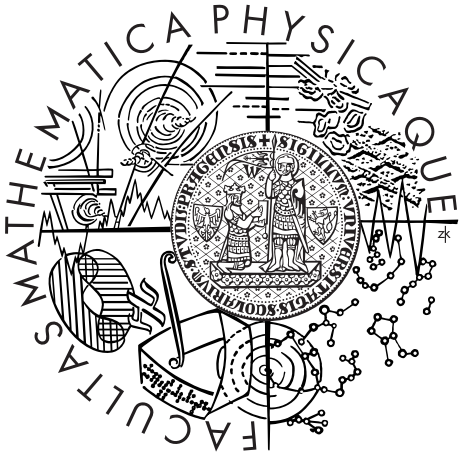
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Bc. Radek Solnický

Metody statistické inference založené na matici vzdáleností

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	Ing. Omelka Marek, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie
Studijní plán:	Matematická statistika

Praha 2011

Na tomto místě bych chtěl poděkovat vedoucímu mé práce Ing. Marku Omelkovi, Ph.D. za cenné rady a připomínky a dále také všem ostatním, kteří přispěli svou radou nebo podporou ke zdárnému dokončení práce.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 19. července 2011 Podpis autora

Název práce: Metody statistické inference založené na matici vzdáleností

Autor: Bc. Radek Solnický

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Při analýze dat pocházejících z oblasti ekologie často nelze použít tradičních mnohorozměrných metod. Použití koeficientů nepodobnosti a matice vzdáleností představuje způsob, jak tento problém vyřešit. V této práci představujeme některé z těchto koeficientů a následně testy založené na matici vzdáleností: Mantelův test, varianty testů ANOSIM a MRPP a test homogenity disperzí. Zkoumáme vztahy mezi těmito testy a předvádíme jejich použití na reálných datech. Na simulacích pak upozorňujeme na problematiku interpretace těchto testů.

Klíčová slova: nepodobnost, testování založené na nepodobnostech, mnohorozměrná analýza, Mantelův test, test ANOSIM, test MRPP

Title: Distance-based testing

Author: Bc. Radek Solnický

Department: Department of Probability and Mathematical Statistics

Supervisor: Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: When analyzing ecological data, one considers traditional multivariate techniques to be unsuitable. The use of dissimilarity coefficients and distance matrices is a way, how to solve this problem. In this work we present some of these coefficients and distance-based tests: Mantel test, several versions of ANOSIM and MRPP tests and distance-based test for homogeneity of multivariate dispersions. We focus on relationships among these tests and illustrate the use with an example. We also discuss the difficulties of interpretation of the results of these tests.

Keywords: dissimilarity, distance-based analysis, multivariate analysis, Mantel test, ANOSIM test, MRPP test

Obsah

1. Úvod

Matice vzdáleností a statistické metody na ní založené se často využívají při analýze sociologických, biologických a ekologických dat. V této práci se zaměříme na data pocházející z oblasti životního prostředí. Lidé se čím dál více zajímají nejen o to, v jakém prostředí žijí, ale také o to, jak je svými činnostmi ovlivňují, a to jak v pozitivním, tak v negativním smyslu. S tím také souvisí otázka, jak případné vlivy a změny rozpoznat, prokázat a kvantifikovat. Data, která obvykle bývají k dispozici, popisují druhovou skladbu v daných stanovištích a charakteristiky těchto stanovišť jako jsou geografická poloha, typ krajiny, vlastnosti půdy, množství slunečního záření, vlhkost a podobně. Otázky, kterými se ekologové zabývají, mohou být například: Má továrna u řeky vliv na živočichy žijící níže po proudu? Ovlivňuje ropný vrt v moři své okolí? Je druhová skladba v dané oblasti nějak provázána s vlastnostmi krajiny a pokud ano, jak? V posledních 25 letech se k řešení těchto úkolů začaly používat metody, které využívají matici vzdáleností. Cílem této práce je představit některé z nich.

Přirozenou otázkou je, proč zavádět matici vzdáleností a metody pro ni vhodné, když máme k dispozici přesná data a klasické statistické metody. Odpovědí je hned několik:

- Ekologická data často obsahují i kategoriální data nebo pořadí, pro které klasické metody nemusí být použitelné. Koeficienty nepodobnosti, pomocí nichž je matice vzdáleností zkonstruována nejsou omezeny typem proměnných.
- Může se stát, že k dispozici bude méně pozorování (stanovišť) než vysvětlujících proměnných. To nám znemožní použití některých klasických metod.
- Spíše než konkrétní četnosti zastoupených druhů nás zajímá celková druhová skladeb na stanovištích. Pomocí koeficientů nepodobnosti pracujeme s jednorozměrnými veličinami namísto mnohorozměrných.
- Klasické metody pracují implicitně s eukleidovskou metrikou. Ta nemusí být vždy vhodná.
- Tyto metody dále mívají požadavky, které ekologická data zpravidla nesplňují, např. normalitu dat.

Matice vzdáleností nám tedy může zjednodušit práci a umožní upravit postup tak, aby více vyhovoval našim představám a interpretacím. Na druhou stranu zde tak vzniká prostor, kdy zvolené postupy (např. volba koeficientu nepodobnosti) značně ovlivňují výsledky analýzy.

V kapitole 2 představíme některé často užívané koeficienty nepodobností a některé jejich vlastnosti, jako například eukleidovskost. Dále zavedeme pojem matice vzdáleností a představíme metodu hlavních koordinát jako způsob reprezentace dat, na jejichž základě byla matice vzdáleností vytvořena.

Kapitola 3 se zabývá testováním nezávislosti a jeho modifikací pro matici vzdáleností - Mantelovým testem. Zároveň je tento test ilustrován na datech `mouchy`, které jsou podrobněji představeny v příloze.

V kapitole 4 představíme metody zabývající se problémem jednoduchého třídění ANOSIM a MRPP. Ukážeme jejich varianty, vzájemné vztahy a vztah k Mantelovu testu. Dále uvedeme test homogenity disperzí, který s těmito metodami souvisí. Vše opět ilustrujeme na datech `mouchy`.

Kapitola 5 je věnována simulacím týkajícím se metod uvedených v kapitole 4. Simulacemi poukazujeme na případná rizika použití těchto metod.

V příloze jsou pak popsána data `mouchy` a obsah CD přiloženého k této práci.

2. Koeficienty nepodobnosti, matice vzdáleností, PCoA

2.1 Koeficienty nepodobností a matice vzdáleností

Ke konstrukci matice vzdáleností se používají koeficienty nepodobnosti (*dissimilarity measures*). Ty mají za úkol měřit, jak jsou si daná pozorování v požadovaném smyslu „vzdálená“, jinými slovy měří jejich odlišnost. Jsou konstruovány tak, aby tuto míru odlišnosti měřily nějakým vhodným způsobem

Definice 2.1 *Funkce* $\Delta : R^p \times R^p \to R_0^+$*, splňující pro všechna* $\boldsymbol{x}, \boldsymbol{y} \in R^p$

- $\Delta(\boldsymbol{x}, \boldsymbol{x}) = 0$,
- $\Delta(\boldsymbol{x}, \boldsymbol{y}) \geq 0$,
- $\Delta(\boldsymbol{x}, \boldsymbol{y}) = \Delta(\boldsymbol{y}, \boldsymbol{x})$,

se nazývá koeficient nepodobnosti.

V případě, že koeficient nepodobnosti splňuje také trojúhelníkovou nerovnost

$$\Delta(\mathbf{x}, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{z}) \geq \Delta(\mathbf{x}, \mathbf{z}), \qquad \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathrm{R}^p,$$

a

$$\Delta(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y},$$

je zároveň metrikou.

Definice 2.2 *Nechť* $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in R^p, \boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ *a* Δ *je koeficient nepodobnosti. Maticí vzdáleností rozumíme čtvercovou matici* $\boldsymbol{D}$*, o rozměrech* $n \times n$ *a prvcích* $[\boldsymbol{D}]_{ij}$*,* $i, j = 1, ..., n$*, pro kterou* $[\boldsymbol{D}]_{ij} = \Delta(\boldsymbol{x}_i, \boldsymbol{x}_j)$*.*

Z vlastností koeficientů nepodobnosti pak vyplývá, ža matice vzdáleností je symetrická a na hlavní diagonále má pouze nuly. Obecně tedy matice vzdáleností vypadají následovně:

$$\mathbf{D} = \begin{pmatrix} 0 & \Delta(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \Delta(\mathbf{x}_1, \mathbf{x}_n) \\ \Delta(\mathbf{x}_1, \mathbf{x}_2) & 0 & \cdots & \Delta(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & & \vdots \\ \Delta(\mathbf{x}_1, \mathbf{x}_n) & \Delta(\mathbf{x}_2, \mathbf{x}_n) & \cdots & 0 \end{pmatrix}.$$

Zde je na místě poznamenat, že v literatuře se často vedle sebe používají pojmy vzdálenost (*distance*) a nepodobnost (*dissimilarity*). Někteří autoři je volně zaměňují, jiní pojem vzdálenost používají výhradně pro ty nepodobnosti, ktere jsou zároveň i metrikami. Také v této práci budeme používat pro koeficienty nepodobnosti pojem nepodobnost a pojem vzdálenost vyhradíme pro metriky (nejčastěji

eukleidovskou). Matice vzdáleností je tedy často přes svůj název maticí nepodobností mezi pozorováními.

Jedním z důvodů, proč se používají metody založené na matici vzdáleností, je nevhodnost eukleidovské metriky Δ_{EU}, se kterou pracují klasické metody.

$$\Delta_{EU}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{k=1}^{p} (x_{1k} - x_{2k})^2}, \qquad \mathbf{x}_1, \mathbf{x}_2 \in R^p.$$

Eukleidovskou vzdálenost můžeme využít pokud například měříme geografické vzdálenosti mezi stanovišti nebo hodnotu nějakých fyziologických proměnných (např. délku křídel vrabců (Anderson (2006)). Je už ale méně vhodná, pokud potřebujeme porovnávat druhovou skladbu na stanovištích, kdy spíše než na vlastních počtech druhů záleží více na jejich přítomnosti či nepřítomnosti. Použitím jiných koeficientů nepodobnosti můžeme analýzu přizpůsobit vlastním potřebám. Volba koeficientu nepodobnosti je tedy velmi důležitým krokem, který ovlivňuje výsledky celé analýzy, a proto by měla být vždy odůvodněna. Za zmínku rovněž stojí možnost sloučit jednotlivé druhy do vyšších taxonomických jednotek, např. čeledí nebo tříd (Olsgard a kol. (1997)).

Na příkladu uvedeném v Legendre a Gallagher (2001) ukážeme, proč nemusí být eukleidovská vzdálenost vždy vhodná. Mějme tři stanoviště, na kterých sledujeme četnosti tří druhů, viz tabulka 2.1. Požadujeme, aby námi zvolená nepodobnost reflektovala to, že na stanovišti B se vyskytuje pouze druh 1, zatímco na zbylých stanovištích se tento druh nevyskytuje. Stanoviště B by se tedy mělo od zbylých dvou lišit co nejvíce.

	Druh 1	Druh 2	Druh 3
Stanoviště A	0	1	1
Stanoviště B	1	0	0
Stanoviště C	0	4	8

Tabulka 2.1: Příklad: Výskyty druhů na stanovištích.

V tabulce 2.2 jsou uvedeny eukleidovské vzdálenosti mezi jednotlivými stanovišti. Vidíme, že nejvíce podobné (ve smyslu nejméně nepodobné) jsou stanoviště A a B, zatímco stanoviště A a C jsou si velmi nepodobné. To ale neodpovídá našim požadavkům.

$\Delta_{EU}(\mathrm{A},\mathrm{B})$	$\Delta_{EU}(\mathrm{A},\mathrm{C})$	$\Delta_{EU}(\mathrm{B},\mathrm{C})$
1,73	7,65	9,00

Tabulka 2.2: Příklad: Eukleidovské vzdálenosti mezi stanovišti.

Jednou z variant, jak přistoupit k tomuto problému, je transformovat vlastní data a posléze použít eukleidovskou popř. váženou eukleidovskou vzdálenost $\Delta^w_{EU}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{k=1}^{p} w_k (x_{1k} - x_{2k})^2}$. Některé z užívaných koeficientů nepodobností vznikly právě takovýmto způsobem (Legendre a Gallagher (2001)).

- *Chord distance*, navrhovaná Orlóci (1967), je eukleidovská vzdálenost spočítaná pro vektory pozorování po normování na velikost 1, neboli nepodobnost $\Delta_{Chord}(\mathbf{x}_1, \mathbf{x}_2)$ je rovna $\Delta_{EU}(\mathbf{x}'_1, \mathbf{x}'_2)$, kde $\mathbf{x}'_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}$ a $\|\cdot\|_2$ eukleidovská

norma.

$$\Delta_{Chord}(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad \sqrt{\sum_{k=1}^{p} \left(\frac{x_{1k}}{\sqrt{\sum_{l=1}^{p} x_{1l}^2}} - \frac{x_{2k}}{\sqrt{\sum_{l=1}^{p} x_{2l}^2}} \right)^2}.$$

Tento koeficient nepodobnosti je užitečný v situaci, kdy nás nezajímají vlastní četnosti výskytu, ale jen poměry v zastoupeních jednotlivých druhů. Dvojici pozorování, z nichž jedno je nenulovým násobkem druhého přiřadí tento koeficient nepodobnost 0. Tento koeficient je shora omezen hodnotou $\sqrt{2}$ a úzce souvisí s úhlem, který svírají vektory pozorování ve výběrovém prostoru (Legendre a Legendre (1998), kapitola 7.4).

- χ^2 *metrika* $\Delta_{\chi^2 metric}$ je určena pro situace, kdy požadujeme, aby rozdíly v méně početných druzích více přispívaly k nepodobnostem mezi stanovišti. Jedná se o váženou eukleidovskou vzdálenost na vektorů normovaných na velikost 1 pomocí normy $\|\mathbf{x}_i\|_1 = \sum_{l=1}^{p} x_{il}$, kde vahami jsou převrácené hodnoty celkových četností jednotlivých druhů na všech stanovištích. Je-li k-tý druh málo zastoupený, je jeho příslušná váha velká.

$$\Delta_{\chi^2 metric}(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad \sqrt{\sum_{k=1}^{p} \frac{1}{\sum_{i=1}^{n} x_{ik}} \left(\frac{x_{1k}}{\sum_{l=1}^{p} x_{1l}} - \frac{x_{2k}}{\sum_{l=1}^{p} x_{2l}} \right)^2}$$

- S χ^2 metrikou se pojí χ^2 *distance* $\Delta_{\chi^2 distance}$. Od předchozí se liší pouze multiplikativní konstantou $\sqrt{\sum_{i=1}^{n} \sum_{k=1}^{p} x_{ik}}$, tj. součtem všech četností výskytu všech druhů.

$$\Delta_{\chi^2 distance}(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad \Delta_{\chi^2 metric}(\mathbf{x}_1, \mathbf{x}_2) \sqrt{\sum_{i=1}^{n} \sum_{k=1}^{p} x_{ik}}$$

- *Hellinger distance* je další z doporučovaných koeficientů nepodobnosti (Rao (1995)). Vzorec pro výpočet je

$$\Delta_{Hellinger}(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad \sqrt{\sum_{k=1}^{p} \left(\sqrt{\frac{x_{1k}}{\sum_{l=1}^{p} x_{1l}}} - \sqrt{\frac{x_{2k}}{\sum_{l=1}^{p} x_{2l}}} \right)^2}$$

Je zřejmé, že $\Delta_{Hellinger}(\mathbf{x}_1, \mathbf{x}_2) = \Delta_{EU}(\mathbf{x}'_1, \mathbf{x}'_2)$, kde nyní $\mathbf{x}'_i = \sqrt{\frac{x_{ik}}{\sum_{k=1}^{p} x_{ik}}}$.

Dále se často používají i speciálně navržené koeficienty nepodobnosti pro četnosti, procentuální zastoupení, nebo přítomnosti a nepřítomnosti jednotlivých druhů.

- *Bray-Curtisův koeficient nepodobnosti* Δ_{BC} se stal pro svou jednoduchost a snadnou interpretaci velmi populárním.

$$\Delta_{BC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{k=1}^{p} |x_{1k} - x_{2k}|}{\sum_{k=1}^{p} (x_{1k} + x_{2k})}$$

Δ_{BC} nabývá hodnot pouze v intervalu $[0, 1]$, kde hodnota 1 je přiřazena stanovištím, která nemají žádné společné druhy. Tento koeficient je vhodný i pro situace, kdy data nemají formát četností jednotlivých druhů, ale zaznamenávají pouze jejich přítomnost či absenci. Jeho nevýhodou je, že pokud se na stanovištích $\mathbf{x}_1$ a $\mathbf{x}_2$ nevyskytují žádné druhy, vyskytne se v jeho jmenovateli nula. Pak se obvykle definuje $\Delta_{BC} = 1$, případně se tato stanoviště vyloučí z analýzy.

- Koeficient *Canberra metric*

$$\Delta_{Canberra}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{\mathrm{NZ}} \sum_{k=1}^{p} \frac{|x_{1k} - x_{2k}|}{|x_{1k}| + |x_{2k}|},$$

kde NZ označuje počet nenulových sčítanců, je citlivější na rozdíly v méně zastoupených druzích než u početnějších druhů. Pro stanoviště, na kterých se nevyskytují žádné druhy se definuje $\frac{0}{0} = 0$. Tento koeficient nepodobnosti je méně ovlivněn odlehlými hodnotami než Manhattan distance.

- *Manhattan distance*

$$\Delta_{Manhattan}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{k=1}^{p} |x_{1k} - x_{2k}|.$$

Výše uvedené koeficienty nepodobnosti nejsou rozhodně jedinými, které lze použít. Mnoho dalších je uvedeno v knize Legendre a Legendre (1998), kapitola 7.4. Tamtéž lze nalézt i informace o koeficientech podobnosti, které lze rovněž využít pro konstrukci matice vzdáleností.

Vraťme se nyní k našemu příkladu se stanovišti A, B a C. V tabulce 2.3 jsou uvedeny nepodobnosti vypočtené pomocí právě uvedených koeficientů. Na první pohled je zřejmé, že pro všechny uvedené koeficienty až na $\Delta_{Manhattan}$ a Δ_{EU} je $\Delta(A, C) < \Delta(B, C)$ a $\Delta(A, C) < \Delta(A, B)$, což jsme požadovali.

	$\Delta(A,B)$	$\Delta(A,C)$	$\Delta(B,C)$
Δ_{EU}	1,73	7,62	9,00
Δ_{Chord}	1,41	0,32	1,41
$\Delta_{\chi^2 metric}$	1,04	0,09	1,03
$\Delta_{\chi^2 distance}$	4,02	0,36	4,01
$\Delta_{Hellinger}$	1,41	0,17	1,41
Δ_{BC}	1,00	0,71	1,00
$\Delta_{Canberra}$	3,00	1,38	3,00
$\Delta_{Manhattan}$	3,00	10,00	13,00

Tabulka 2.3: Příklad: Nepodobnosti mezi stanovišti.

Na závěr definujme stejně jako v Legendre a Legendre (1998) eukleidovskost matic vzdáleností, potažmo koeficientů nepodobností.

Definice 2.3 *Matici $\boldsymbol{D}_{n\times n}$ nazveme eukleidovskou, pokud existuje $p \in N$ a n-tice bodů $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in R^p$, že $[\boldsymbol{D}]^2_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j)$, kde $'$ označuje transpozici.*

Matice $\mathbf{D}$ je tedy eukleidovská, pokud existuje nějaká konfigurace n bodů v prostoru R^p, že eukleidovské vzdálenosti mezi těmito body odpovídají příslušným prvkům matice $\mathbf{D}$. Koeficient nepodobnosti se pak nazývá eukleidovský, pokud jeho použitím vznikají pouze eukleidovské matice. Je zřejmé, že eukleidovská vzdálenost Δ_{EU} jako koeficient nepodobnosti má tuto vlastnost triviálně. Eukleidovské jsou rovněž koeficienty $\Delta_{\chi^2 metric}$ a $\Delta_{\chi^2 distance}$. Naopak Δ_{BC} a $\Delta_{Manhattan}$ eukleidovské nejsou a v případě Δ_{Chord}, $\Delta_{Hellinger}$ a $\Delta_{Canberra}$ toto není známo (Legendre a Legendre (1998), kapitola 9.4, str. 433). Informace o eukleidovskosti dalších koeficientů nepodobnosti lze nalézt v kapitole 7.4 tamtéž.

2.2 PCoA

Metoda hlavních koordinát (*principal coordinate analysis*, PCoA, *multidimensional scaling*, MDS, (Gower (1966))) je postup, jak pro matici vzdáleností n bodů nalézt reprezentaci (konfiguraci) v mnohorozměrném eukleidovském prostoru R^p tak, aby eukleidovské vzdálenosti mezi body této reprezentace odpovídaly původně zadaným nepodobnostem mezi příslušnými body.

V první řadě tento postup umožňuje grafické znázornění jinak mnohorozměrných dat, přičemž díky použití koeficientů nepodobnosti lze takto zpracovat i data ordinální nebo nominální. Ve druhé pak umožňuje rozšíření testů založených na eukleidovské vzdálenosti na obecné koeficienty nepodobnosti (viz kapitola 4, test homogenity disperzí). Jediným, dílčím omezením je eukleidovskost matice vzdáleností se kterou se pracuje. Pro neeukleidovské matice vzdáleností přesná reprezentace v eukleidovském prostoru neexistuje. V takovém případě pak reprezentace v R^p zachytí pouze část obsažených nepodobností.

Postup metody hlavních koordinát je následující:

- Mějme maticí vzdáleností $\mathbf{D}_{n\times n}$ mezi n prvky $\mathbf{x}_1, ..., \mathbf{x}_n$ získanou pomocí nějakého koeficientu nepodobností Δ, tedy $[\mathbf{D}]_{ij} = \Delta(\mathbf{x}_i, \mathbf{x}_j)$.

- Definujme matici $\mathbf{A}$

$$[\mathbf{A}]_{ij} = -\frac{1}{2}[\mathbf{D}]^2_{ij}. \tag{2.1}$$

- Matici $\mathbf{A}$ vycentrujme tak, aby řádkové a sloupcové průměry matice $\mathbf{A}$ byly nulové:

$$[\mathbf{A}^*]_{ij} = [\mathbf{A}]_{ij} - [\mathbf{A}]_{i\cdot} - [\mathbf{A}]_{\cdot j} + [\mathbf{A}]_{\cdot\cdot}, \tag{2.2}$$

kde

$$\begin{aligned}
[\mathbf{A}]_{i\cdot} &= \frac{1}{n}\sum_{j=1}^{n}[\mathbf{A}]_{ij}\,, \\
[\mathbf{A}]_{\cdot j} &= \frac{1}{n}\sum_{i=1}^{n}[\mathbf{A}]_{ij}\,, \\
[\mathbf{A}]_{\cdot\cdot} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[\mathbf{A}]_{ij}\,.
\end{aligned}$$

Maticově lze tuto operaci zapsat jako

$$\mathbf{A}^* = \left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right)\mathbf{A}\left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right),$$

kde $\mathbf{I}$ značí jednotkovou matici o velikosti $n \times n$ a $\mathbf{1}$ vektor $\underbrace{(1, ..., 1)}_{n\times}'$.

- Spočtěme vlastní čísla matice $\mathbf{A}^*$, $\lambda_1 \geq ... \geq \lambda_n$, a znormujme vlastní vektory příslušné nenulovým vlastním číslům $\mathbf{u}_i$ tak, aby jejich velikost byla rovna odmocnině z daného vlastního čísla λ_i. Pro negativní vlastní čísla se pro normování použije odmocnina absolutní hodnoty vlastního čísla. Tj. $\sqrt{\mathbf{u}_k'\mathbf{u}_k} = \sqrt{|\lambda_k|}$.
- Zapišme tyto vlastní vektory jako sloupce matice $\mathbf{U}$. Vektory odpovídající nulovým vlastním číslům vynecháme.

Protože pro hodnost matic platí $r(\mathbf{BC}) \leq \min(r(\mathbf{B}), r(\mathbf{C}))$ a hodnost matice $\left(\mathbf{I} - \frac{\mathbf{11}'}{n}\right)$ je $n-1$, má matice $\mathbf{A}^*$ nutně alespoň jedno nulové vlastní číslo - to odpovídá tomu, že n bodů tvoří podprostor dimenze $n-1$. $\mathbf{U}$ má tedy rozměr $n \times q$, kde q je počet nenulových vlastních čísel matice $\mathbf{A}^*$ a $\mathbf{A}^* = \mathbf{UU}'$. Řádky matice $\mathbf{U}$ jsou pak souřadnice (koordináty) reprezentací prvků ze kterých byla zkonstruována $\mathbf{D}$. Sloupce odpovídající p kladným vlastním číslům (označme je $\mathbf{U}^+$) udávají koordináty v reálném eukleidovském prostoru R^p, sloupce odpovídající $q-p$ záporným vlastním číslům ($\mathbf{U}^-$) pak v imaginárním iR^{q-p}. Pokud je matice $\mathbf{D}$ eukleidovská, je $\mathbf{A}^* = \mathbf{U}^+\mathbf{U}^{+\prime}$ pozitivně semidefinitní. Všechna nenulová vlastní čísla matice $\mathbf{A}^*$ jsou kladná a jejich počet p určuje dimenzi potřebného eukleidovského prostoru R^p ($p = q$ a $q \leq n-1$). V případě, že matice $\mathbf{D}$ eukleidovská není, je možné v ní provést korekce tak, aby se eukleidovskou stala (Lingoes (1971), Cailliez (1983)), případně problém ignorovat, pokud je reprezentace v eukleidovském prostoru R^p dostatečně kvalitní (viz níže).

Eukleidovské vzdálenosti mezi reprezentacemi pozorování v uplné reprezentaci v $R^p \times iR^{p-q}$ jsou rovny zadaným nepodobnostem mezi původními pozorováními:

$$\Delta_{EU}(\mathbf{u}_i, \mathbf{u}_j) = \sqrt{\sum_{k=1}^{p}(u_{ik} - u_{jk})^2 - \sum_{k=p+1}^{q}(u_{ik} - u_{jk})^2} = \Delta(\mathbf{x}_i, \mathbf{x}_j),$$

kde $\mathbf{u}_i = (u_{i1}, ..., u_{iq})$ a $\mathbf{u}_j = (u_{j1}, ..., u_{jq})$ jsou koordináty reprezentací dvou pozorování $\mathbf{x}_i$ a $\mathbf{x}_j$. Δ označuje použitý koeficient nepodobnosti.

