STATISTICA 2

PRIMA RELAZIONE

Mazzei Lorenzo

ANALISI

PROBLEMA

Si propone di studiare l'incidenza che hanno i mezzi pubblici e privati sugli spostamenti dei lavoratori, nello specifico l'interesse è da concentrarsi nel tempo impiegato da questi per andare a lavoro tramite, appunto, i mezzi. I dati sono stati raccolti nelle varie regioni Italiane.

La tabella trovata avrebbe in realtà 2 possibili fattori d'uscita utili allo scopo: una colonna riguardante un tempo impiegato inferiore a 15 min e una colonna su un tempo maggiore di 31 minuti. Si è scelto come caso di studio quella riguardante più di 31 minuti.

DESCRIZIONE DELLE COMPONENTI DELLA TABELLA

Dimensioni tabella: 108 osservazioni, 14 variabili

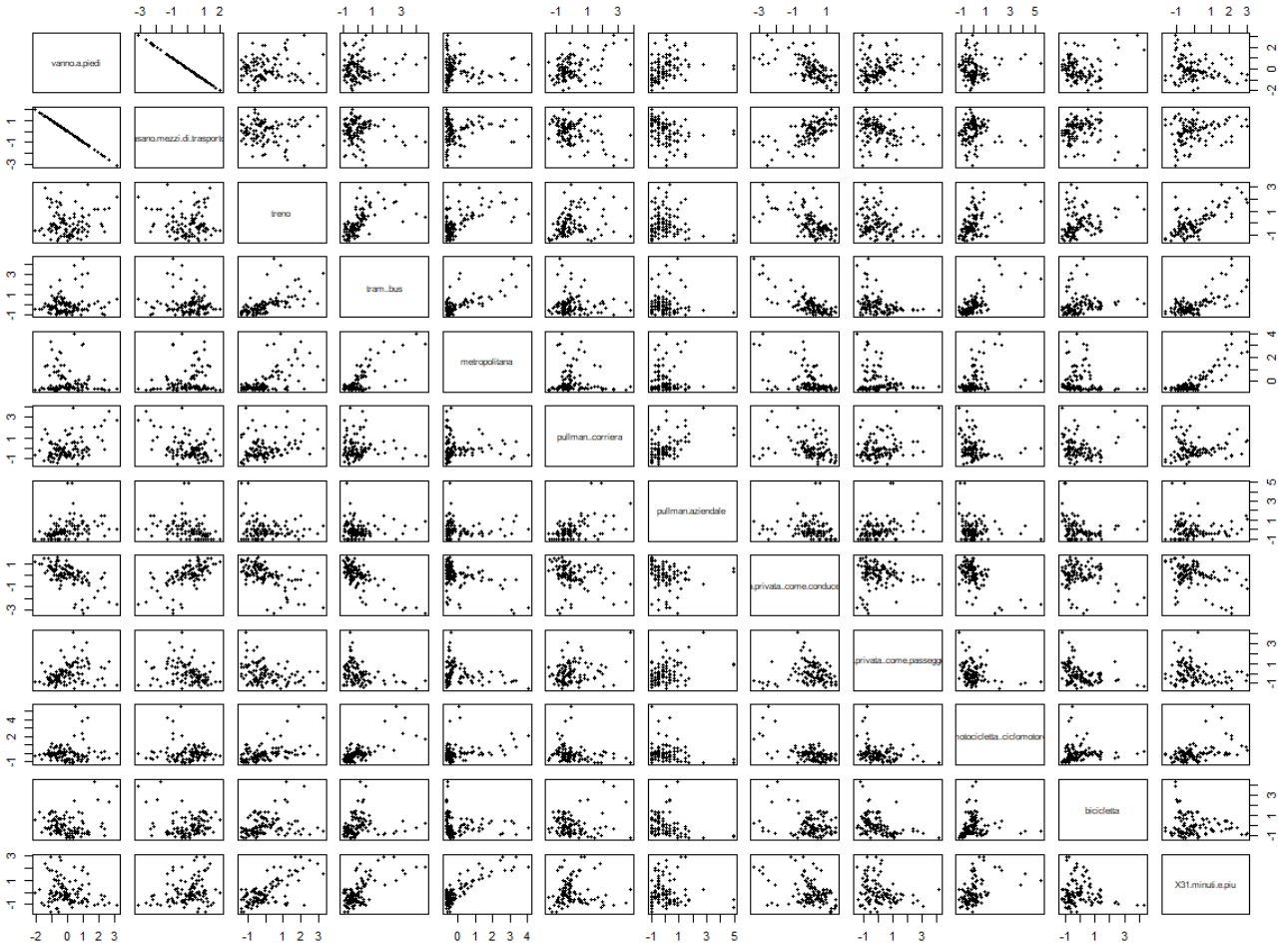
Le feature della tabella rappresentano i mezzi usati dai lavoratori, c'è da prestare attenzione al fatto che sono presenti features che concernono lo stesso mezzo (auto e pullman) ma facendo distinzione tra Auto guidata come conducente e come passeggero, Pullman pubblico e privato. Inoltre si è scelto di selezionare come periodo di studio dei datigli ultimi 3 anni (2019-2021). 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": esso infatti è categorico e dunque non compatibile col metodo di analisi scelto per studiare la tabella, ovvero la PCA.

