Mnohorozměrná analýza ekologických dat

Jan Lepš & Petr Šmilauer

**Překlad: Dana Vašková,
s následnými korekcemi autorů**

**Biologická fakulta
Jihočeské univerzity v Českých Budějovicích
České Budějovice, 2000**

Předmluva

Tato kniha je překladem starší verze studijního materiálu pro účastníky kurzu “Multivariate Analysis of Ecological Data”, který na naší univerzitě přednášíme již od roku 1997. Materiál zde předkládaný by měl sloužit jak pro kurz úvodní, tak pro jeho pokročilou variantu. Připouštíme, že některým kapitolám by prospělo vylepšení. V jeho brzké uskutečnění společně s vámi doufáme. Zatím tento překlad poskytne alespoň základní použitelné materiály pro českou verzi kurzu a pro posluchače kurzu "Plánování a hodnocení ekologických experimentů".

Chtěli bychom, aby vám tato kniha poskytla čtivější doplněk k přesnějším a podrobnějším publikacím, jako jsou například články Dr. Ter Braaka nebo manuál k programu Canoco for Windows 4.0. Vedle témat popisovaných ve jmenovaných publikacích se tento text věnuje také klasifikačním metodám mnohorozměrné analýzy dat a představuje moderní regresní metody užívané v ekologickém výzkumu.

Kdykoli se zmiňujeme o nějakých komerčních softwarových produktech, jsou implikovány příslušné obchodní nebo registrační značky vlastněné jejich výrobci.

Na kvalitě této publikace je znát, že to stále ještě není finální verze: některá témata jsou zde probírána opakovaně, nicméně doufáme, že to čtenáři přijmou jako vítanou možnost podívat se na stejné téma i z jiného úhlu.

Český překlad má svou specifickou problematiku, a tou je česká terminologie. Oba autoři ji ve vlastních pracech užívají minimálně, ale nebylo možné do českého textu jen přepisovat neskloňované anglické pojmy (alespoň u běžněji užívaných termínů). Některé překlady mohou znít dosti nezvykle, jsou vždy uvedeny nejprve s anglickým ekvivalentem. V předmluvě zdůrazníme jen nejfrekventovanější tři termíny: Pro termín *constrained* [*analysis, axis*] užíváme termín *omezená*, i když na rozdíl od delší alternativy *s omezením* má toto slovo občas nevhodný emotivní nádech. Pro termín specifický v daném kontextu programu CANOCO, a to *environmental variables* (zde znamená vysvětlující proměnné, na které zaměřujeme interpretační pozornost), jsme se nakonec přiklonili k termínu *charakteristika prostředí* (i když oba používáme spíše více slangový termín *environmentální proměnná*), naproti tomu ve spojení *environmental data* se slova *environmentální* držíme více, vzácně užíváme zde více matoucí *data o prostředí*. Termín *samples* (pro odběrové jednotky, popisné plochy, zkrátka pozorování) zvolila překladatelka termín *vzorky*, který v překladu ponecháváme.

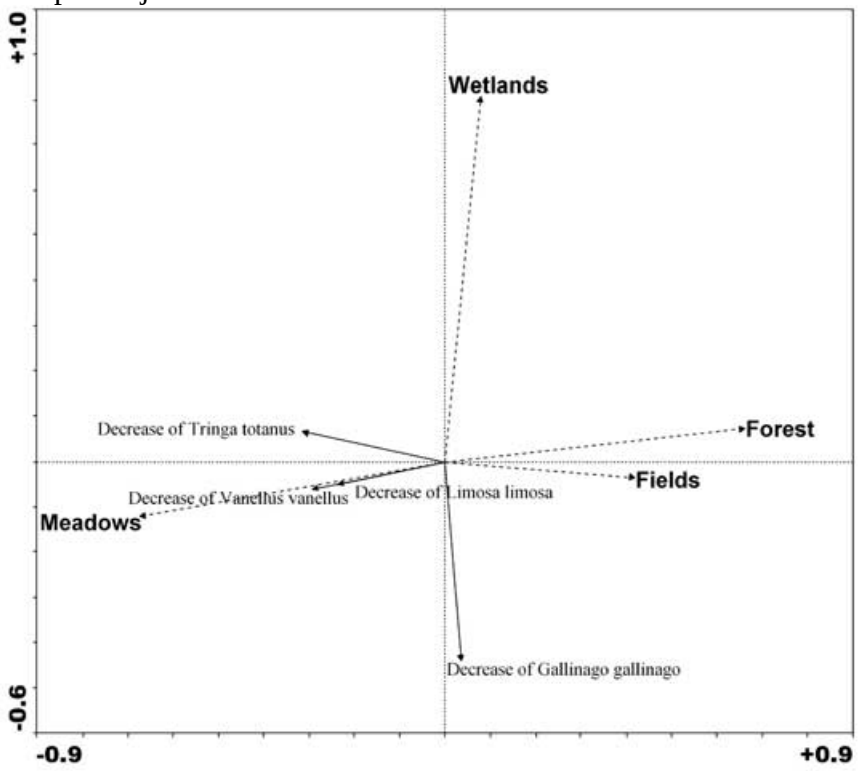
Obsah

1. Úvod a práce s daty

1.1. Příklady témat výzkumu

Metody mnohorozměrné statistické analýzy již nejsou omezeny jen na zpracování mnohorozměrných dat. Můžeme testovat složité hypotézy a během těchto analýz brát v úvahu velmi i komplexní uspořádání pokusů. Následují dva příklady, ve kterých se použití mnohorozměrné analýzy dat ukázalo jako mimořádne užitečné:

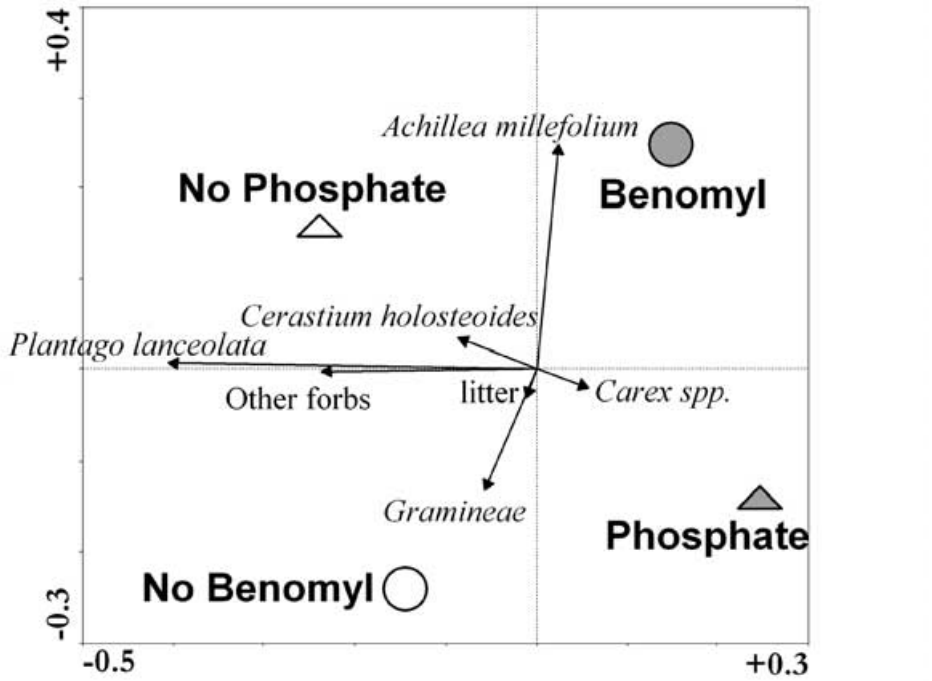
- Můžeme na základě aktuálního stavu krajiny předpovědět zánik hnízdiště ohroženého vodního ptáka? Které ze složek krajiny jsou pro tuto předpověď nejdůležitější?
  Následující diagram ukazuje výsledky statistických analýz, které na tyto otázky odpovídají:



Obrázek 1-1 Ordinační diagram, který ukazuje první dvě osy redundanční analýzy (RDA) údajů o hnízdních preferencích bahňáků.

  Diagram naznačuje, že tři studované ptačí druhy v krajině s vyšším procentem luk svou hnídní frekvenci snižují, zatímco čtvrtý druh (*Gallinago gallinago*) ustoupil do krajiny, kde bylo poslední dobou malé procento mokřadů. Když jsme ale významnost naznačených vztahů otestovali, tak se ukázalo, že žádný z nich nedosáhl statisticky významné úrovně.
  V tomto případě jsme sledovali závislost (semi-)kvantitativních vysvětlovaných proměnných (rozsah ústupu jednotlivých druhů ptáků) na procentickém zastoupení jednotlivých složek krajiny. Ordinační metoda zde představuje rozšíření regresní analýzy, kdy modelujeme současně odpověď několika proměnných.
- Jak reagují jednotlivé druhy rostlin na přídavek fosforu a/nebo na vyloučení mykorhizní symbiózy? Naznačuje odpověď společenstva interakci mezi těmito dvěma faktory?

Tento druh otázek býval tradičně řešen za použití nějakého typu analýzy variance (ANOVA). Její mnohorozměrné rozšíření pak dovoluje řešení podobných problémů pro více vysvětlovaných proměnných současně. Výsledky analýz současně ukazují i korelace mezi výskytem jednotlivých rostlinných druhů.



Obrázek 1-2 Ordinační diagram, který ukazuje první dvě osy redundanční analýzy (RDA) shrnující vliv fungicidu a fosforu na rostlinné společenstvo.

Tento ordinační diagram naznačuje snížení biomasy mnoha bylin ať už po aplikaci fungicidu (*benomyl*) nebo fosforu. Zdá se, že žebříček (*Achillea millefolium*) z aplikace fungicidu vytěžil, zatímco trávy na týž zásah odpovídají negativně. V tomto případě je efekt naznačený na diagramu podpořen i statistickým testem, který vede k zamítnutí nulové hypotézy na hladině významnosti $\alpha = 0.05$.

1.2. Terminologie

Terminologie mnohorozměrných statistických metod je poměrně komplikovaná, takže se jí musíme trochu věnovat. Běžně používaná terminologie je přinejmenším dvojí. První – obecnější a abstraktnější – obsahuje čistě statistické výrazy použitelné pro všechny vědní obory. V této části uvádíme takové výrazy převážně v závorkách a kurzívou. Druhá sada terminologických výrazů vychází z ekologické statistiky s nejběžnějšími příklady z ekologie společenstev. Toto jsou výrazy, na které se zde zaměříme a které jsou též užívány programem Canoco. Výrazy z první skupiny budeme používat jen k odkazům na obecnější statistickou teorii.

Ve všech případech máme soubor s **primárními daty**. V tomto souboru jsou obsažena pozorování – **vzorky (samples)** (*sampling units*)[+]. Každý vzorek zahrnuje hodnoty pro víc druhů, nebo eventuelně tzv. **charakteristiky prostředí (***environmental variables***,**

[+] Zde je terminogie značně nejednotná: v klasické statistice znamená **sample** soubor pozorovacích jednotek (sampling units), z populace zpravidla vybraných náhodně. V ekologii společenstev je však **sample** zpravidla používán pro popis pozorovací jednotky. Takto budeme toto slovo používat i my v tomto textu. Obecné statistické balíky užívají v témže významu slůvka **case**.

variables). Primární data mohou být ve formě obdélníkové matice, kde v řádcích bývají jednotlivé vzorky a ve sloupcích proměnné (druhy, chemické či fyzikální vlastnosti vody nebo půdy, atd.).

Naše primární data (obsahující vysvětlované – *response* - proměnné) jsou velmi často doprovázena dalším datovým souborem s vysvětlujícími (*explanatory*) proměnnými. Pokud primární data vyjadřují složení společenstva, bývají vysvětlujícími proměnnými velmi často půdní vlastnosti, semi-kvantitativní vyjádření vlivu lidské činnosti apod. Vysvětlující proměnné používané k předpovědi primárních dat (např. složení společenstva) můžeme rozdělit do dvou různých skupin. První skupinu nazýváme (někdy poněkud nevhodně) **charakteristikami prostředí** (*environmental variables*) a jsou v ní zahrnuty nejdůležitější a nejvlivnější proměnné. Ve druhé skupině jsou tzv. **kovariáty** (*covariables*, v jiných statistických publikacích zvány též *covariates*), které jsou taktéž vysvětlujícími proměnnými se známým (nebo alespoň předpokládaným) vlivem na proměnné vysvětlované (*response*). Jejich vliv však obvykle chceme oddělit, a to ještě před tím, než se zaměříme na důležitější proměnné.

Jako příklad si uvedeme situaci, kdy v určitém území sledujeme vliv vlastností půd a typu obhospodařování (senoseč nebo pastva) na druhovou skladbu luk. V jedné z analýz nás může zajímat jen vliv půdních vlastností, zatímco typu obhospodařování si vůbec všímat nechceme. Pak tedy použijeme složení louky jako **druhová (species) data** (tj. primární soubor dat), kde jednotlivé rostlinné druhy vystupují jako jednotlivé vysvětlované (*response)* proměnné a naměřené vlastnosti půd jako **charakteristiky prostředí** (vysvětlující proměnné – *explanatory variables*). Na základě výsledků pak můžeme odvodit závěry o preferencích jednotlivých rostlinných populací ve vztahu ke gradientům prostředí popsaným (více či méně přesně) naměřenými vlastnostmi půdy. Podobně se můžeme ptát i na vliv obhospodařování. Ten v tomto případě vystupuje jako charakteristika prostředí. Lze očekávat, že typ obhospodařování bude vlastnosti půdy také ovlivňovat. Na základě tohoto předpokladu se můžeme ptát na vliv obhospodařování **kromě** vlivu, jenž je způsobován změnou půdních vlastností. Abychom na tuto otázku dostali odpověď, použijeme jako charakteristiku prostředí popis typu obhospodařování a jako *kovariáty* naměřené půdní vlastnosti.

Pro pochopení terminologie používané programem Canoco je klíčové uvědomit si, že data, která jsou v programu nazývána **druhová data** (*species data*)**,** mohou být ve skutečnosti jakýmikoli daty s proměnnými, jejichž hodnoty chceme předpovídat. Tak pokud bychom chtěli kupříkladu předpovídat obsah různých iontů kovů v říční vodě na základě skladby krajiny v povodí řeky, pak by v terminologii programu Canoco vystupovaly koncentrace jednotlivých iontů jako jednotlivé "druhy". V případě, že druhová data opravdu vyjadřují druhové složení společenstva, používáme obvykle pro jejich vyčíslení různá měřítka, včetně počtů, odhadů pokryvnosti a odhadů biomasy. Jindy můžeme mít k dispozici pouze informaci o přítomnosti či nepřítomnosti jednotlivých druhů. Také mezi vysvětlujícími poměnnými (tento termín je v programu Canoco používán jak pro charakteristiky prostředí, tak pro kovariáty) můžeme mít proměnné kvantitativní nebo typu přítomen / nepřítomen. S těmito různými typy dat se seznámíme podrobněji v jiné části této kapitoly.

1.3. Analýzy

Pokud se snažíme modelovat hodnoty jedné nebo více vysvětlovaných proměnných, závisí výběr vhodné statistické metody na tom, zda modelujeme každou vysvětlovanou proměnnou zvlášť a zda máme při tvorbě modelu k dispozici vysvětlující proměnné (prediktory).

Následující tabulka shrnuje nejdůležitější statistické metody používané v různých situacích:

Vysvětlovaná proměnná ...	Prediktor(y)	
	Nemáme	**Máme**
... je jedna	• shrnutí distribučních vlastností	• regresní model s.l.
... je jich více	• nepřímá gradientová analýza (indirect gradient analysis - PCA, DCA, NMDS) • klastrová analýza	• přímá gradientová analýza • omezená klastrová analýza • diskriminační analýza (discriminant analysis - CVA)

Tabulka 1-1 Typy statistických modelů

Pokud se chceme podívat jen na jednu vysvětlovanou proměnnou a nemáme k dispozici žádné prediktory, tak zmůžeme stěží něco víc než shrnutí jejích distribučních vlastností. U mnohorozměrných dat můžeme použít buď ordinační přístup představovaný metodami **nepřímé gradientové analýzy (indirect gradient analysis,** z nichž nejvýznamnější jsou analýza hlavních komponent - *principal components analysis, PCA*, detrendovaná korespondenční analýza - *detrended correspondence analysis*, DCA, a nemetrické mnohorozměrné škálování - *nonmetric multidimensional scaling*, NMDS) nebo se můžeme pokusit naše vzorky (hierarchicky) rozdělit do kompaktních oddělených skupin metodami klastrové analýzy s.l. – viz kapitola 5.

Pokud máme k dispozici jeden nebo více prediktorů a modelujeme-li očekávané hodnoty jediné vysvětlované proměnné, můžeme použít **regresní modely** v širším slova smyslu, tedy jak včetně tradičních regresních metod, tak včetně analýzy variance (*analysis of variance*, ANOVA) a analýzy kovariance (*analysis of covariance* , ANOCOV). Tato skupina metod je spojena do tzv. **obecných (general) lineárních modelů** a byla dále rozšířena a vylepšena metodami **zobecněných lineárních modelů (GLM)** a **zobecněných aditivních modelů (GAM)**. Další informace na toto téma jsou v kapitole 11.

1.4. Vysvětlovaná (druhová) data

Naše primární data (často zvaná podle nejtypičtějšího obsahu dat z biologických společenstev daty **druhovými**) lze často určovat poměrně přesným (kvantitativním) způsobem. Příkladem může být suchá váha nadzemní biomasy rostlin, počty jedinců jednotlivých druhů hmyzu chycených do zemních pastí nebo procentická pokryvnost jednotlivých vegetačních typů v určité krajině. Různé hodnoty můžeme porovnávat nejen prostým "větší než", "menší než" nebo "rovno", ale také použitím poměrů ("tato hodnota je dvakrát větší naž tato").

V ostatních případech odhadujeme hodnoty primárních dat podle jednoduché semi-kvantitativní stupnice. Dobrým příkladem je určování skladby rostlinných společenstev, kde se užívá např. původní Braun-Blanquetova stupnice nebo její nejrůznější modifikace. Nejjednoduší variantou takových odhadů jsou data typu přítomen / nepřítomen (0 / 1).

Pokud studujeme vliv různých faktorů na chemické či fyzikální prostředí (popsané například koncentrací různých iontů nebo složitějších sloučenin ve vodě, kyselostí půdy, teplotou vody, apod.), dostáváme obvykle kvantitativní odhady s jedním dodatečným omezením: tyto charakteristiky nejsou vyjádřeny ve stejných jednotkách. Tento fakt

vylučuje použití unimodálních ordinačních metod a určuje způsob standardizace dat pro jejich případné použití v metodách lineární ordinace.

1.5. Vysvětlující proměnné

Vysvětlující proměnné (někdy zvané též prediktory) poskytují informaci, kterou můžeme použít k předpovězení hodnot vysvětlovaných proměnných. Můžeme se například pokusit předpovědět složení rostlinného společenstva na základě znalosti půdních charakteristik a typu obhospodařování. Předpovídání samo obvykle nebývá hlavním cílem. Mnohem cennější jsou "předpovídací pravidla" (*prediction rules*) odvozená v případě ordinačních metod z ordinačních diagramů, která můžeme použít k tomu, abychom rozšířili svoje znalosti o studovaných organismech nebo systémech.

Prediktory mohou být proměnné kvantitativní (jako je koncentrace dusíkatých iontů v půdě), semi-kvantitativní (např. stupeň ovlivnění člověkem odhadovaný na stupnici od 0 do 3) nebo faktoriální (kategoriální proměnné).

Faktory jsou přirozeným způsobem, jak vyjádřit zařazení našich vzorků. Pro louky můžeme mít třídy typů obhospodařování, pro studium znečištění řek typy toků, nebo indikátor přítomnosti či nepřítomnosti osídlení v blízkosti. Chceme-li však faktory používat v Canocu, musíme je překódovat do tzv. **indikátorových proměnných** (*dummy variables*, též *indicator variables*). Pro každou hladinu faktoru (tj. pro všechny různé hodnoty) existuje zvláštní proměnná. V případě, že určitý vzorek (pozorování) má určitou hodnotu faktoru, nabývá příslušná indikátorová proměnná hodnoty 1.0. Všechny ostatní hodnoty tvořící faktor nabývají hodnoty 0.0. Ukážeme si to ještě na příkladu: představme si, že pro každý vzorek z travinného společenstva zaznamenáváme, jestli byl z pastviny, z louky nebo z opuštěné louky. Pro takový faktor potřebujeme tři indikátorové proměnné a pro louku pak budou příslušné hodnoty 0.0, 1.0 a 0.0.

Toto rozkládání faktorů do indikátorových proměnných umožňuje navíc využití tzv. **fuzzy kódování**, které si vysvětlíme na předchozím příkladu. Do pokusu jsme teoreticky mohli zahrnout plochu, která byla do loňského roku využívána jako senosečná louka, ale letos již jako pastvina. Je velmi pravděpodobné, že na současné složení rostlinného společenstva měly vliv oba druhy obhospodařování. Takže oběma indikátorovým proměnným můžeme přidělit nějakou hodnotu mezi 0.0 a 1.0. Jedinou podmínkou je, že součet těchto hodnot musí být opět 1.0. Pokud neumíme relativní důležitost aplikovaných typů obhospodařování kvantifikovat, je nejlepším odhadem použití hodnot 0.5, 0.5 a 0.0.

Pokud vytváříme model, ve kterém chceme pomocí vysvětlujících proměnných (*environmental data*) předpovídat hodnoty vysvětlovaných proměnných (*species data*), můžeme často narazit na situaci, kdy určitá vysvětlující proměnná má sice prokazatelný vliv na druhová data, ale pro nás je z různých příčin nezajímavá. Její vliv nechceme interpretovat, chceme ho jen vzít v úvahu při hodnocení vlivu jiných proměnných. Takovým proměnným říkáme **kovariáty** (*covariables*, nebo často také *covariates*). Typickým příkladem jsou data z pokusu uspořádaného do logických nebo fyzických bloků. Hodnoty vysvětlovaných proměnných celé skupiny ploch mohou být podobné čistě díky tomu, že spolu tyto plochy sousedí. Tento vliv potřebujeme do našeho modelu zahrnout a počítat s ním při zpracování našich dat. Snažíme se oddělit ("partial-out") odlišnosti vysvětlovaných proměnných vzniklé dík příslušnosti vzorků do různých bloků.

Vpodstatě může ale kovariátou být jakákoli vysvětlující proměnná – například ve studii vlivu typu obhospodařování na společenstvo motýlů můžeme mít jednotlivé lokality umístěné v různé nadmořské výšce. Nadmořská výška může mít na populace motýlů veliký

vliv, ale v této situaci se zaměřujeme jen na vliv typu obhospodařování, který uvidíme mnohem zřetelněji, dokážeme-li se efektu nadmořské výšky zbavit.

1.6. Co s chybějícími údaji

Jakkoliv jsme opatrní, může se přihodit, že nám některá data budou chybět. Vzorek zaslaný k analýze do laboratoře se může ztratit, stejně tak, jako můžeme zapomenout vyplnit příslušnou kolonku v datovém souboru.

Velmi často už není možné vrátit se zpátky a chybějící údaj prostě doplnit. Hlavně proto, že studované předměty se časem mění. Můžeme se pokusit nechat ono místo prázdné, ale ani to nebývá tím nejlepším řešením. Například při sběru dat ze společenstev (kde máme celkem, řekněme, 300 druhů, ale průměrný počet druhů na ploše je výrazně menší) používáme prázdná místa v tabulce pro nepřítomnost, tj. jako hodnotu nula. Jenže nepřítomnost druhu je přeci výrazně jiným případem, než když hodnotu zapomeneme vyplnit! Některé statistické programy chybějící údaj nějakým způsobem označují (například písmeny „NA“), ale to je stále jen vylepšení týkající se označení a nevypovídá o tom, jak se s nimi aktuální statistická metoda vypořádá. Nabízí se několik řešení:

Vzorky, v nichž hodnoty chybějí, můžeme prostě vypustit. To se vyplatí zejména tehdy, pokud jsou chybějící data zakoncentrována v několika málo vzorcích. Máme-li kupříkladu datový soubor s 30 proměnnými a 500 vzorky, kde 20 chybějících hodnot pochází pouze ze tří vzorků, bude zřejmě moudré tyto tři vzorky před analýzou vypustit. Tuto strategii často používají obecné statistické balíky a obvykle jí říkají „case-wise deletion“.

Pokud naopak patří chybějící hodnoty k několika málo proměnným, bez nichž se obejdeme, můžeme je z našeho datového souboru také vypustit. Taková situace nastává často v datech z chemických analýz. Když určujeme koncentrace téměř všech možných iontů, zjistíme, že je mezi nimi obvykle velmi silná korelace. Tak ze znalosti koncentrace iontů kadmia v usazeninách můžeme obvykle rozumně odhadnout koncentraci iontů rtuti (ačkoli to závisí na druhu zdroje znečištění). Silná korelace mezi těmito dvěma charakteristikami pak naznačuje, že bychom mohli vystačit jen s jednou z nich. A tak pokud bychom měli hodně chybějících hodnot právě v koncentracích kadmia, bylo by nejlepší tuto proměnnou z datového souboru vypustit.

Jmenované dvě metody vypořádání se s chybějícími údaji se mohou zdát poměrně drastické. Vždyť se při nich vzdáváme tolika výsledků, jejichž sběr byl často velmi nákladný. Proto existují také nejrůznější „doplňovací“ (imputation) metody. Nejjednodušší z nich je doplnění průměru na místa prázdných polí (průměr však musí být samozřejmě počítán jen ze vzorků, pro něž jsou hodnoty k dispozici). Jiným, důmyslnějším, přístupem je vytvoření mnohonásobného regresního modelu za použití vzorků bez chybějících hodnot k odhadnutí chybějících hodnot vysvětlované proměnné ve vzorcích, kde hodnoty vybraného prediktoru nechybí. Tak můžeme vyplnit mezery v naší tabulce, aniž bychom museli vypustit jediný vzorek nebo proměnnou. Musíme si ale přiznat, že se vlastně podvádíme – jen duplikujeme informaci, kterou již máme. Stupně volnosti, o které takto přijdeme, již nikdy zpět nezískáme. Když potom analyzujeme takto doplněná data nějakým statistickým testem, má tento test naprosto mylnou představu o skutečných stupních volnosti (o počtu nezávislých pozorování v datech). Odhady hladiny významnosti pak nejsou zcela přesné (jsou „příliš optimistické“). Tento problém můžeme částečně zmírnit tím, že vzorkům, pro než jsme hodnoty dopočítávali, přidělíme tím či oním způsobem menší statistickou váhu. Výpočet lze provést poměrně jednoduše: v souboru s 20 proměnnými dostane vzorek s pěti

doplněnými hodnotami váhu 0.75 (=1.00 - 5/20). Nicméně ani toto řešení není zcela ideální, protože i když pracujeme jen s výběrem proměnných (jako např. při postupném výběru vysvětlujících proměnných), mezi kterými ony doplňované již vůbec být nemusí, jsou vzorky, pro něž bylo cokoli dopočítáváno, stále penalizovány.

1.7. Import dat z tabulek - program CanoImp

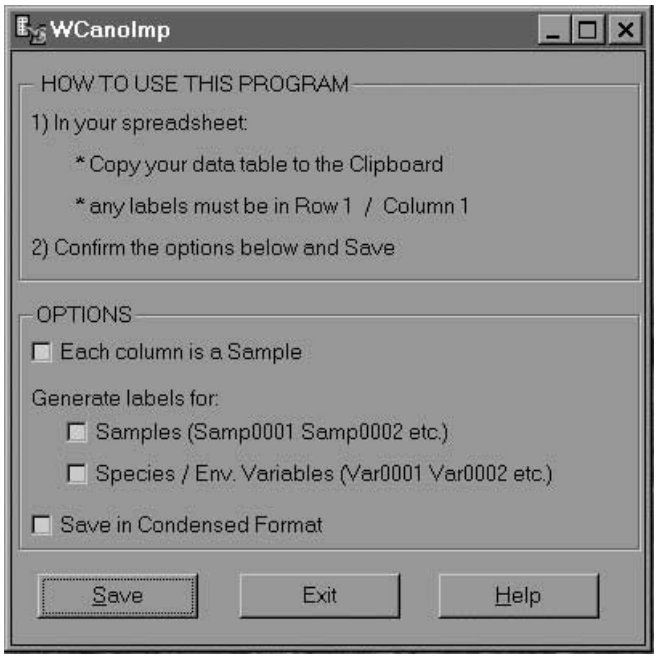
Příprava vstupních dat pro mnohorozměrné analýzy byla vždy největší překážkou jejich efektivního využití. Ve starších verzích Canoca bylo nutné rozumět velmi složitému a nesmlouvavému formátu, který byl důsledkem požadavků programovacího jazyka FORTRAN, ve kterém byl program Canoco napsán. Verze Canoco 4.0 zmírňuje tento problém hned dvěma alternativními způsoby. Jednak je k dispozici jednodušší formát dat s minimálními požadavky na obsah souboru, a navíc (a to je zřejmě důležitější) existuje nová a snadná cesta transformace dat uložených v tabulkách do přesného formátu Canoca. V této části si ukážeme, jak pro tento účel využít program WCanoImp.

Musíme začít s daty ve vašem tabulkovém programu. Ačkoli většina uživatelů zřejmě bude používat program Microsoft Excel, je popsaný postup použitelný i pro jakýkoli jiný tabulkový program spustitelný pod Microsoft Windows. Pokud máte data uložena v nějaké relační databázi (Oracle, FoxBase, Access, atd.), využijeme nejdřív schopnoti tabulkového programu načíst (importovat) tato data. V tabulce pak musíme data uspořádat do obdélníkové matice. V obvyklém rozložení odpovídají jednotlivé řádky vzorkům a sloupce proměnným. Navíc máme pro řádky i sloupce jednoduché hlavičky: první řádek obsahuje (krom prvního prázdného pole) jména proměnných a první sloupec označení jednotlivých vzorků. Použití hlaviček není povinné, WCanoImp je schopen jednoduchá jména vygenerovat. Pokud hlavičky používáme, musíme respektovat určitá omezení programu Canoco. Jména nemohou být delší než 8 znaků a rovněž výběr znaků je poněkud omezen: nejjistější strategií je použít pouze jednoduchá písmena anglické abecedy, číslice, spojovníky a mezery. WCanoImp sice zakázané znaky nahrazuje tečkou a jména delší než 8 znaků zkracuje, v takovém případě ale hrozí, že přijdeme o jednoznačnost (a tedy interpretovatelnost) našich jmen, a proto je lépe s tímto omezením počítat již od samého začátku.

Zbývající pole tabulky musí být vyplněna pouze čísly (celými nebo desetinnými), nebo musí být prázdná. Kódování jiným způsobem není přijatelné. Kvalitativní proměnné (faktory) musí být pro Canoco kódovány systémem indikátorových proměnných (viz oddíly 1.5 a 2.6).

Jakmile je naše tabulka v tabulkovém programu hotova, vybereme tuto obdélníkovou matici (např. myší) a zkopírujeme ji do paměti (Windows Clipboard). WCanoImp si ji z Clipboardu převezme, určí její vlastnosti (rozsah hodnot, počet desetinných míst, atd.) a umožní nám vytvořit nový soubor s těmito hodnotami, který odpovídá požadavkům programu Canoco. Doufejme, že nyní je jasné, že omezení popsaná výše, se vztahují pouze na formát dat obsažených v oblasti kopírované do Clipboardu. Mimo tuto oblast můžeme umístit jakoukoli hodnotu, graf nebo objekt.

Poté, co umístíme data do Clipboardu, musíme spustit program WCanoImp. Lze to učinit z programové nabídky Canoco for Windows (**Start/Programs/**[složka *Canoco for Windows*]). Tato importovací utilita má velmi jednoduché uživatelské rozhraní, reprezentované především jedním dialogovým oknem:



Obrázek 1-3 Hlavní okno programu WCanoImp.

V horní části dialogového okna je zkrácená verze pokynů popsaných v tomto materiálu. Když máme data v Clipboardu, musíme zkontrolovat další nastavení WCanoImp, zda odpovídají našim požadavkům. První volba (**Each column is a Sample**) se týká jen situace, kdy máme matici transponovanou vůči výše popsanému formátu. Může to být užitečné, když nemáme mnoho vzorků, ale zato máme mnoho proměnných (např. MS Excel omezuje počet sloupců na 256). Pokud nemáme jména vzorků v prvním sloupci, musíme zaškrtnout druhé políčko (**Generate labels for:** ... **Samples**), podobně pokud v prvním řádku jsou přímo hodnoty z prvního vzorku a nikoliv jména proměnných, zaškrtneme třetí políčko. Poslední políčko (**Save in Condensed Format**) řídí skutečný formát, který se při tvorbě souboru použije. Pokud nejsme omezeni místem na pevném disku, je vcelku jedno, co vybereme (výsledky statistických analýz by měly být stejné).

Poté, co jsme zkontrolovali nastavení voleb, můžeme pokračovat stiskem tlačítka **Save**. Teď vybereme jméno souboru a jeho umístění ve struktuře existujících adresářů. WCanoImp pak požaduje jednoduchý popis (jeden řádek ASCII textu) zpracovávaných dat. Tento popis se posléze objevuje ve výstupu z analýz a připomíná nám, s jakými daty jsme pracovali. V případě, že nám na tomto popisu nesejde, je nabízen neutrální text. WCanoImp pak soubor uloží a o úspěšném provedení nás informuje jednoduchým dialogovým oknem.

1.8. Plný formát datových souborů pro Canoco

V předchozí části jsme si ukázali, jak jednoduché je vytvořit soubor v Canocu z dat v tabulce. V ideálním světě by nás vůbec nemuselo zajímat, jak vlastně soubory vytvořené programem WCanoImp vypadají. Uživatelé Canoca však bohužel často v ideálním světě nežijí. Někdy nelze tabulky použít, a pak nezbývá nic jiného, než se obejít bez pomoci WCanoImp. To se může přihodit například v situaci, kdy máme víc než 255 druhů a 255 vzorků současně. V tomto případě je jednoduchá metoda, kterou jsme si popsali, nepoužitelná. Pokud lze vytvořit soubor s hodnotami oddělenými tabelátory (*TAB-separated values format*), můžeme využít z příkazové řádky ovládanou alternativu programu WCanoImp zvanou CanoImp, která dokáže zpracovat data o značně větším rozsahu než je oněch 255 sloupců. I WCanoImp je ve skutečnosti schopen pracovat s více sloupci, takže

pokud máte tabulkový program, který podporuje větší počet sloupců, máte vyhráno a můžete zůstat klidně ve sféře uživatelsky příjemnějšího rozhraní programu Windows (větší počet sloupců než Microsoft Excel podporoval např. program Quattro for Windows).

Ale v ostatních případech musíme vytvořit soubory pro Canoco buďto „od ruky“ nebo si musíme napsat programy, které zvládají konverzi dat z formátu uživatele do formátu Canoca. V těchto případech je nezbytné znát pravidla určující obsah datových souborů. Začneme nejdříve specifikací tzv. plného formátu (**full format**).

```
WCanoImp produced data file
(I5,1X,21F3.0)
21
----1---1--1--0--1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    2   1  0  0  1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    3   0  1  0  1  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
...
   48   1  1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
    0   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
PhosphatBenlate Year94  Year95  Year98  B01     B02     B03     B04     B05
B06     B07     B08     B09     B10     B11     B12     B13     B14     B15
B16
PD01    PD02    PD03    PD04    PD05    PD06    PD07    PD08    PD09    PD10
PD11    PD12    PD13    PD14    PD15    PD16    C01     C02     C03     C04
...
```

Obrázek 1-4 Část datového souboru Canoca v plném formátu. Pomlčky v prvním řádku dat ukazují přítomnost mezer a ve skutečném souboru se neobjevují.

První tři řádky mají podobný vzhled jak v plném, tak v kondenzovaném formátu. V první řádce je krátký textový popis datového souboru o maximálně 80 znacích. Ve druhé řádce je přesný popis formátu pro hodnoty dat, které se v souboru objevují (počínaje čtvrtým řádkem). Tato řádka je podrobněji popsána v části 1.10. Ve třetí řádce je jedno jediné číslo, ale jeho význam se mezi plným a kondenzovaným formátem liší. V plném formátu vypovídá o celkovém počtu proměnných v matici.

Obecně vzato – soubor v plném formátu obsahuje celou matici dat, včetně nulových hodnot. Je tedy srozumitelnější, když se na něj podíváme, ale je mnohem zdlouhavější a pracnější jej vytvořit (většina hodnot pro společenstva bývá nula).

V plném formátu je každý vzorek zastoupen neměnným počtem řádků – ve vzorové ukázce to je jeden řádek na vzorek. První vzorek (na čtvrtém řádku) začíná svým číslem (**1**), po němž následuje 21 hodnot (máme-li 21 proměnných). Podotýkáme, že počet mezer mezi hodnotami je pro všechny řádky stejný, pole dat je na svém pravém okraji zarovnáno. V každém poli je počet pozic („sloupců“) přesně určen číslem ve formátovém řádku. Pokud se všechny proměnné do jednoho řádku (kde je maximálně 127 sloupců) nevejdou, můžeme použít pro každý vzorek řádek (řádky) další. To se potom vyznačí v popisu formátu ve formátovém řádku lomítkem. Po posledním vzorku následuje jakýsi „falešný“ vzorek, jehož číslo je nula.

Pak následují jména („labels“) proměnných, jejichž formát je velmi striktní: každé jméno má přesně 8 znaků (je-li třeba, zleva či zprava doplněných mezerami) a je jich přesně 10 na řádku (mimo řádku posledního, který nemusí být vyplněn až do konce). Všimněte si, že počet požadovaných položek odpovídá počtu proměnných, který je uveden ve třetím řádku plného formátu. V našem případě máme tedy dva úplné řádky jmenovek, po kterých následuje třetí jen s jedním jménem.

Jména vzorků následují za blokem se jmény proměnných. Jejich počet je určen nejvyšším číslem vzorku obsaženého v souboru. Dokonce i když některé položky mezi 1 a tímto nejvyšším číslem chybí, příslušné pozice pro ně musí být v bloku jmen vymezeny.

Měli bychom poznamenat, že použití tabelátoru (TAB) v datovém souboru není vhodné – tabelátor je Canocem počítán jako jeden znak, i když ve většině textových editorů je reprezentován několika mezerami. Dále podotýkáme, že pokud tvoříte datový soubor „od ruky“, neměli byste používat žádný editor, který do dokumentu vkládá informace o formátování (jako Microsoft Word nebo Wordperfect). Notepad je tím nejjednodušším programem, jaký pro psaní dat ve formátu programu Canoco můžete najít.

1.9. Kondenzovaný formát dat pro Canoco

Tento formát je nejvýhodnější pro "řídká" data ze společenstev organismů. Soubor v tomto formátu obsahuje pouze nenulové hodnoty. Každá hodnota musí být proto doplněna indexem, který určí, ke které proměnné se vztahuje.

```
WCanoImp produced data file
(I5,1X,8(I6,F3.0))
8
----1-----23--1----25-10----36--3    41  4    53  5    57  3    70  5    85  6
    1     89 70   100  1   102  1   115  2   121  1
    2     11  1    26  1    38  5    42 20    50  1    55 30    57  7    58  5
    2     62  2    69  1    70  5    74  1    77  1    86  7    87  2    89 30
...
   79    131 15
    0
TanaVulgSeneAquaAvenPratLoliMultSalxPurpErioAnguStelPaluSphagnumCarxCaneSalxAuri
...
SangOffiCalaArunGlycFlui
PRESEK  SATLAV  CERLK   CERJIH  CERTOP  CERSEV  ROZ13   ROZ24   ROZC5   ROZR10
...
```

Obrázek 1-5 Část datového souboru Canoca v kondenzovaném formátu. Pomlčky v prvním řádku dat ukazují přítomnost mezer a ve skutečném souboru se neobjevují.

V tomto formátu se počet řádků potřebných na zaznamenání hodnot liší vzorek od vzorku. Každý řádek tedy musí začínat indexem vzorku a také formátový řádek popisuje formát jednoho řádku. V příkladě na obrázku Obrázek 1-5, jsou pro první vzorek použity dva řádky a tento vzorek obsahuje 13 druhů. Například druh s indexem 23 má hodnotu 1.0 a hodnota druhu s indexem 25 je 10. Vyhledáním druhu s největším indexem zjistíme, že celkem je v datech informace o 131 druzích. Hodnota na třetím řádku toto číslo neurčuje – zde, v souboru v kondenzovaném formátu, obsahuje informaci o maximálním počtu dvojic (index proměnné – hodnota proměnné) na řádku. Po posledním vzorku následuje opět vzorek „falešný“ s indexem rovným nule. Formát bloků se jmény proměnných a označením vzorků je stejný jako v souboru v plném formátu.

1.10. Formátový řádek

Následující příklad obsahuje všechny důležité části formátového řádku a vztahuje se k souborům v kondenzovaném formátu.

