



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



Multikriteriální analýzy CANOCO

Studijní opory k předmětu - rozšířené osnovy přednášek

Hana Švehláková



1. PRÁCE S DATY. TYPY POUŽÍVANÝCH DAT. SBĚR DAT. PRIMÁRNÍ DATA. PŘEPIS A KONTROLA DAT. EDA. CDA. TRANSFORMACE DAT

Proměnné – charakteristiky prvků základního souboru, mohou nabývat více hodnot

Data – záznam skutečnosti, aktuální hodnoty proměnných

Klademe si otázky:

Co chceme zkoumat? Je to smysluplné? Jak kvalitní máme data?

Potřebné znalosti z „ekologické terminologie“

Ekologická data:

druh

populace

společenstvo

faktory prostředí

a „statistické terminologie“

Typy proměnných a dat :

Závislé proměnné/data – odpověďové, kritériální, vysvětlované, často druhové; např. počty jedinců druhů cévnatých rostlin, hmotnost biomasy, pokryvnost vegetačních typů v různých územích

Nezávislé proměnné/data – explanační, prediktory; využíváme k předpovězení hodnot závislých proměnných (kvalitativní, kvantitativní)

Rušivá proměnná (matoucí) – moderující, má vztah k závislé proměnné a její působení zkresluje námi sledovaný vztah závislá – nezávislá proměnná; můžeme nebo nemusíme o nich vědět

Kovarianční proměnná – kovariáta, rušivá proměnná, kterou známe a jsme schopni ošetřit (zahrneme do plánu výzkumu – metoda vytváření bloků, matchování apod., randomizace, statistická eliminace)

Kvalitativní (kategoriální) proměnné/data

Binární – data typu ano/ne, přítomnost/absence, 0/1; nelze seřadit

Nominální – data popisná typu půdní druh, půdní typ, biologický taxon; nelze seřadit

Ordinální – data obsahující více kategorií typu Braun -Blanquetova stupnice pokryvnosti, lze seřadit

Kvantitativní data

Spojité – mohou nabývat jakýchkoli hodnot v určitém rozmezí; lze řadit

Diskrétní – počty, např. počet druhů na ploše; lze řadit

Způsob sběru dat

- a) Cenzus – do zkoumání se zahrnují všechny jednotky populace; velmi náročné
- b) Výběrové (statistické) šetření - sběr informací standardizovaným postupem, náhodný (pravděpodobnostní) výběr je optimální

Chyby v datech:

Data – záznam skutečnosti není dokonalý, obsahují chyby a zkreslení

- a) Chyby výběrové náhodné - vznikají náhodně při výběru vzorku (lze ji posoudit např. tzv. intervalem spolehlivosti, statistickým testováním hypotézy)
- b) Chyby výběrové systematické – nemají náhodný charakter, jde o chyby v designu výzkumu
- c) Chyby nevýběrové - vznikají rovněž při statistické analýze dat (např. chyby přístrojů, aritmetické chyby atd.)

Chybějící data

- a) Lze si je opatřit znovu – jdeme do terénu a provedeme nový fytoecologický snímek, odebereme vzorek a opětovně analyzujeme apod.
- b) Nelze si je opatřit – necháme dané místo prázdné (pozor na 0 – nese i informaci o nulové hodnotě/nepřítomnosti měřeného parametru); vzorky s chybějící hodnotou vypustíme; doplníme hodnoty průměrem, dopočítáme hodnoty pomocí mnohonásobného regresního modelu

Kontrola dat – analýza odlehlých bodů, odstranění šumu v datech, EDA

EDA (Exploratory data analysis)

Skupina statistických technik kladoucí důraz na grafické znázornění dat, „detektivní práce“ v souboru dat

Základní prvky:

- Vizualizace dat
- Analýza reziduálních hodnot
- Distribuce a transformace dat a jejich změna
- Robustní a rezistentní procedury

Grafické znázornění – krabicové grafy – box plots (jednorozměrná data); bodové grafy (scatter plots) pro dvou a vícerozměrná data

Lze využít programů Excel, Statistica, OriginLab, Canoco

Transformace a standardizace dat

Změna relativní vzdálenosti mezi jednotlivými hodnotami – změna tvaru jejich distribuce

Funkční transformace lineární

- a) Přičítání/ odečítání konstanty: ke všem datům přičteme kladnou nebo zápornou konstantu (časté odečtení aritmetického průměru od všech získaných skóre dané proměnné = centrovaná data)
- b) Násobení/dělení konstantou – při přechodu mezi jednotkami měření (metry, centimetry, gramy, kilogramy atd.)
- c) kombinace odčítání a násobení

Funkční transformace nelineární

Logaritmická transformace, odmocninová transformace, trans. obrácenou hodnotou – linearizace nelineárních dat, úprava tvaru rozdělení dat, aby se více podobalo Gaussovu rozdělení

2. EKOLOGICKÁ DATA A JEJICH VYUŽITÍ. EKOLOGICKÁ PODOBNOST. INDEXY BIODIVERZITY. ELLENBERGOVY INDIKAČNÍ HODNOTY. FUNKČNÍ VLASTNOSTI DRUHŮ