STUDIO INIZIALE

Iniziamo l'analisi con un plot della tabella per avere un'idea iniziale della relazione tra i vari dati:

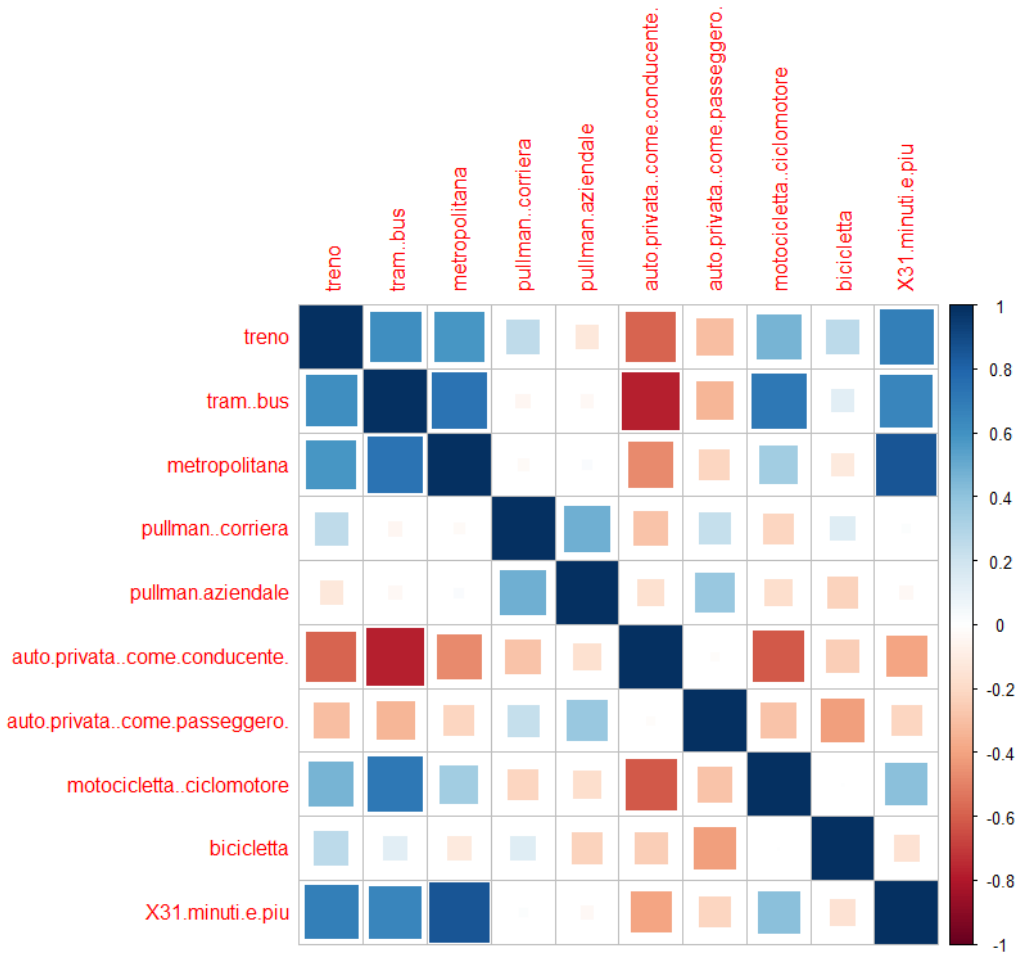


Il fattore di uscita che vogliamo osservare corrisponde all'ultima colonna del grafico. Il fattore "fino a 15 minuti" è stato tolto per concentrarsi su "X31 minuti e piu", su cui tra l'altro non influisce. Si può notare una corrispondenza abbastanza lineare con alcuni dei fattori mentre è meno marcata con i rimanenti. Da notare che invece i fattori "Vanno a piedi" e "Usano mezzi di trasporto" hanno una relazione perfettamente lineare: se osserviamo la tabella infatti emerge che questi fattori sono complementari tra loro, cioè la loro somma dà sempre 100 (ovvero 100% visto che sono dati in percentuale).

Ovviamente con solo questo grafico non possiamo dire molto quindi andiamo a vedere le correlazioni tra i vari fattori.

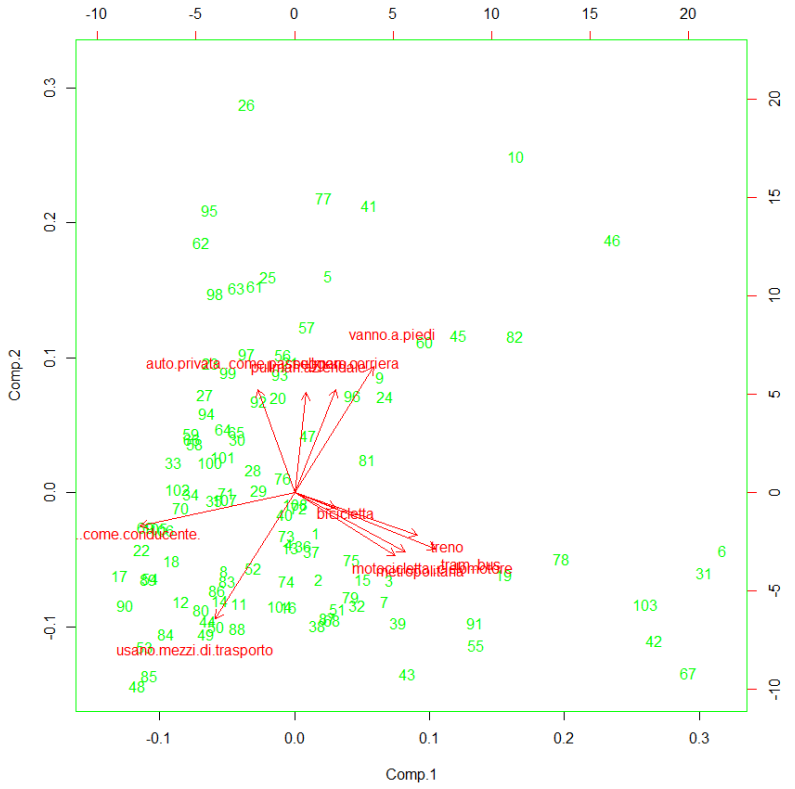
CORRELAZIONI

Nella matrice delle correlazioni si è scelto di non riportare a livello grafico le colonne “vanno a piedi” e “usano mezzi pubblici” per favorire una resa grafica più pulita e per il fatto che, come detto in precedenza, questi 2 fattori sono complementari: questo comporta una correlazione significativa tra di loro a scapito di quella con gli altri fattori. Dato lo scopo del nostro studio, ha senso concentrarsi (solo a livello grafico) solo sulle correlazioni dei mezzi di trasporto. Escluso questo caso grafico, i 2 fattori sono comunque inclusi nell’ analisi globale dei dati per fornire una corretta interpretazione di quest’ultimi.



