Válasz Szabó Szilárd, az MTA doktora opponensi véleményére

Megköszönöm Tisztelt Opponensemnek értekezésem részletes bírálatát, elismerő szavait, gondolatébresztő kérdéseit. Megjegyzéseire és kérdéseire az alábbiakban válaszolok.

Formai felvetések:

„A szöveg viszont, bár információtartalma értékes, meglehetősen zsúfolt (betűméret, tagolás), ebben a formában nem tette könnyűvé az olvasást. A helyesírási hibák rendszerint csak egybeírás-különírás témakörben fordultak elő. A Szerző fogalmazásmódja gördülékeny, ugyanakkor, nem hibaként megemlítve, kedveli az idegen kifejezéseket: számos olyan kifejezést találtam, ami magyarul is megfelelő lehetett volna."

Sajnálom, hogy a dolgozat zsúfoltságával megnehezítettem opponenseim dolgát. Ahogy Bírálóm is írta, egy szintetizáló jellegű munkát állítottam össze, amelyből bármely rész elhagyása felborította volna annak koncepcióját. A formátum volt az az eszköz, amellyel a terjedelmi korlátok túllépése nélkül tudtam átadni a legtöbbet abból, amit terveztem. Többé-kevésbé tisztában vagyok stílusom sajátosságaival, és folyamatosan küzdök annak vadhajtásaival, mint amilyen az idegen kifejezések esetenkénti túlzott használata. A dolgozat szerkesztése során is próbáltam odafigyelni, ezek szerint nem jártam teljes sikerrel. Mindenesetre köszönöm az építő kritikát, amely remélem, jó hatással lesz rám a későbbiekben.

Bírálóm megjegyzi, hogy a természeti hátrányokkal érintett területek térképzése kapcsán helyes lett volna az első előfordulásnál egy kicsit több információt szolgáltatni.

Sok fejtörést okozott a dolgozatom tagolása, az abban bemutatandó munkák többszálú kapcsolatrendszere miatt. Főképp azonban azért, mivel a módszerek fejlődésével és/vagy a rendelkezésre álló adatok bővülésével az adott kihívásra megújított válaszok adhatók, amely időnként együtt jár a kérdések ismételt feltevésével. Ez tipikusan igaz a természeti hátrányokkal érintett területek lehatárolásának problémakörére. Úgy éreztem, ennek részletes kifejtése megbontaná a 2. fejezet gondolatmenetét, viszont az indikátor krigelésen alapuló első eredményeket nem szerettem volna kihagyni az azt tárgyaló alfejezetből. Az ilyen, tematikusan több fejezetbe is illeszkedő munkálatokat részletesebben csak az egyikben mutattam be, azonban egyes aspektusait kiemelve a másikban is röviden tárgyaltam, a teljesebb kép bemutatása céljából. A fejezetek közti átjárást kereszthivatkozások használatával próbáltam megkönnyíteni. Ebben az esetben úgy látszik, ez nem volt eléggé sikeres megoldás.

Tartalmi felvetések:

Bírálóm 13 számozott pontban foglalta össze megjegyzéseit és kérdéseit. Bár az ezeket összekötő részekben is előfordulnak reakciót igénylő felvetések, ezeket nem veszem tételesen sorra, hanem a vonatkozó kérdéscsoportokra adott válaszaimba foglalom bele mondandómat. A kiegészítő megjegyzéseket köszönöm, az azokban tett kiegészítésekkel, illetve pontosításokkal teljes mértékben egyetértek.

A Digitális Kreybig Talajinformációs rendszer reambulációja és információtartalma

„1. A DKTIR adatbázis további felhasználása kapcsán merült fel bennem kérdésként, hogy az alkalmazott módszer (48. oldal utolsó bekezdése) véletlen fa (Decision Tree, DT) vagy véletlen erdő (Random Forest, RF) volt? Nincs okom feltételezni, hogy a Szerző ne lett volna precíz az eddigiek, alapján (vagyis ha jól értem a DT módszert alkalmazta), továbbá mindkettő módszer alkalmazása indokolt lehet, de részemről megbízhatóbbnak tartom az RF módszert az ismétlések miatt. Azt elismerem, hogy az RF esetében nem lehetséges a döntési fa felrajzolása (vagy szöveges leírása), a kérdés az, hogy szükséges-e."

