

9

Eredmények összehasonlító értékelése

(Nincs megállás!)

Osztályozások, ordinációk és más eredmények előállításával rendszerint nem fejeződik be többváltozós adataink elemzése. Jó néhány okát lehet felsorolni annak, hogy számításainkat tovább kell folytatnunk. Mindig fennáll ugyanis a lehetőség, hogy sokféle módszert alkalmazva többé-kevésbé eltérő eredményeket kapjunk ugyanazon objektumokra! Bár módszereink önmagukban “objektívnek” tekinthetők, a vizsgálat során számos ponton kell szubjektív döntéseket hoznunk a mintavételezést (pl. kvadrátnagyság), a változók kiválasztását, az adat-típust és az adatok standardizálását és transzformálását, az alkalmazandó távolsági- vagy különbözőségi függvényt és az osztályozó vagy ordinációs algoritmust illetően – és ezt a sort még folytathatnánk. Ahhoz, hogy következtetéseinket eme döntések ne befolyásolhassák döntő mértékben, célszerű többféle elemzést végrehajtani. Ezáltal kiszűrhető az eredményekből az, amit csupán a módszertani változtatásoknak tudhatunk be. A párhuzamos elemzések *alternatív* eredményeket hoznak létre, s ezek *összehasonlítása* adhatja meg a keresett választ. Az összehasonlítás lényeges, bár nem mindig “kötelező” – s nem mindig nyilvánvaló – eleme az értékelésnek. Az is előfordul, hogy csupán egyetlen egy módszer ad sok és egyformán jó végeredményt (mint például a kladisztika parszimónia eljárásai), s ezek összesítése vezethet el a végső konklúzióhoz. Erre a problémakörre már többször utaltunk az előző fejezetekben, de részletezése mostanra, könyvünk záró fejezetére maradt.

Az összehasonlítás alapegységei

Az alternatív eredményeket most ugyanúgy objektumoknak tekintjük, ahogy tettük ezt a taxonokkal, a mintavételi egységekkel és másokkal a vizsgálat megelőző, fő szakaszában. Ezekre az objektumokra – a numerikus taxonómia OTU-jaival analóg módon – az OUC (“*operational unit of comparison*”, Podani 1989d), vagyis *összehasonlítási alapegység* néven hivatkozhatunk. Egy OUC tehát lehet távolságmátrix, egy dendrogram, egy partíció, egy ordináció, stb. Speciális objektumokról van szó, amelyek különleges eljárásokat, pontosabban: az objektumok sajátosságait tükröző leírást igényelnek. Két dendrogram például ugyanúgy

összevethető az euklidészi távolság alapján, mint két taxon, de az mindenképpen új számunkra, hogy a dendrogramok miféle jellemzői szolgáltatják az alapot eme távolság kiszámításához. Az előző fejezetek módszereit is bevethetjük majd, hiszen sok OUC közötti páronkénti távolságok mátrixa már hasonló módon értékelhető, mint a hagyományos OTU-k távolságmátrixai.

9.1 Választási lehetőségek

Az összehasonlításokat sokféleképpen végrehajthatjuk, s ennek megfelelően a téma meglehetősen szövevényes. Didaktikai szempontból is, és a jobb tájékozódás érdekében is célszerű az összehasonlítási lehetőségeket csoportosítani. Számos ponton két-két lehetőség közül kell választanunk – akár tudatos, akár nem ez a döntés – (Podani 1989d). A választások egy része szépen beilleszthető egy dichotomikus döntési fába (9.1 ábra), az utolsó három pedig általános érvényű, a fa több pontján is számításba jöhet.

9.1.1. Az alapegységek típusa azonos vagy eltérő?

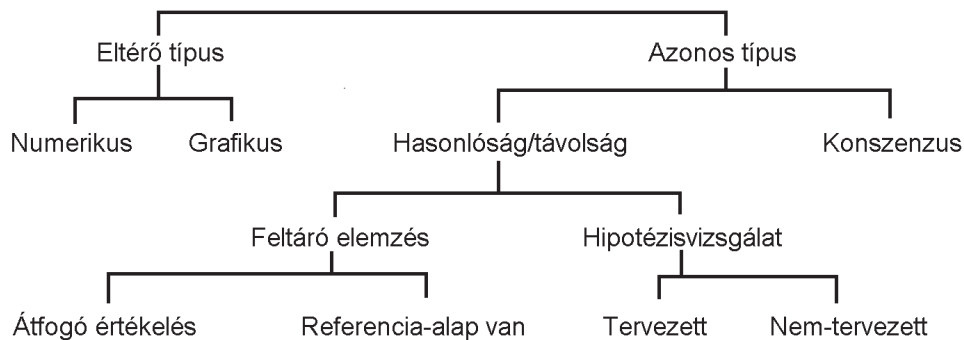
Az összehasonlítást megtehetjük pl. két partíció, vagyis *ugyanolyan típusú* eredmény között. Ez még ismeretlen, újszerű feladat számunkra, arra viszont már eddig is láttunk példákat, hogy az összehasonlított objektumok *nem azonos jellegűek*. A dendrogram összehasonlítása a mátrixszal, amiből származik (kofenetikus korreláció, 5.5.1 rész) voltaképpen eltérő típusú eredmények hasonlóságának kvantitatív kifejezése. Nem is annyira numerikus összehasonlítás, hanem inkább együttes ábrázolás, vagyis grafikus összehasonlítás volt a plexus gráf ábrázolása az ordinációs térben (8.10b ábra). További példákat látunk majd a 9.5.2 részben.

9.1.2. Hasonlóság/távolság vagy konszenzus?

Erre a kontrasztra már utaltunk az imént. Két OUC hasonlósága vagy távolsága egy számértékkel fejezhető ki, s $k > 2$ OUC összevetése minden lehetséges párosításban (*többszörös összehasonlítás*) pedig egy mátrixot eredményez, amelyből pl. osztályozhatunk (akár osztályozások osztályozása, egyfajta “meta-analízis” is elképzelhető). Ugyanazt a k darab eredményt azonban egyesíthetjük is egy $k+1$ -edik eredményben, ami az egyezéseket és különbözőségeket egyaránt mutathatja. Ez a szintézis a *konszenzus* OUC.

9.1.3. Hipotézisvizsgálat vagy feltáró elemzés?

A kutatót igen komolyan érdekelheti az a kérdés, hogy két OUC hasonlósága statisztikailag *szignifikáns*-e vagy sem, vagyis a feltáró szakaszt lezárhatjuk egy hipotézisvizsgálattal is – ami igazán nem volt jellemző az eddigiekre. Ehhez azonban két fontos kívánalomnak kell teljesülnie. Először is a két összehasonlítandó OUC-nak *független*nek kell lennie egymástól, például nem származhatnak ugyanabból az adathalmazból (ez alapvető követelmény általában). Két dendrogram hasonlóságának a szignifikanciáját tehát akkor érdemes tesztelni, ha mondjuk az egyik az A változócsoporthon, a másik pedig a B változócsoporthon alapuló osztályozást képvisel. Ekkor az a kérdés, hogy szignifikáns-e a két dendrogram hasonlósága, mert ha igen, akkor kimondható: a két változócsoporthon hasonló klasszifikációt implikál. A második feltétel a hasonlóságot mérő statisztika *eloszlásának* az ismerete. Mivel – néhány kivétellel – a statisztikák eloszlását nem ismerjük, ezt Monte Carlo szimulációval közelítjük, hogy a kérdéses



9.1 ábra. A többváltozós eredmények összehasonlítása során felmerülő választási lehetőségek fa-diagramja.

hasonlósági érték szignifikanciáját eldönthessük. A módszer lényege az, hogy több száz vagy annál is több véletlenszerűen generált OUC-párt előállítunk, mindegyikre kiszámítjuk a hasonlóságot, majd a kategóriákba osztott értékek gyakoriságeloszlását hisztogramban ábrázoljuk, s ebben megkeressük a tesztelni kívánt érték helyét.

Ha a szignifikancia-próbáról a függetlenség feltételének megsértése miatt nem lehet szó, a hasonlóságértékek *feltáró* funkciója még megmarad. $k > 2$ esetén tovább mehetünk a 9.1.5 választás szerint. Ha azonban csupán két OUC összevetésére szorítkozik a vizsgálat, akkor a kapott hasonlóságérték önmagában szinte semmit sem mond nekünk. Ezért ekkor megtehetjük – sőt meg is kell tennünk – azt, hogy mégis összehasonlítjuk az eloszlással, s annak a várható értékével, de ekkor tartózkodjunk mindennemű, az eredmény “szignifikanciájára” utaló megállapítástól.

9.1.4 Tervezett és nem-tervezett összehasonlítások

Ha a szignifikancia-próba alkalmazható, és többszörös összehasonlítást végzünk, akkor még egy dologra kell ügyelnünk. Az egyszempontos és többmintás variancia-elemzés utáni szignifikáns differencia meghatározásával (Sokal & Rohlf 1981a) analóg helyzettel állunk szemben: a szignifikáns hasonlóságok kiválogatása a mátrixból egyértelműen az I. típusú (“elsőfajú”) hiba halmozódására vezet, vagyis több párosításban találunk szignifikáns hasonlóságot, mint amennyi egy adott szignifikancia szinten (pl. $p < 0,05$) fennáll. Amennyiben a teszt végrehajtása előtt kimondjuk, hogy csak néhány kitüntetett, egymástól független párosítás érdekel bennünket (*tervezett* összehasonlítások, “*planned comparisons*”), akkor elkerüljük ezt a problémát, s a szimulált eloszlás nyugodtan alkalmazható ama néhány érték szignifikanciájának eldöntésére. Ha viszont nincsenek kitüntetett párok (*nem tervezett* összehasonlítások, “*unplanned comparisons*”), vagyis minden párosítás érdekel bennünket, akkor a problémát a teszt szigorításával oldhatjuk meg, pl. a $k(k-1)$ összehasonlításra kapható *minimumok eloszlásának* a szimulációjával (részleteket lásd a 9.3.6 részben).

9.1.5 Átfogó összehasonlítás vagy referenciához való viszonyítás?

Az első esetben az eredmények között *egyik sincs kitüntetve*, ezért az összehasonlításokat minden lehetséges párban elvégezzük. A kapott értékek relatív nagysága lesz a fő információhor-

dozó a soronkövetkező meta-elemzésben. Erre nincs feltétlenül szükség, ha az egyik eredményt valamilyen szempontból kitüntetjük, és a többieket csak ezzel hasonlítjuk össze. Például egy, a változók összességén alapuló ordináció lehet a viszonyítási alap, és megvizsgálhatjuk, hogy az – előzőekben rangsorolt – változók fokozatos elhagyogatása miképpen befolyásolja a referenciától való eltérést. A kitüntetett eredmény tehát a “*kontroll*” szerepét tölti be az összehasonlításban.

9.1.6. Kongruencia-elemzés vagy algoritmikus hatások kiszűrése?

Erre a megkülönböztetésre Rohlf & Sokal (1981b) és Gower (1983) munkái hívták fel figyelmünket, és a dichotomikus döntési fában már nem érdemes feltüntetni, mert több ponton is releváns lehet. Lényegében véve az elméleti és a technikai/módszertani aspektusok különválasztásáról van szó. Az alternatív eredmények közötti eltéréseket biológiai megfontolások eredményezhetik, ekkor *kongruencia*-elemzést végzünk (pl. rovaraxonok klasszifikációinak összevetése az imágó, ill. a lárvastádium alapján, a *taxonómiai kongruencia* értékelésére). Ezt az esetet célszerű elválasztani attól, amikor az eredmények közötti különbségeket csupán az alkalmazott algoritmus eltérései vagy más technikai változtatások okozzák.

9.1.7. Elemi vagy komplex összehasonlítás?

Ha az értékelendő eredmények eltéréseit csupán egyetlen tényező megváltozása okozza, akkor *elemi* összehasonlításról beszélünk. Annak vizsgálatában, hogy mondjuk az osztályozás stratégiája miképpen befolyásolja a kapott eredményt, az elemzés más szempontjait (pl. adat-típust, távolságfüggvényt) nem szabad közben változtatgatni – vagyis minden más legyen konstans. Ha nem vagyunk körültekintőek, akkor több hatás keveredése jelentkezik az eredményben és ez hamis következtetésekre vezet (vö. Kenkel & Orlóci 1986). Az ilyen keveredés többször is előfordul a szakirodalomban, mint gondolnánk. Ha azonban két (v. több) tényezőt szisztematikusan, vagyis minden kombinációban kipróbálunk, akkor ezek relatív fontossága kimutathatóvá válik (*komplex* összehasonlítások, Podani 1989d).

9.1.8. Egy- vagy többváltozós értékelés?

Az összehasonlítás egyváltozós, ha az értékelt eredmények egyetlen egy tulajdonságát vesszük csak figyelembe (pl. dendrogramok összehasonlítása a topológiai differenciák szerint, l. 9.2.3). Meglepő módon az elemzések túlnyomó többsége ilyen típusú, holott a vizsgálat összes megelőző lépése lényegében véve többváltozós volt! Podani & Dickinson (1984) mutattak rá, hogy dendrogramok esetében például az összehasonlítás több szempontot is figyelembe vehet egyszerre, s ezáltal az egész vizsgálat sorozat többváltozós jellegűvé alakítható. Ez más típusú, viszonylag bonyolult szerkezetű eredmények (kladogramok, additív fák) esetében is elképzelhető.

9.2 Eredmények páronkénti összevetése

A legtöbb vizsgálat alapja a páros összehasonlítás, ami még a konszenzus eredmények létrehozásában is szerepet játszhat, így ezt a témát érintjük először (holott a 9.1 ábra döntési fájában nem ez az első választás). Az összehasonlítás módszereit sorjában, az eredmények típusai szerint vesszük át. Bár a távolságmátrixok eddig a vizsgálat közbülső állomásaként, s nem

végeredményként jöttek számításba, elsőnek mégis róluk szólunk, hiszen fejezetünk témáján belül sok minden a mátrixok összevetésére vezethető vissza.

9.2.1 Mátrixok összehasonlítása

Numerikus módszerek. Két szimmetrikus távolság- (hasonlóság-, stb.) mátrixot, amelyet jelöljön \mathbf{D} és \mathbf{E} , úgy hasonlíthatunk össze számszerűen, hogy mindkettő felső félmátrixát képzeletben egy-egy oszlopvektorra “nyújtjuk”, s ezután már válogathatunk is a 3. fejezetben megismert összehasonlítási lehetőségek közül. Legszélesebb körben a korrelációs együtthatót (3.70) alkalmazzák, *mátrix korreláció* (Sneath & Sokal 1973: 280) néven:

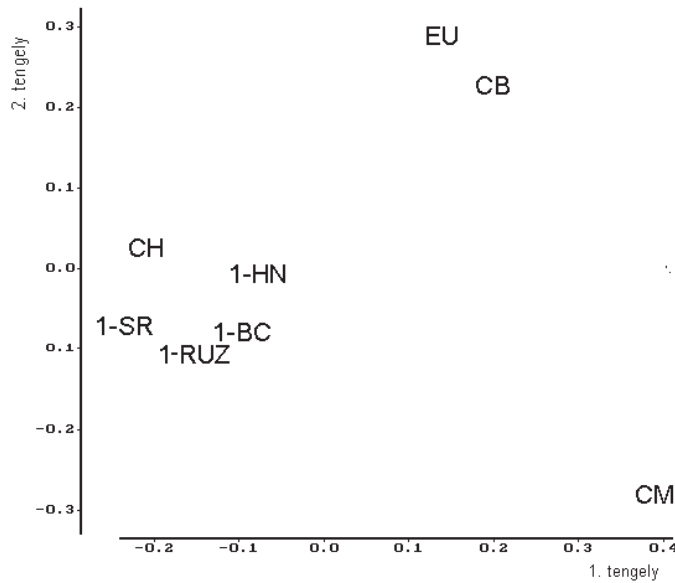
$$r_{DE} = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^m (d_{ij} - \bar{d})(e_{ij} - \bar{e})}{\sqrt{\sum_{i=1}^{m-1} \sum_{j=i+1}^m (d_{ij} - \bar{d})^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m (e_{ij} - \bar{e})^2}} \quad (9.1)$$

amelyben \bar{d} és \bar{e} a \mathbf{D} ill. \mathbf{E} mátrixok értékeinek az átlaga. Az átlós értékeket kihagyjuk az összehasonlításból. Ha \mathbf{D} és \mathbf{E} hasonló tendenciákat tükröz az eredeti objektumok összehasonlításában, függetlenül a távolságértékek abszolút nagyságrendjétől, akkor a mátrix korreláció értéke 1-hez közeli lesz. r_{DE} a $[-1, 1]$ tartományba eső értékeket vehet fel egyébként. Hangsúlyozandó, hogy a korreláció alkalmazása teljesen formális ebben az esetben, hiszen a mátrix egyes értékei egymástól természetesen nem függetlenek, így r_{DE} “szignifikanciája” a hagyományos módon semmiképpen sem értékelhető (l. a 9.3.1 részt). \mathbf{D} és \mathbf{E} euklidészi távolságot is kiszámíthatjuk az alábbi képlet felhasználásával:

$$d_{DE} = \left(\sum_{i=1}^{m-1} \sum_{j=i+1}^m (d_{ij} - e_{ij})^2 \right)^{1/2} \quad (9.2)$$

Emellett más lehetőségek is elképzelhetők, de a gyakorlatban ez a két függvény szerepel a legtöbbször. A korreláció leginkább a grafikus értékeléssel együttesen alkalmazva informatív, és amikor az összehasonlítandó mátrixok nem összemérhetők egymással. Többszörös összehasonlítás utáni meta-elemzésre a korreláció komplementjét alkalmazzuk, amint az alábbi példa szemlélteti. Az euklidészi távolságnak csak akkor van ugyanis értelme, ha az összehasonlítandó mátrixok értékei azonos skálán mozognak.

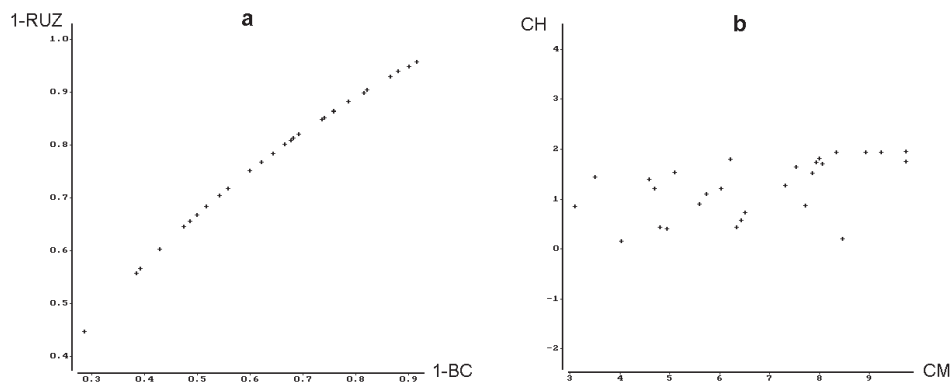
Az A1 táblázat objektumainak különbözőségeit többféle együttható felhasználásával kiszámítjuk, majd a kapott mátrixokat minden párosításban összehasonlítjuk. A 9.1 formulát választjuk, mivel a nyolc függvény (l. a 9.2 ábra jelmagyarázatát) eltérő skálákon méri a különbözőséget. A komplementek előállítását után eredményül a mátrixok különbözőségi mátrixát kapjuk, amelyet főkoordináta-módszerrel elemzünk. Az ordinációban az első két tengely a távolságviszonyok 84 %-át magyarázza, ami elég jó információ-sűrítésnek számít. A diagram jól szemlélteti a nyolc együttható egymás közötti viszonyát az illetQQQ vizsgálatban. A baloldali ötös csoport egymáshoz nagyon hasonlóan viselkedik, csakúgy mint az euklidészi és a Manhattan metrika. Ez utóbbi például azért érdekes, mert az euklidészi távolságban az eltérések négyzetesen, a másikban pedig abszolút értékben összegződnek. A Canberra metrika viszont, az összegzésen belüli osztás miatt, mindegyiktől külön áll. Egészen más adatokra is könnyen kapunk hasonló eredményt (mint pl. Podani 1994: 191), ami jelen következtetéseink általános érvényességét sugallja.



