STATISTICA 2

SECONDA RELAZIONE

Mazzei Lorenzo

ANALISI

PROBLEMA

Si propone di studiare l'utilizzo dei mezzi pubblici e privati per gli spostamenti dei lavoratori, nello specifico l'interesse è vedere se si possono trarre delle conclusioni sulle preferenze di alcune regioni sui mezzi: vorremmo cercare di capire dal punto di vista del trasporto su quali mezzi dovrebbero essere fatti più investimenti per agevolare gli spostamenti verso il luogo di lavoro. Come metodo di studio del problema si è scelto il Clustering: vediamo se ci sono scelte comuni a seconda delle regioni in cui ci troviamo.

DESCRIZIONE DELLE COMPONENTI DELLA TABELLA

Dimensioni tabella: 36 osservazioni, 14 variabili
Link tabella: http://dati.istat.it/index.aspx?queryid=16502

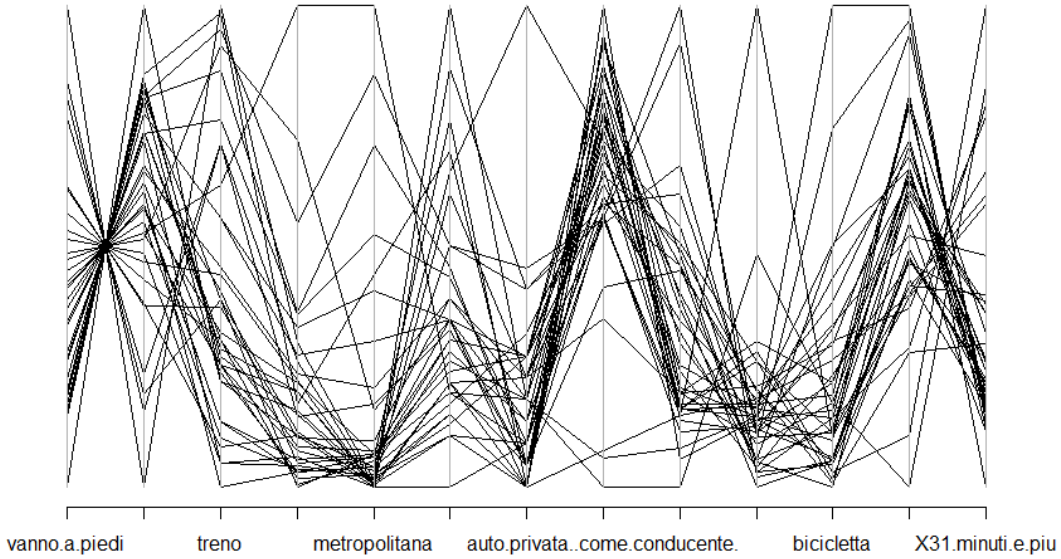
Le features della tabella rappresentano i mezzi usati dai lavoratori, c'è da prestare attenzione al fatto che sono presenti features che concernono lo stesso mezzo (auto e pullman) ma facendo distinzione tra Auto guidata come conducente e come passeggero, Pullman pubblico e privato. Si è scelto di selezionare come periodo di studio l'anno del 2020. Di seguito vengono riportate le colonne:

1) **Territorio**: regioni/parti d'Italia in cui sono stati raccolti i dati
2) **Vanno a piedi**: tasso di persone che va a piedi a lavoro
3) **Usano mezzi di trasporto**: tasso di persone che usa i mezzi di trasporto per andare a lavoro
4) **Treno:** tasso di persone che usa il treno per andare a lavoro
5) **Tram/Bus**: tasso di persone che usa il tram per andare a lavoro (il bus è stato assimilato al tram da chi ha realizzato la tabella)
6) **Metropolitana**: tasso di persone che usa la metropolitana per andare a lavoro
7) **Pullman/Corriera**: tasso di persone che usa il pullman per andare a lavoro (la Corriera è stata assimilata al Pullman da chi ha realizzato la tabella)
8) **Pullman Aziendale**: tasso di persone che usa un pullman privato dell'azienda per andare a lavoro
9) **Auto Privata** (**come conducente)**: tasso di persone che usa la propria auto per andare a lavoro
10) **Auto Privata (come passeggero)**: tasso di persone che si fa dare un passaggio in auto per andare a lavoro
11) **Motocicletta/Ciclomotore**: tasso di persone che usa la motocicletta per andare a lavoro (il Ciclomotore è stata assimilato alla Motocicletta da chi ha realizzato la tabella)
12) **Bicicletta**: tasso di persone che usa una bicicletta per andare a lavoro
13) **Fino a 15 minuti**: tasso di persone che impiega un tempo minore o uguale a 15 minuti per andare a lavoro
14) **31 minuti e più**: tasso di persone che impiega un tempo maggiore o uguale a 31 minuti per andare a lavoro

Apporto subito una modifica alla tabella andando a togliere il fattore "Territorio" in quanto categorico; rispetto alla relazione precedente si è scelto di togliere anche l'osservazione "Italia" perché è un dato generale nato da una media che non è utile allo studio delle singole regioni.