Indexy podobnosti

- vyjádření podobnosti mezi vzorky
- interval (0,1); 0 – vzorky nesdílejí žádný společný druh, 1 – vzorky jsou identické

Problém „dvojitě nuly“ – počet druhů chybějících v obou vzorcích zároveň (v porovnání s druhy, které jsou zároveň přítomny v obou vzorcích):

- asymetrické – dvojitě nepřítomnosti se ignorují, vychází z rozložení organismů v optimálních podmínkách gradientu, na stanovištích, které nemají optimální hodnoty podmínek, může být daný druh vzácný nebo se zde nemusí vůbec vyskytovat. Pokud se druh vyskytuje na dvou stanovištích, poukazuje to na určitou podobnost mezi těmito dvěma lokalitami (mají podobné podmínky – optimum \pm rozpětí). Avšak když se druh nevyskytuje ani na jednom stanovišti, mohou mít stanoviště úplně rozdílné podmínky (jedna nebo druhá strana optima) Dvojitá přítomnost je lepším ukazatelem podobnosti lokalit než dvojitá nepřítomnost.
- symetrické – dvojitě nepřítomnosti se hodnotí stejně jako dvojitě přítomnosti, v ekologické praxi nepoužívané

Binární indexy podobnosti

Prezence nebo absence druhů (1-1), (0-0)

Jaccardův koeficient podobnosti (eliminace vlivu dvojitě nuly):

$$S(x_1, x_2) = \frac{a}{a+b+c}$$

x_1 a x_2 jsou srovnávaná společenstva, a je počet současných výskytů a b , c je počet neshodujících se taxonů ve srovnávaných společenstvech

Sørensenův koeficient jako varianta předchozího koeficientu dává dvojnásobnou váhu dvojitým prezencím

$$S(x_1, x_2) = \frac{2a}{2a+b+c} \quad (1.8)$$

protože se může zdát, že přítomnost druhů je více informativní než jejich absence, která může být způsobena různými faktory a nemusí nutně odrážet rozdílnost prostředí

Simpsonův koeficient je vhodný pro vzorky s velmi odlišným počtem druhů

$$S_i = a / [a + \min(b, c)]$$

Kvantitativní indexy podobnosti

Euklidovská vzdálenost

Lze využít v případě nepřítomnosti dvojité nuly, je vhodná standardizace proměnných

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2} \quad (1.17)$$

kde x_1 a x_2 jsou srovnávaná společenstva, P je počet taxonů ve společenstvu a y_{1j} a y_{2j} jejich abundance ve společenstvech. V případě, že $D = 0$ jsou vzorky identické

Bray – Curtis index

porovnává dvě společenstva z hlediska minimální abundance u každého druhu.

$$C_{BC}(x_1, x_2) = \frac{\sum_{j=1}^p |y_{1j} - y_{2j}|}{\sum_{j=1}^p (y_{1j} + y_{2j})} = 1 - \frac{2W}{A + B},$$

kde W je součet minimálních abundancí druhu, A a B je součet abundancí jednotlivých společenstev

Sørensenův modifikovaný index

$$C_N = \frac{2jN}{(aN + bN)}$$

aN a bN jsou celkové počty jedinců ve společenstvu na lokalitách A a B a jN je suma abundancí druhů, pokud se druh vyskytuje v obou společenstvech.

Indexy diverzity

Druhová bohatost – počet druhů ve vzorku

Vyrovnanost – relativní zastoupení druhů ve vzorku

Alfa diverzita – druhová bohatost vzorku

Beta diverzita – rozdíly mezi vzorky (změna druhového složení)

Gama diverzita – celková bohatost širšího území (regionu, kraje...)

Shannonův - Wienerův index

Index vychází z informační teorie. Jeho předpokladem je náhodný výběr jedinců z teoreticky neomezeného množství a přítomnost všech druhů společenstva ve vzorku. Jeho exponenciální hodnota vyjadřuje, kolik stejně početných druhů by vytvořilo Shannonův - Wienerův index o stejné hodnotě. Základní vztah pro výpočet Shannonova - Wienerova indexu je:

$$H = - \sum_{i=1}^S p_i \ln p_i \quad p_i = \frac{n_i}{N}$$

kde S je celkový počet taxonů, n_i je počet jedinců i -tého druhu a N celkový počet jedinců. Podstatná je také volba základu použitého logaritmu, která ovlivňuje číselný výsledek výpočtu; není tak možné srovnávat hodnoty indexů počítané s různou bází logaritmu

Maximální hodnota Shannonova-Wienerova indexu pro dané společenstvo odpovídá logaritmu počtu druhů a ukazuje, jaké hodnoty by index nabyl při shodné početnosti všech druhů společenstva.

$$H_{max} = -\ln S$$

Shannon – Wienerova vyrovnanost (ekvitabilita)

$$E = \frac{H}{H_{max}} = \frac{H}{\ln S}$$

Simpsonův index

Patří do skupiny indexů založené na dominanci. Je silně závislý na nejpočetnějším druhu a méně citlivý ke vzácným druhům. Může nabývat hodnot od nuly do jedné..

