Analiza și modelarea statistică multivariată a comunităţilor; Analiza de ordonare

Foarte frecvent în ecologie analizăm comunităţi (sisteme suprapopulaționale), în habitate caracterizate prin diferiţi parametri calitativi și cantitativi, în dorinţa de a analiza modul în care acestea răspund la factorii de mediu. În acest gen de studii dorim să răspundem la diverse întrebări sau probleme, cum ar fi:

- modul în care fiecare populaţie reacţionează (îşi modifică parametri ecologici cantitativi) la variaţia graduală a unui factor de mediu (sau mai mulți);
- modificarea unei populaţii analizată pe diferiţi gradienţi ai mediului;
- răspunsul mai multor populaţii (specii) la variaţia graduală a unui număr oarecare de caracteristici ale habitatului;
- care factori ai mediului explică mai bine și în ce mod se reflectă aceștia în compoziţia comunităţii (cum influenţează gradienții mediului comunitatea)?;
- cum se clasifică variabilele de mediu în funcţie de gradul de explicare a compoziţiei speciilor (structura comunităţii)?
- cât din variaţia compoziţiei specifice (a variabilelor dependente sau de răspuns) este explicată de către variabilele de mediu evaluate?

... și multe altele ...

La aceste întrebări putem prea rar răspunde dacă le descompunem în părţi sau în elemente constituente; frecvent nici nu putem să adoptăm un model reducţionist, în care să analizăm separat fiecare variabilă sau să căutăm modele de regresie între variaţia fiecărei specii în relaţie cu modificarea fiecărei variabile de mediu.

Dacă studiem modul în care o variabilă ecologică (abundenţa unei populaţii de exemplu) este relaţionată de o variabilă de mediu (altitudinea, salinitatea etc.) o analiză de regresie, liniară sau neliniară, ar putea fi suficientă, iar relaţia s-ar putea reprezenta grafic sub forma unui plan bidimensional (2D), în care ordonata este prima, iar abscisa definește cea de-a doua variabilă (care descrie variaţia mediului). Când dorim să analizăm relaţii multidimensionale (nD) între diverse specii și mai multe variabile de mediu (independente sau de habitat), precum și să reprezentăm grafic aceste relaţii, vom fi constrânși să înţelegem și să aplicăm metode sau modele statistice multivariate, dintre care analizele de ordonare fac parte integrantă.

Analiza de ordonare este o metodă de studiu a modalităţilor prin care comunităţile (ansambluri de populaţii simpatrice și sincronice care interacţionează între ele și cu factorii de mediu) răspund la variaţia graduală a condiţiilor de habitat. Deoarece factorii de mediu se modifică gradual, diversele specii vor răspunde prin variaţia treptată, continuă, a abundenţei sau a altui parametru ecologic. Prin urmare o

analiză de ordonare se poate asemăna cu un studiu de gradient, respectiv se pretează la o analiză a schimbărior factorilor de mediu care le condiționează pe cele ale parametrilor populaţiilor. O altă idee este aceea că factorii de mediu sunt frecvent mai mult sau mai puţin corelaţi, sau sunt grupaţi în funcţie de gradul de interacţiune, speciile modificându-și parametri cantitativi în relaţie mai puternică cu unii dintre aceșta și mai puţin cu alţii, uneori demonstrând chiar o relativă indiferenţă la variaţia unora dintre ei.

Comunitatea se modifică neuniform și treptat, deoarece populaţiile din care este alcătuită sunt adaptate în mod diferit la factorii de mediu și își modifică relativ independent parametri cantitativi în funcţie de aceștia. Dar, atunci cînd analizăm aceste schimbări, constatăm adesea că răspunsul grupelor de specii, respectiv a comunităţii întregi, pot fi prevăzute în manifestarea lor, iar gradienţii de-a lungul cărora se înregistrează variaţiile pot fi sintetizaţi sub forma unui număr redus de axe sau de gradienţi imaginari, care redau sintetic corelaţia dintre diverșii factori ai mediului.

Câteodată gradienţii sau axele de-a lungul cărora se modifică abundenţa și compoziţia speciilor pot fi recunoscuţi într-un peisaj, de exemplu sub forma umidităţii (trecerea de la un habitat acvatic la unul uscat, terestru) sau a altitudinii. Alteori aceste axe nu pot fi identificate ca atare și atunci vorbim de gradienţi de modificare (schimbare) a compoziţiei speciilor.

Studiul variaţiei comunităţilor se poate face în multe feluri (există metode diverse), dar atunci când analizăm schimbarea treptată și continuă (graduală sau pe unul sau mai mulţi gradienţi) metodele de ordonare sunt printre cele mai eficiente și adesea singurele capabile să răspundă la întrebările pe care le formulăm. Expresia grafică a acestor metode poartă numele de **diagrame de ordonare** și acestea implică proiectarea speciilor, a probelor (eșantionelor) și/sau a variabilelor (factorilor) de mediu într-un **spaţiu de ordonare**, adică pe un plan construit dintr-un număr redus de axe. Regulile de reprezentare și interpretare vor fi studiate în cele ce urmează, dar esenţa înţelegerii diagramelor este schimbarea continuă a compoziţiei comunităţilor de-a lungul unor gradienţi, care sunt reprezentaţi sub forma unor axe, precum și ideea că proximitatea implică similaritate (ceea ce este mai aproape, este mai asemănător sau între entităţi sunt relaţii mai strânse). Reprezentarea și interpretarea graficelor (a diagramelor de ordonare) este numai o mică parte a problemei: cel puţin la fel de importantă este modelarea statistică, în sensul testării diverselor ipoteze (de exemplu semnificaţia relaţiei speciilor cu diverși factori ai mediului, clasificarea acestora în funcţie de importanţa lor predictivă etc.), al conceptualizării și al identificării relaţiilor multiple între diversele variabile, îndepărtarea variaţiei aleatoare și identificarea

modelelor repetitive, legice, de tip corelativ și de tip cauză-efect, între diversele categorii de fenomene.

De la bun început este bine să înţelegem că metodele descrise aici includ o componentă grafică, descriptivă sau explorativă, dar nu se limitează nici pe departe cu aceasta. Tehnicile expuse includ și permit dezvoltarea unor unelte complexe de testare a ipotezelor multiple de lucru, analize de regresie, de varianţă, relaţii sau dependenţă de timp, măsurători multiple repetate asupra acelorași obiecte, analiza datelor structurate spaţial, analiză ierarhică de varianţă (ANOVA și MANOVA) etc.. Aceste tehnici permit cercetătorior să conceapă studii și să ridice întrebări complexe și frecvent mai adecvate problemelor abordate în studiile moderne de ecologie. Prin urmare: deși sunt metode excelente pentru **analiza exploratorie a datelor,** acestea sunt extrem de utile (uneori de neînlocuit) și în **analiza confirmatorie, adică în studiile ipotetic-deductive, respectiv cele bazate pe design experimental complex.**

Amănuntele și formulele matematice vor fi eludate în cea mai mare parte, pentru a simplifica cât mai mult aspectele mecanismelor și pentru a ne concentra mai ales asupra efectelor analizelor datelor ecologice, utilizând pachetul de programe cunoscut sub denumirea de *Canoco.* Acest produs software este concentrat pe reducerea dimensiunii datelor (ordonare), analiza de regresie și combinarea celor două, tehnică cunoscută sub denumirea de **ordonare constrânsă, condiţionată** sau **ordonare canonică.** Obiectivul general este să ajute cercetătorii să înţeleagă și să identifice modele și structuri în datele multivariate, prin relaţii între variabile și articole. Numele softului *Canoco* provine de la acronimul ***CANOnical Community Ordination*** și provine din domeniul ecologiei comunităţilor. Din acest motiv a adoptat o serie de termeni și denumiri din această disciplină, cum ar fi **probă** pentru un articol sau o linie, **specie** pentru o variabilă de răspuns sau variabilă dependentă, precum și **variabilă de mediu** pentru o variabilă explicativă, independentă sau suplimentară.

Scopul ordonării este cel de a identifica **axe ale variabilităţii maxime în compoziţia comunităţii** (numite **axe de ordonare**), pentru un set de probe, precum și de a vizualiza (utilizând **diagrame de ordonare**) structurile de similaritate pentru probe și specii. Este de așteptat ca axele de ordonare să coincidă cu unele variabile de mediu, iar atunci cînd aceste variabile sunt măsurate, le vom corela cu axele de ordonare. Scopul **ordonării constrânse sau canonice** este de a găsi variabilitatea în compoziţia speciilor care poate fi explicată prin măsurarea variabilelor de mediu. Ordonarea poate fi și **parţial constrânsă (sau parţial canonică):** în analizele parţiale mai întâi se îndepărtează variabilitatea în compoziţia speciilor explicată de **covariate sau covariabile** (variabile de mediu definite ca atare), după care se realizează o ordonare canonică a **variabilităţii reziduale** sau remanente. Există și analiza de

ordonare hibridă, în care primele *x* axe (*x* este uzual 1, 2 sau 3) de ordonare sunt canonice (constrânse) iar celelalte sunt neconstrânse sau necanonice.

Tipuri de variabile incluse în analiza multivariată a comunităţilor

Orice set de date primare (observaţionale sau experimentale) este format din:
- **variabile ale comunităţii, pe care le mai numim (poate cam impropriu) "de specii", dependente, sau de răspuns** (aşezate pe coloane de obicei);
- **valorile acestora în probe (eșantioane, experienţe) numite și articole** (pe linii de obicei)**;** (în termeni de prezenţă-absenţă, densitate, efectiv, abundenţa relativă, indicele Braun-Blanquet etc.). Valorile asociate cu speciile pot fi cantitative dar pot fi și date de prezenţă-absenţă (calitative sau de tip 1/0). Corespunzător modului în care definim și evaluăm speciile se va determina și scara variabilelor dependente sau de răspuns: cantitative, semi-cantitative (de exemplu mijlocul claselor de abundenţă-dominanţă relativă prin scara Braun-Blanquet) sau de tip "boolean" sau binare (adică 1/0, da/nu, prezenţă/absenţă).

Frecvent dispunem de valori (articole) și pentru:
- variabile de mediu, pe care le mai numim și externe, independente, canonice, explicative, de condiţionare, de constrângere etc.
Variabilele independente (în mod asemănător cu cele dependente) **pot fi:**
- **cantitative** (numere reale sau numere naturale etc.), de exemplu altitudinea;
- **semicantitative** (ex. impactul antropic pe o scară de la 0 - absent la 10 - foarte puternic); în analize acestea sunt de obicei tratate similar cu variabilele numerice, cantitative;
- **nominale** (de tip **categorie**, numite și **factori**) - pot fi de tip prezenţă/absenţă (1/0 sau **variabile binare**), sau pot fi de tip **variabile artificiale, substitutive** (*dummy variables*) sau de tip **indicator**. Legat de ultima categorie - dacă vrem să codificăm tipul de folosinţă al unui teren - dacă este fâneaţă, pajiște, alte folosinţe sau este abandonat, în *Canoco* 4.x nu putem defini o variabilă cu 4 categorii (adică cele enunţate) ci trebuie să definim 4 variabile, fiecare cu valoare de 0 pentru fals (sau negaţie) și 1 pentru adevăr (sau afirmaţie). Astfel, folosinţa terenului se va codifica prin variabile, cum ar fi **fâneaţă (da/nu, 1/0), pajiște (1/0), altele (1/0) și părăsit (1/0)**. În ***Canoco* 5.x** această problemă s-a rezolvat prin posibilitatea de definire a variabilelor factoriale sau a **factorilor.** Astfel, în cazul de mai sus o variabilă de tip *folosinţa terenului* va fi codificată (de exemplu sub numele sau **eticheta** *FOLOS*) și va cuprinde valori alfanumerice (deci litere sau combinaţii), cum ar fi *F* pentru articolele care conţin date din probele sau suprafeţele de eșantionare situate în fâneţe, *P* pentru pajiști, *A* pentru alte folosinţe și *P* pentru teren părăsit.

Variabilele explicative pot fi heterogene: unele sunt cantitative, altele pot fi ordinale sau binare, în același fișier.

În analize (și implicit în *Canoco*) fișierele sunt de tip rectangular, fiind constituite **din tabele de specii (ale comunităţilor)** și **tabele de variabile explicative** sau de **mediu**. De obicei o **coloană este o variabilă** (specie sau parametru de mediu), iar o linie este **un articol, sau o probă unitară,** adică valorile tuturor variabilelor care provin dintr-un anumit eșantion individual. Relaţia de corespondență dintre cele două tabele este dată de articole sau de linii, motiv pentru care la fiecare tabel, prima coloană se referă la numărul curent al probelor, iar ordinea (poziţia acestora) trebuie să fie aceeași. Adică: un anumit rând trebuie să conţină atât în tabelul de specii cât și în cel de mediu valorile variabilelor din aceeași probă.

Și acum, ceva **teorie despre metodele aplicate în ecologia comunităţilor** și locul, respectiv rolul, tehnicilor multivariate de analiză a gradientului, construirea și interpretarea diagramelor de ordonare și unele metode relaţionate.

Aceste metode trebuie puse în context; nu sunt apărute din senin și studenţii de la ecologie sau biologie au avut de-a face atât cu metode de studiu ale comunităţilor cât și cu tehnici și metode de analiză multivariată. În timpul studiilor de licenţă, la laboratoarele de Ecologie generală, Ecologia populaţiilor, Ecosisteme: structură și funcţii, precum și la Modelarea proceselor și sistemelor ecologice (se presupune că) s-au învăţat:

- Definirea, abordarea și natura sistemelor ecologice pe diferite niveluri de organizare (abordarea teoretică; identificarea sistemelor ecologice);
- Parametrii ecologici cantitativi ai populațiilor și ai comunităţilor;
- Grafice: reprezentarea (cantitativă a) comunităţilor și a dinamicii acestora;
- Analiza de regresie univariată și multivariată, liniară și neliniară.

(dacă aţi uitat despre ce vorbim, recapitulaţi temele din volumul de Ecologie Practică, Sîrbu și Benedek, ediţia a 3-a, 2012)

- Analiza de corelaţie;
- Testarea ipotezelor;
- Tehnici și metode speciale aplicate în studiul comunităţilor de păsări și de mamifere;
- Nișa ecologică: lăţimea, preferinţele și suprapunerea nișelor (ultima fiind o metodă excelentă pentru investigarea modului în care sunt alcătuite și cum funcţionează comunităţile);
- Biodiversitatea - Diversitatea ecologică;
- Analiza de asociere;
- Compararea sistemelor ecologice - Indici de similitudine;

- **Metode de ordonare şi clasificare a sistemelor ecologice** - construirea dendrogramelor.
- Metode de studiu ale productivităţii și analiza reţelelor trofice...
și multe altele...

Prin urmare tot ce conţine acest capitol este pus într-un context clar; tema de faţă completează ceea ce (se presupune că) s-a învăţat (cel puţin) la laboratoarele de ecologie și de modelare de până acum.

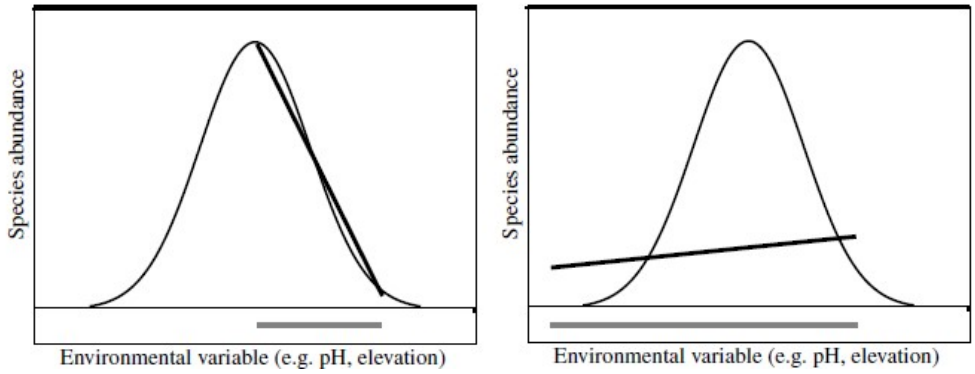
Comunitatea, definită ca un ansamblu de populaţii simpatrice și sincronice care interacţionează pe mai multe niveluri, poate fi abordată teoretic și metodologic prin prisma a două puncte de vedere; ambele sunt valabile în anumite condiţii și se completează reciproc:

- **Abordarea discontinuă** (a comunităţilor unitare, a cartării comunităţilor, abordarea inegralistă *sensu* Clements) - comunităţile sunt identificabile, individualizate, delimitate, cu graniţe bine definite între ele etc. Principalele metode de studiu au fost amintite anterior (cele de clasificare sunt în mod special folosite la această categorie) (*Discontinuum Approach*, în engleză) fiind totodată concepţia clasică în fitocenotaxonomie (taxonomia asociaţiilor de plante).
- **Abordarea continuă** (*Continuum Approach* în engleză), sau abordarea individualistă *sensu* Gleason**, analiza de gradient** - conform căreia nu există graniţe distincte între comunităţi, se presupune că sunt zone largi de tranziţie între diferitele sisteme ecologice adiacente etc. Metodele de ordonare prezentate aici sunt bazate pe ideea analizei de gradient a comunităţilor.