Jako grafický výstup PCoA se používá graf prvních m (v praxi vetšinou dvou) složek souřadnic objektů. Kvalita této reprezentace, stejně jako kvalita reprezentace v reálném R^p případě, že se ve výpočtu objevila záporná vlastní čísla se měří následovně (Legendre a Legendre (1998), kapitola 9.2):

Pokud jsou všechna záporná vlastní čísla v absolutní hodnotě menší než prvních m vlastních čísel

$$R_1 = \frac{\sum_{k=1}^{m}\lambda_k}{\sum_{k=1}^{p}\lambda_k},$$

pokud ne

$$R_2 = \frac{\sum_{k=1}^{m}\lambda_k + m\,|\lambda_q|}{\sum_{k=1}^{q}\lambda_k + (q-1)\,|\lambda_q|},$$

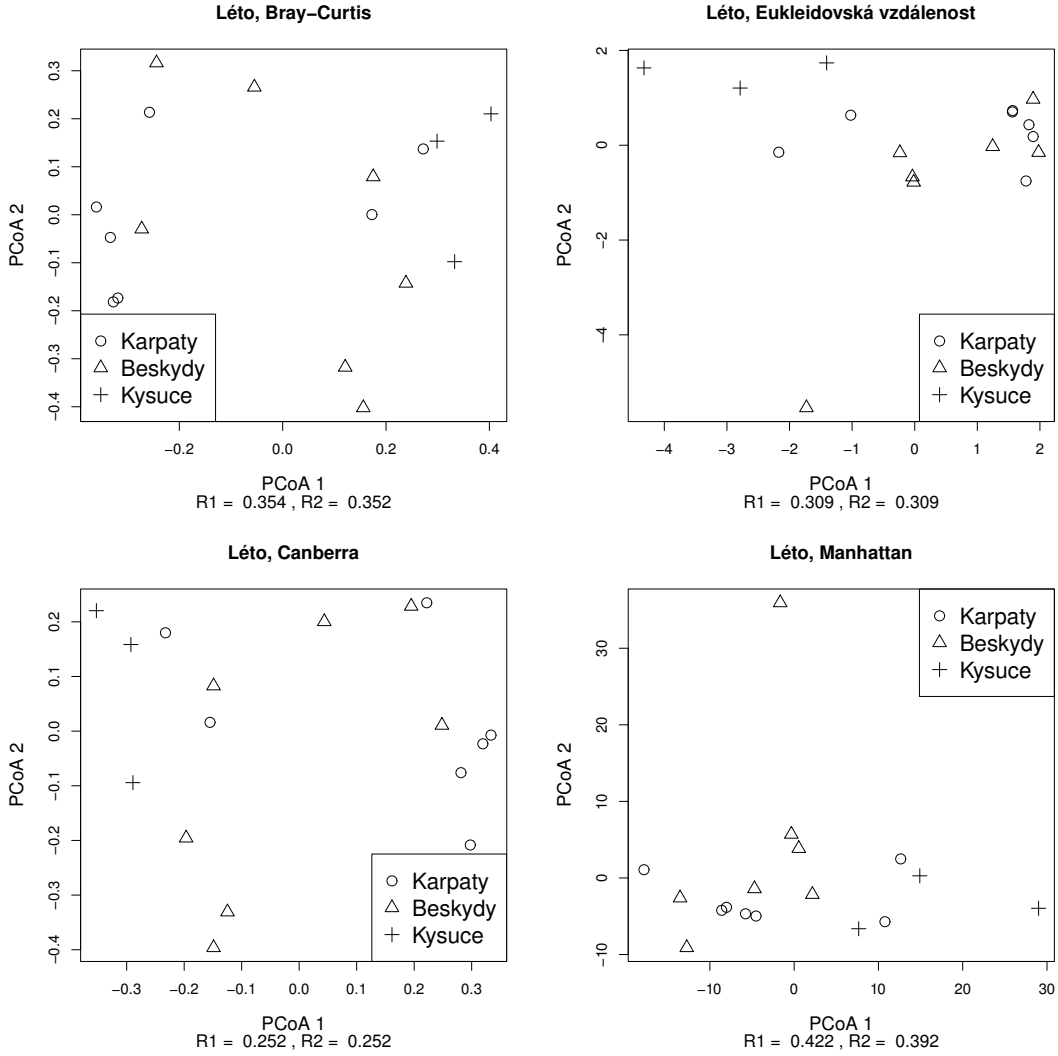
	λ_1	λ_2	λ_{15}	λ_{16}	R_1	R_2
Δ_{BC}	1,246	0,669	0,038	-0,003	0,354	0,352
Δ_{EU}	61,3	42,1	4,3	3,6	0,309	0,309
$\Delta_{Canberra}$	1,00	0,62	0,19	0,16	0,252	0,252
$\Delta_{Manhattan}$	2283	1614	-7	-81	0,422	0,392

Tabulka 2.4: Vybraná vlastní čísla a hodnoty R_1 a R_2 pro Δ_{BC}, Δ_{EU}, $\Delta_{Canberra}$ a $\Delta_{Manhattan}$.

kde λ_q je nejmenší vlastní číslo z uspořádání $\lambda_1 \geq ... \geq \lambda_q$, tedy záporné vlastní číslo s největší absolutní hodnotou. Tyto koeficienty se interpretují jako podíl variability zachycený reprezentací. Poznamenejme ještě, že konfigurace, kterou získáme pomocí metody hlavních koordinát, je co do kvality ekvivalentní reprezentacím, které bychom získali její rotací, zrcadlením nebo posunutím.

Metoda hlavních koordinát a tvorba reprezentací je, jak dále uvidíme, součástí testu homogenity disperzí. Klasické statistické metody, jako je např. ANOVA, implicitně používají eukleidovskou vzdálenost. Vyvořením takovéto reprezentace, která zachovává vztahy mezi původními pozorováními získáme možnost „podpsunout“ těmto metodám data na způsob, jaký potřebujeme. Navíc není nutné znát data samotná, ale stačí mít k dispozici pouze matici vzdáleností $\mathbf{D}$.

Metodu PCoA ilustrujme na datech `mouchy` (viz příloha), konkrétně na letních pozorováních druhových četností. Data jsou rozdělena do třech skupin - oblastí Beskydy (7 stanovišť), Karpaty (7 stanovišť) a Kysuce (3 stanoviště) a na každém stanovišti byly druhové četnosti měřeny dvakrát. Tato dvě pozorování jsme vždy nahradili jejich průměrem a poté použili transformaci $\sqrt[4]{x}$, (viz příloha). Na obrázku 2.1 je reprezentace stanovišť ve dvou dimenzích. Vzdálenosti mezi jednotlivými body odpovídají nepodobnostem druhových skladeb stanovišť při použití Bray-Curtisova koeficientu nepodobnosti Δ_{BC} , eukleidovské vzdálenosti Δ_{EU} , $\Delta_{Canberra}$ a $\Delta_{Manhattan}$ na druhové skladby jednotlivých stanovišť (proměnné `druh1`,..., `druh156`). Ve všech případech je pro úplnou reprezetaci potřeba 16 dimenzí. Pro Bray-Curtisův koeficient nepodobnosti vzniklo jedno záporné vlastní číslo ($\lambda_{16} = -0{,}003$), hodnoty $R_1 = 0{,}354$ a $R_2 = 0{,}352$ se od sebe ale mnoho neliší. Pro eukleidovské vzdálenosti jsou všechna vlastní čísla nezáporná a $R_1 = R_2 = 0{,}309$. Pro $\Delta_{Canberra}$ opět nevznikla žádná záporná vlastní čísla a $R_1 = R_2 = 0{,}252$. Pro $\Delta_{Manhattan}$ vznikla dvě záporná vlastní čísla ($\lambda_{15} = -7, \lambda_{16} = -81$) a $R_1 = 0{,}422$ a $R_2 = 0{,}392$. V tabulce 2.4 jsou uvedena vždy první a poslední dvě vlastní čísla z uspořádání $\lambda_1 \geq ... \geq \lambda_q$ a hodnoty R_1 a R_2 pro výše uvedené koeficienty nepodobnosti. V závislosti na použitém koeficientu nepodobnosti první dvě koordináty zachytí od 25 do 40 % variability. Pro Bray-Curtisův koeficient nepodobnosti a eukleidovskou vzdálenost, se kterými budeme v této práci dále pracovat, se zdá skupina pozorování příslušející Kysucím být umístěna odděleně od zbylých dvou skupin pozorování. To nás vede k domněnce, že druhová skladba v Kysucích se liší od druhové skladby v Karpatech a Beskydech. Právě na těchto datech budeme v kapitole 4 ilustrovat použití testů ANOSIM a MRPP.



Obrázek 2.1: PCoA, grafické znázornění letních pozorování pro Bray-Curtisův koeficient nepodobnosti, eukleidovskou vzdálenost, $\Delta_{Canberra}$ a $\Delta_{Manhattan}$.

3. Mantelův test

3.1 Testování nezávislosti, permutační test

Mějme náhodný výběr $(X_1, Y_1), ..., (X_n, Y_n)$ z nějakého dvojrozměrného rozdělení. Výběrový korelační koeficient r

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

se používá pro měření závislosti veličin X a Y. Pokud (X, Y) pocházejí z dvojrozměrného normálního rozdělení s konečnými kladnými rozptyly a korelačním koeficientem $\rho = 0$, je $T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t_{n-2}$, čehož se využívá pro testování hypotézy nulovosti korelačního koeficientu ρ.

$$\begin{aligned} &\mathrm{H}_0: \quad \rho = 0, \\ &\mathrm{H}_1: \quad \rho \neq 0. \end{aligned}$$

Poznamenejme, že pro dvojrozměrné normální rozdělení nekorelovanost složek již implikuje jejich nezávislost.

Tento test se používá jako test hypotézy nezávislosti rozdělení jednotlivých složek vektoru (X, Y).

$$\begin{aligned} &\mathrm{H}_0: \quad \text{veličiny X a Y jsou nezávislé,} \\ &\mathrm{H}_1: \quad \text{veličiny X a Y nejsou nezávislé.} \end{aligned}$$

P-hodnotu tohoto testu můžeme získat také pomocí permutací (Fisher (1935), Pitman (1937)). Předpokladem pro tento postup je zaměnitelnost permutovaných objektů za platnosti nulové hypotézy.

Definice 3.1 *Nechť $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ jsou náhodné vektory. Označme $\mathcal{L}(\boldsymbol{X}_i)$ rozdělení vektoru $\boldsymbol{X}_i$. Tyto vektory se nazývají zaměnitelné, pokud pro všechny permutace π na konečně mnoha prvcích $1, ..., n$ platí*

$$\mathcal{L}(\boldsymbol{X}_{\pi(1)}, ..., \boldsymbol{X}_{\pi(n)}) = \mathcal{L}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n).$$

Je zřejmé, že prvky náhodného výběru - nezávislé stejně rozdělené náhodné vektory - jsou zaměnitelné. Tyto prvky budeme nazývat zaměnitelné objekty.

P-hodnotu pomocí permutací zjišťujeme následujícím způsobem. Nechť $X_1, ..., X_n$ je náhodný výběr z nějakého rozdělení a $S = S(X_1, ..., X_n)$ je nějaká používaná statistika. Označme S_0 napozorovanou hodnotu statistiky $S(X_1, ..., X_n)$. Dále označme Π množinu všech permutací na n prvcích. Π obsahuje $n!$ prvků a ty označme π_i, $i = 1, ..., n!$. Označme S_i^* hodnoty statistiky S vypočtené na permutovaných výběrech, tedy $S_i^* = S(X_{\pi_i(1)}, ..., X_{\pi_i(n)})$. Takto jsme získali (podmíněné)

rozdělení statistiky S při daných pozorováních, se kterým pak porovnáváme hodnotu S_0. Pokud je test oboustranný, vypočteme jeho P-hodnotu jako

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\{|S_i^*| \geq |S_0|\}}{n!},$$

kde $\mathbb{I}\{|S_i^*| \geq |S_0|\}$ je indikátor jevu, že hodnota S_i^* je v absolutní hodnotě větší než $|S_0|$. Pro jednostrannou variantu se P-hodnota vypočte jako

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\{S_i^* \geq S_0\}}{n!}, \tag{3.1}$$

popřípadě

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\{S_i^* \leq S_0\}}{n!}. \tag{3.2}$$

Výhodou tohoto postupu je, že jeho jediným předpokladem je zaměnitelnost permutovaných objektů za nulové hypotézy. Není tedy třeba předpokládat, že náhodný výběr pochází z nějakého konkrétního rozdělení. Nevýhodou tohoto postupu je, že s rostoucím počtem pozorování velmi rychle roste jeho výpočetní náročnost. V případě, že je celkový počet možných permutací příliš velký, používá se Monte Carlo aproximací k odhadu skutečné P-hodnoty testu. V závislosti na požadované přesnosti tohoto odhadu se zvolí dostatečný počet náhodných permutací na kterých se provede výše uvedený postup. Přibližná P-hodnota oboustranného testu se pak vypočte jako

$$\frac{\sum_{i=1}^{M} \mathbb{I}\{|S_i^*| \geq |S_0|\} + 1}{M+1},$$

kde M je předem zvolený počet permutací. P-hodnoty jednostranných testů se pak vypočtou analogicky z (3.1) a (3.2). Směrodatná odchylka tohoto odhadu P-hodnoty je rovna $\sqrt{\frac{p(1-p)}{M+1}}$, kde p je skutečná P-hodnota testu. Pro $p = 0{,}05$ a $M = 999$ je $\mathrm{se}(\hat{p}) = 0{,}0067$, pro $p = 0{,}05$ a $M = 9999$ je $\mathrm{se}(\hat{p}) = 0{,}0022$.

Vypočtěme tedy P-hodnotu testu nezávislosti složek vektorů ze začátku této kapitoly pomocí permutací. Za platnosti nulové hypotézy jsou X_i a Y_i nezávislé.

$$\mathcal{L}(X_1, Y_1) = \mathcal{L}(X_i, Y_j), \quad i, j = 1, ..., n.$$

Budeme tedy zaměňovat přiřazení Y_j k jednotlivým X_i. Označme T_0 hodnotu statistiky $T((X_1, Y_1),...,(X_n, Y_n))$ a T_i^* hodnoty statistiky vypočtené na permutovaných výběrech $T((X_1, Y_{\pi_i(1)}),...,(X_n, Y_{\pi_i(n)}))$, $\pi_i \in \Pi$. Přesnou, resp. přibližnou P-hodnotu tohoto testu vypočteme jako

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\{|T_i^*| \geq |T_0|\}}{n!} \quad \text{resp.} \quad \frac{\sum_{i=1}^{M} \mathbb{I}\{|T_i^*| \geq |T_0|\} + 1}{M+1}.$$

Tento test lze také sestavit jako jednostranný.

H_0:	veličiny X a Y jsou nezávislé,
H_1:	veličiny X a Y jsou kladně korelované.

Přesná resp. přibližná P-hodnota tohoto testu jsou pak

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\{T_i^* \geq T_0\}}{n!} \quad \text{resp.} \quad \frac{\sum_{i=1}^{M} \mathbb{I}\{T_i^* \geq T_0\} + 1}{M+1}.$$

Zaměřme se nyní na tento jednostanný test. Stejné P-hodnoty dostaneme, použijeme-li namísto testové statistiky T samotný výběrový korelační koeficient r. Protože T je ryze rostoucí funkcí r je

$$\mathbb{I}\{T_i^* \geq T_0\} = \mathbb{I}\{r_i^* \geq r_0\}.$$

Protože permutace nemají vliv na členy $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\sum_{i=1}^n Y_i$ a $\sum_{i=1}^n Y_i^2$, je také

$$\mathbb{I}\{T_i^* \geq T_0\} = \mathbb{I}\{r_i^* \geq r_0\} = \mathbb{I}\{m_i^* \geq m_0\},$$

kde

$$m = \sum_{i=1}^{n} X_i Y_i. \tag{3.3}$$

Pokud tedy zjišťujeme P-hodnotu pomocí permutací, můžeme namísto statistiky T pracovat se statistikou m.

Situace se zkomplikuje, pokud je náhodný výběr $(\mathbf{X}_1, \mathbf{Y}_1), ..., (\mathbf{X}_n, \mathbf{Y}_n)$, $i = 1, ..., n$ složen z vektorů $\mathbf{X}_i$ a $\mathbf{Y}_i$ s nějakými mnohorozměrnými rozděleními. Mantelův test (Mantel (1967)), který popíšeme v následující části, využívá právě uvedené statistiky m (3.3).

3.2 Aplikace pro matici vzdáleností, Mantelův test

Nechť $(\mathbf{X}_1, \mathbf{Y}_1), ..., (\mathbf{X}_n, \mathbf{Y}_n)$, $i = 1, ..., n$ je náhodný výběr, kde vektor $\mathbf{X}_i$ má nějaké mnohorozměrné rozdělení a popisuje nějaké vlastnosti stanovišť. Nechť $\mathbf{Y}_i$ pochází rovněž z nějakého mnohorozměrného rozdělení a popisuje četnosti jednotlivých druhů na daných stanovištích. Testujeme hypotézu

H_0:	náhodné vektory $\mathbf{X}$ a $\mathbf{Y}$ jsou nezávislé,
H_1:	vektory $\mathbf{X}$ a $\mathbf{Y}$ nejsou nezávislé.

Myšlenka Mantelova testu je: Označme $\mathbf{A}$ matici vzdáleností n stanovišť spočtenou na základě $(\mathbf{X}_1, ..., \mathbf{X}_n)$ a koeficientu nepodobnosti Δ_X a $\mathbf{B}$ matici vzdáleností druhových skladeb na těchto stanovištích spočtenou na základě $(\mathbf{Y}_1, ..., \mathbf{Y}_n)$ a koeficientu nepodobnosti Δ_Y. Jsou-li vektory $\mathbf{X}$ a $\mathbf{Y}$ nezávislé, pak i vypočtené nepodobnosti $\Delta_X(\mathbf{X}_i, \mathbf{X}_j) = [\mathbf{A}]_{ij} = \mathbf{a}_{ij}$ a $\Delta_Y(\mathbf{Y}_i, \mathbf{Y}_j) = [\mathbf{B}]_{ij} = \mathbf{b}_{ij}$, $i, j = 1, ..., n$ jsou nezávislé. Testovou statistikou je v minulé části představená statistika m, kterou je nutno upravit pro práci s maticemi vzdáleností na tvar

$$z_1 = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{a}_{ij} \mathbf{b}_{ij},$$

kde $\mathbf{a}_{ij}$ a $\mathbf{b}_{ij}$ jsou prvky příslušných matic. Vzhledem k vlastnostem matice vzdáleností lze z_1 dále upravit na

$$z_2 = \sum_{i=2}^{n}\sum_{j=1}^{i-1} \mathbf{a}_{ij}\mathbf{b}_{ij}.$$

Díky symetrii a přítomnosti nul na hlavní diagonále se jedná jen o vydělení hodnoty z_1 dvěma. Toto je také tvar, ve kterém se tato statistika uvádí nejčastěji a nazývá se Mantelova statistika (Legendre a Legendre (1998), kapitola 10.5). Někdy se používá i normovaná varianta Mantelovy statistiky z_2, kterou označíme z_3. Ta nabývá, stejně jako korelační koeficient, hodnot pouze v intervalu $[-1, 1]$ a stejná je i její interpretace jako míry lineární závislosti.

$$z_3 = \frac{1}{\frac{n(n-1)}{2} - 1}\sum_{i=2}^{n}\sum_{j=1}^{i-1} \frac{\mathbf{a}_{ij} - \bar{\mathbf{a}}}{s_a}\frac{\mathbf{b}_{ij} - \bar{\mathbf{b}}}{s_b},$$

kde

$$\bar{\mathbf{a}} = \frac{1}{\frac{n(n-1)}{2}}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbf{a}_{ij}, \qquad \bar{\mathbf{b}} = \frac{1}{\frac{n(n-1)}{2}}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathbf{b}_{ij},$$

$$s_a^2 = \frac{1}{\frac{n(n-1)}{2} - 1}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\mathbf{a}_{ij} - \bar{\mathbf{a}})^2, \qquad s_b^2 = \frac{1}{\frac{n(n-1)}{2} - 1}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\mathbf{b}_{ij} - \bar{\mathbf{b}})^2.$$

Za platnosti nulové hypotézy jsou podmínky na stanovištích (ze kterých se získávavají prvky $\mathbf{A}$) a četnosti výskytu druhů (ze které se získávají prvky $\mathbf{B}$) nezávislé a

$$\mathcal{L}(\mathbf{X}_1, \mathbf{Y}_1) = \mathcal{L}(\mathbf{X}_i, \mathbf{Y}_j), \quad i, j = 1, ..., n.$$

Budeme tedy permutovat přiřazení četností druhů jednotlivým stanovištím. Označme $(z_2)_0$ hodnotu Mantelovy statistiky vypočtené na původním výběru $(\mathbf{X}_1, \mathbf{Y}_1), ..., (\mathbf{X}_n, \mathbf{Y}_n)$ a $\mathbf{A}_0$ a $\mathbf{B}_0$ příslušné matice vzdáleností. Permutované výběry mají tvar $(\mathbf{X}_1, \mathbf{Y}_{\pi_i(1)}), ..., (\mathbf{X}_n, \mathbf{Y}_{\pi_i(n)})$, $\pi_i \in \Pi$. Pro tyto výběry se opakovaně počítají matice vzdáleností $\mathbf{A}_i^*$ a $\mathbf{B}_i^*$ a na základě těchto matic hodnoty statistiky $(z_2)_i^*$. Protože

$$\mathbb{I}\left\{(z_1)_i^* \geq (z_1)_0\right\} = \mathbb{I}\left\{(z_2)_i^* \geq (z_2)_0\right\} = \mathbb{I}\left\{(z_3)_i^* \geq (z_3)_0\right\},$$

jsou P-hodnoty všech variant tohoto testu stejné a to

$$\frac{\sum_{i=1}^{n!}\mathbb{I}\left\{(z_2)_i^* \geq z_0^2\right\}}{n!} \quad \text{resp.} \quad \frac{\sum_{i=1}^{M}\mathbb{I}\left\{(z_2)_i^* \geq z_0^2\right\} + 1}{M+1}.$$

V praxi však není potřeba opakovaně provádět výpočty matic vzdáleností. Matice $\mathbf{A}_i^*$ jsou pro všechny permutace stejné a matice $\mathbf{B}_i^*$ vznikly z matice $\mathbf{B}_0$ současnou permutací řádků a sloupců.

$$\begin{aligned} \mathbf{A}_i^* &= \mathbf{A}_0, \\ [\mathbf{B}_i^*]_{jk} &= [\mathbf{B}_0]_{\pi_i(j)\pi_i(k)}, \quad \pi_i \in \Pi, j, k = 1, ..., n \end{aligned}$$

Pro výpočet P-hodnoty tohoto testu tedy stačí současně permutovat řádky a sloupce matice **B**. Poznamenejme, že výpočet P-hodnoty odpovídá jednostrannému testu. Důvodem je, že nás obvykle zajímá, jestli na stanovištích s podobnými vlastnostmi žijí podobné druhy a na stanovištích s rozdílnými vlastnostmi ddruhy rozdílné, tedy jestli mezi prvky **A** a **B** existuje kladná závislost. Výsledky Mantelova testu pro matice **A** a **B** tak, jak jsme je zvolili my, se pak interpretují takto: Pokud nezamítneme nulovou hypotézu, „četnosti výskytů jednotlivých druhů na stanovištích nezávisí na vlastnostech stanovišť". Pokud nulovou hypotézu zamítneme, „mají stanoviště s podobnými vlastnostmi podobnou druhovou strukturu".

Podle toho, jaké matice do Mantelova testu zvolíme, jej můžeme použít pro následující situace. Tím se ukazuje i flexibilita tohoto postupu.

- Zajímá-li nás zda blízká stanoviště mají podobnou druhovou skladbu, použijeme jako matici **A** matici geografických vzdáleností.

- Zajímá-li nás vztah mezi druhovou skladbou a přírodními podmínkami na stanovištích, tj. zda na podobných stanovištích žijí podobné druhy, použijeme jako matici **A** matici vzdáleností stanovišť ve smyslu naměřených přírodních podmínek s volbou vhodného koeficientu nepodobnosti.

- Jako matici **A** můžeme použít i matici námi požadovaného modelu a porovnávat naměřenou nepodobnost s nepodobností předpovězenou modelem. Například můžeme pozorování rozdělit do skupin. Matice modelu se pak bude skládat z 0 a 1, přičemž patří-li dvojice pozorování do stejné skupiny, je jim přiřazena nepodobnost 0, v opačném případě pak 1. Tento postup vede na metodu ANOSIM (viz kapitola 4).

Ve Smouse a kol. (1986) lze nalézt rozšíření Mantelova testu inspirované parciálním korelačním koeficientem - parciální Mantelův test, který umožňuje například porovnávat druhovou skladbu s podmínkami na stanovištích při kontrole geografického rozmístění (Zuur a kol. (2007)). Mantelův test může posloužit jako výchozí bod pro podrobnější analýzu, jako například analýzu vzdáleností (*analysis of distance*, Gower a Krzanowski (1999)) nebo redundanční analýzu (*distance-based redundancy analysis*, Legendre a Anderson (1999), McArdle a Anderson (2001)), které jsou již nad rámec této práce. Možnou modifikací Mantelova testu je použití pořadí vzdáleností namísto jejich vlastních hodnot, což odpovídá Spearmanovu korelačnímu koeficientu (Dietz (1983)). Problém nezávislosti dvou vektorů lze řešit i jinými způsoby, např. v Deheuvels (1981) lze nalézt přístup přes zobecněné χ^2 testy nebo použít test významnosti kanonických korelací (Rencher (2002), kapitola 7.4.1 a 11.4.1). Pro tento test je však mimo jiné nutné, aby výběrové varianční matice jednotlivých vektorů měly plnou hodnost (což například v datech `mouchy` není splněno).

Mantelův test ilustrujeme na datech `mouchy`. Zaměřili jsme se na oblast Karpaty a jarní pozorování. V této oblasti se nachází 7 stanovišť. Pro výpočet nepodobností druhových skladeb na stanovištích jsme použili Bray-Curtisův koeficient nepodobnosti. Hodnoty jednotlivých fyzikálních a chemických vlastností stanovišť jsou uvedeny v tabulce 3.1 a pro výpočet nepodobností podmínek na stanovištích jsme použili eukleidovskou vzdálenost poté, co byla provedena standardizace na nulovou střední hodnotu a jednotkový rozptyl, protože tyto veličiny byly měřeny v rozdílných jednotkách (jednotky pH, °C, mg/l). Dále uvádíme matici vzdáleností

druhových skladeb (tabulka 3.2) a matici vzdáleností pro podmínky jednotlivých stanovišť (tabulka 3.3). Tyto matice jsme spočítali v programu `R` pomocí funkcí `vegdist` z knihovny `vegan` a `daisy` z knihovny `cluster`.

