dc_16_10

Cönológiai adatbázisok alkalmazása a vegetációkutatásban

MTA Doktori értekezés tézisei

Botta-Dukát Zoltán

2010

Bevezetés

Egy-egy tudományterület fellendülése az új elméletek, vagy az új adatforrások megjelenéséhez kötődik. Az új elmélet magyarázatot adhat a korábban is ismert, de a régebbi elméletek kereteibe nem illeszkedő adatokra (például az általános relativitáselmélet alapján jól magyarázható a Merkúr pályájának eltérése a korábbi elméletek alapján várttól), de ugyanakkor újabb adatok gyűjtésére is inspirál. A hatás azonban nem csak egyirányú. Hubble felfedezése előtt a táguló világegyetem csak Einstein kozmológiai modelljeinek egy lehetséges, de valószerűtlennek tartott megoldása volt. A vörös-eltolódás felfedezésével, amit a rádiócsillagászat megjelenése tett lehetővé, viszont minden kozmológiai elmélet központi elemévé vált.

A fenti – szándékosan igen távoli – példában az új adatforrások megjelenése az adatgyűjtés módszerének megváltozását és a hozzáférhető adatok körének ebből következő kibővülését jelentette. A vegetációtudományban az adatgyűjtés módjában az elmúlt 100 évben nem volt ilyen drámai változás. Már a 20. század elején készültek cönológiai felvételek, amelyek a mintaterület fajlistáját és a fajok tömegességét rögzítették, és mai napig is ez a leggyakrabban használt adattípus. A módszer részletei – mintavételi egység mérete, becslési skála stb. – ugyan változhattak, de alapvetően ma is ugyanolyan szerkezetű adatokat gyűjtünk.

Jelentős fejlődés zajlott le viszont az adatok kezelésének módjában, és ezzel új adatforrás jelent meg: a vegetációs adatbázisok. A felvételek kézi átrendezése csak kisszámú felvétel kezelését tette lehetővé. Sok esetben a felvételek inkább csak a levont következtetések illusztrációi a cikkekben, és nem az információk elsődleges forrásai. A szintetizáló munkák is a fajok konstanciáin és nem az egyedi felvételeken alapultak. Az elemzésre használható számítógépes programok az 1960-as évektől indultak fejlődésnek, de csak az 1980-as évektől, a személyi számítógépek

megjelenésével váltak széles körben alkalmazott eszközzé a vegetációkutatásban. Kezdetben még az elemezhető adatsor mérete korlátozott volt, így előfordult, hogy az elemzés objektumai nem az egyes felvételek, hanem a korábban elkülönített szüntaxonok voltak. A személyi számítógépek kapacitásának gyors bővülésével azonban a feldolgozó programok egyre kevésbé korlátozták a feldolgozható adatok mennyiségét. Ma már egy több száz felvételből álló adatsor elemzésénél lényegesen nagyobb munka az adatbevitel, mint magának az elemzésnek a végrehajtása. Ezért érdemes az adatbevitelt csak egyszer elvégezni, a digitalizált adatokat adatbázisokban tárolni és az elemzéseknél elsősorban az adatbázisokban tárolt adatokból dolgozni. Az adatbázisokban tárolt felvételek lehetővé teszik, hogy több száz vagy esetleg több ezer felvétel alapján vizsgáljuk meg az aktuális problémát, esetleg távoli vidékek adatait is bevonva.

Schaminée és munkatársai (2009) becslése szerint Európában több mint 4,3 millió vegetációs felvétel készült, amelyből legalább 1,8 millió már elektronikus adatbázisban is hozzáférhető. A nemzeti adatbázisok kialakításához jelentős lökést adott a Nemzetközi Vegetációtudományi Társaság (IAVS) European Vegetation Survey (EVS) nevű munkacsoportjának megalakulása 1992-ben. A munkacsoport célja az európai vegetáció egységes szemléletű leírása.

Az EVS munkacsoport célja egy európai léptékű, modern szüntaxonómiai szintézis, de az adatbázisokban tárolt felvételeket nem csak ilyen célra lehet felhasználni, hanem más ökológiai kutatásoknak is hasznos eszközei lehetnek. Felhasználhatók például a klímaváltozás hatásának detektálására, fajok termőhely-preferenciáinak vizsgálatára és ehhez kapcsolódóan a nicheszélesség becslésére, a diszperzál-limitáltság, vagy társulások elözönölhetősége vizsgálatára, hosszú távú változások detektálására vagy a környezeti tényezők diverzitásra gyakorolt hatásának leírására.

A dolgozat szerkezete

A dolgozat először bemutatja a hazai nemzeti cönológiai adatbázis fejlesztésére tett eddigi erőfeszítéseket és az ezen a téren elért eddigi eredményeket (2. fejezet). Az adatbázis nem csak a dolgozatban szereplő módszerfejlesztések és esettanulmányok alapját jelenti, de kiépítésében személyesen is jelentős szerepem volt. A harmadik fejezet azt vizsgálja meg, hogy a döntő részben szubjektív mintavétellel gyűjtött adatok mennyire használhatók ökológiai hipotézisek statisztikai tesztelésére.

A számítógépek kapacitásának növekedése megteremtette a hardver feltételeket a nagy mennyiségű cönológiai adat feldolgozásához és a cönológiai adatbázisokban rendelkezésre is állnak az elemzéshez az adatok. A feldolgozott felvételek számának emelkedése azonban a feldolgozás során használható módszerek továbbfejlesztését is szükségessé tette. A dolgozat 4-8. fejezete a feldolgozó módszerek fejlesztésében elvégzett munkámat mutatja be.