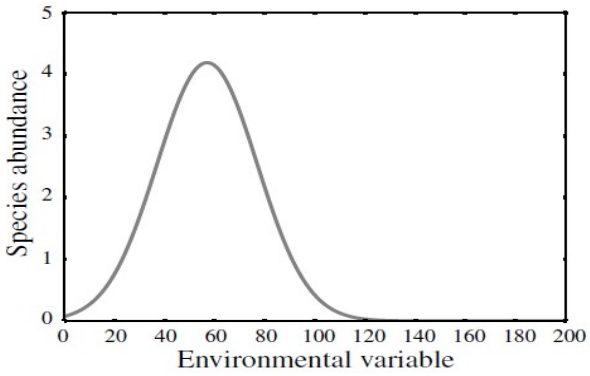
O întrebare foarte importantă în analiza bazată pe abordarea continuă este:

Cum răspunde o specie (cum variază un parametru al acesteia) de-a lungul gradientului (variaţia treptată, continuă a) unui factor (variabilă, parametru) al mediului?

- dacă segmentul de variaţie al gradientului este scurt, presupunem că abundenţa speciei variază liniar. Atunci aplicăm **modele liniare** (PCA, RDA);
- dacă segmentul este lung, presupunem că abundenţa speciei variază după o curbă unimodală (adică are o singură valoare maximă, un singur maxim sau mod), caz în care aplicăm **modele unimodale** (CA, CCA)



Analiza de gradient pe baza modelului liniar, se pretează la segmente scurte ale domeniului de variaţie a variabilei de mediu (stânga sus), nu însă şi dacă segmentul este lung (dreapta sus).



Răspunsul (şi modelul) unimodal reprezentat mai sus, este mult mai corect atunci când domeniul de variaţie al factorului (a variabilei de mediu) delimitează sau include amplitudinea valenţelor speciei (gradient lung).

Răspunsul liniar al speciei pe un gradient scurt al variabilei de mediu este evaluat prin modelul liniar de regresie, respectiv prin metoda pătratelor minime. Pentru modelul unimodal de răspuns, cel mai simplu mod de a estima optimul speciei este de calcul a mediei ponderate (*Mp(Sp)*) a valorilor variabilelor de mediu în cele n probe în care este prezentă specia. Valorile de abundenţă sau importanţă a speciei sunt utilizate ca ponderi în calcularea mediei:

$$MP(Sp) = \frac{\Sigma_{i=1}^{n} X_i * Abundenta_i}{\Sigma_{i=1}^{n} Abundenta_i}$$

unde *n* este numărul de probe, *i* este incrementul care ia valori succesive de la 1 la n, X_i este valoarea variabilei de mediu *X* în proba *i*, iar $Abundenta_i$ este valoarea abundenţei speciei în proba i.

Când este necesar, toleranţa speciei (lăţimea curbei de tip clopot) poate fi calculată ca rădăcina pătrată a mediei ponderate a diferenţelor pătrate dintre optimumul speciei și valoarea actuală dintr-o probă (*SD(Sp)*). Valoarea este analogă cu abaterea (deviaţia) standard și servește ca bază pentru definirea unităţilor de SD pentru măsurarea lungimii axelor de ordonare.

$$SD(Sp) - \sqrt{\frac{\Sigma_{i=1}^{n}(X_i - MP(Sp))^2 * Abundenta_i}{\Sigma_{i=1}^{n} Abundenta_i}}$$

În funcţie de prezenţa sau absenţa variabilelor de mediu, precum și de câte variabile de specii sunt incluse în analize, metodele alese pot fi diferite.

Tipuri de analize statistice

Variabila(e) de răspuns...(specii)	Variabila(e) de prognoză (de mediu)	
	Absentă	Prezente
... nu este niciuna	Atunci nu este nici ecologie și nici biologie -nu își are locul în acest capitol, volum și nici în disciplina studiată! Metoda: aţi greșit facultatea și/sau specializarea!	
... este una	- rezumat de distribuţie (al proprietăţilor de distribuţie, adică grafice, frecvenţe, histograme, statistică descriptivă etc.)	- modele de regresie *sensu lato*
... sunt mai multe	- analiza indirectă de gradient (PCA, CA, DCA, NMDS) - analiză de clasificare ierarhică (dendrograme)	- analiză directă de gradient (RDA, CCA) - analiza discriminantă (CVA)

Abrevierile provin din engleză:
PCA = *"principal component analysis"* (analiza componentelor principale)
DCA = *"detrended correspondence analysis"* (analiza de corespondenţă segmentată)
NMDS = *"non-metric multidimensional scaling"* (scalare multidimensională non-metrică)
RDA = "*redundancy analysis*" (analiza de redundanţă)
CCA = "*canonical correspondence analysis*" (analiză de corespondenţă canonică)
CVA = "*canonical variate analysis*" (analiza variabilei canonice sau analiza funcţiei discriminante)

Metodele de analiză a gradientului sau de ordonare sunt de diverse categorii, cele mai importante fiind redate mai jos.

Nr. var. mediu	**Nr. specii**	**Cunoştinţe a priori asupra relaţiei dintre specii şi mediu?**	**Metoda**	**Rezultate**
1, n	1	NU	Regresie	Dependenţa speciei de (relaţiile cu) variabilele de mediu
Nici una	n	DA	Calibrare	Estimarea valorilor mediului
Nici una	n	NU	Ordonare necondiţionată (independentă, ne-canonică)	Axele de variabilitate în compoziţia speciilor; analiza indirectă de gradient
1, n	n	NU	Ordonare condiţionată (constrânsă, dependentă, canonică)	Variabilitatea din compoziţia speciilor explicată de variabilele de mediu. Relaţiile variabilelor de mediu cu axele speciilor. Analiza directă de gradient.

Tehnicile fundamentale de ordonare (analize şi diagrame de ordonare) se clasifică şi se aleg în funcţie de modul în care răspund speciile la variaţia gradientului (scurt sau lung, respectiv dacă sunt modele liniare sau unimodale) precum şi dacă se consideră sau nu relaţia variabilelor dependente cu cele de mediu (canonice). Tabelul de mai jos reprezintă o sinteză a acestora.

	Metode liniare	**Metode unimodale (medii ponderate)**
Necondiţionate, non-canonice, neconstrânse (nu se ţine seama sau nu avem variabile de mediu) = analiza indirectă de gradient	Analiza componentelor principale (**PCA** = *Principal components analysis*)	- Analiza de corespondenţă (**CA** = *Correspondence analysis*) - **DCA** = *Detrended correspondence analysis*)
Condiţionate, constrânse sau canonice (există şi sunt considerate variabilele de mediu, independente) = analiza directă de gradient	Analiza de redundanţă (**RDA** = *Redundancy analysis*)	Analiza de corespondenţă canonică (**CCA** = *Canonical correspondence analysis*)

Ne oprim aici cu teoria, pentru a vedea cum se realizează o analiză de ordonare.

Realizarea unei analize de ordonare, alcătuirea și interpretarea unei diagrame de ordonare în Canoco 4.5

Un exemplu simplu (redus ca număr de articole sau probe), prea puţin realist, dar sugestiv și un bun punct de plecare pentru a învăţa organizarea datelor și analiza de ordonare a comunităților în *Canoco*.

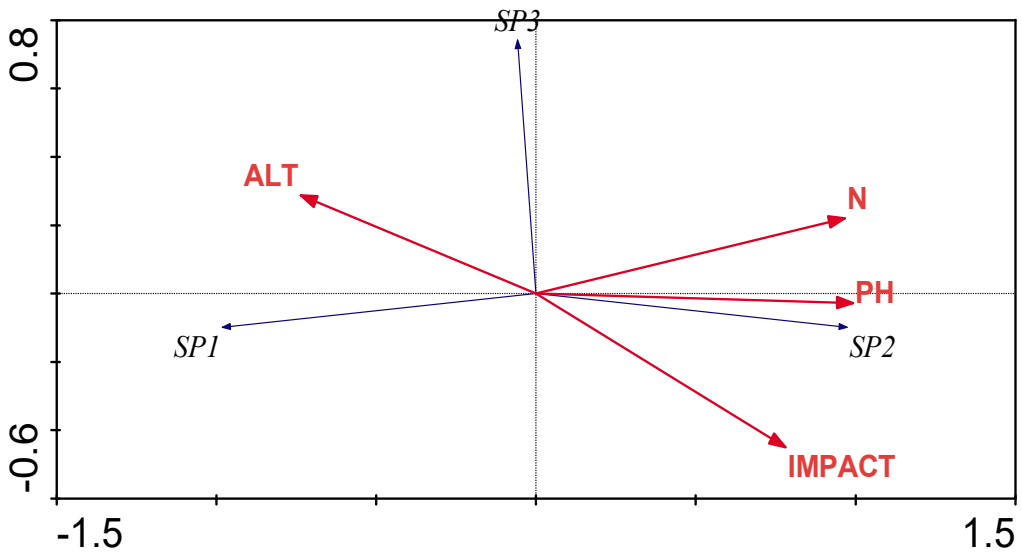
Fie o comunitate de plante descrisă prin 9 probe (SAMPLE), 3 specii de cormofite (SP1, SP2, SP3) - acestea alcătuind variabilele care se referă la comunitate (variabile dependente sau ”de specii”), iar separat scriem din nou probele (SAMPLE) și variabilele independente sau ”de mediu”: IMPACT - evaluarea mai mult sau mai puţin obiectivă a impactului antropic, pe o scară de la 1 (cel mai redus impact) la 6 (impactul cel mai puternic), PH - valoarea pH a solului din probe, N - conţinutul de azot din sol și ALT - altitudinea la care se situează staţia de eșantionare. Speciile sunt evaluate în termeni de abundenţă-dominanță relativă (sunt alese din asociaţii numai speciile cu abundenţă ridicată). Datele sunt introduse într-un program de foi de calcul (Excel sau altul):

E3

	A	B	C	D	E	F	G	H	I	J
1	SAMPLE	SP1	SP2	SP3		SAMPLE	IMPACT	PH	N	ALT
2	1	35	2	51		1	1	2.1	12	563
3	2	23	4	32		2	3	2	13	612
4	3	21	3	24		3	2	2.5	15	620
5	4	34	5	56		4	3	2.1	11	510
6	5	12	6	21		5	4	3.2	23	462
7	6	2	45	16		6	5	6.7	56	374
8	7	0	56	58		7	4	7.9	89	451
9	8	4	45	35		8	6	5.3	45	304
10	9	11	4	27		9	3	2.8	22	450

Ce am dori să obţinem în cele din urmă? De exemplu (ca mai jos) o diagramă care rezultă în urma unei analize de redundanţă (RDA - *Redundancy Analysis*), în care toate datele sunt proiectate într-un spaţiu 2D, definit de două axe care extrag cea mai mare parte din variabilitatea sau varianţa datelor, săgeţile albastre indicând sensul în care crește abundenţa speciilor, cele roșii indică sensul de creștere al valorilor variabilelor de mediu, iar direcţiile și unghiurile dintre săgeţi, precum și dintre variabile și axe, indică gradul de corelaţie între diferitele entităţi.

Menţionăm că această analiză și diagramă nu este obligatorie și nici singura posibilă!

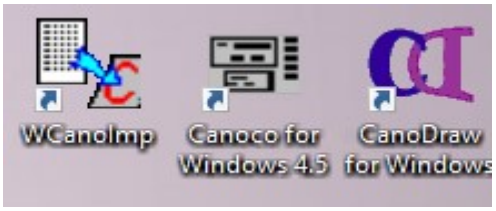


Diagramă de ordonare (RDA) realizată pe baza datelor din fişierul Excel prezentat anterior.

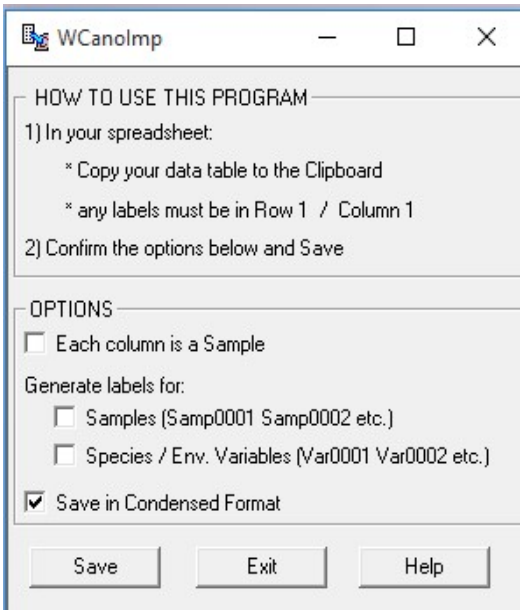
Rezolvare şi (totodată) ghid succint de lucru în Canoco 4.5

- Instalăm *Canoco* şi aşezăm (astfel încât să le avem la îndemână) icoanele pentru: introducerea datelor (*WCanoImp*), programul care realizează prelucrările şi analizele (*Canoco for Windows*) şi cel care realizează graficele (*CanoDraw for Windows*). Menţinem deschis fişierul cu datele primare ale comunităţii şi variabilele de mediu din programul respectiv (Excel sau altul).

Menţionăm că această etapă este valabilă pentru versiunea 4.5 şi cele inferioare. Începând cu *Canoco 5.x* toate aplicaţiile sunt adunate (grupate) într-un singur program de bază, motiv pentru care se instalează numai acesta.



- Dublu clic pe WCanoImp → apare fereastra:



Faceţi ce scrie acolo:

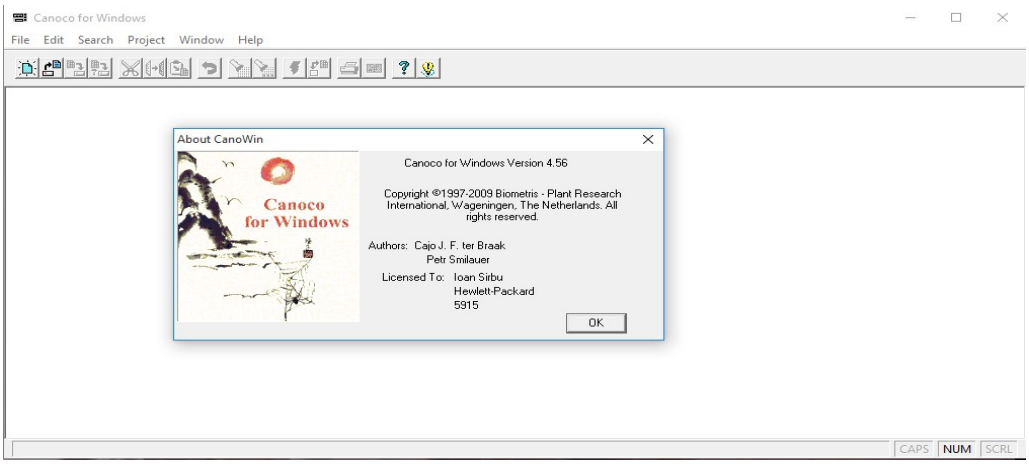
- Datele să fie astfel concepute încât pe prima linie să fie etichetele variabilelor, iar pe prima coloană să fie numărul curent al probelor sau al staţiilor de eşantionare.
- Dacă probele sunt pe coloane și speciile pe linii, selectaţi *Each column is a Sample*
- Copiaţi (Ctrl+C) sau puneţi în Clipboard toate datele de specii (trasaţi un dreptunghi în jurul secţiunii fișierului care se referă la specii). Ignoraţi variabilele de mediu deocamdată.
- Clic pe Save.

	A	B	C	D	E	F	G	H	I	J
1	SAMPLE	SP1	SP2	SP3		SAMPLE	IMPACT	PH	N	ALT
2	1	35	2	51		1	1	2.1	12	563
3	2	23	4	32		2	3	2	13	612
4	3	21	3	24		3	2	2.5	15	620
5	4	34	5	56		4	3	2.1	11	510
6	5	12	6	21		5	4	3.2	23	462
7	6	2	45	16		6	5	6.7	56	374
8	7	0	56	58		7	4	7.9	89	451
9	8	4	45	35		8	6	5.3	45	304
10	9	11	4	27		9	3	2.8	22	450
11										

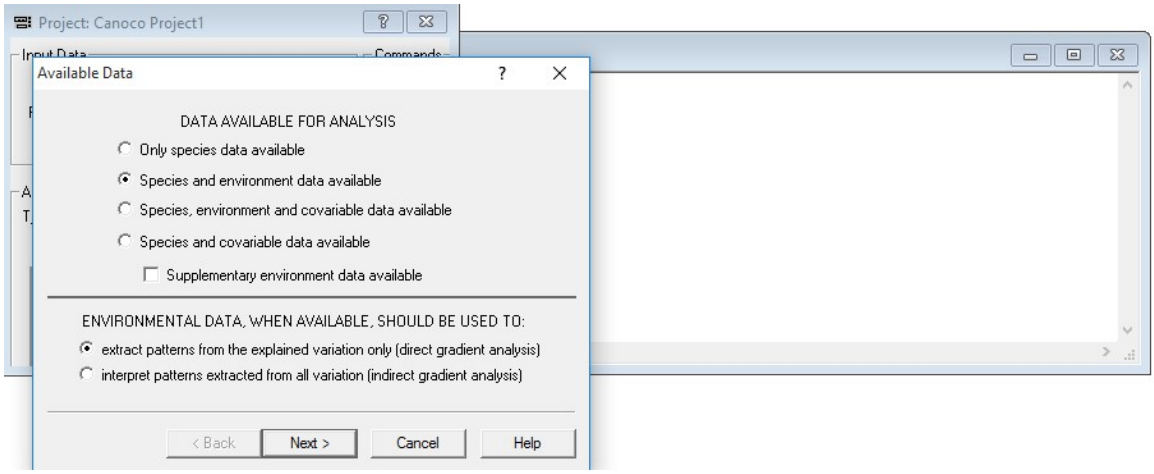
- Am marcat și am copiat (Ctrl+C)
- În *WCanoImp* alegem locaţia (directorul, fișierul) și dăm un nume sugestiv pentru acest fișier, cu extensia .dta. De exemplu: specii.dta. Scrierea extensiei nu este obligatorie!
- Clic pe Save
- Programul ne deschide o fereastră pentru a introduce, dacă dorim, un anumit titlu pe prima linie a acestui fișier (Title Line Text)

- Clic pe OK
- Programul ne anunţă că a creat fișierul dorit.
- Apare fosta fereastră de introducere și salvare a fișierelor.
- Noi selectăm acum din fișierul de date primare variabilele și datele de mediu (SAMPLE, IMPACT, PH, B, ALT). Trasaţi un dreptunghi în jurul acestor date și variabile și copiaţi (Ctrl+C).
- Ne întoarcem la WCanoImp în fereastra veche și clic din nou pe Save.
- Se deschide fereastra de salvare, scriem un nou nume cu extensia adecvată (de exemplu mediu.dta) și dăm clic pe Save. Suntem anunţaţi că fișierul a fost salvat.
- Clic pe OK și acum putem închide și fișierul de date și pe WCanoImp (Exit).

