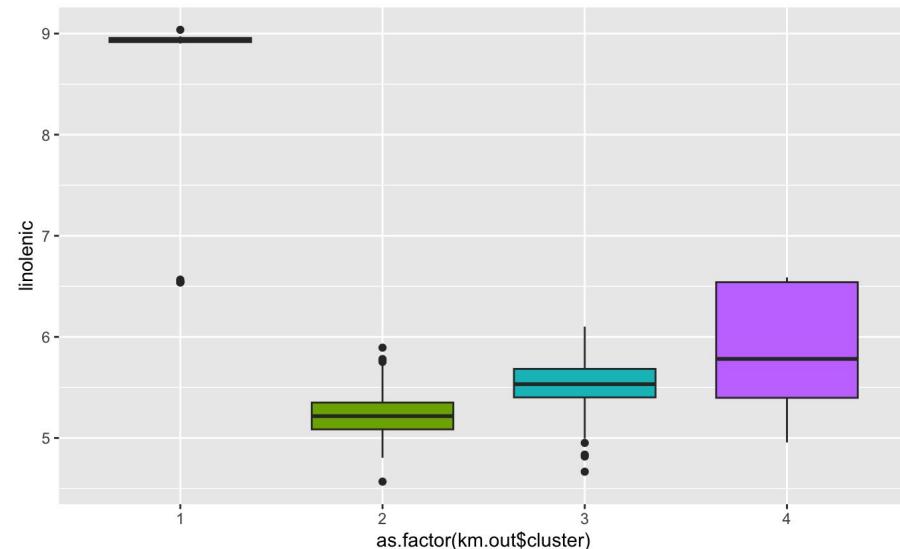
- Cluster 1, 2 e 3 hanno mediane vicine tra loro
- Cluster 4 ha una mediana di 3.67

- Cluster 1 contiene tutti i valori nulli di questi acidi
- Varianza del cluster 4 più elevata dei cluster 2 e 3
- Cluster 2 e 3 contengono una distribuzione e una mediana simili

Acido Arachidico

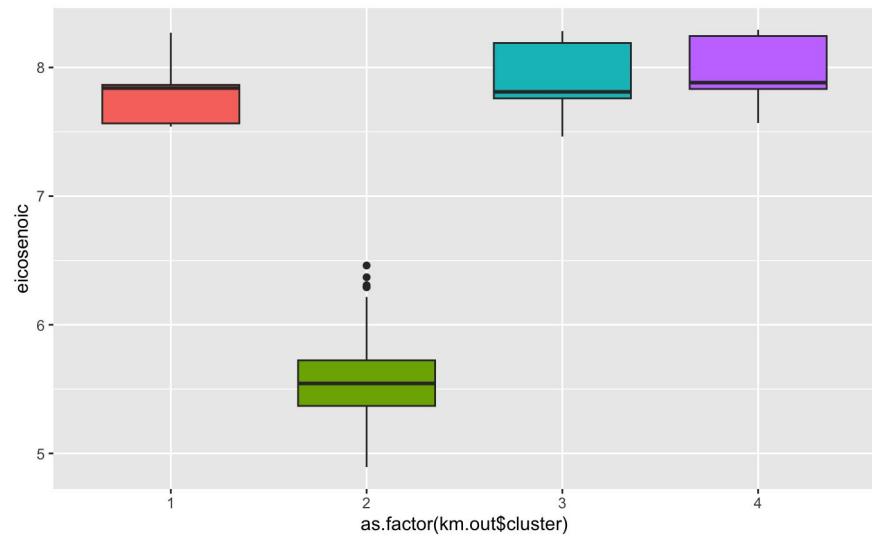


Acido Linolenico



Acido Eicosenoico

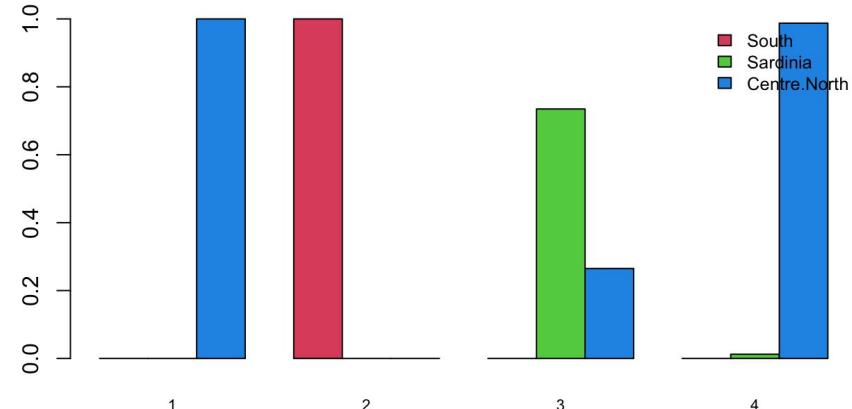
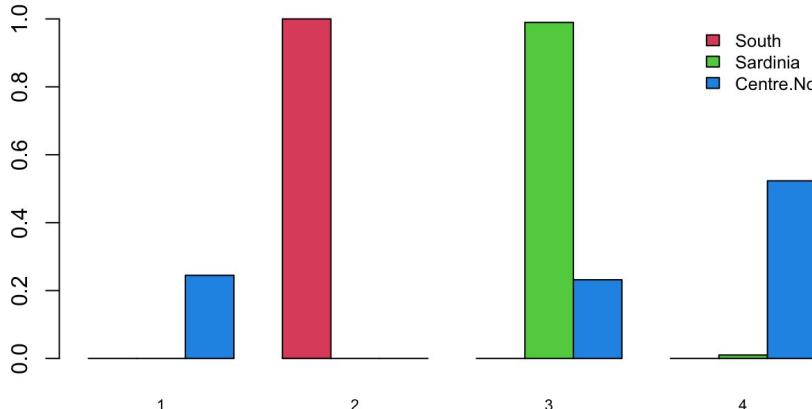
- Oli con percentuali molto basse di acido Eicosenoico sono raggruppati sotto i cluster 1, 3 e 4
- Gli oli del cluster 2 sono notevolmente separati dagli altri raggruppamenti, in quanto contengono una percentuale di acido eicosenoico più elevata
- Cluster 3 e 4 presentano una distribuzione simile



- Gli oli del sud si trovano al 100% nel cluster 2
- Gli oli della Sardegna al 99% nel cluster 3
- Gli oli del centro nord sono divisi tra i cluster

	1	2	3	4
South	0.000	1.000	0.000	0.000
Sardinia	0.000	0.000	0.990	0.010
Centre.North	0.245	0.000	0.232	0.523