A cönológiai adatok felhasználásának leggyakoribb formája a numerikus szüntaxonómiai elemzés, amelynek eszköze a numerikus klasszifikáció. A vegetációs adatok mindig zajosak, így a numerikus klasszifikáció hatékonyságát erősen növelheti az előzetes zajszűrés, amelyre korábban másokhoz hasonlóan a metrikus ordinációt javasoltuk (Botta-Dukát et al. 2005), amelynek hatékonyságát elsőként teszteltem (4. fejezet). Több száz – esetleg több ezer – felvétel klasszifikációjakor már nincs lehetőség az eredmények részletes áttekintésére, a csoportok értelmezése a karakterfajokon alapul, amelynek objektív eszköze a fidelitás kiszámítása. Az 5. fejezet áttekinti és értékeli az erre a célra javasolt módszereket, amelyek egy részét szerzőtársaimmal dolgoztuk ki (Botta-Dukát & Borhidi 1999, Chytrý et al. 2002). A numerikus klasszifikáció számos módszere (beleértve a távolságfüggvényeket és az összevonási algoritmust)

közötti választás gyakran a kutató döntésén múlik. Bár szemben a szubjektív csoportosítással ezek a döntések jól dokumentáltak, így az eredmények reprodukálhatók, szükség lenne a módszerválasztást segítő objektív kritériumokra. A 6. fejezet ezeket a kritériumokat tekinti át, illetve alkalmazhatóságukat vizsgálja a vegetációs adatok elemzésében. Ugyanezek a módszerek egy másik központi kérdés, az optimális csoportszám megállapítására is használhatók.

A következő két fejezet a cönológiai felvételek fajelőfordulási-adatainak két ökológiai alkalmazásával – fajkészlet becslése (7. fejezet) és nicheszélesség vizsgálata (8. fejezet) – kapcsolatos módszertani kérdéseket vizsgál. Végül az utolsó két fejezet két esettanulmányt mutat be: a közép-európai mocsárrétek változatosságának vizsgálatát (9. fejezet) és a hazai fajok cönológiai adatbázisból becsült nicheszélessége és areájuk mérete közötti kapcsolat elemzését (10. fejezet).

CoenoDAT Referencia Adatbázis

Habár a cönológiának nagy tradíciója van Magyarországon, az 1980-as, 90-es években lemaradásba kerültünk ezen a területen. Miközben Európa szerte elindultak a cönológiai adatbázis-építő programok, Magyarországon 2002-ig nem volt működő cönológiai adatbázis.

2002-ben a „Magyarország természetes növényzeti örökségének felmérése és összehasonlító értékelése” program keretében indulhatott meg a CoenoDAT Referencia Adatbázis kiépítése. Az adatbázis létrehozása során jelentős számú új felvétel elkészítését terveztük, amihez részletes módszertani útmutatót és adatlapot készítettünk. A tervezett számú új felvétel ugyan nem készült el, de a tervezettnél jóval több felvételt digitalizáltunk a szakirodalomból, így az adatbázis a tervezett 7200 helyett 9200 felvételt tartalmaz, amelyek együttesen többé-kevésbé reprezentálják a hazai vegetációt,

és bár az adatbázis további bővítését tervezzük, már így is jól használható országos elemzésekre.

A cönológiai adatbázisok felhasználásának lehetőségei és korlátai

A vegetációs adatbázisok nem kizárólag, de jelenleg döntő többségében preferenciális – Nyugat- és Közép-Európában a Braun-Blanquet cönológiai iskola szabályai szerint készült – felvételekből állnak. Felmerülhet a kétség, hogy alkalmasak-e ezek az adatbázisok a valós vegetáció reprezentálására, illetve, hogy lehetnek-e ezek a nem random mintavétellel gyűjtött adatok statisztikai elemzés tárgyai.

Az első kérdésre a válasz, hogy részben igen, és még ha nem is tökéletes, de a rendelkezésünkre álló legjobb eszközök a durva léptékű vegetációs vizsgálatokban. A Braun-Blanquet iskola felfogása szerint a vegetáció jól elkülönülő egységekből áll, alapegysége az asszociáció a fajjal analóg módon leírható és tipizálható természetes egység. Ezért, hasonlóan az idiotaxonómiához, ahol a faj leírása a típusegyeden alapul, a növényszociológusok is a tipikus állományok felvételezésére törekedtek. Emiatt hiányoznak, vagy legalábbis alulreprezentáltak az „atipikus”, „jellegtelen”, „átmeneti” vegetációtípusok az adatbázisban. Másfelől viszont, mivel a preferenciális mintavételezés során a felmérő igyekszik a területen előforduló minden típust megmintázni ("catch as much variation as possible"), a preferenciális minta több ritka típust tartalmaz és így jobban reprezentálja a vegetáció változatosságát, mint az azonos méretű random minta.

A második kérdést illetően, hogy lehet-e statisztikai tesztekben a preferenciális mintavétellel gyűjtött adatokat használni, megoszlanak a vélemények. Egyesek szerint a válasz egyértelműen nem, mert nem teljesül az adatok függetlensége. Mások viszont sokkal megengedőbbek ebben a kérdésben, bár elismerik, hogy szigorú

matematikai értelemben a függetlenség nem teljesül. vitatják, hogy ebből az következik, hogy egyáltalán nem lehet statisztikai teszteket alkalmazni. Magam is az utóbbi állásponttal értek egyet.

Véleményem szerint, habár a cönológiában alkalmazott mintavétel nem teljesíti a statisztikai elemzés formális követelményeit, a cönológiai felvételek használhatók statisztikai elemzésekben, ha a szokványos (informális) feltételek teljesülnek. Sajnos a biológusoknak szóló statisztikai kézikönyvek csak a formális kritériumokat tárgyalják, a szokványos kritériumokról sem ezekben, sem a cönológiai szakirodalomban nem olvashatunk.

A legfontosabb ilyen kritériumok az interpretálás szabályai. Mi az alapsokaság, amit a cönológiai felvételekből álló minta reprezentál, és amire az eredmények vonatkoznak? Úgy gondolom ez nem egy egyszerű kérdés, amire létezik általános érvényű válasz. Ha a kutató adatbázisból vagy irodalomból származó felvételeket használ, mindig alaposan át kell gondolnia mik lehettek a felvételek készítőjének a preferenciái. Ezt általában könnyen meg tudjuk tenni, de a következtetéseinket nem tudjuk ellenőrizni.

A dolgozatban bemutatok egy olyan szerzőtársaimmal készített esettanulmányt, amelyben lehetőség volt erre az ellenőrzésre. Az esettanulmányban két helyszínen (Csévharaszt és Fülöpháza) hasonlítottuk össze a nyílt homoki gyepben készült random és preferenciális felvételeket. A felvételek fajösszetétele csak Fülöpházán különbözött szignifikánsan, és a különbséget csak két – a mintavétel időszakában nehezen észlelhető – egyéves faj okozta. A származtatott jellemzők tekintetében is meglepően kevés szignifikáns eltérést találtunk. Kiemelném, hogy a várttal ellentétben a fajszám a preferenciális felvételekben nem magasabb, hanem alacsonyabb volt, ami a vizsgált objektum tulajdonságaival magyarázható.