STUDIO INIZIALE

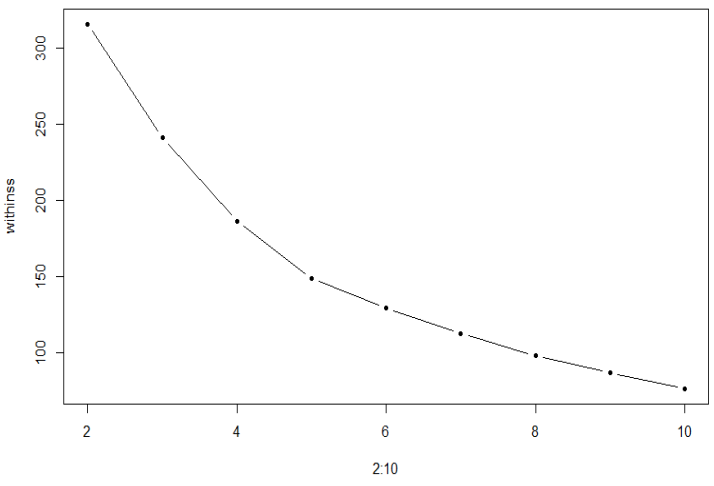
Eseguiamo un'analisi iniziale con una visione globale dell'andamento dei dati utilizzando il grafico del Parcoord cercando di capire se sono già notabili interessanti andamenti delle osservazioni.

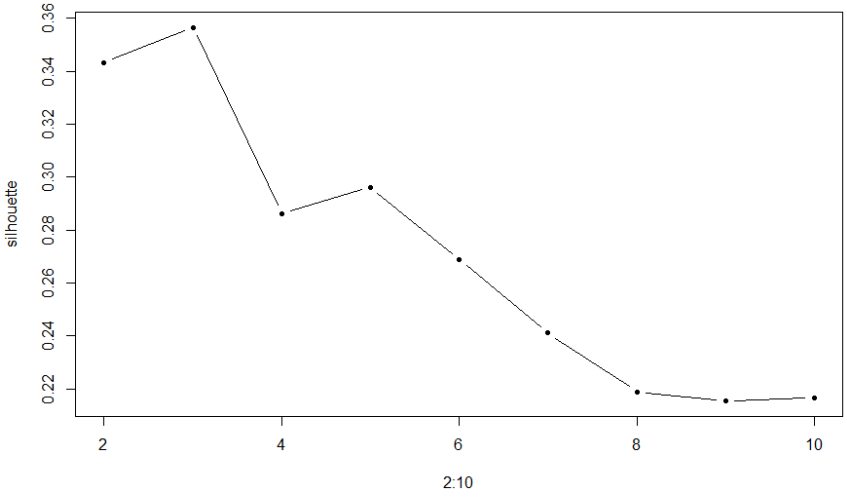


Il Parcoord non ci dice molto su eventuali Cluster, possiamo solamente osservare dei picchi sul fattore "auto come passeggero" e altre concentrazioni di valori su altri fattori. Possiamo però notare una suddivisione delle osservazioni per l'ultimo fattore "X31 minuti": un agglomerato di elementi per valori bassi e un fascio di osservazioni per valori più alti del fattore; possiamo allora aspettarci nelle successive partizioni in Cluster che queste divisioni appartengano a Cluster differenti.

K-MEANS

Possiamo allora iniziare un'analisi preliminare del metodo K-Means andando a guardare l'andamento della somma totale dei quadrati delle distanze tra gli elementi dei Cluster e al contempo l'andamento della Silhouette.