Deschidem *Canoco for Windows* (dublu clic pe icoană). Apare fereastra**:**

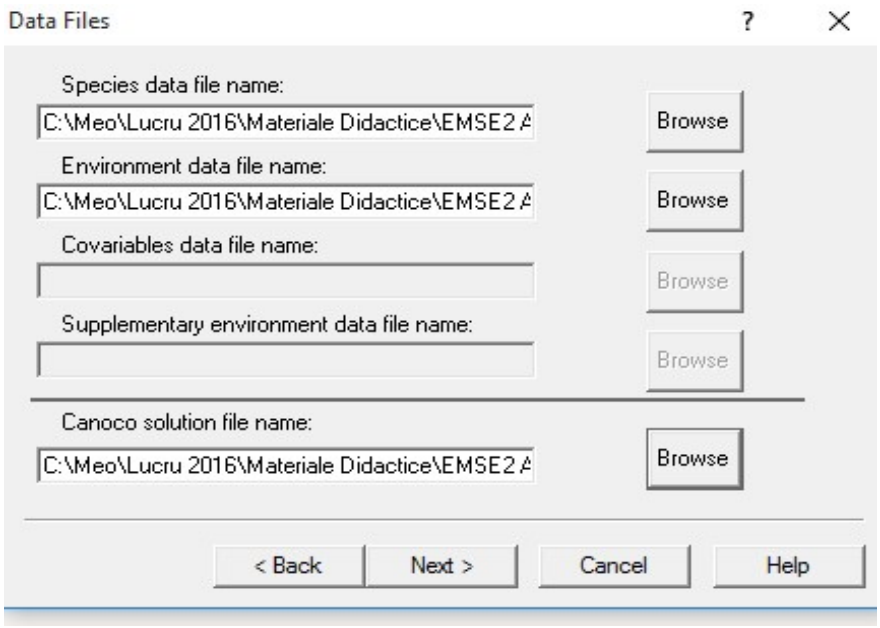


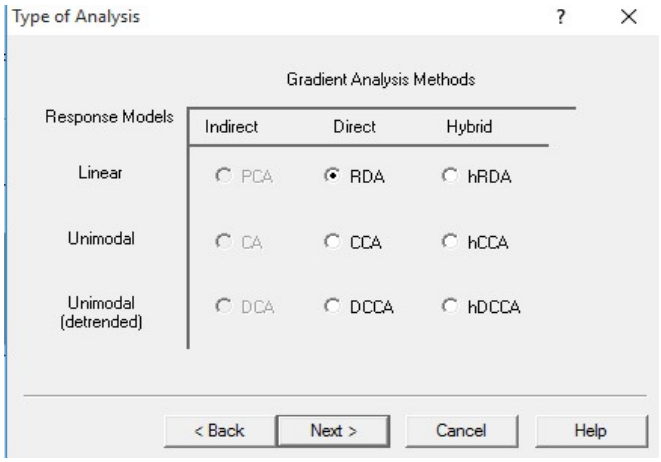
- Selectăm: File, apoi New Project.
- Apare o fereastră care ne întreabă ce date sunt disponibile (Available Data)
- Într-un studiu real, mai întâi vom face câteva interogări numai pe datele de specii, pentru a afla informaţii despre acestea, precum și pentru a selecta tipul de analiză care se pretează la setul de date în cauză (vom afla dacă se poate realiza o analiză de redundanţă RDA, sau una de corespondenţă canonică CCA, în funcţie de lungimea segmentelor care rezultă dintr-o analiză preliminară DCA). Acum, pentru a simplifica, presupunem că am trecut de această etapă, care va fi învăţată mai târziu, și realizăm o RDA. Vom testa setul de date și vom reprezenta grafic diagrama corespunzătoare.



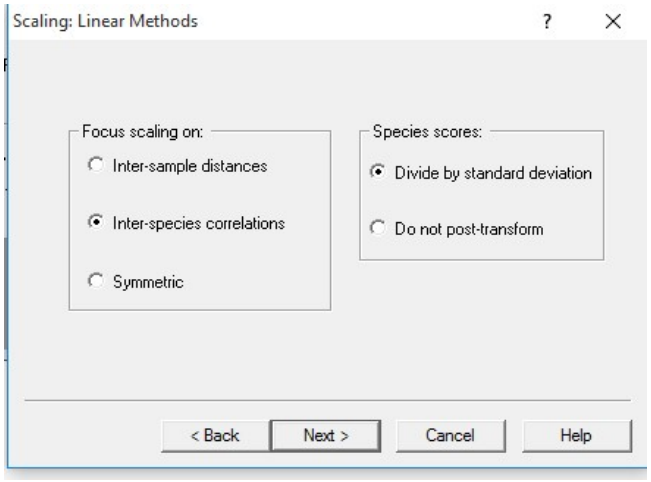
- (acum, la început) selectăm ”Species and environment data available” (sunt disponibile date despre specii și mediu) din meniul de DATE DISPONIBILE PENTRU ANALIZĂ (DATA AVAILABLE FOR ANALYSIS)
- din meniul DATELE DE MEDIU, DACĂ SUNT DISPONIBLE, POT FI UTILIZATE PENTRU (ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO), noi (în acest studiu de caz) vom alege să extragem modelul numai din variaţia explicată de către variabilele de mediu alese (adică analiza directă de gradient) - în original: *extract patterns from the explained variation only (direct gradient analysis).*
- Cealaltă variantă (atunci când nu avem sau nu vrem să ţinem seama decât de variabilele de specii, deci cele care se referă la comunitate), este de a selecta analiza indirectă de gradient (*indirect gradient analysis*).
- Clic pe Next

- Programul ne solicită să încărcăm fișierele în următoarea fereastră. Cu Browse căutăm folderul unde avem datele și (sus) selectăm fișierul cu variabilele de specii (Species data file name), apoi (mai jos) fișierul cu date de mediu (Environment data file name), celelalte două câmpuri le lăsăm în pace, iar în ultima linie trebuie să alegem o cale și să scriem numele unui fișier care va conţine soluţiile analizelor, fișier care trebuie să aibă extensia .sol. De exemplu vom alege numele exercitiu.sol și amplasăm noul fișier în aceeași locaţie cu fișierele de date.
- Clic pe Next



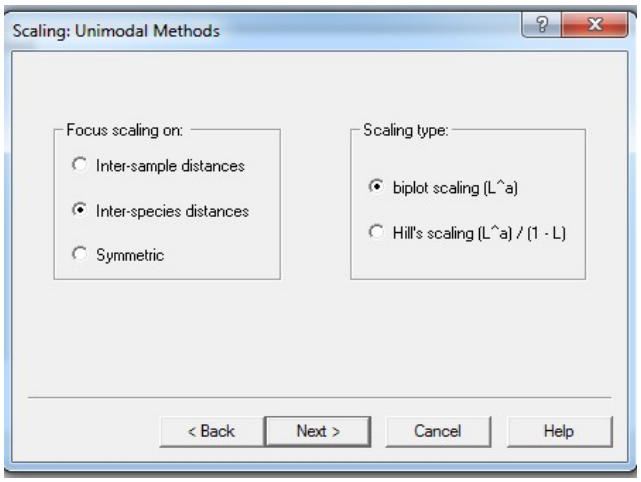


- Alegem tipul de analiză pe care dorim să o realizăm.
- Vom discuta mai târziu care sunt metodele, când și în ce condiţii le alegem.
- Metoda de analiză a gradientului este indirectă, dacă nu avem decât variabile de specii (nu dispunem de date cu privire la mediu) sau nu vrem să le luăm în calcul, la început cel puțin. Analiza este directă în celălalt caz, dar există și posibilitatea să efectuăm o analiză hibridă.
- Deocamdată selectăm RDA (*Redundancy Analysis*) considerând că modelul de răspuns liniar este cel mai potrivit (ceea ce de fapt și este, vom vedea în continuare motivele).
- Next. Va aduce pe ecran **fereastra de scalare a scorurilor de ordonare**



- Deocamdată lăsaţi opţiunile din această fereastră așa cum sunt.

Ce sunt și cum utilizăm ferestrele de scalare? Mai sus am redat fereastra de scalare pentru metoda liniară. Pentru metodele unimodale fereastra de scalare este redată mai jos:



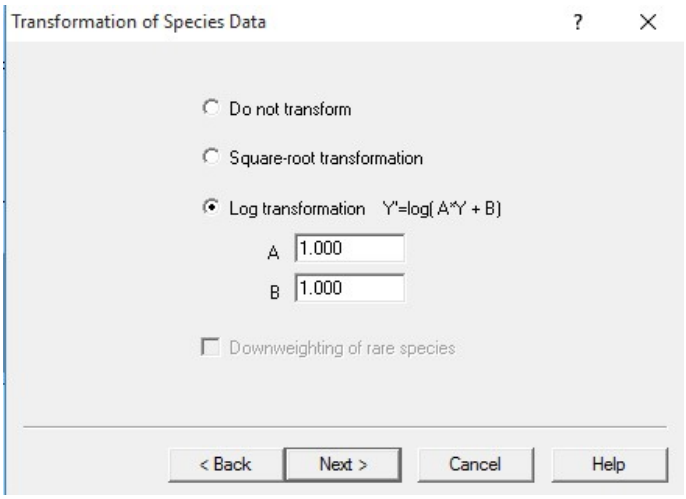
Precizia concluziilor cu privire la similaritatea probelor, a relaţiilor dintre specii și/sau variabilele de mediu, depinde parţial și de scalarea relativă a scorurilor axelor individuale de ordonare. Opţiunile din ferestrele de mai sus permit alternative în interpretarea vizuală a rezultatelor ordonărilor, fără a afecta alte rezultate sau aspecte ale analizelor (cum ar fi valoarea variabilităţii explicate de fiecare axă, puterea și direcţia relaţiei între axe și variabile explicative, rezultatele testelor de semnificaţie sau de eroare de tip I în analizele canonice etc.). Când selectăm opţiunile trebuie să decidem dacă atenţia se va concentra asupra probelor sau asupra speciilor în interpretarea diagramelor de ordonare. Apoi, la modele liniare, se ia decizia dacă lungimea săgeţilor trebuie să reflecte diferenţele dintre abundenţele speciilor (speciile mai abundente să aibă săgeţi mai lungi) sau abundenţa speciilor individuale să fie transformată și raportată la o scară comparabilă. A doua opţiune corespunde la un *biplot de corelaţie*. Dacă (în modelul liniar) selectăm ca scorurile speciilor să fie

împărţite la abaterea standard (*Species scores: divide by standard deviation*), atunci lungimea fiecărei săgeţi a unei specii exprimă cât de bine valorile speciei respective sunt aproximate de către diagrama de ordonare. În cazul diagramelor de ordonare unimodală, vom alege scalarea Hill numai dacă gradienţii compoziţionali sunt foarte lungi (valoare mare a diversităţii beta dintre probe), în celelalte cazuri (și recomand utilizarea acestei opţiuni) scalarea de tip biplot oferă diagrame care pot fi interpretate într-o manieră cantitativă.

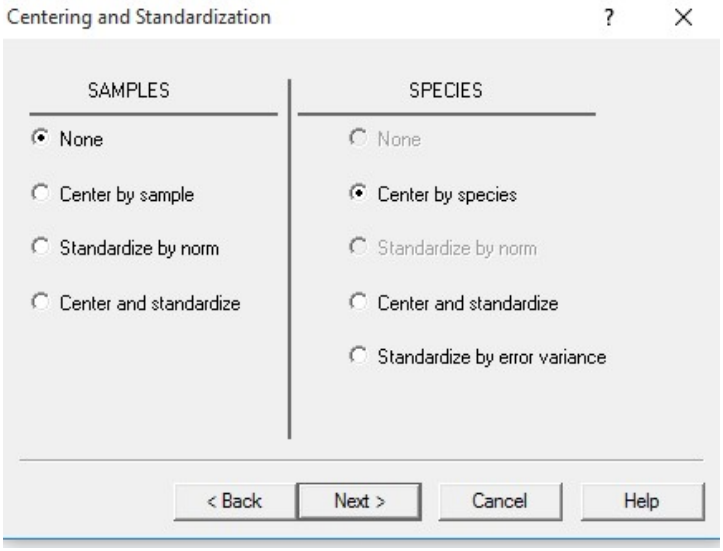
- Daţi clic pe: Next.

- Teoria recomandă ca (în unele cazuri) datele comunităţii, care redau speciile în termeni de abundenţă relativă, să fie transformate. Se poate opta pentru transformare prin extragerea rădăcinii pătrate sau prin logaritmare. Dar, în cele mai multe date de comunităţi, sunt multe specii care apar rar (adică abundenţă zero în multe probe), motiv pentru care transformarea logaritmică nu se poate face direct (logaritm de zero este minus infinit). Variabilele se transformă după expresia:Y' = log(A*Y + B), unde - implicit - A și B sunt 1 (deci Y' = log (Y + 1)). Clic pe Next.

Motivul transformărilor este legat mai ales de relaţia dintre variabilele de mediu și cele de specii. În timp ce primele se modifică treptat, cele dependente prezintă o relaţie neliniară cu primele, de obicei. Multe valori ale speciilor în probe sunt zero, iar în majoritatea cazurilor valorile de abundenţă sunt fie mici, fie (atunci când speciile sunt în zona de optimum) foarte mari. Motivul pentru care datele de specii sunt transformate este de a defini o scară și o relație similară de variaţie cu cea a variabilelor de mediu, adică de a încerca identificarea unei relaţii de tip liniar (sau aproape) între cele două categorii. Transformarea logaritmică nu este singura posibilă: se poate opta și pentru transformarea prin extragerea radicalului sau alta.



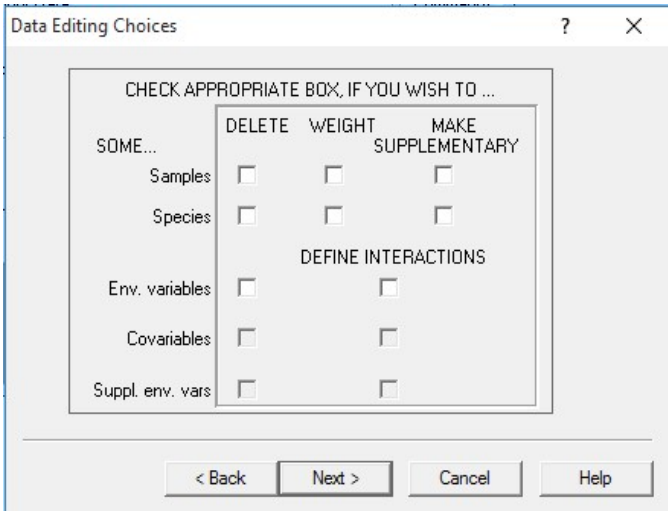
- Următoarea fereastră poate fi lăsată ca atare, și mergem mai departe (Next).