Módszerfejlesztések

A dolgozatban tárgyalt módszerfejlesztések részben új módszerek kidolgozását jelenti, részben az új vagy a szakirodalomból már ismert módszerek összehasonlítását, hatékonyságuk tesztelését. A tesztelés során ahol az lehetséges volt ismert szerkezetű, szimulált adatsorokat használtam, így az előre ismert várt eredményhez hasonlítottam az egyes módszerekkel kapott eredményeket.

Zajszűrés

A klasszifikációt megelőző zajszűrésre a metrikus sokdimenziós skálázás alkalmazását javasoltam: a klasszifikáció bemenő adata ebben az esetben a fontos (szignifikáns) tengelyek koordinátái.

A metrikus sokdimenziós skálázás lehetővé teszi az ökológiai szempontból leginkább releváns távolságfüggvény alkalmazását (bár nem metrikus különbözőségfüggvények esetén a negatív sajátértékek kiküszöbölésére korrekcióra lehet szükség). Mivel az ordináció a felvételeket egy euklideszi térbe képezi le, a klasszifikáció során már minden esetben euklideszi távolságot használunk, így a csak ezzel a távolságfüggvénnyel kompatibilis összevonási algoritmusok (pl. Ward módszer) is minden esetben használhatók.

A szignifikáns tengelyek számának megállapítására megfelelő módszernek bizonyult a törött pálca modellel való összehasonlítás, így nincs szükség számításigényes randomizációs tesztre, amelynek csak a bináris adatok esetén van jól kidolgozott módszere.

A metrikus sokdimenziós skálázás segítségével végzett zajszűrés mind szimulált, mind terepi adatok esetén növelte a klasszifikációk átlagos jóságát, és csökkentette a klasszifikációs módszerek jósága közötti különbségeket. Szimulált adatok elemzése alapján elmondható, hogy nincs olyan klasszifikációs algoritmus, amely minden szituációban felülmúlná a többit: teljesítményük a csoportok méretétől és alakjától függ. Ha az adataink ezen tulajdonságait nem

ismerjük (és a klasszifikáció előtt erről legfeljebb sejtéseink vannak, de általában még az sem), akkor az összevonási algoritmus kiválasztása az elemző szubjektív döntése. A zajszűrés ennek a szubjektív döntésnek a súlyát csökkenti.

Fidelitás

A vegetációs adatok bármilyen csoportosítása esetén feltehető az kérdés, hogy mely fajokat találunk egy csoportban a véletlenül vártnál gyakrabban, nagyobb egyedszámmal, nagyobb borítással, vagy éppen mely fajok kerülik el a csoportot. A véletlen alapján várt gyakoriságtól való eltérést méri a fajok fidelitása.

A fidelitáshoz szorosan kötődő karakterfaj-koncepció központi szerepet játszik a Braun-Blanquet-i cönológiai iskola módszertanában, a fidelitás alkalmazásának elterjedését azonban gátolta, hogy az objektív mérésére vonatkozó régebbi javaslatok nem váltak széles körben ismerté. A fidelitás indexeket bemutató és összehasonlító cikkünk és a Juice program nyomán mára elterjedt a fidelitás alkalmazása a klasszifikációval kapott csoportok jelentésének feltárására.

Az összehasonlító vizsgálatok alapján az egy faj és egy felvételcsoport kölcsönös fidelitását mérő függvények általában nagyon hasonló eredményt adnak, kivéve az IndVal, amely nem jelzi a negatív fidelitásokat. Az IndVal Podani és Csányi által javasolt módosított változata jelzi a negatív fidelitásokat is, de nem a kölcsönös hűséget méri, hanem faj hűségét a felvételcsoporthoz.

Ha a fajok fidelitását minden egyes felvételcsoporthoz külön-külön teszteljük, az ahhoz hasonló, mint ha az egytényezős ANOVA helyett kétmintás t-próbákat csinálnánk minden párosításban: mindkét esetben megnő az elsőfajú hiba elkövetésének valószínűsége. A másik probléma a felvételcsoportok egyenkénti vizsgálatánál, hogy a generalistább fajok esetleg nem fidélisek egyetlen csoporthoz sem, de fidélisek a csoportok egy részéhez, ha

azokat összevonjuk.

A problémát csak az utóbbi években ismerték fel a vegetációkutatók, és a megoldására eddig született javaslatok azon alapulnak, hogy ha a felvételcsoportokat összevonva két csoportot alakítunk ki, akkor továbbra is használhatók a páros összehasonlításokra kidolgozott módszerek.

Az eddig javasolt megoldások hátránya, hogy vagy egy merev hierarchia (pl. szüntaxonómiai rendszer) szerint történik az összevonás, és ezért a hierarchikus rendszerhez nem illeszkedő fajok fidelitását alulbecsli, vagy minden lehetséges összevonást ki kell próbálni, de ebben az esetben a csoportok számának növekedésével exponenciálisan nő a számítások elvégzésének időigénye. Ráadásul előfordulhat, hogy a faj gyakorisága a csoportban kisebb, mint a teljes adatsorban, mégis úgy tűnik mintha preferálná a csoportot.

Ezeknek a problémáknak a megoldására a nagyobb méretű (2 x n-es) kontingenciatáblákra is kiszámítható statisztikák (Khi-négyzet, G, F, általánosított Fisher egzakt teszt) alkalmazását javasoltam a fidelitás erősségének mérésére. Ennek kapcsán kimutattam, hogy a Φ-koefficiens és a lineáris korrelációs koefficiens (point-biserial correlation) közötti kapcsolat általánosítható több csoport esetére is: a bináris adatokra kiszámolt általános lineáris modell determinációs koefficiensének (R^2) a négyzetgyöke egyenlő a Φ-koefficienssel. A javasolt statisztikák csak a fidelitás erősségét mérik, de nem adnak információt arról, hogy a faj mely csoportokat preferálja és melyeket kerüli el. Ennek megállapítására bináris adatok esetén a Freeman-Tukey eltérések (’deviates’) alkalmazását javaslom.

