Podani János

Bevezetés a többváltozós biológiai adatfeltárás rejtelmeibe

avagy

“Mit is kezdjünk azzal a rengeteg adattal?”

Scientia Kiadó, Budapest
1997

ISBN 963 8326 06 9
Scientia Kiadó
pf. 658
1365 Budapest

Kinyomtattatott az 1997-es esztendőben,
a szerzőnek nyújtott OTKA P18941 könyvkiadási
támogatásnak köszönhetően.

Tartalomjegyzék

0

Bevezetés

(Miről is lesz szó, miért és hogyan?)

A biológusok számára örömök és nehézségek forrását jelentő tény, hogy vizsgálati objektumaik az esetek jelentős részében értelmes módon csak számos, esetleg igen sok bélyeggel (tulajdonsággal, változóval, stb) jellemezhetők. A biológus kutató vizsgálódásai során rengeteg hasznos információhoz jut, amely gyakran áttekinthetetlen masszaként rejti el a mélyebb összefüggéseket. Ha maga a kutató tisztában is van bizonyos összefüggésekkel – hiszen elég sokat dolgozott az adatgyűjtés során ahhoz, hogy ez így legyen –, nemigen tudja azokat mások számára is érthető, egyszerű formába hozni a napjainkban rendkívül széles körben alkalmazott többváltozós módszerek segítsége nélkül.

E módszerek alkalmazási lehetőségeit két – csak a célkitűzéseket tekintve élesen elváló – fő csoportba oszthatjuk. A többváltozós eljárások egy része voltaképpen a biometriában tárgyalt egyváltozós módszerek[1] kiterjesztése sok változóra. Feladatuk ennek megfelelően megegyező: szignifikancia-próbák segítségével adnak lehetőséget statisztikai következtetésekre. Tipikus példa a többváltozós variancia-analízis vagy MANOVA (amelyben az egyes "kezelések" hatását egyidejűleg több változón mérjük le) és a többszörös regresszió (egy "függő" változó és számos "független" ható tényező közötti függvénykapcsolatot keressük). A statisztikai hipotézis-vizsgálatok szerves része a "populáció" (=statisztikai alapsokaság, tehát nem keverendő össze a genetikai populációval) valamilyen *paraméter*ének (pl. többszörös korreláció) *becslése*, melynek alapján később oksági összefüggéseket kereshetünk, és előrejelzésre (predikcióra) alkalmas modelleket építhetünk. Így például a becsült regressziós koefficiensek alkalmasak lehetnek a függő változó értékének megjóslására a független változók olyan kombinációira is, amelyek eredetileg nem állanak rendelkezésünkre a vizsgálatban. Az ilyen módszerekre legcélszerűbben többváltozós *statisztikai* eljárások néven hivatkozhatunk.

A becslés mellett a biológusok számára éppen olyan fontos – a biológia történetét áttekintve bátran állíthatjuk: valójában jóval fontosabb – a másik lehetőség, a többváltozós

1 Ebben a témában a legjobb kiindulás Izsák et al. (1981) könyve, melyet nagy haszonnal forgathat – mintegy megalapozásként – a kizárólag többváltozós módszerek iránt érdeklődő Olvasó is.

módszerek *mintázat*-, vagy *adatstruktúra-feltáró* funkciója. Ebben az esetben feladatunk a lényegkiemelés, a látens struktúrák felismerése, láthatóvá tétele, vagy egyszerűen csak a biológiai mintázatok leírása (deszkripció) és tömör összefoglalása, megmagyarázása. Mindezt többnyire matematikai konstrukciók, mint például osztályok, gráfok, mesterséges dimenziók stb. bevezetésével érjük el. A lényeg tehát az adatfeltárás, amelyre a szakirodalom rendszerint az "*exploratory data analysis*" cimkével hivatkozik, és elsősorban a klasszifikáció és az ordináció módszereit érti alatta. A becslés, és ennek következtében a statisztikai következtetés ekkor elhanyagolhatóvá vagy legalábbis másodlagossá válik.

Jelen könyvben a többváltozós módszerek második csoportjáról lesz elsősorban szó, az adatszerkezetet feltáró módszerek mellett a hipotézisek ellenőrzésére alkalmas próbák legfeljebb segédeszközként jönnek számításba. Számos olvasó úgy érezheti majd, hogy sok – a hagyományos biometriából megszokott – fogalom, pl. eloszlás, szignifikancia-szint, becslés, null-hipotézis, statisztikai próba, "hiba", paraméter, stb. "túlságosan" ritkán vagy egyáltalán nem szerepel a könyvben. Ez is mutatja a többváltozós módszerek két célkitűzése közötti jelentős különbségeket.