S jeho zvyšující se hodnotou stoupá dominance a klesá vyrovnanost společenstva, proto se často používá jeho převrácená hodnota nebo odpočet od jedné; v případě interpretace publikovaných výsledků je vždy nezbytné ověřit, v jaké formě byl tento index použit. Vztah mezi tímto indexem a počtem druhů, pro vzorky s více než 10 druhy, je silně závislý na rozložení abundancí druhů (na modelech rozdělení abundance taxonů, tedy species abundance models) ve vzorku. Výpočet základní verze nabývající nejvyšších hodnot při vysoké dominanci a nejnižších při vyrovnaném společenstvu je dán vztahem:

$$D = \sum_{i=1}^S \frac{n_i(n_i-1)}{N(N-1)}$$

kde S je počet taxonů, n_i počet jedinců i -tého taxonu a N celkový počet jedinců.

Simpsonova vyrovnanost (ekvitabilita)

$$E = (1/D)/S$$

Funkční diverzita

Zohledňuje diverzitu funkčních typů vyskytujících se ve vzorku. Problém – jak definovat funkční typy (skupiny):

RAO index

$$FD = \sum_i \sum_j dij p_i p_j$$

Proměnná dij je mírou (funkční) vzdálenosti tj. nepodobnosti druhů i a j . Problematika výpočtu tohoto indexu tedy spočívá v tom, jak měřit právě tuto vzdálenost. P je relativní abundance druhů i a j .

Elenbergovy indikační hodnoty

Optima druhů rostlin na gradientu živin, vlhosti, půdní reakce, kontinentality, teploty, světla a salinity, původně určené pro Německo. Vychází z empirických zkušeností – jde ovšem o průměrné hodnoty, které se nemusí shodovat se skutečnými.

Nabývají hodnot 1 - 9 (12).

Pokud nemáme data o podmínkách prostředí naměřena, lze použít EIH, ALE nelze je použít jako vysvětlující proměnné v přímých gradientových analýzách, v případě použití s naměřenými hodnotami není vhodné znázorňovat v ordinačních diagramech – EIH vykazují vyšší korelační koeficient korelaci (např. v případě DCA bude mít delší šipky)

4. REGRESE. LINEÁRNÍ MODEL. REGRESNÍ KŘIVKY

Korelace

míra stupně asociace dvou proměnných – proměnné jsou korelovány, pokud určité hodnoty jedné proměnné mají tendenci vyskytovat se společně s určitými hodnotami druhé proměnné. Není uvažován kauzální vztah.

Pearsonův korelační koeficient

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}},$$

kde \bar{x} a \bar{y} jsou výběrové průměry,

v jednodušší verzi

$$r = \frac{s_{xy}}{s_x s_y},$$

kde s_x je směrodatná odchylka proměnné X, s_y směrodatná odchylka proměnné Y a s_{xy} takzvaná kovariance proměnných X a Y.

Správná interpretace Pearsonova korelačního koeficientu předpokládá, že obě proměnné jsou náhodné veličiny a mají společné dvourozměrné normální rozdělení. Potom nulový korelační koeficient znamená, že veličiny jsou nezávislé. Pokud není splněn předpoklad dvourozměrné normality, z nulové hodnoty korelačního koeficientu nelze usuzovat na nic víc, než že veličiny jsou nekorelované.

Regrese

Předpokládá vztah mezi vysvětlující (x) proměnnou a vysvětlovanou (y) proměnnou

Co nás zajímá?

- Testování statistických odhadů neznámých parametrů regresní funkce
- Testování hypotéz o těchto parametrech
- Ověřování předpokladů regresního modelu

Lineární regresní model

V případě lineárního vztahu mezi závislou proměnnou na nezávislých proměnných :

Lineární regrese

- Lze použít je-li závislost veličiny y na x lineární,
- v praxi: proložení bodů v grafu regresní **přímkou** $y = a + bx$ tak, aby součet druhých mocnin odchylek jednotlivých bodů od přímky byl minimální (metoda nejmenších čtverců),
- a, b – regresní koeficienty,

- a – posun na ose y (místo kde regresní přímka protíná svislou osu),
- b – sklon regresní přímky.

pozn. čtverec = druhá mocnina

Kvadratická regrese

- Je speciálním případem regrese lineární, kdy soubor dat proložíme kvadratickou funkcí (**parabola**)
 $y = ax^2 + bx + c$,
- a, b, c jsou regresní koeficienty, které můžeme v praxi odhadnout opět metodou nejmenších čtverců.

Logaritmická regrese

- Je speciálním případem regrese lineární, kdy soubor dat proložíme **logaritmickou** funkcí
 $y = a + b \cdot \ln(x)$.

3. ZÁKLADNÍ TERMINOLOGIE MNOHOROZMĚRNÝCH STATISTICKÝCH METOD

Základní pojmy z „čisté“ a „ekologické“ statistiky

Viz rovněž přednáška 1.

Primární data

- popisují společenstvo a jeho prostředí
- data z terénu – mnohorozměrný tvar
- náhodnost ekologických procesů
- zkrácení dat (např. přehlédnutí druhu, ztráta vzorku....)

Problém s dlouhodobým uchováváním dat!

