

Opponensi vélemény

Dr. Botta-Dukát Zoltán

Cönológiai adatbázisok alkalmazása a vegetációkutatásban: Módszerfejlesztések és esettanulmányok

c. doktori disszertációjáról

Az adatbázisok fontos szerepet játszanak a különféle tudományterületeken, és az „adatbázis” szó hallatán elsősorban inkább különféle számítástechnikai rutinfeladatokra asszociálhatunk. Magam is erre gondoltam először, de hamar kiderült, a dolgozatban ilyesmiről viszonylag kevés szó esik, s a figyelem a cönológiai adatok feldolgozásának lehetőségeire összpontosul, különös tekintettel az osztályozásokra és azok értékelésére. A kidolgozott téma tehát valamivel általánosabb, mint amit a cím sejtet. A nyolc fő téma némelyike valójában egyáltalán nem igényli, hogy adataink adatbázisból származzanak.

Az adatbázisok bizonyos értelemben kockázatos kiindulást jelentenek, hiszen tartalmuk igen heterogén lehet: számos kutató különböző taxonómiai és ökológiai ismeretekkel, eltérő erőfeszítéssel és igen változatos feltételek mellett rögzített megfigyelését összesítik. A veszélyekkel a Szerző is tisztában van, kiemelve a preferenciális mintavétellel és a változók függetlenségével kapcsolatos gondokat. Egyik témája éppen a preferenciális és a random mintavétel összehasonlítása. Ugyanakkor szemérmesen hallgat egy másik problémáról, az alkalmazott mérési skála nehézségeiről. Köztudott módon, Braun-Blanquet akkor dolgozta ki módszerét, amikor objektív mintavételezésről, statisztikákról, numerikus osztályozásról szinte szó sem volt, nem is nagyon lehetett. Csak a számítógépes módszerek elterjedése során derült ki, hogy nemcsak a preferenciális mintavételezési stratégia, hanem az abundancia-dominancia skála ordinális jellege is jelentős akadály az objektív értékelésben. A dolgozat, úgy látom, ezt a gondot általában a prezencia-abszencia adatokra egyszerűsítéssel oldja meg. Ez önmagában helyeselhető, de azért az A-D skála problémájára én részletesen kitértem volna, hiszen az archivált adatok jelentős része ilyen jellegű – főleg a 20 évnél régebbiek.

A Szerző nagy statisztikai apparátust vesz igénybe, s e módszerek alkalmazásában, az eredmények interpretációjában különleges jártasságot árul el. Sokfélék a feldolgozott adatok is, hiszen változatos társulás-típusokat képviselnek. A vizsgált ökológiai/cönológiai problémák köre a mintavételezéssel kapcsolatos döntések értékelésétől a cönológiai fidelitáson át a niche szélességig terjed. Módszertanilag a hangsúly a statisztikai hipotézisvizsgálatokon van, magának az osztályozásnak, ill. ordinációnak az eredménye ritkán tűnik fel a dolgozatban. A statisztikai tesztek főleg a standard eloszlásokra visszavezethető klasszikus próbák, illetve azok változatai, bár randomizációs módszerek is előfordulnak a dolgozatban (pl. a fidelitásról szóló részben). A hipotézisvizsgálatok eredménye többnyire táblázatos vagy oszlopdiagramos formában jelenik meg. Egyes esetekben talán a sok diagram helyett az eredmények többváltozós meta-analízise (pl. ordinációja) szemléletesebb lett volna, amire egyébként a Szerző maga is jó példával szolgál a fidelitás-indexek összehasonlító vizsgálatában. Ugyanakkor például a 24-26. ábrákon, három oldalon szereplő információ hatékonyan sűrítető lett volna egy hasonló meta-analízis felhasználásával.

A szakirodalom ismeretében a Szerző nagy jártasságot árul el, mondhatnánk: ebben a tekintetben is a dolgozat „up to date”. Ami a külsőségeket illeti, itt sincs gond, hiszen szépen

összeállított, formailag is kiemelkedően jó dolgozatról van szó. Apró kifogás, hogy néhány hosszabb táblázatot inkább egy függelékben helyeztem volna el.

Megjegyzések, apró kérdések, amelyekre nem feltétlenül várok választ:

1. Kár az egész dokumentumon végig vonuló „dc_16_10” fejlécért.
 5. A cönológiai „felvétel” sokszor nem is egy konkrét méretű kvadrátként jelent meg, főleg a cönológia kezdeti időszakaiban, hanem csak az éppen észlelt fajok egy listája volt.
 13. Bár ma is sokan hivatkoznak rá, a „minimum area” továbbra is egy igen rosszul definiált alapfogalom – objektivizálására voltak és vannak törekvések, de általánosan alkalmazható eljárás a minimum area meghatározására nem ismeretes.
 14. A 7-fokozatú Braun-Blanquet skála megengedése egy rossz gyakorlat konzerválása.
 17. Magam is egyetértek azzal, hogy a függetlenség megsértése nem jelenti azonnal az összes statisztikai teszt érvénytelenségét. A randomizációs teszteket ez például egyáltalán nem zavarja.
 19. Az „alapsokaság”, annak „elemei” stb. egy standard statisztikai zsargon kifejezései, ahol valóban diszkrét „egyedekből” áll a „populáció”. Jelen dolgozat témája azonban nem ilyen, hiszen a társulások térbeli kontinuumot alkotnak, amelyben a kutató maga határol el kisebb részleteket, hogy adatokhoz jusson. Ezért célszerű lett volna utalni a diszkrét és a folytonos jellegű univerzumhalmazok eltéréseire, arra, hogy az utóbbiban egyáltalán nem elemek „kiválasztásáról” van szó (vö. Kenkel et al. 1983, Vegetatio). Ugyancsak ki lehetett volna térni arra a fontos mozzanatra, hogy a mintavételezéssel foglalkozó munkák túlnyomó része eleve feltételezi, hogy a mintavétel és a későbbi adatelemzés célja becsléssel kapcsolatos, tehát variancia-minimalizálásra törekszik. A társulások (s általában a biológiai diverzitás) vizsgálata ugyanakkor olyan mintavétel esetén optimális, amely varianciát maximalizál, vagyis az alapsokaságot a lehető legnagyobb változatosságában mutatja meg. Éppen emiatt a térbeli autokorreláció (vagy éppen egy filogenetikai autokorreláció) eleve szükségszerű – s ezen a ponton már teljesen egyetértek a Szerzővel: az autokorreláció léte nem érinti a cönológiai érvelés lényegét. Autokorreláció nélkül pl. nem volna mit osztályozni.
 21. Az „adatfeltárás”-on, Hallgren és mtsai véleményével ellentétben, én egészen mást értek, s nem leíró statisztikát.
 24. A p-a adatokra való áttérés itt is helyeselhető, mint általában mindenütt, különösen, ha a BB skála alapján rögzítették az adatokat.
 28. A mintavétel időpontja azonos volt a két vizsgálatban? Erre nem találtam utalást a bevezető részben.
- Az oldal alján lévő mondatot nem értem: „más, mint a preferenciális mintavétel esetén”? A BB módszer is preferenciális.
30. A zajszűrés jelen módszerei egyértelműen egy lineáris-ortogonális világot tekintenek vonatkoztatási alapnak, s ami abba nem illeszkedik, azt automatikusan zajnak veszik.

Érdemes lenne egy kicsit filozófiai területre is átmennünk, hiszen erősen kérdéses, hogy maga a természet is „így gondolja-e”. Általában pedig a „zaj” valamilyen sztochasztikus komponens, ami eleve szükségszerű a biológiai adatokban. Gauch definíciója számomra eléggé nehezen értelmezhető.

A PCA alkalmazhatóságának nem alapfeltétele a normalitás – a közhiedelemmel ellentétben. Ennek erőteljes „megsértése” sem feltétlenül akadály a PCA eredmény értelmezhetőségének.

31. Az itt leírt munkamenetet nem teljesen értem. Először kiderül, hogy a 4. fejezet példáiban a Jaccard különbség szerepel, ami prezencia-abszencia koefficiens. Az adatgenerálás módszere (5. táblázat) ehhez képest túlságosan szofisztikáltnak tűnik.

Egyébként a „zajszűrés” alkalmazása érdekes. A kérdés az, hogy a Jaccard mátrixból milyen osztályozást kapunk, ha először ordinációs módszerrel elemezzük azt – de: hány tengelyt érdemes figyelembe venni? A törött-pálca modell alkalmazása itt jó ötletnek tűnik.

33. Valójában csak HÁROM hierarchikus módszerről van szó, hiszen UPGMA=átlagos lánc. Korrekt, hogy a Ward módszer csak az ordinációs tengelyek esetén alkalmazható.

34. Azt azért érdemes lenne alaposabban is megvizsgálni, miért van az, hogy a leggyakrabban alkalmazott hierarchikus módszer, az UPGMA, „zajszűrés” után miért ad kisebb hasonlóságot a várt osztályozással, mint zajos adatokból. Hogyan segít ebben a belső jósági mérőszámok használata?

36. Az alsó táblázatban a jobb oszlopban helyesen e_{10} és e_{00} szerepel.

37. Az alapeloszlást valóban χ^2 -nek nevezzük, de a próba-statisztikát már X^2 -tel jelölném – hiszen csak később derül ki, hogy származhat-e abból az eloszlásból.

Minek a szórásáról van szó a legelső képletben? Ez csak később derül ki...

38. Mi az N a legelső képletben? (Jó lett volna a képleteket megszámozni, hogy könnyebben hivatkozhatunk rájuk.)

39. Mi a g jelentése?

40. A rang-korrelációk mátrixából én nem-metrikus ordinációt (is) készítettem volna. Ehhez hasonló meta-elemzés – ahogy már említettem – más esetekben is segíthetett volna. Jelen esetben viszont a korrelációs mátrixok (8-11. táblázat) talán már feleslegesek is. Magunk egyébként nem tapasztaltunk ekkora eltéréseket az IndVal változatok között.