A fidelitás vizsgálatokhoz kapcsolódóan feltehetjük a következő kérdéseket is: A felvételek több alternatív csoportosítása közül melyik alapján mennyire jósolható meg a faj előfordulása (abundanciája)? A faj jelenléte (tömegessége) alapján mennyire jósolható meg, hogy a felvétel melyik csoportba tartozik? Ezekre a kérdésekre a fajok

válogatóképességének (separation power) a mérésével adhatunk választ. Egy faj válogatóképessége annál nagyobb, minél inkább megjósolható a faj jelenléte a felvétel csoportja alapján, illetve minél inkább megjósolható a felvételcsoport a faj jelenlétéből/hiányából. Az átfogó mérőszámok (χ^2-, G-statisztika, R^2) alkalmasak a fajok válogatóképességének mérésére, de értékük függ a csoportok számától. Az elméleti eloszlásokon alapuló standardizálások általában nem adnak megfelelő eredményt, a véletlen előfordulás esetén várható értéket és szórást megfelelő számú random minta alapján kell becsülni.

A klasszifikációk értékelése

A numerikus klasszifikáció során a kutatónak számos döntést kell hoznia – Milyen adat-transzformációt, távolságfüggvényt, klasszifikációs algoritmust használjon? Hány csoportot különböztessen meg? Minden csoportot értelmezzen, vagy vannak műtermék-csoportok is? – amelyek befolyásolják a kapott eredményeket. A lehetőségek száma óriási, még akkor is, ha a kézikönyvekben szereplő elméleti megfontolások alapján a konkrét esetben szűkíthető a szóba jöhető módszerek köre. Jó lenne valamilyen objektív módszer, ami segít az eredmények értékelésében, a „legjobb” módszer kiválasztásában. Bár történtek próbálkozások a módszerek általános érvényű összehasonlítására, ezek alapján csak az általában jó és az általában kevésbé jó módszerek különíthetők el. Nincs egyetlen, minden adatsorra optimális módszer, ezért érdemes mindig a konkrét adatsorra legjobb módszert megkeresni.

A dolgozatban áttekintést adtam a klasszifikációk értékelése során használható módszerekről, 8 csoportba sorolva azokat aszerint, hogy milyen szempontból vizsgálják a csoportosítás „jóságát”:

1. a csoportok mennyire tükrözik a távolságmátrixban levő információt

2. a csoportok tömörsége (compactedness), összekötöttsége (connectedness) és elkülönülése (separation)
3. az eredmények robosztussága
    a. az elemzés során hozott döntésekkel szemben
    b. a figyelembevett változókkal szemben
4. az eredmények stabilitása az adatok kismértékű megváltozásakor
5. ismétlődő csoportok ugyanabból az alapsokaságból vett párhuzamos minták esetén (repetitivitás)
6. a csoportok értelmezhetősége
7. a csoportok prediktív ereje az osztályozás során nem használt változókra
8. a csoportosítás összehasonlítása a random adatok csoportjaival

A fenti csoportosítás új, bár egyes elemei mások korábbi munkáiban is megjelennek, sokkal gyakoribb, hogy a vizsgált szempontot nem definiálják elég világosan. Különösen az eredmények robosztussága, stabilitása és repetitivitása terén sok a különböző szempontok összemosásából adódó félreértés. Az áttekintés végén tárgyalom a módszereknek a szakirodalomban előforduló egyéb szempontok szerinti (pl. external/internal, geometriai/nem geometriai, lokális/globális) csoportosításait is.

A nagy adatbázisok elemzésekor, kisebb számításigényük miatt, érdemes az egyszerű – randomizációt vagy újramintavételezést nem igénylő – indexeket választani. A dolgozatban 23 ilyen index tulajdonságait értékelem szimulált adatok alapján. Hasonló vizsgálat ökológiai adatokkal tudomásom szerint eddig nem történt, a más tudományterületeken elvégzett – nem túl nagy számú – vizsgálat eredményeinek adaptálhatósága pedig kétséges. Ennek egyik oka a vegetációs adatoktól eltérő szerkezetű adatok (például szférikus csoportok euklideszi térben) használata. A másik ok, hogy a

kifejezetten a vegetációs adatokra kifejlesztett módszerek, ezekben a más területeken dolgozó kutatók – például pszichológusok – által készített összehasonlításokban érthető módon nem szerepelnek.

A vizsgálatban a távolságmátrixban rejlő információ torzulását mérő indexek közül a Baker & Hubert és McClaine & Rao, csoportok tömörségét és elválását vizsgáló indexek közül a Popma-index súlyozott verziója, a csoportok értelmezhetőségét számszerűsítő indexek közül a relatív divergencia, az átlagos nicheszélesség és a karakterfajok száma (OptimClass) adta a legjobb eredményt. Míg a korábbi vizsgálatokban a pont-biszeriális korreláció és az átlagos sziluett a legjobb indexek között volt, addig ebben a vizsgálatban kifejezetten rossz eredményt adtak.

Az eredmények alapján a következő ajánlások fogalmazhatók meg:

1. A bemutatott összehasonlításokban jó eredményt elért indexek bátran alkalmazhatók, de az itt sikertelenül szereplők sem feltétlenül használhatatlanok, különösen ha élesebben elváló fajkészletű csoportok várhatók.
2. Az indexek ugyan nem annak tesztelésére szolgálnak, hogy vannak-e egyáltalán csoportok, de trendjük (esetenként értékük) jelzi az adatstruktúra hiányát. Érdemes ezekre a jelekre odafigyelni, és nem egyszerűen az index hibás viselkedésének tekinteni őket.
3. Érdemes ugyanarra a problémára több indexet is kipróbálni, lehetőleg a három nagy csoport mindegyikéből.
4. Ha távolságfüggvények összehasonlítása a cél, csak nem-geometriai indexek használhatók.
5. Amennyiben a rendelkezésre álló számítási kapacitás és az adatsor mérete megengedi, érdemes az egyszerű indexek mellett, a csoportosítás repetitivitását és stabilitását is vizsgálni.
6. További hasonló, mesterséges adatokat használó

összehasonlításokra lenne szükség, ahol a csoportok elkülönülése nagyobb, elhelyezkedésük a gradiens mentén nem egyenletes és esetleg nem csak egy környezeti gradiens van, hogy feltárjuk, hogyan befolyásolják ezek a beállítások a kapott eredményeket.