Az exploratív többváltozós módszerek biológiai alkalmazásairól már legalább száz, központi fontosságú könyv áll rendelkezésünkre az – angol nyelvű – irodalomban. Ezzel csak rá szeretnék mutatni arra, hogy teljességre még csak távolról sem törekedhettem, nemcsak terjedelmi, hanem majdhogynem elvi okokból sem. A tárgyalt tematika megválasztásában mindenesetre szem előtt tartottam a sokféleséget, azt, hogy minél több lehetőséget villantsak fel az Olvasó előtt. Az egyes fejezetek irodalmi összefoglalói, a kötet végén található terjedelmes bibliográfia figyelembevételével elősegítik a tájékozódást, ha valaki valamely részterülethez különösképpen kedvet érez[2]. Különösen fontosak a számításokat megkönnyítő, ill. egyáltalán lehetővé tevő számítógépes programok, amelyekre minden fejezetben kitérek.

A hangsúly talán a növényökológián, cönológián és rendszertanon van, s ez némiképpen mutatja a szerző elfogultságát is eme tipikusan "többváltozós" diszciplínák mellett. A többváltozós alaphelyzet azonban a biológiában jóval általánosabban jelentkezik, amint azt a 0.1 táblázat is szemlélteti. A könyvben leírtak szerencsére kis erőfeszítéssel a biológia bármely más területére is "lefordíthatók" és adaptálhatók. Az olvasónak jut az a – remélhetően kis – feladat, hogy a szakzsargont a maga szakterületéhez igazítsa. Ha például a cönológus "nevében" kvadrátról vagy mintavételi egységről, ill. az őket jellemző "fajokról" beszélünk, akkor ezek helyett gondolatban a saját témánknak megfelelő objektumtípust és változót kell csupán alkalmaznunk.

A módszerek biológiai jelentőségére már sokan rámutattak korábban is. Viszonylag friss James & McCulloch (1990) áttekintése, amely – bizonyos fenntartások megfogalmazása mellett – leszögezi, hogy "a rendszertan és az ökológia teljes megértése a többváltozós módszerek némi ismerete nélkül ma már lehetetlen, és megfordítva: a módszerek félreértése a tudomány[ág] előrehaladásának akadályozója lehet." Mindezt hét, a rendszertanban és ökoló-

2 Jelent már meg Magyarországon biológiai tematikájú könyv(fordítás), nem is egy, amely – "helyhiányra" hivatkozva – teljesen mellőzte az irodalomjegyzéket, nagymértékben csökkentve ezzel a könyv használhatóságát. Véleményem szerint egy jó érzékkel összeállított, kiegyensúlyozott bibliográfia csaknem olyan értékes lehet, mint maga a könyv, amelyben megjelenik.

0.1 táblázat. Többváltozós alaphelyzetek a biológia különböző (határ-)területein.

Tudományterület	Objektumok	Változók
Etológia	fajok	viselkedési jellemzők
Paleontológia	rétegek	fajok
Antropológia	leletek	morfológiai ismérvek
Biogeográfia	fajok	elterjedési információ
Orvostudomány	betegségek	tünetek
Genetika	populációk	géngyakoriságok
Molekuláris biológia	fehérjék	aminosav szekvencia
Ökofiziológia	fajok	fotoszintézis-jellemzők
Növénytermesztés	fajták	termésmutatók
Erdészet	fafajok	életkori megoszlás
Hidrobiológia	tavak, folyók	vízmin•ségi jellemzők
Pszichológia	kísérleti személyek	tesztre adott válaszok
Mikrobiológia	baktérium-törzsek	szubsztrátumok
Talajtan	talajprofilok	%-os összetétel
Bioklimatika	él•helyek	éghajlati jellemzők

giában elismerten központi fontosságú folyóirat 1983-1988 közötti évfolyamainak tematikus elemzésével támasztja alá a két szerző: a cikkekben a többváltozós módszerek több, mint 500 alkalmazására sikerült rábukkanniuk. (A gyakoriságokat tekintve "dobogós" helyezések: 1. főkomponens analízis, 2. diszkriminancia elemzés, 3. numerikus osztályozás).