(I5,1X,8(I6,F3.0))

Nejdříve si všimněte, že popis formátu musí být uzavřen v závorkách. V tomto příkladě jsou použita tři písmena (jmenovitě **I**, **F** a **X**), která jsou dostačující pro popis jakýchkoliv dat v kondenzovaném formátu. V plném formátu se může objevit ještě znak pro rozdělení řádku (new-line), kterým je lomítko (/).

Písmeno **I** se užívá ke specifikaci formátu indexů (*indices*). Ty se používají jak v plném, tak v kondenzovaném formátu pro očíslování vzorků a pouze v kondenzovaném ještě pro číslování druhů. Spočtete-li tedy výskyt písmene **I** ve formátovém řádku, můžete snadno odvodit, vztahuje-li se k plnému (je tam pouze jedenkrát), či kondenzovanému (víc než jednou) formátu. Pokud se toto písmeno nevyskytne ani jednou, je formát obvykle neplatný. Taková situace může nastat i v případě, že jako vstup pro další analýzu je použit výsledek předchozích analýz programu Canoco (viz část 10.2). Specifikátor **I** je ve formě **I*w***, kde *w* je číslo udávající šířku rezervovanou pro index v datovém poli. Je to počet sloupců určených pro index. Pokud je počet cifer potřebných k úplnému popisu menší než tato šířka, je číslo zarovnáno doprava (zleva tedy vyplněno potřebným počtem mezer).

Vlastní hodnoty dat jsou ve formátu určeném specifikátorem **F*w.d***, tj. písmenem F následovaným dvěma čísly oddělenými tečkou. První číslo udává celkovou šířku pole (počet sloupců) vymezeného pro hodnoty, zatímco druhé číslo se týká počtu desetinných míst. Hodnoty jsou ve vymezeném prostoru zarovnány doprava (doplněním mezer na jejich levou část). Tak máme-li specifikátor **F5.2**, víme, že dva sloupce nejvíce vpravo obsahují číslice za desetinnou tečkou, ve třetím sloupci zprava je desetinná tečka a na celou část zbývají dva sloupce. Pokud máme hodnoty větší než 9.99, vyplní celé pole pro ně vymezené a od sousední hodnoty je vizuálně nebude nic oddělovat. Pak můžeme buď zvýšit hodnotu *w*, nebo přidat specifikátor **X**.

Specifikátor ***n*X** říká, že *n* sloupců obsahuje toliko mezery a mělo by být tedy při čtení přeskočeno. Alternativní způsob zápisu je přehození místa pro číslo udávající šířku a písmene **X** (**X*n***).

Teď už tedy můžeme rozebrat příklad formátového řádku z počátku kapitoly. V prvních pěti sloupcích je označení vzorku. Pamatujte, že toto číslo musí být zarovnáno doprava, takže označení prvního vzorku jsou 4 mezery a číslice '**1**'. V šestém sloupci jsou mezery a Canoco ho při čtení dat přeskočí. Další hodnota, která je před vloženou závorkou, je **specifikátor opakování**, a udává, že formát popsaný v závorce se bude opakovat osmkrát. Uvnitř závorky je šířka indexu druhu (6 sloupců) a hodnoty (3 sloupce). V případě kondenzovaného formátu lze mít na řádce i menší počet dvojic (index druhu – hodnota) než osm. Představme si, že máme vzorek, ve kterém je přítomno 10 druhů. Při použití vzorového formátu bude tento vzorek zapsán na dvou řádcích, z nichž první bude zaplněn zcela a druhý jen částečně (dvěma dvojicemi).

Jak jsme se zmínili již v části 1.8, je pro vzorky v úplném formátu vymezen stálý počet řádků. Specifikace formátu obsahuje tedy na své druhé řádce popis všech řádků jednoho vzorku. Je zde jen jedno pole **I**, které se vztahuje k číslu vzorku (je to ten popis **I**, jímž specifikace formátu začíná). Další informace určují pozici jednotlivých polí, ve kterých jsou hodnoty proměnných. Ke specifikaci místa, kde musí Canoco při čtení přejít na další řádek, se používá lomítko.

1.11. Transformace druhových dat

Jak si ukážeme v kapitole 2, hledají ordinační metody osy, které reprezentují regresní prediktory, jež jsou pro předpověď vysvětlovaných proměnných (tj. hodnot v druhových datech) v určitém smyslu nejlepší. Tudíž problém výběru vhodné transformace těchto proměnných je velmi podobný problému, který bychom museli řešit v případě použití kteréhokoli druhu jako vysvětlované proměnné v (mnohonásobné) regresní analýze. Další omezení, se kterým se potýkáme, je nutnost použít pro všechny vysvětlované proměnné

(“druhy”) stejnou transformaci. V unimodální (weighted averaging) ordinaci (viz sekce 2.4) nelze navíc použít záporné hodnoty, což přináší další omezení možných transformací.

Toto omezení je obzvlášť patrné při logaritmické transformaci. Logaritmus jedničky je nula a logaritmus hodnot mezi nulou a jedničkou je záporné číslo. Proto Canoco nabízí přizpůsobitelný vzorec pro log-transformaci:

$$y' = \log(A*y + C)$$

Čísla A a C bychom měli zvolit tak, aby po jejich použití v kombinaci s našimi daty (*y*) byl výsledek vždy větší nebo roven 1.0. Default hodnoty A i C jsou rovny 1.0, což elegantně mění nulové hodnoty opět na nuly, přičemž ostatní hodnoty jsou kladné. Nicméně pokud jsou naše originální data malá (řekněme v rozmezí od 0.0 do 0.1), bude posun hodnot způsobený přičítáním relativně velké hodnoty 1.0 ve výsledné struktuře matice dat převládat. V takovém případě upravíme transformaci zvýšením hodnoty A, v našem případě třeba na 10.0. Default log-transformace (i.e. *log(y+1)*) se hodí výborně například na procentuální data na stupnici 0-100.

Otázka, kdy použít log-transformaci a kdy data ponechat v původní podobě, není vůbec jednoduchá a existuje na ní téměř tolik odpovědí, kolik je statistiků. Já osobně nemám valné mínění o distribučních charakteristikách, alespon ne v tom smyslu, že by se frekvenční histogramy našich proměnných nutně musely porovnávat s „ideální“ Gaussovou (normální) distribucí. O použití či nepoužití log-transformace se raději rozhoduji podle postavení problému, který řeším. Jak už jsme před chvílí uvedli, lze ordinační metody vnímat jako rozšíření metod mnohonásobné regrese, takže si nyní tento přístup předvedeme v kontextu regrese. Zde bychom se mohli snažit předpovědět zastoupení určitého druhu ve vzorku na základě hodnot jednoho nebo více prediktorů (charakteristik prostředí a / nebo ordinačních os v případě ordinačních metod). V regresním modelu, kde pro jednoduchost pracujeme pouze s jedním prediktorem, pak může otázka znít: „Jak se změní průměrná hodnota druhu Y se změnou hodnoty prediktoru o jednotku?“. Pokud netransformujeme ani vysvětlovanou proměnnou, ani prediktor, může být odpověď: „Hodnota druhu Y se zvětší o **B**, pokud se hodnota prediktoru zvýší o jednu jednotku“. B je samozřejmě regresním koeficientem z rovnice lineárního modelu $Y = B_0 + B*X + E$. V jiných případech ale můžeme potřebovat spíše odpověď stylu: „Pokud se hodnota vysvětlující proměnné (prediktoru) zvýší o jednotku, zvedne se průměrná hodnota výskytu druhu o 10 % (nebo alternativně řečeno 1.10 krát)“. Zde uvažujeme v násobné stupnici, která není lineárnímu modelu vlastní. V takovém případě bych tedy použit log-transformaci vysvětlované proměnné.

Podobně, pokud chceme mluvit o vlivu změny prediktoru v násobcích, měli bychom tento prediktor log-transformovat. Jako příklad si uvedeme situaci, kdy používáme jako prediktor koncentraci dusíkatých iontů v půdě. Po našem modelu zřejmě nebudeme chtít, aby řešil otázku, co se stane v případě zvýšení koncentrace o 1 mmol/l, protože pak by se zvýšení z 1 na 2 bralo za stejný skok jako z 20 na 21.

Data o složení rostlinného společenstva jsou velmi často semi-kvantitativní odhady na Braun-Blanquetově stupnici se sedmi stupni (*r*, +, *1*, *2*, *3*, *4*, *5*). Taková stupnice je v tabulkách často kvantifikována odpovídajícími pořadovými hodnotami (v tomto případě od 1 do 7). Všimněte si, že toto kódování už v sobě logaritmickou transformaci obsahuje, protože skutečné rozdíly v početnosti / pokryvnosti jsou mezi jednotlivými stupni více či méně rostoucí. Alternativním přístupem využití takových dat v analýzách je, že je nahradíme odhadnutými hodnotami, které jsou přibližně ve středu rozsahů pokryvností vyjádřených v procentech. Narazíme sice na problém s hodnotami *r* a +, protože ty odrážejí spíše než

pokryvnost prostý počet jedinců, nicméně náhrada 0.1 za *r* a 0.5 za + obvykle pracuje uspokojivě.

Další šikovnou transformací dostupnou v CANOCU je odmocninná transformace. Tato transformace může být nejlepším řešením pro data vyjadřující počty (počty jedinců jednoho druhu, které se podařilo chytit do zemní pasti; počty jedinců různých druhů mravenců, kteří přejdou přes označenou „sčítací linii“, apod.). Na tato data však můžeme s klidem použít i log-transformaci.

Textová ("console") verze programu CANOCO 4.0 nabízí také poměrně obecnou metodu „lineární segmentoé transformace“, která nám dovoluje aproximovat mnohem komplikovanější transformační funkce pomocí lomené čáry s definovanými souřadnicemi zlomových bodů. Tato obecná transformace však ve verzi Canoco for Windows není.

Navíc, pokud potřebujeme jakýkoli typ transformace, kterou Canoco nenabízí, můžeme ji provést v tabulkovém procesoru a transformovaná data do formátu programu Canoco vyexportovat. To je obzvlášť užitečné, pokud naše „druhová“ data nepopisují složení společenstva, ale něco jako chemické či fyzikální vlastnosti půdy. V takovém případě mívají proměnné různé jednotky a pro každou z nich může být vhodná jiná transformace.

1.12. Transformace vysvětlujících proměnných

U vysvětlujících proměnných (v terminologii programu Canoco charakteristik prostředí a kovariát) se předpokládá, že nemají jednotnou stupnici a že pro každou z nich musíme volit vhodnou transformaci (včetně časté volby – netransformovat). Canoco ale takový postup neumožňuje, takže jakoukoli případnou transformaci vysvětlujících proměnných musíme provést před jejich exportem do souboru v Canoco formátu.

V každém případě však Canoco poté, co charakteristiky prostředí a / nebo kovariáty načte, je transformuje tak, aby měly nulový průměr a jednotkovou varianci. Tomuto postupu se často říká *normalizace*.

2. Metody gradientové analýzy

Úvodní terminologická poznámka: Výraz **gradientová analýza** je zde používán v širším slova smyslu pro jakoukoli metodu, která se pokouší dát do vztahu druhovou skladbu a gradienty prostředí (měřené nebo hypotetické). Vysvětlující proměnné jsou nazývány charakteristikami prostředí (obdobně termínu *environmental variables*, užívanému v programu Canoco). Kvantifikované druhové složení (vysvětlované proměnné) je ve shodě se středo-evropskou tradicí zvanou „snímek - relevé“ . Výraz **ordinace** je zde rezervován pro skupinu metod gradientové analýzy.

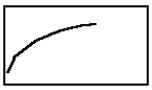
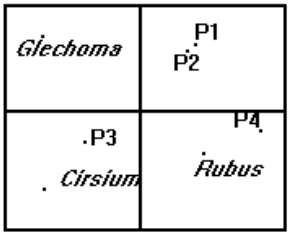
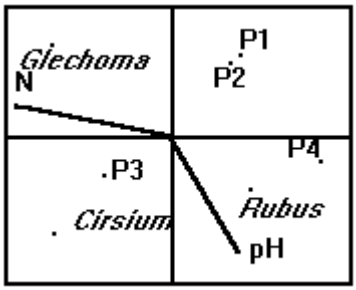
Metody pro analýzu druhového složení jsou často rozdělovány na analýzy gradientové (ordinace) a klasifikaci. Historicky jsou klasifikační metody spojovány s diskontinuálním přístupem založeným na vegetačních jednotkách nebo dokonce s Clemensovským organismálním přístupem k ekologickým společenstvům, zatímco metody gradientové analýzy bývají spojovány s pojetím kontinuálním, nebo dokonce s individualistickým pojetím rostlinných společenstev. Toto rozdělení (částečně) odráží historii vývoje metod, ale dnes již tak úplně neplatí. Zmíněné metody se doplňují a jejich použití závisí přednostně na cíli studia. Uvedeme si příklad na mapování vegetace, kde je klasifikace nezbytná. Dokonce i když mezi sousedními vegetačními typy žádné výrazné hranice neexistují, musíme toto kontinuum rozdělit, abychom pro účely mapování jednotlivé vegetační jednotky vymezili. Metody ordinace mohou napomoci při hledání opakujících se typů ve vegetaci, diskontinuit ve druhovém složení, nebo mohou ukázat přechodné typy atd., a dnes jsou užívány dokonce i ve fytocenologických studiích.

2.1. Techniky gradientové analýzy

Tabulka 2-1 shrnuje problémy, které se můžeme pokusit řešit za použití té či oné statistické metody. Kategorie se liší hlavně podle typu informace, kterou máme k dispozici.

Dále jsme do tabulky mohli zahrnout *parciální* ordinaci a *parciální* přímou (constrained) ordinaci, kde máme krom vysvětlujících proměnných ještě tzv. kovariáty (covariables, covariates). V parciální ordinaci nejdříve oddělíme závislost druhového složení na těchto kovariátách a potom provedeme (přímou nebo nepřímou) ordinaci.

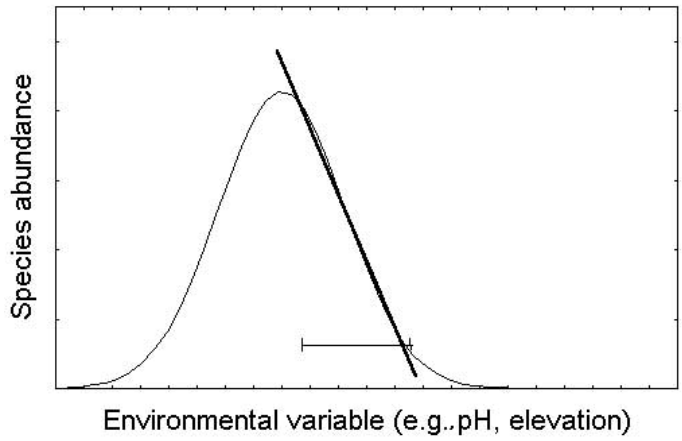
Charakteristikami prostředí a kovariátami jsou jak kvantitativní, tak kategoriální data.

Data, která máme		Apriorní znalost vztahů mezi druhy a prostředím	Použijeme	Dostaneme
Počet charakt. prostředí	**Počet druhů**			
1, *n*	1	ne	**Regrese**	Závislost druhu na prostředí 
žádné	*n*	ano	**Kalibrace**	Odhady hodnot charakteristik prostředí
žádné	*n*	ne	**Nepřímá ordinace**	Osy variability v druhovém složení (mohou být a měly by být poté vztaženy k naměřeným charakteristikám prostředí, pokud jsou tyto k dispozici) 
1, *n*	*n*	ne	**Přímá ordinace**	Variabilita ve druhovém složení vysvětlená charakteristikami prostředí. Vztah těchto charakteristik k osám druhů 

Tabulka 2-1 Vztah typů statistických přístupů k datům, která máme k dispozici, a otázkám, na které se ptáme.

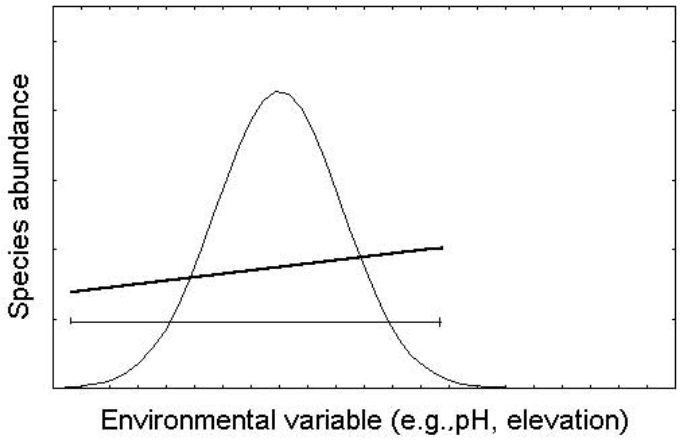
2.2. Modely odpovědi druhů na gradienty prostředí

Používají se dva typy modelů odpovědi druhů na gradienty prostředí: lineární (*linear*) a jednocestné (*unimodal*). Lineární odpověď je nejjednodušším odhadem, předpokladem unimodální odpovědi je, že druh má na gradientu prostředí své optimum.



Obrázek 2-1 Lineární aproximace unimodální odpovědi na krátkém výseku gradientu

Na krátkém gradientu funguje dobře lineární aproximace jakékoliv funkce (včetně funkce unimodální, Obrázek 2-1).



Obrázek 2-2 Lineární aproximace unimodální odpovědi na dlouhé části gradientu

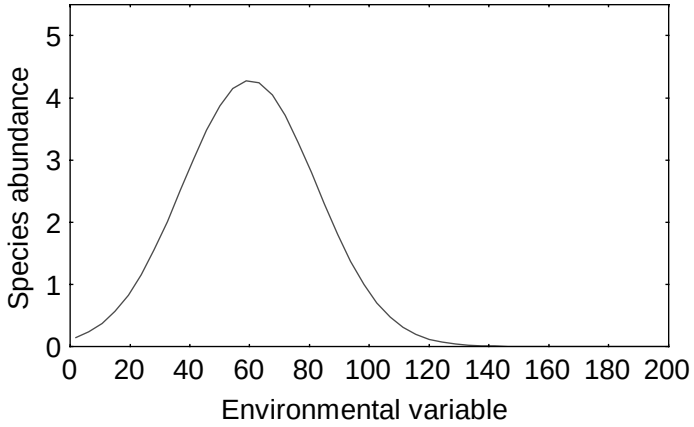
Na dlouhém gradientu je aproximace lineární funkcí velmi špatná (Obrázek 2-2). Měli bychom zde poznamenat, že i křivka unimodální odpovědi je zjednodušením: ve skutečnosti bývá odpověď zřídkakdy symetrická a nacházené odpovědi bývají rovněž mnohem složitější (např. dvouvrcholová - *bimodální*).

2.3. Odhad optima druhů metodou váženého průměrování

Lineární odpověď je obvykle odhadována klasickými metodami regrese (metoda nejmenších čtverců). Nejjednodušší cestou, jak odhadnout optimum druhu pro unimodální model, je spočítat vážený průměr těch hodnot charakteristik prostředí, při kterých se druh vyskytuje. Jako vah se při výpočtu používá četnosti či jiné důležitostní hodnoty druhu:

$$WA = \frac{\sum Env \times Abund}{\sum Abund}$$

kde *Env* je hodnota charakteristiky prostředí a *Abund* je četnost (případně jiná důležitostní hodnota) druhu v odpovídajícím vzorku.[+] Metoda váženého průměrování je vyhovující, pokud vzorky pokryjí celou distribuční křivku (Obrázek 2-3).



Obrázek 2-3 Příklad rozsahu, který pokrývá celou distribuční křivku

Celý rozsah pokrytý:

Hodnota charakteristiky prostředí	Druhová abundance	Součin
0	0.1	0
20	0.5	10
40	2.0	80
60	4.2	252
80	2.0	160
100	0.5	50
120	0.1	12
Total	**9.4**	**564**

$$WA = \frac{\sum Env \times Abund}{\sum Abund} = 564 / 9.4 = 60$$

Naopak pokud je pokryta jen část rozsahu, je odhad zkreslený:

Pokrytá jen část rozsahu:

Hodnota charakteristiky prostředí	Abundance druhu	Součin
60	4.2	252
80	2.0	160
100	0.5	50
120	0.1	12
Total	**6.8**	**472**

$$WA = \frac{\sum Env \times Abund}{\sum Abund} = 472 / 6.8 = 69.4$$

Čím delší osa, tím víc druhů bude mít svoje optimum odhadnuto správně.

[+] Jinou možností je odhadnout parametry unimodální křivky přímo. Tato možnost je ale mnohem složitější a není vhodná pro výpočty probíhající současně, což se v ordinačních metodách velmi často používá.

Techniky založené na modelu lineární odpovědi (*linear response*) jsou vhodné pro homogenní datové soubory, zatímco pro heterogenní data jsou lepší techniky váženého průměrování (*weighted averaging*).

2.4. Ordinace

Problém nepřímé ordinace můžeme formulovat několika různými způsoby:

1. Najdi takové rozložení vzorků v ordinačním prostoru, kde vzdálenosti vzorků v ordinačním prostoru odpovídají nejlépe rozdílům ve druhovém složení. Toto činí explicitně nemetrické (ale i metrické) mnohorozměrné škálování (*non-metric multidimensional scaling*, *NMDS*).

2. Najdi teoretické ("latentní") proměnné (= ordinační osy), pro které je celková závislost všech druhů nejtěsnější. Tento model vyžaduje, aby byl typ odpovědi druhů na proměnné explicitně specifikován: lineární odpověď pro lineární metody, unimodální odpověď pro metody založené na vážených průměrech (eplicitní „Gaussova ordinace“ se pro výpočty běžně neužívá). V lineárních metodách je skóre vzorku lineární kombinací (váženým součtem) skóre druhů. V metodách váženého průměru se pak ke skóre vzorku dochází váženým průměrem druhových skóre (po určitých úpravách).

> **Poznámka**: při váženém průměrování je implicitně zahrnuta standardizace po vzorcích i po druzích. U lineárních metod si můžeme vybrat mezi standardizovanou a nestandardizovanou verzí.

3. Představme si vzorky jako body v mnohorozměrném prostoru, kde jsou druhy osami a pozice každého vzorku odpovídá četnosti příslušného druhu. Potom je cílem ordinace najít takové promítnutí tohoto mnohorozměrného prostoru do prostoru o méně rozměrech, které způsobí jen minimální zkreslení prostorových vazeb. Všimněte si, že výsledek závisí na tom, jak definujeme “minimální zkreslení”.

Měli bychom poznamenat, že různé formulace mohou vést ke stejným výsledkům. Například analýza hlavních komponent (*principal component analysis*, *PCA*) můžeme formulovat kterýmkoliv z výše uvedených způsobů.

2.5. Přímá ordinace

Tyto ordinace můžeme nejlépe vysvětlit v rámci ordinací definovaných jako hledání nejlepších vysvětlujících proměnných (tj. druhá formulace v předcházející sekci). Zatímco v nepřímých ordinacích hledáme jakoukoli proměnnou, která je schopna vysvětlit nejlépe druhové složení (a tu potom vezmeme jako ordinační osu), v ordinacích přímých jsou ordinačními osami vážené charakteristik prostředí. Proto čím méně těchto charakteristik máme, tím přísnější bude omezení. Pokud je jejich počet větší než počet vzorků zmenšený o jedna, pak se ordinace stává nepřímou.

Neomezené (*unconstrained*) ordinační osy odpovídají směru největší variability v souboru dat. Omezené (*constrained*) ordinační osy odpovídají směru největší variability v datovém souboru, která může být vysvětlena charakteristikami prostředí. Počet omezených os nemůže být větší než počet charakteristik prostředí.

2.6. Kódování charakteristik prostředí

Charakteristiky prostředí mohou být buď kvantitativní (pH, převýšení, vlhkost) nebo kvalitativní (kategoriální). Kategoriální proměnné s více než dvěma kategoriemi se kódují jako několik indikátorových proměnných (*dummy variables, indicator variables*); indikátorová proměnná nabývá obvykle hodnot jedna nebo nula. Předpokládejme, že máme pět ploch; plochy 1 a 2 jsou na vápenci, plochy 3 a 4 na žule a plocha 5 na čediči. Podloží pak bude popsáno třemi charakteristikami prostředí (vápenec, žula, čedič) takto:

	vápenec	žula	čedič
Plocha 1	1	0	0
Plocha 2	1	0	0
Plocha 3	0	1	0
Plocha 4	0	1	0
Plocha 5	0	0	1

Proměnná čedič není nezbytně nutná, protože je lineární kombinací předcházejících dvou: čedič = 1 - vápenec - žula. Pro konstrukci grafů je však užitečné tuto kategorii mít.

2.7. Základní techniky

Existují čtyři základní ordinační techniky, založené na modelu druhové odpovědi a na tom, zda je ordinace přímá či nepřímá (Ter Braak & Prentice, 1998):

	Metody lineární	**Metody váženého průměrování**
neomezené	analýza hlavních komponent (PCA)	korespondenční analýza (CA)
omezené	redundanční analýza (RDA)	kanonická korespondenční analýza (CCA)

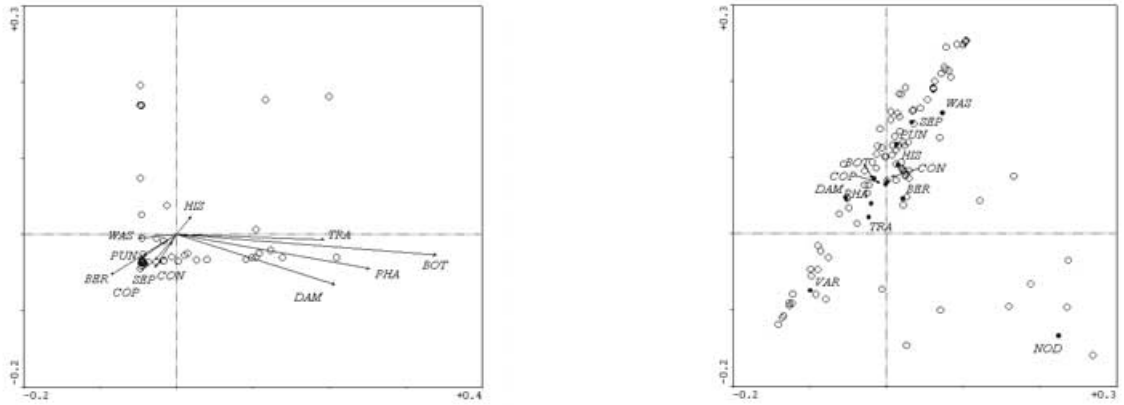
Tabulka 2-2 Typy gradientové analýzy sensu *Ter Braak & Prentice, 1998*

Pro metody váženého průměrování existují trendu zbavené (detrended) verze (tj. Detrended Correspondence Analysis, DCA, známá DECORANA, and Detrended Canonical Correspondence Analysis, DCCA, viz také oddíl 3.5). Pro všechny metody lze použít dílčí (parciální) analýzy. V parciálních analýzách je nejdříve oddělen vliv kovariát a analýza je pak provedena jen na zbývající variabilitě.

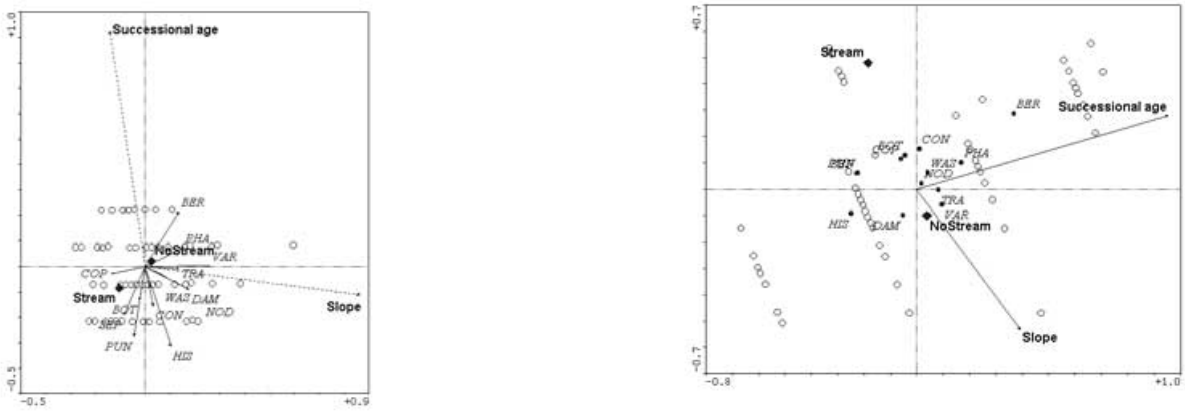
2.8. Ordinační diagramy

Výsledky ordinací se obvykle prezentují jako ordinační diagramy. Ve všech metodách jsou plochy (vzorky) zastoupeny body (symboly). Druhy jsou v lineárních metodách zobrazeny jako šipky ve směru, v jakém roste abundace druhu a jako body (symboly) v metodách váženého průměru (pak označují optimum druhu). Kvantitativní charakteristiky prostředí jsou značeny jako šipky ve směru, v jakém roste jejich hodnota. Kvalitativní charakteristiky prostředí jsou pro jednotlivé kategorie, ve kterých se vyskytují, značeny jako centroidy.

PCA **CA**



RDA **CCA**



Obrázek 2-4: Příklady typických ordinačních diagramů. Analýza zastoupení druhů rodu *Ficus* v lesích různého sukcesního stáří na Papui Nové Guinei. Druhy jsou značeny takto: *F. bernaysii* - BER, *F. botryocarpa* - BOT, *F. conocephalifolia* - CON, *F. copiosa* - COP, *F. damaropsis* - DAM, *F. hispidoides* - HIS, *F. nodosa* - *NOD, F. phaeosyce* - PHA, *F. pungens* -PUN, *F. septica* - *SEP, F. trachypison* - TRA, *F. variegata* - VAR, a *F. wassa* – WAS. Kvantitativními charakteristikami prostředí jsou sklon (*slope*) a sukcesní stáří (*successional age*), kvalitativní jsou přítomnost malého toku (*NoStream*, *Stream*). Snímky jsou vyznačeny jako prázdná kolečka.

2.9. Dva přístupy

Pokud máte jak data o prostředí, tak informace o druhové skladbě společenstev, můžete použít oba přístupy: spočítat nejdřív nepřímou ordinaci s následnou regresí ordinačních os na měřených charakteristikách prostředí (tj. promítnutím těchto charakteristik do ordinačního diagramu), a také můžete spočítat přímou (omezenou) ordinaci. **Tyto přístupy jsou komplementární a měly by se použít oba!** Pokud spočítáte nejdřív nepřímou ordinaci, určitě postihnete hlavní složku variability ve druhovém složení, ale zato byste mohli opominout tu část variability, která je vztažena k charakteristikám prostředí. Když naopak počítáte přímou ordinaci, zcela jistě vám neujde hlavní část variability vysvětlená

charakteristikami prostředí, ale mohli byste opomenout variabilitu nemající vztah k těmto měřeným proměnným.

Dbejte na to, abyste vždy uvedli metodu, kterou jste při analýze použili. Z ordinačního grafu totiž sice můžete vyčíst, zda se jednalo o lineární nebo unimodální analýzu, použití přímé či nepřímé ordinace však vyčíst nelze.

Hybridní analýzy jsou jakýmsi „křížencem“ mezi přímými a nepřímými ordinacemi. Ve standardní přímé ordinaci je tolik omezených (kanonických) os, kolik je nezávislých vysvětlujících proměnných a pouze další ordinační osy jsou neomezené. V hybridní analýze se spočítá jen předem daný počet omezených os a jakékoliv další ordinační osy jsou neomezené.

2.10. Parciální analýzy

Někdy potřebujeme nejdříve oddělit variabilitu vysvětlenou jedním souborem vysvětlujích proměnných a teprve potom analyzovat variabilitu zbývající (tj. provést analýzu na zbylé variabilitě). Toho lze dosáhnout **parciální (dílčí) analýzou**, kde nejdřív odstraníme variabilitu vysvětlenou **kovariátami** (*covariables*) (tj. vliv proměnných, který si přejeme vyloučit) a potom provedeme (přímou nebo nepřímou) ordinaci. Kovariáty jsou často kontinuální nebo kategoriální proměnné, jejichž vliv je nezajímavý, např. bloky v experimentálním designu. Pokud máme více vysvětlujících proměnných, provádíme několik analýz, kde je vždy jedna z proměnných vysvětlující proměnnou a zbytek jsou kovariáty, což nám umožní testovat parciální (dílčí) vlivy (analogicky parciálním regresním koeficientům v mnohonásobné regresi).

2.11. Testování významnosti vztahů s charakteristikami prostředí

V běžných statistických testech je hodnota testové statistiky spočtené z dat porovnávána s očekávanou distribucí statistiky, jaká by byla získána v případě platnosti nulové hypotézy, kterou testujeme. Na základě tohoto porovnání pak odhadujeme pravděpodobnost toho, že bychom získali tak od nulového modelu odlišná (nebo dokonce ještě odlišnější) data, než jaká ta naše skutečně byla. Distribuci testové statistiky odvozujeme z odhadu distribuce našich originálních dat. V Canocu distribuce statistik testu[*] v případě platnosti nulové hypotézy o nezávisloti neznáme. Tyto distribuce závisí nejen na počtu charakteristik prostředí, ale také na rozložení pokryvností druhů a dalších vlastnostech (korelační struktuře) souboru. Podmínky, jaké vzniknou v případě platnosti nulové hypotézy, však můžeme simulovat **Monte Carlo permutačním testem.**

V tomto testu získáme distribuci statistik následujícím způsobem: Nulová hypotéza zní tak, že odpověď (druhové složení) je na charakteristikách prostředí nezávislé. Pokud je to pravda, pak je jedno, který soubor vysvětlujících proměnných je přiřazen ke kterému snímku. Následně tedy přiřadíme hodnoty charakteristik prostředí náhodně k jednotlivým snímkům a spočítáme hodnoty testových statistik. Výsledná hladina významnosti je počítána takto: $P = \frac{1+m}{1+n}$; Kde *m* je počet náhodných permutací, pro které byla statistika testu vyšší než pro originální data, a *n* je celkový počet permutací. Tento test nezávisí na žádném předpokladu o distribuci hodnot pokryvností druhů. Permutační schéma si můžeme upravit podle použitého designu pokusu. Toto je základní verze Monte Carlo permutačního testu.

[*] F-poměr v nejnovějších verzích Canoca – mnohorozměrný protějšek běžného F-poměru, v předcházejících verzích bylo použito charakteristické číslo příslušné k dané ose, tzv. *eigenvalue*

V programu Canocu se používají sofistikovanější přístupy, zejména pokud jsou použity kovariáty – viz manuál programu Canoco for Windows (*Ter Braak & Šmilauer, 1998*).

2.12. Jednoduchý příklad Monte Carlo permutačního testu pro významnost korelace

Máme výšku pěti rostlin a obsah dusíku v půdě, kde rostly. Jejich vztah je popsán korelačním koeficientem. Za určitých podmínek známe distribuci korelačního koeficientu v případě platnosti nulové hypotézy o nezávisloti. Předpokládejme ale, že z nějakého důvodu tuto distribuci neznáme. Distribuci můžeme simulovat náhodným přiřazením hodnot dusíku k výškám rostlin. Uděláme mnoho náhodných permutací a pro každou spočteme koeficient korelace množství dusíku s výškou rostlin. Protože hodnoty dusíku byly k výškám rostlin přiřazeny náhodně, odpovídá distribuce korelačních koeficientů nulové hypotéze o nezávislosti.

Výška rostlin	**Dusík (v datech)**	**První perm.**	**Druhá perm.**	**Třetí perm.**	**Čtvrtá perm.**	**Pátá etc....**
5	**3**	3	8	5	5	
7	**5**	8	5	5	8	
6	**5**	4	4	3	4	
10	**8**	5	3	8	5	
3	**4**	5	5	4	3	
Korelace	**0.878**	0.258	-0.568	0.774	0.465	0.###

Významnost odchylky od distribuce při nulovém modelu pro jednostranný test je pak odhadována jako:

$$\frac{1 + \text{počet permutací, kde } (r>0.878)}{1 + \text{celkový počet permutací}}$$

Pro oboustranný test je to:

$$\frac{1 + \text{počet permutací, kde } (|r|>0.878)}{1 + \text{celkový počet permutací}}$$

Všimněte si, že F test používaný ANOVOU (a podobně F-poměr používaný programem Canoco) jsou testy jednostranné (*one-sided, one-tailed*).

3. Používání programového souboru Canoco for Windows 4

3.1. Přehled programů

Programový soubor Canoco for Windows se skládá z několika samostatných programů, které si v této části popíšeme. V dalších kapitolách jsou pak typické příklady použití. Tato kapitola však v žádném případě není plnohodnotnou náhradou dokumentace programu Canoco for Windows.

Canoco for Windows 4.0

Tento program je centrální částí balíku. Zde specifikujeme data, se kterými chceme pracovat, ordinační model a nastavení testů. Zde také můžeme provádět výběry z vysvětlujících a vysvětlovaných proměnných nebo měnit váhy jednotlivých vzorků.

Canoco for Windows nám umožňuje analyzovat datové soubory s až 25000 vzorky, 5000 druhy, 750 charakteristikami prostředí a 1000 kovariátami. Na celkový počet hodnot v datech se vztahují ještě další omezení – pro druhová data se tato omezení týkají jen nenulových hodnot, tj. absence jsou vypuštěny, protože ty program neukládá.

Canoco for Windows pracuje s poměrně širokým okruhem ordinačních metod. Hlavní jsou lineární (PCA a RDA) a unimodální (CA, DCA a CCA) metody. Kromě nich můžeme použít i další metody, jako je diskriminační analýza (CVA) nebo metrické mnohorozměrné škálování (*principal coordinate analysis*, PCoA). V seznamu použitelných metod chybí jen nemetrické mnohorozměrné škálování (NMDS).

CANOCO 4.0

Toto je alternativa k uživatelsky příjemnějšímu, ale o něco jednoduššímu programu Canoco for Windows. Je to textová podoba tohoto softwaru ovládaná z příkazové řádky (tzv. *console version*). Uživatelské rozhraní je shodné s předchozími verzemi programu Canoco (hlavně s verzemi 3.x), ale výkonnost programu byla rozšířena.

Tato varianta je méně pohodlná než verze pod Windows – pokud totiž uděláme chybu a zvolíme špatné nastavení, nelze se ke špatně zodpovězené otázce vrátit a nezbývá, než program ukončit.

Na druhou stranu má tato varianta několik výhod. Podle mého názoru stojí mezi nimi za zmínku vpodstatě jen možnost zadání „nepravidelného“ designu. Můžete mít například data z trvalých ploch sbíraná opakovaně ze tří lokalit. Pokud byla data sbírána různý počet let, nelze to ve verzi pro Windows nijak zadat, takže v Monte Carlo permutačních testech jsou pak nesprávná omezení permutací. *Console version* umožňuje specifikovat uspořádání vzorků (druhovou či časovou strukturu a / nebo obecný split-plot design) pro každý blok vzorků nezávisle.

Další výhodou této varianty je její schopnost číst zadání analýz (které jsou normálně zadávány uživatelem jako odpovědi na jednotlivé otázky programu) z dávkového souboru. Potom je tedy možné naprogramovat takovou dávku a provést rychle více analýz. Toto nastavení je výhodou hlavně pro zkušené uživatele.

WCanoImp a CanoImp.exe

Funkce programu WCanoImp jsme si popsali už v části 1.7. Jediným větším nedostatkem této malé, uživatelsky příjemné aplikace, je její omezení velikostí schránky (Clipboard) ve Windows. Nyní již toto omezení není tak závažné, jako bývalo v Microsoft Windows 3.1 a 3.11. Důležitější je omezení možnostmi našeho tabulkového procesoru. V programu Microsoft Excel nelze mít víc než 255 sloupců dat, takže se musíme omezit buď na 255 proměnných nebo 255 vzorků. Druhý rozměr je mnohem shovívavější – ve verzi Microsoft Excel 97 je to 65536 řádků.

Pokud se naše data do těchto limitů nevejdou, můžeme se pokusit omezení obejít rozdělením tabulky, exportem částí a následným spojením Canoco souborů (což ovšem vůbec není jednoduchý úkol) nebo můžeme použít *console version* (ovládanou z příkazové řádky) programu WCanoImp – program **canoimp.exe**. Oba programy mají stejnou funkci, odlišují je však dvě věci. První z nich je, že vstupní data musí být uložena v textovém souboru. Obsah souboru je shodný s tím, co tabulkový procesor ukládá do schránky (Clipboard). Je to textová reprezentace obsahu tabulek, kde je přechod mezi sloupci naznačen tabelátorem a přechod mezi řádky znakem pro nový řádek. Nejjednodušším způsobem, jak vytvořit vstupní soubor pro program canoimp.exe, je postupovat stejně, jako když používáme program WCanoImp, až do bodu, kdy jsou data kopírována do schránky (Clipboard). V této chvíli přeskočíme do programu WordPad (ve Windows 9x) nebo do programu Notepad (ve Windows NT 4.x a Windows 2000), tam vytvoříme nový dokument a zvolíme příkaz **Edit/Paste**. Potom dokument uložíme jako soubor ASCII (v Notepadu to ostatně ani jinak nejde, ale WordPad podporuje i jiné formáty). Alternativním postupem je uložení tabulky přímo v tabulkovém procesoru použitím funkce File/Save as… a takovým výběrem formátu, který je charakterizován jako *Text file (Tab separated)* nebo podobným textem. Tento postup ovšem funguje bez obtíží jen tehdy, pokud krom tabulky v tabulkovém dokumentu již nic jiného není.

