PROGETTO DI STATISTICA
PARTE II

Corso di Laurea Magistrale in Artificial Intelligence and Data Engineering
Cocchella Ludovica | matricola: 532189

Sommario

1. Introduzione al caso di studio e obiettivo dell'analisi

Il contesto di analisi rimane il medesimo dell'analisi *PCA,* che è stata effettuata per la prima relazione, ovvero l'Istituto superiore di sanità conduce molteplici attività dedicate alla salute mentale: attività di ricerca, di sorveglianza, di prevenzione e comunicazione. L'obiettivo della nostra analisi risulta essere la ricerca di gruppi sostanzialmente omogenei di individui (nel nostro caso le regioni italiane) su cui definire, programmare e attuare con priorità e dove più necessarie campagne di intervento e sensibilizzazione a favore di tutti i cittadini e puntare anche su nuovi investimenti per aiutare chi sta soffrendo di questo disagio cercando di migliorare e prevenire varie condizioni che vanno ad influire sullo stato di salute mentale delle persone.

1.2 Reperimento dei dati

La tabella su cui viene effettuata la nostra analisi è la stessa che è stata utilizzata nell'analisi *PCA* (effettuata nella prima relazione). Per maggiore chiarezza riportiamo brevemente quanto già era stato detto nella precedente relazione riguardo al reperimento dei dati che compongono la tabella. Tali dati sono stati ottenuti dalla banca dati ISTAT con un attento lavoro di selezione. Il link da cui sono stati estrapolati i dati è il seguente http://dati.istat.it/. La procedura di estrapolazione dei dati e di creazione della tabella è stata esposta in maniera dettagliata nella prima relazione. Di seguito ci limitiamo a riportare solo i fattori che caratterizzano la nostra tabella e il loro significato:

- I.Sedentari: fattore che rappresenta gli individui sedentari.
- Buona.Salute e A.Disturbi.Nervosi: fattori che rappresentano rispettivamente gli individui che godono di buona salute e gli individui affetti da disturbi nervosi (quali ansia, depressione, psicosi ecc.).
- I.Economica, I.Tempo.Libero, I.Relazioni.Amicali, I.Salute: fattori che rappresentano vari aspetti riguardanti la soddisfazione delle persone per la vita.
- Obesi: fattore che rappresenta gli individui obesi.

2. Cluster Analysis

Per raggiugere il nostro obiettivo di analisi è stata utilizzata la *Cluster Analysis* che consiste in un insieme di tecniche di analisi multivariata dei dati per la selezione e il raggruppamento di elementi omogenei in un insieme di dati basandosi su misure relative alla somiglianza tra questi e utilizzando come supporto l'analisi *PCA* (teniamo in considerazione le sole prime due componenti principali).

2.1 Elaborazione dei dati

La nostra analisi è stata effettuata tramite il supporto del software "R" e di seguito riportiamo tutti i vari step e comandi necessari per il caricamento della tabella sul software e la sua relativa analisi.

Impostiamo la directory da cui importare la tabella tramite il comando: `> setwd()`

Dopo di che carichiamo la tabella tramite il comando: `> tab=read.csv2("tabella.csv", row.names = 1)`. Prima di proseguire con la nostra analisi è necessario standardizzare la nostra tabella per evitare che problemi di scala possano andare a falsare il risultato del clustering.

Possiamo adesso procedere con la *Cluster Analysis*. Il problema del clustering può essere affrontato per mezzo di metodi di tipo punti prototipo e per mezzo di metodi gerarchici. Quest'ultimi sono tutti metodi esplorativi quindi non vi è il metodo giusto per un set di dati. Risulta dunque necessario capire quali di questi metodi che abbiamo a disposizione risulti il più adatto per il fenomeno in esame. Pertanto, applichiamo vari metodi di clustering e valutiamo per ognuno di questi la bontà dell'analisi per poi decidere quale sia il migliore per il raggiungimento del nostro obiettivo.

2.1.1 Analisi con l'algoritmo k-means

Esaminiamo il problema utilizzando il metodo k-means. Il primo passo consiste nell'identificare qual è il numero di cluster ottimale per ottenere delle partizioni il più simile tra loro. Tale scelta può essere effettuata osservando il grafico, che mostra l'andamento della silhouette media globale al variare del numero di cluster, che riportiamo di seguito:

```
> as=rep(0,15)

> for(k in 2:15){

+
cl=kmeans(tab.st,k,nstart=25)$cluster

+
as[k]=mean(silhouette(cl,dist(tab.st))[,
3])

+ }

> plot(2:15,as[2:15],type="b",pch=20)
```