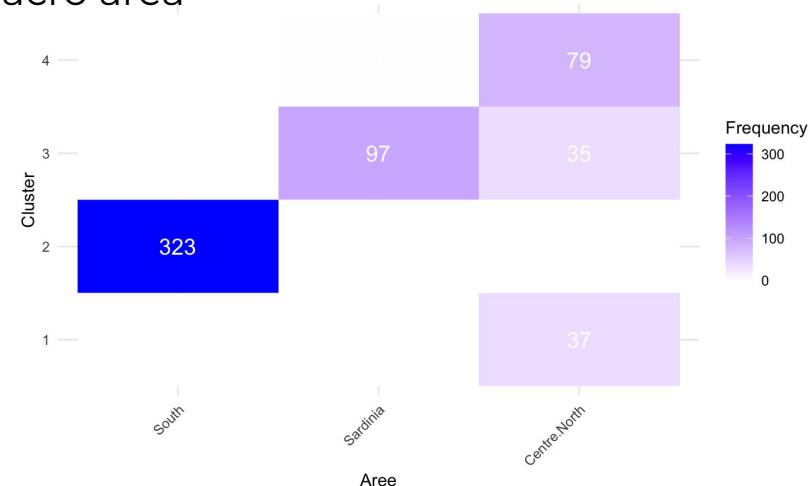
Poporzione all'interno dei cluster



ARI e Confusion Matrix

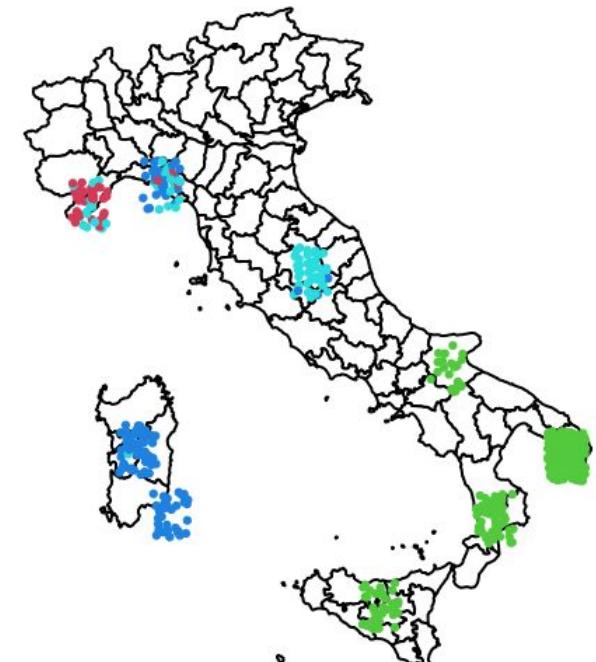
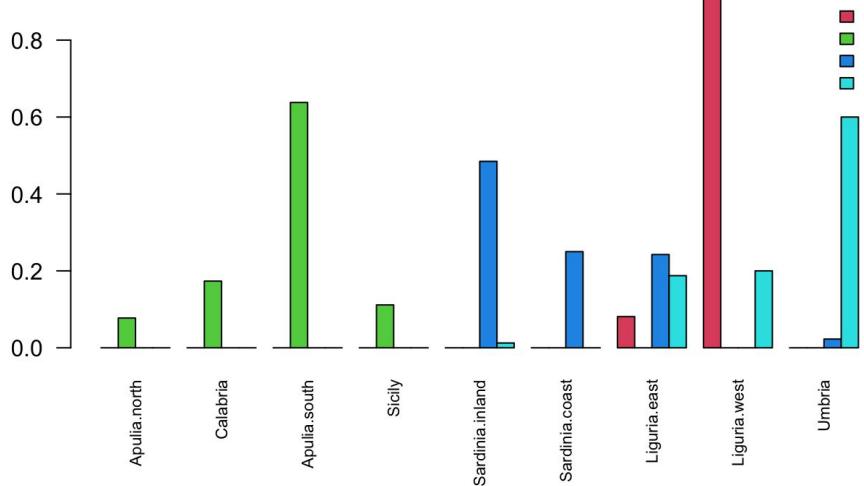
Adjusted Rand Index: si confrontano i cluster con le macro aree.
Si ottiene un indice di 0.8660622

Confusion Matrix: si vedono le frequenze assolute di ogni cluster,
condizionato alle macro area



- Il cluster 2 ha individuato con molta precisione gli oli del sud
- Liguria est contiene oli con le qualità più eterogenee
- Il cluster 1 contiene principalmente oli della liguria ovest

Poporzione all'interno dei cluster





PAM

PAM

1. Si selezionano k punti (**medoidi**) nel dataset e si associa ogni punto al medoide più vicino
2. Si selezionano in modo casuale nuovi candidati come medoidi
3. Si calcola tot.withins con i nuovi centri, e se la nuova distanza è minore della precedente allora i centri diventano i nuovi medoidi
4. Si itera dal punto 2 fino a quando non ci sono cambiamenti nell'insieme di medoidi.

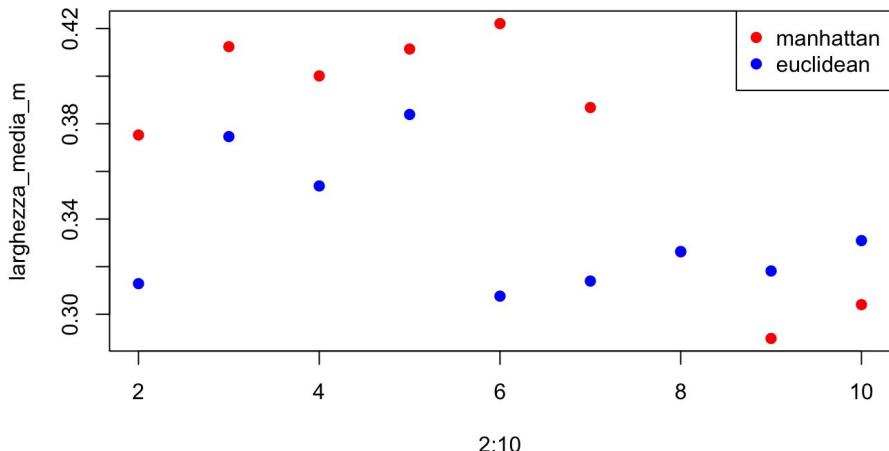
Scelta di K

Distanza euclidea : $d_E(p, q) = \left(\sum_{i=1}^n (p_i - q_i)^2 \right)^{\frac{1}{2}}$

Distanza di Manhattan : $d_M(p, q) = \sum_{i=1}^n |p_i - q_i|$

Si testa l'algoritmo con diverse distanze e diversi valori di k

6 cluster, distanza euclidea

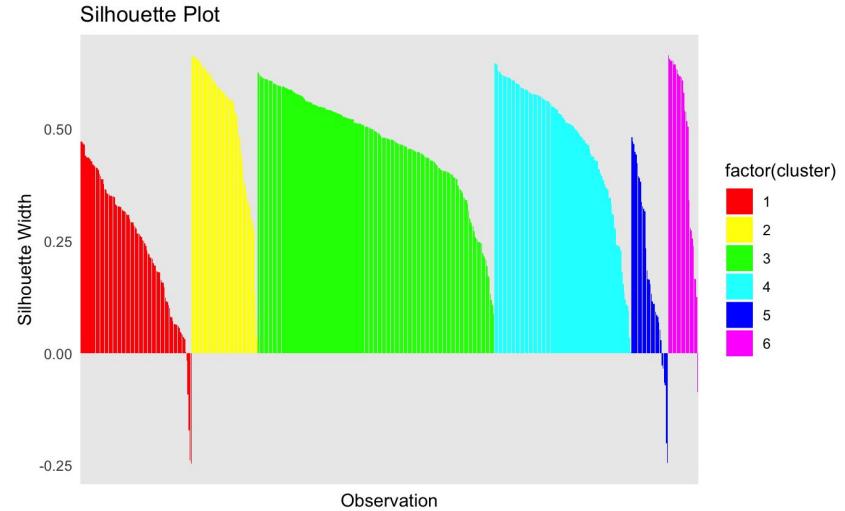


Cluster

```
pam.out<-pam(oliveALR[,3:9], 6, metric="manhattan", stand=TRUE, nstart = 10)
```

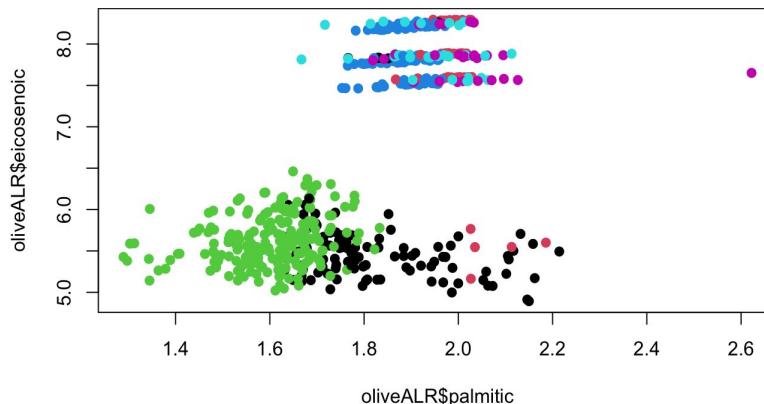
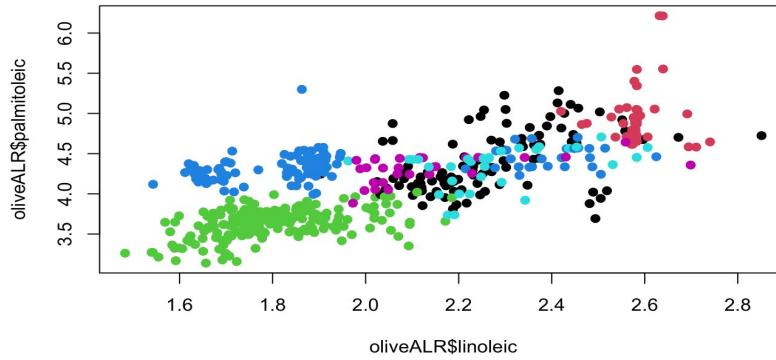
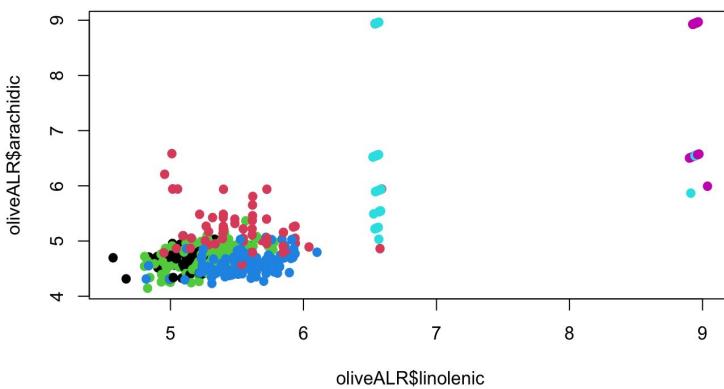
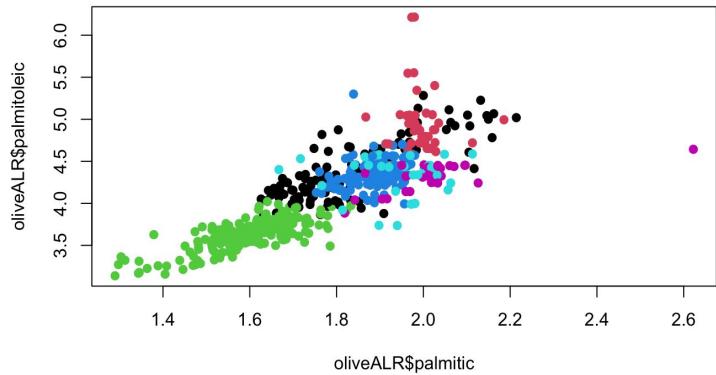
Silhouette : si nota che