Druhým rozdílem mezi utilitou WCanoImp a programem canoimp.exe je, že volby, které se v programu WCanoImp nastavují v hlavním okně, se musí u canoimp.exe zadat (spolu se jménem vstupního a požadovaným jménem výstupního souboru) na příkazové řádce při spouštění programu conoimp. Následující příklad nám ukáže, jak tedy může spuštění programu vypadat:

```
d:\canoco\canoimp.exe -C -P inputdta.txt output.dta
```

kde volba **–C** znamená výstup v kondenzovaném formátu, a volba **–P** transpozici vstupních dat (tj. v řádcích jsou ve vstupním souboru proměnné). Data oddělená tabelátorem se načtou ze souboru **inputdta.txt** a CanoImp poté vytvoří nový datový soubor (nebo přepíše stávající soubor s tímto jménem), který pojmenuje **output.dta**.

Pokud chcete zjistit, jak přesně má příkazová řádka vypadat, spusťte canoimp.exe bez jakýchkoli parametrů (tj. i beze jmen vstupního a výstupního souboru). Program pak vypíše krátkou informaci o tom, v jakém formátu má specifikace parametrů být.

CEDIT

Program CEDIT je k dispozici v instalačním programu Canoco for Windows jako jeho volitelná součást. Pro operační systémy Windows NT (a Windows 2000) jeho instalaci příliš nedoporučujeme, protože bezchybná instalace vyžaduje hlubší znalost těchto systémů.

V prostředí Windows 9x by však instalace měla proběhnout hladce (alespoň pokud ji provedete do default adresáře **c:\canoco**).

Dostupnost tohoto programu je věcí speciální úmluvy s jejím autorem, a proto není k dispozici uživatelská podpora pro případ jakýchkoli komplikací. Pokud si jej ale nainstalujete, získáte dokumentaci k programu v instalačním adresáři, včetně instrukcí pro správné nastavení.

Další nevýhodou je v očích mnoha uživatelů jeho stručné textové rozhraní, dokonce ještě mnohem víc složitější než to, které je k dispozici u console verze programu Canoco. Pokud ale rádi používáte textové editory pod UNIXem, kde je takové rozhraní s příkazy zkrácenými na několik písmen běžné, potom si určitě oblíbíte i CEDIT.

Co je tedy pro nás zbylé na tomto programu přitažlivé? Je to jeho mimořádná síla v provádění poměrně složitých operací s daty, která už jsou ve formátu Canoco. Není pochyb o tom, že většina těchto operací se dá udělat (téměř) stejně snadno i v tabulkovém procesoru pod Windows, potíž je v tom, že data vždy ve vhodném formátu nemáme (hlavně u již existujících souborů). CEDIT může transformovat proměnné, rozdělovat nebo spojovat datové soubory, transponovat data, překódovat různé faktory (přeměnou faktoriální proměnné na soubor indikátorových proměnných) a ještě mnohem víc.

CanoDraw 3.1

Program CanoDraw 3.1 je distribuován s programovým balíkem Canoco for Windows a je založen na původní verzi 3.0, která byla k dispozici jako přídavek k softwaru CANOCO 3.1x (“*lite*” verze CanoDraw 3.0 byla distribuována s každou kopií programu CANOCO 3.1x pro PC).

Mezi verzemi 3.0 a 3.1 nastalo jen málo změn a protože jejich kořeny sahají až k roku 1992, uživatelské rozhraní se z pohledu dnešních standardů zdá poněkud neohrabané. CanoDraw nemá textové rozhraní a jeho grafický mód je omezen na standardní VGA rozlišení (640x480 bodů) a běží obvykle jen v celoobrazovkovém režimu. Většinou je však možné jej rozběhnout přímo z Windows, takže můžeme při závěrečných úpravách diagramů podle potřeby přeskakovat z programu Canoco do programů CanoDraw nebo CanoPost.

CanoDraw disponuje rozsáhlou funkčností na "malé ploše". To je také důvodem, proč někdy jeho použití není právě lehké. Krom zobrazení jednoduchých ordinačních diagramů a příslušného vzájemného přeškálování skóre pří přípravě projekčních diagramů, umožňuje CanoDraw také další prozkoumání našich dat na základě ordinačních výsledků. K těmto účelům poskytuje celou řadu metod – včetně zobecněných lineárních modelů, vyhlazovací metody *loess* a zobrazování výsledků těchto metod pomocí vrstevnicových diagramů. Dále můžeme naše ordinační data kombinovat se zeměpisnými souřadnicemi jednotlivých vzorků, roztřídit je do jednotlivých skupin a výsledné roztřídění také zobrazit, porovnat skóre vzorků mezi jednotlivými ordinačními metodami a tak dále.

Co se týče nastavení výstupu, CanoDraw podporuje přímý výstup na několik typů tiskáren, včetně tiskáren typu HP LaserJet, ale dnes bychom uživatelům doporučili spíše uložení jejich grafů buď do formátu Adobe Illustrator (AI) nebo PostScript (PSC), které potom ještě mohou vylepšit v programu CanoPost. Ačkoliv Adobe Illustrator je beze sporu velmi silným prostředkem pro další úpravu jakýchkoli grafů, má bohužel také svá omezení. Tento program totiž vůbec netuší, co to taková ordinační metoda je: neví, že se symboly a šipkami se hýbat nesmí, na rozdíl od popisků, nebo že měřítko svislé osy by se nemělo

měnit nezávisle na ose vodorovné. A v neposlední řadě je používání programu Adobe Illustrator spojeno s nutností další licence, zatímco CanoPost je poskytován přímo v programovém balíku Canoco for Windows. AI soubory mohou být navíc exportovány i z programu CanoPost, takže uživatelé o lepší vlastnosti Illustrátoru stejně nepřijdou.

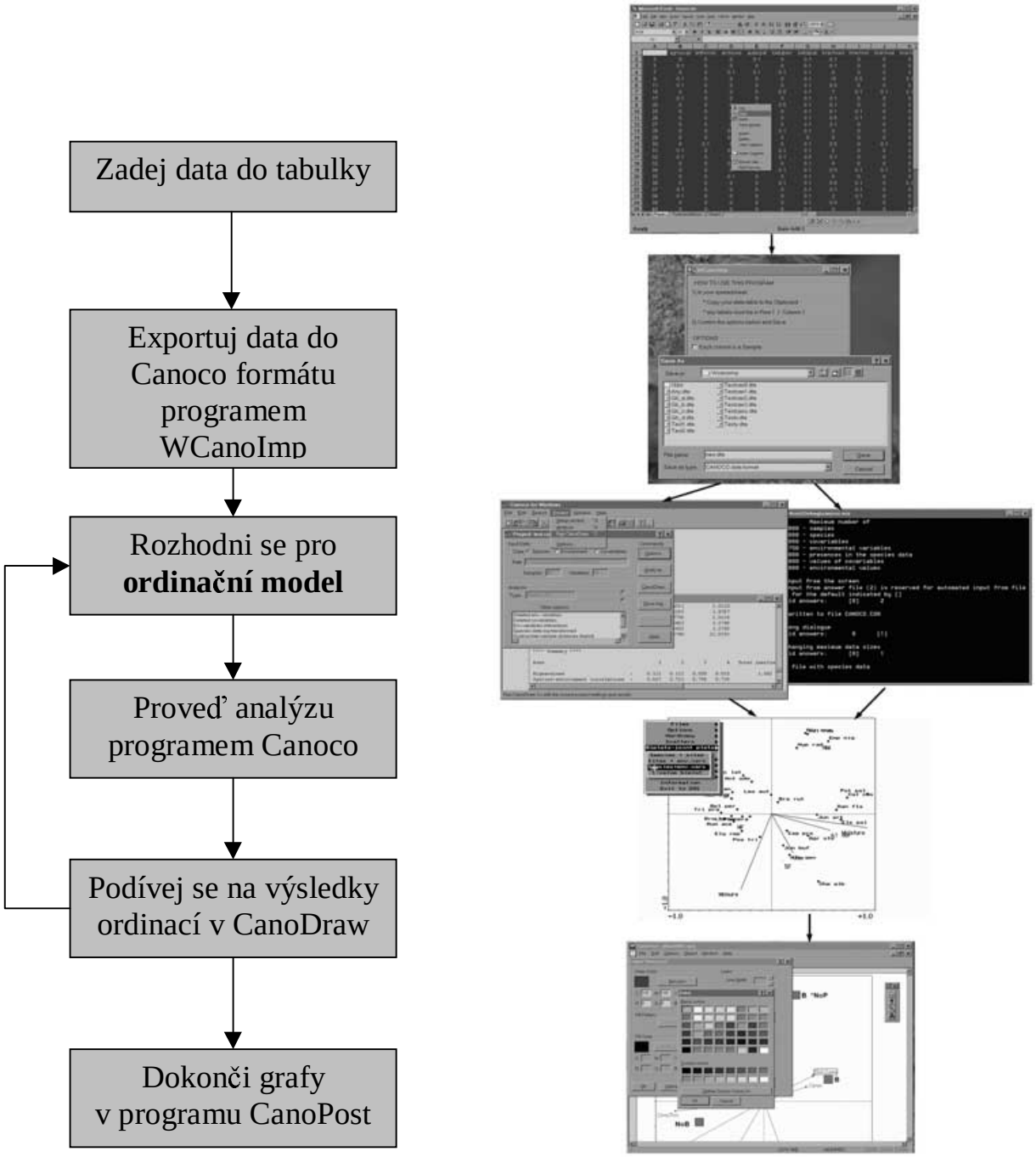
CanoPost for Windows 1.0

Tento program pracuje se soubory vytvořenými programem CanoDraw, které byly uloženy v PostScriptovém formátu (obvykle s příponou **.psc**). Za zmínku stojí, že takové soubory můžete také tisknout přímo na laserové tiskárně podporující PostScript. K použití souborů v programu CanoPost však PostScriptovou tiskárnu nepotřebujete! V CanoPostu ale můžete pracovat pouze se soubory vytvořenými programem CanoDraw, nikoli s jinými typy PostScriptových souborů.

CanoPost umožňuje další úpravy grafů, včetně změn textu, stylu jmenovek, symbolů, čar nebo šipek. Umístění jmenovek lze měnit tažením myší. Změny, které se do jednotlivých diagramů udělají, lze uložit do stylových souborů, a potom je použít pro jakýkoliv jiný ordinační diagram. Kromě úprav jednotlivých diagramů lze také provádět slučování několika grafů do jednoho.

Upravené diagramy můžete uložit v nativním formátu programu CanoPost (s příponou **.cps**), tisknout na libovolném rastrovém výstupním zařízení podporovaném instalací Windows nebo exportovat jako bitmapu (.BMP) nebo soubor ve formátu programu Adobe Illustrator.

3.2. Typický postup analýz s programem Canoco for Windows 4.0



Obrázek 3-1 Zjednodušený postup při použití programu Canoco for Windows

Na obrázku 3-1 je typické pořadí akcí, které provádíme, když analyzujeme mnohorozměrná data. Začínáme s daty v tabulce, která exportujeme s pomocí programu WCanoImp do souborů kompatibilních s Canocem. Potom v programu Canoco for Windows buď vytvoříme nový projekt nebo naklonujeme nějaký stávající, a to příkazem *File/Save as...* Díky naklonování zdědíme nastavení všech parametrů, z nichž pak upravíme jen ty, které potřebujeme změnit. Je zřejmé, že změnou jmen zdrojových souborů jsou postiženy i výběry, které na nich bezprostředně závisí (jako například seznam charakteristik prostředí, které se mají vypustit).

V každém projektu jsou dvě okna (views). **Project view** shrnuje nejdůležitější charakteristiky projektu (např. typ ordinační metody, rozměry tabulek s daty a jména souborů, ve kterých jsou data uložena). Navíc je v tomto zobrazení sloupec s tlačítky pro nejpoužívanější příkazy při práci s projekty: provedení analýzy, změna nastavení projektu, spuštění CanoDraw, uložení záznamu analýzy, atd. V **Log view** se ukládají záznamy o provedených akcích a výsledky analýz. Některé ze statistických výsledků, které Canoco spočte, jsou dostupné jen v tomto zobrazení. Jiné výsledky jsou ukládány do „SOL“ souborů, ve kterých jsou aktuální ordinační skóre. Obsah Log view lze rozšířit vložením nového textu (poznámek): Log view funguje jako jednoduchý textový editor.

Nastavení projektu můžeme definovat pomocí průvodce (*Project Setup wizard*). Můžeme ho vyvolat například kliknutím na tlačítko **Options** v Project view. Canoco pak zobrazí první stránku ze serie stran s různými informacemi, které průvodce potřebuje znát, aby použil vhodnou ordinační metodu. Tato série není statická – zobrazení určité stránky závisí na předchozích volbách. Některá nastavení se například vztahují jen na lineární ordinační metody, a proto jsou tyto stránky zobrazeny pouze v případě, že tyto metody (PCA nebo RDA) vybereme. Z jedné stránky na druhou se dostaneme po stisknutí tlačítka **Next** ("další"). K předchozím stránkám se lze vrátit kliknutím na tlačítko **Back** ("zpět"). Některé z kritických výběrů, které při používání průvodce musíme učinit, si v této kapitole ještě později popíšeme podrobněji. Na poslední stránce je tlačítko **Next** nahrazeno tlačítkem **Finish** ("dokončit"). Po jeho stisknutí se změny nastavení aplikují na náš projekt. Pokud začínáme nový projekt, bude se Canoco dožadovat jména pro tento projekt, aby ho mohl uložit.

Po nadefinování celého projektu můžete spustit analýzu kliknutím na tlačítko **Analyze** v Project view (nebo případně použitím ikony z lišty nebo nabídky příkazů). Pokud analýza proběhne úspěšně, zobrazí se její výsledky ve výsledkovém souboru (jeho jméno se určí na druhé průvodcovské stránce při definování projektu) a další informace se objeví v Log view, kde si je můžete prohlédnout. Jsou zde statistická shrnutí pro první čtyři ordinační osy, informace o korelaci mezi charakteristikami prostředí a ordinačními osami, určení odlehlých pozorování (*outliers*) a výsledky Monte Carlo permutačních testů. Část těchto výsledků je důležitá pro zvládnutí dalších úkolů, ale nic z toho není třeba zachovat pro vykreslení ordinačních diagramů programem CanoDraw. CanoDraw potřebuje pouze výsledky z výsledkového souboru.

Programem CanoDraw je možné prozkoumat výsledky ordinací a zkombinovat je s původními daty. Zde si určujeme základní obsah ordinačních diagramů (rozsah os, které položky vynést a které ne, obsahy diagramů atributů, atd.). Výsledné diagramy můžeme dál upravovat programem CanoPost (měnit typ symbolů, barvy nebo velikost, písmo a umístění popisků, styl čar, atd.) do podoby přijatelné pro publikace.

3.3. Rozhodnutí o ordinačním modelu: unimodální nebo lineární?

Tuto část lze brát jako jednoduchou "kuchařku" pro rozhodování mezi ordinačními metodami založenými na modelu lineární odpovědi druhů na gradient prostředí a ordinačními metodami váženého průměru (weighted averaging - WA), které odpovídají modelu jednovrcholové (unimodální) odpovědi druhů. Takový recept je ale nevyhnutelně zjednodušením, takže ho nelze sledovat slepě.

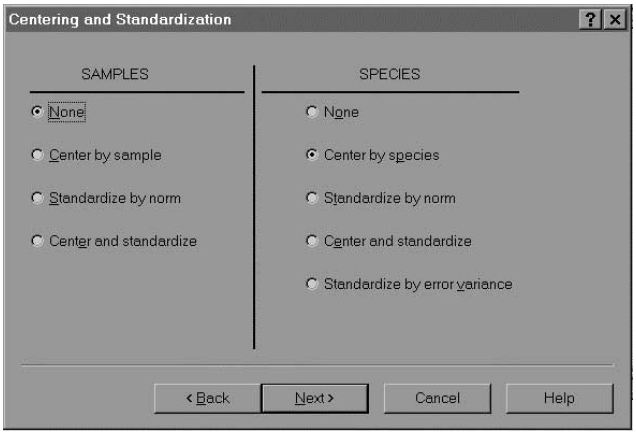
V projektu programu Canoco for Windows, pomocí kterého se rozhodujeme mezi unimodálními a lineárními metodami, se snažíme udělat co nejvíc stejných rozhodnutí, jako v konečných analýzách. Máme-li kovariáty, použijeme je zde také, chceme-li pracovat pouze s výběrem charakteristik prostředí, uděláme tentýž výběr i zde. Plánujeme-li logaritmickou transformaci (nebo odmocninnou transformaci) našich dat, provedeme ji i tady.

V tomto zkušebním projektu zvolíme metodu váženého průměru s odstraněním trendu. To znamená metodu DCA pro nepřímou gradientovou analýzu nebo DCCA pro přímou gradientovou analýzu (tj. s omezením). Pak použijeme metodu odstranění trendu po segmentech (což v sobě zahrnuje také Hillovo škálování ordinačních skóre) a následně zvolíme i ostatní nastavení stejná jako v závěrečných analýzách a analýzu spustíme. Poté se podíváme do Log view na výsledky. Na konci výpisu je souhrnná tabulka (Summary table) a v ní řádek, který začíná slovy *Lengths of gradient* (tj. "délky gradientu"). Ten může vypadat podobně jako následující příklad:

```
Lengths of gradient     :   2.990  1.324   .812   .681
```

Nyní najdeme největší hodnotu (nejdelší gradient) a pokud je tato hodnota větší než 4.0, měli bychom použít unimodální metodu (DCA, CA nebo CCA). Použití lineární metody by v tomto případě nebylo vhodné, protože data jsou příliš heterogenní a od předpokládaného lineárního modelu se odchyluje příliš mnoho druhů. Pokud je nejdelší gradient kratší než 3.0, bude použití lineární metody pravděpodobně vhodnější (ovšem ne nutně – viz Ter Braak et Šmilauer 1998, část 3.4 na straně 37).

3.4. Provádění ordinací - PCA: centrování a standardizace



Obrázek 3-2 Nastavení vycentrování a standardizace v průvodci setupu projektu

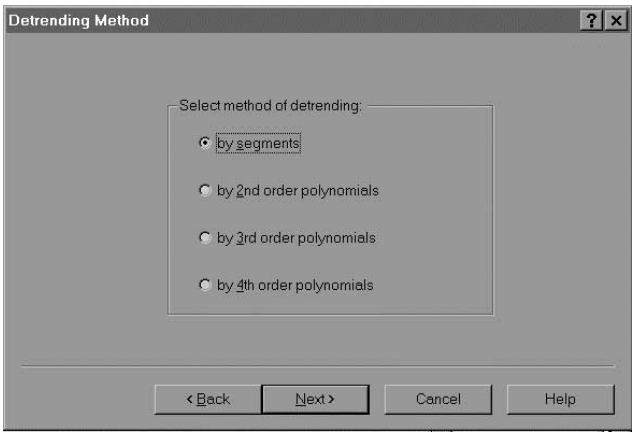
Tato stránka průvodce se zobrazí pro lineární ordinační metody (PCA nebo RDA) a týká se manipulace s tabulkami druhových dat před vlastním počítáním ordinace.

Vycentrování vzorků (volba v levé polovině okna) způsobí to, že průměr každého řádku bude roven nule. Podobně vycentrováním druhů (v pravé polovině) dosáhneme nulového průměru každého sloupce. Vycentrování druhů je nutné pro lineární metody s omezením (tj. RDA) nebo pro jakoukoliv **parciální** lineární ordinační metodu (tj. při použití kovariát).

Standardizace (vzorků nebo druhů) vyústí v to, že norma každého řádku nebo sloupce bude rovna jedné. Tato **norma** je odmocnina ze sumy čtverců hodnot v řádku nebo sloupci. Pokud použijeme jak centrování, tak standardizaci, provede se centrování jako první. Pak tedy po vycentrování a standardizaci druhů budou ve sloupcích proměnné s nulovým průměrem a jednotkovou variancí. Z toho tedy plyne, že PCA provedená na druhových datech bude odpovídat „PCA na matici korelací“ (mezi druhy).

Pokud máme v ordinačních metodách k dispozici charakteristiky prostředí (vždy v RDA a volitelně v PCA), můžeme zvolit standardizaci chybovou variancí (**error variance**). V tomto případě program Canoco odhaduje pro každý druh zvlášť varianci v druhových datech, která zůstane nevysvětlena po fitování závislosti hodnot tohoto druhu na vybraných charakteristikách prostředí (a kovariátách, pokud je máme). Převrácená hodnota této variance se pak použije jako váha druhu. Pak tedy čím lépe bude druh popsán charakteristikami prostředí, tím vyšší bude mít váhu v analýzách.

3.5. Provádění ordinací - DCA: odstraňování trendu



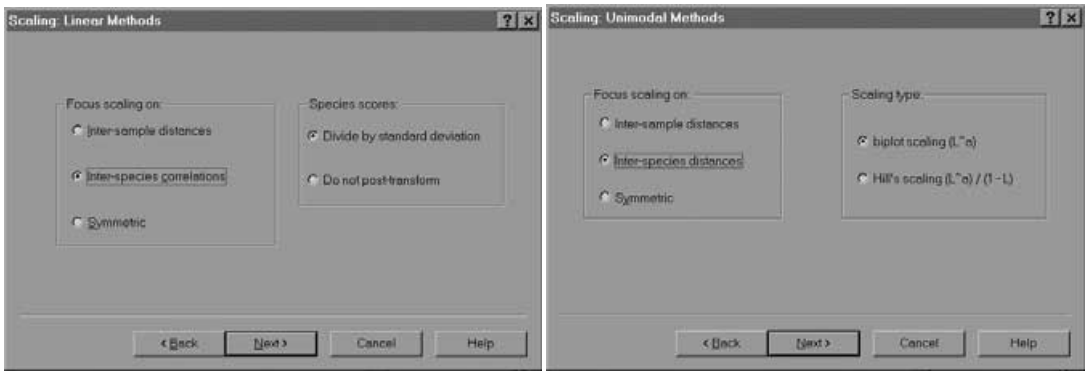
Obrázek 3-3: Výběr metody odstraňování trendu v průvodci definicí projektu

Původní metoda korespondenční analýzy (correspondence analysis) často trpí tzv. obloukovým efektem (*arch effect*). Tento efekt lze popsat tak, že skóre vzorků (a druhů) na druhé ordinační ose jsou kvadratickou funkcí těchto skóre na ose první. *Hill et Gauch (1980)* navrhli heuristickou, ale obvykle dobře fungující metodu na odstranění tohoto efektu, která se jmenuje **odstraňování trendu po segmentech** (*detrending by segments*). Tato metoda je sice některými autory kritizována (např. *Knox, 1989*), ale v současné době v podstatě nemáme žádnou lepší alternativu, jak se popsaného artefaktu zbavit. Odstraňování trendu po segmentech se nedoporučuje pro unimodální ordinační metody, kde jsou užívány kovariáty nebo charakteristiky prostředí. V těchto případech použijeme (pokud to ale opravdu potřebujeme!) **odstraňování trendu polynomy** (*detrending by polynomials*). Na tomto místě doporučíme čtenářům, aby se podívali do manuálu pro Canoco for Windows na detaily o použití polynomů druhého, třetího nebo čtvrtého stupně.

Pro unimodální ordinace s omezením není obvykle detrendování nutné. Pokud se v CCA obloukový efekt přeci jen objeví, je to známkou určité nadbytečnosti v souboru zvolených charakteristik prostředí. Mohou zde být dvě nebo více charakteristik prostředí,

které jsou navzájem silně korelovany (ať už negativně nebo pozitivně). Pokud z takové skupiny ponecháme pouze jedinou proměnnou, obloukový efekt zmizí. Výběr charakteristik prostředí, které jsou mezi sebou korelovány jen minimálně, lze provést postupnou selekcí charakteristik prostředí (*forward selection of environmental variables*) v programu Canoco for Windows.

3.6. Provádění ordinací – škálování ordinačních skóre



Obrázek 3-4 Nastavení škálování pro lineární a unimodální metody v průvodci definicí projektu

Nejdůležitějším výsledkem ordinačních metod je ordinační diagram. Teoreticky bychom z něj měli být schopni zrekonstruovat (s určitou chybou) nejen tabulku primárních dat (druhová data), ale i matici (ne)podobností mezi vzorky a korelační matici druhů. Obvykle se o zrekonstruování všech dat nepokoušíme (protože je máme k dispozici), tento postup ale přesto částečně používáme při interpretaci výsledků a vymýšlení zajímavých vědeckých hypotéz. Přesnost závěrů o podobnosti druhů, vztazích mezi druhy a / nebo charakteristikami prostředí závisí z části na relativních měřítcích (škálách) na jednotlivých ordinačních osách. Jeden druh škálování je výhodnější, pokud se zaměříme na vztahy mezi vzorky, jiný pak v případě, že interpretujeme vztahy mezi druhy.

Nastavení je podobné jak pro lineární, tak pro unimodální ordinační metody. Na počátku se musíme rozhodnout, zda se při interpretacích zaměříme na vzorky (včetně srovnání tříd vzorků, tak, jak je popisují nominální charakteristiky prostředí), nebo na druhy.

Potom se v případě lineárních modelů musíme rozhodnout, zda chceme, aby se rozdíly v četnosti jednotlivých druhů odrážely v délce jejich šipek (dominantní druhy by pak měly šipky delší než druhy vzácnější), nebo zda chceme, aby byl každý druh zrelativizován. Zrelativizování je vhodné pro tzv. korelační projekční diagramy (*correlation biplots*).

V případě unimodálních ordinačních metod bychom měli vybrat metodu interpretace ordinačních diagramů. Pro data s velmi dlouhými kompozičními gradienty (s velkou beta diverzitou vzorků) je vhodná interpretační metoda "pravidla vzdálenosti" (*distance rule)* a Hillovo škálování (*Hill' scaling*). V ostatních případech dá metoda projekčního škálování výsledky, které se lépe kvantitativně interpretují.

3.7. Spuštění CanoDraw 3.1

Nová verze programového balíku Canoco (Canoco for Windows 4.0) nabízí uživatelům další program – CanoPost, který umožňuje podstatná vylepšení diagramů vytvořených programem CanoDraw. Proto je záhodno soustředit se při používání programu CanoDraw (verze 3.1) v tomto prostředí pouze na vytvoření **obsahu** diagramů. Vlastní úpravu vzhledu je lépe odložit a udělat ji až v CanoPostu (viz část 3.8).

Program CanoDraw lze spustit přímo z programu Canoco for Windows tlačítkem **CanoDraw**, které je umístěno v okně Project view. Nicméně toto tlačítko není vždy aktivní, protože Canoco počítá s omezeními danými programem CanoDraw. CanoDraw 3.1 je totiž programem běžícím pod DOSem, takže nemůže využívat celou paměť, kterou mají k dispozici Windows. Proto s ním můžeme vizualizovat analýzy nanejvýš pro 500 vzorků, 500 druhů a 30 nebo méně charakteristik prostředí. Dalším omezením je, že CanoDraw neotevře takovou analýzu programu Canoco, kde byl explicitně proveden přímý výběr charakteristik prostředí.[+]

Pokud spustíme CanoDraw 3.1 z programu Canoco for Windows, spouští se s několika parametry, které určují:

- jméno souboru projektu v programu Canoco. Po otevření CanoDraw (pracuje pouze v celoobrazovkovém grafickém režimu DOSu) tedy už není potřeba otevírat projektový soubor (s příponou .CON), protože ten je již otevřen. CanoDraw také najde výsledkový souboru ("solution" file) s výsledky ordinací – samozřejmě pouze za předpokladu, že nenastanou nějaké komplikace.
- typ výstupního zařízení jako Postscriptovou tiskárnu. Toto nastavení je důležité pro uložení grafů kompatibilních s programem CanoPost. Programu CanoDraw se tento parametr nepředá, pouze pokud je cesta k souboru, ve kterém je projekt uložen, příliš dlouhá. Příkazová řádka používaná ke spuštění CanoDraw je omezena na 127 znaků a to zahrnuje celou cestu k programu CanoDraw. Jediná změna, kterou musíme v nastavení programu CanoDraw udělat explicitně, je změna výstupu z prvního paralelního portu do souboru.

Jak už jsme se zmínili dříve, je vhodné se při používání programu CanoDraw soustředit pouze na obsah grafů: na rozsah os diagramu, kterou typ objektů zobrazit, zda ke všem položkám zobrazovat jejich popisky, atd. V této fázi není třeba se starat o barvu a typ symbolů nebo čar, ani o přesnou pozici popisů – to vše se může upravit až později v programu CanoPost.

První a hlavní rozhodnutí, které však musíme učinit, je o typu diagramu, který chceme vytvořit. Pokud shrnujeme výsůedly projektu, kde byly použity jak druhy (složení společenstva), tak data o prostředí, bude nejvhodnější počáteční volba projekčního diagramu s druhy i charakteristikami prostředí. Před jeho vytvořením se však musíme rozhodnout o jeho obsahu:

- kvalitativní vysvětlující proměnné jsou nejlépe znázorněny pomocí centroidů pro jednotlivé indikátorové proměnné. Symboly pak vyznačují centroidy ("těžiště") pro skóre vzorků, které patří k příslušné třídě této kvalitativní proměnné. CanoDraw používá tato skóre pro proměnné označené jako nominální. Můžeme je vybrat příkazem **File/Nominal env. variables.**
- při práci s druhovou skladbou společenstva obvykle není vhodné zobrazovat skóre všech druhů. Některé druhy jsou v datech tak vzácné, že o jejich ekologických preferencích nemáme dostatek informací. Také jiné druhy mohou být našimi vysvětlujícími proměnnými charakterizovány příliš slabě. Výběr druhů pro ordinační diagramy se obvykle snažíme definovat pomocí kombinace dvou kritérií. Obě jsou dostupná z dialogového okna vyvolaného příkazem **Options/Restrictions/Rules of visibility**. **Minimum fit** vyjadřuje nejmenší procento variability v hodnotách druhů, které je

[+] Canoco for Windows zde nabízí určité řešení: nejdříve provedeme analýzu s "ručním" (interaktivním) postupným výběrem. Potom nám Canoco nabídne, učinit výsledky výběru explicitními. To znamená, že proces výběru je z nastavení projektu odstraněn a všechny charakteristiky prostředí, které v tomto projektu nebyly vybrány, jsou z analýzy explicitně vymazány. Potom musíme analýzu spustit znovu (ordinační výsledky budou stejné, jako při analýze s postupným výběrem), načež můžeme použít program CanoDraw.

vysvětleno ordinačním podprostorem, do kterého se skóre druhů promítnou (obvykle první dvě ordinační osy); tato charakteristika není použitelná pro analýzy založené na unimodálním modelu, ve kterých bylo užito odstraňování trendu po segmentech. **Minimum weight** je k dispozici pouze pro unimodální ordinační model a určuje minimální váhu (jako procento váhy druhu s váhou největší), kterou musí druh mít, aby byl zobrazen.

3.8. Úprava diagramů programem CanoPost

Práci na diagramu (obvykle ordinačním) v programu CanoPost začínáme importem souboru vytvořeném programem CanoDraw 3.1. Takový soubor má obvykle příponu **.psc**. Potom můžeme změnit vzhled diagramu a upravenou verzi uložit ve clastním formátu programu CanoPost (přípona **.cps**). Soubor lze později v programu Canopost znovu otevřít.

Při úpravě diagramů je nejlépe začít od globálních vlastností před tím, než začneme upravovat detaily. Nejdřív bychom se měli rozhodnout o výsledné velikosti diagramu. Pro její vyjádření se používá rozměr vytištěného obrázku, i přes to, že někdy má být konečným výsledkem pouze obrázek vložený do elektronického dokumentu. CanoPost zakládá odhad velikosti tisku na default nastavení tiskárny ve Windows, můžeme ji ale měnit také příkazem **File/Page setup...** Velikost výstupu se v CanoPostu zobrazí jako bílý obdélník, do něhož se graf kreslí. Je vhodné vyplnit tento prostor grafem, jak jen to jde nejvíc. K tomuto účelu slouží kombinace dvou příkazů (**Edit/Scale graph** a **Edit/Shift graph**), které lze vyvolat rovněž klávesovými zkratkami.

Poté, co uděláme tyto zásadní změny, můžeme upravovat velikost symbolů a typ a velikost písma popisků. Teprve potom má smysl upravovat umístění popisků. Snažíme se je umístit tak, abychom minimalizovali jejich překryvy a zlepšili tak čitelnost obsahu grafu.

3.9. Nové analýzy, které poskytují nové pohledy na soubory dat

Práce s mnohorozměrnými daty má obvykle iterativní charakter. Výsledky úvodních analýz mohou například naznačit, že by bylo vhodné upravit vysvětlující proměnné. Úprava se týká výběru té správné kombinace (vysvětlujících) charakteristik prostředí podle výsledků postupného výběru těchto charakteristik v programu Canoco.

Speciálním případem těchto výběrů jsou kvalitativní proměnné. Každá jejich hladina se posuzuje nezávisle na ostatních (jako samostatná indikátorová proměnná), a díky tomu máme možnost zjistit, která z nich ovlivňuje složení společenstva nejvíce.

Podíváme-li se dobře na vztah vysvětlujících proměnných k ordinačním osám a k jednotlivým druhům, můžeme charakteristiky prostředí transformovat tak, abychom maximalizovali jejich lineární vztah ke gradientu vytvořeném ordinačním modelem.

V některých případech se na ordinačním diagramu objeví dvě nebo více skupin vzorků nápadně oddělených díky rozdílnému druhovému složení. Potom se hodí výsledky doplnit dalšími samostatnými analýzami každé takové skupiny. V přímé gradientové analýze se pak výsledný obrázek často výrazně změní: charakteristiky prostředí, které vystihují rozdíly **mezi** jednotlivými skupinami se totiž často liší od proměnných, které se uplatňují **v rámci** těchto skupin.

Dalším typem jsou analýzy následující po zjištění, že určitá charakteristika prostředí vysvětluje část struktury v druhových datech. Této proměnné můžeme přidělit statut kovariáty a potom testovat dodatečný vliv ostatních potenciálně zajímavých vysvětlujících proměnných (dílčí přímá gradientová analýza, *partial direct gradient analysis*) nebo se

můžeme podívat na "zbytkovou" variabilitu a snažit se vysvětlit pattern, který takto najdeme (dílčí nepřímá gradientová analýza).

3.10. Lineární diskriminační analýza

(Fisherova) lineární diskriminační analýza (LDA), které se také říká Canonical Variate Analysis (CVA), je metodou, s jejíž pomocí hledáme skóre vzorků, vyjádřené lineární kombinací vysvětlujících proměnných, které (v určitém smyslu) optimálně oddělí předem nadefinované skupiny. Metodu můžeme provést v rámci programu Canoco, dokonce s určitými doplňkovými funkcemi, které v klasických implementacích této metody nejsou k dispozici.

Pro provedení LDA v programu Canoco je nutné, aby klasifikace vzorků byla použita jako druhová data. Každá třída je zastoupena jednou proměnnou ("druhem") a vzorky do ní patřící mají hodnotu 1.0, zatímco pro všechny ostatní proměnné ("druhy") mají tyto vzorky hodnotu nulovou. Popsané kódování odpovídá "*crisp*" variantě, pro klasickou diskriminační analýzu. Zakódování lze ale provést i "*fuzzy*" způsobem, kdy (některé) vzorky mohou patřit do několika skupin současně. Taková situace nastává při skutečné přítomnosti vzorku v několika skupinách nebo i když si pouze nejsme jisti, do které skupiny by vlastně vzorek měl patřit. Jedinou podmínkou fuzzy kódování je, aby součet všech hodnot pro jeden vzorek byl roven jedné.

Proměnné, které pro diskriminaci chceme použít, vstupují do programu Canoco jako charakteristiky prostředí. Použijeme kanonickou korespondenční analýzu (CCA) s Hillovým škálováním a se zaměřením na mezidruhové vzdálenosti (hodnota -2 v console verzi Canoca).

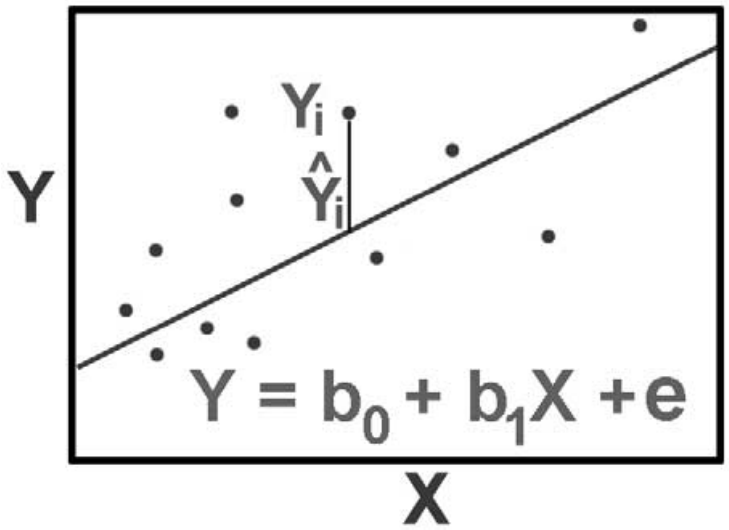
Výraznou výhodou LDA v programu Canoco je možnost provádět **parciální** diskriminační analýzu. V této analýze hledáme (navíc k již známým rozlišujícím proměnným) další vysvětlující proměnné, které nám umožní rozlišení daných tříd.

Další důležité rysy LDA v programu Canoco jsou schopnost vybírat diskriminační (vysvětlující) proměnné na základě postupného výběru a dále možnost testovat diskriminační schopnosti jednotlivých proměnných neparametrickým Monte-Carlo permutačním testem.

Při kreslení výsledků se vynesou jako průměry (centroidy) jednotlivých tříd v diskriminačním prostoru skóre druhů. Jako diskriminačních skóre jednotlivých pozorování slouží skóre vzorků (**SamE**). Projekční diagram obsahující projekční skóre charakteristik prostředí (*biplot scores of environmental variables*, **BipE**) představuje tabulku průměrů jednotlivých vysvětlujících proměnných uvnitř jednotlivých tříd, zatímco regresní / kanonické koeficienty (**Regr**) charakteristik prostředí představují "zátěže" (*loadings*) jednotlivých proměnných na diskriminačních skóre.

4. Přímá gradientová analýza a Monte-Carlo permutační testy

4.1. Model mnohonásobné lineární regrese



Obrázek 4-1 Grafické znázornění jednoduchého lineárního regresního modelu

Tuto kapitolu začneme jednoduchým shrnutím klasických lineárních regresních modelů, protože jejich znalost je klíčová, pokud chceme porozumět významu přímých gradientových analýz.

Obrázek 4-1 znázorňuje jednoduchý lineární regresní model závislosti hodnot proměnné Y na hodnotách proměnné X. Jsou zde vidět rozdíly mezi skutečnými (pozorovanými) hodnotami vysvětlované proměnné Y a hodnotami předpovězenými modelem (Y se stříškou). Tomuto rozdílu se říká regresní reziduál a v obrázku 4-1 je značen jako **e**.

Všechny statistické modely (včetně modelů regresních) mají dvě důležité složky:

- první z nich se říká systematická a popisuje tu část variability vysvětlovaných proměnných, kterou můžeme vysvětlit jednou nebo několika vysvětlujícími proměnnými (prediktory) pomocí zvolené parametrické funkce. Nejjednodušším případem je lineární kombinace vysvětlujících proměnných a využívá se v (obecných) lineárních regresních modelech.
- **stochastická složka** popisuje zbývající část variability hodnot vysvětlované proměnné, kterou nebylo možno předpovědět systematickou částí modelu. Stochastická složka se obvykle definuje pomocí svých předpokládaných pravděpodobnostních a distribučních vlastností.

