2

Az adatmátrix, az adatok átalakítása

(Az első bátortalan lépések... de még sok minden rejtve marad)

A mintavételezés során, mint láttuk, a mintavételi egységeket változók segítségével írjuk le. A kapott adatok célszerűen egy téglalap alakú táblázatba írhatók; mondjuk úgy, hogy a sorok felelnek meg a változóknak, az oszlopok pedig a mintavételi egységeknek. Erre már láttunk is példát az előző fejezetben, amikor a binarizálás módszerét illusztráltuk. A biológus egy ilyen táblázatot leggyakrabban a következő formátumban készít el:

	1. egyed	2. egyed	3. egyed
Hossz	12	14	10
Szélesség	7	9	8
Magasság	10	9	12

Ebben az egyszerű példában 3 változó jellemez 3 mintavételi egységet, egy faj három egyedét. E táblázat "letisztult" formában, cimkézés nélkül adja az *adatmátrix*ot. Könyvünkben az adatmátrix jele **X** (konvenció szerint: kövér betűvel), azaz:

$$\mathbf{X}_{n,m} = \begin{bmatrix} 12 & 14 & 10 \\ 7 & 9 & 8 \\ 10 & 9 & 12 \end{bmatrix} \tag{2.1}$$

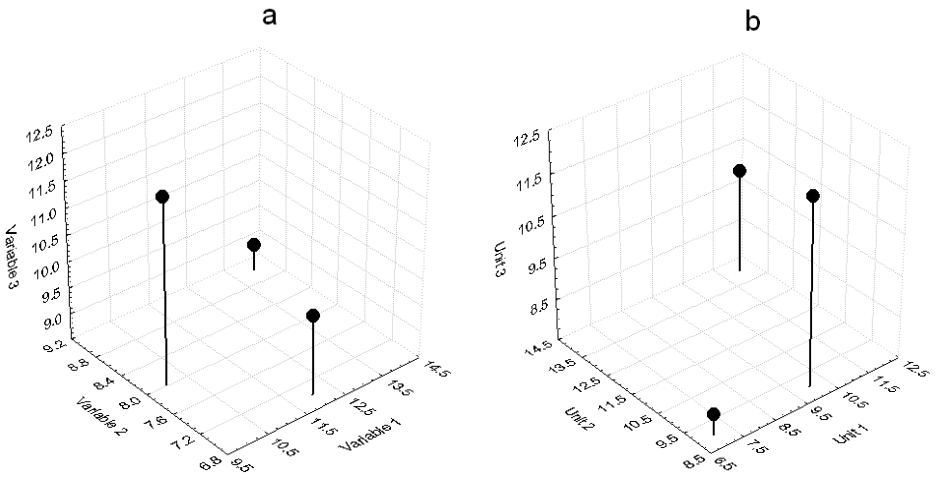
Mint látjuk, az egész mátrixot szögletes zárójelbe kell tenni, de nem nagy baj, ha a hagyományos, ívelt zárójelet alkalmazzuk. (Ugyanakkor vigyázzunk: ha a mátrixot két függőleges vonal közé írjuk, az már mást jelent, lásd a C függeléket.) A mátrix *i*-edik sorában és *j*-edik oszlopában található értéket x_{ij} jelöli. A sorok száma ezentúl *n*, az oszlopok száma pedig *m* lesz a könyv hátralévő részében. Erre utal az alsó *n,m* index. Az A függelékben megadunk több, nagyobb méretű valós és mesterséges adatokat tartalmazó mátrixot is, melyeket a módszerek illusztrálásához fogunk majd felhasználni.

Felhívjuk a figyelmét azoknak az Olvasóinknak, akik más könyvekben is utánanéznek az itt leírtaknak, hogy minden esetben tisztázzák még az elején: a változók a sorokban vagy az oszlopokban vannak-e. Ezzel elkerülhetők a képletek értelmezésekor adódó esetleges félreértések. A többváltozós elemzést elsősorban matematikai szempontok szerint tárgyaló könyvek egy része (pl. Chatfield & Collins 1980, Dillon & Goldstein 1984, Mardia et al. 1979, Reyment & Jöreskog 1993) a változókat oszlopokként szerepelteti, mások (pl. Anderson 1958, Kendall 1975) sorokként. Ez utóbbi az általános a biológiai témájú könyvekben is, hiszen a fajok ill. karakterek rendszerint a sorokban szerepelnek, pl. Pielou (1984), Orlóci (1978), Pimentel (1979), Sneath & Sokal (1973), hogy csak néhányat említsünk.

2.1 Az attribútumok dualitása és az adatmátrix geometriai jelentése

Először is tisztázzuk, hogy a továbbiakban *objektumnak* nevezzük majd az elemzés alapegységeit (vagyis amit osztályozunk, stb). Egy rendszertani vizsgálatban szereplő állategyedek általában tehát objektumként, tulajdonságaik pedig *változóként* szerepelnek. Hasonlóképpen, a növényzetben elhelyezett kvadrátok jelentik a későbbi analízis objektumait, a bennük talált fajok pedig a változóit. Ez összhangban is van az eddig elmondottakkal: a *mintavételezés egységei egyben az elemzés objektumai is, a mintavételi egységek jellemzői pedig az elemzés változói*. Ebben az esetben a mintavételi egységeket pontokként képzelhetjük el a változók mint tengelyek alkotta sokdimenziós térben: az **X** mátrix *m* számú pont *n*-dimenziós (hiper)térbeli koordinátáit tartalmazza (*n*=3 esetre lásd a 2.1a ábrát).

A kutatót persze az is érdekelheti, hogy milyen összefüggések rejlenek a tulajdonságok között, például: milyen fajcsoportok ismerhetők fel egy növénytársulásban? Ilyenkor a fenti felállás megfordul: a tulajdonságok ill. fajok most az elemzés objektumai lesznek, az egyedek ill. kvadrátok pedig változóként jönnek számításba. A mintavételi egységek voltaképpen egyszerű ismétlésként szerepelnek ahhoz, hogy a változók hasonlósági struktúráját megismerhessük. Ekkor ugyanaz az adatmátrix most úgy értelmezendő, hogy *n* számú pont *m*-dimenziós térbeli koordinátáit tartalmazza (2.1b ábra).



2.1 ábra. A 2.1 adatmátrix kétféle térbeli reprezentációja. **a**: a tengelyek a mátrix sorai, a pontok a mátrix oszlopai. **b**: a tengelyek a mátrix oszlopai, a pontok pedig a sorai.

A módszerek szempontjából – az esetek túlnyomó többségében – valójában mindegy, hogy mit tekintünk objektumnak és mit változónak. Az adatstruktúra két különböző térbeli reprezentációban vizsgálható, a változók és az objektumok felcserélhetők – mondja ki az *attribútum-dualitás* néven ismert alapelv (Williams & Dale 1965). Ennek megfelelően az ökológusok (pl. Gittins 1965) "mintatérről" ("*sample space*") beszélnek, amikor is a mintavételi egységek a tengelyek, és "fajok teréről" ("*species space*"), amelynek fajok a tengelyei. Ezzel analóg terek nevezhetők meg más tudományterületeken is (pl. "taxonómiai tér" a rendszertani vizsgálatokban).

Gyakran találkozhatunk az "R-" és "Q-típusú elemzés" elnevezésekkel, amely a fenti két eset megkülönböztetésére szolgál. Ez azonban csak kettővel növeli a megjegyzendő kifejezések számát, s – enyhén szólva – nem járul hozzá a tisztánlátáshoz, hanem felesleges ismételgetésekhez vezet. Jelen kötetben sehol sem használjuk ezeket a terminusokat, de felhívjuk a figyelmet azokra az esetekre, amikor az objektumok és változók felcserélhetősége kérdéses vagy el sem fogadható.

Ilyen pl. a lineáris (szorzat-momentum) korreláció (3.70 formula), amelynek valóban csak a tulajdonságoknál, a statisztikai értelemben vett változóknál van értelme, a benne szereplő átlag és variancia miatt. Cönológiai kvadrátok vagy két növényegyed lineáris korrelációjáról beszélni viszont nemigen lehet, hiszen az átlagnak és főleg a varianciának rájuk nézve nincs világos jelentése. (Formailag persze kiszámítható a korreláció bármit is hasonlítunk össze. Ekkor például 1-es "korrelációt" kapunk két kvadrát között, ha az egyikben éppen kétszer annyi van minden fajból, mint a másikban. Két növényegyed "korrelációja" is 1 lesz, ha az első minden testmérete éppen a fele a másodikénak. A korreláció tehát valamiféle arányosságbeli hasonlóság kifejezésére alkalmasnak tűnik, de ennek ellenére talán érezzük, hogy ezzel valami nem stimmel.) További fontos különbség az, hogy két változó korrelációja megvizsgálható szignifikancia teszttel is – ha a mintavételi egységek random mintából származnak, ezáltal függetlenek – két objektumnál viszont nem, hiszen a változók nyilvánvalóan nem jelentenek random "mintát" (vö. Pielou 1984:8).

Biztosan nincs értelme viszont a hasonlósági koefficienseket – attól függően, hogy milyen típusú térben dolgozunk – külön-külön elnevezni, amint ezt sok szakkönyv teszi. A számos példa egyike a Dice és a Sorensen indexek. Ezek formailag megegyeznek (3.25 képlet), az egyik fajokra alkalmazva, mint asszociációs koefficiens kapta elnevezését, a másik cönológiai mintavételi egységek összevetésére használatos. Goodall (1973a,b) még sok ilyen párhuzamosságot ismertet.