Primární data obsahují

- a) vysvětlované proměnné (např. druhová data)
- b) vysvětlující proměnné - prediktory (např. charakteristiky prostředí)
- c) kovariáty

Výběr analýzy pro jednu vysvětlovanou proměnnou

- a) prediktory nemáme k dispozici – analýza distribuce
- b) prediktory máme k dispozici – regresní model (včetně analýzy rozptylu ANOVA, analýzy kovariance ANOCOV)

Výběr analýzy pro více vysvětlovaných proměnných

- a) prediktory nemáme k dispozici – nepřímá gradientová analýza (analýza hlavních komponent PCA, detrendovaná korespondenční analýza DCA, nemetrické mnohorozměrné škálování NMDS); klastrová analýza
- b) prediktory máme k dispozici – přímá gradientová analýza, klastrová analýza, diskriminační analýza

Výhody mnohorozměrné analýzy

- Přesnější popis vícerozměrné podstaty ekologických dat
- Možnost pracovat s velkými soubory dat s mnoha proměnnými, analyzuje i jejich redundanci
- Poskytuje pravidla pro takovou kombinaci proměnných tak, aby co nejlépe popsaly daný ekologický problém
- Možnost uplatnit na stejných datech různé metody – řešení různých ekologických otázek

Nevýhody mnohorozměrné analýzy

- Náročnost na vstupní data
- Matematická náročnost
- Nemožnost použití pro nezávislé proměnné
- Někdy náročná interpretace

- Možnost přeparametrizování modelu (problém s velkým počtem nul, problém s „dummy“ proměnnými, redundancemi, s interpretací variability při velkém množství proměnných)

5. ORDINAČNÍ ANALÝZA. MODEL Y ODPOVĚDÍ DRUHŮ NA GRADIENT PROSTŘEDÍ. ZÁKLADNÍ ORDINAČNÍ TECHNIKY A METODY

Gradientová analýza – jakákoli metoda umožňující dát do vztahu druhovou skladbu a gradienty prostředí – ty mohou být námi měřené, nebo hypotetické, zkušenostní (Ellenbergovy indikační hodnoty). Ordinační analýzy tvoří podmnožinu gradientové analýzy, podobně jako další gradientové analýzy (klasifikace), pracují s mnohorozměrným prostorem.

Mnohorozměrný prostor je určen:

- a) Druhy
- b) Vzorky (environmentální proměnné)
- c) Ekologickými gradienty

Proč ordinace?

Ordinační metody jsou skupina matematických metod, které mají za úkol zjednodušit interpretaci velkého souboru dat tak, že vzorky (objekty) zobrazí podél gradientu, který je definován kombinací proměnných. Především se snaží odhalit nejdůležitější faktory, které vypočítá z originálních proměnných. Těchto několik málo hlavních gradientů (většinou 2 – 3) nám má vysvětlit co nejvíce variability v datech, aniž by se ztratila informace skrytá v datech. Hlavním principem ordinačních metod v ekologii je, že variabilita vícerozměrných dat je koncentrovaná do několika málo dimenzí, a tyto hlavní gradienty jsou spojeny s environmentálními (nezávislými) proměnnými. Ordinační techniky používají tento nadbytek k nalezení a vysvětlení hlavního nezávislého gradientu ve vícerozměrných datech.

Charakteristiky ordinačních metod

- Organizuje vzorky (druhy, lokality, pozorování) podél ekologického gradientu
- Hodnotí vztahy v souboru proměnných, kde není definováno, která proměnná je závislá a která nezávislá
- Nalézají dominantní, základní gradient variability mezi vzorky (objekty) na základě pozorování. Zdůrazňují rozdíl mezi vzorky, raději při tom klade důraz na rozdíl mezi vzorky než na jejich podobnost
- Redukují dimenzionalitu ve vícerozměrných datech „kondenzací“ velkého počtu originálních proměnných do menšího počtu nových proměnných (např. Analýza hlavních komponent) s minimální ztrátou informace
- Definují nové proměnné jako váženou lineární kombinaci originálních proměnných
- Počítá redundanci v datech podle blízkosti jednotlivých objektů v ordinačním prostoru a vytvoří nám jednodušší prostředky k pochopení a vizualizaci dat.

Nepřímé ordinační metody (unconstrained ordination)

- Metody nejsou ovlivněny žádnou proměnnou, která se nenachází v souboru dat.
- Vychází z matice vzorky x druhy
- Vytvoří nový gradient, který nejlépe popisuje druhová data dle lineárního nebo unimodálního modelu.
- Slouží k tvorbě hypotéz, nelze jimi hypotézy testovat

Přímé ordinační metody (constrained ordination)

- Gradient je kombinací konkrétních environmentálních proměnných
- Vychází z matic vzorky x druhy a vzorky x proměnné prostředí
- Ordinační osy představují směr největší variability v druhových datech
- Složí k testování hypotéz o vlivu environmentálních proměnných na druhová data, nesloží k popisu dat

Modely odpovědi druhů na gradient prostředí

- a) Lineární – nejjednodušší odpověď, funguje na krátkém gradientu (užším než je toleranční rozpětí druhu)
- b) Unimodální – předpoklad, že druhy mají na gradientu prostředí své optimum (gradient je širší než je toleranční rozpětí organismu)

Jak odhadnout optima druhů tj. typ odpovědi?

