

Válasz Podani János akadémikus opponensi véleményére

Szeretném megköszönni Podani Jánosnak az opponensi munkát, dolgozatom részletes értékelésére fordított idejét és fáradságát. Külön köszönöm aprólékos megjegyzéseit, amelyek nagyon hasznosak a még nem publikált eredmények közzétételénél.

Opponensem szerint a dolgozat címénél általánosabb annak témája, mert a módszertani fejlesztések alkalmazásakor nem feltétlenül adatbázisokból származnak az adatok. Ezzel teljesen egyetértek: a dolgozatban több olyan módszer is szerepel, amelyek viszonylag kis adathalmazokra is alkalmazhatók. Ugyanakkor úgy gondolom, hogy ezen a területen a módszertani fejlesztések „motorja” a nagy adatbázisok megjelenése. Ezek – a nagy teljesítményű személyi számítógépekkel együtt – olyan mennyiségű cönológiai felvétel együttes elemzését teszik lehetővé, amelynél az eredmények a hagyományos módszerekkel már nehezen áttekinthetőek, így mindenképpen szükség van az új értékelő módszerekre. A módszerek más része (pl. fajkészlet és specialistaság becslése) kifejezetten igényli a nagy adatmennyiségeket. Ezért jelenik meg az „adatbázis” szó a dolgozat címében is.

Az adatbázisokban szereplő adatok minősége sok szempontból heterogén lehet, ami téves eredményekre vezethet. A heterogenitás egyik forrása az eltérő mérési skálák alkalmazása. Opponensem hiányolja, hogy ezzel a valóban fontos kérdéssel a dolgozat nem foglalkozik. Mivel a dolgozat megírásakor erre vonatkozó saját eredményeim nem voltak, a dolgozatban legtöbbször bináris adatokat használtam, amelyeknél ez a kérdés nem merül fel. Az egyik tanulsága a munkának, hogy ha elég sok felvétellel dolgozunk, még ilyen alacsony információtartalmú skála esetén is milyen sok érdekes összefüggést tárhatunk fel. Időközben a Masaryk Egyetem kutatóival, Milan Chytrý-vel és Lubomír Tichý-vel együttműködésben megvizsgáltuk a legproblémásabb skálátípus, az egyszerűsített A-D skálák (például Braun-Blanquet skála) feldolgozásának lehetőségeit. A témában írt cikkünk jelenleg elbírálás alatt van a Journal of Vegetation Science-nél. Röviden összefoglalva: a legelterjedtebb feldolgozási módja ennek a skálának az intervallumok helyettesítése azok középvértékeivel. Opponensem több cikkében is bírálta ezt az eljárást, véleménye szerint csak az AD értékek sorrendje ismert, így azokat ordinális skálának kell tekinteni és ennek megfelelően feldolgozni. Ez azonban csak akkor igaz, ha a skálaértékek terepi megállapításánál valóban figyelembe veszik az egyedszámokat is. A valóságban ez szinte soha nem történik meg, vagyis az AD értékek a borítás intervallumok kódjai. Mivel az intervallumok határai ismertek, több információval rendelkezünk, mint az ordinális skála esetén. Mike Dale hívta fel a figyelmünket, hogy a szimbolikus adatok és azon belül az intervallum adatok (nem intervallum skálájú adatok!) elemzésére kidolgozott módszerek itt is használhatóak lehetnek. Mivel az összehasonlítás célja nem a mérési pontosság (az intervallumok szélességének), hanem az intervallumok helyzetének összevetése, a szimbolikus adatokra kidolgozott távolságfüggvények alkalmazása egyenértékű azzal, ha az intervallumokat azok közepével helyettesítjük. Vagyis az általában alkalmazott módszer matematikai szempontból is megalapozott.

Válaszaim Opponensem kisebb megjegyzéseire:

- A cönológiai felvétel területe erősen befolyásolhatja az elemzés eredményét (pl. Otypková & Chytrý 2006, J. Veg. Sci. 17: 465–472), így nem célszerű nagyon eltérő területű felvételeket együtt elemezni. Emiatt azok a nagyon régi felvételek, amelyek nem egy konkrét méretű kvadrátban készültek általában nem is kerülnek kiválasztásra. Mivel a minimál area megállapítására, ahogy Opponensem is írja, nincs általánosan alkalmazható objektív módszer, az egyes vegetációtípusokban alkalmazott kvadrát méretek szakmai közmegegyezés alapján alakultak ki.
- A 7 fokozatú Braun-Blanquet skálát az adatgyűjtés során azért engedték meg, mert attól tartottunk, hogy egy ennél szigorúbb (például a százalékos borításbecslést előíró) mintavételi protokoll esetén nem lesz elegendő felvételezésre jelentkező botanikus. Annak ellenére, hogy nem volt kötelező alkalmazni, a legtöbb új felvétel százalékos borításbecsléssel készült.
- A vegetáció vizsgálatok az alapsokaság valóban folytonos. A dolgozatban (és a fejezet alapjául szolgáló cikkben) az egyszerűbb tárgyalás miatt csak a diszkrét alapsokasággal foglalkoztunk. A kettő között azonban könnyen kapcsolat teremthető: ha a mintavételi terület méretét és alakját rögzítettük, elvben megállapíthatjuk a lehetséges mintavételi egységek halmazát, ami már egy diszkrét alapsokaság. Mivel a mintavételi egység méretével és alakjával nem foglalkozunk, a térbeli függőség pedig a diszkrét alapsokaság esetén is tárgyalható, ezt az egyszerűbb megoldást választottuk. Abban azonban teljesen igaza van Bírálómnak, hogy erre az átalakításra (folytonosból diszkrét alapsokasággá) a szövegben fel kellett volna hívni a figyelmet. Teljesen egyetértek Opponensemmel, hogy míg becslésnél a variancia minimalizálása, addig mintázatelemzésnél annak maximalizálása a cél a mintavétel során, és a vegetációs vizsgálatok legtöbbször ebbe a második csoportba tartoznak. Magam nem foglalkoztam ugyan ezzel a kérdéssel (és ezért nem szerepel a dolgozatban sem), de PhD hallgatóm Lengyel Attila a heterogenitás maximalizálásán alapuló újra-mintavételezési (re-sampling) módszert dolgozott ki a felvételek kiválasztására az adatbázisokból (Lengyel et al. 2011, J. Veg. Sci. 22: 175-183).
- Ha adatfeltárás alatt az adathalmaz potenciálisan értelmezhető jellemzőinek kiemelését értjük, akkor ebbe a definícióba véleményem szerint egyaránt beleférnek a leíró statisztikák és az Opponensem könyvében tárgyalt sokváltozós módszerek is.
- A preferenciális és a random mintavétel egy időben történt. A homoki gyepek évek közötti nagy változatossága miatt egyébként nem lennének összehasonlíthatók. A szövegben az szerepel, hogy „párhuzamosan készültek” a kétféle típusú felvételek. Valóban lehetett volna egyértelműbben is kimondani, hogy egy időben.
- A 28. oldal alján egy elírás van. Helyesen: „más, mint a **nem** preferenciális mintavétel esetén”.
- A stochaszticitásból adódó zaj szükségszerűen jelen van biológiai adatokban. Esetenként (például kis populációk dinamikájának modellezésekor) figyelmen kívül hagyása téves következtetésre vezet. Máskor viszont a célunk éppen a trendek kiemelése, a stochaszticitás elhanyagolásával. Mivel nem létezik egy Véletlen nevű erő, ami a stochaszticitást okozza, hanem az a figyelembe nem vett hatásokból adódik, mindig a vizsgálat kérdése szabja meg, hogy mit tekintünk zajnak. A lineáris-ortogonális megközelítés alkalmazása nem szükségszerű, csak egy lehetőség. Opponensi véleményében Tóthmérész Béla például a nem metrikus ordinációs módszerek alkalmazásának lehetőségét vetette fel.

- Bár a dolgozat szövegéből így tűnik, nem a PCA-nak, hanem a fontos tengelyek számának megállapítására javasolt módszereknek előfeltétele a sokváltozós normális eloszlás. A főkomponens analízis akkor is elvégezhető, ha ezek a feltételek nem teljesülnek.
- A szimulált adatok előállítására a dolgozatban mindenütt ugyanazt az algoritmust alkalmaztam, csak a paramétereket (elsősorban a mintavételt befolyásoló beállításokat) változtatva. Az algoritmus a Fridley és munkatársai által a niche szélesség becslésére kidolgozott módszert ismertető cikkből származik. Mivel ezzel kapcsolatos eredményeim az ő módszerük továbbfejlesztése, szükségesnek tűnt, hogy a szimulált adatok előállításánál is az ő módszerüket használjam. Ezután úgy gondoltam, hogy egyszerűbb ezt a szakirodalomban már elfogadott módszert használnom a dolgozat többi részében is, még akkor is, ha a jelenlét-hiány adatok előállításához valóban túlságosan szofisztikált.
- Négy hierarchikus klasszifikációs módszert használtam, de a felsorolásban valóban csak három szerepel, mert a Ward-módszer kimaradt.
- A sziklagyepi adatoknál az UPGMA a „zajszűrés” után valójában a várttal teljesen megegyező klasszifikációt ad, a táblázatban szereplő érték hibás. A scriptet az R program jelenlegi verziójával lefuttatva nem sikerült a hibás értéket reprodukálnom. Elképzelhető, hogy programhiba következménye volt, de az is, hogy elírás történt.
- A 37. oldal alján szereplő képletben a szórás a faj adott szüntaxonon belüli frekvenciájának a szórása.
- A 38 oldalon a N – következetlen módon – az összes felvétel számát jelöli. Sajnos az általában használt kétféle jelölési konvenció ezen a helyen keveredik.
- A „g” a csoportméret hatásának kiküszöbölésére utaló, az angol nyelvű cikkből átvett jelölés, amit jobb lett volna alkalmazása előtt és nem csak a 39. oldal közepén elmagyarázni.
- A 9-12. ábrák alternatívájaként elkészítettem az ordinációkat sokdimenziós skálázással is, ahogyan azt Opponensem javasolta. A megfigyelhető mintázatok nem változtak meg lényegesen egyik esetben sem. A legfontosabb eltérés, hogy a 11. ábrán megfigyelhető patkó-alakú elrendeződés helyett egy gradiens rajzolódik ki, ahol a Podani-Csányi féle módosított IndVal köztes helyett foglal el az eredeti IndVal és a Phi koefficiens között.
- A csoportok „tömörsege” és „összekötöttsége” valóban nem teljesen független változók, hiszen tömör, de nem megfelelően összekötött csoport nem létezik. Azonos összekötöttség mellett a nagyobb tömörség a gömbszerűbb, a kisebb tömörség az elnyújtottabb csoportra jellemző. Mivel nem csak a gömbszerű csoportok lehetnek biológiailag értelmesek, az összekötöttség, mint jósági szempont alkalmazása lehetővé tenné, hogy az elnyújtott, nem tömör de erősen összekötött csoportok is magas jóság értéket kapjanak. Míg a klasszifikáció során az egyszerű lánc algoritmus használja az összekötöttséget, addig az értékelő módszerek között nem találtam olyat, ami ezen alapult volna.
- Az ordinációs meta-analízis a jósági mérőszámokat egymáshoz hasonlította volna, megmutatva, hogy melyek viselkednek hasonlóan. Azonban nem válaszolt volna arra a kérdésre, hogy melyek viselkednek az elvárásainknak megfelelően, ami viszont az alkalmazott ábrákról jól leolvasható.
- A relatív gyakoriságok hasonló aránya két csoportban természetesen önmagában nem garantálja, hogy a két csoport felvételei is hasonlóak. Ha azonban figyelembe vesszük, hogy a fajok együttlőfordulása nem véletlenszerű, és a csoportokat nem random, hanem numerikus klasszifikációval alakítottuk ki két párhuzamos mintában, akkor kicsi annak az esélye, hogy ilyen műterméket kapjunk.

