



Ανάλυση δεδομένων οικολογίας σε περιβάλλον R

Τμήμα Βιολογίας

Γιατί R;

- Μέσω της γλώσσας R είναι δυνατή η δημιουργία εξειδικευμένων «πακέτων» που μπορούν να πραγματοποιούν στατιστικές αναλύσεις (και όχι μόνο...) που είναι απαραίτητες για την διερεύνηση των διάφορων ερευνητικών ερωτημάτων και υποθέσεων.
- Τα «πακέτα» καθώς και το εκπαιδευτικό υλικό σχετικό με τη σωστή χρήση τους είναι ελεύθερα διαθέσιμα από τις ιστοσελίδες <http://cran.r-project.org/> και <http://www.r-project.org/>. Επιπλέον στο διαδίκτυο υπάρχουν δεκάδες διαφορετικές πηγές (ιστοσελίδες και video tutorials) σχετικά με τις δυνατότητες και τα πακέτα που έχουν δημιουργηθεί σε περιβάλλον R.
- Κατά συνέπεια η γλώσσα προγραμματισμού R παρουσιάζει κάποια σημαντικά πλεονεκτήματα για την εκτέλεση στατιστικών αναλύσεων σε σχέση με άλλα στατιστικά προγράμματα:
 - Είναι open-source. Όλοι έχουν πρόσβαση στον κώδικά της και μπορούν να κάνουν διορθώσεις και βελτιώσεις.
 - Οποιοσδήποτε μπορεί να αναπτύξει κάποιο πακέτο επεξεργασίας/ανάλυσης δεδομένων
 - Υπάρχουν πάνω από 20000 πακέτα (εργαλεία επεξεργασίας δεδομένων)
 - Είναι δωρεάν!!!

Γιατί R;

- **Ανοιχτού Κώδικα και Δωρεάν:** Δεν έχει κόστος άδειας και μπορεί να χρησιμοποιηθεί από οποιονδήποτε.
- **Προσαρμοστικότητα:** Εκατοντάδες πακέτα επιτρέπουν την επέκταση της λειτουργικότητάς της.
- **Διαχείριση Δεδομένων:** Δυνατότητες για ανάγνωση, καθαρισμό και επεξεργασία δεδομένων.
- **Ανάλυση Δεδομένων:** Υποστηρίζει στατιστική, μηχανική μάθηση, και προχωρημένες τεχνικές μοντελοποίησης.
- **Οπτικοποίηση:** Παρέχει εξαιρετικά εργαλεία για δημιουργία γραφημάτων υψηλής ποιότητας (π.χ. ggplot2).
- **Αναπαραγωγικότητα:** Εύκολη τεκμηρίωση κώδικα για αναπαραγωγή ανάλυσης.

Εφαρμογές της R σε Διάφορα Πεδία

1. Επιστήμη Δεδομένων και Ανάλυση Επιχειρήσεων

- Πρόβλεψη πωλήσεων μέσω χρονολογικών σειρών.
- Ανάλυση συμπεριφοράς πελατών.
- Εξόρυξη δεδομένων (Data Mining).

2. Οικολογία και Περιβαλλοντικές Επιστήμες

- Ανάλυση πληθυσμών ειδών.
- Πρόβλεψη αλλαγών στο κλίμα.

3. Υγεία και Βιοστατιστική

- Ανάλυση κλινικών δεδομένων.
- Μελέτες επιδημιολογίας.
- Γονιδιωματική ανάλυση

4. Χρηματοοικονομικά

- Ανάλυση χρηματιστηριακών δεδομένων.
- Υπολογισμός ρίσκου επενδύσεων.
- Μοντελοποίηση χρονολογικών σειρών (π.χ. ARIMA).

5. Εκπαίδευση και Έρευνα

- Ανάλυση ερωτηματολογίων.
- Εκπαιδευτικά εργαλεία για στατιστική.
- Τεκμηρίωση και δημοσίευση επιστημονικών εργασιών (R Markdown)

6. Γεωχωρική Ανάλυση

- Επεξεργασία γεωχωρικών δεδομένων (GIS).
- Χαρτογράφηση πληθυσμιακών τάσεων.
- Ανάλυση δορυφορικών δεδομένων.