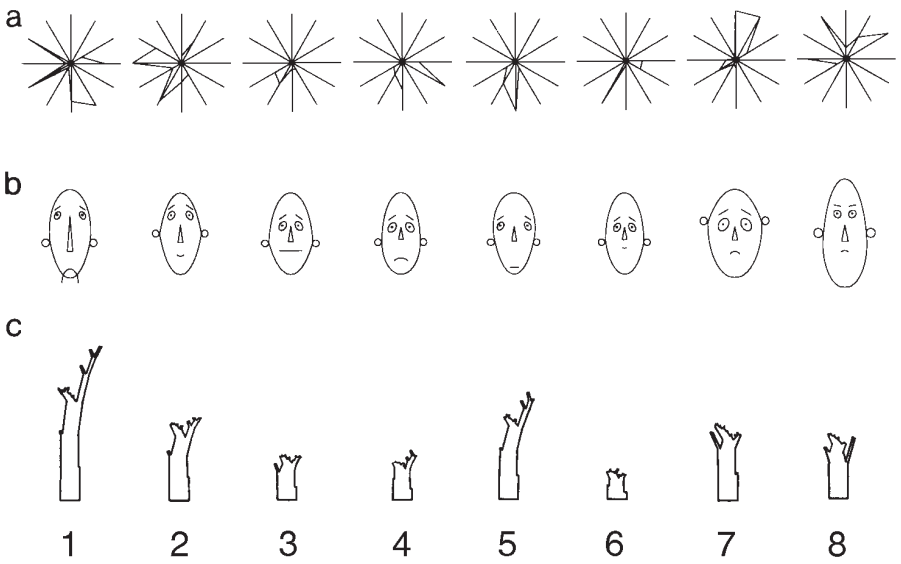
2.2 Bepillantási lehetőségek a többváltozós adatstruktúrákba

A papír síkjában csak két dimenziót tudunk feltüntetni, mégpedig a jól ismert koordinátarendszert alkalmazva. A 2.1 ábra viszont a pontok elhelyezkedését egy 3-dimenziós térben próbálja meg feltüntetni, több-kevesebb sikerrel. A pontok közötti távolságok, az adatok struktúrája itt nem érzékelhető tökéletesen, sőt, ha több pontunk lenne a diagram teljesen áttekinthetetlenné válna. Négy vagy több dimenziót pedig már semmiképpen sem tudunk ábrázolni. A könyv nagy része éppen erről szól: miként lehet egy sokdimenzionalitású térből az általunk érzékelhető kisdimenzionalitású térbe áttérni, s így "láthatóvá tenni a láthatatlant"? A bonyolult módszerek ismertetése előtt érdemes azonban néhány egyszerűbb ábrázolási lehetőséget megismerni. Előrebocsátjuk, e módszerek túl sok változóra kevéssé alkalmasak és nem oldják meg a dimenzionalitás problémáját sem.

2.2.1 Képes ábrázolások (piktogramok)

E módszerek alapelve, hogy az objektumokat kis *képekkel* helyettesítjük, melyek tulajdonságai az eredeti változóktól függenek. Ez különösen akkor lehet szemléletes, ha az eredeti objektumok absztrakt jellegűek voltak, s kevéssé érdekes – mondjuk – növény- vagy állategyedek esetében (hiszen ekkor valójában csupán az egyik – a valós – képet helyettesítenénk be egy másikkal). Önmagukban talán nem mindig alkalmasak, de jól használhatók pl. ordinációs diagramokon az egyedek azonosítására (amennyiben nincs túl sok pontunk). Megjegyzendő, hogy a változókat nem feltétlenül eredeti formájukban vesszük figyelembe, hanem terjedelmük szerint standardizálhatjuk is (2.3 formula), hogy összemérhetők legyenek.

A legegyszerűbb képes ábrázolások a *csillagdiagramok* különféle válfajai és a *Chernoff-arcok*. A csillagdiagramoknál sugárirányban elhelyezkedő vonalak felelnek meg a változóknak, ezen mérjük fel a változó standardizált értékét (ami akkor éri el az ág végét, ha éppen a mintában lévő maximumról van szó). A szemléletesség fokozására a sugarak kijelölt pontjait össze is köthetjük (2.2a ábra). Érdekesebbek talán – éppen "humán" vonatkozásuk miatt is – a Chernoff-arcok (Chernoff 1973), melyek az ember jó arcmegkülönböztető képességét próbálják kiaknázni. A karikatúraszerű rajzok tulajdonságai az eredeti változóknak felelnek meg, pl. a száj hossza az első változóval arányos, íveltsége a másodikkal, és így tovább (2.2b ábra). Az arcok megrajzolását szigorú szabályok irányítják, de az arcvonások közötti összjáték esetleg kedvezőtlenül befolyásolhatja az eredményt (pl. nagyon kicsi szájnál annak alakja már nem jól látható, stb).

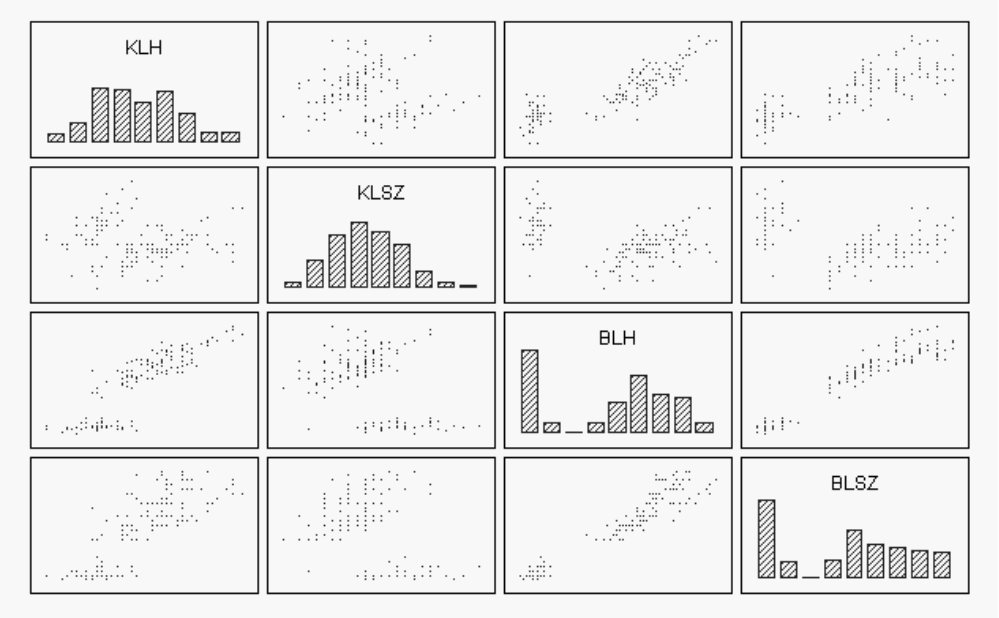


2.2 ábra. Képes ábrázolások a csillagdiagramokkal (**a**), Chernoff arcokkal (**b**) és Kleiner - Hartigan féle fákkal (**c**) az A1 táblázat oszlopaira. A **c** ábra fái a standardizálatlan borításértékek alapján készültek, a 12 változó el•zetes osztályozása a teljes lánc módszerrel készült euklideszi távolságmátrixból (l. a 3. fejezetet).

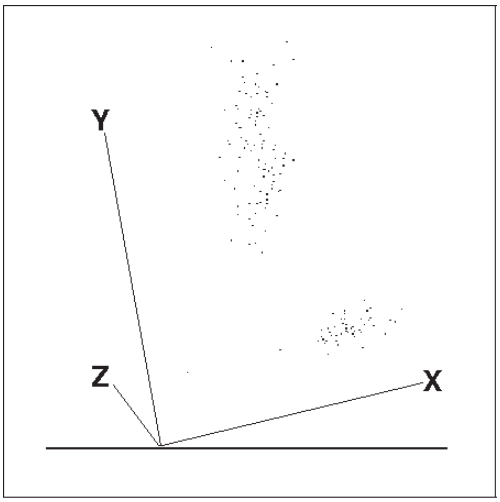
A fenti ábrázolási formák közös hiányossága, hogy a változók és a képeken látható tulajdonságok közötti megfeleltetés teljesen önkényes, ezért egy más "kiosztás" egészen eltérő összképet nyújthat. Ezt oldják meg a Kleiner - Hartigan (1981) féle *fák*. A végágak hossza arányos egy-egy tulajdonsággal, egy köztes ág hossza pedig az összes hozzátartozó végágtól függ, csakúgy mint a törzsé (2.2c ábra). A végágak és a változók közötti megfeleltetés azonban már a változók hierarchikus osztályozásából származó dendrogramból (5. fejezet) adódik (egyébként ugyanúgy önkényes lenne, mint a többi kép esetében). E módszerrel tehát valójában nem kerültük meg a többváltozós elemzést.

2.2.2 Kétváltozós szórásdiagramok mátrixa

Elemi ábrázolási lehetőség az is, amikor a sokdimenziós adatstruktúrát az összes lehetséges, két változóval definiált síkra levetítjük. Ehhez, ha n változónk van, éppen $n(n-1)/2$ koordináta-rendszerre van szükség. Egy 4-dimenziós adatstruktúra tehát 6 különböző nézettel vizsgálható meg. Az ilyen kétdimenziós szórásdiagramok kiválóan alkalmasak arra, hogy vizuálisan meggyőződjünk két-két változó összefüggéséről. Ha megengedjük a tengelyek felcserélését, akkor kétszer ennyi diagramot kapunk, amelyeket mátrix formában is elrendezhetünk (2.3 ábra). Azért nem kell n^2 diagram, mert azokat a koordináta-rendszereket, amelyekben mindkét tengely ugyanaz a változó, felesleges lenne feltüntetni. Ezek helyett a mátrix átlójában rendszerint a változók gyakorisági hisztogramját (Hartigan 1975) vagy gyakorisági poligonját



2.3 ábra. Kétváltozós szórásdiagramok mátrixa az Anderson-féle *Iris* adatokra (A2 táblázat). Rövidítések: K=külső, B=belső, L=lepel, H=hossz, SZ=szélesség. Az egyedek érzékelhetően két csoportra bonthatók, és jól láthatók az eloszlásbeli sajátságok is. KLSZ áll legközelebb a normális eloszláshoz, viszont éppen ez az a változó, melyre nézve a legelmosódottabbak a különbségek a fajok között. A többi változó hisztogramjának többé-kevésbé *bimodális* jellege a taxonok elválására utal.