A téma magyar nyelvű irodalma eléggé szűk, s könyvem kimondott célja bizonyos "fehér foltok" eltüntetése a hazai biológia módszertanának térképéről. Természetesen vannak már magyar nyelvű kiadványok, de ezek egyike sem teszi – úgy érzem – feleslegessé a speciálisan biológusok számára írt kézikönyv megírását. Sváb (1979) elsősorban a többváltozós módszerek agrár-alkalmazásaiban lehet segítségünkre. Könyvének témája azonban lényegében véve a jelen kötet 7. fejezetében tárgyalt ordinációs módszerekre szorítkozik, különös hangsúlyt fektetve a főkomponens-elemzés és a diszkriminancia-analízis elméletére és gyakorlatára. A Móri & Székely (1986) szerkesztésében megjelent cikkgyűjtemény a többváltozós statisztika kemény, matematikai megalapozását adja számos szerző tollából. Ez semmiképpen sem ajánlható a témával most ismerkedőknek, de haszonnal forgathatja mindenki, aki jóval mélyebben akar leásni a többváltozós statisztikában annál, amire e könyv lehetőséget nyújt. A feltétlenül megemlítendő művek sorából nem hagyhatjuk ki Füstös et al. (1986) munkáját, amely – tematikáját tekintve – nagyobb átfedésben van jelen könyvvel, mint a másik kettő. Az ordináció módszereit, különösképpen a nem-metrikus eljárásokat rendkívül részletesen tárgyalják a szerzBUBUk. A legtöbb nehézséget a biológus olvasó számára itt a terminológiai "másság" okozza: a bemutatott – meglehetősen komplikált – példák kizárólag szociológiai és közgazdasági vizsgálatokat illusztrálnak[3]. Megemlítendő még Füstös & Kovács (1989) egyetemi tankönyve, amelyben ugyancsak jelentős terjedelmi hányad esik a többváltozós

0.1 ábra. A legfontosabb módszertani útvonalak a többváltozós adatfeltáró biológiai vizsgálatokban (szemközti oldal).

módszerekre, míg a példák társadalomtudományi jellegűek. Szinte természetes módon, a tartalom jelentős átfedésben van Füstös et al. (1986) tematikájával. Mind a négy kötettel – különösen a másodikkal – kapcsolatban megállapítható, hogy a terjedelmet és a tematika sajátosságait figyelembe véve aránytalanul kevés ábra található bennük. Mivel a biológus Olvasó – feltételezhetően – kifejezetten vizuális típus, könyvemben sokkal több ábrával és diagrammal (összesen 137) igyekszem elősegíteni az elmélet megértését és az interpretációs lehetőségek bemutatását.