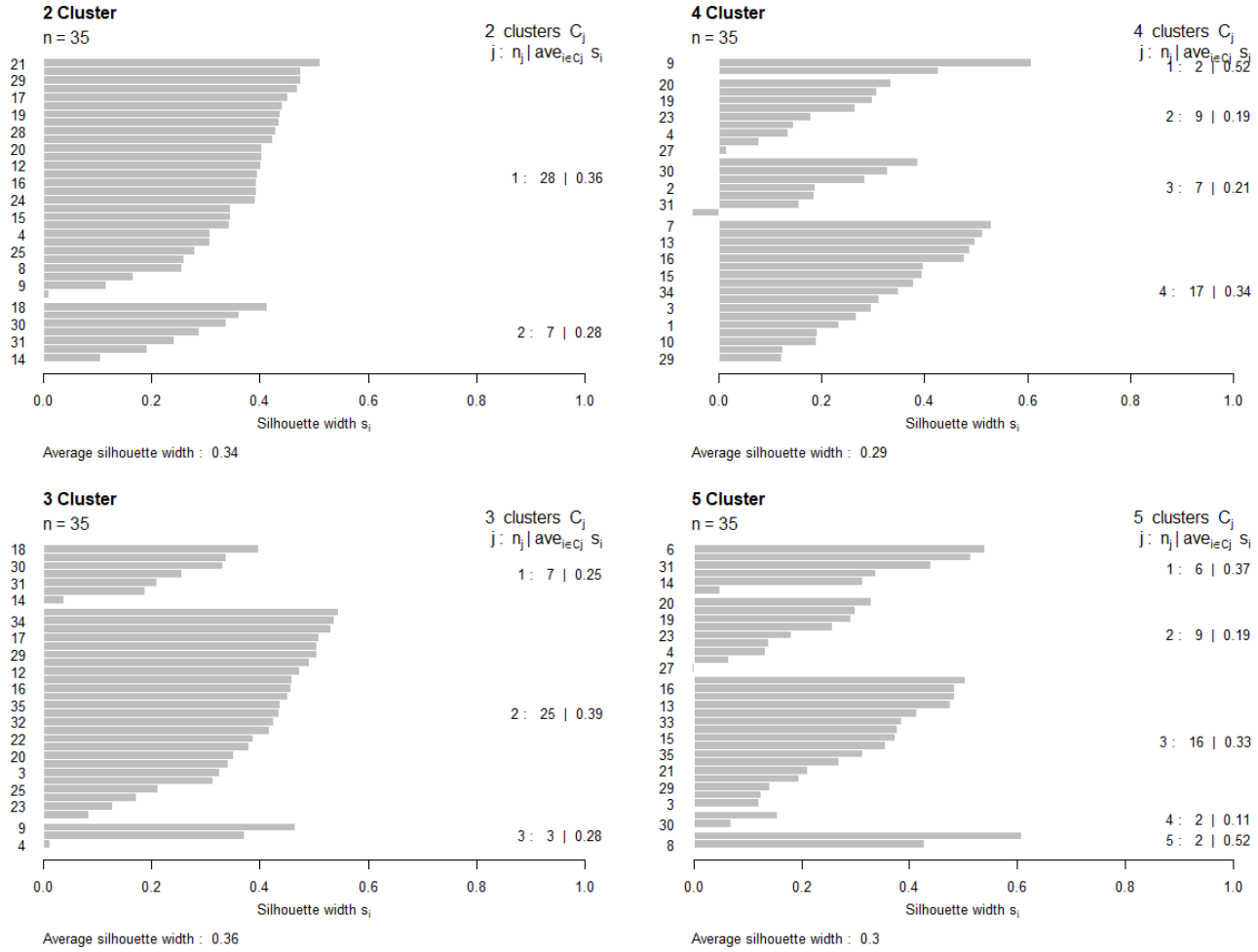




Si può osservare che il numero di Cluster da prendere in considerazione varia tra 2 e 5; la Silhouette per un numero di Cluster superiore inizia ad essere troppo bassa e dunque non conveniente da analizzare.

SILHOUETTE

Andiamo dunque ad osservare la conformazione dei Cluster nei casi corrispondenti a K =2, K = 3, K = 4 e K = 5. Studiamo quindi la Silhouette tramite i grafici riportati qui di seguito:

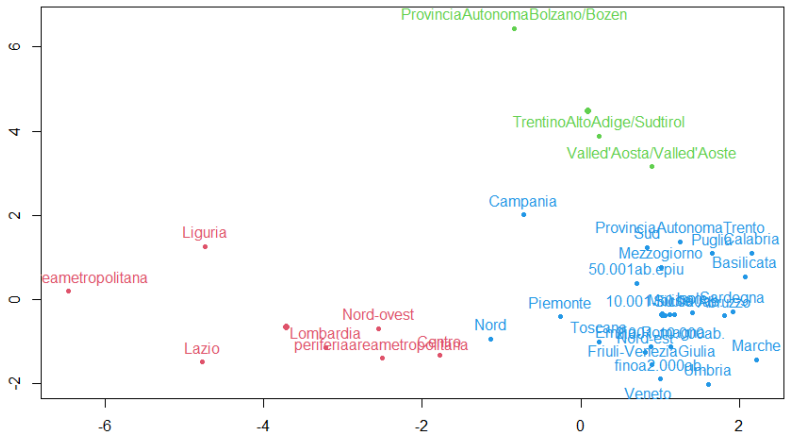


Considerando il motivo dello studio del problema, una suddivisione in almeno 3 Cluster sarebbe auspicabile per poter trarre delle interpretazioni su qualche regione, dunque i casi che verranno esaminati sono quelli per K = 3, K = 4 e K = 5, questa scelta è possibile anche perché la differenza di silhouette globale dal caso K = 2 non è così significativa (anzi con tre partizioni la silhouette è in realtà superiore).