- Cluster 3 ha valori tutti molto distanti dagli altri gruppi, in quanto l'indice cala drasticamente
- Cluster 5 e 6 contengono meno elementi rispetto agli altri
- Cluster 1 e 5 contengono valori inseriti nel cluster incorrecto



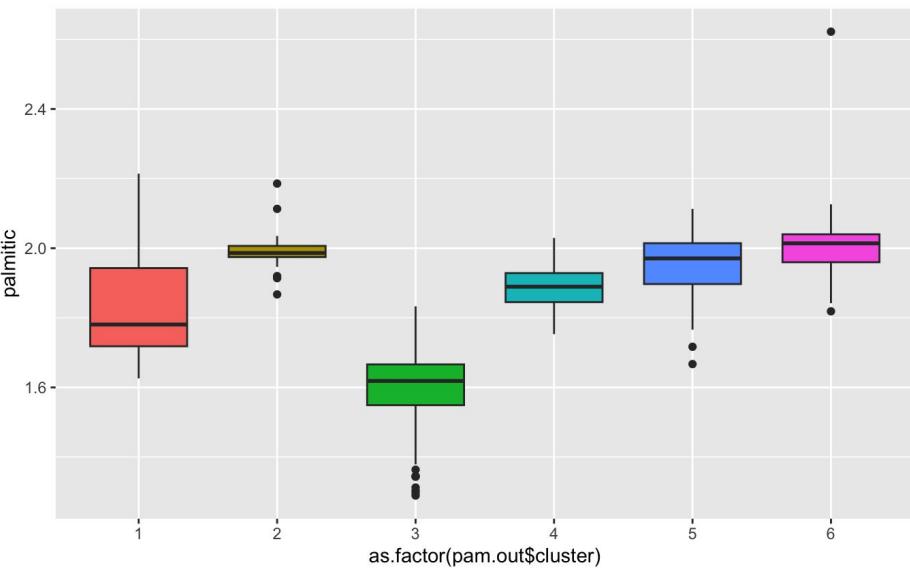
Cluster	1	2	3	4	5	6
Size	103	61	219	127	34	28

Cluster e correlazione

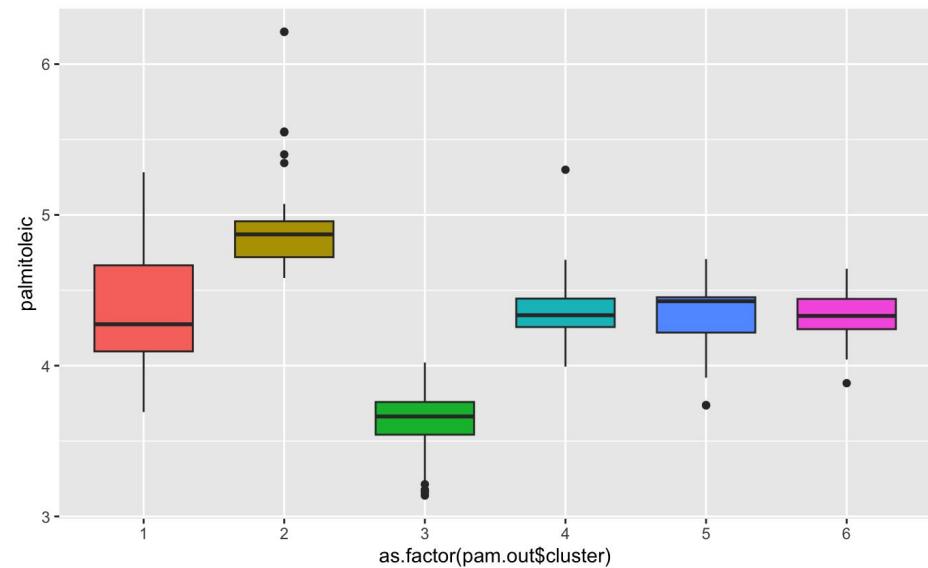


- Cluster 3 contengono la concentrazione di acido maggiore
- Cluster 4, 5 e 6 hanno mediana simile
- Cluster 1 ha variabilità più elevata

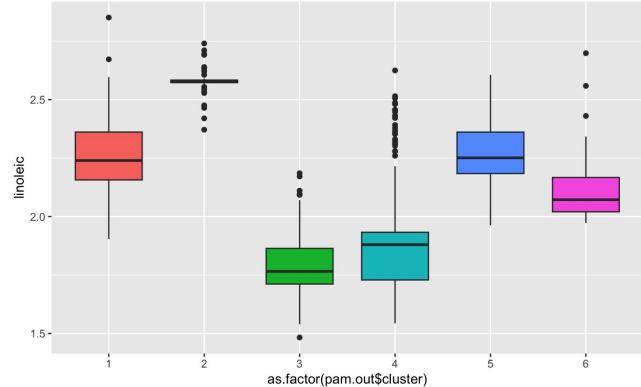
Acido Palmitico



Acido Palmitoleico

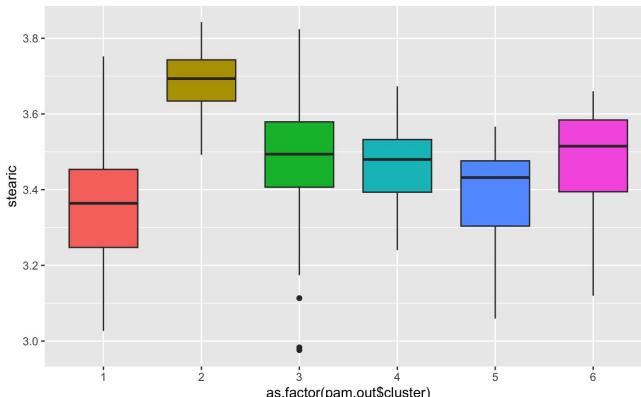


Acido Linoleico



- Devianza fra i gruppi alta
- Cluster 3 e 4 hanno mediane vicine, il cluster 4 presenta un discreto numero di valori outlier nel range 2.25-2.5.
- I cluster 1, 5 e 6 hanno la mediana nel range 2-2.25.