Bár az eredmények repetitivitása, mint elvárás már az 1980-as években megjelent, viszonylag ritkán alkalmazzák a csoportosítások értékelésére. Pedig talán az egyik legfontosabb szempont, hiszen a cél általános érvényű, nem csak a vizsgált adatsorra érvényes csoportosítások felállítása. McIntyre and Blashfield (1980) megfogalmazása szerint: "*If a cluster solution is repeatedly discovered across different samples from the same general population, it is plausible to conclude that this solution has some generality. Certainly, a cluster solution which is not stable is unlikely to have general utility*." A repetitivitást vizsgáló módszerek elterjedésének egyik gátja a párhuzamos minták hiánya volt, de a nagy adatbázisok esetén ez már nem jelent problémát. A másik korlát a megfelelő algoritmusok hiánya. A legtöbb módszer a következő sémát követi: az egyik mintát klasszifikáljuk, majd valamilyen osztályozási szabály alapján a második minta elemeit is besoroljuk ezekbe a csoportokba. Ezután a második mintát is klasszifikáljuk ugyanazzal a módszerrel, mint az első mintát. Minél nagyobb a második minta elemeinek kétféle csoportosítása közötti hasonlóság, annál jobb a klasszifikáció. A módszer gyenge pontja, hogy nagyon sokféle szabály közül választhatunk, és sajnos a választásunk nagymértékben befolyásolhatja az eredményt. Ezért egy olyan új módszert dolgoztam ki, amely a csoportok jelentésén alapul, ezért nem alkalmaz klasszifikációs szabályt. Míg a korábbi módszerek csak a teljes partíció értékelésére használhatók, ez az új módszer az egyes csoportokat külön-külön értékeli, így kiválaszthatók azok a repetitív csoportok,

amelyeket érdemes értelmezni. A módszert két esettanulmányban is sikerrel alkalmaztuk a repetitív, azaz értelmezendő csoportok kiválasztására, ezzel elkerülve a klasszifikáció eredményeinek túlinterpretálását.

Fajkészlet becslése

Az ökológiai szakirodalomban általánosan elfogadott tény, hogy a közösségek fajösszetételét részben a rendelkezésre álló fajkészlet, és az abból való válogatás szabályai, a társulási szabályok (assembly rules) együttesen határozzák meg. A fajkészlet hatására vonatkozó vizsgálatok egyik korlátja a fajkészlet becslésére szolgáló általánosan használható, objektív módszerek hiánya. Számos módszer van, amely az azonos közösségben készült felvételek alapján becslést ad a fajkészlet méretére. Ezek a módszerek azonban nem mondanak semmit arról, hogy melyek azok a fajai a fajkészletnek, amelyeket nem sikerült egyetlen felvételben sem megfigyelnünk, csak ezeknek a fajoknak a számára adnak becslést.

A probléma megoldására Ewald (2002) a Beals simítás alkalmazását javasolta. Ez az eredetileg a sokváltozós elemzések előtti adattranszformációra javasolt módszer az eredeti bináris adatokat a fajok együttes előfordulási valószínűségei alapján számolt 0 és 1 közötti értékekkel helyettesíti, amelyek azt fejezik ki, hogy a termőhely mennyire alkalmas a faj számára. A módszert viszonylag ritkán alkalmazzák, aminek egyik oka lehet, hogy a szakirodalomban nem található egyértelmű megoldás arra, hogy a folytonos értékeket hogyan alakítsuk át bináris skálává (1 = a fajkészlet eleme, 0 = nem tartozik bele a fajkészletbe).

A CoenoDat adatbázison elvégzett elemzés megerősítette a szakirodalomban már korábban is megfogalmazott sejtést, hogy fajonként eltérő határértéket kell alkalmazni a bináris skálára való átalakításkor. Ha az a célunk, hogy minden előforduló fajt a fajkészlet részének tekintsünk, akkor a faj előfordulásaihoz tartozó simított

értékek minimumát kellene határértéknek választani. Azonban esetenként egy vagy néhány kilógó érték miatt a minimum nagyon alacsony lehet. Ezeknek a kilógó értékeknek a hátterében állhat adatbázis-hiba, vagy olyan szituáció, amikor a korábbi állapot alkalmas volt a fajnak, a jelenlegi már nem, de a faj néhány egyede még képes túlélni. Ha a minimumot használjuk határértéknek, akkor a kilógó értékekkel rendelkező fajoknál nagyon magas lehet a másodfajú hiba valószínűsége. Ezért célszerű a kilógó értékek kizárása utáni minimumot használni.

Niche szélesség

Habár a “specialista” és „generalista” fajok széles körben használt fogalmak, mégsincs a „specialistaság”-nak elfogadott mérőszáma. Emiatt a fajok kategorizálása általában szubjektív módon történik, jellemzően két kategóriát használva: specialisták és generalisták. Egy folytonos skálájú, objektív mérőszám sokkal részletesebb képet adna a fajok viselkedéséről, és az így nyert információkat számos ökológiai kérdés vizsgálatában fel lehetne használni.

Ha rendelkezünk környezeti adatokkal azokról a helyekről ahol a faj előfordul, megbecsülhetjük a nicheszélességet. Ez azonban nem használható a „specialistaság” egyértelmű mérőszámaként, mert egy faj niche lehet keskeny az egyik környezeti tényező tekintetében, de széles, ha egy másik változót vizsgálunk.

A vegetációs adatbázisokban rejlő információk kiaknázása megoldást kínál erre a problémára. A Fridley és munkatársai (2007) által javasolt módszer alapgondolata egyszerű: a generalista faj sokféle környezetben, változatos vegetációtípusokban fordul elő, míg a specialista faj csak szűk környezeti tartományban él, ezért kevésbé változatosak azok a felvételek, amelyekben megtaláljuk. Vagyis a faj „specialistaságának” megméréséhez azt kell megvizsgálni, hogy mennyire különböző összetételűek azok a felvételek, amelyekben

előfordul, vagyis mekkora a béta diverzitás. Az így kapott értéket egy minden releváns környezeti tényezőt együttesen figyelembe vevő (realizált) nicheszélességnek is tekinthetjük, annak ellenére, hogy a számítások során nem használunk környezeti adatokat.

