ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Μεταπτυχιακό Πρόγραμμα Συμπληρωματικής Ειδίκευσης
Εφαρμοσμένη Στατιστική

Μάθημα: Στατιστική Μάθηση
Statistical Learning

Εργασία 2η: Clustering
PCA
Factor Analysis

Καθηγήτρια: Παπαγεωργίου Ιουλία

Ημερομηνία Παράδοσης εργασίας: 7/6/2020

Κωνσταντίνα Τσάμη (p3621817)

ΠΕΡΙΕΧΟΜΕΝΑ

Περιγραφή του Προβλήματος

Τα δεδομένα που έχουμε στη διάθεσή μας αφορούν διαφορετικά είδη πουλιών τα οποία παρατηρήθηκαν σε διαφορετικές περιοχές στις οποίες μετρήθηκαν κάποια χαρακτηριστικά. Τα χαρακτηριστικά αυτά αφορούν τρείς μεγάλες κατηγορίες. Την κατηγορία της σύνθεσης του εδάφους (γεωμορφολογίας), την κατηγορία της σύνθεσης του τοπίου της περιοχής και την κατηγορία της ετερογένειας του τοπίου.

Στην πρώτη κατηγορία έχουμε 14 μεταβλητές στο συνολό τους, στην δεύτερη 20 και στη τελευταία κατηγορία 14. Άλλη μία μεταβλητή η οποία περέχεταιι στην τελευταία στήλη των δεδομένων μας αφορά την συχνότητα που παρατηρήθηκαν τα εκάστοτε ήδη πουλιών σε κάποιο ορισμένο χρονικό διάστημα. Στη διάθεσή μας έχουμε επίσης και τις κωδικές ονομασίες των ειδών των πτηνών στην πρώτη στήλη των δεδομένων μας.

Στόχος μας, στην παρούσα εργασία είναι να κατηγοριοποιήσουμε τα είδη αυτά ως προς κάποια κοινά περιβαλλοντολογικά χαρακτηριστικά των κατηγοριών που αναφέραμε. Η κατηγοριοποίηση αυτή μας είναι άγνωστη και καλούμαστε εμείς να τη διερευνήσουμε. Το πρόβλημά μας δηλαδή, δεν περιλαμβάνει κάποια μεταβλητή ενδιαφέροντος την οποία προσπαθούμε να προβλέψουμε και βάση της οποίας να ταξινομήσουμε τις παρατηρήσεις

. Η μεταβλητή αυτή αν υπάρχει δεν είναι καταγεγραμμένη σε αυτό το στάδιο της έρευνας. Θα πρέπει λοιπόν να ανιχνεύσουμε κάποια στοιχεία – patterns στα δεδομένα μας που να μας οδηγούν σε κάποια ομαδοποίηση αυτών των ειδών.

Αναλυτικότερη περιγραφή των δεδομένων μας μπορεί να βρεθεί στο παρακάτω αρχείο.

¸ããñáöï
êåéìÝíïõ

Clustering στις αρχικές μεταβλητές

Ξεκινάμε την ανάλυσή μας κάνοντας Clustering και έτσι ομαδοποιώντας τις παρατηρήσεις μας με βάση τις αρχικές μεταβλητές. Θα χρησιμοποιήσουμε **Ιεραρχική** και **k-means** μέθοδο και θα αξιολογήσουμε τα αποτελέσματα.

Περνάμε λοιπόν τα δεδομένα μας στην R και αφαιρούμε τις παρατηρήσεις με απόλυτη συχνότητα μικρότερη του 4. Τα είδη πουλιών εκείνα δηλαδή, που παρατηρήθηκαν λιγότερο από 4 φορές στην εκάστοτε περιοχή.

Διαβάζουμε το αρχείο δεδομένων.

```
set <- read.csv(file = file.choose(), header = T)
```

Αφαιρούμε τις αντίστοιχες γραμμές με συχνότητα μικρότερη του 4.

```
set <- set[set$NOBLOCKS>=4,]
```

- **Για την Ιεραρχική μέθοδο ακολουθούμε την παρακάτω διαδικασία.**

Υπολογίζουμε τον πίνακα dissimilarities (D), ο οποίος περιέχει τις αποστάσεις όλων των παρατηρήσεων μεταξύ τους. Ως μέτρο επιλέγουμε την ευκλείδια απόσταση: `method = euclidian`. Επίσης, **τυποποιούμε** τα δεδομένα μας (`scale`) ώστε να βρίσκονται όλες οι μεταβλητές στην ίδια κλίμακα, καθώς η ιεραρχική μέθοδος όπως και η k-means είναι ευαίσθητες στην διαφοροποίηση των διακυμάνσεων.

```
d <- dist(scale(set[,-1]), method = "euclidian")
```

Εν συνεχεία, θα εφαρμόσουμε τη μέθοδο για διάφορα “linkages”-μεθόδους.

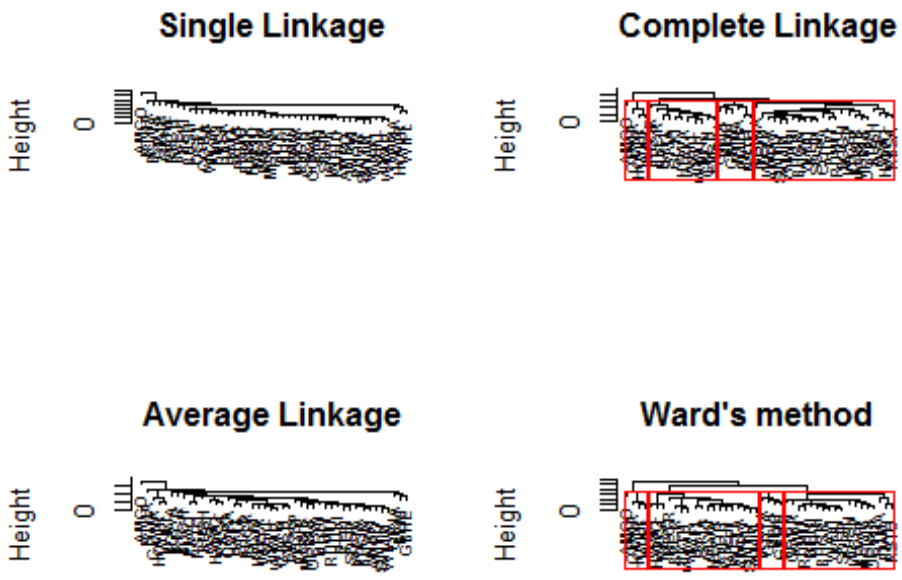
```
fit.s <- hclust(d, method = "single")
fit.c <- hclust(d, method = "complete")
fit.a <- hclust(d, method = "average")
fit.d <- hclust(d,method="ward.D")
fit.d2 <- hclust(d,method="ward.D2")
```

Κατασκευάζουμε τα δενδροδιαγράμματα που προκύπτουν από κάθε μία μέθοδο.

```
par(mfrow=c(2,2))
plot(fit.s, main = "Single Linkage", sub = "", xlab = "",
labels=set[,1], cex=.6)
plot(fit.c, main = "Complete Linkage", sub = "", xlab = "",
labels=set[,1], cex=.6)
rect.hclust(fit.c, 4)
plot(fit.a, main = "Average Linkage", sub = "", xlab = "",
labels=set[,1], cex=.6)
plot(fit.d2, main = "Ward's method", sub = "", xlab = "",
labels=set[,1], cex=.6)
rect.hclust(fit.d2, 4)
```

Τα δενδροδιαγράμματα που προκύπτουν φαίνονται παρακάτω στο *Διάγραμμα 1*.

Διάγραμμα 1- Δενδροδιαγράμματα διαφορετικών Ιεραρχικών μεθόδων με μέτρο απόστασης την ευκλείδια.



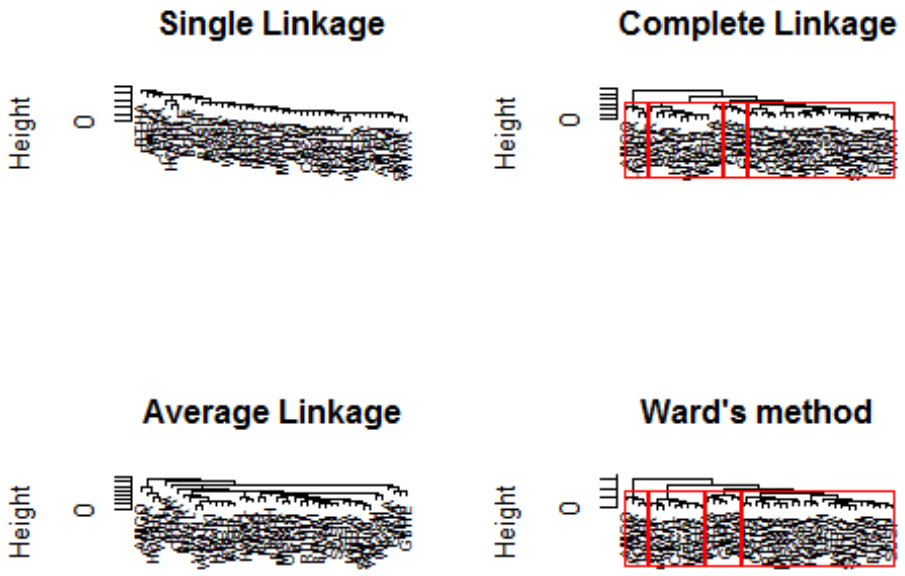
Μπορούμε να παρατηρήσουμε 4 κυρίως ομάδες στο *Διάγραμμα 1* και στη μέθοδο "Complete Linkage" και στη μέθοδο "Ward" και επίσης παρατηρούμε κάποια κοινά μοτίβα.

Θα επαναλάβουμε τη διαδικασία επιλέγοντας τώρα ως μέτρο απόστασης την **"Manhattan"**,`method = manhattan`.

