Analisi PM10

Stefano e Cucca

Questo è un file r Markdown che permette di editare file di varia natura (PDF, Presentazioni di slide, pagine HTML, App) direttamente in un ambiente misto con linee di testo e linee di codice. Sotta "Help" trovi la voce "Markdown Quick Reference" e altre cose utili ma su internet ci sono pagine e pagine che ti spiegano come usarlo.

Qui sotto trovi un "chunk" di codice (tasto rapido ctrl+alt+i per inserirla) dove puoi scrivere i codici in r. Per far girare un codice funziona come uno script normale. Mentre se premi il tastino verde a destra ti runna tutto quello che sta in quella chunk.

Poi magari ci mettiamo una volta su skype e ti spiego un po meglio ma intanto i commenti te li metto qui che è più comodo.

PCA Analysis (con gli # crei i titoli con una scala gerarchica # ## ### #### ##### etc..). Guardati anche le cheatsheets sull'Help

```
### PCA ANALYSIS ----
require(ade4)
```

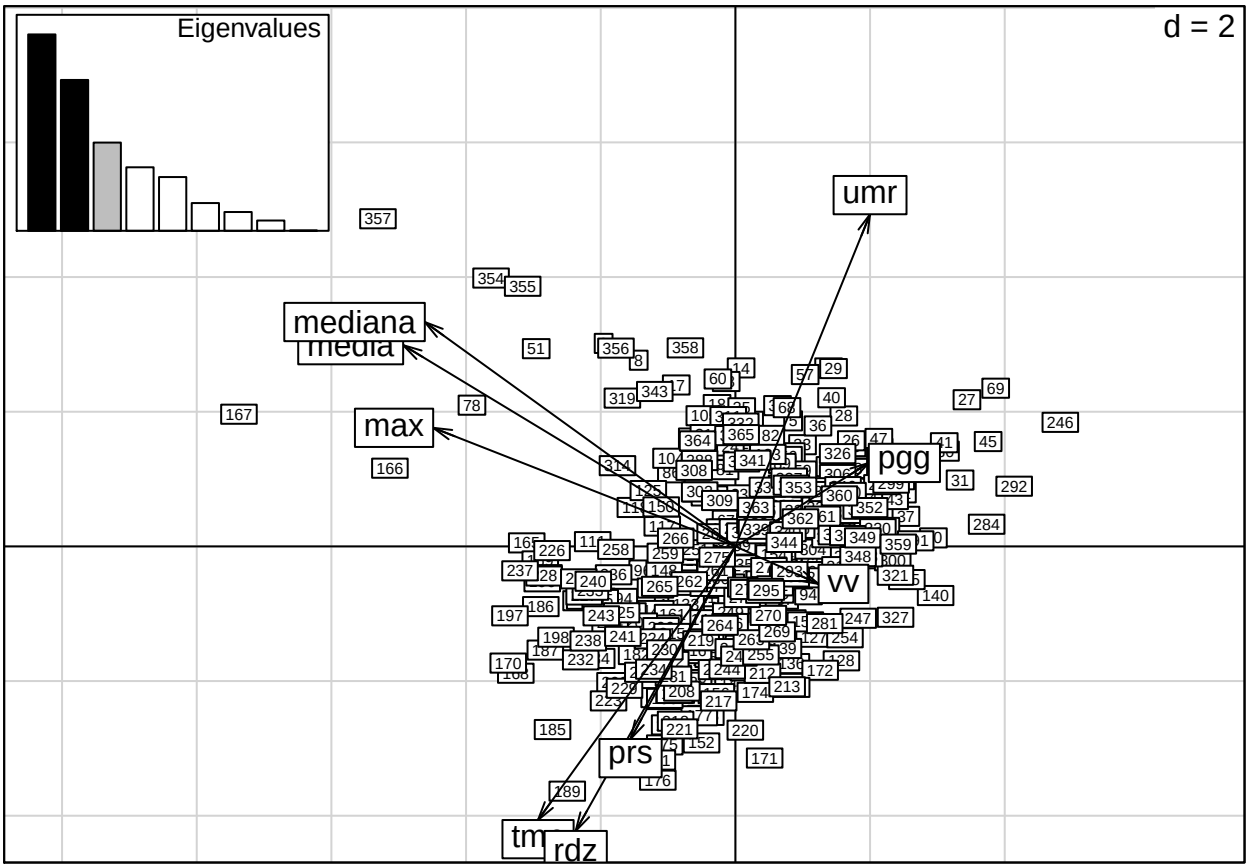
```
## Loading required package: ade4
```

```
## Warning: package 'ade4' was built under R version 3.3.2
```

```
names(pm10_a)
```

```
##  [1] "media"    "mediana"  "max"      "tmp"      "umr"      "pgg"
##  [7] "rdz"      "prs"      "vv"       "dv"       "anno"     "mese"
## [13] "giorno"   "mese2"    "giorno2"  "Stagione"
```

```
#estraggo da pm10 solo le variabili numeriche da inserire nella PCA escludendo giorno mese e anno che s
data_pca <- pm10_a[,c(1:9)]

# con nf = 3 gli stò dicendo di tenere tre assi
pcapm <- dudi.pca(data_pca, scannf = FALSE, nf = 3)

#faccio il biplot, cos'è clab.row?
# clab.row serve a gestire la grandezza dei quadratini delle osservazioni
# clab.col invece controlla le variabili
scatter(pcapm, clab.row = 0.5)
```



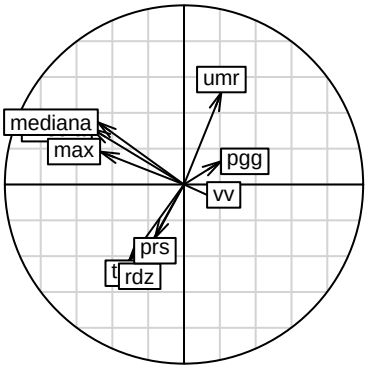

```
#valori dei punteggi per le variabili
pcapm$c1
```

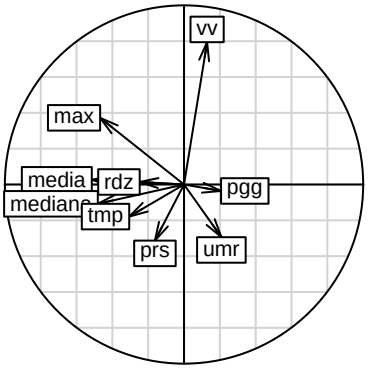
```
##                CS1         CS2         CS3
## media    -0.5148158  0.31181443  0.02609813
## mediana -0.4812498  0.34765592 -0.10785132
## max      -0.4683100  0.18372295  0.37203784
## tmp      -0.3058405 -0.42397086 -0.17915070
## umr       0.2086416  0.51464826 -0.29406707
## pgg       0.2067228  0.12934962 -0.03608484
## rdz      -0.2474513 -0.44087121  0.01414590
## prs      -0.1624832 -0.29843062 -0.31322712
## vv        0.1288881 -0.05813173  0.79441234
```

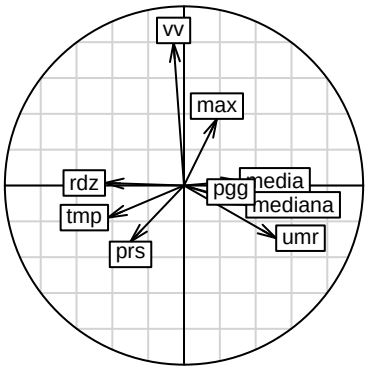
CARATTERIZZIAMO GLI ASSI

se vedi come ci si aspetta mediana media e max stanno insieme perchè sono molto correlati e vanno tutti sulla prima componente sulla seconda ci vanno i fattori ambientali tmp, umr, rdz, pgg e prs. Solo che tmp, prs e rdz hanno correlazione negativa rispetto a umr e pgg [pgg è al limite. ha un punteggio abbastanza basso] sulla terza si prende tutto il vento.

```
#cerchio di correlazione, i punteggi delle varibili sono plottati su una circonferenza di raggio unitar
par(mfrow = c(2,2))
s.corcircle(pcapm$c1, xax = 1, yax = 2) #plotto prima e seconda
s.corcircle(pcapm$c1, xax = 1, yax = 3) #plotto prima e terza
s.corcircle(pcapm$c1, xax = 2, yax = 3) #plotto seconda e terza
```







Autovalori

```
#calcolo gli autovalori
pcapm$eig / sum(pcapm$eig)
```

```
## [1] 0.3219695419 0.2473643668 0.1446821326 0.1041546381 0.0881580394
## [6] 0.0455184307 0.0308126012 0.0168601788 0.0004800706
```

```
#calcolo la somma cumulata degli autovalori
cumsum(pcapm$eig / sum(pcapm$eig)) # con i primi tre assi spiego circa il 70% che non è male
```

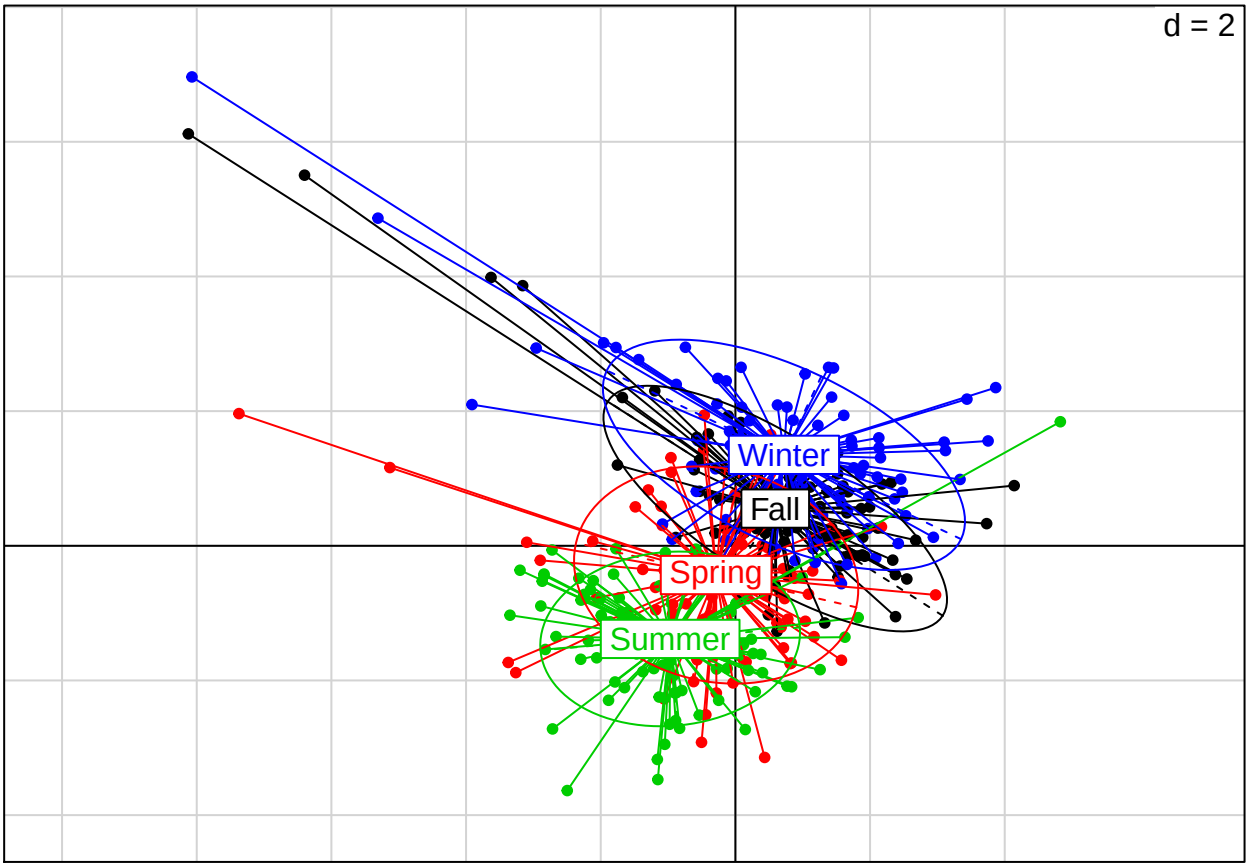
```
## [1] 0.3219695 0.5693339 0.7140160 0.8181707 0.9063287 0.9518471 0.9826598
## [8] 0.9995199 1.0000000
```

da cosa lo capisci??? perchè il 3° dice 0.71? gli autovalori corrispondono alla quantità di variabilità spiegata lungo quella direzione (autovettore) se divido ogni singolo autovalore per la somma di tutti gli autovalori (riga 80) ottengo la % che ogni singola componente spiega. Facendo la somma cumulata (riga 83) non faccio altro che: 1c , 1c + 2c, 1c + 2c + 3c. La somma di tutte chiaramente mi darà 100 (o meglio 1) quindi in questo modo io so che tenedo le prime tre componenti io riesco a vedere il 71% dell'informazione presente nel mio dataset.