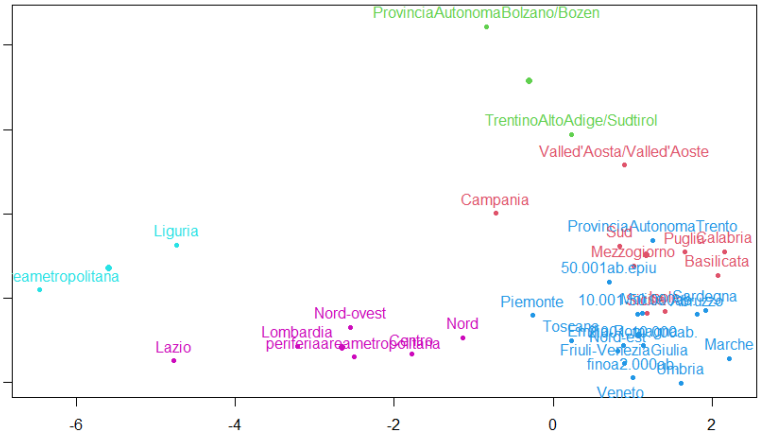
In tutti e tre questi scenari compare almeno un Cluster con 2-3 elementi. Può risultare interessante studiare queste partizioni con così pochi elementi per capire se rappresentano una serie di Outlier rispetto al cluster più numeroso e quindi se ha effettivamente senso separarli: ciò può essere fatto tramite un plot dei Cluster da cui possiamo cercare di trarre delle possibili interpretazioni.

PLOT DEI CLUSTER

3 Cluster (K = 3)



5 Cluster (K = 5)



INTERPRETAZIONI

K = 3

Notiamo come il cluster piccolo è composto dagli elementi evidenziati in verde nel grafico. L'osservazione "**Provincia Bolzano**" è molto distante da tutte le altre osservazioni e questo potrebbe portare a "tirare" verso di sé alcuni elementi nella realizzazione dei vari Cluster (in questo caso "**Trentino**"). A livello di interpretazione questo potrebbe essere spiegabile col fatto che il dato non rappresenta una regione o un dato generale su una zona d'Italia o ancora un range di numero di popolazione ma bensì una provincia autonoma, quindi è possibile che abbia una distribuzione dei dati significativamente diversa da quella delle regioni. Può risultare interessante allora ricompiere l'analisi provando a togliere il dato per osservare eventuali cambiamenti.

Se si effettua questa cosa risulta in realtà che la Silhouette diminuisce sia se decidiamo di fare 3 cluster o più (nel caso di 3 cluster la Silhouette diminuisce addirittura di 0.1): questa cosa non ci garantisce assolutamente una migliore interpretazione del problema e risulta anche in una minore coesione delle partizioni, quindi si può concludere che l'osservazione che abbiamo provato a togliere ci dia in realtà una visione utile. In effetti Bolzano si trova in Trentino e il fatto che compaia nello stesso Cluster di quest'ultimo può essere una conferma che i dati globali della regione sono coerenti con quelli delle singole città.

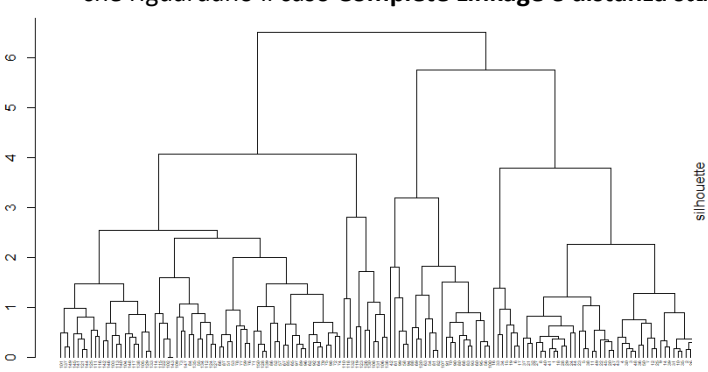
Si può notare inoltre che esiste un'altra osservazione che rappresenta una provincia autonoma ("**Trento**") che invece appartiene al Cluster più numeroso invece che essere in quello del Trentino, ma non avrebbe senso considerarlo un Outlier dato che si integra bene col Cluster, dunque non eliminiamo il dato.