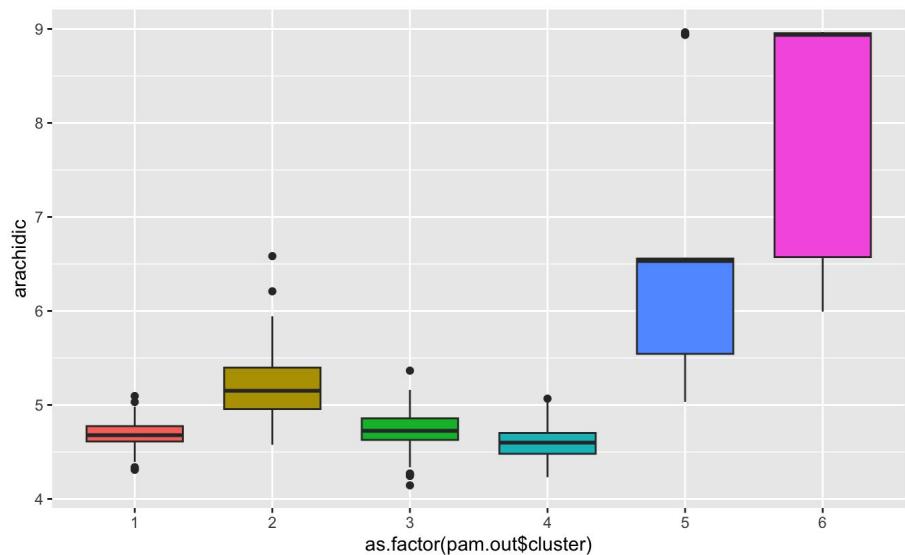
Acido Stearico



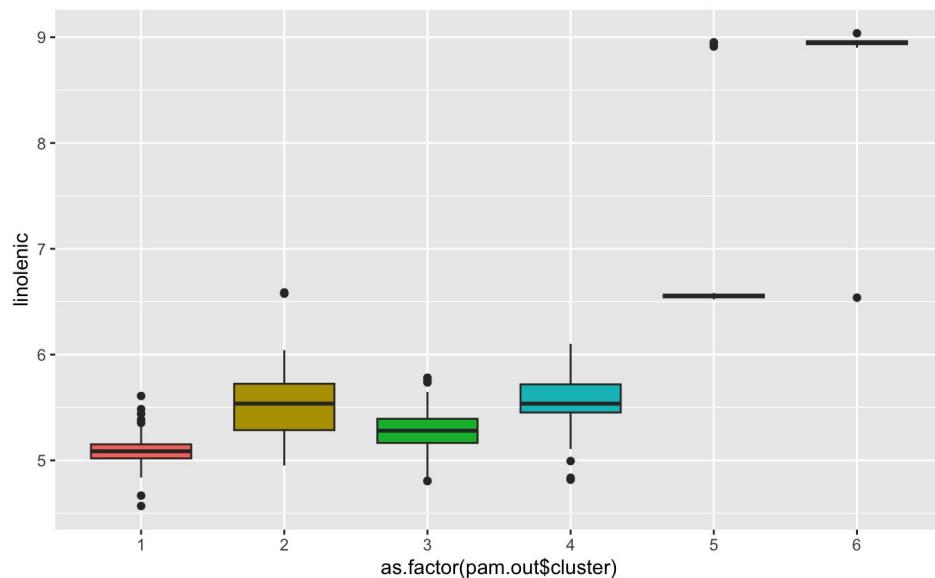
- Cluster 3,4,5,6 hanno una mediana vicina
- Cluster 1 e 2 si discostano significativamente

- I cluster 5 e 6 contengono tutti i valori di acido arachidico e linolenico pari a zero
- I primi 4 cluster presentano delle mediane simili
- Entrambi i cluster 5 ha una mediana di circa 6.5

Acido Arachidico

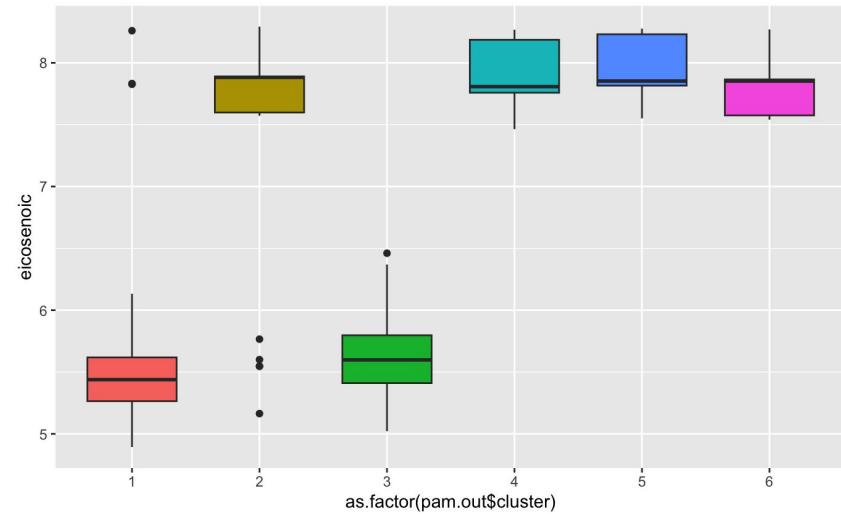


Acido Linolenico



Acido Eicosenoico

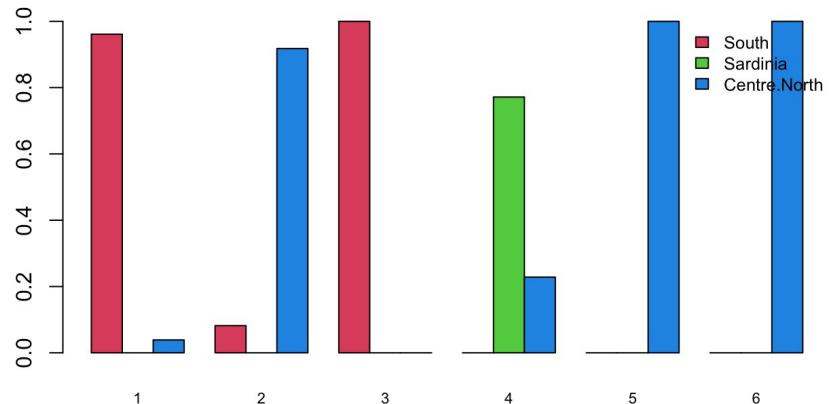
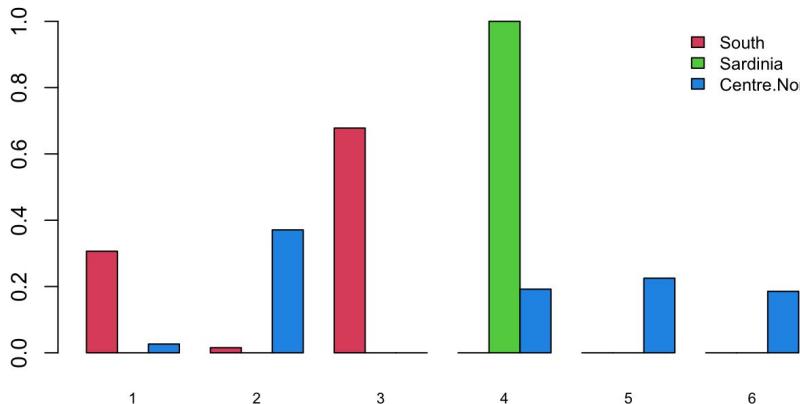
- Cluster 1 e 2 contengono valori alti di acido eicosenoico
- Si notano similarità nei cluster
 - 4 e 5
 - 1 e 3
 - 2 e 6
- Le distribuzioni dei cluster 1 e 2 presentano alcuni valori outlier



- Gli oli della Sardegna vengono inseriti interamente nel cluster 4.
- Gli oli del centro nord invece non sono ben identificati
- Cluster 1, 2, 3, 5 e 6 omogenei

	1	2	3	4	5	6
South	0.307	0.015	0.678	0.000	0.000	0.000
Sardinia	0.000	0.000	0.000	1.000	0.000	0.000
Centre.North	0.026	0.371	0.000	0.192	0.225	0.185