Εγκατάσταση της R και R studio

<https://cran.r-project.org/bin/windows/base/R-4.4.3-win.exe>

<https://download1.rstudio.org/electron/windows/RStudio-2024.12.1-563.exe>

*(Για MacOS και Linux πηγαίνετε στις διευθύνσεις <http://cran.r-project.org/> και <https://posit.co/download/rstudio-desktop/> και επιλέξετε τον αντίστοιχο installer)

Στατιστική ανάλυση

Έλεγχος κανονικότητας: Ανεξάρτητα από το τι θέλουμε να κάνουμε πρέπει να γνωρίζουμε αν τα δεδομένα μας ακολουθούν κανονική κατανομή. Η διάγνωση της κανονικής κατανομής γίνεται οπτικά, π.χ παρατηρώντας το ιστόγραμμα συχνοτήτων ή/και το Q-Q διάγραμμα ή για μικρό δείγμα (π.χ <100) γίνεται με το Shapiro-Wilk τεστ.

```
hist(H,probability=T)
```

```
lines(density(H),col=2)
```

```
qqnorm(H)
```

```
qqline(H)
```

```
shapiro.test(H)
```

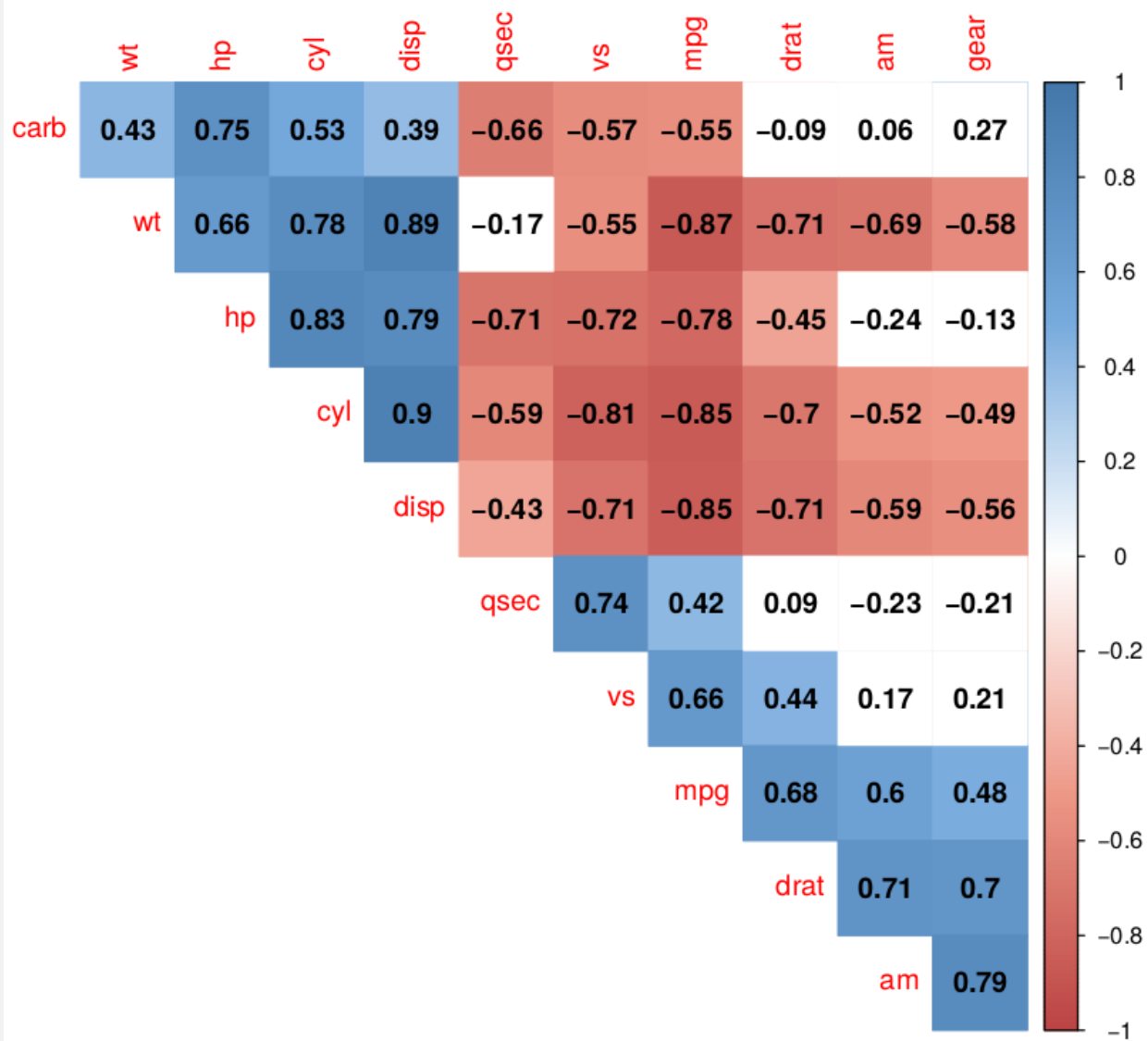
Στατιστική ανάλυση

Συσχέτιση μεταξύ δύο παραμέτρων

- Με τους συντελεστές συσχέτισης εκτιμούμε κατά πόσο μία μεταβλητή αυξάνεται ή μειώνεται σε σχέση με κάποια άλλη. Δηλαδή αν υπάρχει θετική συσχέτιση μεταξύ δύο παραμέτρων αν αυξάνει η μία τότε αυξάνεται και η άλλη. Αν υπάρχει αρνητική συσχέτιση τότε η αύξηση της μίας παραμέτρου ακολουθεί μείωση της άλλης.
- **Προσοχή:** Συσχέτιση δε σημαίνει αιτιότητα.
- Οι πιο διαδεδομένοι συντελεστές συσχέτισης είναι οι Pearson (για δεδομένα με κανονική κατανομή) και Spearman (για δεδομένα χωρίς κανονική κατανομή)
- Στην R ο υπολογισμός των συντελεστών συσχέτισης γίνεται με τα παρακάτω functions

```
cor(x, y, method = c("pearson", "kendall", "spearman"))  
cor.test(x, y, method=c("pearson", "kendall", "spearman"))
```