9.2 ábra. Nyolc különbözőségi index révén kapott távolságmátrixok főkoordináta-elemzése. Jelmagyarázat: 1-BC: Bray - Curtis, 1-RUZ: Ruzicka, 1-SR: hasonl. hányados, 1-HN: Horn index, CM: Manhattan m., CB: Canberra m., EU: euklidészi távolság, CH: hűrtávolság.

Grafikus eljárás. Egy koordináta-rendszert készítünk, amelyben a pontok a j_k objektum-párokat képviselik, d_{jk} a vízszintes tengelyen, e_{jk} pedig a függőleges tengelyen felvett koordináták. A kapott szórásdiagram (*matrix plot*, Rohlf 1993a) értelmezése hasonló a Shepard-diagraméhoz (7.4.2 rész). A két mátrix hasonlósága – vagyis lineáris korrelációja – annál nagyobb, minél inkább illeszkednek a pontok egy képzeletbeli egyenesre.

Mindezt be is mutatjuk a fenti példából származó két párosításban, az egymáshoz közel álló 1-BC, 1-RUZ ($r=0,994$) és a CM, CH ($r=0,522$) párokat kiemelve. Az első esetben tehát csaknem tökéletesen lineáris az összefüggés, a második példában viszont jóval gyengébb (pl. a



9.3 ábra. Mátrixok grafikus összehasonlítása. **a:** 1-Bray Curtis és 1-Ruzicka, **b:** Canberra metrika (CM) és hűrtávolság (CH) az A1 mátrix oszlopaira számítva.

legkisebb hűrtávolságot adó két objektumpár a Canberra metrikát tekintve már egyáltalán nem azonos távolságra van egymástól).

9.2.2 Partíciók összehasonlítása

Erre a feladatra a módszerek egy része a mátrix összehasonlítás stratégiáját alkalmazza. Más eljárások a kereszt-partíciókból indulnak ki, míg a módszerek egy harmadik csoportja partíciók egymásba alakításának lehetőségeit értékeli.

Mátrix összehasonlítások. Bármely P partíciót felírhatunk egy $m \times m$ -es szimmetrikus incidencia mátrix alakjában, jele C_P , amelyben $c_{gh}=1$ ha a g és h objektumok ugyanabba az osztályba tartoznak, más esetben pedig $c_{gh}=0$. Ezután a P és Q partíciók hasonlósága (vagy különbözősége) a megfelelő C_P és C_Q incidencia mátrixok összehasonlításával kapható meg, a bináris adatokra alkalmas koeficiens (szinte) bármelyikének felhasználásával (3.2 alfejezet). Ebben az esetben a 2×2 -es kontingencia-táblázat a értéke azoknak az objektumpároknak a száma, amelyek mind a két összehasonlítandó partícióban együtt vannak. A partíciók összehasonlítására leggyakrabban alkalmazott formulákat nem ismételjük meg, csak felsoroljuk azokat, megemlítve az irodalomban szereplő neveket (amelyek “természetesen” nem egyeznek meg a különbözőségi indexek irodalmából ismert elnevezésekkel):

- egyezési koeficiens (3.6 formula, = “Rand” index, Rand 1971),
- euklidészi távolság (3.7 formula, = “PAIRBONDS”, Arabie & Boorman 1973),
- Jaccard index (3.24 formula, Downton & Brennan 1980),
- Sorensen index (3.25 formula, = “percent mutual matches”, Arabie & Boorman 1973) és
- Ochiai index (3.26 formula, Fowlkes & Mallows 1980).

Az euklidészi távolság kivételével mindegyiket érdemes a *komplement* formájában kifejezni. Ekkor a partíciók teljes azonosságára 0 értéket kapunk minden formulával. A maximális eltérésre azonban nem kapunk 1-es különbözőséget, mert a partíciók szükségképpen meg egyeznek bizonyos objektumpároknak, vagyis a értéke sosem lehet 0. Nem tudunk ugyanis m objektumra két olyan partíciót generálni (a triviális eseteket leszámítva), amelyekben ne lennének mindkét partícióban együtt lévő objektumpárok. Annak érdekében, hogy a különböző szituációkból származó különbözőség-értékek egyáltalán összemérhetőek legyenek egymással, standardizálást kell alkalmaznunk, felhasználva a random partíciókra kapható *várható* értéket, és a lehetséges *maximumot*, az alábbiak szerint:

$$\frac{\text{Aktuális érték} - \text{Várható érték}}{\text{Lehetséges maximum} - \text{Várható érték}}, \quad (9.3)$$

(vö. Hubert & Arabie 1985). A szerzők rámutatnak, hogy a maximum megállapítása a kombinatorikus optimalizálás (nehéz) témakörébe tartozik. Podani (1986) heurisztikus kereső módszereket javasolt a maximum közelítésére. Részleges, de jó megoldást jelenthet az aktuális értékek összehasonlítása a szimulációból származó várható értékkel és egy adott szignifikancia-szinten adódó küszöbértékkel.

Kereszt-partíciók. A blokk-osztályozásból már megismert kereszt-partíciók most egy kontingencia-táblázatként értelmezendők, amelyben a sorok a P-szerinti osztályozásnak, míg az oszlopok a Q-szerinti osztályozásnak felelnek meg. Ha a két partícióban lévő osztályok száma s illetve t , akkor ez egy $s \times t$ méretű táblázat, amelyben az ij cella értéke azoknak az objektu-

moknak a száma lesz, amelyek P-ben az i osztályba, Q-ban pedig a j osztályba tartoznak (vagyis mindig együtt vannak). Például, 10 objektumra az alábbi két partíció:

$$1: \{ 1, 2, 3, 4, 5 \} \{ 6, 7, 8, 9, 10 \}$$

$$2: \{ 1, 2, 3, 6, 7 \} \{ 4, 5, 8, 9, 10 \}$$

kereszt-partíciós táblája a következő

	2. Partíció	
	1 osztály:	2 osztály:
1. Partíció	1 osztály:	2 osztály:
	3	2
	2	3

amely megfelel az alábbi részhalmazoknak:

$$\{ 1, 2, 3 \} \{ 4, 5 \}$$

$$\{ 6, 7 \} \{ 8, 9, 10 \}$$

Általánosságban a kontingencia-táblázat a következőképpen írható fel:

		Q			
		q_1	q_j	q_t	
P	p_1				
	p_i	n_{ij}			n_i
	p_s				
		n_j			$n_{\cdot} = m$

Amelyben a marginális összegek a P és Q osztályméretei, m pedig a főösszeg, vagyis az osztályozott objektumok száma. A táblázatot a jól ismert χ^2 statisztikával (3.36 formula) értékelhetjük ki, amely a fenti jelölésekkel a következő:

$$X^2 = \sum_{i=1}^s \sum_{j=1}^t \frac{(n_{ij} - n_i n_j / m)^2}{n_i n_j / m} \quad (9.4)$$

Zérus értéket kapunk, ha a P partíció összes csoportja egyenlő mértékben szét van szórva Q csoportjai között, míg maximális érték adódik ha $P=Q$. Ez a maximum $m \times \min [(s-1), (t-1)]$, amellyel a Cramér-indexnél (3.37) ismert módon standardizálhatunk, s a koefficiens értéke a $[0,1]$ intervallumba kerül.

A már szintén bemutatott Goodman - Kruskal (1954) féle lambda (3.38-3.39) is alkalmas partíciók összehasonlítására. A függvény interpretációja most a következő. Gondoljuk el, hogy először egy objektum Q-beli osztályát szeretnénk eltalálni anélkül, hogy ismernénk, melyik osztályba is tartozik P-ben. A legjobb próbálkozás nyilván a Q legnagyobb osztálya lesz, vagyis megkeressük a $\max_j [n_j]$ -t, hiszen ez minimalizálja a téves találatok számát. Ha azonban tudjuk, hogy az objektum P-nek az i -edik osztályába tartozik, akkor csak az i -edik sor lesz érdekes a táblázatban, és ennek a sornak a legnagyobb értékét kell kiválasztanunk, vagyis

$\max_j [n_{ij}]$ -t. P-t ismerve tehát a Q-beli osztályba tartozás eltalálásának hibája az alábbiak szerint csökken:

$$LAS_{PQ} = \frac{\sum_{j=1}^s \max_j [n_{ij}] - \max_j [n_{.j}]}{m - \max_j [n_{.j}]} \quad (9.5)$$

amely a *megjósolhatóság (prediktabilitás)* vagy a *predikciós erő* egy aszimmetrikus mértéke. Értéke 0, ha P semmiféle információt nem ad Q-ról, és 1 ha a két partíció azonos. Ez tehát akkor hasznos, ha az összehasonlításokat egy referenciával végezzük (vö. 9.1.5). A *kölcsönös prediktabilitás* szimmetrikus mértékszámát, vagyis magát a Goodman-Kruskal féle lambdát, a következőképpen állítjuk elő:

$$\Lambda_{PQ} = \frac{\sum_{i=1}^s \max_j [n_{ij}] + \sum_{j=1}^t \max_j [n_{ij}] - \max_j [n_{.j}] - \max_i [n_{i.}]}{2m - \max_j [n_{.j}] - \max_i [n_{i.}]} \quad (9.6)$$

melynek értéke szintén 0-tól 1-ig terjed. Ennek komplementje P és Q távolságaként fogható fel.

Megjegyzendő, hogy a kereszt-partíciók és a mátrix összehasonlítások között természetesen van formális kapcsolat, s az egyik kifejezhető a másikból. Például, a hasonlóságoknál figyelembe vett a érték a kereszt-partíció jelöléseivel $[\sum \sum n_{ij}^2 - m]/2$.

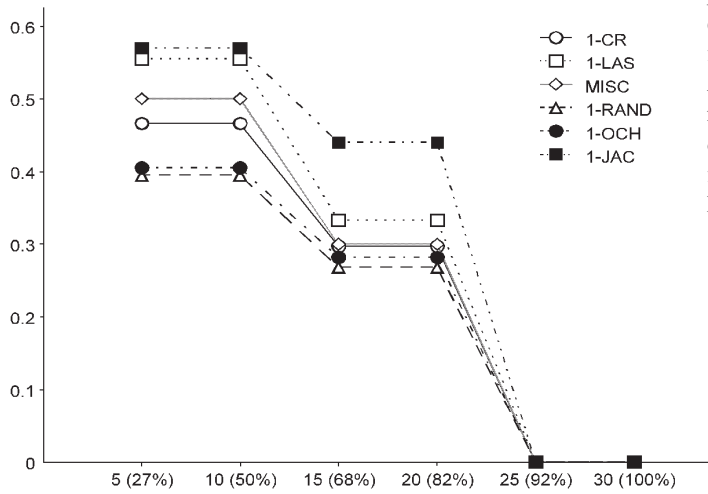
Transzformációs metrikák. A partíciókra speciálisan alkalmas eljárások azt vizsgálják, hogy hány elemi lépésre van szükség ahhoz, hogy P-t átalakítsuk Q-vá. A Day-féle (Day 1981) *átmeneti metrika (MINDMT, "min. divisions, mergences and transfers")* a legegyszerűbb, mert ez a P osztályai között átcsoportosítandó objektumok száma ahhoz, hogy Q-t előállítsuk. Ha $s=t$, akkor a kereszt-partíciós táblázatot egy \mathbf{Z} mátrixba transzformáljuk oly módon, hogy az átlós elemek összege maximális legyen, tehát az áthelyezendő objektumok minimális száma:

$$MINDMT_{PQ} = m - \text{tr} \{ \mathbf{Z} \} \quad (9.7)$$

Ha $s \neq t$, akkor a fenti formula alkalmazásához üres osztályokat adunk a kevesebb csoportot tartalmazó partícióhoz. A 9.7 formula maximumát is érdemes meghatározni, mert ezáltal egy standardizált mértéket kaphatunk, ami a különböző szituációból származó eredményeket összehasonlíthatóvá teszi:

$$MISC_{PQ} = \frac{m - \text{tr} \{ \mathbf{Z} \}}{\max \{ MINDMT \}} \quad (9.8)$$

("misclassification" index). Ennek szélső értékei 0 (teljes azonosság esetén) és 1 (ha a maximális számú áthelyezésre van szükség a P partíció Q-ba alakításához). A maximum akkor jelentkezik, amikor a kereszt-partíciós táblázatban a lehető legegyszerűsebbek az értékek. Day (1981) sok más mértékszámot is javasolt, például a *szigma-metrikát*, amely az eddig tárgyalt módszereket kombinálja:



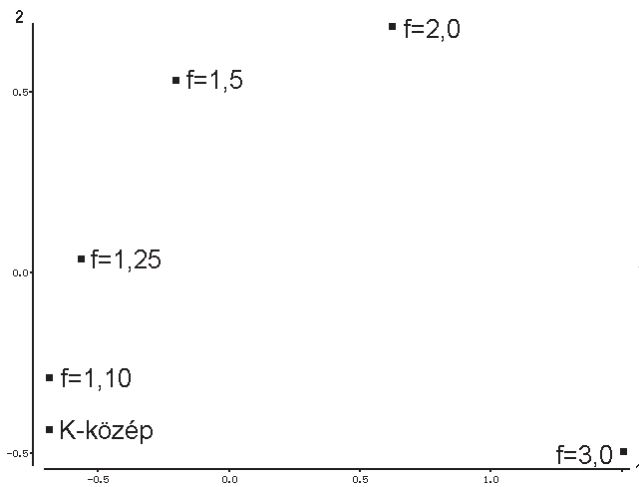
9.4 ábra. Csökkenő fajszám alapján nyert partíciók összehasonlítása egy referencia-alaphoz (30 faj, 100 %) hat módszerrel. Az osztályozások a binarizált A4 táblázat oszlopaira készültek. A függőleges tengelyen a különbözőséget mértük fel.

$$\sigma_{PQ} = 2a + b + c - 2 \sum_{i=1}^s z_{ii}^2 + 2 \sum_{i=1}^s z_{ii} \quad (9.9)$$

amelyet ugyancsak a $\text{tr} \{Z\}$ maximalizálásával kapunk meg. A megoldás azonban nem feltétlenül unikális, mert a keresztpartíciós tábla átlós celláira többféleképpen adódhat egyforma négyzetösszeg, s ezekhez különféle a , b és c értékek tartoznak.

Példaképpen a következő vizsgálatsorozatot hajtjuk végre. Az A4 táblázat adatait először binárisá alakítjuk. A 30 fajt egyszerűen rangsoroljuk az 5.8 kritérium alapján (lásd 8.1.1 rész). Ezután az index-független módszerrel (4.1.2 rész), az egyezési index (3.6 formula) szerint osztályozzuk a 20 objektumot a teljes fajkészletből, illetve a kevésbé fontos fajok fokozatos elhagyásával kapott részmatrixokból 2 osztályba. Az összes fajon alapuló klasszifikáció a referencia alap, amihez a többi viszonyítva az eredmény egy vonaldiagrammal ábrázolható (9.4 ábra). Ily módon szemléltethető a figyelembe vett változók száma és a klasszifikáció közötti összefüggés, és összehasonlíthatók a különféle indexek is, amelyek közül csak a $[0,1]$ intervallumba eső függvényeket illusztráljuk (kimaradt a PAIRBONDS és a szigma metrika). Az első 5 faj kiesése nem változtatja meg az osztályozást, a következő ötös csoporté viszont igen, s 20 és 15 fajra azonos eredményt kapunk, ezután 10 és 5 fajra is. Különösen érzékenyen reagál a kezdeti változásra az 1-JAC, amely később már csak kismértékben tud tovább növekedni. Mivel a maximum áthelyezések száma 10, a MISC együttható értékéből megkapjuk, hogy először 3, majd 5 objektum helyeződött át a kezdeti partícióhoz képest. Leginkább ez az együttható tükrözi a többi index "átlagos" viselkedését.

Lágy partíciók összehasonlítása. Fuzzy osztályozások összehasonlítása az $U_{m,c}$ matrixokon alapszik, amelyben $0 \leq u_{jk} \leq 1$ a j objektum k -adik osztályba tartozásának a mértékét fejezi ki. Adott F és G lágy partíciók tehát megfeleltethetők az U_F és U_G matrixokkal. Podani (1990) javaslata szerint F és G különbözőségének mértékét az U_F -ben lévő értékek minimális megváltoztatásával (ill. ennek négyzetösszegével) mérhetjük, amely ahhoz szükséges, hogy F-et G-be alakítsuk át. A mértékszám szimmetrikus, és az U_F oszlopai összes permutációjának előállításával kapható meg, miközben U_G változatlan marad. Minimalizálandó tehát a



9.5 ábra. Osztályozási sor illusztrációja főkoordináta ordinációval. Az *Iris* egyedek (A2 táblázat) k -közép osztályozása az első lépés, amely a lágy osztályozással az f paraméter növelése mellett folytatva egyre inkább eltér a kiindulástól. A patkó jelenség, mint látjuk, az eredmények összehasonlításában is felbukkanhat.

$$\Delta_{FG} = \left(\sum_{j=1}^m \sum_{k=1}^c (u_{Fjk} - u_{Gjk})^2 \right)^{1/2} \quad (9.10)$$

mennyiség, amelyben c az osztályok száma. A permutációk gond nélkül előállíthatók hét osztályig, nagyobb c érték amúgy sem fordul gyakran elő. A fenti függvény akkor is használható, ha F-ben és G-ben nem egyforma az osztályok száma, csak üres osztály(oka)t kell adnunk a kisebb osztályszámú partícióhoz. A 9.10 függvény “kemény” partíciókra is alkalmazható természetesen, hiszen azok a lágy osztályozások speciális eseteinek tekinthetők¹ (minden objektumra nézve egy súlyérték 1, a többi 0). Megmutatható a 9.10 és 9.7 között fennálló egyszerű kapcsolat is: $\Delta^2 = 2MINDMT$. A 9.7 formula maximuma felhasználható tehát a 9.10 mennyiség standardizálására.

A módszert az *Iris* adatok felhasználásával illusztráljuk. A 150 egyed 3 osztályba soroljuk a fuzzy osztályozás segítségével, a lágysági paraméter változtatásával (az alkalmazott értékek 1,10, 1,25, 1,5, 2,0 és 3,0). Mivel a paraméternek 1-es értéket – szingularitási problémák miatt – nem adhatunk, kemény partícióként a k -közép osztályozást vesszük tekintetbe. Ily módon egy klasszifikációs sort kapunk. Az eredményeket minden párosításban összehasonlítjuk, és PCoA ordinációval ábrázoljuk (9.5 ábra). Az első két tengely 80 %-os részesedése azt jelenti, hogy az osztályozások kapcsolatrendszere hatékonyan megmutatható a koordinátarendszerben. A kemény osztályozás felől a leginkább fuzzy felé haladva valójában egy “grádiens”, fokozatos eltávolodás alakul ki, amelyet a már jól ismert patkó-jelenség ellenére könnyen felismerhetünk.

9.2.3 Dendrogramok, kladogramok összehasonlítása

A dendrogramok és kladogramok az eddigi OUC típusoknál bonyolultabb struktúrák, és összehasonlításuk szerteágazó, nehezen felöllelhető problémakör. Topológiai tulajdonságaikat tekintve a dendrogram és a gyökeres kladogram hasonlóak, mindkettő valamilyen hierarchiát összegző, rendszerint – de nem feltétlenül – dichotomikus fa-gráf, belső szögpontokkal, és az

¹ Ezért nem tárgyaljuk a témát külön számozott fejezet részben.