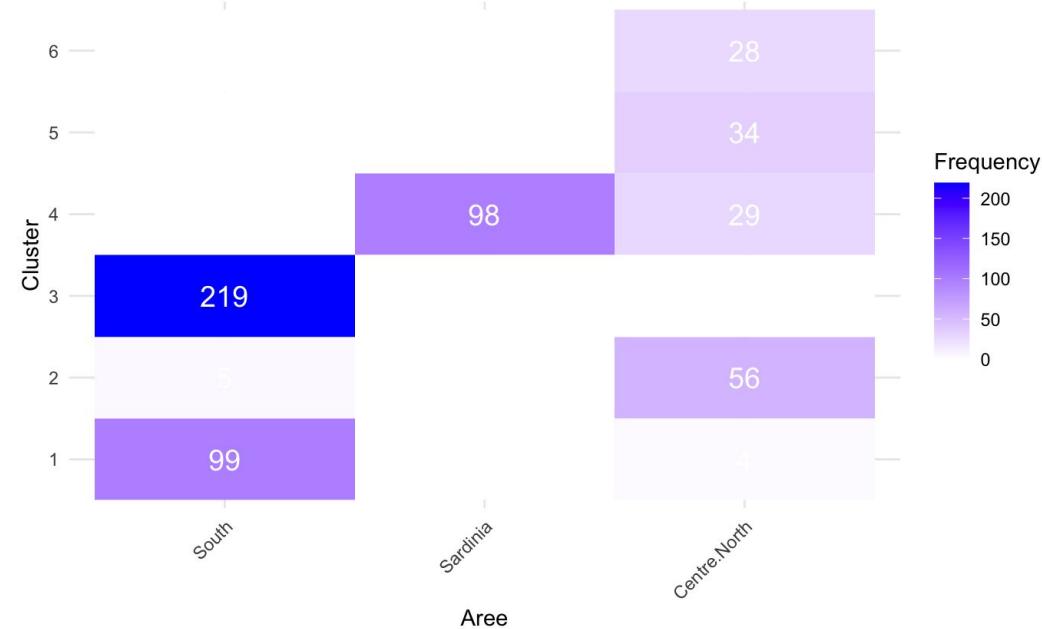
Poporzione all'interno dei cluster



ARI e Confusion Matrix

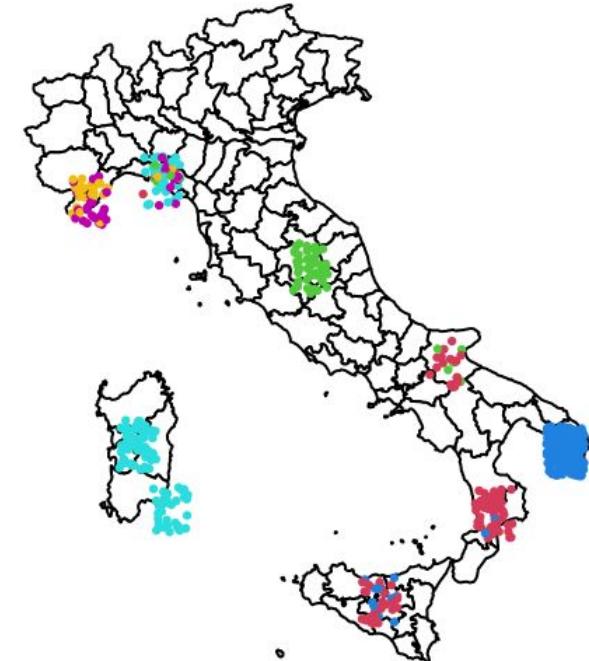
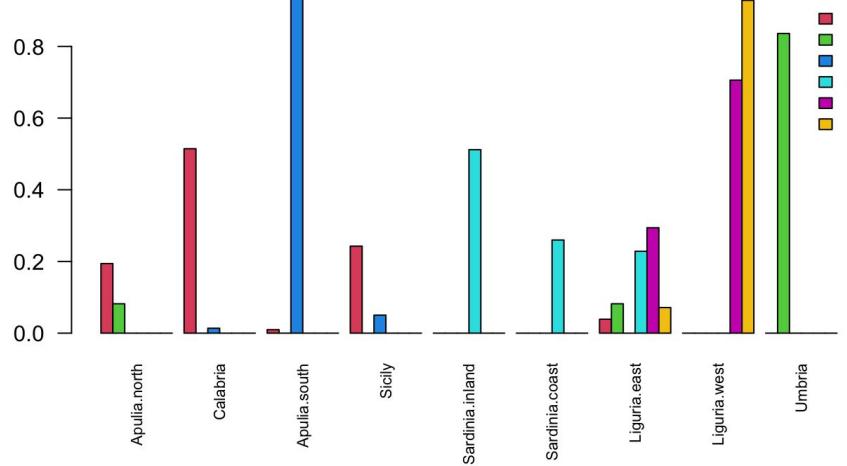
Adjusted Rand Index: Si confrontano i cluster con le macro aree.
Si ottiene un indice di 0.5285497

Confusion Matrix: si vedono le frequenze assolute di ogni cluster, condizionato alle macro area



- Oli della puglia sud vengono inseriti interamente nel cluster 3
- Oli dell'umbria vengono inseriti interamente nel cluster 2
- Oli della liguria est sono divisi tra i cluster
- il cluster 5 e 6 contiene esclusivamente oli della liguria

Poporzione all'interno dei cluster





DBSCAN

COME FUNZIONA?

```
db.out <- dbscan(oliveALR[,3:9], eps = 0.5, minPts = 18)
```

eps: Raggio per individuare Core Points

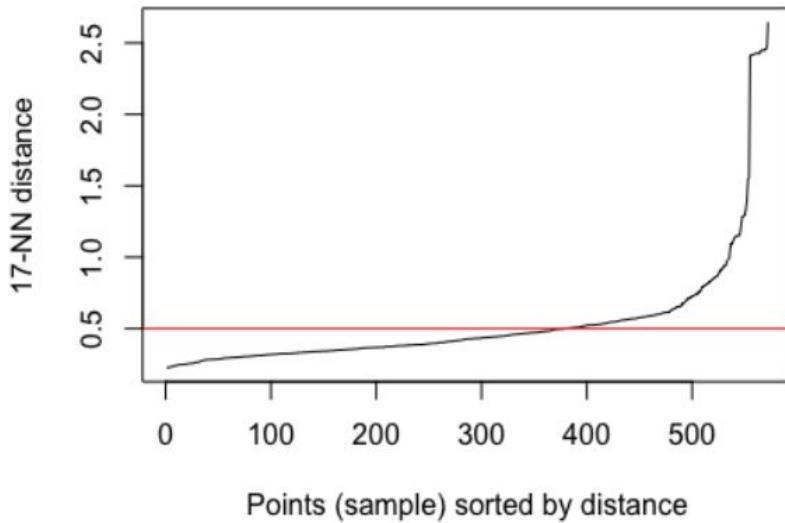
minPts: Minimo di punti in un Cluster

1. Individua i Core Points.
2. Trova i componenti connessi dei Core Points.
3. Associa i punti rimasti a Cluster che distano meno di eps, se esistono.

```
db.out

## DBSCAN clustering for 572 objects.
## Parameters: eps = 0.5, minPts = 18
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 3 cluster(s) and 87 noise points.
##
##    0   1   2   3
##  87 315 125  45
```

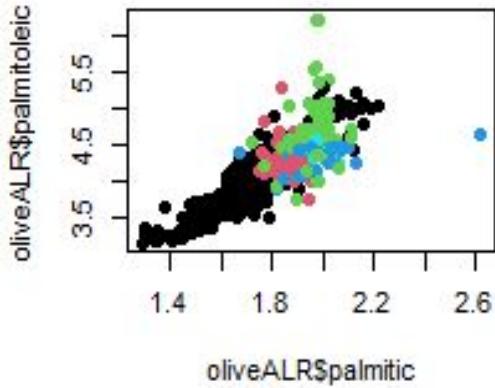
```
kNNdistplot(oliveALR[,3:9], k = 17)  
abline(h=0.5, col = "red")
```



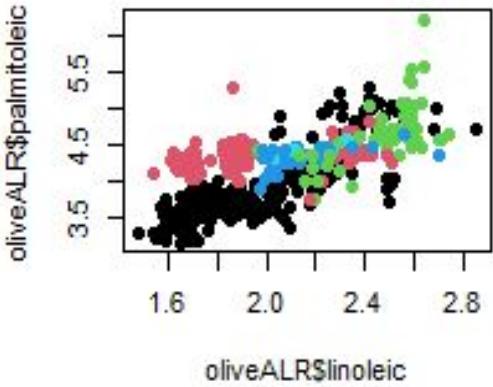
Svolgendo diverse prove
abbiamo avuto risultati migliori
con $\text{eps} = 0.5$ e $\text{minPts} = 18$