Az eddig a nicheszélesség mérésére javasolt béta-diverzitás indexek egyike sem ad torzítatlan és robosztus becslést. A torzítás ebben az esetben azt jelenti, hogy a nicheszélesség függ a fajkészlet méretétől. Ez a hatás úgy küszöbölhető ki teljesen, hogy a megfigyelt fajlisták helyett a lokális fajkészletet használjuk a Whittaker-féle béta-diverzitás kiszámításakor. Ebben az esetben az alfa-diverzitás a lokális fajkészlet mérete, míg a gamma-divertitás azoknak a fajoknak a száma amelyek a környezeti változók vizsgált tartományának legalább egy pontján a fajkészlet részei. A gamma-diverzitás multiplikatív felbontása, azaz a Whittaker-féle béta-diverzitás használata, garantálja, hogy a béta diverzitás minden esetben független lesz az alfa-diverzitástól, vagyis a lokális fajkészlet méretétől. Az így kiszámolt nicheszélesség további előnye, hogy a kapott értékek könnyen értelmezhetők. Az elméleti minimum 1, ami azt jelenti, hogy a fajkészlet azonos a vizsgált faj minden előfordulási helyén, vagyis a faj extrém specialista. Ha a faj nicheszélessége 2, akkor az azt jelenti, hogy épp annyira generalista, mint egy olyan faj, amely két olyan termőhelyen képes előfordulni, amelyek fajkészlete – a vizsgált fajtól eltekintve – nem fed át.

A robosztus becslés azt jelenti, hogy az adatok kis mértékű megváltoztatása az eredményekben is csak kis mértékű változást okoz. Sajnos a béta-diverzitás becslési módszerek érzékenyek a kilógó (outlier) fajkészletű felvételekre, ezért a robosztus becslés érdekében ezeket ki kell zárni az elemzésből.

Esettanulmányok

A cönológiai adatbázisok alkalmazásának lehetőségeit két

esettanulmányban mutatom be: az első a nedves rétek (*Molinietalia* rend) változatosságát vizsgálja, míg a második a nicheszélesség és az area mérete közötti összefüggést elemzi.

A közép-európai nedves rétek klasszifikációja

A European Vegetation Survey nevű nemzetközi munkacsoport fő célkitűzése egy európai léptékű szintézis megalkotása, amelynek első lépése az országos léptékű szintézisek összeszerkesztése volt. Pusztán az irodalmak feldolgozásával azonban nem lehet teljesen feloldani a különböző rendszerek közötti ellenmondásokat és különbségeket. Ehhez nagyobb területekről származó eredeti felvételek újraelemzésére is szükség van.

A nehezen összeegyeztethető nemzeti vegetációosztályozási rendszerekre jó példa a közép-európai síkságok és dombvidékek nedves rétjeinek osztályozása. Míg a szubóceánikus típusú nedves rétek osztályozása többé-kevésbé stabil, és olyan nemzetközileg elfogadott asszociációcsoportokat használ, mint például *Arrhenatherion elatioris, Polygono-Trisetion, Calthion palustris* és *Molinion coeruleae*, a szubóceánikus-szubkontinentális elterjedésű típusok esetén nem jött létre ilyen konszenzus. Ezeket a réteket általában a következő asszociációcsoportokba sorolják: *Agrostion albae, Alopecurion pratensis, Cnidion venosi, Deschampsion cespitosae* és *Veronico longifoliae-Lysimachion vulgaris*. Még ha a használt név meg is egyezik a különböző szerzőknél, az asszociációcsoportok pontos jelentése országonként és szerzőnként változó.

Részben nemzeti adatbázisokból, részben az irodalomból gyűjtöttük össze az alföldi nedves rétek felvételeit egy Közép-Európán keresztül húzódó északnyugat-délkeleti gradiens mentén. Azokat a Csehországban, Kelet-Ausztriában, Szlovákiában, Magyarországon és Észak-Horvátországban, 350 m-es tengerszint feletti magasság alatti

területeken készült felvételeket választottuk ki, amelyeket a készítőik a *Molinietalia* rendbe soroltak. Az egyes területek túlreprezentáltságából adódó esetleges problémák kiküszöbölésére rétegzett újramintavételezést alkalmaztunk.

A felvételek távolságait relatív Manhattan távolsággal mértük. Zajszűrést követően Ward módszerrel klasszifikáltuk a felvételeket. Az optimális csoportszám megállapításához kiszámítottuk a csoportok élességét (crispness). A csoportok jellemzéséhez az u_{hyp} függvénnyel megállapított fidelitás értékeket használtuk.

Az esettanumány legfontosabb megállapításai a következők:

- A klasszifikáció élességének (crispness) alapján az optimális csoportszám három. Ez a három csoport a korábbi szüntaxonómiai rendszerek asszociációcsoportjainak feleltethetők meg.
- További csúcsok figyelhetők meg öt és kilenc csoportnál, amelyek a finomabb felosztások optimális csoportszámai. Ezek a felvételcsoportok a szüntaxonómiai rendszerben az asszociáció-alcsoportoknak (suballiance), illetve az asszociációknak feleltethetők meg.
- A három csoport közül kettő (2. és 3. csoport) egyértelműen megfeleltethető a szüntaxonómiában nemzetközileg elfogadott asszociáció-csoportoknak: *Calthion* (2.csoport), illetve *Molinion* (3.csoport)
- Az első csoportba tartozó rétek jellemzően a nagy folyók alföldi árterein fordulnak elő. Ezeket a területeket ugyan rendszeresen elönti a víz, de szemben a *Calthion* rétek termőhelyével nyáron a talajuk kiszárad, és a kontinentális klíma miatt a légkör páratartalma is alacsony. A termőhelyükön a talajvíz dinamikája hasonló a kékperjés rétekéhez, de különböznek azoktól a talaj magasabb tápanyagtartalmában. Ezeket a gyepeket a különböző szerzők különböző asszociációcsoportokba sorolták. Habár az a

csoport heterogénebb, mint a másik kettő, belső struktúrája egyiket sem támasztja alá a korábbi szakirodalomban megjelenő felosztásoknak, ezért azt javasoltuk, hogy tekintsük a csoportot egyetlen asszociációcsoportnak, amire a legrégebbi érvényes nevet – *Deschampsion cespitosae* Horvatić 1930 – kell használni.