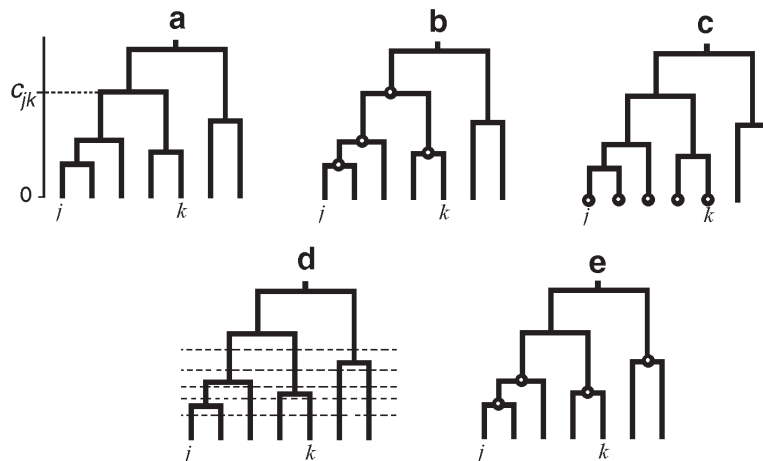
objektumoknak megfelelő terminális szögpontokkal. (A gyökér nélküli kladogramokkal nem foglalkozunk, mivel könyvünk szempontjából jóval kisebb a jelentőségük.) A dendrogramnál a belső szögpontokhoz rendelünk értékeket (hierarchikus szintek), a kladogramoknál az élekhez (vagy azokhoz sem). Az eltérések ellenére célszerű együtt kezelni ezt a két eredménytípust.

Összehasonlítás mátrixok alapján. A klasszikus módszerek mátrix összehasonlításra vezetnek vissza két dendrogram összevetésének feladatát. Egy dendrogram azonban sokféleképpen jellemezhető az objektumpárokra vonatkozó dendrogram deszkriptorok (9.6 ábra) felhasználásával, s ennek megfelelően többféle mátrix jöhet számításba. Podani & Dickinson (1984) áttekintésében a következő deszkriptorok szerepelnek (kivéve a hatodikat, de bizonyos elképzelhetők még mások is):

1. *Kofenetikus differencia:* az a legalacsonyabb hierarchikus szint, amelyen a jk objektumpár egy osztályba tartozik. Az összes lehetséges objektumpárhoz tartozó szintek egy **C** kofenetikus mátrixba írhatók, amely egyértelműen leírja a dendrogramot (pl. Sokal & Rohlf 1962).

2. *Topológiai differencia:* a j és k objektumok közötti úton lévő belső szögpontok száma a fában, ami eggyel kevesebb az összekötő él számánál (pl. Farris 1973, Phipps 1971, Williams & Clifford 1971). Más néven kladisztikus távolságként is említik, de ez kissé félrevezető, mert nemcsak kladogramokról lehet szó. A differenciák **T** mátrixa tartalmazza az elágazások rendszerét, hiszen négy-pont metrika, de a hierarchikus szintek természetesen nem maradnak meg.

3. *Osztályba-tartozási divergencia:* az objektumok száma a legkisebb osztályban, amelynek j és k is eleme. A divergenciák **M** mátrixa kétségkívül alkalmas a fa szerkezetének reprodukálására (ugyanis ultrametrikus jellegű).



9.6 ábra. Dendrogram deszkriptorok a j és k objektumok példáján. **a:** kofenetikus differencia, **b:** topológiai differencia (=4), **c:** osztályba-tartozási divergencia (=5), **d:** partícióba-tartozási divergencia (=5), **e:** részfába-tartozási divergencia (=4).

4. Partícióba-tartozási divergencia: ez a deszkriptor azt a tulajdonságot használja fel, hogy egy dendrogram partíciók sorozataként fogható fel. Kizárva az összes objektumot tartalmazó esetet, a dendrogramban legfeljebb $m-1$ partíció van (de kevesebb, ha vannak egyenlő hierarchikus szintek és többszörös elágazások). A j és k objektumok relatív helyzete a dendrogramban kifejezhető tehát azoknak a partícióknak a számával, amelyben j és k külön osztályba tartozik. Bár topológiai jellegű, mégis megtartja a hierarchikus szintek szerinti információt is.

5. Részfába-tartozási divergencia: Ez a deszkriptor a dendrogramot a belső fa-szerkezet szerint jellemzi. Egy bináris fában (csak dichotomikus elágazásokkal) összesen $m-1$ részfa van (beleértve magát a teljes dendrogramot is), minden egyes belső szögpontra jut egy. A jk objektumpár kapcsolata kifejezhető azoknak a részfáknak a számával, amelyekben nem jelennek meg együtt.

6. Úthossz (patrisztikus távolság): amennyiben a fa (pl. kladogram) egyes éleit külön-külön súlyozzuk, akkor a két objektum közötti út hosszúsága egy újabb deszkriptort szolgáltat, amelyet a \mathbf{P} patrisztikus mátrixba írhatunk.

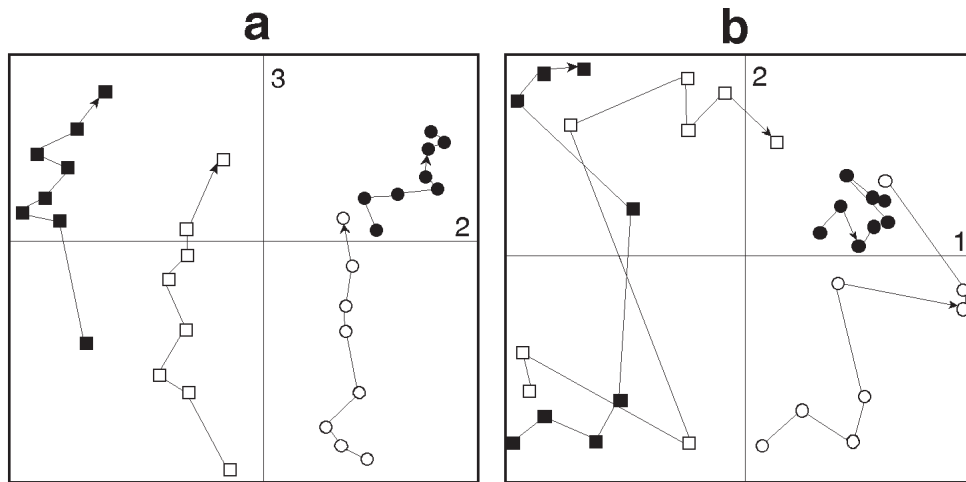
A deszkriptorok tehát a fák más és más sajátosságait emelik ki, ennek következtében a megfelelő sajátosságot érintő "anomáliákra" ügyelnünk kell. Ha ilyen anomáliák jelentkeznek az összehasonlítandó fák valamelyikében, akkor az arra érzékeny deszkriptort célszerű mellőzni. A dendrogramban esetlegesen jelentkező visszafordulások az 1. és 4. deszkriptor használatát kérdőjelezzik meg. A többszörös elágazások (a dichotómia megsértése) a 2. és az 5. deszkriptorra hatnak kedvezőtlenül. A 3. viszont nem érzékeny sem a visszafordulásokra sem a többszörös elágazásokra. A kofenetikus differencia esetében arra is figyelniük kell, hogy a kérdéses két dendrogramban hasonló jellegű-e a szintek növekedése. Az 5.7a és az 5.11a ábra dendrogramjai például nagyon különböznek ilyen szempontból (az egyik fokozatosan, a másik hirtelen, "exponenciálisan" emelkedik), s a kofenetikus differencia alapján történő összehasonlítás félrevezető lehet: voltaképpen a két dendrogram eszerint nem összehasonlítható.

A deszkriptorok egy része már nem új számunkra, a kofenetikus szintekről már volt szó a dendrogramok értékelésével kapcsolatban (5.5.1 rész), a patrisztikus távolságot pedig az additív fáknál (5.4.4 rész) említettük. Értelemes, hogy a fenti deszkriptorok nem alkalmazhatók mindenféle fára; az első öt dendrogramok leírására használható, míg kladogramoknál (additív fáknál) a 2-3 és 5-6 deszkriptorok jöhetnek számításba.

Dendrogramok vagy kladogramok összehasonlítása a fenti deszkriptorok valamelyikének kiválasztásával, s – a már ismert módon – a két dendrogramhoz tartozó deszkriptormátrixok korrelációjának v . távolságának kiszámításával történhet. Mint már említettük a 9.1.8 részben, ez tipikusan egyváltozós eljárás, ugyanis a dendrogramok egy önkényesen kiragadott sajátosságán alapszik csupán. *Többváltozós* tehető például a D_1 és D_2 dendrogramok összehasonlítása, ha az alábbi formula (négyzetre emelt euklidészi távolság) alapján az 1-5. deszkriptorokat egyszerre figyelembe vesszük:

$$\delta_{12}^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m \sum_{a=1}^5 [x(a)_{1ij} - x(a)_{2ij}]^2 \quad (9.11)$$

amelyben $x(a)$ egyes állapotai felelnek meg a deszkriptoroknak. (Megjegyezzük, hogy kladogramok összehasonlítása a 2, 3, 5 és 6. deszkriptorok bevonásával tehető többváltozós.) Annak érdekében, hogy az egyes deszkriptorok közötti skála-eltéréseket kiegyenlítsük először külön-külön standardizálnunk kell őket. A kofenetikus differenciát átskálázzuk, hogy minden



9.7 ábra. Példa a komplex összehasonlításra. Az elemzés a kvadrátnagyság (a növekedés irányát a nyíl mutatja 0,25-től 16 m²-ig) és az adattípus (■=borítás, □= Clymo $c=3$, ○=Clymo $c=15$, ●=prezencia/abszencia) együttes hatását értékeli. A távolságmátrixok PCoA ordinációja (a) szebben mutatja a trendeket, mint a dendrogramoké (b), Podani (1989d).

egyes dendrogramra a [0,1] intervallumba essen. Az osztályba-tartozási divergenciát egyszerűen m -mel osztjuk. A partícióba-tartozási divergencia hasonló standardizálásához megkereshető a dendrogramban lévő partíciók száma, vagy magunk adjuk meg előre a figyelembe veendő partíciók számát (ami persze legfeljebb $m-1$ lehet). A topológiai differencia és a rész-fába-tartozási divergencia standardizálása a dendrogramban talált maximummal történhet.

A most következő illusztráció nem az A függelék mátrixainak valamelyikén, hanem egy szélesebb körű vizsgálaton alapszik (Podani 1985), s egyúttal a komplex összehasonlításokat is példázza. A feladat annak kimutatása, hogy egy növényökológiai tanulmány eredményére milyen hatással van a mintavételezés (a mintavételi egység mérete) és az alkalmazott adattípus. Az adatokat (fajok borítási százaléka) nyolc kvadrátnagyság mellett vettük fel, 20 kvadrátra. Ebből még másik három adattípust állítottunk elő, kettőt a Clymo függvénnyel (2.16 formula, $c=3$ és $c=15$), valamint a prezencia/abszencia mátrixot. Ez egy adattípus transzformációs sort jelent, két átmenettel a tisztán "kvantitatív" és a bináris adattípus között. A kvadrátnagyság és az adattípus együttes változtatása 32 kombinációt eredményez, s ennek alapján 32 különböző távolságmátrix, majd abból pedig 32 különböző dendrogram volt előállítható (az eltérésnégyzetösszeg-növekedés minimalizáló módszerrel). A 32 mátrixot minden párosításban (mátrix korrelációval) összehasonlítva egy 32×32-es újabb mátrix állt elő, amelynek PCoA ordinációjában a 2. és a 3. tengely (9.7a ábra) 9, ill. 2,3 %-ot magyaráz az összes távolságból (az 1. tengely erősen egypólusú volt). A diagram egyértelműen mutatja, hogy az adattípus a fontosabbik tényező, és a kvadrátnagyság szerepe a prezencia/abszencia esetben a legkisebb. Az osztályozásokban (a 32 dendrogram többváltozós összevetése a 9.11 formula szerint) a trendek kevésbé ugyan, de felismerhetők (9.7b ábra; a két tengely 21, ill. 16 %-nak felel meg). Az 1. tengelyen az adatátalakítási sor két első, ill két utolsó lépcsője nem különböztethető meg, s a bináris esetben itt is a legkisebb a kvadrátnagyság hatása. A többváltozós dendrogram összevetésre másik példát találunk Csontos & Lőkös (1992) cikkében.

Grafikus összehasonlítás. Két dendrogram vagy kladogram grafikus módon is összehasonlítható egymással, az alkalmas deskriptorok valamelyikének felhasználásával felírt két deskriptormátrix alapján (lásd 9.2.1 rész).

Ultrametrikus viszonyok. A dendrogramok összehasonlításának egy, az eddigiektől teljesen eltérő megoldását javasolta Dobson (1975). A módszer minden objektumhármásra megvizsgálja az ultrametrikus egyenlőtlenséget, majd megszámlálja azokat a tripleteket, melyekre az egyenlőtlenség különböző a két dendrogramban. Más szóval, az $\{i, j, k\}$ tripletet beszámítjuk, ha a $c_{ij} < c_{ik} = c_{jk}$ fennáll D_1 -ben, de vagy $c_{ik} < c_{ij} = c_{jk}$ vagy $c_{jk} < c_{ij} = c_{ik}$ teljesül D_2 -ben. Az összeget célszerű a lehetséges maximummal, $\binom{m}{3}$ -mal osztani (ez a tripletek száma m objektumra), s így egy 0-tól 1-ig terjedő különbözőséget kapunk (*ultrametric dissimilarity*).

Élek és élhosszak. A módszerek egy másik csoportja a gráf éleivel, ill. a hozzájuk rendelhető értékekkel (hosszakkal) dolgozik. Az alapvető koncepciót Robinson & Foulds (1979, 1981) dolgozta ki, elsősorban gyökér nélküli fákra, amely kis változtatásokkal dendrogramokra is adaptálható. A legegyszerűbb ilyen mértéket említjük meg, amely a belső élek egyenkénti eltávolításával (*edge removal*) nyújt lehetőséget az összehasonlításra. Egy teljesen dichotomikus dendrogramban a belső élek száma $m-2$. Ha egyiküket eltávolítjuk, akkor az objektumok egy két osztályra történő felosztását kapjuk². A D_1 dendrogram egy belső éle, valamint a D_2 dendrogram egy belső éle *megfelel egymásnak*, ha eltávolításukkal éppen ugyanazokat a partíciókat kapjuk. (A kis változtatás abban áll, hogy dendrogramok esetében a gyökértől induló két élt egyként kezelhetjük, a megvizsgálandó élek száma tehát dendrogramként $m-3$ -ra csökken.) Vagyis az egymásnak *nem* megfelelő élek száma a dendrogramok távolságának a mérésére alkalmas (*symmetric-difference distance* vagy *partition metric*, vö. Robinson & Foulds 1979, 1981). Ezt a maximummal osztva az *élegyezési indexet* kapjuk:

$$EM_{12} = \frac{\text{a nem megfelelő élek száma } D_1 \text{ - ben és } D_2 \text{ - ben}}{2m - 6} \quad (9.12)$$

melynek értéke a $[0, 1]$ tartományba esik, 0 a teljes megegyezés, 1 a teljes különbözőség esetén adódik; és $m > 3$. A 9.12 mérőszámában minden él egyformán fontosnak számít, függetlenül attól, hogy milyen hosszú, vagy mekkora a hierarchikus szintbeli különbség a hozzátartozó két szögponthoz. Kidolgozhatók olyan – a 9.12-höz hasonló – indexek is, amelyek az él hossza szerint súlyoznak a számlálóban és a nevezőben is.

Gyökér nélküli fákra a legismertebb mértékszám a *legközelebbi szomszéd felcserélési metrika* (*nearest neighbour interchange [nni] metric*, Waterman & Smith 1978; *crossover* Robinson 1971). A kladogramok szerkesztésénél már láttuk (6.3.1.2 rész), hogy a fák átalakításának egy lehetséges módja egy adott belső élhez tartozó részfák felcserélése. Az *nni* metrika szerint két fa távolsága azoknak az ilyen felcseréléseknek a *minimális száma*, amely ahhoz szükséges, hogy az egyiket a másikba alakítsuk. A számítási feladat nagy m -re igen nehéz, ezért Brown & Day (1984) egy közelítő metrikát javasolt, mely viszonylag rövid idő alatt is kiszámítható.

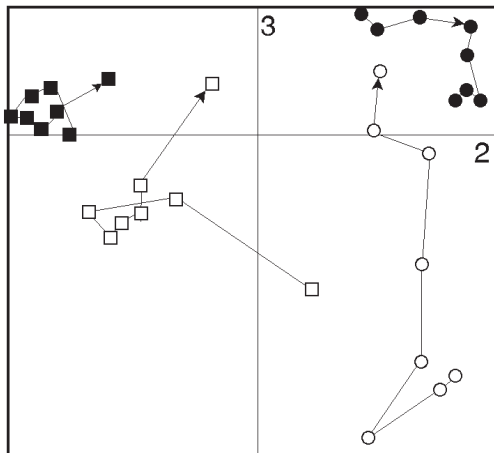
9.2.4 Ordinációk összehasonlítása

Egy ordináció m objektum t -dimenzióbeli koordinátaival reprezentálható, bár rendszerint csak a legfontosabb p dimenzió érdekel bennünket az összehasonlításban, mert – mondjuk – a többi már nem hordoz számottevő információt (a megmagyarázott variancia %-os értelmében). Az

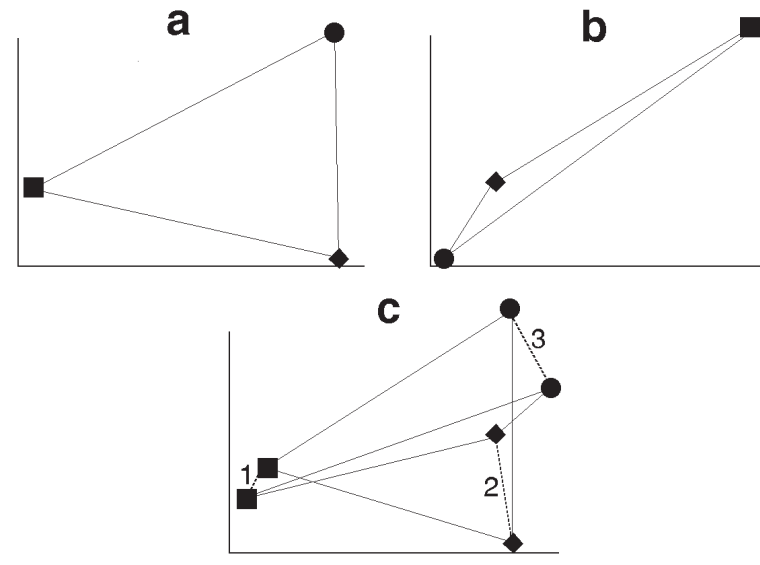
2 A terminális szögpontokhoz vezető éleket nem vesszük figyelembe, hiszen ezek eltávolítása eleve azonos, és érdektelen partíciókra (egy objektum plusz a többi) vezet mindkét dendrogramban.

ordinációk páronkénti összehasonlítása, csakúgy mint más típusú OUC-ok esetében, érdekes lehet annak feltárásában, hogy a változók, az adattípusok, a különbözőségi függvények és az ordinációs módszerek kiválasztása milyen mértékben hat az eredményre. Például a taxonómiában érdemes megnézni azt, hogy az OTU-k vegetatív ill. reprodukzív bélyegein alapuló ordinációi mennyire kongruensek egymással. A 7.6.2 részben – a szuperpozíciós módszerek elnevezéssel – pedig már említettük, hogy a biológiai formák leírása koordináták segítségével szintén ordináció, így a formák összehasonlítása az ordinációk összehasonlítására (is) visszavezethető probléma. Egy dimenzióban az ordinációk összevetése egyszerűen megoldható a korrelációs koefficiens felhasználásával. Ha azonban $p \geq 2$, s rendszerint ez a helyzet, akkor a korrelációs együttható ilyenét alkalmazása már nem megy, s más eszközhöz kell folyamodnunk. A *mátrix korreláció* jelentheti például a megoldást, mégpedig úgy, hogy az egyes ordinációkat az m pont közötti távolságok mátrixával írjuk le a kívánt p dimenzió figyelembevételével, s ezeket vetjük össze (pl. Podani 1989d, 9.8 ábra). Ha ez lehetséges, akkor nyilván menni fog két ordináció grafikus összehasonlítása is. Annak ellenére, hogy a mátrix korreláció alkalmazhatósága nem vitatható ordinációk esetében sem, a gyakorlatban egy matematikailag jóval “elegánsabb” geometriai eljárás terjedt el, a *Prokrusztész módszer*, melynek kifejlesztése – részben függetlenül egymástól – számos szerzőnek is betudható (Green 1952, Gower 1971a, Schönemann & Carroll 1970). A név a görög mitológia hírhedt alakját idézi, aki vendégeit vagy halálra nyújtotta, vagy lábaik levágásával megrövidítette, hogy pontosan beférjenek a szerinte szabályos ágyba (innen a “Prokrusztész ágy”, az általunk elviselni kénytelen kényszerhelyzet kifejezése). A név nyilvánvaló utalás arra, hogy az ordinációkat bizonyos átalakításoknak kell alávetnünk, mielőtt bármiféle értelmes összehasonlítást tehetnénk velük. A módszer alapfeltevései a következők:

- két ordináció azonosnak tekinthető, ha az egyik a másiknak az elcsúsztatásával (a koordináták egy konstanssal való megnövelésével) állítható elő;
- két ordináció ugyancsak azonos, ha az egyiknek a koordinátái a másik koordinátáiból egy konstanssal való szorzással kaphatók meg;
- az ordinációk akkor is azonosak, ha létezik egy olyan α szög, amellyel az egyiket elforgatva éppen megkapjuk a másikat (beleértve az $\alpha=180^\circ$ esetet, a tükrözést).