2.4 ábra. Az Anderson-féle *Iris* adatok (A2 táblázat) 150 egyedének rotációs diagramja. A forgatást abban a pillanatban állítottuk le, amikor a csoportok közötti különbségek a legjobban érzékelhetők. X=külső lepel szélessége, Y=belső lepel hossza, Z=belső lepel szélessége. A vízszintes vonal a forgástengely.

(Tukey & Tukey 1981a) szokták elhelyezni, ahogy azt sok programcsomag is teszi. A gyakorisági eloszlást érdemes legalább ránézésre megvizsgálni, különösen akkor, ha a normális eloszlás alapfeltétele az elemzésnek. A terjedelemmel rendszerint itt is standardizálunk (mint ahogy a 2.2a,b ábra diagramjain is).

2.2.3 Rotációs diagramok

A rotációs diagram nagyon szemléletes, a számítógép aktív közreműködését igénylő módszer három-dimenziós ponteloszlás szemléltetésére a képernyő síkjában (Tukey et al. 1976). A koordinátarendszer a pontokkal együtt egy vízszintes tengely körül forog, s jó felbontású képernyőn a három dimenzió illúzióját kelti. Néhány forgás után már érzékelhetjük a pontfelhő alakját. A tengelyeknek a forgástengellyel alkotott szöge is változtatható, s ilymódon olyan síkokat kereshetünk a háromdimenziós térben, melyek legjobban láttatják az adatfelhő bizonyos tulajdonságait, pl. pontok csoportosulásait, lineáris trendeket stb. (2.4 ábra).

2.3 Az adatok átalakítása

A változókat – mint az előző fejezetben láttuk – sokszor más és más mértékegységben fejezzük ki (összemérhetőség hiánya), de a nagyságrendbeli eltérések is jelentősek lehetnek (belső súlyozás). Ezért a többváltozós adatokat gyakran nem az eredeti, a mintavételezésből származó formájukban elemezzük. Ha nem alakítjuk át az adatokat, akkor a nagy különbségek miatt az egyes változók nagyon különböző mértékben járulhatnak hozzá a végeredményhez, ami – hacsak valami oknál fogva éppen ezt akarjuk – mindenképpen kiküszöbölendő. Sőt, ökológiai adatok feldolgozásában még az objektumok közötti nagyságrendi különbségek eltüntetése is kívánatos lehet! Adatok átalakításának másik fontos indoka a változók eloszlásának módosítása (elsősorban a normalitás elérése), hogy az eloszlás milyenségére érzékenyebb módszerek is végrehajthatók legyenek.

Megjegyzendő: most változókról ill. objektumokról a hagyományos statisztikai értelemben beszélünk (azaz objektum = mintavételi egység). Ez azért fontos, mert – mint rövidesen látjuk – bizonyos adatátalakításoknak voltaképpen csak változók esetében van értelme: az

attribútum-dualitás érvényessége korlátozott. Az adatátalakítási eljárásokat tehát külön-külön soroljuk fel változókra és objektumokra.

Az adatátalakítás két alaptípusát különböztetjük meg: a *standardizálást* és a *transzformációt*. (Persze, most rögtön megjegyezheti az Olvasó: *transzformáció = átalakítás.* Annyi szabadságunk azonban van, hogy az idegen eredetű kifejezéssel egy kicsit speciálisabb dologra utaljunk, mint annak magyar megfelelőjével.) Standardizálás során az átalakítás az *adatokból számított* valamilyen statisztika figyelembevételével történik, az eljárás tehát adat-függő. Ilyen statisztika például a variancia, a terjedelem, az átlag, vagy egyszerűen a maximális érték. A standardizálás elsősorban a súlyozásbeli eltérések feloldására alkalmas. Transzformáció során viszont a függvény és annak paraméterei nem az adatokból számított statisztikákra alapoznak. Ezek például a változók eloszlásának a normálishoz való közelítésére jók.

Az eredeti x_{ij} érték átalakításával kapott új értéket x'_{ij} jelöli a továbbiakban. A változók súlyozását befolyásoló módszereket a 2.5a ábra koordináta-rendszerébe helyezett egyszerű fenyőfával szemléltetjük. A fa alakját két változó írja le: objektumok, azaz a fa kerületén jellegzetes helyeken kiválasztott mérőpontok (= "landmark", vö. Bookstein et al. 1985) vízszintes ill. függőleges koordinátája. (Állatok és növények alakjának ilyen típusú leírása általános gyakorlat a numerikus taxonómián belül, a morfometria szakterületén.) A fenyőfa alakjának változása illusztrálja a *súlyozásbeli* különbségeket. A változók *eloszlásának átalakítására alkalmas* eljárásokat viszont az eredeti és a módosított gyakorisági eloszlások hisztogramjai szemléltetik majd (2.7 ábra).

A fenyőfát leíró nyers adatok, a mérőpontok koordinátái az alábbi táblázatban foglalhatók össze:

2.65 3.35 0.00 2.70 3.30 6.00 1.00 2.75 3.25 5.00 1.75 2.80 3.20 4.25 2.25 2.85 3.15 3.75 3.00

0.00 0.00 2.00 2.25 2.25 2.00 3.80 4.00 4.00 3.80 5.25 5.40 5.40 5.25 6.75 7.00 7.00 6.75 8.00

A következő fejezetben felsorolt hasonlósági együtthatók jelentős része eleve tartalmaz bizonyos adatátalakítást (pl. korreláció, húrtávolság). Ha tehát az elemzés során majd ilyen függvényt alkalmazunk, akkor adataink előzetes standardizálására természetesen nincs szükség.

2.3.1 Változók standardizálása

Centrálás. A legegyszerűbb standardizálási módszer: az eredeti értékekből kivonjuk az adott változó átlagértékét:

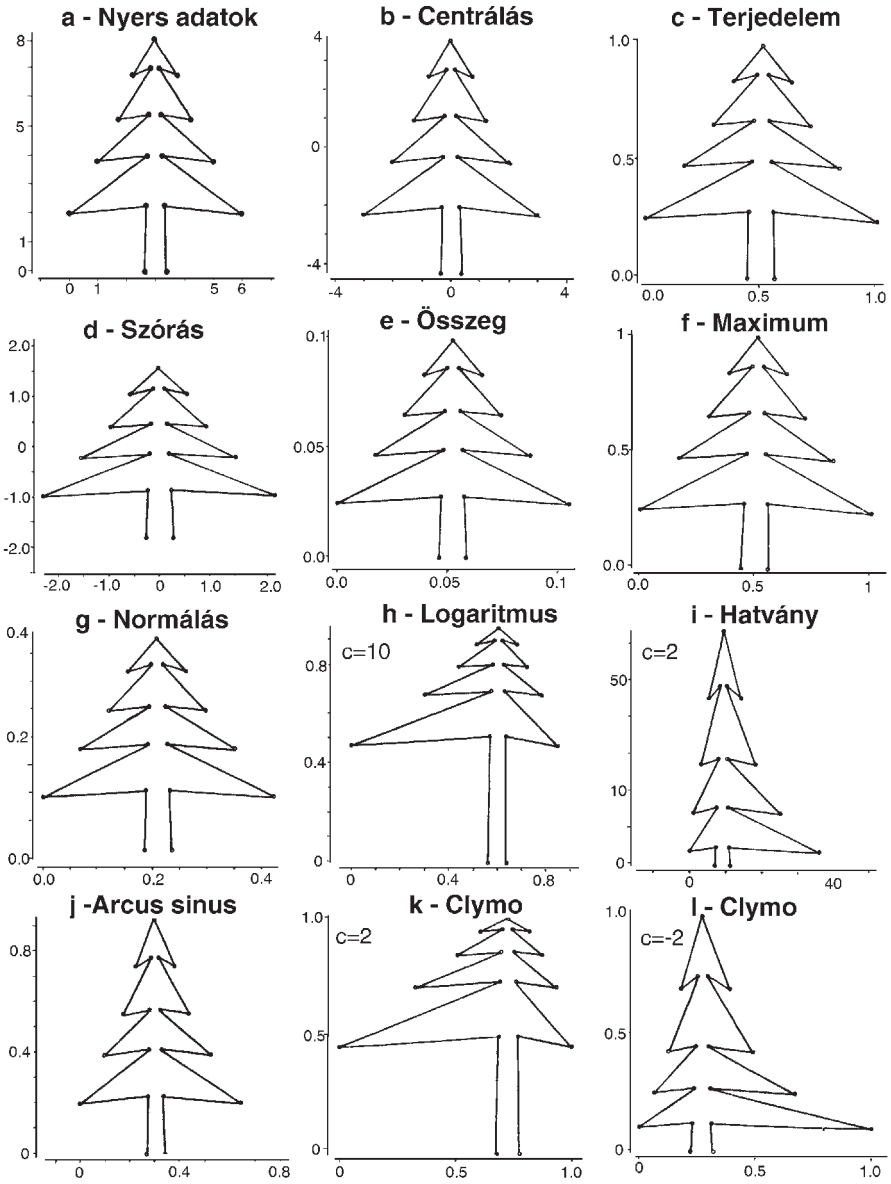
$$x'_{ij} = x_{ij} - \bar{x}_i \qquad (2.2)$$

Valójában a fenyőfa alakjával semmi sem történik, csupán a tengelyek csúsznak el úgy, hogy az origó a fenyőfa súlypontjába kerül (2.5b ábra). A centrálás önmagában ritkán használatos, viszont jelen van más standardizálási eljárásokban ill. függvényekben. A centrálás része a kovariancia- vagy korrelációszámításnak (a főkomponens- és a kanonikus korrelációelemzésben, lásd a 7. fejezetet).