Lineární odpověď – metoda nejmenších čtverců

Unimodální odpověď – výpočet váženého průměru těch hodnot environmentálních charakteristik, při kterých se druh vyskytuje

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

Kde *Env* je hodnota charakteristiky, *abund* četnost daného druhu v odpovídajícím vzorku.

6. NEPŘÍMÁ ORDINACE. PCA (ANALÝZA HLAVNÍCH KOMPONENT). CA (KORESPONDENČNÍ ANALÝZA). DCA (DETRENDOVANÁ KORESPONDENČNÍ ANALÝZA)

Analýza hlavních komponent (PCA)

Základní, poměrně jednoduchá metoda. Účelem je zestručnit informaci obsaženou ve velkém množství proměnných do menšího počtu dimenzí (gradientů) s co nejmenší ztrátou informací. N – dimenzí (každá dimenze odpovídá jedné proměnné) redukuje na několik málo dimenzí, přičemž každá nová dimenze je lineární kombinací původních n – proměnných.

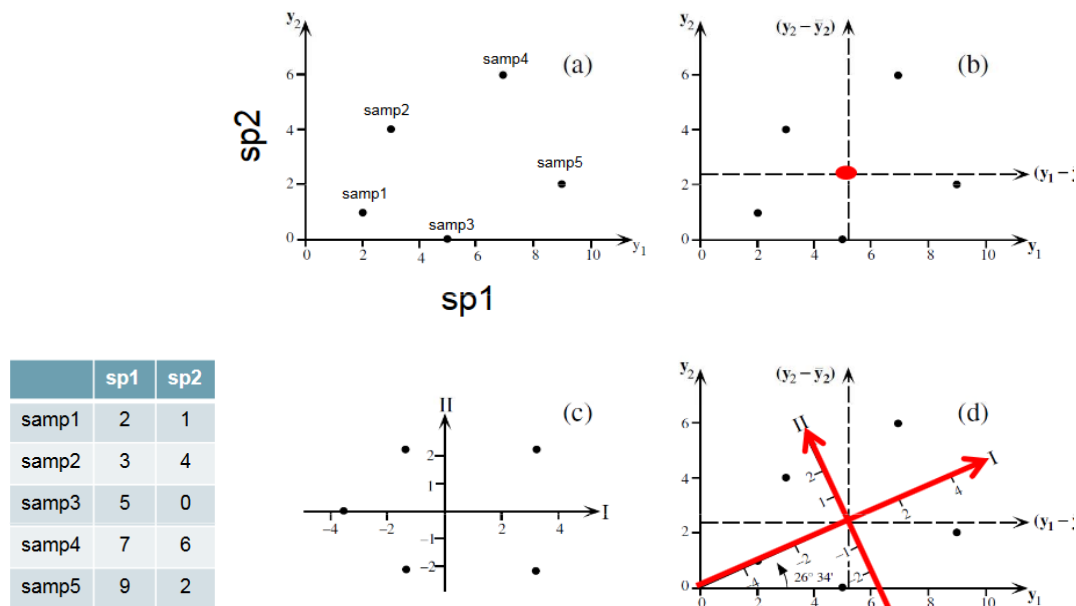
Tyto lineární kombinace se nazývají **hlavní komponenty**.

Nové gradienty (proměnné) vytváří PCA tak, že maximalizuje rozdíly mezi vzorky podél os. Gradient vzniklý kombinací původních dat odpovídá maximálnímu rozptylu těchto dat.

Vícerozměrná data jsou zobrazována jako „mrak“ v mnohazměrném prostoru, kde každá osa je definována jednou proměnnou. PCA lze chápat jako projekci skrz mrak vzorků orientovanou dle největší variability podél každé z os. První hlavní komponenta se znázorní přes střed mraku v jeho nejdelší části (vysvětluje tudíž nejvíce variability). Druhá hlavní komponenta musí být kolmá na první a vysvětluje co nejvíce zbylé variability.

Pro PCA je důležité, aby velká část variability byla koncentrována do několika málo dimenzí. Proměnné musejí být na sobě závislé tak, že se dochází k jejich změně v závislosti na jiných proměnných. Pokud jsou nezávislé, pak je „datový mrak“ rozprostřen rovnoměrně podél všech originálních os. Pak žádná projekce není lepší než druhá a analyzovaný problém nejlépe vysvětlují originální data.

Původně navržena pro kvantitativní znaky, lze využít i pro binární a částečně kvalitativní.



Obr.1 Zpracování dat v ekologii společenstev (Zelený)

Korespondenční analýza (CA) a detrendovaná korespondenční analýza (DCA)

CA je v podstatě analýzou kontingenčních tabulek (lze provádět i např. v Excelu)
Všechny proměnné v CA mají stejnou dimenzi a obsahují výhradně kladné hodnoty nebo nuly.