ECG p-value significance



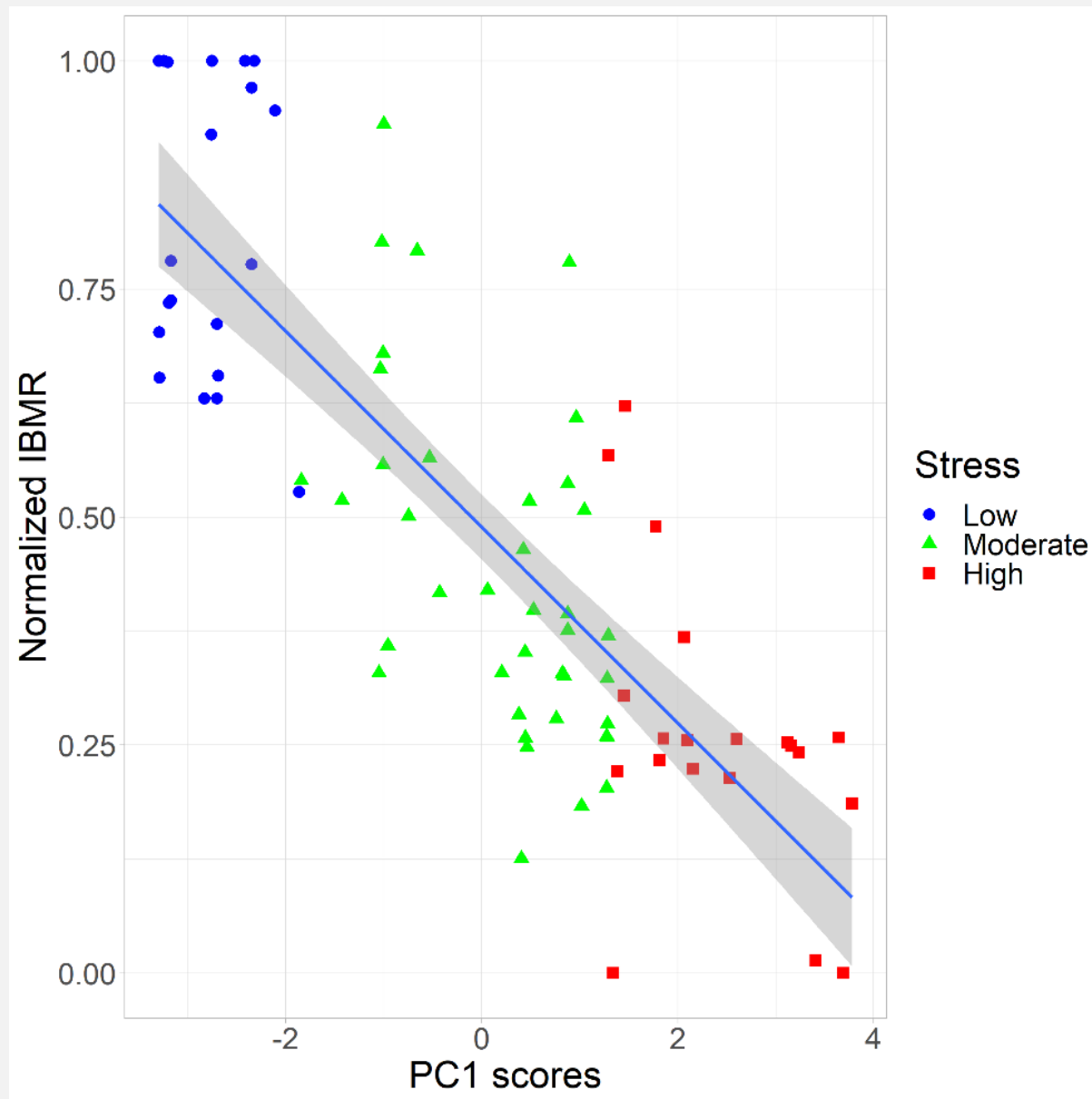
Γραμμική παλινδρόμηση

- Με τη γραμμική παλινδρόμηση περιγράφουμε μαθηματικά τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων επεξηγηματικών μεταβλητών. Γραμμική παλινδρόμηση πρακτικά σημαίνει πως η σχέση μεταξύ των δύο μεταβλητών είναι γραμμική. Δηλαδή αν y είναι η εξαρτημένη μεταβλητή (π.χ η αφθονία των ειδών) και x η επεξηγηματική μεταβλητή (π.χ η υγρασία στο έδαφος ή η θερμοκρασία του αέρα ή η έκταση της δειγματοληπτικής επιφάνειας) τότε η γραμμική σχέση μεταξύ τους θα περιγράφεται από τη σχέση:
- **$y = ax + b$**
- Στην R χρησιμοποιούμε την εντολή `lm` για τη δημιουργία γραμμικών μοντέλων