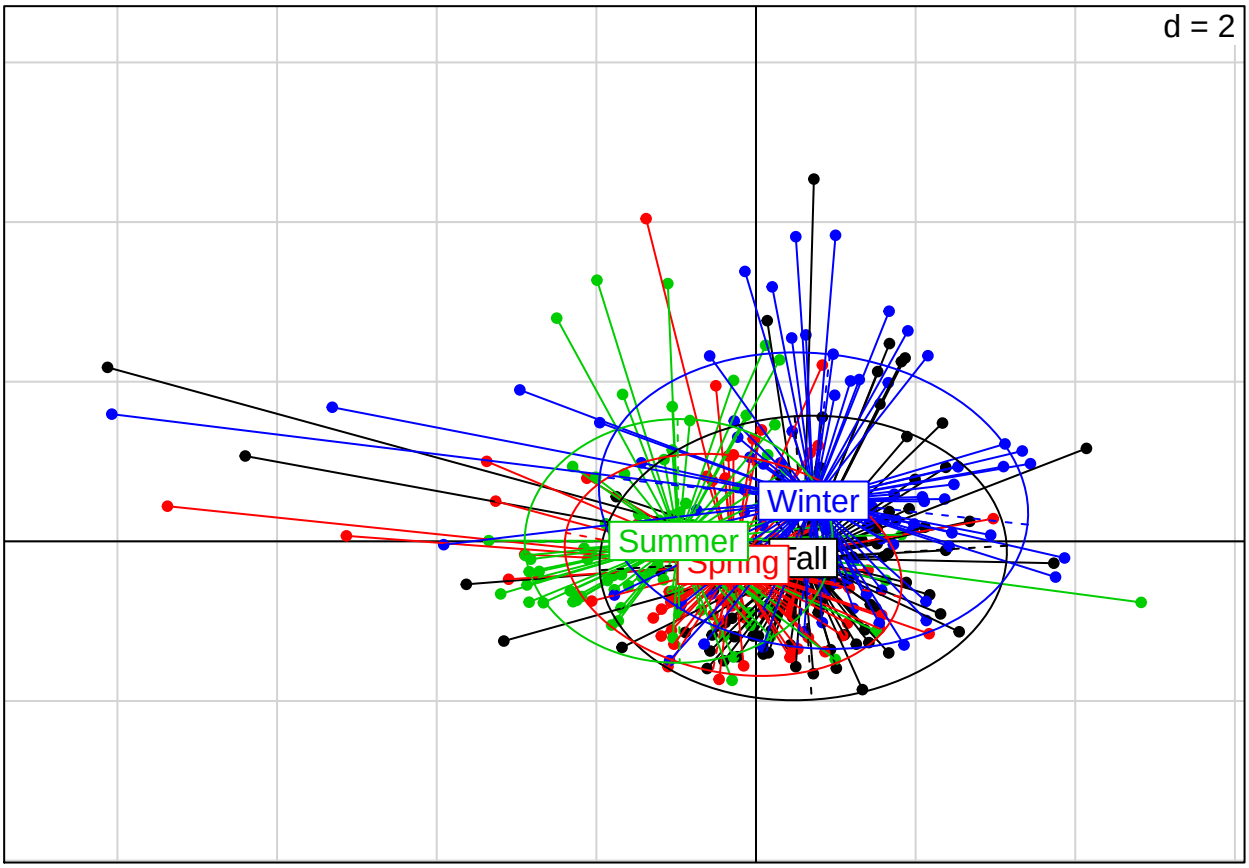
Stagione

```
#vediamo adesso come si comportano le variabili qualitative - lascio a te l'interpretazione!!!!!!
# PCA STAGIONE ----
```

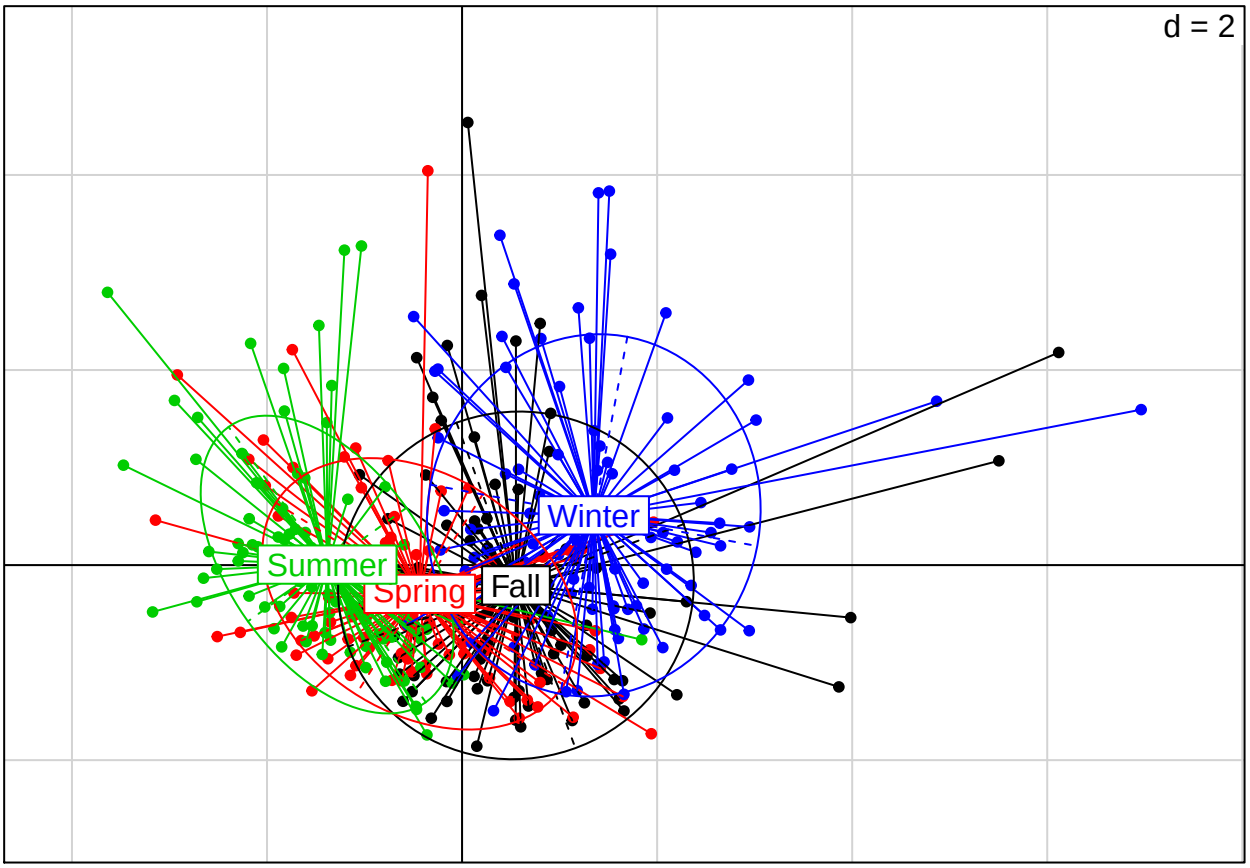
```
par(mfrow= c(1,1))
s.class(pcapm$li, factor(pm10_a$Stagione), xax = 1, yax = 2, col = c(1,2,3,4))
```



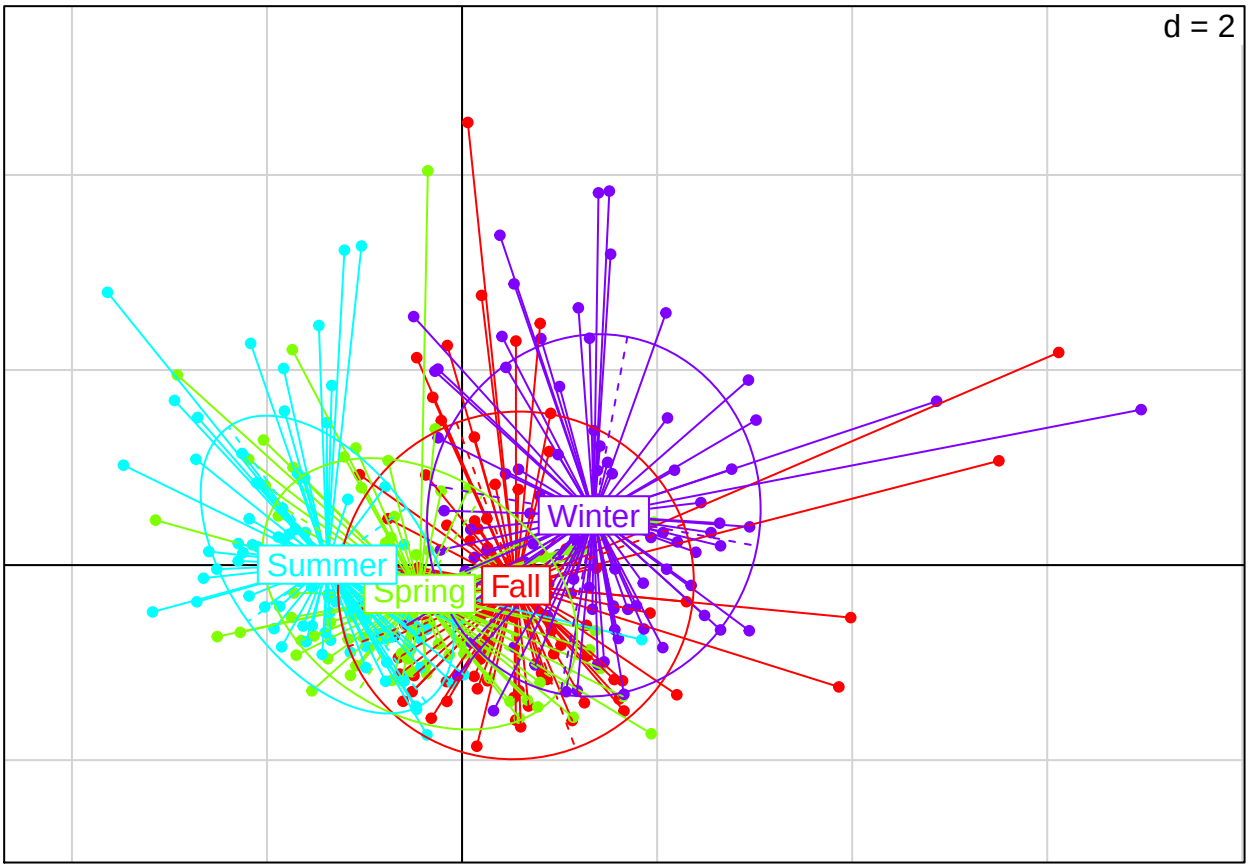

```
s.class(pcapm$li, factor(pm10_a$Stagione), xax = 1, yax = 3, col = c(1,2,3,4))
```




```
s.class(pcapm$li, factor(pm10_a$Stagione), xax = 2, yax = 3, col = c(1,2,3,4))
```




```
s.class(pcapm$li, factor(pm10_a$Stagione), xax = 2, yax = 3, col = rainbow(4)) #se volessi usare la fun
```




```
#un modo carino per selezionare i colori
color = grDevices::colors()[grep('gr(a|e)y', grDevices::colors(), invert = T)]

# questo comando crea un oggetto che si chiama color che contiene tutti e 433 i colori base di R esclus
# poi basta fare col = sample(color, n) all'intero di una funzione che usa come argument col
# dove n è il numero di colori che devi utilizzare

#alternativa se 433 colori sono troppi
library(RColorBrewer)
qual_col_pals = brewer.pal.info[brewer.pal.info$category == 'qual',]
col_vector = unlist(mapply(brewer.pal, qual_col_pals$maxcolors, rownames(qual_col_pals)))
# questa funzione produce un vettore di 74 colori
```