Notiamo che le correlazioni più forti per la nostra variabile di uscita sono con “treno”, ”tram”, ”metropolitana”; minore per gli altri fattori. Importante è tenere presente che non necessariamente (anche se in buona parte dei casi sarà così) l’utilizzo di un mezzo esclude l’uso anche di un altro mezzo. Potremmo ipotizzare che alcune correlazioni tra i mezzi vadano a rappresentare proprio questo: realisticamente la metropolitana, il treno, i pullman ecc… possono difficilmente essere presenti vicino alla propria abitazione tranne che in pochi casi, di conseguenza l’ausilio di un altro mezzo può essere necessario per raggiungerli.

BIPLOT



Dal grafico del Biplot osserviamo che alcuni fattori sono abbastanza “agglomerati” nella parte positiva della prima componente principale: “bicicletta”,”treno”,”tram”,”metropolitana”,”motocicletta”; il loro allineamento come detto è verso la prima componente principale ma non è molto marcato, invece il fattore “Auto privata come conducente” è abbastanza allineato ad essa ma negativamente. I rimanenti osserviamo che invece sono più o meno orientati sulla seconda componente.

Da questo grafico in realtà non siamo in grado di trarre molti ragionamenti su come interpretare i dati, complice anche la complessità nell’assegnazione delle variabili. Dobbiamo vedere la matrice dei Loading per fornire un’interpretazione del problema. Prima di fare ciò analizziamo la proporzione di varianza spiegata fornite dalle varie componenti.