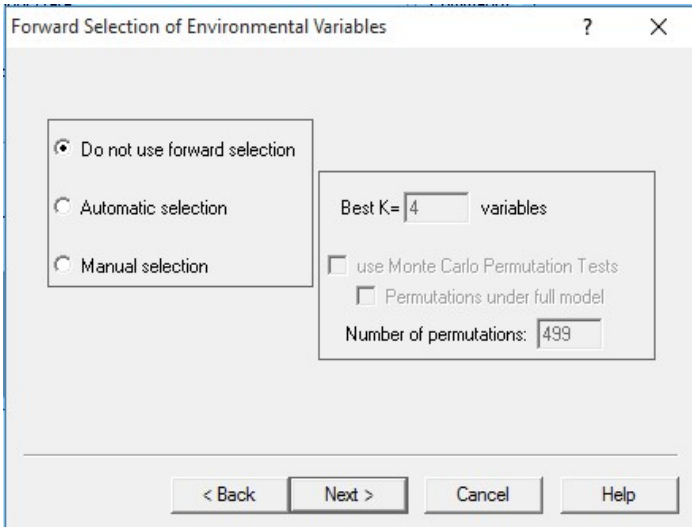
Explicaţii suplimentare:

Această fereastră apare numai la metodele liniare (PCA sau RDA) și se referă la manipularea datelor din matricea comunităţii (variabilele de specii) înainte de calcularea ordonării. Centrarea pe probe (*Center by sample*) în coloana din stânga, imaginea de mai sus, produce o medie de zero pentru fiecare rând (articol). Centrarea pe specii (coloana din dreapta: *Center by species*) produce un rezultat al mediei de zero pentru fiecare coloană. Centrarea pe specii este obligatorie pentru metodele liniare canonice (RDA) sau pentru orice metodă liniară de ordonare parţială (adică acolo unde sunt utilizate covariate). Standardizarea pe specii (sau pe probe) produce ca norma fiecărei coloane (sau rând) să fie egală cu unu. Norma este rădăcina pătrată a sumei de pătrate ale valorilor din coloane (sau rânduri în cel de-al doilea caz). Dacă se optează atât pentru centrare cât și pentru standardizare, centrarea este realizată mai întâi. După centrarea și standardizarea pe specii, coloanele reprezintă variabile cu medie zero și varianţă egală cu unitatea. Atunci, o PCA realizată pe datele de specii va corespunde la o ”PCA pe o matrice de corelaţii” (între specii). Dacă nu se standardizează prin norma speciei, va rezulta o ”PCA pe matrice de varianţă-covarianţă”. De reţinut faptul că standardizarea pe specii (și nu standardizare și nici centrare pe probe) trebuie realizată atunci când variabilele de specii (de răspuns) diferă în ceea ce privește scara de măsură. Dacă sunt variabile de mediu disponibile în metoda de ordonare (RDA), se poate selecta standardizarea prin eroarea varianţei (*Standardize by error variance*), care va calcula, separat pentru fiecare specie, cât din varianţa speciei nu a fost explicată de variabilele de mediu. Inversa erorii de varianţă este **ponderea speciei**. Astfel, cu cât o specie este explicată mai bine de către variabilele de mediu disponibile, cu atât ponderea acesteia este mai mare în analiza finală.

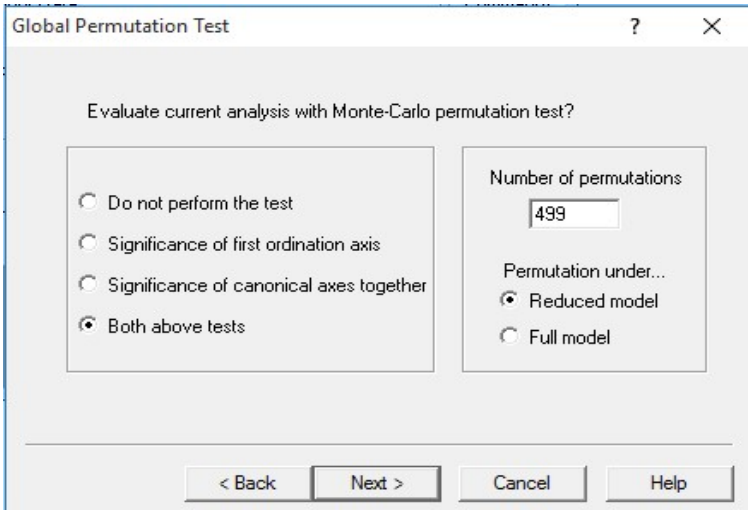
- Next, va aduce fereastra opţiunilor de editare a datelor.



- Dacă avem opţiuni legate de editarea datelor, de a defini interacţiuni între variabile, de a stabili ponderi sau să eliminăm anumite articole și/sau variabile, putem să facem toate acestea și multe altele în această fereastră. Acum nu dorim să facem nimic de acest gen, așa că lăsăm în pace fereastra.
- Next.



- Aceeași decizie. Implicit este selectată opţiunea ”*Do not use forward selection*”.
Dacă dorim să aflăm importanţa fiecărei variabile de mediu, precum și a grupului de variabile, în ceea ce privește explicarea variabilelor dependente, vom selecta opţiunea *Automatic selection*, care va ordona variabilele de mediu în ceea ce privește importanţa lor în explicarea variaţiei compoziţiei speciilor.
- Next



- Mai importantă decât multe altele este testarea independenţei dintre variabilele comunităţii (de specii) și cele de mediu, cu ajutorul testului de permutări prin metoda Monte-Carlo.

Din păcate am văzut că testarea aceasta este frecvent ignorată sau necunoscută, ceea ce nu este corect. De aceea putem alege ambele teste de semnificaţie (*Both above tests*), pentru a testa atât prima axă de ordonare cât și toate axele. În *Canoco 5* putem realiza câte un test separat pentru fiecare axă în parte, nu însă și în această versiune.

- Ipoteza nulă afirmă că nu există suficientă evidenţă pentru a susţine o relaţie semnificativă (suficient de strânsă), de dependenţă, între variabilele de mediu și cele ale speciilor (altfel spus ***H₀: variabilele independente (de mediu) nu explică variaţia compoziţiei comunităţilor***).

- Programul calculează un parametru statistic de tip F (Fischer) între cele două categorii de variabile, după care începe să facă permutări, adică amestecă seturile de date, la întâmplare. Aceeași linie (articol) de valori ale speciilor, este pusă în faţa, și comparată cu alte și alte seturi (linii, articole) de date ale variabilelor de mediu, care au caracterizat originar alte comunităţi, și se calculează în fiecare caz parametrul F al testului Fischer. Ideea de bază este că, dacă într-adevăr nu există relaţii între cele două categorii de variabile, atunci oricare set de date de specii poate fi explicat la fel de bine de oricare altă combinaţie de valori ale variabilelor de mediu.

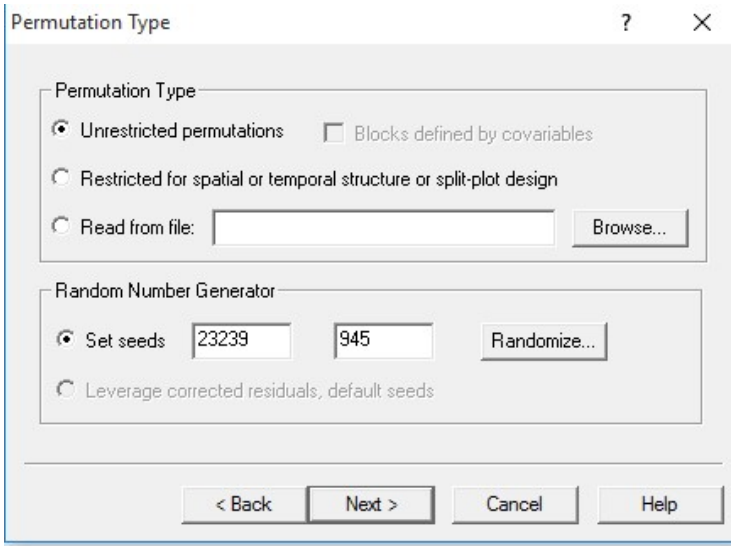
- Softul calculează de câte ori, în urma combinaţiilor întâmplătoare, a rezultat o valoare a lui *F* mai mare decât cea calculată pe datele originale (setul inițial, real, de date).

$$p = \frac{n_x + 1}{N + 1}$$

unde: n_x = de câte ori (în urma câtor permutări) a rezultat o valoare F mai mare decât cea originală, N = numărul total de permutări, p = probabilitatea ca ipoteza nulă să fie adevărată.

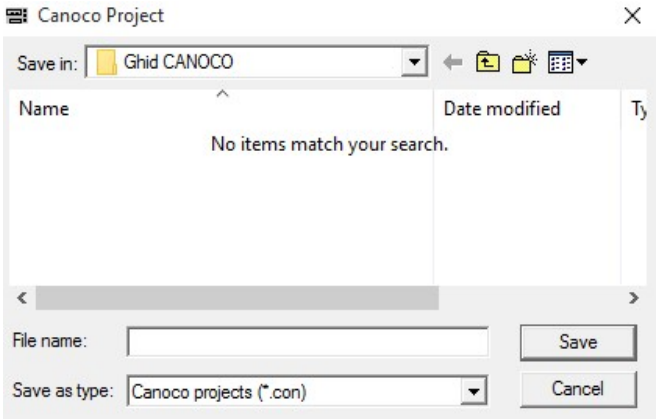
Dacă p < 0.05 (sau nivelul de asigurare ales), respingem ipoteza nulă, caz în care afirmăm că există o relaţie statistic semnificativă între variabilele explicative și cele dependente (între cele de specii și de mediu), respectiv că variaţia compoziţiei comunităţilor este bine și semnificativ explicată de variabilele de mediu alese, în ceea ce privește (sau de către) o axă sau de toate axele de ordonare.

- Next.

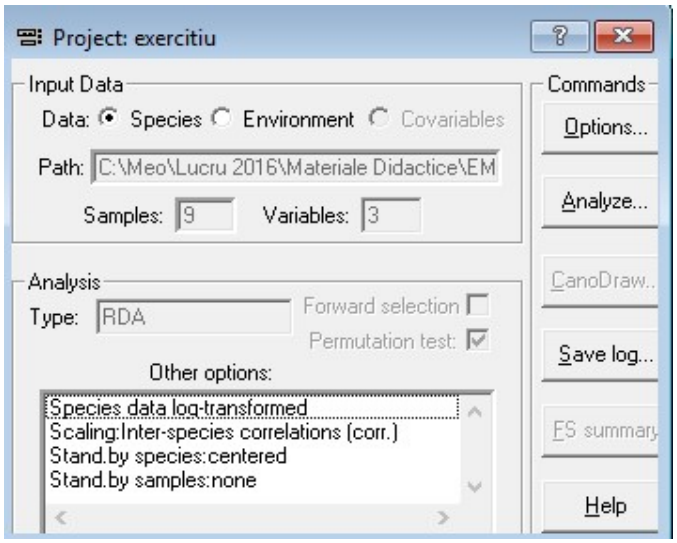


- Dacă dorim să modificăm sau să restricţionăm permutările etc. putem să o facem din următoarea fereastră, dacă nu, lăsăm opţiunile implicite și trecem cu ”Next” mai departe.

- Suntem anunţaţi că proiectul a fost încheiat, dar dacă mai dorim să revenim asupra unei opţiuni, acum este momentul, dacă nu, selectăm ”Finish” și va apare fereastra de denumire a produsului (a rezultatului) proiectului nostru.



- Alegem un nume și eventual tastăm extensia pentru analiză sau proiect *.con*. De exemplu putem opta pentru *exercitiu.con*.
- Clic pe Save.



- Vedem din nou opţiunile și caracteristicile proiectului nostru, ne putem întoarce (Options) pentru a modifica proiectul, putem realiza analizele (Analyze), putem salva rezultatele (Save log) etc. Noi dăm clic pe Analyze și derulăm fereastra de rezultate (jurnalul).

```
**** Summary of Monte Carlo test ****

Test of significance of first canonical axis: eigenvalue =    0.897
                                              F-ratio    =   34.882
                                              P-value    =    0.0020

Test of significance of all canonical axes  : Trace      =    0.948
                                              F-ratio    =   18.376
                                              P-value    =    0.0040

(  499 permutations under reduced model)
```

- Primim mai multe informaţii. La sfârșitul analizei (ca mai sus) aflăm dacă ipotezele nule care susţin independenţa dintre cele două categorii de variabile, respectiv faptul că prima și, separat, toate axele canonice, sunt semnificative. După cum vedem în ambele cazuri probabilitatea este foarte mică (sub 0.005), deci ipoteza nulă este respinsă. Acest fapt înseamnă că există relaţii semnificative între seturile noastre de variabile, deci studiul poate continua.
- Derulăm rezultatele analizei pentru a primi o sumedenie de informaţii, dar foarte importante, acum că știm că dispunem de un studiu (proiect) semnificativ, sunt cele redate astfel:

Axes		1	2	3	4	Total variance
Eigenvalues	:	0.897	0.047	0.004	0.050	1.000
Species-environment correlations	:	0.999	0.965	0.279	0.000	
Cumulative percentage variance						
of species data	:	89.7	94.4	94.8	99.8	
of species-environment relation:		94.6	99.6	100.0	0.0	

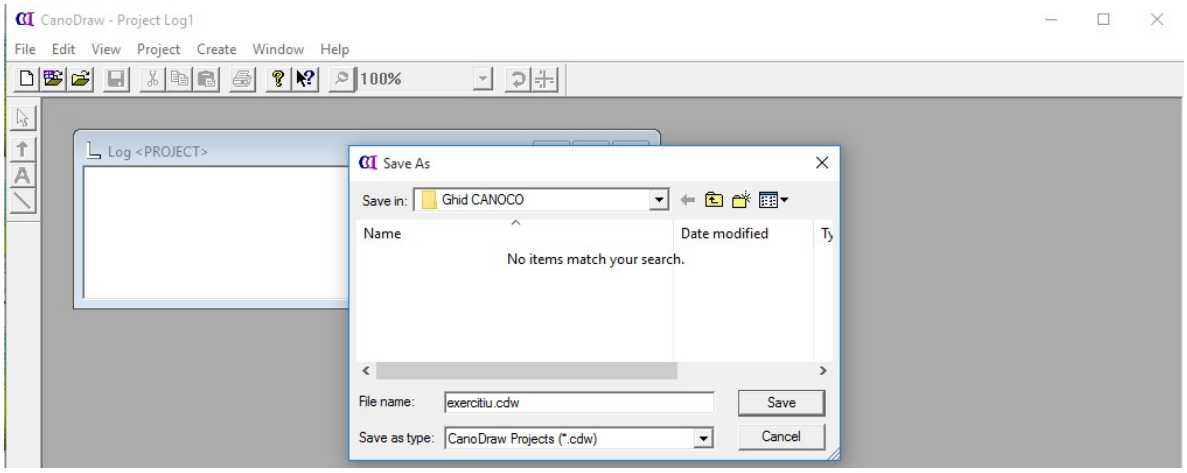
Sum of all	eigenvalues	1.000
Sum of all canonical	eigenvalues	0.948

Se observă din linia de valori ale proporţiei de varianţă extrasă de către axe (respectiv din **valorile rădăcinilor latente** = ”*Eigenvalues*” ale matricii) că prima axă este responsabilă pentru extragerea a 89.7% din variaţie, iar primele două axe explică 94.4% din toată variaţia conţinută de către variabile. Prin urmare este foarte justificată reprezentarea variabilelor pe primele 2 axe.

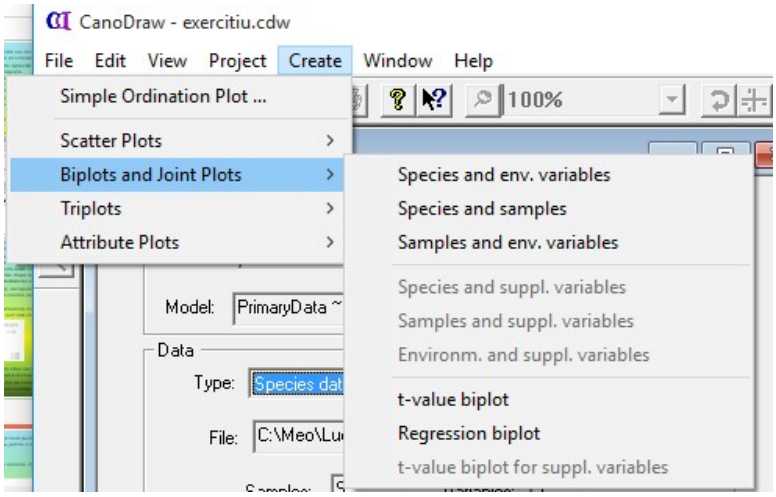
La întrebarea ”câte procente din variaţia în compoziţia comunităţii sunt explicate de variaţia variabilelor independente” (altfel spus care este echivalentul coeficientului de determinare din analiza de regresie pentru această analiză multivariată), răspunsul este dat de suma rădăcinilor latente canonice înmulţit cu 100 (*sum of all canonical eigenvalues* = 0.948 * 100 = 94.8%). Prin urmare 94.8% din compoziţia specifică a comunităţii este explicată de variabilele de mediu considerate. Concluzionăm că am obţinut un model foarte bun!

- Ne întoarcem (*switch views*) la ultima fereastră a proiectului și acum putem selecta să intrăm în al treilea modul al acestui soft, și anume în *CanoDraw*, pentru a reprezenta relaţiile și diagramele dorite. Clic pe *CanoDraw.*

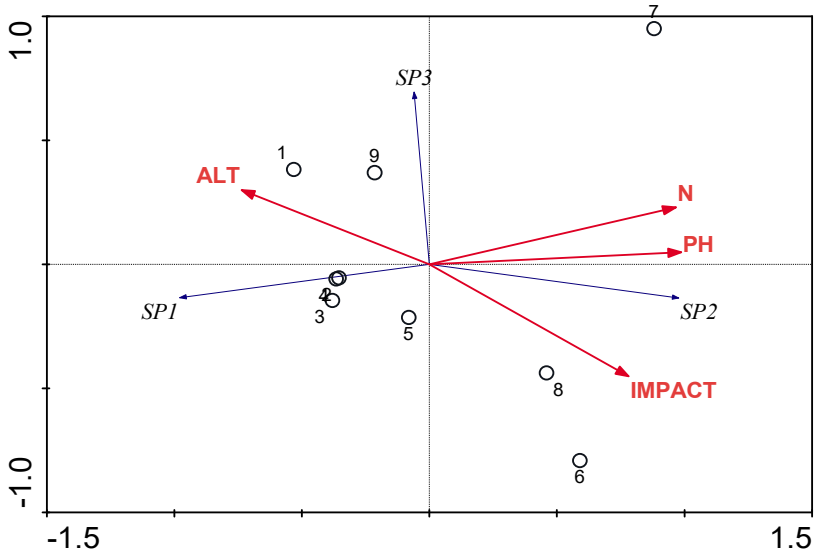
- Vom alege un nou nume pentru fișierul din modulul grafic, cu extensia .cdw. De exemplu exercitiu.cdw (ca în fereastra de mai jos). Clic pe Save.