Interpretazione (la devi fare tu), stai attenta! :

- 1 - come si dispongono le ellissi rispetto alle tre componenti?
- 2 - ci sono delle direzioni in cui la variazione è più evodente e altre invece dove non sembra esserci?
- 3 - cosa rappresentano le tre componenti?
- 4 - l'asse maggiore delle ellissi su che componente si allinea?
- 5 - che cosa mi racconta quindi il grafico?

facile dire la devi fare tu gne gne gne gne XD alloraaaa(sto rosicando che l'avevo già fatta ma cambiando computer ho fatto un panico... dunque) rispondo alle domande per stagioni:

1) le ellissi praticamente si sovrappongono rispetto alle componenti, in particolare le componenti 1 e 3 , mentre sono trasversali tra 1 e 2
2) se per variazione intendi i picchi, in inverno c'è variazione
3) le tre componenti sono la 1a la concentrazione del pm10 la 2a le variabili ambientali, la 3a la velocità

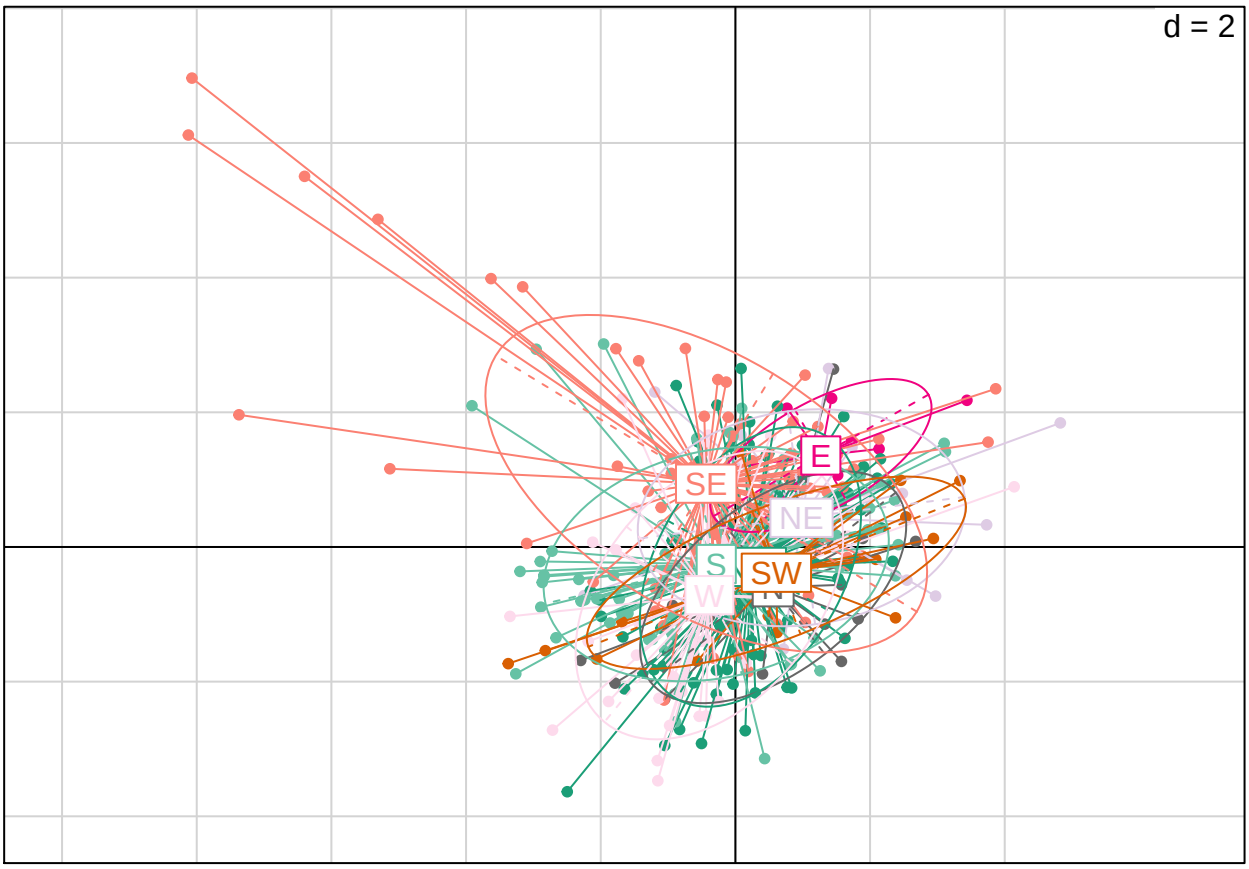
del vento
4) l'asse maggiore è parallelo alla componente che rappresenta la concentrazione
5) essendo così tutte attaccate mi sa che il fattore stagione posso anche non cagarlo posso invece indagare su st'inverno che esce fuori dagli schemi no??

SPIEGONE STAGIONE

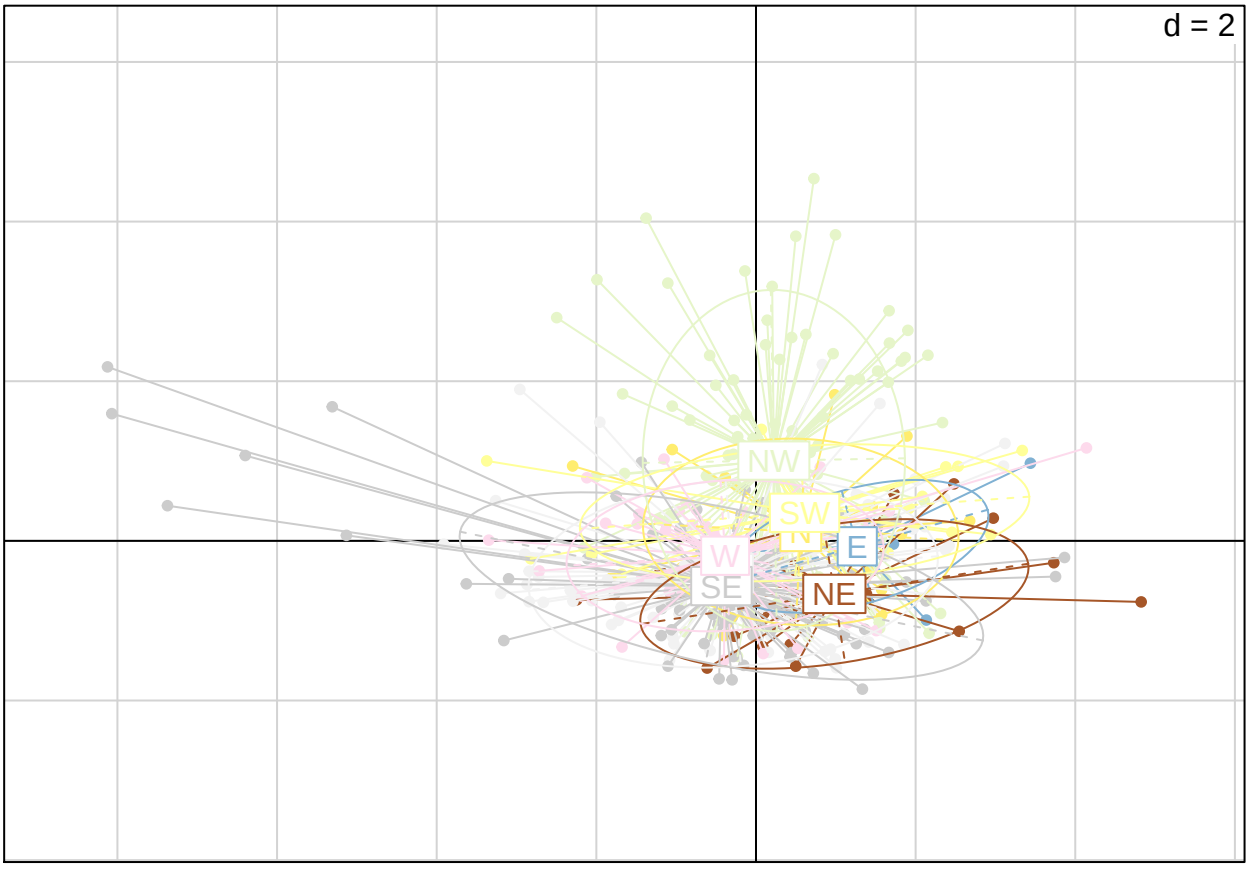
Primo piano fattoriale comp 1-2: le ellissi si dispongono trasversalmente. Tuttavia cosa si nota: - sembrerebbe esserci una maggiore variazione sulla direzione della seconda componente rispetto alla prima. Il che significa che c'è una certa variazione dei fattori ambientali durante l'anno (come ci aspettiamo), temperatura , radiazione solare e pressione elevate in estate e invece umidità e forse precipitazioni (ha punteggio basso) elevate in inverno. - in termini di concentrazione di pm10 c'è una variazione ma sembrerebbe essere inferiore rispetto a quelle delle variabili ambientali. Tuttavia, l'asse maggiore delle ellissi per autunno e inverno è lungo la direzione della prima componente il che indica che in quei mesi c'è una maggiore varibilità ma in termini di media stagionale sono identici. - La primavera è praticamente un cerchio il cui centro combacia con il centro del piano fattoriale indicando che le condizioni corrispondono alla media del sistema.

Con la PCA non stai cercando risposte stai semplicemente osservano che informazioni sono contenute nei tuoi dati. Le ellissi non sono poi così sovrapposte e il fattore stagionale in termini di conc di pm10 e fattori ambientali è piuttosto evidente. Per quanto rigarda il vento sembrerebbe essere maggiore solo in inverno il che corrisponde ad una conc più bassa di pm10 (come ci aspettiamo perchè il vento sparge i pm10 mentre la pioggia li fa depositare a terra). L'inverno non esce dagli schemi ma semplicemente sembra mostrare una maggiore variabilità (cosa del tutto naturale).

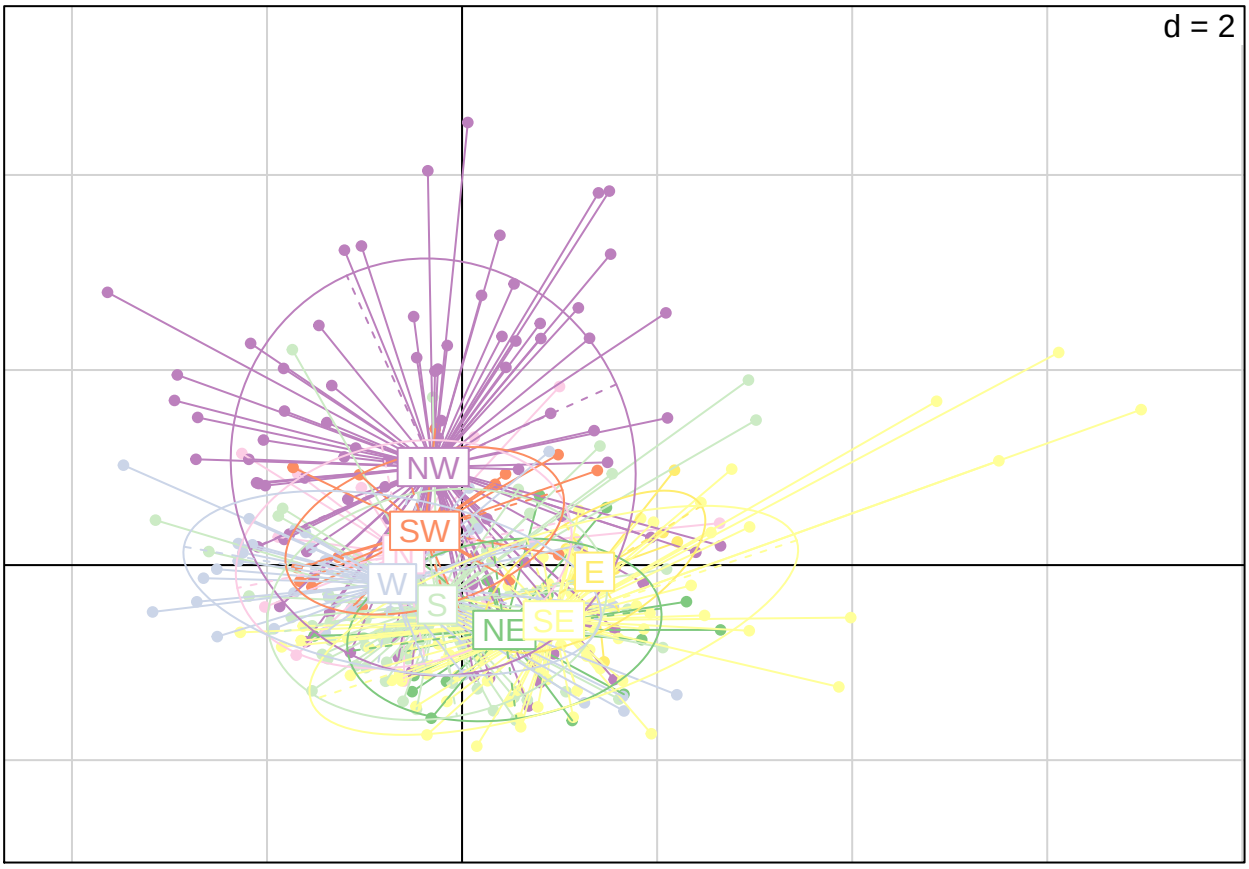
```
#DIREZIONE DEL VENTO ----
s.class(pcapm$li, factor(pm10_a$dv), xax = 1, yax = 2, col = sample(col_vector, nlevels(pm10$dv)))
```