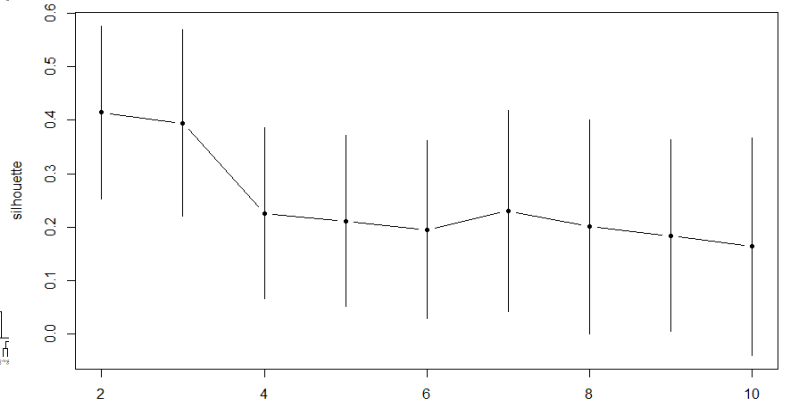
K = 5

Oltre a quando detto per il caso K = 3, qua c'è da considerare anche che la Clusterizzazione non risulta efficace: come si evince dal grafico le partizioni non sono ben distinte, anzi il Cluster rosso è inglobato in parte in quello blu, questo di fatto rende inutile questo partizionamento perché se i gruppi non sono abbastanza distinti tra loro non possiamo concludere niente. La stessa situazione si verifica in realtà anche per K = 4, il cui plot non è stato quindi riportato per pulizia grafica (unica differenza il cluster di "Liguria" era accorpato al Cluster di "Lazio") per cui dobbiamo scartare entrambi i casi K = 4 e K = 5.

METODI GERARCHICI

Proviamo a vedere ora se possiamo trarre conclusioni differenti col Cluster gerarchico. Dopo aver calcolato il Dendrogramma, vediamo come varia la Silhouette al variare del numero di Cluster tramite i due grafici seguenti che riguardano il caso **Complete Linkage** e **distanza standard**:



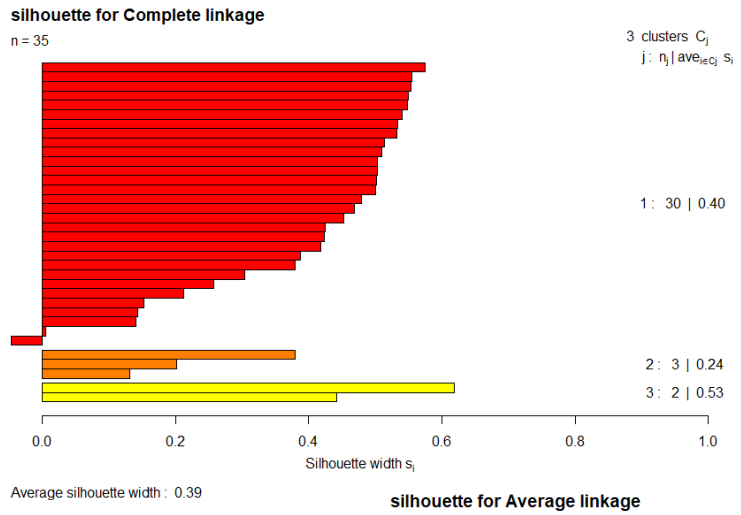


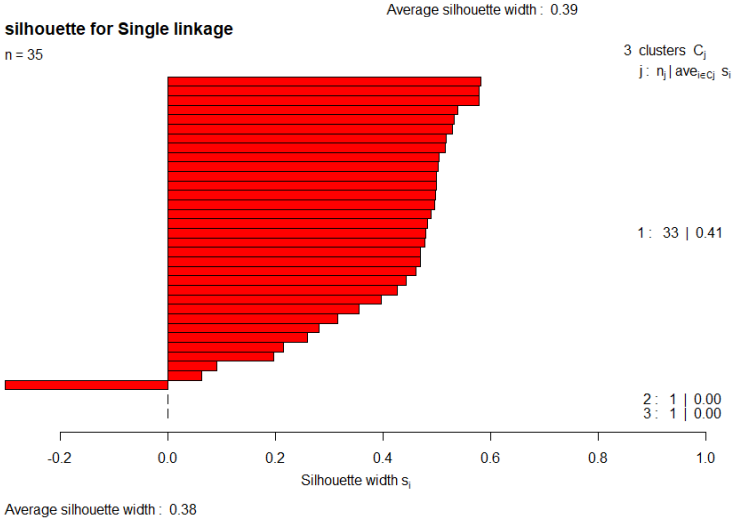
In questo caso 2 o 3 cluster sono la scelta migliore. Provando a considerare anche le alternative con il Single e l'Average Linkage l'alternativa valida rimane comunque tra 2 o 3. Risulta chiaro dunque che dobbiamo concentrarci a prescindere su queste due opzioni.