Mantelův test jsme počítali pomocí funkce `mantel` z knihovny `vegan`. Počet permutací byl 9999. Hodnota Mantelovy statistiky z_3 je 0,68, P-hodnota tohoto testu je 0,0038, tedy zamítáme hypotézu nezávislosti druhové skladby na podmínkách na stanovišti. Stanoviště s podobnými podmínkami mají podobnou druhovou skladbu a hodnota statistiky ukazuje na silný vztah. Funkce `mantel` rovněž počítá kvantily rozdělení získaného permutacemi, které uvádíme spolu s ostatními výsledky v tabulce 3.4. Pokud namísto vlastních nepodobností použijeme pořadí, dostaneme obdobné výsledky.

stanoviště	pH	T	NH_4	NO_3	Cu	Ca	Mg	Na	K
K1	7,4	9,0	0,03	8,00	1,0	86,9	2,4	6,3	1,6
K2	7,5	7,0	0,08	9,60	6,8	105,0	7,7	15,6	8,7
K3	7,3	14,0	0,04	5,56	1,0	65,5	16,1	6,8	4,9
K4	8,2	10,0	0,03	12,00	1,0	60,6	16,5	16,3	6,9
K5	7,9	9,0	0,03	20,50	1,0	103,0	18,6	7,0	1,4
K6	7,3	10,5	0,13	0,15	1,3	107,0	14,4	3,8	2,7
K7	8,0	9,5	0,03	0,17	1,0	112,0	8,6	3,8	1,0

Tabulka 3.1: Jarní podmínky na stanovištích, skupina Karpaty.

K1	0,82	0,53	0,72	0,74	0,76	0,71
0,82	K2	0,86	0,86	0,78	0,89	0,84
0,53	0,86	K3	0,61	0,65	0,75	0,73
0,72	0,86	0,61	K4	0,73	0,82	0,88
0,74	0,78	0,65	0,73	K5	0,72	0,56
0,76	0,89	0,75	0,82	0,72	K6	0,59
0,71	0,84	0,73	0,88	0,56	0,59	K7

Tabulka 3.2: Matice vzdáleností druhových skladeb na stanovištích, Bray-Curtisovy nepodobnosti.

K1	4,48	3,66	4,39	3,62	3,73	2,59
4,48	K2	5,43	4,64	5,02	4,81	5,05
3,66	5,43	K3	3,76	4,16	3,69	4,15
4,39	4,64	3,76	K4	3,61	5,31	4,52
3,62	5,02	4,16	3,61	K5	4,38	3,43
3,73	4,81	3,69	5,31	4,38	K6	3,45
2,59	5,05	4,15	4,52	3,43	3,45	K7

Tabulka 3.3: Matice vzdáleností podmínek na stanovištích, eukleidovské vzdálenosti.

Mantelův test	z_3	P-hodnota	90%	95%	97,5%	99%
pearson	0,6845	0,0038	0,372	0,467	0,536	0,620
spearman	0,7662	0,0009	0,397	0,501	0,570	0,662

Tabulka 3.4: Mantelův test: pearson - výpočet založený na nepodobnostech, spearman - výpočet založený na pořadí nepodobností. Kvantily rozdělení statistiky z_3 za nulové hypotézy.

4. Jednoduché třídění

V této kapitole uvedeme dvě metody, které se používají při řešení problému jednoduchého třídění v souvislosti s maticí vzdáleností. Popíšeme metodu ANOSIM (Clarke (1993)) založenou na porovnávání průměrného pořadí nepodobností mezi prvky stejných a rozdílných skupin a dále metodu MRPP (Mielke a Berry (2007), kap. 2) inspirovanou klasickým F testem analýzy rozptylu jednoduchého třídění. Dále se budeme zabývat vztahem mezi nimi a vztahem těchto metod k Mantelově testu a představíme test homogenity disperzí, který s těmito testy souvisí.

4.1 ANOSIM

První metodou, kterou si představíme je ANOSIM (*analysis of similarities*, Clarke (1993)). Nechť $F_1, ..., F_g$ jsou distribuční funkce nějakých p-rozměrných rozdělení. Nechť je dáno g nezávislých náhodných výběrů z těchto distribučních funkcí.

$$\begin{aligned} \mathbf{X}_{11}, ..., \mathbf{X}_{1n_1} &\sim F_1, \\ &\vdots \\ \mathbf{X}_{g1}, ..., \mathbf{X}_{gn_g} &\sim F_g. \end{aligned} \tag{4.1}$$

Označme $n = \sum_{i=1}^{g} n_i$. Nadále budeme používat značení G_k pro k-tý výběr, $k = 1, ..., g$ a $\mathbf{X}_i, i \in G_k$ pro vektory tohoto výběru $\mathbf{X}_{k1}, ..., \mathbf{X}_{kn_k}$.

Budeme testovat hypotézu, že rozdělení, ze kterých pochází tyto výběry jsou shodná.

$$\begin{aligned} &\mathrm{H}_0: \quad F_1 \equiv ... \equiv F_g, \\ &\mathrm{H}_1: \quad \text{non } \mathrm{H}_0. \end{aligned} \tag{4.2}$$

Základní myšlenka metody ANOSIM je: Pocházejí-li skupiny ze stejného rozdělení, je rozdělení nepodobností pozorování mezi skupinami (meziskupinové nepodobnosti) a nepodobností pozorování uvnitř skupin (vnitroskupinové nepodobnosti) stejné. Tedy pro $\mathbf{X}_i, \mathbf{X}_j, i, j \in G_k$ a $\mathbf{X}_m, m \in G_l$, $k, l = 1, ..., g$, platí

$$\mathcal{L}(\mathbf{X}_i) = \mathcal{L}(\mathbf{X}_m) \Rightarrow \mathcal{L}(\Delta(\mathbf{X}_i, \mathbf{X}_j)) = \mathcal{L}(\Delta(\mathbf{X}_i, \mathbf{X}_m)), \tag{4.3}$$

kde Δ označuje použitou nepodobnost.

Při testu ANOSIM se porovnávají meziskupinové a vnitroskupinové nepodobnosti a na základě výsledku se pak dělají závěry o původních pozorováních. Je zřejmé, že zatímco rozdělení vektorů $\mathbf{X}$ a zvolený koeficient nepodobnosti Δ již určují rozdělení nepodobností mezi jednotlivými prvky, obrácená implikace obecně neplatí. Vztah (4.3) platí jako ekvivalence pokud hustoty $\mathbf{X}$ a $\mathbf{Y}$ splňují určité požadavky (viz Maa a kol. (1996)), například $\Delta(a\mathbf{x} + \mathbf{b}, a\mathbf{y} + \mathbf{b}) = |a|\,\Delta(\mathbf{x}, \mathbf{y})$, $a \in \mathbf{R}, \mathbf{b} \in \mathbf{R}^p$. Tuto vlastnost má například Eukleidovská vzdálenost Δ_{EU}, nebo $\Delta_{Manhattan}$. Naopak například Bray-Curtisův koeficient nepodobnosti tuto vlastnost nesplňuje.

Podívejme se nyní na samotný test ANOSIM. Zvolme koeficient nepodobnosti Δ, uspořádejme pozorování po skupinách a spočtěme matici vzdáleností $\mathbf{D}_{n\times n}$ mezi jednotlivými pozorováními. Matice $\mathbf{D}$ se skládá z čtvercových

bloků nepodobností mezi prvky ležícími ve stejných skupinách a obecně obdélníkových bloků nepodobností prvků příslušejících skupinám různým. Označme $\mathbf{D}^{rs}_{n_r \times n_s}$ matici vzdáleností prvků r-té a s-té skupiny. Díky symetrii nepodobností jsou matice typu $\mathbf{D}^{rr}_{n_r \times n_r}$ symetrické a na jejich hlavních diagonálách se vyskytují pouze nuly. Dále platí $\mathbf{D}^{rs} = (\mathbf{D}^{sr})'$, kde $'$ označuje transpozici. Vzhledem k těmto vlastnostem se dále pracuje pouze s jednou polovinou prvků matice $\mathbf{D}$. Bez újmy na obecnosti budeme dále pracovat s prvky matice $\mathbf{D}$ pod hlavní diagonálou. Metoda ANOSIM pokračuje přiřazením pořadí r_{ij} těmto prvkům (indexy i a j nyní příslušejí řádkům a sloupcům matice $\mathbf{D}$, avšak zároveň určují i příslušnost do jednotlivých skupin G_k). Tato pořadí nabývají hodnot od 1 do $\frac{n(n-1)}{2}$, což je celkový počet prvků $\mathbf{D}$ pod hlavní diagonálou. Označme $\bar{r}_W$ průměné pořadí nepodobností prvků náležejících stejným skupinám (z bloků $\mathbf{D}^{rr}$) a $\bar{r}_B$ průměrné pořadí nepodobností prvků z různých skupin (bloky $\mathbf{D}^{rs}$, $r > s$).

$$\begin{aligned} \bar{r}_W &= \frac{1}{\sum_{i=1}^{g} \frac{n_i(n_i-1)}{2}} \sum_{k=1}^{g} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij}, \\ \bar{r}_B &= \frac{1}{\frac{n(n-1)}{2} - \sum_{i=1}^{g} \frac{n_i(n_i-1)}{2}} \sum_{k=1}^{g} \sum_{l=k+1}^{g} \sum_{i \in G_k} \sum_{j \in G_l} r_{ij}. \end{aligned}$$

Statistika ANOSIM je pak definována takto:

$$R = \frac{\bar{r}_B - \bar{r}_W}{\frac{n(n-1)}{4}}.$$

Statistika R nabývá hodnot v intervalu [-1,1]. Její významnost se stejně jako u Mantelova testu určuje permutacemi. Připomeňme, že testujeme hypotézu shody rozdělení $F_1, ..., F_g$ (4.2). Za platnosti této hypotézy pocházejí všechna pozorování ze stejného rozdělení, tudíž i veškeré nepodobnosti mezi nimi pocházejí z nějakého stejného rozdělení (toto rozdělení samozřejmě závisí na původním rozdělení pozorování a volbě koeficientu nepodobnosti). Za platnosti nulové hypotézy jsou vlastní pozorování $\mathbf{X}_{ij}$ zaměnitelnými objekty a tedy podobně jako v Mantelově testu i zde se permutuje jejich přiřazení do jednotlivých skupin. Namísto opakovaných výpočtů matice vzdálenosti lze opět provádět současně příslušné permutace řádků a sloupců matice vzdáleností $\mathbf{D}$. Těchto permutací je $n!$. Pokud pozorování pocházejí ze stejných rozdělení, je hodnota statistiky R blízká nule. Nulovou hypotézu zamítáme pro velké hodnoty R, přesná resp. přibližná P-hodnota testu ANOSIM je rovna

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\{R_i^* \geq R_0\}}{n!} \quad \text{resp.} \quad \frac{\sum_{i=1}^{M} \mathbb{I}\{R_i^* \geq R_0\} + 1}{M+1},$$

kde R_0 je hodnota statistiky R na původnách datech, R_i^* jsou hodnoty statistiky R na permutovaných datech a M zvolený počet permutací.

Záporné hodnoty statistiky R se vyskytují zřídka. Tyto hodnoty znamenají, že nepodobnosti uvnitř skupin jsou větší než meziskupinové nepodobnosti. Chapman a Underwood (1999) ukázali, že pravidelně se negativní hodnoty R vyskytují v následujících situacích:

- ve výběrech se vyskytují odlehlá pozorování

- jednotlivé výběry $(X_{11}, ..., X_{1n_1}), ..., (X_{g1}, ..., X_{gn_g})$ na sobě nejsou nezávislé (např. časová autokorelace)

- pracujeme s daty, kdy v rámci jedné skupiny se druhová skladba stanovišť velmi liší, ale ve všech skupinách je podobná, tzv. *patchy habitats*.

Jak již bylo řečeno, test ANOSIM je testem shody rozdělení vnitroskupinových a meziskupinových nepodobností a jeho výsledky se pak vztahují na rozdělení původních pozorování. Tento test se často interpretuje jako test shody polohy rozdělení, pokud lze předpokládat, že pozorování pocházejí z rodiny rozdělení s parametrem polohy.

Definice 4.1 *Nechť $g(\boldsymbol{x})$ je hustota vhledem k nějaké σ-konečné míře ν. Množina všech hustot $\mathcal{F}_1 = \{f | f(\boldsymbol{x}, \boldsymbol{\mu}) = g(\boldsymbol{x} - \boldsymbol{\mu}),\ \boldsymbol{\mu} \in \boldsymbol{R}^p\}$ se nazývá rodina rozdělení s parametrem polohy (location family).*

Zamítá-li test ANOSIM nulovou hypotézu shody rozdělení, interpretuje se tento výsledek jako: „Mezi skupinami existují dvě, které se liší v poloze“. Problémem je, že ANOSIM je testem na celkovou shodu rozdělení, tedy zamítnutí nulové hypotézy nemusí být způsobeno rozdílnými parametry polohy. To ilustruje následující příklad, pro jehož potřeby potřebujeme následující definici.

Definice 4.2 *Veličina U je stochasticky větší než veličina V, pokud $\forall x \in R$ $F_U(x) \leq F_V(x)$ a $\exists x\ F_V(x) \neq F_U(x)$.*

Nechť pro $i = 1, ..., n$ jsou X_i a Y_i nezávislé náhodné veličiny s normálním rozdělením $N(\mu_1, \sigma_1^2)$ resp. $N(\mu_2, \sigma_2^2)$ a Δ je eukleidovská vzdálenost. Parametry polohy jsou zde zřejmě μ_1 a μ_2 a dále se zde vyskytují parametry σ_1 a σ_2, rozptyly jednotlivých rozdělení. Můžeme rozlišit následující čtyři situace:

1. $\mu_1 = \mu_2$ a $\sigma_1^2 = \sigma_2^2 = \sigma^2$
   Obě skupiny pocházejí ze stejného rozdělení. Rozdělení vnitroskupinových a meziskupinových vzdáleností je stejné a to $\Delta(X_i, X_j) = \Delta(Y_i, Y_j) = \Delta(X_i, Y_j) = |N(0, 2\sigma^2)|$, $i, j = 1, ..., n$, kde označením $|N(0, 2\sigma^2)|$ míníme rozdělení veličiny $|Z|$, kde $Z \sim N(0, 2\sigma^2)$.

2. $\mu_1 \neq \mu_2$ a $\sigma_1^2 = \sigma_2^2 = \sigma^2$
   Vzdálenosti prvků z různých skupin $\Delta(X_i, Y_j) = |N(\mu_1 - \mu_2, 2\sigma^2)|$ jsou stochasticky větší než vzdálenosti mezi prvky stejných skupin $\Delta(X_i, X_j) = \Delta(Y_i, Y_j) = |N(0, 2\sigma^2)|$.

3. $\mu_1 = \mu_2$ a $\sigma_1^2 < \sigma_2^2$
   $\Delta(X_i, X_j) = |N(0, 2\sigma_1^2)| < \Delta(X_i, Y_j) = |N(0, \sigma_1^2 + \sigma_2^2)| < \Delta(Y_i, Y_j) = |N(0, 2\sigma_2^2)|$ ve smyslu stochastického uspořádání (definice 4.2), neboli vzdálenosti mezi prvky skupiny X jsou nejmenší, následovány vzdálenostmi mezi prvky různých skupin a vzdálenostmi mezi prvky skupiny Y.

4. $\mu_1 \neq \mu_2$ a $\sigma_1^2 < \sigma_2^2$
   Pak $\Delta(X_i, X_j) = |N(0, 2\sigma_1^2)| < \Delta(X_i, Y_j) = |N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)|$ a $\Delta(X_i, X_j) = |N(0, 2\sigma_1^2)| < \Delta(Y_i, Y_j) = |N(0, 2\sigma_2^2)|$, ale vztah $\Delta(X_i, Y_j)$ a $\Delta(Y_i, Y_j)$ závisí na konkrétních hodnotách středních hodnot a rozptylů.

Body (1.) a (2.) popisují situaci, které jsme se před chvílí věnovali, tedy pokud se rozdělení liší pouze v poloze. Interpretace testu ANOSIM pak jsou korektní. Pro situace v bodech (3.) a (4.) může být interpretace pomocí polohy chybná, neboť zamítnutí nulové hypotézy může být způsobeno právě rozdílností parametrů σ_1^2 a σ_2^2. Stejný problém se řeší i v ANOVě, kde se pro odlišení situací (3.) a (4.) od (1.) a (2.) používá Leveneův test (Levene (1960). Na tomto testu je založen test homogenity disperzí Anderson (2006), který v této kapitole také představíme (viz část 4.4).

4.2 Varianty testu ANOSIM, vztah k Mantelovu testu

Díky tomu, že se signifikance R získává permutacemi lze veškeré její části, které zůstávají při permutování konstantní, vynechat. Mají vliv jen na hodnotu statistiky, ale ne na její významnost. To je případ normovací konstanty $\frac{n(n-1)}{4}$. Vynecháme-li ji, vypadá statistika R následovně:

$$\bar{r}_B - \bar{r}_W = \frac{1}{n_B} \sum_{k=1}^{g} \sum_{l=k+1}^{g} \sum_{i \in G_k} \sum_{j \in G_l} r_{ij} - \frac{1}{n_W} \sum_{k=1}^{g} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij},$$

kde n_B a n_W odpovídají počtu meziskupinových a vnitroskupinových nepodobností.

$$n_W = \sum_{i=1}^{g} \binom{n_i}{2}, \qquad n_B = \sum_{i=1}^{g} \sum_{j=i+1}^{g} n_i n_j = \binom{n}{2} - n_W.$$

Tento výraz lze dále upravit na

$$\bar{r}_B - \bar{r}_W = \frac{1}{n_B} \sum_{k=1}^{g} \sum_{l=1}^{g} \sum_{i \in G_k} \sum_{j \in G_l} r_{ij} - \left(\frac{1}{n_B} + \frac{1}{n_W} \right) \sum_{k=1}^{g} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij},$$

kde definujeme $r_{ii} = 0$. První část tohoto výrazu je invariantní na permutace, stejně jako násobící konstanta před druhou částí, proto stačí pro výpočet P-hodnoty jen výraz

$$r = \sum_{k=1}^{g} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij},$$

součet pořadí vnitroskupinových nepodobností ve všech skupinách. Zatímco u R hovořily ve prospěch alternativy velké hodnoty, pro r to jsou hodnoty malé.

$$\mathbb{I}\left\{R_i^* \geq R_0\right\} = \mathbb{I}\left\{r_i^* \leq r_0\right\},$$

kde r_0 označuje napozorovanou hodnotu r a r_i^* její hodnoty na permutovaných datech. Samotný výraz r lze přepsat jako

$$r = \sum_{k=1}^{g} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij} = \sum_{k=1}^{g} \binom{n_k}{2} \bar{r}_k,$$

ANOSIM 1	ANOSIM 2	ANOSIM 3
$w_k = \frac{1}{g}$	$w_k = \frac{n_k - 1}{n - g}$	$w_k = \frac{\binom{n_k}{2}}{\sum_{i=1}^{g} \binom{n_i}{2}}$

Obrázek 4.1: Váhy jednotlivých variant testu ANOSIM

kde $\bar{r}_k$ je průměrné pořadí vnitroskupinových nepodobností v k-té skupině.

$$\bar{r}_k = \frac{1}{\binom{n_k}{2}} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij}.$$

Jedná se tedy o vážený průměr takovýchto průměrných pořadí, kde vahami jsou počty dvojic prvků v té které skupině. Aby se váhy vyščítaly na 1, lze celý výraz podělit jejich součtem (opět aniž by to mělo vliv na signifikanci testové statistiky). Dojdeme tedy ke tvaru

$$r = \sum_{k=1}^{g} w_k \bar{r}_k, \quad \bar{r}_k = \frac{1}{\binom{n_k}{2}} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij}, \quad w_k = \frac{\binom{n_k}{2}}{\sum_{l=1}^{g} \binom{n_l}{2}}. \tag{4.4}$$

Jak uvidíme dále, má ANOSIM mnoho společného s metodami MRPP (viz kapitola 4.3) a inspirováni těmito metodami, mohli bychom uvážit i použití jiných typů vah. Právě uvedený typ, se nazývá ANOSIM 3 ($w_k = \frac{\binom{n_k}{2}}{\sum_{i=1}^{g} \binom{n_i}{2}}$), dále se používají varianty vah $w_k = \frac{1}{g}$, která všem skupinám přikládá stejnou váhu (ANOSIM 1) a $w_k = \frac{n_k - 1}{n - g}$, kdy jsou váhy proporcionální velikostem skupin (ANOSIM 2). Váhy ANOSIMu 3 nejvíce favorizují velké skupiny, zatímco ANOSIM 1 na velikost skupin vůbec nebere ohled. Jaký to má význam, ukážeme v kapitole věnované simulacím (část 5.1.2, alternativa měřítka). Samozřejmě tyto varianty testu ANOSIM již nutně dávají rozdílné výsledky s výjimkou vyváženého designu, tj. s vyjímkou situací, kdy všechny skupiny mají stejnou velikost. Tehdy všechny tři typy vah splývají. Varianta ANOSIM 3 byla zpočátku používána, kdy Mantel a Valand (1970) chybně předpokládali, že asymptoticky má r (4.4) při použití $\Delta_{Manhattan}$ za platnosti nulové hypotézy normální rozdělení, což bylo následně opraveno (Mielke (1978), Mielke (1979)).

Vraťme se nyní zpět ke statistice ANOSIM, kterou jsme upravili do tvaru

$$r = \sum_{k=1}^{g} w_k \bar{r}_k = \sum_{k=1}^{g} \frac{w_k}{\binom{n_k}{2}} \sum_{i \in G_k} \sum_{j \in G_k, j \neq i} r_{ij}.$$

To odpovídá, po vhodném přeuspořádání, Mantelově testu pro matici pořadí nepodobností a matici vah, která obsahuje ve vnitřněskupinových blocích $\mathbf{W}^{rr}$ upravené hodnoty vah a v meziskupinových blocích $\mathbf{W}^{rs}$ nuly.

$$[W]_{ij} = \begin{cases} \frac{w_k}{\binom{n_k}{2}} & i, j \in G_k, \\ 0 & \text{jinak.} \end{cases} \tag{4.5}$$

Zde je příklad pro 2 skupiny po dvou prvcích:

$$W = \begin{pmatrix} & & & \\ *_1 & & & \\ *_1 & *_1 & & \\ 0 & 0 & *_2 & \\ 0 & 0 & *_2 & *_2 \end{pmatrix},$$

kde $*_k = \frac{w_k}{\binom{n_k}{2}}$. Pro ANOSIM 1 je $*_k = \frac{1}{g\binom{n_k}{2}}$, pro ANOSIM 2 $* = \frac{2}{n_k(n-g)}$ a pro ANOSIM 3 $*_k = \frac{1}{\sum_{k=1}^{g}\binom{n_k}{2}}$. Tedy vlastní výpočet testu ANOSIM lze provést výpočtem Mantelova testu.

4.3 MRPP a jeho vztah k ANOSIMu

Postup metody MRPP (Mielke a kol. (1976), Mielke a Berry (2007), kapitola 2, *multi response permutation procedures*) je inspirován klasickým F testem analýzy rozptylu jednoduchého třídění. Nechť je dáno g nezávislých náhodných výběrů z normálních rozdělení se stejnými konečnými kladnými rozptyly.

$$\begin{aligned} X_{11}, ..., X_{1n_1} &\sim \mathrm{N}(\mu_1, \sigma^2) \\ &\vdots \\ X_{g1}, ..., X_{gn_g} &\sim \mathrm{N}(\mu_g, \sigma^2) \end{aligned}$$

Označme opět $n = \sum_{i=1}^{g} n_i$. F testem testujeme hypotézu

$$\begin{aligned} \mathrm{H}_0\text{:} &\quad \mu_1 = ... = \mu_g, \\ \mathrm{H}_1\text{:} &\quad \text{non } \mathrm{H}_0, \end{aligned}$$

a testovou statistikou tohoto testu je

$$F = \frac{\frac{SS_{Between}}{g-1}}{\frac{SS_{Within}}{n-g}},$$

kde

$$\begin{aligned} SS_{Between} &= \sum_{i=1}^{g} n_i (X_{i\cdot} - X_{\cdot\cdot})^2, \\ SS_{Within} &= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - X_{i\cdot})^2, \end{aligned}$$

$$\begin{aligned} X_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \\ X_{\cdot\cdot} &= \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} X_{ij}. \end{aligned}$$

Za nulové hypotézy pocházejí všechna pozorování X_{ij} ze stejného rozdělení a jsou vzájemně nezávislá, jsou tedy zaměnitelnými objekty. Při výpočtu P-hodnoty pomocí permutací se permutují přiřazení jednotlivých pozorování do skupin. Takto získáme podmíněné rozdělení statistiky F při daných datech. Označíme-li F_0 napozorovanou hodnotu F statistiky, je přesná resp. přibližná P-hodnota rovna

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\{F_i^* \geq F_0\}}{n!} \quad \text{resp.} \quad \frac{\sum_{i=1}^{M} \mathbb{I}\{F_i^* \geq F_0\} + 1}{M+1},$$

kde F_i^* označuje hodnoty F statistiky vypočtené na permutovaných datech a M je počet permutací.