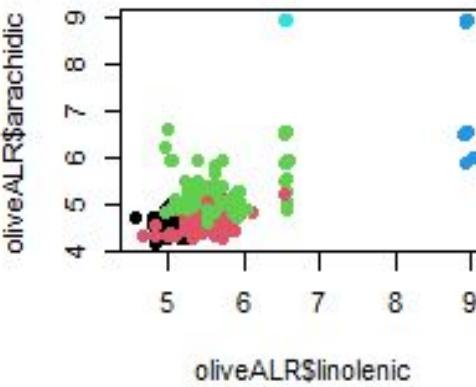
GRAFICI



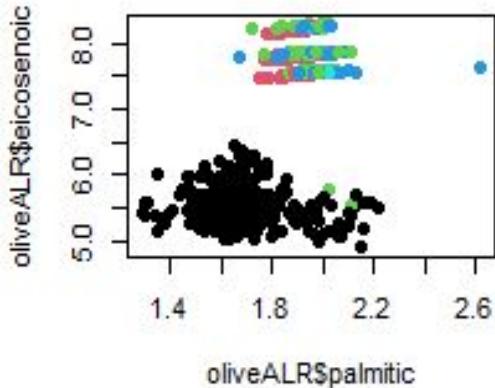
oliveALR\$palmitic



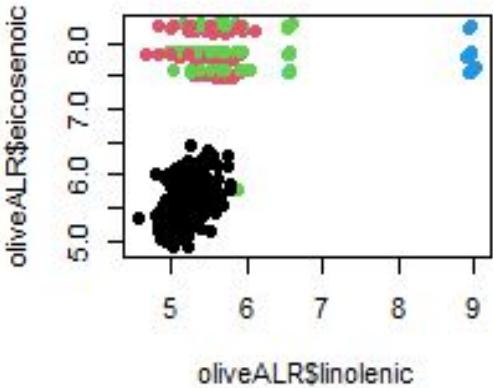
oliveALR\$linoleic



oliveALR\$linolenic



oliveALR\$palmitic

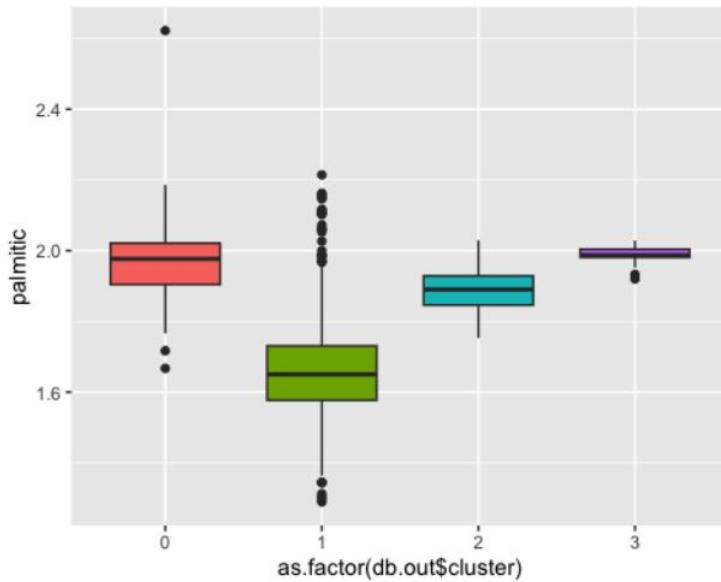


oliveALR\$eicosenoic

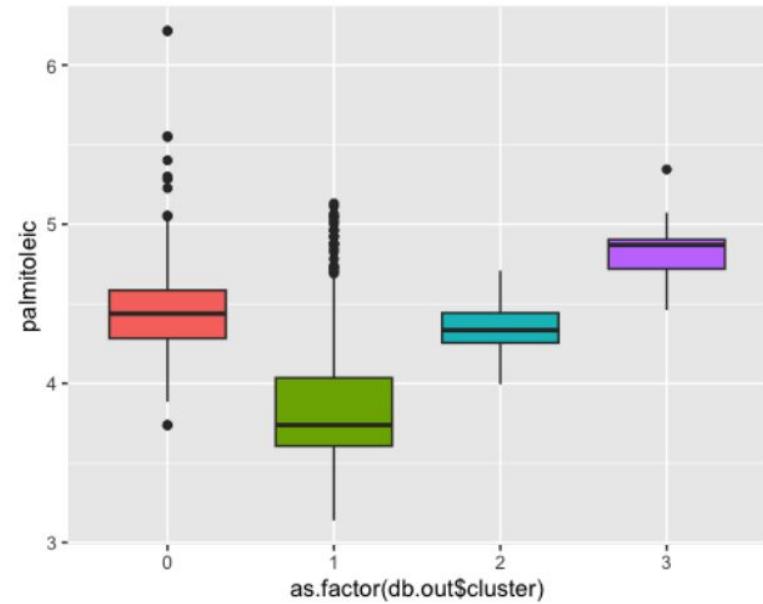
Continuiamo ad utilizzare le stesse coppie di acidi per dare una prima idea dei cluster.

- Acidi fortemente Correlati
- Il Palmitoleico in percentuali minori
- I cluster si discostano l'uno dall'altro
- Range di punti di Noise discretamente ampio

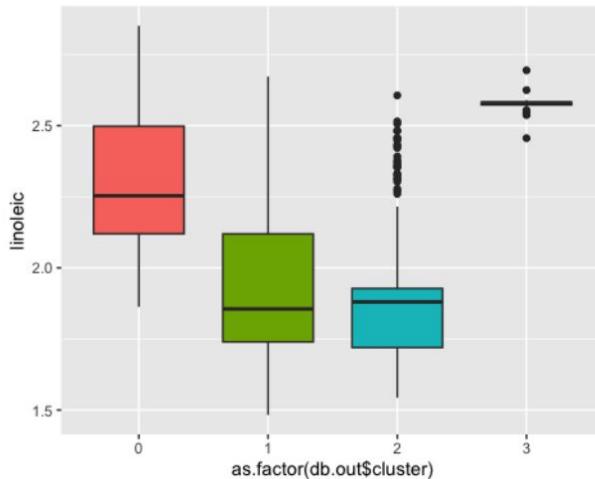
Acido Palmitico



Acido Palmitoleico

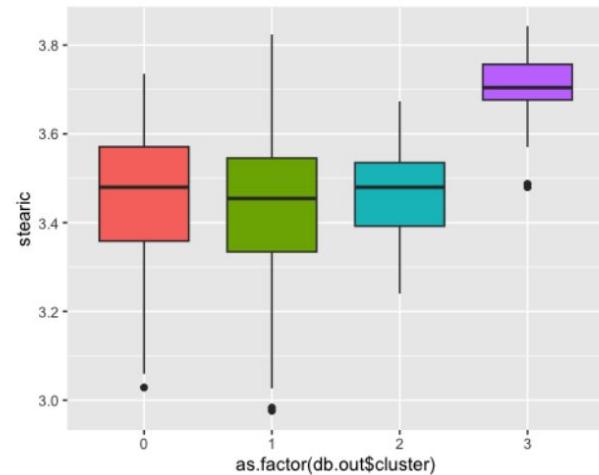


Acido Linoleico



- Il range di valori del cluster 3 è il più basso
- Questo è anche il cluster con percentuali di questi acidi minore

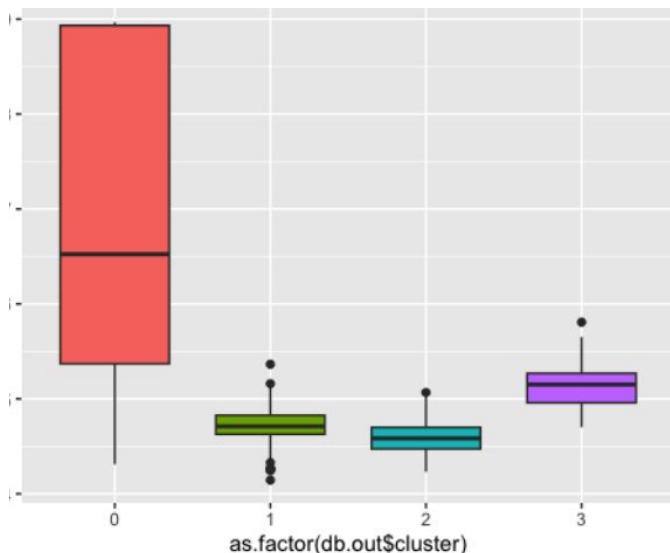
Acido Stearico



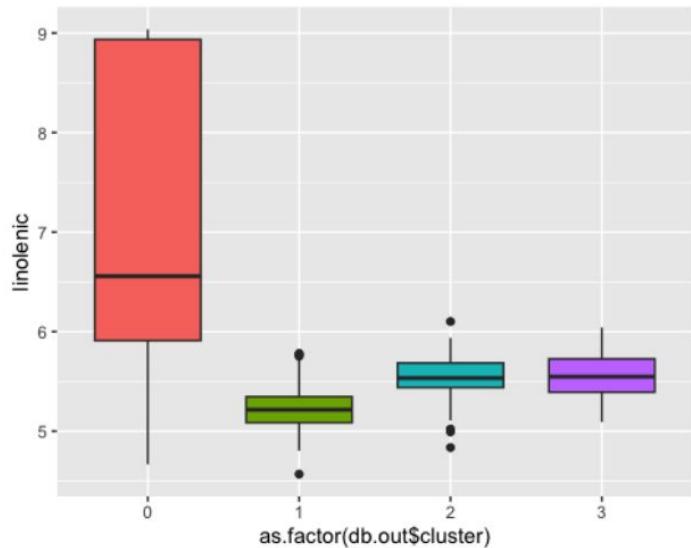
- I cluster 1 e 2 hanno in questo caso varianze molto alte
- Stessa cosa per l'insieme dei punti di noise