Παρατηρούμε όμοια σχεδόν αποτελέσματα κυρίως στις δύο μεθόδους που επιλέξαμε ("Ward's" και "Complete Linkage") σύμφωνα με το *Διάγραμμα 2*. Κυρίως η πρώτη μικρή ομάδα και η τελευταία μεγαλύτερη ομάδα βλέπουμε να επαναλαμβάνεται. Παρατηρούμε επίσης στο *Διάγραμμα 2* ότι και οι δύο μέθοδοι χωρίζουν ακριβώς με τον ίδιο τρόπο τις παρατηρήσεις αν κόψουμε τα δενδρο-διαγράμματα λίγο πιο ψηλά και δημιουργήσουμε **3** ομάδες. Θα ενωθούν έτσι η δεύτερη και η τρίτη ομάδα σχηματίζοντας μια κοινή ισοπληθή ομάδα και στις δύο περιπτώσεις. Το ίδιο παρατηρούμε και στα αποτελέσματα με μέτρο απόστασης της "ευκλείδια" *(Διάγραμμα 1),* αν ενωθούν η 2η και η 3η ομάδα στην "Complete Linkage" και οι 2 τελευταίες στην "Ward's method".

Μπορούμε να πούμε λοιπόν με μεγαλύτερη βεβαιώτητα ότι οι **ομάδες** που μπορούν να χωριστούν τα είδη πουλιών είναι **3**. Ενδεχομένως, να μπορούν να χωριστούν και σε 4 ή και 5 μικρότερες. Τα είδη πουλιών που είναι πιο κοντά στο διάγραμμα, όπως για παράδειγμα τα *Purple Finch (*PUFI) και *Rufous Humingbird (*RUHU) είναι αυτά που εμφανίζουν τη μεγαλύτερη ομοιογένεια ως προς τα χαρακτηριστικά του περιβάλλοντος που ζουν.

Διάγραμμα 2- Δενδροδιαγράμματα διαφορετικών Ιεραρχικών μεθόδων με μέτρο απόστασης την ***"manhattan"****, έχοντας αφαιρέσει ακραίες τιμές .*



Από την Ward's μέθοδο (*Διάγραμμα 2*) μπορούμε να παρατηρήσουμε και 2 υπο-ομάδες ίσως οι οποίες δεν ξέρουμε πόσο πολύ διαφέρουν.

```
cluster.d2.4 <- cutree(fit.d2,3)
cluster.d2.5 <- cutree(fit.d2,4)
table(cluster.d2.3)
table(cluster.d2.4)
```

- **Για την k-means ακολουθούμε την παρακάτω διαδικασία.**

Στη μέθοδο k-means χρειάζεται να ορίσουμε εκ των προτέρων τον αριθμό των ομάδων στις οποίες θα ταξινομήσει τις παρατηρήσεις. Εμείς για αρχή θα ορίσουμε K=3 ομάδες όπως αξιολογήσαμε στην προηγούμενη μέθοδο και θα επανεξετάσουμε τα αποτελέσματα. Το συγκριτικό πλεονέκτημα αυτής της μεθόδου σε σχέση με την Ιεραρχική που χρησιμοποιήσαμε προηγουμένως, είναι ότι σε οποιοδήποτε στάδιο της διαδικασίας κάποια παρατήρηση που έχει τοποθετηθεί σε μία ομάδα μπορεί να αλλάξει.

Πρέπει να είμαστε ιδιαίτερα προσεκτικοί στη περίπτωση που έχουμε **ακράιες τιμές** καθώς ο αλγόριθμος k-means είναι ευαίσθητος στην ύπαρξη ακραίων τιμών. Επίσης, χρησιμοποιούμε και εδώ, όπως προηγουμένως, τα **τυποποιημένα** δεδομένα καθώς και αυτή η μέθοδος επηρεάζεται από τις διαφοροποιήσεις στην κλίμακα μεταξύ των μεταβλητών.

```
set.seed(999)
kmeans.3 <- kmeans(scale(set[,-1]), centers = 3, iter.max = 25,
trace = TRUE)
table(kmeans.4$cluster)
```

Πίνακας 1- Πλήθος παρατηρήσεων στις ομάδες με την Ιεραρχική μέθοδο με μέτρο απόστασης "manhattan" και "Ward's" linkage και με τη μέθοδο k-means

```
> table(cluster.d2.3 )

cluster.d2.3

 1  2  3

 4 39  9

> table(kmeans.3$cluster)

 1  2  3

 4 41  7
```

Βλέπουμε διαφορές στο τρόπο που ταξινομήθηκαν οι παρατηρήσεις στις δύο μεθόδους αλλά όχι πολύ μεγάλες.

Σε γενικές γραμμές, δεν μπορούμε να αξιολογήσουμε τα αποτελέσματα σε αυτό το σημείο καθώς οι παρατηρήσεις που έχουμε συνολικά μετά και από την αφαίρεση των ειδών που παρατηρήθηκαν με μικρή συχνότητα είναι 52, οι επεξηγηματικές μεταβλητές είναι 48 συνολικά. Θα προσπαθήσουμε λοιπόν να μειώσουμε την διάσταση του σετ δεδομένων ούτως ώστε να απλουστεύσουμε την διαδικασία, να αποδώσουν οι αλγόριθμοι καλύτερα (κυρίως ο k-means) και να μπορέσουμε να οπτικοποιήσουμε και τα δεδομένα.

Είναι απαραίτητος έτσι, ένας μετασχηματισμός των επεξηγηματικών μεταβλητών σε κύριες συνιστώσες (Principal Components), ώστε να μειώσουμε τη διάσταση.

Clustering στις PCA μεταβλητές

Σε αυτό το σημείο θα δοκιμάσουμε να ομαδοποιήσουμε τα δεδομένα μας, χρησιμοποιώντας, όμως, τις κύριες συνιστώσες των αρχικών μεταβλητών μας. Η κίνηση αυτή θα μας βοηθήσει να αποφύγουμε προβλήματα που μπορεί να οφείλονται σε πιθανή "υπερ-μετροποίηση" (όταν p >= =n) ή συσχέτιση μεταξύ των μεταβλητών. Στην περίπτωσή μας, θα χρησιμοποιήσουμε κύριες συνιστώσες οι οποίες θα αντιστοιχούν στην κάθε ομάδα μεταβλητών: γεωμορφολογία (geomorphology), σύνθεση τοπίου (landscape composition) και ετερογένεια του τοπίου της περιοχής όπου συναντόνται τα διάφορα είδη (landscape heterogeneity).

Εφαρμόζουμε λοιπόν Principal Components ετελώντας τις παρακάτω εντολές.

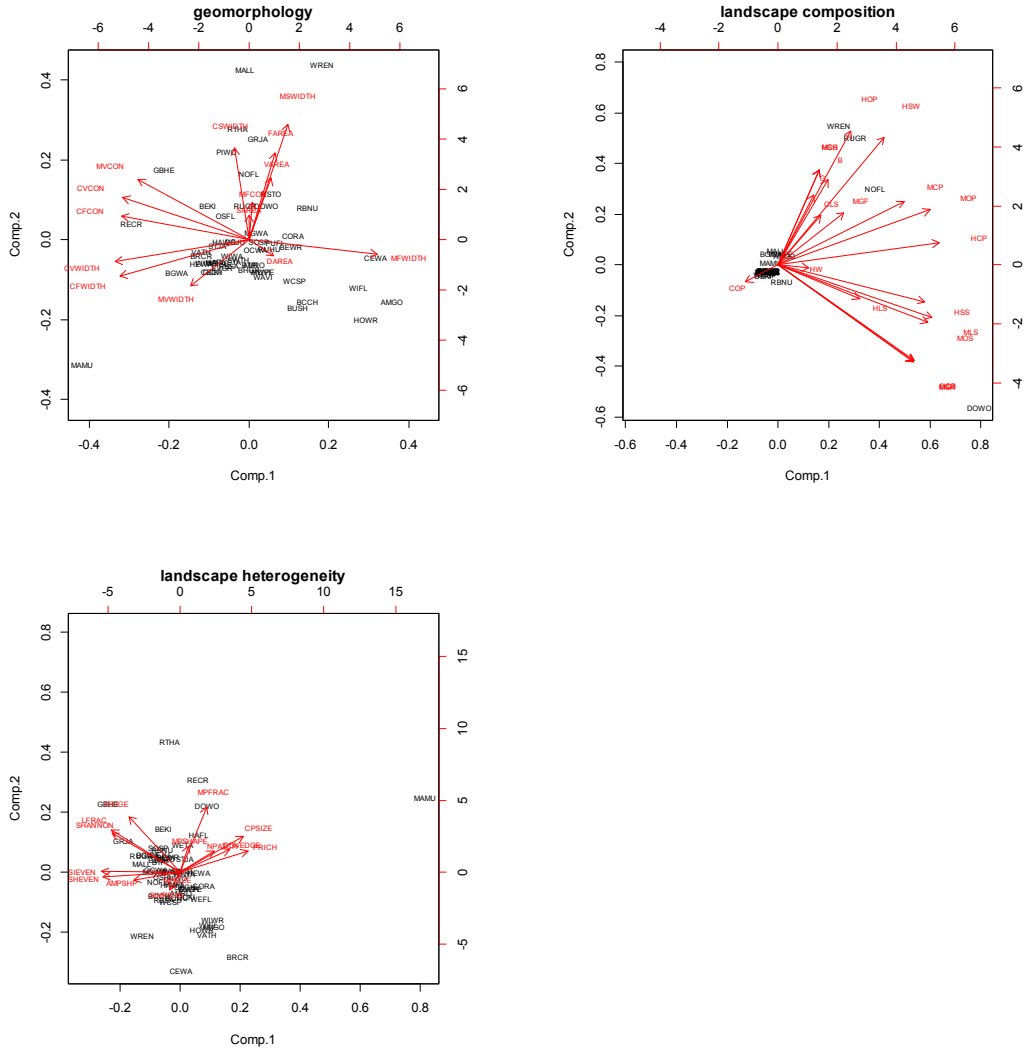
```
set.pr1<-princomp(scale(set[,2:15]),cor=T)
summary(set.pr1)
set.pr1$loadings
screeplot(set.pr1, npcs = 15, type = "lines")

set.pr2<-princomp(scale(set[,16:35]),cor=T)  #5
summary(set.pr2)
set.pr2$loadings
screeplot(set.pr2, npcs = 15, type = "lines") #5

set.pr3<-princomp(scale(set[,36:49]),cor=T)   #5
summary(set.pr3)
set.pr3$loadings
screeplot(set.pr3, npcs = 15, type = "lines")
```