Lineáris standardizálás. Ennek során az *i* változó értékeit a változóra vonatkozó összes megfigyelés alapján nyert valamely konstans értékkel szorozzuk. Ez, a fenyőfa példáján, azt jelenti, hogy a szimmetriaviszonyok érintetlenül maradnak, az alak nem torzul el, csak

valamelyik irányban megnyúlik v. összezsugorodik. Ez a változás fordított arányban van a változó éppen alkalmazott statisztikai jellemzőjével (terjedelem, szórás, stb.).

Az első két eljárást nem befolyásolja, ha a változó összes értékéhez egy konstanst adunk (azaz standardizálás előtt a fenyőfát eltoljuk mondjuk 3 egységgel jobbra). Ez azt jelenti, hogy intervallum és arányskálán mért változókra egyaránt alkalmazhatók (hiszen nem függenek a



2.5 ábra. Különböző adatátalakítási módszerek hatásának szemléltetése. A fenyőfa megváltozása elsősorban a súlyozásbeli változásokat szemlélteti (Podani 1994). A mér•pontok csak az **a** ábrán látszanak.

0 pont helyétől). A többi módszernél azonban a konstans hozzáadása már megváltoztatja a standardizálás mértékét, így intervallum-skála esetén már nem alkalmazhatók.

– *Standardizálás a terjedelemmel.* Ennek során a változó értékei a [0,1] intervallumba kerülnek:

$$x'_{ij} = [\, x_{ij} - \min_j \{ x_{ij} \} \,] / [\, \max_j \{ x_{ij} \} - \min_j \{ x_{ij} \} \,] \qquad (2.3)$$

azaz a minimumot és maximumot, valamint ezek különbségét kell meghatároznunk minden egyes változóra. A terjedelemmel való standardizálás elsősorban a belső súlyozás kiegyenlítésére alkalmas, de természetesen az össze nem mérhető változók is azonos skálára alakíthatók vele.

A fenyőfa alakja a standardizálás hatására némiképp megváltozik, mert a két változó terjedelme eltérő volt (6 ill. 8). Az *x* változó irányában ható növekedés a fa kiterebélyesedését okozza (2.5c ábra). Ez a standardizálási művelet a kevert típusú adatokra kidolgozott 3.103 és 3.104 függvényekben már megvan.

– *Standardizálás a szórással.* Ennek hatására a változók szórása 1, átlaga pedig 0 lesz:

$$x'_{ij} = \{ x_{ij} - \bar{x}_i \} / \; s_i \qquad (2.4)$$

ahol

$$s_i = \left[\frac{\sum_{j=1}^{m} (x_{ij} - \bar{x}_i)^2}{m-1} \right]^{1/2} \qquad (2.5)$$

az *i* változó empirikus (mintából számított) szórása. A számlálóban az eltérésnégyzet-összeg, a nevezőben a szabadsági fok szerepel. Ezt az eljárást elsősorban akkor ajánljuk, amikor az eredeti változókat egészen eltérő mértékegységekben fejezzük ki (pl. pH, koncentráció, hőmérséklet stb., ugyanabban mintában). Standardizálás hatására az új mértékegység az egységnyi szórás lesz, s ezután minden változó összemérhető lesz egymással. A korreláció (3.70 egyenlet) ezt a standardizálást eleve tartalmazza.

Miután a fenyőfát leíró *x* és *y* változók között *y* javára a szórást tekintve még nagyobb a különbség, mint a terjedelemben, a fa még lapítottabb lesz (2.5d ábra).

– *Standardizálás az összeggel.* Minden egyes értéket elosztunk a változóra vonatkozó összeggel:

$$x'_{ij} = x_{ij} / \sum_{j=1}^{m} x_{ij} \qquad (2.6)$$

Ilymódon a nagy értékekkel jellemzett változókat lefelé, a kis értékekkel rendelkezőket felfelé súlyozzuk. Csak akkor logikus a használata, ha az összegnek értelme van, mint a cönológiai kvadrátok esetén, amikor az összeg pl. az *i* faj összes egyedszámát jelenti a mintában. Az egyedszámban mutatkozó nagy abszolút különbségek ezáltal lecsökkennek.

Bár a fenyőfa esetében ilyen standardizálásnak nincs igazán értelme, a szemléltetés kedvéért mégis bemutatjuk (2.5e ábra). Mint látható, az eredetileg nagyobb értékekkel jellemzett *y* változó új értékei kisebbek lettek, mint az *x*-é, s a fa alakja nagyon hasonló a 2.5c fához.

– *Standardizálás a maximummal.* Minden értéket elosztunk a megfelelő változó mintabeli maximumával:

$$x'_{ij} = x_{ij} / \max_j \{ x_{ij} \} \qquad (2.7)$$

Ha a mintában szereplő értékek minimuma 0, akkor ez a módszer és a terjedelemmel való standardizálás azonos eredményt ad, mint az a 2.5c és 2.5f ábrák összehasonlításából is látszik.

– *Standardizálás egységnyi vektorhosszra (normálás*[1]*).* A változóknak megfelelő tengelyekkel jellemzett térben az origóból vektorokat irányíthatunk az objektumokat képviselő pontok felé. E vektorok hosszúságához a változók különböző mértékben járulnak hozzá. Ezt a hozzájárulást teljes mértékben kiegyenlíti a következő standardizálás:

$$x'_{ij} = x_{ij} / \left[\sum_{j=1}^{m} x_{ij}^2 \right]^{1/2} \qquad (2.8)$$

Ennek hatására az egyes változók értékeinek négyzetösszege 1 lesz. (Vagyis, az objektumok mint tengelyek alkotta térben a változókhoz mint pontokhoz mutató vektorok hossza egységnyi). A 2.5g ábra tanúsága szerint e módszer a változók hatását kiegyenlítő többi eljáráshoz hasonló eredményt ad.

> További, ritkán alkalmazott standardizálási lehetőségek: 1. minden érték osztása a változó *terjedelmével* (2.3 képlet, de a számlálóban nem szerepel a minimum kivonása), 2. osztás a változó eltérésnégyzet-összegének négyzetgyökével, 3. osztás a változó *összegének a négyzetgyökével* (azaz a 2.6 egyenlet, de a nevező négyzetgyök alatt), és 4. osztás a *szórással* (azaz a 2.4 egyenlet, az átlag kivonása nélkül).

2.3.2 Transzformáció

Mint már említettük, transzformáción olyan átalakítást értünk, amely nem az adatokból számított statisztikán alapul. Teljesen önkényesen magunk adjuk meg a transzformáló függvény kitevőjét vagy valamilyen paraméterét. Néhány módszert az előző részben alkalmazott fenyőfa példával illusztrálunk, és így lehetővé válik a standardizálással való összehasonlítás is.

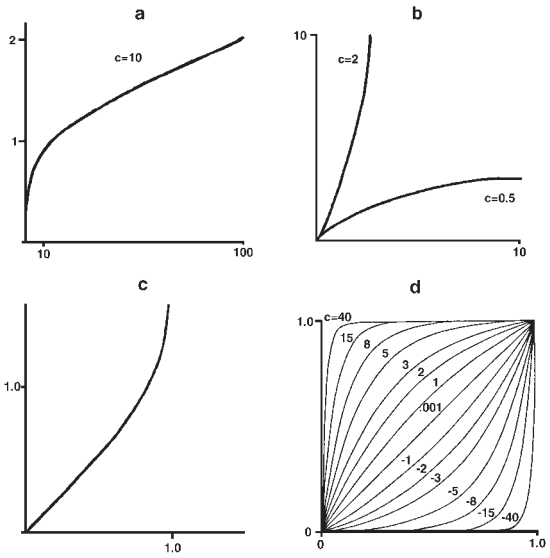
Lineáris transzformáció. Ez a többváltozós elemzés legtöbb módszerére csak elvi lehetőség. Az eredményeket ugyanis az összes értékre egyöntetűen alkalmazott lineáris transzformációk (pl. szorzás egy konstanssal) általában nem változtatják meg. Ha viszont a szorzást egyes változókra korlátozzuk, akkor valójában külső súlyozást hajtunk végre.

Nemlineáris transzformáció. E módszerek – a fentiekkel ellentétben – "eltorzítják" az adatstruktúrát, amint az a fenyőfa szimmetriaviszonyainak a megváltozásában is látható lesz. A "torzítás" persze sok szempontból hasznos jelenség lehet, amint azt az egyes függvények ismertetésénél is látni fogjuk.

– *Logaritmikus transzformáció.* Az összes értéket annak logaritmusával helyettesítjük:

$$x'_{ij} = \log_c x_{ij} \qquad (2.9)$$

1 A *normálás* nem tévesztendő össze a *normalizálással,* ami a változó eloszlásának normálishoz való közelítését jelentő transzformáció.