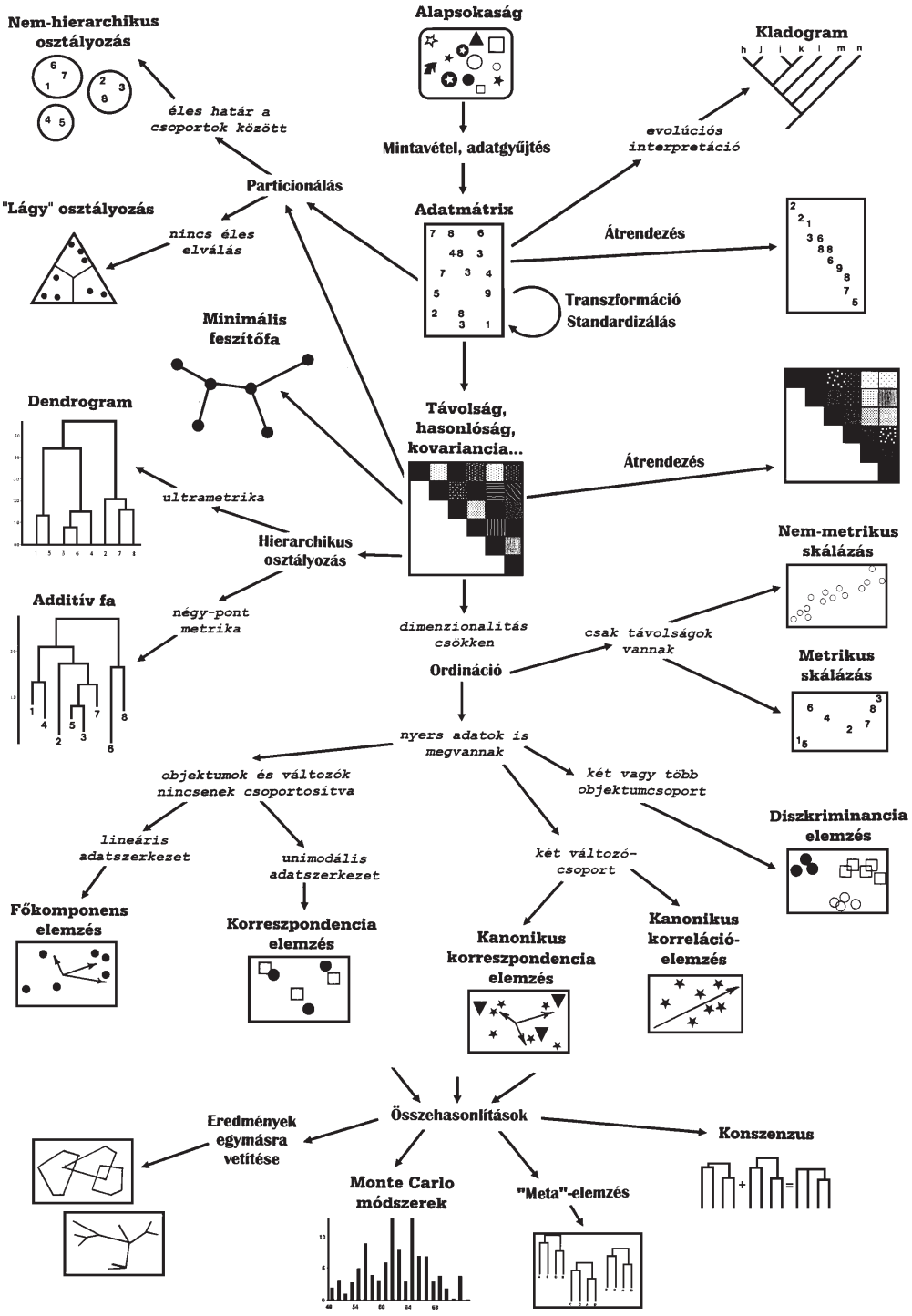
Legyen az első, 0.1 számú ábra mindjárt a könyv tematikájának, a legfontosabb módszertani útvonalaknak a summázata[4]. Természetesen nem mutat, és nem is mutathat be minden lehetőséget, de talán támpontot nyújt az Olvasó számára, hogy nagyjából mire számíthat ebben a könyvben. Nem valószínű, hogy az általa alkalmazott módszereket éppen ennek alapján fogja kiválasztani, de néhány fontos döntési lehetőséget megtalál benne. A séma fő tengelye az "*alapsokaság* → *adatmátrix* → *távolság...*" útvonal, amelyet – valamilyen formában – szinte mindenki megtapasztal. Az igazi választási lehetőségek ezután nyílnak, az osztályozás és az ordináció irányokban. Az ábra alsó részére voltaképpen mindenhonnan mutathatna nyíl (csak három van, jelképesen): itt arra utalok, hogy az ordinációs és klasszifikációs eredményekkel nem mindig elégedhetünk meg, és szükség van valamilyen, az alternatív eredmények összehasonlítására alkalmas metodológiára is.

A könyv felépítése

A bevezetőt követő kilenc fejezet tárgyalja a többváltozós módszereket, a téma előnyösnek vélt felbontásában. A fejezeteket persze nem feltétlenül kell pontosan ilyen sorrendben olvasni: bár sok keresztutalás található a fejezetek között, valójában mindegyikük külön olvasmányként is kezelhető. Aki a kladisztika iránt érdeklődik például, annak az előző részek – néhány bekezdéstől eltekintve – vajmi keveset mondanak, s közvetlenül belefoghat a 6. fejezet olvasásába. Az ordinációs módszerekhez sem feltétlenül szükséges a terjedelmes 3. fejezet ismerete, és így tovább. Leginkább a 9. fejezet az, amely erőteljesen támaszkodik az előző részekre, s ez nem véletlen, hiszen az eredmények értékeléséről és összehasonlításáról van benne szó. Minden fejezet szerkezete azonos: a módszertani alfejezeteket követően rövid irodalmi/program összefoglalót találunk, majd a száraz tényanyagot a *Kérdezz-Válaszolok!* alfejezet kötetlen és képzeletbeli dialógusai zárják. A fejezeteket követi a négy függelék az

3 E mű egyébként – szemben a másik kettővel – szisztematikusan a sokváltozós és nem a többváltozós megjelölést alkalmazza. A szóhasználat nyilván ízlés kérdése, nem feladatunk eldönteni, hogy a több-e a "sok" mint a "több" vagy sem. Mindenesetre igyekszem a "többváltozós" elnevezést következetesen alkalmazni.