- Și acum (în sfârșit) putem realiza diagrame de ordonare și multe alte grafice. De asemenea putem selecta diverse opţiuni și condiţii de reprezentare grafică (putem include sau exclude variabile și specii, de exemplu cele foarte rare etc.)



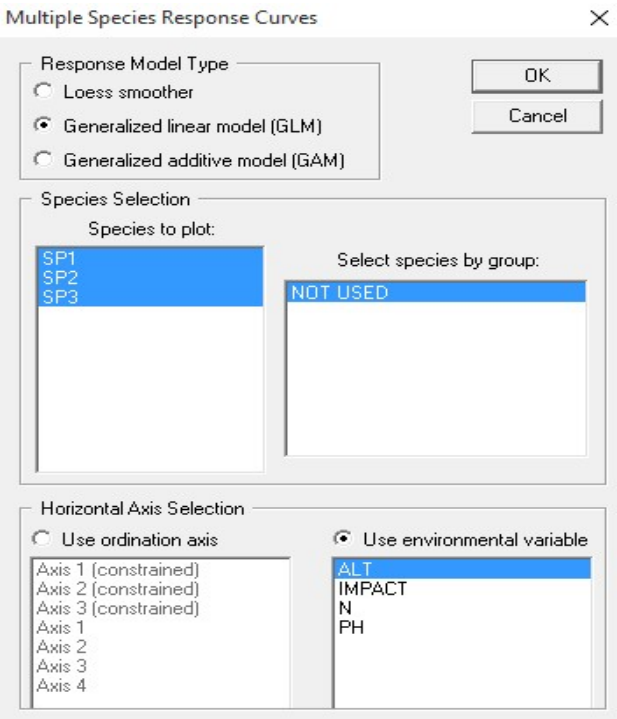
- Prin selecţia: Create → Biplots and Joint Plots → Species and env. variables - rezultă diagrama de ordonare care corespunde unei analize de redundanţă (RDA), adică cea pe care am dorit să o realizăm;
- urmează interpretarea...
- sau putem să selectăm Triplots și apoi clic pe ”*with environmental variables*”, obţinând:



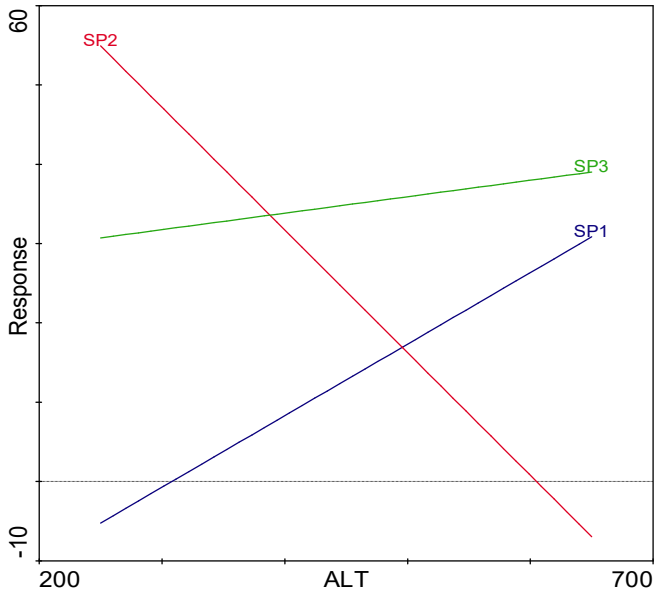
În aceasta vedem poziţia relativă a speciilor, a variabilelor de mediu (independente sau explicative), dar și probele (staţiile de eșantionare) sau comunităţile

individuale și caracteristicile mediului lor. Această analiză permite interpretări de amănunt cu privire la preferinţele speciilor pentru diferitele valori (domenii de variaţie) ale variabilelor de mediu, asemănări și deosebiri între valenţele ecologice dar și între dimensiunea de habitat a nișelor ecologice etc.

Răspunsul speciilor la o anumită variabilă se face prin analiza de regresie; *Canoco* permite o serie de opţiuni pentru aceasta (regresie liniară, pătratică, cubică etc.), precum și testarea semnificaţiei acestora. De exemplu dacă dorim să vedem răspunsul după un model liniar al variaţiei abundenţei speciilor la altitudine, selectăm:
Create → Attribute Plots → Species response curves...
urmează alte câteva opţiuni ....



... și (după câteva ferestre de rezultate) rezultă graficul de mai jos. Din rezultatele intermediare aflăm că relaţiile dintre abundenţa speciei 1, a cărei abundenţă crește cu altitudinea și cea a speciei 2, care descrește cu altitudinea, sunt semnificative, nu însă și a speciei 3 care pare să fie independentă de aceasta.



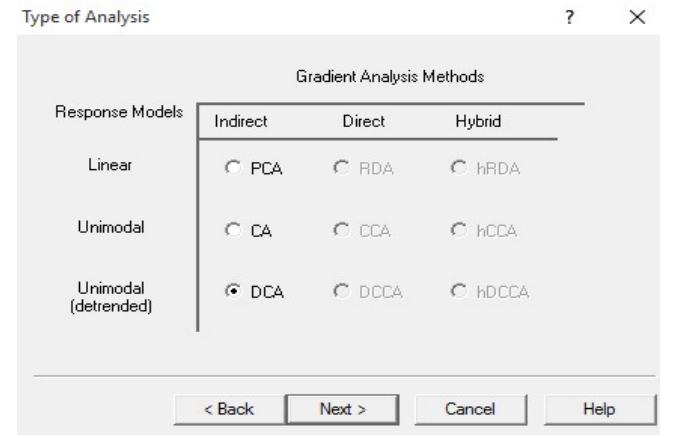
Cum selectăm metoda de analiză de ordonare (model liniar sau unimodal)?

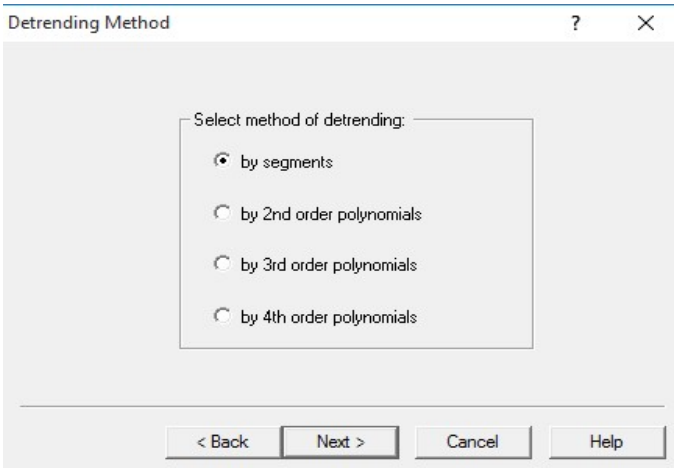
Metodele liniare sunt:
- potrivite pentru analiza gradienţilor scurţi de mediu,
- pentru analiza experimentelor manipulative cu mulţi factori,
- analiza seturilor de date omogene.

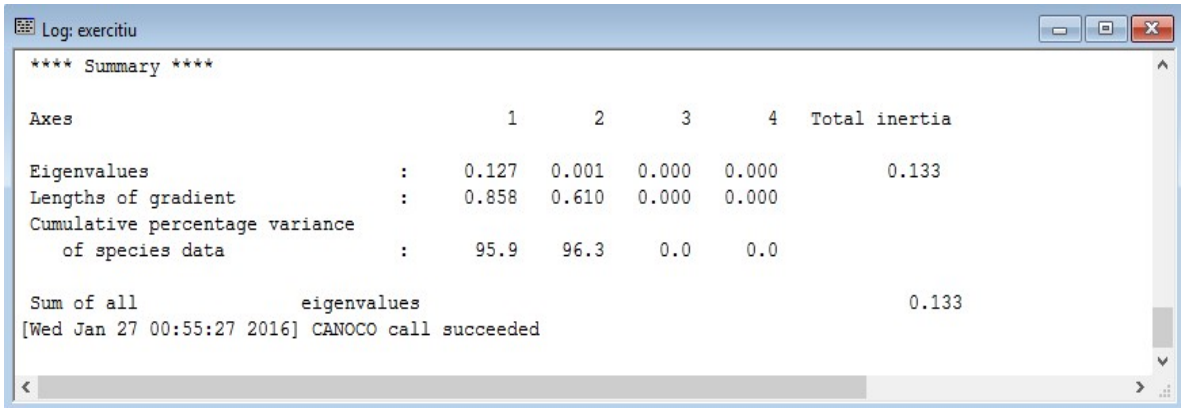
Metodele unimodale sunt:
- potrivite pentru analiza gradienţilor lungi de mediu,
- a seturilor de date heterogene,
- a variabilelor numerice continue de mediu.

Algoritmul simplificat pentru selecţia metodei:
1. Rulăm iniţial o DCA (*detrended correspondence analysis*) și ne uităm la lungimea cea mai mare a gradienţilor (căutăm și notăm cea mai mare valoare a *Length of gradients*);
2. Dacă lungimea maximă este **sub valoarea 3, utilizaţi metode liniare**;
3. Dacă lungimea este **peste 4, utilizaţi metode unimodale**;
4. Dacă lungimea este **între 3 și 4, puteţi utiliza oricare dintre metode**.

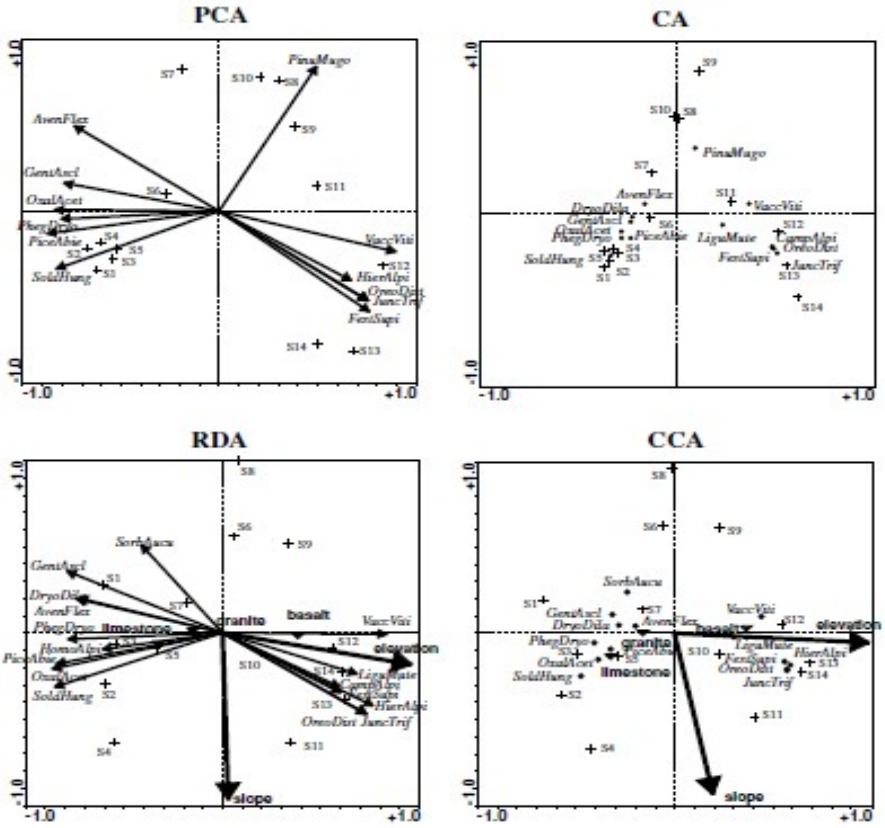
Aplicaţia DCA și evaluarea lungimii gradientului primei axe de ordonare la exerciţiul nostru; observăm că lungimea este de 0.858, mult sub valoarea de 3, motiv pentru care am ales o analiză RDA și nu CCA!


Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
< Back
Next >
Cancel
Help


Detrending Method
Select method of detrending:
by segments
by 2nd order polynomials
by 3rd order polynomials
by 4th order polynomials
< Back
Next >
Cancel
Help

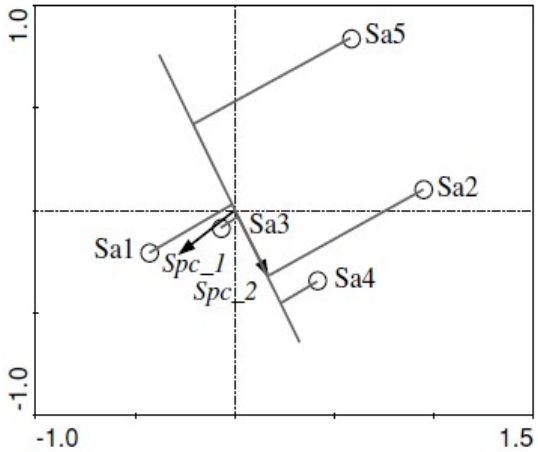

Log: exercitiu
**** Summary ****
Axes 1 2 3 4 Total inertia
Eigenvalues : 0.127 0.001 0.000 0.000 0.133
Lengths of gradient : 0.858 0.610 0.000 0.000
Cumulative percentage variance
of species data : 95.9 96.3 0.0 0.0
Sum of all eigenvalues 0.133
[Wed Jan 27 00:55:27 2016] CANOCO call succeeded

Exemple de diagrame de ordonare



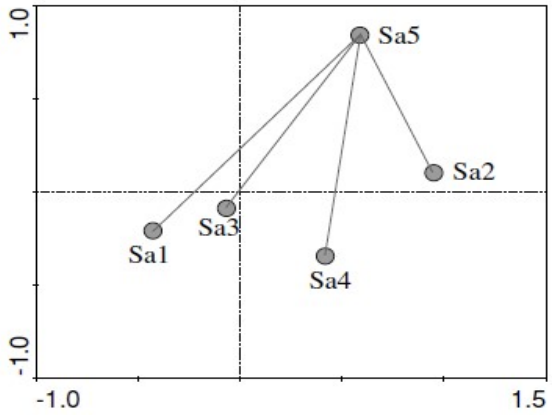
Diagrame de ordonare şi interpretarea acestora - metode liniare

Mai jos redăm diverse scheme de diagrame de ordonare, sub care specificăm câteva reguli de interpretare.

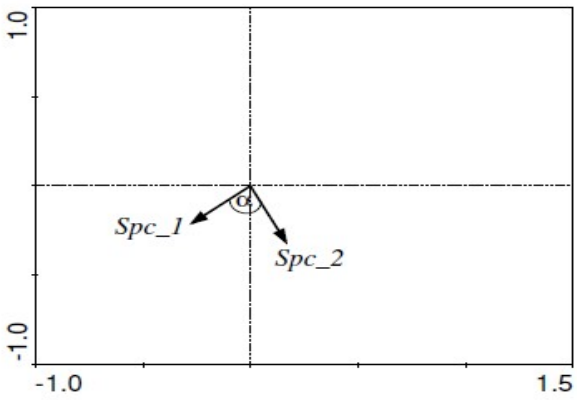


Proiectarea probelor pe vectorii speciilor (pe Spc_2) într-un biplot dintr-o diagramă de ordonare liniară. Aici prognozăm cea mai mare abundenţă a Spc_2 în probele Sa4

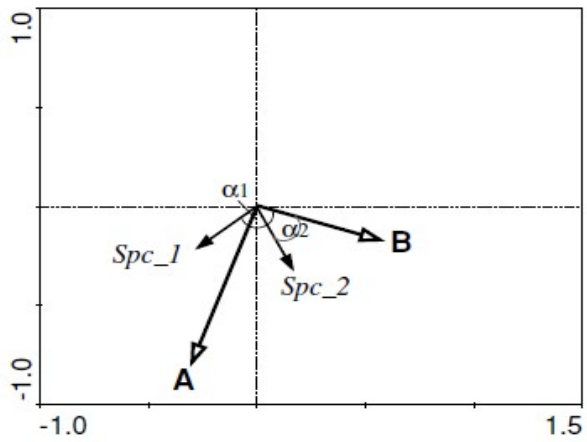
urmată de Sa2, apoi în Sa3 abundenţa speciei este aproape de valoarea medie din probe (proiecţia este în zona de origine), valoarea cea mai mică fiind în Sa5.