Θα κρατήσουμε **5** κύριες συνιστώσες για κάθε σετ μεταβλητών. Κρατάω όσες έχουν ιδιοτιμή μεγαλύτερη του 1 (standard deviation) και εξηγούν ένα ικανοποιητικό ποσοστό της διακύμανσης των αρχικών μεταβλητών (Cumulative Proportion) (επίσης παρατηρώντας τα αντίστοιχα screeplots, σημεία μετά τα οποία ο ρυθμός αύξησης της διακύμανσης μειώνεται σημαντικά) . Στην πρώτη περίπτωση και στο σετ μεταβλητών που αφορούν την γεωμορφολογία, οι 5 πρώτες κύριες συνιστώσες εξηγούν το **77.82%** της αρχικής διακύμανσης των μεταβλητών. Στην περίπτωση του σετ μεταβλητών για τη σύνθεση του τοπίου της περιοχής , οι 5 πρώτες κύριες συνιστώσες εξηγούν το **88.47%** της αρχικής διακύμανσης, ενώ στην περίπτωση των μεταβλητών που αφορούν την ετερογένεια του τοπίου οι κύριες συνιστώσες εξηγούν το **84.69%** της αρχικής διακύμανσης. (Σύμφωνα με τους περιληπτικούς πίνακες αποτελεσμάτων που μας έσωσε η R, *βλέπε Πίνακες 1-3, ΠΑΡΑΡΤΗΜΑ*).

Κατασκευάζουμε και τα “biplots” των παρατηρήσεων ως προς τις 2 πρώτες κύριες συνιστώσες (PCA) κάθε μίας από τις 3 κατηγορίες μεταβλητών

```
par(mfrow=c(1,2))
biplot(set.pr1, choices=c(1,2), xlabs=set[,1], cex=.6,
        main="geomorphology")
biplot(set.pr2, choices=c(1,2), xlabs=set[,1], cex=.6,
       main="landscape composition")
par(mfrow=c(1,2))
biplot(set.pr3, choices=c(1,2), xlabs=set[,1], cex=.6,
       main="landscape heterogeneity")
```

*Διάγραμμα 3- **“biplots”** ως προς της δύο πρώτες κύριες συνιστώσες κάθε κατηγορίας.*



Στο *Διάγραμμα 3* και στα διαφορετικά “biplots” για κάθε κατηγορία μεταβλητών παρατηρείται η ύπαρξη κάποιων ακραίων τιμών. Κάποιες παρατηρήσεις είναι πολύ μακριά σε σχέση με όλες της υπόλοιπες. Για παράδειγμα, στην κατηγορία της γεωμορφολογίας (geomorphology) παρατηρούμε ότι το είδος Marbled Murrelet (MAMU) διαφέρει πολύ ως προς τα γεωμορφολογικά χαρακτηριστικά από όλα τα άλλα είδη (εξαιρετικά ακραία τιμή). Εξίσου πολύ διαφέρει και το είδος Downy Woodpecker (DOWO) ως προς τα χαρακτηριστικά

της σύνθεσης του τοπίου της περιοχής (landscape composition) το οποίο έχει εξαιρετικά υψηλές τιμές σε δύο μετρήσεις των χαρακτηριστικών της κατηγορίας αυτής.

Στην κατηγορία της ετερογένειας του του τοπίου της περιοχής (landscape heterogeneity) μια εξαιρετικά ακραία τιμή αποτελεί το είδος Marbled Murrelet (MAMU) και εντοπίζουμε και άλλες ακρείες τιμές σε όλες τις κατηγορίες. Τα είδη που εμφανίζονται πιο κοντά θα λέγαμε πως έχουν παρόμοια γωμορφολογικά χαρακτηριστικά ή χαρακτηριστικά σύνθεσης ή ετερογένειας του τοπίου, αναλόγως την κατηγορία. Κάτι επίσης ενδιαφέρον που παρατηρούμε στο διάγραμμα της σύνθεσης τοπίου (landscape composition) είναι ότι όλες οι παρατηρήσεις φαίνονται συγκεντρωμένες στο κέντρο με 4 μόνο εξαιρέσεις. Δηλαδή, όλα τα είδη μοιράζονται παρόμοια χαρακτηριστικά ως προς την σύνθεση του τοπίου και μόνο 4 είδη διαφέρουν σημαντικά παρουσιάζοντας πολύ υψηλές τιμές σε κάποιες μετρήσεις χαρακτηριστικών-μεταβλητές της κατηγορίας.

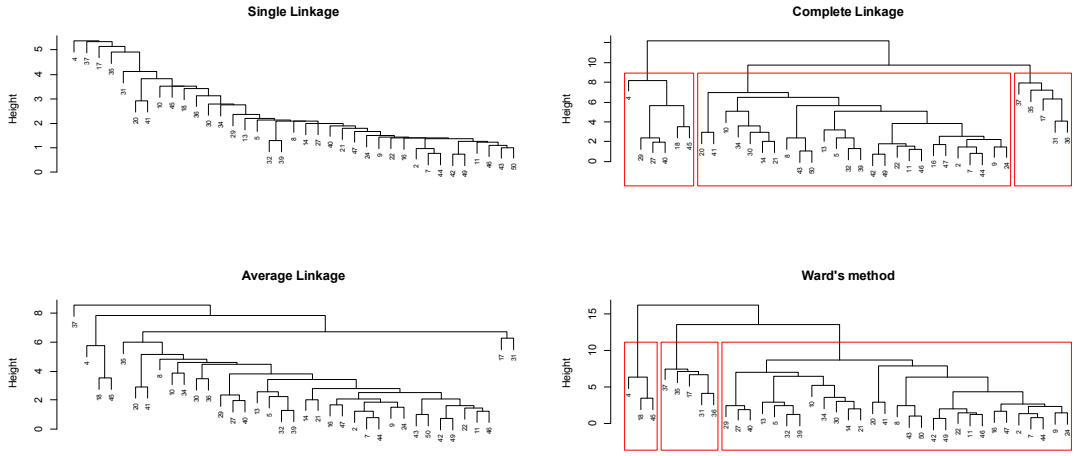
Να σημειώσουμε εδώ, ότι οι ακραίες τιμές που ταυτοποιήσαμε δεν έχουν να κάνουν με το ευρύτερο σετ δεδομένων μας ούτε χρειάζεται να εξαιρεθούν. Θα έπρεπε κάποια ακραία τιμή να εμφανίζεται και στις 3 κατηγορίες για να πούμε ότι είναι όντως ακραία και δε μπορεί να ομαδοποιηθεί με άλλες παρατηρήσεις. Με άλλα λόγια μπορεί ένα είδος να μην έχει παρόμοια γεωμορφολογικά στοιχεία με άλλα είδη ,αλλά να μοιράζεται παρόμοια χαρακτηριστικά σύνθεσης και ετερογένειας τοπίου.

Κάνουμε εκ νέου clustering χρησιμοποιώντας αυτή τη φορά τα “scores” των κυρίων συνιστωσών (PCA) που επιλέξαμε.

```
new.set<-cbind(set.pr1$scores[,1:5], set.pr2$scores[,1:5],
               set.pr3$scores[,1:5], set[,1])

#dissimilarities matrix
d <- dist(new.set[,-ncol(new.set)], method = "euclidean")

# hierarchical, different linkages
fit.s <- hclust(d, method = "single")
fit.c <- hclust(d, method = "complete")
fit.a <- hclust(d, method = "average")
fit.d<-hclust(d,method="ward.D")
fit.d2<-hclust(d,method="ward.D2")

par(mfrow=c(2,2))
plot(fit.s, main = "Single Linkage", sub = "", xlab = "",  cex=.6)
plot(fit.c, main = "Complete Linkage", sub = "", xlab = "",  cex=.6)
rect.hclust(fit.c, 3)
plot(fit.a, main = "Average Linkage", sub = "", xlab = "",  cex=.6)
plot(fit.d2, main = "Ward's method", sub = "", xlab = "",  cex=.6)
rect.hclust(fit.d2, 3)

#afairesh outliers
new.set<-new.set[-c(51,28,15,52,26,38,19,6,3,33,1,12,23,48,25),]
```

Κατασκευάζουμε εκ νέου τα δεδνδροδιαγράμματα για διαφορετικά μέτρα απόστασης. Συγκεκριμένα για “manhattan”, “euclidean” και “minkowski”. Και στις 3 περιπτώσεις τα διαγράμματα παρουσιάζουν την ίδια εικόνα και στα linkages “Ward’s” και “Complete” διακρίνουμε **3** ομάδες *(Διάγραμμα 3)*.