V ekologii se tato metoda nejvíce používá na analýzu druhových dat (přítomnost/nepřítomnost nebo abundance). Je založena na unimodálním modelu a dá se použít na dlouhém gradientu)

Grafické znázornění vztahů, které obdržíme z CA, je založené na myšlence jak prezentovat všechny sloupce a řádky a interpretovat relativní pozice bodů, jako váhy patřící k danému sloupci a řádku. Systém takto získaných indexů nám bude poskytovat souřadnice každého sloupce a řádku, které následně vykreslíme do grafu. Z něj pak můžeme odvodit, které sloupcové kategorie jsou důležité vůči řádkovým kategoriím a naopak.

Arch effect a DCA

Vyplývá z nelineárního efektu gradientů na vzdálenosti vztahu mezi jednotlivými objekty (druhovými daty). Tato nelinearita je ordinačními metodami „překládána“ do Euklidovského dvourozměrného prostoru – vzniká tzv. arch effect.

Jak ho odstranit?

Zbavením se trendu CA pomocí DCA a to detrendováním dle segmentů (nejčastější), nebo polynomů (např. v případě použití kovariát)

V rámci DCA dojde k nelineárnímu přeškálování první osy – odstranění „nahloučení“ vzorků na okrajích osy. Výsledný diagram má osy v jednotkách směrodatné odchylky (SD). Platí, že druhové složení se změní na gradientu o délce 1 – 1,4 SD o polovinu.

Výhody DCA

- Ekologicky dobře interpretovatelné výsledky
- Osy jsou v jednotkách směrodatné odchylky – lze zjistit, jak dlouhý gradient je pokryt daty

Nevýhody DCA

- Neelegantní (rozdělení osy na segmenty)
- Výsledek bývá ovlivněn zvolením počtu segmentů
- Pokud jsou v datech dva a více gradientů, DCA se s nimi nevypořádá (druhou a vyšší osy poškodí)

Výběr metody dle DCA

Lineární metody jsou vhodné pro homogenní data, unimodální pro heterogenní data.

Vhodnost metody lze zjistit z délky 1. osy DCA:

- a) Menší než 3 SD – vhodná lineární metoda
- b) Větší než 4 SD – vhodná unimodální metoda
- c) V rozmezí 3 – 4 – lze zvolit obě (doporučení provést obě)

Pokud nevím, nejsem si jistý, použiji raději unimodální metodu.

7. PŘÍMÁ ORDINACE. RDA (REDUNDANČNÍ ANALÝZA). CCA (KANONICKÁ KORESPONDENČNÍ ANALÝZA)

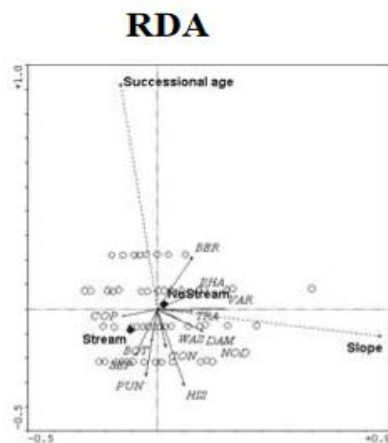
Z nepřímých ordinací nezjistíme, co model nalezený v datech znamená a čím je způsoben. Osy v PCA i CA odpovídají nejdůležitějším gradientům, ale jejich interpretace bývá někdy velmi složitá. Proto je vhodné v určitých případech zvolit přímou ordinaci, kdy je jedna část proměnných umístěna na osu, jež je kombinací druhé části proměnných, obzvlášť pokud jsou první proměnné druhová data (např. početnost) a druhé environmentální proměnné. Definujeme tedy závislé proměnné (druhová data) a nezávislé proměnné (environmentální data), která tvoří 2 skupiny datové matice – matice Y = druhová data, matice X = environmentální data.

Redundanční analýza (RDA)

Jde v podstatě o rozšířenou PCA – přímé rozšíření vícenásobné regrese modelováním odpovědi vícerozměrných dat.

RDA je vhodná, když vysvětlovaná data Y jsou analyzovaná analýzou hlavních komponent, to znamená, když proměnné y jsou lineárně spojené s jinými a Euklidovská vzdálenost je považována za vhodnou k popisu vztahu mezi objekty ve faktorovém prostoru.

Graficky je RDA vyjádřena 3 skupinami bodů a) poloha skóre, b) vysvětlovaná proměnná z Y , c) vysvětlující proměnná z X .



Obr.2 ordinační diagram RDA (Lepš, Šmilauer: Mnohorozměrná analýza ekologických dat, 2000)

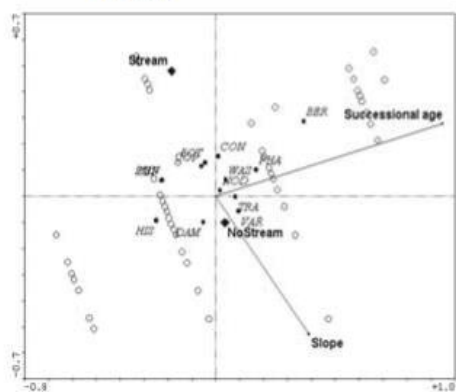
Kanonická korespondenční analýza (CCA)

V podstatě hybrid mezi ordinací a mnohorozměrnou regresí.