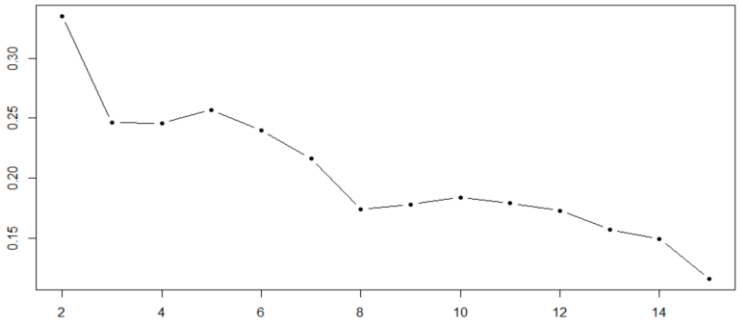

Figura 1: Andamento silhouette media globale al variare del numero di cluster

Dal grafico in figura 1 possiamo notare un picco in corrispondenza di 2 cluster, in cui si ha il valore più alto di silhouette media globale, un andamento pressoché stabile dai 3 ai 6 cluster per poi decrescere fino a tendere quasi a zero. Risulta dunque interessante andare a visualizzare più nel dettaglio le suddivisioni nel caso di 2,3,4,5 cluster (notiamo in corrispondenza di 5 cluster una leggera variazione di pendenza, per questo motivo vale la pena vedere cosa succede). Valutiamo i risultati dell'applicazione dell'algoritmo k-means, specificando tra i parametri il numero di cluster che vogliamo ottenere:

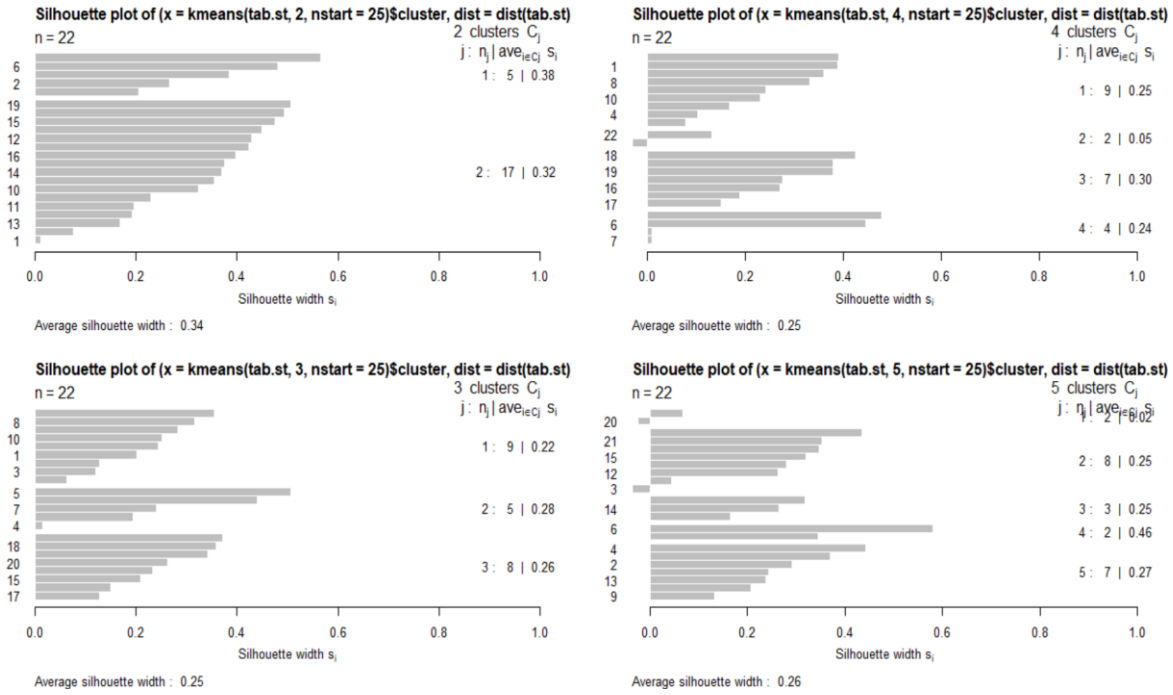

Figura 2: Silhouette per k=2,3,4,5

Da una prima visualizzazione possiamo notare che la suddivisione in 2 cluster presenta la silhouette media globale più alta mentre per quanta riguarda la suddivisione in 3,4 e 5 cluster notiamo che la silhouette media globale risulta essere invece sostanzialmente uguale ma di valore inferiore rispetto a quella della suddivisione in due cluster (indice di un aumento di incertezza nella decisione di appartenenza ai vari cluster). Nel caso di k=2,3 notiamo che i valori della silhouette media di ogni cluster siano paragonabili e inoltre non ci sono elementi che risultano essere mal classificati ovvero che hanno silhouette negativa. Non possiamo dire lo

stesso sia per quanto riguarda k=5 che k=4. Infatti, guardando le silhouette medie relative ad ogni cluster notiamo che vi è una maggiore variabilità probabilmente dovuta al fatto che ci sono valori che risultano essere mal classificati in quanto caratterizzati da una silhouette negativa. Oltre a ciò, notiamo una disomogeneità nel numero di individui per ogni classe, infatti, sia nel caso k=4 che nel caso k=5 vediamo che è presente un cluster costituito da soli due individui di cui uno con silhouette negativa il che porta a pensare che questo cluster sia decisamente sbagliato. Dalle considerazioni appena fatte si evince che la suddivisione in quattro o cinque cluster non sia quella più giusta mentre risulta interessante la suddivisione nei casi di k=2,3.