Statistiku F můžeme také vyjádřit následovně:

$$\begin{aligned} F &= \frac{(n-g)SS_{Between}}{(g-1)SS_{Within}} \\ &= \frac{(n-g)SS_{Total} - (n-g)SS_{Within}}{(g-1)SS_{Within}} \\ &= \frac{(n-g)SS_{Total}}{(g-1)SS_{Within}} - \frac{n-g}{g-1}, \end{aligned}$$

kde $SS_{Total} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - X_{\cdot\cdot})^2$. Konstanta $\frac{n-g}{g-1}$ se permutacemi nemění a totéž platí i pro $\frac{n-g}{g-1}SS_{Total}$. Pro výpočet P-hodnot F testu pomocí permutací nám tedy stačí pouze $\frac{1}{SS_{Within}}$. Použijme vztah

$$\sum_{j=1}^{n_i}(X_{ij} - X_{i\cdot})^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1,\, k>j}^{n_i}(X_{ij} - X_{ik})^2.$$

Nyní můžeme psát

$$\begin{aligned} SS_{Within} &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - X_{i\cdot})^2 \\ &= \sum_{i=1}^{g}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1,\, k>j}^{n_i}(X_{ij} - X_{ik})^2. \end{aligned}$$

Označme $\xi_i = \binom{n_i}{2}^{-1}\sum_{k>j}(X_{ij} - X_{ik})^2$ a $C_i = \frac{n_i - 1}{n-g}$. Získáme

$$SS_{Within} = \frac{n-g}{2}\sum_{i=1}^{g} C_i\xi_i,$$

kde $\frac{n-g}{2}$ je opět konstanta bez vlivu na P-hodnotu. Tedy permutační test založený na statistice $\sum_{i=1}^{g} C_i\xi_i$ (kterou označíme δ) je ekvivalentní permutačnímu F testu analýzy rozptylu jednoduchého třídění, protože

$$\mathbb{I}\{F_i^* \geq F_0\} = \mathbb{I}\{\delta_i^* \leq \delta_0\}.$$

Zatímco v případě statistiky F svědčily proti hypotéze její velké hodnoty, pro δ jsou to hodnoty malé. δ samo o sobě je pak váženým průměrem veličin ξ_i s vahami C_i. Veličiny ξ_i mají dvojí význam. Zaprvé je

$$\xi_i = \binom{n_i}{2}^{-1}\sum_{k>j}(X_{ij} - X_{ik})^2 = \frac{2}{n_i(n_i-1)}n_i\sum_{j=1}^{n_i}(X_{ij} - X_{i\cdot})^2 = 2S_i^2,$$

tedy ξ_i jsou veličiny popisující variabilitu v jednotlivých skupinách. Zároveň na ně lze pohlížet jako na součty kvadrátů eukleidovských vzdáleností. ξ_i je pak veličinou popisující průměrnou nepodobnost mezi dvěma prvky stejné skupiny při pužití Δ_{EU}. Úprava do tvaru

$$\delta = \sum_{i=1}^{g} C_i \xi_i$$

nabízí prostor pro použití různých koeficientů nepodobnosti namísto $(X_{ij} - X_{ik})^2$ a tím aplikaci pro vícerozměrná data.

Právě na statistice δ jsou založeny testy MRPP, které jsou stejně jako ANOSIM testy shody rozdělení. Označme stejně jako v části 4.1 $F_1, ..., F_g$ distribuční funkce nějakých p-rozměrných rozdělení a nechť je dáno g nezávislých náhodných výběrů z těchto distribučních funkcí.

$$\begin{aligned} \mathbf{X}_{11}, ..., \mathbf{X}_{1n_1} &\sim F_1, \\ &\vdots \\ \mathbf{X}_{g1}, ..., \mathbf{X}_{gn_g} &\sim F_g. \end{aligned}$$

Označme dále $n = \sum_{i=1}^{g} n_i$. Testujeme hypotézu

$$\begin{aligned} &\mathrm{H}_0\text{:} \quad F_1 \equiv ... \equiv F_g, \\ &\mathrm{H}_1\text{:} \quad \text{non } \mathrm{H}_0. \end{aligned}$$

Testovou statistikou je

$$\delta = \sum_{i=1}^{g} C_i \xi_i,$$

kde

$$\begin{aligned} \sum_{i=1}^{g} C_i &= 1, \\ C_i &> 0, \quad i = 1, ..., g, \\ \xi_i &= \binom{n_i}{2}^{-1} \sum_{k>j} \Delta(\mathbf{X}_{ij}, \mathbf{X}_{ik}), \end{aligned}$$

kde Δ označuje koeficient nepodobnosti a C_i nějaké váhy.

Jak již bylo řečeno, je ξ_i průměr vnitroskupinových nepodobností v i-té skupině, přeznačme jej proto pro větší názornost na $\bar{\Delta}_i$. Porovnáme-li tvar testové statistiky MRPP δ s testovou statistikou testu ANOSIM r

$$\begin{aligned} &\text{MRPP} && \text{ANOSIM} \\ &\delta = \sum_{i=1}^{g} C_i \bar{\Delta}_i, && r = \sum_{k=1}^{g} w_k \bar{r}_k, \end{aligned}$$

MRPP 1	MRPP 2	MRPP 3
$C_i = \frac{1}{g}$	$C_i = \frac{n_i - 1}{n - g}$	$C_i = \frac{\binom{n_i}{2}}{\sum_{i=1}^{g} \binom{n_i}{2}}$

Obrázek 4.2: Váhy jednotlivých variant testu MRPP

zjistíme, že MRPP je totéž, co ANOSIM (konkrétní varianta závisí na zvolených vahách) provedený na původní vzdálenosti. Tím pádem stejně jako v případě ANOSIMu, lze i MRPP vyjádřit v řeči Mantelova testu pro matici vzdáleností a matici vah W (viz (4.5)). Při zjišťování signifikance pak opět permutujeme přiřazení prvků do skupin (současně permutujeme řádky a sloupce matice vah, resp. matice vzdáleností).

Vlastní volba vah a koeficientu nepodobnosti pak specifikuje konkrétní variantu MRPP. O volbě koeficientu nepodobnosti bylo již pojednáno dříve, pro volbu vah navrhují Mielke a Berry (2007) (kapitola 2.3) následující možnosti: $C_i = \frac{1}{g}$, tato metoda se nazývá MRPP 1, varianta s vahami $C_i = \frac{n_i - 1}{n - g}$ se nazývá MRPP 2 a varianta s vahami $C_i = \frac{\binom{n_i}{2}}{\sum_{i=1}^{g} \binom{n_i}{2}}$ se nazývá MRPP 3. Jedná se o tytéž váhy, jaké jsme zavedli pro varianty testu ANOSIM. Pokud jsou vektory $\mathbf{X}_{ij}$ jednorozměrné a jako koeficient nepodobnosti použijeme eukleidovskou vzdálenost, je varianta MRPP 2 ekvivalentní permutačnímu F testu v analýze jednoduchého třídění. Mielke a Berry (2007) ve své knize poukazují, že právě varianta MRPP 2 má nejvýhodnější vlastnosti.

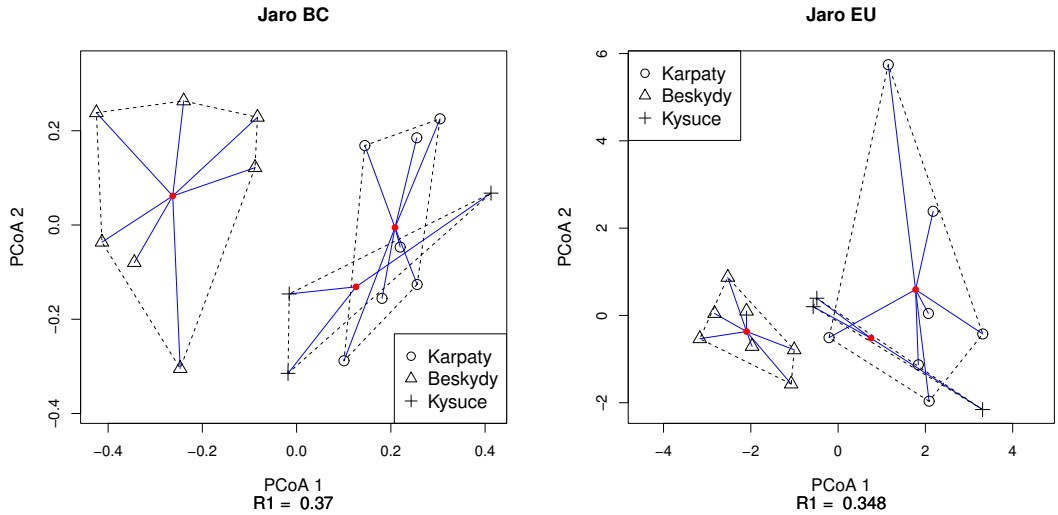
V Mielke a Berry (2007), kapitola 2.1, je uvedena definice MRPP s $g + 1$ skupinami, kde poslední skupina obsahuje nezařazené prvky. V mnoha situacích - i v situaci jednoduchého třídění tak, jak jsme jej popsali zde - je $g + 1$ skupina prázdná. Na závěr ještě poznamenejme, že nezávisle na autorech MRPP byly vyvinuty testy pro prostorovou autokorelaci (Friedman a Rafsky (1979), Cliff a Ord (1973)), které jsou speciálními případy MRPP při vhodné volbě vah a koeficientu nepodobnosti (Whaley (1983)).

Stejně jako u ANOSIMu i v případě MRPP se na základě testu založeného na nepodobnostech mezi jednotlivými pozorováními dělají závěry o původních pozorováních a zamítnutí nulové hypotézy shody rozdělení se interpretuje jako rozdíl v parametru polohy. To s sebou nese rizika, o kterých jsme již hovořili v části věnované testu ANOSIM.

Testy ANOSIM a MRPP nejsou jediné testy založené na nepodobnostech, které lze použít pro problém jednoduchého třídění. Z dalších postupů zmiňme například metody NPMANOVA (*nonparametric multivariate analysis of variance* v Anderson (2001)), kdy pro jednoduché třídění je NPMANOVA ekvivalentní MRPP 2, dále AMOVA (*analysis of molecular variance*) v Excoffier a kol. (1992), *analysis of distance* popsaná v Gower a Krzanowski (1999), nebo statistiky navržené v Pillar a Orlóci (1996) nebo Smith a kol. (1990). Z klasických postupů zmiňme metodu MANOVA (*multivariate analysis of variance*, Rencher (2002), kapitola 6.1.3). Tato metoda kromě mnohorozměrné normality vyžaduje, aby počet pozorování v každé skupině byl větší, než rozměr pozorování (např. data `mouchy` jsou 156-rozměrná, ale pozorování je k dispozici pro každou skupinu nejvýše 7). Warton a Hudson (2004) provedli rozsáhlé simulační studie, kde porovnávali vlastnosti metod založených na nepodobnostech s metodami založený-

mi přímo na (transformovaných) datech (např. F-statistiky z Edgington (1995), str. 188-190 nebo LR-IND statistika v Warton a Hudson (2004)). Upozorňují, že výsledky těchto metod se lépe interpretují než výsledky metod založených na nepodobnostech, protože nebývá zřejmé, co přesně se daným testem založeným na nepodobnostech testuje. Na simulacích rovněž ukazují, že metody založené na nepodobnostech nemají, co se síly testů týče, žádné zjevné výhody v porovnání s těmito metodami.

Ilustrujme nyní varianty testů ANOSIM a MRPP na datech `mouchy`, konkrétně na četnostech jednotlivých druhů. Tato data jsou rozdělena do tří skupin - oblastí Beskydy (7 stanovišť), Karpaty (7 stanovišť) a Kysuce (3 stanoviště). Abychom se vyhnuli problémům s nezávislostí pozorování, nahradili jsme vždy po dvojicích pozorování naměřená ve stejné lokalitě ve stejnou dobu jejich průměrem a s každým ročním obdobím (jaro, léto a podzim) jsme pracovali zvlášť. Nyní lze předpokládat, že pozorování z různých stanovišť v rámci oblasti a pozorování z různých oblastí jsou nezávislá. Na vytvořené průměrné druhové četnosti jsme použili transformaci $\sqrt[4]{x}$. Jako koeficienty nepodobností Δ jsme použili postupně Bray-Curtisův koeficient nepodobnosti (BC) a eukleidovskou vzdálenost (EU).

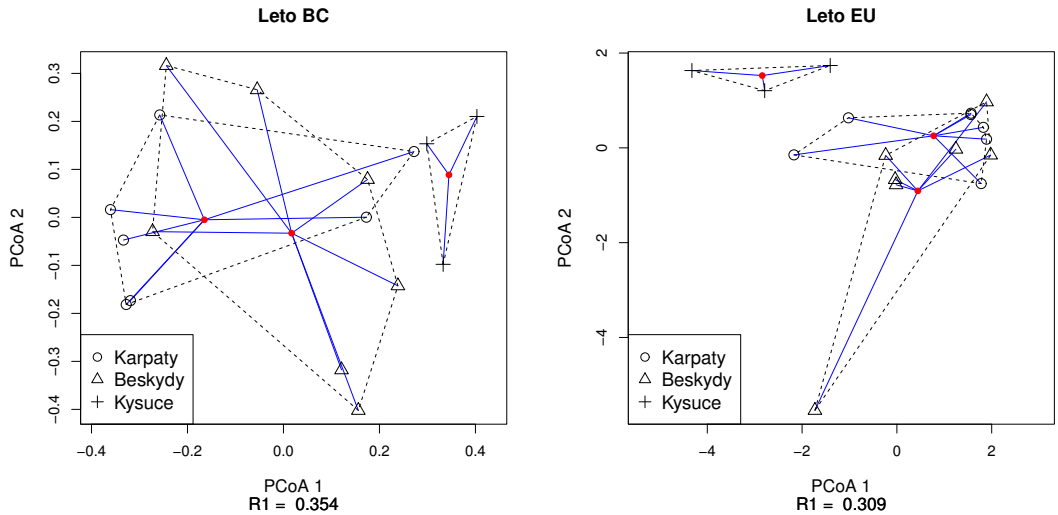


	BC	EU		BC	EU
ANOSIM 1	0,039	0,038	MRPP 1	0,033	0,056
ANOSIM 2	0,001	0,002	MRPP 2	0,001	0,002
ANOSIM 3	0,001	0,002	MRPP 3	0,001	0,004

Obrázek 4.3: Reprezentace druhových skladeb jarních měření pomocí PCoA při použití Δ_{EU} a Δ_{BC}, P-hodnoty variant testů ANOSIM a MRPP.

Výpočet byl opět proveden v programu `R`. V knihovně `vegan` lze nalézt funkci `mrpp`, která počítá všechny tři zmíněné varianty testu MRPP a funkci `anosim`, která však počítá pouze test ANOSIM 3. Proto jsme vytvořili vlastní funkci pro současný výpočet všech variant těchto testů, která navíc umožnila použití stejných permutací pro jednotlivé varianty obou testů. Počet permutací pro výpočet P-hodnot testů byl 999. Tabulky v obrázcích 4.3, 4.4 a 4.5 uvádějí P-hodnoty

jednotlivých variant testů pro obě použité nepodobnosti. Přidány jsou grafické reprezentace druhové skladby stanovišť pomocí prvních dvou os získaných metodou hlavních koordinát, které zachytily přibližně 35 % variablity. Pro každou skupinu je zakreslen také vypočtený centroid vzhledem k použitému koeficientu nepodobnosti (viz test homogenity disperzí (4.6)) a pro snazší orientaci jsou pozorování stejných skupin graficky propojena.

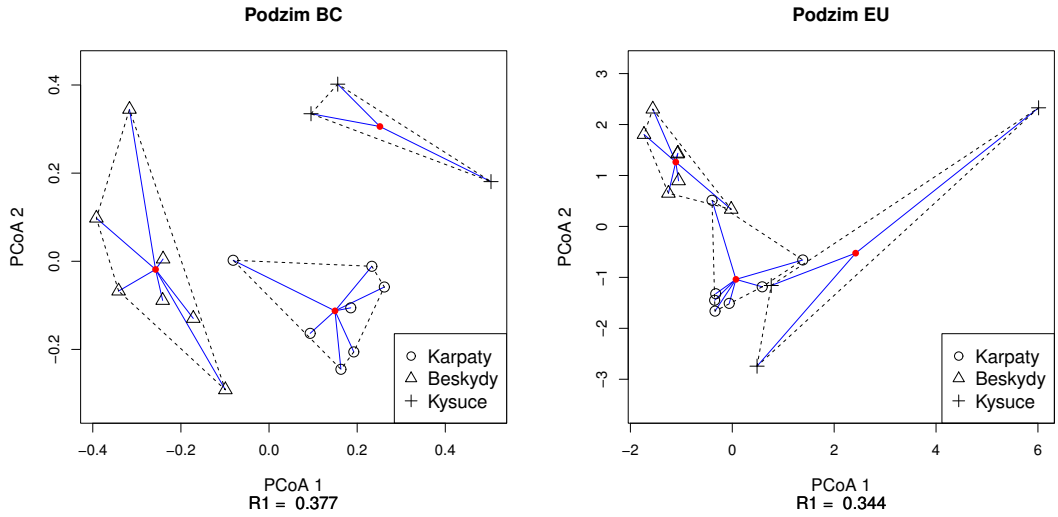


	BC	EU		BC	EU
ANOSIM 1	0,001	0,257	MRPP 1	0,001	0,117
ANOSIM 2	0,008	0,011	MRPP 2	0,002	0,001
ANOSIM 3	0,013	0,029	MRPP 3	0,007	0,009

Obrázek 4.4: Reprezentace druhových skladeb letních měření pomocí PCoA při použití Δ_{EU} a Δ_{BC}, P-hodnoty variant testů ANOSIM a MRPP.

Pro jarní měření (obrázek 4.3) jsme až na jeden případ vždy zamítli nulovou hypotézu, P-hodnoty příslušných variant testů ANOSIM a MRPP jsou velice podobné. Výjimkou byl test ANOSIM 1 při použití eukleidovské vzdálenosti s P-hodnotou 0,056. Pro letní měření (obrázek 4.4) testy ANOSIM 1 a MRPP 1 při použití eukleidovské vzdálenosti nezamítají shodu rozdělení (P-hodnoty 0,257 a 0,117). Tyto hodnoty se také výrazně liší od P-hodnot ostatních testů. Odpovídající si testy ANOSIM a MRPP opět mají podobné P-hodnoty. Pro podzimní měření (obrázek 4.5) opět testy ANOSIM 1 a MRPP 1 při použití eukleidovské vzdálenosti nezamítají shodu rozdělení (P-hodnoty 0,069 a 0,460). Výrazně odlišná P-hodnota 0,460 je pravděpodobně způsobena větším rozptýlením skupiny Kysuce, které snižuje sílu varianty 1 (viz část 4.4 a kapitola 5). Více je ovlivněn test MRPP, který pracuje přímo s hodnotami nepodobností. I zde mají odpovídající si testy ANOSIM a MRPP podobné P-hodnoty s výjimkou právě uvedených ANOSIM 1 a MRPP 1 pro eukleidovskou vzdálenost. Pokud se omezíme pouze na Bray-Curtisův koeficient nepodobnosti, zamítli jsme pro všechna roční období hypotézu stejného rozdělení pozorování.

Testy ANOSIM a MRPP jsme použili i pro párová porovnávání. Protože se jedná o mnohonásobná porovnávání, použili jsme korekci pomocí Bonferoniho nerovnosti, za signifikantní považujeme jen P-hodnoty menší než 0,0167. Pro tyto



	BC	EU		BC	EU
ANOSIM 1	0,015	0,069	MRPP 1	0,019	0,460
ANOSIM 2	0,001	0,001	MRPP 2	0,001	0,001
ANOSIM 3	0,001	0,001	MRPP 3	0,001	0,001

Obrázek 4.5: Reprezentace druhových skladeb podzimních měření pomocí PCoA při použití Δ_{EU} a Δ_{BC}, P-hodnoty variant testů ANOSIM a MRPP.

testy jsme počet permutací zvýšili na 9999. V tabulce 4.1 jsou uvedeny P-hodnoty jednotlivých testů při použití Bray-Curtisova koeficientu nepodobnosti. Pro jarní pozorování se významně liší Karpaty a Beskydy a to pro všechny varianty testů (jedná se o vyvážený design). Pro letní pozorování se významně liší Karpaty a Kysuce pro varianty testů 1, naopak P-hodnoty variant 3 jsou 0,24 resp. 0,18. Tento výsledek je zřejmě opět způsoben rozdílnou disperzí těchto dvou skupin (viz část 4.4 a kapitola 5). Varianty 2 a 3 pak zamítají shodu rozdělení pozorování pocházejících z Beskyd a Kysuc. Pro podzimní pozorování všechny varianty testů zamítnou shodu rozdělení pozorování pocházejících z Karpat a Beskyd. Pro ostatní dvojice je nulová hypotéza zamítnuta vždy variantami 2 a 3. Důvodem opět může být snížená síla varianty 1, protože skupina Kysuce má větší disperzi (viz kapitola 4.4)

4.4 Test homogenity disperzí

Jak již bylo zmíněno, jak ANOSIM tak MRPP jsou testy shody rozdělení několika skupin, které se pak často interpretují jako případný rozdíl v poloze. Pokud se rozdělení liší nějakým jiným způsobem a tyto testy zamítnou nulovou hypotézu, je pak tato interpretace chybná. Jedním z takových rozdílů je „rozptýlení“ pozorování v jednotlivých skupinách. Anderson (2006) inspirována Leveneovým testem (Levene (1960)) navrhla test homogenity disperzí.

Nechť $\mathbf{X}_{ij}$, $i = 1, ..., g$, $j = 1, ..., n_i$ jsou nezávislé náhodné výběry o velikostech n_i z rozdělení s distribuční funkcí F_i stejně jako v (4.1), Δ koeficient nepodobnosti a $\mathbf{D}$ matice vzdáleností. Rozdělení nepodobností mezi prvky ve stejné skupině $\mathcal{L}(\Delta(X_{ij}, X_{ij'}))$, $i = 1, ..., g$, $j, j' = 1, ..., n_i$ jsou určena rozděleními F_i a použitým

Jaro	ANOSIM 1	ANOSIM 2	ANOSIM 3
Karpaty vs. Beskydy	0,0008	0,0008	0,0008
Karpaty vs. Kysuce	0,2054	0,0288	0,0426
Beskydy vs. Kysuce	0,8522	0,2480	0,0760
Léto	ANOSIM 1	ANOSIM 2	ANOSIM 3
Karpaty vs. Beskydy	0,0789	0,0789	0,0789
Karpaty vs. Kysuce	0,0112	0,0299	0,2381
Beskydy vs. Kysuce	0,0221	0,0062	0,0062
Podzim	ANOSIM 1	ANOSIM 2	ANOSIM 3
Karpaty vs. Beskydy	0,0008	0,0008	0,0008
Karpaty vs. Kysuce	0,1637	0,0112	0,0112
Beskydy vs. Kysuce	0,3576	0,0062	0,0062
Jaro	MRPP 1	MRPP 2	MRPP 3
Karpaty vs. Beskydy	0,0008	0,0008	0,0008
Karpaty vs. Kysuce	0,1590	0,0208	0,0356
Beskydy vs. Kysuce	0,8689	0,2055	0,0587
Léto	MRPP 1	MRPP 2	MRPP 3
Karpaty vs. Beskydy	0,0513	0,0513	0,0513
Karpaty vs. Kysuce	0,0112	0,0208	0,1778
Beskydy vs. Kysuce	0,0221	0,0062	0,0062
Podzim	MRPP 1	MRPP 2	MRPP 3
Karpaty vs. Beskydy	0,0008	0,0008	0,0008
Karpaty vs. Kysuce	0,1788	0,0112	0,0112
Beskydy vs. Kysuce	0,3024	0,0062	0,0062

Tabulka 4.1: Párová porovnávání

koeficientem nepodobnosti. Označme tato rozdělení $\mathcal{L}(\Delta_i)$, $i = 1, ..., g$. Hypotéza, která se testuje testem homogenity disperzí, je

$$\begin{aligned} \mathrm{H}_0: &\quad \mathcal{L}(\Delta_1) = ... = \mathcal{L}(\Delta_g), \\ \mathrm{H}_1: &\quad \text{non } \mathrm{H}_0. \end{aligned}$$

Myšlenka tohoto testu je: Pokud jednotlivé skupiny pocházejí až na parametr polohy ze stejných rozdělení, jsou rozdělení vnitroskupinových nepodobností ve všech skupinách stejné. Tato implikace ovšem platí pouze pro koeficienty nepodobnosti Δ invariantní na posunutí jako je například eukleidovská vzdálenost. O problému tvoření závěrů o původním rozdělení na základě nepodobností mezi prvky jsme již pojednali v části věnované testu ANOSIM. Pokud test nezamítne nulovou hyptézu, interpretujeme výsledek tak, že „skupiny pozorování pocházejí až na parametr polohy ze stejných rozdělení".

Na základě Leveneova testu navrhla Anderson (2006) jeho rozšíření pomocí metody hlavních koordinát (viz kapitola 2.2) pro libovolný koeficient nepodobnosti - test homogenity disperzí. Pojem disperze označuje střední hodnotu nepodobností mezi prvky dané skupiny a nějakým jejím středem, přičemž předpokládáme, že tato střední hodnota existuje.

V případě jednoduchého třídění v jednorozměrném případě vypadá Leveneův test tak, že se provede ANOVA na absolutní odchylky jednotlivých pozorování ve skupině od příslušného průměru $\bar{X}_i$ (Levene (1960)) resp. mediánu $\hat{X}_i$ (Brown a Forsythe (1974)).

$$\begin{aligned} Z_{ij}^c &= \left| X_{ij} - \bar{X}_i \right|, \\ Z_{ij}^m &= \left| X_{ij} - \hat{X}_i \right|, \\ i = 1, ..., g, &\qquad j = 1, ..., n_i, \end{aligned}$$

Zároveň jsou Z_{ij}^c a Z_{ij}^m eukleidovskými vzdálenostmi X_{ij} od příslušného středu skupiny, tedy

$$\begin{aligned} Z_{ij}^c &= \Delta_{EU}(X_{ij}, \bar{X}_i), \\ Z_{ij}^m &= \Delta_{EU}(X_{ij}, \hat{X}_i), \end{aligned}$$

kde Δ_{EU} označuje eukleidovskou vzdálenost. Přirozeně se nabízí rozšíření tohoto testu pro eukleidovskou vzdálenost ve více dimenzích (Van Valen (1978), O'Brien (1992)) popřípadě použití jiných koeficientů nepodobnosti (Anderson (2006)), kdy nahradíme eukleidovskou vzdálenost obecným koeficientem nepodobnosti Δ. Problémem je, jak nalézt středy jednotlivých skupin vzhledem k použitému koeficientu nepodobnosti. K tomu využijeme metodu hlavních koordinát. Pomocí této metody - pokud nedojde ke vzniku záporných vlastních čísel - získáme z matice vzdáleností $\mathbf{D}$ reprezentaci jednotlivých pozorování v eukleidovském prostoru nějaké dimenze p.