Největší výhodou je, že osy jsou lineární kombinací nezávislých proměnných tak, že vysvětlují nejvíce variability v závislých proměnných. Tímto lze snadněji pochopit a interpretovat význam osy. Vysvětlující proměnné jsou graficky znázorněny jako šipky, délka šipky odpovídá významu proměnné. Kosinus úhlu mezi šipkou a osou odpovídá korelačnímu koeficientu mezi proměnnou a osou. Úhly mezi šipkami odpovídají korelacím proměnných.

Druhy jsou označeny jako body, jejich vztah k „šipkám“ ukazuje na jejich rozdělení na každé environmentální proměnné.

CCA



Obr.3 ordinační diagram CCA (Lepš, Šmilauer: Mnohorozměrná analýza ekologických dat, 2000)

8. MODEL NULOVÉ HYPOTÉZY. MONTE CARLO PERMUTAČNÍ TEST. TESTOVACÍ STATISTIKY.

Nulová hypotéza

V podstatě hypotéza, že druhová data nejsou závislá na vysvětlujících proměnných.

Základem jsou permutační testy.

Nulovou hypotézu zamítneme v případě, že dostaneme uspořádání dat, které je velmi nepravděpodobné za předpokladu nulové hypotézy.

Monte Carlo permutační test

Metodami Monte Carlo se nazývají numerické metody řešení matematických úloh pomocí modelování náhodných veličin a statistického odhadu jejich charakteristik.

Jde o randomizační typ testu. Určuje statistickou významnost použitých vysvětlujících proměnných. Použití u přímých ordinací

Randomizace (znáhodnění)

Lze např. randomizovat plochy s vegetačními zápisy – každé ploše se náhodně přiřadí jeden zápis, tím se zruší závislost mezi environmentální proměnnou (např. vlhkostí) a druhovými daty (např. pokryvnost). Z hodnot druhových dat (např. pokryvnosti) se pak vypočte množství variability vysvětlené proměnnými prostředí (tzv. F statistika) pro původní i randomizovaná data.

Testovací statistiky (dle Lepš, Šmilauer: Mnohorozměrná analýza dat. 2000)

F statistika

Základem je fakt, že variabilita vysvětlovaná proměnné (druhová data) je popsána vysvětlujícími proměnnými (data podmínek prostředí) a rozložena mezi více kanonických os. Nejvýznamnější je první osa, ale lze použít i další osy. V CANOCO existují 2 testovací statistiky

a) Test využívající první kanonickou osu – vliv jen jedné environmentální proměnné

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

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 .

b) Test využívající všechny kanonické osy – vliv všech proměnných

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

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

9. PRAKTICKÉ PŘÍKLADY POUŽITÍ ORDINACÍ. PŘÍPADOVÉ STUDIE

Bude zpracováváno v průběhu výuky v rámci projektu. Studenti budou pracovat a vyhodnocovat vlastní data (v rámci BP nebo DP, případně poskytnuta vyučujícím)

10. KLASIFIKAČNÍ ANALÝZA. NEHIERARCHICKÁ KLASIFIKACE.

Cílem těchto metod je získat skupinu dat (vzorky nebo druhy), které jsou vnitřně homogenní a odlišné od jiných skupin. Pokud analyzujeme vzorky, obsahuje daná skupina vzorky s podobnými druhy (druhovým složením); pokud analyzujeme druhy, obsahuje daná skupina druhy s podobným ekologickým chováním (podobnými nároky na podmínky prostředí)

Nehierarchická klasifikace – K - means clustering

Všechny shluky jsou si rovny

Shlukování metodou k průměrů; minimalizuje sumy čtverců vzdáleností mezi vzorky uvnitř shluků, přičemž počet shluků je určen na začátku

Jde o iterativní metodu, začne od náhodného přiřazení vzorků do shluků, postupně přehazuje vzorky mezi shluky a hledá optimální řešení

Výsledek do určité míry záleží na počátečním rozmístění shluků do vzorků a je proto dobré proces mnohokrát zopakovat (najít stabilní řešení).

Lze použít program STATISTICA.

11. KLASIFIKAČNÍ ANALÝZA. HIERARCHICKÁ KLASIFIKACE (CLUSTER ANALYSIS). DIVIZIVNÍ 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 (Lepš, Šmilauer, 2000)

- a) Divisivní
 - monotetická
 - polytetická (např. TWINSpan)
- b) Aglomerativní (shluková analýza)

Shluková analýza (cluster analysis)