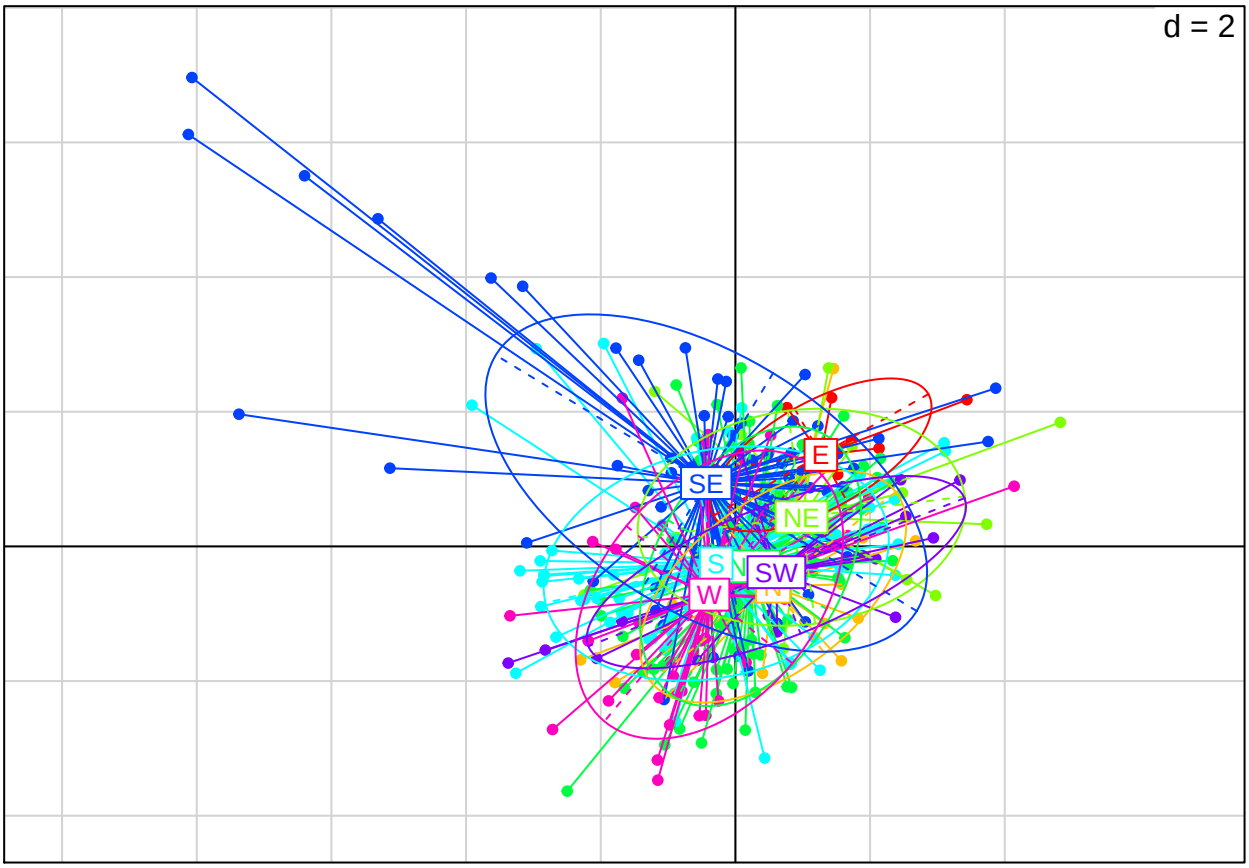
```
s.class(pcapm$li, factor(pm10_a$dv), xax = 1, yax = 3, col = sample(col_vector, nlevels(pm10$dv)))
```



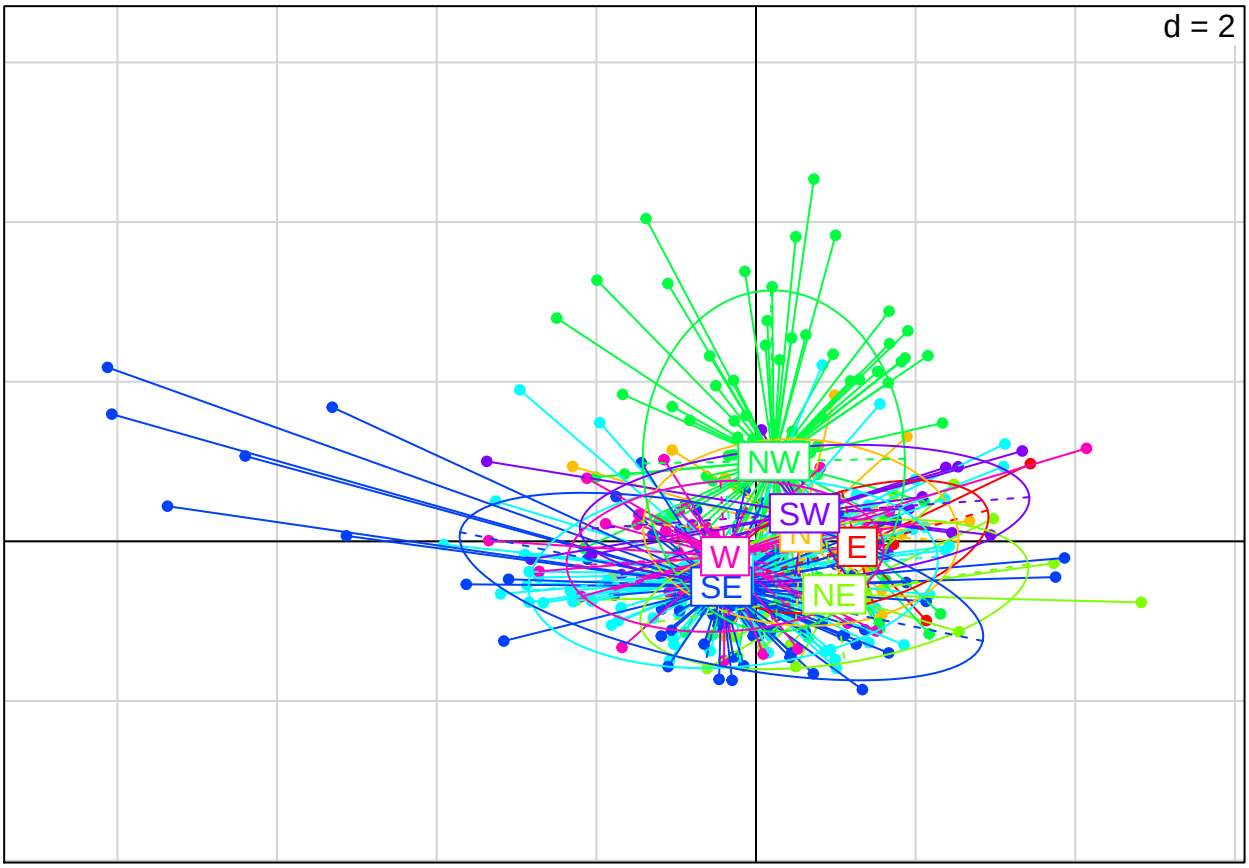

```
s.class(pcapm$li, factor(pm10_a$dv), xax = 2, yax = 3, col = sample(col_vector, nlevels(pm10$dv)))
```



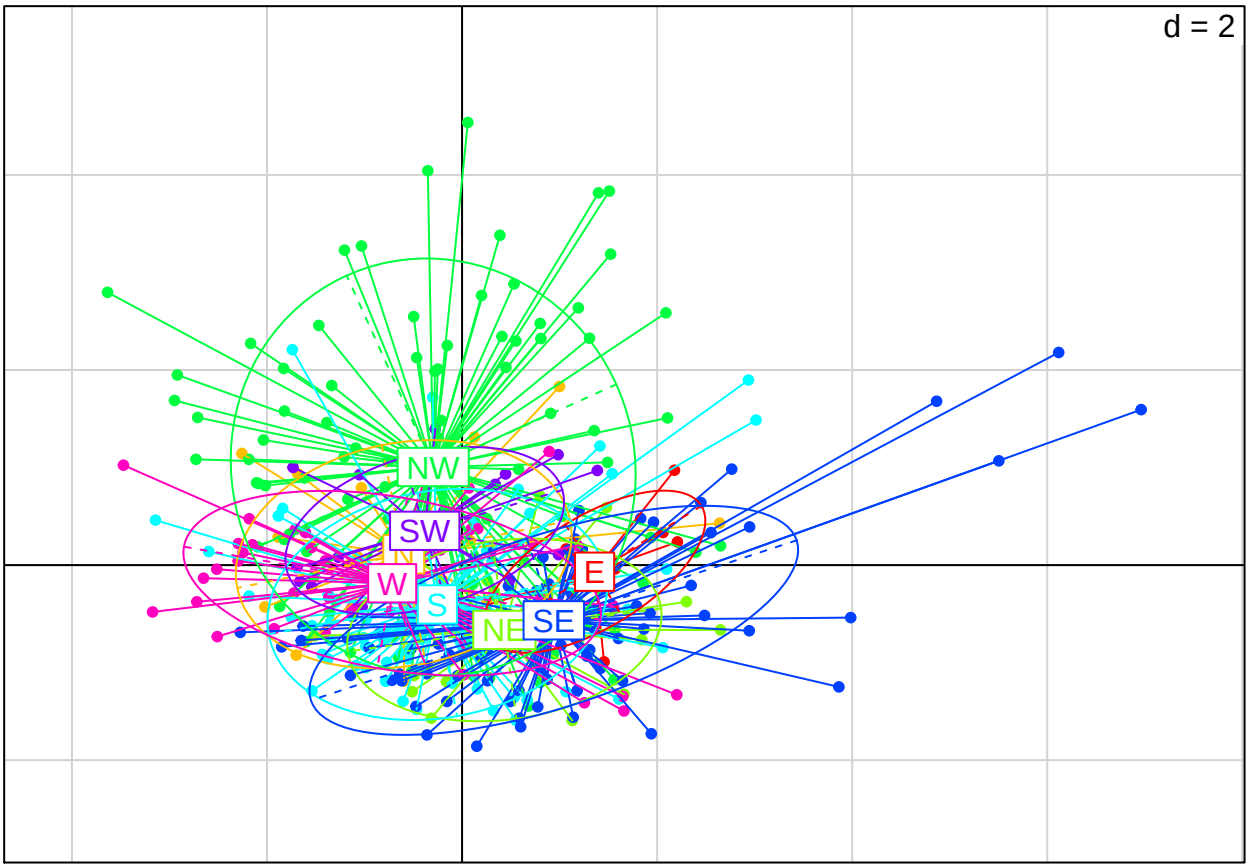

```
s.class(pcapm$li, factor(pm10_a$dv), xax = 1, yax = 2, col = rainbow(nlevels(pm10$dv)),clabel = .8)
```




```
s.class(pcapm$li, factor(pm10_a$dv), xax = 1, yax = 3, col = rainbow(nlevels(pm10$dv)))
```




```
s.class(pcapm$li, factor(pm10_a$dv), xax = 2, yax = 3, col = rainbow(nlevels(pm10$dv)))
```




```
# ARI-DOMANDE
# 1) le ellissi sono sempre accozzagliate l'una sull'altra rispetto alle componenti
# 2) mi sembra che il vento da sudest ha sempre sti picchi malefici che escono dagli schemi
# 3) le componenti sono sempre le stesse
# 4) l'asse maggiore sta sempre sulla concentrazione di pm10
# 5) anche qui stanno tutte una sull'altra non ci sta niente di "eclatante" apparte i picchi di vento
# da sudest

# fatti le stesse domande che ti ho scritto prima
```

Sui venti c'è un po' di casino ma qualcosa si nota: - i venti da SE hanno una variabilità estrema in termini di pm10 e fattori ambientali - i venti da NW sono i più intensi ma non sembrano corrispondere ad una riduzione di conc di pm10.