Ha most kezdenék bele kategória típusú talajtérképek dezaggregálásába, vagy akár másfél évvel ezelőtt, magam is véletlen erdőt használnék a modellezés során. Az ebben a fejezetben közölt vizsgálatok és eredmények azonban a 2013-ra datálódnak, amikor a digitális talajtérképezésben az egyszerű osztályozó és regressziós fákat még nem szorították háttérbe az azok továbbfejlesztésével előállított módszerek, mint amilyen a véletlen erdő. Azóta, de még a dolgozat írásának megkezdése óta eltelt több mint másfél évben is több területen jóval előrébb jutottunk. Húzni kellett azonban egy

határvonalat, az addig elért és kiforrott (azaz publikált) eredményeket próbáltam összedolgozni egy egységes egésszé, szem előtt tartva, hogy mindegyik megfelelőképpen alá legyen támasztva lehetőleg elsőszerzős publikációkkal. A dezaggregálás módszereinek továbbfejlesztése és eredményeinek finomítása tavaly fokozatot szerzett doktoranduszom, Laborczi Annamária kutatási területébe tartozik.

„2. További kérdésem a fejezet egészéhez, hogy mennyire tekinthető egységesnek maga a referencia talajtani adatbázis? Ez azért lehet fontos, mert ha nem egy laborban történtek a vizsgálatok vagy pl. a szemcseösszetétel/textúra osztályok meghatározása empirikus volt, akkor felmerülhet a kérdés, hogy az egzakt, kvantitatív módon levezetett térképek pontossága nem azért tér-e el a referenciától, mert egy-egy kategória helytelenül lett megállapítva még a térképezés idején?"

A Kreybig térképezés anyaga hatalmas és egyedülállóan értékes örökséget képvisel, de mint minden egyéb adatbázisról erről sem feltételezhető, hogy hibátlan lenne. A felvételezés és a laborvizsgálatok során követett módszertani protokollok célja az adatgyűjtés hibáinak, bizonytalanságainak mérséklése, de a szubjektivitásból származó eltérések, illetve a mérési hibák ezek megfelelő alkalmazása esetén sem zárhatók ki. Utólagos azonosításukra, feltárásukra minimális az esély. Az összes ezekkel kapcsolatos, pontosan meg nem határozott hibát a referencia adatok bizonytalanságába értjük bele, amelynek értelmezését és kezelését megkönnyíti a statisztikai környezet. Nem tekintjük a felhasznált referencia adatokat abszolút pontosságúnak, csupán egy (remélhetőleg) kis szórású valószínűségi változó egy reprezentációjának, amelynek pontatlanságáért a modell eredmények bizonytalanságának egy részével fizetünk.

Az elhúzódó munkálatok ellenére is a Kreybig felvételezés anyaga egységesnek tekinthető, mivel a vizsgálatok azonos módszertan szerint és laborban folytak (már amennyire ez utóbbi egy világháború előtt és után annak tekinthető).

A lehatárolás funkcionális talajtérképi vonatkozásai

„3. Az ehhez kapcsolódó kérdésem az, hogy a bevont segédváltozók között mennyire volt zavaró a geometriai felbontások különbözősége. Az alkalmazott 100 m-es általános újraskálázás az EU-DEM esetében a felbontás rontását, míg a MODIS NDVI esetében a felbontás javítását jelentette. Míg az előbbi esetében a jobb minőségű adatból rosszabb lett, a MODIS NDVI-nál már nem valós térbeli részletességet eredményezett."