2.1.2 Analisi con l'algoritmo PAM

Ripetiamo la nostra analisi applicando l'algoritmo PAM. Anche in questo caso è necessario prima di tutto identificare il numero di cluster ottimale. Riportiamo di seguito il grafico che mostra l'andamento della silhouette media globale per ogni suddivisione:

```
> c=rep(0,15)
> for(i in 2:15){
+   c[i]=pam(tab.st,i)$silinfo$avg.width
+ }
> plot(2:15,c[2:15],type="b",pch=19)
```

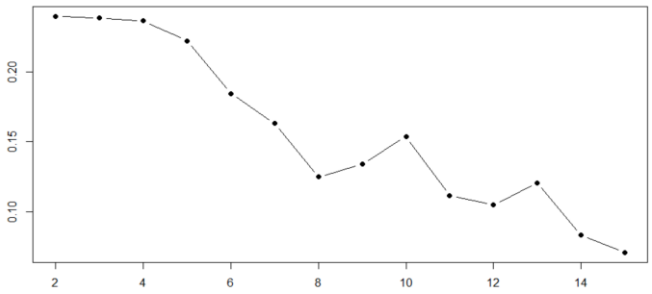

Figura 3: Variazione silhouette al variare del numero di cluster

Dalla figura 4 possiamo notare un picco in corrispondenza di due cluster in cui si ha il valore più alto di silhouette. La silhouette risulta avere una andamento decrescente fino al cluster 8 per poi aumentare e diminuire più volte presentando due punti di picco in corrispondenza della suddivisione in 10 e 13 cluster. Risulta quindi interessante andare ad esaminare più nel dettaglio le suddivisioni nei casi k=2 e k=3 poiché presentano i valori più alti di silhouette media globale. Nonostante per k=10 e k=13 siano presenti nel grafico di figura 3 delle nette variazioni di pendenza non risulta aver senso esaminare la situazione più nel dettaglio visto che, essendo la nostra tabella composta da 22 osservazioni, saranno presenti diversi cluster formati da un solo elemento che non sono molto significativi visto che noi vogliamo gruppi di elementi omogenei.

Valutiamo i risultati dell'applicazione dell'algoritmo PAM, specificando tra i parametri il numero di cluster che vogliamo ottenere. Di seguito riportiamo i grafici relativi alle suddivisioni in 2 e 3 cluster:

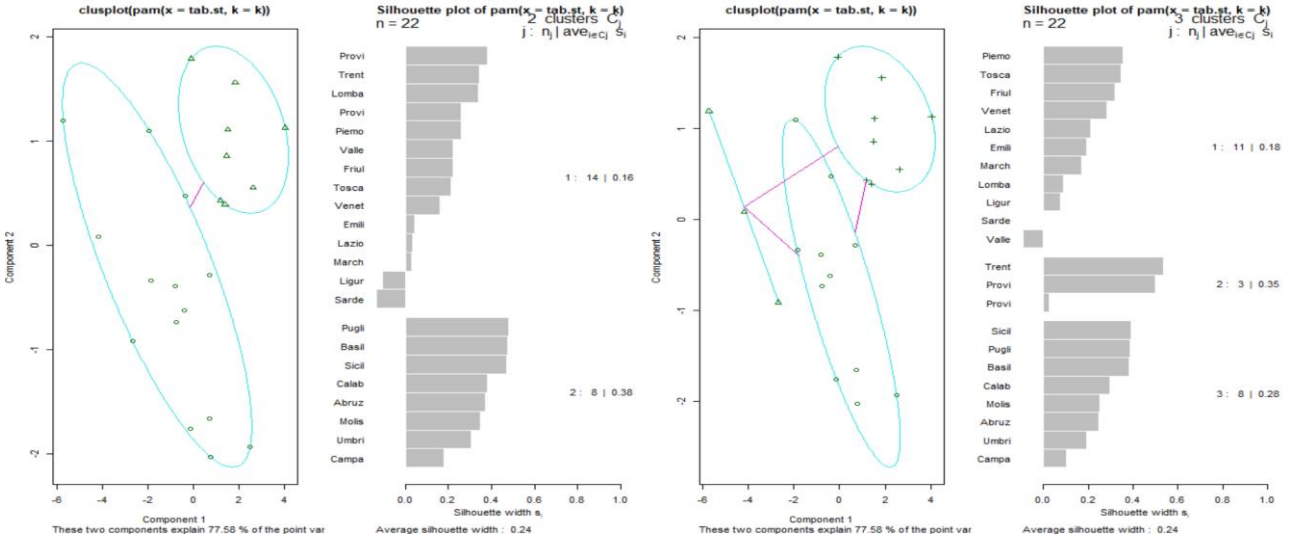

Figura 4: grafici a sinistra suddivisioni per k=2 e grafici a destra suddivisioni per k=3

Notiamo che il valore della silhouette media globale (0.24) risulta essere la stessa sia per k=2 che per k=3. Possiamo inoltre affermare che la suddivisione sia in due che in tre cluster non risulta essere particolarmente soddisfacente poiché in entrambe le suddivisioni si ha variabilità nei valori delle silhouette medie relative ad ogni cluster e ciò è dovuto al fatto che vi sono individui mal classificati (silhouette relativa al singolo elemento negativa) che vanno infatti ad abbassare il valore della silhouette media globale. Questi valori sono quindi indice di una certa incertezza nella classificazione.