- Differenza estremamente grande tra i range dei cluster e dell'insieme dei punti di noise
- Stiamo osservando valori trasformati molto alti
- I cluster sono molto vicini tra di loro

Acido Arachidico

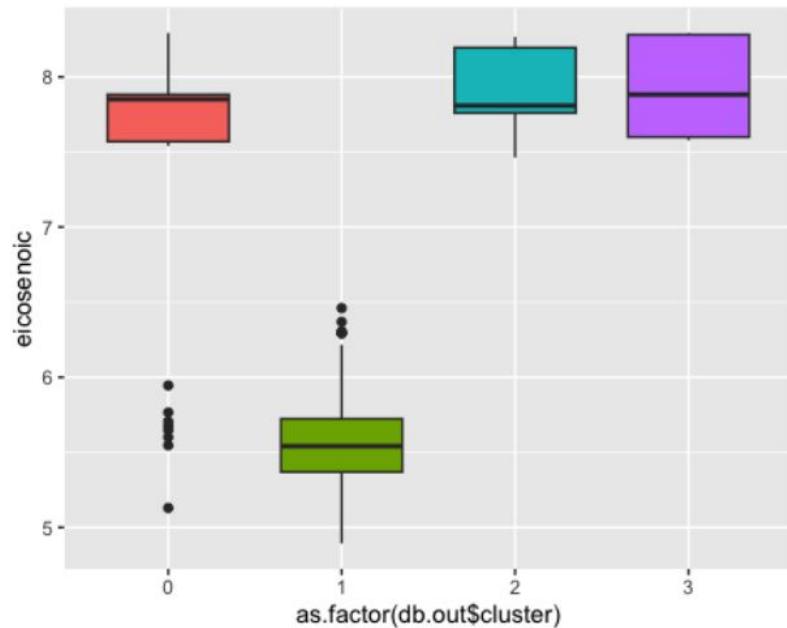


Acido Linolenico



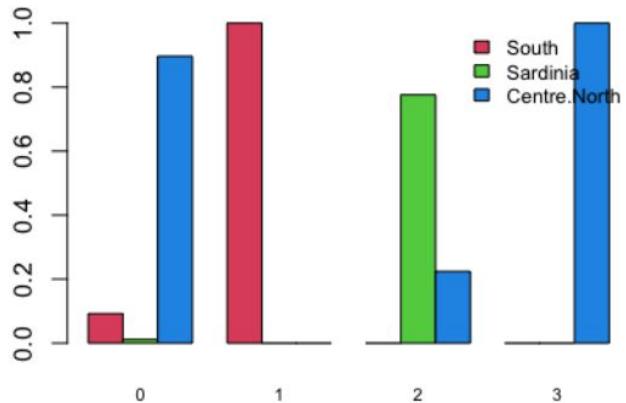
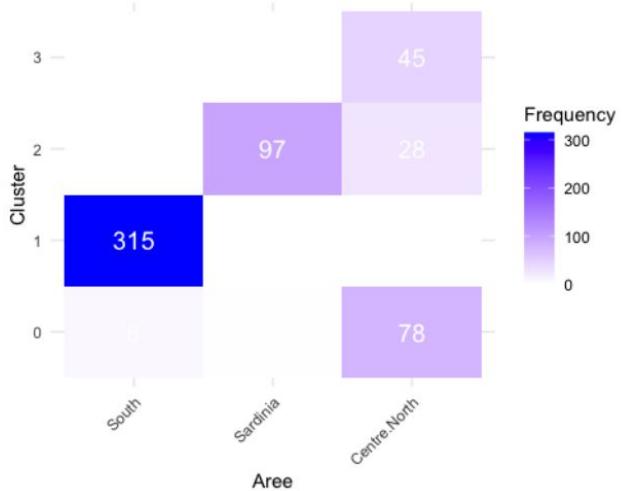
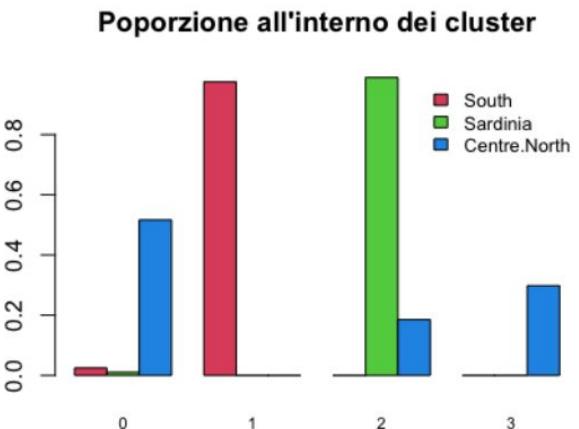
Acido Eicosenoico

- Anche in questo caso gli oli con percentuali molto basse di acido Eicosenoico sono raggruppati sotto diversi Cluster
- Gli oli del cluster 1 sono notevolmente separati dagli altri raggruppamenti
- Questo è uno dei fattori più importanti per la definizione di questo cluster

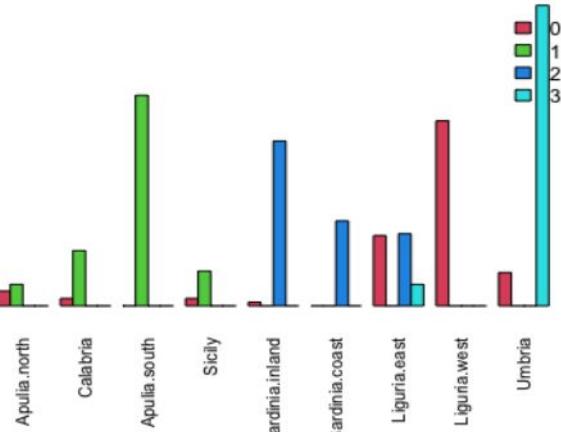


- Il sud e la Sardegna vengono individuati estremamente bene
- Il centro nord si divide invece in diversi cluster
- Questo ha infatti molti punti di noise

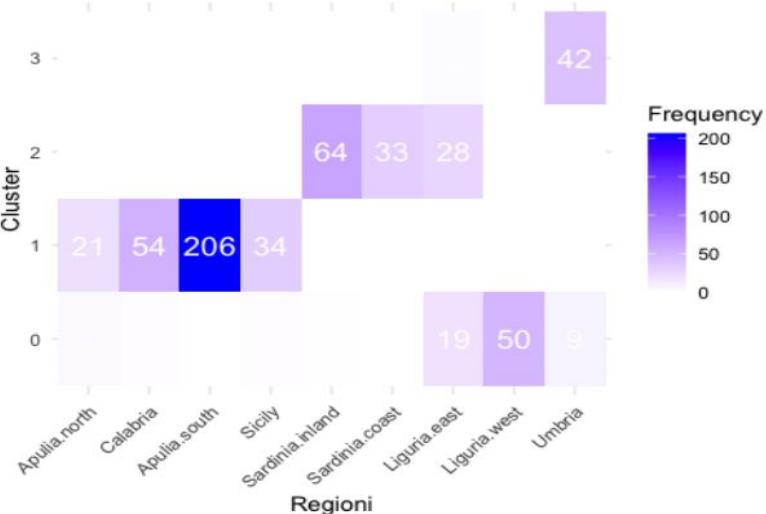
Cosa causa questa divisione negli oli del Centro Nord?



Poporzione all'interno dei cluster



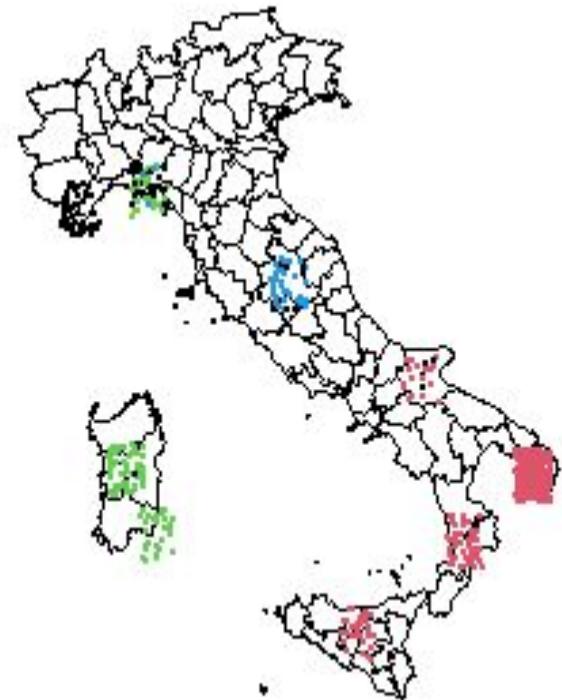
- Umbria completamente isolata
- Liguria che comprende principalmente punti di noise
- Sud e Sardegna perfettamente individuati



CONCLUSIONI:

- Sardegna sempre isolata
- La Liguria è consistentemente la regione che presenta più varianza al suo interno
- L'Umbria è nella gran parte dei casi individuata perfettamente
- Il sud presenta caratteristiche molto simili
- La Puglia del Sud viene considerata cluster separato in alcuni casi

ARI : 0.8341706





**DATI NON
TRASFORMATI**

Il raggruppamento in cluster risulta decisamente peggiore. Ogni distribuzione ha le sue caratteristiche particolari.