*Διάγραμμα 3 - Δενδροδιαγράμματα των PCA μεταβλητών για διάφορα "linkages" και με μέτρο απόστασης την **"manhattan"**.*

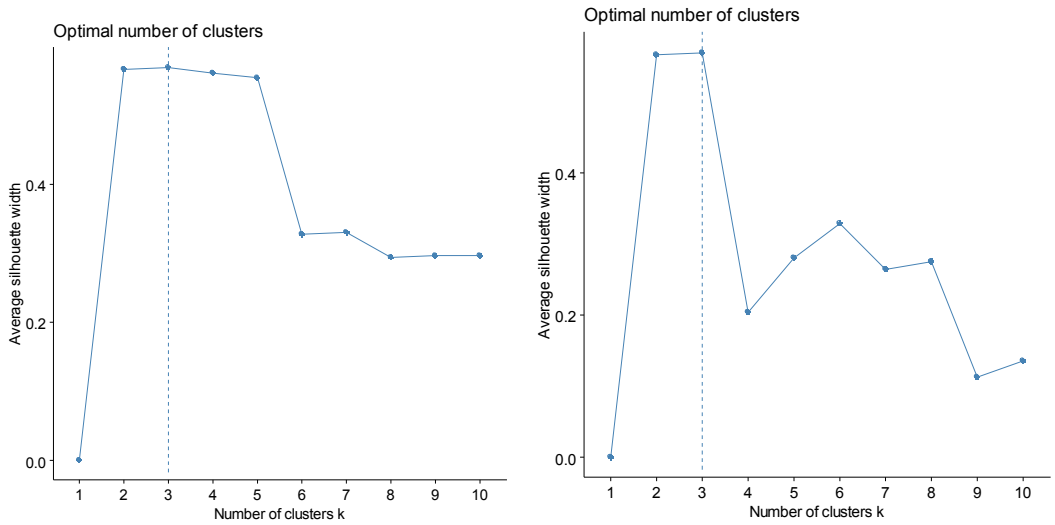


Προχωρούμε με τη μέθοδο **k-means** όπου θα ορίσουμε 3 ομάδες. Στη συνέχεια, θα αξιολογήσουμε τον ιδανικό αριθμό ομάδων, χρησιμοποιώντας διαφορετικούς στατιστικούς δείκτες.

```
set.seed(777)
kmeans.3 <- kmeans(new.set[,-ncol(new.set)], centers = 3, iter.max =
25, trace = TRUE)
table(kmeans.3$cluster)
kmeans.3

library(factoextra)
library(NbClust)
fviz_nbclust(new.set[,-ncol(new.set)],kmeans, method = "wss")
fviz_nbclust(new.set[,-ncol(new.set)],kmeans, method = "silhouette")

hier_cluster_fun <- function(x,k){
    list( cluster = cutree(hclust(dist(x,method="euclidean"),
                      method="ward.D2"),k=k))

fviz_nbclust(new.set[,-ncol(new.set)]),hier_cluster_fun , method =
"silhouette")

gap_stat <- clusGap(new.set[,-ncol(new.set)], FUN = kmeans, nstart =
25, K.max = 10, B = 50)
fviz_gap_stat(gap_stat)
```

Από το *Διάγραμμα 4*, βλέπουμε πως ο ιδανικός αριθμός ομάδων που μας προτείνουν οι μέθοδοι k-means και Ιεραρχική είναι **3**.

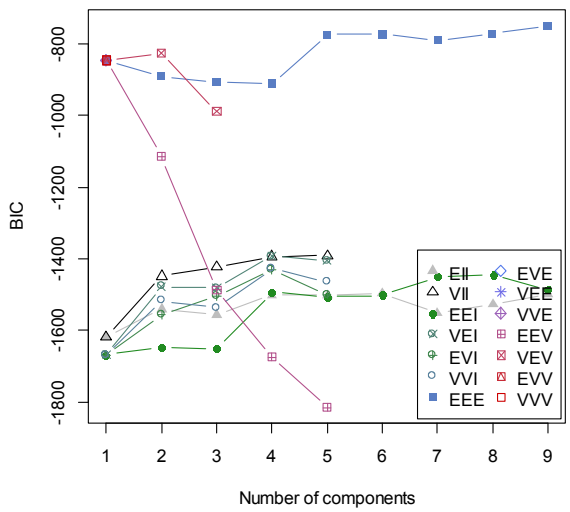
*Διάγραμμα 4- Διάγραμμα silhouette της μεθόδου **k-means** και της Ιεραρχικής μεθόδου των PCA μεταβλητών.*



Θα εφαρμόσουμε επίσης και **model based** μέθοδο (αλγόριθμο **EM**).

```
library(mclust)
fit.m <- Mclust(scale(new.set[,-ncol(new.set)]))
names(fit.m)
plot(fit.m, new.set[,-ncol(new.set)], what = 'BIC')
print(fit.m)
```

Διάγραμμα 5- Διάγραμμα της τιμής του κριτηρίου BIC για τα διαφορετικό πλήθος ομάδων (components) και διαφορετικά μοντέλα .



Το μοντέλο που μας προτείνεται ως καλύτερο από τη μέθοδο model based είναι όπως φαίνεται στο *Διάγραμμα 5* το EEE (υποθέτει ότι τα χαρακτηριστικά volume, shape, orientation είναι ίσα ανάμεσα στα groups) για 9 components-clusters. Παρατηρούμε ωστόσο, ότι για πλήθος components ίσο με **5** η διαφορά της τιμής BIC είναι πολύ μικρή, μας παρέχει δηλαδή παρόμοια πληροφορία με μικρότερο αριθμό clusters, σε αντίθεση με 9 clusters και πολλές ομάδες να περιέχουν μόλις μία παρατήρηση όπως φαίνεται στον *Πίνακα 2* . Λόγω του ότι το μέγεθος του δείγματος δεν είναι αρκετά μεγάλο και ο αριθμός μεταβλητών είναι αρκετός (15 μεταβλητές στο σύνολο), το μοντέλο EEE είναι και το μοναδικό που μπορεί να υπολογιστεί με ακρίβεια λόγω των λιγότερων παραμέτρων. Τα υπόλοιπα μοντέλα απαιτούν ένα επαρκές μέγεθος δείγματος σε σχέση με τον αριθμό των μεταβλητών και των παραμέτρων.

Πίνακας 2 - Πίνακας συχνοτήτων στις 9 ομάδες που προέκυψαν με την model based μέθοδο.

```
> table(fit.m$classification)

 1  2  3  4  5  6  7  8  9

 8  3 13  5  3  1  2  1  1
```

```
fit.m2 <- Mclust(new.set[,-ncol(new.set)],G=5)
print(fit.m2)
```

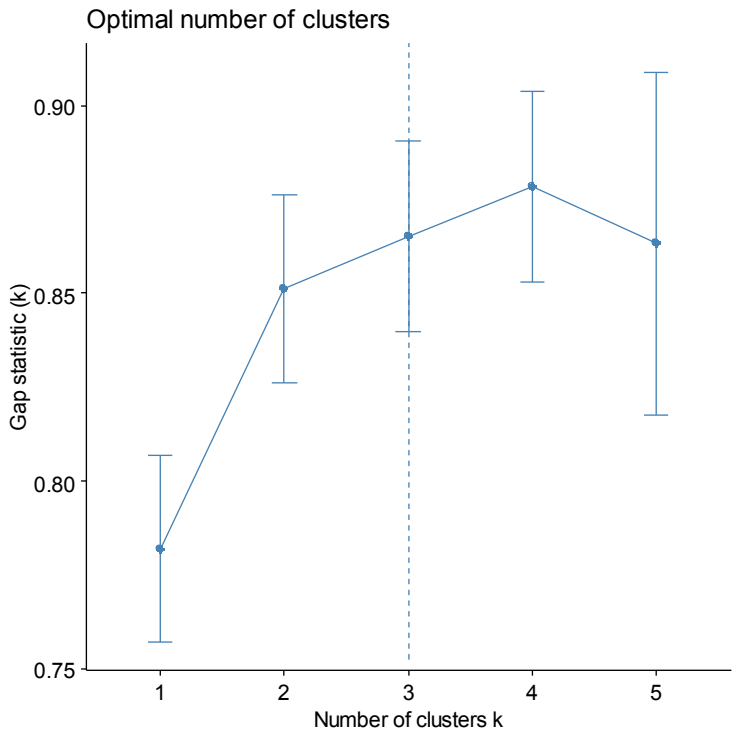
Υπολογίζουμε τον ιδανικό αριθμό ομάδων σύμφωνα με τον δείκτη **gap-statistic** για την model based μέθοδο ως εξής.

```
m_cluster_fun <- function(x,k){
                      list(cluster=Mclust(x,G=k)$classification)
                }
gap_stat.mclust <- clusGap(scale(set[,-1]), m_cluster_fun, K.max =
5, B = 50)

fviz_gap_stat(gap_stat.mclust)
```

Το αποτέλεσμα που παίρνουμε είναι το παρακάτω διάγραμμα.

Βλέπουμε λοιπόν, σύμφωνα με το *Διάγραμμα 6* ότι ο ιδανικός αριθμός clusters που μας προτείνει ο δείκτης gap-statistic είναι **3**.

*Διάγραμμα 6- Διάγραμμα gap-statistic της **model based** μεθόδου των PCA μεταβλητών.*



Σε αυτό το σημείο, σειρά έχει να συγκρίνουμε την συσταδοποίηση που προέκυψε από τις τρείς μεθόδους clustering που εφαρμόσαμε (hierarchical, k-means και model-based) για τον ιδανικό αριθμό clustes που προέκυψε από την κάθε μία (3 ομάδες ο ιδανικός αριθμός clusters για όλες τις μεθόδους).

Για τον σκοπό αυτό θα χρησιμοποιήσουμε πάλι τον δείκτη gap-statistic. Η υψηλότερη τιμή του σημαίνει και καλύτερη ταξινόμηση των παρατηρήσεων στις τρεις ομάδες με την αντίστοιχη μέθοδο. Επίσης, θα χρησιμοποιήσουμε και τον δείκτη silhouette ο οποίος παίρνει τιμές στο διάστημα [-1, 1]. Όσο πιο κοντά στο 1 είναι η τιμή του τόσο μεγαλύτερος ο βαθμός συμφωνείας.

```
gap_stat.kmeans <- clusGap(scale(new.set[,-ncol(new.set)]), FUN =
kmeans, nstart = 25, K.max = 5, B = 50)

hier_cluster_fun <- function(x,k){
    list( cluster = cutree(hclust(dist(x,method="euclidean"),
                      method="ward.D2"),k=k))
                }
gap_stat.hier <- clusGap(new.set[,-ncol(new.set)], FUN =
hier_cluster_fun, K.max = 5, B = 50)

print(list("hierarchical"=gap_stat.hier$Tab[3,3],
           "k-means"=gap_stat.kmeans$Tab[3,3],
           "model-based"=gap_stat.mclust$Tab[3,3]))
```

Πίνακας 3 - Πίνακας της τιμής του gap-statistic για τις τρεις διαφορετικές μεθόδους των αρχικών και των pca μετασχηματισμένων μεταβλητών.