Proviamo a cambiare la distanza, visto che la distanza standard utilizzata dal comando pam risulta essere quella euclidea. Utilizziamo la distanza manhattan. Anche in questo caso valutiamo la silhouette media globale tramite il grafico che mostra l'andamento della silhouette media globale per ogni suddivisione:

```
> c=rep(0,15)
> for(i in 2:15){
+
c[i]=pam(tab.st,i,metric="manhattan")$
silinfo$avg.width
+ }
> plot(2:15,c[2:15],type="b",pch=19)
```

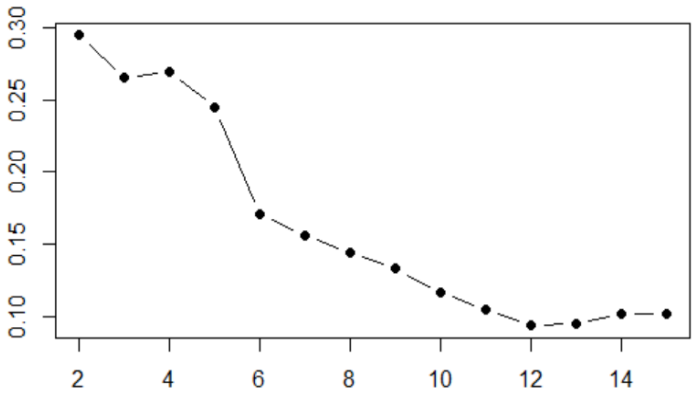

Figura 5: Variazione della silhouette al variare del numero di cluster

Dal grafico di figura 6 notiamo un picco in corrispondenza di 2 cluster in cui si registra il valore più alto di silhouette media globale. Da due cluster in poi la silhouette tende a decrescere fino ad arrivare a zero. Risulta interessante osservare più in dettaglio i casi per k=2,3,4 (in k=4 la silhouette aumenta leggermente rispetto al caso k=3 dunque risulta interessante analizzare la situazione per 4 cluster).

Applichiamo l'algoritmo di PAM specificando tra i parametri il numero di cluster che vogliamo ottenere e la distanza manhattan. Riportiamo di seguito i grafici ottenuti per k=2,3,4:

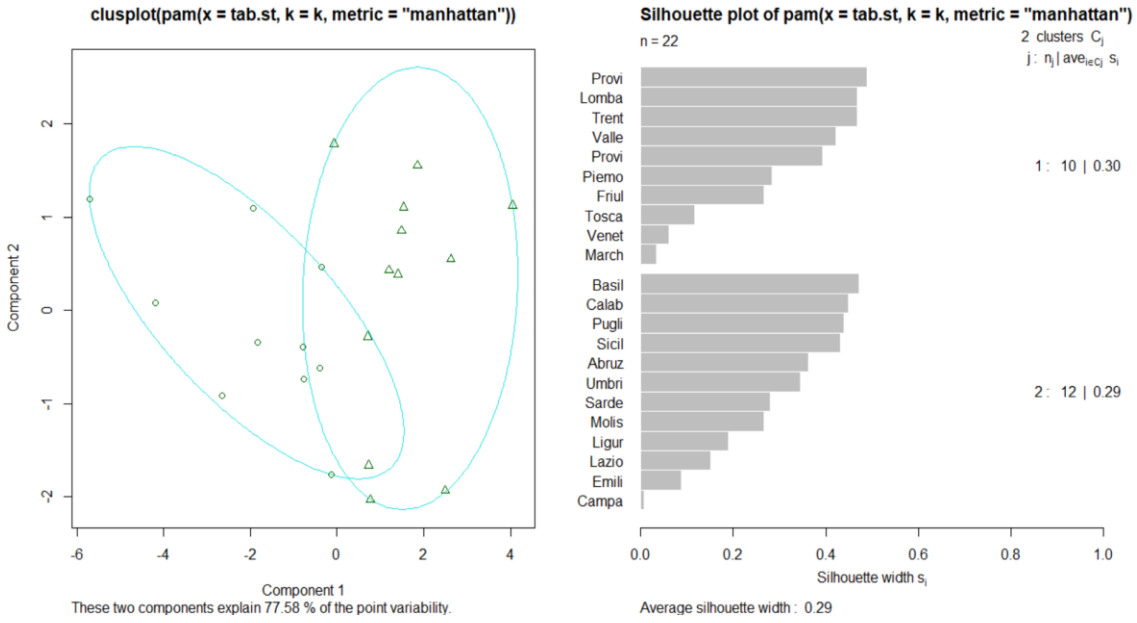

Figura 6: Grafici ottenuti per k=2

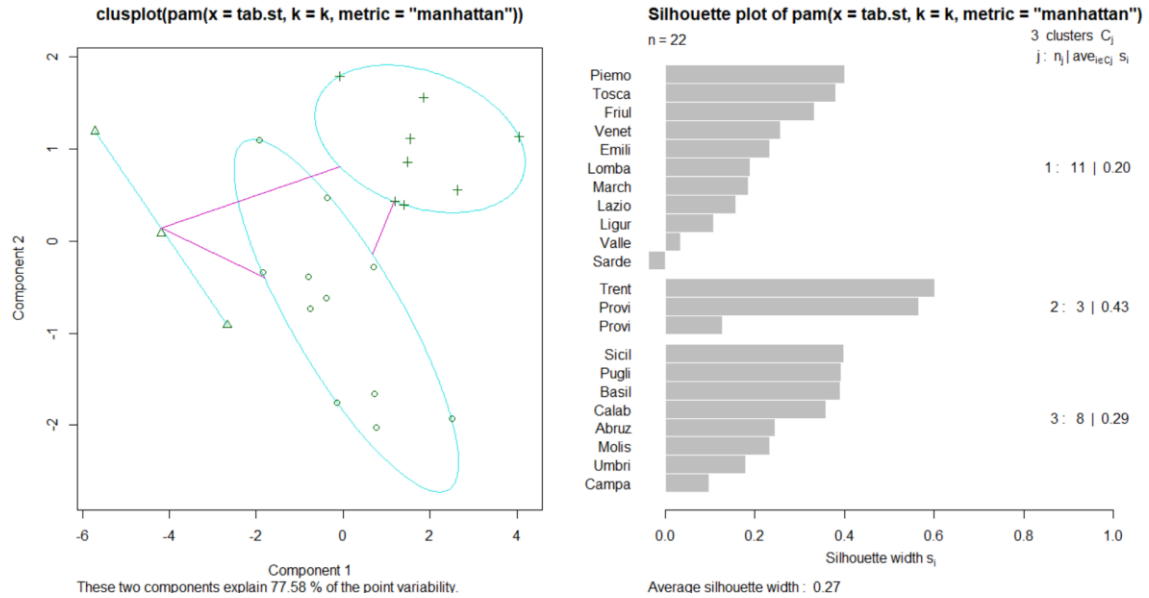

Figura 7: Grafici ottenuti per k=3

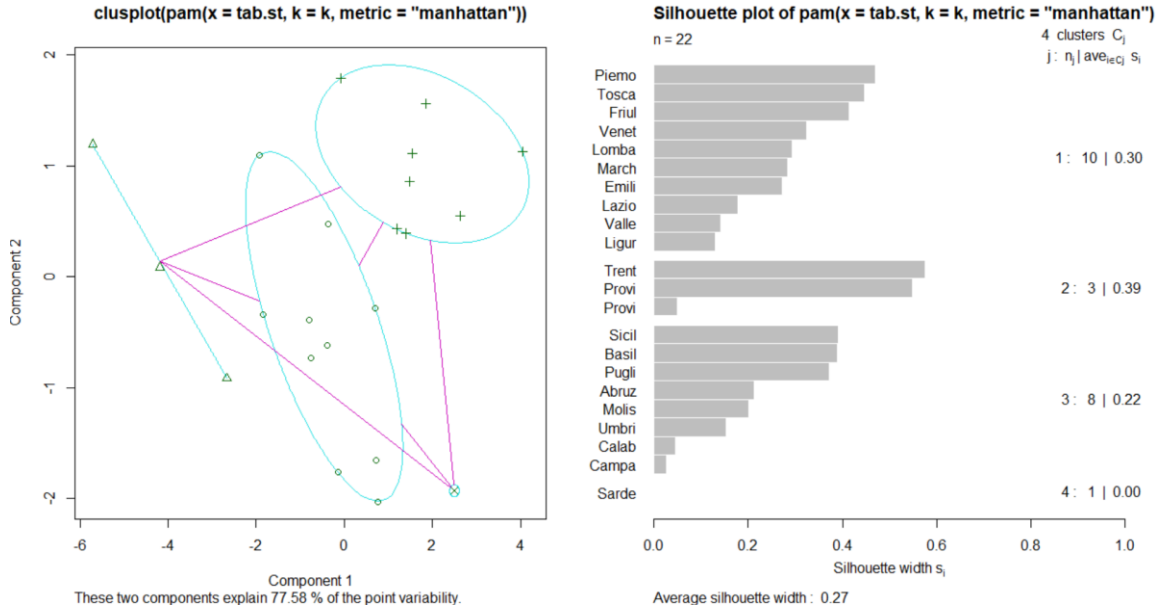

Figura 8: Grafici ottenuti per k=4

Dal grafico di figura 6 possiamo dire che per k=2 i valori della silhouette media relativa ad ogni cluster risultano paragonabili, Infatti, notiamo l'assenza di individui con silhouette negativa. Possiamo invece notare dalla figura 7 e dalla figura 8 come per k=3 e k=4 la silhouette rimanga sostanzialmente la stessa mentre in entrambi i casi risulta essere leggermente inferiore rispetto al caso k=2. Per k=3 vediamo che è presente un elemento mal classificato mentre per k=4 non sono presenti individui mal classificati anche se possiamo notare la presenza di un cluster singolo che non risulta significativo poiché noi vogliamo avere gruppi di elementi omogenei.

Nel caso dell'utilizzo dell'algoritmo PAM con distanza manhattan possiamo dire che la suddivisione più significativa risulta essere quella per k=2.

2.1.3 Analisi con i metodi gerarchici

Analizziamo la nostra tabella usando i metodi di clustering gerarchico. Cominciamo con il vedere il dendogramma. Per prima cosa è necessario definire la distanza (in questo caso la distanza euclidea) del dataset tramite il comando: `> d<-dist(tab.st)`

La distanza che abbiamo appena definito ci dirà quanto differenti sono due osservazioni fra loro. Importante sarà anche il tipo di distanza fra cluster che impostiamo che ci dirà la somiglianza fra cluster. Otteniamo i vari dendogrammi applicando il metodo per ognuna delle possibili distanze tra cluster: complete linkage, average linkage e single linkage.