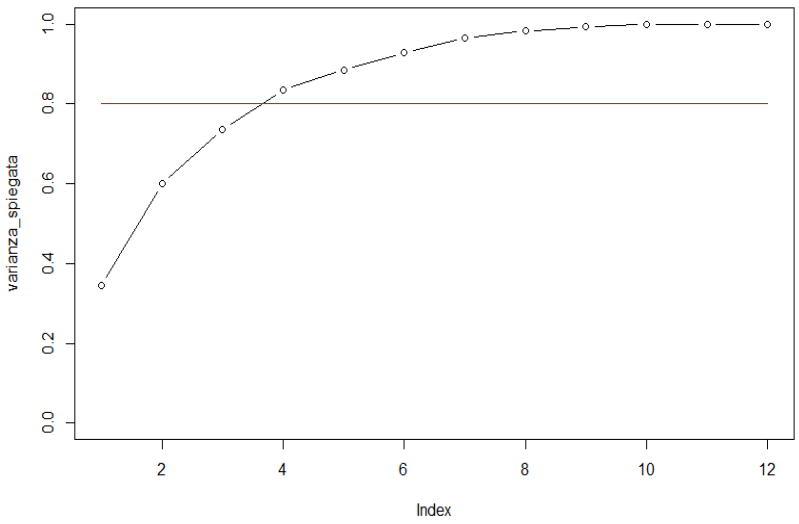
IMPORTANZA DELLE COMPONENTI

Studiamo l’importanza delle componenti per capire quali dobbiamo considerare per “spiegare” il problema.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6     Comp.7     Comp.8      Comp.9
Standard deviation     1.922186 1.6976803 1.2072760 1.0566955 0.76583935 0.70652587 0.66362120 0.47349546 0.288018786
Proportion of Variance 0.335891 0.2620108 0.1325014 0.1015096 0.05331908 0.04537989 0.04003574 0.02038163 0.007541347
Cumulative Proportion  0.335891 0.5979018 0.7304032 0.8319127 0.88523182 0.93061171 0.97064744 0.99102908 0.998570423
                           Comp.10      Comp.11
Standard deviation     0.125367835 2.873309e-03
Proportion of Variance 0.001428827 7.505365e-07
Cumulative Proportion  0.999999249 1.000000e+00
```

Notiamo che con le prime 3 componenti raggiungiamo una soglia vicina a quella empirica da noi cercata (risulta circa 73%), se prendiamo anche in considerazione la quarta allora la superiamo (83%). Questo ci fa osservare che le prime 4 componenti possono essere buone per “spiegare” il problema.

Per approfondire questa osservazione possiamo usare la matrice dei loading.

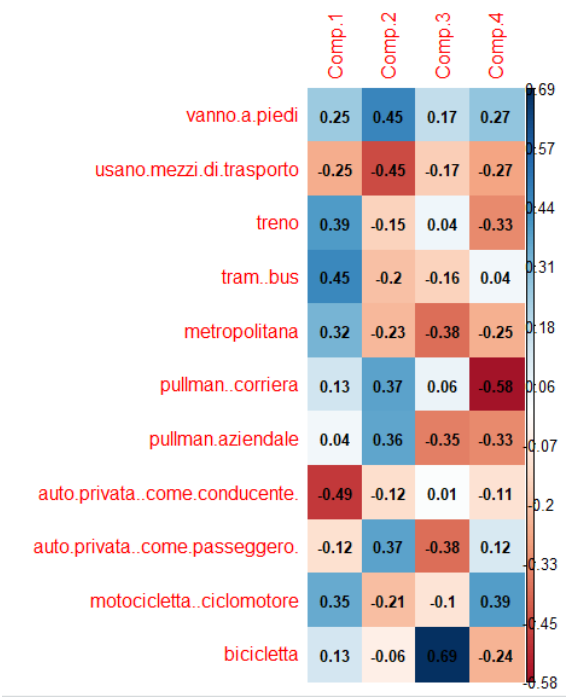


Con la resa grafica della varianza spiegata in base alle componenti possiamo notare come non ci sia un gomito marcato nella curva

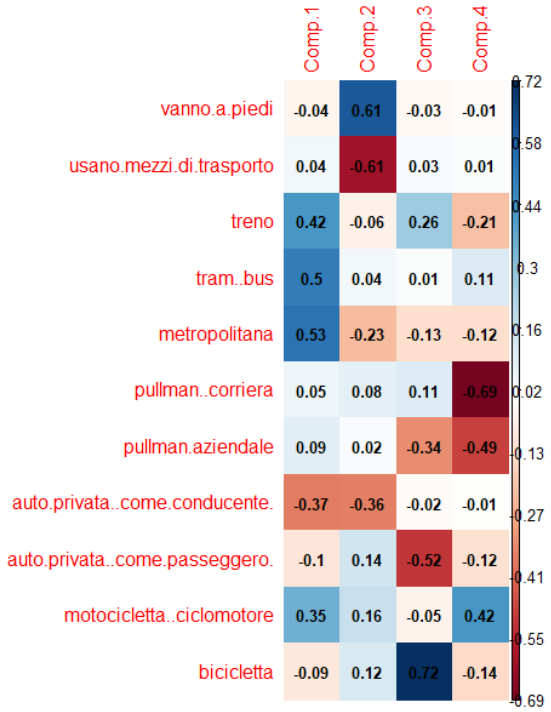
MATRICE DEI LOADING

Passiamo dunque alla matrice dei Loading per analizzare le possibili associazioni:

SENZA ROTAZIONE



CON ROTAZIONE



SENZA ROTAZIONE

Da questa prima osservazione non emerge un quadro ben definito, o meglio, alcune associazioni possono essere già delineate (“bicicletta”,”pullman/corriera”..) ma per le altre possiamo utilizzare il metodo delle rotazioni per vedere se siamo in grado di ottenere un risultato più definito.

CON ROTAZIONE

Si può osservare che le rotazioni hanno dato un contributo significativo per le associazioni: la matrice risulta significativamente più sparsa e di conseguenza l’assegnazione dei “ruoli” è più chiara.

POSSIBILE INTERPRETAZIONE