Kvalitu modelu posuzujeme podle množství variability popsané systématickou složkou (obvykle v poměru k nevysvětlené variabilitě = stochastické složce). Snažíme se prezentovat jen takové modely, ve kterých **každý** z prediktorů významně přispívá k jejich kvalitě. Soubor takových prediktorů můžeme vybrat pomocí postupného výběru (*stepwise selection*). Jeho nejpoužívanějším druhem je tzv. výběr s postupným přidáváním (*forward selection*). Zde začínáme s nulovým modelem bez prediktorů, čímž předpokládáme, že variabilitu vysvětlované proměnné nelze předpovědět, takže ji popisuje jen stochastická složka. Potom vybereme z dostupných proměnných jediný prediktor - ten, který v regresním modelu vysvětlí nejvíc variability. Ale i když zvolíme ten nejlepší prediktor, může se stát, že

jeho příspěvek je dán pouhou náhodou: když totiž náhodně přeházíme hodnoty tohoto prediktoru, stejně i takto nesmyslné hodnoty vysvětlí nenulové množství variability vysvětlované proměnné. Proto musíme příspěvek uvažovaných kandidátů na prediktor testovat. Pokud je příspěvek vybraného prediktoru signifikantní, opakujeme celý proces a snažíme se mezi zbývajícími proměnnými najít další dobrý prediktor. Opět otestujeme jeho příspěvek a končíme v okamžiku, kdy "nejlepší" ze zbývajících kandidátů již není "dostatečně dobrý".

4.2. Ordinační model s omezením (constrained model)

V kapitole 2 jsme si definovali lineární a unimodální metody **nepřímé gradientové analýzy (indirect gradient analysis)** (PCA, resp. CA) jako metody hledající jeden nebo více (vzájemně nezávislých) gradientů, které budou "optimálními" prediktory v regresních modelech lineární či unimodální odpovědi druhů. Optimalita je omezena předpoklady těchto metod a hodnotí se přes všechny druhy v primárních datech.

Metody **přímé gradientové analýzy (direct gradient analysis)** (zvané též *constrained* nebo *canonical ordination methods*) se snaží o totéž, ale gradienty, které je těmto metodám "dovoleno najít", jsou více omezené. Tyto gradienty musí být lineární kombinací předložených vysvětlujících proměnných (charakteristik prostředí). Snažíme se tedy vysvětlit abundanci (všech) jednotlivých druhů pomocí složených proměnných, ale tyto proměnné jsou definovány na základě hodnot pozorovaných charakteristik.

V tom se podobají metody přímé gradientová analýzy mnohorozměrné násobné regresi. V takovéto regresi s **m** vysvětlovanými proměnnými (*species* v programu Canoco) a **p** prediktory (charakteristikami prostředí v programu Canoco) musíme z dat odhadnout **m*p** parametrů (regresních koeficientů). Taková situace však v ordinacích s omezením nenastává. Zde je vliv prediktorů na vysvětlované proměnné prezentován přes několik "zprostředkujících" gradientů - ordinačních os, zde nazývaných **kanonické osy** (*canonical axes, constrained axes*). Existuje vždy tolik kanonických os, kolik je nezávislých **vysvětlujících** proměnných.

V programu Canoco velmi často provádíme tzv. **dílčí (parciální)** analýzy, kde kromě (vysvětlujících) charakteristik prostředí zadáváme ještě tzv. **kovariáty** (*covariables*). Ty zastupují vlivy, se kterými chceme počítat, a z řešení ordinačního modelu je oddělit. Kovariáty se takto používají i v analýzách variance. Tam bývají krom faktoriálních ještě kvantitativní kovariáty. V regresních analýzách *sensu stricto* se pojem kovariáta příliš nepoužívá, ale rozdíl mezi nimi a "skutečnými" vysvětlujícími proměnnými je pouze ve způsobu, jakým se na ně díváme. Obě jsou v regresních (a ordinačních) modelech vysvětlujícími proměnnými a liší se pouze rolí, kterou jim přisoudíme.

4.3. RDA: PCA s omezením

Téma předcházející části si budeme ilustrovat na příkladě redundanční analýzy (RDA, *redundancy analysis*), což je omezená forma lineární ordinační metody - analýzy hlavních komponent (PCA, *principal component analysis*). Použijeme velmi jednoduché podmínky - budeme se snažit najít pouze první ordinační osu (první hlavní komponentu) a budeme užívat jen dvě charakteristiky prostředí ($\mathbf{z_1}$ a $\mathbf{z_2}$) pro omezení ordinačních os RDA.

Jak metoda PCA, tak RDA se snaží najít hodnoty nové proměnné (pojmenujeme si ji **x**), která by byla "optimálním" prediktorem pro hodnoty všech druhů. Hodnota této proměnné pro **i**-tý vzorek je $\mathbf{x_i}$ a používáme ji pro předpověď hodnoty (abundance, pokryvnosti, atd.) **k**-tého druhu v **i**-tém vzorku podle následujícího vzorce:

$$y_{ik} = b_{0k} + b_{1k}x_i + e_{ik}$$

Tady musí obě metody odhadnout dvě sady parametrů: hodnoty $\mathbf{x_i}$, což jsou **skóre vzorků** na první ordinační ose a regresní koeficienty pro každý druh ($\mathbf{b_{1k}}$), které na první ordinační ose zastupují **skóre druhů**. Další parametr pro každý druh ($\mathbf{b_{0k}}$) je průsečík (*intercept*) hledané regresní přímky a jeho odhadu se můžeme zbavit vycentrováním primárních dat přes druhy (viz kapitola 3.4).

Tady podobnost PCA a RDA končí, protože v omezené metodě podléhají hodnoty skóre vzorků další omezující podmínce: definují se zde jako lineární kombinace vysvětlujících proměnných, které jsou v našem případě dvě, takže toto omezení lze formulovat následovně*:

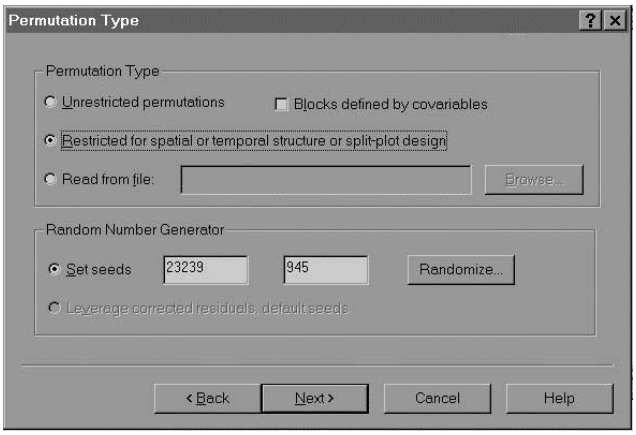
$$x_i = c_1 z_{i1} + c_2 z_{i2}$$

Oba předchozí vztahy můžeme sloučit do jediného, který ještě více zvýrazní podobnost omezených ordinací s mnohorozměrnou násobnou regresí:

$$y_{ik} = b_{0k} + b_{1k}c_1 z_{i1} + b_{1k}c_2 z_{i2} + e_{ik}$$

Součiny $\mathbf{b_k * c_j}$ jsou vlastně koeficienty mnohorozměrného násobného regresního modelu, ale tyto koeficienty jsou omezené již ze své definice: jsou definovány pomocí dvou menších skupin druhových skóre a skóre charakteristik prostředí.

4.4. Monte Carlo permutační test: úvod



Obrázek 4-2 Úvodní strana průvodce definicí projektu pro volbu typu permutací

Program Canoco umí testovat statistickou významnost omezených ordinačních modelů, popsaných v předchozí kapitole, pomocí **Monte-Carlo permutačních testů** (*Monte Carlo permutation tests*). Tyto statistické testy se vztahují k nulové hypotéze, že primární (druhová) data jsou nezávislá na vysvětlujících proměnných. Canoco disponuje četnými způsoby konkrétního nastavení testů pro data s určitou prostorovou, časovou a logickou vnitřní strukturou, v závislosti na uspořádání pokusu a odběru vzorků. Na obrázku 4-2 je

* Všimněte si, že parametry, které tu odhadujeme ($\mathbf{c_j}$ pro j-tou charakteristiku prostředí), neodpovídají těm skóre charakteristik prostředí, která jsou obvykle vynášena do ordinačního diagramu. Reprezentují spíš **Regr** skóre z výstupu programu Canoco. Skóre, která jsou normálně vynášena (**BipE** - projekční skóre charakteristik prostředí, *biplot scores of environmental variables*), jsou regresními koeficienty regresního modelu, kde jsou skóre vzorků (x_i) fitována samostatně pro jednotlivé vysvětlující proměnné. Odpovídají tedy marginálním (nezávislým) efektům jednotlivých charakteristik prostředí.

první stránka ze série stran průvodce definicí projektu, které se vztahují k určení vlastností permutačních testů.

4.5. Model nulové hypotézy

Základem permutačních testů v programu Canoco je nulová hypotéza o nezávislosti mezi odpovídajícími řádky matice druhových dat a matice environmentálních dat (vztahujícím se k témuž vzorku). Program Canoco sice zde popsaným postupem napracuje, my ho však k ozřejmení permutačních testů použijeme.

- Začneme náhodným přeuspořádáním (permutací) vzorků (řádků) v tabulce environmentálních dat, přičemž druhová data ponecháme netknutá. Pokud platí nulová hypotéza, bude jakákoli takto vzniklá kombinace hodnot druhů a hodnot charakteristik prostředí stejně pravděpodobná jako "skutečný" soubor dat.
- Pro každý takto vytvořený soubor dat spočteme omezený ordinační model a jeho kvalitu vyhodnotíme podobně jako při posuzování kvality regresních modelů. V regresním modelu používáme F statistiku, což je poměr variability vysvětlené regresním modelem (dělené počtem parametrů modelu) a variability nevysvětlené (dělené počtem reziduálních stupňů volnosti). V případě omezených ordinačních metod používáme podobnou statistiku, jejíž podrobnější popis je v následující sekci.
- Hodnotu statistiky v každé permutaci zaznamenáme. Rozložení těchto hodnot určuje rozložení této statistiky v případě platnosti nulové hypotézy. Nulovou hypotézu zamítáme, pokud je velmi nepravděpodobné, že by hodnota testové statistiky skutečných dat (bez jakékoli permutace řádků tabulky environmentálních dat) z tohoto rozložení mohla pocházet (pravděpodobně bude mnohem vyšší, což odpovídá mnohem kvalitnějšímu ordinačnímu modelu). Pravděpodobnost, že hodnota statistiky testu "odvozená z dat" přeci jen pochází z rozložení spočteného pro nulovou hypotézu, pak odpovídá chybě Typu I, tj. zamítnutí správné nulové hypotézy.

Skutečný algoritmus nemění vlastní tabulku environmentálních dat. Pracuje s reziduály ze spočtených závislostí druhových dat na charakteristikách prostředí, na základě každé takové permutace vytvoří "novou" tabulku dat. Celý algoritmus je však ještě mnohem složitější, vstoupí-li do hry kovariáty.

4.6. Testovací statistiky

V předchozí kapitole jsme si popsali obecný princip permutačních testů a nyní je čas pro popis možných výběrů testovacích statistik pro tyto testy. Už jsme se zmínili, že tyto statistiky odpovídají F statistice, která se používá v parametrickém testování významnosti regresních modelů. Výběr správné definice této statistiky v omezených ordinacích je ale kvůli mnohorozměrnosti získaného řešení složitější. Obecným rysem je, že variabilita vysvětlované proměnné (druhových dat) popsaná vysvětlujícími proměnnými (charakteristikami prostředí) je rozložena mezi více kanonických os. Jejich relativní význam sice od první osy k poslední klesá, ale jen zřídka kdy si můžeme dovolit všechny osy mimo první vypustit. Z řečeného vyplývá, že variabilitu vyjádříme buď pomocí sumy všech kanonických os nebo užijeme jen jednou z nich, obvykle první. To odpovídá i tomu, že v programu Canoco (verze 3.x i 4.x) jsou dvě testovací statistiky a jim odpovídající dva permutační testy:

- Test využívající pouze první kanonickou osu má statistiku definovanou následujícím vztahem:

$$F_{\lambda} = \frac{\lambda_1}{RSS_{X+1}/(n-p-q)}$$

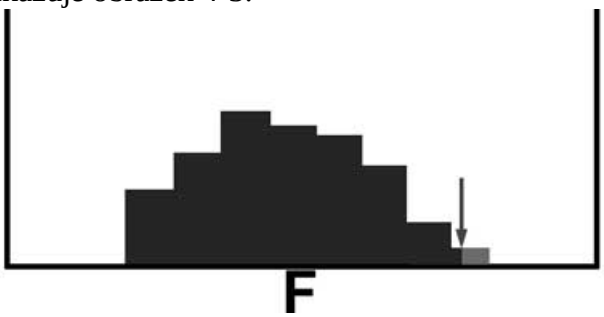
variabilita vysvětlená první (kanonickou) osou je vyjádřena jejím charakteristickým číslem (*eigenvalue*, λ_1). Reziduální suma čtverců (*residual sum of squares*, RSS) odpovídá rozdílu mezi celkovou variabilitou v druhových dat a množstvím variability vysvětlené první kanonickou osou (a kovariátami, pokud jsou tyto v analýze přítomny). Počet kovariát je označen jako *q*.

- Test založený na sumě kanonických os. Zde pracujeme s celkovým vlivem *p* vysvětlujících proměnných, popsaným (až) *p* kanonickými osami:

$$F_{trace} = \frac{\sum_{i=1}^{p} \lambda_i / p}{RSS_{X+Z}/(n-p-q)}$$

Zkratka RSS se zde vztahuje k rozdílu mezi celkovou variabilitou druhových dat a sumě charakteristických čísel všech kanonických ordinačních os.

Jak už jsme se zmínili v předchozí kapitole je hodnota testové statistiky pocházející z původních dat porovnávána s rozložením této statistiky platné za předpokladu platnosti nulové hypotézy. To ukazuje obrázek 4-3.



Obrázek 4-3: Rozložení hodnot F-statistiky z Monte Carlo permutačního testu, porovnávané s hodnotou F-statistiky ze skutečného uspořádání dat

Na tomto obrázku vidíme histogram, který ukazuje tvar rozložení testové statistiky. Svislá šipka je v místě hodnoty spočtené pro "skutečná" data. Permutace, pro které vyšla hodnota vyšší, pak mluví proti zamítnutí nulové hypotézy. Chybu Typu I spočteme takto:

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

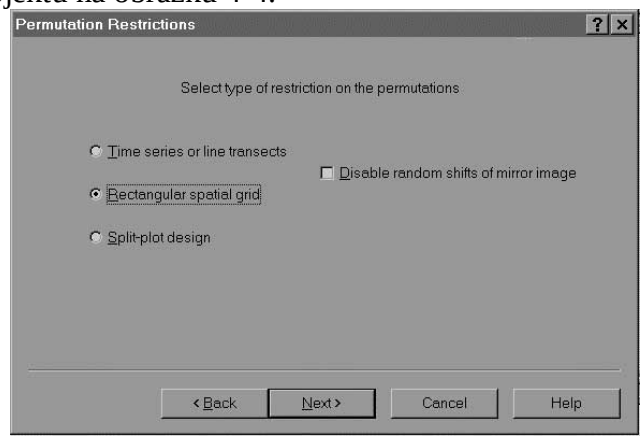
kde $\mathbf{n_x}$ je počet permutací s výsledkem stejným nebo vyšším než dají "skutečná" data a **N** je celkový počet provedených permutací. Hodnotu 1 přičítáme jak k čitateli, tak ke jmenovateli, protože (v případě platnosti nulové hypotézy) pochází z rozložení daného nulovou hypotézou i hodnota statistiky ze skutečně pozorovaných dat, a tím i ona přispívá k nezamítnutí nulové hypotézy. Tento specifický rys vzorce se odráží v obvyklém výběru počtu permutací (99, 199 nebo 499).

4.7. Prostorová a časová omezení

Permutační testy, jak jsme si je popsali v předchozí části, lze takto provádět jen tehdy, nemají-li data žádnou vnitřní strukturu, tedy hlavně jsou-li vzorky odebírány náhodně

a nezávisle. Jen v tomto případě můžeme vzorky přehazovat (permutovat) zcela náhodně, protože v případě platnosti nulové hypotézy je stejně pravděpodobné, že hodnoty charakteristik prostředí (vysvětlujících proměnných) každého vzorku odpovídají druhovým datům jakéhokoli jiného vzorku.

To už ale neplatí, jsou-li si některé vzorky navzájem "příbuznější" než jiné. Nejčastější tři typy vnitřních struktur dat, se kterými pracuje Canoco, jsou ve volbách průvodce definicí projektu na obrázku 4-4.



Obrázek 4-4 Stránka průvodce definicí projektu s výběrem omezení permutačních testů

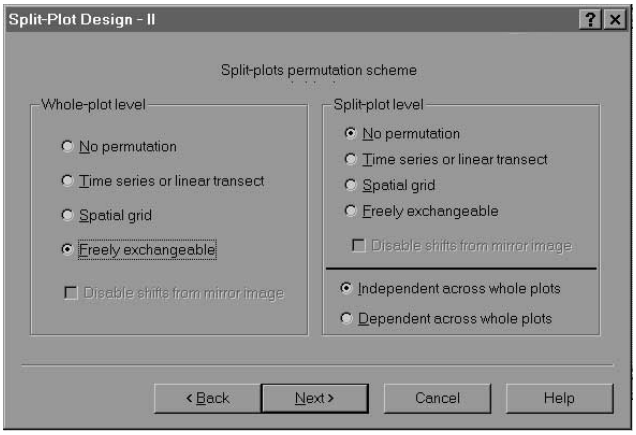
Vzorky mohou být uspořádány podél lineárního transektu nebo podél časové osy. Potom je ale nemůžeme prohazovat náhodně, protože do hry vstupuje autokorelace mezi jednotlivými pozorováními, jak na úrovni druhových, tak environmentálních dat. Tento korelační pattern bychom při testu neměli porušit, protože naše hypotéza se zabývá vztahem mezi druhy a charakteristikami prostředí, nikoliv mezi těmito pozorováními samotnými. Canoco se s tímto problémem vyrovnává tak, že (pomyslně) "zakroutí" sekvenci vzorků jak v druhových, tak v environmentálních datech do kruhové formy a přeuspořádání pak provádí "rotováním" kruhu environmentálních dat vůči kruhu dat druhových. Pokud se chcete o detailech a dalších popsaných omezeních permutací dozvědět více, nahlédněte do manuálu programu Canoco.

Podobná prostorová autokorelace nastává, pokud zobecňujeme umístění vzorku na lineárním transektu do obecného umístění vzorku v prostoru. Canoco toto však nedovoluje a podporuje pouze umístění vzorku do pravoúhlé, homogenně rozmístěné mřížky.

Nejobecnější model vnitřní struktury dat je na obrázku 4-4 jako poslední položka se jménem **split-plot design**. Právě tento typ omezení permutací si popíšeme v následující části.

Všechna jmenovaná omezení můžeme aplikovat i na další hladiny bloků. Bloky se v programu Canoco obvykle definují pomocí (podmnožiny) nominálních kovariát a jsou to skupiny vzorků podobných více sobě navzájem než vzorkům z jiných bloků. V permutačních testech jsou vzorky přeuspořádávány jen v rámci těchto bloků (někdy ještě s dalšími omezeními). Když porovnáme omezený ordinační model s modelem analýzy variance, můžeme bloky chápat jako faktor s náhodným efektem (*random effect factor*), jehož vliv je pro interpretaci nezajímavý.

4.8. Omezení daná designem



Obrázek 4-5 Strana průvodce definicí projektu, která specifikuje omezení permutačních testů pro split-plot design

Omezení split-plot designu používaná v programu Canoco for Windows 4.0 nám dovolují popsat strukturu se dvěma hladinami variability (se dvěma "hladinami chyb").[†]

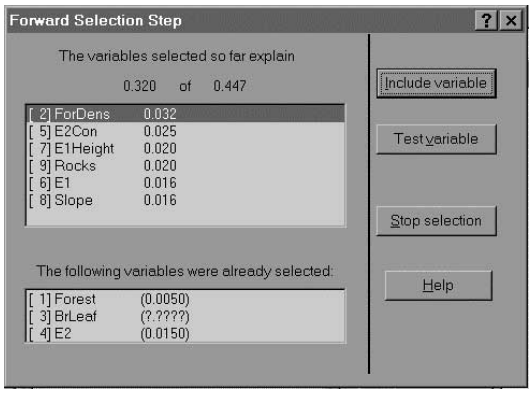
Horní hladina je reprezentována tzv. **whole-plots** ("celé plochy"). Každý z nich obsahuje stejný počet **split-plots** ("podplochy" nebo "dílčí plochy"), které jsou na nižší hladině designu. Canoco je v možnostech permutací jednotlivých hladin velmi pružné. V obou hladinách nabízí možnost nepermutovat vůbec, prostorově a časově omezené permutace, i zcela volnou zaměnitelnost. Navíc Canoco nabízí pojem závislosti uspořádání *split-plots* přes jednotlivé *whole-plots*. V tomto případě jsou v každém nezávislém permutačním testu jednotlivé permutace pro *split-plots* v rámci všech *whole-plots* stejné.

4.9. Postupný výběr modelu

Na konci sekce 4.1, jsme si podrobněji popsali přímý výběr vysvětlujících proměnných pro regresní model. Postupný výběr v programu Canoco má stejný účel i metodiku. Pro hodnocení kvality každého z potenciálních prediktorů rozšiřujícího výběr vysvětlujících proměnných v omezeném ordinačním modelu používá Monte Carlo permutační test.

Pokud zvolíme interaktivní ("manuální") postupný výběr, objeví se během analýzy následující okno (Obrázek 4-6).

[†] Další hladinu můžeme v některých případech přidat permutací v rámci bloků definovaných pomocí kovariát (viz předcházející sekce).



Obrázek 4-6 Dialogový box pro postupný výběr (*forward selection*) charakteristik prostředí. Otazníky u proměnné *BrLeaf* znamenají, že tato proměnná se permutačním testem během výběru netestovala.

Obrázek 4-6 ukazuje stav procedury postupného výběru s přidáváním (*forward selection*), kde již byly tři nejlepší vysvětlující proměnné (*Forest*, *BrLeaf*, *E2*) vybrány (jsou zobrazeny v dolní části okna). Hodnoty v horní části okna ukazují, že tyto tři vybrané proměnné odpovídají přibližně za 72% celkové variability vysvětlitelné všemi charakteristikami prostředí (tj. 0.320 z 0.447).

V horní části okna je seznam zbývajících "kandidátů na prediktory" seřazených podle klesajícího příspěvku, který by tyto proměnné měly po přidání k již vybraným proměnným. Vidíme, že proměnná "ForDens" je horkým kandidátem. Zvýšila by množství vysvětlené variability z 0.320 na 0.352 (0.320 + 0.032).

K posouzení, zda je tento příspěvek větší než příspěvek náhodný, můžeme použít parciální Monte Carlo permutační test. V tomto testu použijeme "kandidující" proměnnou jako jedinou vysvětlující proměnnou (čímž dostaneme ordinační model s jedinou kanonickou osou). Již vybrané charakteristiky prostředí (v tomto případě *Forest*, *BrLeaf* a *E2*) použijeme jako kovariáty, společně s jinými *a priori* vybranými kovariátami. Zamítneme-li v tomto parciálním (dílčím) testu nulovou hypotézu, můžeme testovanou proměnnou do výběru přidat.

Vlivu proměnné, který takto testujeme, se říká vliv **podmíněný** (*conditional*), nebo též částečný (*partial*). Jeho hodnota velmi silně závisí na přesném pořadí výběrů. Ale na začátku procesu přímého výběru, kdy ještě žádná charakteristika prostředí vybrána nebyla, můžeme testovat každou proměnnou zvlášť, abychom odhadli její nezávislý, **marginální vliv** (*marginal effect*). Je to variabilita v druhových datech, kterou bychom vysvětlili omezeným ordinačním modelem s touto proměnnou jako jedinou vysvětlující proměnnou. Rozpor mezi pořadím proměnných seřazených podle jejich marginálních vlivů a pořadím odpovídajícím "slepému" přímému výběru (vybíráme-li vždy nejlepšího kandidáta) je způsoben korelacemi mezi vysvětlujícími proměnnými. Kdyby byly tyto proměnné úplně lineárně nezávislé, byla by obě pořadí stejná.

Je-li hlavním důvodem postupného výběru hledání dostatečné podmnožiny vysvětlujících proměnných, která bude reprezentovat vztah mezi druhy a charakteristikami prostředí, narazíme na problém "globální" hladiny významnosti vztahující se k celému zvolenému výběru, kdy k němu přistupujeme jako k celku. Pokud pokračujeme s výběrem charakteristik prostředí tak dlouho, dokud má ten nejlepší kandidát odhad chyby Typu I *P* menší než je nějaká předem zvolená hladina významnosti α, bude ve skutečnosti "kolektivní" pravděpodobnost chyby Typu I větší než tato hladina. Nevíme přesně, jak bude tato pravděpodobnost velká, známe ale horní mez, kterou je $N_c^* \alpha$, kde N_c je maximální počet testů (porovnání) proveditelných během výběru. Vhodnou úpravou prahu hladin

významnosti každého parciálního testu je **Bonferroniho korekce** (vybírají se jen proměnné, které dosáhnou odhadu pravděpodobnosti chyby Typu I menšího než α / N_c). N_c je maximální možný počet kroků během přímého výběru (tj. obvykle počet **nezávislých** charakteristik prostředí). Použití Bonferroniho korekce je však velmi kontroverzním tématem.

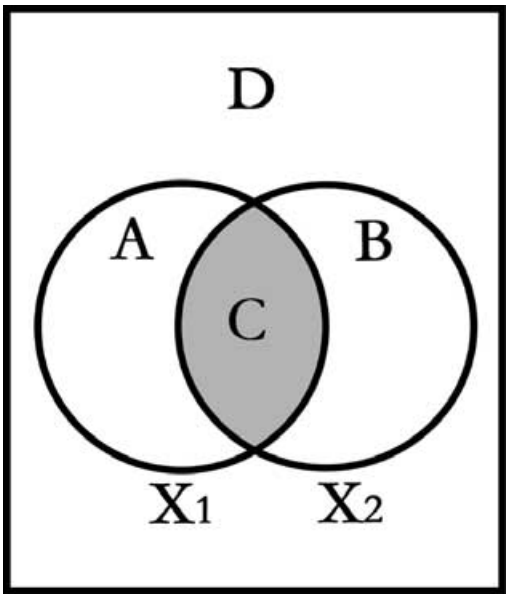
Další problém, který při procesu přímého výběru můžeme potkat, nastává, máme-li některé faktory kódovány jako indikátorové proměnné a používáme-li je jako charakteristiky prostředí. Při přímém výběru se každá indikátorová proměnná bere jako nezávislý prediktor, takže nelze zjistit příspěvek celého faktoru najednou. Je to hlavně proto, že celý faktor přispívá do omezeného ordinačního modelu více než jedním stupněm volnosti (podobně v regresním modelu přispívá faktor s **K** hladinami **K** - 1 stupni volnosti). V omezené ordinaci je třeba k vyjádření příspěvku takového faktoru **K** - 1 kanonických os. Na druhou stranu umožňuje nezávislé pojetí hladin faktoru ohodnotit míru rozdílů mezi jednotlivými skupinami vymezenými na základě takového faktoru. Toto je z části analogické procedurám mnohonásobných porovnání v analýze variance.

4.10. Rozklad variance (variance partitioning)

V předcházející sekci jsme si vysvětlili rozdíl podmíněných a marginálních vlivů jednotlivých vysvětlujících proměnných (charakteristik prostředí) na druhová data (vysvětlované proměnné). Řekli jsme si také, že nesrovnalosti v důležitosti vysvětlujících proměnných určených jejich marginálními a podmíněnými vlivy jsou způsobeny korelacemi těchto proměnných. Jakékoli dvě korelované vysvětlující proměnné sdílejí část svého vlivu na druhová data (přinejmenším ve statistickém slova smyslu). Sdílené množství vysvětlující schopnosti takové dvojice je rovno rozdílu mezi marginálním vlivem proměnné A a jejím podmíněným vlivem hodnoceným po přidání k proměnné B.

Toto je základem procedury zvané **variance partitioning** (snad lze užít český ekvivalent "rozklad variance", viz Borcard, Legendre & Drapeau, 1992), kde se však většinou neporovnávají pouze dvě vysvětlující proměnné, ale spíše dvě nebo více **skupin** charakteristik prostředí zastupujících nějaký odlišný, ekologicky interpretovatelný jev. Jako typický příklad si můžeme uvést oddělení časové a prostorové variability.

Rozklad variance si popíšeme na nejjednodušším příkladě se dvěma skupinami vysvětlujících proměnných (X_1 a X_2). V každé skupině je jedna nebo více samostatných charakteristik prostředí. Diagram na obrázku 4-7 ukazuje dělení celkové variability druhových dat na základě těchto dvou skupin proměnných.



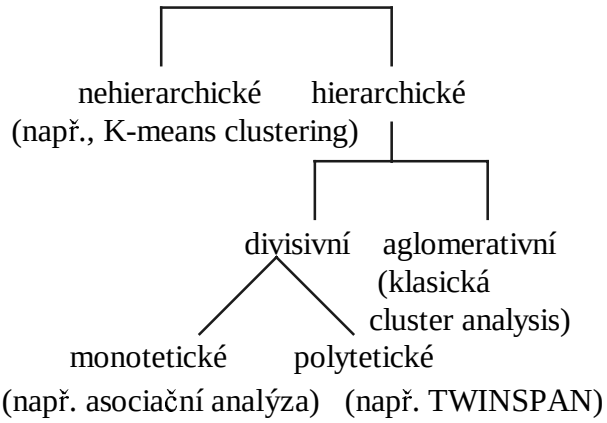
Obrázek 4-7 Rozdělení variability druhových dat do příspěvků dvou skupin charakteristik prostředí (A, B a sdílená část C) a variability reziduální (D).

Oblast označená písmenem D odpovídá reziduální variabilitě, tj. variabilitě nevysvětlené ordinačním modelem zahrnujícím skupiny X_1 a X_2 charakteristik prostředí. V části **A** je parciální vliv proměnných ze skupiny X_1, podobně, jako je v části **B** parciální vliv skupiny X_2. Vliv sdílený oběma skupinami je v části **C.** Je zřejmé, že variabilita vysvětlená skupinou X_1 (ignorujeme-li proměnné skupiny X_2) je rovna **A+C**. S odhady postoupíme dále použitím parciálních analýz.

A odhadneme z analýzy, kde proměnné z X_1 použijeme jako charakteristiky prostředí, a proměnné z X_2 jako kovariáty. Podobně **B** odhadneme jako sumu charakteristických čísel kanonických os z analýzy, kde X_2 vystupují jako charakteristiky prostředí a X_1 jako kovariáty. Potom spočteme velikost **C** odečtením sumy **A** a **B** od variability vysvětlené ordinačním modelem, ve kterém vystupují X_1 i X_2 jako vysvětlující proměnné. Pro praktický příklad užití tohoto postup viz závěr kapitoly 7.

5. Klasifikační metody

Cílem klasifikace je získat skupiny objektů (vzorků nebo druhů) vnitřně homogenní a odlišné od jiných skupin. Pokud klasifikujeme druhy, znamená homogenita podobné ekologické chování, které se projevuje podobností distribuce druhů. Klasifikační metody se obvykle rozdělují tak, jak je tomu na obrázku 5-1.



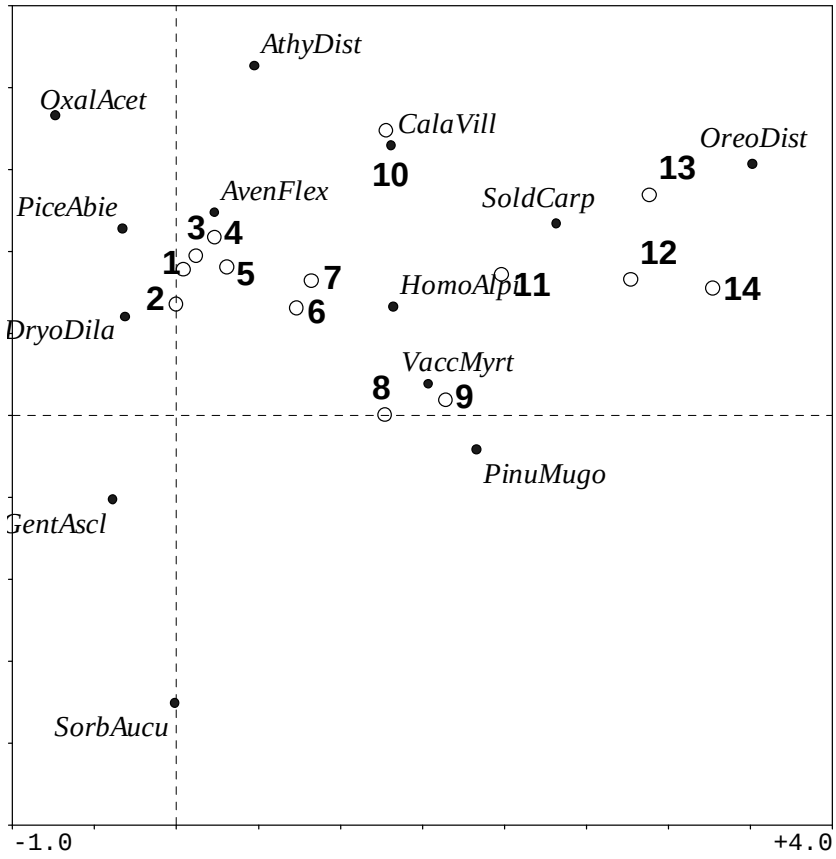
Obrázek 5-1 Typy klasifikačních metod

Původně byla numerická klasifikace považována za objektivní alternativu k subjektivní klasifikaci (např. Z-M fytocenologického systému). Objektivní je v tom smyslu, že stejná metoda dá (obvykle) stejný výsledek. Měli bychom ale mít na paměti, že výsledky numerické klasifikace jsou **vždy** závislé na výběru metody.

5.1. Soubor dat

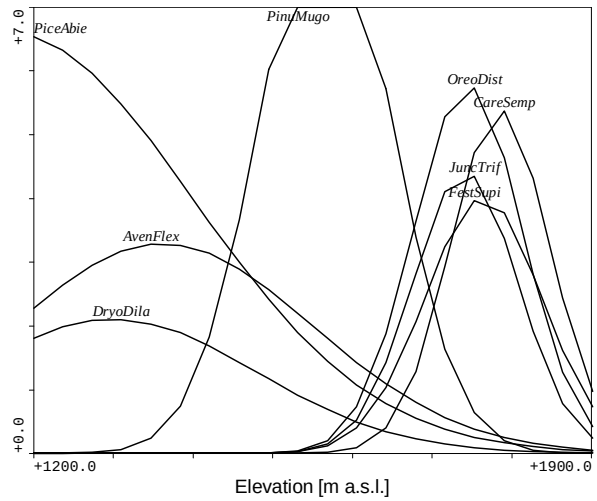
Různé možnosti klasifikace si ukážeme na datech ze 14 snímků z transektu na nadmořské výšce v Nízkých Tatrách. Snímek 1 je z nadmořské výšky 1200 m n. m., snímek 14 z 1830 m n. m. K pořízení snímků byla použita Braun-Blanquetova stupnice (r, +, 1 až 5). Pro výpočty byla pak tato stupnice převedena na hodnoty 1 až 7 (ordinální transformace podle *van der Maarel, 1979*). Data jsou ve formátu běžné fytocenologické tabulky (soubor **tatry.xls**) a data s druhovým složením byla importována pomocí WCanoImp do kondenzovaného (Cornell, CANOCO) formátu, aby bylo možné použít programy CANOCO a TWINSPAN. Data byla také importována do souboru programu Statistica.

Nejdřív se podíváme na strukturu podobností spočítanou pomocí metody DCA (*detrended correspondence analysis*). Získali jsme následující projekční diagram druhů a vzorků (*species-samples biplot*, obrázek 5-2).



Obrázek 5-2: Projekční diagram druhů a vzorků z výsledků DCA transektu na nadmořské výšce v Nízkých Tatrách. Ukázány jsou jen druhy s největším významem (nejčastěji zastoupené).

Obrázek ukazuje, že v datech je lineární variabilita, která odpovídá gradientu nadmořské výšky. Snímky 1 až 5 jsou ze smrkového lesa (charakterizovaného např. *Picea abies, Dryopteris dilatata, Avenella [=Deschampsia] flexuosa*), a snímky (11-) 12 až 14 jsou z typické alpinské louky. Mezi nimi je zóna kleče (Obrázek 5-3). Nyní se podíváme, jak bude gradient rozdělen do jednotlivých vegetačních typů různými klasifikačními metodami, a také si ukážeme, jak to technicky provést.



Obrázek 5-3: Křivky odpovědí důležitých druhů na gradient nadmořské výšky, tak jak byly spočteny programem CanoDraw při užití lineárního prediktoru 2. řádu (zobecněné lineární modely) - viz kapitola 11

5.2. Nehierarchická klasifikace (K-means clustering)

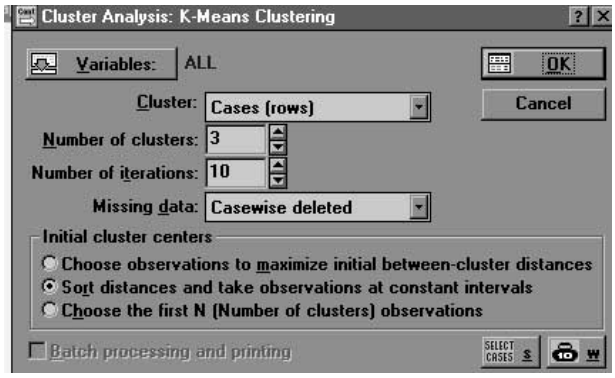
Cílem metod je vytvořit předem určený počet skupin objektů. Skupiny by měly být vnitřně homogenní a odlišné od sebe navzájem. Všechny skupiny jsou na stejné úrovni, není zde žádná hierarchie. Zde si předvedeme K-means clustering jako ukázku nehierarchické klasifikace.

Pro výpočet se používá opakovaná relokační procedura. Procedura začíná s *k* (požadovaný počet skupin) skupinám a pak přesouvá objekty tak, aby minimalizovala variabilitu uvnitř skupiny a naopak maximalizovala variabilitu mezi skupinami. Pokud jsou skupiny různé, pak ukazuje ANOVA pro jednotlivé druhy signifikantní rozdíly, takže proceduru můžeme brát jako „ANOVA in reverse“. Jedná se o tvoření skupin s "nejprůkaznějšími" ANOVA výsledky pro většinu druhů (proměnných).

Relokační procedura se ukončí, když žádný další přesun už kriteria nezlepší. Měli bychom si uvědomit, že takto můžeme získat jen extrém lokální, o kterém nemáme jistotu, že je zároveň extrémem globálním. Proto se doporučuje začínat s různými počátečními skupinami a sledovat, zda jsou výsledky těchto analýz vždy stejné.

Použijeme program Statistica (proceduru *K-means clustering*). Ve Statistice zvolte *Cluster Analysis* a pak *K-means clustering*. Na panelu pak zvolte:

Variables: *ALL* (lze vybrat také jen část a spočítat klasifikaci jen pro omezený počet druhů – např. jen pro byliny).



Obrázek 5-4 Dialog specifikující K-means clustering v programu Statistica for Windows

V následujícím panelu se budeme nejdřív ptát na *Members of each cluster & distances.* Zde bychom se měli dozvědět, že v první skupině jsou vzorky 1 až 6, ve druhé 7 až 11 a ve třetí 12 až 14 (to, že skupiny jsou tvořeny vzorky jdoucími po sobě, je způsobeno lineárním charakterem proměnlivosti našich dat; v jiném případě bychom mohli dostat skupinu se vzorky 1, 3 a 5 nebo jinou se vzorky 2 a 4). Pro každý objekt vidíme vzdálenosti od středu příslušné skupiny, např. pro skupinu 1:

Members of Cluster Number 1 (tatry.sta)
and Distances from Respective Cluster Center
Cluster contains 6 cases

	Case No. C_1	Case No. C_2	Case No. C_3	Case No. C_4	Case No. C_5	Case No. C_6
Distance	0.49637	0.577225	0.456277	0.600805	0.520278	1.017142

Vzorky jsou poměrně homogenní, jen vzorek 6 stojí trošku bokem (stejně tak jako vzorek 10 je stranou ve skupině 2). Výsledek se dobře shoduje s DCA analýzou (včetně určení vzorků "stojících stranou", tzv. *outliers*).

Teď se můžeme ptát na *Cluster means and Euclidean distances. Euclidean distances* vypovídají o podobnosti mezi jednotlivými skupinami a *Cluster means* poskytují informaci o průměrných hodnotách druhů v jednotlivých skupinách.

Vidíme, že nejpodobnější jsou si skupiny 1 a 2 a nejméně podobné jsou si 1 a 3 (přesně podle očekávání).