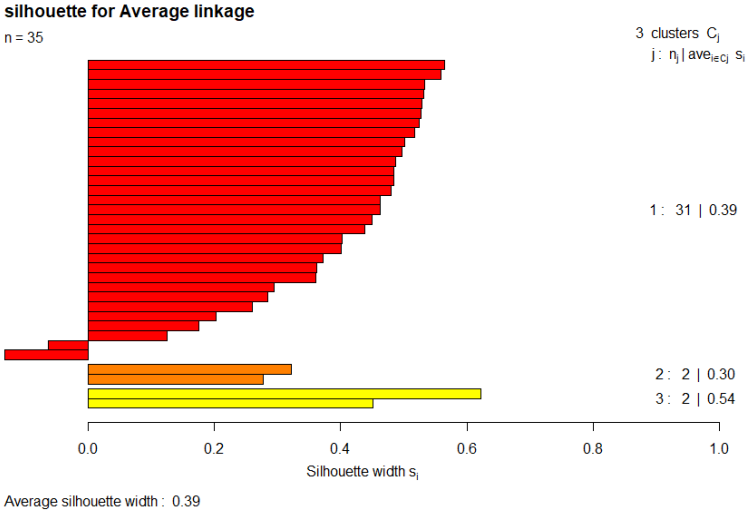
Single Linkage: con due Cluster uno dei due risulta di una sola osservazione, con tre cluster due dei tre risultano con una osservazione; è chiaro che questo output non ci serve a niente.

Average Linkage: migliore del caso precedente anche de di poco, difatti ora ci sono 2 osservazioni nei cluster che nel Single Linkage ne avevano solo una.

Complete Linkage: Caso migliore, con due cluster ne abbiamo uno da tre osservazioni mentre con tre Cluster ne abbiamo uno da due e uno da tre.



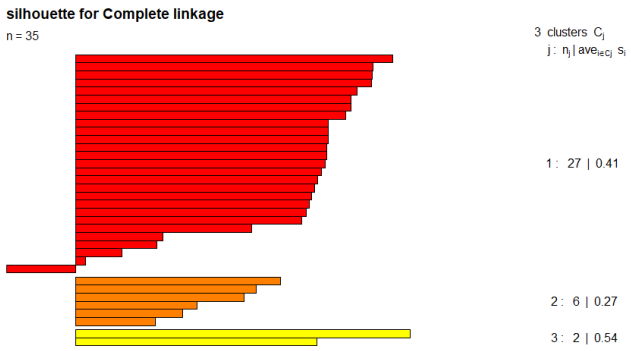




DISTANZE

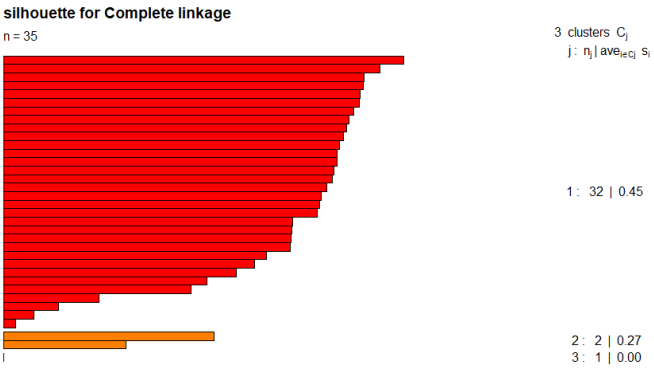
Quanto visto finora sono i risultati eseguiti con la distanza standard, vediamo per esempio i risultati con la distanza Manhattan e quella del Massimo per il **Complete Linkage**:

MANHATTAN



La distanza Manhattan fornisce risultati molto simili a quelli ottenuti dal K-Means, l'unica differenza sostanziale è che la Valle d'Aosta viene accorpata col Cluster più numeroso. Per le considerazioni fatte in precedenza però, risulta più coerente l'accorpamento della Valle d'Aosta col cluster più piccolo

MASSIMO



La distanza del Massimo invece ha come output una combinazione di Cluster pessima: una delle partizioni è una singola osservazione e l'altra è di due elementi. Si evince che questa distanza non ci dà alcuna informazione utile.

Per quanto riguarda gli altri casi, per i 2 Cluster meno numerosi abbiamo:

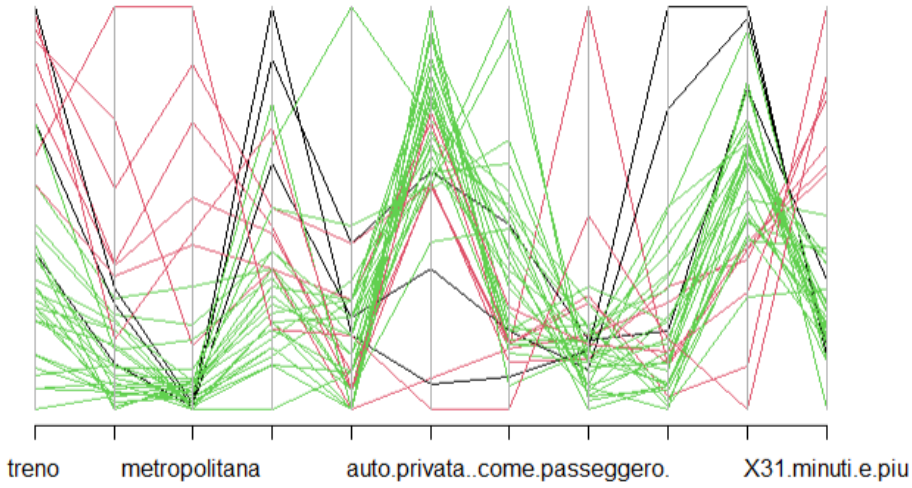
- **Single linkage**
    a) Maximum: entrambi i Cluster hanno 1 elemento
    b) Manhattan: entrambi i Cluster hanno 1 elemento
- **Average linkage**
    a) Maximum: un Cluster da 2 elementi e l'altro da 1 elemento
    b) Manhattan: entrambi i Cluster hanno 2 elementi

Dunque per Single linkage e Average Linkage il computo di queste distanze non dà alcun risultato degno di nota.

OSSERVAZIONE

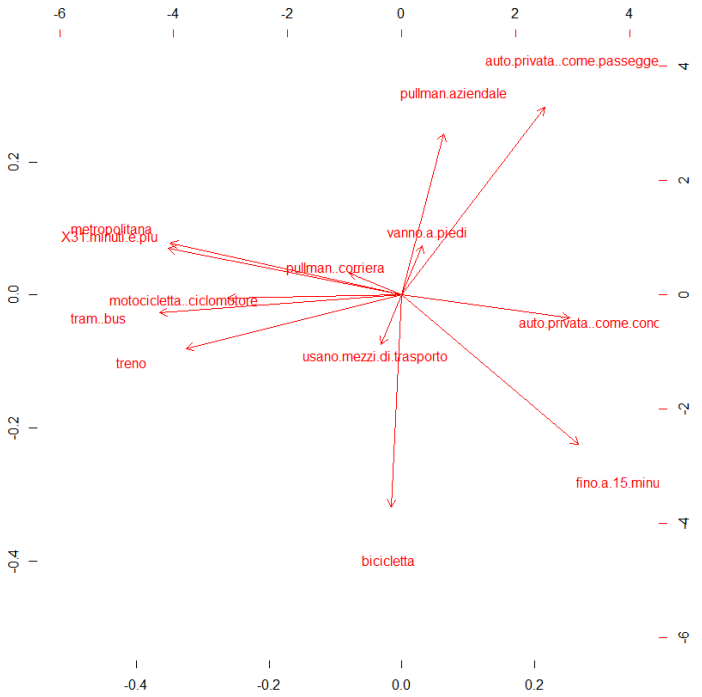
Abbiamo visto che la scelta migliore è quella con 3 Cluster sia per K-Means che con i metodi Gerarchici.
Per il resto i metodi gerarchici non forniscono risultati molto soddisfacenti, l'unico output più interessante è quello relativo alla distanza Manhattan che è in realtà molto simile come detto a quello ottenuto col K-Means.
Tra queste due possibili soluzioni possiamo concludere che il modello che risulta migliore è quello del K-Means: ha una silhouette globale più piccola ma la differenza non è significativa (0.03), ma ci garantisce un elemento in più in entrambi i Cluster più piccoli e favorisce l'interpretazione finale.

Andiamo allora a trarre le conclusioni finali: eseguiamo il Parcoord di nuovo ma a seguito delle analisi fatte in precedenza e visualizzando l'andamento dei tre Cluster ottenuti dal metodo K-Means.



Notiamo che il cluster rosso (quello di 7 elementi) rappresenta i territori in cui viene impiegato più tempo per andare a lavoro mentre quelle verde e nero ci mettono un tempo minore. Potremmo cercare di spiegare i cluster in funzione delle distanze da percorrere.

In tal senso possiamo usare il Biplot per aiutarci con una possibile interpretazione (si noti che la figura risulta differente rispetto a quella della passata relazione perché si ha un numero significativamente minore di osservazioni)



