Seconda Relazione di Statistica

Clustering

Manni Matteo

PRESENTAZIONE DEL PROBLEMA

Lo scopo di questa relazione è cercare di raggruppare alcuni Stati del continente europeo in cluster, sulla base di dati che riguardano la produzione di carne destinata al consumo umano. Lo studio potrebbe essere rilevante, per esempio, per campagne di sensibilizzazione o pubblicitarie. Si forniranno delle conclusioni mediante differenti metodologie di clustering.

DATASET

Le analisi si svilupperanno utilizzando un dataset composto da quattro diverse tabelle acquisite dal sito dell'Eurostat. Nelle tabelle è presente il peso totale delle carcasse macellate in un anno in ogni Stato (unità di misura: 10^6 Kg). Links:

https://ec.europa.eu/eurostat/databrowser/view/tag00042/default/table?lang=en
https://ec.europa.eu/eurostat/databrowser/view/tag00043/default/table?lang=en
https://ec.europa.eu/eurostat/databrowser/view/tag00044/default/bar?lang=en
https://ec.europa.eu/eurostat/databrowser/view/tag00045/default/table?lang=en

Sono stati presi i dati relativi al 2019 da ogni tabella per ottenere un dataset di 39 osservazioni e 5 fattori, che sono:

1. **geo**: Lo Stato dal quale provengono i dati.
2. **cattle**: Indicatore che riguarda i bovini (vitelli, vitelli, tori, giovenche e vacche) macellati.
3. **pig**: Indicatore relativo ai suini macellati.
4. **sheep_and_goats**: Indicatore che riguarda pecore e capre macellate. Sono compresi gli agnelli.
5. **poultry**: Indicatore del pollame macellato. Sono compresi i seguenti volatili: galline, polli, anatre, tacchini, faraone, oche.

Dopo aver ripulito il dataset, rimuovendo le osservazioni con almeno un valore mancante, l'analisi si restringe a 27 osservazioni.

ANALISI

Per raggiungere gli scopi andremo ad applicare due diverse metodologie di clustering per punti prototipo: k-means e partitions around medoids (pam). Confronteremo poi i risultati con quelli del clustering gerarchico.

K-MEANS

Innanzitutto, per tentativi, decidiamo che 15 è un buon numero di volte per "far girare" l'algoritmo k-means. Impostando un numero minore di iterazioni il metodo potrebbe fermarsi a minimi locali inadeguati.

Cerchiamo di capire quale sia il numero di cluster ottimali per il nostro problema (k = numero di cluster). Rappresentiamo graficamente l'andamento della somma delle mutue distanze tra gli elementi di uno stesso cluster all'aumentare di k (fig.1).

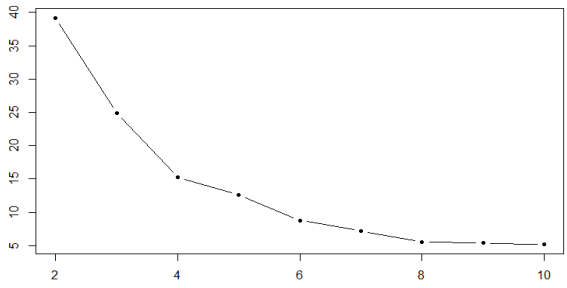

fig.1

Vediamo che le possibili opzioni sono 2, 3 e 4. Con 5 e successivi non si hanno variazioni sostanziali rispetto a 4. Esaminiamo la silhouette nel dettaglio per indicazioni più affidabili. Vediamo che con k = 2 il risultato non è pessimo, i due cluster sono sbilanciati in termini di numero di osservazioni ma comunque accettabili in tal senso (fig.2). La silhouette per il cluster più grande è molto buona, peggiore invece quella del secondo, solo l'osservazione 9 (corrispondente alla Spagna) da un contributo negativo. Impostando k = 3 i due cluster rimangono quasi invariati se non per la sola osservazione 22 (corrispondente al Regno Unito), che viene separata e messa in un cluster al quale non vengono assegnate altre osservazioni (fig.3).