Munkáink során próbáljuk az elérhető adatok térbeli felbontásában meglevő eltéréseket minimalizálni, miközben figyelembe kell venni az azokhoz való hozzáférés lehetőségeit, az általuk biztosított adatminőséget, a tematikus tartalmat, a térbeli lefedettséget és nem utolsó sorban a számítási kapacitásokat. Mindezen területeken nagyon gyorsak a változások. A dolgozatban bemutatott térbeli modellezések elvégzése idején (néhány évvel ezelőtt) országos és akár globális digitális domborzat modellek már nagy térbeli felbontásban, ingyenesen is hozzáférhetők voltak, miközben az ingyenesen elérhető, több időpontú és többsávos távérzékelt képi információkat országos fedettséggel a MODIS adatok képviselték. A nemzetközi gyakorlathoz igazodóan mi is ezekre támaszkodtunk a GlobalSoilMap szabványban is követett 100 méter felbontású predikciók kidolgozása során. Jelen munkáinkban már magunk is áttértünk az azóta elérhetővé vált, nagyfelbontású Sentinel adatok használatára, ezzel is javítva az újabb predikciók eredményét.

„4. A segédváltozókat főkomponens analízis alkalmazása után vonták be a vizsgálatba, amivel kapcsolatban kérdésem az, hogy történt-e a bevont változókon standardizálás és/vagy a korrelációs mátrix alapján történt-e a modell futtatása. Mi alapján ítélték meg a modell illeszkedését (pl. KMO, RMSR, AGFI)?"

A kvantitatív típusú környezeti segédváltozókon (prediktorokon) a regresszió krigelés előkészítéseként végeztünk főkomponens analízist, mellyel a multikollinearitás csökkentése volt a célunk a többszörös lineáris regresszió analízis érvényes elvégzése érdekében. A főkomponens analízist megelőzően a környezeti segédadatok fedvényeinek pixelértékeit 0-255 skálára transzformáltuk. A regressziós vizsgálatok során független változóként a PCA eredményeként kapott főkomponenseket, illetve az indikátor változókat használtuk fel. A regressziós modell illeszkedését a determinációs együtthatóval mértük ugyan, de nem tartjuk olyan nagy jelentőségűnek az értékét, mivel ezzel a hibrid modellben csak a determinisztikus komponens illeszkedését jellemezzük. Ezt követi ugyanis a regressziós modellből fennmaradó reziduumok krigelése, ami egyrészt a térbeli predikció pontosságára vonatkozó

becslést is szolgáltatja, másrészt a determinisztikus taggal kombinálva eredményezi a tematikára vonatkozó térképet.

„5. A 3.6. táblázatban nem teljesen világos, hogy mit látunk. A táblázat címe szerint területi arányok szerepelnek az egyes mezőkben, de a termőréteg vastagság és lejtés metszetében 870 szerepel, ami kétségessé teszi, hogy ez arány lenne."

A táblázatban a keresztfeltételek együttes előfordulásainak területi kiterjedései szerepelnek. A cellákban szereplő értékek hektárban értendők, a mértékegység sajnálatos módon lemaradt.

„6. A talajbonitációs térkép dezaggregálása kapcsán az alkalmazott módszer látványosan feljavította a térbeli információtartalmat, amivel kapcsolatban az a kérdésem, hogy a javulás mennyire megbízható – megvizsgálta-e a Szerző a tematikus pontosságát az így kapott térképnek?"

A dezaggregált talajbonitás térképre nem történt bizonytalansági modellezés, ezért nem is szerepel a dolgozatban becslés a térkép globális és lokális megbízhatóságára, csakúgy mint a fejezet többi predikciója esetén sem. Térbeli predikcióinkat lehetőség szerint kiegészítjük a térbeli megbízhatóságra vonatkozó becsléssel, a szóban forgó fejezetben azonban több esetben erre nem volt mód. A tematikus térképi dezaggregáláshoz a későbbiekben kidolgoztunk egy módszert a becsült térkép bizonytalanságának modellezésére, amelyet szerepeltettem is a 2.5.2 fejezetben, az AGROTOPO genetikai talajtípus fedvény Duna-Tisza-köze területére történő leskálázásának bemutatására. A talajbonitációs térképnél ezt a becslést azonban utólag nem használtuk. A megbízhatóság kvantitatív elemzésének hiánya ellenére is joggal bízhatunk a dezaggregált térkép által szolgáltatott predikció javulásában a következők miatt. Az Agrotopográfiai térképek döntően a Kreybig-féle talajfelvételezés adataira alapozva és a Kreybig térképek talajfoltjainak térbeli és tematikus generalizálásával történő szintetizáló munka eredményeként születtek. Mivel a generalizálás összevonással és elhanyagolással, ennek megfelelően információvesztéssel jár, az így született agroökológiai egységek talajtani szempontból inhomogének, amelyeket az akkor segédeszközül választott 1:100.000-es topográfiai térképek által szolgáltatott domborzathoz illesztettek. Így a Kreybig mintázat és egy megfelelő részletességű domborzat modell hatékony eszköze a foltokon belüli heterogenitás elemzésének, amennyiben az agroökológiai egységek lehatárolása során alkalmazott mentális talaj-táj modelleket egy arra alkalmas adatbányászati eszközzel formalizálni tudjuk.