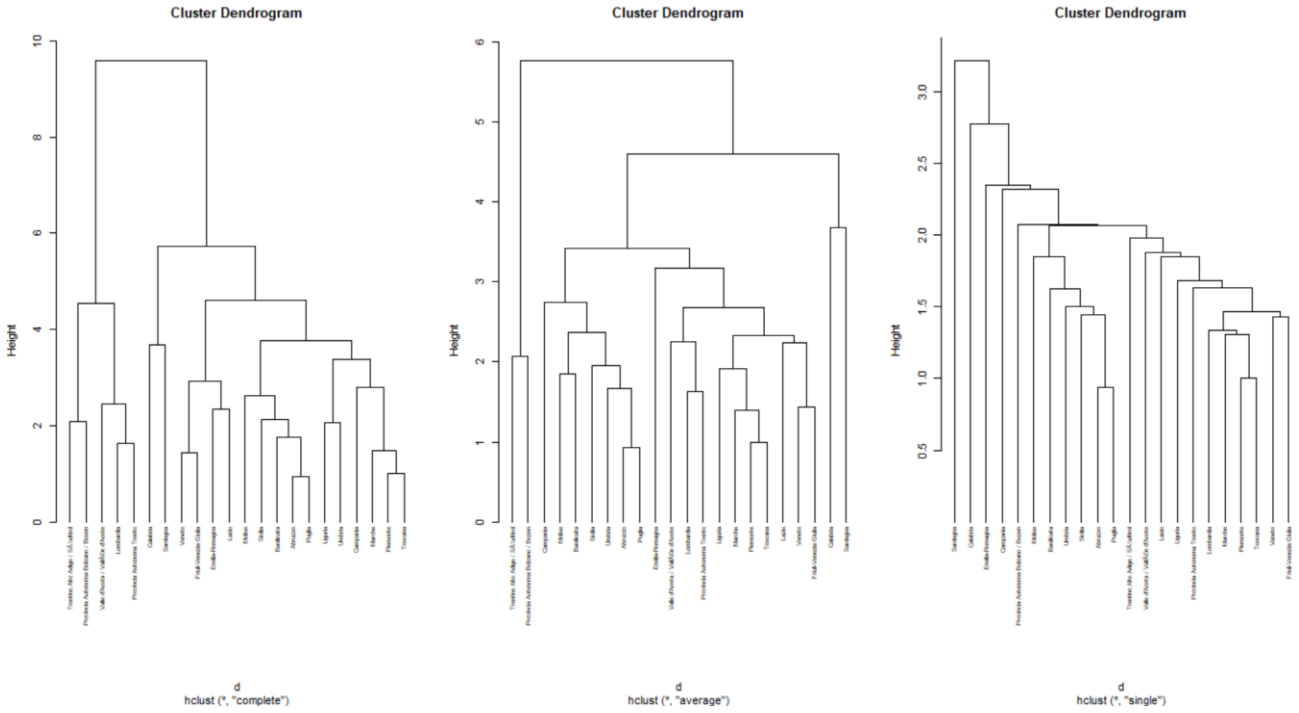

Figura 9: Dendogramma per complete,average e single linkage

Possiamo affermare che il single linkage non risulta molto interessante per il rilevamento di cluster sostanziali poiché vi è la presenza di molti cluster singoli. Vediamo che se chiediamo per esempio 10 cluster si nota che la maggior parte sono tutti costituiti da un singolo individuo e se diminuiamo il numero di cluster questi piano piano si fondono:

```
> table(cutree(tab.hcs,10))

 1  2  3  4  5  6  7  8  9 10
 9  1  1  1  1  5  1  1  1  1
> table(cutree(tab.hcs,6))

 1  2  3  4  5  6
17  1  1  1  1  1
```

Esaminiamo la silhouette globale per il metodo complete e average:

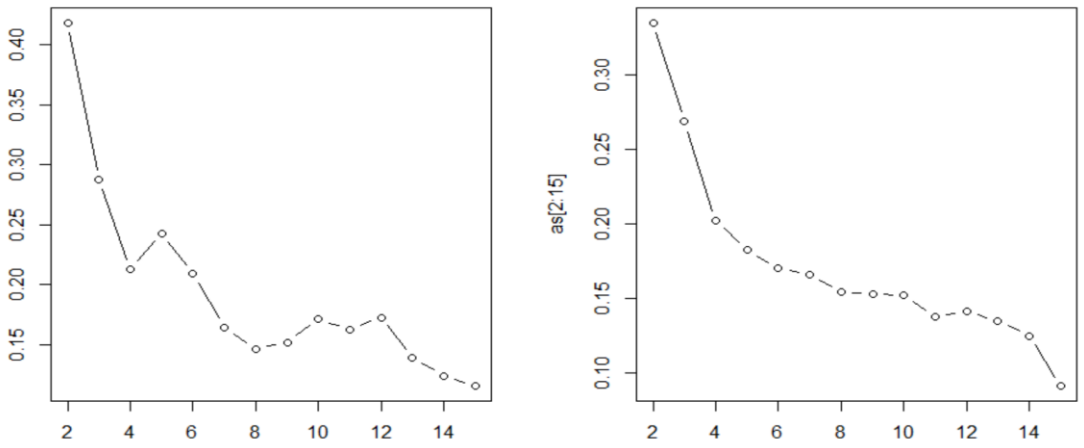

Figura 10: Andamento silhouette

Dalla figura 10 è possibile notare che sia nel caso del complete linkage (grafico a sinistra) che nel caso di average linkage (grafico a destra) si ha un picco in corrispondenza di 2 cluster in cui si registra in entrambi i casi il valore di silhouette più alto. Osserviamo in entrambi i grafici un andamento decrescente fino a tendere quasi a zero con l'unica differenza che nel caso del complete linkage si registrano diversi punti di picco però con un valore di silhouette più basso rispetto a 2 e 3 cluster. Quindi esaminiamo le suddivisioni per k=2,3. Riportiamo di seguito i grafici relativi alle silhouette:

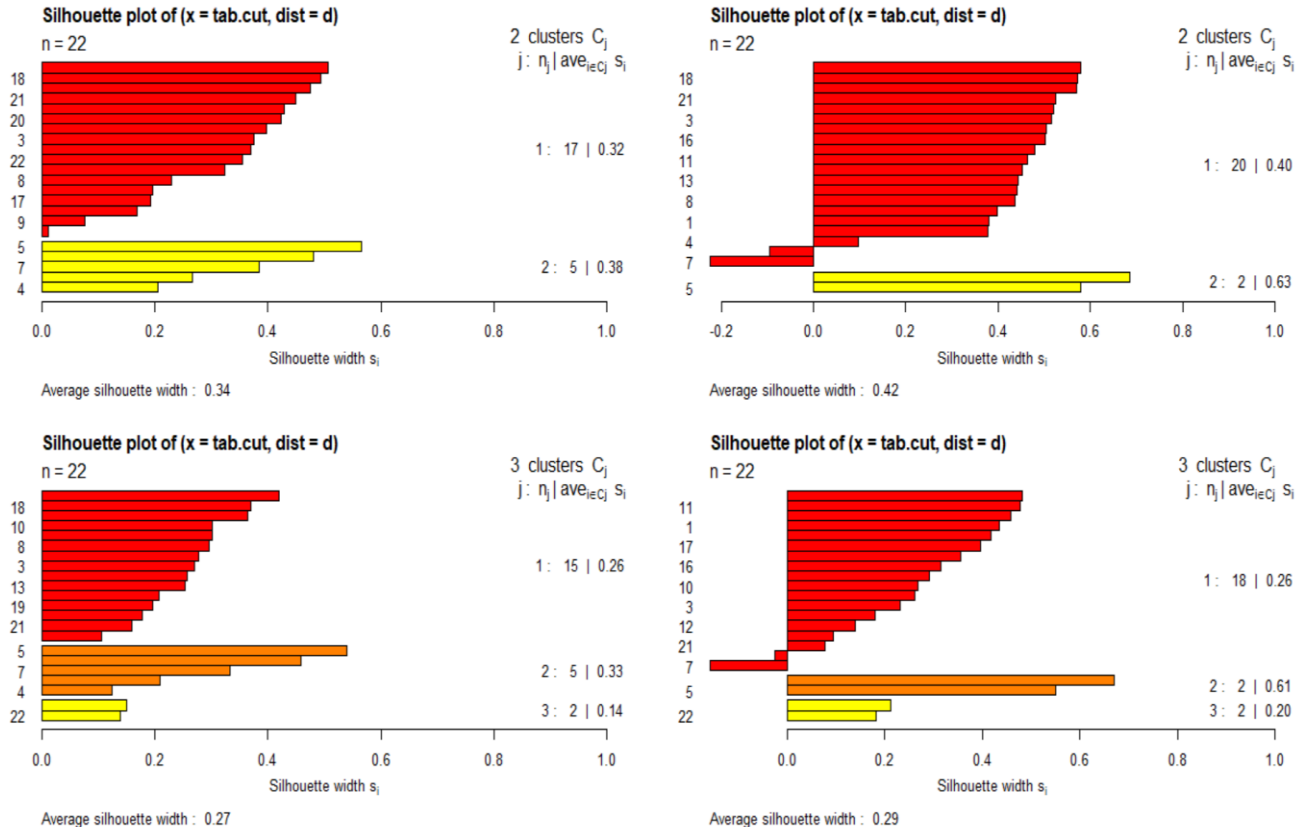

Figura 11: Grafici a sinistra metodo complete e grafici a destra metodo average

Dalla figura 11 possiamo notare, in entrambi i metodi per entrambi i casi, una certa disomogeneità nel numero di individui per ogni cluster. Osservando i grafici vediamo che il metodo average registra valori di silhouette media globale più alti in entrambi i casi rispetto al metodo complete. Vediamo però che nel caso dell'average linkage sono presenti, sia nella suddivisione in due cluster che nella suddivisione in tre cluster, individui mal classificati. Ad eccezione del caso k=2 del metodo complete (grafico in alto a sinistra) negli altri casi i valori delle silhouette per ogni cluster non risultano essere paragonabili. Quindi possiamo dire che il metodo average, nonostante presenti valori di silhouette apparentemente più buoni rispetto al metodo complete, non risulta soddisfacente.

Per completezza è stata effettuata l'analisi utilizzando la distanza manhattan. Non riportiamo i grafici ottenuti poiché notiamo una certa coerenza con quanto abbiamo ottenuto utilizzando la distanza euclidea.