48. Az „érezem” szó a lábjegyzetben egyike a nagyon kevés számú elírásoknak. (L. még „csoportba” az 56. oldalon)

55. és 59. A „tömörség” és az „összekötöttség” feltehetően nem teljesen független szempontok, hiszen ami tömör, az megfelelően összekötött is. A clusterek sokdimenziós térbeli alakja lehet inkább döntő és jobb szempont, vagyis, hogy gömbszerű-e, vagy megnyúlt.

57. Ez a rész a különféle jósági kritériumok egyfajta review-szerű összefoglalása, és nagyon hasznos.

72. A 23. ábrán a jelek szerint 3, 5 illetve 3 csoportba vannak a pontok osztva, és nem 4, 3 ill. 5 csoportba.

75-77. Itt tehát ordinációs meta-analízis segített volna a jobb szemléltetésben. Ugyanígy, az indexek összevetését is elkészítettem volna ordinációkkal, ami a 27-29. ábrák egyfajta összesítése lehetett volna.

88. Megjegyzés: A relatív gyakoriságok hasonló aránya persze még sokféle florisztikai kiosztásban is megvalósulhat. Tehát két gyakoriságeloszlás megegyezése közel sem jelenti azt, hogy a kiinduló felvételek közül bármelyik pár hasonló lenne.

90. Az 1 - Sörensen index nem metrika, a korábban is használt Jaccard talán jobb lett volna itt is.

95. C nem az i és j FAJOK együttes előfordulásainak a száma?

96. Érdekes eredmény a bináris skálára való áttéréssel kapcsolatban....

100. Érdekes a niche-szélességnek eme közvetett mérése, talán „indirekt niche szélesség”-nek is lehetne mevezni.

105. A 35.d ábra pontjaira majdnem biztosan szignifikáns lineáris korreláció jön ki. A c és d ábrák között alig van különbség.

107. A bináris euklidészi távolság persze csak közvetett módon veszi figyelembe a felvételek fajszerkezetét, hiszen $\sqrt{b+c}$ ugyanaz lehet a bármely értékére.

111. A BB skála értékeinek behelyettesítése intervallum-közepekkel bevett gyakorlat, jobb, mintha enélkül számolt volna az adatokkal.

A SYNTAX 2000 voltaképpen 2001-re datálódik, semmiképp se 1994-re.

112. Érdekes lenne ugyanezt megvizsgálni az általam javasolt módszerrel (In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock & Y. Baba (eds.), *Data Science, Classification, and Related Methods*. Springer, Tokyo) amely egyben megmutatja az is, hogy mely fajok felelősek a kialakult csoportokért, és azt a csoportszámot tekinti optimálisnak, amelyet a legtöbb faj támogat.

115. Fajösszetételen vagy inkább borításon alapult az elemzés?

119-125. Ez a táblázat nagyon hosszú, megtöri a szöveg lendületét. Jobb lett volna függelékbe tenni.

129-130. „A felvételek ordinációja...” Be kellett volna írni, hogy MELY felvételek... Jó lett volna látni a százalékos varianciákat.

Egy új cikk, ami majd a jövőben érdekes lehet: E. Feoli, P. Ganis, R. Venanzoni & V. Zuccarello 2011. Toward a framework of integrated knowledge of terrestrial vegetation system: The role of databases of phytosociological relevés. Plant Biosystems 145, Supplement 1, 74-84.

Összefoglaló értékelés

Örvendetes tény, hogy a Szerző életben tartja a cönológiai hagyományokat azáltal, hogy szüntaxonómiai osztályozásokat pontosít, értékelési lehetőségeiket kibővíti és nomenklaturai vitákba is belemegy. Ugyanakkor a cönológia eszköztárát erőteljesen modernizálja, a mai kor igényeinek megfelelően átalakítja, s ebben jelentős előrehaladást ér el még a pár évvel ezelőtti helyzethez képest is. Megmutatja, hogy adatbázisok, ill. általában a nagy adathalmazok mennyire sokféleképpen vizsgálhatók, és milyen fontos ökológiai háttér-összefüggések mutathatók ki belőlük. Olyanok is, amire bizony nemigen gondolna a legtöbb kutató.

A legfontosabb eredmények véleményem szerint a következők:

1. A preferenciális és random mintavétel összehasonlításának újszerű módja;
2. A zajsűrűs alkalmazása osztályozások értékelésében;
3. Új módszert ajánlott a csoportok ismétlődésének vizsgálatára dendrogramokból kapott partíciók alapján;
4. Új algoritmust dolgozott ki a niche-szélesség becslésére; és
5. Mocsárrétek osztályozásának elkészítése, közvetlen gyakorlati alkalmazhatósággal.

Egyetlen kérdésre szeretnék választ kapni, ami inkább arra vonatkozik, ami nincs is benne a dolgozatban, azaz: mi a véleménye az adatbázisokban összegzett adatok heterogén skáláiról, különös tekintettel a Braun-Blanquet féle AD skálára?

A fentieket összefoglalva elmondhatom, hogy a dolgozat tudományos eredményei elegendők az MTA doktori fokozat megszerzéséhez, és így egyértelműen javaslom a nyilvános vita kitűzését.

Budapest, 2011. szept. 5.

Dr. Podani János