„7. A terméseredmények termőhelyi adottságainak integrált felhasználása az elérhető adatok és technológiák tükrében kreativitást és magas fokú problémamegoldó képességet tükröznek. A kérdésem az, hogy a kapott eredmények validálhatók-e valamilyen formában?"

Az adott térképre vonatkozóan nem történt közvetlen validációs vizsgálat. Ennek legfőbb oka, hogy a térkép alapjában véve a modellfuttatások eredményeinek visszavetítése a földrajzi térbe. Ennek megfelelően egy-egy pixelre vonatkozóan a bizonytalanság az alkalmazott 4M növény szimulációs modell eredményének bizonytalanságából fakad, amely integráltan tartalmazza az adott pixelről a modell számára bemenő adatként használt talajtani és meteorológiai tematikák lokális bizonytalanságán túl a modell felépítéséből és paraméterezéséből származóakat is. Ezek együttes elemzése egyelőre nem történt meg, bár éppen mostanság kezdtünk egy vizsgálatba, melynek eredményei várhatólag felhasználhatók lesznek a későbbiekben ezen kérdés legalább részleges megválaszolására.

Az eredmények validálására egy másik lehetőségként adódna valós terméseredmények felhasználása. Erre vonatkozó, operatív, lehetőség szerint idősoros, országos kiterjedésű, reprezentatív adatgyűjtésről azonban nem tudunk. Éppen ennek hiánya miatt merült fel a termőhelyen szimulált növénynövekedési modellek alkalmazása a produktivitás becslésére.

A hazai térbeli talajinformációk előállításának és szolgáltatásának megújítása

„8. A talajok felső rétegének szervesanyag-tartalmához kapcsolódó kérdésem itt is a PCA-hoz kapcsolódik: milyen eredménye lett a PCA-nak, milyen értékelés szerint fogadták el az eredményét? Ennél a lépésnél érdekes lett volna látni, hogy a regressziós modellekbe mely főkomponensek kerültek és azok mely változókkal korreláltak. Az eredményeket természetesen elfogadom, mert meggyőzők, a fenti megjegyzés a bíráló érdeklődési köréből és

ehhez kapcsolódó kíváncsiságából ered. Ez módszertanilag segített volna tisztázni azt is, hogy az A-L modellek esetében minden esetben külön PCA futtatására került sor? Másik kérdésem az, hogy a túl sok változó nem okoz túlillesztést (overfitting)?"

A regresszió krigeléssel végzett predikciók során a főkomponens analízist alapvetően a környezeti segédváltozók információszolgáltatásának optimalizálására, illetve multikollinearitásuk csökkentésére használjuk, lényegében egy adat-előfeldolgozási lépésként. Ezért nem minden esetben szoktuk vizsgálni, hogy a többszörös lineáris regresszió analízisben felhasznált főkomponensek mely környezeti segédváltozókból és milyen módon jöttek létre. A digitális talajtérképezésben használt SCORPAN modell megalkotói és legtöbb használója szerint eredendően nem a talajképződés tényezőit próbálja magyarázni, hanem a kvantitatív és kvalitatív talajjellemzők térbeli predikciójához nyújt eszközt. Az utóbbi években kezd hangot kapni az a vélemény, amelyre Opponensem felvetése is utal, mely szerint a matematikailag formalizált összefüggések értelmezése hozzásegíthet a mentális talaj-táj modellek jobb megértéséhez.