Euclidean Distances between Clusters
(tatry.sta)
Distances below diagonal
Squared distances above diagonal

	No. 1	No. 2	No. 3
No. 1	0	1.616661	3.178675
No. 2	1.27148	0	2.212292
No. 3	1.782884	1.487377	0

Zastoupení druhů v jednotlivých skupinách je patrné z průměrů:

Cluster Means (tatry.sta)

	Cluster No. 1	Cluster No. 2	Cluster No. 3
PICEABIE	5.416667	1.6	0
SORBAUCU	1.75	1.4	0
PINUMUGO	0.333333	6.4	1.333333
SALISILE	0	0.8	0

etc……

Picea abies je běžná ve skupině 1 (smrkový les) a chybí ve skupině 3 (alpinská louka); kleč (*Pinus mugo*) je vzácná mimo tzv. „krumbholz“ zónu, atd. Užitečné informace získáme v *Analysis of variance:*

	Between SS	df	Within SS	df	F	signif. p
PICEABIE	71.68095	2	21.40831	11	18.41552	0.000309
SORBAUCU	6.3	2	23.575	11	1.469778	0.271826
PINUMUGO	107.6571	2	15.2	11	38.9549	1.02E-05
SALISILE	2.057143	2	12.8	11	0.883929	0.440566
JUNICOMM	6.547619	2	4.666667	11	7.716836	0.00805
VACCMYRT	0.32381	2	20.03333	11	0.0889	0.915588
OXALACET	16.16667	2	19.33333	11	4.599139	0.035353
HOMOALPI	1.414286	2	4.3	11	1.808971	0.209308
SOLDHUNG	18.66666	2	7.333336	11	13.99999	0.000948
AVENFLEX	16.89524	2	4.033331	11	23.03897	0.000117
CALAVILL	0.928572	2	40.875	11	0.124945	0.88378
GENTASCL	7.295238	2	5.633332	11	7.122571	0.010368
DRYODILA	8.914286	2	4.8	11	10.21428	0.003107

etc.

Pro každý druh se počítá ANOVA, která porovnává průměry ve všech třech skupinách. Upozorňuji, že p-hodnota by se teď neměla vykládat jako chyba Typu I, protože skupiny byly vytvořeny tak, aby rozdíly mezi nimi byly co největší. Nicméně přesto může p-hodnota naznačit, které druhy se mezi skupinami výrazně liší (podle kterých druhů byly vlastně skupiny vymezeny). *Picea abies, Pinus mugo* nebo *Soldanella hungarica* se mezi jednotlivými skupinami liší velmi výrazně, zatímco *Vaccinium myrtillus* ne (byla relativně běžná podél celého transektu).

5.3. Hierarchické klasifikace

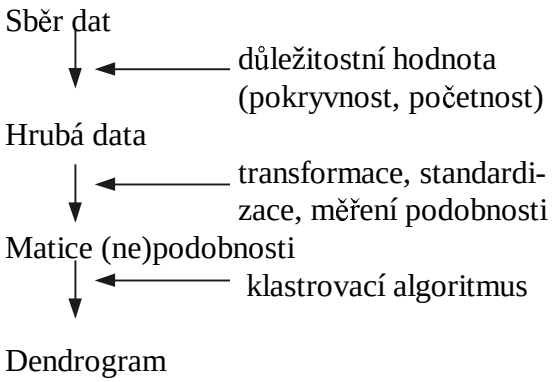
V hierarchických klasifikacích se tvoří skupiny, které obsahují podskupiny, takže tu existuje určitá hierarchie hladin. Pokud se skupiny tvoří zezdola (tedy slučováním těch nejpodobnějších vzorků), mluvíme o klasifikacích **aglomerativních**. Když klasifikace začíná s celým souborem, který se nejdříve rozdělí na dvě skupiny a ty pak na další a další, mluvíme o klasifikacích **divizivních**. Označení **klastrová analýza** (*cluster analysis*) se často používá jen pro aglomerativní metody.

Aglomerativní hierarchické klasifikace (Cluster analysis)

Cílem je vytvořit hierarchickou klasifikaci (tj. skupiny s podskupinami), která se nejčastěji prezentuje jako dendrogram. Skupiny se tvoří „zespoda“. To znamená, že nejdříve se spojí dva nejpodobnější objekty do shluku, který se potom považuje za jeden objekt, a spojování

pokračuje až do té doby, než jsou všechny objekty spojeny do jedné velké skupiny. Procedura má dva základní kroky: v prvním se spočte pro všechny páry objektů matice podobností (tato matice je symetrická a na úhlopříčce jsou buďto nuly – pro nepodobnosti – nebo čísla vyjadřující maximální možnou podobnost). Ve druhém kroku jsou objekty spojovány (*joined, amalgamated*) do skupin a podobnost všech objektů vůči těmto nově vzniklým skupinám je přepočítána. Jednotlivé algoritmy se liší ve způsobu, jakým přepočítávají podobnosti.

Měli byste vědět, že výsledek klasifikace ovlivňuje několik metodologických rozhodnutí:



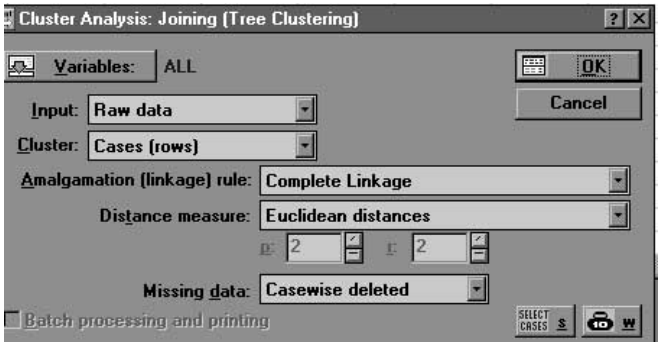
Obrázek 5-5 Naše rozhodnutí ovlivňující výsledky aglomerativních hierachických klasifikací

Aglomerativní klasifikace jsou běžnou součástí většiny statistických programů. My si ukážeme jejich použití v programu Statistika, ale pravidla jsou podobná ve většině programů. Postup je opět rozdělen do dvou kroků: výpočet matice (ne)podobnosti a potom samotné klastrování. Program Statistika dovoluje (tak jako většina programů) přímý vstup matice podobnosti. To je velmi užitečné, protože Statistika obsahuje jen velmi omezený počet metod pro určení (ne)podobnosti. Neobsahuje například metodu v ekologii používanou velmi často – procentuální podobnost (*percent similarity)*. Není však složité připravit si (např. v Excelu) jednoduché makro, které podobnosti spočítá a výslednou matici pak naimportovat do Statistiky.[+]

V základní formě je postup ve Statistice velmi jednoduchý:

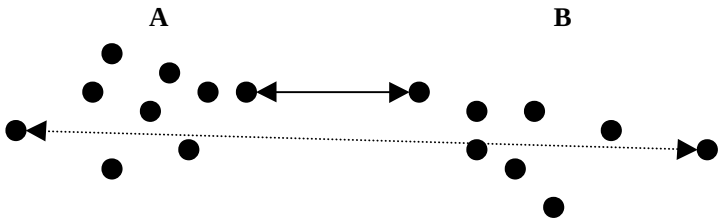
V úvodním panelu vyberte *Joining (tree clustering)*: pro shlukovou analýzu ploch použijte nastavení jako na obrázku 5-6.

[+] Soubor však musí mít všechny náležitosti Statistikou vyžadované – tedy symetrickou čtvercovou matici, kde jména sloupců jsou stejná jako jména řádků, plus další čtyři řádky s průměry a standartními odchylkami proměnných (toto je nutné jen pro matice korelace a kovariance), v prvním sloupci třetího řádku počet porovnávaných položek a konečně v prvním sloupci čtvrtého řádku typ matice: 1 = korelace, 2 = podobnost, 3 = nepodobnost, 4 = kovariance.



Obrázek 5-6 Dialog hierarchického, aglomerativního klastrování v programu Statistica

Pro *Variables* vyberte *ALL* (analýzu lze provést i pro omezený počet proměnných). *Raw data* znamená, že ve vašem souboru jsou hrubá data, nikoli matice podobnosti. Pro spojování vybereme proceduru *Complete linkage.* Máme na výběr z několika alternativ a naše volba má výrazný vliv na výsledný dendrogram. Jsou tu tzv. „short hand" metody (např. *Single Linkage*), které definují vzdálenost skupin jako vzdálenost mezi nejbližšími body těchto skupin[*]. Výsledkem takových metod je dendrogram charakteristický velkým zřetězováním (*chaining*). „Long hand" metody, např. *Complete Linkage* definují vzdálenost mezi skupinami podle nejvzdálenějších bodů skupin, výsledné skupiny jsou pak kompaktní a přibližně stejně velké (obrázek 5-7).



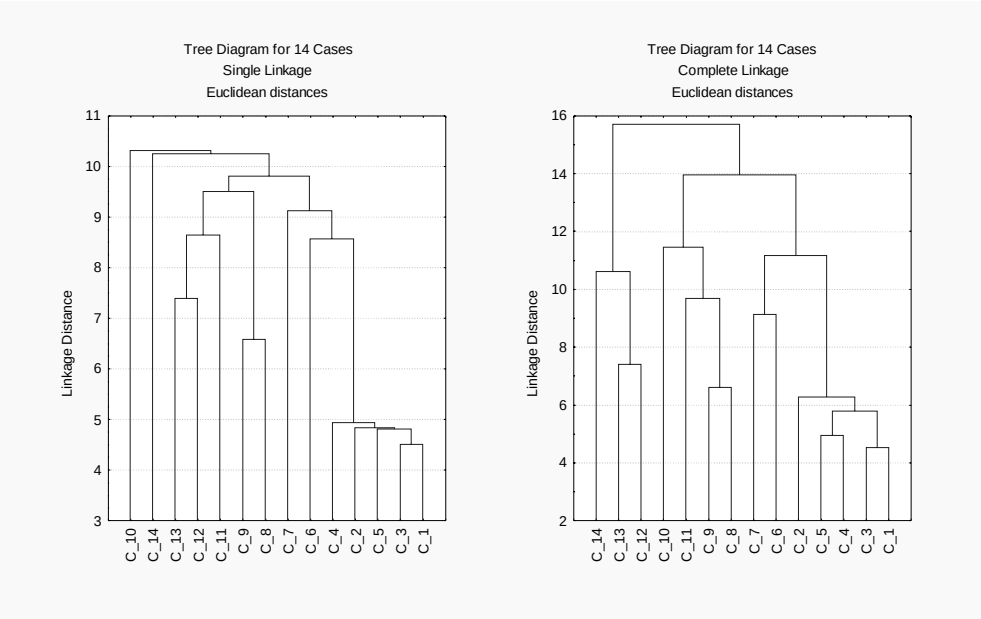
Obrázek 5-7: Vzdálenosti dvou skupin (A a B) definované metodami *single linkage* (plná čára) a *complete linkage* (přerušovaná čára).

Existuje ještě mnoho dalších metod, které jsou někde mezi dvěma popsanými případy: velmi populární bývaly *average linkage* metody. Tento termín však nebyl používán důsledně. Výraz *unweighted-pair groups method* (UPGMA) popisuje nejčastější variantu. „Short hand" metody se v ekologii obvykle moc neuplatňují. Srovnáním dvou výsledných dendrogramů (obrázek 5-8) zjistíte, že lze lépe ekologicky interpretovat výsledky *Complete Linkage* – jsou tu vytvořeny skupiny, které podle našich předchozích znalostí popisují gradient nadmořské výšky lépe.

Pro určení podobnosti vzorků vybereme *Euclidean distance*. Ve Statistice moc dalších možností nemáme (PCORD má kupříkladu výběr mnohem větší). Lze ale použít matici vytvořenou jinde (např. předvýpočtem v jakémkoliv tabulkovém procesoru). Na hodnotách podobnosti se odrážejí velmi významně také použité standardizace a transformace dat. Podle našich zkušeností (Kovář & Lepš 1986) mají transformace na výslednou klasifikaci větší vliv než metody klastrování.

[*] Existuje ne jedna, ale nejméně 3 české terminologie – což dokazuje, že i pokud někdo českou terminologii vymyslí, ostatní ji ignorují: v nich se single linkage nazývá buď jednospojná, nebo metoda jednoduché cesty, nebo metoda nejbližšího souseda; budeme se raději držet terminologie anglické, i v té je ale trochu zmatek.

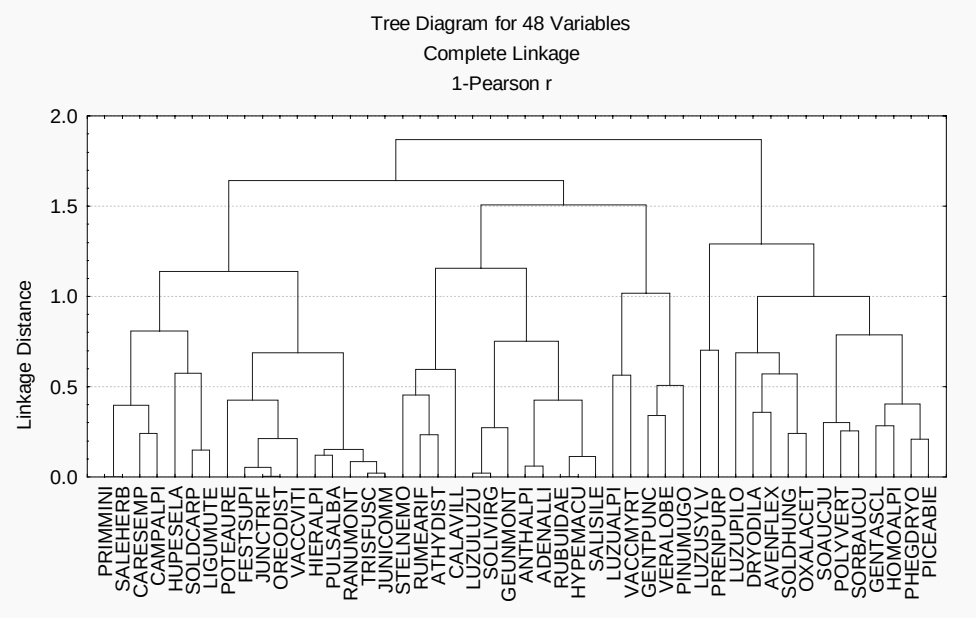
Povšimněte si také, že na rozdíl od metody TWINSPAN (viz níže) je orientace podskupin v dendrogramu náhodná, a tudíž ji nemůžeme nijak interpretovat.



Obrázek 5-8: Srovnání výsledků single linkage a complete linkage analýz. Všimněte si vyššího rozsahu zřetězování (vzorky 14 a 10 nepatří do žádného klastru, ale jsou zřetězeny s největší skupinou obsahující všechny ostatní vzorky). Výsledky complete linkage jsou snadněji interpretovatelné.

Podobně můžeme počítat aglomerativní hierarchickou klasifikaci (cluster analysis) pro proměnné (např. pro druhy). V tomto případě bude zřejmě rozumným měřítkem distribuční podobnosti druhů korelační koeficient (všimněte si, že měřítko rozumné podobnosti se liší podle toho, zda srovnáváme vzorky nebo druhy).

Výsledek klastrování druhů vypadá takto:



Obrázek 5-2: Klasifikace druhů. Procedura rozumně oddělila druhy alpinské louky (*Primula minima, Salix herbacea, Carex sempervirens, Campanula alpina*) na levou stranu diagramu a druhy smrkového lesa na stranu pravou.

Divizivní klasifikace

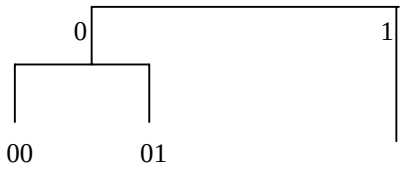
V divizivních klasifikacích je celý soubor dat dělen „shora“: nejdříve se rozdělí na dvě části, z nichž s každou se pak pracuje odděleně, atd. Pokud je klasifikace založena jen na jednom atributu (např. na jednom druhu), mluvíme o klasifikaci **monothetické**, při rozdělování podle více atributů pak o klasifikaci **polythetické**. Význam monothetických metod je hlavně historický – jednou z nich je např. klasická „association analysis“. Výhodou divizivních metod je, že ke každému dělení je připojeno kriterium, podle kterého dělení proběhlo (např. soubor druhů typických pro některou ze skupin).

Nejoblíbenější divizivní metodou (a programem s týmž jménem) je TWINSPAN (Two Way INdicator SPecies ANalysis), který byl z části inspirován klasifikačními metodami klasické fytocenologie (užívajícími indikátory pro definici vegetačních typů). V důsledku toho, že myšlenka indikátorových druhů je principiálně kvalitativní, pracuje tato metoda jen s kvalitativními daty. Aby se neztratila informace o kvantitě druhů, zavádí se pojetí pseudodruhů (*pseudospecies)* a mezních hodnot pseudodruhů (*cut levels*). Každý druh může být nahrazen několika pseudodruhy, v závislosti na zastoupení ve vzorku. Pseudodruh je přítomen, pokud zastoupení druhu přesáhne tzv. cut level. Představte si, že pro pseudodruhy zvolíme cut levels 0, 1, 5 a 20. Potom bude původní tabulka dat převedena TWINSPANem do následující formy:

	Species	Sample 1	Sample 2
Original table	*Cirsium oleraceum*	0	1
	Glechoma hederacea	6	0
	Juncus tenuis	15	25
Table with pseudo-species used in TWINSPAN	Cirsoler1	0	1
	Glechede1	1	0
	Glechede2	1	0
	Junctenu1	1	1
	Junctenu2	1	1
	Junctenu3	1	1
	Junctenu4	0	1

Tak přetransformujeme kvantitativní data na kvalitativní (přítomen / nepřítomen).

V TWINSPANu se dělení (dichotomy, division) děje podle výsledků ordinace korespondenční analýzou (*correspondence analysis*). Spočte se ordinace a vzorky se rozdělí na levou (zápornou) a pravou (kladnou) stranu dichotomie podle jejich skóre na první CA ose. Osa je rozdělena v těžišti (centroidu). V těžišti bývá ale mnoho vzorků, následkem čehož je mnoho vzorků blízko hranici a jejich zařazení záleží na mnoha náhodných činitelích. Poté se tedy udělá nová ordinace, která dá větší váhu tzv. „preferentials“, tj. druhům, které upřednostňují jednu nebo druhou stranu dichotomie. Vlastní algoritmus je poměrně komplikovaný, ale v principu jde o zisk polarizovaných ordinací, tj. takových ordinací, kde je většina vzorků mimo těžiště. Potom tedy klasifikace není založena na druzích zastoupených v obou částech, ale hlavně na druzích pro jednu či druhou stranu typických (ty mohou být následně - v dobré shodě s fytocenologickou tradicí – považovány za dobré indikátory ekologických podmínek). V prvním dělení je rozhodnutí o polaritě (tj. která strana bude pozitivní a která negativní) náhodné; v dalších děleních už je polarita určena podobností částí k sesterské skupině vyššího dělení. Například v dendrogramu na obrázku 5-10, je skupina 01 podobnější skupině 1 než 00. Díky tomu jsou vzorky srovnány v závěrečné tabulce, která je velmi podobná uspořádané fytocenologické tabulce.



Obrázek 5-10 Rozdělení vzorků TWINSPAN klasifikací

V každém kroku je programem vytištěno rovněž kriterium použité pro to které dělení. Tento fakt přispívá velmi výrazně k interpretovatelnosti výsledků. Klasifikace vzorků je doplněna klasifikací druhů a závěrečná tabulka se vytvoří na základě těchto dvou klasifikací.

Analýza vzorků z Tater

V následující části si ukážeme použití TWINSPANu pro analýzu 14 snímků z transektu na nadmořské výšce v Nízkých Tatrách.. TWINSPAN je užitečný hlavně při analýze velkých datových souborů, my užíváme tento malý soubor jen pro jednoduchost. Použijeme soubor *tatry.spe*, tj. soubor importovaný pomocí WCanoImp do Cornell (CANOCO) formátu. V tomto případě jsme požadovali v programu TWINSPAN dlouhý výstup, abychom si ukázali, co všechno z programu můžeme dostat (normálně chceme jen krátký výstup a i ten je vesměs pěkně dlouhý).

Nejdřív se zobrazí hlavičky a program vypíše použitá nastavení. Důležité jsou:

```
Cut levels:
      .00    1.00    2.00    3.00    4.00    5.00
```

Data jsme transformovali ordinální transformací z Br.-Bl stupnice, (která má vzhledem k pokryvnosti druhů přibližně logaritmický charakter), a tak není důvod dále snižovat důležitost druhů s vysokou pokryvností. Pokud jsou data ve formě odhadů pokryvnosti, pak je rozumné použít implicitních mezních hodnot (cut levels), tj. 0 2 5 … Hodnoty 0 1 10 100 1000 dají výsledky odpovídající logaritmické transformaci a jsou užitečné, pokud data vyjadřují počty individuí lišící se řádově.

Z dalších voleb si povšimněte následujících (všechny jsou default):

```
1. Minimum group size for division:                    5
2. Maximum number of indicators per division:          7
3. Maximum number of species in final tabulation:    100
4. Maximum level of divisions:                         6
```

1. Znamená, že skupiny obsahující méně než 5 snímků jsou konečné, tj. už nejsou dále rozdělovány. Pro malé soubory dat je rozumné snížit tuto hodnotu na čtyři. **2.** Počet indikátorů na dělení: obvykle je default vyhovující. **3.** Pokud máte více než 100 druhů, tak se v závěrečné tabulce objeví jen 100 těch nejčastějších. Hodnotu můžete zvýšit, zdá-li se vám malá. **4.** Jiný způsob, jak kontrolovat, kdy rozdělování skončí (kontrola počtem druhů - ve volbě **1.** - je lepším řešením). Pro data o rozumné velikosti je hodnota 6 obvykle dostačující.

Potom je první dělení popsáno takto:

```
DIVISION    1  (N=  14)         I.E. GROUP *
 Eigenvalue  .565  at iteration   1
 INDICATORS, together with their SIGN
 Oreo Dist1(+)
```

Indikátorem prvního dělení je *Oreochloe disticha* (jednička na konci znamená, že byla použita mezní hodnota prvního pseudodruhu, tj. že k rozhodnutí o přítomnosti indikátoru stačila pouhá přítomnost tohoto druhu, nikoli jeho množství).

```
Maximum indicator skóre for negative group  0     Minimum indicator skóre for positive group
1

 Items in NEGATIVE group   2  (N=   10)         i.e. group *0
Samp0001  Samp0002  Samp0003  Samp0004  Samp0005  Samp0006  Samp0007  Samp0008  Samp0009
Samp0010

 BORDERLINE negatives    (N=    1)
 Samp0009

 Items in POSITIVE group   3  (N=    4)         i.e. group *1
 Samp0011  Samp0012  Samp0013  Samp0014
```

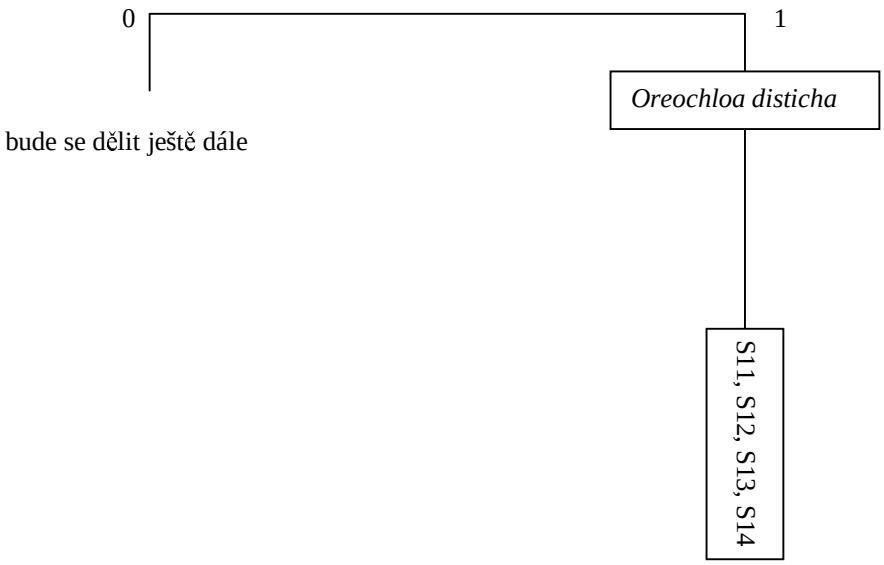
Rozdělení vzorků. Všimněte si u vzorku 9 varování, že se nachází na hranici mezi dvěma skupinami (toto varování se objevuje jen v dlouhém výpisu).

Nyní jsou vypsány druhy preferující jednu ze stran dichotomie (preferentials) spolu s jejich zastoupením v každé skupině. (Např. *Picea abies* v 7 vzorcích negativní skupiny a v jednom vzorku skupiny pozitivní. Za pozornost stojí, že preferentials jsou určeny v závislosti na počtu vzorků v každé skupině a zvlášť také pro všechny mezní hodnoty pseudodruhů.) Preferentials se uvádějí jen ve dlouhém výpisu (zaberou velmi mnoho místa a i zde je značná část informace vypuštěna).

```
 NEGATIVE PREFERENTIALS
Pice Abie1( 7,  1) Sorb Aucu1( 7,  0) Oxal Acet1( 7,  0) Sold Hung1( 5,  0) Aven Flex1(10,  1) Gent Ascl1( 8,  0)
Dryo Dila1( 8,  0) Pheg Dryo1( 6,  0) Pren Purp1( 2,  0) Poly Vert1( 3,  0) SoAu cJu 1( 3,  0) Luzu Pilo1( 2,  0)
etc

 POSITIVE PREFERENTIALS
Juni Comm1( 0,  2) Ligu Mute1( 4,  4) Sold Carp1( 2,  4) Ranu Mont1( 1,  2) Hupe Sela1( 3,  3) Geun Mont1( 2,  2)
Vacc Viti1( 2,  4) Puls Alba1( 0,  2) Gent Punc1( 2,  3) Soli Virg1( 1,  1) Luzu Luzu1( 1,  1) Oreo Dist1( 0,  4)
etc.
```

```
 NON-PREFERENTIALS
Pinu Mugo1( 5,  2) Vacc Myrt1(10,  4) Homo Alpi1(10,  4) Cala Vill1( 8,  3) Rume Arif1( 4,  1) Vera Lobe1( 5,  3)
Pinu Mugo2( 5,  2) Vacc Myrt2(10,  4) Homo Alpi2(10,  4) Cala Vill2( 8,  3) Rume Arif2( 4,  1) Vera Lobe2( 5,  3)
etc.

    End of level   1
```

Nyní můžeme začít kreslit dendrogram. Část, která je už jasná, vypadá takto:

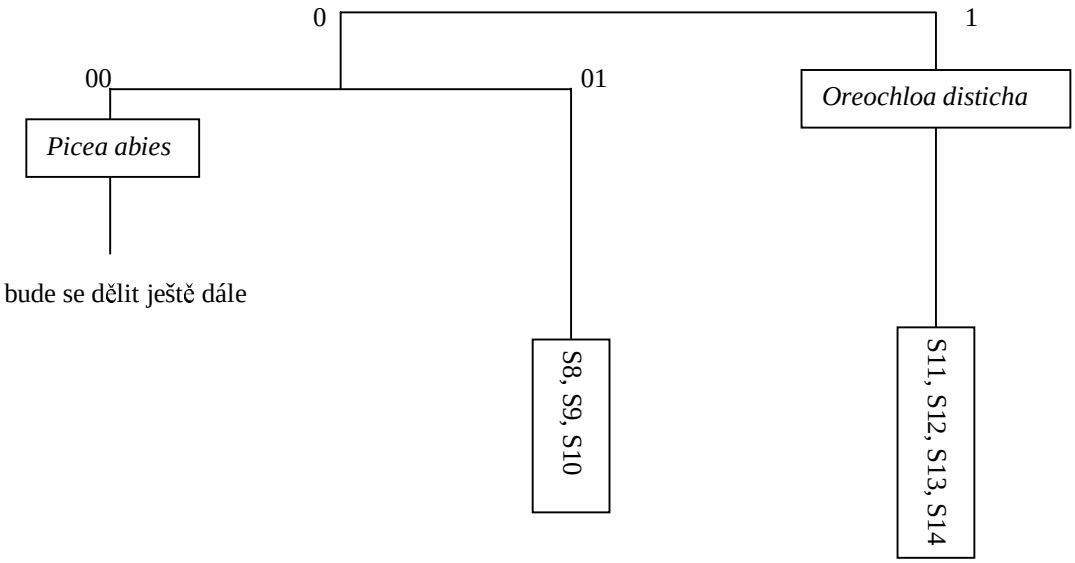


Obrázek 5-11 První dělení v příkladu pro program TWINSPAN

Pozitivní skupina už se měnit nebude, protože je menší než pět, což je minimální velikost pro dělení. Další hladina (bez preferentials):

```
DIVISION    2  (N=  10)         I.E. GROUP *0
 Eigenvalue  .344  at iteration   1
 INDICATORS, together with their SIGN
 Pice Abie1(-)
Maximum indicator skóre for negative group  -1     Minimum indicator skóre for positive group
0

 Items in NEGATIVE group   4  (N=    7)         i.e. group *00
 Samp0001  Samp0002  Samp0003  Samp0004  Samp0005  Samp0006  Samp0007

 Items in POSITIVE group   5  (N=    3)         i.e. group *01
 Samp0008  Samp0009  Samp0010

DIVISION    3  (N=   4)         I.E. GROUP *1
 DIVISION FAILS - There are too few items

    End of level   2
```

Podobně můžeme pokračovat dál v konstrukci dendrogramu (jediný indikátor na rozdělení je spíše výjimkou než pravidlem).



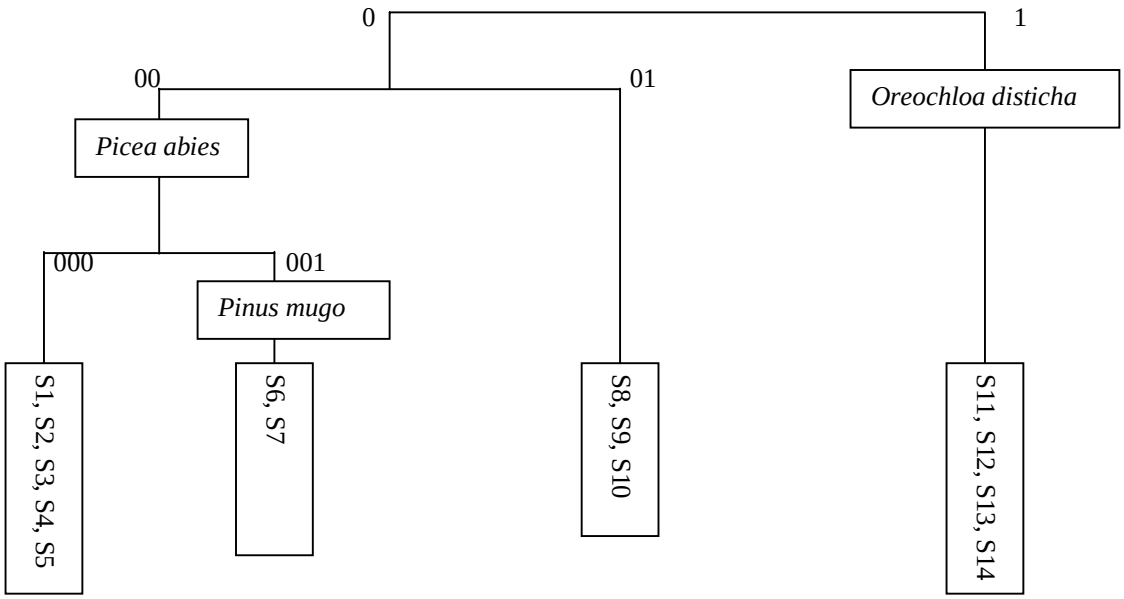
Obrázek 5-12 Rozhodnutí na druhé hladině příkladu v TWINSPANu

Další hladina:

```
DIVISION    4  (N=   7)         I.E. GROUP *00
 Eigenvalue  .279  at iteration   1
 INDICATORS, together with their SIGN
 Pinu Mugo1(+)
Maximum indicator skóre for negative group   0     Minimum indicator  skóre for positive group
1

 Items in NEGATIVE group   8  (N=    5)         i.e. group *000
 Samp0001  Samp0002  Samp0003  Samp0004  Samp0005

 Items in POSITIVE group   9  (N=    2)         i.e. group *001
 Samp0006  Samp0007
```

Povšimněte si zde významu indikátorového druhu *Pinus mugo*. Je indikátorem při dělení skupiny 00 (která má 7 vzorků), kde je přítomen jen ve dvou z nich, 6 a 7 (čímž tvoří pozitivní skupinu 001). Je ale také běžný ve vzorcích 8, 9, 10, 11 a 12. Z toho plyne, že indikátory jsou určeny a měly by být interpretovány jen pro určité dělení, nikoliv pro celou skupinu. Například skupina 001 je charakterizovaná přítomností *Pinus mugo* oproti skupině 000, a ne v rámci celého datového souboru. Zde je patrná výhoda TWINSPANu, kde orientace skupiny (tj. která část dělení bude pozitivní a která negativní) záleží na podobnosti ke skupině 01 (a tato skupina má, mimo jiné, *Pinus mugo* ve všech svých vzorcích).

Další dělení si už ukazovat nebudeme (skupina 000 má pět vzorků a mohla by se dělit ještě dál) a dokončíme dendrogram na této úrovni:



Obrázek 5-13 Závěrečný stav klasifikačního příkladu v TWINSPANu

Podobně připravíme klasifikaci (dendrogram) druhů. Indikátory jsou opět stejné jako v předchozích děleních, jen jsou omezeny co do počtu. Kdybyste chtěli mít dělení charakterizované více druhy, museli byste použít preferentials.

Závěrem se vytiskne tabulka připomínající klasickou fytocenologickou tabulku doplněná klasifikací vzorků i druhů.

```
               SSSSSSSSSSSSSS
               aaaaaaaaaaaaaa
               mmmmmmmmmmmmmm
               pppppppppppppp
               00000000000000
               00000000000000
               00000000011111
               21345678901234

 4 Sali SIle  ---------5----  0000
29 Anth Alpi  -----2---3----  0000
30 Hype Macu  ------2--3----  0000
31 Rubu Idae  ------2--3----  0000
28 Aden Alli  -----2---2----  0001
 1 Pice Abie  6666665---5---  001000
 7 Oxal Acet  55344-4--3----  001001
 9 Sold Hung  43444---------  001001
18 Luzu Pilo  2-2----------  001001
20 Luzu Sylv  --3243--------  001001
12 Gent Ascl  23333333------  001010
14 Pheg Dryo  4333-33-------  001010
15 Pren Purp  2----3--------  001011
16 Poly Vert  3----33-------  001011
22 Stel Nemo  ---2--3-------  001011
 2 Sorb Aucu  42-23444------  00110
10 Aven Flex  34543443343---  00110
13 Dryo Dila  333333-3-3----  00110
17 SoAu cJu   3-----32------  00110
19 Athy Dist  3-23335--53---  00111
 6 Vacc Myrt  54646666636653  01
 8 Homo Alpi  44454454334334  01
11 Cala Vill  33365-54-6445-  01
21 Rume Arif  --23--4--33---  01
 3 Pinu Mugo  -----3666665--  10
23 Vera Lobe  ----2333-4322-  10
27 Hupe Sela  -----223--22-3  10
36 Soli Virg  ---------2--2-  10
33 Vacc Viti  -------33-3343  1100
35 Gent Punc  -------3-4333-  1100
37 Luzu Luzu  ---------3--4-  1100
24 Ligu Mute  ----233--23333  1101
25 Sold Carp  -----54---3333  1101
26 Ranu Mont  -----2-----33-  1101
32 Geun Mont  ------2--3-33-  1101
 5 Juni Comm  -----------24-  111
34 Puls Alba  -----------32-  111
38 Oreo Dist  ----------5564  111
39 Fest Supi  ----------3444  111
40 Camp Alpi  ----------34-4  111
41 Junc Trif  ----------4453  111
42 Luzu Alpi  ----------33--  111
43 Hier Alpi  ----------233-  111
44 Care Semp  -----------545  111
45 Tris Fusc  -----------33-  111
46 Pote Aure  ------------32  111
47 Sale Herb  -------------5  111
48 Prim Mini  -------------4  111

               00000000001111
               0000000111
               0000011
```

Z posledních tří řádků lze vyčíst tři hladiny dělení a spojování vzorků do odpovídajících skupin, např. že vzorky 11 až 14 jsou ve skupině 1 (která už se dále nedělí).

6. Vizualizace mnohorozměrných dat programy CanoDraw 3.1 a CanoPost 1.0 for Windows

Základním prostředkem pro znázornění výsledků ordinačních modelů je ordinační diagram. Obsah ordinačních diagramů můžeme použít pro aproximaci matice druhových dat, matic vzdáleností mezi jednotlivými vzorky a / nebo matic korelací nebo nepodobností mezi jednotlivými druhy. V případě ordinačních analýz s charakteristikami prostředí můžeme použít ordinační diagram k aproximaci obsahu tabulek environmentálních dat, korelací mezi druhy a charakteristikami prostředí, korelací mezi jednotlivými charakteristikami prostředí, atd. Vše, co lze z ordinačních diagramů vyčíst, je shrnuto v následujících dvou sekcích, zvlášť pro lineární a zvlášť pro unimodální ordinační metody.

6.1. Co lze vyčíst z ordinačních diagramů: Lineární metody

Ordinační diagram založený na lineárních ordinačních metodách (PCA nebo RDA) může zobrazovat skóre vzorků (body), druhů (šipky), kvantitativních charakteristik prostředí (šipky) a nominálních indikátorových proměnných (body - centroids - odpovídající jednotlivým hladinám faktoru). Tabulka 6-1 (podle Ter Braak, 1994) shrnuje, co vše se dá z ordinačních diagramů na základě těchto skóre zjistit.

Porovnávané jednotky	**Škálování 1** **Zaměřeno na vzdálenosti vzorků**	**Škálování 2** **Zaměřeno na korelace druhů**
druhy X *vzorky*	(fitované) hodnoty abundance v druhových datech	
vzorky X *vzorky*	Eukleidovské vzdálenosti mezi vzorky	**xxx**
druhy X *druhy*	**xxx**	lineární korelace mezi druhy
druhy X *charakt. prostředí*	lineární korelace mezi druhy a charakteristikami prostředí	
vzorky X *charakt. prostředí*	**xxx**	hodnoty environmentálních proměnných
char. prostředí X *char. prostředí*	marginální vlivy charakteristik prostředí na ordinační skóre	korelace mezi charakteristikami prostředí
druhy X *nominální char. prostředí.*	průměrné hodnoty abundance druhů v rámci tříd	
vzorky X *nominální char.prostředí.*	příslušnost vzorků ke třídám	
nom. char. pros. X *nom. char.pros.*	Eukleidovské vzdálenosti mezi třídami vzorků	**xxx**
char. prostředí X *nom. char.. prost.*	**xxx**	průměry charakteristik prostředí v rámci tříd

Tabulka 6-1

Promítneme-li body vzorků kolmo na šipku vybraného druhu, aproximujeme pořadí hodnot tohoto druhu v promítaných vzorcích. Pokud použijeme skóre vzorků, která jsou lineární kombinací charakteristik prostředí (skóre **SamE**, typicky v metodě RDA), aproximujeme **fitované**, nikoliv pozorované hodnoty těchto abundancí. To platí pro oba typy škálování. Pokud jsme provedli centrování druhy, pak pro daný druh předpovídáme ve vzorku promítaném do středu osy souřadnic (kolmo na šipku druhu) jeho průměrné zastoupení. Vzorky promítané od nuly dále ve směru šipky budou mít abundance tohoto druhu

nadprůměrné, zatímco vzorky promítané v opačném směru budou mít hodnoty podprůměrné.

Vzdálenost mezi body vzorků vypovídá o nepodobnosti vzorků jen při škálování, které je na vzdálenost vzorků zaměřeno. Nepodobnost se pak vyjadřuje Eukleidovskou vzdáleností.

Podobně lze odhadovat (lineární) korelační koeficienty mezi druhy z úhlů mezi druhovými šipkami jen při škálování zaměřeném na korelace druhů. Aproximovaná korelace mezi dvěma proměnnými se rovná cosinu úhlu mezi odpovídajícími šipkami: šipky mířící stejným směrem odpovídají druhům, pro které je předpovězena velká pozitivní korelace, naopak šipky ukazující do opačných směrů patří druhům s velkou negativní korelací. Pro dvojici druhů, jejichž šipky se setkávají v pravém úhlu, je očekávána nízká (lineární) korelace.

Podobnou aproximační metodu můžeme použít i při vzájemném srovnávání šipek druhů a charakteristik prostředí. Například pokud šipka kvantitativní charakteristiky prostředí ukazuje stejným směrem jako šipka druhu, předpokládá se, že hodnoty tohoto druhu jsou v pozitivní korelaci s hodnotami této charakteristiky prostředí. Tento výklad platí pro oba druhy škálování ordinačních skóre.