- Ha kilenc csoportot különítünk el, akkor azok nagyjából megfeleltethetők a hagyományos cönológiai leírások asszociációinak:
    - 1.1 csoport: Egész évben jó vízellátottságú területek (általában árterek) mocsárrétjei
    - 1.2. csoport: Tavasszal elárasztott, de nyáron erősen kiszáradó területek mocsárrétjei
    - 1.3. csoport: Átmenet a kaszálórétek (Arrhenatherion) felé szubóceánikus klímájú területen
    - 1.4. csoport: Átmenet a kaszálórétek (Arrhenatherion) felé szubkontinentális klímájú területen
    - 2.1. csoport: Kaszált Calthion gyepek
    - 2.2. csoport: Filipendula ulmaria dominálta felhagyott Calthion gyepek
    - 2.3. csoport: Scirpus sylvaticus dominálta felhagyott Calthion gyepek
    - 3.1 és 3.2 csoport Molinion gyepek
- A *Molinion* gyepek látszólag a magas éves középhőmérsékletű és magas éves hőingású helyekhez kötődnek , azonban ez a mintázat valószínűleg műtermék, ami annak következtében alakult ki, hogy kékperjés rétek ritkák Csehország sík- és dombvidéki részén, mert ott inkább nagyobb magasságokban fordulnak elő, gyakoriak viszont a kontinentálisabb klímájú Magyarországon és Dél-Szlovákiában. A valóságban azonban ezeken a helyeken is általában a makroklímánál hűvösebb mezoklímájú helyeken fordulnak elő. A *Calthion* gyepek a hűvösebb, csapadékosabb,

kiegyenlítettebb (szub)óceáni klímához kötődnek. A *Deschampsion* gyepek viszont Közép-Európa aridabb területeire jellemzők: talajuk tavasszal az áradásoknak köszönhetően vizes, de nyáron kiszárad.

- A csoportok klímáját Kruskal-Wallis teszttel összehasonlítva megállapítható, hogy a de Martonne féle humiditás index jobban jellemzi a vízellátottságot, mint a csapadék mennyisége. A *Calthion* csoportok szignifikánsan humidabb területeken fordulnak elő, mint a *Deschampsion* asszociációcsoportba tartozó csoportok. Az éves középhőmérséklet tekintetében két nagy csoport különböztethető meg: a *Calthion* gyepek és a szubóceánikus területek *Deschampsion* gyepjei a hűvösebb, míg a többi *Deschampsion* csoport és az alföldi *Molinion* gyepek a melegebb területekre jellemzőek. Az éves hőingás is az első csoportban alacsonyabb és a másodikban magasabb, ami azt jelenti, hogy a szubatlantikus *Calthion* asszociációcsoportnak és a kontinentális *Deschampsion* asszociációcsoport leginkább szubóceánikus alcsoportjának termőhelye nem különbözik a kontinentalitás hőmérsékleti komponense szempontjából, de eltér a humiditásuk.

Nicheszélesség és az area méret kapcsolata

A nicheszélesség vizsgálatára kidolgozott egyszerű és objektív módszer lehetővé teszi, hogy eddig nehezen tesztelhető hipotéziseket is megvizsgáljunk. Ilyen az a hipotézis is, hogy a generalista fajok, mivel kevésbé válogatnak a termőhelyben, nagyobb areájúak. Számos a hipotézist bizonyító példát lehet hozni, de találhatunk ellenpéldákat is. Vagyis a feltételezett összefüggés csak tendencia jellegű, statisztikusan érvényesülő szabály, aminek a meglétét a magyar flóra fajain teszteltem.

Ha a hipotézis igaz, akkor is várható, hogy vannak kis áreájú generalista és nagy areájú specialista fajok. Érdekes,kérdés hogy vajon ezek a fajok egyformán gyakoriak minden élőhelyen, vagy

egyes közösségekben gyakrabban, másutt ritkábban fordulnak elő?

Bár az egyes areaméret kategóriákon belül erősen szórnak a nicheszélesség értékek, az elvégzett elemzés igazolta a várt pozitív összefüggést a nicheszélesség és az areaméret között (Kendall tau = 0.278, $p<0.1\%$). Ez jól magyarázható azzal, hogy ezek a fajok könnyebben alkalmazkodnak a különböző környezeti körülményekhez, így nagy területen találhatnak a számukra megfelelő élőhelyet. Egy specialista faj is lehet azonban tág elterjedésű, ha a számára alkalmas speciális élőhely elég nagy földrajzi területen előfordul. Ez a helyzet például a mocsári és lápi fajoknál, amelyek csak megfelelő vízellátás, trofitás és pH viszonyok mellett fordulnak elő, de ezek a termőhelyek az északi mérsékelt övben általánosan elterjedtek. A rendszeres emberi bolygatáshoz alkalmazkodott gyomok életfeltételeit világszerte megteremti az emberi tevékenység. A leginkább meglepő az üde erdei fajok magas száma a nagy areájú specialisták között. Az ide fajok többsége (a 25 fajból 18) eurázsiai flóraelem. Ebben az esetben a specializálódást a zárt erdők fényszegény körülményeihez való alkalmazkodás jelenti és ezek a körülmények Eurázsia nagy területein megtalálhatók, mégha az erdők uralkodó fafaja területenként eltérő is.

Míg a tág areájú specialisták élőhelypreferenciáit elméleti megfontolások alapján – legalább részben – előzetesen is megjósolhattuk, sokkal nehezebb lenne előrejelezni a kis areájú, generalista szárazgyepi fajok magas számát. A jelenség hátterében valószínűleg az áll, hogy a Kárpát-medencében a szárazgyepek nagy kiterjedésben és változatos termőhelyi körülmények (pl. különböző alapkőzetek) között fordulnak elő, ami lehetővé teszi, hogy sok generalista szárazgyepi fajunk legyen. Ezek között magas a pontusi és szubmediterrán fajok aránya. Érdemes lenne filogeográfiai módszerekkel megvizsgálni ezeknek a fajoknak az elterjedéstörténetét: valószínűleg a balkáni és Fekete-tenger

környéki jégkorszaki refúgiumokból rajzottak szét és a refúgiumokban lezajló speciáció is feltételezhető.