Označme $\mathbf{Y}_{ij}$ souřadnice (koordináty) reprezentace bodu $\mathbf{X}_{ij}$. Pak

$$\Delta_{EU}(\mathbf{Y}_{ij}, \mathbf{Y}_{i'j'}) = \Delta(\mathbf{X}_{ij}, \mathbf{X}_{i'j'}).$$

Označme dále $\tilde{\boldsymbol{C}}_i$ centroid i-té skupiny reprezentací

$$\tilde{\boldsymbol{C}}_i = \arg\min_{\boldsymbol{\theta} \in R^p} \sum_{j=1}^{n_i} \Delta_{EU}^2(\mathbf{Y}_{ij}, \boldsymbol{\theta}), \tag{4.6}$$

což odpovídá průměrům po složkách $\tilde{\boldsymbol{C}}_i = \frac{1}{n_i}\sum_{l=1}^{n_i}\mathbf{Y}_{ij}$ a dále $\tilde{\boldsymbol{M}}_i$ medián i-té skupiny reprezentací (*geometric median*, v Gower (1974) označený jako *median-centre*). Tento však nemusí být dán jednoznačně.

$$\tilde{\boldsymbol{M}}_i = \arg\min_{\boldsymbol{\theta}\in R^p}\sum_{j=1}^{n_i}\Delta_{EU}(\mathbf{Y}_{ij},\boldsymbol{\theta}).$$

S těmito body budeme pracovat jako s reprezentacemi středů skupin (označme je $\mathbf{C}_i$ a $\mathbf{M}_i$). Jejich eukleidovské vzdálenosti od jednotlivých reprezentací bodů budou určovat příslušné nepodobnosti mezi středy skupin a jednotlivými pozorováními.

$$\begin{aligned}\Delta(\mathbf{X}_{ij},\mathbf{C}_i) &:= \Delta_{EU}(\mathbf{Y}_{ij},\tilde{\boldsymbol{C}}_i),\\ \Delta(\mathbf{X}_{ij},\mathbf{M}_i) &:= \Delta_{EU}(\mathbf{Y}_{ij},\tilde{\boldsymbol{M}}_i).\end{aligned}$$

Máme tedy

$$\begin{aligned}Z^c_{ij} &= \Delta_{EU}(\mathbf{Y}_{ij},\tilde{\boldsymbol{C}}_i),\\ Z^m_{ij} &= \Delta_{EU}(\mathbf{Y}_{ij},\tilde{\boldsymbol{M}}_i)\end{aligned}$$

a namísto nepodobností mezi prvky samotnými nyní pracujeme s nepodobnostmi prvků a (vypočítaných) středů skupin.

V případě, že se při provádění PCoA záporná vlastní čísla vyskytla, jsou jejich příslušné osy imaginární, tedy reprezentace v reálném eukleidovském prostoru R^p není přesná. S využitím vlastnosti $i^2 = -1$ ale můžeme spočítat eukleidovské vzdálenosti v úplné reprezentaci (viz část 2.2). Označme $\mathbf{Y}^+_{ij}$ a $\mathbf{Y}^-_{ij}$ souřadnice reprezentace bodu $\mathbf{X}_{ij}$ vzhledem k reálným a imaginárním osám odpovídající kladným a záporným vlastním číslům. Pak se i souřadnice centroidu a mediánu rozpadají na dvě části: $\tilde{\boldsymbol{C}}^+_i$ a $\tilde{\boldsymbol{C}}^-_i$ resp. $\tilde{\boldsymbol{M}}^+_i$ a $\tilde{\boldsymbol{M}}^-_i$. Platí, že nepodobnost mezi body $\mathbf{X}_{ij}$ a $\mathbf{X}_{i'j'}$ je

$$\Delta(\mathbf{X}_{ij},\mathbf{X}_{i'j'}) = \sqrt{\Delta^2_{EU}(\mathbf{Y}^+_{ij},\mathbf{Y}^+_{i'j'}) - \Delta^2_{EU}(\mathbf{Y}^-_{ij},\mathbf{Y}^-_{i'j'})}.$$

Stejným způsobem určíme i nepodobnosti bodů a jednotlivých středů skupin.

$$\begin{aligned}Z^c_{ij} &= \sqrt{\Delta^2_{EU}(\mathbf{Y}^+_{ij},\tilde{\boldsymbol{C}}^+_i) - \Delta^2_{EU}(\mathbf{Y}^-_{ij},\tilde{\boldsymbol{C}}^-_i)},\\ Z^m_{ij} &= \sqrt{\Delta^2_{EU}(\mathbf{Y}^+_{ij},\tilde{\boldsymbol{M}}^+_i) - \Delta^2_{EU}(\mathbf{Y}^-_{ij},\tilde{\boldsymbol{M}}^-_i)}.\end{aligned}$$

Poznamenejme, že tyto kroky lze provést i pokud nemáme k dispozici vlastní pozorování $\mathbf{X}_{ij}$, ale pouze matici vzdáleností $\mathbf{D}$.

Vlastní test homogenity disperzí pak probíhá stejně jako Leveneův test. Provede se ANOVA na vzdálenosti jednotlivých pozorování od středů příslušných skupin Z^c_{ij} nebo Z^m_{ij}. F statistika pro případ volby centroidů pak vypadá následovně:

$$F^c = \frac{MS_{Between}}{MS_{Within}}, \tag{4.7}$$

kde

$$\begin{aligned} MS_{Between} &= \frac{1}{g-1}\sum_{i=1}^{g} n_i (Z^c_{i\cdot} - Z^c_{\cdot\cdot})^2, \\ MS_{Within} &= \frac{1}{N-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i} (Z^c_{ij} - Z^c_{i\cdot})^2, \\ Z^c_{i\cdot} &= \frac{1}{n_i}\sum_{j=1}^{n_i} Z^c_{ij}, \\ Z^c_{\cdot\cdot} &= \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} Z^c_{ij}. \end{aligned}$$

P-hodnotu tohoto testu lze získat dvěma způsoby. Můžeme využít permutací. Problémem je, že za platnosti nulové hypotézy sice pocházejí všechny vnitroskupinové nepodobnosti $\Delta(\mathbf{X}_{ij}, \mathbf{X}_{ij'})$, $i = 1, ..., g$, $j, j' = 1, ..., n_i$ ze stejných rozdělení, ale pro nepodobnosti jednotlivých pozorování a středů skupin tomu již tak není. Rozdělení $\tilde{\boldsymbol{C}}_i$ a $\tilde{\boldsymbol{M}}_i$ jsou totiž ovlivněna také velikostmi skupin, které jsou obecně různé. Tedy Z^c_{ij} a Z^m_{ij} nemají za nulové hypotézy stejná rozdělení a pokud jsou rozdíly ve velikostech skupin velké, může dojít k nedodržení předepsané hladiny testu (viz část 5.2.3). Přesto se k určení P-hodnot permutace používají, ovšem tento test je již pouze přibližný (viz kapitola 5). Jeho P-hodnoty se pak počítají

$$\frac{\sum_{i=1}^{n!} \mathbb{I}\left\{(F^c)^*_i \leq F^c_0\right\}}{n!} \quad \text{resp.} \quad \frac{\sum_{i=1}^{M} \mathbb{I}\left\{(F^c)^*_i \leq F^c_0\right\} + 1}{M+1},$$

kde F^c_0 je napozorovaná hodnota statistiky F^c, $(F^c)^*_i$ označuje hodnoty statistiky vypotečné na permutovaných datech a M je počet permutací. Druhou variantou je užití klasického F rozdělení (Anderson (2006)). Přestože nelze předpokládat, že $\mathcal{L}(\Delta_i)$ nebo $\mathcal{L}(Z_{ij})$ jsou normální rozdělení (nepodobnosti jsou pouze nezáporné), a jednotlivé Z^c_{ij} a Z^m_{ij} nejsou nezávislé, dává tato varianta velmi podobné výsledky jako permutační test. Anderson (2006) dále poznamenává, že varianta s centroidy je anti-konzervativní, to znamená, že nedodržuje předepsanou hladinu testu a její hladina je vyšší. Obojí se ukázalo i na našich simulacích (viz kapitola 5).

Připomeňme na závěr, že i tento test je testem shody rozdělení vnitroskupinových nepodobností $\mathcal{L}(\Delta_i)$ na základě kterého pak usuzujeme zda jsou původní pozorování stejně „rozptýlená“. Protože však pracujeme s mnohorozměrnými pozorováními není toto rozptýlení jediný způsob, jak se mohou rozdělení rozdělení lišit. Jako příklad uveďme situaci náhodných vektorů $\mathbf{X} = (X_1, X_2)$ a $\mathbf{Y} = (Y_1, Y_2)$ s dvojrozměrným normálním rozdělením.

$$\begin{aligned} \mathbf{X} &\sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}9 \\ 0{,}9 & 1 \end{pmatrix}\right), \\ \mathbf{Y} &\sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0{,}9 \\ -0{,}9 & 1 \end{pmatrix}\right). \end{aligned}$$

Mějme stejný počet pozorování obou náhodných vektorů a jako koeficient nepodobnosti použijme eukleidovskou vzdálenost. Ta je invariantní na otočení. Protože $\mathcal{L}(X_1, X_2) = \mathcal{L}(Y_2, Y_1)$, je rozdělení nepodobností pozorování od středů skupin

v obou skupinách stejné a test homogenity disperzí tento rozdíl neodhalí. Testy ANOSIM a MRPP jsou však na tento rozdíl v rozdělení citlivé, což následně může vést k chybným závěrům, že se skupiny liší polohou (viz kapitola 5.1.2, alternativa tvaru).

Závěrem, při použití testu homogenity disperzí je třeba si uvědomit následující:

- Pracujeme s reprezentací bodů a pokud nepoužijeme úplnou reprezentaci, nemusí být tato vždy dostatečně kvalitní.
- Namísto nepodobností mezi jednotlivými pozorováními pracujeme s nepodobnostmi od vypočteného středu skupiny. Pokud mají skupiny rozdílnou velikost, nejsou tyto nepodobnosti zaměnitelné za nulové hypotézy. Jak se ukáže v kapitole věnované simulacím (část 5.2.3), má nesplnění tohoto požadavku výrazný vliv na chování tohoto testu
- Mnoho koeficientů nepodobnosti není invariantní na posunutí (např. Δ_{BC}).

Z těchto důvodů musíme být při používání tohoto testu obezřetní.

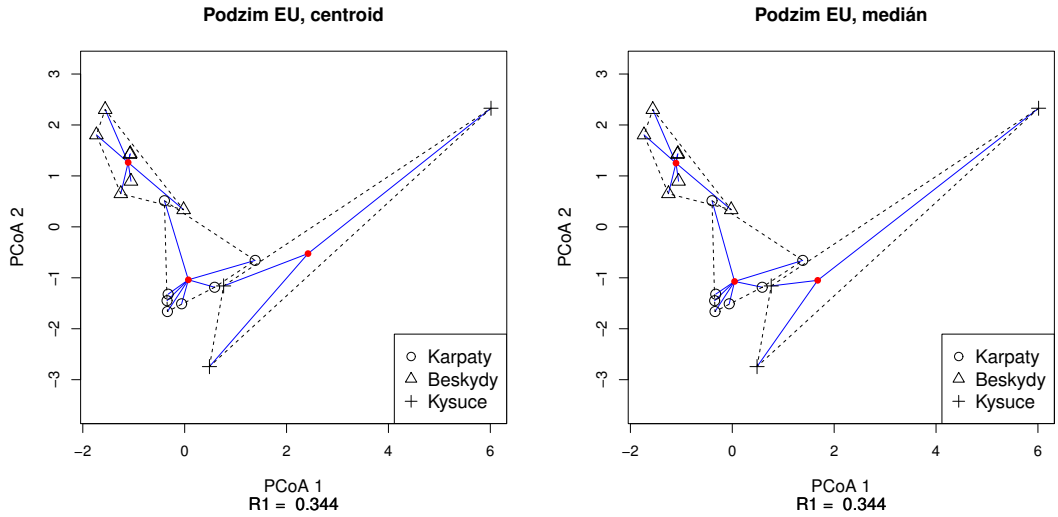
Vraťme se nyní k datům `mouchy`. Pro jednotlivá roční období ověříme homogenitu disperzí skupin. Data jsou tatáž jako v případě testů ANOSIM a MRPP a výpočet jsme opět provedli v programu `R`. Použili jsme funkce `anova`, `betadisper` a `permutest` z knihoven `vegan` a `stats`. V tabulce 4.2 jsou uvedeny P-hodnoty jednotlivých variant testů homogenity disperzí skupin pro jednotlivá roční období. Vidíme, že při použití Bray-Curtisova koeficientu nepodobnosti za-

Skupina	Nepodobnost	FC	PC	FM	PM
Jaro	BC	0,231	0,213	0,374	0,379
Jaro	EU	0,177	0,170	0,351	0,358
Léto	BC	0,010	0,007	0,066	0,060
Léto	EU	0,417	0,407	0,596	0,564
Podzim	BC	0,578	0,586	0,734	0,725
Podzim	EU	0,033	0,036	0,309	0,338

Tabulka 4.2: P-hodnoty testů homogenity disperzí skupin pro jednotlivá roční období. BC - Bray-Curtisova nepodobnost, EU - eukleidovská vzdálenost, varianty středů skupin: C - centroid, M - medián, varianty určení P-hodnoty: F - F rozdělení, P - 999 permutací.

mítáme shodu rozdělení nepodobností pro letní data a to pro varianty testu s centroidem (P-hodnoty 0,010, 0,007). Pro variantu s mediánem však již nulovou hypotézu nezamítáme (P-hodnoty 0,066 a 0,060). Pro podzimní data a eukleidovskou vzdálenost při použití centroidů jsou P-hodnoty 0,033 a 0,036, zatímco při užití mediánů 0,309 a 0,338. Tento výrazný rozdíl je pravděpodobně způsoben tím, že zatímco pro Beskydy a Karpaty se pozice centroidu a mediánu výrazně neliší, ve skupině Kysuce, která obsahuje pouze tři pozorování, z nichž jedno je velmi vzdálené od ostatních, je rozdíl výrazný (viz obrázek 4.6). Vypočtené vzdálenosti ke středu skupiny se pak ve skupině Kysuce při použití centroidu a mediánu výrazně liší (viz tabulka tamtéž). Totéž v menší míře platí i pro letní pozorování při použití eukleidovské vzdálenosti a skupinu Beskydy (4.4). P-hodnoty

jednotlivých variant testu homogenity disperzí se pak velmi neliší, používáme-li pro jejich určení F rozdělení nebo permutace. Pokud se rozhodneme pro variantu s mediánem a Bray-Curtisovým koeficientem nepodobnosti, nezamítáme ani v jednom ročním období nulovou hypotézu a výsledek tohoto testu intepretujeme tak, že všechny skupiny pocházejí až na parametr polohy ze stejného rozdělení.



Karpaty							
centroid	2,99	3,49	3,17	2,56	2,97	3,26	3,06
medián	3,00	3,57	3,12	2,45	2,96	3,31	3,06
Beskydy							
centroid	2,59	3,36	3,66	2,94	3,46	3,66	3,73
medián	2,45	3,35	3,71	2,89	3,44	3,74	3,79
Kysuce							
centroid	3,74	3,35	4,99				
medián	3,27	2,56	6,02				

Obrázek 4.6: Grafická reprezentace druhové skladby stanovišť pro podzimních měření s vypočtenými středy skupin. Vzdálenosti k vypočteným středům skupin pro podzimní pozorování při použití eukleidovské vzdálenosti.

5. Simulace

Obsahem následující kapitoly je prezentace výsledků simulací týkajících se variant testů ANOSIM, MRPP a testu homogenity disperzí popsaných v kapitole 4. Zabývali jsme se tím, jaký vliv na tyto testy mají konkrétní rozdíly v rozděleních, pokud data pocházejí z dvojrozměrného normálního rozdělení při použití eukleidovské vzdálenosti jako koeficientu nepodobnosti. Posléze jsme podobně jako Anderson (2006) použili Monte Carlo simulace pro porovnání síly a chyby 1. druhu těchto testů na datech pocházejících z mnohorozměrného Poisson-log normálního rozdělení.

Připomeňme, že testové statistiky testů ANOSIM a MRPP mají tvar:

$$\begin{array}{cc} \text{MRPP} & \text{ANOSIM} \\ \delta = \sum_{i=1}^{g} C_i \, \bar{\Delta}_i, & r = \sum_{k=1}^{g} w_i \, \bar{r}_i, \end{array}$$

kde $\bar{\Delta}_i$ resp. $\bar{r}_i$, $i = 1, ..., g$ označují průměrnou hodnotu resp. průměrné pořadí vnitroskupinových nepodobností v i-té skupině a C_i resp. w_i označují váhy přiřazené jednotlivým skupinám. Podle typu vah pak rozlišujeme tři varianty těchto testů:

MRPP 1/ANOSIM 1	MRPP 2/ANOSIM 2	MRPP 3/ANOSIM 3
$C_i = w_i = \frac{1}{g}$	$C_i = w_i = \frac{n_i - 1}{n - g}$	$C_i = w_i = \frac{\binom{n_i}{2}}{\sum_{l=1}^{g} \binom{n_l}{2}}$

Testy ANOSIM a MRPP testujeme hypotézu shody rozdělení g skupin pozorování s distribučními funkcemi F_i, $i = 1, ..., g$

$$\begin{array}{ll} \mathrm{H}_0: & F_1 = ... = F_g, \\ \mathrm{H}_1: & \text{non } \mathrm{H}_0. \end{array}$$

a P-hodnoty těchto testů se získávají permutacemi původních pozorování mezi skupinami. Zamítnutí hypotézy se pak často interpretuje jako rozdíl v poloze.

Test homogenity disperzí je obdobou Leveneova testu. Testujeme hypotézu

$$\begin{array}{ll} \mathrm{H}_0: & \mathcal{L}(\Delta_1) = ... = \mathcal{L}(\Delta_g), \\ \mathrm{H}_1: & \text{non } \mathrm{H}_0. \end{array}$$

pomocí statistiky F^c nebo F^m (4.7), podle zvolené varianty testu. P-hodnota tohoto testu se určuje buď pomocí F rozdělení nebo permutacemi, kdy se permutuje přiřazení nepodobností mezi jednotlivými prvky a příslušných středů skupin.

5.1 Simulace - normální rozdělení

5.1.1 Postup při simulování

Aby byla situace co nejjednodušší, pracovali jsme pouze s pozorováními rozdělenými do dvou skupin a tato pozorování jsme generovali z dvojrozměrného normálního rozdělení. Dvojrozměrné rozdělení bylo zvoleno proto, že se data dají dobře

graficky znázornit. Normální rozdělení pak proto, že nám umožní od sebe oddělit parametr polohy a parametr měřítka. Jako koeficient nepodobnosti byla zvolena eukleidovská vzdálenost, protože v případě zobrazení dat je intuitivně zřetelná a protože je invariantní na posun a otočení, což umožnilo zjednodušení volby parametrů.

Pozorování pro první skupinu jsme generovali z rozdělení

$$\mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}\right)$$

a pozorování pro druhou skupinu z rozdělení

$$\mathcal{N}_2\left(\begin{pmatrix} \Delta\mu \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho_2\sigma^2 \\ \rho_2\sigma^2 & \sigma^2 \end{pmatrix}\right). \tag{5.1}$$

Voleny byly tři velikosti skupin: $(n_1, n_2) = (10, 10), (15, 10)$ a $(10, 15)$. Varianční matice z (5.1) tedy postupně příslušela stejně početné, méně početné a početnější skupině.

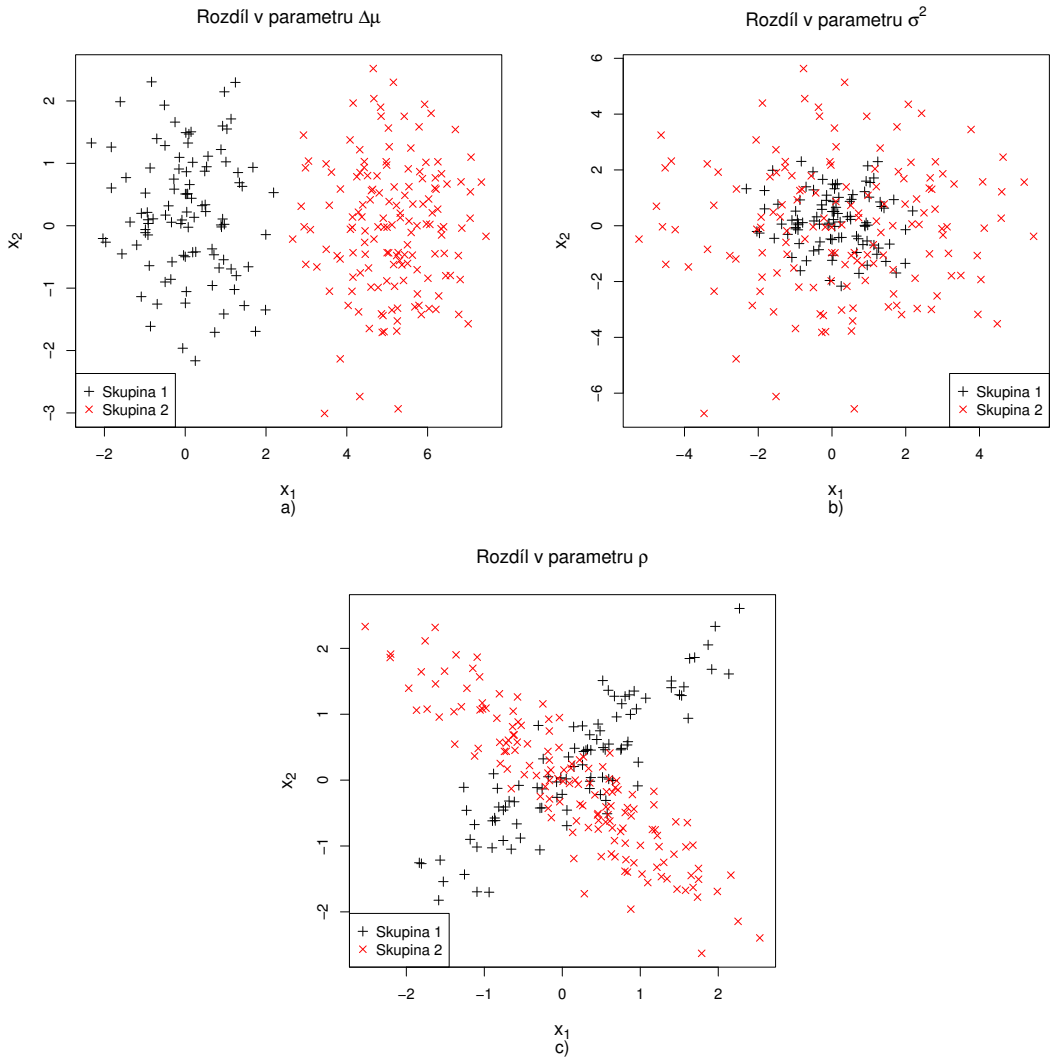
Jednotlivé parametry jsme volili následujícím způsobem:

$$\begin{aligned} \Delta\mu &= 0,0{,}2,0{,}4,0{,}6,0{,}8,1 \\ \sigma^2 &= 1,1{,}2,1{,}4,1{,}6,1{,}8,2{,}0 \\ (\rho_1,\rho_2) &= (0{,}9,0{,}9),(0{,}5,0{,}5),(0{,}25,0{,}25),(0,0) \\ &\quad\ (0{,}25,-0{,}25),(0{,}5,-0{,}5),(0{,}9,-0{,}9). \end{aligned}$$

Parametr $\Delta\mu$ budeme nazývat parametr polohy, σ^2 parametr měřítka. Pro zkrácený zápis dvojice parametrů (ρ_1, ρ_2) jsme zavedli označení ρ. Tento parametr budeme nazývat parametr tvaru a nabývá hodnot $\rho = 0{,}9, 0{,}5, 0{,}25, 0, -0{,}25, -0{,}5,$ a $-0{,}9$, které odpovídají výše uvedeným kombinacím parametrů ρ_1 a ρ_2. Kladné hodnoty tohoto parametru znamenají, že korelační koeficienty v obou skupinách jsou stejné a rovny ρ. Záporné hodnoty pak znamenají, že $\rho_1 = \rho = -\rho_2$, tedy korelační koeficienty ve skupinách mají opačné znaménko a v absolutní hodnotě jsou rovny ρ. Na obrázku 5.1 je graficky znázorněn efekt jednotlivých parametrů.

Změnami jednotlivých parametrů jsme zkoumali sílu jednotlivých variant testů ANOSIM a MRPP pro jednotlivé alternativy a sledovali, za jakých podmínek je interpretace testů chybná. Výpočty jsme provedli pro všech $6 \times 6 \times 7 = 252$ kombinací parametrů. Poznamenejme, že pokud $\Delta\mu = 0$, $\sigma^2 = 1$ a $\rho = 0, 0{,}25, 0{,}5, 0{,}9$, pocházejí obě skupiny ze stejného rozdělení, tedy simulacemi jsme ověřovali hladinu testů. Pokud $\Delta\mu \neq 0$, $\sigma^2 = 1$ a $\rho = 0, 0{,}25, 0{,}5, 0{,}9$, pak se rozdělení liší pouze v parametru polohy. Zároveň jsme zkoumali i test homogenity disperzí.

Všechny simulace byly prováděny v programu `R` pomocí funkcí uvedených v částech věnovaných příslušným testům. Pro určení P-hodnoty testů bylo prováděno 999 permutací. Výpočty síly jsou založeny na 1000 souborech dat vygenerovaných z příslušných rozdělení, pro ověřování hladiny testů jsme tento počet zvýšili na 10000. Veškeré výsledky pro všechny zmíněné konfigurace a varianty testů se nacházejí na přiloženém CD v souborech `VysledkySimulaciAnosim.txt` a `VysledkySimulaciLevene.txt` (viz příloha). Zde upozorníme na jejich základní rysy.



Obrázek 5.1: Efekt jednotlivých parametrů: a) parametr polohy $\Delta\mu$, b) parametr měřítka σ^2, c) parametr tvaru ρ.

5.1.2 ANOSIM a MRPP

Hladina testu

Pokud $\Delta\mu = 0$, $\sigma^2 = 1$ a $\rho = 0, 0{,}25, 0{,}5, 0{,}9$, skupiny měly stejné rozdělení. Simulacemi jsme tedy ověřovali, zda testy dodržují předepsanou hladinu, která je 0,05. V tabulce 5.1 jsou uvedeny hladiny jednotlivých testů získané simulacemi. Pro situaci $n_1 = n_2 = 10$ se většina těchto hodnot pohybuje pod hodnotou 0,05. Pro skupiny o rozdílných velikostech jsou hladiny mírně vyšší, ovšem stále velice blízké této hodnotě.