2.2 Interpretazione

Dalle analisi che sono state effettuate riteniamo che gli algoritmi hanno restituito i risultati migliori specificando 2 come numero di cluster in output. Nonostante che il metodo PAM, con l'utilizzo della distanza manhattan, presenti valori di silhouette minori rispetto al metodo k-means e ai metodi di clustering gerarchico, restituisce delle suddivisioni più omogenee nel numero di individui per ogni cluster. Inoltre, il metodo PAM risulta essere anche più robusto rispetto agli altri metodi e quindi conseguentemente possiamo affermare che il metodo risulta essere il compromesso migliore per quanto riguarda la silhouette media

globale, le silhouette dei singoli cluster e l'omogeneità nel numero di individui per ogni cluster. In sintesi, riteniamo che il metodo PAM, con l'utilizzo della distanza manhattan, sia il più adatto per il raggiungimento del nostro obbiettivo.

Entriamo nel merito delle osservazioni ed esaminiamo come gli individui (ovvero le regioni italiane) vengono partizionati dall'algoritmo:

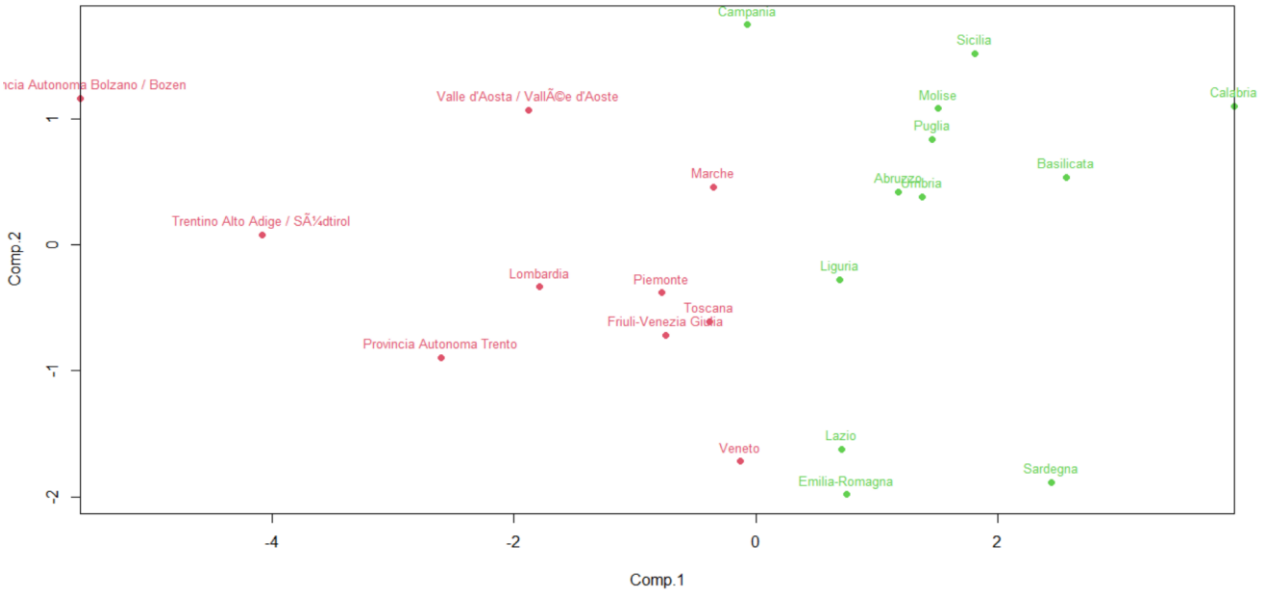

Figura 12: rappresentazione delle osservazioni partizionate dall'algoritmo pam

Vediamo che la distinzione tra i due gruppi sembra essere relativa alla posizione geografica delle regioni, notiamo infatti come il cluster 1 (gruppo definito dal colore rosso) sia costituito prevalentemente dalle regioni del nord Italia, il cluster 2 (gruppo definito dal colore verde) risulta essere costituito in gran parte dalle regioni del sud Italia mentre le regioni del centro Italia risultano essere distribuite fra questi due gruppi. Quindi l'interpretazione che possiamo dare in seguito alla nostra analisi è che regioni vicine hanno delle condizioni e stili di vita molto simili (in particolare la condizione economica) che vanno quindi ad influire in maniera significativa sullo stato di salute mentale degli individui. Infatti, anche sulla base dei risultati della prima relazione si evidenzi come il cluster 1 (gruppo rosso) sia composto da regioni che godono di una migliore condizione economica e di vita e quindi di salute mentale rispetto alle regioni del cluster 2 (gruppo verde).

3. Conclusioni

Abbiamo visto come la suddivisione in 2 cluster sembra essere quella che fornisce un prima risposta, nel contesto della nostra analisi e dell'obiettivo che vogliamo raggiungere, suddividendo le regioni a seconda delle caratteristiche delle condizioni di vita che vanno a influire in maniera significativa sullo stato di salute mentale generale degli individui nelle varie regioni italiane.

Possiamo quindi concludere come questa analisi risulta essere un buon punto di partenza da fornire a un team di persone, che disporranno così di una visione generale della situazione. Conseguentemente avranno uno strumento che permetterà loro di poter decidere dove intervenire con una maggiore urgenza attraverso campagne di intervento e sensibilizzazione per aiutare chi sta soffrendo cercando quindi di migliorare e prevenire varie condizioni che vanno ad influire sullo stato di salute mentale delle persone.