Body vzorků můžeme promítnout kolmo na šipky charakteristik prostředí. To nám poskytne přibližné seřazení vzorků podle rostoucí hodnoty této charakteristiky prostředí (pohybujeme-li se směrem ke špičce šipky a za ni). Charakteristiky prostředí jsou (stejně jako kovariáty) vždy vycentrovány (a standardizovány) ještě před tím, než je ordinační model vytvořen, takže podobně jako při promítání bodů vzorků na šipky druhů odpovídá promítnutí blízko nule (počátku souřadnicového systému) průměrným hodnotám jednotlivých charakteristik prostředí v promítaném vzorku.

Úhel mezi šipkami charakteristik prostředí můžeme použít k odhadu korelací mezi těmito proměnnými pouze při škálování zaměřeném na korelace druhů. Podotýkáme ale, že tento odhad není tak dobrý jako při analýze tabulky dat o prostředí v PCA. Pokud je škálování zaměřeno na mezidruhové vzdálenosti, lze každou jednotlivou šipku interpretovat tak, že ukazuje směrem, kterým by se skóre vzorku posunulo při vyšší hodnotě této charakteristiky prostředí. Z délky šipek lze vyvodit srovnání velikosti tohoto vlivu charakteristik prostředí (znovu zde podotýkáme, že všechny tyto charakteristiky do analýz vstupují s nulovým průměrem a jednotkovou variancí).

Výstup programu Canoco umožňuje jinou interpretaci skóre nominálních charakteristik prostředí. Jsou to obvykle indikátorové proměnné s hodnotami 0 nebo 1, které jsou vytvořeny z původních faktoriálních proměnných. Tyto nominální charakteristiky prostředí jsou v diagramech zastoupeny body, které jsou centroidy pro skóre těch vzorků, jež mají v dané indikátorové proměnné hodnotu 1. Na původní faktoriální proměnnou se můžeme dívat jako na proměnnou klasifikační, a jednotlivé nominální (indikátorové) proměnné pak vymezují jednotlivé třídy vzorků. Takže lze říci, že centroidové skóre této nominální proměnné je průměrem skóre vzorků příslušné třídy.

Promítneme-li centroidy nominálních charakteristik prostředí na šipky druhů, můžeme odhadnout průměrnou hodnotu tohoto druhu v jednotlivých třídách vzorků. Podobně vzdálenost mezi centroidy nominálních proměnných bude (při škálování zaměřeném na mezivzorkové vzdálenosti) ukazatelem nepodobnosti v jejich druhovém složení vyjádřené pomocí Eukleidovské vzdálenosti.

V obou typech škálování lze z umístění centroidů jednotlivých tříd vzorků a jednotlivých bodů vzorků určit zařazení toho kterého vzorku. Vzorek patří s největší pravděpodobností do té třídy, jejíž centroid je mu nejblíže.

Promítneme-li centroidy nominálních charakteristik prostředí na šipky kvantitativních charakteristik prostředí, můžeme vydedukovat průměrné hodnoty těchto proměnných pro jednotlivé třídy vzorků.

6.2. Co lze vyčíst z ordinačních diagramů: Unimodální metody

Interpretace ordinačních diagramů pocházejících z unimodálních ordinačních modelů je shrnuta v Tabulce 6-2. Je zde mnoho podobností s interpretací lineárních ordinačních modelů, kterou jsme podrobně popisovali v minulé sekci, takže se na ni v případě potřeby odkážeme.

Hlavní rozdíl v interpretacích lineárních a unimodálních ordinačních diagramů je v odlišnosti modelů odpovědi druhu na vytvořený gradient (ordinační osu). Zatímco v minulé části jsme uvažovali pouze lineární (monotónní změnu), zde se předpokládá, že druh bude mít optimum na každé z ordinačních os a že jeho abundance (pro data typu přítomen / nepřítomen pak pravděpodobnost výskytu) bude ve všech směrech od tohoto bodu symetricky klesat. Odhadnutá pozice optima druhu se zobrazí jako skóre druhu, tedy jako bod. Optimum se počítá jako vážený průměr z pozic vzorků, kde váhami jsou relativní abundance druhu v jednotlivých vzorcích.

Další důležitý rozdíl je ve výpočtu nepodobnosti. Ten je založen na chi-square metrice, což znamená, že jakékoliv dva vzorky se stejným **relativním** zastoupením (např. 3 druhy ve 2 vzorcích s hodnotami 1 2 1 a 10 20 10) jsou unimodálním modelem vyhodnoceny jako stejné. Nepodobnost v distribuci různých druhů se hodnotí stejnou metrikou, jen ji aplikujeme na transponovanou matici primárních dat.

Porovnávané jednotky	Škálování 1 Zaměřeno na vzdálenosti druhů a Hillovo škálování	Škálování 2 Zaměřeno na vzdálenosti druhů a projekční škálování
druhy X *vzorky*	(fitované) relativní abundance v tabulce druhových dat	
vzorky X *vzorky*	vzdálenosti v jednotkách druhové změny* mezi vzorky	χ^2 vzdálenosti mezi vzorky (pokud jsou λ srovnatelná)
druhy X *druhy*	**xxx**	χ^2 vzdálenosti mezi druhy (z fitovaných abundancí)
druhy X *char. prostředí*	vážené průměry - optima druhů ve vztahu k jednotlivým charakteristikám prostředí	
vzorky X *char. prostředí*	**xxx**	hodnoty charakteristik prostředí ve vzorcích
char. prostředí X *char. prostředí*	marginální vlivy charakteristik prostředí	korelace mezi charakteristikami prostředí
druhy X *nominální char. prostředí*	relativní celkové abundance ve třídách vzorků	
vzorky X *nominální char. prostředí*	zařazení vzorků do tříd	
nom.char.pr.. X *nom.char.pr.*.	vzdálenosti v jednotkách druhové změny* mezi třídami vzorků	χ^2 vzdálenosti (pokud jsou λ srovnatelná) mezi třídami vzorků
char. prostředí X *nom.char. prostř.*	**xxx**	průměry charakteristik prostředí v rámci tříd vzorků

Tabulka 6-2

Vzájemné umístění bodů vzorků a druhů vypovídá o relativních abundancích v matici druhových dat. Skóre druhů bývají poblíž bodů těch vzorků, ve kterých je jejich relativní abundance nejvyšší a podobně body vzorků jsou blízko pozicí druhů, které se v nich vyskytují. Tomuto způsobu interpretace se říká **centroidový princip**. Chceme-li však výsledky kvantifikovat, budeme pracovat přímo se vzdálenostmi mezi body. Pokud seřadíme vzorky podle rostoucí vzdálenosti od bodů určitého druhu, bude toto seřazení odpovídat klesající relativní abundanci druhu ve vzorcích.

Pro kratší gradientové vzdálenosti můžeme pozice druhů a vzorků interpretovat pomocí **projekčního pravidla**, podobně jako v lineárních ordinačních diagramech. Spojíme jednoduše bod druhu s počátkem souřadnicového systému a body vzorků promítneme kolmo na tuto čáru.

Vzdálenost mezi body vzorků aproximuje chi-square vzdálenost mezi vzorky v projekčním škálování se zaměřením na druhy, ale jen v případě, kdy použité ordinační osy vysvětlují podobné množství variability (mají-li srovnatelné hodnoty charakteristických čísel).

Použijeme-li Hillovo škálování se zaměřením na vzdálenosti vzorků, je vzdálenost mezi vzorky v jednotkách "výměny druhů" (*species turnover units*, říká se jim také **SD jednotky** od *standard deviation* křivek odpovědí druhů), které odpovídají označení ordinačních os. Vzorky vzdálené od sebe více než 4 jednotky budou stěží sdílet jeden jediný druh, protože během jedné jednotky nastane "poloviční výměna" druhového složení.

* Rychlost změny zde odpovídá anglickému termínu *turnover rate*, v kombinaci se vzdáleností pak termínu *turnover distance*.

Vzdálenosti mezi body druhů v projekčním škálování (se zaměřením na vzdálenosti druhů) aproximuje chi-square vzdálenost mezi distribucemi druhů.

Pokud promítneme body druhů na šipku kvantitativní charakteristiky prostředí, dostaneme pořadí optim těchto druhů ve vztahu k této charakteristice. Podobně (ale pouze v projekčním škálování) promítnutí bodů vzorků na šipku charakteristiky prostředí aproximuje hodnoty z tabulky environmentálních dat.

Interpretace vztahů mezi šipkami charakteristik prostředí (buď pomocí úhlů nebo srovnáním relativních směrů a velikosti vlivu) se podobá lineárním metodám popsaným v minulé kapitole.

Vzdálenost mezi body druhů a centroidy nominálních charakteristik prostředí (na význam tohoto výrazu se podívejte do minulé sekce, je-li třeba) můžeme použít k aproximaci relativní **celkové** abundance druhu ve vzorcích příslušné třídy (sumujeme přes všechny vzorky třídy).

Porovnání mezi body vzorků a centroidy nominálních proměnných a mezi centroidy a šipkami kvantitativních charakteristik prostředí probíhá stejně jako v lineárních modelech.

Vzdálenost centroidů dvou nominálních proměnných lze interpretovat podobně jako při určování vzdálenosti (nepodobnosti) bodů vzorků. Tato vzdálenost se však vztahuje k *chi-square* (nebo k *turnover* - to závisí na škálování) vzdálenostem mezi vzorky odpovídajících dvou tříd.

6.3. Regresní modely v programu CanoDraw

Program CanoDraw používáme k vytvoření ordinačních diagramů se skóre pro vybranou kombinaci položek (např. vzorků nebo druhů). Mimo to lze ale tento program použít i k prozkoumání mnohorozměrných dat v kontextu ordinačního prostoru, abychom posoudili trendy naznačené ordinačním diagramem a zkontrolovali předpoklady vytvořené ordinačním modelem v programu Canoco. O využití těchto funkcí budou následující sekce. Technická hlediska však podrobněji popisuje CanoDraw 3.0 User's Guide (*Šmilauer 1992*).

Většina metod určených pro vizuální hodnocení je v nabídce **Attributes**. Tady si můžeme vybrat kupříkladu zobrazení hodnot určité charakteristiky prostředí v ordinačním prostoru. Hodnoty této proměnné jsou zobrazeny na pozicích jednotlivých vzorků pomocí proporcionálně velkých symbolů. Očekáváme-li v ordinačním prostoru monotónní (nebo v omezeném ordinačním modelu dokonce lineární) změnu hodnot této proměnné, měl by takový pattern být vidět i v zobrazení zvaném **symbolový diagram** (*symbol plot*). Často ale (a to dokonce i pro středně velké soubory) vynesení hodnot jednotlivých vzorků k efektivní abstrakci patternu příliš nepřispěje. Originální body však můžeme nahradit jednodušším statistickým modelem. V CanoDraw jsou tři velké skupiny regresních modelů, které můžeme používat záměnně v diagramech atributů (*attribute plots*), navíc k symbolovým diagramům.

Zobecněné lineární modely (GLM, *generalized linear models*) jsou rozšířením tradičních lineárních modelů s větší tolerancí pro různé distribuční vlastnosti vysvětlovaných proměnných. Více se dozvíte v části 11-3. CanoDraw umožňuje fitovat GLM s užitím jedné ze čtyř distribučních rodin (předpokládaná Poissonova, binomická, Gamma nebo Gaussovou statistická distribuce vysvětlované proměnné). U link funkcí však možnost výběru není - pro vybranou distribuční rodinu se používají tzv. kanonické link funkce (*canonical link functions*)**.** Systematickou složku modelu můžeme specifikovat buď přesně fitováním tzv. **fixním modelem** (takže požadujeme například fitování polynomu druhého stupně

prediktoru) nebo necháme CanoDraw zvolit kompexitu modelu pomocí postupného výběru na základě testů analýzy deviance. Specifickým rysem implementace GLM v programu CanoDraw je, že při fitování modelu kvadratického polynomu s jedním prediktorem předpokládanou Poissonovou distribucí CanoDraw rozpozná v tomto modelu popis unimodální odezvové křivky (*unimodal response curve*) druhu ve vztahu ke gradientu prostředí a odhaduje hodnoty jejího optima a také šířky (pomocí parametru **tolerance**), je-li to možné (viz Ter Braak & Looman, 1986).

Generalized loess smoother je rozšířením dnes již klasické metody *loess* (*Cleveland & Devlin, 1988*), což je akronym pro *locally weighted regression*. Rozdíl mezi standardním a zde implementovaným loess modelem je v tom, že CanoDraw do něj zapojuje zobecněné lineární modely. Fitované odezvové křivky a povrchy jsou pak např. pro binární data mnohem smysluplnější.

Generalized kriging je regresní metoda odvozená z běžného modelu nejmenších čtverců. Zohledňuje však i (modelované) prostorové autokorelace mezi vzorky sebranými v prostoru. Ve srovnání se standardní metodou odhaduje *generalized kriging* (zvané též *universal kriging*) i lineární nebo polynomiální trend v hodnotách vysvětlované proměnné. Prediktory zde používané jsou prostorovými souřadnicemi vzorků. Tato metoda je vhodná nejen pro fitování abundancí druhů nebo hodnot charakteristik prostředí na prostorové souřadnice vzorků (což se dá udělat také s CanoDraw 3.x), ale i pro jejich fitování do ordinačního prostoru, protože tam jsou skóre vzorků také mezi sebou korelována.

Ať zvolíme kterýkoliv z těchto modelů, budou výsledky shrnuty buď fitovanou čarou či křivkou (máme-li jediný prediktor, jako např. při vynášení hodnot charakteristiky prostředí proti skóre na jedné z ordinačních os) nebo vrstevnicovým diagramem (*contour plot* nebo *isolines plot*), máme-li prediktory dva (nejčastěji souřadnice vzorku na dvou ordinačních osách). CanoDraw umožňuje upravení hladin fitované vysvětlované proměnné, které jsou v tomto diagramu vynášeny.

6.4. Ordinační diagnostika

Diagramy (ordinační a atributové) shrnuté v předchozí části lze použít ke kontrole, jak dalece splňují data analyzovaná ordinační metodou její předpoklady.

První možná kontrola se týká našeho předpokladu o tvaru druhové odpovědi podél ordinačních os, které reprezentují "objevené" gradienty ve změnách složení společenstva (v případě omezených ordinačních metod jsou interpretovatelné pomocí charakteristik prostředí). V zobecněných lineárních modelech (a to jak v případě jejich užití u lineárních, tak u unimodálních ordinačních metod) bychom pravděpodobně měli nechat CanoDraw zvolit specifikaci regresního modelu (tj. model nulové hypotézy vs. lineární model vs. model polynomiální regrese). Uvědomte si ale, že omezená rodina polynomiálních specifikací vysvětlované křivky vylučuje to, aby byl nalezený tvar křivky asymetrický nebo bimodální (tímto způsobem bychom ale mohli interpretovat výsledky, kdy je vybrán polynom druhého stupně s minimem ve středu gradientu). *Loess smoother* poskytuje obecnější rodinu regresních modelů, které nemají žádný určitý předpoklad o tvaru fitované křivky.

Dalším předpokladem omezených ordinačních metod (jak RDA, tak CCA) je, že kanonické (omezené) ordinační osy jsou **lineární** kombinací zadaných charakteristik prostředí. Pro kvantitativní charakteristiku prostředí můžeme její hodnoty vynést proti skóre vzorků na jedné ordinační ose, abychom viděli, je-li změna podél ordinační osy lineární. Ačkoli v mnoha případech tomu tak je (přinejmenším v omezených ordinačních metodách), může se stát, že najdeme průběh sice monotónní, ale zdaleka ne lineární. Často to odpovídá

odpovědi ve složení společenstva podél gradientů jako jsou sukcesní čas nebo nadmořská výška a může to naznačovat potřebu vhodné transformace hodnot vysvětlující proměnné (např. logaritmické transformace).

Manuál programu Canoco se o **ordinačních diagnostikách** zmiňuje trošku jinak, a to jako o statistikách počítaných ordinačními metodami. Tyto statistiky v obecném smyslu popisují, jak dobře charakterizuje fitovaný ordinační model vlastnosti jednotlivých vzorků a druhů. Tato statistika používaná pro druhy se jmenuje **kumulativní fit - CFit** (viz *Ter Braak et Šmilauer 1998*, p. 174) a v programu CanoDraw ji lze použít k výběru těch druhů, které se zobrazují v ordinačním diagramu.

6.5. Interpretace projekčního diagramu T statistik

Projekční diagram T statistik (*T-value biplot)* je diagram s úsečkami pro druhy a charakteristiky prostředí. Jeho primárním účelem je najít statisticky významné párové vztahy mezi druhy a charakteristikami prostředí (tj. který druh závisí na které charakteristice prostředí). Tento diagram interpretujeme pomocí projekčního pravidla a aproximujeme tak tabulku T-hodnot regresních koeficientů pro mnohonásobnou (váženou) regresi závislosti jednotlivých druhů na všech charakteristikách prostředí.

Při interpretaci promítáme špičky šipek charakteristik prostředí na čáry překrývající šipky jednotlivých druhů. Pozice získaná tímto promítnutím aproximuje T statistiku regresních koeficientů jednotlivých charakteristik prostředí, pokud jsou tyto použity jako prediktory pro hodnoty druhů. Promítne-li se špička šipky charakteristiky prostředí na linii dále od počátku než je špička šipky druhu, je odpovídající regresní koeficient větší než 2.0. Při podobném promítnutí do opačného směru je T statistika menší než -2.0. Body promítnuté mezi tyto hranice (včetně počátku souřadnicového systému) odpovídají T hodnotám mezi -2 a +2.

Chceme-li zjistit, které druhy reagují na určitou charakteristiku prostředí signifikantně **pozitivně**, nakreslíme kruh se středem v polovině vzdálenosti špičky šipky této proměnné a středu souřadnic, jehož průměr je stejný jako délka šipky. Čáry, které v tomto kruhu končí, patří druhům s pozitivním regresním koeficientem vůči dané charakteristice prostředí. Odpovídající T hodnota je vyšší než 2.0. Podobný kruh můžeme nakreslit i v druhém směru, a ten pak odpovídá signifikantnímu negativnímu regresnímu koeficientu.

Další informace o těchto diagramech jsou v *Ter Braak & Šmilauer* (1998), str. 177 - 180.[+]

[+] Projekční diagramy T statistik vytvořené programem CanoDraw 3.x mají dva problémy: 1) u čar reprezentujících charakteristik prostředí je správná jen část vyznačená plnou čarou. Přerušovaně vyznačené segmenty byste měli v programu CanoPost odstranit. 2) Default rozsah os také není správný. Před vytvořením diagramu (na "virtuálním zařízení" reprezentujícím PostScriptový výstup), by se mělo zvolit z menu **Zoom-out**. Tím vznikne rozsah os pro diagram mnohem vhodnější.

7. Případová studie 1: Oddělení vlivu vysvětlujících proměnných

7.1. Úvod

V mnoha případech potřebujeme oddělit vliv několika vysvětlujících proměnných, dokonce i v případech, kdy jsou tyto proměnné spolu korelovány. Tento příklad pochází ze zemědělského pokusu s hnojením (*Pyšek a Lepš 1991*). Pole ječmene bylo hnojeno třemi dusíkatými hnojivy (síran amonný, dusičnan amonný a tekutá močovina) ve dvou různých celkových dávkách dusíku. Kvůli praktickým omezením nebyl pro založení pokusu použit korektní design; tzn. že plochy jsou pseudoreplikacemi. Původní pokus byl navržen k hodnocení splachu živin, takže malé plochy nebyly použitelné. My jsme využili pokus s velkými plochami.* Druhové složení plevelů ve 122 plochách jsme charakterizovali pomocí klasické Braun-Blanketovy stupnice (upravené na ordinální škálu - t.j. čísla 1 až 7 pro jednotlivé stupně Br.-Bl. stupnice). Odhadována byla také pokryvnost ječmene.

Očekávali jsme, že plevele budou ovlivněny jak přímo hnojivy, tak nepřímo kompeticí ječmene. Vliv hnojiv můžeme hodnotit díky experimentálním zásahům. Pokryvnost ječmene je vysoce korelována s dávkou hnojiva a přímý vliv jeho pokryvnosti nelze zjistit, protože tu jsme v pokusu cíleně neměnili. Data nám však umožní přímý vliv hnojení od nepřímého vlivu pokryvnosti ječmene alespoň částečně oddělit. Dělá se to podobným způsobem, jako se odděluje vliv korelovaných prediktorů (nezávisle proměnných) na jednorozměrnou proměnnou v mnohonásobné regresi. Oddělení se provede tak, že proměnnou, kterou sledujeme, označíme jako vysvětlující (*environmental*), a ostatní jako kovariáty.

7.2. Data

Pro zjednodušení budeme pracovat jen s upravenými daty - nebudeme brát v potaz typ hnojení a zaměříme se pouze na celkovou dávku. Data jsou v souborech *fertil.spe* (snímky) a *fertil.env* (hnojiva a pokryvnost ječmene), nebo v excelovských souborech *fertspe.xls* a *fertenv.xls* (to proto, abyste si mohli vytvořit vlastní soubory pro CANOCO užitím programu WCanoImp). Dávky hnojiv jsou 0 pro nehnojené plochy, 1 pro 70 kg dusíku na hektar a 2 pro 140 kg N/ha.

7.3. Analýza dat

Doporučuji užít následující postup:

1. Spočtěte nepřímou ordinaci (*unconstrained ordination*), nejlépe DCA s Hillovým škálováním os. To ukáže celkovou variabilitu podél délky os, podle čehož můžeme usoudit na celkovou heterogenitu vegetačních dat. Délka první osy v Hillově škále je 3.8, což zapadá do „šedé zóny“, kde by se jak lineární, tak unimodální metody měly chovat vcelku rozumně. Poté užijte pasivní analýzu environmentálních dat (ta jsou promítnuta na ordinační osy *ex post*).

2. Spočtěte přímou ordinaci (*constrained ordination*) se všemi charakteristikami prostředí, které máte k dispozici. Délka gradientu z první analýzy slouží jako vodítko pro výběr mezi CCA a RDA. Dalším vodítkem by mělo být, zda očekávaná odezva na hnojení a pokryvnost ječmene je pro většinu druhů lineární, nebo zda pro některé druhy (eventuelně pro všechny) čekáme spíš optimum na gradientu. Zhruba se dá říci toto: pokud druhy na

gradientu mění svá zastoupení, pak užijte lineární aproximaci; pokud však očekáváme spíše kvalitativní změny druhového složení (objevení a zmizení mnoha druhů), pak jsou metody založené na unimodální odpovědi (*weighted averaging*) lepší.

Ve článku (Pyšek & Lepš 1991) jsme užili CCA. V tomto případě však užijeme RDA. RDA umožňuje standardizovné / ne-standardizované analýzy. Starndardizací vzorků můžeme oddělit rozdíly v celkové pokryvnosti od rozdílů v druhovém složení. Užijte RDA (s dávkami hnojiv a pokryvností ječmene jako charakteristikami prostředí) **ne**standardizovanou přes vzorky (pak výsledek odráží **jak** rozdíly v celkové pokryvnosti, **tak** rozdíly v relativním druhovém složení) i RDA přes vzorky standardizovanou (pak výsledek odráží **jen** rozdíly v relativním zastoupení jednotlivých druhů). Měli byste si být ale vědomi, že interpretace standardizovaných dat, založených na odhadu z ordinální stupnice, může být problematická.

Otestujte významnost Monte Carlo permutačním testem - *unconstrained permutations*. Nulová hypotéza tohoto testu zní: charakteristiky prostředí nemají žádný vliv na druhové složení, t.j. vliv obou proměnných je nulový. Zamítnutí nulové hypotézy pak znamená, že alespoň jedna proměnná na druhové složení rostlinného společenstva vliv má. Význam testu je analogický celkové ANOVě modelu v mnohonásobné regresi.

3. Spočtěte dvě oddělené dílčí omezené ordinace (*partial constrained ordinations*). V první užijte dávku hnojiv jako charakteristiku prostředí a pokryvnost ječmene jako kovariátu (*covariable*). Tato analýza odráží vliv hnojení, který nemůže být vysvětlen vlivem způsobeným větší pokryvností ječmene.

Ve druhé analýze užijte naopak pokryvnost ječmene jako charakteristiku prostředí a dávku hnojiv jako kovariátu. Tato analýza vypovídá o variabilitě způsobené rozdíly v pokryvnosti ječmene, kterou nelze přiřknout vlivu hnojení. Užijte jak analýzu standardizovanou vzorky, tak analýzu nestandardizovanou. V Monte Carlo permutačním testu je pak oprávněné použít permutace uvnitř bloků (*permutations within blocks*), bloky jsou definované jako plochy se stejnou hodnotou kovariáty (tj. se stejnou dávkou hnojiva). Je to proto, že dávka může nabývat tří hodnot a plochy ošetřené stejnou dávkou můžeme považovat za blok. Význam testu je analogický testu významnosti parciálních regresních koeficientů v mnohonásobné regresi.

Poznámka: v určitých případech se může přihodit (ne v tomto případě), že analýza s oběma proměnnými jako vysvětlujícími, je (vysoce) významná, ale při použití jedné jako vysvětlující a druhé jako kovariáty výsledek signifikantní není. Takový případ je podobný situaci v mnohonásobné lineární regresi, kdy ANOVA celkového regresního modelu signifikantní je, ale žádný z regresních koeficientů se významně neliší od nuly. To se stává, pokud jsou prediktory silně korelovány. Pak můžeme říct, že oba dohromady vysvětlují značnou část celkové variability, že ale nelze rozhodnout, který z nich je ten důležitější.

4. Rozklad variability (*variation partitioning*) (viz též sekce 4-10). Může nás zajímat, jaká část variability se dá vysvětlit čistě dávkou hnojiva, a jaká pouze pokryvností ječmene. Vzhledem k tomu, že obě proměnné jsou silně korelovány, je přirozené, že bude existovat část variability, o které sice budeme moct říct, že je způsobena jednou z těchto proměnných, ale nebude možné rozhodnout, kterou. Rozklad variability lze provést takto:

1. RDA s oběma proměnnými jako vysvětlujícími poskytne takovýto výsledek:

```
Axes                                    1       2       3       4  Total variance

 Eigenvalues                       :    .097    .046    .200    .131         1.000
 Species-environment correlations  :    .705    .513    .000    .000
 Cumulative percentage variance
Cumulative percentage variance
    of species data                :     9.7    14.3    34.4    47.5
    of species-environment relation:    67.8   100.0      .0      .0

 Sum of all unconstrained eigenvalues                                    1.000
 Sum of all canonical     eigenvalues                                     .143
```

Do CANOCA se vloudila malá chybička, *the sum of all unconstrained eigenvalues* má být zřejmě správně *sum of all eigenvalues*.

Výsledky ukazují, že 14.3% celkové variability se dá vysvětlit oběma proměnnými dohromady.

2. Nyní spočtěte parciální analýzu s pokryvností jako vysvětlující (environmentální) proměnnou a dávkou jako kovariátou:

```
Axes                                    1       2       3       4  Total variance

 Eigenvalues                       :   .074    .200    .131    .102         1.000
 Species-environment correlations  :    .590    .000    .000    .000
 Cumulative percentage variance
    of species data                :     8.0    29.5    43.6    54.5
    of species-environment relation:   100.0      .0      .0      .0

 Sum of all unconstrained eigenvalues                                     .931
 Sum of all canonical     eigenvalues                                     .074


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

Toto ukazuje, že pokryvnost vysvětluje 7.4% celkové variability, která nemůže být vysvětlena dávkou. Všimněte si, že nepoužíváme procento uvedené v *cumulative percentage variance* druhových dat (8%), protože to je spočteno z variability po odečtení vlivu kovariát, ale počítáme přímo s hodnotami kanonických charakteristických čísel.

3. Spočítáme parciální ordinaci s dávkou jako vysvětlující proměnnou a pokryvností jako kovariátou:

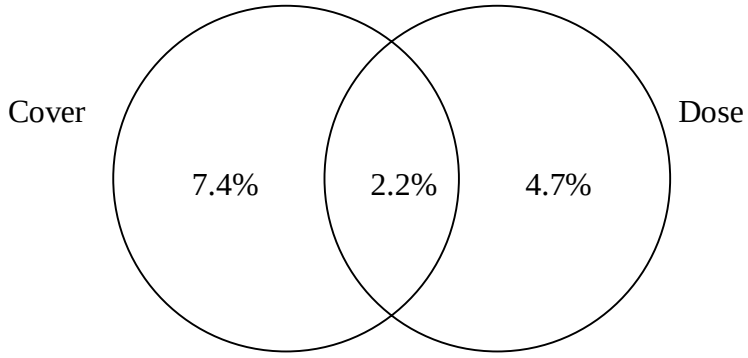
```
Axes                                    1      2      3      4  Total variance

 Eigenvalues                      :    .047   .200   .131   .102         1.000
 Species-environment correlations :    .500   .000   .000   .000
 Cumulative percentage variance
    of species data               :     5.2   27.4   41.9   53.2
    of species-environment relation:  100.0     .0     .0     .0

 Sum of all unconstrained eigenvalues                                    .904
 Sum of all canonical     eigenvalues                                    .047
```

Dávka vysvětluje 4.7% druhové variability, jež nemůže být připsána pokryvnosti.

4. Nyní můžeme počítat:

Celková vysvětlená variabilita je 14.3%. Z ní lze 7.4% vysvětlit čistě vlivem pokryvnosti ječmene, 4.7% dávkou hnojiva; to dává 12.1%. Pro zbývající 2.2% nelze rozhodnout, která z vysvětlujících proměnných je za ně odpovědná.

Rozdělení variability si můžeme ukázat na následujícím diagramu:



8. Případová studie 2: Hodnocení pokusů v úplných znáhodněných blocích

8.1. Úvod

Pro hodnocení jednorozměrné odpovědi (např. počtu druhů, celkové biomasy), použijeme pro hodnocení pokusů v úplných znáhodněných blocích dvoucestnou ANOVU bez interakcí (mean square (MS) interakce se použije jako *error term* - jmenovatel při výpočtu F-statistiky). V následujícím textu užijeme pro podobná hodnocení odpovědí společenstva CANOCO (mnohorozměrnou odpovědí je teď např. druhové složení společenstva). V tomto případě byl pokus založen jako čtyři úplné znáhodněné bloky - zásah byl čtyřúrovňový a odpověď byla zaznamenána jen jednou. Pokus byl detailně popsán v článku Špačková, Kotorová a Lepš (1998). Zde uvádím jen mírně zjednodušený popis experimentu.

Pokus jsme založili ve čtyřech úplných znáhodněných blocích v březnu roku 1994, krátce po roztání sněhu. Každý blok obsahoval čtyři zásahy: (1) odstranění stařiny, (2) odstranění stařiny a mechů, (3) odstranění dominantního druhu *Nardus stricta* a (4) kontrolu, kde vegetace zůstala nenarušena. Rozměr každé plochy byl 1 x 1 m. Původní pokryvnost *Nardus stricta* byla asi 25%. Její odstranění bylo velmi úspěšné s téměř nulovým obnovením. Odstranění *Nardus* na jaře způsobilo jen malé narušení půdy, které pak už v létě nebylo vůbec patrné.

8.2. Data

V každé ploše o velikosti 1 m^2 byla v srpnu 1994 vizuálně odhadnuta pokryvnost dospělých rostlin a mechů. V této době byl také uprostřed každé plochy vymezen další čtverec (0.5 x 0.5 m), který byl rozdělen na 25 plošek velikých 0.1 x 0.1 m. Pro každou plošku byla zaznamenána pokryvnost dospělých rostlin a počet semenáčků. V následujících příkladech použijeme jen souhrnné počty semenáčků na plochách 0.5 x 0.5 m. Data jsou ve formátu CANOCO (*seedl.spe*) nebo EXCEL (*seedlspe.xls*).

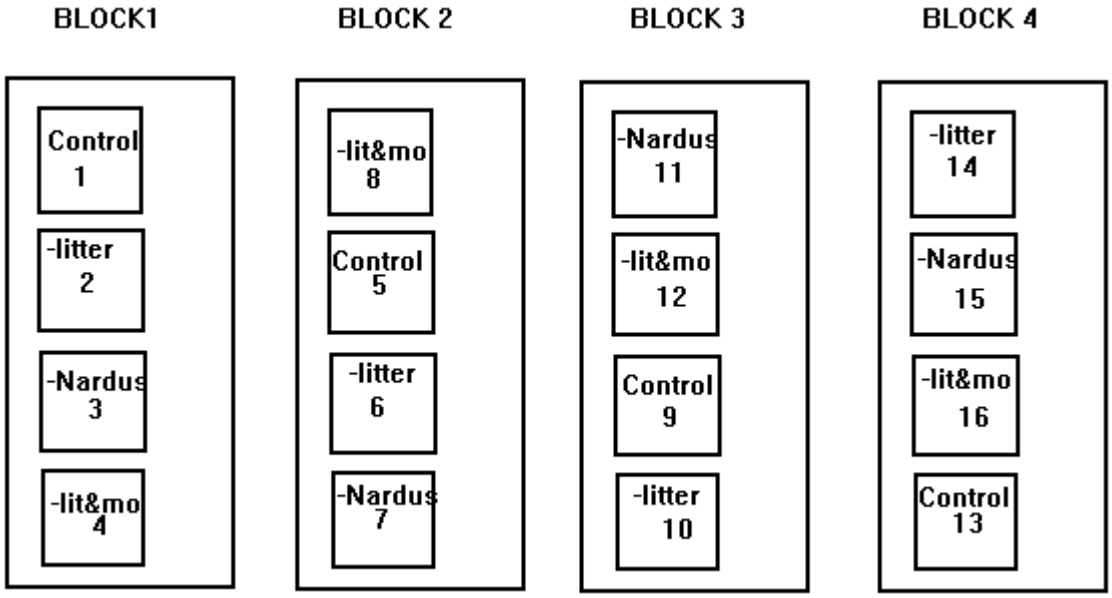
8.3. Analýza dat

Otázka: Je opravdu nutné používat mnohorozměrné metody? Nebylo by lepší testovat vliv na každý druh samostatně jednorozměrnou metodou (např. ANOVOU nebo Analýzou deviance[1])?

Odpověď: Mnohorozměrné metody jsou zde mnohem lepší. Při použití mnoha jednorozměrných testů totiž hrozí určité riziko: pokud provedeme několik testů s nominální hladinou významnosti $\alpha = 0.05$, pravděpodobnost chyby prvního druhu (Type I error) je 0.05 **pro každý jednorozměrný test**. To znamená, že pokud otestujeme, řekněme, 40 druhů, můžeme očekávat dva signifikantní výsledky čistě jako výsledek chyby prvního druhu. To může vést k „statistical fishing“, kdy si někdo vybere jen výsledky statisticky významných testů a snaží se je interpretovat. Riziko, o kterém jsme mluvili, lze obejít také tzv. Bonferroniho korekcí, kdy se nominální hladina významnosti dělí počtem provedených testů a ty se potom provedou na vypočtené hladině významnosti. Tento postup zajistí, že celková pravděpodobnost chyby prvního druhu je přinejmenším v jednom testu menší nebo

[1] Analýza deviance je metoda v zobecněných lineárních modelech, je to zobecnění ANOVA umožňující zpracovat data, která nemají normální distribuci (např. mají Poissonovo rozdělení).

rovna α, ale zároveň to vede k extrémně slabým testům. Já osobně (i když mnozí statistici by se mnou pravděpodobně nesouhlasili) považuji za vcelku korektní užití jednorozměrných metod pro jednotlivé druhy (bez korekce), když zjistíme, že celková odpověď společenstva je průkazná. Měli bychom mít ale stále na paměti, že pravděpodobnost chyby prvního druhu je v každém provedeném testu rovna α. Měli bychom si také uvědomit, že když vybereme druhy s nejsilnější odpovědí podle výsledků ordinace, pak tyto druhy poskytnou velmi pravděpodobně výrazně signifikantní rozdíly a jednorozměrné testy příliš mnoho nových informací k výsledkům ordinace nepřidají.



Obrázek 8-1: Design experimentu

Design pokusu je na obrázku 8-1. Každá pokusná plocha je charakterizována jednak **(1)** typem zásahu, a pak také **(2)** číslem bloku. Obě hodnoty jsou kategoriálními proměnnými a v programu CANOCO musí být kódovány jako indikátorové proměnné. Proměnné popisující strukturu bloku použijeme jako kovariáty a zásahy použijeme jako charakteristiky prostředí.

Odpovídající tabulka dat je v obrázku 8-2. Program Canoco se ptá zvlášť na environmentální data a zvlášť na kovariáty, takže tato data mohou být v oddělených souborech. Pokud použijeme společný soubor pro oba typy dat, musíme zadat stejný soubor (v tomto případě *seedl.env*; importovali jsme ho z Excelu, kde je k dispozici jako *seedlenv.xls*) jak pro charakteristiky prostředí, tak pro kovariáty. Pak vyřadíme první čtyři proměnné ze seznamu kovariát a poslední čtyři ze seznamu charakteristik prostředí. Společný soubor je velmi výhodný - v různých parciálních analýzách můžeme volit různé výběry charakteristik prostředí i kovariát.

V tomto případě není čtvrtá charakteristika prostředí nutná: pro kódování jedné kategoriální proměnné se čtyřmi kategoriemi postačí tři indikátorové proměnné. Čtvrtá proměnná bude z výpočtů vypuštěna, protože je lineární kombinací předešlých třech proměnných ("lit+moss" = 1 - "control" - "litter-r" - "nardus-r"). Přesto je ale vhodné ji do tabulky dat zahrnout, protože ji budeme potřebovat při kreslení výsledků programem CanoDraw. Pak budeme vynášet centroidy pro všechny čtyři kategorie. Podobně není nezbytně nutná ani zvláštní proměnná pro čtvrtý blok.