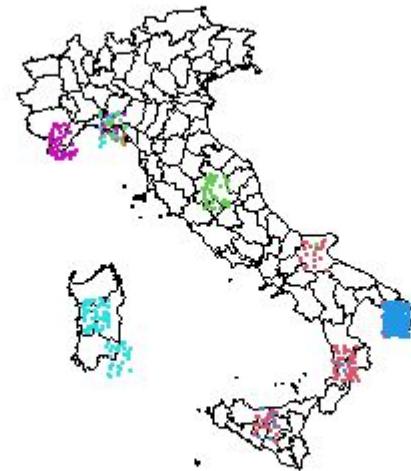
ARI : 0.430368

Kmeans



ARI : 0.5467989

Pam



ARI : 0.3906598

DBScan

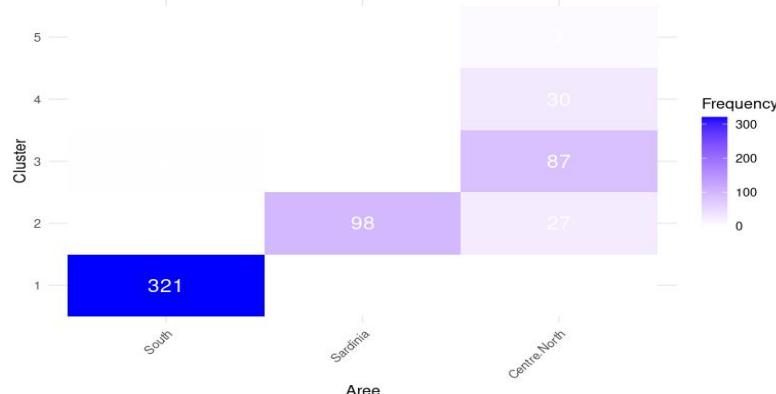




PDFCLUSTER

PDFCluster

- Metodo basato sulla densità.
- Basato sulla ricerca di mode nella distribuzione dei dati.

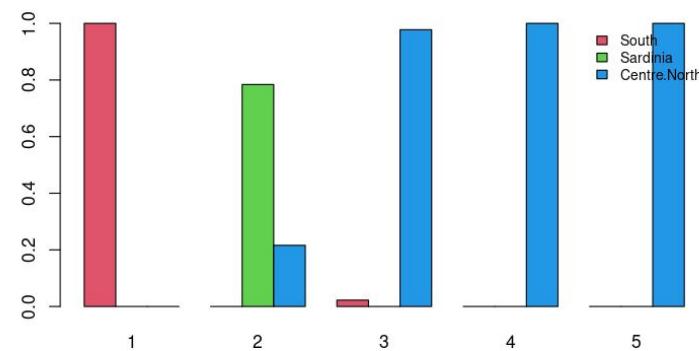
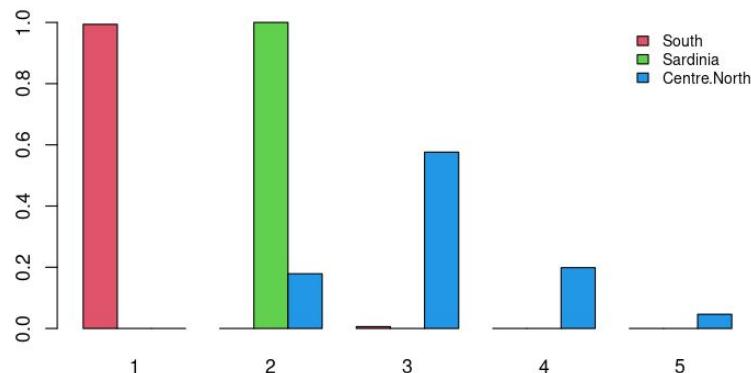


Il cluster 1 individua perfettamente gli oli del sud.

Il cluster 2 individua tutti gli oli della Sardegna.

Cluster 3, 4 e 5 contengono la maggior parte degli oli del centro-nord.

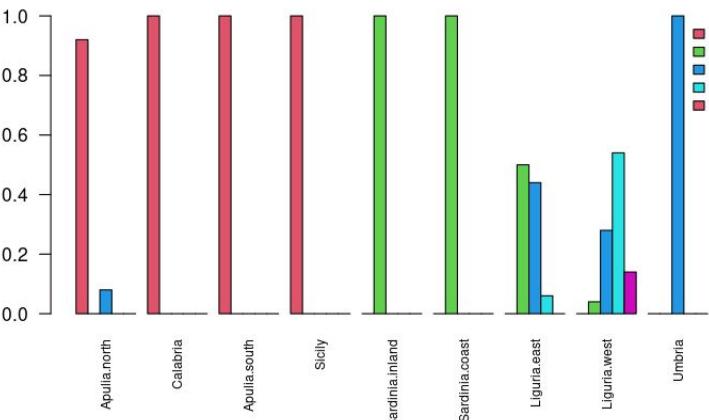
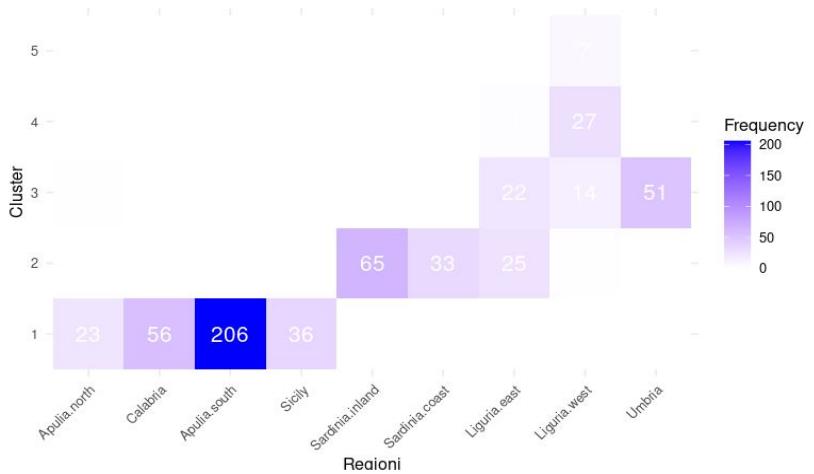
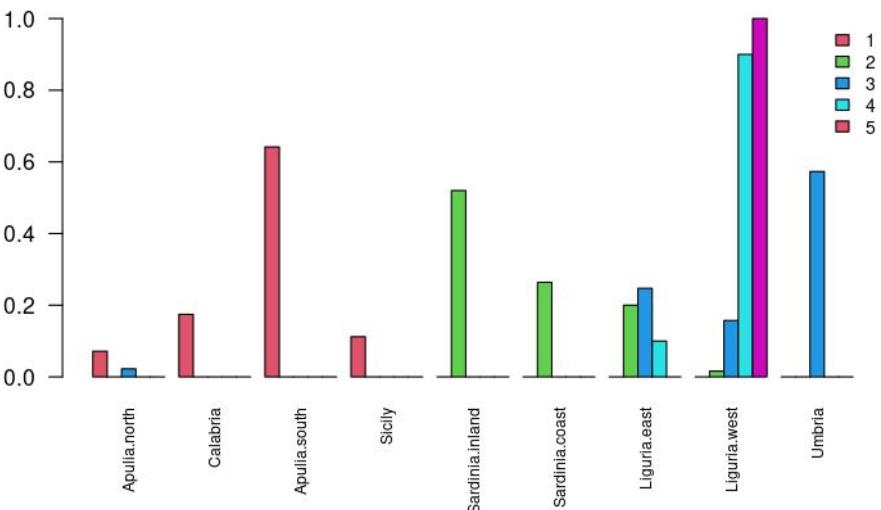
Popozione all'interno dei cluster



Oli dell'Umbria perfettamente individuati dal cluster 3.

La Liguria è la regione divisa su più cluster.

Poporzione all'interno dei cluster



ARI:

Macro area:

0.869787

Regioni:

0.4787563