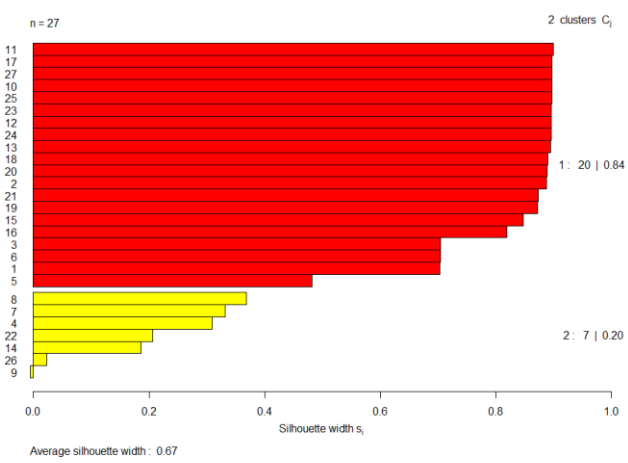

fig.2

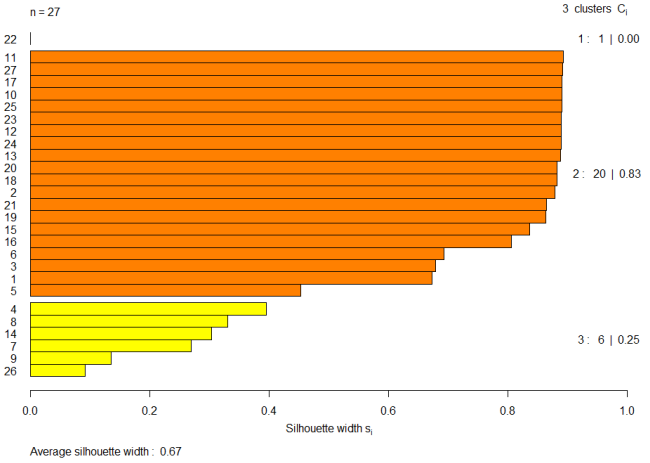

fig.3

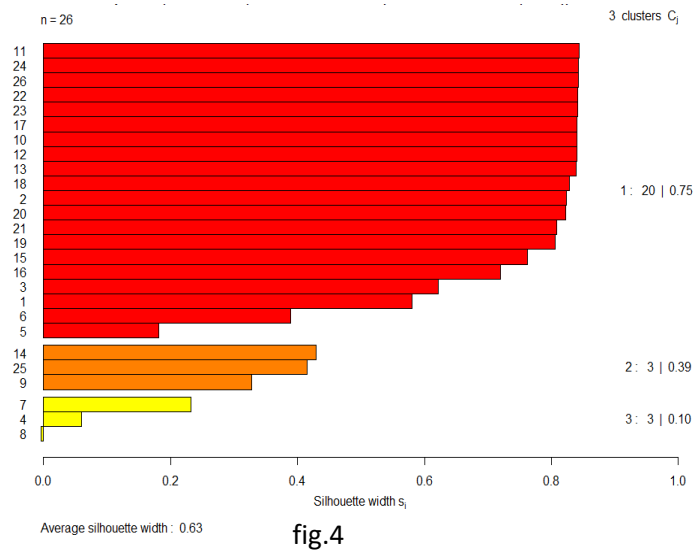

fig.4

Quest'ultimo comportamento non è sicuramente qualcosa che speravamo di ottenere. Proviamo a vedere quindi come reagisce l'algoritmo se escludiamo l'osservazione 22 (fig.4).

Notiamo che le cose non migliorano di molto. Purtroppo, rimane il forte sbilanciamento tra i cluster sia in termini di numero di osservazioni che di silhouette. Con k = 4 quello che otteniamo è un risultato ben peggiore e quindi decidiamo di fermarci a 3.

Visualizziamo sul piano principale in che modo vengono divise le osservazioni per cercare di dare un'interpretazione ai cluster.

Esaminiamo il caso k=2 con ancora presente l'osservazione riguardante il Regno Unito (fig.5). Dalla composizione delle componenti principali (fig.6) notiamo che i fattori **poultry** e **cattle** sono allineati quasi perfettamente alla prima componente. Per quanto riguarda la seconda componente abbiamo che i valori positivi sono caratterizzati dal fattore **pig** mentre quelli negativi dal fattore **sheep_and_goats**. Possiamo quindi solo dire che il cluster di destra (in verde) racchiude gli Stati con una produzione maggiore di carne rispetto agli Stati raggruppati nel cluster di sinistra (in rosso). Con l'integrazione del grafico a Coordinate Parallele non sono possibili commenti particolarmente più ricchi di quelli già fatti, in ogni caso riprenderemo l'argomento nelle conclusioni.

Purtroppo, con questi soli due cluster non si riesce a fare una distinzione netta sulla tipologia di carne prodotta, bensì solo sulla quantità. Sarebbe stata un'informazione interessante da estrapolare.