`model <- lm(formula = y ~ x)`

`summary(model)`

Με την εντολή `summary` μπορούμε να δούμε τις τιμές των συντελεστών a και b



Ανάλυση διακύμανσης

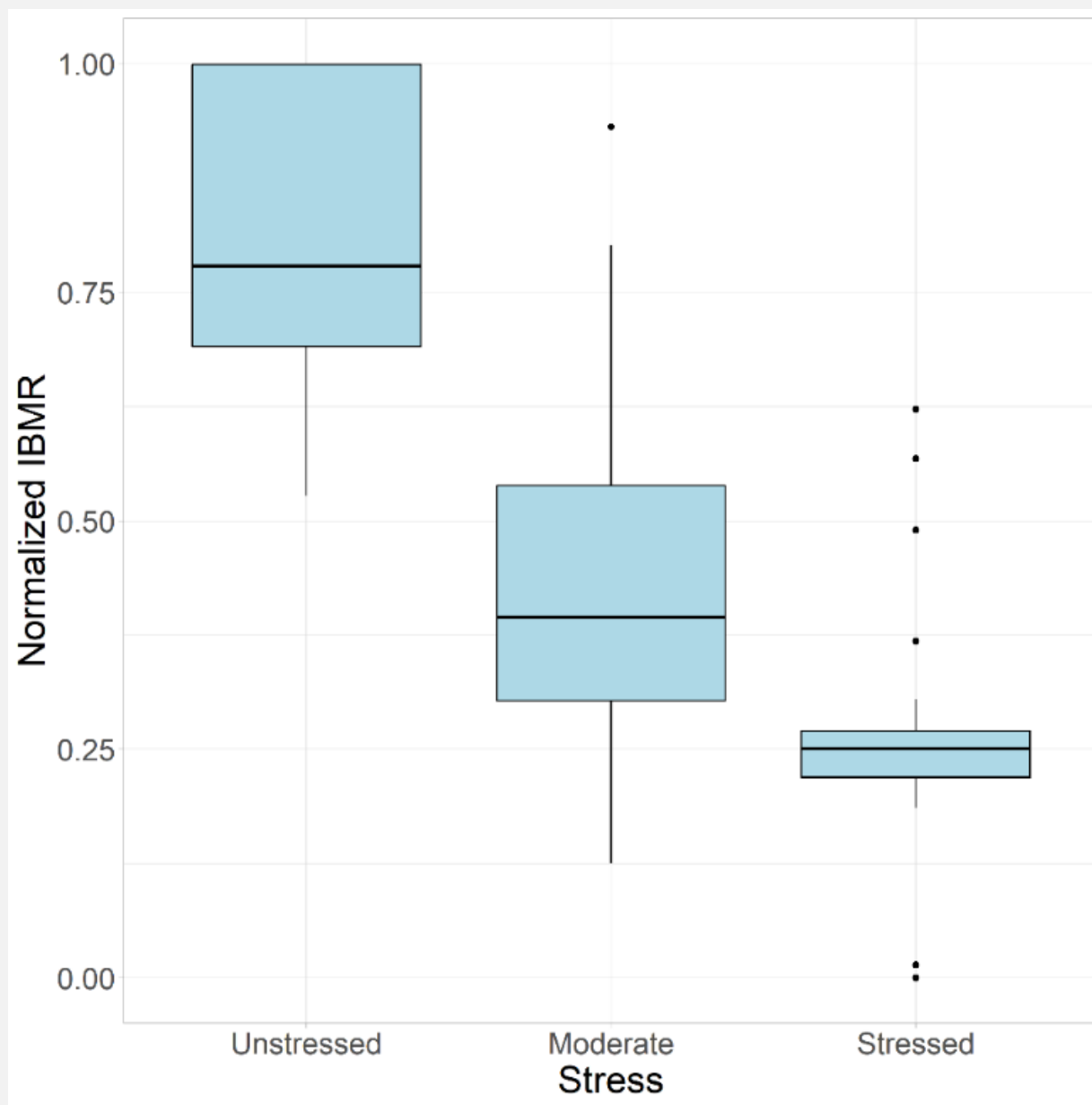
- Η ανάλυση διακύμανσης (Analysis of Variance, ANOVA) αποτελεί έναν τρόπο σύγκρισης των μέσων τιμών μεταξύ δύο δειγμάτων. Η ανάλυση αυτή είναι χρήσιμη για παράδειγμα αν θέλουμε να διερευνήσουμε πειραματικά την επίδραση μιας ένωσης σε μια βιολογική διεργασία ή να συγκρίνουμε τις μέσες τιμές μιας παραμέτρου που έχουν προέλθει από δύο διαφορετικά «υποσύνολα» (Π.χ μέσες τιμές αφθονίας ειδών μεταξύ δύο διαφορετικών οικοσυστημάτων).
- **Σημείωση:** Προϋπόθεση για την ανάλυση διακύμανσης είναι τα δεδομένα μας να ακολουθούν κανονική κατανομή (βλ. έλεγχο κανονικότητας). Αλλιώς χρησιμοποιείται η ανάλυση διακύμανσης *Kruskal-Wallis*

Η ανάλυση διακύμανσης στην R γίνεται με την εντολή:

`aov(y~x, data=data)`

και για την ανάλυση διακύμανσης *Kruskal-Wallis*

`kruskal.test(y~x, data=data)`



Μη-παραμετρική πολυμεταβλητή ανάλυση διασποράς

Η ανάλυση διακύμανσης μας δίνει τη δυνατότητα να διερευνήσουμε αν κάποια μετρική α-ποικιλότητας διαφέρει μεταξύ ομάδων

Για να δούμε αν η β-ποικιλότητα διαφέρει μεταξύ διαφορετικών ομάδων τότε θα πρέπει να χρησιμοποιήσουμε διαφορετικά εργαλεία.

Η PERMANOVA ή μη-παραμετρική πολυμεταβλητή ανάλυση διασποράς (permutational multivariate analysis of variance) → συγκρίνουμε ομάδες τιμών ελέγχοντας τη μηδενική υπόθεση ότι τα κεντροειδή και οι διασπορές των τιμών μέσα στις ομάδες είναι ίσες.