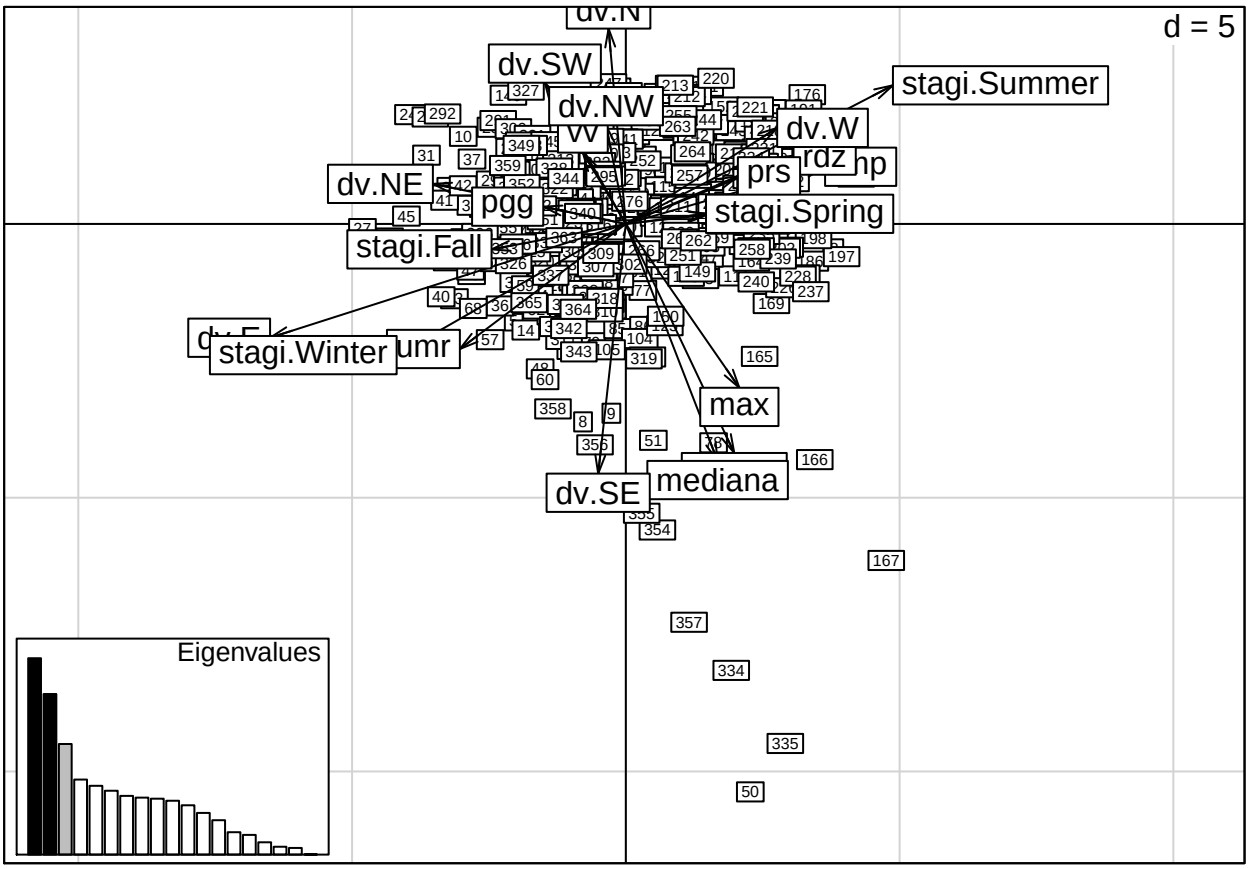
HillSmith

```
#### HILLSMITH SU DATI ----

data_hills<-(pm10_a[,1:10])
data_hills$stagione<-pm10_a$Stagione


str(data_hills)
```

```
## 'data.frame':    365 obs. of  11 variables:
```

```
##  $ media   : num  30.4 15.2 18.3 21.8 27.6 ...
##  $ mediana : num  26 14.5 14.7 22.5 27.5 ...
##  $ max     : num  46 24 35 29 32 38 38 48 61 17 ...
##  $ tmp     : int  13 12 8 6 11 13 13 14 12 9 ...
##  $ umr     : int  64 58 55 61 84 73 72 78 67 63 ...
##  $ pgg     : num  0 1.6 0 11.4 0.4 ...
##  $ rdz     : num  1 1 1 1 1 2 1 0 0 1 ...
##  $ prs     : int  1001 1007 1018 1022 1010 1007 1014 1009 1002 1001 ...
##  $ vv      : num  3.7 4.1 7.3 1.7 2.6 ...
##  $ dv      : Factor w/ 8 levels "E","N","NE","NW",..: 7 7 4 3 6 5 5 6 5 7 ...
##  $ stagione: chr  "Winter" "Winter" "Winter" "Winter" ...
```

```
data_hills$stagione<-as.factor(data_hills$stagione)
dd1 <- dudi.hillsmith(data_hills, scannf = FALSE, nf = 3)
par(mfrow = c(1,1))

scatter(dd1, clab.row = 0.5, posieig = "bottom")
```

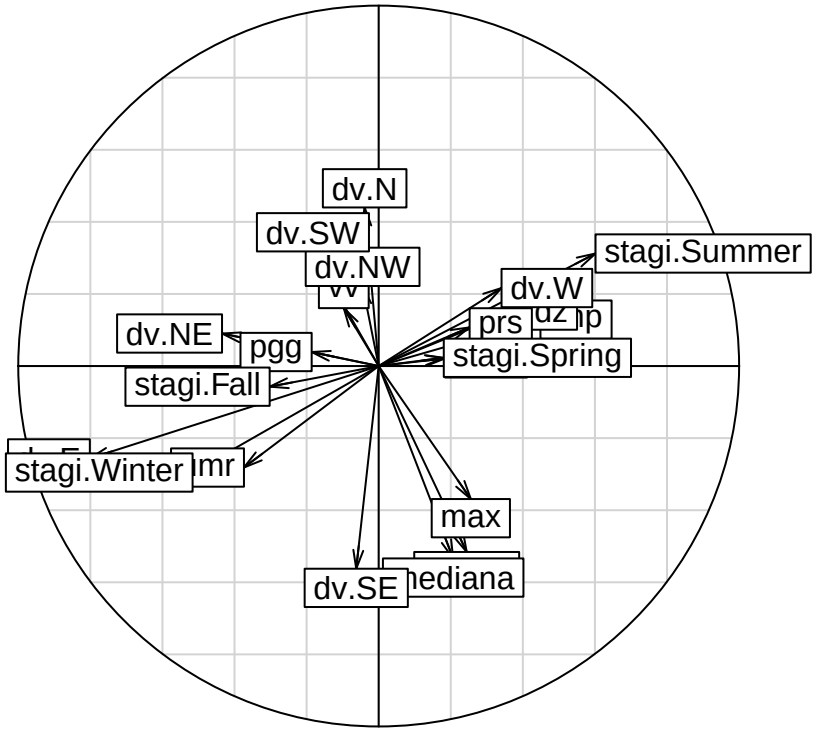



```
# perchè lo scatter sta a fanculo in alto?
# stavi pensando a cosa rappresentano i valori della hills e quali sono le
# componenti principali in questo caso...
dd1$c1
```

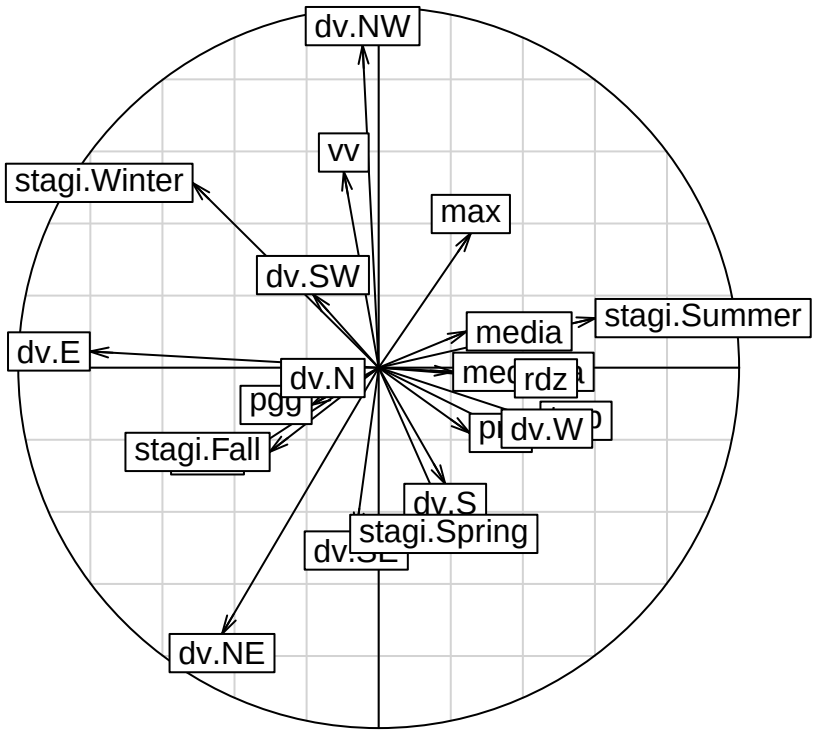
```
##                    CS1         CS2         CS3
## media        0.24420926 -0.51588959  0.10067313
## mediana      0.20668324 -0.53375687 -0.01137620
## max          0.25567159 -0.36899455  0.37319907
## tmp          0.44879826  0.12879803 -0.14724194
```

```
## umr          -0.37384765 -0.28162825 -0.24322914
## pgg          -0.18598709  0.03891771 -0.10339526
## rdz           0.37769169  0.15390185 -0.02969131
## prs           0.25234425  0.10728023 -0.18101382
## vv           -0.09683114  0.16061103  0.54382640
## dv.E         -0.80133330 -0.25628655  0.04498849
## dv.N         -0.03888944  0.44002296 -0.02872326
## dv.NE        -0.43463498  0.09028357 -0.73880250
## dv.NW        -0.04459080  0.22215773  0.89453602
## dv.S          0.18454104  0.02120006 -0.32213588
## dv.SE        -0.06227084 -0.56236213 -0.45413332
## dv.SW        -0.18241175  0.31804808  0.20379361
## dv.W          0.34088653  0.21639485 -0.16912079
## stagi.Fall   -0.30247894 -0.05730287 -0.23374887
## stagi.Spring  0.18062484  0.02286136 -0.40799016
## stagi.Summer  0.60164421  0.31224041  0.13755502
## stagi.Winter -0.51628018 -0.29546725  0.51283506
```