9.8 ábra. Ordinációk meta-elemzése ugyanazokból az adatokból kiindulva, mint a 9.7 ábra esetében. Egy pont a kvadrátnagyság és az adattípus egy kombinációjának a főkoordináta ordinációját képviseli az 1-2 dimenziókra, s az ordinációk ordinációja is a PCoA révén készült, mátrix korrelációk komplementjéből kiindulva. Az első tengely egypólusú volt, a következő két tengely relatív fontossága 14 ill. 7%. Szemben a mátrixok 9.7a ordinációjával, most az adattípus két szélső esete tűnik kevésbé érzékenynek a kvadrát méretére (Podani 1989d).



9.9 ábra. Prokrusztész módszer. Három objektum két ordinációjának (a és b) különbözősége az optimális egymásra illesztés (c) révén a számokkal jelölt távolságok négyzetösszegével mérhető.

Eme feltételek szerint az összehasonlítás a kérdéses ordinációk optimális egymásra illesztését jelenti, centrálással, elforgatással és a koordináták átskálázásával. A legjobb illeszkedés a megfelelő pontok közötti távolságok (9.9 ábra) négyzetösszegének a minimalizálásával található meg.

Formálisabban, ha az m pont p -dimenzióbeli, előzetesen centrált koordinátáit az \mathbf{X} és \mathbf{Y} mátrixok tartalmazzák, akkor a következő függvény értékét kell minimalizálnunk:

$$\sum_{i=1}^m \sum_{j=1}^p (x_{ij} - y_{ij})^2 = \text{tr}[(\mathbf{X} - \mathbf{Y})(\mathbf{X} - \mathbf{Y})']. \quad (9.13)$$

Ennek előállításához \mathbf{X} érintetlenül hagyása mellett \mathbf{Y} -t forgatjuk el a $p \times p$ méretű \mathbf{H} ortogonális *forogatómátrix* felhasználásával:

$$\mathbf{H} = \mathbf{V}\mathbf{U}' \quad (9.14)$$

amelyben \mathbf{U} és \mathbf{V} az $\mathbf{X}'\mathbf{Y}$ szorzatmátrix szingulárisérték-felbontásából (C függelék) származik:

$$\mathbf{X}'\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}' \quad (9.15)$$

(\mathbf{S} a sajátértékek diagonálmátrixa). Az illeszkedés jósága ezek alapján a következőképpen fejezhető ki:

$$\begin{aligned} R^2_E &= \text{tr}[(\mathbf{X} - \mathbf{Y}\mathbf{H})(\mathbf{X} - \mathbf{Y}\mathbf{H})'] = \\ &= \text{tr}(\mathbf{X}\mathbf{X}') + \text{tr}(\mathbf{Y}\mathbf{Y}') - 2\text{tr}(\mathbf{Y}\mathbf{X}'\mathbf{X}\mathbf{Y}')^{1/2} \end{aligned} \quad (9.16)$$

amely egy szimmetrikus mérőszám ugyan, de az ordinációk skálázása (a koordináták aktuális nagysága) befolyásolja. Önkényes skálázású ordinációkra egy c szorzófaktor veendő figyelembe az \mathbf{Y} mátrix $c\mathbf{Y}\mathbf{H}$ -ba történő transzformációjánál:

$$c = \text{tr}(\mathbf{Y}\mathbf{H}\mathbf{X}') / \text{tr}(\mathbf{Y}\mathbf{Y}'), \quad (9.17)$$

melynek révén a következő statisztikát kapjuk:

$$R^2_S = \text{tr}(\mathbf{XX}') - 2(\text{tr}(\mathbf{YX}'\mathbf{XY}')^{1/2} / \text{tr}(\mathbf{YY}')). \quad (9.18)$$

Ez a koefficiens viszont már nem lesz szimmetrikus, és ezért Gower (1971a) javaslata szerint centrálás után voltaképpen az ordinációkat egységnyi négyzetösszegűre kellene standardizálnunk:

$$\text{tr}(\mathbf{XX}') = \text{tr}(\mathbf{YY}') = 1, \quad (9.19)$$

(ami azt jelenti, hogy mindkét ordinációban a pontoknak az origótól vett távolságnégyzetösszege egységnyi). A 9.18 függvényt ebből a kiindulásból előállítva jelöljük d^2 -tel. Amint arra Sibson (1978) rámutatott, R^2_S közvetlenül is standardizálható:

$$\gamma_S = R^2_S / \text{tr}(\mathbf{XX}'), \quad (9.20)$$

melynek értéktartománya a $[0, 1]$ intervallum. A d^2 és γ_S összefüggése pedig a következő:

$$d^2 = 2(1 - (1 - \gamma_S)^{1/2}). \quad (9.21)$$

Amelyből már látszik, hogy d^2 értéke 0-tól 2-ig terjed, 0 felel meg a tökéletes illeszkedésnek, vagyis az ordinációk azonosságának, 2 pedig a konfigurációk maximális különbözőségének.

Példaképpen összehasonlítjuk az A1 táblázat objektumainak (oszlopainak) PCA és COA ordinációit. Az első két tengely figyelembevételével (7.2 és 7.14 ábra) a d^2 értéke 0,1, ami alacsonynak tűnik. Többet erről csak akkor mondhatunk, ha viszonyítási alappal is rendelkezünk (l. a következő alfejezetet). Megjegyezzük, hogy az első három dimenzióra nagyobb távolságot kapunk ($d^2 = 0,309$), jelezve a harmadik tengelyen mutatkozó erős eltérést.

9.2.5 Átrendezett tabellák összehasonlítása

Ez a probléma viszonylag ritkán merült még fel a szakirodalomban, holott a különféle módszerekkel kapott – s akár a kézzel végzett – átrendezések összehasonlítása éppúgy érdekes feladat lehet, mint a többi eredményfajta értékelése. Példaképpen először a keresztpartíciós *blokk* osztályozások összehasonlítására alkalmas eljárást (Podani & Feoli 1991) ismertetjük röviden. A partíciók átmeneti metrikáját (9.7 formula) tesszük alkalmassá a feladat elvégzésére. Legyen \mathbf{X}_i és \mathbf{X}_j a két átrendezett, $n \times m$ -es méretű táblázat, mindkettőben p osztály a sorokra és q osztály az oszlopokra. A feladat annak meghatározása, hogy minimálisan hány érték besorolását kell megváltoztatnunk \mathbf{X}_i -ben, hogy \mathbf{X}_j -t előállítsuk (vagy fordítva). A sorok két osztályozását és az oszlopok két osztályozását külön-külön összehasonlítjuk, amelyből megkapjuk a $M_{ij}(\text{sorok})$ és az $M_{ij}(\text{oszlopok})$ értékeket (a rövideg kedvéért *MINDMT* helyett *M*-et írunk). Ebből az áthelyezendő adatok száma a következő:

$$K_{ij} = m M_{ij}(\text{sorok}) + n M_{ij}(\text{oszlopok}) - M_{ij}(\text{sorok}) M_{ij}(\text{oszlopok}) \quad (9.22)$$

melyet a lehetséges maximummal osztva kapjuk a κ indexet:

$$\kappa_{ij} = K_{ij} / [m \max M_{ij}(\text{sorok}) + n \max M_{ij}(\text{oszlopok}) - \max M_{ij}(\text{sorok}) \max M_{ij}(\text{oszlopok})] \quad (9.23)$$

Ennek értéke 0 és 1 közé esik, tükrözve a teljes azonosságot, ill. maximális különbözőséget.

Szeriálással átrendezett mátrixok összevetése a sorok és oszlopok permutációinak (adatmátrixok esetén), ill. a sorok permutációinak (távolságmátrixok esetén) az összehasonlításával történhet. Ezt elvégezhetjük úgy, hogy meghatározzuk az összes sorra és az összes oszlopra (távolságmátrixoknál csak a sorokra) a sorrendek közötti eltéréseket, és ezeket összegezzük. Ha a sorok és oszlopok indexe az első mátrixban i és j , és a megfelelő sorok, ill. oszlopok indexe a másodikban $y(i)$ és $y(j)$, akkor a keresett mennyiség adatmátrixokra:

$$\kappa_{1,2} = \sum_{i=1}^n |i - y(i)| + \sum_{j=1}^m |j - y(j)|. \quad (9.24)$$

Távolságmátrixok esetében a fenti formulának csak az első tagját kell meghatároznunk. A 9.24 mennyiségnél persze jóval elegánsabb az a mérőszám, amely analóg a *MINDMT*-vel vagy az *mi* metrikával: hány *szomszédos* sort és oszlopot kell minimálisan felcserélni egy iterációs sorozatban az első mátrixban, hogy megkapjuk a másodikat. Ennek meghatározása – mint sejtjük – már nehezebb feladat a 9.24 kiszámításánál.

9.3 Hipotézisvizsgálatok, várható értékek, eloszlások

Eredményeink páronkénti numerikus összehasonlítását elvégezve egy különbözőségi értéket kapunk, amelynek jó esetben ismert az értéktartománya, de az is lehet, hogy nincs egyértelműen megadható felső korlátja (mint pl. a 9.11 függvénynek). Ez nem okozhat gondot, amíg egy adott vizsgálaton (pl. meta-analízisen) belül maradunk. A különbözőségi értékek összehasonlíthatóságának a problémája azonban rögvest felmerül, ha meg szeretnénk állapítani, hogy – mondjuk – negyven objektum két ordinációjára kapott $d^2=1,42$ érték nagyobb eltérést jelent-e, mint tíz objektum két ordinációjára kapott $d^2=1,40$. A felületes szemlélő azt mondhatná, hogy d^2 értéke 0-tól 2-ig terjed bármekkora m esetén, következésképpen az első érték valóban kissé nagyobb eltérést jelez, mint a második. Ez azonban semmiképpen sem biztos! Ilyen típusú összehasonlítást ugyanis csak akkor tehetünk korrekt módon, ha ismerjük az összes lehetséges (véletlen ordináció-párokra kapott) különbözőség-érték *eloszlását*, s annak legfontosabb paraméterét, a *várható értéket* (“átlag”). Visszatérve a konkrét példára: 40 objektumra az 1,42-es különbözőség jóval a várható érték alatt van, míg tíz objektumra az 1,40 jóval felette! Az 1,42-es érték a 40 objektum esetében relatíve nagy egyezésnek számít, az 1,40-es érték 10 objektumra pedig már nagy eltérésnek. Kezdeti ítéletünk elhamarkodott volt tehát, s be kell látnunk: a különbözőségi értékek nagysága igenis viszonylagos. Az eloszlás ismerete még inkább nélkülözhetetlen, ha a különbözőség *szignifikanciáját* keressük, vagyis meg akarjuk mondani, hogy ugyanazon objektumhalmazra egymástól függetlenül kapott két eredmény különbözősége jóval alacsonyabb-e a vártnál (szignifikánsan kicsiny) vagy beleesik abba a tartományba, amelybe a véletlenszerűen generált eredménypárok különbözőségeinek nagy része (pl. 95 %-a) is tartozik (nem szignifikáns). Ezekkel a kérdésekkel egy igen izgalmas, jelenleg is erősen kutatott témakörhöz érkeztünk.

Gondjainkat növeli, hogy két OUC különbözőségének az elméleti eloszlása rendszerint ismeretlen. Meghatározták már például a partíciós metrika (9.12) eloszlását 16 objektumig (Hendy et al. 1984), és kladogramok közötti néhány metrikára is ismertek az eloszlás paraméterei (Steel & Penny 1993), de ez a ritka kivétel: kellő nagyságú, s a gyakorlatban is rendszerint előforduló objektumszámokra nem tudjuk pontosan az eloszlásokat, vagy ha tudjuk is az elvet, a számítások rendkívüli nehézséget okoznak. Praktikus segítséget ekkor a Monte Carlo szimuláció módszerei, vagyis a – sokak szerint e módszerek speciális eseteit jelentő – randomizációs vagy permutációs tesztek jelentenek.

Az alapkérdés minden esetben a null-hipotézisnek megfelelő szituáció előzetes megfogalmazása. Ekkor hamar kiderül, hogy a Monte Carlo, ill. a speciálisabb permutációs tesztek közötti választás nem is olyan könnyű. A *Monte Carlo módszerekhez* általában akkor nyúlunk,

ha az eloszlást *teljesen véletlenszerű* OUC-párok felhasználásával közelítjük s kimondjuk: bármilyen OUC előállítása egyformán valószínű. Egy megfelelő random-szám generátor felhasználásával előállítunk mondjuk 999 pár dendrogramot, véletlen hierarchikus szintekkel és véletlen bifurkációkkal (a pontos algoritmus most nem lényeges, vö. Lapointe & Legendre 1991), mindegyikre kiszámítjuk az előzetesen kiválasztott különbözőségi függvény értékét, majd ezeket sorba rendezve és kategóriákba osztva gyakorisági hisztogramot is készíthetünk. Az ezredik érték a tesztelni kívánt aktuális különbözőség lesz, ezt jelöljük d -vel. A szimulált értékek alapján megállapítható: mi annak a (becsült) valószínűsége, hogy véletlen dendrogramokra d -vel egyenlő vagy annál kisebb különbözőséget kapjunk. Ha ez a valószínűség alacsonyabb, mint az általunk előzetesen megszabott α szignifikancia-szint (rendszerint $\alpha=0,05$), akkor a kérdéses két dendrogram szignifikánsan hasonlóknak tekinthető. 1000-es elemszámot választva ez akkor van így, ha d -nél legalább 950 esetben nagyobb értéket eredményez a szimulációs algoritmus. Ellenkező esetben megtartjuk a null-hipotézist, azaz kimondjuk: a d értéket voltaképpen random dendrogramokra is megkaphatjuk (adott α mellett), az összehasonlított két dendrogram hasonlósága tehát nem szignifikáns. A *permutációs tesztek*³ lényegileg ugyanezen az elven alapszanak, eltekintve a párok generálásának módszerétől. Ekkor ugyanis nem teljesen véletlen eredményeket állítunk elő, hanem az összehasonlított dendrogramokon cseréljük fel véletlenszerűen az objektumokat. Más szóval: az *objektumok elrendezését permutáljuk*, megtartva ugyanakkor az eredmény “alapvázat” (az eredményben csupán a “címkézés” változik). Az ún. *exakt permutációs tesztek*, viszonylag alacsony m -re, az összes permutációs lehetőséget kipróbálva állítják elő a koefficiens pontos eloszlását, egyéb esetben viszont be kell érniük az eloszlás becslésével adott számú szimuláció alapján. Ez a szám persze legyen minél nagyobb: dendrogramokra Lapointe & Legendre (1992) az 1000-es számot bőven elegendőnek tartja, míg mátrixok összevetésére Jackson & Somers (1989) tízszázezer lépést javasol minimálisan. Podani (1986) a 9.7 mérőszám eloszlását vizsgálva azt találta, hogy 5000 szimuláció igen jól közelíti az exakt módon kapott eloszlást. Nyilvánvaló, hogy abszolút érvényű szabályokat nehéz lenne felállítani, de célszerű minél nagyobb lépésszámot választani. Mondanuk sem kell talán, hogy műveleteinkhez minden esetben számítógépet kell igénybe vennünk.

A Monte Carlo és permutációs tesztek végrehajtásának módozatait a 9.1 táblázatban összesítjük a fontosabb OUC típusokra. A “tisza” Monte Carlo szimuláció a permutációs teszteknel rendszerint sokkal nehezkesebben végezhető el. A lehetséges OUC-ok száma, amelyből a szimuláció során végül is “mintát veszünk” végtelen vagy igen nagy, szemben a permutációs tesztek relatíve kisebb esetszámával. A Monte Carlo módszerek az általánosabbak, a permutációs tesztek minden esetben az egyedi problémához igazodnak.

9.3.1 Mátrixok

A Monte Carlo szimuláció és a permutációs tesztek eltérését a távolságmátrixok összehasonlítására – csaknem kizárólagosan – alkalmazott mátrix-korreláció szignifikancia próbájával

3 A permutációs teszt és a randomizációs teszt elnevezések gyakorlatilag szinonimák (vö. Manly 1991). A mi esetünkben inkább a permutációs teszt elnevezés tűnik szemléletesebbnek, hiszen ez utal jobban az eloszlás előállításának a módjára.

9.1 táblázat. A Monte Carlo és permutációs módszerek összehasonlítása különféle eredménytípusok esetében.