Clustering μέθοδος		
	gap-statistic	silhouette
Hierarchical	0.9274654	0.6190661
k-means	0.9040119	0.6181528
model-based	0.8352118	0.2549085

Η υψηλότερη τιμή του δείκτη gap-statistic στις αρχικές μεταβλητές είναι ίση με 0.92029599 *(Πίνακας 3)* και αντιστοιχεί στην **Ιεραρχική** μέθοδο. Η υψηλότερη τιμή του δείκτη silhouette εμφανίζεται επίσης στην Ιεραρχική μέθοδο είναι πού κοντά όμως σε αυτή της k-means μεθόδου (silhouette=0.6190661, kmeans=0.6181528) . Επίσης, ο δείκτης silhouette αρκετά μεγαλύτερος από 0 και κοντά στο 1, δείχνει καλό διαχωρισμό των παρατηρήσεων μεταξύ των ομάδων. Σε αντίθεση έρχεται η model-based μέθοδος που αν και έχει την μικρότερη τιμή και στους δύο δείκτες , σύμφωνα με τον δείκτη silhouette η διαχωριστική ικανότητα είναι αρκετά χειρότερη από τις 2 άλλες μεθόδους (πιθανότατα δεν λειτουργεί καλά λόγω των πολλών παραμέτρων παρά την μείωση της διάστασης).

Για να αξιολογήσουμε την καλύτερη ικανότητα συσταδοποίησης μέσω του δείκτη **"silhouette"** εκτελέσαμε τις παρακάτω εντολές.

```
fit.m3 <- Mclust(scale(new.set[,-ncol(new.set)]),G=3)
kmeans.3 <- kmeans(scale(new.set[,-ncol(new.set)]), centers = 3,
iter.max = 25, trace = TRUE)

sil1 <- silhouette(kmeans.3$cluster, dist(scale(new.set[,-
ncol(new.set)]), "euclidean"))
sil2 <- silhouette(cutree(fit.d2,3), dist(scale(new.set[,-
ncol(new.set)]), "euclidean"))
sil3 <- silhouette(fit.m3$classification, dist(scale(new.set[,-
ncol(new.set)]), "euclidean"))

mean(sil1[,3]);mean(sil2[,3]);mean(sil3[,3])
```

Τα ευρήματα αυτά θα τα αξιολογήσουμε και οπτικοποιώντας τον διαχωρισμό στις ομάδες κάθε μεθόδου ως προς τις 2 πρώτες κύριες συνιστώσες κάθε κατηγορίας μεταβλητών (geomorphology, landscape composition and landscape heterogeneity).

Η οπτικοποίηση των δεδομένων θα γίνει με τη βοήθεια του πακέτου **"ggbiplot"** της R και θα την εξετάσουμε στο τελευταίο κεφάλαιο.

Factor Analysis

Σαν τελευταία ανάλυση θα δοκιμάσουμε παραγοντική ανάλυση (Factor Analysis) για την μείωση της διάστασης του προβλήματος και την διερεύνηση κάποιων λανθανουσών μεταβλητών (lattent variables, δηλαδή μεταβλητών που δεν έχουν καταγραφεί σε πρώτο στάδιο της έρευνας και περιγράφουν τις υπόλοιπες μεταβλητές) ή μεταβλητών κοινών παραγόντων (common factors, δηλαδή μεταβλητές που δεν έχουν νόημα, αλλά ερμηνεύουν τις αρχικές).

Ο τρόπος που θα κάνουμε την ανάλυσή μας είναι παρόμοιος με αυτόν των κυρίων συνιστωσών (PCA) που ακολουθήσαμε προηγουμένως. Δηλαδή, θα ερευνήσουμε την ύπαρξη factor ή factors για κάθε μία από τις τρεις κατηγορίες μεταβλητών που έχουμε.

Θα διαβάσουμε αρχικά τις απαραίτητες βιβλιοθήκες στην R.

```
require(FactoMineR)
library(psych)
```

Ξεκινούμε την ανάλυσή μας με τις μεταβλητές που μετρούν τα γεωμορφολογικά χαρακτηριστικά (**geomorphology**).

Ερευνούμε την ύπαρξη συσχετίσεων, απαραίτητο πρώτο βήμα πριν κάνουμε οποιαδήποτε παραγοντική ανάλυση. Εφαρμόζουμε συγκεκριμένα KMO (Kaiser-Meyer-Olkin) τεστ και Barllet's test of sphericity.

```
### geomorphology ###------------------------------

--------------------------------------------------
cor.set1<-cor(set[,2:15])
KMO(cor.set1)
cortest.bartlett(cor.set1, n=45)
--------------------------------------------------
```

Πίνακας 4 - Πίνακας αποτελεσμάτων ελέγχου KMO για την κατηγορία "geomorphology".

```
      Kaiser-Meyer-Olkin factor adequacy

      Call: KMO(r = cor.set1)
      Overall MSA =  0.58
      MSA for each item =
VAREA FAREA SAREA DAREA MVWIDTH MFWIDTH MSWIDTH MVCON MFCON CVWIDTH
0.36  0.42  0.24 0.39   0.68    0.81    0.70    0.65  0.32   0.58
CFWIDTH CSWIDTH   CVCON   CFCON
0.63    0.33      0.56    0.61
```

Πίνακας 5- Πίνακας αποτελεσμάτων ελέγχου Barllet's test of sphericity για την κατηγορία "geomorphology".

```
$chisq
[1] 429.8393

$p.value
[1] 1.985958e-45

$df
[1] 91
```

Ο έλεγχος Barlett *(Πίνακας 5)* δεν αποδέχεται την υπόθεση R=I σε επίπεδο στατιστικής σημαντικότητας 1% (p-value=$1.985958e^{-45}$ <0.01). Επομένως, τα διαγώνια στοιχεία του πίνακα συσχετίσεων R διαφέρουν από το 0 με πιθανότητα 99% και έτσι υπάρχουν σημαντικές συσχετίσεις μεταξύ των μεταβλητών.

Από τον *Πίνακα 4* αποφασίζουμε ποιες μεταβλητές θα κρατήσουμε στο μοντέλο και ποιες όχι. Κρατάμε τις μεταβλητές με MSA > 0.5. Αυτές οι μεταβλητές είναι: **MVWIDTH, MFWIDTH, MSWIDTH, MVCON, CVWIDTH, CFWIDTH, CVCON, CFCON**. Αυτές οι μεταβλητές είναι λίγες σχετικά για αυτό και το συνολικό MSA είναι ίσο με 58% (Overall MSA = 0.58). Πιο συγκεκριμένα η μεταβλητή με πολύ ικανοποιητικό MSA και η οποία έχει ισχυρή συσχέτιση με άλλες μεταβλητές είναι η MFWIDTH (MSA=0.81 >0.80), οι υπόλοιπες έχουν ένα ικανοποιητικό-μέτριο MSA.

Και οι δύο έλεγχοι λοιπόν μας επιτρέπουν να προχωρήσουμε με την παραγοντική ανάλυση. Ξεκινάμε με αριθμό factors=1. Κάθε φορά ξεκινάμε με τον μικρότερο δυνατό αριθμό factors και αν αυτό στη συνέχεια το κρίνουμε μη-ικανοποιητικό τότε αυξάνουμε τον αριθμό factors κάθε φορά κατά 1.

```
fit1 <-   factanal(
      set[,match(names(KMO(cor.set)$MSAi[KMO(cor.set)$MSAi>0.5]),
                       colnames(set))],
                       factors=1)
fit1
```

Στα αποτελέσματα της ανάλυσης του μοντέλου στον *Πίνακα 6* που ακολουθεί, βλέπουμε πως ο ένας παράγοντας που προσαρμόσαμε με το factor μοντέλο ερμηνεύει το **56.2%** της μεταβλητότητας των αρχικών μεταβλητών (`Proportion Var = 0.562`). Το ποσοστό αυτό δεν κρίνεται ιδιαίτερα ικανοποιητικό και θα χρειαστεί να προχωρήσουμε στην προσαρμογή δύο factors. Άλλη μία ένδειξη πως το μοντέλο μας δεν είναι ικανοποιητικό είναι και οι υψηλές τιμές της ιδαιτερότητας (Uniqueness) σε κάποιες μεταβλητές (`MVWIDTH = 0.883, MSWIDTH = 0.940 =` $1\text{-}0.244^2$`)`. Στον έλεγχο X^2 δεν δίνουμε βαρύτητα επί του παρόντος καθώς προϋποθέτει έλεγχο κανονικότητας.

*Πίνακας 6- Πίνακας αποτελεσμάτων ανάλυσης κατα **1** παράγοντα για την κατηγορία "geomorphology".*

```
Uniquenesses:
MVWIDTH MFWIDTH MSWIDTH   MVCON CVWIDTH CFWIDTH   CVCON   CFCON
  0.883   0.225   0.940    0.462   0.155   0.236   0.295   0.303

      Loadings:
              Factor1
      MVWIDTH  0.342
      MFWIDTH -0.880
      MSWIDTH -0.244
      MVCON    0.733
      CVWIDTH  0.919
      CFWIDTH  0.874
      CVCON    0.839
      CFCON    0.835

                     Factor1
      SS loadings      4.500
      Proportion Var   0.562

      Test of the hypothesis that 1 factor is sufficient.
      The chi square statistic is 128.43 on 20 degrees of freedom.
      The p-value is 7.66e-18
```

Αφού ως προς ένα παράγοντα το μοντέλο μας δεν ερμηνεύει ένα ικανοποιητικό ποσοστό διακύμανσης, προχωρούμε προσαρμόζοντας μοντέλο με **2** παράγοντες.

```
fit2 <- update(fit1, factors = 2)
fit2
```

Παρατηρούμε τώρα, λοιπόν, ότι βελτιώθηκε αρκετά το ποσοστό διακύμανσης που ερμηνεύεται από το μοντέλο μας και είναι ίσο με **70%** (*Πίνακας 7,* `Cumulative Var = 0.690`). Επίσης, συγκρίνοντας τον πίνακα συσχετίσεων (R) των αρχικών μεταβλητών με τον εκτιμώμενο πίνακα συσχετίσεων βάσει των κοινών παραγόντων (R^) (`fit$r, fit$model, fit$residual`) παρατηρούμε ότι οι δύο πίνακες δεν διαφέρουν πολύ. Τέλος, δοκιμάζοντας να αυξήσουμε κατά έναν ακόμη τους παράγοντες και να προσαρμόσουμε ακόμη έναν παράγοντα θα παρατηρήσουμε ότι το ποσοστό της μεταβλητότητας που θα ερμηνεύεται θα αυξηθεί μόλις κατά 3% (73%) . Επομένως, θα αρκεστούμε στους 2 παράγοντες.