4 Bevallom, hogy nem igazán szeretem az ilyen típusú "folyamatábrákat", mert elég ritkán sikeresek: sokszor túl részletesek és áttekinthetetlenek – és ezért használhatatlanok –, máskor pedig olyan végtelen egyszerűek, hogy voltaképpen nincs is rájuk szükség. Most úgy éreztem azonban, hogy a kis illusztrációkkal kiegészített diagram elősegítheti a könyv témájának gyors áttekintését.


Alapsokaság
Mintavétel, adatgyűjtés
Adatmátrix
Transzformáció
Standardizálás
Átrendezés
Kladogram
evolúciós interpretáció
Particionálás
éles határ a csoportok között
Nem-hierarchikus osztályozás
"Lágy" osztályozás
nincs éles elválás
Minimális feszítőfa
Dendrogram
ultrametrika
Hierarchikus osztályozás
Távolság, hasonlóság, kovariancia...
Átrendezés
Additív fa
négy-pont metrika
dimenzionalitás csökken
Ordináció
csak távolságok vannak
Nem-metrikus skálázás
Metrikus skálázás
nyers adatok is megvannak
objektumok és változók nincsenek csoportosítva
két vagy több objektumcsoport
lineáris adatszerkezet
unimodális adatszerkezet
két változó-csoport
Diszkriminancia elemzés
Főkomponens elemzés
Korreszpondencia elemzés
Kanonikus korreszpondencia elemzés
Kanonikus korreláció-elemzés
Összehasonlítások
Eredmények egymásra vetítése
Monte Carlo módszerek
"Meta"-elemzés
Konszenzus

adattáblázatokkal, a programok beszerzési forrásaival, a mátrixalgebrai összefoglalóval és az "első" angol-magyar "többváltozós-elemzéstani" kisszótárral és kislexikonnal. Az irodalomjegyzék nemcsak bibliográfia, hanem egyben a szerzők mutatója is, így a záró tárgymutatóban már csak valóban a "tárgyak" és fogalmak szerepelnek. (Elnézést kell kérnünk tehát minden második és további szerzőtől, ill. az őket kereső Olvasóktól, mert az irodalomjegyzékben természetszerűleg az első szerzők szerint készül a sorrend, így sokan kimaradnak a visszakeresés lehetőségéből.)

Köszönetnyilvánítások

A kötetben leírtakat többen átolvasták, hozzájárulva a félreértések és hibák számának csökkentéséhez. Külön köszönettel tartozom Kontra Györgynek a részletes kritikáért, s azért, hogy mindenféle gyengeségekre még idejekorán rámutatott. Értékes megjegyzéseket fűzött a kézirathoz, ill. a "hibavadászatban" segített sokat Tóthmérész Béla, Garay József, Ódor Péter, Demeter András, Kontra Klára, Peregovits László, Czárán Tamás, Scheuring István és id. Podani János. Megköszönöm hallgatóimnak a kérdező odafigyelést, s azt, hogy egy ideig "áldozatai" voltak e készülő munkának. A kötet nem jöhetett volna létre hazai és külföldi kollégáim, és természetesen az e témában dolgozó összes biológus és matematikus kutató közvetett "közreműködése" nélkül.

Köszönet illeti egyes, a könyvben említett programcsomagok fejlesztőit és terjesztőit a térítés nélkül rendelkezésemre bocsátott programokért: **Statistica** (StatSoft Inc., Tulsa, Oklahoma, USA), **BMDP** (Statistical Software Ltd., Cork, Írország) és **PHYLIP** (J. Felsenstein, University of Washington, Seattle, USA).

E kötet elkészítését az OTKA T6032 sz. pályázat tette lehetővé (a pályázat futamideje időközben már lejárt), míg a könyv megjelenéséhez az OTKA a P18941 sz. könyvkiadási pályázatom elfogadásával járult hozzá. Enélkül a könyv megírására még gondolni sem mertem volna; s ezúttal fejezem ki köszönetemet az anyagi támogatásért.

Fontos megjegyzés

Hibamentes könyv valószínűleg nem létezik, így – minden erőfeszítés ellenére – ez a kötet sem az. A szerző előre is megköszöni minden olyan Olvasójának javításait, észrevételeit, bárminemű megjegyzéseit és kérdéseit, aki mindezt eljuttatja a podani@ludens.elte.hu "drótposta" címre. Az esetlegesen felmerülő hibák állandóan frissített jegyzéke, a téma lényegét érintő megjegyzések összefoglalója, a *Kérdezz – Válaszolok!* alfejezetekből "kimaradt" – mert újonnan felvetődő – problémák, és a példaadatok mátrixai az interneten, a http://ramet.elte.hu/~podani címen találhatók meg.