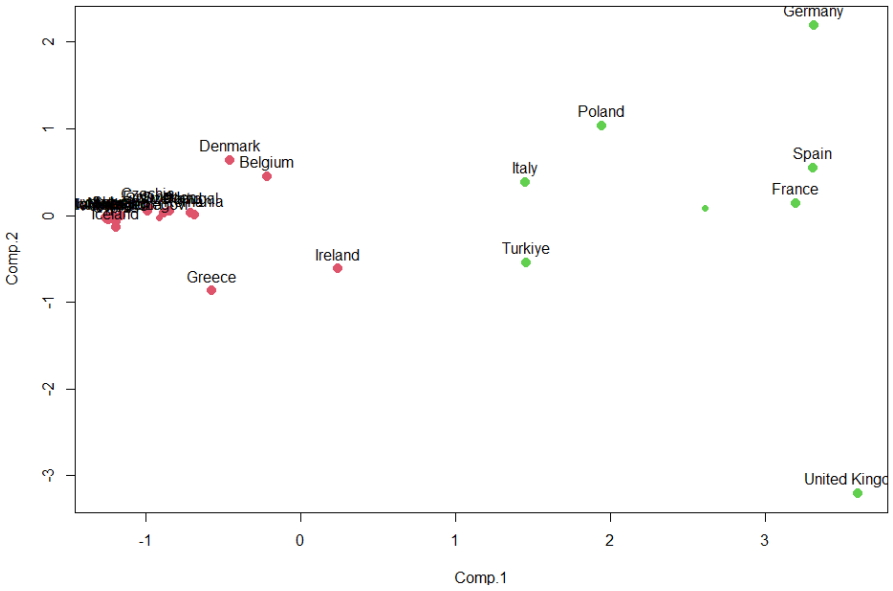

fig.5

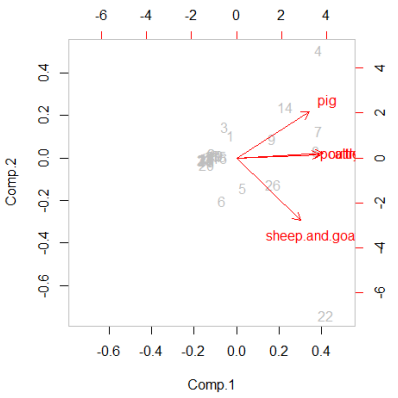

fig.6

Passiamo ad analizzare il risultato prodotto con k=3 e senza l'osservazione che riguarda il Regno Unito (fig.7). Questa volta i fattori quasi allineati alla prima componente sono tre, l'unico che sembra dare un contributo sostanziale alla seconda rimane **sheep_and_goats**.

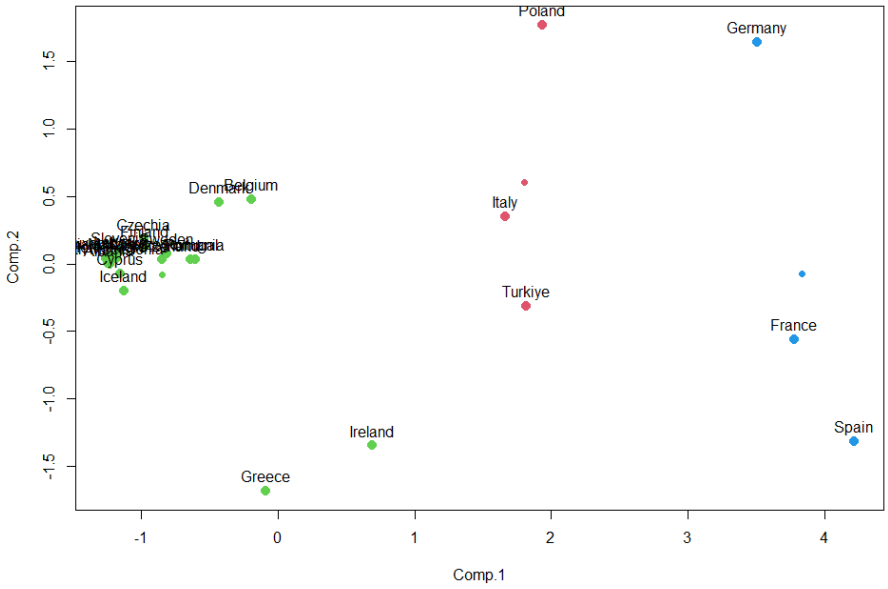

fig.7

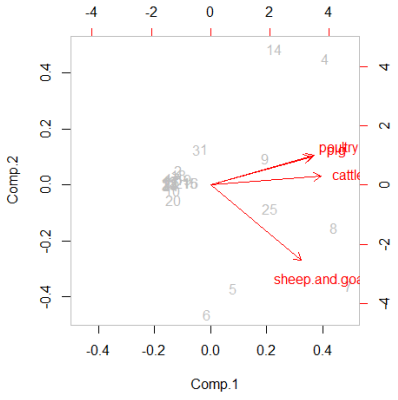

fig.8

Purtroppo, anche in questo caso non riusciamo ad attribuire ai cluster una distinzione sulla tipologia di carne prodotta. Per il momento ci limitiamo, analogamente a come abbiamo fatto prima, ad assegnare ai cluster un identificatore di quantità di produzione, che però questa volta è più granulare perché su tre livelli anziché su due.

PARTITION AROUND MEDOIDS

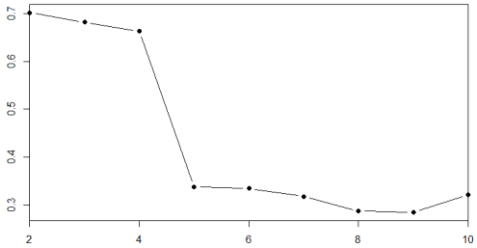

fig.9

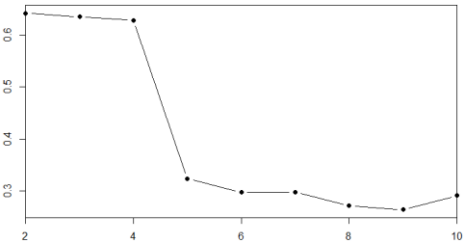

fig.10

Iniziamo l'analisi con un grafico che rappresenta la silhouette media globale al crescere del numero di cluster (fig.9).