Χρειαζόμαστε έναν πίνακα με “αποστάσεις” όπως π.χ. είναι ένας πίνακας ανομοιοτήτων που προκύπτει με την εφαρμογή της λειτουργίας *vegdist*.

```
adonis2(bray~grazing, data=environment, permutations = 999,  
method="bray")
```

Αναλύσεις κατάταξης

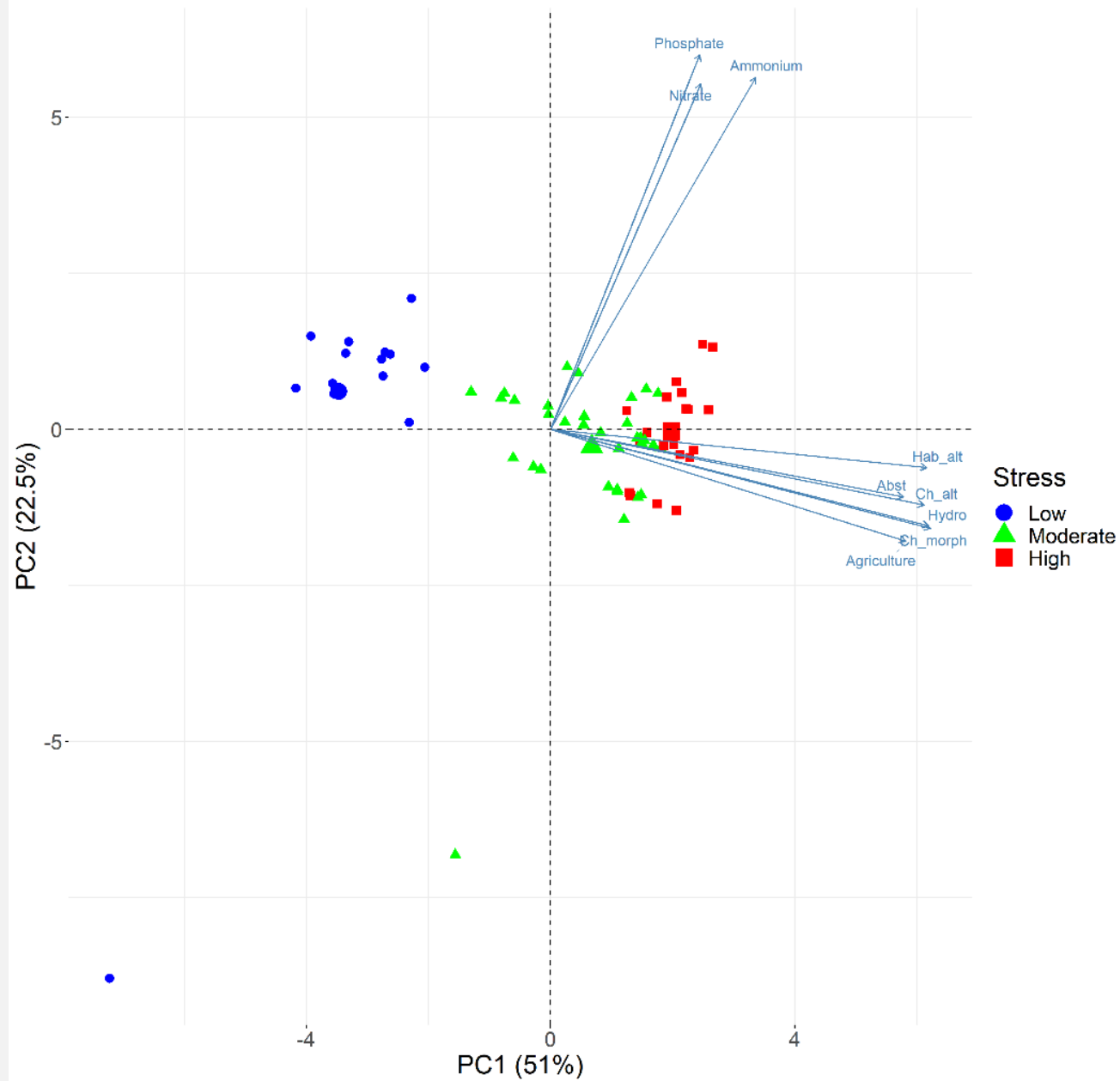
Γενικά, ως αναλύσεις κατάταξης μπορούμε να χαρακτηρίσουμε το σύνολο των τεχνικών που μας βοηθούν να απεικονίσουμε πολυδιάστατα δεδομένα σε 2 ή 3 διαστάσεις.