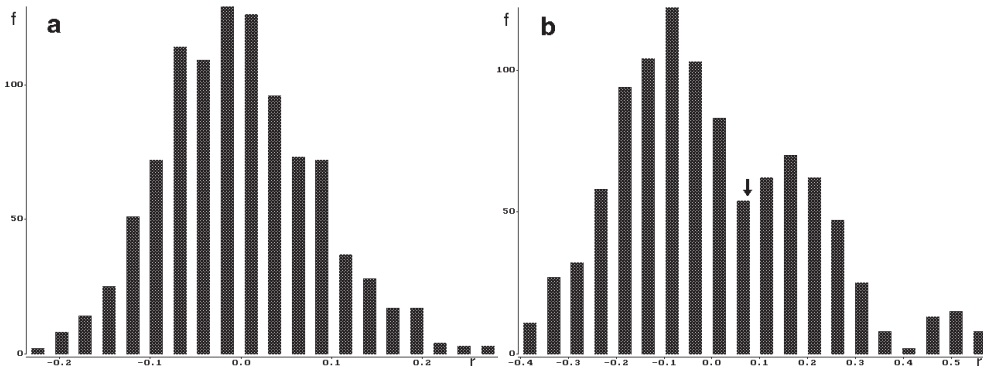
Eredmény (OUC) típusa	“Tiszta” Monte Carlo szimuláció	Permutációs teszt
Távolságmátrix	Teljesen véletlen távolságok	Oszlopok (és sorok) random felcserélése
Kemény partíció	Objektumok véletlenszerű besorolása k számú osztályba, tekintet nélkül azok nagyságára	Objektumok véletlenszerű besorolása k számú osztályba, megtartva az eredeti osztályméreteket
Lágy partíció	Random súlyértékek minden objektumra, az alapfeltétel betartásával (összegük ui. 1)	Az eredeti súlyértékeket megtartjuk, de random módon soroljuk melléjük az objektumokat
Dendrogram	Véletlen szintek, véletlen bifurkációk, véletlenszerűen elrendezett objektumok	A fa végágain lévő objektumokat cserélgetjük véletlenszerűen
Gyökeres kladogram	Véletlen élhosszak, bifurkációk és objektum elrendezések	Mint fent.
Ordináció	Random koordináták minden dimenzióban	Az eredeti koordinátákat megtartva az objektumokat keverjük össze.

illusztráljuk. Mantel (1967) javasolta először, hogy két távolságmátrix, \mathbf{D} és \mathbf{E} korrelációjának (9.1 függvény) kiszámítása után a kapcsolat erősségét ne a táblázatbeli szignifikanciaszinthez viszonyítsuk (ez balgaság volna, hiszen a mátrixokon belüli értékek nem függetlenek egymástól!), hanem – most jön a lényeg – a \mathbf{D} mátrix sorainak (és ezáltal oszlopainak) állítsuk elő sok-sok permutációját, s az így megkevert mátrixokat is vessük össze \mathbf{E} -vel. r_{DE} szignifikanciáját a permutációkból származó korrelációk eloszlása alapján értékelhetjük (Mantel teszt).⁴ A null-hipótesis szerint a \mathbf{D} -t generáló mechanizmusok függetlenek az \mathbf{E} -t létrehozó mechanizmusoktól, vagyis r_{DE} valószínűen megkapható akkor is, ha a mátrix értékeit permutációk révén “összezavarjuk”. Ha azonban ez nem áll fenn, akkor a két mátrix mögöttes hatásai nem függetlenek, s a null hipotézist elvetjük.

A permutációs teszt mellett felmerülhet a teljesen randomizált távolságmátrixokra kapott korrelációk előállításának a lehetősége is. Random mátrixok előállítása oly módon, hogy azok valóban távolságstruktúrát tükrözzenek (s ne csak teljesen véletlen számok legyenek) már nehézkesebb. Megoldható például úgy, hogy az eredeti adatokat keverjük össze, s ebből számolunk távolságokat (ez viszont az eredeti adatokra vonatkozó permutálást jelenti). Azt is figyelembe kellene venni, hogy milyen formulával számoltuk ki a mátrixot, mert az egyes különbözőségi indexek eleve más és más eloszlásúak (Hajdu 1981, Gower & Legendre 1986). A Monte Carlo szimulációs modellbe ezt be kellene építeni, ami már túlságosan megnehezítené a dolgunkat.

Példaképpen vizsgáljuk először meg a dűne adatokat. Az egyik mátrixban a mintavételi helyek távolságait a fajok alapján számoljuk ki, a másikban pedig a “környezeti” változók

4 Pontosabban: Mantel nem a korrelációs formulát, hanem a keresztszorzatot vette figyelembe, hiszen ez a korrelációnak a permutációkra érzékeny része. Ha az összehasonlítás előtt a mátrixok értékeit a szórással standardizáljuk, akkor a keresztszorzat számszerűen megegyezik a \mathbf{D} -nek \mathbf{E} -re vonatkozó regressziós együtthatójával, és az \mathbf{E} -nek \mathbf{D} -re vonatkozó regressziós együtthatójával is.



9.10 ábra. A mátrix korreláció eloszlásának hisztogramja 1000 permutáció alapján **a:** a dűne fajadatokra és környezeti adatokra (A4 táblázat) számolt mátrixok között; **b:** az A1 táblázatból az egyszikűekre és a kétszikűekre számolt mátrixok között. Az utóbbin nyíl mutatja a tesztelt érték helyét.

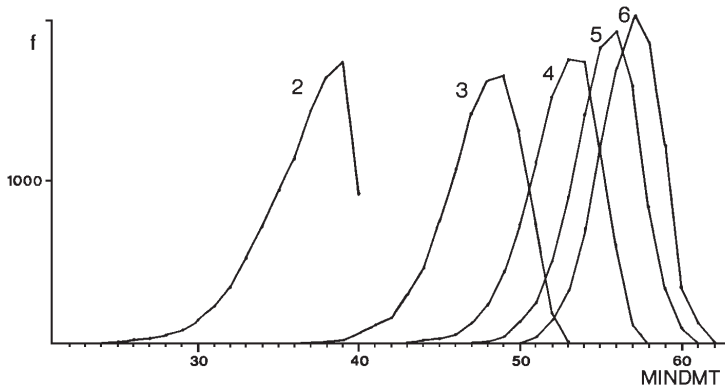
alapján. A nullhipotézis szerint a két változócsoport független egymástól, vagyis a mátrix korreláció értéke nem nagyobb, mint a permutált mátrixokra számoltoefficiensnek 95 %-a. A mátrix korreláció értéke 0,44-nek bizonyult, ami a szimulált értékek (9.10a ábra) mindegyikénél nagyobb, a nullhipotézis tehát elvethető: a két változócsoport szignifikánsan hasonló távolságmátrixokat generál.

A második példa kissé mesterkétebb, de jól illusztrálja a fentitől eltérő eseteket. Az A1 táblázatból kiindulva vizsgáljuk meg, hogy a felvételeknek az egyszikűek (7 faj), illetve a kétszikűek (5 faj) alapján számolt távolságmátrixai korrelálnak-e egymással!? A mátrix korreláció értéke 0,091, ami már sejteti, hogy a két mátrixnak nem sok köze van egymáshoz, s ez a permutációkkal kapott eloszlás (9.10b) alapján be is igazolódik: az értékek 31 %-a nagyobb, 69 %-a kisebb 0,091-nél. A permutálás során minden harmadik esetben nagyobb korrelációt kaptunk az aktuálisnál, a nullhipotézis így megtartható: a két rendszertani faj-csoport nem ad szignifikánsan hasonló távolságstruktúrát. Ez természetesen csak erre a példára igaz, a Mantel teszt érvényessége ugyanis mindig az éppen vizsgált mátrixokra korlátozódik!

9.3.2 Kemény partíciók

Nem-hierarchikus osztályozások különbözőségének eloszlása ugyancsak a Monte Carlo szimulációval vizsgálható legegyszerűbben, annak ellenére, hogy egyes esetekben pontosan ismert az eloszlás néhány paramétere (vö. Hubert & Arabie 1985). Tételezzük fel, hogy a P és Q partíciók összevetéséből eredő d értéket teszteljük, a két partíció osztályszáma s és t , p_i és q_j osztályméretekkel. A legtisztább Monte Carlo eset az, amikor az osztályok száma és nagysága is véletlen hatások eredője, de ennél értelmesebbnek látszik, ha legalább az s és t értékét konstansnak vesszük a szimulációban. Általában azonban a permutációs teszt a legelfogadottabb, vagyis amikor az osztályméretek is kötöttek, a véletlen partíciókat a kereszt-partíciós tábla peremeinek változatlanul hagyása mellett állítjuk elő (Hubert & Arabie 1985).

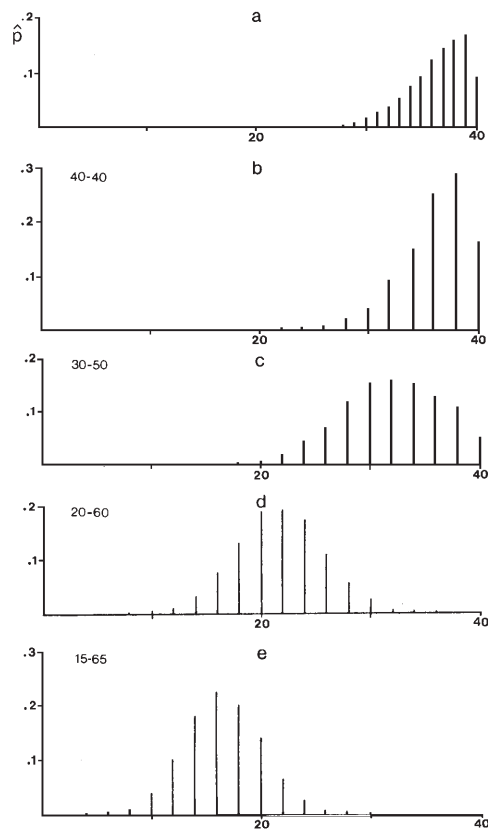
Az általánosabb esetet, amikor s és t rögzített (mégpedig $s=t$), illusztrálja a *MINDMT* függvényre kapott eloszlások sorozata s különböző értékeire nyolcvan objektumra (9.11 ábra). Mint látható, s növelésével a különbözőség várható értéke is nő, de egyre kisebb mértékben, míg maga az eloszlás egyre szimmetrikusabbá válik. Az osztályszámok rögzítésének hatását a permutációk segítségével kapott eloszlások mutatják (9.12 ábra). Amikor az osztályok mérete szabadon alakul a szimuláció során (9.12a), az eloszlás jól megegyezik a 9.11 ábra $s=2$ esetére kapott eloszlásával. Ha azonban rögzítjük az osztályszámot, és mindkét partícióban fokoza-



9.11 ábra. A partíciók összehasonlítására alkalmas *MINDMT* mértékszám Monte Carlo módszerrel készült gyakoriságeeloszlása $m=80$, $s=t=2, 3, 4, 5, 6$ mellett. A pontokat a jobb áttekinthetőség kedvéért összekötöttük. Minden eloszlás 10000 partíció-pár összevetésén alapszik (Podani 1986).

tosan növeljük a két osztály méretbeli különbségét (9.12b-e ábra), a várható érték csökken, az eloszlás egyre szimmetrikusabb lesz.

Érdeemes megnézni, hogy a 9.4 ábrán bemutatott függvényváltozások mikor, milyen fajszámredukciónál válnak esetleg "szignifikánssá". A szót idézőjelbe tettük, hiszen statisztikai értelemben vett szignifikanciáról nem lehet szó: az összehasonlított partíciók nem függetlenek egymástól, mivel részben ugyanazokon a fajokon alapszanak. Ennek ellenére jó tájékozódási alapot ad a kritikus értékek meghatározása – mondjuk – a 95 %-os szignifikancia-szinten. Szimulációval megkapjuk mindegyik koefficiensre azt az értéket, amelyre, ill. ez alatt a két



9.12 ábra. A *MINDMT* becsült valószínűség-eloszlása az $m=80$, $s=t=2$ esetre az osztályméretek rögzítése nélkül (a) és rögzített osztályméretekkel (b-e) előállítva. Az utóbbi esetekben az osztályméretek a bal felső sarokban tüntettük fel. Minden eloszlás 10000 partíció-pár összevetésével készült (Podani 1986).

összehasonlítandó partíció hasonlósága szignifikáns (a szimuláció során a kötetlen csoportmértet használtuk, mert az osztályok mérete más és más lett a fajsám redukció során). Ezek a küszöbértékek a következők: 0,5 (MISC), 0,655 (1–LAS), 0,445 (1–RAND), 0,585 (1–JAC) és 0,415 (1–OCH). Így már mindjárt érhetővé válik, hogy miért is annyira eltérőek az ábra konkrét értékei. A koefficiensek *eloszlása* más és más, és ezért mindegyiküket a saját kritikus értékükhöz kell viszonyítani, s nem a többihez. A viszonyításból kiderül, hogy a fajsám redukciója 5-re sem elegendő ahhoz, hogy a partíciókban akkora változást okozzon, amekkora a “szignifikáns” eltéréshez szükséges, bármelyik koefficienst nézzük is.

9.3.3 Lágy partíciók

Lágy (fuzzy) partíciókat a súly- (vagy osztályba tartozási-) mátrixok írják le, amelyekre igaz, hogy az értékek összege minden objektumra 1. Az általános Monte Carlo esetben e feltétel betartásával teljesen véletlenszerű súlyokat állítunk elő. A permutációs tesztek – logikus módon – megtartják az összehasonlítandó két osztályozás eredeti súlyértékeit, és az objektumokat permutálják. Más szóval, az U_1 súlymátrix (vö. 4.3 alfejezet) sorait keverik véletlenszerűen össze, míg az U_2 mátrix változatlan marad. Az oszlopokat, vagyis az egy osztályra jutó súlyértékek összegét változatlanul hagyjuk, hiszen így lesz a teszt kompatibilis a kemény partíciók permutációs próbájával. A téma részletesebb kidolgozása időszerűnek látszik.

9.3.4 Dendrogramok és kladogramok

A dendrogramok és kladogramok, vagyis általában a fák, az összes OUC típus közül a legbonyolultabb szerkezetű objektumok, így összehasonlításuk (mint már láttuk) és szimulációjuk is meglehetősen komplikált feladat. A téma kimeríthetetlen, de egy adott d különbözőségi index szignifikancia próbájához alapot adó eloszlás előállításánál az alábbiakat kell elsősorban tekintetbe vennünk:

- A szimulált dendrogramok halmaza valójában egy minta az összes lehetséges dendrogramokból. Biztosítani kell azt, hogy a lehetséges dendrogramok mindegyike egyforma eséllyel kerüljön bele ebbe a mintába. (Hogy mi az, ami egyáltalán lehetséges, azt lásd lentebb.)
- A szimuláció folyamata legyen összhangban a tesztelni kívánt d függvénnyel. Ha például d érzéketlen a hierarchikus szintekre (mint a topológiai differencia vagy a részfába-tartozási divergencia alapján számított függvények) akkor az összes lehetséges dendrogramok – és kladogramok – száma nyilván V_m (5.16 képlet), s ezeket kell véletlenszerűen előállítanunk. Erre példa Shao & Rohlf (1983). Ha d figyelembe veszi a hierarchikus szintek sorrendiségét is (partícióba-tartozási divergencia), akkor már a referencia-eloszlásba H_m féle dendrogramot vonhatunk bele (5.17 formula) – s ez jóval nagyobb az előzőnél. Ilyen szimulációs eseteket tárgyal Lapointe & Legendre (“double permutation algorithm”, 1991) és – kisebb hangsúllyal – Steel & Penny (“*D_{tip}*”, 1993)⁵. Legnehezebb a feladat a kofenetikus szintek alapján történő összehasonlításoknál. Itt is H_m -féle dendrogram-szerkezet képzelhető el, de a szintek teljesen véletlenszerű megadása végtelen számú különböző dendrogramra vezet. Bár Lapointe

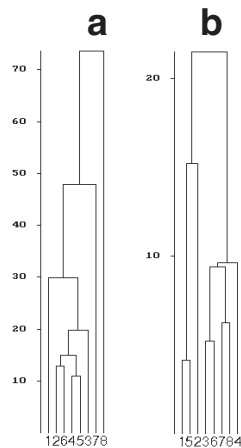
5 A **SYN-TAX** 5.02 program dendrogram szimuláló rutinja ezt állítja elő a “véletlen agglomerációs algoritmus” alapján, amely eredményét tekintve ekvivalens a többi módszerrel.

& Legendre (1991: 189) elképzelhetőnek tartják a hierarchikus szintek véletlenszerű kiválasztását, maguk is elismerik, hogy a lehetséges szinteket célszerűbb a két összehasonlítandó dendrogramban talált értékekre korlátozni, azaz maradjon a H_m darab lehetséges dendrogram.

- A fenti két bekezdés a “tisztá” Monte Carlo szimulációkat érinti. Nagyrészt tisztázatlan még ezek viszonya a permutációs tesztekhez. Két dendrogram összehasonlítása voltaképpen akkor felel meg logikailag a Mantel-tesztben is alkalmazott elgondolásoknak, ha a topológiát és a szinteket sem változtatgatnánk, s csak az objektumokat permutálnánk az összehasonlítandó dendrogramok egyikén (vagy mind a kettőn). Lapointe & Legendre (1995) láthatólag a teljes randomizációt részesíti előnyben a permutációs módszerrel szemben. Ugyanakkor nagyon is elképzelhető, hogy nem tekintjük az összes lehetőséget egyformán valószínűnek, mondjuk azért, mert a dendrogramot előállító módszer eleve kiemel bizonyos formákat (pl. a láncszerű elrendezést).

A 6. fejezet első ábráján (6.1) látható két – gyökér nélküli – kladogram összehasonlítása a permutációs logikát követi: az egyik fa teljesen bináris, a másikban pedig többszörös elágazások is vannak, s emiatt nem lenne indokolt a teljes randomizációt választani (Penny et al. 1993). A két fát topológiai differencia mátrixokkal felírva, majd közöttük euklidészi távolságot számítva a szerzők 126-ot kaptak eredményül. Ez pedig tízmillió permutációt elvégezve rendkívül szignifikánsnak bizonyult, hiszen ilyen vagy ennél alacsonyabb érték csak az esetek egy százzezred részében ($p < 0.00001$) adódott! Következésképpen a nyelvi és a genetikai törzsfák szignifikánsan megegyeznek, persze az ok megállapítása ennél jóval bonyolultabb “ügy”.

További illusztrációként a 9.3.1 rész második példájában említett két távolságmátrixból a csoportátlag módszerrel osztályozást hajtottunk végre (9.13 ábra). Miután a távolságmátrixok sem voltak szignifikánsan hasonlóak, várhatóan a dendrogramok sem lesznek azok. Két értéket számoltunk ki, a mátrix korrelációt a dendrogramokat leíró topológiai differenciák mátrixából, ill. a particióba-tartozási divergenciák mátrixából. Az első korreláció $-0,01$, a második pedig $-0,22$ amelyeket 1000-1000 teljesen véletlenszerűen előállított dendrogram-párra kapott értékkel összevetve megállapítható, hogy valóban: a dendrogramok éppen annyira hasonlóak, amennyire a véletlenszerű dendrogramoktól elvárható. A két korreláció között azonban van némi különbség: a topológiai differenciákra kapott érték majdnem megfelel a szimulált várható értéknek ($-0,004$), míg a másik korreláció igen közel van a szignifikáns *eltérés* küszöb-



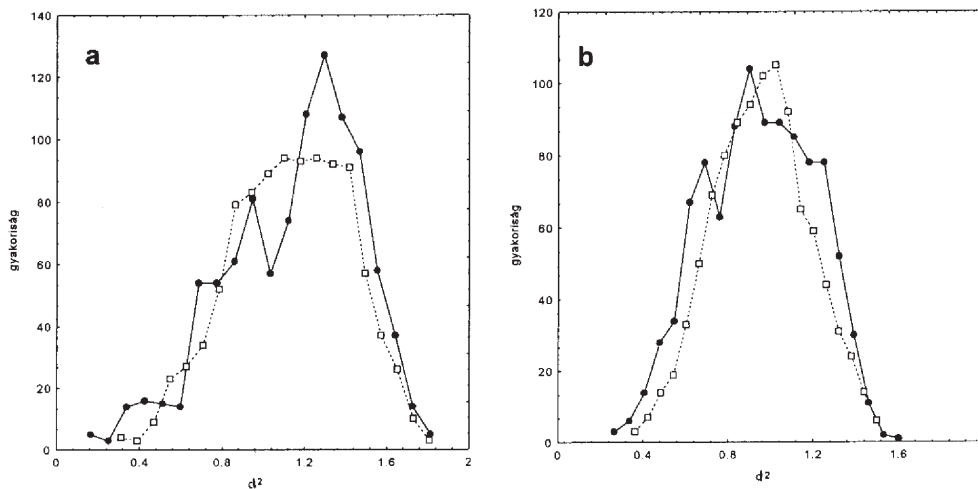
9.13 ábra. Az A1 táblázat mintavételi egységeinek osztályozása a csoportátlag módszerrel, euklidészi távolságokból, az egyszikűek (a) és a kétszikűek (b) figyelembevételével.