*Πίνακας 7- Πίνακας αποτελεσμάτων ανάλυσης κατα **2** παράγοντες για την κατηγορία "geomorphology".*

```
      Uniquenesses:
MVWIDTH MFWIDTH MSWIDTH   MVCON CVWIDTH CFWIDTH   CVCON   CFCON
0.746   0.215   0.620   0.306   0.168   0.005   0.177   0.239

      Loadings:
              Factor1 Factor2
      MVWIDTH  0.153   0.480
      MFWIDTH -0.849  -0.253
      MSWIDTH         -0.615
      MVCON    0.832
      CVWIDTH  0.796   0.445
      CFWIDTH  0.646   0.760
      CVCON    0.906
      CFCON    0.773   0.404

                     Factor1 Factor2
      SS loadings      3.907   1.616
      Proportion Var   0.488   0.202
      Cumulative Var   0.488   0.690
```

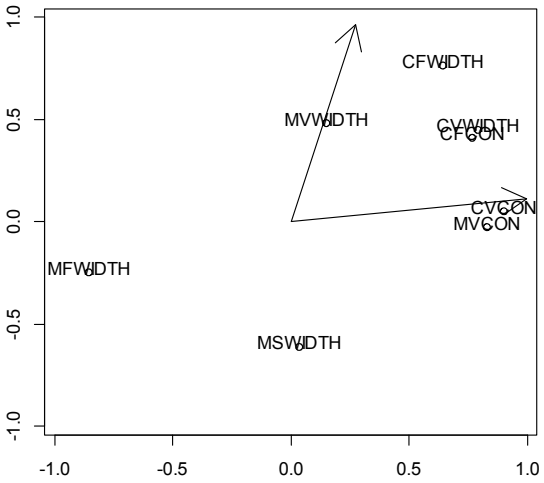
Αν θέλουμε μια πιο εύκολη – πιο απλή ερμηνεία του μοντέλου των 2 factors, τότε χρησιμοποιούμε την μέθοδο της **"περιστροφής"** των κοινών παραγόντων όπου στόχος είναι η εύρεση λύσης που δίνει απλή δομή. Αυτό μπορούμε να το κάνουμε εκτελώντας τις ακόλουθες εντολές στην R.

```
#rotation
update(fit2, rotation = 'promax')
```

Επίσης, στην ερμηνεία του ποιες μεταβλητές συμμετέχουν περισσότερο σε ποιον παράγοντα , μπορεί να μας βοηθήσει η κατασκευή ενός απλού γραφήματος όπως στη συνέχεια.

```
### plot
L <- loadings(fit2)
asvd <- svd(loadings(fit2)[1:8,])
adiag <- diag(1/asvd$d)
solution <- asvd$v %*% adiag %*% t(asvd$u) %*% loadings(update(fit2,
rotation = 'promax'))[,]
naxes <- solve(solution)
plot(L, xlim = c(-1.04,1.04), ylim = c(-1.04,1.04), xaxs = 'i', yaxs
= 'i', xlab = '', ylab = '')
text(L[,1]-0.005, L[,2]+0.035, labels = dimnames(L)[[1]])
arrows(rep(0, 2), rep(0, 2), naxes[,1], naxes[,2])
```

Διάγραμμα 7- Διάγραμμα των loadings των αρχικών μεταβλητών ως προς τους 2 παράγοντες. Τα δύο βέλη είναι οι παράγοντες και τα σημεία τα loadings των μεταβλητών ως προς τους δύο παράγοντες.



Θα λέγαμε, με βάση το *Διάγραμμα 7,* πως με πιο ισχυρή συσχέτιση συμμετέχουν στον ένα παράγοντα οι μεταβλητές CVCON, MVCON και MFWIDTH. Η τελευταία μάλιστα συσχετίζεται αρνητικά. Στον άλλο παράγοντα συμμετέχουν πιο πολύ οι μεταβλητές MVWIDTH και MSWIDTH(αρνητικά).

Για να αποθηκεύσουμε τα **"scores"** κάθε παράγοντα για κάθε παρατήρηση χρησιμοποιούμε τις παρακάτω εντολές.

```
scores<-factor.scores(set[,match(c("MVWIDTH","MFWIDTH",
                                   "MSWIDTH","MVCON",
                                   "CVWIDTH","CFWIDTH",
                                   "CVCON","CFCON"),
                            colnames(set))],
                            fit2)$scores
```

Ομοίως, προχωράμε και στην προσαρμογή factors μοντέλου και στις υπόλοιπες 2 κατηγορίες μεταβλητών. Οι εντολές συνοψίζονται στον παρακάτω κώδικα, ενώ οι πίνακες που περιγράφουν τα μοντέλα με τον αριθμό factors στους οποίους καταλήξαμε σε κάθε κατηγορία βρίσκονται στο *Παράρτημα, Πίνακες 4 & 5*.

```
### landscape composition ###----------------------

-------------------------------------------------
cor.set2<-cor(set[,16:35])
KMO(cor.set2)
cortest.bartlett(cor.set2, n=45)
-------------------------------------------------

fit1 <- factanal (
    set[,match(names(KMO(cor.set2)$MSAi[KMO(cor.set2)$MSAi>0.5]),
                  colnames(set))],
                  factors=1)
fit1
fit2 <- update(fit1, factors = 2)
fit2
fit3 <- update(fit2, factors = 3)
fit3

#scores
scores.lc←factor.scores(
       set[,match(names(KMO(cor.set2)$MSAi[KMO(cor.set2)$MSAi>0.5]),
                        colnames(set))],
                        fit3)$scores


#residuals
round(factor.residuals(cor(
       set[,match(names(KMO(cor.set2)$MSAi[KMO(cor.set2)$MSAi>0.5]),
                        colnames(set))]), fit3),2)

pr=princomp(factor.residuals(cor(set[,
            match(names(KMO(cor.set2)$MSAi[KMO(cor.set2)$MSAi>0.5]),
                        colnames(set))]), fit3),cor=T)

summary(pr)
```

```
### landscape heterogeneity ###----------------------

-------------------------------------------------
cor.set3<-cor(set[,36:49])
KMO(cor.set3)
cortest.bartlett(cor.set3, n=45)
-------------------------------------------------

fit1 <-
factanal(set[,match(names(KMO(cor.set3)$MSAi[KMO(cor.set3)$MSAi>0.5]
),
                  colnames(set))],
                  factors=1)
fit1
```

```
fit2 <- update(fit1, factors = 2)
fit2
fit3 <- update(fit2, factors = 3)
fit3

#scores
scores.lh←factor.scores(set[,
        match(c(names(KMO(cor.set3)$MSAi[KMO(cor.set3)$MSAi>0.5])),
                    colnames(set))],
                    fit3)$scores

#residuals
round(factor.residuals(cor(set[,
        match(c(names(KMO(cor.set3)$MSAi[KMO(cor.set3)$MSAi>0.5])),
                    colnames(set))]), fit3),2)

pr=princomp(factor.residuals(cor(set[,
        match(c(names(KMO(cor.set3)$MSAi[KMO(cor.set3)$MSAi>0.5])),
                    colnames(set))]), fit3),cor=T)

summary(pr)
```

Οπτικοποίηση δεδομένων

Η οπτικοποίηση των δεδομένων θα μας βοηθήσει να διακρίνουμε οπτικά το πόσο καλά διαχωρισμένες είναι οι ομάδες. Επίσης, θα μας βοηθήσει να καταλάβουμε καλύτερα τις ομοιότητες που υπάρχουν στα είδη που ανήκουν σε κοινό group αλλά και που διαφέρουν τα groups μεταξύ τους.

Αρχικά, θα κατασκευάσουμε τα κατάλληλα διαγράμματα ώστε να οπτικοποιήσουμε τα δεδομένα στις δύο διαστάσεις των 2 πρώτων κυρίων συνιστωσών κάθε κατηγορίας μεταβλητών.