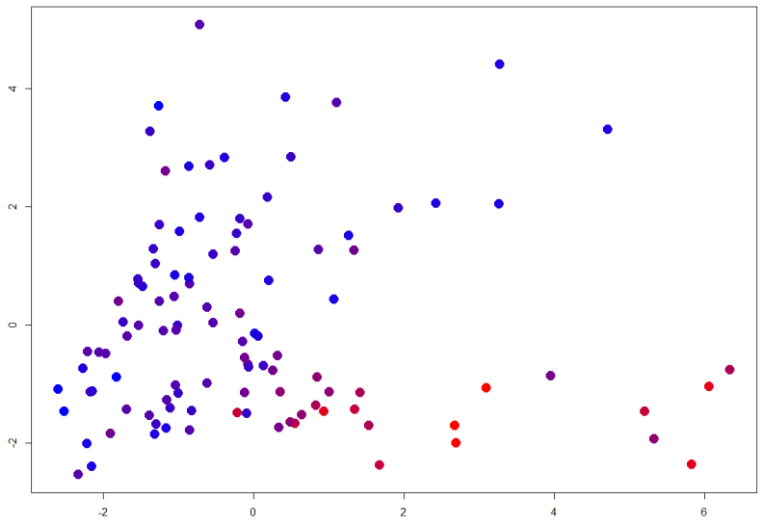
1ma comp. (treno,tram,metropolitana,auto come conducente): Si può vedere questa componente come quella che rappresenta i mezzi di trasporto più veloci e pensati per raggiungere distanze considerevoli (l'auto in realtà è opinabile da questo punto di vista, il suo utilizzo può essere infatti molto variegato in termini di distanze).

2da comp. (vanno a piedi,usano mezzi di trasporto): questa componente spiega il tasso di utilizzo di mezzi pubblici o meno da parte dei lavoratori. In particolare i 2 fattori hanno lo stesso "peso" a livello di associazione in quanto come già detto sono l'uno il complementare dell'altro.

3za comp. (auto come passeggero,bicicletta): questa componente è la più difficile da inquadrare, si potrebbe comunque ipotizzare che rappresenti la parte dei lavoratori che deve percorrere distanza abbastanza brevi, da cui l'utilizzo della bicicletta e/o il farsi dare un passaggio in macchina per essere lasciato a lavoro da un/una conoscente che data la breve distanza non ha problemi quindi a impiegare la sua vettura per questo scopo.

4ta comp. (pullman/corriera,pullman aziendale,motocicletta): Quest'ultima componente invece può essere vista come quella dedicata alle medie distanza, da cui l'utilizzo dei pullman che risultano molto comodi nello spostamento per esempio all'interno di città o zone limitrofe ad esse. La moto è allo stesso modo plausibile in questa ipotesi.

PIANO PRINCIPALE

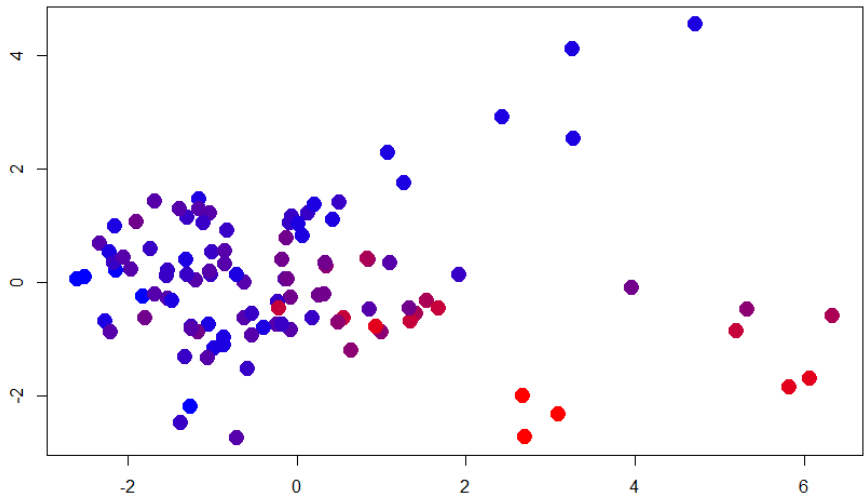


Qua possiamo osservare il piano principale e come i campioni si correlano ad esso. Nello specifico la tonalità del colore ci dà informazione del fattore "31 minuti e piu" rispetto al piano principale: il blu rappresenta i tassi più bassi di quel fattore mentre in rosso quelli più alti e coerentemente con quanto abbiamo osservato fin'ora, notiamo che più ci spostiamo a destra nel grafico più i punti tendono a diventare rossi ad indicare che il crescere delle componenti principali fa aumentare anche il tasso del fattore osservato.

OSSERVAZIONE PER L'INTERPRETAZIONE