CONCLUSIONE

Proviamo a trarre delle spiegazioni finali per il problema: sicuramente il cluster di 7 elementi (rosso) segue l'andamento della prima componente principale (in seguito al plot degli scores infatti questa conclusione resta valida anche col piano tra la prima e la terza componente principale). Si potrebbe concludere quindi che in "**Liguria**","**Lombardia**" e "**Lazio**" ci sia la tendenza ad utilizzare i mezzi in generale più veloci perché si devono percorrere distanze significative per andare a lavoro (da cui "**più di 31 minuti**" è concorde a questa componente). Per quanto riguarda il Cluster di 3 elementi, può essere spiegato con la seconda componente

principale (il plot tra questa e la terza componente principale conferma questa ipotesi): esso appartiene alla sola parte positiva della componente, quindi si potrebbe concludere che il "**Pullman aziendale**" e l"**auto da passeggero**" siano mezzi usati per la maggiore in "**Valle d'Aosta**" e in "**Trentino**", con una componente anche nell'andare "**a piedi**" a lavoro: in effetti da alcune ricerche personali sul web risulta che in Valle d'Aosta siano stati creati dei programmi per ridurre l'emissione di gas serra avvalendosi del fatto che Aosta non è una città molto grande, quindi incentivando l'andare a piedi o fornendo dei siti web specifici per prendere più familiarità con i pullman di linea. Infine per il cluster più numeroso non a riusciamo ad avere un'interpretazione abbastanza solida a differenza dei casi precedenti: dobbiamo prendere la parte positiva della prima componente principale e l'altra componente, esse rientrano nella categoria dei "fino a 15 minuti" e considerando quanto già detto prima per il cluster di 3 elementi, potremmo ipotizzare che in queste rimanenti regioni prevalga l'utilizzo di mezzi in effetti non ad alta velocità (i mezzi del cluster meno numeroso , l'auto e la bicicletta) in concordanza con il tempo medio per andare a lavoro minore, anche se era auspicabile ottenere un'interpretazione più definita.

APPENDICE

CODICE (commenti segnati con il simbolo '#')

#caricamento tabella,eliminazione fattore categorico "territori"; parcoord dei dati

```
tabellaOriginale = tabellaOriginale[-c(1),]

tabellaOsservata = data.frame(scale(tabellaOriginale[,-c(1)]))

parcoord(tabellaOsservata)
```

#Plot della silhouette

```
wss=rep(0,10)

for(k in 2:10){

wss[k]=kmeans(tabellaOsservata,k,nstart=30)$tot.withinss }

plot(2:10,wss[2:10],type="b",ylab = "withinss",pch=20)

as=rep(0,10)

for(k in 2:10){

  cl=kmeans(tabellaOsservata,k,nstart=30)$cluster

  as[k]=mean(silhouette(cl,dist(tabellaOsservata))[,3]) }

plot(2:10,as[2:10],type="b",ylab = "silhouette",pch=20)
```

#plot dell'andamento dei Cluster

```
layout(matrix(1:4,2,2))

plot(silhouette(kmeans(tabellaOsservata,2,nstart=30)$cluster,dist(tabellaOsserva
ta)),main = "2 Cluster")

plot(silhouette(kmeans(tabellaOsservata,3,nstart=30)$cluster,dist(tabellaOsserva
ta)),main = "3 Cluster")

plot(silhouette(kmeans(tabellaOsservata,4,nstart=30)$cluster,dist(tabellaOsserva
ta)),main = "4 Cluster")
```

```
plot(silhouette(kmeans(tabellaOsservata,5,nstart=30)$cluster,dist(tabellaOsserva
ta)),main = "5 Cluster")

layout(1)

#plot dei Cluster

k=5 #fatto anche per k = 3

rownames(tabellaOriginale) <- tabellaOriginale$territorio

tabellaOsservata.km=kmeans(tabellaOsservata,k,nstart=30)

tabellaOsservata.pca=princomp(tabellaOsservata)

plot(tabellaOsservata.pca$scores,col=1+tabellaOsservata.km$cluster,ylim=c(-
2,6.6),pch=20)

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

text(tabellaOsservata.pca$scores,labels=(as.character(rownames(tabellaOriginale)
)),col=1+tabellaOsservata.km$cluster,pos=3)

#Dendrogramma

d<-dist(tabellaOsservata)

tabellaOsservata.hc=hclust(d)

plot(ir.hc,hang=-1,cex=0.3)

#Silhouette

tabellaOsservata.hc=hclust(d) #complete linkage,fatto anche per single e average

as=matrix(ncol=2,nrow=10)

for(i in 2:10){

  tabellaOsservata.cut=cutree(tabellaOsservata.hc,i)

  as[i,1]=mean(silhouette(tabellaOsservata.cut,d)[,3])

  as[i,2]=sd(silhouette(tabellaOsservata.cut,d)[,3])

}

as2=as[2:10,]

ymin=min(as2[,1]- as2[,2])

ymax=max(as2[,1] + as2[,2])

plot(2:10,as2[,1],ylab = "silhouette",type="b",pch=20,ylim=c(ymin,ymax))

segments(2:10,as2[,1]-as2[,2],2:10,as2[,1]+as2[,2])

tabellaOsservata.cut=cutree(tabellaOsservata.hc,3)

plot(silhouette(tabellaOsservata.cut,d),col=heat.colors(3),border=par("fg"))

#BIPLOT
```