2.6 ábra. Adatok transzformációja. **a**: logaritmikus transzformáció, **b**: hatványozás, **c:** arc sin transzfor-máció, **d**: Clymo transzformáció. x-tengely: nyers adat, y-tengely: transzformált adat.

ahol *c* a logaritmus alapja (rendszerint *e* – a természetes logaritmus esetén –, vagy 10). Ez a transzformáció nagyságrendbeli különbségek eltüntetésére alkalmas, és jól alkalmazható egyedszám-adatok átalakítására, ha az abszolút mennyiségi különbségek helyett a nagyságrendbeli különbségeket tartjuk fontosnak. 10-es alapú logaritmus esetében például az 1 és 10 közötti különbség ugyanakkora lesz, mint a 10 és 100 közötti (2.6a ábra). Más jellegű, bármilyen arányskálán mért változónál is értelmes lehet ez az átalakítás, ha a változó eloszlása erősen jobbra ferdül (azaz jobbra elnyújtott, 2.7a ábra). A transzformáció eredményeképpen az eloszlás közelítően szimmetrikussá tehető, s ekkor már közelebb állunk a sok módszer által "megkövetelt" normalitási feltételhez (2.7b ábra).

A logaritmikus transzformáció szerves része az alak elemzését célzó többváltozós allometriának (lásd később). Egyes vélemények ugyanakkor azt sugallják, hogy a logaritmikus transzformáció nem minden esetben előnyös (Reyment 1971, 1991), s megnehezítheti az eredmények interpretálását.

A logaritmusfüggvény csak pozitív értékekre számítható ki, s mivel a 0 értékek igen gyakoriak a biológiai adattáblázatokban, a fenti formula a következővel helyettesíthető:

$$x'_{ij} = \log_c (x_{ij}+1) \qquad (2.10)$$

A 2.5h ábra jól illusztrálja a logaritmikus transzformáció hatását: kis értékkel kódolt részek (a baloldali ágak és a törzs) nagyobb súlyt kapnak, a nagyobb értékűek fontossága pedig csökken.

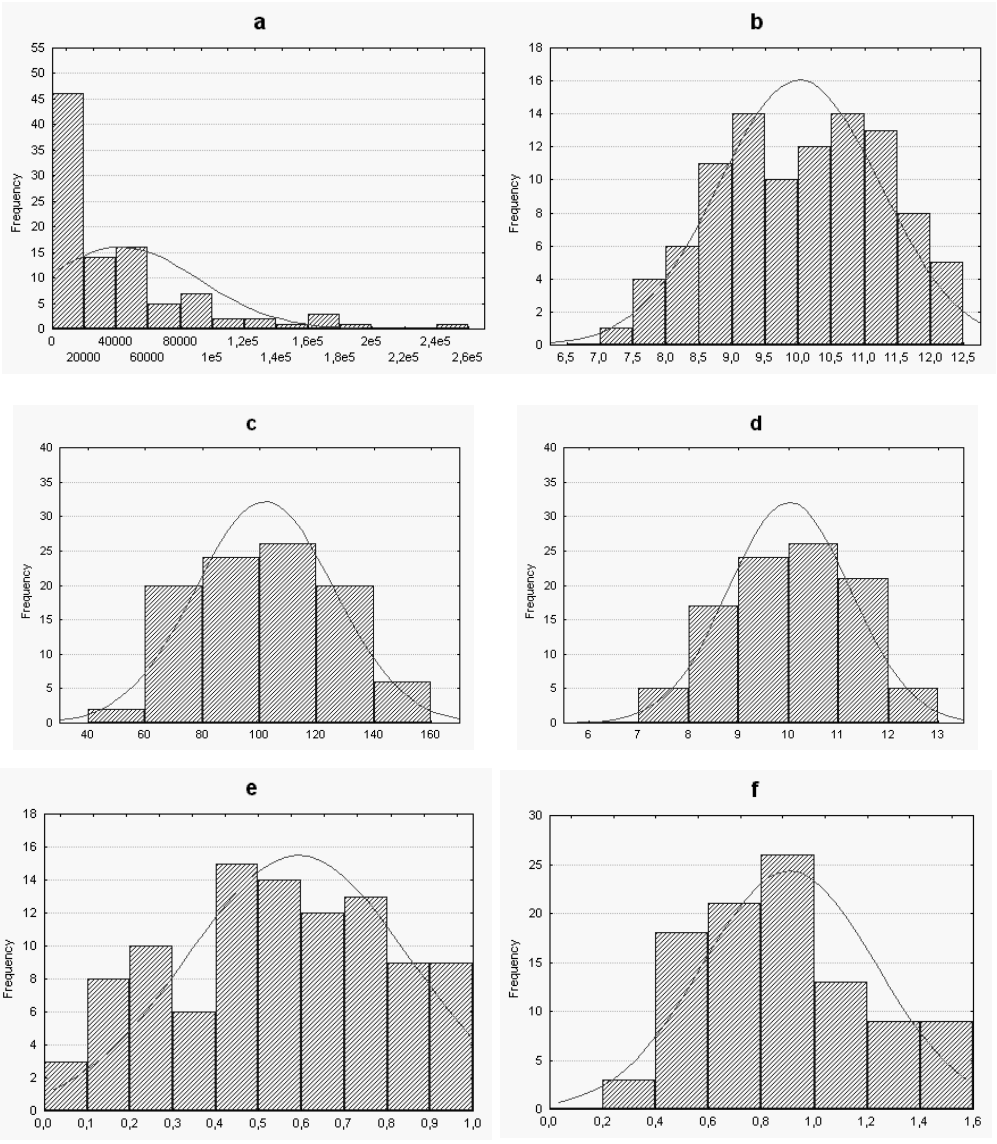
– *Hatványozás.* Az eredeti értékeket az alábbi hatványfüggvény segítségével alakítjuk át:

$$x'_{ij} = x_{ij}^{c} \qquad (2.11)$$

Az eredmény erősen függ *c* értékének a megválasztásától (2.6b ábra). Ha $c>1$, akkor a nagy értékeket még inkább fontosnak tekintjük, erre azonban igen ritkán lehet szükség (2.5i ábra). Sokkal fontosabbak a $c<1$ feltétel melletti transzformációk, elsősorban a $c=0.5$ (azaz a *négyzet-*

gyök transzformáció). Az átalakítás eredményeképpen a nagy értékek túlsúlya csökken. Poisson eloszlású egyedszámadatok esetén a négyzetgyök transzformációval jól közelíthető a normális eloszlás (2.7c-d ábra), bár a transzformáció hagyományos alkalmazási területe a varianciák stabilizálása. A hatványozás c=–1 esetén a *reciprok értéknek* felel meg.

A fenti transzformációk egy függvénycsaládba egyesíthetők Box & Cox (1964) javaslata szerint:



2.7 ábra. Transzformációk hatása változók eloszlására. **a-b**: logaritmikus transzformáció erősen jobbra ferde eloszlásból, **c-d**: négyzetgyök transzformáció, **e-f**: arc sin - négyzetgyök transzformáció relatív gyakorisági adatokból. A folytonos vonal az adatokra illesztett normális eloszlásnak felel meg.

$$x'_{ij} = (x_{ij} - 1)^{\lambda}, \text{ ha } \lambda \neq 0; \tag{2.12a}$$

$$x'_{ij} = \ln x_{ij}, \text{ ha } \lambda = 0. \tag{2.12b}$$

Amikor λ=1 egy egyszerű elcsúsztatásról van szó. Ez semmi lényeges következménnyel nem jár. Ha λ=0,5, a négyzetgyök transzformációt kapjuk, λ=0 pedig megfelel a logaritmikus transzformációnak. A függvénycsalád arra használható, hogy λ szisztematikus változtatásával megállapíthassuk a normális eloszlásra adott legjobb illeszkedést, az alábbi ún. *log likelihood* becslőfüggvény alapján (Sokal & Rohlf 1981a):

$$L_i = -\frac{\nu}{2} \ln s_T^2 + (\lambda - 1) \frac{\nu}{m} \sum_j \ln x_{ij} \tag{2.13}$$

ahol s^2_T a transzformált adatok varianciája, ν a szabadsági fokok száma, *m* a mintanagyság. Azt a λ-t, melyre nézve a fenti összefüggés maximumot ad, lesz célszerű alkalmazni a transzformációban. Az eljárás, relatíve nagy számítási igénye és a többváltozós módszerek viszonylagos robusztussága miatt, inkább az egyváltozós statisztikában használatos.

Mivel a 2.11 függvény $x_{ij} = 0$ és c=0,5 esetén nem értelmezhető, helyette a következő formulát alkalmazhatjuk:

$$x'_{ij} = \sqrt{x_{ij} + 0.5} \tag{2.14}$$

– *Arcus sinus transzformáció.* Ez a függvény 0 és 1 közé eső értékek átalakítására alkalma (2.15)

s: $x'_{ij} = \arcsin x_{ij}$ de nem ebben a formában használjuk, hanem a négyzetgyökkel kombinálva (következő oldal). A teljesség kedvéért azonban bemutatjuk a transzformáció hatását (2.5j és 2.6c ábra)

– *Clymo-féle transzformáció.* Ez a függvény feltételezi, hogy az adatok arányokat fejeznek ki, és 0-tól 1-ig terjednek. (Ha adataink nem ilyenek, akkor az összeggel standardizálunk először a 2.6 egyenlet alapján). A függvény alakja a következő:

$$x'_{ij} = (1 - e^{-cx_{ij}}) / (1 - e^{-c}) \tag{2.16}$$

(van der Maarel 1979). A függvény segítségével egy transzformációsorozat állítható elő, pl. cönológiai adatsorok vizsgálatára. A *c* paraméter változtatásának hatását a 2.6d ábrán láthatjuk. Nagy *c* értékekre a prezencia/abszencia típust közelítjük a transzformációval. 0-hoz közeli *c* értékeknek gyakorlatilag nincs befolyásuk az adatokra. (A *c*=0 esetre a függvény nincs értelmezve.) Növekvő negatív *c* értékekre pedig a nagy számok túlhangsúlyozása és a kicsik negligálása érhető el. Mindez a megfelelően módosított fenyőfapéldán is jól látható (2.5k-l ábra).