```
Ohrazeni 1994-seedlings, design of experiment
  (i3,8(f3.0))
  8
  1  1  0  0  0  1  0  0  0
  2  0  1  0  0  1  0  0  0
  3  0  0  1  0  1  0  0  0
  4  0  0  0  1  1  0  0  0
  5  1  0  0  0  0  1  0  0
  6  0  1  0  0  0  1  0  0
  7  0  0  1  0  0  1  0  0
  8  0  0  0  1  0  1  0  0
  9  1  0  0  0  0  0  1  0
 10  0  1  0  0  0  0  1  0
 11  0  0  1  0  0  0  1  0
 12  0  0  0  1  0  0  1  0
 13  1  0  0  0  0  0  0  1
 14  0  1  0  0  0  0  0  1
 15  0  0  1  0  0  0  0  1
 16  0  0  0  1  0  0  0  1

 control litter-rnardus-rlit+mossblock1  block2 bl.
 rel1    rel2    rel3    rel4    rel5    rel6   re.
 rel11   rel12   rel13   rel14   rel15   rel16
```

Obrázek 8-2: Environmentální data charakterizující uspořádání experimentu (soubor programu Canoco v plném formátu). Soubor má název **seedl.env**

Protože vegetace na ploše je velmi homogenní a my máme jen kategoriální vysvětlující proměnné, použijeme redundanční analýzu (RDA), metodu postavenou na lineárním modelu. Nyní můžeme otestovat přinejmenším dvě hypotézy.

První hypotéza může znít takto: **Zásahy nemají na semenáčky žádný vliv.** K zamítnutí hypotézy stačí, pokud se celkový počet semenáčků mezi jednotlivými zásahy bude lišit. Relativní zastoupení druhů semenáčků může zůstat nezměněno. Pokud se změní i zastoupení druhů, nulová hypotéza bude samozřejmě také zamítnuta.

Druhou hypotézu můžeme formulovat takto: **Relativní zastoupení semenáčků se neliší mezi jednotlivými zásahy** (to znamená, že semenáčky různých druhů se neliší ve svých reakcích na jednotlivé zásahy). První hypotézu můžeme testovat, pouze pokud nepoužíváme žádnou standardizaci po vzorcích - snímcích (v Canocu nabízeno jako default volba). Pokud použijeme standardizaci přes vzorky (nejčastěji *by sample norm*, tj. na jednotkovou délku vektoru), tak testujeme druhou hypotézu, totiž že zastoupení druhů mezi jednotlivými zásahy je neměnné. Všimněte si, že standardizace je součástí výpočtu vážených průměrů v kanonické korespondenční analýze (CCA): z toho vyplývá, že CCA nenajde rozdíl mezi plochami lišícími se v celkovém počtu semenáčů ale s konstantním zastoupením jednotlivých druhů. První test je silnější, ale signifikantní výsledek druhého testu je zase ekologicky zajímavější: fakt, že semenáče různých druhů reagují na jednotlivé zásahy různě, je dobrým argumentem pro význam regenerační niky pro udržení druhové rozmanitosti.

Výpočet RDA se provede klasickým postupem (viz manuál k programu CANOCO). Doporučujeme dát si pozor na následující:

(1) Pokud máte pro charakteristiky prostředí a kovariáty společný soubor, nezapomeňte vypustit příslušné proměnné. Vyloučení kovariát je důležitější, protože pokud je jedna a ta samá proměnná jak v kovariátách, tak v charakteristikách prostředí, je z charakteristik prostředí vypuštěna automaticky, neboť nevysvětluje žádnou variabilitu.

(2) Pokud chcete provést standardizaci přes vzorky a spouštíte CANOCO pod DOSem, užijte tzv. "dlouhý dialog" (*long dialog*).

original	block	perm1	perm2	perm3	perm 4	perm 5
1	1	2	4	1	etc.	
2	1	4	3	4	etc.	
3	1	3	2	2	etc.	
4	1	1	1	3	etc.	
5	2	7	5	7	etc.	
6	2	8	8	6	etc.	
7	2	5	7	8	etc.	
8	2	6	6	5	etc.	
9	3	11	9	9	etc.	
10	3	9	12	12	etc.	
11	3	10	10	11	etc.	
12	3	12	11	10	etc.	
13	4	14	16	14	etc.	
14	4	15	13	15	etc.	
15	4	16	15	13	etc.	
16	4	13	14	16	etc.	

Obrázek 8-3: Schéma permutace v rámci bloků

(3) Pokud provádíte Monte Carlo test, můžete žádat permutace v rámci bloků, podmíněné všemi třemi kovariátami (čtvrtá je kolineární a z výpočtů je vypuštěna). Každá permutační třída pak bude odpovídat jednomu bloku v pokusu. Permutace v rámci bloků jsou ukázány na obrázku 8-3. Tento postup se doporučoval hlavně u předchozích verzí programu Canoco (za předpokladu platnosti nulové hypotézy jsou zásahy v rámci bloku volně zaměnitelné). V novějších verzích jsou permutovány residuály po odečtení vlivu kovariát, pokud je vybrán tzv. redukovaný (*reduced*) model (residuály by měly být v případě platnosti nulové hypotézy volně zaměnitelné). Je tedy možné vybrat redukovaný model a neomezené permutace.

9. Případová studie 3: Analýza opakovaných pozorování druhové skladby ve faktoriálním pokusu: vliv hnojení, kosení a odstranění dominantního druhu v oligotrofní vlhké louce

9.1. Úvod

Opakovaná pozorování jsou pro mnohé oblasti ekologického výzkumu typická. První data jsou většinou sbírána na úplném začátku pokusu - ještě než se na plochách udělají experimentální zásahy. Tak získáme tzv. „baseline“ data, tj. data, kde jsou odlišnosti mezi plochami dány pouze náhodnou variabilitou. Po založení pokusu jsou data sbírána ještě jednou, eventuelně několikrát, abychom zachytili odlišnosti ve vývoji (dynamice) experimentálních a pokusných ploch. Takovému designu pokusu se někdy říká opakovaný (replikovaný) BACI (Before After Control Impact). Pro analýzu jednorozměrných odpovědí (např. počtu druhů nebo celkové biomasy) pak obvykle používáme model opakovaných měření (*repeated measurements*) ANOVy.

V podstatě máme dvě možnosti, jak data analyzovat: můžeme použít *split-plot* model ANOVy, kde je čas (tzv. *repeated measures factor*) považován za *within-plot* factor (*univariate repeated measurements ANOVA*), nebo lze použít také model MANOVA. Ačkoliv teoretické odlišení těchto dvou přístupů je složité, dá se říct, že obvykle je volen ten první, protože dává silnější test (je ovšem citlivější k narušení předpokladů, či spíše má více předpokladů). Vzájemný vztah mezi časem a zásahem odráží rozdíl ve vývoji pokusných ploch. CANOCO umí analyzovat opakovaná pozorování druhového složení podobným způsobem, jako to činí *repeated measurements ANOVA*. Odlišnost je v tom, že ANOVA testuje všechny efekty zároveň, zatímco v programu CANOCO se musí každý efekt testovat zvlášť. Popsaný přístup si ukážeme na analýze faktoriálního pokusu s přihnojováním, kosením a odstraněním dominantního druhu na oligotrofní vlhké louce. Popis pokusu je tady zjednodušen, jeho plná verze je v Lepš (1999).

9.2. Design pokusu

Pokus byl založen v roce 1994 ve faktoriálním uspořádání ve třech opakováních každé kombinace zásahů (obr. 9-1). Zásahy byly: hnojení, kosení a odstranění dominantního druhu (*Molinia caerulea*), to je 8 kombinací ve třech opakováních, tedy 24 ploch o rozměrech 2 x 2 m. Hnojení bylo prováděno každý rok aplikací 65 g/m^2 komerčního NPK hnojiva. Kosení probíhalo každoročně na konci června nebo na začátku července a posekaná tráva byla vždy z ploch odstraněna. *Molinia caerulea* byla odstraněna šroubovákem v dubnu roku 1995 s minimálním poškozením půdy. Noví jedinci byli z ploch odstraňováni každoročně.

Obrázek 9-1: Design pokusu

9.3. Snímkování

Snímkování ploch probíhalo každoročně (od roku 1994) v červnu nebo v červenci. Všimněte si, že první snímkování bylo provedeno ještě před započetím pokusných zásahů, aby byla k dispozici „baseline“ data pro každou plochu. Pokryvnost cévnatých rostlin a mechů byla vizuálně odhadnuta v centrálním 1 m^2 každé plochy o velikosti 2x2 m.

9.4. Analýza dat

Data jsou ve formě opakovaných měření; každá plocha byla osnímkována čtyřikrát. Pro výpočet jednorozměrných charakteristik (počet druhů) byl použit příslušný model *repeated measurements* ANOVA (*von Ende 1993*). Druhové složení jsem analyzoval za pomoci redundanční analýzy (RDA, *Renundancy Analysis*) v programu CANOCO s Monte Carlo permutačním testem. Grafická prezentace a výsledky ordinací jsou vytvořeny programy CanoDraw a CanoPost. RDA je metoda založená na lineární odpovědi druhů a byla použita proto, že druhové složení v plochách bylo poměrně homogenní a vysvětlující proměnné jsou kategoriální data. **Důležité je, že *Molinia* byla v analýzách použita jako pasívní druh**, neboť její odstranění byl jeden z experimentálních zásahů**. Pokud by jako pasívní použita nebyla, prokázalo by se (a to velmi významně), že *Molinia* má větší pokryvnost v plochách, ze kterých nebyla odstraněna...** Měli bychom také podotknout, že použitím různých kombinací vysvětlujících (v terminologii programu CANOCO environmentálních) proměnných a kovariát v RDA, spolu s vhodným permutačním schématem v Monte Carlo permutačním testu, jsme schopni udělat testy analogické těm, jaké pro testování významnosti jednotlivých efektů používá ANOVA (včetně *repeated measurements*). Díky tomu, že máme k dispozici „baseline“ data, je **interakce experimentálního zásahu a času** tím, co nás zajímá nejvíce. Při testování této interakce jsou charakteristiky ploch (kódované jako mnoho indikátorových proměnných) použity jako kovariáty. Tím od každé plochy odečteme průměr spočtený z několika let, a pro jednotlivé plochy analyzujeme jen tyto rozdíly. Hodnoty času jsou 0, 1, 2 a 3 pro roky 1994, 1995, 1996 a 1997 (v tomto pořadí). Toto odpovídá takovému modelu, kdy se plochy příslušející k jednotlivým zásahům na počátku pokusu (1994) od sebe neliší a analýzou je fitován lineární nárůst odlišnosti (tento

přístup je analogický spíše „single degree polynomial contrast“ než běžnému testování efektů v „repeated measurement“ ANOVA). Jinou možností je považovat čas za kategoriální (nominální) proměnnou (každý rok je zvláštní kategorií) a kódovat jej jako několik indikátorových proměnných (to odpovídá klasickému modelu ANOVA pro *repeated measurements*).

9.5. Technický popis

Druhová data jsou v Excelovském souboru *ohrazspe.xls*, design pokusu je v *ohrazenv.xls*. Použitím WCanoImp si připravíme soubor pro CANOCO (*ohraz.spe*) s druhy a soubor s environmentálními daty (*ohraz.env*). V obou souborech jsou vzorky (tj. druhy a jejich pokryvnosti) v následujícím pořadí: vzorky z roku 1994 mají čísla 1 až 24, z roku 1995 čísla 25 až 48, atd. Pořadí bude důležité pro popis permutačního schématu. Jména vzorků jsou *r94p1*, což znamená sebráno 1994 na ploše 1. V souboru s environmentálními daty popisují první tři proměnné zásah, který byl použit: 1 – použit, 0 – nepoužit. Další proměnná, *Year*, je čas do začátku pokusu, tj. čas jako kvantitativní proměnná. Další čtyři proměnné, Yr0, Yr1, Yr2 a Yr3, jsou indikátorové proměnné, které popisují rok jako kategoriální proměnnou (např. pro všechny záznamy z roku 1994 je Yr0=1 a Yr1=0, Yr2=0 a Yr3=0). Další proměnné, P1 až P24, jsou identifikátory ploch (tj. např. pro všechny záznamy z první plochy P1=1 a P2 až P24 jsou nula). Tento soubor použijeme jak pro charakteristiky prostředí, tak pro kovariáty, s tím, že z něj vždy vybereme (tedy spíše vypustíme) příslušné proměnné.

Pro testování vybraných hypotéz použijeme následující kombinace charakteristik prostředí a kovariát:

Tabulka 9-1: Výsledky RDA analýzy pokryvností druhů odhadnutých na plochách 1m x 1m. Data jsou centrována přes druhy. Standardizace přes vzorky použita nebyla (Standardization N). Vysvětlující proměnné jsou v Canoco terminologii charakteristiky prostředí. **% expl. 1-st axis**: procento druhové variability vysvětlené první osou, míra vysvětlující síly vysvětlujících proměnných. **R 1-st axis**: korelace druhu a prostředí na první ose. **F-ratio**: F statistika pro sledovaný test. **P**: příslušná pravděpodobnost získaná Monte Carlo permutačním testem, 499 náhodných permutací (tj. pravděpodobnost chyby Typu I při testování hypotézy, že vliv všech vybraných vysvětlujících proměnných je nulový). **Yr** – pořadové číslo roku, **M** – kosení, **F** – hnojení, **R** – odstranění Molinie, **PlotID** – identifikace plochy. ***** mezi dvěma efekty znamená interakci.

Analysis	Explanatory variables	Covariables	Standardization	% expl. 1-st axis	R 1-st axis	F-ratio	P
C1	Yr, Yr*M, Yr*F, Yr*R	PlotID	N	16.0	0.862	5.38	0.002
C2	Yr*M, Yr*F, Yr*R	Yr, PlotID	N	7.0	0.834	2.76	0.002
C3	Yr*F	Yr, Yr*M, Yr*R, PlotID	N	6.1	0.824	4.40	0.002
C4	Yr*M	Yr, Yr*F, Yr*R, PlotID	N	3.5	0.683	2.50	0.002
C5	Yr*R	Yr, Yr*M, Yr*F, PlotID	N	2.0	0.458	1.37	0.040

Nulové hypotézy testů pro jednotlivé analýzy:

- **C1**: V druhové skladbě nejsou žádné směrované změny, a to ani změny společné pro všechny zásahy, ani změny pro jednotlivé zásahy specifické (toto odpovídá testu všech „within subject“ efektů v *repeated measured ANOVA*).
- **C2**: Časový posun v druhové skladbě je nezávislý na zásahu.
- **C3** (**C4**, **C5**): Hnojení (resp. odstranění, kosení) nemá na změny druhové skladby žádný vliv (toto odpovídá testům jednotlivých efektů v *repeated measures ANOVA*).

Všimněte si, že pokud je PlotID použito jako kovariáty, potom hlavní vlivy (jako M, F, a R) nevysvětlují žádnou variabilitu a je zbytečné je používat, ať už jako kovariáty, nebo jako vysvětlující proměnné.

V této analýze považujeme čas za kvantitativní (kontinuální) proměnnou. To znamená, že použijeme *Year* a vypustíme Yr0 až Yr3. Odpovídá to hledání lineárního trendu v datech. Pokud budeme hledat obecné rozdíly v dynamice, budeme čas považovat za proměnnou kategoriální a použijeme Yr0 až Yr3 (jedna z nich je sice nadbytečná, ale zato užitečná pro ordinační diagramy), přičemž vypustíme *Year*. V tomto případě ale musí být interakce mezi zásahem a časem definována jako interakce navzájem mezi všemi indikátorovými proměnnými charakterizujícími rok a všemi indikátorovými proměnnými charakterizujícími zásahy.

Nulová hypotéza C1 je trošku komplikovaná a její ekologická interpretace je složitější; tato analýza je užitečná pro srovnání s dalšími analýzami vysvětlujícími variabilitu a korelaci druhů a prostředí na první ose. Rovněž permutační schéma není úplně jednoznačné. U ostatních analýz je v případě platnosti nulové hypotézy dynamika nezávislá na aplikovaném zásahu. To znamená, že pokud je nulová hypotéza pravdivá, pak jsou plochy zaměnitelné; přičemž ale záznamy z téže plochy musí zůstat pohromadě (vyjádřeno technicky, záznamy z různých let na téže ploše jsou podplochami této plochy a jen hlavní

plochy jsou permutovány). K tomu, abychom toho dosáhli, by měly být v dialozích zvoleny následující volby:

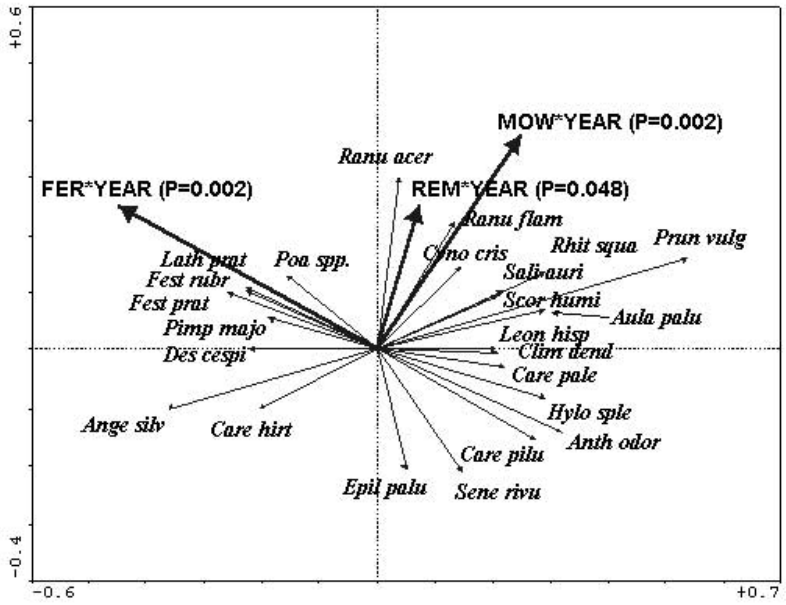
K provedení testu obvykle používám volbu *Both above tests*, ačkoli postup, kdy se udělají oba testy a vybere se ten s lepšími výsledky, není úplně korektní (vlastně není korektní vůbec, ale většina lidí to tak dělá). Lze očekávat, že test první ordinační osy bude silnější v přítomnosti jednoho dominantního gradientu, zatímco test všech omezených (*constrained*) os bude silnější, pokud je v datech několik nezávislých gradientů. Všimněte si, že s jedinou vysvětlující proměnnou budou test pro první osu i test pro všechny osy stejné; logicky: jedna osa jsou všechny osy. Doporučuji použít test založený na redukovaném modelu (*reduced model*). Pokud budete mít čas, dobrý počítač a úkol o přiměřené velikosti, můžete počet permutací zvýšit. Permutace by měly být *Restricted for spatial or temporal structure or split-plot design*, potom v následujícím okně *Split-plot design* má být *Number of split-plots in whole-plot* 4 (tj. 4 záznamy z každé plochy). Budou vybírány pravidlem *Take 1 Skip 23*. To odpovídá pořadí záznamů v našem souboru: záznam z každé plochy je oddělen 23 záznamy z jiných ploch.[*] Celé plochy (*whole-plots*) jsou volně zaměnitelné a na *split-plot* úrovni se nedoporučuje dělat žádné permutace (podle manuálu lze permutovat i čas, ale základ pro to je poněkud méně jistý než pro permutaci míst). Po proběhnutí analýzy je vhodné překontrolovat log soubor permutačního schématu. V našem případě by měl vypadat takto:

```
*** Sample arrangement in the permutation test ***

 Whole plot     1 :
     1    25    49    73
 Whole plot     2 :
     2    26    50    74
 Whole plot     3 :
     3    27    51    75
etc...
```

což potvrzuje, že *whole-plots* jsou složeny ze záznamů ze stejných ploch.

Vztah mezi jednotlivými druhy a experimentálními zásahy můžeme znázornit ordinačním diagramem. Pravděpodobně nejlepší možností je zobrazení výsledků analýzy C2 pomocí projekčního diagramu s druhy a charakteristikami prostředí (*environmental variables – species biplot*, Obr. 9-2). Protože čas je brán jako kontinuální proměnná, je interakce času se zásahy také proměnnou kontinuální, což je vidět z použití šipek. A protože čas je zde použit také jako covariable, budou trendy znázorněny relativně k průměrnému trendu ve všech typech zásahů. Například souhlasný směr šipky druhu s šipkou **FER*YEAR** znamená buď že se pokryvnost druhu zvětšuje v hnojených plochách nebo že klesá v plochách nehnojených (nebo oboje).

[*] Dejte si pozor na jiný případ, kdy děláte permutace uvnitř bloků (*within blocks*). Potom počet ploch k přeskočení je počet ploch k přeskočení ***within the block***.



Obrázek 9-2: Výsledky analýzy na základě hypotézy C2

9.6. Další užití ordinačních výsledků

Výsledky analýz mohou být dále využity. Jedna z možností je použít druhová skóre. Pokud byly interakce **čas*zásah** jedinými vysvětlujícími proměnnými a ostatní faktory byly použity jako kovariáty(C3 až C5) mohou být skóre druhů na omezené ose považována za charakteristickou odpověď druhu na zásah. Potom jako prediktory této odpovědi byly testovány následující biologické charakteristiky druhů:

1. Výška druhu, braná jako střed hodnot uváděných pro místní květenu.

2. Přítomnost arbuskulární mykorrhizy (AM), založená na datech Grime *et al.* (1988) a na ekologické databázi flory (už k ní nemáme přístup).

3. Relativní rychlost růstu semenáčů (RGR) z Grime *et al.* (1988).

Protože jsem očekával, že druhy podobné *Molinii* vytěží z jejího odstranění nejvíc, použil jsem pro předpověď vlivu odstranění Molinie ještě čtvrtou (ordinální) proměnnou – odlišnost druhu od Molinie. Podobnost jedna byla přiřazena graminoidům vyšším než 50 cm, dvojka širokolistým travám nižším než 50 cm, trojka úzkolistým travám nižším než 50 cm a čtyřka dvouděložným. Pro analýzu vztahu těchto hodnot s RDA skóre odstranění Molinie byla použita Spearmanova korelace.

Technické provedení: výsledky analýz můžeme přetáhnout ze *.sol* souboru do listu Excelu, tj. načíst soubor jako „delimited“ (tzn. s oddělovači). Potom můžeme použít skóre druhů a dodat tu informaci, kterou potřebujeme (tj. výšku rostlin, mykorrhizu, atd.). V tomto případě je rozumné vynechat druhy v datech řídce zastoupené (tyto druhy ve většině vzorků nedávají žádnou odpověď, a to prostě proto, že nebyly přítomny ani před ani po zásahu).

10. Triky a pravidla při používání ordinačních metod

V této kapitole jsou praktické rady, které nám přišly při analýzách mnohorozměrných dat ordinačními metodami užitečné. Jsou roztříděny do několika sekcí.

10.1. Volby škálování

- CanoDraw 3.x nebo nějaký jiný, obecnější statistický balík můžeme použít k fitování křivky unimodální odpovědi pro vybranou podmnožinu druhů ve vztahu k vybraným ordinačním osám. Abychom toho dosáhli, budeme fitovat zobecněný lineární model (GLM) s předpokládanou Poissonovou distribucí a logaritmickou link funkcí za použití abundancí příslušných druhů jako vysvětlované proměnné a polynomu druhého stupně skóre vzorků na vybrané ose jako proměnné vysvětlující. Nejlepších výsledků dosáhneme, zvolíme-li (v DCA nebo CCA) Hillovo škálování se zaměřením na vzdálenosti mezi vzorky.
- Pokud v lineární ordinační metodě (PCA, RDA) zvolíme, že skóre druhů se mají dělit jejich standardními odchylkami, bude délka šipky každého druhu vyjadřovat kvalitu aproximace hodnoty druhu ordinačním diagramem. Pokud se ale naopak rozhodneme dělení neprovádět, budou délky šipek vyjadřovat variabilitu hodnot druhů v zobrazeném ordinačním (pod)prostoru.

10.2. Permutační testy

- Potřebujeme-li otestovat významnost druhé kanonické osy, měli bychom vytvořit nový projekt v programu Canoco, který bude podobný původnímu, ale skóre vzorků na první kanonické ose bude specifikováno jako (další) kovariáta. Pokud jsme původně žádné kovariáty neměli, uděláme to jednoduše tak, že určíme celý výsledkový (**.sol**) soubor původní analýzy jako soubor s kovariátami. Program Canoco je schopen výsledkový soubor rozebrat a najít si skóre vzorků. Uvědomte si ale, že použitelná skóre vzorků jsou zde ta, která jsou **lineární** kombinací charakteristik prostředí. Proto chceme-li po programu Canoco, aby našel správná skóre, musíme změnit text "*Sample score*" před sekcí **Samp** na něco jiného (např. "*xample score*", jak navrhuje manuál). Pak program Canoco při analyzování výsledkového souboru "nenajde" první sekci a skóre čte až ze sekce **SamE**. Po úpravě souboru bychom měli v projektu určit, že jako kovariáta se má použít pouze skóre **AX1** a proměnné **AX2**, **AX3** a **AX4** se mají ignorovat. Při testování významnosti třetí kanonické osy bychom jako kovariáty měli zachovat jak skóre **AX1**, tak i **AX2**.
- Při fitování omezeného ordinačního modelu bychom měli snížit vzájemnou korelaci mezi vysvětlujícími proměnnými (charakteristikami prostředí). Přítomnost několika silně korelovaných vysvětlujících proměnných činí výsledný omezený ordinační model nevěrohodným, podobně jako stejná situace v regresních modelech. Přibližným kritériem, kterým bychom se měli řídit, je VIF (*variance inflation factor*, viz Ter Braak et Šmilauer, 1998, str. 119), který by pro žádnou proměnnou neměl překročit hodnotu zhruba 20.

- K udržení rozumné síly (relativně nízké pravděpodobnosti chyby Typu II dané pravděpodobností chyby Typu I) Monte Carlo permutačního testu, lze odhadnout doporučený počet permutací zvolením požadované přesnosti odhadu hladiny významnosti - $\mathbf{P_0}$ (řekněme 0.01) a spočítáním počtu permutací podle vztahu $N = (10/P_0)-1$, tj. v našem případě 999.

10.3. Další problémy

- Uživatelé mají často tendenci ignorovat třetí a čtvrtou ordinační osu. Tyto osy sice bývají obvykle neinterpretovatelné a v případě přímých gradientových analýz také často bez průkazného vlivu, tyto závěry se ale musí vždy prozkoumat a / nebo otestovat Monte Carlo permutačním testem.
- Neměli bychom propadat panice, pokud je variabilita vysvětlená předloženými charakteristikami prostředí v omezeném ordinačním modelu nízká, zvlášť pokud analyzujeme data typu přítomen / nepřítomen nebo data s mnoha nulami. V takových případech lze často nalézt dobře interpretovatelné struktury, i když množství vysvětlené variability nepřekročí 10 %. Měli bychom rozlišovat mezi statisticky významným vztahem a silou takového vztahu.
- Nejjednodušším způsobem, jak spočítat matici korelačních koeficientů mezi druhy a charakteristikami prostředí je nadefinovat projekt programu Canoco pomocí redundanční analýzy (RDA) a vycentrovat a standardizovat přes druhy. Poté, co projekt zanalyzujeme, bude v adresáři s výsledky soubor **spec_env.tab** s těmito koeficienty ve vstupním formátu Canoca.
- Výsledky dvou analýz vycházející ze stejného souboru dat můžeme porovnat tak, že výsledkový soubor prvního projektu určíme jako soubor s doplňkovými (pasivními) charakteristikami prostředí v rámci druhého Canoco projektu. Podotýkáme ale, že CanoDraw 3.x doplňkové (supplementary) proměnné nerozeznává, a tedy ani nezobrazuje diagramy s jejich skóre.
- Měli bychom si dát dobrý pozor na pojmenování charakteristik prostředí a kovariát, které se používají pro definici interakčních členů (*interaction terms*). Význam proměnné bychom měli zachytit v prvních čtyřech znacích její jmenovky, protože pouze tyto znaky se použijí při definici jmen interakcí prvního stupně. V opačném případě můžeme skončit s několika nerozlišitelnými jmény, která brání snadné interpretaci ordinačních diagramů.

11. Moderní regrese: úvod

Díky regresním modelům můžeme modelovat závislost (obvykle) jedné vysvětlované proměnné (kvantitativní či kvalitativní) na jednom nebo více prediktorech. Prediktory mohou být opět kvantitativní a / nebo faktoriální proměnné. Podle této definice pak regresní modely zahrnují i metody jako je analýza variance (ANOVA) nebo analýzy kontingenčních tabulek.

Během 80. let se objevilo více nových typů regresních modelů, které ve větší či menší míře rozšiřují modely předcházející. V této kapitole najdete krátké shrnutí těch metod, které jsme pro studie ekologických dat pokládali za nejužitečnější.

11.1. Regresní modely obecně

Všechny regresní modely sdílejí určité základní požadavky na vysvětlované proměnné a prediktory. Pro jednoduchost se omezíme jen na nejčastější typy regresních modelů, kde je pomocí jednoho nebo několika prediktorů modelována právě jedna vysvětlovaná proměnná.

Nejjednodušší způsob, jak popsat jakýkoliv typ takového regresního modelu je tento:

$$\mathbf{Y = EY + e}$$

kde **Y** označuje hodnoty vysvětlované proměnné, **EY** je hodnota vysvětlované proměnné, kterou očekáváme pro určité hodnoty prediktorů a **e** je variabilita skutečných hodnot okolo očekávané hodnoty EY. Očekávanou hodnotu vysvětlované proměnné můžeme formálně popsat jako funkci hodnot prediktorů:

$$\mathbf{EY = f(X_1, ..., X_p)}$$

Složce EY říkáme také **systematická**, zatímco e je **stochastickou složkou** modelu. Jejich obecné vlastnosti a odlišné role v regresních modelech popisuje tabulka 11-1.

Stochastická složka	Systematická složka
Odráží *a priori* předpoklady modelu	Je určená naší hypotézou
Její parametr(y) se odhadují během nebo po fitování (variance vysvětlované proměnné)	Její parametry (regresní koeficienty) se odhadují při fitování modelu
Používáme ji k odhadu kvality modelu (**k regresní diagnostice**)	Interpretujeme ji a její parametry vizualizujeme a testujeme

Tabulka 11-1

Když na naše data fitujeme určitý regresní model, jsou naše předpoklady o stochastické složce (obvykle o distribučních vlastnostech a o nezávislosti či o určitém typu závislosti mezi jednotlivými pozorováními) již dány, měníme jen obsah a komplexitu systematické složky. Systematickou složku můžeme vyjádřit v nejjednodušším lineárním regresním modelu s jednou vysvětlovanou proměnnou Y a jedním prediktorem X takto:

$$\mathbf{EY = f(X) = \beta_0 + \beta_1 * X}$$

Ve skutečnosti lze dosáhnout větší komplexity modelu závislosti proměnné Y na X, a to díky polynomiálním modelům. Komplexita modelu se pohybuje na gradientu počínajícím **nulovým modelem** $EY = \beta_0$, přes lineární závislost popsanou výše, kvadratickou polynomiální závislost $EY = \beta_0 + \beta_1 * X + \beta_2 * X^2$, až k polynomům **n**-tého stupně, kde **n** je

počet pozorování snížený o jedna. Tento nejkomplexnější model (zde **model plný**) prochází všemi daty velmi přesně, ale neposkytuje žádné zjednodušení skutečnosti (což je u modelů jeden ze základních účelů). Prostě jen nahrazuje *n* hodnot *n* regresními koeficienty (β_0, ..., β_{n-1}). Nulový model však na druhé straně zjednodušuje situaci až příliš, takže ani z něho se o datech nic nového nedozvíme (a rozšiřování našich poznatků je další z podstatných funkcí modelů).
Z naší diskuze o těchto extrémních případech vyplývá, že při výběru komplexity modelu se pohybujeme od jednoduchých, nepříliš přesných modelů k modelům (až příliš) komplexním. Složité modely mají ještě jednu nevýhodu: jsou sice velmi dobře fitovány na naše data, pro nesebranou část statistické populace však mohou poskytnout zkreslené (*biased*) předpovědi. Naší snahou je nalezení nějakého kompromisního modelu - modelu dostatečně jednoduchého, avšak jen tak, aby byl stále ještě funkční. Takovému modelu se často říká *parsimonious model***.**

11.2. Obecný lineární model: pojmy

První důležitou zastávkou na naší pouti typy regresních modelů jsou metody obecného lineárního modelu. Všimněte si slova **obecného** (*general*) - další typy modelů používají na stejném místě slova **zobecněné** (*generalized*), což však znamená něco jiného. Ve skutečnosti jsou zobecněné lineární modely (GLM), o nichž je řeč v příští kapitole, na obecných lineárních modelech založeny a jsou jejich zobecněním*.

Podíváme-li se na hlavní odlišnosti obecných lineárních modelů a tradičních lineárních modelů z hlediska uživatele, zjistíme především, že jako prediktory lze použít jak kvantitativní, tak kvalitativní proměnné (faktory). ANOVA tedy patří do rodiny obecných lineárních metod. Pro jednoduchost si můžeme představit, že každý faktor s **k+1** hladinami nahradíme **k** indikátorovými proměnnými. Všimněte si, že metoda, kterou používá Canoco (indikátorové proměnné s hodnotami 1 pro přítomnost v odpovídající třídě a 0 pro nepřítomnost), není jediným možným řešením.

Pak tedy můžeme obecný lineární model popsat následujícím vztahem:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j * X_{ji} + \varepsilon$$

musíme si však uvědomit, že faktor bývá obvykle zastoupen více než jedním prediktorem X_j, a tedy více než jedním regresním koeficientem. Symbol ε se vztahuje k náhodné, stochastické proměnné zastupující stochastickou složku regresního modelu. V kontextu obecných lineárních modelů se předpokládá, že tato proměnná má nulový průměr a konstantní varianci.

Tento popis modelu nám také odhaluje jednu velmi důležitou vlastnost obecných lineárních modelů - jejich **aditivitu**. Vlivy jednotlivých prediktorů jsou vzájemně nezávislé. Pokud například zvýšíme hodnotu jednoho prediktoru o jednu jednotku, má to konstantní vliv (vyjádřený hodnotou regresního koeficientu odpovídajícího této proměnné), nezávislý na hodnotách jiných proměnných a nezávislý dokonce na původní hodnotě proměnné, kterou zvyšujeme.

Uvedený vztah popisuje model pro (hypotetickou) populaci všech možných pozorování, ze které vybíráme při sběru dat. Z takového výběru konečné velikosti pak odhadujeme skutečné hodnoty regresních koeficientů $\boldsymbol{\beta_j}$, značené obvykle $\mathbf{b_j}$. Vezmeme-li

* Tato věta byla opravdu míněna vážně :}

pozorované hodnoty prediktorů, lze **fitované** (předpovězené) hodnoty vysvětlované proměnné spočítat takto:

$$\hat{Y} = b_0 + \sum_{j=1}^{p} b_j * X_j$$

Fitované hodnoty nám dovolují odhadnout "realizace" náhodné proměnné reprezentující stochastickou složku - taková "realizace" se nazývá **regresní reziduál** a značí se $\mathbf{e_i}$:

$$e_i = Y_i - \hat{Y}_i$$

Reziduál je tedy rozdíl mezi pozorovanou hodnotou vysvětlované proměnné a odpovídající hodnotou předpovězenou fitovaným regresním modelem.

Variabilitu hodnot vysvětlované proměnné můžeme vyjádřit jako **celkovou sumu čtverců (total sum of squares)**, definovanou jako

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

V kontextu fitovaného regresního modelu je možné tuto variabilitu rozdělit na dvě části - na variabilitu vysvětlenou fitovaným modelem - tzv. **sumu čtverců modelu** (*model sum of squares*), definovanou jako

$$MSS = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$$

a tzv. **reziduální sumu čtverců** (*residual sum of squares*) definovanou jako

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2$$

Platí, že TSS = MSS + RSS. Tyto statistiky můžeme použít k otestování významnosti modelu. V případě platnosti celkové nulové hypotézy ("vysvětlovaná proměnná je na prediktorech nezávislá") se MSS od RSS nebude lišit, zrelativizujeme-li oba počtem jejich stupňů volnosti. †

11.3. Zobecněné lineární modely

Zobecněné lineární modely (*generalized linear models,* GLM, *McCullagh & Nelder, 1989*) rozšiřují obecné lineární modely ve dvou významných, navzájem svázaných aspektech. Zaprvé, **očekávané** hodnoty vysvětlované proměnné (**EY**) nemusí být vždy lineární kombinací prediktorů. Závislost škály vysvětlované proměnné na škále prediktorů je definována jednoduchou parametrickou funkcí, které se říká **link funkce:**

$$g(EY) = \eta$$

kde η je **lineární prediktor** a definuje se podobně jako celá systematická složka **obecného** lineárního modelu, tedy:

$$\eta = \beta_0 + \sum \beta_j X_j$$

† Na podrobnosti se podívejte do jakékoliv slušné učebnice, např. do Sokal & Rohlf (1995)

Výhodou link funkcí je, že umožňují "přemapování" hodnot ze skutečné škály lineárního prediktoru (obecně od -∞ do +∞) do specifického intervalu, který je pro vysvětlující proměnnou smysluplnější (např. hodnoty v intervalu 0 až 1 pro určení pravděpodobností).

Druhým odlišujícím aspektem je, že zobecněné lineární modely nemají tak specifické předpoklady o stochastické složce jako obecné lineární modely. Její variance nemusí být konstantní (může být závislá na očekávané hodnotě vysvětlované proměnné EY) a pro její předpokládanou statistickou distribuci (a tím i pro vysvětlovanou proměnnou) jsou zde různé volby.

Avšak volby link funkce a předpokládaný typ distribuce nemůžeme kombinovat zcela nezávisle. Například **logit** link funkce přemapuje skutečnou škálu do intervalu mezi 0 a +1, což není zcela vhodné řekněme pro předpoklad Poissonovy distribuce. V tabulce 11-2 jsou typické kombinace link funkcí a očekávaných distribucí vysvětlované proměnné, spolu s charakterizací vysvětlovaných proměnných, které daným předpokladům odpovídají.

Typ proměnné	"Typická" link funkce	Referenční distribuce
počty (frekvence)	log	Poissonova
pravděpodobnost (relativní frekvence)	logit nebo probit	binomická
rozměry, poměry	inverze nebo log	gamma
vzácné typy měření	identita	Gaussova ("normální")

Tabulka 11-2

S tím, co jsme se již dozvěděli, můžeme shrnout, jaké typy regresních modelů jsou v GLM obsaženy:

- "klasické" obecné lineární modely (včetně většiny typů analýzy variance)
- rozšíření těchto klasických lineárních modelů na proměnné s nekonstantní variancí (počty, relativní frekvence)
- analýzy kontingenčních tabulek (pomocí log-lineárních modelů)
- modely pravděpodobností přežití používané v toxikologii (probit analýzy)

Zobecněné lineární modely nepoužívají pojem reziduální sumy čtverců. Místo toho používají k vyjádření odlišnosti skutečných hodnot vysvětlované proměnné a hodnot předpovězených modelem **devianci**. Takže na hodnocení kvality modelu se používají testy založené na **analýze deviance**, což je koncepčně podobné analýze variance v klasickém regresním modelu.

Linearita, jakožto důležitá vlastnost obecných lineárních modelů, je ve zobecněných lineárních modelech zachována na škále lineárního prediktoru. Vliv určitého prediktoru se vyjadřuje jediným parametrem – lineárním transformačním koeficientem (regresním koeficientem). Podobně je na lineární škále prediktoru zachována i aditivita. Škála vysvětlované proměnné však může vypadat zcela jinak. Například u logaritmické link funkce odpovídá aditivita na škále prediktoru násobnému vlivu na škále odpovědi (užíváno např. při analýze kontingenčních tabulek).

11.4. Loess smoother

Termín **smoother** se používá pro regresní model, který se pokouší popsat (obvykle neparametricky) očekávané hodnoty odpovědi pro jednotlivé hodnoty prediktoru. Takto získané hodnoty mají menší variabilitu než hodnoty skutečně pozorované (proto termín *smoother*).

Existuje několik typů těchto "vyhlazujících" funkcí, z nichž některé nejsou příliš dobré, zato jsou ale jednoduché na pochopení. Jedním z nich je klouzavý průměr (*moving average*). Příkladem lepší metody je tzv. **loess smoother** (dříve zvaný též *lowess*). Je založen na tzv. lokálně vážené lineární regresi (*Cleveland & Devlin 1988, Hastie & Tibshirani 1990*). Pro odhad hodnoty vysvětlované proměnné používá lineární regresní model a kombinaci několika hodnot prediktorů. Pro fitování modelu používá jen taková pozorování, kdy jsou hodnoty prediktorů dostatečně blízko odhadovanému bodu. Taková oblast (u jediného prediktoru pás) kolem odhadovaného bodu, ze které se vybírají data pro fitování lokálního regresního modelu, je určena parametrem, který se nazývá **bandwidth** ("šířka pásu") a je určena jako frakce všech dat, jež jsou k dispozici. Pak tedy bandwidth hodnota α=0.5 říká, že pro každý odhadovaný bod se v regresi použije polovina pozorování (ta nejbližší uvažované kombinaci hodnot prediktorů). Komplexita lokálního regresního modelu se určuje druhým parametrem, kterým je **stupeň** (*degree*, λ). Obvykle se používají jen dvě hodnoty: pro lineární regresní model 1 a pro model polynomu druhého stupně 2.

Mimoto mají data používaná k fitování modelu lokální regrese různou váhu. Ta závisí na jejich vzdálenosti od uvažovaného odhadovaného bodu v prostoru prediktorů. Mají-li data přesně hodnotu požadovaných prediktorů, je jejich váha rovna 1.0 a váha směrem k hranicím pásu postupně klesá k 0.0.

Důležitou vlastností loess modelu je, že jeho komplexitu lze vyjádřit - stejně jako v tradičních lineárních regresních modelech - počtem stupňů volnosti (DF) odebraných z dat fitovaným modelem. Říká se jim také **ekvivalentní počet parametrů**. Dále, protože loess model předpovídá hodnoty vysvětlované proměnné (tak jako jiné modely), můžeme variabilitu vysvětlenou modelem oddělit a porovnat ji s reziduální sumou čtverců. Protože máme odhadnutý počet DF modelu, můžeme spočítat reziduální DF a sumu čtverců připadající na jeden stupeň volnosti (to odpovídá *mean square* v analýzách variance v klasickém regresním modelu). Následně tedy lze porovnávat regresní modely analýzou variance podobně, jako to provádíme pro obecné lineární modely.

11.5. Zobecněné aditivní modely

Zobecněné aditivní modely (*generalized additive models*, GAM, *Hastie & Tibshirani, 1990*) poskytují zajímavé rozšíření zobecněných lineárních modelů (GLM). Lineární prediktor GLM je zde nahrazen tzv. **aditivním prediktorem**. Je to také suma nezávislých příspěvků jednotlivých prediktorů, nicméně vliv konkrétních prediktorů se nevyjadřuje jako jednoduchý regresní koeficient. Místo toho je pro **j**-tý prediktor určen tento vliv hladkou funkcí $\mathbf{s_j}$, která popisuje transformaci hodnot prediktoru na (aditivní) vliv tohoto prediktoru na očekávané hodnoty vysvětlované proměnné.

$$\eta_A = \beta_0 + \sum s_j(X_j)$$

Aditivní škála prediktoru je opět spojena se škálou vysvětlované proměnné link funkcí.

Vidíme, že zobecněné aditivní modely zahrnují zobecněné lineární modely jako svůj zvláštní případ, kde je transformační funkce pro každý prediktor definována takto:

$$s_j(X_j) = \beta_j * X_j$$

V obecnějších případech jsou však transformační funkce (zvané obvykle *smooth terms*) fitovány pomocí neparametrických modelů typu *loess* nebo *cubic spline*. Při fitování zobecněného aditivního modelu nepředpisujeme tvar těchto funkcí jednotlivých prediktorů, musíme ale předem zvolit složitost jednotlivých křivek pomocí jejich stupňů volnosti. Pro

jednotlivé prediktory obvykle musíme zvolit typ transformační funkce, který použijeme pro vyhledání tvaru příslušného "hladkého členu" (*smooth term*).

Se zobecněnými aditivními modely můžeme také provádět postupný výběr, a to nejen výběr prediktorů použitých v systematické části modelu, ale také výběr složitosti jejich *smooth terms*.

Zobecněné aditivní modely nelze jednoduše numericky shrnout, na rozdíl od zobecněných lineárních modelů, kde jejich primární parametry - regresní koeficienty - charakterizují podobu regresního modelu. Nejlepším způsobem, jak sumarizovat zobecněné aditivní modely, je vynést odhadnuté hladké transformační funkce, reprezentující vztah mezi hodnotami prediktoru a jeho vlivem na danou vysvětlovanou proměnnou.

11.6. Klasifikační a regresní stromy

Regresní metody založené na tvorbě stromů jsou snad nejvíce neparametrickými metodami, jaké můžeme najít pro popis závislosti hodnot vysvětlované proměnné na hodnotách prediktorů. Definují se rekurzivním binárním dělením souboru dat do podskupin, které jsou postupně v hodnotách vysvětlované proměnné čím dál tím homogennější. V každém dělícím kroku se k binárnímu rozdělení používá právě jeden z prediktorů - a to buď prediktor kvantitativní nebo kvalitativní. Vybírá se takové rozdělení, které maximalizuje homogenitu dat v rámci skupin a rozdílnost mezi nimi.

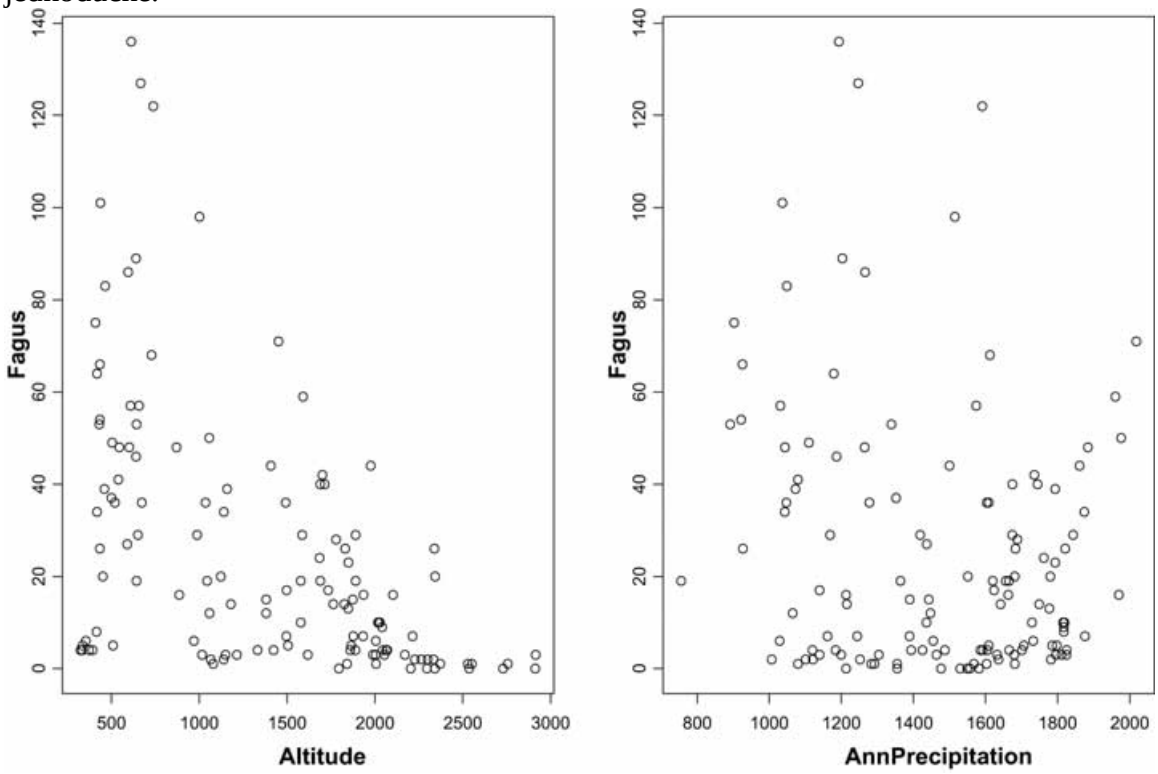
Vysvětlovaná proměnná může být kvantitativní (v případě **regresních stromů**) nebo kvalitativní (pro **stromy klasifikační**). Výsledky fitování se popisují "stromem" znázorňujícím postupná dělení. Každá větev je popsána specifickým dělícím pravidlem: toto pravidlo má pro kvantitativní prediktory formu nerovnosti (např. "*ProměnnáX2 < 3.7*") a pro kvalitativní proměnné (faktory) formu výčtu možných hodnot (např. "*ProměnnáX4 má hodnoty a, c nebo d*"). Takto vytvořené dvě podskupiny se dělí dále, dokud nejsou příliš malé, nebo v hodnotách vysvětlované proměnné dostatečně homogenní. Výsledné skupiny (listy) se potom identifikují předpovězenou hodnotou vysvětlované proměnné (pokud je tato proměnnou kvantitativní) nebo předpovědí příslušnosti listu k určité skupině (je-li vysvětlovaná proměnná faktoriální).

Při fitování modelu založeného na stromech na naše data vytvoříme obvykle nejprve strom příliš složitý. V další fázi se pak snažíme najít optimální velikost stromu pro predikci hodnot vysvětlované proměnné. Určení "optimální" velikosti děláme procedurou **cross-validation**, a to tak, že vytvoříme sérii významně redukovaných ("prořezaných") stromů jen z určité části dat a zbylá data použijeme na hodnocení "výkonnosti" vytvořeného stromu: tato pozorování "proženeme" hierarchickou soustavou rozdělovacích pravidel a porovnáme predikovanou hodnotu s hodnotou pozorovanou. Pro každou velikost ("komplexitu") stromu to uděláme několikrát (pro jednotlivé podskupiny dat). Obvykle soubor dat rozdělujeme na deset přibližně stejně velkých skupin a každou z nich použijeme k hodnocení výkonu modelu o dané komplexitě, který jsme fitovali ze zbývajících devíti částí dat. Graf závislosti "kvality" stromu na jeho komplexitě (velikosti modelu) má obvykle minimum, a právě to odpovídá optimální velikosti. Použijeme-li strom větší, dostaneme model "přefitovaný", který nám sice výborně aproximuje sesbírané vzorky, pro celou populaci však poskytne jen nejednoznačný popis.

11.7. Modelování křivek druhové odpovědi: srovnání modelů

V této části nás čeká příklad hledání vhodného popis odpovědi vybraného druhu (buku - *Fagus sylvatica*) na mezoklimatické faktory, jmenovitě na nadmořskou výšku a roční

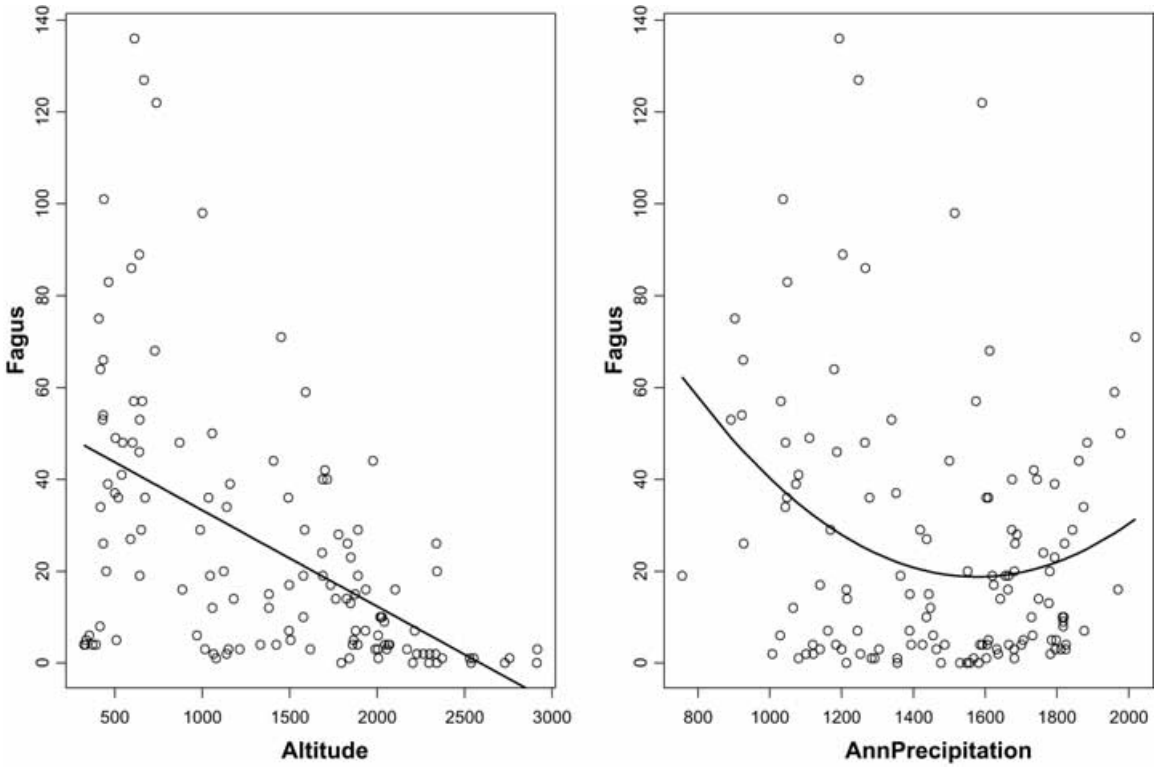
množství srážek. Obrázek 11-1 ukazuje původní data ve formě dvou XY diagramů. Všimněte si, že vysvětlovaná proměnná (frekvence buku) je ve formě jednotlivých počtů, tedy striktně nezáporná. Regresní modely uvedené v této sekci byly vytvořeny programem S-Plus for Windows, verze 4.5. Příkazy, které jsme při fitování použili, zde nejsou uvedeny, neboť naučit se pracovat s tímto pokročilým statistickým programem není nikterak jednoduché.



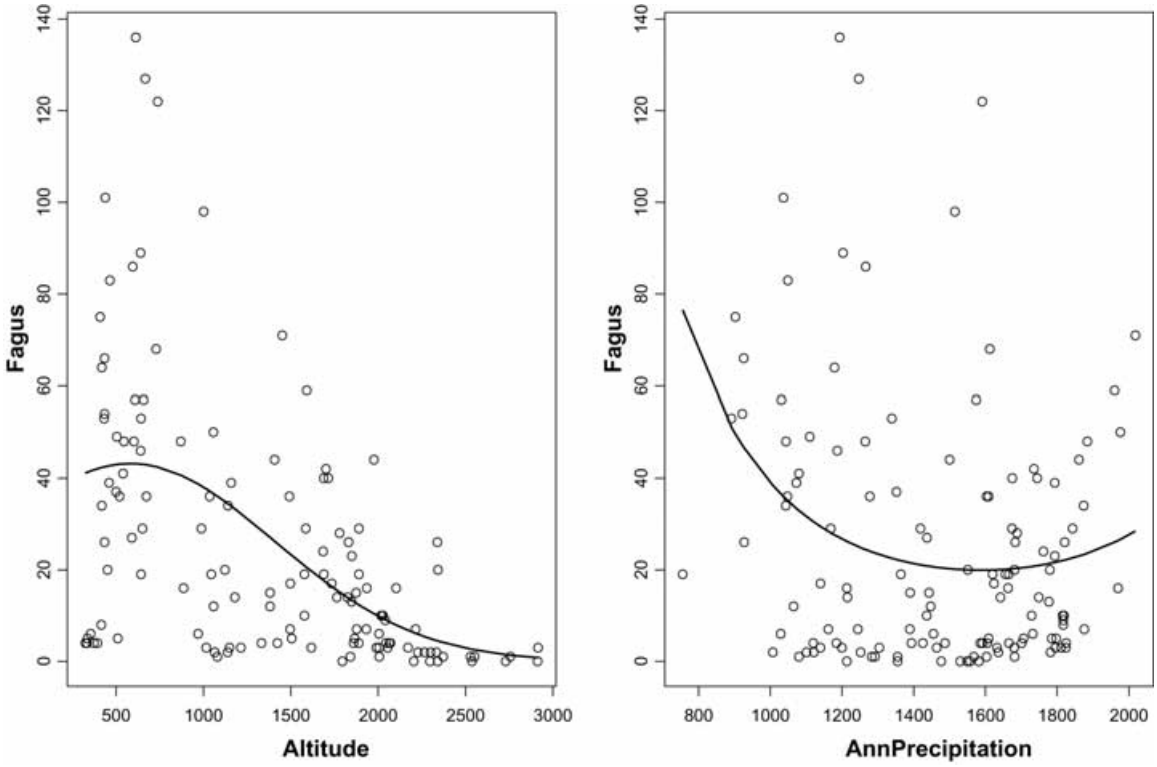
Obrázek 11-1 Závislost frekvence druhu *Fagus sylvatica* na nadmořské výšce a ročních srážkách

Zdá se, že frekvence buku s rostoucí nadmořskou výškou klesá, ale pro nejníže položená místa jsou příslušné frekvenční hodnoty také nízké. Hodnocení závislosti frekvence buku na ročních srážkách se zdá téměř nemožné.

První model, který použijeme, je tradiční lineární model fitovaný pomocí algoritmu založeném na kriteriu "nejmenších čtverců". Komplexitu modelu jsme vybrali postupným výběrem, kde testované varianty modelu byly: nulový model (nezávislost frekvence na uvažované proměnné), lineární model, polynom druhého stupně (kvadratická závislost) a polynom třetího stupně. Vybrané modely jsou na obrázku 11-2. Pro fitování závislosti frekvence buku na nadmořské výšce byl nejlepší lineární model. Můžeme si však na něm všimnout omezení, jaké lineární modely mají: frekvence predikovaná pro nadmořskou výšku 3000 m n. m. je záporná, což není příliš realistické. Dále, pro vyšší očekávanou frekvenci buku (v nižších nadmořských výškých) je rozptyl skutečných hodnot kolem modelem předpovídané hodnoty mnohem větší než pro očekávané nízké frekvence. To se příčí předpokladům testů, které na klasickém regresním modelu provádíme (jde o tzv. předpoklad konstance variability - *homoskedasticity*). Modelem pro vliv ročních srážek na frekvenci buku je polynom druhého stupně, který indikuje lokální minimum pro srážky okolo 1500 mm za rok.



Obrázek 11-2 Dva lineární modely, které odděleně popisují závislost frekvence buku na nadmořské výšce a úhrnu ročních srážek

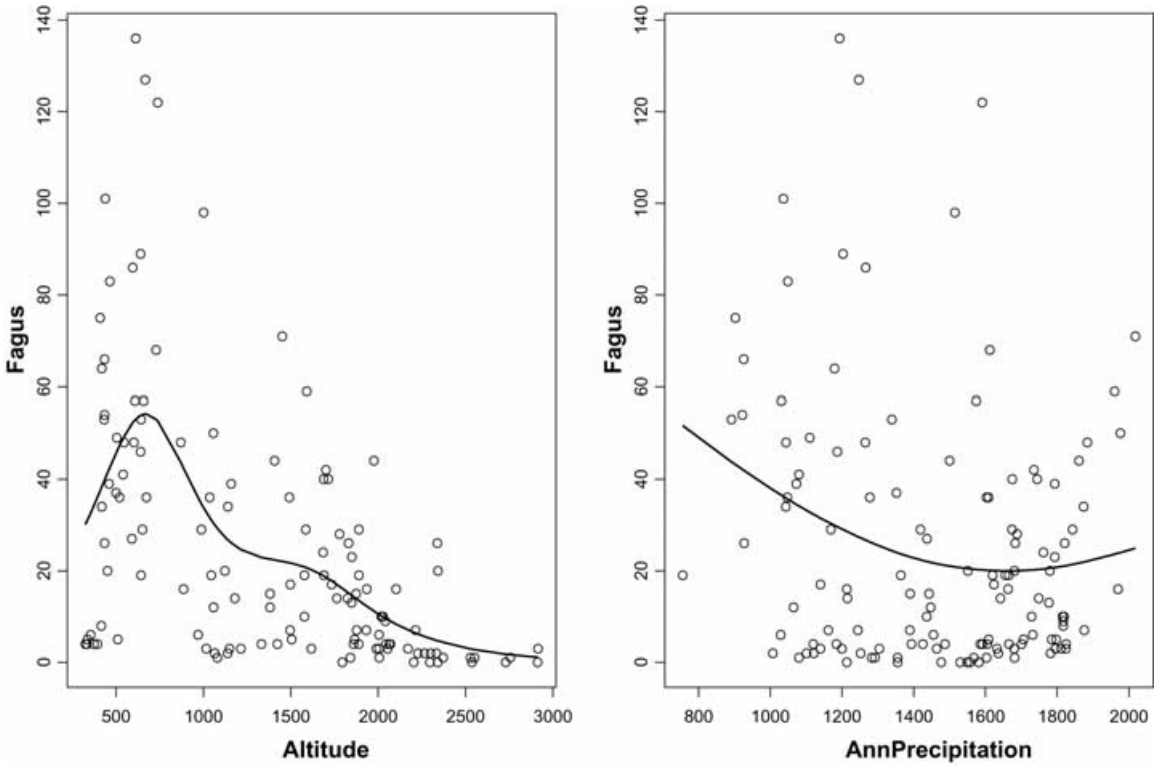


Obrázek 11-3 Tvar dvou zobecněných lineárních modelů popisujících odděleně závislost frekvence buku na nadmořské výšce a úhrnu ročních srážek

Obrázek 11-3 shrnuje podobné regresní modely, jaké jsou k vidění na obrázku 11-2, tentokrát však fitované zobecněnými lineárními modely. Pro distribuci vysvětlované proměnné byla navržena Poissnova distribuce a byla použita logaritmická link funkce. Opět

jsme použili přímý výběr, tentokrát založený na testech analýzy deviance (jsou to aproximativní testy založené na předpokládané chi-square distribuci reziduálních deviancí porovnávaných modelů). Nyní byl pro oba faktory prostředí vybrán model polynomu druhého stupně. Ten zřetelně ukazuje příkré snížení frekvence výskytu v nízkých nadmořských výškách. Naproti tomu pro vysoké nadmořské výšky klesají očekávané frekvence pomalu k nule.

Poslední rodinou regresních modelů, kde jsme pro dva uvažované prediktory fitovali dva různé modely, byly zobecněné aditivní modely. Jako v GLM, tak i zde jsme předpokládali Poissonovu distribuci a vybrali jsme logaritmickou link funkci. Opět jsme k rozhodnutí o komplexitě použili postupný výběr, ale tentokrát byl výběr založen na statistice vyjadřující přímo *model parsimony*.* Během výběru byly postupnými kandidáty nulový model, zobecněný lineární model s lineární podobou prediktoru, zobecněný lineární model s prediktorem ve tvaru polynomu druhého stupně a zobecněný aditivní model s hladkou funkcí s jedním, dvěma a třemi stupni volnosti. Užitý typ hladké funkce byl *cubic spline*.

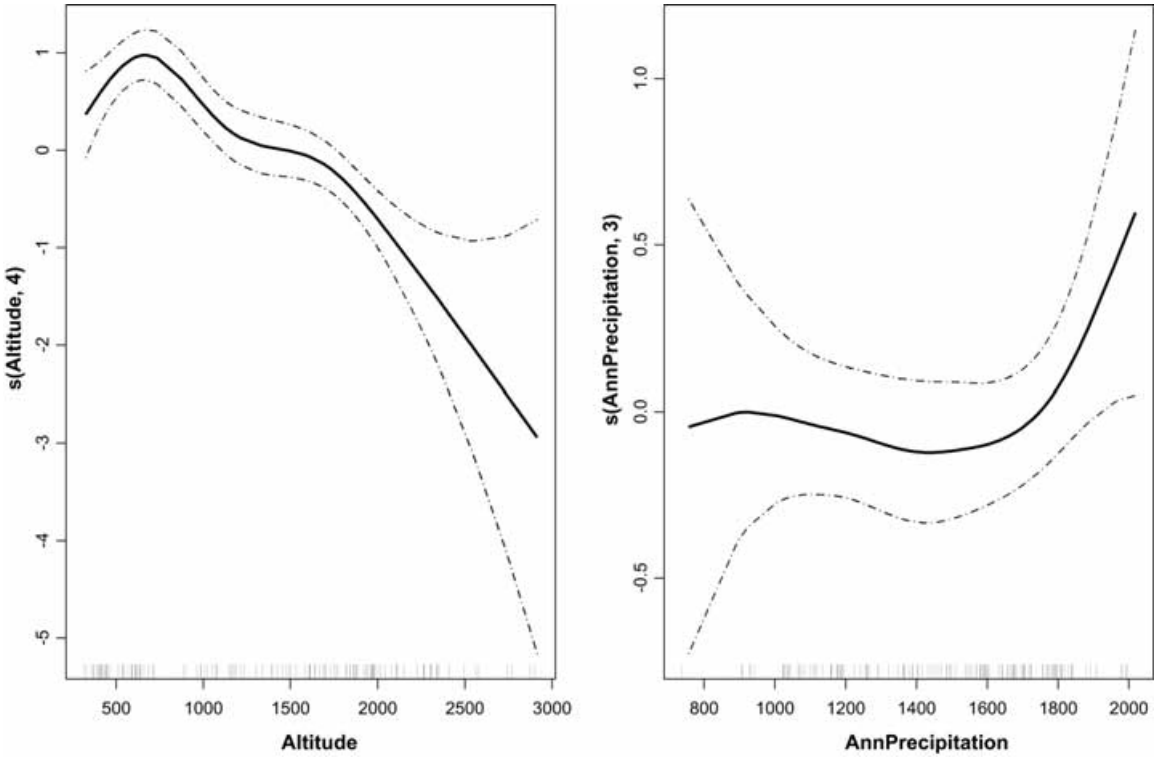


Obrázek 11-4 Dva zobecněné aditivní modely, které odděleně fitují závislost frekvence buku na nadmořské výšce a ročních srážkách

Postupným výběrem jsme pro nadmořskou výšku jako prediktor došli k aditivnímu modelu s hladkou transformační funkcí se třemi stupni volnosti, zatímco pro srážky byl zvolen mnohem jednodušší model (hladký člen s jedním stupněm volnosti). Tak pro závislost frekvence buku na ročních srážkách zůstává tvar křivky podobný, ale pro nadmořskou výšku odhaluje nový model křivku jinou, pravděpodobně realističtější. Je to asymetrická unimodální křivka, kde frekvence buku okamžitě roste od nadmořské výšky přibližně 400 m do 700 m (optimum?), pak stejně rychle klesá až k 1000 m n. m. Nicméně po dosažení této výšky již frekvence klesá mnohem pomaleji a ubývá postupně až do 2500 metrů.

* Této statistice se říká *Akaike information criterion (AIC*) a je popsána například v *Hastie & Tibshirani (1990).*

Nyní se můžeme podívat, co se se zobecněným aditivním modelem stane, použijeme-li obě vysvětlující proměnné najednou. Měli bychom si uvědomit, že i když GAM modeluje vliv obou prediktorů nezávisle, aditivně, bude charakter jejich příspěvku jiný, než když použijeme každý zvlášť. Také volbu komplexity modelu musíme udělat znovu. A ta tentokrát překvapivě dochází k mnohem komplexnějším modelům. [+] Hladká funkce pro nadmořskou výšku používá čtyři stupně volnosti, funkce pro srážky tři stupně volnosti.

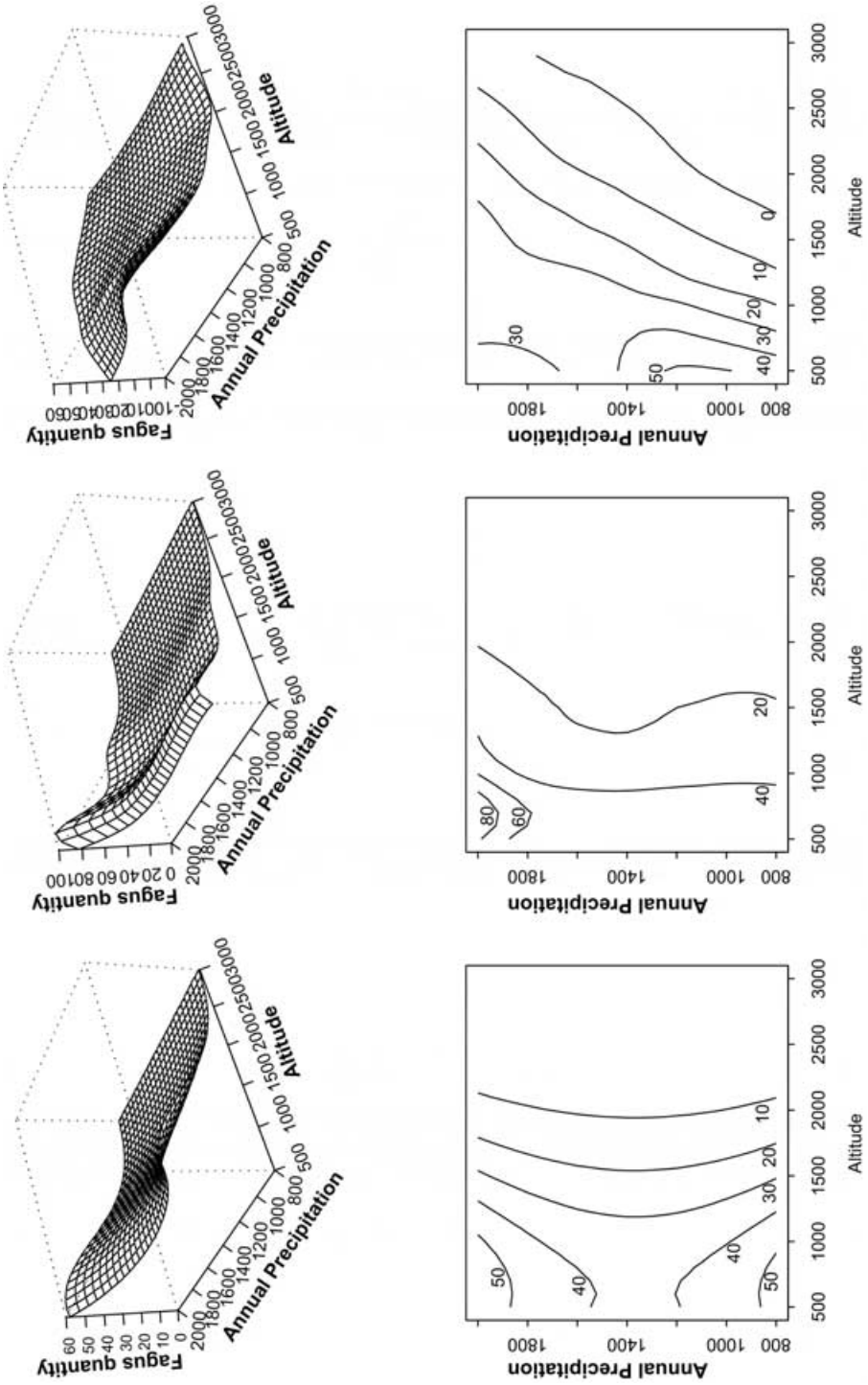


Obrázek 11-5 Zobecněný aditivní model závislosti frekvence buku jak na nadmořské výšce, tak na ročním úhrnu srážek

Na obrázku 11-5 je použit nejběžnější typ reprezentace zobecněných aditivních modelů: pro jednotlivé prediktory se vynáší hladká funkce $\mathbf{s_j}$, jak je popsána v sekci 11.5. Svislá osa tedy ukazuje příspěvek konkrétní transformační funkce k hodnotě aditivního prediktoru ve zobecněném aditivním modelu. Jsou znázorněny i odhadnuté konfidenční intervaly pro fitované transformační křivky. Transformační funkce pro nadmořskou výšku vypadá podobně jako v samostatném GAM pro nadmořskou výšku, který jsme již viděli, ale funkce pro srážky indikuje ztrátu klesající části v oblasti nízkých hodnot.

Na tomto místě bychom si mohli porovnat, jak vypadá výsledek fitování jednotlivých typů regresních modelů. Zobecněný lineární model, zobecněný aditivní model a *loess smoother* jsou na obrázku 11-6, kde jsou ukázány jako 3-D povrchy i jako vrstevnicové diagramy.

[+] Může to být díky nově "objevenému" patternu vztahujícímu se k možné interakci těchto dvou prediktorů, nebo to může být způsobeno nedostatky AIC statistiky použité k výběru parsimonního modelu.



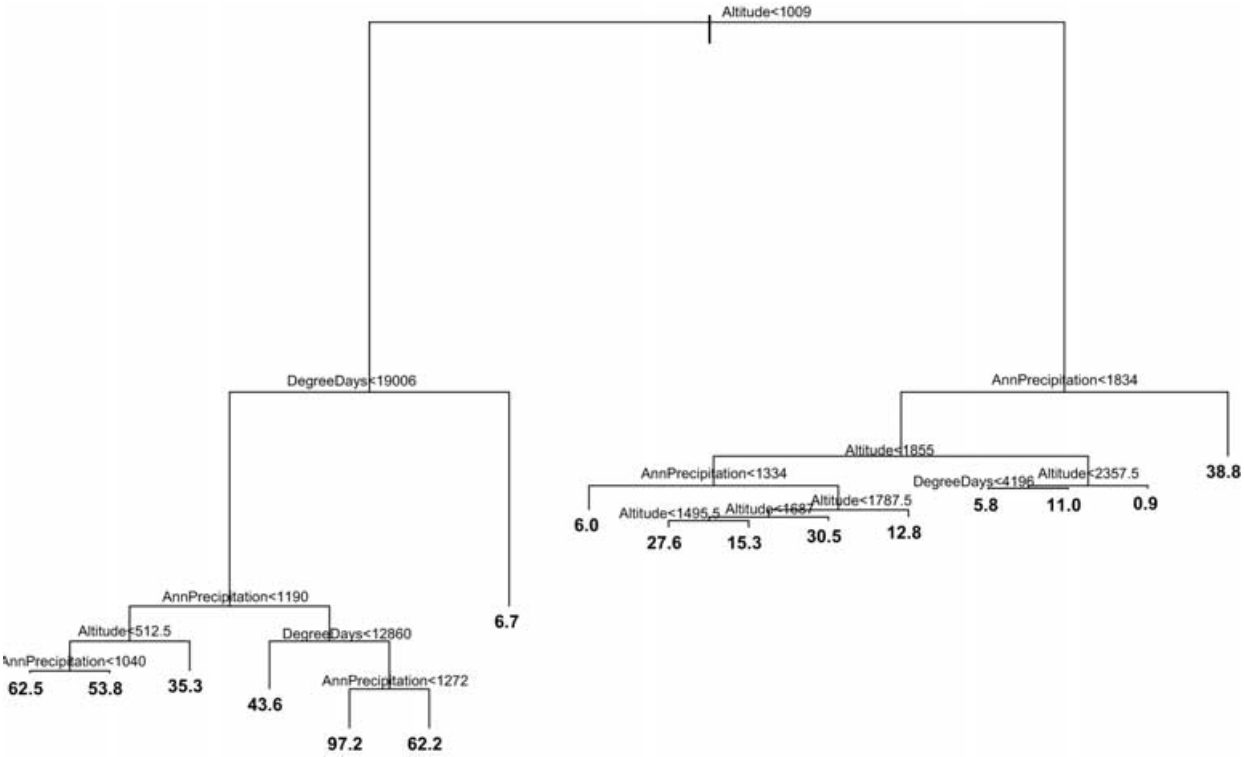
Obrázek 11-6 Srovnání třech odezvových povrchů modelujících frekvenci buku pomocí nadmořské výšky a ročních srážek jako prediktorů. Zleva doprava (při pohledu z boku) to jsou GLM, GAM a *loess smoother*

Loess model neudržuje aditivitu vlivu obou prediktorů, takže je v modelování jejich interakce mnohem flexibilnější. Za to však musíme zaplatit tím, že nemůžeme hodnotit oba vlivy nezávisle: tvar křivky odpovědi podél, řekněme, gradientu nadmořské výšky závisí na hodnotách srážek. Navíc pro takovýto flexibilnější model obvykle potřebujeme mnohem více dat (neboť "sežere" více stupňů volnosti).

Nakonec si ukážeme, co může při modelování odpovědi frekvence buku na mezoklimatické faktory nabídnout metoda regresních stromů. V tomto případě přidáme ke

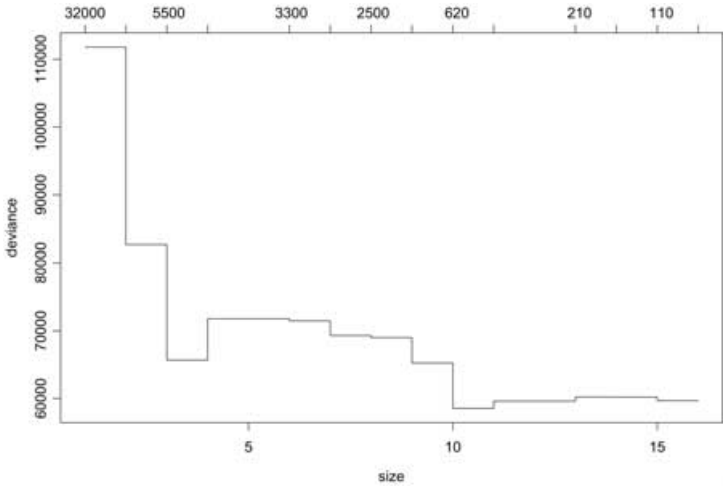
dvěma dosud používaným prediktorům ještě třetí - proměnná *DegreeDays* obsahuje záznamy o teplotě konkrétních míst.

Jak si vzpomínáme ze sekce 11.6, necháme algoritmus nejdříve vytvořit co největší regresní strom. Výsledek je na obrázku 11-7. Výška větví je úměrná poklesu deviance modelu, nebo (jinak řečeno) rozdílu mezi hodnotami odpovědi obou skupin, které se konkrétním rozdělením vytvoří. Vidíme zcela jasně, že rozdíl v nadmořské výšce (s prahovou hodnotou okolo 1000 m n. m.) je nejdůležitější. Buk preferuje nižší nadmořské výšky (větve na levé straně), kde je jeho frekvence obecně vyšší, krom extrémně teplých míst (hodnota DegreeDays vyšší než asi 19000).



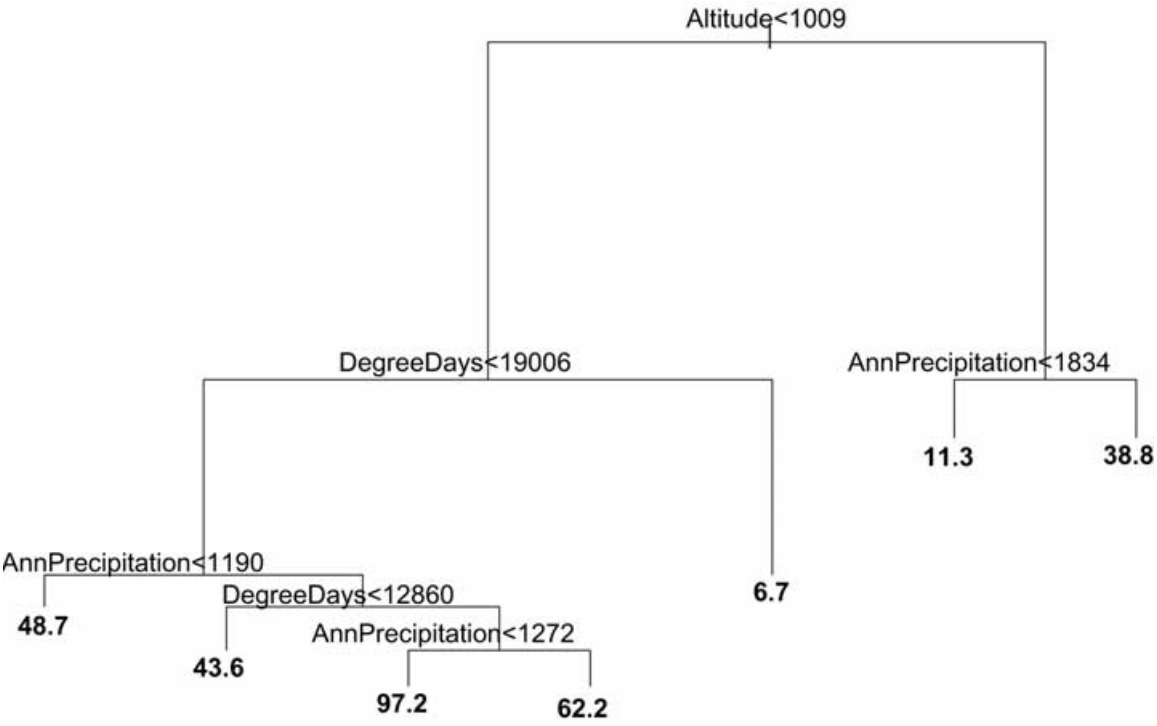
Obrázek 11-7 Nově vytvořený regresní strom, který je příliš složitý

Než s interpretací pokročíme dál, měli bychom se podívat, jak komplikovaný strom je pro predikci nových hodnot potřeba. To můžeme udělat procedurou *cross-validation*, jejíž výsledky jsou graficky znázorněny na obrázku 11-8.



Obrázek 11-8 Výsledky procedury *cross-validation*

Reziduální deviance modelu regresního stromu je vynesena proti komplexitě stromu vyjádřené dvěma různými charakteristikami na dolní a horní vodorovné ose. Minimum predikčních chyb je zřejmě u stromu s 10 terminálními větvemi (s 10 "listy"), nicméně pro snadnější orientaci jsme nakonec vytvořili ještě jednodušší strom, který je na obrázku 11-9.



Obrázek 11-9 Výsledný regresní strom

Všimněte si, že výsledky fitování regresního stromu nejsou tak kvantitativní, jako u jiných uvedených modelů, ale za to mají jiné výhody. Umožňují například jednoduše znázornit interakce mezi prediktory, stejně tak jako vlivy prediktorů s různým prostorovým rozlišením, atd.

12. Literatura

- Borcard D., Legendre P. & Drapeau P. (1992): Partialling out the spatial component of ecological variation. - Ecology, 73: 1045 - 1055
- Cleveland W.S. & Devlin S.J. (1988): Locally-weighted regression: an approach to regression analysis by local fitting. - J. Am. Statist. Assoc., 83: 597 - 610
- Grime J.P., Hodgson J.G. & Hunt R. (1988): Comparative plant ecology. - Unwin Hyman, London.
- Hastie T.J. & Tibshirani R.J. (1990): Generalized Additive Models. - Chapman and Hall, London. 335 pp.
- Hill M.O. & Gauch H.G. (1980): Detrended correspondence analysis, an improved ordination technique. - Vegetatio, 42: 47 - 58
- Knox R.G. (1989): Effects of detrending and rescaling on correspondence analysis: solution stability and accuracy. - Vegetatio, 83: 129 - 136
- Kovář P. & Lepš J. (1986): Ruderal communities of the railway station Ceska Trebova (Eastern Bohemia, Czechoslovakia) - remarks on the application of classical and numerical methods of classification. - Preslia, 58: 141 - 163
- Lepš J. (1999): Nutrient status, disturbance and competition: an experimental test of relationship in a wet meadow. - Journal of Vegetation Science, in press.
- McCullagh P. & Nelder J.A. (1989): Generalised Linear Models. Second Edition. - Chapman and Hall, London. 511 pp.
- Pyšek P. & Lepš J. (1991): Response of a weed community to nitrogen fertilization: a multivariate analysis. - Journal of Vegetation Science, 2: 237 - 244
- Sokal R.R. & Rohlf F.J. (1995): Biometry. - 3rd edition, W.H. Freeman, New York, USA. 887 pp.
- Šmilauer P. (1992): CanoDraw User's Guide v. 3.0 - Microcomputer Power, Ithaca, USA. 118 pp
- Špačková, I., Kotorová, I. & Lepš, J. (1998): Sensitivity of seedling recruitment to moss, litter and dominant removal in an oligotrophic wet meadow. - Folia Geobot., 33: 17 - 30
- Ter Braak C.J.F. (1994): Canonical community ordination. Part I: Basic theory and linear methods. - Ecoscience, 1: 127 - 140
- Ter Braak C.J.F. & Looman C.W.N. (1986): Weighted averaging, logistic regression and the Gaussian response model. - Vegetatio, 65: 3 - 11
- Ter Braak C.J.F. & Prentice I.C. (1988): A theory of gradient analysis. - Advances in Ecological Research, 18: 93 - 138
- Ter Braak C.J.F. & Šmilauer P. (1998): CANOCO Reference Manual and User's Guide to Canoco for Windows. Microcomputer Power, Ithaca, USA. 352 pp
- Van der Maarel E. (1979): Transformation of cover-abundance values in phytosociology and its effect on community similarity. - Vegetatio, 38: 97 - 114
- Von Ende C.N. (1993): Repeated measures analysis: growth and other time-dependent measures. In: Scheiner S.M. & Gurevitch J. [eds]: Design and analysis of ecological experiments, pp. 113 - 117. Chapman and Hall, New York, NY, USA.