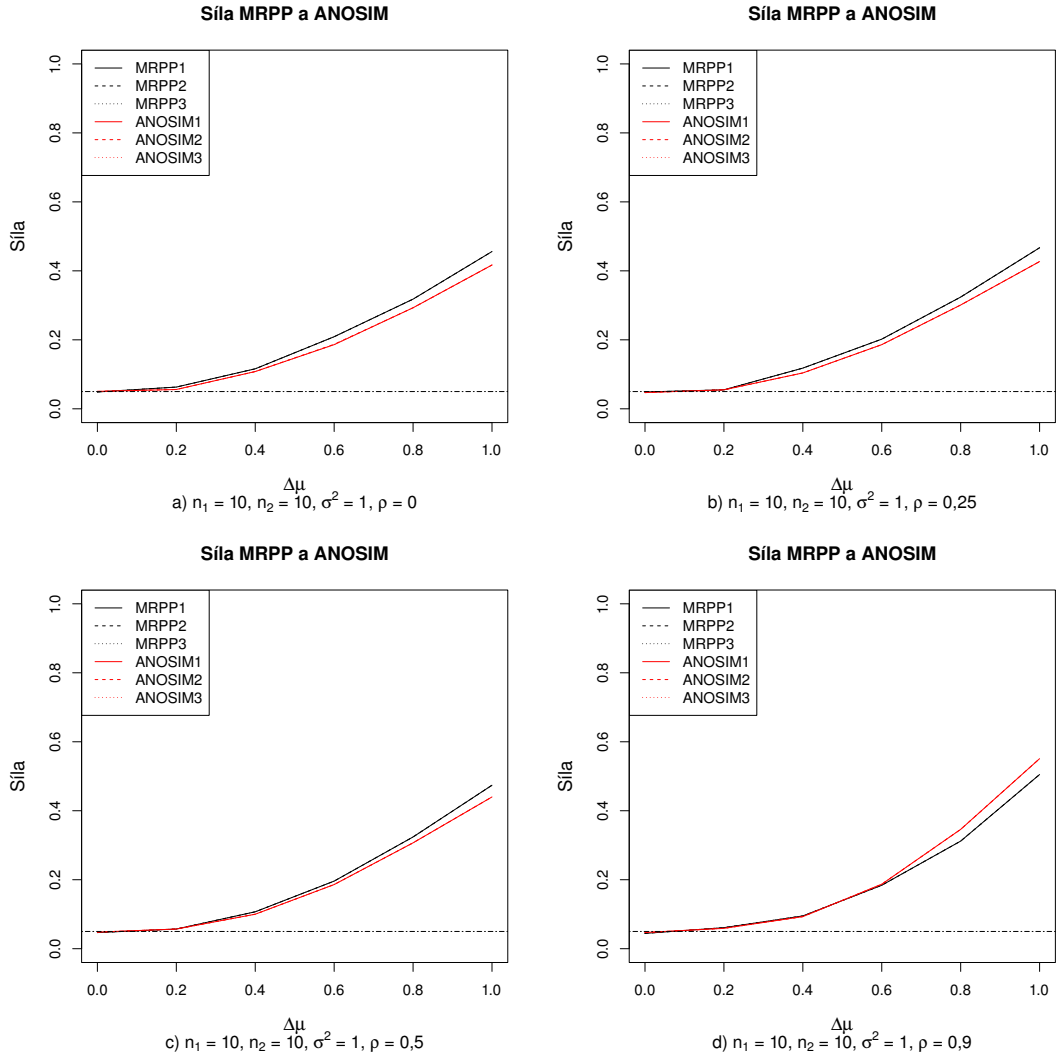
Alternativa polohy

Pokud $\Delta\mu \neq 0$, $\sigma^2 = 1$ a $\rho = 0, 0{,}25, 0{,}5, 0{,}9$, pocházejí skupiny z rozdělení, které se liší pouze polohou. Simulacemi tedy zjišťujeme sílu testu pro alternativy různě velkých rozdílů v poloze. Pro velikosti skupin $n_1 = n_2 = 10$ je síla

$\rho = 0$	$n_1 = n_2 = 10$	$n_1 = 10, n_2 = 15$	$n_1 = 15, n_2 = 10$
MRPP 1	0,0488	0,0512	0,0501
MRPP 2	0,0488	0,0516	0,0513
MRPP 3	0,0488	0,0524	0,0508
ANOSIM 1	0,0505	0,0499	0,0487
ANOSIM 2	0,0505	0,0509	0,0491
ANOSIM 3	0,0505	0,0503	0,0511
$\rho = 0{,}25$	$n_1 = n_2 = 10$	$n_1 = 10, n_2 = 15$	$n_1 = 15, n_2 = 10$
MRPP 1	0,0480	0,0504	0,0505
MRPP 2	0,0480	0,0507	0,0511
MRPP 3	0,0480	0,0501	0,0508
ANOSIM 1	0,0470	0,0506	0,0512
ANOSIM 2	0,0470	0,0505	0,0503
ANOSIM 3	0,0470	0,0495	0,0517
$\rho = 0{,}5$	$n_1 = n_2 = 10$	$n_1 = 10, n_2 = 15$	$n_1 = 15, n_2 = 10$
MRPP 1	0,0468	0,0502	0,0504
MRPP 2	0,0468	0,0494	0,0520
MRPP 3	0,0468	0,0494	0,0512
ANOSIM 1	0,0481	0,0524	0,0502
ANOSIM 2	0,0481	0,0499	0,0531
ANOSIM 3	0,0481	0,0489	0,0507
$\rho = 0{,}9$	$n_1 = n_2 = 10$	$n_1 = 10, n_2 = 15$	$n_1 = 15, n_2 = 10$
MRPP 1	0,0444	0,0488	0,0488
MRPP 2	0,0444	0,0483	0,0484
MRPP 3	0,0444	0,0502	0,0498
ANOSIM 1	0,0470	0,0501	0,0529
ANOSIM 2	0,0470	0,0480	0,0509
ANOSIM 3	0,0470	0,0486	0,0502

Tabulka 5.1: Hladiny testů ANOSIM a MRPP pro $\Delta\mu = 0$, $\sigma^2 = 1$ a $\rho = 0, 0{,}25, 0{,}5, 0{,}9$.

testů MRPP větší než síla testů ANOSIM s výjimkou $\rho = 0{,}9$, kde je situace opačná (obrázek 5.2). Připomeňme, že pokud jsou obě skupiny stejně velké, jsou varianty jednotlivých testů shodné. Pro velikosti skupin $n_1 = 10, n_2 = 15$ se již

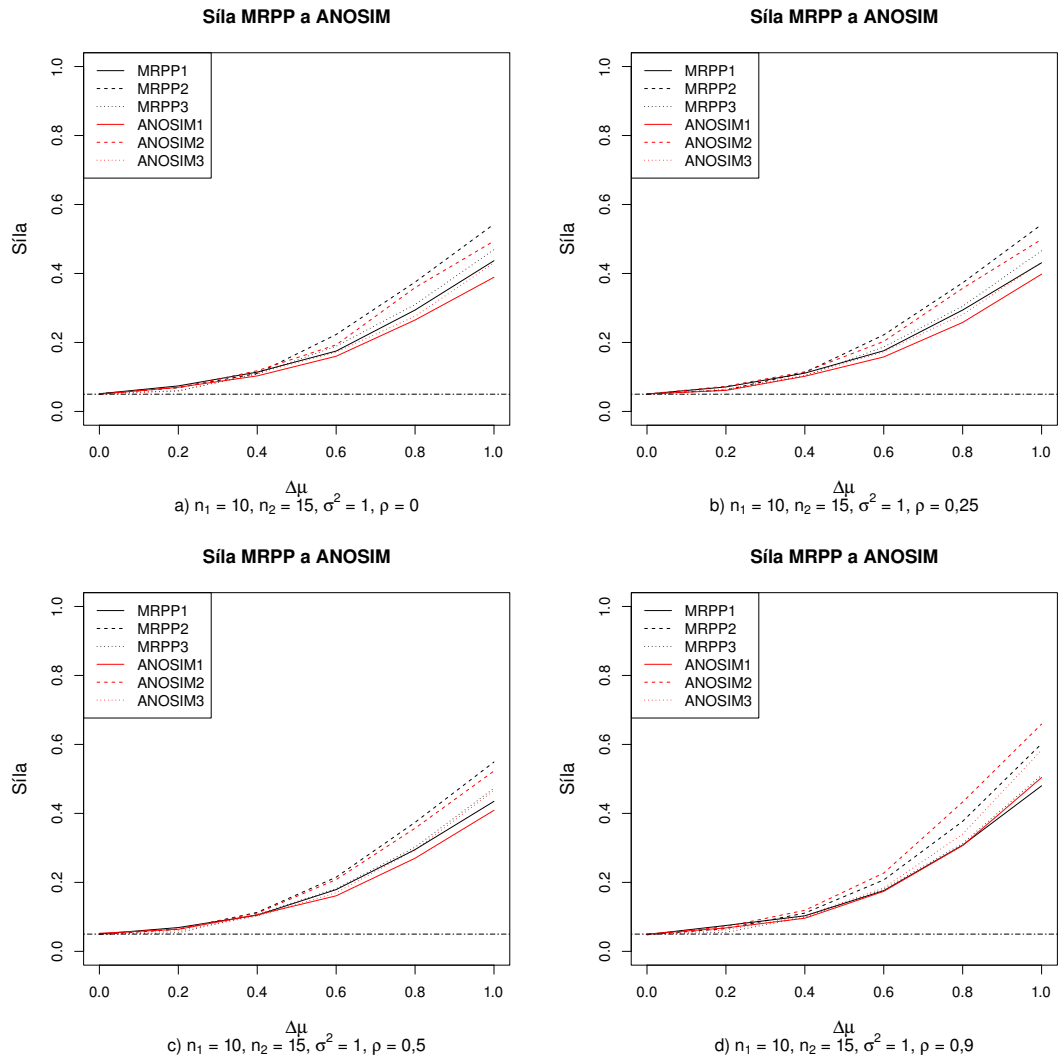


Obrázek 5.2: Síla testů MRPP a ANOSIM pro $n_1 = n_2 = 10$, $\sigma^2 = 1$, a) $\rho = 0$, b) $\rho = 0{,}25$, c) $\rho = 0{,}5$, d) $\rho = 0{,}9$.

síly jednotlivých variant testů liší, ale opět platí, že síla testů MRPP je větší než příslušných variant testů ANOSIM s výjimkou situace $\rho = 0{,}9$ (obrázek 5.3). Z variant pak má vždy největší sílu varianta 2 následována variantami 3 a 1. Velmi podobné výsledky jsme dostali pro $n_1 = 15, n_2 = 10$, proto je zde neuvádíme.

Alternativa měřítka

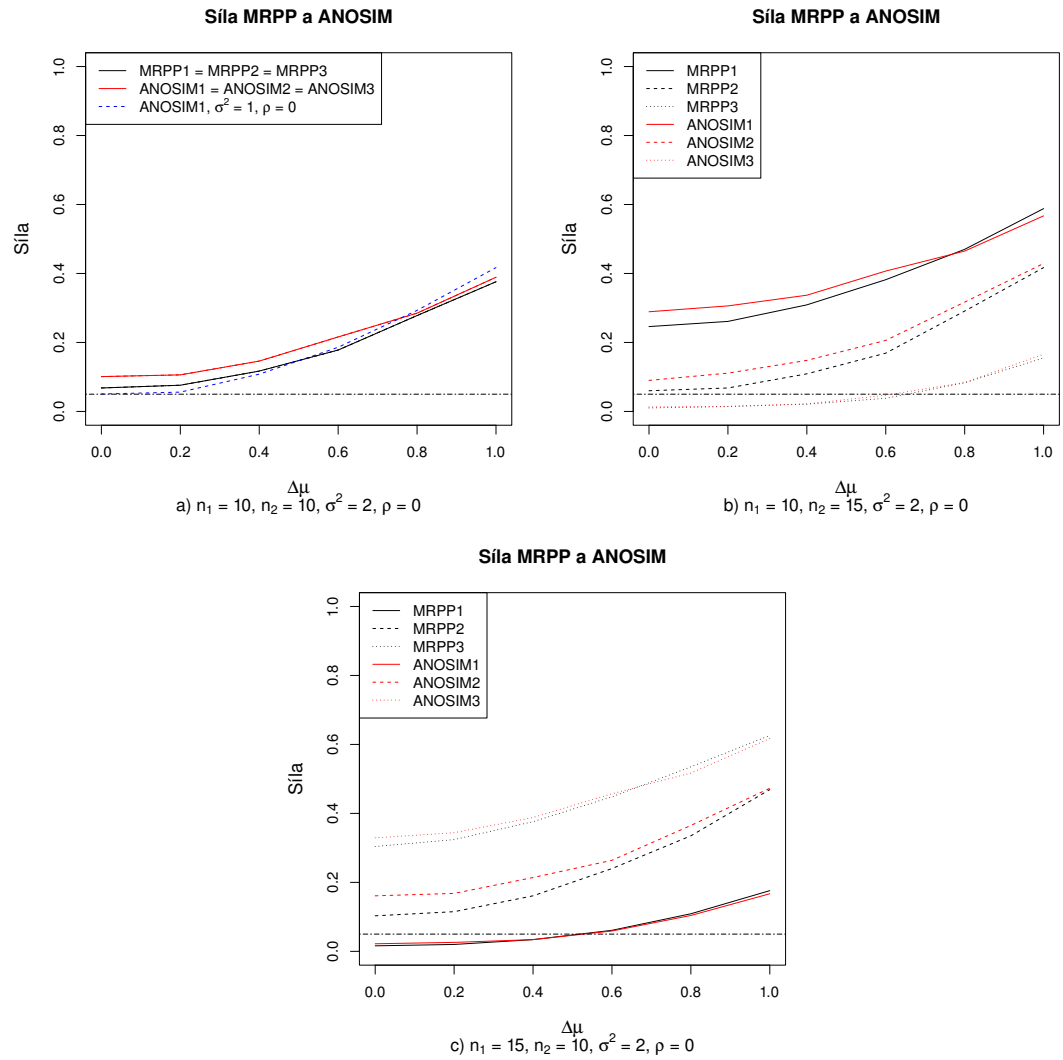
I když se jak ANOSIM tak MRPP interpretují často jako testy rozdílu polohy, jsou to testy celkové shody rozdělení. Důvodem k zamítnutí nulové hypotézy pak může být mj. rozdíl v parametru měřítka σ^2. Na grafech v obrázku 5.4 je zobrazena síla



Obrázek 5.3: Síla testů MRPP a ANOSIM pro $n_1 = 10, n_2 = 15$, $\sigma^2 = 1$, a) $\rho = 0$, b) $\rho = 0{,}25$, c) $\rho = 0{,}5$, d) $\rho = 0{,}9$.

jednotlivých testů proti altenativám polohy, pokud navíc parametr $\sigma^2 = 2$, tedy jedna ze skupin je více rozptýlená. Parametr ρ jsme ponechali roven 0.

Pokud $n_1 = n_2 = 10$, je síla testů pro nízké hodnoty $\Delta\mu$ vyšší, než v situacích $\sigma^2 = 1$ (testy detekují rozdílná měřítka), roste ale pomaleji (rozdíl v poloze se v rozptýlenějších datech obtížněji detekuje). Pro porovnání je do grafu vložena síla testu ANOSIM 1 pokud $\sigma^2 = 1$. Pro skupiny o rozdílných velikostech se vyskytly mezi jednotlivými variantami testů výrazné rozdíly. V situaci $n_1 = 10, n_2 = 15$ (početnější skupina je více rozptýlená) má největší sílu varianta 1, pro varianty 2 je síla nižší a varianty 3 mají sílu nejnižší. Rozdíl v měřítku nejvíce ovlivnil právě varianty 1 a 3. U první vedl k výraznému nárůstu síly, u druhé naopak k poklesu. V tabulce 5.2 jsou uvedeny síly jendotlivých variant testů pro $\Delta\mu = 0$. V situaci $n_1 = 15, n_2 = 10$ (početnější skupina je méně rozptýlená) jsou pak výsledky zcela obrácené (viz tabulka 5.3). Nejslabší je varianta 1 a nejsilnější je varianta 3. Velmi rozdílné P-hodnoty jednotlivých variant testů mohou naznačit, že kromě polohy se rozdělení, ze kterých pocházejí pozorování, liší právě v parametru měřítka. Pro



Obrázek 5.4: Síla testů MRPP a ANOSIM pro $\sigma^2 = 2$, $\rho = 0$ a velikosti skupin a) $n_1 = 10, n_2 = 10$, b) $n_1 = 10, n_2 = 15$, c) $n_1 = 15, n_2 = 10$.

tyto situace je pak zkonstruován test homogenity disperzí.

Podobné výsledky jako pro $\rho = 0$ jsme získali i pro hodnoty parametru $\rho =$ 0,25, 0,5 a 0,9. Uvádíme proto opět pouze výsledky pro situaci $\rho = 0{,}9$ (obrázek 5.5). Do prvního grafu je opět pro srovnání zakreslena síla testu ANOSIM 1 při $\sigma^2 = 1$ a $\rho = 0.9$.

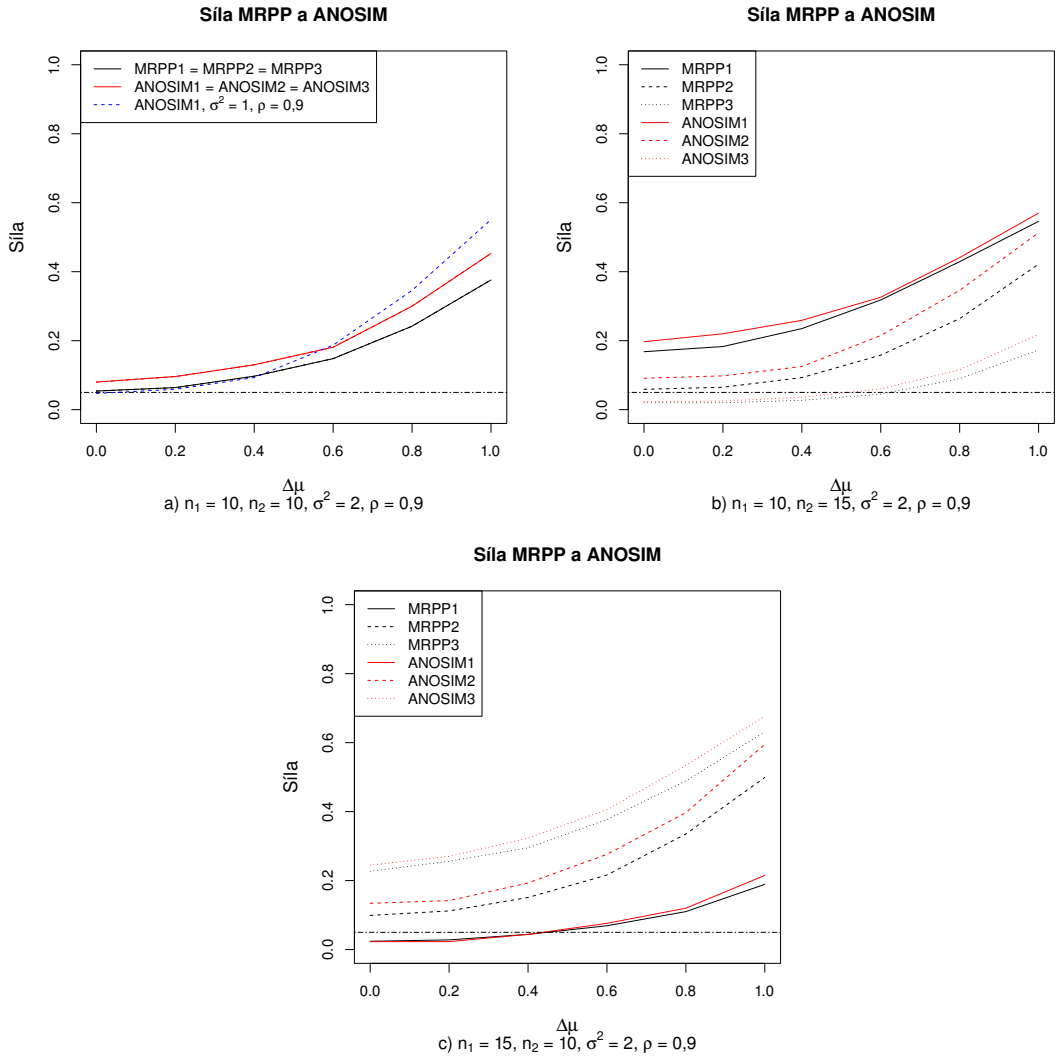
Obrázek 5.6 je ilustruje důvod rozdílných sil jednotlivých variant. Mějme situaci, kdy $\mu = 0, \sigma = 2$ a $\rho = 0$. Jedna skupina pozorování je tedy rozptýlenější než druhá, ale obě mají stejnou polohu (obrázek 5.6, a)). Skupina 1 obsahuje 100 pozorování, skupina 2 je rozptýlenější a má 150 pozorování. Dodány jsou boxploty vnitroskupinových a meziskupinových nepodobností (obrázek 5.6, c)). Na obrázku 5.6, b) je přiřazení do skupin náhodně permutováno a opět jsou dodány boxploty jednotlivých nepodobností (obrázek 5.6, d)). Zatímco v původních datech se průměrné vnitroskupinové nepodobnosti od sebe výrazně liší, v permutovaných datech tomu tak není. Testové statistiky testů ANOSIM a MRPP jsou váženými průměry těchto průměrných vnitroskupinových nepodobností. Zatímco pro

MRPP 1	0,246	ANOSIM 1	0,289
MRPP 2	0,060	ANOSIM 2	0,090
MRPP 3	0,010	ANOSIM 3	0,013

Tabulka 5.2: Síly testů ANOSIM a MRPP pro $n_1 = 10$, $n_2 = 15$, $\Delta\mu = 0$ a $\sigma^2 = 2$

MRPP 1	0,016	ANOSIM 1	0,022
MRPP 2	0,103	ANOSIM 2	0,161
MRPP 3	0,304	ANOSIM 3	0,329

Tabulka 5.3: Síly testů ANOSIM a MRPP pro $n_1 = 15$, $n_2 = 10$, $\Delta\mu = 0$ a $\sigma^2 = 2$



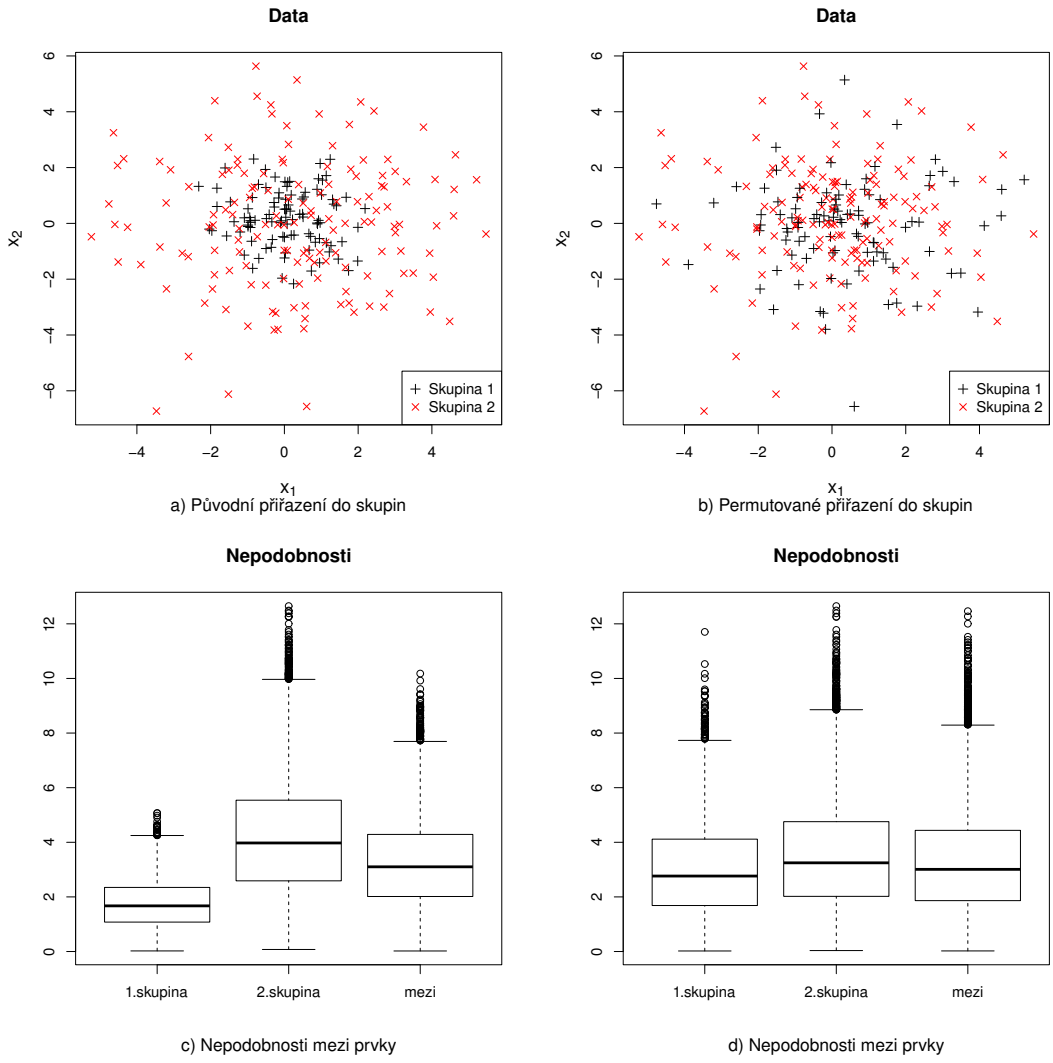
Obrázek 5.5: Síla testů MRPP a ANOSIM pro $\sigma^2 = 2$, $\rho = 0{,}9$ a velikosti skupin a) $n_1 = 10, n_2 = 10$, b) $n_1 = 10, n_2 = 15$, c) $n_1 = 15, n_2 = 10$.

permutovaná data volba vah výrazně neovlivní hodnotu testové statistiky, pro původní data ji ovlivní výrazně. A protože v těchto testech zamítáme nulovou hypotézu pro malé hodnoty testové statistiky, vedou velké váhy u skupin s nižšími vnitroskupinovými nepodobnostmi k častějšímu zamítání nulové hypotézy a

tím k větší síle. Pro případ 100 a 150 pozorování jsou váhy jednotlivých variant následující:

- varianta 1 - váhy 0,5 a 0,5
- varianta 2 - váhy 0,4 a 0,6
- varianta 3 - váhy 0,3 a 0,7

Varianta 1 dává první skupině největší váhu (0,5) a má tedy i největší sílu. Pokud by velikosti skupin byly obrácené, dávala by největší váhu méně rozptýlené skupině varianta 3 (0,7). Tedy nejsilnější by byla tato varianta. To přesně odpovídá výsledkům simulací (viz obrázky 5.4, 5.5 a tabulky 5.2, 5.3).