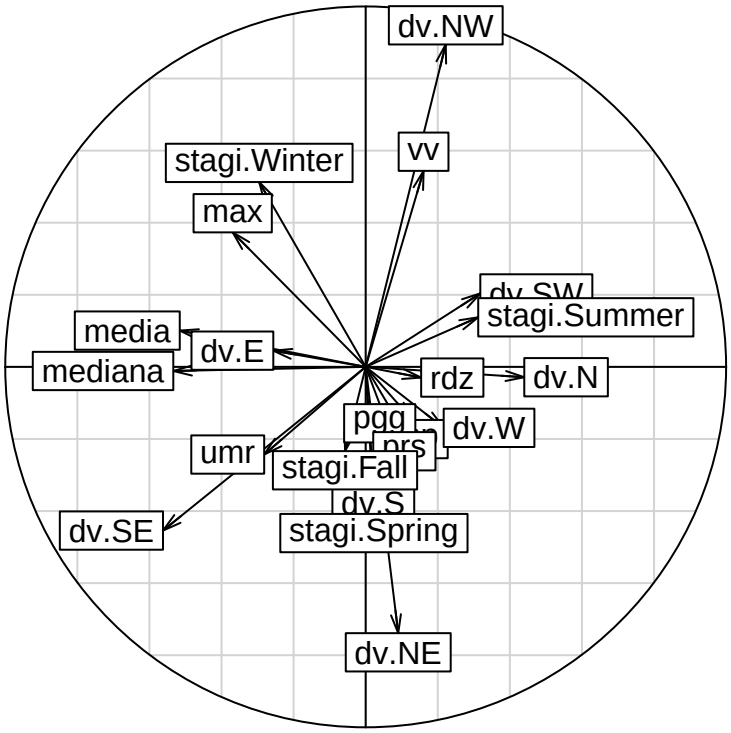
```
par(mfrow=c(1,1))
s.corcircle(dd1$c1, xax = 1, yax = 2) #plotto prima e seconda
```




```
s.corcircle(dd1$c1, xax = 1, yax = 3) #plotto prima e terza
```




```
s.corcircle(dd1$c1, xax = 2, yax = 3) #plotto seconda e terza
```




```
dd1$eig/sum(dd1$eig)
```

```
##  [1] 0.1792674765 0.1466015157 0.1009806688 0.0685059244 0.0627992002
##  [6] 0.0583123136 0.0536131425 0.0520145735 0.0510403206 0.0491774667
## [11] 0.0450266734 0.0380627101 0.0314160318 0.0204608379 0.0179993801
## [16] 0.0112376258 0.0071854322 0.0060806861 0.0002180201
```

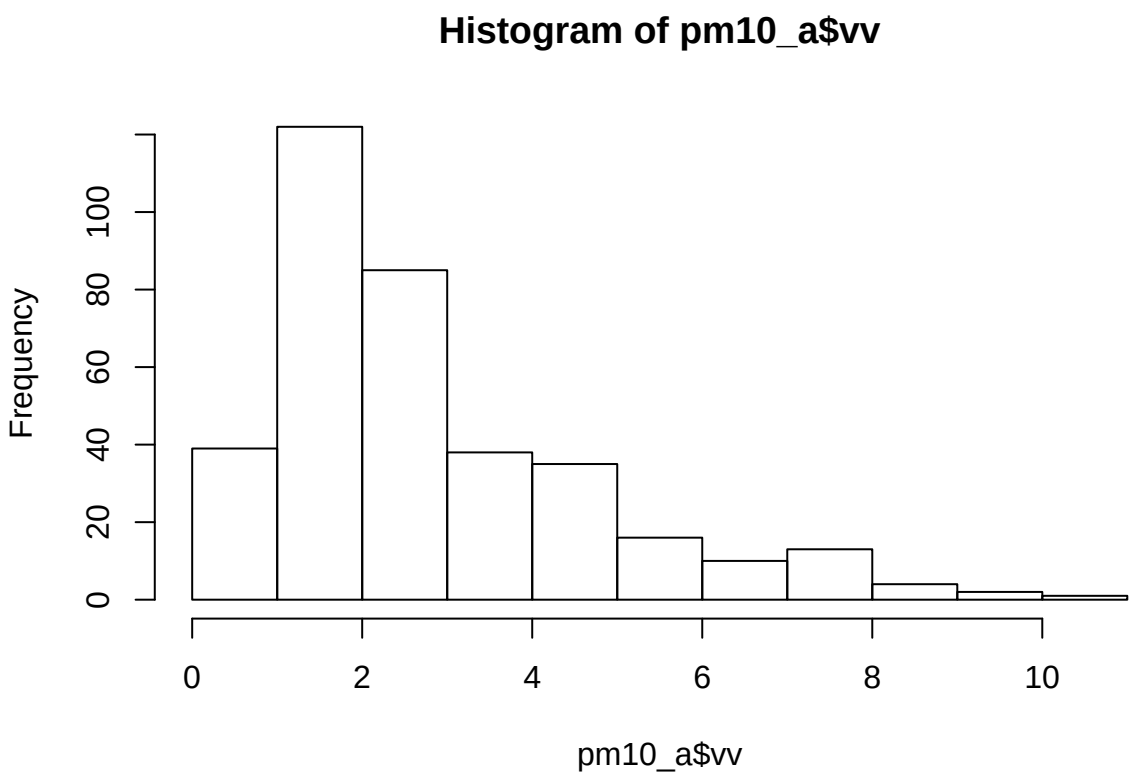
```r
cumsum(dd1$eig/sum(dd1$eig))
```

```
##  [1] 0.1792675 0.3258690 0.4268497 0.4953556 0.5581548 0.6164671 0.6700802
##  [8] 0.7220948 0.7731351 0.8223126 0.8673393 0.9054020 0.9368180 0.9572789
## [15] 0.9752782 0.9865159 0.9937013 0.9997820 1.0000000
```

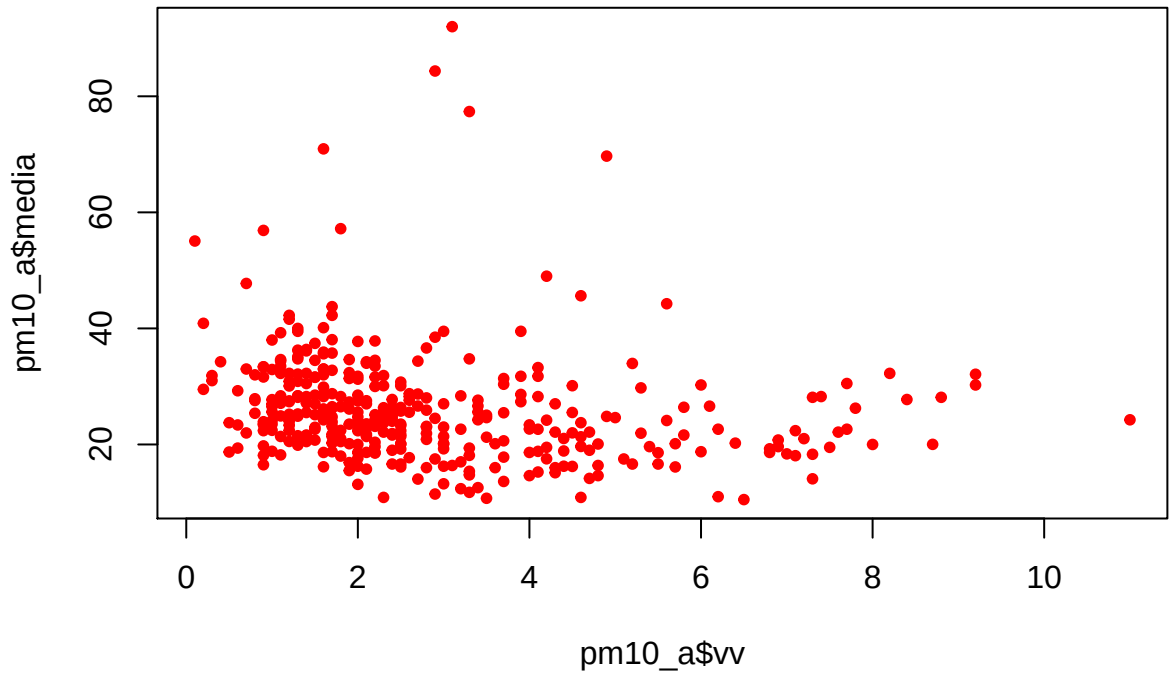
```r
# quindi aggiungendo la stagione  e la dv e facendo la hills
# i 3 assi mi perdono quantità di informazione e spiegano quasi il 50% dei dati.


# ma sopratutto ora che so che:
# media mediana e max sono super correlate, quindi mi conviene tenerne una sola
# tmp,rdz,prs sono inversamente prop all'umidità e che la pioggia sta a se
#velocità del vento va per fatti suoi e tira i miei dati
# d'inverno ho dei picchi particolari che non so se hanno valenza
# e mi sembra che la direzione NW può avere una valenza
```

La hillsmith è chiaramente più incasinata e vedi meno (42% rispetto al 71%). Io lascerei la PCA normale e basta poi vedete voi. [I venti da SE sembrano interessanti sono molto correlati alla concentrazione di pm10]

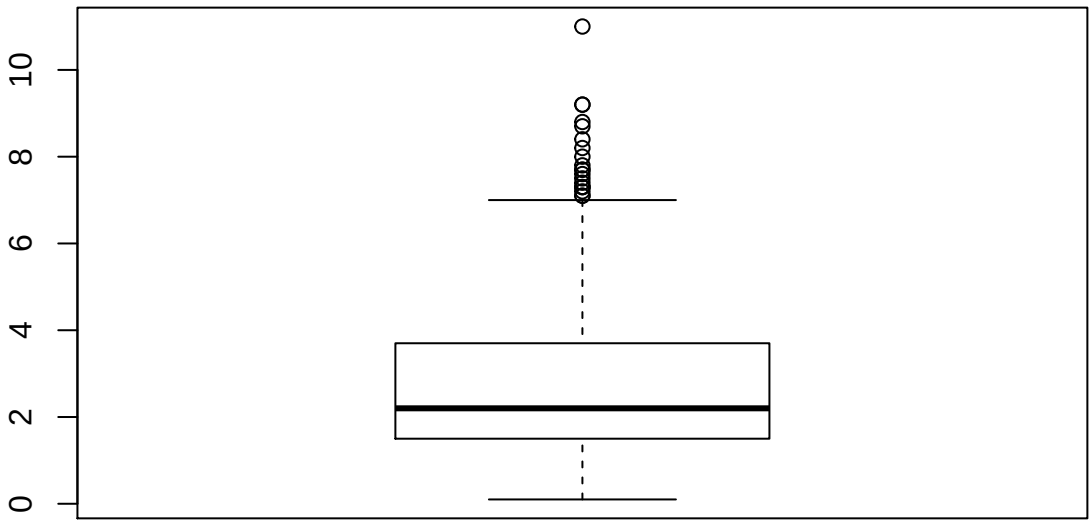
```r
# provo a vedere sta velocità del vento che problemi ha


hist(pm10_a$vv)
```




```r
# sembrerebbe non esserci correlazione

plot(pm10_a$media~pm10_a$vv,col = "red", pch = 20)
```




```r
# sembrerebbe non esserci correlazione
boxplot(pm10_a$vv)
```




```r
#faccio un modello di regressione multipla con tutto dentro
modvv<-lm(vv~.,data=data_hills)
summary(modvv)
```

```
##
## Call:
## lm(formula = vv ~ ., data = data_hills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1712 -0.7949 -0.1700  0.5783  6.5330
##
## Coefficients:
```