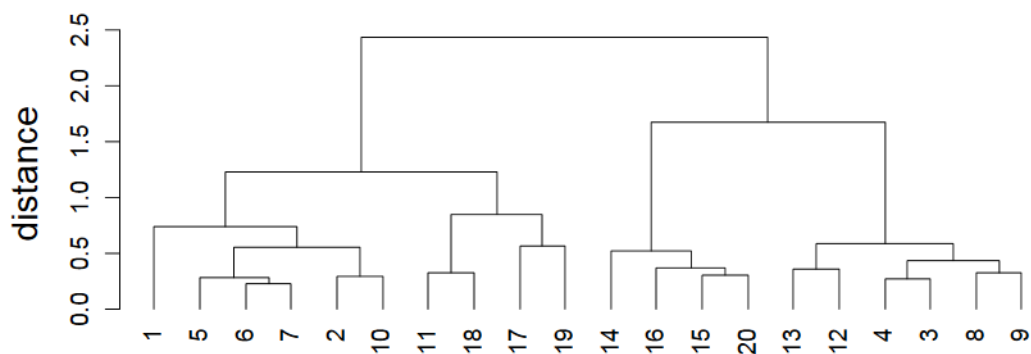
Shluky jsou tvořeny zespod, tj. postupným shlukováním jednotlivých vzorků do skupin a shlukování pokračuje, dokud se všechny shluky nespojí do jedné velké skupiny. Vzorky se pojí do shluku, kde je jim nejpodobnější vzorek nebo se připojí, až když shluk obsahuje všechny podobné vzorky.

Grafickým výstupem je tzv. dendrogram.

Provádí se ve 2 krocích

- a) Spočítání matice podobnosti pro všechny páry dat – matice je symetrická a uhlopříčku tvoří nuly (při nepodobnosti) nebo čísla vyjadřující max. podobnost
- b) Shlukování objektů do skupin, přepočítání podobnosti všech objektů k těmto skupinám

Nejde o metodu objektivní, je ovlivněna našim rozhodováním na různé úrovni zpracování dat.



Obr. 4 Dendrogram (zdroj Zelený, Zpracování ekologických dat)

Nezáleží na tom, jestli je vzorek vpravo nebo vlevo, ale na tom, s kterými vzorky a na které úrovni je spojen.

Divizivní klasifikace

Soubor dat je dělen shora. Pokud je klasifikace založena jen na jednom atributu (např. druhu), jde o klasifikaci monotetickou. Pokud je založena na více atributech, jde o klasifikaci polytetickou – v současnosti používanější.

TWINSPAN

(Two Way INDicator SPecies ANalysis)

Jde o polytetickou klasifikaci vycházející ze zpracování fytocenologických dat. Vzorky jsou uspořádány podle první osy korespondenční analýzy (CA, DCA) a podle ní se rozdělují do shluků (vzorky s pozitivním a negativním skóre).

Výhodou je schopnost ošetřit vzorky blízko středu osy – bývají často špatně klasifikovány. Metoda pracuje s kvalitativními daty, aby se neztratila informace o jejich kvantitě, zavedlo se pojetí pseudodruhů a mezních hodnot pseudodruhů. Každý druh může být nahrazen několika pseudodruhy. Pseudodruh je přítomen, pokud zastoupení druhu přesáhne mezní hodnotu. Metoda je vhodná v případě, že jsou data rozložená podle jednoho výrazného gradientu.

Lze využít JUICE, PC-ORD

Výsledkem je tabulka podobná fytocenologické.

```

graph TD
    A[ ] --- B[ ]
    A --- C[ ]
    style A fill:none,stroke:#000
    style B fill:none,stroke:#000
    style C fill:none,stroke:#000
  
```

[illegible]

Obr.5 TWINSPAN (zdroj Zelený, Zpracování ekologických dat)

12. PRAKTICKÉ PŘÍPADY POUŽITÍ KLASIFIKACÍ. PŘÍPADOVÁ STUDIE.

13. VIZUALIZACE MNOHOROZMĚRNÝCH DAT. INTERPRETACE DIAGRAMŮ.

Jde o praktickou činnost ve cvičeních. Studenti pracují s vlastními (nebo vyučujícím poskytnutými daty). Výsledkem je vhodná grafická vizualizace výsledků a jejich správná interpretace, která je následně diskutována v plénu dané studijní skupiny.

14. DESIGN EXPERIMENTŮ - MANIPULAČNÍ VS. PŘÍRODNÍ EXPERIMENTY

Pracovní metody experimentů

- Popis
- srovnání
- Analýza
- Syntéza
- Experiment
- Tvorba modelů
- Interpretace výsledků

Manipulační experimenty

- Uměle manipulujeme vysvětlující proměnnou X a sledujeme reakci vysvětlované proměnné Y
- Umožňuje testování hypotéz
- Zjistíme kauzalitu

Press experimenty

Sledovaná vysvětlovaná proměnná je pod stálým tlakem (např. kosení)

Pulse experimenty

Zásah je proveden jen jednou – na začátku experimentu (např. sledování resilience)

Důležité rozmístění ploch

Randomizace ploch

- a) Úplné znáhodnění
- b) Znáhodnění do bloků dle gradientu prostředí

Přírodní experimenty

- Důležitá manipulace „přírodou“, ta ovlivňuje vysvětlující proměnnou
- Nejistíme kauzální vztah, pouze korelaci
- Slouží spíše k vytváření hypotéz

Snapshot experimenty

- Opakují se v prostoru, ne v čase
- Sběr vzorků na mnoha lokalitách v průběhu krátké doby
- Většina přírodních experimentů

Trajectory experimenty

- Opakování v čase (někdy i prostoru)
- Sběr vzorků na pevně daných lokalitách, delší doba mezi sběry
- Trvalé monitorovací plochy (u nás často v lesích)