Visualizziamo anche il grafico analogo utilizzando la distanza *manhattan* anziché quella *euclidea* standard (fig.10). Notiamo che più o meno gli andamenti sono simili e non dipendono dalla distanza utilizzata e che i valori sono leggermente migliori per la distanza *manhattan*. Ipotizziamo che per k = 5 e successivi non si otterranno risultati sensati.

Procediamo con l'analizzare nel dettaglio il caso k = 2 e distanza *euclidea* (fig.11/fig.12). Notiamo subito che l'osservazione 5 dà un contributo molto negativo alla silhouette del cluster più piccolo, si tratta del record sull'Irlanda. In ogni caso oltre al poco bilanciamento in termini di popolazione dei cluster non osserviamo altre criticità.

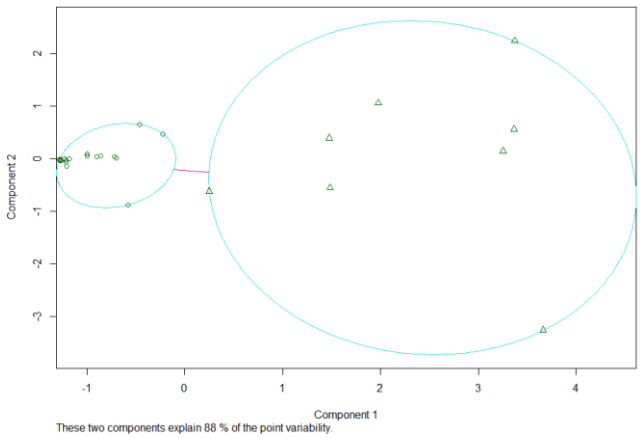

fig.11

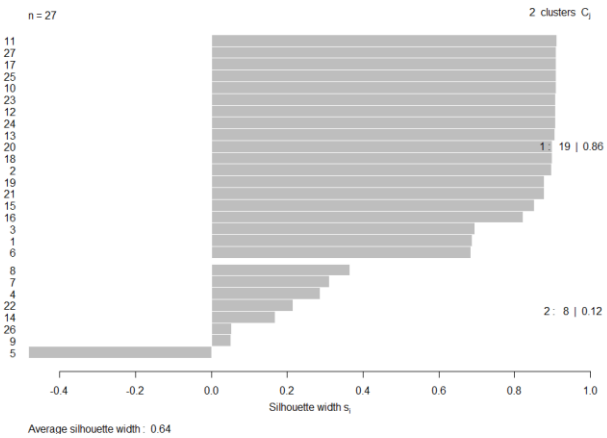

fig.12

Andiamo ad analizzare ciò che succede con la metrica *manhattan* (fig.13/fig.14).

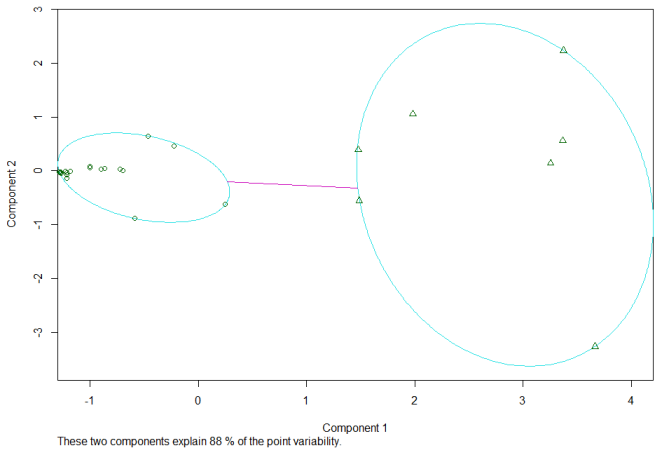

fig.13

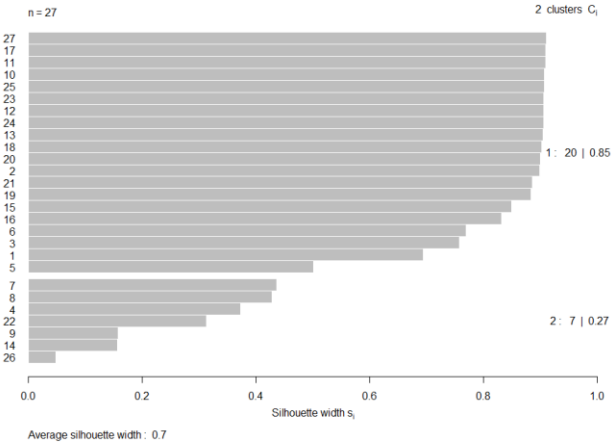

fig.14

Quello che otteniamo è un risultato migliore rispetto al precedente, la differenza sostanziale è stata il passaggio dell'osservazione riguardante lo Stato dell'Irlanda dal cluster più piccolo al più grande. Inoltre, la divisione delle osservazioni è la stessa ottenuta con il metodo k-means, in questo caso però abbiamo una conclusione leggermente migliore in termini di silhouette.

Se proviamo ad applicare l'algoritmo per tre cluster non otteniamo risultati soddisfacenti o comunque sensati, sia utilizzando la distanza *euclidea* che la distanza *manhattan*. Il problema lo si può ricondurre facilmente al medesimo fenomeno che causava problemi nel metodo k-means, ovvero l'osservazione del Regno Unito. In quel caso avevamo risolto il problema rimuovendo il record dall'analisi, in questo però non è sufficiente. Anche per k = 4 i risultati non sono sensati. Inoltre, come avevamo ipotizzato, aumentare il numero di cluster al di sopra di 5 non migliora la situazione.