értékéhez ($\alpha = 0,05$ mellett ez $-0,267$, míg a várható érték itt is $-0,004$). Más szóval, az elágazási mintázatot tekintve annyira tér el a két dendrogram egymástól, amennyire a véletlen dendrogramok általában, a hierarchikus szintek sorrendiségét is tekintetbe véve viszont majdnem annyira eltérnek, ami random esetben is ritkaságnak számít. A dendrogramok összehasonlítása tehát önmagában nem mond sokat, de a felhasznált mérőszám eloszlását ismerve már többet hámozhatunk ki az eredményből.

9.3.5 Ordinációk

Most jóval könnyebb dolgunk van, mint a dendrogramok esetében, mert ordinációkat viszonylag könnyű előállítani. Podani (1991) javaslata szerint például teljesen véletlenszerű, egyenletes eloszlású koordinátákból indulhatunk ki a kívánt k dimenzióban. Ha a Prokrusztész módszer 9.21 formuláját alkalmazzuk, akkor ez amúgy is tartalmazza az egységnyi négyzetösszegre történő standardizálást, így a léptékkal nem kell törődnünk. Emellett azonban szóba jöhet az objektumok permutálása is a két összehasonlított ordinációk koordinátáinak a rögzítése mellett. A teljesen random eset és a permutációs átrendezés természetesen eltérő nullhipotézisnek felel meg, s ezt nem szabad elfelejtenünk az értékelés során. Az alábbi példa sejteti, hogy ez az eltérés a dimenziók számának növelésével csökken.

A 9.2.4 rész végén említett összehasonlításhoz állítsuk elő a d^2 referencia-eloszlásokat két és három dimenzióra, a tiszta random és a permutációs esetre is (9.14 ábra). “Valódi” szignifikancia-próba itt sem lehetséges, hiszen a két szóban forgó ordináció nem független egymástól: ugyanazokon az adatokon alapszik. Mindazonáltal érdemes az eloszlásokat figyelembe venni, hogy eldöntsük: a random ordináció-párokhoz miképpen viszonyul a kérdéses PCA és COA ordinációk távolsága. A várható értékek két, ill. három dimenzióra 1,14, ill. 0,95, függetlenül attól, hogy teljes vagy permutációs randomizálást hajtottunk-e végre. Ehhez viszonyítva a 0,1 és 0,309-es d^2 értékeket, már látszik, hogy két dimenzióban sokkal nagyobb az egyezés, mint háromban: jóval nagyobb, mint amit a két érték abszolút eltérése sugall. A teljesen véletlen, ill. a permutált esetek egyébként elsősorban abban térnek el, hogy az utóbbinál szélesebb a kapott értékek tartománya.



9.14 ábra. Ordinációk Prokrusztész-távolságának (9.21 formula) eloszlása 8 objektumra, két (a) illetve három dimenzióra (b), a teljesen véletlen (szaggatott vonal) és a permutált esetekben (vastag vonal). A vonalas összekötést a jobb áttekinthetőség kedvéért alkalmazzuk.

9.3.6 Nem-tervezett esetek – általában

A 9.3 alfejezetben eddig csak olyan hipotézisvizsgálatokat, ill. viszonyításokat említettünk, amikor előre tudtuk, hogy melyik OUC párokat fogjuk összehasonlítani, s többnyire eleve kettő pár volt csupán. Amint erre már rámutattunk, más a helyzet akkor, amikor mondjuk k számú OUC-t minden párosításban összehasonlítunk, s *ezután* szándékozunk kiválogatni a szignifikánsan eltérő (vagy hasonló) párokat. Az “elsőfajú” hiba halmozódását elkerülendő, egy szigorított próbát kell elvégeznünk, vagyis a kritikus küszöbértékeket kell növelnünk (jobboldali próbánál), ill. csökkentenünk (baloldali próbánál), amire többféle lehetőség is kínálkozik. Az alábbiakban kettőt mutatunk be.

- A küszöbértéket úgy választjuk meg, hogy az elsőfajú hiba elkövetésének a valószínűsége a végrehajtandó $g = k(k-1)/2$ számú összehasonlításokat elvégezve összesen se haladja meg az α -t. Ehhez az egy párra jutó α' értéke csökkentendő a következő módon:

$$\alpha' = 1 - (1 - \alpha)^{1/g} \quad (9.25)$$

A módszert eredetileg sok csoport átlagának többszörös összehasonlítására javasolták (vö. Sokal & Rohlf 1981a), de a mi esetünkben is alkalmazható. A tervezett esetre szóló eloszlásban (azaz a 9.3.1-5 részek bármelyikében említett, tiszta Monte Carlo módszerrel kapott eloszlásokban) ekkor az α' szintre határozzuk meg a küszöbértéket, s ehhez viszonyítjuk az egyes konkrét különbözőség-értékeket.

- A második lehetőség a szélsőértékek eloszlásának szimulációja. Tiszta Monte Carlo módszerrel véletlenszerűen előállítunk k darab OUC-t, s ezeket minden, vagyis g számú párosításban összehasonlítva megkeressük a legkisebb különbözőséget. (Elképzeltető az is, hogy ehelyett a randomizáció során az összes OUC-t permutáljuk.) Ezt megismételjük mondjuk 1000-szer, s ezzel előállítottuk a g számú, egymástól nem független összehasonlításra kapható *minimumok* eloszlását. Ezt követően megvizsgáljuk, hogy az eloszlásban milyen küszöbérték tartozik az α szinthez. Ezt vesszük alapul annak eldöntésére, hogy a g számú konkrét értékből melyek a szignifikánsan alacsonyak (következésképpen a megfelelő párok szignifikánsan hasonlóak).

Vizsgáljuk most meg, hogy partíciók összehasonlításában, a *MINDMT* függvény alapján, mit is jelent a fenti szigorítás. Tegyük fel, hogy 80 objektumunk van, s ezeket két osztályba soroltuk, ötféle – egymástól független – klasszifikációban. A tervezett esetben a 9.11 ábra bal szélső eloszlása jelentheti az alapot, ami $\alpha=0,05$ mellett azt mutatja, hogy a *MINDMT*=31 lesz a küszöbérték, s az ezzel megegyező vagy kisebb értékek szignifikánsan hasonló partíciókat jeleznek. Az öt eredmény összevetése 10 párban lehetséges, de ehhez már nem használhatjuk a fenti kritikus értéket. A minimumok szimulált eloszlásából $\alpha=0,05$ szinten a *MINDMT*=27 új küszöbértéket kapjuk, vagyis a 10 párból kevesebb lesz szignifikáns, mintha figyelmen kívül hagytuk volna a hibák akkumulációját. A 9.25 függvény szerint számolt α' (=0,005) szerint is egyébként ugyanez a küszöbérték adódik a 9.11 ábra eloszlásából.

9.4 Konszenzus eredmények

A “konszenzus” mostanában divatos szó, s ez a divat mintha az eredmények értékelésében is jelentkezne: az egy objektumhalmazra kapható alternatív, s egyformán elfogadható eredmények szintézisét értjük alatta. Eme új eredmény a k alternatíva hasonlóságait domborítja ki elsősorban, s általános – de nem kizárólagos – vélemény szerint jobban alkalmas az objektu-

mok jellemzésére, mint bármelyik kiinduló eredmény önmagában. Alternatív eredmények az objektumoknak hasonló módszerekkel történő elemzéséből származhatnak, s ekkor a módszerek kiválasztásában jelentkező szubjektív döntéseinket semlegesíthetjük a konszenzussal. A másik felhasználási lehetőség akkor adódik, amikor egy adott módszer több különböző végeredményre vezet (pl. egyformán optimális kladogramok), s ezeket akarjuk egyesíteni. Konszenzus módszerekre – áttekintve a vastkos irodalmat – elsősorban partíciók, fák (dendrogramok és kladogramok), valamint ordinációk esetében van szükségünk.

9.4.1 Konszenzus partíciók

Először a kemény partíciókról szólnunk, amikor minden partíció páronként elkülönült (diszjunkt) osztályokból áll. A feladat az, hogy m objektum $k \geq 2$ számú partíciójának (amelyekben az osztályok száma nem feltétlenül azonos) állítsuk elő a konszenzusát. Bár vannak próbálkozások arra nézve, hogy egyetlen ilyen konszenzust keressünk (lásd lentebb), Neumann & Norton (1986) rámutatott, hogy valójában számos, elvileg egyformán “jogos” konszenzus partíció adható meg. Ezek mindegyike előállítható az úgynevezett *szoros* (“*strict*”) konszenzus partícióból, amely a következőképpen definiálható: a szoros konszenzus partíció bármely j osztálya azokat és csak azokat az objektumokat tartalmazza, amelyek a k partíció mindegyikében egy osztályba tartoznak. Ez tehát nincs ellentmondásban egyik partícióval sem, de “hibája”, hogy rendszerint nagyon sok osztályból áll (végletes esetben éppen m -ből), s praktikus célokra nem mindig alkalmas. A szoros konszenzus partíció osztályainak fokozatos egyesítésével kapjuk azokat a *köztes* konszenzus partíciókat, amelyekben már az egy osztályba tartozó objektumok csak $k-1$, $k-2$ stb. partícióban voltak eredetileg együtt. Az egyesítésekkel eljutunk a konszenzus partíciók másik szélsőségéhez, a *tág* konszenzushoz, melynek j osztályában mindazok az objektumok szerepelnek, amelyek az osztály legalább egy tagjával legalább egyszer együvé tartoznak a k partícióban. Ennek “hátránya” viszont az, hogy néhány bizonytalan helyzetű objektum miatt a konszenzus osztályok száma 1-re csökken, ami semmire sem alkalmas. Ha egynél több ilyen osztály van, azok viszont abszolút érvényűek, hiszen a többtől teljesen izolált objektumokat egyesítenek magukban.

Az elmondottak illusztrálására vizsgáljuk meg 10 objektum alábbi felosztásait:

$$\begin{aligned} P_1 &= \{1, 2, 3, 4\} \quad \{5, 6, 7, 8, 9, 10\} \\ P_2 &= \{1, 2, 3, 4, 5, 6\} \quad \{7, 8, 9, 10\} \\ P_3 &= \{1, 2, 3, 4, 5\} \quad \{6, 7, 8, 9, 10\} \\ P_4 &= \{1, 2, 3, 7\} \quad \{4, 5, 6, 8, 9, 10\} \end{aligned} \quad (9.26)$$

Ezekből az alábbi *szoros konszenzus* adódik:

$$P_s = \{1, 2, 3\} \{4\} \{5\} \{6\} \{7\} \{8, 9, 10\}$$

Ezen osztályok egyesítéséből számos lehetőség írható fel a *köztes* konszenzusra, például:

$$P_c = \{1, 2, 3\} \{4, 5, 6\} \{7, 8, 9, 10\},$$

de szerepeltethetjük volna akár az összes kiinduló partíciót is. Végezetül a *tág* konszenzus is előállítható, ami ebben az esetben csupán egy triviális osztályból áll.

A fenti példából nyilvánvaló, hogy a szóba jövő lehetőségek száma már ilyen egyszerű esetekben is túlságosan nagy. A sok konszenzus-jelölt azonban nem tűnik egyformán egyenrangúnak. A fenti esetben már ránézésre látszik, hogy előállítható olyan partíció is, amely az

alternatívák többségének tükrözi a “véleményét”. Ez az úgynevezett *többségi* (“*majority rule*”) konszenzus, amely a k alternatíva több, mint a felével egyező. A fenti példában ez két osztályra nem megy az 5. objektum ellentmondásos helyzete miatt (és k értéke páros, ami megnehezíti a döntést), de három osztályra már megoldható:

$$P_T = \{1, 2, 3, 4\} \{5, 6\} \{7, 8, 9, 10\}$$

A P_T -ben egy osztályba csak olyan objektumok tartoznak, amelyek legalább három kiinduló partícióban is együtt vannak. Ez természetesen egyike a köztes konszenzus partícióknak, s ugyanakkor nem biztos, hogy csak egyetlen egy ilyen többségi konszenzus eredmény állítható elő. A köztes konszenzusok közül történő választás másik lehetőségét a *medián* konszenzus (vö. Barthélemy & Monjardet 1981) jelenti, amelyhez alkalmaznunk kell valamilyen, partíciók összehasonlítására alkalmas d különbözőségi függvényt. A P_m partíció a k számú partícióra nézve akkor tekinthető medián konszenzusnak, ha teljesül az alábbi feltétel:

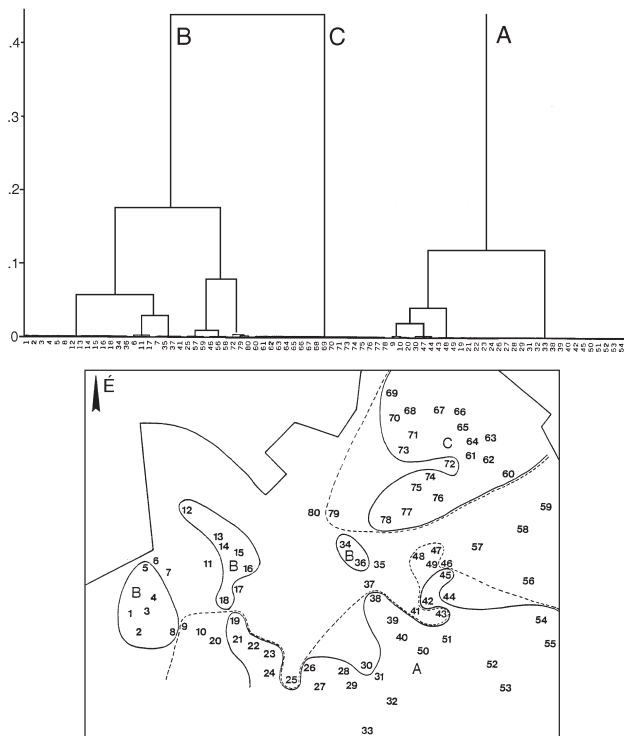
$$\sum_{i=1}^k d(P_m, P_i) = \min_c \sum_{i=1}^k d(P_c, P_i) \quad (9.27)$$

amelyben a c index utal az összes lehetséges köztes konszenzusra. Szavakban: a medián konszenzus esik átlagosan a legközelebb az összes többihez, s persze most sem biztos, hogy csak egyféle ilyen létezik.

A szoros konszenzus meghatározása könnyen megy, de a többségi és a medián konszenzus előállítására már bonyolultabb (valójában *NP*-nehéz) probléma, főleg nagy m esetén. A gyakorlatban ezért egyszerűbben alkalmazható a Podani (1989a) javasolta heurisztikus jellegű agglomeratív osztályozó módszer. Kiindulásképpen előállítunk egy $\mathbf{D}_{m,m}$ távolságmátrixot, amelyben d_{jk} azoknak a partícióknak a száma, amelyekben a j és k objektumok *nem* tartoznak egy osztályba. Ezt a globális optimalizálás (5.2.4 rész) módszerével elemezzük, amely egyidejűleg veszi figyelembe a konszenzus szorosságát (osztályokon belüli távolságtátlagok) és az izolációt is (osztályok közötti távolságtátlagok). A kapott hierarchia egy köztes konszenzus-sorozat, amelyben a kívánt osztályszámra kért konszenzus eredmény könnyen meghatározható. A hierarchia szemléletbeli többlete egyébként az, hogy megmutatja: a k alternatív partíciónak több szintű konszenzusa lehetséges. Diday & Simon (1976) – a konszenzusra való utalás nélkül – már felhasználta a \mathbf{D} mátrixot a teljes lánc-módszerrel történő osztályozásban.

A hierarchikus konszenzus generálást egy valós példával illusztráljuk (Podani 1989a). Az volt a feladat, hogy 80 cönológiai mintavételi egység (dolomit-sziklagyepek, Sashegy, Budapest) hat alternatív partíciójából (mindegyikben 3-3 osztály) állítsuk elő a konszenzus partíciót, amely ezután a kvadrátok helyét feltüntető térképre vetítve egy általánosabb érvényű Y vegetációtérképet ad, mint bármelyik kiinduló osztályozás (9.15 ábra). Az A, B és C jelű osztályok a nyílt, a rozsnokos zártabb illetve a nyúlfarkfüves zárt gyepársulásokkal azonosíthatók.

Lágy partíciókra a *medián* konszenzus-konceptió adaptálható a legegyszerűbben (Podani 1990). Adott k számú lágy osztályozás medián konszenzusa egy olyan osztályozás, amelynek a többitől vett távolság-négyzetösszege minimális. Ha u_{ijh} jelöli a j objektum h -adik osztályba tartozási súlyát az i -edik partícióban, u_{cjh} pedig a súlyérték a keresett konszenzus partícióban (F_c), valamint p az osztályok száma minden partícióban, akkor a minimalizálandó mennyiség



9.15 ábra. Partíciók hierarchikus konszenzusa. A 80 objektum 6, három-osztályos partíciójának értékelése a globális optimalizálás módszerével (felső ábra), majd a kapott szoros konszenzus három fő csoportjának (vastag vonal) és a közelítőleg optimális 3-osztály szintű konszenzusnak (szaggatott vonal) térképes ábrázolása (alsó ábra).

a következő:

$$SSQ_c = \sum_{i=1}^k \sum_{j=1}^m \sum_{h=1}^p (u_{ijh} - u_{cjh})^2 \quad (9.28)$$

Ennek megtalálása teljes kikereséssel is történhet, vagyis a k osztályozást az osztályok minden lehetséges permutációjában egymáshoz illesztjük. (Ha az osztályszámok különbözőek, akkor szükség szerint üres osztályokat adunk a partíciókhoz.) Az illesztés összesen $p!^{k-1}$ féle módon történhet, ezért a teljes kikeresés nemigen működik sok osztályra és sok osztályozásra. A centroid módszer (Podani 1990) egy heurisztikus közelítő eljárás arra az esetre, ha a probléma teljes kikereséssel nem oldható meg.

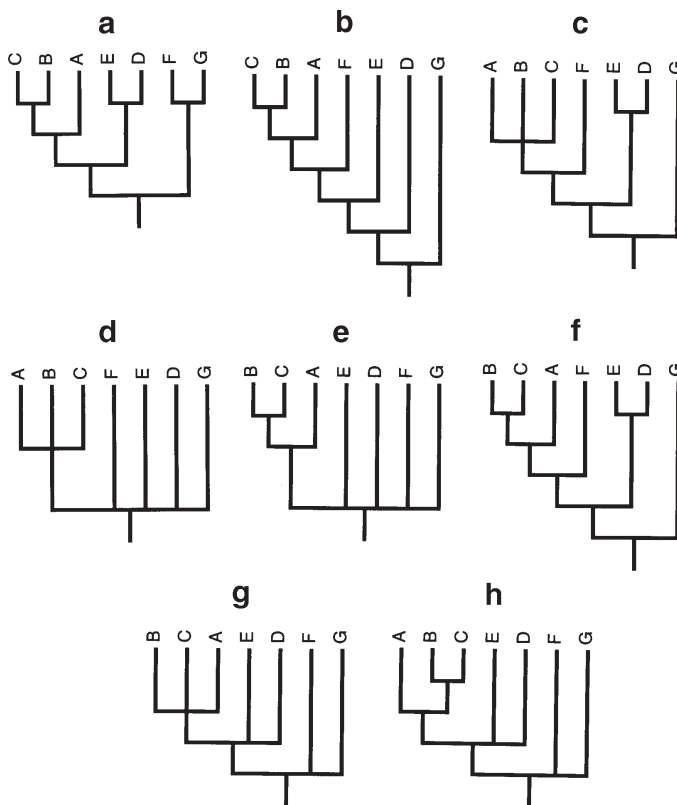
Mivel a kemény partíciók a lágy osztályozások speciális esetei, felmerül a lehetőség, hogy az előzőek konszenzusaként egy lágy partíciót keressünk. A kemény partíciók – nevükhöz híven – túl merevek, nem mindig alkalmasak a finomságok bemutatására, és sokszor túl nagy osztályszámot igényelnek az egyértelmű megoldáshoz (vö. a 9.26 példa 5. objektumának esetével). A kemény partíciók lágy konszenzusa viszont – a meglehetősen szabadon változtatható súlyértékek segítségével – jobban kifejezheti a hasonlóságokat és eltéréseket.