A főkomponens analízist mindig az adott modellben szereplő numerikus változókon végeztük el. Ennek megfelelően a zalai vizsgálatban az azonos változócsoportokon alapuló C-D-E-F-G, illetve H-I-J-K-L modellek esetén csak egyszer történt meg a PCA futtatás. Az együttesen a variancia 99%-át magyarázó főkomponenseket használtuk a többszörös lineáris regresszió analízis során a kvalitatív segédváltozókból képzett indikátor változókkal együtt.

Jelen esetben nem merülhet fel a túlillesztés veszélye, mivel a felhasznált 31 térbeli segédváltozó csaknem 2.000 adatponton lehetett kalibrálni, ráadásul az egyes modellekben az összes prediktornak csak egy részét használtuk.

„9. Az országos genetikus talajtérképhez kapcsolódó kérdésem az, hogy az új módszerrel készült, az AGROTOPO, valamint MÉM NAK genetikus talajtérkép összevetése (4.11. táblázat), mennyiben releváns? Van-e jelentősége a nagyon kis egyezésnek, illetve lehetséges-e jobb eredményt elérni bármilyen egyéb módszerrel?"

Az új és a két korábbi térkép közötti összehasonlító vizsgálatokat azért gondoltuk fontosnak, mivel az új térkép a korábbaiktól teljesen eltérő koncepció mentén született, miközben az ország talajtakarójának ugyanazon jellemzőjét hivatott bemutatni, mint elődei. Jelen kérdés kapcsán nem látom relevanciáját valamiféle jobb eredmény elérésének, mivel az új, illetve a régi térképek közti eltérések viszonylag jól értelmezhetők a két korábbi térkép szerkesztési módszertanának fényében. Az összehasonlítások eredményei lényegében a hasonlóságaik és különbségeik számszerűsített jelzéseiként tekinthetők.

„10. A vízerózió térképezésével kapcsolatban a kérdésem, hogy mit ért a Szerző „egyszerű ensemble modellezés" alatt (101. oldal, 3. bekezdés)?"

A lehető legegyszerűbb ensemble modellezést alkalmaztuk, azaz a két modell eredményeit pixel szinten átlagoltuk. Azért fogalmaztam mégis általánosabban, mert egyre több munka mutat arra, hogy a különbözőképpen elvégzett predikciók eredményeit érdemes összedolgozni, az egyes predikciók előnyeit kihasználó, összesített modelleket előállítani, amelyre vonatkozóan folyamatosan születnek javaslatok. Számomra a legígéretesebb az egyes predikciók lokális megbízhatóságukkal történő súlyozása. Mivel az alkalmazott eróziós modellekre nem született bizonytalanság elemzés, így a térképek lokális megbízhatóságára nem állt rendelkezésre olyan becslés, amit fel lehetett volna használni a két modell megfelelő súlyozására.

A mintavétel tervezéstől a célspecifikus talajtérképekig

„11. A komoly tervezés ellenére helyenként mégis gyengébb lett a bemutatott termőréteg-vastagsági térkép megbízhatósága, bár láthatóan sokat javult a második kör mintavételét követően. Szemmel is láthatóan kevesebb minta begyűjtésére került sor a mintázandó terület ÉK-i sarkában (van olyan önálló földrészlet ahonnan egy sem), és ezeken a helyeken a második körben sem javult a megbízhatóság, sőt nem meglepően a legkisebb ott lett, ahonnan nem lett begyűjtve minta. A kérdésem ez esetben az, hogy ebben volt-e szándékosság, vagy az SSA módszerrel kapott mintavételi sémát használták, annak felülbírálata nélkül? Van-e esetleg kritikai megjegyzése a módszerrel kapcsolatban, vagy még ezzel együtt is (azaz egyes területek látható kihagyásával) ez a legjobb mintavételi tervező eljárás?"