Possiamo concludere che per quando riguarda il Partition Around Medoids non possiamo spingerci oltre i due cluster purtroppo. L'interpretazione rimane quindi la stessa che avevamo dato nel caso del k-means. Le analisi svolte non sono state tuttavia inutili perché la silhouette è risultata migliore, anche se in minima parte. Preferiamo quindi la partizione del Partition Around Medoids con la distanza *manhattan*.

CLUSTERING GERARCHICO

Per quanto riguarda il clustering gerarchico abbiamo diverse possibilità. Iniziamo a studiare il taglio a due cluster. Le soluzioni prodotte tramite distanza *euclidea* utilizzando *complete*, *single* e *average* linkage non sono accettabili (fig.16). Quello che succede è che viene dedicato un cluster per la sola osservazione del Regno Unito, la stessa osservazione che risultava essere problematica anche per gli altri metodi. In questo caso però, se proviamo a rimuoverla, otteniamo risultati sì migliori ma non sorprendenti e comunque peggiori di quelli già ottenuti con k-means e pam. Quindi la distanza *euclidea* è da scartare.

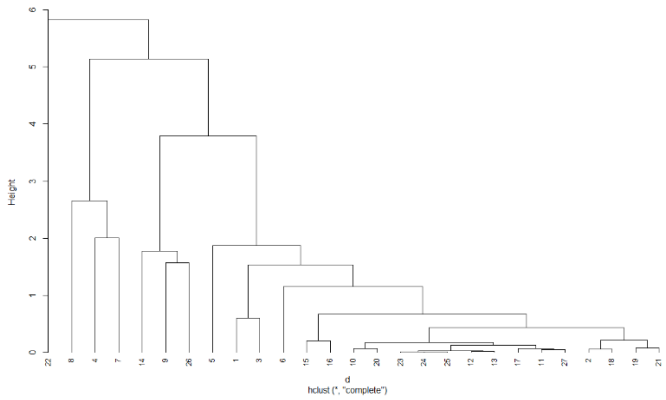

fig.15

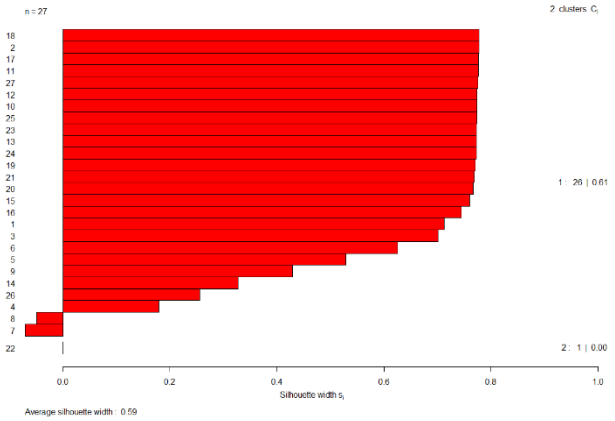

fig.16

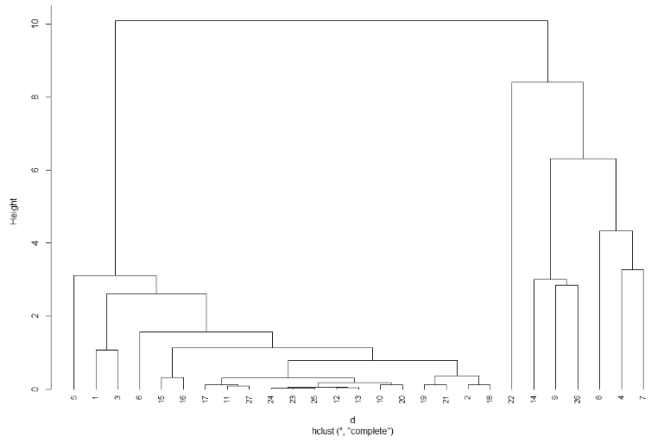

fig.17

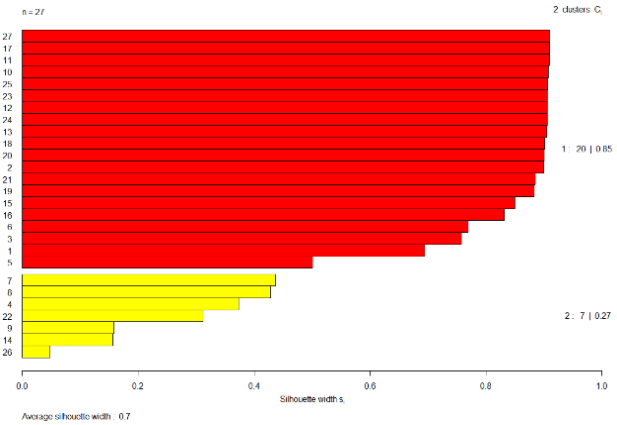

fig.18

Le soluzioni prodotte tramite distanza *manhattan* utilizzando *complete* (fig.18) e *average* linkage sono abbastanza buone. Tramite *single* linkage invece ci imbattiamo nello stesso problema della distanza *euclidea*. I risultati prodotti con *complete* e *average* sono tra loro quasi uguali e inoltre entrambi i due cluster racchiudono rispettivamente le stesse osservazioni che gli venivano assegnate anche con i metodi k-means e pam. Quindi purtroppo non si può dire nulla di diverso né cercare di dare interpretazioni più interessanti.