A 9.26-ban megadott négy partíció két osztályos lágy konszenzusaként a teljes kikeresés módszerével a következő – a már ismert módon értékelhető – eredményt kapjuk (helykímélés végett az alábbi súlymátrixban kivételesen az objektumok az oszlopok):

$$\begin{matrix} 1,0 & 1,0 & 1,0 & 0,75 & 0,5 & 0,25 & 0,75 & 0,0 & 0,0 & 0,0 \\ 0,0 & 0,0 & 0,0 & 0,25 & 0,5 & 0,75 & 0,25 & 1,0 & 1,0 & 1,0 \end{matrix}$$

9.4.2 Konszenzus fák

Alternatív eredmények konszenzus formában történő összesítése napjainkban elsősorban a szisztematikában jelent izgalmas feladatot. Akár dendrogramról, akár kladogramról van szó, az esetek túlnyomó többségében a fa topológiai szerkezetének az egyesítése a feladat, az élek hossza, illetve a hierarchikus szintek általában “feledésbe merülnek” (a kivételt lásd lentebb). Így voltaképpen rendszerint mindegy, hogy dendrogramokról vagy kladogramokról beszélünk. A konszenzus fák általában csak nem bináris elágazások megengedésével képesek magukba sűríteni az információt (eltekintve persze attól, amikor minden eredmény egyforma). Mindez különösen jól megfigyelhető a *szoros konszenzus fa* (“*strict consensus tree*”, Sokal & Rohlf 1981b, Swofford 1991) esetében, amely – hasonlóan a szoros konszenzus partíciókhoz – csak olyan osztályokat tartalmazhat, amelyek mindegyik eredményben megvoltak. Más szóval: ha az objektumok egy csoportja megjelenik a konszenzus fán, akkor az bizonyosan megvan az összes többi eredményben is, és ez hallatlanul megkönnyíti a konszenzus értékelését. (Vizsgáljuk meg a 9.16 ábrán látható **a**, **b** és **c** fákat, amelyeket a **d** szoros konszenzus fa egyesít. Az {A,B,C} osztály az egyetlen, amely mindhárom kiindulásban megvan. Ebből látszik a szoros konszenzus hátránya, a politómiák esetleges túlbujjánzása.) Szélsőséges esetben, mint pl. a 6.18 ábra kladogramjainál, a szoros konszenzus fa minden ága közvetlenül a gyökérből indul ki (ún. “*bokor*”), tehát nem túlzottan érdekes a számunkra. Voltaképpen minél kevesebb politómia van a konszenzus dendrogramon, annál nagyobb az egyetértés a

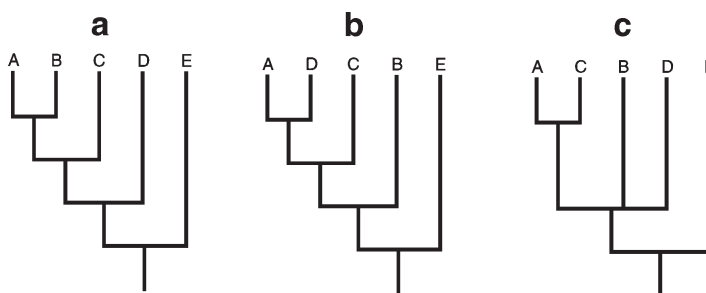


9.16 ábra. Konszenzus fák. Az alternatív eredmények (**a-c**) egyesítése a szoros konszenzus (**d**), a félig szoros konszenzus (**e**), a többségi (>50 %) konszenzus (**f**), az Adams konszenzus (**g**) és a “durchschnitt” konszenzus (**h**) szerint. Figyeljük meg, mennyire “sikeresek” az egyes konszenzus fák az **a-c** dendrogramok egyesítésében!

kiinduló dendrogramok között (ezt méri az egyik konszenzus index, l. lentebb).

A *félíg szoros* (“*semi-strict*”) vagy *kombinációs* konszenzus (Bremer 1990, Swofford 1991, Quicke 1993) némiképpen toleránsabb az előzőnél. Itt az a kikötés, hogy a konszenzus eredmény osztályai *nem lehetnek ellentmondásban* a kiinduló fák egyikével sem. Mivel az {ABC} csoport léte a 9.16c fában *nem zárja ki* az {AB} csoportot, a konszenzus fa e három objektumra nézve már bináris szerkezetű lesz (9.16d). Ha a kiinduló fák teljesen binárisak, akkor a szoros és a félíg-szoros konszenzus szükségképpen megegyező. A *többségi* konszenzus fa (Margush & McMorris 1981) még több eltérést enged meg, mert egy konszenzus osztály megjelenésének az a feltétele csupán, hogy a kiinduló fák több mint a felében meglegyen. Ennek megfelelően a példa dendrogramok többségi konszenzusa (9.16e ábra) már teljesen bináris szerkezetű: osztályai ({BC}, {ABC}, {ABCF}, {DE}, {ABCDE} és {G}) a kiinduló eredmények közül legalább kettőben felfedezhetők. A többségi feltétel persze változtatható, sok dendrogram esetén 50 %-nál magasabbra is emelhetjük a tőrésarányt. Könnyen belátható, hogy két fa esetében a szoros és a többségi konszenzus megegyező. Megemlítendő még a *medián* konszenzus fa is (Barthélemy & Monjardet 1981, Barthélemy & McMorris 1986), amelynek meghatározása ugyanazon az elven alapszik, mint a medián konszenzus partícióé (9.27), csupán egy megfelelő függvényt kell találnunk a fák összevetésére. Ha ez a partíciós metrika (9.12), akkor az 50 %-os többségi konszenzus *is* egy medián fa (Barthélemy & McMorris 1986), amely persze nem feltétlenül bináris. Ha ragaszkodunk ahhoz, hogy a medián konszenzus mindenképpen teljesen dichotomikus legyen, akkor Penny et al. (1982) tanácsait követhetjük (“medián bináris fa”, Swofford 1991).

Konszenzus fák előállítására a történetileg legelső javaslat érdekes módon nem egyezik meg az egyik eddig tárgyalt, az egyes részfákra ill. osztályokra “kiélezett” módszerrel sem. Adams (1972) elképzelése az volt, hogy megvizsgálja: miképpen “hasadnak” a nagyobb osztályok egyre kisebb csoportokra, ahogy a gyökértől kezdve a fa végágai felé közeledünk. Először a gyökérszint meghatározott partíciókat állítjuk elő, s képezzük ezeknek a szoros konszenzusát. A 9.16a-c dendrogramokon az így kapott partíciók: {ABCDE}{FG}, és {ABCDEF}{G} kétszer. Ezekből az {ABCDE}{F}{G} szoros konszenzus partíció adódik, a konszenzus fában tehát egy trifurkációval kezdünk. Ezután már csak az {ABCDE} csoport további elemzése van hátra, hasonló elvek szerint, a 9.16g ábrán látható eredménnyel. Az Adams-féle módszerrel szemben felhozott leggyakoribb ellenérv az, hogy olyan osztályok is jelentkezhetnek benne, amelyek az eredeti fák egyikében sem szerepeltek (9.17 ábra).



9.17 ábra. Az Adams konszenzus (c) fő “hibája”, hogy a kiinduló fák (a-b) egyikében sem szereplő csoport(ok) is lehet(nek) benne, viszont elég jól kifejezi a legtöbb objektumpár relatív közelségi viszonyait.

A hierarchikus szinteket is figyelembe vevő, tehát dendrogramokra és nem kladogramokra alkalmas módszer az ún. “*durchschnitt*” (keresztmetszet) konszenzus (Neumann 1983, Smith & Phipps 1984). Ez felülről lefelé haladva összeveti a dendrogramok elmetszéséből kapott partíciókat, és minden szintre előállítja a szoros konszenzus partíciót. Ezek osztályai alkotják az egyes szinteknek megfelelő elágazásokat a konszenzus fában. Kellő mélységig lemelve a dendrogramokon megkapjuk az objektumok triviális felosztását m osztályra, s a konszenzus fa szerkesztése befejeződik (9.16h ábra). Mint a leírásból is látszik, a hierarchikus szintek tényleges értékei végül is nem számítanak, csak a szintek sorrendisége. Itt is könnyen kapunk olyan csoportokat, amelyek a kiinduló dendrogramok egyikében sem figyelhetők meg.

Végezetül a *metszéses* módszert (Finden & Gordon 1985) érdemes megemlíteni, ami nem törekszik arra, hogy az összes objektumot szerepeltesse a konszenzus fában. Ennek az a célja, hogy a “kilógó”, vagyis az ellentmondó helyzetű ágak fokozatos eltávolításával egy olyan fa maradjon, amely már minden tekintetben összhangban van a kiinduló fákkal. Ilyen fa persze sokféle lehet, s ezek közül azokat érdemes kikeresni, amelyek a legnagyobb számú objektumot tartják meg (“*largest common pruned tree*”). Ennek akkor van igazán értelme, ha néhány objektum rendkívül ingadozó helyzete okozza csupán az eltéréseket. A módszer hátránya az, hogy nem ismerjük a legnagyobb fa biztos megtalálásának a módját (az összes lehetőség megvizsgálása vagy m -re – a már ismert elvi okok miatt – kizárt).

9.4.2.1. *Konszenzus indexek.* Amennyiben sok fa konszenzusát nem egy másik fa formájában, hanem csupán egy mérőszámmal akarjuk kifejezni, akkor a konszenzus indexeket (Rohlf 1982) alkalmazhatjuk. Ezek valójában hasonlóságot fejeznek ki 0 és 1 között, és $k=2$ esetére már tárgyalhattuk volna őket a dendrogramok összehasonlításáról szóló fejezetben is (9.2.3). Miután azonban igazi értelmük akkor van, ha az alternatívák száma nagy, inkább most kerítünk rájuk sort. A témáról Rohlf imént idézett cikke és Swofford (1991) eléggé kimerítő összesítést ad, így csak néhány indexet említünk most meg.

A legegyszerűbb közülük a Colless (1980)-féle “*consensus fork index*”, ami a konszenzus fa bináristól való eltérésének a jelzője. Ez a konszenzus fában lévő nem triviális⁶ osztályok száma, osztva $m-2$ -vel, az összes lehetséges nem triviális osztályok számával. Az index értéke a 9.16d dendrogram esetén 0,2 (mert csak egy osztály jelent meg az öt lehetségesből), 0,4 a 9.16g fára, és természetesen 1 a teljesen bináris 9.16f fára. A Mickevich-index (1980) az egyes konszenzus osztályokat méretük szerint súlyozza, és e tekintetben továbblépést jelent az előző indexhez képest. Ha az i osztály mérete n_i , akkor ennek fontossága $N_i = \min\{n_i - 1, m - n_i\}$. Eme fontosság értékeket összegezve, majd elosztva azt az index lehetséges maximumával, megkapjuk a kívánt mértékszámot. A fent említett három esetben az index értékei: 0,222, 0,444, ill. 0,888. Végezetül, Schuh & Farris (1981) egészen más súlyozást ajánl: minden osztályra számoljuk ki az objektumpárok számát, vagyis $N_i = n_i(n_i - 1)/2$, s ezeket összegezzük (“*levels sum*”). A példákban (a 9.16 d, g és f fákra) 3, 26 és 13 adódik. Az összegek a maximummal akár oszthatók is, hogy az index értéke a $[0,1]$ intervallumba kerüljön. A normalizálás problémája minden említett index esetében az, hogy a maximum nagymértékben függ a fa alakjától (“láncszerű”-nél nagyobb, mint a teljesen szimmetrikus fában).

6 Triviális egy osztály, ha az összes objektumot magában foglalja vagy csak egy objektumot tartalmaz.

9.4.3 Konszenzus ordinációk

Ha ugyanannak az m objektumnak k különféle ordinációja készült el, akkor érdekes lehet ezek bizonyos értelemben véve konszenzus – vagy pontosabban megnevezve – *átlagos* ordinációja is⁷. A dendrogramokhoz képest megintcsak könnyű dolgunk van, mert elegendőnek látszik a medián konszenzus elvének alkalmazása: a keresett ordináció legyen egy olyan konfiguráció, amelynek a többitől vett távolságnégyzet-összege (9.13 formula) minimális. Más szóval, a $k+1$ -edik ordinációra a lehető legjobban illeszkedjenek a kiinduló ordinációk. Ezt a Gower (1975) javasolta általánosított Prokrusztész elemzés (“*generalized Procrustes analysis*”) segítségével állíthatjuk elő. A módszer, mint sok más többváltozós eljárás, iteratív lépések sorozata. Ez alól csak a $k=2$ eset kivétel, mert ekkor az illesztés gond nélkül megoldható, s a konszenzus koordináták átlagolással kiszámíthatók.

Először minden ordinációt centrálunk és egységnyi négyzetösszege normálunk (egyébként az átlagos ordinációnak nem volna sok értelme). Ha \mathbf{X}_i jelöli az i -edik ordinációt, és \mathbf{Y} a keresett konszenzus eredményt, akkor a feladat az egyes \mathbf{X}_i konfigurációk legjobb illesztése (elforgatása) \mathbf{Y} -hoz, melynek eredménye \mathbf{Y}_i :

$$\mathbf{Y}_i = \rho_i \mathbf{X}_i \mathbf{H}_i \quad (9.29)$$

Ebben ρ_i egy skálaparaméter, és \mathbf{H}_i a rotációs mátrix, ami az alábbi függvény minimalizálásával kapható meg:

$$RES = \sum_{i=1}^k \text{tr}[(\mathbf{Y} - \mathbf{Y}_i)'(\mathbf{Y} - \mathbf{Y}_i)] \quad (9.30)$$

(RES jelöli a reziduális négyzetösszeget). A rotációs illesztést úgy hajtjuk végre, hogy az eredeti k ordináció teljes négyzetösszege változatlan maradjon:

$$SSQ = \sum_i \text{tr}(\mathbf{X}_i' \mathbf{X}_i) = \sum_i \text{tr}(\mathbf{Y}_i' \mathbf{Y}_i) = \sum_i \rho_i^2 \text{tr}(\mathbf{X}_i' \mathbf{X}_i) \quad (9.31)$$

A konszenzus konfiguráció az alábbi átlagot jelenti:

$$\mathbf{Y} = 1/k \sum_i \mathbf{Y}_i \quad (9.32)$$

Kiindulásképpen az \mathbf{X}_2 -t illesztjük \mathbf{X}_1 -re, majd \mathbf{X}_3 -at az \mathbf{X}_2 és \mathbf{X}_1 átlagára, és így tovább mindaddig, amíg \mathbf{X}_k is sorra nem kerül: ezt az összes előzőnek az átlagához illesztjük. Ez jelenti a kezdeti konszenzus konfigurációt, amely sok iterációs illesztéssel javítandó, amíg a RES változása elhanyagolhatóvá nem válik. Minden egyes lépés forgatásokat és skálafaktor-át-számítást tartalmaz, de az utóbbi akár el is hagyható (bár ezt nem igazán ajánljuk). A végső ordinációnak teljesen önkényes az irányultsága, s emiatt célszerű előállítani az \mathbf{Y} PCA-ordinációját, majd az összes kiinduló ordinációt ehhez illeszteni.

A teljes négyzetösszeg (SSQ) a reziduális négyzetösszegeből (RES) és a konszenzus négyzetösszegeből ($SSQ-RES$) tevődik össze. Az előző a konszenzussal nem magyarázott rész, ami annál nagyobb, minél rosszabb az ordinációk átfedése. Meghatározható az egyes objektumok, illetve az egyes ordinációk százalékos hozzájárulása RES értékéhez, s így azonosíthatók a “kilógó” (az átlagtól leginkább eltérő) objektumok és ordinációk.

A módszert illusztráló példában a kvadrátnagyság ordinációkra gyakorolt hatását igyekszünk kiküszöbölni az általánosított Prokrusztész módszerrel. A 9.4.1 részben említett példáról

7 A módszer fontos alkalmazási területe még a morfometriai értékelés, amint arról szuperpozíciós módszerek néven a 7.6.2 részben már meg is emlékeztünk.



9.18 ábra. Általánosított Prokrusztész analízis. Az összehasonlított hat ordináció ugyanannak a 80 cönológiai felvételnek az elemzése változó kvadrátnagyság mellett. Szabálytalan idomok jelzik az egy-egy felvételhez tartozó hat méret pozícióit. A konszenzus ordináció így csak közvetve szerepel a rajzon, a pontokat nem ábrázoltuk.

van szó, de most a 80 db kvadrátot hat méretre ($1,5 \times 1,5 \text{ m}^2$ -től $4 \times 4 \text{ m}^2$ -ig, félméteres oldalhossz-növekedéssel minden lépésben) elemezzük a PCoA módszerrel. A kapott hat ordináció optimális egymásra illesztését – az első két dimenzióra – az általánosított Prokrusztész módszerrel érjük el. Az ordinációs diagram túlságosan áttekinthetetlen lenne, ha minden pontot feltüntetnénk, ehelyett csupán egy szabálytalan síkidom jelzi minden egyes kvadrát “mozgásterét”, amit a méretbeli változás határozott meg. A 9.18 ábrán jól látható, hogy a mintavételezés általában nem hatott lényegesen a kvadrátok helyzetére, és a – nyílt gyeptől a zártig húzódó – háttérgradiens íves torzulása is lényegileg érintetlen maradt.

9.5 Különböző típusú eredmények összevetése

9.5.1 Numerikus összehasonlítások

Eltérő típusú eredmények összehasonlító értékelése numerikus módon csak akkor sikerülhet, ha azokat megegyező formába tudjuk alakítani. Ez a standard típus pedig a *szimmetrikus mátrix*, amely az m objektum kapcsolatrendszerét összesíti minden lehetséges párosításban. Két mátrix a korrelációs együtthatóval (9.1 formula) hasonlítható össze legcélszerűbben; az euklidészi távolság vagy más függvények kevésbé használhatók, hacsak valamilyen standardizálással azonos skálára nem hozzuk a két mátrix értékeit. Szóba jöhet esetleg a rang-korreláció is. Eltérő típusú eredmények mátrixos megjelenítésére és korrelációs összehasonlítására már láttunk példát az 5.5.1 részben: a kofenetikus mátrix és a hierarchikus osztályozás alapjául szolgáló különbözőségek mátrixa közötti kapcsolatot, ezáltal a “torzítást” mértük ily módon. Egy additív fában lévő patrisztikus távolságok és a kiinduló távolságmátrix korrelációja is kiszámítható, hogy megtudjuk: milyen hű ábrázolása az illető gráf az objektumok távolságviszonyainak. Miután partíciók és ordinációk is leírhatók $m \times m$ -es mátrixok formájában, a korreláció alkalmazása más eredmény-kombinációban is szóba jöhet.