A többváltozós elemzésben ritkán alkalmazott további transzformációk az *exponenciális* függvény ($x'_{ij} = e^{x_{ij}}$) és az *arcus cosinus* függvény ($x'_{ij} = \arccos x_{ij}$).

Binarizálás. Intervallum- vagy arányskálán mért változókat gyakran át kell alakítanunk bináris (prezencia/abszencia) adatokká (pl. ha mindenképpen ki akarunk próbálni egy ilyen adattípust igénylő módszert). Ekkor

$$x'_{ij} = 1, \text{ ha } x_{ij} > p; \tag{2.17a}$$

$$x'_{ij} = 0, \text{ ha } x_{ij} \leq p \tag{2.17b}$$

ahol *p* a binarizálás küszöbértéke, amelyet többnyire 0-nak választunk (minden pozitív érték "jelenlét"-nek számít).

Összetett transzformációk. A fentiekben ún. elemi transzformációs függvényeket mutattunk be. Vannak esetek, amikor két vagy több függvényt kombinálunk a transzformáció során, s így érjük el a kívánt eredményt.

– *Alaktranszformáció*. Ha adataink valamilyen alak körvonalait írják le[2] (többváltozós allometria), akkor főkomponens vagy kanonikus korreláció elemzés előtt Darroch & Mosimann (1985) javaslatára a következő kombinált transzformációt célszerű elvégezni. Először az adatokat logaritmikus transzformációnak vetjük alá, majd standardizáljuk az új átlagértékek kivonásával: azaz először a 2.9, majd a transzformált adatokra a 2.2 függvényt alkalmazzuk. (Megjegyzendő, hogy a centrálás "benne van" a fent említett elemzésekben, így voltaképpen az elemzést megelőzően elegendő a logaritmikus transzformációt végrehajtani.)

– *Arcus sinus - négyzetgyök transzformáció arányokra.* Csak *relatív gyakoriságokra* alkalmazható, amikor az adatok pl. arányokat fejeznek ki a [0,1] intervallumban. Először az összes érték négyzetgyökét vesszük, majd végrehajtjuk a 2.15 transzformációt. A módszer a többváltozós elemzésben legfeljebb a normális eloszlás közelítésére jöhet számításba. A transzformáció hatása kevéssé olyan erőteljes, mint a logaritmikus tanszformációé (2.7e-f ábra).

2.3.3 Objektumok standardizálása

Változók átalakítása általánosan elterjedt, rutinszerű művelet, az objektumok szerinti standardizálásra viszont elsősorban az ökológiában kerülhet sor (bár ennek igénye a taxonómiában is felmerülhet, vö. Sneath & Sokal 1973:156). Ennek célja például az lehet, hogy a mintavételi egységek közötti borításbeli különbségeket csökkentsük. Azaz, egy kvadrát amelyben sok faj, de viszonylag kis mennyiségben van jelen, olyan fontos legyen, mint amelyben ugyanannyi faj sok egyeddel van képviselve.

A standardizálás hatását három objektummal, cönológiai "kvadráttal" illusztráljuk, amelyekben négy faj található. Ezek borítása – a szemléletesség kedvéért – a magasságukkal lesz arányos a 2.8 ábrán. A nyers adatmátrix a következő:

1,0 0,5 5,0
5,0 2,5 3,0
3,0 1,5 1,5
1,0 0,5 0,75

Az objektumok standardizálásának geometriai értelmezését próbálja elősegíteni a 2.9 ábra is. A tengelyek két változónak felelnek meg, a pontok pedig négy objektumot képviselnek. Az adatokat nem adjuk meg, a koordináták leolvashatók az ábráról.

– *Centrálás.* Az objektum átlagértékét vonjuk ki az összes adatból:

$$x'_{ij} = x_{ij} - \bar{x}_j \qquad (2.18)$$

Mivel itt negatív értékeket is kapunk, az eredményt nem mutatjuk be a 2.8 ábrán. Jól illusztrálható viszont a centrálás hatása két dimenziónál (2.9a ábra): az összes pont egy átlószerű

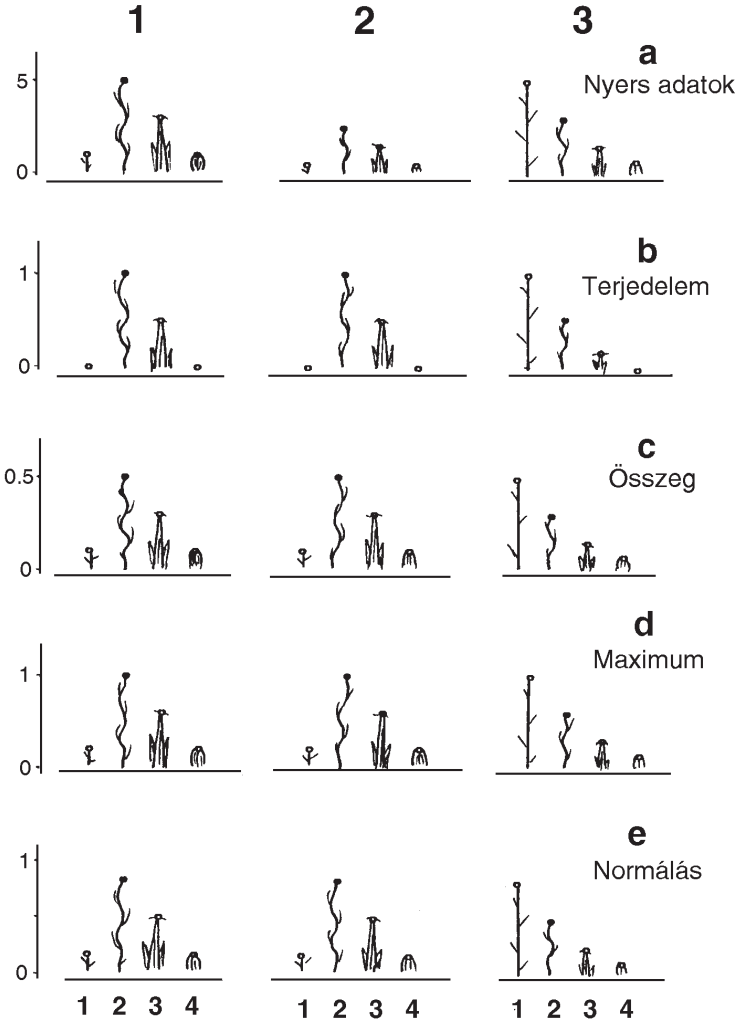
2 A 7.6 alfejezetben bemutatott módszerek ilyen standardizálást nem tesznek szükségessé.

egyenesre kerül. Három dimenziónál egy síkra, még több dimenzió esetén hipersíkra vetül minden pont. A centrálás műveletével voltaképpen *egy dimenzió kiesik,* az "átlóra" merőleges irányú nagyságrendi hatás eltűnik.

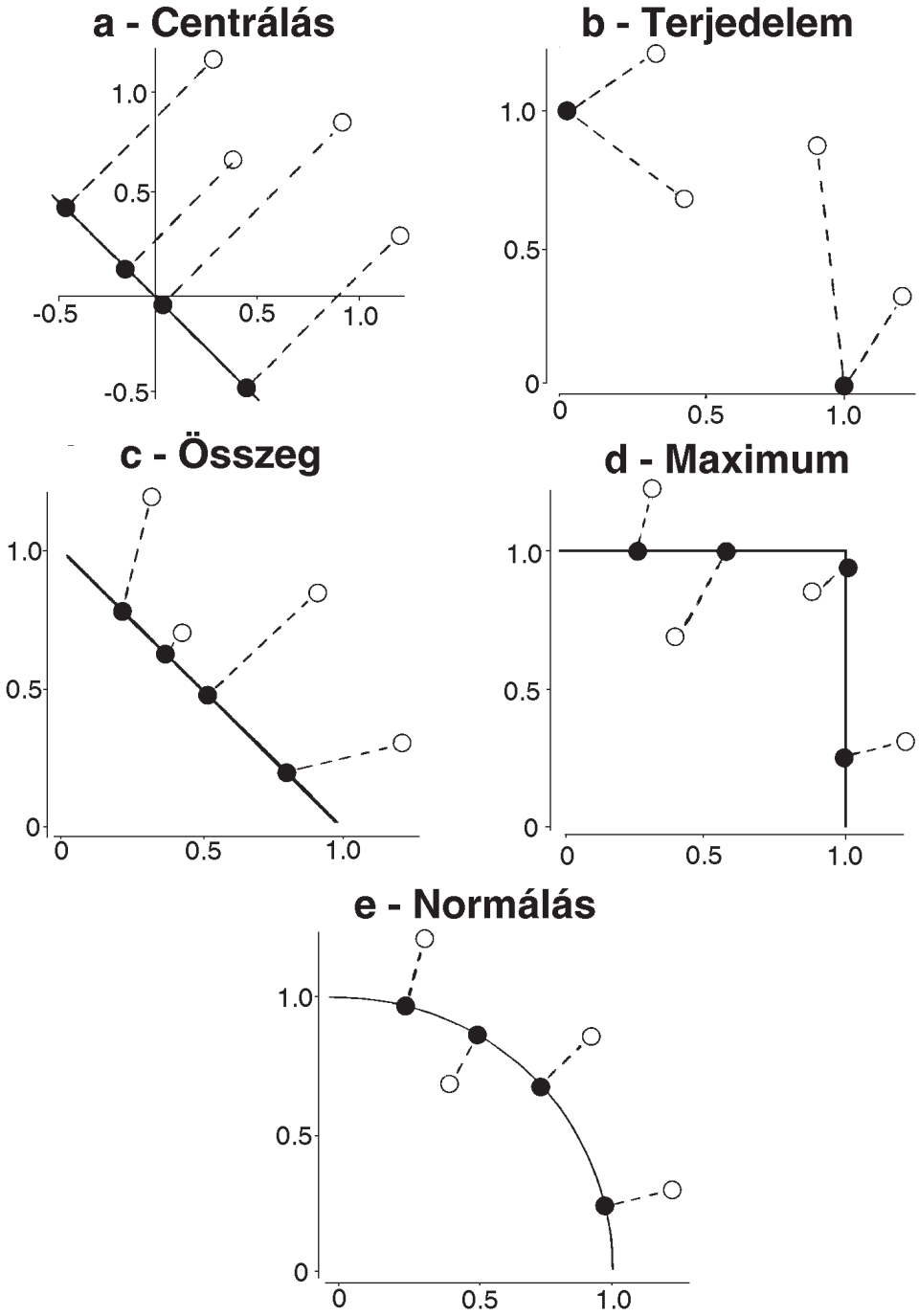
– *Standardizálás a terjedelemmel.* Az eredeti értékekből kivonjuk a minimumot, majd elosztjuk az objektum terjedelmével.