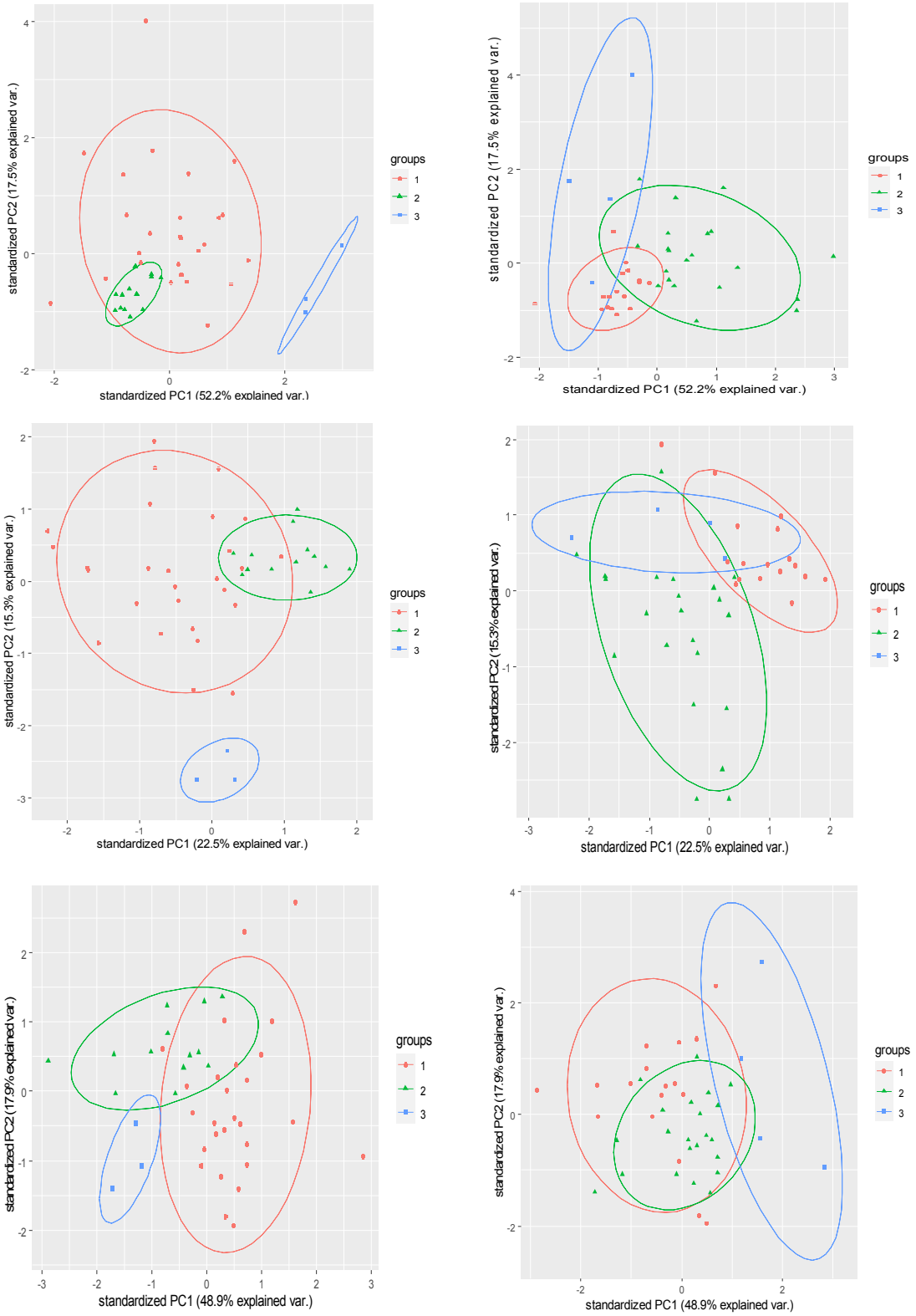
Για τα biplots θα χρησιμοποιήσουμε το πακέτο “ggbiplot” της R και θα εκτελέσουμε τις παρακάτω εντολές.

```
library(ggbiplot)

### geomorphology ###[1]

#hierarchical
d <- dist(scale(set[,-1]), method = "minkowski")
fit.d2 <- hclust(d,method="ward.D2")
g<-ggbiplot(set.pr1, choices = c(1,2), pc.biplot = TRUE,
            groups = as.factor(cutree(fit.d2,3)), ellipse = TRUE,
            ellipse.prob = 0.85, var.axes=FALSE, varname.size = 3,
alpha=0)
g<-g+geom_point(aes(colour=groups,shape=groups),size=1.3)
g<-g+scale_color_discrete(name = 'groups')
g

#k means
set.seed(995)
kmeans.3 <- kmeans(scale(set[,-1]), centers = 3, iter.max = 25,
trace = TRUE)
g<-ggbiplot(set.pr1, choices = c(1,2), pc.biplot = TRUE,
            groups = as.factor(kmeans.3$cluster), ellipse = TRUE,
            ellipse.prob = 0.85, var.axes=FALSE, varname.size = 4,
alpha=0)
g<-g+geom_point(aes(colour=groups,shape=groups),size=1.3)
g<-g+scale_color_discrete(name = 'groups')
g
```

Αυτό που παρατηρούμε στα διαγράμματα *(Διάγραμμα 8)* είναι πως πιο καλά διαχωρίζονται οι ομάδες όταν χρησιμοποιούμε την Ιεραρχική μέθοδο, όπως και αναμέναμε. Παρατηρούμε επίσης, πως η μία ομάδα (πράσινο χρώμα) παρεμβάλεται με την μεγαλύτερη ομάδα (κόκκινο χρώμα). Στην πρώτη κατηγορία αποτελεί μία υπο-ομάδα της μεγαλύτερης ομάδας χωρίς να διαχωρίζεται θα λέγαμε από αυτή (εντος του διαστήματος εμπιστοσύνης, έλειψη με κόκκινο χρώμα). Ξεκάθαρα διακρίνουμε δύο ομάδες, μία μεγάλη και μια μικρότερη ομάδα. Κυρίως, οι ομάδες αυτές διαφέρουν περισσότερο ως προς τα χαρακτηριστικά ετερογένειας τοπίου (landscape heterogeneity) και τα γεωμορφολογικά χαρακτηριστικά (geomorphology).

1 Εκτελούμε τον ίδιο κώδικα αλλάζοντας το σετ PCA, δηλαδή αντικαθιστώντας το set.pr1

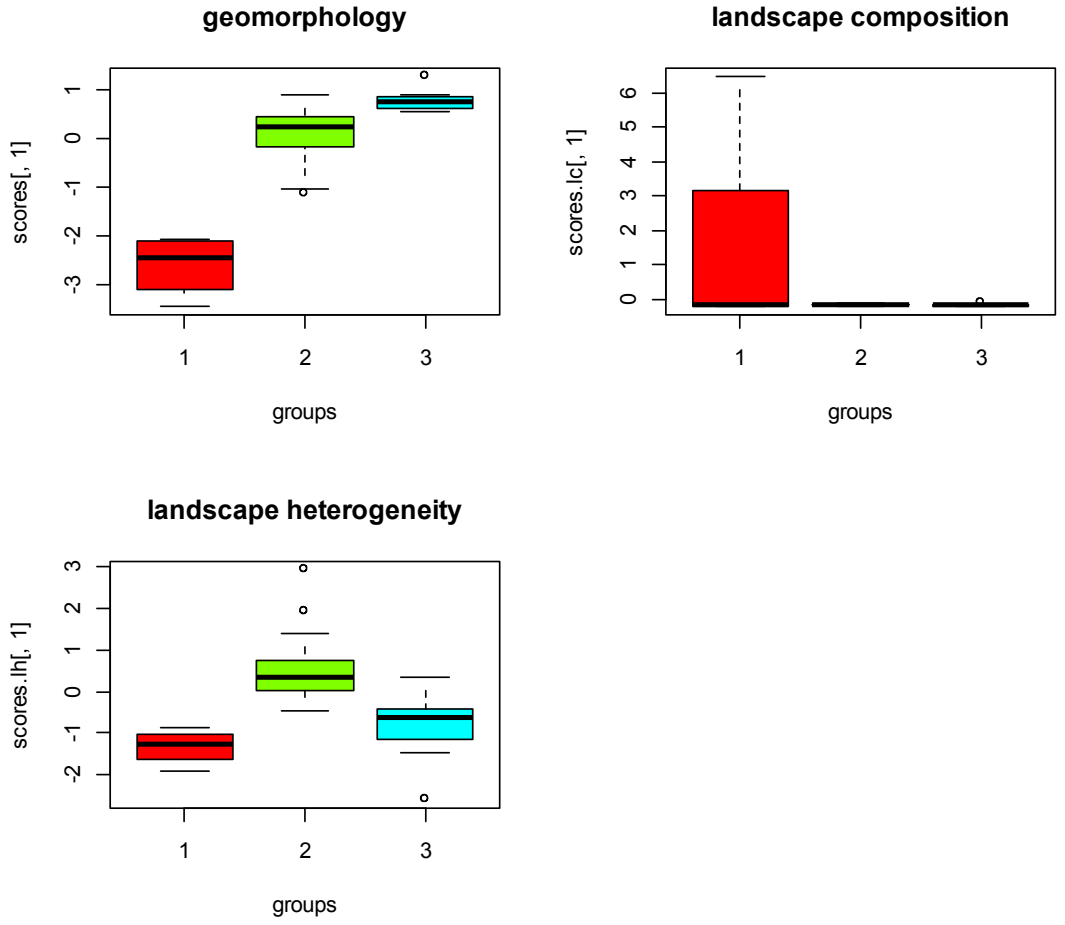
Διάγραμμα 8- Διάγραμματα biplots των ομαδοποιημένων μεταβλητών σύμφωνα με την Ιεραρχική μέθοδο (αριστερά), τη μέθοδο k-means(δεξιά) ως προς τις 2 πρώτες κύριες συνιστώσες κάθε κατηγορίας μεταβλητών. Η πρώτη γραμμή αφορά τη κατηγορία "geomorphology", η δεύτερη τη "landscape heterogeneity" και η τρίτη τη "landscape composition ".



Τέλος, θα κατασκευάσουμε και τα κατάλληλα boxplots των scores του πρώτου factor από αυτούς που επιλέξαμε για κάθε κατηγορία και group από αυτά που κατασκευάσαμε σύμφωνα με την Ιεραρχική μέθοδο.

```
groups <- cutree(fit.d2,3)
par(mfrow=c(2,2))
boxplot(scores[,1]~groups, col=rainbow(4), main ="geomorphology")
boxplot(scores.lc[,1]~groups, col=rainbow(4),
              main ="landscape composition")
boxplot(scores.lh[,1]~groups, col=rainbow(4),
              main ="landscape heterogeneity")
```

Διάγραμμα 9- Διάγραμματα boxplots των ομαδοποιημένων μεταβλητών σύμφωνα με την Ιεραρχική μέθοδο ως προς τα scores του πρώτου κοινού παράγοντα (factor) κάθε κατηγορίας.