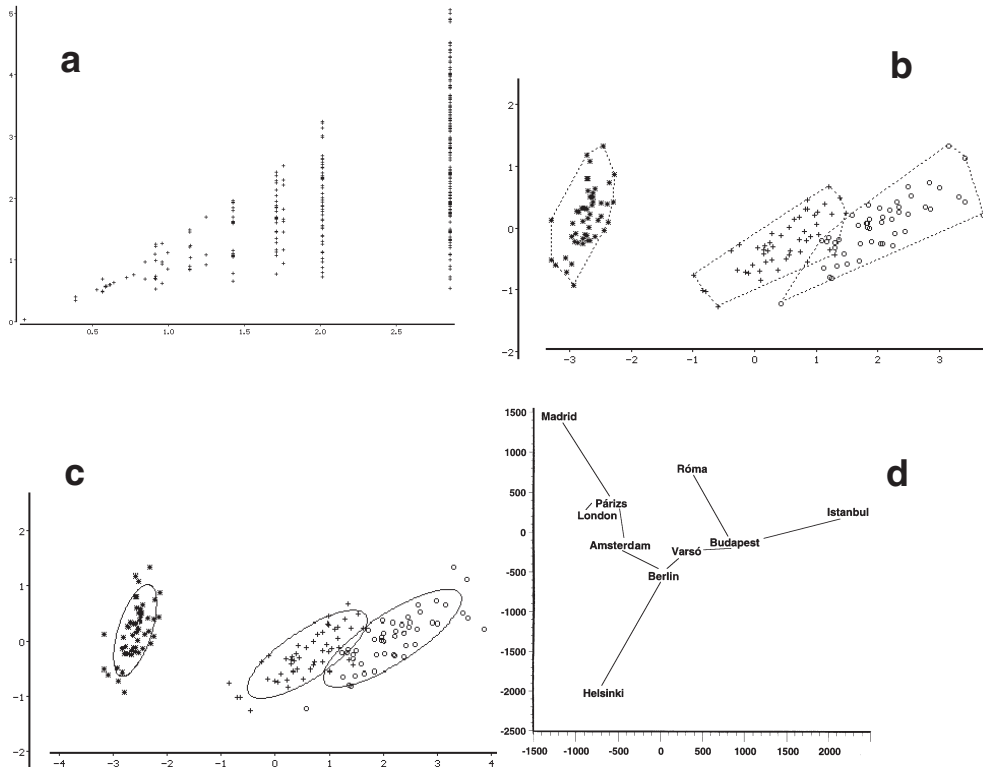
9.5.2 Grafikus összehasonlítások

Mátrix formába hozott különböző típusú eredmények grafikus összehasonlítására tökéletesen alkalmas lehet a 9.2.1 rész végén említett koordináta-rendszeres módszer. Az eredmények

együttes ábrázolása azonban ennél is gyakrabban szerepel a gyakorlatban. Az alapot rendszerint egy kétdimenziós ordináció szolgáltatja, amelyre a másik eredményt egyszerűen rárajzoljuk (már láttunk ilyent, a 7.2 és a 8.10b ábrán). Az egymásra vetítés lehetővé teszi az egyik típusú eredmény feltételezhető hiányosságainak a kiküszöbölését a másik típusúval, amint azt az alábbi példák is bemutatják. A grafikus értékelést, és főleg az együttes ábrázolást *melegen ajánlhatjuk* a többváltozós adatelemzés szinte minden területén.

Először egy távolságmátrix és a belőle származó dendrogram grafikus értékelését mutatjuk be a "mátrix plot" segítségével. Vegyük ismét elő – mondjuk – az 5.8 ábra **c** dendrogramját, amely a 4.3c ábra pontjainak a csoportátlag módszerrel történt osztályozását szemlélteti. Az 5.5.1 részben már kiszámítottuk a kofenetikus korrelációt (0,662). Most megvizsgáljuk, hogy mi is áll ennek az értéknek a háttérében. A 9.19a ábrán a vízszintes tengelyen a dendrogram-beli hierarchikus szintek, a függőleges tengelyen pedig a távolságmátrix értékei szerepelnek. Mivel összesen legfeljebb $m-1$ különböző kofenetikus értékünk van, a diagram pontjai oszlopokban sorakoznak. A háromszögszerű elrendeződés pedig jól mutatja, hogy egy adott hierarchikus szint valójában az eredeti távolságértékek tartományát helyettesíti, amely különösen széles lehet az utolsó fúziók esetében.

Nem-hierarchikus osztályozásokat ellenőrizhetünk, ha azokat az ordinációs síkban (leggyakrabban az első két dimenzióban), az egy osztályba tartozó objektumok körülrajzolásával



9.19 ábra. Eltérő típusú eredmények grafikus összehasonlítása (**a**: dendrogram *versus* távolságmátrix), illetve *együttes ábrázolása* (**b**: ordináció és osztályozás konvex burkokkal, **c**: ordináció és osztályozás valószínűségi ellipszisekkel, **d**: ordináció és a minimális feszítőfa).

azonosítjuk. Ezt megtehetjük csak úgy “szemre” is, de legjobb, ha a pontok köré rajzolható minimális konvex burkot vagy osztály-poligont (“*convex hull*”, ill. “*classification polygon*”) ábrázoljuk: ez minden pontot körülölel, s nincsenek rajta “horpadások” (belső szögei nem nagyobbak 180° -nál). Nyilvánvaló, hogy az osztályok annál “jobbak”, minél kisebb területűek a burkok, s minél kevésbé fednek át egymással. Ha erőteljes átfedés jelentkezik, akkor a partíció kérdőjeleződik meg. Példaképpen az *Iris* egyedek *a priori* osztályozását (három fajra) és a PCA ordináció első két tengelyének együttes ábrázolását mutatja be a 9.19b ábra. A diagram persze csak megerősíti a már jól ismert tényt: az egyik faj jól elkülönül, a másik kettő viszont némi átfedésben van egymással (a vizsgált négy tulajdonság alapján).

Másik lehetőség partíciók és kétdimenziós ordinációk összevetésére, az osztályokhoz tartozó *valószínűségi ellipszisek* (“*ellipses of equal concentration*”) berajzolása (Mardia et al. 1979, Lagonegro & Feoli 1985). Az ellipszis olyan területet jelöl ki, amely az illető osztály (pl. taxon) egyedeinek $100(1-\alpha)$ százalékát tartalmazza (α az általunk kiválasztott valószínűségi szint). Ennek akkor van létjogosultsága, ha a koordináták osztályokon belüli eloszlásáról jó okkal feltételezzük, hogy normális, és a mintavételezés (az objektumok kiválasztása) randomizált. Ezek a feltételek ritkán teljesülnek. Ennek ellenére érdemes megvizsgálni az *Iris* példa 95 %-os valószínűségi ellipsziseit a PCA ordináción (9.19c ábra). Az egyes taxonoknak megfelelő ellipszisek átfedése hasonló a klasszifikációs poligonokéhoz.

Az 5.4.3 részben már említettük, hogy maguk az ordinációk a *minimális feszítőfa* berajzolásával ellenőrizhetők, mert két objektum közelsége az ordinációs síkon esetleg megtévesztő lehet. A 9.19d ábrán az európai nagyvárosok PCoA ordinációjára (7.18 ábra) vetített minimális feszítőfa látható. A gráf szépen “terített”, azaz összhangban van az ordinációs elrendeződéssel. Voltaképpen ezt vártuk, hiszen a mutatott dimenziók távolság-ábrázolási hűsége 84 %-os. Ha azonban a feszítőfa élei kereszteződnek, vagy pedig két közel lévő pont között csak több másik pont közbeiktatásával jutunk el a gráfban, akkor az ordináció – a kérdéses pontokra legálábbis – félrevezető, és további dimenziókat is meg kell vizsgálni.

9.6 Irodalmi áttekintés

Szinte nem is gondolnánk, hogy az eredmények összehasonlításának irodalma mennyire jelentős és lassan áttekinthetlenné váló “massza”. Ez persze nincs minden területen így. Ha alaposan szétnézünk a releváns biológiai szakirodalomban, akkor feltűnhet, hogy például az ökológiai/cönológiai művek az eredmények összehasonlítását kevésbé tekintik fontosnak, mint a taxonómiai/evolúciobiológiai tárgyúak, sőt többnyire teljesen negligálják azt. Ez alól üdvözítő kivétel Digby & Kempton (1987), ill. Orlóci (1978). Az elsőben az ordinációk összehasonlítása, a másodikban pedig a partíciók összevetésére alkalmas információelméleti módszerek szerepelnek részletesebben. Az összehasonlító értékelések fontosságát viszont szinte minden rendszertani és kladsztikai monográfia kiemeli, így például Sneath & Sokal (1973). Külön megemlítendő Rohlf & Sokal (1981a) összefoglalása a numerikus taxonómiai összehasonlítások típusairól és módzatairól. A viszonylag új irodalmat áttekintve megállapítható, hogy a konszenzus kladoگرام megtalálása (pl. Swofford 1991) izgatja leginkább a biológusokat, melynek egyik fő oka már ismert előttünk: sok egyformán optimális kladoگرام “bukkanhat” fel a keresgélés közben.

Az összehasonlítás igénye sok esetben úgy jelentkezik, hogy a módszerek közül melyik produkál olyan eredményt, amely legjobban megfelel egy előre megszabott feltételnek. Ezt a lehetőséget nem is említettük eddig, pedig van rá példa bőven (Fasham 1977, Gauch et al. 1977, 1981 stb.). Ilyenkor a feltett kérdések így hangzanak: mely ordinációs módszer van legkevésbé kitéve a patkó-jelenségnek? melyik alkalmas leginkább egy feltételezett (vagy szimulált) háttér-grádiens feltárására?, és más hasonlók. Itt tehát nem az egyes eredmények direkt összevetése a cél, amit jelen fejezetben hangsúlyoztunk, hanem egy külső szempontnak

9.2 táblázat. Eredmények összehasonlítására alkalmas módszerek különféle programcsomagokban.

	NT-SYS	SYN-TAX	PHYLIP	PAUP
Mátrix összehasonlítások	+	+		
Partíciók összehasonlítása		+		
Dendrogramok, fák összehasonlítása		+		+
Prokrusztész módszerek		+		
Konszenzus partíciók		+		
Konszenzus fák	+		+	+
Összehasonlítások szignifikanciája		+		
Különböző típusú eredmények együttes grafikus ábrázolása	+	+		

az érvényesülését vizsgáljuk. Ez lényegileg egészen más kérdés, s inkább a módszertani értékelés témájába tartozik.

Érdeemes néhány forrásra külön felhívni a figyelmet, bár egyiket-másikat már említettük ebben a fejezetben. Az osztályozások összehasonlítását, valamint a konszenzus partíciók és fák előállítását a *Journal of Classification* több cikke, különösképpen a 3(2)-es speciális szám tárgyalja részletesen, komoly matematikai apparátust véve igénybe (így csak "haladóknak" szól). Ezekon kívül Rohlf (1974, 1982) és Day (1988) ajánlható jó kiindulásnak. A fák összehasonlításának evolúciós vonatkozásait Penny et al. (1982, 1991) értékelik, de a *Systematic Biology* szinte minden számában találunk érdekes olvasnivalót. Ordinációk esetében kissé szűkösebb a választék.

Az összehasonlítások szignifikanciáját tekintve a Mantel teszt alkalmazásai járnak legelől. Manly (1991) egy teljes fejezetet szentel a témának, sokféle alkalmazási lehetőséget megemlítve. A próba tipikus felhasználási területei például a fenotípusos és genotípusos információ összevetése (Douglas & Endler 1982), genetikai és antropometriai távolságok összehasonlítása (Dietz 1983, aki a rang korreláció alkalmazását is érdemesnek tartja), pontmintázatok értékelése (Harvey et al. 1988) és faj/alapkőzet kapcsolatok vizsgálata (Burgman 1987). Dendrogramok és kladogramok összehasonlításának a szignifikanciájáról Lapointe & Legendre írt jó néhány cikket (1990, 1991, 1992), partíciók esetében pedig Podani (1986) mutat be számos példát.

9.6.1 Számítógépes programok

A legtöbb programcsomag eléggé mostohán kezeli az eredmények összehasonlító értékelését, pontosabban fogalmazva: az egész témakört mellőzi. Ez csupán részben magyarázható azzal, hogy sok esetben meglehetősen speciális módszerekre van szükségünk. A 9.2 táblázatban soroljuk fel azokat a programokat, amelyek leginkább szóba jöhetnek.

9.7 Kérdezz - Válaszolok!

K: *Először is leszögezném, hogy – a fejezetet elolvastva – minden korábbi kételkedésem elmúlt, mert bizony eleinte nem értettem: miért szentelsz ennyi oldalt ennek a témának!? A példák egy része legalábbis eléggé meggyőző volt, hogy belássam: a többváltozós elemzés folyamata nem zárul le az eredmények, az "OUC"-ok előállításával.*

V: Köszí az elismerést. Amint az irodalmi összefoglalóban is említettem, a téma sok területen sajnos teljesen vagy nagyrészt mellőzött. Ezt próbáltam például a növényökológia és vegetációtudomány területén kissé feloldani egy cikkel (Podani 1989d), amire egyébként azóta is csak 6-8 citációt tudtam “begyűjteni” (ennyiről tudok, legalábbis). A fáradozások később hozzák meg talán az eredményt. Persze az is lehet, hogy nincs igazam, s kár az erőfeszítésért...

K: *Ha jól értettem, akkor a Prokrusztész-módszer csak olyan ordinációk összehasonlítására alkalmas, amelyeknek azonos a dimenzionalitása. Mi van azonban akkor, ha engem éppen az izgat, hogy a dimenzionalitás növelésével, vagyis egyre több tengely figyelembe vételével, mennyire változik egy ordináció? Az is érdekelhet, hogy egy kétdimenziós nem-metrikus ordináció egy metrikus ordinációnak (pl. PCoA) hány dimenziós esetével egyezik meg legjobban.*

V: Valóban, a Prokrusztész módszer erre a feladatra nem vethető be, viszont használhatod a mátrix összehasonlításokat. Az ordinációkat reprezentáló $m \times m$ -es távolságmátrixokat annyi dimenziót bevonva számíthat ki, amennyire akarod, s ezután a kérdéses összehasonlításokat már könnyű elvégezni.

K: *Az eddigi tapasztalatok alapján le merném fogadni, hogy valamiféle összehasonlítási tér-sort is tudsz mondani.*

V: Már említettem is egyet, ha nem vetted volna észre. A többségi konszenzus osztályozások előállításakor a többségi feltétel 50 és 100 % között változtatható, és ezáltal egy konszenzus osztályozás-sort hoz létre. Ilyen jellegű a – még nem tárgyalt – Stinebrickner (1984) -féle módszer. Ez is egy konszenzus család, amelyben egy s -sel jelölt paraméter változtatható. Ha $s=1$, akkor a szoros konszenzust kapjuk, majd s fokozatos csökkentésével egyre több osztály jelenik meg a konszenzus fában. A 9.16 ábrán levő dendrogramok közül a **g**, vagyis az Adams egyezik meg a Stinebrickner-féle konszenzussal, $s=0,5$ mellett.

K: *Ha már volt ultrametrikus különbözőség, akkor kihasználjuk-e a négy-pont metrikát kladogramok (additív fák) összevetésében?*

V: Igen, a “*quartet metric*” valami hasonlót jelent gyökértelen fák esetében (lásd Steel & Penny 1993, és a benne idézett irodalom). Minden lehetséges objektum-négyesre (éppen $\binom{m}{4}$ -féle van) megvizsgáljuk a két összehasonlítandó fát. Azoknak az objektum-négyeseknek a száma, amelyekre nézve a két fa eltérő topológiájú, adja a fák metrikus távolságát.

K: *Úgy tűnik az eddigiekből, hogy a konszenzus fák szerkesztése elsősorban a kladisztikára jellemző. Vajon másutt miért nem alkalmazzák?*

V: A konszenzus fák akkor kellene, ha a teljes fa-struktúra érdekes a számunkra, mint például az evolúciós utakat rekonstruáló fák, azaz a kladogramok. De például egy cönológiai osztályozásban, annak ellenére, hogy teljes dendrogramokat állítunk először elő, végeredményben nem igazán érdekelnek bennünket a fa finomabb részletei. Elsősorban a dendrogramból kapható partíciókra vagyunk figyelemmel néhány kitüntetett osztályszám mellett, s inkább partíciók konszenzusa lehet érdekes, amit a 9.15 ábra illusztrált. Így van ez más tudományterületeken is, ahol nem végcél a fa előállítása, az csupán kiinduló tájékozódásra való, melynek ismeretében más típusú eredményeket optimalizálunk.

K: *Nem veszélyes-e az, hogy jóval több konszenzus módszer van, mint ahány egyesítendő kladogram? Vagyis a sok konszenzus módszer túlbonyolítja a dolgot és még több alternatív eredmény jön ki, mint amennyi eredetileg volt.*

V: Igen, ez helyes észrevétel és nehéz nem észrevenni benne egy kis gúnyt! Valóban nagyon sokféle konszenzus eredmény képzelhető el – számos konszenzus módszerről nem is szóltunk –, és ma is sokat vitatkoznak ezek relatív előnyeinek és hátrányainak. Ez olyan probléma amit nem lehet megkerülni, és én a megoldást csak abban látom, hogy majd egy olyan konszenzust találunk fel, ami nemcsak sok fát egyesít, hanem a fák minél több tulajdonságát is egyidejűleg tekintetbe veszi. (Hadd emlékeztesselek a dendrogram deskriptorokra.) Egy ilyen “többváltozós” konszenzus véleményem szerint általánosabb érvényű lehet, mint az eddig ismert konszenzus eredmények.

K: *Nem vagyok biztos abban, hogy a konszenzus eredményt mindig így kell keresni, ahogy te mondod. Ha például az alternatív fák ugyanazoknak a taxonoknak különböző változócsoporthoz tartozó elemzéseit, akkor valójában miért nem egyesítjük az adatok szintjén az információt, s ezzel megspóroljuk a fák értékelését?*

V: Biztosan olvasod a *TREE-t (Trends in Ecology & Evolution)*, melyben éppen arról zajlik egy vita, hogy miképpen kombinálhatók különféle adatok a kladisztikában (1996-os évfolyam, pl. Ballard 1996). Az egyik irányzat azt mondja, amire most utaltál: minden lehetséges adatot először egyesítsünk, s ebből állítsuk elő a “végleges” fát. A másik lehetőség az, hogy elemezzük: a különböző változócsoporthoz tartozó milyen evolúciós hipotéziseket generálnak, s ezeket próbáljuk meg “közös nevezőre” hozni a konszenzussal. Mindez statisztikai tesztekkel egészíthető ki annak kimutatására, hogy a kapott kladogramok azonosaknak tekinthetők-e (a sztochasztikus ingadozásoktól eltekintve), mert ha igen, akkor az előzetes összevonás teljesen elfogadható volt. Ha a fák nem azonosak, akkor viszont célszerű egy kicsit részletesebben megvizsgálni az okokat is.

K: *Ha a páronkénti összehasonlításoknak van szignifikanciája, akkor a konszenzus eredménynek vajon van-e?*

V: Teljesen jogos a kérdés, bár rendszerint nem pontosan ilyen formában merül fel. A kérdés inkább az, hogy az úgynevezett *bootstrap fák* (amelyek mindegyike a kiinduló változók újramintavételezésén alapszik) többségi konszenzusa milyen csoportokat tartalmaz (Felsenstein 1985). Egyes osztályok megjelenhetnek 100 %-os gyakorisággal, vagyis minden bootstrap kladogramban megváltak. Ezek a leginkább “szignifikáns” csoportok a kladisztikai elemzésben. Más csoportok pedig kisebb (de 50 %-osnál mindig nagyobb) gyakorisággal jelennek meg, azaz egyre kevésbé “szignifikánsak”. Mindezt a konszenzus kladogram egyes ágain (kládjain) kis számokkal szokták jelezni. Ilyen jellegű szekvencia elemzésre szolgáltat jó példát Krajewski & Dickerman (1990) és Cracraft & Helm-Bychowski (1991), hogy csak ketőt említsek a bőséges irodalomból. A módszer alkalmasságáról persze megoszlanak a vélemények; Hillis & Bull (1993) kétségeiket hangoztatják, mert szerintük túl szigorúak a becslések az egyes kládok szignifikanciáját illetően.