Az eredmények földrajzi érvényességi területe is érdekes kérdés. A specialista fajok kisebb várható areamérete valószínűleg általános érvényű. Szintén nagy területen érvényesnek gondolom a gyomok és a mocsári, lápi fajok magas arányát nagy areájú specialisták között. A szárazgyepi fajok nagy száma a kis areájú generalisták között viszont valószínűleg csak a Kárpát-medencére és a szomszédos területekre érvényesek, és más eredményeket kapnánk, ha megismételnénk a vizsgálatokat a nyugat-európai lomboserdő-zónában.

Összegzés és kitekintés

A cönológiai adatbázisok a jövőben várhatóan fontos eszközei lesznek az ökológiai kutatásoknak. Ehhez azonban ki kell dolgozni a megfelelő elemzőmódszereket is. A dolgozat nagyobbik része az ezen a téren elért eredményeimről szól. A fejlesztések egy része a korábban, kisebb adatsorokon is használt numerikus klasszifikációs eljárásokhoz kapcsolódik. Ezeket azt tette szükségessé, hogy ha sok felvételt elemzünk, akkor nehéz az eredmények minden részletét áttekinteni, szükség van az eredmények értékelését segítő módszerekre. Más módszerek olyan kérdésekhez kapcsolódnak (fajkészlet és nicheszélesség becslése), amelyek numerikus vizsgálatára nagy adatbázisok hiányában korábban nem volt lehetőség.

A módszerfejlesztések nem érnek véget az elméletileg jól működő módszerek kitalálásával. Azok alkalmasságát tesztelni kell és ha ugyanarra a problémára több módszer is létezik érdemes azok hatékonyságát is összehasonlítani. A dolgozatban nagy figyelmet fordítottam erre a két utóbbi lépésre. A módszerek tesztelése során a terepi adatok mellett szimulált adatokat is használtam, amelyeknél a várt eredmény előre ismert.

Az adatbázisok használatát bemutató két esettanulmány a dolgozat kisebbik részét teszi ki. Ennek oka, hogy az alkalmazások feltétele a megfelelő méretű adatbázis és a már kidolgozott módszerek. Az adatbázisok kiépülésével (lásd 1. és 2. fejezet) és a módszertani kérdések tisztázásával az ilyen tanulmányok száma várhatóan emelkedni fog mind a nemzetközi, mind a hazai szakirodalomban.

A cönológiai adatbázisok kiépülésével párhuzamosan több országban is új vegetációmonográfiák készültek és hasonló készítését Magyarországra is tervezzük. A klasszifikációhoz kapcsolódó módszertani fejlesztések elsősorban ezekben a munkákban, illetve az ezeket megalapozó tanulmányokban hasznosulnak.

A fajkészlet becslése a társulási szabályok vizsgálatában játszhat a jövőben fontos szerepet, de emellett felhasználható a védett, ritka fajok potenciális termőhelyeinek megkeresésében is, ezzel növelve a restaurációs beavatkozások hatékonyságát.

Számos érdekes és egyelőre kihasználatlan lehetőséget rejt a cönológiai adatbázisok összekapcsolása a növényi tulajdonság (trait) adatbázisokkal. Ezek közül talán a legfontosabb, hogy igen nagy mintákon vizsgálható, hogy az együtt élő fajok tulajdonságainak eloszlása milyen irányban tér el a random esetben várttól. Szintén érdekes kérdés, hogy milyen tulajdonságokban térnek el a specialista és a generalista fajok. Az ilyen hazai vizsgálatoknak jelenleg a legnagyobb korlátja a teljes hazai flórára vonatkozó sok tulajdonságot tartalmazó adatbázis hiánya, mivel a nyugat-európai adatbázisokból számos nálunk gyakori faj is hiányzik.

A cönológiai adatbázisokban található irodalmi felvételekhez általában nincsenek környezeti háttéradatok, de az újabb felvételekhez már egyre több esetben ilyen adatokat is gyűjtenek, ami lehetővé teszi a fajok előfordulását meghatározó környezeti tényezők vizsgálatát is.

Egy teljes lista összeállítása reménytelen feladat lenne, de remélhetőleg a fenti kiragadott példák is jól szemléltetik, hogy mennyire sokféle területen használhatók fel a kutatásban a cönológiai adatbázisok.

Az értekezés témakörében megjelent publikációk

Impakt faktoros közlemények:

Illyés E, Bauer N, Botta-Dukát Z 2009. Classification of semi-dry grassland vegetation in Hungary. *Preslia* **81**: 239-260.

Tichý L, Chytrý M, Hájek M, Talbot S, Botta-Dukát Z 2010. OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation Science* **21**: 287-299.

Illyés E, Chytrý M, Botta-Dukát Z, Jandt U, Skodova I, Janisova M, Willner W, Hajek O 2007. Semi-dry grasslands along a climatic gradient across Central Europe: Vegetation classification with validation. *Journal of Vegetation Science* **18**: 835-846. (2007)

Chytrý M, Tichy L, Holt J, Botta-Dukát Z 2002. Determination of diagnostic species with statistical fidelity measures. *Journal of Vegetation Science* **13**: 79-90.

Botta-Dukát Z, Chytry M, Hajkova P, Havlova M 2005. Vegetation of lowland wet meadows along a climatic continentality gradient in Central Europe. *Preslia* **77**: 89-111. (2005)

Botta-Dukát Z, Kovács-Láng E, Rédei T, Kertész M, Garadnai J 2007. Statistical and biological consequences of preferential sampling in phytosociology: Theoretical considerations and a case study. *Folia Geobotanica* **42**: 141-152.

Nem impakt faktoros cikkek:

Botta-Dukát, Z & Borhidi, A 1999. New objective method for calculating fidelity. Example: The Illyrian beechwoods. *Annali di Botanica (Roma)* **57**: 73-90.

Botta-Dukát Z 2004. A magyarországi mocsárrétek cönológiai irodalmának áttekitése és szüntaxonómiai revíziója. *Kanitzia* **12**: 43-73.

Botta-Dukát Z 2008. Validation of hierarchical classifications by splitting dataset. *Acta Botanica Hungarica* **50**: 73-80.

Lájer, K, Botta-Dukát, Z, Csiky, J, Horváth, F, Szmorad, F, Bagi, I, Dobolyi, K, Hahn, I, Kovács, J A & Rédei, T 2008. Hungarian Phytosociological database (COENODATREF): sampling methodology, nomenclature and its actual stage. *Annali di Botanica (Roma) nuova series* **7**: 197-201.