Coma abbiamo visto in precedenza. La seconda componente principale esprime in realtà informazioni abbastanza generali, ovvero sui tassi di utilizzo dei mezzi di trasporto o meno. Possiamo pensare allora di poter usare un'altra componente al posto della seconda come riassuntiva del nostro problema. Ricordando che ci interessa associare i mezzi di trasporto a tempi superiori alla mezz'ora, possiamo usare la terza componente e vedere se otteniamo comunque risultati simili al piano principale.

PIANO TRA PRIMA E TERZA COMPONENTE



In effetti anche in questo caso la parte destra del grafico si comporta in modo simile a quella del piano principale e coerentemente con quanto detto fin'ora, per l'interpretabilità del problema la terza componente principale risulta preferibile alla seconda nell'aspetto riassuntivo finale.

CONCLUSIONI:

Fattore studiato: X31 minuti e più

Il problema preposto era studiare l'incidenza dei mezzi pubblici e privati sui tempi per andare a lavoro. L'utilizzo di quattro componenti principali ci ha permesso di superare la treshold empirica che volevamo (80%). Si è ipotizzato in questo caso che la prima,la seconda e la quarta componente principale fossero rappresentative delle categorie dei mezzi in base alle distanze da percorrere, mentre la seconda era più legata al rapporto tra l'uso dei mezzi di trasporto e l'andare a piedi. Dato che vogliamo favorire l'interpretabilità, piuttosto che osservare l'andamento dei campioni del fattore studiato sul piano principale può essere più utile usare il piano tra la prima e la terza componente. Da questo grafico possiamo trarre una possibile conclusione finale: la prima e la terza componente principale possono essere considerate come le componenti globale riassuntive per il fattore studiato, in particolare si può notare che il tasso del fattore studiato cresce all'aumentare della prima componente principale e contemporaneamente al diminuire della terza.

APPENDICE

CODICE (commenti segnati con il simbolo '#')

#caricamento tabella,eliminazione fattore categorico e fattore 'fino a 15 minuti'; plot della tabella