$$x'_{ij} = [\, x_{ij} - \min_i \{ x_{ij} \} \,] \;/\; [\, \max_i \{ x_{ij} \} - \min_i \{ x_{ij} \} \,] \qquad (2.19)$$

A standardizálás eredményeképpen minden objektumban 0 és 1 közé kerülnek az értékek (2.8b ábra). A minimális egyedszámú (vagy borítású) fajok (1 és 4) azonban a standardizálás hatására el is "tűnnek", s ez nem feltétlenül kívánatos. Két dimenzió esetén az új értékek vagy 0-val vagy 1-gyel lesznek egyenlőek, így minden pont két új pozícióba "csúszik össze" (2.9b ábra). Több dimenziónál ez természetesen már nem így lesz: a pontok az egységnyi oldalú hiperkocka felületére kerülnek.



2.8 ábra. Standardizálás objektumok szerint. A növények magassága arányos a fajok borításával (Podani 1994).



2.9 ábra. Objektumok standardizálásának hatása két változó esetén. Üres körök: eredeti objektumok, telt körök: standardizált objektumok.

– *Standardizálás az összeggel.* Az objektumhoz tartozó összeggel osztunk minden értéket:

$$x'_{ij} = x_{ij} / \sum_{i=1}^{n} x_{ij} \qquad (2.20)$$

Ilymódon az új értékek összege 1 lesz, és az adatok az objektumbeli arányokat fogják tükrözni (2.8c ábra). Két dimenzióban a pontok az egységsugarú kör húrjára vetülnek (2.9c ábra), három dimenzióban egy egyenlő oldalú háromszögre, sok dimenzióban egy "hipersíkra".

– *Standardizálás a maximummal.* Az objektumhoz tartozó adatok maximumával osztunk minden egyes értéket:

$$x'_{ij} = x_{ij} / \max_i \{ x_{ij} \} \qquad (2.21)$$

A módszer csak akkor tér el a terjedelemmel történő standardizálástól, ha minden változónak 0-nál nagyobb az értéke az objektumban, ahogy a példában is (2.8d ábra). Valós adatok esetében azonban a minimum gyakran 0 (egyedszám, borításadatok sok fajra), így a két módszer egyező eredményt ad. Két változó esetén az objektumokat az egységnyi oldalú négyzet kerületére (2.9d ábra), több dimenzióban pedig az egységnyi oldalhosszúságú "hiperkocka" felületére vetítjük.

– *Standardizálás egységnyi vektorhosszra (normálás).* Ekkor minden értéket elosztunk az objektumra vonatkozó négyzetösszeg gyökével:

$$x'_{ij} = x_{ij} / \left[\sum_{i=1}^{n} x_{ij}^2 \right]^{1/2} \qquad (2.22)$$

A standardizálás hatását a 2.8e ábra is illusztrálja, de ez kevésbé szemléletes. A változókkal mint tengelyekkel jellemzett térben ugyanis a standardizálás azzal a következménnyel jár, hogy minden pont – amelyek tehát most objektumokat jelentenek – egységnyi távolságra lesz az origótól. Azaz, a pontok az egységsugarú hipergömb felületére kerülnek (két dimenzióban az egységsugarú körre, 2.9e ábra). A húrtávolság (3.54 egyenlet) ezt a standardizálást tartalmazza.

Kettős centrálás. Objektumok és változók egyidejű standardizálásáról van szó, a következők szerint:

$$x'_{ij} = x_{ij} - \bar{x}_i - \bar{x}_j - \bar{\bar{x}} \qquad (2.23)$$

ahol $\bar{\bar{x}}$ a főátlag, az adatmátrix összes értékére. Nyilvánvalóan ennek csak akkor van értelme, ha az összes változót ugyanazon a skálán mértük. Ha például a változók fajok borításai, akkor $\bar{\bar{x}}$ a fajok átlagos borításának felel meg. A centrálás eredményeképpen a változókat és az objektumokat egyformán ítéljük meg. Egy ritka faj, ha fajszegény kvadrátban fordult elő nagymértékben súlyozódik, a fajgazdag kvadrátokban talált gyakoribb fajok pedig kis súlyt kapnak. Az "egyedi, unikális" ill. "átlagos" viselkedés ilyen megkülönböztetése értelmes lehet az ökológus szempontjából (vö. Noy-Meir et al. 1975).

Kettős standardizálás az összeggel. Az adatmátrix minden értékét elosztjuk a megfelelő sor- és oszlopösszeggel is. Ez az eljárás a χ^2-távolságba (3.67 formula) van beépítve, és fontos szerepe van a korreszpondencia elemzésben (7.3 alfejezet).

2.4 Irodalmi áttekintés

Többváltozós adatok egyszerűsített grafikus szemléltetéséhez a legtöbb ötletet a Barnett (1981) szerkesztette kötet adja, elsősorban is a 10-12. fejezet (Tukey & Tukey 1981a,b,c). Néhány perspektivikus ábrázolást a fizikából kölcsönzött példák illusztrálnak, de pl. az Anderson (1935, 1936) -féle *Iris* adatokra is találunk olyan módszert, amelyre jelen könyvben már nem jutott hely. Barnett (1981) azonban "csupán" áttekintő munka, ne számítsunk a technikai részletek alapos ismertetésére, ebben inkább a bőséges bibliográfia segíthet. Az Olvasó figyelmébe ajánlható még Everitt & Nicholls (1975), Everitt (1978) és Wegmen et al. (1993).

Két vagy többváltozós ökológiai adatok bemutatási lehetőségeire sok példát említ Digby & Kempton (1987), bár ezek jelentős része éppen a fent említett Barnett-féle kötetből származik. Érdemes lehet még a Green (1979) által összefoglaltakat is áttekinteni, bár a közölt ábrák nem annyira az elemzést megelőző, hanem inkább az elemzést követő illusztrációs lehetőségek sokféleségét szemléltetik. Reyment (1991) is bemutat egy, még nem említett ábrázolásmódot, a háromdimenziós perspektivikus vetületre alkalmazott "drótdiagramot" ("*wireline*" diagram), bár a példák kevéssé meggyőzőek.

Az adatok átalakításáról a legtöbb szakkönyv legalábbis megemlékezik. Pl. Gordon (1981) a standardizálást a változók összemérhetőségével és súlyozásával kapcsolatosan említi meg, de mellőzi a módszerek részletes tárgyalását, s transzformációról egyáltalán nem szól. Hasonló a helyzet Dunn & Everitt (1982) könyvével is, holott a numerikus taxonómia egyik alapvető kérdése a standardizálás, mint a karakterek egyenlő súlyozásának fő lehetősége. Taxonómusoknak ezért még mindig Sneath & Sokal (1973: 153-156) összefoglalóját ajánlhatjuk elsősorban. Mayr & Ashlock (1991) erősen kritizálják és elvetik a szórással történő standardizálást mondván, hogy a kevéssé ingadozó karakterek túl nagy súlyt kapnak az elemzésben, míg a rendkívül élesen elváló karakterek fontossága csökken. Hasonlóan vélekedik Stuessy (1990) is: szerinte nem szabad minden változót egyformán figyelembe venni, ha csak egy részük variabilitása magyarázható biológiai okokkal, másoké pedig elsősorban mérési hibákból származik. Ez valóban egy megfontolásra érdemes szempont mindenki számára; bár annak eldöntése, hogy a változók varianciája honnan származik, nem könnyű feladat. Megjegyezzük, hogy ebben a szemléletben a kladisztika (6. fejezet) erőteljesen differenciáló karakter-súlyozási törekvése ismerhető fel.

A standardizálás és a transzformáció általunk alkalmazott megkülönböztetése összhangban van sok munkával, pl. Sokal & Rohlf (1981a) vagy Rohlf (1993). A matematikai statisztikában jártasabbaknak viszont feltűnhet, hogy a standardizálást itt jóval általánosabb értelemben használtuk, ugyanis a statisztikusok számára a standardizálás csak az átlag kivonását és a szórással történő osztást jelenti (vö. pl. Jánossy et al. 1966).