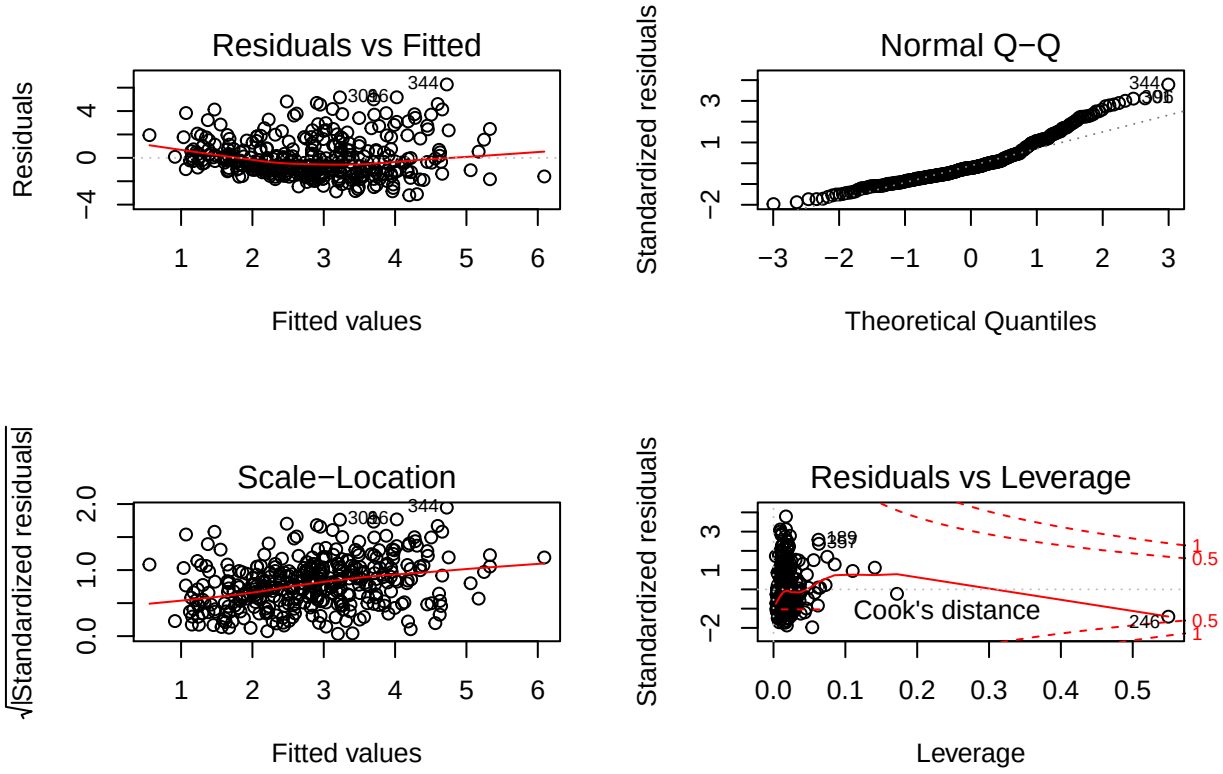
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    42.942248   9.915161   4.331 1.95e-05 ***
## media           0.137213   0.084460   1.625 0.105161
## mediana        -0.213432   0.066343  -3.217 0.001417 **
## max             0.052771   0.018685   2.824 0.005014 **
## tmp            -0.060692   0.025063  -2.422 0.015964 *
## umr            -0.037690   0.010342  -3.644 0.000309 ***
## pgg             0.020369   0.005241   3.886 0.000122 ***
## rdz            -0.003827   0.011622  -0.329 0.742138
## prs            -0.036945   0.009720  -3.801 0.000170 ***
## dvN             0.961099   0.560199   1.716 0.087123 .
## dvNE            0.154381   0.541688   0.285 0.775815
## dvNW            1.052401   0.496737   2.119 0.034834 *
## dvS             0.010887   0.490241   0.022 0.982296
## dvSE            0.456477   0.491151   0.929 0.353328
## dvSW           -0.018092   0.558421  -0.032 0.974173
## dvW            -0.408136   0.509457  -0.801 0.423611
## stagioneSpring -0.188967   0.235104  -0.804 0.422089
## stagioneSummer -0.313412   0.316233  -0.991 0.322340
## stagioneWinter -0.208461   0.254230  -0.820 0.412799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.338 on 346 degrees of freedom
## Multiple R-squared:  0.5356, Adjusted R-squared:  0.5114
## F-statistic: 22.17 on 18 and 346 DF,  p-value: < 2.2e-16
```

```
# la vv del vento con la direzione nord fa sempre vedere qualcosa
# per me c'è qualcosa sotto, dice che sono significative
# umr,pgg,prs

# Questo modello così non va. C'è multicollinearità tra le varie conc di pm10.

modvv2 <- lm(vv~ media + tmp + umr + pgg + rdz + prs, data = data_hills)
summary(modvv2)
```

```
##
## Call:
## lm(formula = vv ~ media + tmp + umr + pgg + rdz + prs, data = data_hills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1986 -1.1069 -0.3701  0.7159  6.2756
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.489083  11.148327   3.901 0.000114 ***
## media       -0.013613   0.009371  -1.453 0.147163
## tmp         -0.113626   0.019095  -5.950 6.37e-09 ***
## umr         -0.086662   0.009780  -8.861  < 2e-16 ***
## pgg          0.022933   0.006230   3.681 0.000268 ***
## rdz         -0.018461   0.013043  -1.415 0.157828
## prs         -0.032773   0.011018  -2.974 0.003134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.672 on 358 degrees of freedom
## Multiple R-squared:  0.2501, Adjusted R-squared:  0.2376
## F-statistic:  19.9 on 6 and 358 DF,  p-value: < 2.2e-16
```

```
# hai un R^2 molto basso. La relazione potrebbe non essere lineare o semplicemente la velocità del vent
par(mfrow=c(2,2))
plot(modvv2)
```




```
names(data_hills)
```

```
##  [1] "media"    "mediana"  "max"      "tmp"      "umr"      "pgg"
##  [7] "rdz"      "prs"      "vv"       "dv"       "stagione"
```

```
corr <- cor(data_hills[,1:9])
corr
```

```
##               media     mediana         max         tmp         umr
## media    1.00000000  0.97461260  0.81658630  0.1640706  0.03296613
## mediana  0.97461260  1.00000000  0.68482834  0.1308403  0.13966079
## max      0.81658630  0.68482834  1.00000000  0.1495194 -0.20466947
## tmp      0.16407064  0.13084026  0.14951935  1.0000000 -0.55164831
## umr      0.03296613  0.13966079 -0.20466947 -0.5516483  1.00000000
## pgg     -0.16103781 -0.13709966 -0.17510053 -0.1119014  0.29451167
## rdz      0.08141269  0.03467562  0.11568679  0.5639915 -0.55246082
## prs      0.02731391  0.01662717  0.04976177  0.3642715 -0.27557772
## vv      -0.18897249 -0.31283959  0.17520774 -0.2014872 -0.20796038
##                 pgg         rdz         prs          vv
## media   -0.16103781  0.08141269  0.02731391 -0.18897249
## mediana -0.13709966  0.03467562  0.01662717 -0.31283959
## max     -0.17510053  0.11568679  0.04976177  0.17520774
```

```
## tmp     -0.11190139  0.56399145  0.36427151 -0.20148723
## umr      0.29451167 -0.55246082 -0.27557772 -0.20796038
## pgg      1.00000000 -0.14352482 -0.10764275  0.09885347
## rdz     -0.14352482  1.00000000  0.20200519 -0.05095261
## prs     -0.10764275  0.20200519  1.00000000 -0.16942295
## vv       0.09885347 -0.05095261 -0.16942295  1.00000000
```

Quello che ti interessa a te è indagare cosa regola la conc di pm10

REGRESSIONE MULTIPLA

```
mod1<-lm(media~.,data=data_hills)

summary(mod1)
```

```
##
## Call:
## lm(formula = media ~ ., data = data_hills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0834 -0.5227  0.0084  0.5206  3.4843
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.5608437  6.4304160   1.642   0.1014
## mediana         0.7766381  0.0089137  87.128   <2e-16 ***
## max             0.1876162  0.0064716  28.991   <2e-16 ***
## tmp             0.0076597  0.0160211   0.478   0.6329
## umr            -0.0045430  0.0066783  -0.680   0.4968
## pgg            -0.0009586  0.0033949  -0.282   0.7778
## rdz             0.0141750  0.0073311   1.934   0.0540 .
## prs            -0.0111419  0.0062621  -1.779   0.0761 .
## vv              0.0551718  0.0339604   1.625   0.1052
## dvN             0.6209598  0.3551674   1.748   0.0813 .
## dvNE            0.2459190  0.3432734   0.716   0.4742
## dvNW            0.4521134  0.3160872   1.430   0.1535
## dvS             0.0991681  0.3108190   0.319   0.7499
## dvSE            0.2318803  0.3115809   0.744   0.4573
## dvSW            0.2903219  0.3537540   0.821   0.4124
## dvW             0.1195275  0.3232848   0.370   0.7118
## stagioneSpring  0.1362894  0.1490400   0.914   0.3611
## stagioneSummer  0.3547270  0.1999019   1.775   0.0769 .
## stagioneWinter -0.0349345  0.1613545  -0.217   0.8287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8486 on 346 degrees of freedom
## Multiple R-squared:  0.9929, Adjusted R-squared:  0.9925
## F-statistic:  2675 on 18 and 346 DF,  p-value: < 2.2e-16
```

```
# non capisco perchè è sparito l'autunno cmq dalle stagioni
AIC(mod1)
```

```
## [1] 936.4839
```

```
# dal risultato del modello mi sembra che ovviamente hanno rilevanza mediana e
# massima che si accavallano sempre con l'intercetta che è la media
# sbaglio o la direzione del vento N-NW ha rilevanza?, e poi l'R2 è un sacco alto
# ma l'AIC è altissimo quindi questo modello non spiega assolutamente nulla ahaha
```

Anche qui non puoi usare media, mediana e max insieme. Vendono chiaramente significative e ti alzano l'R^2 producendo un overfit L'autunno non è sparito ma è finito nell'intercetta come "corner point". Se inserisci una variabile "Dummy" (Qualitativa) in un modello R sceglie automaticamente i primo livello in ordine alpha numerico come corner point. Vuol dire che stai confrontando la variazione dei pm10 in autunno rispetto alle altre variabili e stagioni. L'AIC è un criterio di verosimiglianza che serve a confrontare modelli tra loro. Non importa se sia alto o basso, quello dipende dalla quantità di variabilità presente nel modello. Tu devi "teoricamente" ricercare il modello che abbia l'AIC più basso possibile ma sempre confrontandolo con gli altri modelli che hai fatto. E' esattamente ciò che fa la stepwise.

```
#prima di fare la stepwise elimina mediana e max
data_hills2 <- data_hills[c(1,4:11)]

mod1 <- lm(media~., data = data_hills2)
summary(mod1)
```

```
##
## Call:
## lm(formula = media ~ ., data = data_hills2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.211  -5.560  -1.675   4.297  58.754
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.24700   68.08949  -0.018  0.98540
## tmp             0.69836    0.16555   4.218 3.14e-05 ***
## umr             0.01872    0.06953   0.269  0.78787
## pgg            -0.09598    0.03560  -2.696  0.00736 **
## rdz             0.01340    0.07766   0.173  0.86307
## prs             0.01526    0.06626   0.230  0.81795
## vv             -0.33810    0.32441  -1.042  0.29804
## dvN            -4.00869    3.75173  -1.068  0.28604
## dvNE           -1.26119    3.62946  -0.347  0.72843
## dvNW           -1.20402    3.32374  -0.362  0.71739
## dvS            -1.01559    3.29148  -0.309  0.75785
## dvSE            4.27136    3.29232   1.297  0.19536
## dvSW           -5.31358    3.73461  -1.423  0.15569
## dvW            -1.11943    3.41721  -0.328  0.74342
## stagioneSpring -0.53671    1.57841  -0.340  0.73404
## stagioneSummer -3.38860    2.10639  -1.609  0.10858
## stagioneWinter  6.81371    1.66897   4.083 5.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.991 on 348 degrees of freedom
## Multiple R-squared:  0.1945, Adjusted R-squared:  0.1574
## F-statistic: 5.251 on 16 and 348 DF,  p-value: 6.402e-10
```

```r
mod1s <- step(mod1, direction = "both")
```

```
## Start:  AIC=1619.83
## media ~ tmp + umr + pgg + rdz + prs + vv + dv + stagione
##
##            Df Sum of Sq   RSS    AIC
## - rdz       1      2.41 28133 1617.9
## - prs       1      4.29 28135 1617.9
## - umr       1      5.86 28137 1617.9
## - vv        1     87.80 28219 1619.0
## <none>                  28131 1619.8
## - pgg       1    587.56 28718 1625.4
## - dv        7   1907.93 30039 1629.8
## - stagione  3   1575.89 29707 1633.7
## - tmp       1   1438.53 29569 1636.0
##
## Step:  AIC=1617.86
## media ~ tmp + umr + pgg + prs + vv + dv + stagione
##
##            Df Sum of Sq   RSS    AIC
## - prs       1      3.81 28137 1615.9
## - umr       1      4.34 28138 1615.9
## - vv        1     88.35 28222 1617.0
## <none>                  28133 1617.9
## + rdz       1      2.41 28131 1619.8
## - pgg       1    586.37 28720 1623.4
## - dv        7   1919.03 30052 1627.9
## - stagione  3   1575.50 29709 1631.8
## - tmp       1   1546.45 29680 1635.4
##
## Step:  AIC=1615.91
## media ~ tmp + umr + pgg + vv + dv + stagione
##
##            Df Sum of Sq   RSS    AIC
## - umr       1      3.49 28141 1614.0
## - vv        1    100.88 28238 1615.2
## <none>                  28137 1615.9
## + prs       1      3.81 28133 1617.9
## + rdz       1      1.93 28135 1617.9
## - pgg       1    588.14 28725 1621.5
## - dv        7   1939.40 30076 1626.2
## - stagione  3   1673.02 29810 1631.0
## - tmp       1   1544.21 29681 1633.4
##
## Step:  AIC=1613.95
## media ~ tmp + pgg + vv + dv + stagione
##
##            Df Sum of Sq   RSS    AIC
## - vv        1    123.93 28264 1613.6
## <none>                  28141 1614.0
## + umr       1      3.49 28137 1615.9
## + prs       1      2.96 28138 1615.9
## + rdz       1      0.74 28140 1615.9
## - pgg       1    637.05 28778 1620.1
```