Έχουν δημιουργηθεί πολλές διαφορετικές μέθοδοι (PCA, CCA, NMDS, RDA, CA κ.α) με κύριο στόχο την «απλοποίηση» των πολυπαραγοντικών δεδομένων και την εύρεση σημαντικών σχέσεων μεταξύ περιβαλλοντικών δεδομένων και δεδομένων κοινότητας (community)

Ανάλυση Κύριων Συνιστωσών (PCA - Principal Component Analysis)

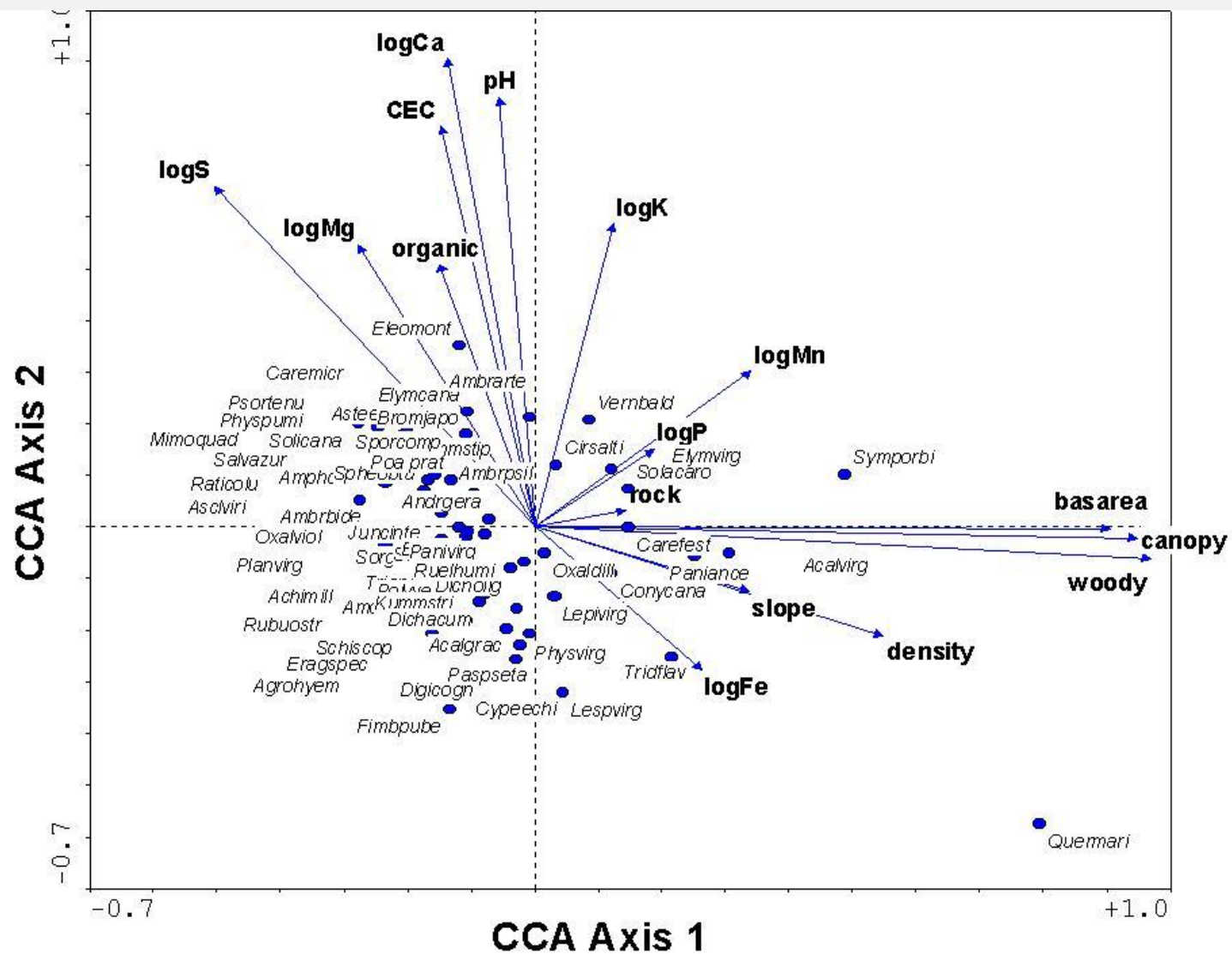
- Σκοπός:** Η PCA είναι μια μέθοδος για τη μείωση της διάστασης των δεδομένων διατηρώντας όσο το δυνατόν περισσότερη πληροφορία. Μετατρέπει συσχετισμένες μεταβλητές σε ένα μικρότερο σύνολο ασυσχέτιστων μεταβλητών που ονομάζονται κύριες συνιστώσες (principal components).
- Χρήση:** Χρησιμοποιείται για την εξερεύνηση προτύπων σε συνεχή δεδομένα, συχνά σε περιβαλλοντικές μελέτες ή για την κατανόηση της ποικιλομορφίας ειδών.
- Ερμηνεία:** Οι πρώτες συνιστώσες εξηγούν τη μεγαλύτερη διακύμανση στα δεδομένα. Τα βάρη (loadings) δείχνουν τη συμβολή κάθε αρχικής μεταβλητής στη συνιστώσα.