- A Sorensen index alkalmazása a Jaccard helyett az idézet tanulmányban nem az én döntésem volt. Tekintettel a kettő közötti szoros korrelációra, valószínűleg nem lett volna minőségileg más az eredmény.
- A 95. oldalon a képletben természetesen a fajok együttes előfordulásainak száma szerepel.
- Az „indirekt niche szélesség” elnevezést nagyon jó ötletnek tartom, a továbbiakban alkalmazni fogom.
- A 107. oldalon az Euklideszi-távolságra vonatkozó mondat helyesen így hangzana: „Előnye az általánosan használt bináris különbségekkel (pl. Jaccard, Sorensen) szemben, hogy **ha a fajkészlet nem fed át**, akkor a felvételek fajsámát is figyelembe veszi.”
- Az eredeti cikkben kerülni igyekeztük a kéziratokra (pl. programok kézikönyve) történő hivatkozást, így kerülhetett a SynTax2000 mögé a programcsomag korábbi változatához kapcsolódó, 1994-es könyv referenciaként.
- Az Opponensem által kidolgozott módszer a geometriai jóság mérőszámok közé tartozik, vagyis feltételezi, hogy az alkalmazott távolságfüggvény megfelelő és a csoportok elválását méri (de nem a szokványos módon, hanem fajonként). Valóban érdemes lett volna ezt is kipróbálni, de a felvételeink száma magasabb, mint a SynTax5-ben megengedett maximum, a módszer más programokban pedig tudomásom szerint nem elérhető.
- A mocsárrétek elemzése borítás adatokon alapult, és az értékeléskor is figyelembe vettük a domináns fajokat.
- A 39.-40. ábrán az első négy tengely 10.8, 6.7, 6.4 és 4 százalékot magyarázott.

Az Opponensem által feltett kérdést két részre bontanám:

- Hogyan használjuk az adatbázisokban már meglevő (vagy elkészült, de még nem digitalizált) Braun-Blanquet skálájú adatokat?
- Gyűjtsünk-e a továbbiakban is ilyen skálával adatokat?

Az első részkérdésre korábban már válaszoltam. Azt csak annyiban egészíteném ki, hogy természetesen csak olyan kérdéseknél használhatók az így gyűjtött adatok, amelyekhez ez a pontosság elegendő. A legtöbb durva léptékű vizsgálat ilyen, hiszen általában elegendő azt figyelembe venni, hogy a faj ritka, szubdomináns vagy domináns. A lokális vizsgálatok azonban ennél pontosabb adatokat igényelhetnek.

A második kérdésre a válaszom egyértelmű nem. A terepi felvételezés során véleményem szerint százalékos borításbecslést érdemes alkalmazni, természetesen figyelembe véve a becslés elérhető pontosságát és a biológia szempontból értelmes különbségeket. Az 1% és 2% közötti különbség terepen felismerhető és biológiailag is fontos lehet (pl. nyílt növényzetben). A 40% és 41% közötti különbség felismerésével viszont nem érdemes küszködni, mert alig van biológiai információja.

Végezetül Opponensem munkáját még egyszer megköszönve tisztelettel kérem válaszaim elfogadását.