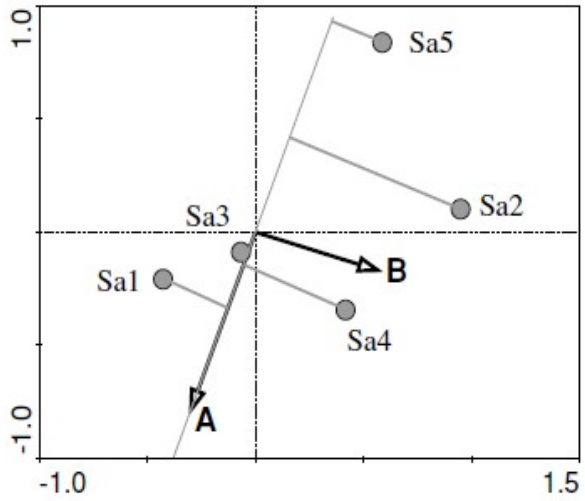
Distanţa dintre probe într-o diagramă de ordonare. Dacă măsurăm disimilarităţile dintre proba Sa5 și celelalte, utilizând distanţa euclidiană, cea mai scurtă distanţă este între Sa5 și Sa2 (sunt cele mai asemănătoare) iar cea mai mare distanţă este între aceasta și Sa1, deci prognozăm că sunt cele mai disimilare (deci diferite ca și compoziţie a speciilor).



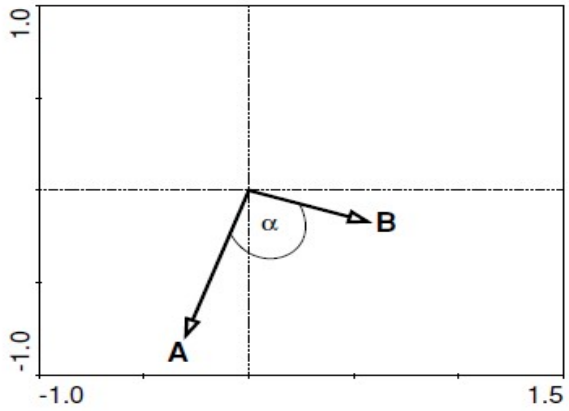
Unghiuri dintre specii într-o diagramă realizată într-o analiză de ordonare liniară. Deoarece săgeţile celor două specii sunt la un unghi de aproape 90º, între cele două specii se prognozează că nu există nici o corelaţie (aproape zero), adică sunt independente în spaţiul de ordonare.



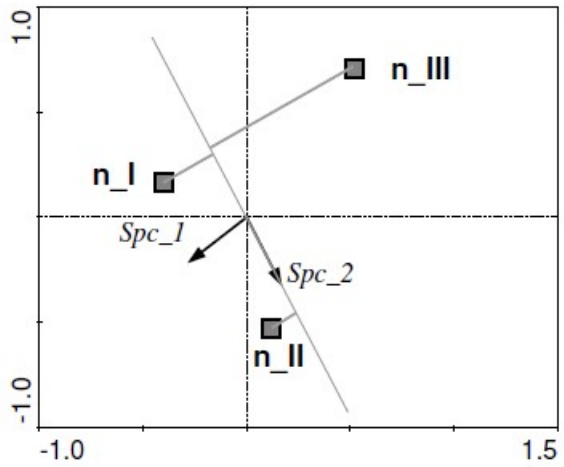
Unghiuri între specii și variabilele de mediu într-o diagramă de ordonare prin model liniar. Unghiul speciei Spc_1 este peste 90^0 fapt care indică o relaţie negativă (abundenţa speciei 1 scade cu creșterea lui B, deci regresie negativă), în timp ce unghiul mic (ascuţit) cu B al lui Spc_2 indică faptul că Spc_2 tinde să crească în abundenţă atunci când B crește. **Regula biplotului (proiectarea vârfurilor săgeţilor speciilor pe săgeţile variabilelor de mediu)** oferă o aproximare și mai precisă.



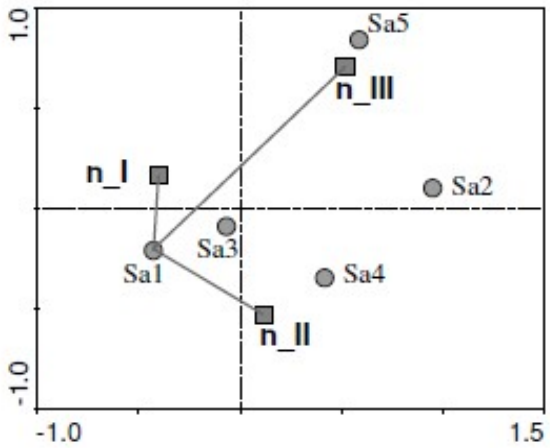
Proiectarea probelor pe săgeţile variabilelor de mediu cantitative. Observaţi că variabila A este prognozată să prezinte valori similare pentru probele Sa3 și Sa4 deși acestea sunt localizate la distanţe și direcţii diferite de linia lui A. A prezintă cele mai mari valori în proba 1 și cele mai mici în proba 5.



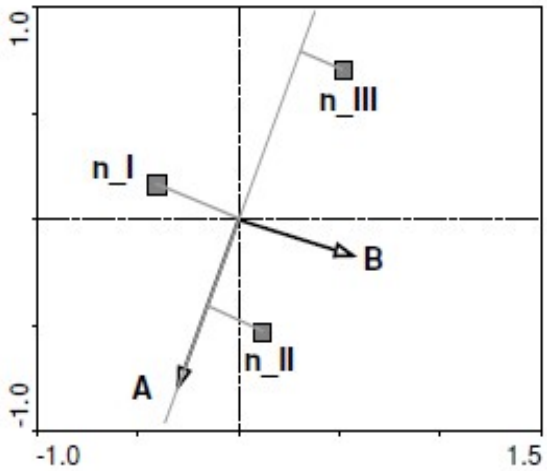
Măsurarea unghiurilor dintre două variabile de mediu. Unghiul aproape drept sugerează că între cele două variabile nu există aproape nici o corelaţie (valoare aproape de zero).



Proiectarea centroizilor variabilelor substitutive (*dummy variables*) pe săgeţi ale speciilor într-o diagramă provenită dintr-o metodă liniară. Prognozăm că cea mai mare abundenţă medie a speciei Spc_2 este în probele care aparţin clasei n_II, în timp ce în cele din clasele n_I și n_III, specia respectivă are o abundenţă medie prognozată mai mică și similară.

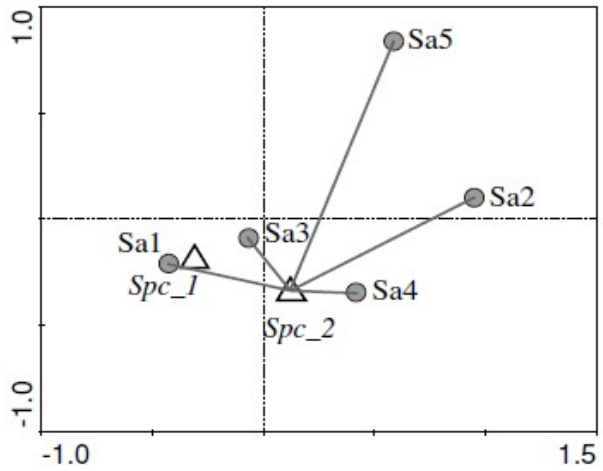


Măsurarea distanţei dintre probe și centroizii variabilelor substitutive. Aici prognozăm că proba Sa1 are cea mai mare probabilitate să aparţină clasei n_I și cea mai mică de a aparţine clasei n_III. Aceleași interpretări le vom realiza și în versiunile superioare ale *Canoco 5.x* unde se lucrează cu niveluri ale variabilelor factoriale și nu cu variabile substitutive.

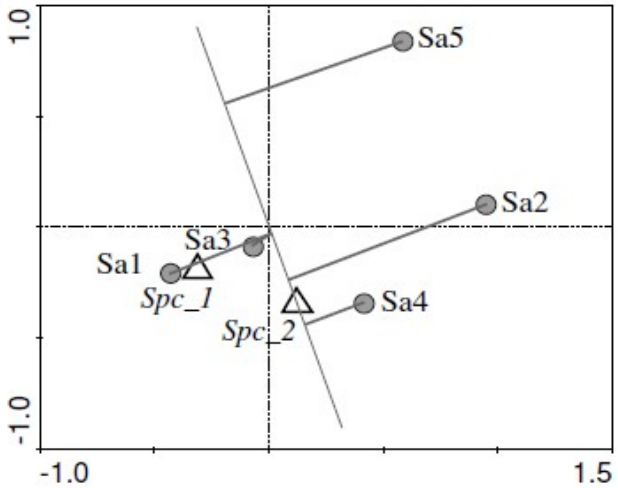


Proiectarea centroizilor variabilelor substitutive pe săgeţile variabilelor cantitative de mediu. Probele aparţinând clasei n_II sunt prognozate să aibă cea mai mare valoare a variabilei A, urmate de cele din clasa n_I și apoi de cele din n_III în care A are cea mai mică valoare medie prognozată.

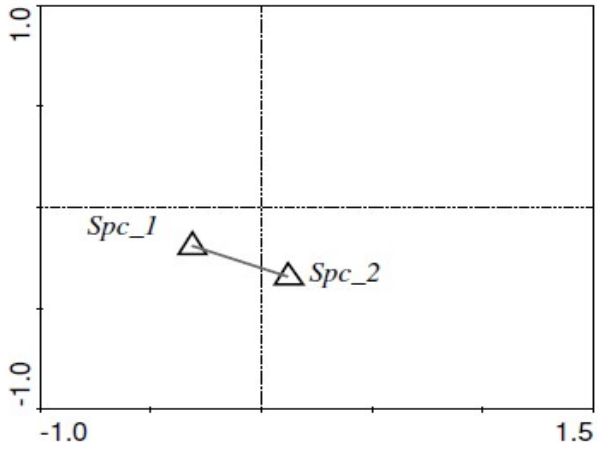
Metode și diagrame unimodale de ordonare



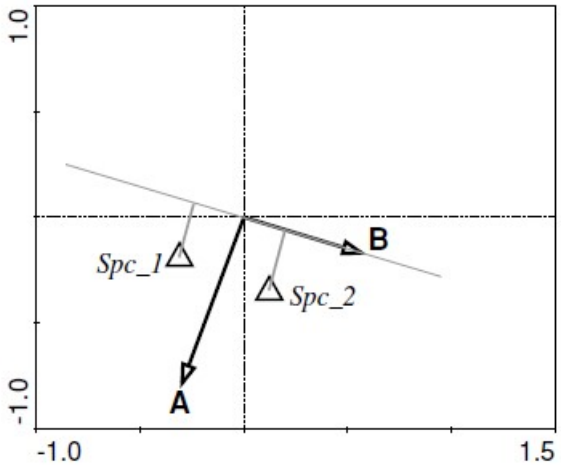
Distanţa dintre specii și probe într-o diagramă unimodală. Specia Spc_2 are probabil cea mai mare abundenţă relativă în probele Sa4 și Sa3 și cea mai mică abundenţă în Sa5 (probabil lipsește din aceasta).



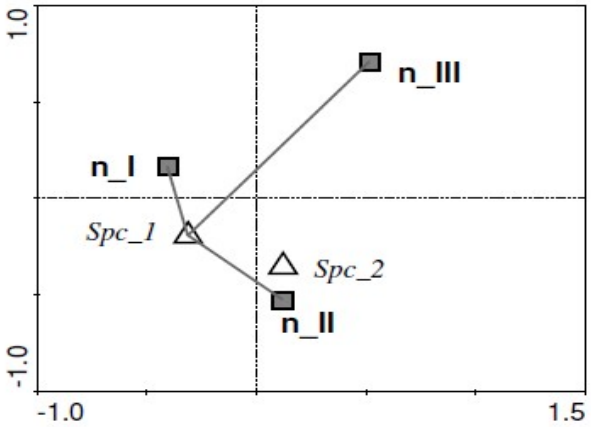
Regula biplotului aplicată la specii și probe în diagrama de ordonare provenită dintr-un model unimodal (o metodă unimodală de ordonare). Specia Spc_2 are prognozabil cea mai mare frecvenţă relativă în probele Sa4 și Sa2 și cea mai mică în Sa5. Această specie este prognozată să apară în probele Sa1 și Sa3 la valoarea medie a frecvenţei sau a abundenţei ei relative. Pentru fiecare specie se poate trasa o linie (imaginară) între punctul de optim și origine, în scopul facilitării aplicării regulii de proiecţie a biplotului.



Măsurarea distanţelor dintre specii într-o diagramă de ordonare realizată printr-o metodă unimodală. Distanţa dintre specii într-o scalare biplot (concentrată pe distanţa dintre specii) aproximează distanţa chi-pătrat dintre distribuţia speciilor.



Proiecţia speciilor pe săgeţile variabilelor cantitative de mediu. Interpretăm graficul de mai sus afirmând că specia Spc_2 are probabil optimumul faţă de variabla B la valori mai mari ale acesteia decât specia Spc_1. Proiectând speciile pe săgeţile variabilelor de mediu putem să le ordonăm în funcţie de valoarea optimă a acestora pe gradientul variabilei. În mod similar putem proiecta probe pe săgeţile variabilelor explicative pentru a aproxima valorile din tabelul datelor de mediu.



Măsurarea distanţelor dintre specii și centroizii variabilelor substitutive de mediu. Frecvenţa medie relativă a speciei Spc_1 este prognozată de a fi maximă în probele din clasa n_I, urmată de cele din clasa n_II și minimă în clasa n_III.

Studiu de caz 2: comunităţi de păsări dintr-o pădure

Scop: studiul variabilităţii comunităţilor de păsări în relaţie cu variaţia variabilelor de habitat (variabilitatea mediului)

Metodă: s-au numărat perechile de păsări cuibăritoare, prin ascultarea cântecelor specifice, de 4 ori, câte două investigaţii (campanii de teren) executate în două sezoane de reproducere succesive, în cadrul unei reţele de puncte echidistante, alese într-o pădure montană. Unele puncte au fost amplasate într-un arboret dominat de molid, altele de fag. Variaţia mediului se referă la gradul de acoperire cu vegetaţie (lemnoasă, tufișuri, pășuni instalate în urma defrișării), altitudine, pantă, expoziţie etc. (studiu realizat de M.E. Šálek în Munţii Velka Fatra, Slovacia).

Variabile: încărcaţi fișierul ***birds.xls*** care conţine două tabele (fișiere) - **fișierul se găsește pe pagina materiei.**

- Variabilele dependente (specii) sunt reprezentate prin valorile medii ale abundenţei (37 de specii respectiv variabile), respectiv suma perechilor (cântecelor) numărate în fiecare pătrat, împărţit la 4 (cele patru studii de teren);
- Variabilele de mediu (care descriu caracteristicile habitatelor) sunt amplasate separat, dar în același document (foaie de calcul) și sunt 13 la număr.

Variabilele de mediu sunt:

- ***Altit*** = singura valoare numerică, reală, semnificând media altitudinii pătratului de eșantionare (sunt 43 de pătrate corespunzătoare);

Celelalte entităţi conţin un amestec de variabile strict ordinale sau semi-cantitative;

- ***Forest*** = acoperirea suprafeţei pătratului de eșantionare cu arbori (megafanerofite);
- ***ForDens*** = densitatea medie a vegetaţiei arborescente;

- ***BrLeaf*** = frecvenţa relativă a foioaselor în arboret (0 = numai molidiș, până la 4 = numai foioase în stratul de arbori)
- ***E2*** = acoperirea cu tufărișuri;
- ***E2Con*** = procentul de conifere (molid) în stratul de tufărișuri;
- ***E1*** = acoperirea cu stratul ierbos și ***E1Height*** = înălţimea medie a acestuia;
- ***Slope*** = panta în grade împărţite la 5;

Urmează două perechi de variabile substitutive (*dummy variables*) care codifică două niveluri posibile pentru doi factori: poziţia unor stânci mari pe suprafaţa pătratului ***Rocks* și *NoRocks*** precum și poziţia sau nu a pătratului pe pante însorite, adică sud-est, sud și sud-vest = ***Warm*,** respectiv ***Cold*** în celelalte cazuri.

Mod de lucru:

- Introducem fișierele (variabilele și valorile corespunzătoare) în WCanoImp (putem în format întreg sau condensat) prin marcare și copiere, alegem numele lor, ex.:

birds_spe.dta pentru datele de specii și
birds_env.dta pentru datele de mediu.