Az SSA módszert hatékonynak és a feladathoz jól illeszkedőnek tartom, de nem feltétlenül a legjobbnak. Ahhoz hogy erről megalapozott állítást lehessen tenni, egyéb mintavételi eljárásokkal történő összehasonlító vizsgálatokat kellene végezni hatékonyságukra vonatkozóan, ami hatalmas munkának ígérkezik. Jelen esetben nem maga az SSA vezetett az Opponensem által említett problémákhoz, hanem alapvetően az a beleavatkozás a mintavétel folyamatába, melynek során egy súlyozás révén (megbízói kérésre) prioritási területeket vezettünk be, melyeknek kiemelt szerepet kellett élvezniük a mintavételezésben, hogy az adott területekre vonatkozó becslések lokális megbízhatósága magasabb legyen. Mindezeken túl az 5.8 ábra két térképén megjelenített információ félreérthető lehet, mivel a rajtuk szereplő, megbízhatóságukra vonatkozó melléktérképek skálája abszolút értékben nem azonos.

Digitális környezeti térképezés a talajtérképezés koncepciója alapján

„12. A Szerző itt is PCA-alapú többváltozós regressziót használt és az ehhez kapcsolódó kérdésem az, hogy a stepwise változószelekcióhoz ez jó választás-e? A PCA első főkomponense magyarázza a variancia legnagyobb hányadát és bár nem valószínű, de mi történik, ha a változószelekció ezt „kidobja”? A dolgozatban leírt állítást helyesen fogalmazva: a főkomponensek a variancia 100%-át magyarázzák, a 99% magyarázott variancia az első 2-5 legfontosabbra vonatkozhat (ez nem derül ki a szövegből), ugyanakkor ez csak akkor marad meg, ha benn hagyunk minden változót a modellben.”

Ahogy a 8. kérdésre adott válaszban is írtam, a regresszió krigeléssel végzett predikciók során a főkomponens analízist alapvetően a kvantitatív típusú környezeti segédváltozók információszolgáltatásának optimalizálására, illetve multikollinearitásuk csökkentésére használjuk, lényegében egy adat-előfeldolgozási lépésként. A hibrid modell első, determinisztikus lépésében, a többváltozós lineáris regresszió során a PCA első néhány főkomponensét, amelyek együttesen a variancia 99%-át magyarázzák, és a kvalitatív változókból képzett indikátor változókat léptetjük be. Magam is úgy vélem, nem valószínű, hogy a stepwise regresszió során az információgazdag főkomponensek rostálódjanak ki. Azonban ha ez történne, sem „lenne ördögtől” való, mivel a PCA eredménye független a referencia adatoktól, amelyek térbeli kiterjesztése a modellezés célja. Ha pedig a kvalitatív prediktor változók megválasztása rosszul sikerül, míg a kategória típusúak nagyon informatívak az adott talaj jellemzőre vonatkozóan, akkor akár az is elképzelhető, hogy a többváltozós lineáris regresszió modell csak ez utóbbiakra épül fel.

„13. A regressziós reziduumok használata sztochasztikus komponensként a krigelésben előre mutató megoldás. A kérdésem arra vonatkozik, hogy az így kapott belvíz veszélyeztetettségi térkép mennyiben lett pontosabb?”

A korábban végzett ún. Komplex Belvíz-veszélyeztetettségi Mutató Térképezés (KBMT) tisztán környezeti korreláción alapult, melynek során az elöntési kockázatot szimplán a többszörös lineáris regresszió egyenletének segítségével azonosítottuk, annak ellenére, hogy a determinációs együttható időnként igen alacsony értéket adott. A digitális talajtérképezésben igen jó eredményekre vezető regresszió krigelés alkalmas eszköznek mutatkozott a KBMT nem kellően kielégítő eredményeinek javítására, mivel alapvetően azt a koncepciót testesíti meg, mely szerint önmagában sem a környezeti korreláció, sem a geostatisztikai interpoláció nem képes számot adni a térképezendő változó teljes térbeli variabilitásáról, azaz elegendően pontos térképi végterméket produkálni. A két, egymást kiegészítő térbeli származtatási módszer együttes használata segít az egyenkénti viszonylagos gyengeségek javításában.

Végezetül még egyszer köszönöm a dolgozat gondos bírálatát és kedvező értékelését.

Budapest, 2019. augusztus 15.

Pásztor László