Proviamo ad esplorare i risultati prodotti dal taglio a tre cluster, per farlo è necessario rimuovere l'osservazione del Regno Unito altrimenti non si ottiene qualcosa di sensato. Valutando tutte le possibili distanze, con tutte le possibilità di linkage, arriviamo alla conclusione che, anche questa volta, la distanza *manhattan* è quella che produce risultati migliori in termini di silhouette. Utilizzare *average* (fig.19) o *complete* linkage è equivalente, da scartare invece il *single*.

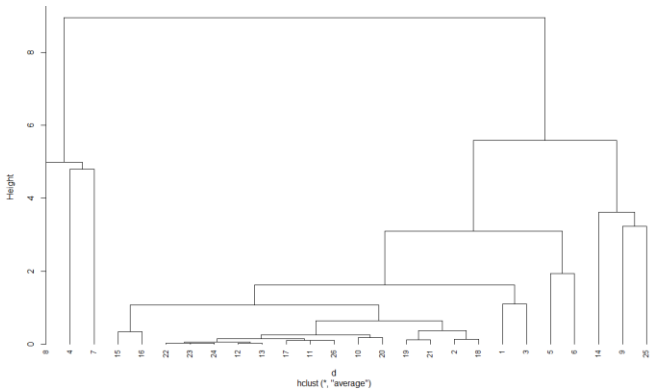

fig.19

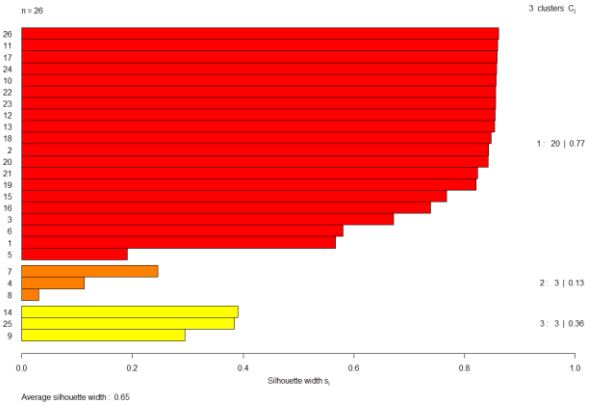

fig.20

Come succedeva per gli altri metodi di clustering, anche qui abbiamo delle partizioni abbastanza sbilanciate. Notiamo la medesima divisione delle osservazioni che avevamo ottenuto durante lo studio del k-means, con la differenza che in questo caso la silhouette risulta leggermente migliore. Quindi, come per il caso dei due cluster, non è possibile dare interpretazioni diverse e più ricche di quelle già fornite.

Andare ad abbassare ulteriormente l'altezza del taglio non ci fa ottenere qualcosa di più significativo, anzi. Quindi ha senso fermarsi a tre cluster.

Sarebbe inoltre inutile andare a visualizzare sul piano principale la divisione delle osservazioni, quello che otterremmo sarebbero rappresentazioni già discusse. Ovvero la fig.5 per la versione con due cluster e la fig.7 per quella a tre.

CONCLUSIONI

In conclusione, quello che si è ottenuto dalle analisi lo possiamo racchiudere in due possibili soluzioni, una sviluppata su due partizioni e una su tre. Conseguenza anche del fatto che utilizzando diverse metodologie abbiamo raggiunto risultati sovrapponibili. Avevamo discusso brevemente una possibile interpretazione durante l'analisi del k-means, rivediamole con l'integrazione dei grafici di tipo Coordinate Parallele.

2 CLUSTER: L'interpretazione più ovvia è quella di attribuire agli Stati appartenenti ai due cluster due diverse produzioni di carne in termini di quantità, banalmente una maggiore dell'altra. Osservando il grafico (fig.21) possiamo notare una differenza netta delle due partizioni per quanto riguarda i valori del pollame prodotto. Anche per gli altri fattori abbiamo una distinzione abbastanza marcata ma con alcune osservazioni che si intrecciano, sporcandola. Possiamo quindi arricchire l'interpretazione distinguendo i due cluster, oltre che sulla base della quantità di carne prodotta, sulla quantità specifica di pollame prodotto.

3 CLUSTER: Con la prima interpretazione approssimativa che avevamo discusso fornivamo sostanzialmente una versione più granulare dell'interpretazione proposta per il caso dei due cluster. La distinzione delle partizioni in tre fasce sulla base della quantità di carne prodotta rimane valida.