- Putem selecta *Species and environmental data available,* dar mai întâi vom decide asupra tipului de model pe care îl vom aplica, motiv pentru care în prima fereastră alegem:
- *indirect gradient analysis.* În fereastra următoare selectaţi fișierele și generaţi numele fișierului de soluţii (de exemplu ***dca.sol***)

În ferestrele următoare vom selecta:

DCA > *by segments* în pagina de *Detrending Method* > *Log transformation* în următoarea > lăsaţi în pace celelalte ferestre clic pe *Finish* și daţi un nume proiectului (ex. ***dca.con***) și
cereţi analiza (*Analyze*)

- treceţi în LogView, observaţi că primele două axe explică cca. 36% din variabilitatea în datele de specii, iar în linia care redă lungimea gradienţilor, vedem cifrele:

Lengths of gradient : 2.001 1.634 1.214 1.613

- Gradienţii sunt relativ scurţi, motiv pentru care optăm pentru un model liniar.
- Începem un nou proiect pe baza celui precedent, deoarece o serie de opţiuni rămân valabile.
- Cu proiectul ***dca.con*** deschis în spaţiul de lucru, selectăm *File* > *Save As...* din meniul de comenzi, după care scriem un nou nume (de exemplu ***rda.con***). Când solicită programul, vom șterge informaţiile din fereastra de rezultate (jurnalul); clic pe *Clear the log window.*
- Clic pe *Options* apoi lăsăm *Species and environment data available* dar apoi selectăm: *direct gradient analysis*
- Vom introduce, atunci când softul o cere, un nou nume pentru fișierul de soluţii

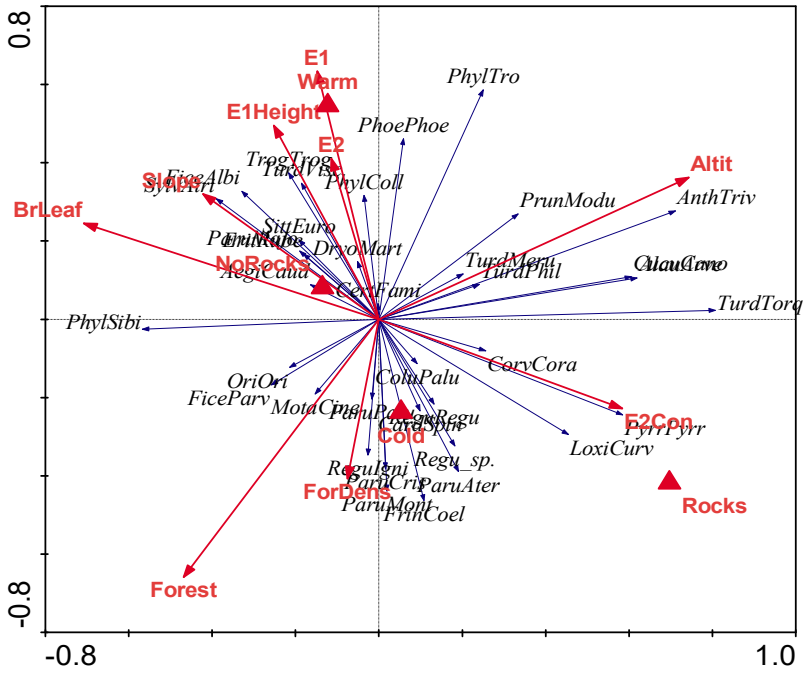
(de exemplu ***rda.sol***).
- În fereastra de *Type of Analysis* selectăm ***RDA*;**
În rest, lăsăm neschimbate toate celelalte selecţii implicite, deoarece acestea sunt potrivite acestei analize.
- În fereastra de *Global Permutation Test* selectăm *Both above tests;*
- Solicităm analiza și ne uităm la valorile afișate în fișierul de rezultate;
- Testele indică respingerea ipotezelor nule (p=0.002), deci există relaţii semnificative între variabilele dependente și cele de mediu (putem merge mai departe);

```
**** Summary ****

Axes                                    1       2       3       4  Total variance

Eigenvalues                        :  0.170   0.088   0.053   0.041          1.000
Species-environment correlations   :  0.919   0.846   0.794   0.821
Cumulative percentage variance
   of species data                 :   17.0    25.8    31.0    35.2
   of species-environment relation:   38.4    58.2    70.1    79.5

Sum of all              eigenvalues                                          1.000
Sum of all canonical    eigenvalues                                          0.442


All four eigenvalues reported above are canonical and correspond to axes that
are constrained by the environmental variables.
```

Constatăm că variabilele de mediu alese explică 44.2% (sum of all eigenvalues * 100) din variaţia compoziţiei în specii.

- fără a mai intra în alte amănunte, deschidem CanoDraw, alegem un nou nume pentru grafic (de exemplu ***rda.cdw* și trebuie să declarăm variabilele de tip indicator sau de substituţie);**
- specificăm variabilele nominale *Rocks*, *NoRocks*, *Warm* și *Cold* (utilizând calea comenzilor: *Project* > *Nominal variables* > *Environmental variables*);
- solicităm graficul pe calea:
Create > *Biplots and Joint Plots* > *Species and env. variables*

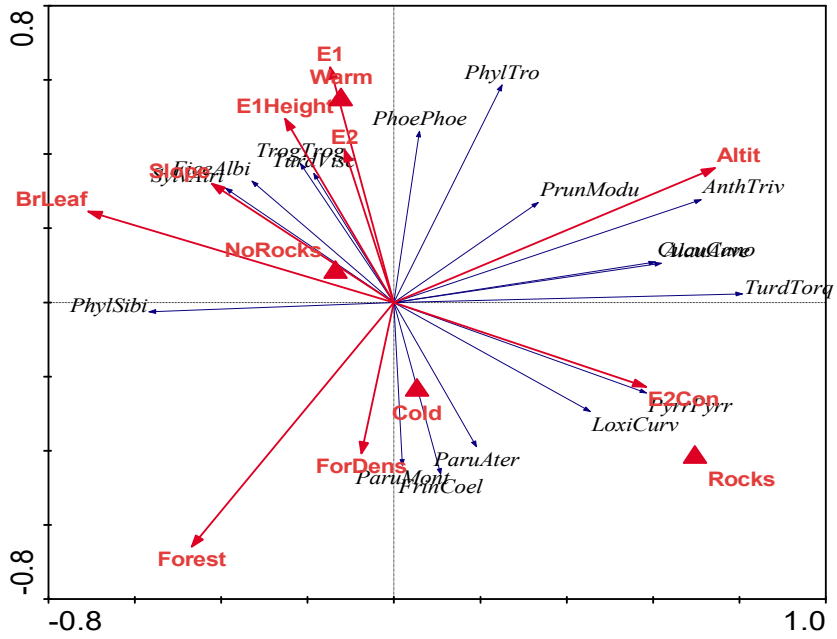
Rezultatul o să ni se pară foarte încărcat:



- deoarece sunt prea multe specii, săgeţi și etichete în grafic, vom selecta numai speciile care sunt optim caracterizate sau descrise de primele două componente principale:

Project > *Settings* iar în fereastra de *Inclusion Rules* vom specifica valoarea 15 în câmpul *From* din aria *Species Fit Range* și închidem dialogul cu *Ok.*

- solicităm redesenarea diagramei cu comenzile *Create* > *Recreate graph.*

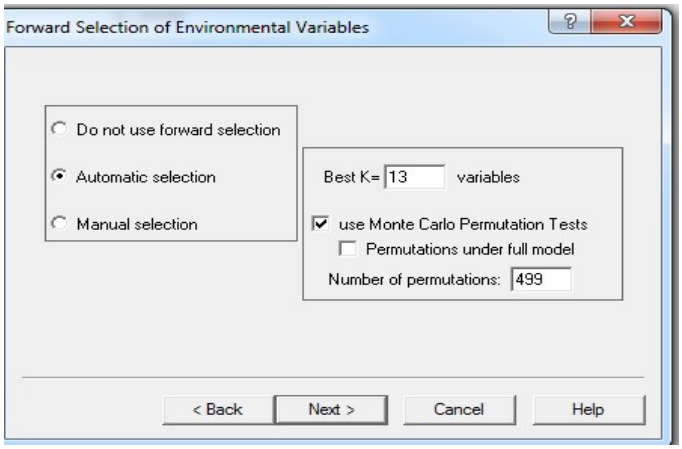


Rezultă o nouă diagramă RDA. Graficul este mai ”aerisit”: cum interpretaţi? (activitate independentă)

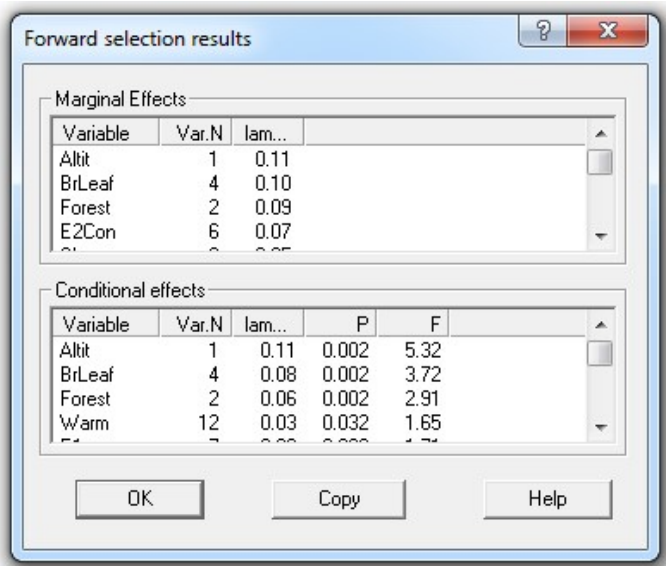
- Cum clasificăm variabilele de mediu în ceea ce privește contribuţia lor la rezultate, respectiv la explicarea compoziţiei specifice a comunităţilor de păsări?

Rezolvare:

Solicităm o nouă analiză (Options), și trecem de ferestrele de opţiuni până la fereastra *Forward selection of environmental variables,* unde selectăm opţiunea *Automatic selection.*



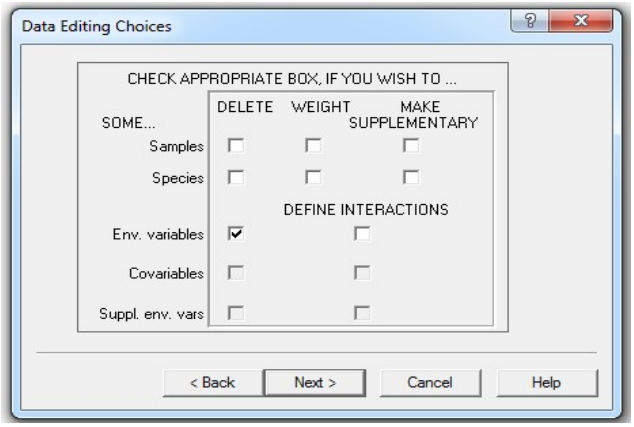
Dăm clic pe Next și apoi selectăm opţiunea *Unrestricted permutation* în fereastra de permutări. Urmează clic pe Finish apoi pe Analyze și vom selecta din ultima fereastră opţiunea ”FS summary” care va aduce un nou ecran, ca mai jos:



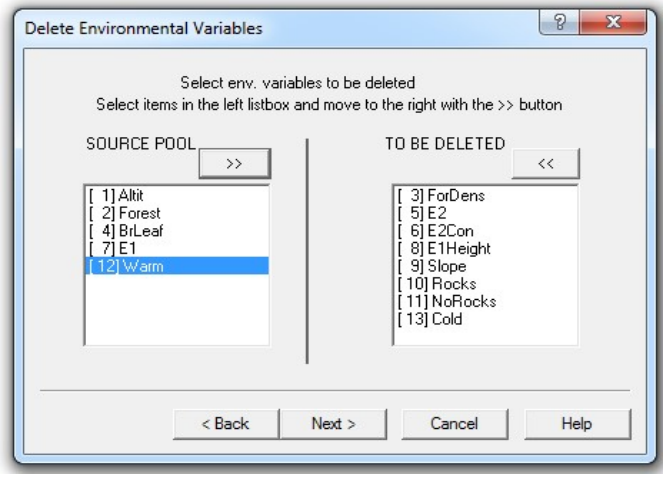
Vom da clic pe *Copy* și vom vizualiza (Ctrl+V sau Paste) rezultatele într-un document Word sau într-un fișier de foi de calcul, sau în Excel. Vom obţien tabelul de mai jos:

Marginal Effects			Conditional Effects				
Variable	Var.N	Lambda1	Variable	Var.N	LambdaA	P	F
Altit	1	0.11	Altit	1	0.11	0.002	5.32
BrLeaf	4	0.1	BrLeaf	4	0.08	0.002	3.72
Forest	2	0.09	Forest	2	0.06	0.002	2.91
E2Con	6	0.07	Warm	12	0.03	0.032	1.65
Slope	9	0.05	E1	7	0.03	0.038	1.71
E1	7	0.05	E2	5	0.03	0.066	1.54
Warm	12	0.05	E2Con	6	0.03	0.064	1.6
Cold	13	0.05	Slope	9	0.02	0.144	1.34
E1Height	8	0.05	ForDens	3	0.02	0.39	1.05
E2	5	0.04	E1Height	8	0.02	0.548	0.94
Rocks	10	0.04	Rocks	10	0.01	0.606	0.88
NoRocks	11	0.04					
ForDens	3	0.03					

Din acest tabel de rezultate vom vedea în coloana din stânga modul în care sunt ordonate variabilele de mediu în sensul descrescător al importanţei lor (Lambda1, de sus în jos) în ceea ce privește explicarea variaţiei în compoziţia de specii, iar în dreapta vom vedea modul de grupare al variabilelor în funcţie de îndepărtarea acumulativă a variaţiei explicate de variabilele de mediu, în ordinea inversă a importanţei lor. Când probabilitatea (P) a grupului de variabile în coloana de *Conditional Effects* depășește nivelul de asigurare ales, de obicei de 0.05, oprim modelul, respectiv selecţia aditivă de variabile, la grupul astfel delimitat. Aceasta înseamnă că variabilele care explică cel mai bine variaţia în compoziţia speciilor de păsări sunt, în ordine aditivă: Altit (altitudinea), la care se adaugă BrLeaf (frecvenţa relativă a foioaselor în arboret), Forest (acoperirea cu arbori), Warm (suprafeţe expuse spre soare, sud și direcţii asemănătoare) și E1 (acoperirea stratului ierbos). Prin adăugarea de alte variabile modelul nu mai are de câștigat, deoarece acestea sunt redundante. Prin urmare putem selecta numai aceste variabile, care alcătuiesc cel mai bun model de explicare a structurii comunităţilor de păsări și să reluăm graficul. Vom elimina celelalte variabile, prin întoarcerea la opţiunile anterioare: în fereastra de *Data Editing Choices* vom selecta *Delete Environmental Variables*, ca mai jos:

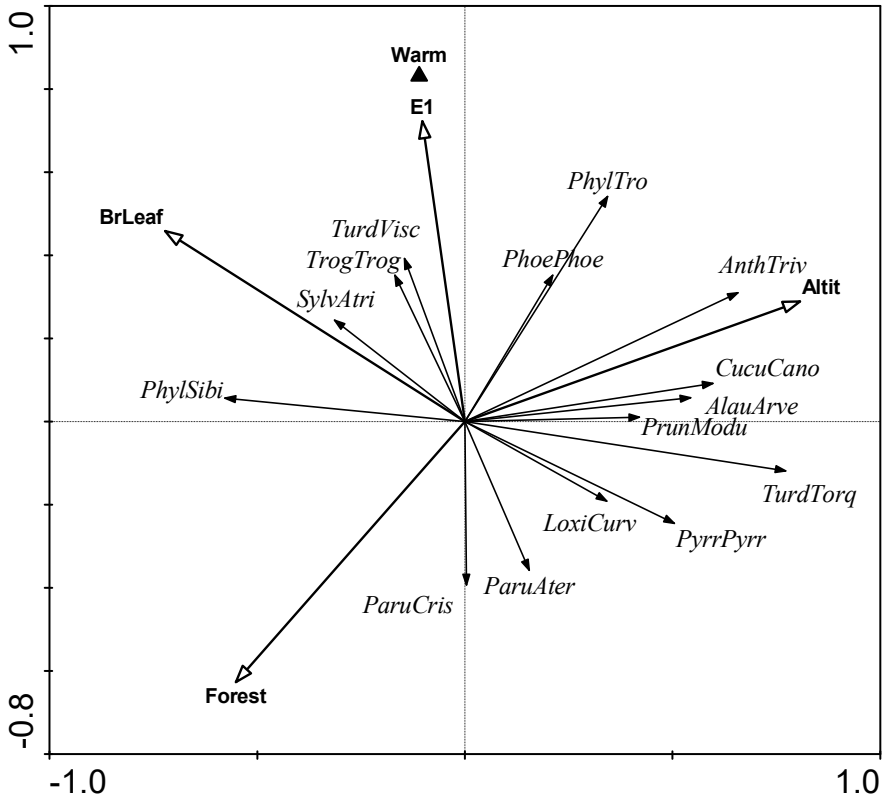


iar din următoarea fereastră vom selecta variabilele pe care dorim să le îndepărtăm, astfel:




```
**** Summary ****

Axes                                1       2       3       4  Total variance

Eigenvalues                  :     0.154   0.071   0.039   0.032          1.000
Species-environment correlations  :     0.877   0.774   0.709   0.715
Cumulative percentage variance
   of species data           :      15.4    22.5    26.4    29.6
   of species-environment relation:      49.7    72.7    85.3    95.5

Sum of all              eigenvalues                                    1.000
Sum of all canonical    eigenvalues                                    0.310
```

Rezultatele ne spun că dacă selectăm numai cele 5 variabile de mediu, acestea explică 31% din compoziţia specifică, faţă de toate variabilele considerate, care adaugă numai cca. 13% în plus. În studiul viitor, sau la continuarea unor studii similare, ne-am putea decide astfel să evaluăm numai grupul celor 5 variabile alese, care realizează cel mai bun și semnificativ model. Graficul, după ce vom explude speciile mai puţin reprezentate cantitativ, va arăta astfel:



Analiza directă a gradientului: efectul altitudinii

Odată realizate aceste analize, putem diversifica întrebările și interpretările datelor noastre. Am aflat că altitudinea este foarte importantă în explicarea compoziţiei specifice a avifaunei. Am putea dori să aflăm cât de importantă este această variabilă, dacă ar fi singura care ar acţiona și cum răspund comunităţile de păsări la celelalte variabile, dacă îndepărtăm varianţa explicată de altitudine?