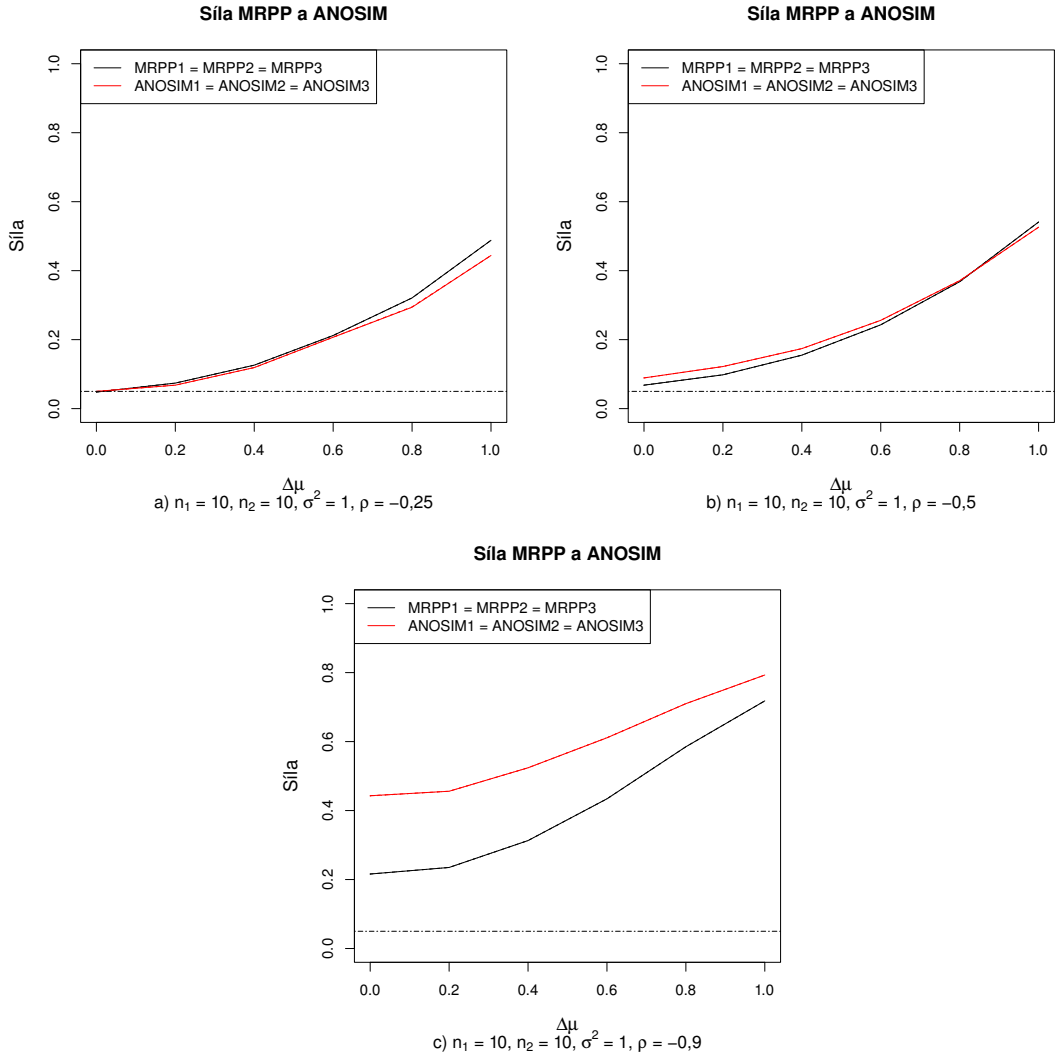
Obrázek 5.6: Ilustrace vlivu vah. Skupiny o velikostech 100 a 150, $\Delta\mu = 0$, $\sigma^2 = 5$, $\rho = 0$. a) Původní nagenerovaná data, b) Data u nichž je permutováno přiřazení do skupin, c) Boxploty nepodobností mezi pozorováními příslušejícími do 1. skupiny, 2. skupiny a rozdílných skupin, d) Tytéž boxploty pro data u nichž je permutováno přiřazení do skupin.

Alternativa tvaru

Posledním parametrem, kterým jsme se zabývali, je parametr tvaru ρ. Jedná se o korelace složek (dvojrozměrných) pozorování, které jsme na začátku této kapitoly označili ρ_1 a ρ_2. Pokud $\Delta\mu = 0$, $\sigma^2 = 1$ a $\rho_1 = -\rho_2$ je

$$\mathcal{L}(X_1, X_2) = \mathcal{L}(-X_1, X_2), \quad \text{kde } (X_1, X_2) \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}\right).$$

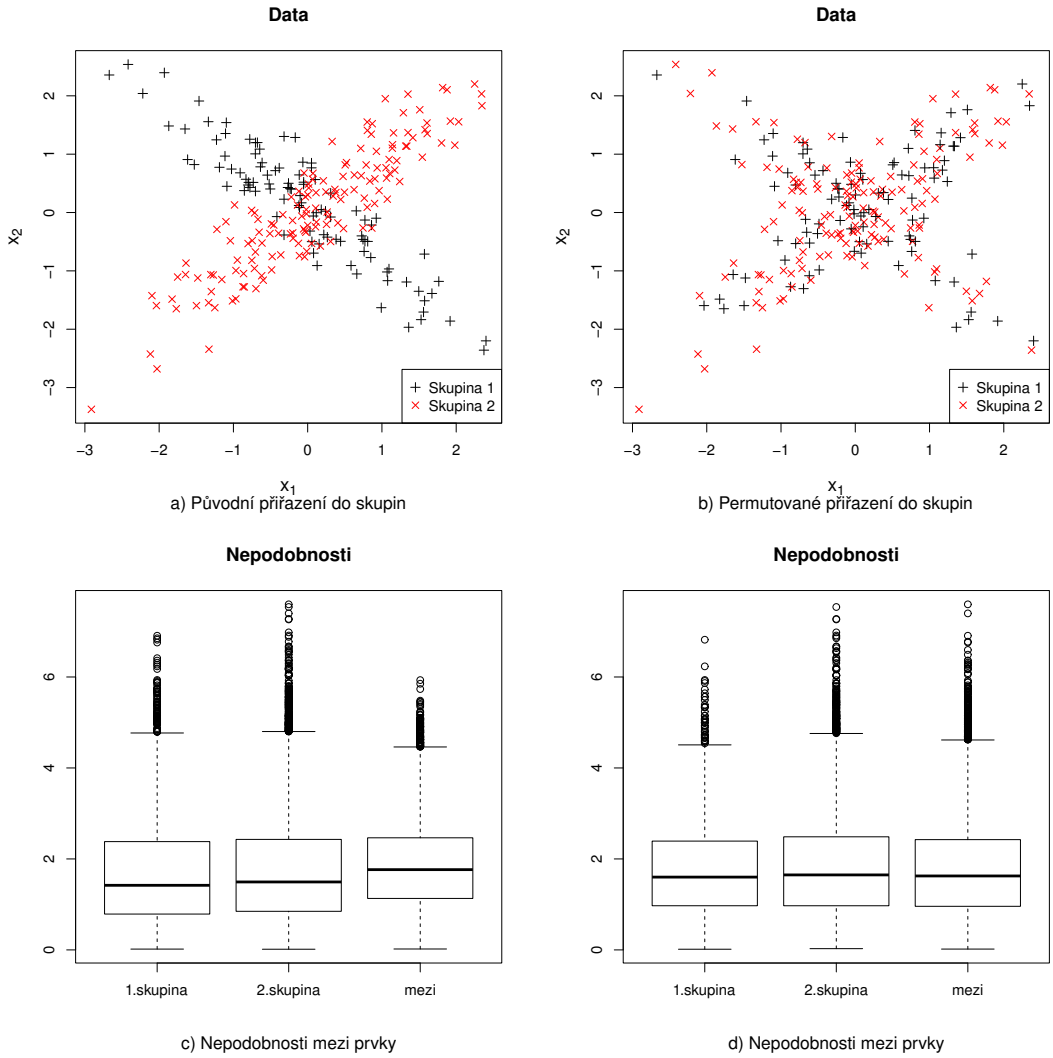
Odsud tedy název parametr tvaru. Pokud je $\rho = 0, 0{,}25, 0{,}5, 0{,}9$, skupiny se v tomto parametru neliší a tyto situace jsme popsali již v předchozích bodech. Nyní se zaměříme na situace, kdy $\rho = -0{,}25, -0{,}5, -0{,}9$ a $\sigma^2 = 1$. Kompletní výsledky se nacházejí na přiloženém CD. Na obrázku pro $n_1 = n_2 = 10$ (obrázek 5.7) vidíme,



Obrázek 5.7: Síla testů MRPP a ANOSIM pro $n_1 = n_2 = 10$, $\sigma^2 = 1$, a) $\rho = -0{,}25$, b) $\rho = -0{,}5$, c) $\rho = -0{,}9$.

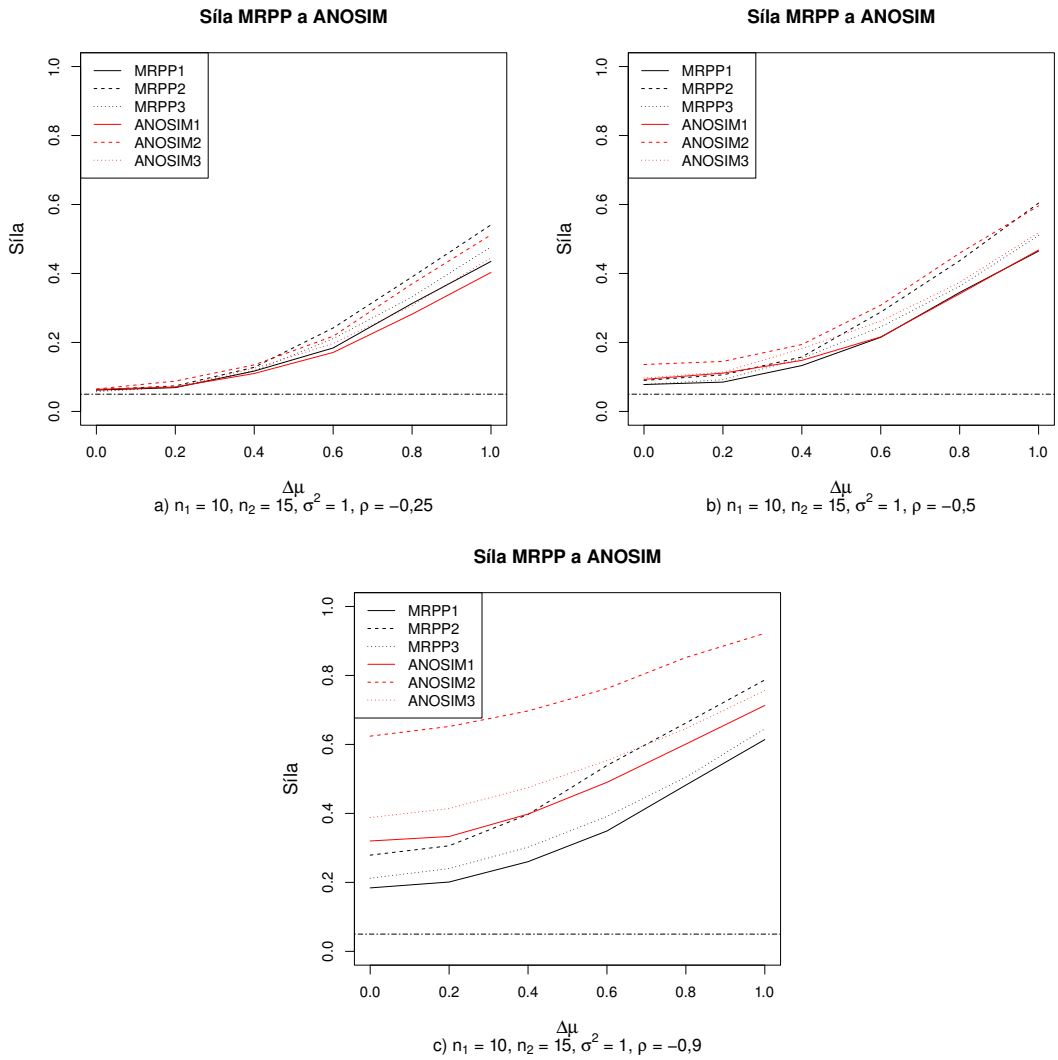
že výrazný rozdíl v tvaru ($\rho = -0.9$) vedl k nárůstu síly testů. Tento fakt je způsoben tím, že permutujeme-li data mezi skupinami, zvětší se v obou skupinách průměrné vzdálenosti ke středům skupin. Původní hodnota statistik ANOSIM

nebo MRPP (r_0 nebo δ_0) je pak „malá“ v porovnání s rozdělením získaným pomocí permutací a to vede k zamítání nulové hypotézy (viz obrázek 5.8).



Obrázek 5.8: Vliv permutací na meziskupinové nepodobnosti. Skupiny o velikostech 100 a 150, $\Delta\mu = 0$, $\sigma^2 = 1$, ρ=-0,9. a) Původní nagenerovaná data, b) Data u nichž je permutováno přiřazení do skupin, c) Boxploty nepodobností mezi pozorováními příslušejícími do 1. skupiny, 2. skupiny a rozdílných skupin, d) Tytéž boxploty pro data u nichž je permutováno přiřazení do skupin.

Nápovědou, že se jednotlivá rozdělení liší „tvarem“ může být stejně jako v případě alternativy měřítka fakt, že P-hodnoty jednotlivých variant testů ANOSIM nebo MRPP vyjdou výrazně rozdílné. Problémem v tomto případě je, že všechny vnitroskupinové nepodobnosti pocházejí ze stejného rozdělení kvůli vlastnostem eukleidovské vzdálenosti a test homogenity disperzí nám v této situaci nepomůže. Pro velikosti skupin $n_1 = 10, n_2 = 15$ opět sledujeme nárůst síly všech testů se změnou parametru ρ (obrázek 5.9). Ovšem tentokrát je nejsilnější varianta 2 následována variantami 3 a 1 (připomeňme, že pořadí síly variant pro alternativu měřítka bylo 1,2,3). Stejné pořadí, na rozdíl od alternativy měřítka, kde se pořadí variant obrátilo, platí i pro velikosti $n_1 = 15, n_2 = 10$ (viz obrázek 5.10).



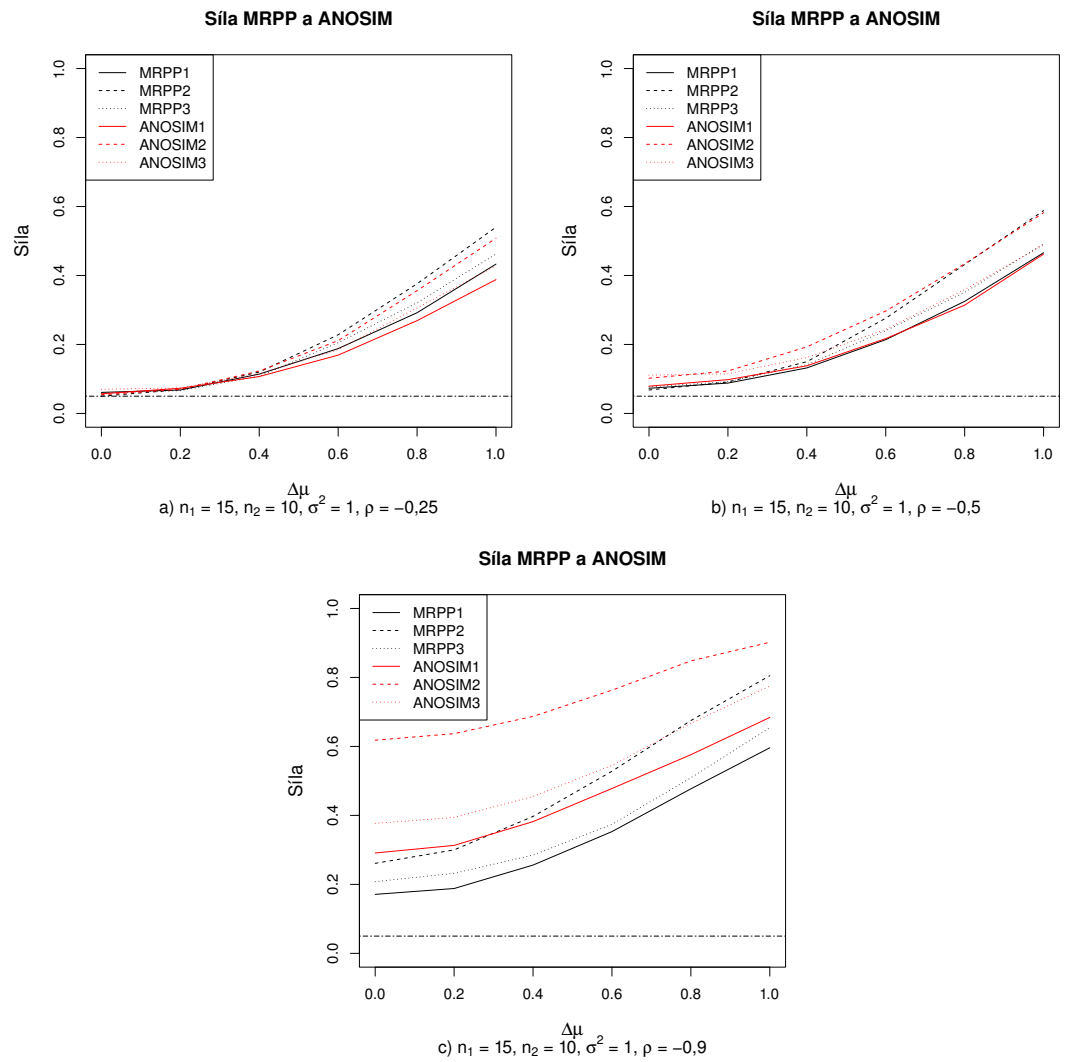
Obrázek 5.9: Síla testů MRPP a ANOSIM pro $n_1 = 10, n_2 = 15$, $\sigma^2 = 1$, a) $\rho = -0{,}25$, b) $\rho = -0{,}5$, c) $\rho = -0{,}9$.

Závěr

Při testování testy ANOSIM a MRPP může být výhodné spočítat všechny tři varianty testu. Pokud zamítneme nulovou hypotézu shody rozdělení doporučujeme porovnat P-hodnoty všech tří variant testů. Pokud pracujeme se skupinami stejných velikostí a pokud si jsou tyto P-hodnoty blízké, zdá se interpretace rozdílu rozdělení jako rozdílu v poloze oprávněná. Pokud se však P-hodnoty jednotlivých variant výrazně liší, může tento fakt vypovídat o jiném rozdílu mezi rozděleními.

5.1.3 Test homogenity disperzí

V simulacích vlastností testu homogenity disperzí jsme se zaměřili na parametry σ^2 a ρ. Pro hodnoty parametru $\sigma^2 = 1$ a $\rho = 0, 0{,}25, 0{,}5$ a $0{,}9$ pocházejí obě skupiny pozorování ze stejných rozdělení, opět tedy ověřujeme hladinu testu (viz tabulka 5.4). Ve všech případech je dosažená hladina varianty s centroidy větší, než dosažená hladina varianty s mediány a pravidelně překračuje předepsanou



Obrázek 5.10: Síla testů MRPP a ANOSIM pro $n_1 = 15, n_2 = 10$, $\sigma^2 = 1$, a) $\rho = -0{,}25$, b) $\rho = -0{,}5$, c) $\rho = -0{,}9$.

hodnotu 0,05. Tato varianta je tedy anti-konzervativní, což odpovídá závěrům v Anderson (2006). Možným důvodem je, že centroid skupiny je více ovlivněn odlehlými pozorováními než prostorový medián. Centroid je pak oproti mediánu více posunut k odlehlým pozorováním. To na jednu stranu snižuje vzdálenost mezi tímto pozorováním a centroidem, na druhou stranu však zvýšuje vzdálenosti centroidu od všech ostatních pozorování ve skupině. To pak může vést k zamítání nulové hypotézy pro variantu s centroidy. Právě tato situace se objevila v kapitole 4 při analýze dat `mouchy` pro letní pozorování a eukleidovskou vzdálenost (obrázek 4.6). Případný rozdíl ve velikosti skupin se více projevil u variant s centroidem, kdy dosahované hladiny testů jsou vyšší pro vyvážený design než pro nevyvážený.

Na obrázku 5.11 je znázorněna síla testu homogenity disperzí proti alternativám měřítka pro různé hodnoty parametru ρ. Výsledky potvrzují, že tento test není vhodný pro odhalování rozdílu v parametru tvaru, protože tento parametr nemá vliv na rozdělení vnitroskupinových nepodobností. Podobné výsledky jsme

$n_1 = n_2 = 10$	FC	FM	PC	PM
$\rho = 0{,}00$	0,0622	0,0340	0,0612	0,0333
$\rho = 0{,}25$	0,0600	0,0327	0,0590	0,0327
$\rho = 0{,}50$	0,0581	0,0318	0,0594	0,0322
$\rho = 0{,}90$	0,0607	0,0338	0,0612	0,0339
$n_1 = 10$, $n_2 = 15$	FC	FM	PC	PM
$\rho = 0{,}00$	0,0556	0,0335	0,0541	0,0322
$\rho = 0{,}25$	0,0545	0,0338	0,0537	0,0342
$\rho = 0{,}50$	0,0535	0,0335	0,0534	0,0334
$\rho = 0{,}90$	0,0564	0,0324	0,0572	0,0338
$n_1 = 15$, $n_2 = 10$	FC	FM	PC	PM
$\rho = 0{,}00$	0,0555	0,0344	0,0537	0,0337
$\rho = 0{,}25$	0,0573	0,0337	0,0550	0,0328
$\rho = 0{,}50$	0,0542	0,0337	0,0546	0,0331
$\rho = 0{,}90$	0,0529	0,0331	0,0531	0,0334

Tabulka 5.4: Hladiny testu homogenity disperzí pro hodnoty parametrů $\sigma^2 = 1$ a $\rho = 0, 0{,}25, 0{,}5, 0{,}9$ a rozdílné velikosti skupin. Varianty testu: C - centroid, M - medián. Určení P-hodnoty: F - F rozdělení, P - permutace.

dostali i pro velikosti skupin $n_1 = 10, n_2 = 15$ a $n_1 = 15, n_2 = 10$.

Závěr

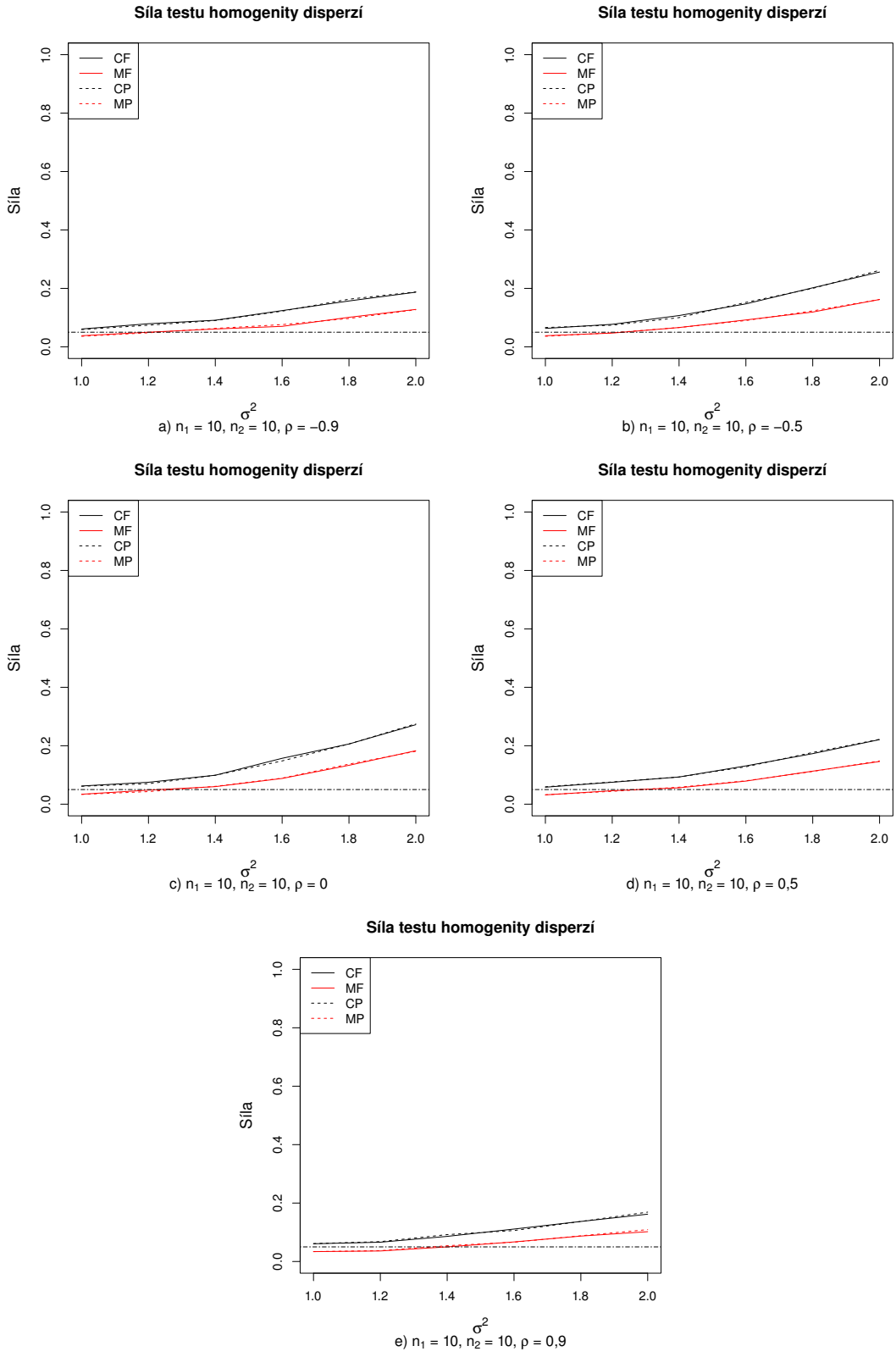
Při testování pomocí testu homogenity disperzí získáme podobné výsledky, použijeme-li k určení P-hodnot F rozdělení nebo permutací. Na druhou stranu je zde rozdíl mezi variantou s centroidy a variantou s mediány. Varianta s centroidy se ukazuje být anti-konzervativní.