Osservando il grafico (fig.22), ottenuto a seguito del clustering gerarchico, possiamo provare a dire qualcosa in più. Notiamo che le osservazioni in verde hanno valori simili alle osservazioni in nero per quanto riguarda i fattori **cattle**, **pig** e **sheep_and_goats**, mentre per il fattore **poultry** si distaccano notevolmente. Abbiamo la situazione opposta confrontando le stesse osservazioni in verde con quelle in rosso; ovvero dei valori di **poultry** che si intrecciano, a differenza degli altri fattori, che rimangono ragionevolmente separati.

Possiamo quindi dire che il cluster avente le osservazioni in verde raggruppa gli Stati con una produzione di carne intermedia, rispetto agli altri due; ed è caratterizzato da una elevata produzione di pollame, grazie alla quale si distingue dal cluster con la produzione di carne minore.

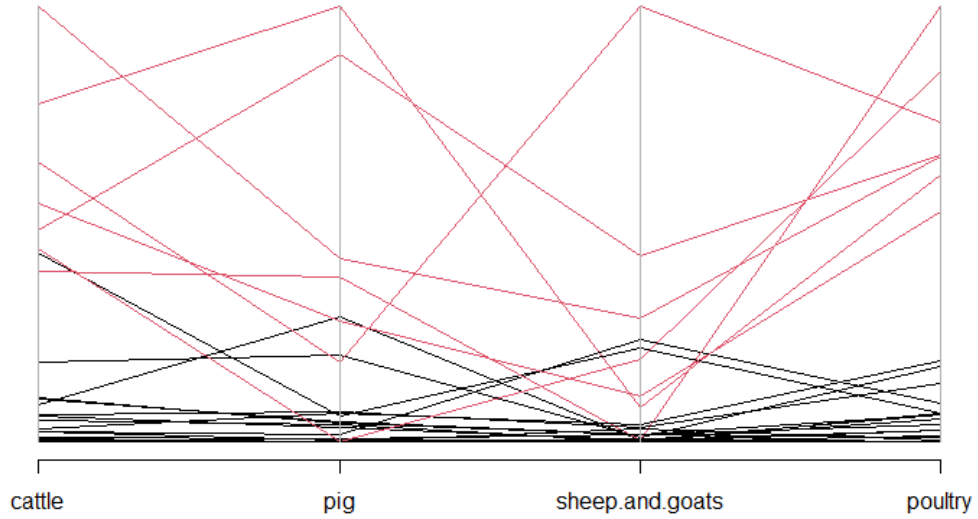

fig.21

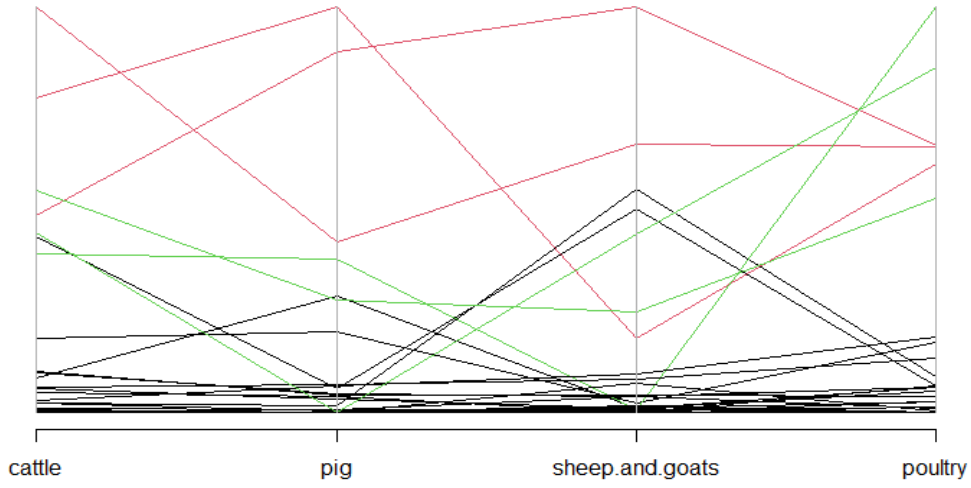

fig.22

APPENDICE

```
data <-read.csv("C:/Users/matte/OneDrive/Desktop/STATISTICA/Seconda_Relazione/tabella.csv",na.strings=c(":"))
data1 <- data[rowSums(is.na(data)) == 0,]
rownames(data1)=NULL
data2 <- scale(data1[,-(1)])
#tolgo l'oss united kingdom
data1.22 <- data1[-22,]
rownames(data1.22)=NULL
data3 <- scale(data1.22[,-(1)])
library(cluster)
library(MASS)
#METODO k-means-------------------------------------------------------
#Osserviamo l'andamento della somma totale dei quadrati delle distanze di ogni cluster dal metodo k-means.
wss=rep(0,10)
for(k in 2:10){
  wss[k]=kmeans(data2,k,nstart=15)$tot.withinss
}
plot(2:10,wss[2:10],type="b",pch=20)
#L'andamento della silhouette
as=matrix(ncol=2,nrow=10)
for(k in 2:10){
  cl=kmeans(data2,k,nstart=15)$cluster
  as[k,1]=mean(silhouette(cl,dist(data2))[,3])
  as[k,2]=sd(silhouette(cl,dist(data2))[,3])
}
as2=as[2:10,]
ymin=min(as2[,1])-max(as2[,2])
ymax=max(as2[,1])+max(as2[,2])
plot(2:10,as2[,1],type="b",pch=20,ylim=c(ymin,ymax))
segments(2:10,as2[,1]-as2[,2],2:10,as2[,1]+as2[,2])
#Esaminiamo la silhouette nei casi k=2,3,4
plot(silhouette(kmeans(data2,2,nstart=15)$cluster,dist(data2)),col=heat.colors(2),border=par("fg"))
plot(silhouette(kmeans(data2,3,nstart=15)$cluster,dist(data2)),col=heat.colors(3),border=par("fg"))
#pca k=2 data2
k=2
data2.km=kmeans(data2,k,nstart=15)
data2.pca=princomp(data2)
plot(data2.pca$scores,col=1+data2.km$cluster,pch=20,cex=2)
points(predict(data2.pca,data2.km$centers),col=2:(k+1),pch=19)
text(data2.pca$scores,labels=as.character(data1$geo),pos=3)
```

```
biplot(data2.pca,col=c("gray","red"))

parcoord(data2,col=as.numeric(data2.km$cluster))

#silhouette k=3 data3

plot(silhouette(kmeans(data3,3,nstart=15)$cluster,dist(data3)),col=heat.colors(3),border=par("fg"))

#pca k=3 data3

k=3

data3.km=kmeans(data3,k,nstart=15)

data3.pca=princomp(data3)

plot(data3.pca$scores,col=1+data3.km$cluster,pch=20,cex=2)

points(predict(data3.pca,data3.km$centers),col=2:(k+1),pch=19)