Παρατηρούμε σύμφωνα με το *Διάγραμμα 9,* ότι οι ομάδες διαφέρουν αρκετά μεταξύ τους ως προς τα γεωμορφολογικά χαρακτηριστικά τα οποία ερμηνεύονται από τον πρώτο παράγοντα. Εδώ ο διαχωρισμός των ομάδων κρίνεται αρκετά καλός, η διάμεσος και των τριών ομάδων διαφέρει σημαντικά. Ειδικά η πρώτη ομάδα η οποία σκοράρει χαμηλά σε σχέση με τις υπόλοιπες δύο ομάδες. Αρκετά διαφέρουν οι ομάδες και ως προς τα χαρακτηριστικά της ετερογένειας του τοπίου (landscape heterogeneity) που ερμηνεύονται από τον πρώτο παράγοντα, με μία μόνο πολύ μικρή επικάλυψη μεταξύ 1ης και 3ης ομάδας.

Σε αντίθεση τώρα, έρχονται τα scores του πρώτου κοινού παράγοντα της σύνθεσης του τοπίου (landscape composition), ως προς τα οποία οι τρεις ομάδες δε φαίνεται να διαφέρουν καθόλου και έχουν διάμεσο score κοντά στο 0 με την πρώτη ομάδα να έχει ένα μεγαλύτερο ενδο-τεταρτημοριακό εύρος. Θα λέγαμε με λίγα λόγια ότι οι τρεις ομάδες στις οποίες χωρίσαμε τα είδη πουλιών πιθανότατα αρέσκονται σε αρκετά διαφορετικά χαρακτηριστικά που συνθέτουν το έδαφος και την ετερογένεια του τοπίου στην περιοχή που ζουν ή επισκέπτονται. Ως προς τα χαρακτηριστικά που συνθέτουν ένα τοπίο θα λέγαμε ότι όλα τα είδη αρέσκονται σε παρόμοια πράγματα .

Συμπεράσματα

Ανακεφαλαιώνοντας, αφού ερευνήσαμε διάφορες μεθόδους clustering καταλήξαμε στο να εφαρμόσουμε συσταδοποίηση μέσω της Ιεραρχικής μεθόδου και σε 3 αριθμούς clusters-ομάδων. Στη συνέχεια, οπτικοποιήσαμε τα δεδομένα μας σύμφωνα με την ομαδοποίηση που εφαρμόσαμε με τη βοήθεια των κυρίων συνιστωσών(Principal Components) και των κοινών παραγόντων (factors).

Από την ανάλυση που κάναμε συμπεράναμε πως οι 3 ομάδες διαφέρουν λιγότερο ή περισσότερο ως προς τα χαρακτηριστικά που αφορούν την γεωμορφολογία, τη σύνθεση του τοπίου και της ετερογένειας του τοπίου. Ειδικότερα, θα μπορούσαμε να πούμε πως οι 3 ομάδες πουλιών αρέσκονται σε διαφορετικά χαρακτηριστικά της μορφολογίας του εδάφους και της ετερογένειας του περιβάλλοντος στο οποίο συναντόνται. Λιγότερο πάλι, φαίνεται να διαφέρουν ως προς τα χαρακτηριστικά σύνθεσης ενός τοπίου στο οποίο αρέσκονται. Επίσης, οι δύο ομάδες από τις τρεις που διακρίναμε ίσως να ανήκουν σε μία κοινή ευρύτερη ομάδα.

ΠΑΡΑΡΤΗΜΑ

ΠΙΝΑΚΕΣ

Πίνακας 1 : PCA για τις μεταβλητές που ανήκουν στην κατηγορία "geomorphology".

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     2.2036 1.4685 1.3006 1.07591
Proportion of Variance 0.3469 0.1540 0.1208 0.08268
Cumulative Proportion  0.3469 0.5009 0.6217 0.70440
                           PC5     PC6    PC7     PC8
Standard deviation     1.01696 0.98026 0.9142 0.65819
Proportion of Variance 0.07387 0.06864 0.0597 0.03094
Cumulative Proportion  0.77827 0.84690 0.9066 0.93755
                           PC9    PC10   PC11   PC12
Standard deviation     0.48704 0.46542 0.4282 0.3627
Proportion of Variance 0.01694 0.01547 0.0131 0.0094
Cumulative Proportion  0.95450 0.96997 0.9831 0.9925
                          PC13    PC14
Standard deviation     0.29652 0.13274
Proportion of Variance 0.00628 0.00126
Cumulative Proportion  0.99874 1.00000
```

Πίνακας 2 : PCA για τις μεταβλητές που ανήκουν στην κατηγορία "landscape composition".

```
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     2.797048 1.9853548 1.5298792
Proportion of Variance 0.391174 0.1970817 0.1170265
Cumulative Proportion  0.391174 0.5882557 0.7052822
                           Comp.4    Comp.5
Standard deviation     1.38190387 1.2959961
Proportion of Variance 0.09548292 0.0839803
Cumulative Proportion  0.80076510 0.8847454
                           Comp.6     Comp.7
Standard deviation     0.99697489 0.85126698
Proportion of Variance 0.04969795 0.03623277
Cumulative Proportion  0.93444335 0.97067612
                           Comp.8     Comp.9
Standard deviation     0.58152190 0.41386545
Proportion of Variance 0.01690839 0.00856423
Cumulative Proportion  0.98758451 0.99614874
                          Comp.10      Comp.11
Standard deviation     0.24598414 0.1178829522
Proportion of Variance 0.00302541 0.0006948195
Cumulative Proportion  0.99917415 0.9998689704
                            Comp.12      Comp.13
Standard deviation     4.184944e-02 0.0231989663
Proportion of Variance 8.756878e-05 0.0000269096
Cumulative Proportion  9.999565e-01 0.9999834488
                            Comp.14      Comp.15
Standard deviation     0.0167333188 5.624043e-03
Proportion of Variance 0.0000140002 1.581493e-06
Cumulative Proportion  0.9999974490 9.999990e-01
                            Comp.16      Comp.17

Standard deviation     3.522848e-03 1.929838e-03
Proportion of Variance 6.205230e-07 1.862138e-07
Cumulative Proportion  9.999997e-01 9.999998e-01
                            Comp.18      Comp.19
Standard deviation     1.383151e-03 9.887492e-04
Proportion of Variance 9.565528e-08 4.888125e-08
Cumulative Proportion  9.999999e-01 1.000000e+00
                            Comp.20
Standard deviation     6.040239e-04
Proportion of Variance 1.824224e-08
Cumulative Proportion  1.000000e+00
```

Πίνακας 3 : PCA για τις μεταβλητές που ανήκουν στην κατηγορία "landscape heterogeneity"

```
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     2.3884483 1.4530017 1.2582387
Proportion of Variance 0.4074775 0.1508010 0.1130832
Cumulative Proportion  0.4074775 0.5582785 0.6713617
                          Comp.4     Comp.5
Standard deviation     1.1838430 1.02764715
Proportion of Variance 0.1001060 0.07543276
Cumulative Proportion  0.7714677 0.84690048
                           Comp.6     Comp.7
Standard deviation     0.83856914 0.78539236
Proportion of Variance 0.05022844 0.04406008
Cumulative Proportion  0.89712893 0.94118901
                           Comp.8    Comp.9
Standard deviation     0.63307892 0.4220881
Proportion of Variance 0.02862778 0.0127256
Cumulative Proportion  0.96981679 0.9825424
                           Comp.10     Comp.11
Standard deviation     0.330576580 0.251021220
Proportion of Variance 0.007805777 0.004500832
Cumulative Proportion  0.990348163 0.994848995
                           Comp.12    Comp.13
Standard deviation     0.195245886 0.17595776
Proportion of Variance 0.002722925 0.00221151
Cumulative Proportion  0.997571921 0.99978343
                            Comp.14
Standard deviation     0.0550633992
Proportion of Variance 0.0002165699
Cumulative Proportion  1.0000000000
```

*Πίνακας 4 : Πίνακας αποτελεσμάτων ανάλυσης κατα **3** παράγοντες για την κατηγορία "landscape composition".*

```
Uniquenesses:
  CCP   HCP   HCS   HLS   HSH   HSS   MCP   MLS   MOP   MSH
0.005 0.178 0.005 0.783 0.005 0.090 0.096 0.085 0.005 0.005

Loadings:
    Factor1 Factor2 Factor3
CCP  0.976   0.209
HCP  0.535   0.727
HCS  0.976   0.214
HLS  0.381   0.264
HSH          0.113   0.989
HSS  0.596   0.741
MCP  0.101   0.945
MLS  0.854   0.430
MOP  0.338   0.832   0.436
MSH  0.977   0.207

                Factor1 Factor2 Factor3
SS loadings       4.503   3.062   1.187
Proportion Var    0.450   0.306   0.119
Cumulative Var    0.450   0.756   0.875
```

87.5 % ποσοστό της διακύμανσης ερμηνεύεται από το μοντέλο των 3 παραγόντων σε σχέση με το **72.2%** που ερμηνεύεται από το μοντέλο των 2 παραγόντων.

*Πίνακας 5 : Πίνακας αποτελεσμάτων ανάλυσης κατα **3** παράγοντες για την κατηγορία "landscape heterogeneiiy".*

```
      Uniquenesses:
NPATCH  CPSIZE   TEDGE   LFRAC   PRICH SHANNON  SHEVEN  SIEVEN
0.767   0.195   0.092   0.053   0.083   0.044   0.018   0.005
AMPSHP CONEDGE
0.508   0.423

      Loadings:
              Factor1 Factor2 Factor3
      NPATCH           -0.376   0.302
      CPSIZE           -0.867   0.210
      TEDGE    0.940           -0.157
      LFRAC    0.883   0.315  -0.263
      PRICH   -0.188  -0.538   0.769
      SHANNON  0.901   0.365  -0.105
      SHEVEN   0.443   0.848  -0.258
      SIEVEN   0.533   0.814  -0.223
      AMPSHP   0.272           -0.645
      CONEDGE          -0.281   0.701

                     Factor1 Factor2 Factor3
      SS loadings      3.079   2.879   1.855
      Proportion Var   0.308   0.288   0.185
      Cumulative Var   0.308   0.596   0.781
```

78.1 % ποσοστό της διακύμανσης ερμηνεύεται από το μοντέλο των 3 παραγόντων σε σχέση με το **66.4%** που ερμηνεύεται από το μοντέλο των 2 παραγόντων.