Prima întrebare: cum se explică compoziţia comunităţilor de păsări dacă considerăm numai altitudinea medie a suprafeţelor de eșantionare, ca singura variabilă independentă?

Mod de lucru:

- salvăm proiectul original sub un nume nou (de exemplu rda1.con), utilizând din meniul principal File → Save as...→rda1.con
- acum trebuie realizate nişte modificări din *Options*. Ne asigurăm că în prima pagină (prima fereastră) alegem *extract patterns from explained variation only*, apoi alegem un nou nume pentru fișierul de soluţii (de exemplu rda1.sol), la *Type of Analysis* vom selecta *RDA*, după care vom lăsa selecţiile de pe celelalte pagini așa cum sunt până la *Data Editing Choices*. Aici vom selecta opţiunea de ștergere a variabilelor de mediu

(*Delete Env. variables*), iar în următoarea fereastră (*Delete Environmental Variables*) vom selecta toate variabilele, cu excepţia primei = altitudine (*Altit*) și le vom muta cu comanda (butonul) > > în coloana din dreapta ferestrei (*to be deleted*). În rest lăsăm selecţiile implicite și la fereastra de *Global Permutation Test* vom selecta oricare dintre teste, deoarece având o singură variabilă explicativă (altitudinea), vom avea implicit și un singur test pentru axa canonică unică astfel definită. În următoarele ferestre vom selecta *Unrestricted permutation* și *Finish.*

După ce vom realiza analiza (*Analyze*) rezultatul testului de permutări Monte Carlo va indica faptul că ipoteza nulă, care afirmă lipsa de relaţie între altitudine și compoziţia specifică, este respinsă la p = 0.002, prin urmare avem o relaţie semnificativă.

**** Summary of Monte Carlo test ****

Test of significance of all canonical axes : Trace = 0.115
F-ratio = 5.323
P-value = 0.0020

(499 permutations under reduced model)

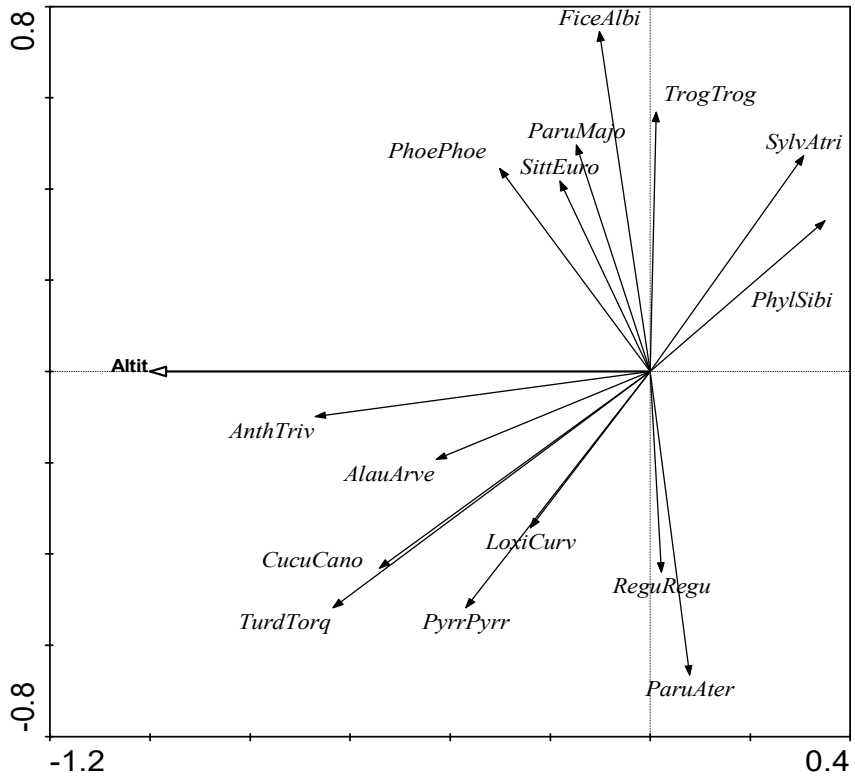
Rezultatele analizei arată ca mai jos:

**** Summary ****

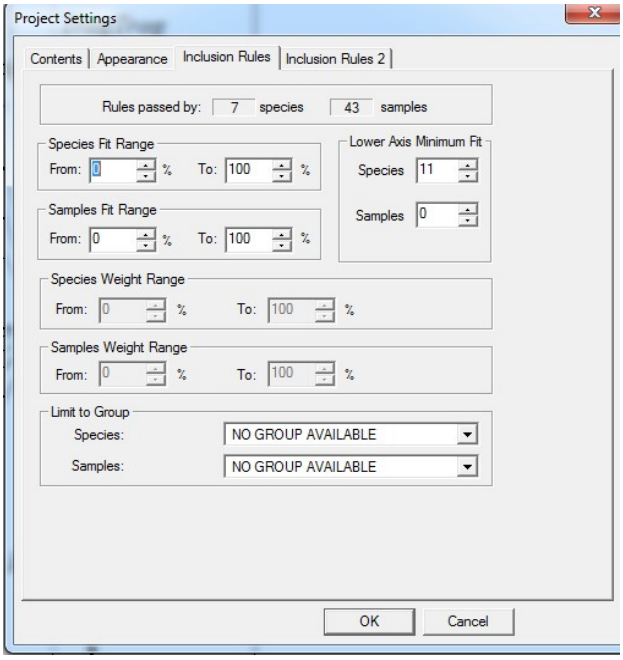
Axes		1	2	3	4	Total variance
Eigenvalues	:	0.115	0.152	0.118	0.075	1.000
Species-environment correlations	:	0.792	0.000	0.000	0.000	
Cumulative percentage variance						
of species data	:	11.5	26.7	38.5	46.0	
of species-environment relation:		100.0	0.0	0.0	0.0	

Sum of all eigenvalues		1.000
Sum of all canonical eigenvalues		0.115

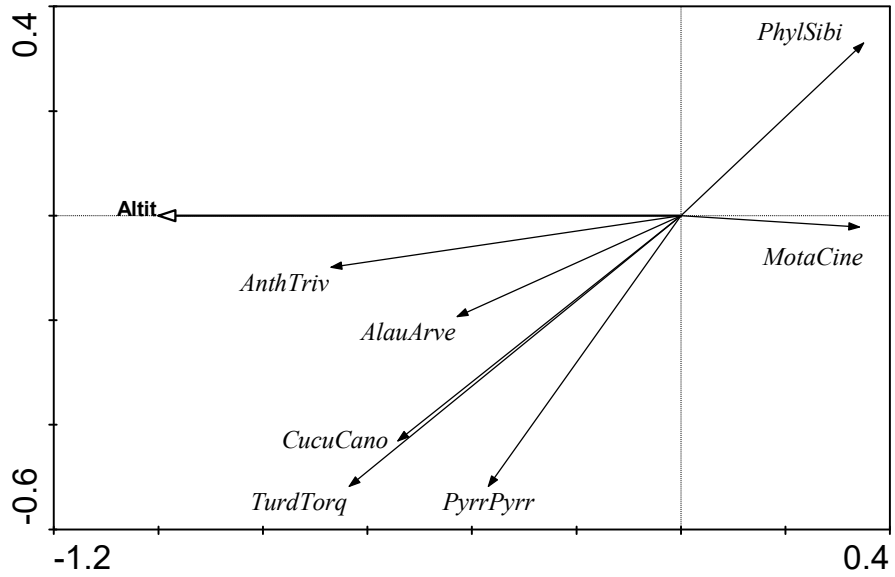
Înmulţind cu 100 valoarea rădăcinii latente a primei axe (deoarece există o singură variabilă, axa canonică și suma valorilor va fi aceeași) respectiv *sum of all canonical eigenvalues* = 0.115*100 = 11.5%, înseamnă că altitudinea explică 11.5% din variaţia compoziţiei în specii a comunităţilor de păsări. Ne aducem aminte că toate variabilele explicau 44.2%, prin urmare altitudinea este răspunzătoare de peste un sfert din variaţia speciilor, dar următoarele două axe (2 și 3) care nu sunt canonice (ne-constrânse) explică mai multă variaţie decât altitudinea (15.2% respectiv 11.8%).



În graficul de mai sus redăm biplotul specii-mediu din RDA care rezumă diferenţele în compoziţia comunităţilor de păsări de-a lungul gradientului altitudinal. Subliniem că acest grafic este de tip mixt (hibrid), în sensul că prima axă este canonică, în timp ce cea de-a doua nu. Se recomandă ca reprezentarea să includă numai speciile care sunt puternic corelate cu axa altitudinală (axa 1) iar celelalte să fie eliminate (adică toate cu unghiuri apropiate de 90^0 faţă de prima axă, sau apropiate de a doua). Deoarece altitudinea explică cca. 11% din variabilitatea datelor de specii, putem selecta numai speciile care sunt explicate în proporţie de cel puţin 11% de către această axă. Această condiţie sau limită se poate fixa în *CanoDraw* din fereastra de dialog *Project* → *Settings* → *Inclusion Rule* → după care vom selecta câmpul *Species* în *Lower Axis Minimum Fit*, deoarece numai axa inferioară (*lower*) este implicată. Numai şapte specii vor trece de această limită sau condiţie.



Graficul este redat mai jos (în meniul *Create* se va alege opţiunea de *Recreate graph)*



Analiză directă de gradient: efectul condiţional al celorlalte caracteristici de habitat

Vom termina acest studiu de caz cu o analiză parţială canonică mai avansată, care va răspunde la întrebarea: **putem să detectăm vreun efect semnificativ al cerlalţi descriptori de habitat, dacă îndepărtăm variabilitatea compoziţională explicată de către altitudinea medie a pătratului de probă?**

La această întrebare vom răspunde printr-o analiză RDA din care altitudinea va fi considerată ca și **covariată**.

Mod de lucru:

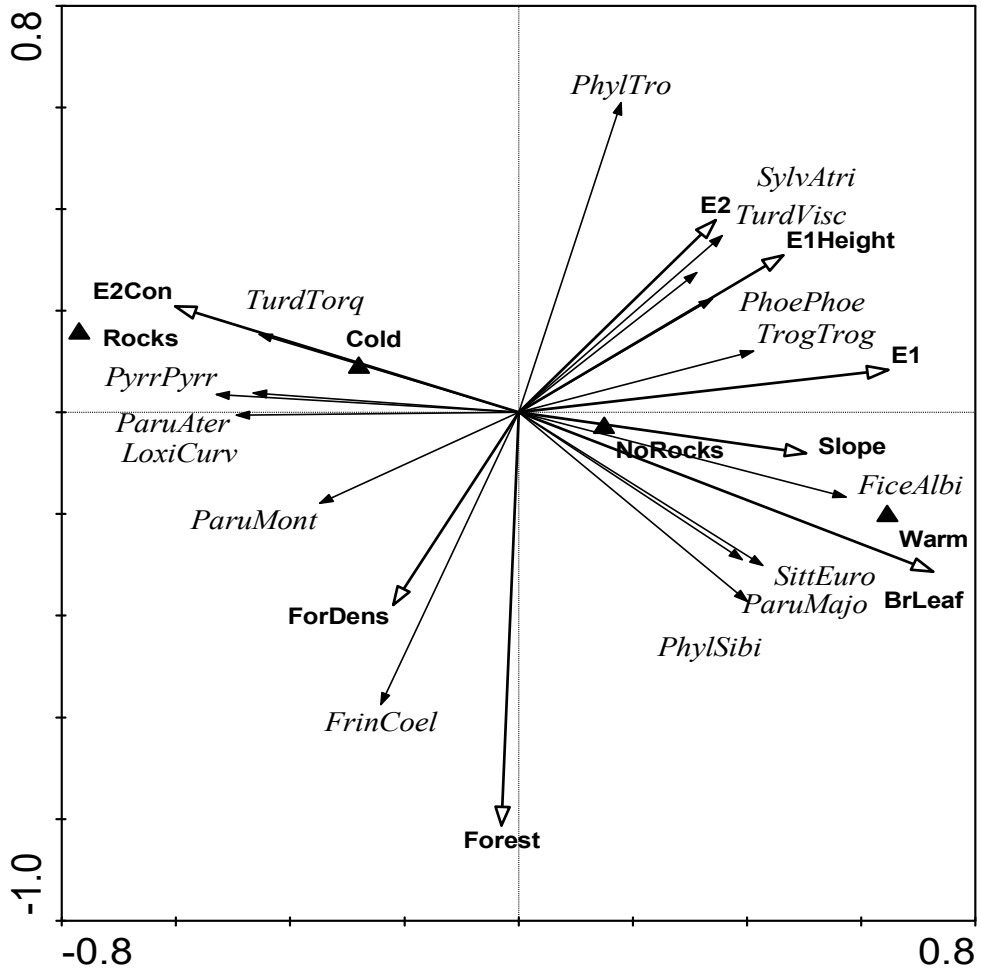
- *File → Save as* (alegem un nou nume de proiect, de exemplu rda2.con) → *clear the project log*;
- *Options* - iar din prima fereastră (*Available Data*) vom selecta opţiunea a treia: *Species, environmental and covariable data available → Next*
- În fereastra de definire a datelor (*Data Files*) va trebui să identificăm fișierul din care programul va alege covariatele. Deoarece acestea sunt tot în fișierul birds_env.dta, vom alege acest fișier, sau vom copia conţinutul câmpului *Environment data file name* la *Covariables data file name*.

Urmează *Next → RDA →*etc. neschimbat până la *Data Editing Choices*. Programul deja a selectat opţiunea de ștergere (*Delete*) a covariatelor (*Covariables*), iar după *Next*, vom selecta variabilele corespunzătoare.

- În fereastra *Delete Environmental Variables* vom șterge (muta în dreapta) numai altitudinea (*Altit*), în timp ce toate celelalte vor fi în stânga (variabilele 2 - 13).
- Next → *Delete covariables*: în această fereastră vom face exact invers. Ștergem (mutăm în dreapta) toate celelalte variabile exceptând altitudinea (variabila 1). Numai variabila *Altit* va rămâne în stânga. În următoarea fereastră (după Next) vom lăsa opţiunea selectată *Do not use forward selection*.
- În fereastra *Global Permutation Test* va fi deja selectată opţiunea *Significance of canonical axes together* (lăsăm ca atare), iar în următoarea lăsăm opţiunea *Unrestricted permutations* selectată. În final închidem proiectul pe următoarea pagină (*Finish*). Soicităm analiza (*Analyze*).
- În fereastra de rezultate sau în jurnal (Log View), vom obţine, alături de alte rezultate, următoarele valori:

```
**** Summary ****

Axes                                    1       2       3       4  Total variance

Eigenvalues                        :    0.112   0.067   0.042   0.034          1.000
Species-environment correlations   :    0.886   0.810   0.818   0.765
Cumulative percentage variance
   of species data                 :     12.6    20.3    25.0    28.8
   of species-environment relation:      34.2    54.7    67.4    77.8

Sum of all              eigenvalues                                            0.885
Sum of all canonical    eigenvalues                                            0.328


The sum of all eigenvalues is after fitting covariables
Percentages are taken with respect to residual variances
               i.e. variances after fitting covariables


All four eigenvalues reported above are canonical and correspond to axes that
are constrained by the environmental variables.
```

Se poate observa că variabilitatea explicată de ceilalţi descriptori de habitat, în adiţie faţă de informaţia adusă de altitudine, este destul de mare (32.8% din totalul variabilităţii în compoziţia specifică a comunităţilor de păsări), prin comparaţie cu variaţia explicată de altitudine (11.5%, vizibilă din acest tabel și ca diferenţa *Total variance - Sum of all eigenvalues = 1.000 - 0.885*). Dar este evident că modelul de faţă este mult mai complex ca număr de grade de libertate. Acest tabel de rezultate este urmat de raportul testului de permutări Monte-Carlo. Se poate vedea că contribuţia adiţională a celor 12 descriptori este extrem de semnificativă (p = 0.002).



RDA de tip biplot care redă efectul descriptorilor de habitat (al variabilelor de mediu) asupra compoziţiei comunităţilor de păsări, după ce a fost îndepărtat efectul gradientului altitudinal (altitudinea este considerată covariată).

Diagrama biplot specii-mediu care rezumă efectul adiţional al celorlalte caracteristici de habitat, după ce efectul altitudinii a fost luat în considerare, este redat mai jos. Nu uităm să declarăm variabilele nominale (*Project → Nominal variables →*

Environmental variables → selecție variabile substitutive sau nominale). Numai speciile cele mai bine reprezentate sau cu ponderi mai ridicate (*Project* → *Settings* → *Species Fit Range > 15*) au fost selectate aici, pentru a facilita interpretarea graficului.

Observăm că prima axă a acestui RDA reflectă parţial diferenţele dintre pătratele dominate de fag (pe partea dreaptă a graficului) și arboretele dominate de molid (pe partea stângă a diagramei), precum și dezvoltarea stratului ierbos (mai dezvoltat în jumătatea dreaptă a figurii). A doua axă de ordonare se corelează negativ cu acoperirea totală cu arbori (*Forest*) a pătratelor de probă și, într-o măsură mai mică, cu densitatea medie a pădurii (*ForDens*).