text(data3.pca$scores,labels=as.character(data1.22$geo),pos=3)

biplot(data3.pca,col=c("gray","red"))

parcoord(data3,col=as.numeric(data3.km$cluster))

#METODO pam----------------------------------------------------

c=rep(0,10)

for(i in 2:10){

  c[i]=pam(data2,i)$silinfo$avg.width

}

plot(2:10,c[2:10],type="b",pch=19)

data2.pam=pam(data2,2)

plot(data2.pam)

#proviamo a cambiare la tipologia di distanza - manhattan

c=rep(0,10)

for(i in 2:10){

  c[i]=pam(data2,i,metric="manhattan")$silinfo$avg.width

}

plot(2:10,c[2:10],type="b",pch=19)

data2.pam2=pam(data2,2,metric="manhattan")

plot(data2.pam2) #meglio perchè l'osservazione 7 viene ceduta all'altro cluster

#con 3 cluster e data3 non vengono risultati sensati come per la distanza euclidea

data3.pam3=pam(data2,2,metric="manhattan") #stesso problema con l'oss 22

plot(data3.pam3)

data3.pam3=pam(data3,3) #proviamo a toglierla

plot(data3.pam3)

data3.pam3=pam(data3,4) #il risultato con 3 non era buono e con 4 neppure lo è

plot(data3.pam3)

plot(data2.pca$scores,col=1+data2.pam2$clustering,pch=20,cex=2) #qui vediamo che funziona come il kmenas con
due cluster


parcoord(data2,col=as.numeric(data2.pam2$cluster))

#biplot

data3.pca=princomp(data3)
```

```
biplot(data3.pca)
#proviamo a dare un senso
parcoord(data2.pca$scores,col=as.numeric(data3.pam2$clustering))
parcoord(data3.pca$scores[,1:2],col=as.numeric(data3.pam2$clustering))
#METODI GERARCHICI----------------------------------------------------
#d<-dist(data2) # distanza euclidea
#d<-dist(data2)^2 # distanza euclidea quadrata
#d<-dist(data2,method="maximum")   # distanza del massimo
d<-dist(data2,method="manhattan") # distanza manhattan
data.hc=hclust(d) # complete linkage
#data.hc=hclust(d,method="single") # single linkage
#data.hc=hclust(d,method="average") # average linkage
plot(data.hc,hang=-1,cex=0.8)
k=2
data.cut=cutree(data.hc,k)
plot(silhouette(data.cut,d),col=heat.colors(k),border=par("fg"))
#----------------------------------------
#d<-dist(data3) # distanza euclidea
#d<-dist(data3)^2 # distanza euclidea quadrata
#d<-dist(data3,method="maximum")   # distanza del massimo
d<-dist(data3,method="manhattan") # distanza manhattan
#data.hc=hclust(d) # complete linkage
#data.hc=hclust(d,method="single") # single linkage
data.hc=hclust(d,method="average") # average linkage
plot(data.hc,hang=-1,cex=0.8)
k=3
data.cut=cutree(data.hc,k)
plot(silhouette(data.cut,d),col=heat.colors(k),border=par("fg"))
#-----------------------------------------------------------------
k=3
d<-dist(data3,method="manhattan")
data.hc=hclust(d,method="average")
data.cut=cutree(data.hc,k)
data3.pca=princomp(data3)
plot(data3.pca$scores,col=1+data.cut,pch=20,cex=2)
#points(data3.pca$scores,col=1+as.integer(iris$Species),pch=20,cex=2)
text(data3.pca$scores,labels=as.character(data1[-22,]$geo),pos=3,cex=1)
parcoord(data3,col=data.cut)
```