MVAOS of main stress variables



Κανονική Ανάλυση Αντιστοίχισης (CCA - Canonical Correspondence Analysis)

- **Σκοπός:** Η CCA είναι μια μέθοδος άμεσης ανάλυσης βαθμίδωσης που συνδέει τη σύνθεση ειδών με περιβαλλοντικές μεταβλητές
- **Χρήση:** Χρησιμοποιείται κυρίως στην οικολογία για τη μελέτη των συσχετίσεων μεταξύ αφθονίας ειδών και περιβαλλοντικών παραγόντων
- **Ερμηνεία:** Στο διάγραμμα της CCA, τα δείγματα και τα είδη τοποθετούνται με βάση τη σχέση τους με τους περιβαλλοντικούς άξονες, ενώ τα βέλη δείχνουν την κατεύθυνση και την ένταση των περιβαλλοντικών βαθμίδων.



Ανάλυση Πλεονασμού (RDA - Redundancy Analysis)

- Σκοπός:** Η RDA είναι μια μέθοδος κατευθυνόμενης ανάλυσης που μοιάζει με την PCA, αλλά περιορίζεται από επεξηγηματικές μεταβλητές. Αξιολογεί τη διακύμανση των δεδομένων ειδών που εξηγείται από ένα σύνολο παραγόντων.

- Χρήση:** Χρησιμοποιείται στην οικολογία για τη διερεύνηση σχέσεων μεταξύ σύνθεσης ειδών και περιβαλλοντικών ή πειραματικών παραγόντων.

- Ερμηνεία:** Τα βέλη των εξηγητικών παραγόντων δείχνουν την επίδρασή τους στους άξονες της ανάλυσης. Τα δείγματα και τα είδη απεικονίζονται για να δείξουν τη σχέση τους υπό την επίδραση των εξηγητικών μεταβλητών.

Μη Μετρική Πολυδιάστατη Κλιμάκωση (NMDS - Non-metric Multidimensional Scaling)

Η NMDS είναι μια μέθοδος πολυδιάστατης απεικόνισης που χρησιμοποιείται για τη μείωση της διάστασης και την οπτικοποίηση δεδομένων. Εστιάζει στη διατήρηση των αποστάσεων μεταξύ σημείων, βασιζόμενη σε μετρικές ομοιότητες (π.χ. Bray-Curtis, Jaccard).

Χαρακτηριστικά:

- **Μη Γραμμική Προσέγγιση:** Σε αντίθεση με τις γραμμικές μεθόδους (π.χ. PCA), η NMDS δεν υποθέτει γραμμικές σχέσεις στα δεδομένα.
- **Διάταξη Βασισμένη σε Αποστάσεις:** Χρησιμοποιεί έναν πίνακα αποστάσεων (dissimilarity matrix) για να τοποθετήσει σημεία σε έναν χώρο χαμηλότερης διάστασης (συνήθως 2 ή 3 διαστάσεων).

Χρήση στην Οικολογία:

Η NMDS χρησιμοποιείται ευρέως για τη μελέτη βιοκοινοτήτων και τη σύγκριση σύνθεσης ειδών μεταξύ διαφορετικών περιοχών ή συνθηκών. Παρέχει οπτική απεικόνιση των ομοιοτήτων και διαφορών στις κοινότητες.

Iris species