```
## - stagione  3   1682.84 29824 1629.2
## - dv        7   2551.80 30692 1631.6
## - tmp       1   1637.03 29778 1632.6
##
## Step:  AIC=1613.56
## media ~ tmp + pgg + dv + stagione
##
##            Df Sum of Sq   RSS    AIC
## <none>                  28264 1613.6
## + vv        1    123.93 28141 1614.0
## + umr       1     26.53 28238 1615.2
## + prs       1     14.16 28250 1615.4
## + rdz       1      0.00 28264 1615.6
## - pgg       1    746.55 29011 1621.1
## - stagione  3   1689.63 29954 1628.8
## - tmp       1   1752.04 30017 1633.5
## - dv        7   2991.92 31256 1636.3
```

```
summary(mod1s)
```

```
##
## Call:
## lm(formula = media ~ tmp + pgg + dv + stagione, data = data_hills2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.787  -5.261  -1.493   4.465  58.489
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.50601    3.85693   3.761 0.000198 ***
## tmp             0.71471    0.15301   4.671 4.27e-06 ***
## pgg            -0.09983    0.03274  -3.049 0.002469 **
## dvN            -4.47059    3.71374  -1.204 0.229477
## dvNE           -1.15285    3.60532  -0.320 0.749338
## dvNW           -2.06169    3.22034  -0.640 0.522454
## dvS            -1.02555    3.27990  -0.313 0.754710
## dvSE            4.44593    3.24115   1.372 0.171026
## dvSW           -5.78110    3.68228  -1.570 0.117319
## dvW            -1.21057    3.39100  -0.357 0.721310
## stagioneSpring -0.51556    1.37708  -0.374 0.708343
## stagioneSummer -3.57923    1.99084  -1.798 0.073057 .
## stagioneWinter  6.71723    1.58779   4.231 2.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.961 on 352 degrees of freedom
## Multiple R-squared:  0.1906, Adjusted R-squared:  0.1631
## F-statistic:  6.91 on 12 and 352 DF,  p-value: 2.966e-11
```

```
# a giudicare da questo credo che eliminerò max e mediana e provo a comparare
# max e media con due modelli differenti in quanto sono sempre correlate e vedo
# se cambia qualcosa
```

```r
mod2 <- lm(media ~ tmp + vv + dv + rdz + pgg + umr + prs + stagione, data = data_hills)
mod2bis <- lm(media ~ tmp + vv + rdz + pgg + umr + prs, data = data_hills)
summary(mod2)
```

```
##
## Call:
## lm(formula = media ~ tmp + vv + dv + rdz + pgg + umr + prs +
##     stagione, data = data_hills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.211  -5.560  -1.675   4.297  58.754
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.24700   68.08949  -0.018  0.98540
## tmp             0.69836    0.16555   4.218 3.14e-05 ***
## vv             -0.33810    0.32441  -1.042  0.29804
## dvN            -4.00869    3.75173  -1.068  0.28604
## dvNE           -1.26119    3.62946  -0.347  0.72843
## dvNW           -1.20402    3.32374  -0.362  0.71739
## dvS            -1.01559    3.29148  -0.309  0.75785
## dvSE            4.27136    3.29232   1.297  0.19536
## dvSW           -5.31358    3.73461  -1.423  0.15569
## dvW            -1.11943    3.41721  -0.328  0.74342
## rdz             0.01340    0.07766   0.173  0.86307
## pgg            -0.09598    0.03560  -2.696  0.00736 **
## umr             0.01872    0.06953   0.269  0.78787
## prs             0.01526    0.06626   0.230  0.81795
## stagioneSpring -0.53671    1.57841  -0.340  0.73404
## stagioneSummer -3.38860    2.10639  -1.609  0.10858
## stagioneWinter  6.81371    1.66897   4.083 5.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.991 on 348 degrees of freedom
## Multiple R-squared:  0.1945, Adjusted R-squared:  0.1574
## F-statistic: 5.251 on 16 and 348 DF,  p-value: 6.402e-10
```

```r
summary(mod2bis)
```

```
##
## Call:
## lm(formula = media ~ tmp + vv + rdz + pgg + umr + prs, data = data_hills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.801  -5.122  -1.579   4.012  64.243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.31744   63.93233   0.943 0.346083
## tmp          0.36427    0.11091   3.284 0.001123 **
## vv          -0.43052    0.29634  -1.453 0.147163
```

```
## rdz          0.04983    0.07351   0.678 0.498238
## pgg         -0.12180    0.03511  -3.469 0.000586 ***
## umr          0.17066    0.06006   2.842 0.004744 **
## prs         -0.04867    0.06267  -0.777 0.437890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.402 on 358 degrees of freedom
## Multiple R-squared:  0.09388,    Adjusted R-squared:  0.07869
## F-statistic: 6.182 on 6 and 358 DF,  p-value: 3.465e-06
```

La stepwise serve a selezionare le variabili che spiegano di più. Chiaramente mettendo sia dv che stagione stai aggiungendo tanta carne al fuoco e hai poche osservazioni per alcune direzioni. Vediamo un attimino come si comporta

```
reg1 <- lm(media ~ dv, data = data_hills)
summary(reg1)
```

```
##
## Call:
## lm(formula = media ~ dv, data = data_hills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.888  -5.482  -1.806   4.177  60.895
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.2097     3.1332   7.727 1.13e-13 ***
## dvN          -2.1603     3.8035  -0.568   0.5704
## dvNE         -1.1854     3.7449  -0.317   0.7518
## dvNW         -0.1367     3.2845  -0.042   0.9668
## dvS           2.4857     3.3112   0.751   0.4533
## dvSE          6.8956     3.3006   2.089   0.0374 *
## dvSW         -2.4032     3.7728  -0.637   0.5246
## dvW           1.3324     3.4260   0.389   0.6976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.4 on 357 degrees of freedom
## Multiple R-squared:  0.09682,    Adjusted R-squared:  0.07911
## F-statistic: 5.467 on 7 and 357 DF,  p-value: 5.572e-06
```

l'unico vento che risulta significativo e manco tanto è SE che sembra incrementare la quantità di pm10 media. In fondo hai sfagiolato un po' e magari poi ne parliamo un po' a voce che ho visto un po' di confusione in giro. Non puoi usare i glm in quel modo e per i tuoi dati i glm non servono.

TI SEI DIMENTICATA L'ANALISI DEI RESIDUI!

```
reg2 <- lm(media ~ stagione, data = data_hills)
summary(reg2)
```

```
##
## Call:
## lm(formula = media ~ stagione, data = data_hills)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -15.320  -5.893  -1.485   3.716  65.341
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24.526      1.032  23.763   <2e-16 ***
## stagioneSpring    1.794      1.452   1.236    0.217
## stagioneSummer    2.356      1.444   1.631    0.104
## stagioneWinter    2.134      1.464   1.458    0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.791 on 361 degrees of freedom
## Multiple R-squared:  0.008955,   Adjusted R-squared:  0.0007191
## F-statistic: 1.087 on 3 and 361 DF,  p-value: 0.3544
```

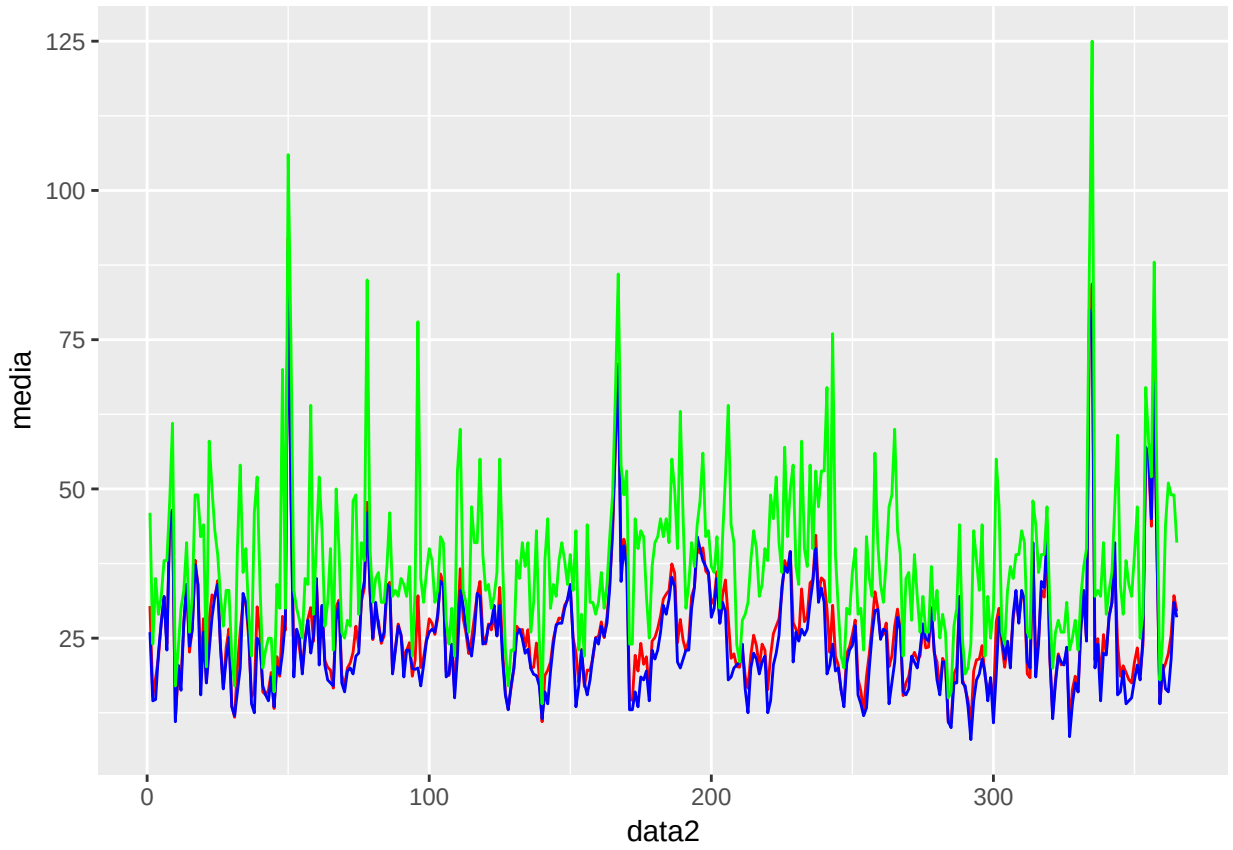
```
#cosi cambi il corner point
type2 <- relevel(data_hills2$stagione, ref = "Summer")

reg2.bis <- lm(media ~ type2, data = data_hills)
summary(reg2.bis)
```

```
##
## Call:
## lm(formula = media ~ type2, data = data_hills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.320  -5.893  -1.485   3.716  65.341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.8811     1.0099  26.617   <2e-16 ***
## type2Fall    -2.3556     1.4440  -1.631    0.104
## type2Spring  -0.5616     1.4360  -0.391    0.696
## type2Winter  -0.2217     1.4481  -0.153    0.878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.791 on 361 degrees of freedom
## Multiple R-squared:  0.008955,   Adjusted R-squared:  0.0007191
## F-statistic: 1.087 on 3 and 361 DF,  p-value: 0.3544
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
pm10$data2 <- seq(1,365,1)


ggplot(pm10, aes(x = data2)) +
  geom_path(aes(y = media), color = "red") +
  geom_path(aes(y = mediana), color = "blue") +
  geom_path(aes(y = max), color = "green")
```



Io proverei a fare i modelli sia usando la media che il max per vedere cosa cambia. Cmq fai l'analisi dei residui anche se sembrerebbe che i modelli lineari sono siano i più appropriati in questo caso. Un'altra cosa che puoi fare per pulire un po' i modelli è standardizzare le variabili anche se non credo cambiarà tanto ma tu comunque provaci che almeno ti rivedi come si fa.