```
tabellaOriginale=read.csv("C:/Users/HP/OneDrive
/Desktop/TabellaPrimaRelazione.csv",header=T)

tabellaOsservata = tabellaOriginale[,-c(1,13)]

plot(tabellaOsservata,pch=20)
```

#Matrice di correlazione

```
corrplot(cor(tabellaOsservata[,-c(1,2)]),"square")
```

#Tolgo il fattore studiato, biplot, importanza delle componenti

```
tabellaOsservata = tabellaOsservata[,-c(12)]

tabellaOsservata.pca=princomp(tabellaOsservata,cor=TRUE)

biplot(tabellaOsservata.pca,choices=c(1,2),col=c("green","red"))

summary(tabellaOsservata.pca)

varianza_spiegata =
cumsum(tabellaOsservata.pca$sdev^2)/sum(tabellaOsservata.pca$sdev^2)

plot(varianza_spiegata,type="b",ylim=c(0,1))

segments(1,0.8,12,0.8,col="red")
```

#Matrice dei Loading (senza rotazione)

```
tabellaOsservata.loadings=loadings(tabellaOsservata.pca)

corrplot(tabellaOsservata.loadings,is.corr=F,method="color", addCoef.col =
"black",  number.digits=2, number.cex=0.8)
```

#Matrice dei Loading (con rotazione)

```
tabellaOsservata.rot=varimax(tabellaOsservata.loadings[,1:4])$loadings

corrplot(tabellaOsservata.rot, is.corr=FALSE,method="color", addCoef.col =
"black",  number.digits=2, number.cex=0.8)
```

#Plot del fattore studiato in relazione alle componenti principali

```
grad<-colorRampPalette(c("blue","red"))

scol=grad(10)

x=(X31minuti-min(X31minuti))/(max(X31minuti)-min(X31minuti))

sidx=1+floor(10*0.99*x)

prediction=tabellaOsservata.pca$scores[,c(1,2)] #ripetuto anche per c(1,3)

plot(prediction,pch=19,cex=2,col=scol[sidx],main=ncol)
```