5.2 Simulace - Poisson-log normální rozdělení

5.2.1 Postup při simulování

Inspirováni Anderson (2006) jsme použili Monte Carlo simulace pro porovnání síly a chyby 1. druhu testů provedných v kapitole 4 na datech `mouchy`. Data jsme simulovali z mnohorozměrného Poisson-log normálního rozdělení (Aitchinson a Ho (1989)), přičemž pro odhad parametrů tohoto rozdělení jsme použili data `mouchy` (před aplikací transformace $\sqrt[4]{x}$). Připomeňme, že jsme sledovali zastoupení 156 hmyzích druhů. Generovali jsme ze 156-rozměrného Poisson-log normálního rozdělení a to následujícím způsobem:

- Ze 156 rozměrného mnohorozměrného normálního rozdělení o parametrech $\boldsymbol{\mu} = (\mu_1, ..., \mu_{156})$ a $\boldsymbol{\Sigma} = (\sigma_{ij}), i, j = 1, ..., 156$ byl vygenerován vektor $\boldsymbol{\theta} = (\theta_1, ..., \theta_{156})$.
- Četnost každého druhu pak byla generována z Poissonova rozdělení s parametrem e^{θ_i}, $i = 1, ..., 156$.



Obrázek 5.11: Síla testů homogenity disperzí pro $n_1 = 10, n_2 = 10$, a) $\rho = -0{,}9$, b) $\rho = -0{,}5$, c) $\rho = 0$, d) $\rho = 0{,}5$, e) $\rho = 0{,}9$. Varianty testu: C - centroid, M - medián. Určení P-hodnoty: F - F rozdělení, P - permutace.

Má-li vektor $\mathbf{X} = (X_1, ..., X_{156})'$ mnohorozměrné Poisson-log normalní rozdělení s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$, je (Aitchinson a Ho (1989)):

$$\begin{aligned} \mathrm{E}X_i &= e^{\mu_i + \frac{1}{2}\sigma_{ii}} = \alpha_i, \\ \mathrm{var}\, X_i &= \alpha_i + \alpha_i^2 e^{\sigma_{ii}-1}, \\ \mathrm{cov}(X_i, X_j) &= \alpha_i \alpha_j e^{\sigma_{ij}-1}. \end{aligned}$$

Hodnoty $\mathrm{E}X_i, \mathrm{var}\, X_i$ a $\mathrm{cov}(X_i, X_j)$ byly odhadnuty momentově a parametry normálního rozdělení vypočteny jako:

$$\begin{aligned} \hat{\sigma}_{ii} &= \log \frac{\widehat{\mathrm{var}}\, X_i - \hat{\mathrm{E}}X_i}{(\hat{\mathrm{E}}X_i)^2} + 1, \\ \hat{\mu}_i &= \log \hat{\mathrm{E}}X_i - \frac{1}{2}\hat{\sigma_{ii}}, \\ \hat{\sigma}_{ij} &= \log \frac{\widehat{\mathrm{cov}}(X_i, X_j)}{\hat{\mathrm{E}}X_i \hat{\mathrm{E}}X_j} + 1. \end{aligned}$$

Pokud se při výpočtu odhadů prvků $\boldsymbol{\Sigma}$ v argumentu logaritmů vyskytla nekladná hodnota, nebo hodnota $\hat{\sigma}_{ii}$ byla záporná, zadefinovali jsme příslušnou hodnotu $\hat{\sigma}_{ij}$ nebo $\hat{\sigma}_{ii}$ jako 0. Toto se týkalo především druhů, které se na daných stanovištích vůbec nevyskytovaly (potom $\hat{\mu}_i = -\infty$ a $\theta_i = 0$). Veškeré simulace jsme prováděli zvlášť pro každé roční období.

Pro odhad chyb prvního druhu jsme simulovali 10 000 datových sad o stejných velikostech jako data původní, přičemž parametry byly odhadnuty ze všech 17 pozorování (tedy za platnosti nulové hypotézy). Pro výpočet P-hodnot bylo použito 999 permutací. Počet zamítnutí nulové hypotézy (na hladině 0,05) je pak odhadem chyby prvního druhu. Stejný postup jsme použili i pro odhad síly s tím rozdílem, že simulováno bylo pouze 1000 datových sad a parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ byly odhadnuty pro každou ze tří skupin (o velikostech 7, 7 a 3) zvlášť. Použili jsme Bray-Curtisův koeficient nepodobnosti.

5.2.2 ANOSIM a MRPP

Výsledky srovnání variant testů ANOSIM a MRPP pro jednotlivá roční období uvádíme v tabulce 5.5. Hladina 0,05 je dodržena pro všechny varianty testů ANOSIM i MRPP a jejich síla byla stejná (100 %).

Protože síla všech testů v alternativě byla rovna 1, rozhodli jsme se nalézt takovou lineární kombinaci hypotézy a alternativy, aby síly jednotlivých variant testů byly menší než 1 a umožnily porovnat jednotlivé varianty. Připomeňme, že za platnosti hypotézy jsme parametry odhadovali ze všech 17 pozorování pro každé roční období (označme tyto odhady $\boldsymbol{\mu}_0^{SP}, \boldsymbol{\mu}_0^{SU}, \boldsymbol{\mu}_0^{AU}$ a $\boldsymbol{\Sigma}_0^{SP}, \boldsymbol{\Sigma}_0^{SU}, \boldsymbol{\Sigma}_0^{AU}$). V alternativě jsme pak tyto parametry odhadovali zvlášť pro každou skupinu (označme je $\boldsymbol{\mu}_{Karpaty}^{SP}, \boldsymbol{\mu}_{Beskydy}^{SP}, \boldsymbol{\mu}_{Kysuce}^{SP}, \boldsymbol{\Sigma}_{Karpaty}^{SP}$ atd.) Pro každou skupinu a každé roční období jsme vytvořili n8sledující lineární kombinaci těchto parametrů.

$$\begin{aligned} \boldsymbol{\mu}_{j,mix}^{i} &= 0{,}975\, \boldsymbol{\mu}_0^i + 0{,}025\, \boldsymbol{\mu}_j^i \\ \boldsymbol{\Sigma}_{j,mix}^{i} &= 0{,}975\, \boldsymbol{\Sigma}_0^i + 0{,}025\, \boldsymbol{\Sigma}_j^i \\ i &= \text{SP, SU, AU} \\ j &= \text{Karpaty, Beskydy, Kysuce} \end{aligned}$$

Jaro	MRPP 1	MRPP 2	MRPP 3
chyba 1. druhu	0,051	0,048	0,048
síla	1,000	1,000	1,000
Jaro	ANOSIM 1	ANOSIM 2	ANOSIM 3
chyba 1. druhu	0,051	0,049	0,047
síla	1,000	1,000	1,000
Léto	MRPP 1	MRPP 2	MRPP 3
chyba 1. druhu	0,049	0,043	0,041
síla	1,000	1,000	1,000
Léto	ANOSIM 1	ANOSIM 2	ANOSIM 3
chyba 1. druhu	0,049	0,042	0,043
síla	1,000	1,000	1,000
Podzim	MRPP 1	MRPP 2	MRPP 3
chyba 1. druhu	0,046	0,046	0,046
síla	1,000	1,000	1,000
Podzim	ANOSIM 1	ANOSIM 2	ANOSIM 3
chyba 1. druhu	0,044	0,044	0,044
síla	1,000	1,000	1,000

Tabulka 5.5: Odhady chyb 1. druhu a síly testu v alternativě pro jednotlivá roční období.

V parametrech $\boldsymbol{\mu}_j^i$ se často vyskytovaly hodnoty $-\infty$, které odpovídaly tomu, že v dané skupině se daný druh nevyskytoval. Tato hodnota však představuje problém při výpočtech lineárních kombinací. Odpovídající $\boldsymbol{\mu}_{j,mix}^i$ má poté také hodnotu $-\infty$ a tedy v datech generovaných z mnohorozměrného Poisson-log normálního rozdělení s parametrem $\boldsymbol{\mu}_{j,mix}^i$ jsou zastoupeny tytéž druhy jako při hodnotě parametru $\boldsymbol{\mu}_j^i$ (liší se pouze v četnostech). Vzhledem k tomu, že jsme na data používali transformaci $\sqrt[4]{x}$, je Bray-Curtisův koeficient nepodobnosti mnohem více ovlivněn druhovou skladbou než konkrétními četnostmi. Nepodobnosti takto vygenerovaných dat byly velmi podobné těm, které jsme generovali za platnosti alternativy a síly jednotlivých testů byly opět 1. Proto jsme nahradili hodnotu $-\infty$ hodnotou -100. Ta sama o sobě příliš neovlivní parametr Poissonova rozdělení, ze kterého se generuje četnost ($e^{-100} = 10^{-44}$), avšak umožní generovat v jednotlivých skupinách druhy, které se v nich původně nevyskytovaly. Hodnota -100 samozřejmě ovlivnila koeficienty lineární kombinace (zde 0,975 a 0,025). Parametry $\boldsymbol{\Sigma}_{j,mix}^i$ jsou tedy velmi blízké parametru $\boldsymbol{\Sigma}_0^i$. Rovněž jsme vypočetli testy homogenity disperze pro tuto alternativu síly (viz následující část). Pro odhad síly bylo generováno 1000 datových sad.

Výsledky jsou shrnuty v tabulce 5.6. Ve všech případech má největší sílu varianta 2 následovaná variantou 3. Nejslabší je ve všech případech varianta 1, přičemž pro podzimní pozorování je tento rozdíl největší. Síly odpovídajících si variant MRPP a ANOSIM jsou velmi podobné. Pořadí P-hodnot variant (2,3,1, přičemž 2 a 3 jsou si blízké) naznačuje větší disperzi u skupiny s nejmenším počtem pozorování (Kysuce).

Jaro	MRPP 1	MRPP 2	MRPP 3
síla	0,532	0,759	0,655
Jaro	ANOSIM 1	ANOSIM 2	ANOSIM 3
síla	0,519	0,728	0,633
Léto	MRPP 1	MRPP 2	MRPP 3
síla	0,404	0,755	0,747
Léto	ANOSIM 1	ANOSIM 2	ANOSIM 3
síla	0,397	0,740	0,731
Podzim	MRPP 1	MRPP 2	MRPP 3
síla	0,441	0,833	0,827
Podzim	ANOSIM 1	ANOSIM 2	ANOSIM 3
síla	0,438	0,821	0,822

Tabulka 5.6: Odhady síly testů MRPP a ANOSIM v alternativě lineární kombinace pro jednotlivá roční období.

5.2.3 Test homogenity disperzí

V tabulce 5.7 uvádíme výsledky simulací týkající se variant testu homogenity disperzí. Tyto testy výrazně nedodržely předepsanou hladinu 0,05 (chyba 1. druhu až 0,31 pro varianty s centroidy a 0,13 pro varianty s mediány). Důvodem je rozdílná velikost skupin. Rozdělení nepodobností pozorování a středů skupin totiž závisí také na počtu pozorování ve skupinách. Skupiny Karpaty a Beskydy mají po sedmi pozorováních, ale skupina Kysuce má pozorování pouze tři. Proto jsme provedli ještě jedny simulace, kdy jsme pro skupinu Kysuce generovali 7 pozorování. Varianta s centroidy se pak ukázala být pouze mírně anti-konzervativní a dosažená hladina variant s mediány poklesla pod hodnotu 0,05. (viz tabulka 5.8). Až na letní pozorování pak síla testu v alternativě pro všechny varianty klesla přibližně o 15%. Ukazuje se tedy, jaký výrazný vliv má rozdílná velikost skupin na tento test. Mezi určením P-hodnot pomocí permutací nebo pomocí F-rozdělení se neobjevily výrazné rozdíly a síly variant s centroidy byly vždy vyšší než síly variant s mediány.

V tabulkách 5.9 a 5.10 uvádíme odhady sil testu homogenity disperzí pro lineární kombinaci hypotézy a alternativy zmíněnou v předchozí části. Opět jsme provedli výpočty jak pro původní velikosti skupin (7, 7 a 3), tak pro vyvážený design (skupiny mají po sedmi pozorováních). Pro nevyvážený design jsou síly v alternativě lineární kombinace menší než v původní alternativě, pro variantu s mediánem jsou pak některé hodnoty dokonce nižší než dosažená hladina v hypotéze. Opět se zde projevuje rozdílná velikost skupin. Pro vyvážený design jsou výsledky podobné výsledkům pro původní hypotézu. Tím, že je alternativa lineární kombinace blízká nulové hypotéze, pohybují se dosažené síly kolem hodnoty 0,075 pro varianty s centroidem a 0,030 pro varianty s mediánem. Ve všech případech je pak síla varianty s centroidy větší než varianty s mediány.

Závěr

Pro data generovaná z Poisson-log normálního rozdělení testy ANOSIM a MRPP, na rozdíl od testu homogenity disperze, dodržely předepsané hladiny. Hladina

Jaro	FC	PC	FM	PM
chyba 1. druhu	0,296	0,294	0,131	0,126
síla	0,595	0,594	0,360	0,363
Léto	FC	PC	FM	PM
chyba 1. druhu	0,312	0,312	0,131	0,130
síla	0,831	0,823	0,624	0,628
Podzim	FC	PC	FM	PM
chyba 1. druhu	0,279	0,277	0,116	0,114
síla	0,502	0,497	0,315	0,310

Tabulka 5.7: Odhady chyb 1. druhu a síly testu v alternativě. Varianty středů skupin: C - centroid, M - medián, varianty určení P-hodnoty: F - F rozdělení, P - 999 permutací.

Jaro	FC	PC	FM	PM
chyba 1. druhu	0,082	0,081	0,033	0,031
síla	0,451	0,448	0,280	0,281
Léto	FC	PC	FM	PM
chyba 1. druhu	0,071	0,069	0,029	0,028
síla	0,824	0,819	0,647	0,645
Podzim	FC	PC	FM	PM
chyba 1. druhu	0,068	0,067	0,021	0,021
síla	0,340	0,336	0,194	0,192

Tabulka 5.8: Odhady chyb 1. druhu a síly testu v alternativě, všechny skupiny mají po 7 pozorováních. Varianty středů skupin: C - centroid, M - medián, varianty určení P-hodnoty: F - F rozdělení, P - 999 permutací.

Jaro	FC	PC	FM	PM
síla	0,359	0,357	0,166	0,162
Léto	FC	PC	FM	PM
síla	0,262	0,261	0,094	0,093
Podzim	FC	PC	FM	PM
síla	0,223	0,220	0,084	0,084

Tabulka 5.9: Odhady síly testu v alternativě lineární kombinace. Varianty středů skupin: C - centroid, M - medián, varianty určení P-hodnoty: F - F rozdělení, P - 999 permutací.

Jaro	FC	PC	FM	PM
síla	0,094	0,096	0,045	0,042
Léto	FC	PC	FM	PM
síla	0,065	0,059	0,018	0,020
Podzim	FC	PC	FM	PM
síla	0,073	0,077	0,028	0,029

Tabulka 5.10: Odhady síly testu v alternativě lineární kombinace, všechny skupiny mají po 7 pozorováních. Varianty středů skupin: C - centroid, M - medián, varianty určení P-hodnoty: F - F rozdělení, P - 999 permutací.

testu homogenity disperze z důvodů nestejné velikosti skupin dodržena nebyla. Po doplnění skupin na stejnou velikost byla hladina (alespoň pro variantu s mediánem) testu dodržena. To opět ukazuje na nevhodnost tohoto testu pro skupiny nestejných velikostí.

6. Závěr

V předchozích kapitolách této práce jsme se zabývali problematikou testování hypotéz pomocí testů, jež jsou založeny na matici vzdáleností. Takovéto testy se používají například pro data pocházející z oblasti ekologie, u kterých často není možné předpokládat splnění předpokladů klasických mnohorozměrných metod, nebo tyto metody vůbec není možné použít.

Představili jsme několik koeficientů nepodobnosti (např. hojně užívaný Bray-Curtisův koeficient), které se používají při analýze ekologických dat a upozornili na některé jejich výhody oproti eukleidovské vzdálenosti. Pomocí těchto koeficientů jsme pak zkonstruovali matici vzdáleností, na níž jsou založeny testy, kterými jsme se zabývali. Zároveň jsme v krátkosti představili metodu hlavních koordinát, kterou jsme využili jednak pro zobrazení dat a jednak jako součást testu homogenity disperzí.

Prvním testem, kterým jsme se zabývali, je Mantelův test pro testování nezávislosti vektorů. Ukázali jsme jeho původ v testu nezávislosti dvou náhodných veličin a poukázali na jeho mnohostrannost. Použití tohoto testu jsme pak ilustrovali na datech `mouchy`.

Dále jsme se zabývali testy ANOSIM a MRPP jako alternativou k analýze rozptylu jednoduchého třídění. Ukázali jsme, že tyto testy jsou ve skutečnosti variantami Mantelova testu pro matici vzdáleností (MRPP) resp. matici jejich pořadí (ANOSIM) a vhodně zvolenou maticí vah. Právě typ použitých vah výrazně ovlivňuje vlastnosti příslušných variant těchto testů. Dále jsme upozornili na problematiku interpretace těchto testů. Zatímco se tyto testy používají pro test shody polohy, jedná se ve skutečnosti o testy celkové shody rozdělení. S testy MRPP a ANOSIM pak úzce souvisí test homogenity disperzí, kterým jsme se rovněž zabývali. Použití všech těchto testů jsme opět ilustrovali na datech `mouchy`.

V poslední části jsme pomocí simulací zkoumali vlastnosti těchto testů na změny parametrů ve dvojrozměrném normálním rozdělení. To nám umožnilo od sebe rozdělit parametry polohy a parametry rozptylu a dodat korelaci mezi složkami. Upozornili jsme na rozdílné chování variant testů ANOSIM a MRPP pro rozdílné velikosti skupin a citlivost těchto testů na rozdíly jak v parametrech rozptylu tak v korelacích složek. Zároveň jsme poukázali na rozdíly mezi testem homogenity disperzí při použití centroidů a při použití mediánů, kdy se varianta s centroidy ukázala být anti-konzervativní. Rovněž jsme ověřili výhodnost použití F rozdělení v tomto testu. Na závěr jsme zkoumali chování těchto testů na datech generovaných z Poisson-log normálního rozdělení, kdy jsme upozornili na chování testu homogenity disperzí pro skupiny nestejných velikostí.

Příloha

Data mouchy.txt

Data `mouchy` byla získána v rámci výzkumného záměru MŠMT MSM 002162241. Obsahují četnosti výskytů 156 hmyzích druhů ve vodních plochách. Jednotlivá stanoviště (lokality) pocházejí ze tří oblastí: Karpaty - 7 stanovišť, Beskydy - 7 stanovišť, Kysuce - 3 stanoviště. Veškerá měření byla na každém stanovišti prováděna na jaře, v létě a na podzim, a to pokaždé dvakrát. Tyto informace jsou obsaženy v proměnných `oblast` (hodnoty Karpaty, Beskydy, Kysuce), `lokalita` (hodnoty 1,...,7 resp. 1,...,3), `cas` (SP - jaro, SU - léto, AU - podzim) a `mereni`(hodnoty 1,2). Proměnné `druh1`,..., `druh156` obsahují četnosti jednotlivých hmyzích druhů na příslušných stanovištích v příslušnou dobu. Spolu s druhovými četnostmi byly měřeny i fyzikální a chemické vlastnosti stanovišť, konkrétně pH, teplota a obsahy chemických látek (NH_4, NO_3, Cu, Ca, Mg, Na, K). Hodnoty jsou obsaženy v proměnných `pH` (stupně pH), `T` (°C), `NH4` , `NO3` , `Cu` , `Ca` , `Mg` , `Na` a `K` (vše mg/l).

Abychom se vyhnuli problémům s nezávislostí jednotlivých pozorování nahradili jsme vždy dvojici pozorování pocházejících ze stejného stanoviště a stejného ročního období jejich průměrem. Na četnosti jsme následně aplikovali transformaci $\sqrt[4]{x}$ (navrhována např. ve Warton a Hudson (2004)), abychom snížili vliv početně zastoupených druhů na výslednou nepodobnost. Na takto upravených datech jsou ilustrovány veškeré postupy v této práci.

oblast	lokalita	cas	mereni	pH	T	NH4	NO3	Cu
Karpaty	1	SP	1	7,4	9	0,03	8	1
Karpaty	1	SP	2	7,4	9	0,03	8	1
Karpaty	1	SU	1	7,8	9	0,03	8	1
Karpaty	1	SU	2	7,8	9	0,03	8	1
Karpaty	1	AU	1	7,8	11	0,03	8	1
Karpaty	1	AU	2	7,8	11	0,03	8	1

Ca	Mg	Na	K	druh1	druh2	···	druh155	druh156
86,9	2,4	6,3	1,6	1	0	···	0	0
86,9	2,4	6,3	1,6	0	0	···	0	0
86,9	2,4	6,3	1,6	0	0	···	0	0
86,9	2,4	6,3	1,6	1	0	···	0	0
86,9	2,4	6,3	1,6	0	0	···	0	2
86,9	2,4	6,3	1,6	0	0	···	0	0

Tabulka 6.1: Ukázka dat v souboru mouchy.txt .

Obsah přiloženého CD

Na CD přiloženém k této práci se nacházejí zdrojové kódy k programu `R`. Soubory `Kapitola2.r`, `Kapitola3.r`, `Kapitola4.r` a `Kapitola5.r` obsahují kódy k výpočtům a simulacím v příslušných kapitolách, soubor `MujAnosim.R` obsahuje kód pro výpočet testů ANOSIM a MRPP. V souboru `mouchy.txt` jsou obsažena data. Soubor `VysledkySimulaci.R` obsahuje kód k otevření souborů obsahuyjících výsledky simulací `VysledkySimulaciAnosim.txt` a `VysledkySimulaciLevene.txt`, ve kterých jsou uloženy kompletní výsledky simulací testů ANOSIM, MRPP a testu homogenity disperzí pro dvojrozměrné normální rozdělení z kapitoly 5.

Literatura

Aitchinson, J. a Ho, C. H. (1989). The multivariate poisson-log normal distribution. *Biometrika*, **76**, 643–653.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, **62**, 245–253.

Brown, M. B. a Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, **69**, 364–376.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika*, **48**, 305–308.

Chapman, M. G. a Underwood, A. J. (1999). Ecological patterns in multivariate assemblages: information and interpretation of negative values in ANOSIM tests. *Marine Ecology Progress Series*, **180**, 257–265.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, **18**, 117–143.

Cliff, A. D. a Ord, J. K. (1973). *Spatial autocorrelation*. Pion, London.

Deheuvels, P. (1981). Multivariate tests of independence. *Lecture Notes in Mathematics*, **861**, 42–50.

Dietz, E. J. (1983). Permutation tests for association between two distance matrices. *Systematic Zoology*, **32**, 21–26.

Edgington, E. S. (1995). *Randomization tests*. Marcel Dekker, New York, Third edition. ISBN 978-0824796693.

Excoffier, L., Smouse, P. E., a Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among dna haplotypes: application to human mitochondrial dna restriction data. *Genetics*, **131**, 479–491.

Fisher, R. A. (1935). *The Design of experiments*. Oliver Boyd, Edinburgh.

Friedman, J. H. a Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample test. *Annals of Statistics*, **7**, 697–717.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Gower, J. C. (1974). Algorithm AS 78: The mediancentre. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, **23**, 466–470.

Gower, J. C. a Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis. *Journal of th Royal Statistical Society. Series C, Applied Statistics*, **48**, 505–519.

Legendre, P. a Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69(1)**, 1–24.

Legendre, P. a Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.

Legendre, P. a Legendre, L. (1998). *Numerical Ecology*. Elsevier Science BV, Amsterdam, 2nd english edition edition. ISBN 0-444-89250-8.

Levene, H. (1960). *Contributions to Probability and Statistics*. Stanford University Press, Palo Alto, California. ISBN 0-471-36357-X.

Lingoes, J. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, **36**, 195–203.

Maa, J. F., Pearl, D. K., a Bartoszubski, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The Annals of Statistics*, **24**, 1069–1074.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Mantel, N. a Valand, R. S. (1970). A technique of nonparametric multivariate analysis. *Biometrics*, **26**, 547–558.

McArdle, B. H. a Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, **82(1)**, 290–297.

Mielke, P. W. (1978). Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique. *Biometrics*, **34**, 277–282.

Mielke, P. W. (1979). On asymptotic non-normality of null distributions of MRPP statistics. *Communication in Statistics - Theory and Methods*, **8**, 1541–1550.

Mielke, P. W., Berry, K. J., a Johnson, E. S. (1976). Multi-response permutational procedures for a priori classifications. *Communications in Statistics, Series A*, **5**, 1409–1424.

Mielke, P. a Berry, K. (2007). *Permutation Methods, A Distance Function Approach*. Springer, New York, Second edition. ISBN 978-0-387-69811-3.

O'Brien, P. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics*, **48**, 819–827.

Olsgard, F., Somerfield, P. J., a Carr, M. R. (1997). Relationships between taxonomic resolution and data transformation in analyses of a macrobenthic community along an estabilished pollution gradient. *Marine Ecology Progress Series*, **149**, 173–181.

Orlóci, L. (1967). An agglomerative method for clasification of plant communities. *Journal of Ecology*, **55**, 193–205.

Pillar, V. D. P. a Orlóci, L. (1996). On randomization testing in vegetation science: multifactor comparisons of relevé groups. *Journal of Vegetation Science*, **7**, 585–592.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *The Journal of the Royal Statistical Society*, **4**, 119–130.

Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Qüestiió*, **19**, 23–63.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons, Chichester, Second edition. ISBN 0-471-41889-7.

Smith, E. P., Pontasch, K. W., a Cairns Jr., J. (1990). Community similarity and the analysis of multispecies environmental data: a unified statistical appriach. *Water Research*, **24**, 507–514.

Smouse, P., Long, J. C., a Sokal, R. R. (1986). Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627–632.

Van Valen, L. (1978). The statistics of variation. *Evolutionary Theory*, **4**, 33–43.

Warton, D. I. a Hudson, H. M. (2004). A Manova statistic is just as powerful as distance-based statistics, for multivariate abundancies. *Ecology*, **85**, 858–874.

Whaley, F. S. (1983). The equivalence of three independently derived pemutation procedures for testing the homogeneity of multidimensional samples. *Biometrics*, **39**, 741–745.

Zuur, A. F., Ieno, E. N., a Smith, G. M. (2007). *Analysing Ecological Data*. Springer, New York. ISBN 0-387-45967-7.