Az adatok átalakításának hatását vegetáció-ökológiai kontextusban Austin & Greig-Smith (1968), Noy-Meir (1973) és Noy-Meir et al. (1975) vizsgálták. Bár ezek viszonylag régebbi publikációk, a témával foglalkozó kutatók ma is haszonnal olvashatják. Az ökológiai tárgyú könyvek egy sora, pl. Digby & Kempton (1987), Jongman et al. (1987), Pielou (1984), Ludvig & Reynolds (1988) viszonylag keveset szentel e témának. Orlóci (1978) a változók standardizálását az összemérhetőség szempontjából veszi szemügyre, az objektumok standardizálását pedig úgy vizsgálja, hogy azok milyen hasonlósági ill. távolság-függvényekben (3. fejezet) szerepelnek.

2.1 táblázat. Adatstruktúrák grafikus illusztrációja és adatok átalakítása különféle programcsomagokban (B függelék). + jelöli a közvetlenül elérhető módszert, * pedig a függvény definiálásával, kissé bonyolultabban, változónként külön-külön elvégezhető átalakítást. A Kleiner-Hartigan féle fák rajzolására nem találtam programot, a 2.2c ábra kézzel készült.

	Statistica	NT-SYS	SYN-TAX	BMDP	NuCoSA
Szórásdiagramok mátrixa	+	+			
Rotációs diagram			+		
Chernoff-arcok	+				
Csillagdiagramok	+				
Hisztogramok	+			+	+
3-dimenziós persp. rajzok	+	+			
Centrálás	*	+	+	*	+
Terjedelem	*	+	+	*	
Szórás	+	+	+	*	+
Összeg	*	+	+	*	+
Maximum		+	+	*	+
Normálás		+	+	*	
Log x	*	+		*	+
Log (x+1)	*	+	+	*	+
Hatvány (általános formula)	*		+	*	+
Négyzetgyök	*	+	+	*	+
Négyzetgyök (x+0.5)	*	+			
Négyzetre emelés	*	+	+	*	+
Arc sin	*	+	+	*	
Clymo			+		+
Binarizáció		+	+	*	+
Kettős centrálás		+	+	*	+

2.4.1 Számítógépes programok

A 2.1 táblázat sorolja fel az ebben a fejezetben ismertetett módszereket és jelzi, hogy azok mely programcsomagokban találhatók meg. A programok listája természetesen nem teljes, hiszen lehetetlen lenne minden szóba jöhető programcsomagot fellelni és értékelni. Az összeállításban ezért elsősorban olyan programok szerepelnek, amelyek személyi számítógépeken futtathatók, és Magyarországon már elterjedtek, viszonylag könnyen beszerezhetők vagy megrendelhetők, és a könyvben tárgyalt más módszereket is tartalmaznak (B függelék). Reméljük, hogy ezzel is megkönnyítjük az esetleges felhasználók munkáját, bár a táblázat tartalmáért "üzleti értelemben" nem vállalhatjuk a felelősséget.

Az adatátalakítás stratégiája az egyes programcsomagokban többféle lehet. Nagy adattáblázatokra a **Statistica** és a **BMDP** használata viszonylag kényelmetlen, hiszen minden egyes változóra külön-külön kell elvégeznünk a műveleteket, rendszerint a fő elemzést megelőzően. Az **NT-SYS** pedig nagy mátrixokra is alkalmazható, megtartva azt a lehetőséget, hogy az egyes változókat különféleképpen kezeljük. A **SYN-TAX** és a **NuCoSA** viszont egyöntetűen

alkalmazzák az átalakítást minden változóra, ennek megfelelően gyors és kényelmes a használatuk.

2.5 Kérdezz - válaszolok

K: *Mire végigolvastam ezt a fejezetet, már egy kicsit meg is zavarodtam: mikor van szó mintavételi egységről, mikor változóról, mikor objektumról; mit lehet felcserélni mivel, és így tovább. Lehet, persze, hogy én vagyok a hibás, de jó lenne még egyszer tisztázni a dolgokat.*

V: Ez elől nem zárkózhatom el; én se szeretném ha homályos maradna ez a kérdés. Foglaljuk tehát össze: mintavétel során technikai értelemben beszélünk mintavételi egységekről, amelyeket az alapsokaságból *kiválasztunk*, vagy a kontinuumban *elhatárolunk*. Ezeket – statisztikai értelemben vett – változók segítségével írjuk le. Természetesen ezek még nem keverhetők össze! Az elemzés során a mintavételi egységek helyett viszont már objektumokról beszéltünk, a változókra újabb elnevezést nem kerestünk. Ettől fogva az attribútum-dualitás elve értelmében az objektumok és változók felcserélhetők lesznek (kivéve azt a néhány esetet, amikor ennek jogossága vitatható, illetve a szignifikancia próbáknál).

K: *Amikor előzetesen megvizsgálom az adataimat, könnyen találhatok olyan változókat, amelyek csak logaritmikus transzformáció után közelítik a normális eloszlást. Ugyanabban a mátrixban más változók viszont eleve normális eloszlásúnak tűnnek. Van-e annak értelme, ha bizonyos változókat átalakítok, másokat pedig nem?*

V: Ennek nincs elvi akadálya, csak jól át kell gondolnunk, mit is akarunk elérni. Adatok átalakításának, mint láttuk, kétféle célja lehet: a változók súlyozásának megváltoztatása ill. az eloszlás módosítása. A logaritmikus transzformáció egyszerre normalizál és "egalizál" is, holott meglehet: csak az egyikre lenne szükség. Bizonyos egyensúlyt kell tehát a súlyozás és normalizálás között megteremteni. A többváltozós elemzésben inkább a súlyozás megváltoztatása a fontosabb, ez szinte minden módszernél számításba jöhet. Normalizálásra ritkábban van szükség, s ez egyáltalán nem érinti pl. a klasszifikációs módszereket.

Annak, hogy más és más módon alakítjuk át a változókat, persze van egy fontos következménye: a közöttük lévő kapcsolatok (pl. korreláció) is megváltoznak! Objektumok standardizálását pedig csak a teljes objektumhalmazra egyöntetűen érdemes elvégezni.

K: *Ha jól értettem az előző fejezet alapján, a térsorelemzés a valós térben a mintavételezés paramétereinek apró megváltoztatásával próbál hasznos következtetésekre jutni. Ebben a fejezetben újabb tereket ismertünk meg, pl. a fajok mint dimenziók alkotta teret. Logikus lenne, ha itt is tudnánk térsorokat definiálni.*

V: Úgy van. A valós térbeli *sor*ok (vagy *sorozat*ok, ha így jobban tetszik) csak a kezdetet jelentik. Az adatmátrix elkészítésével és a későbbi elemzések során már elvont, konceptuális terekkel van dolgunk, és sorokat mindegyikben lehet definiálni. Gondoljunk például a Clymo függvény, a logaritmus és a hatványfüggvény *c* paraméterének, vagy a Box - Cox transzformáció λ paraméterének a fokozatos megváltoztatására.

K: *Mi lehet ennek az értelme?*

V: Ahogy a valós térbeli sorok a mintavételezés paraméterei önkényes megválasztásának hatását képesek illusztrálni, az *adattérbeli* sorok (mondjuk így) pedig az adatátalakítási

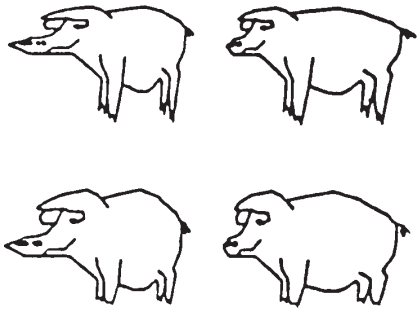
"önkényeskedések" hatását mutathatják meg. Pl. a 10-es alapú logaritmus sokkal erőteljesebben redukálja a nagy egyedszámadatokat, mint a természetes alapú, vagy pláne a 2-es alapú logaritmus. A Clymo transzformációsor, amelyet azt hiszem a 2.6 ábra elég szemléletesen illusztrál, jól használható az adattípusok fokozatos változtatására. Megjegyzendő, hogy mostanában egyre többen vizsgálnak ilyen sorokat, bár nem elegen...

K: *Ami nyilván senkit sem ment fel a lustaság vádja alól!*

V: Igen, meg kell "sajnos" szoknunk, hogy az elemzés során nagyon sok minden saját döntéseinkre van bízva. A mintavételezés, az adattípus és adatátalakítás megtervezése ránk vár. És akkor még nem is említettük a hátralévő számos választási lehetőséget, amelyekre persze kitérünk a későbbiekben. Döntéseink hatását egy kicsit komolyabban kellene vennünk, mint eddig, s ilyen irányban a térsorok sokat segíthetnek. Több konkrét példát láthatsz majd a könyv záró fejezetében.

K: *Nagyon szemléletesnek tartom a fenyőfás ábrát...*

V: Ennek örülök, de rögtön be kell vallanom, hogy az ötlet bizony nem teljesen eredeti. Egyes transzformációk kombinált hatását illusztrálta malacok alakváltoztatásával a *Münch. med. Wschr.* 124. kötete 13. számának 15. oldala. Be is mutatok neked néhányat, íme:



Ezek a rajzok azonban túl jól, túlságosan is mulatságosra sikeredtek, a lényeget a fenyőfák talán jobban láttatják. Az egyes irányokban pedig eltérő a transzformáció típusa, és ezt nem igazán ajánlom. A változók sokféle átalakítása végül is kavarodást okozhat, de erre már fentebb is utaltam, mikor a normalizálásról kérdeztél.