Opponensi vélemény

Dr. Pásztor László
Célspecifikus térbeli predikciók kidolgozása feladatorientált, térképi alapú talajinformációk előállítására c.
MTA Doktori értékezéséről

A dolgozat felépítése, stílusa, arányai és eredetisége

A dolgozat fő erénye, hogy egy komplex, kvantitatív alapokon nyugvó és sokrétű képet kaphatunk a digitális talajtérképezésről a lehető legszélesebb értelemben.

A dolgozat szerkezetileg nem követi a természettudományokban megszokott hagyományos bevezetés – szakirodalmi áttekintés – anyag és módszer – eredmények (diszkusszió) – következtetések sorrendet, ugyanakkor ez csak akkor lenne zavaró, ha nem lehetne elkülöníteni a már meglévő eredményeket a Szerző által végzett munkától. A bevezetés fejezetbe került a problémafelvetés és szakirodalmi áttekintés, ami a nemzetközi cikkekben bevett szokás, itt talán jobb lett volna a szakirodalmi áttekintést külön fejezetbe tenni, de mint később ki fogok térni rá, a fejezet alapossága és szakszerűsége feledteti ezt a kritikai megjegyzést. Az eredmények öt külön fejezetet kaptak, és az alkalmazott módszerek, felhasznált adatok problémaorientált tematikában ezekben a fejezetekben kerültek bemutatásra. Arányaiban a szakirodalmi áttekintés 20%-ot tesz ki és 80% jut a módszertanra és eredményekre, ami ideális eloszlásnak számít.

A dolgozat szakirodalmi feldolgozása már-már zavarba ejtően részletes. Összesen 617 tétel került a szakirodalmi listába, ami önmagában sokat elmond a Szerző alaposságáról, de ehhez hozzátehetjük, hogy dicséretesen idézte mind a legfontosabb magyar és angol nyelvű szerzőket a nemzetközi folyóiratcikkektől a konferenciakiadványokig. A szakirodalmi áttekintés egy akadémiai doktori dolgozattól elvárhatóan értékelő jellegű, azaz nem egyszerűen csak ismerteti a vonatkozó elméleti, vagy történeti hátterét a szóban forgó digitális térbeli talajtérképezési módszernek, hanem saját szempontrendszere alapján ki is fejti, hogy mennyiben hasznos, mi az előnye és hátránya, illetve a tudomány jelen állása szerint mennyiben számít megbízható eljárásnak.

A szöveg viszont, bár információtartalma értékes, meglehetősen zsúfolt (betűméret, tagolás), ebben a formában nem tette könnyűvé az olvasást. Nyelvi stílusa kifejezetten szakmai, ugyanakkor könnyen értelmezhető és olvasható. 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 (pl. források allokációja, omnipotens).

A szakirodalmi részben néhány kérdés/megjegyzés merült fel bennem:

1. Az osztályozó/regressziós fák esetében egyetértek a robosztusságot összefoglaló gondolatokkal (24. oldal), azonban érdemes tovább gondolni a következőket:
    - bár a lényegtelen változók a modellt nem befolyásolják, elhagyásuk nem haszontalan, mert csökkenti a potenciális túlillesztést (overfitting);
    - a prediktor változók multikollinearitására a módszer valóban nem érzékeny, ugyanakkor a végső modellben érdemes megszüntetni, mert egyrészt befolyásolja a változók fontosságának a százalékos megoszlását, másrészt nehezíti a modell interpretációját.
2. A véletlen erdő (RF) modellek esetében egyetértek a Szerző azon megjegyzésével, hogy „nem véletlenül használtak véletlen erdőket a számos globális, kontinentális talajtulajdonság adatbázis létrehozásához”. Saját tapasztalatom szerint az RF más osztályozókkal szemben sokszor ad jobb eredményt, továbbá az alapbeállításokkal, a hiperparaméterek finomhangolása nélkül is kiváló eredményt hoz. Ahol az RF nem működik, ott más módszer sem szokott jó eredményt adni. (Ezzel nem akarom azt mondani, hogy az RF a legjobb választás, hanem azt, hogy használata mellett szól a pontossága, akár robosztus, akár nem.) Az RF-fel kapcsolatban rendszerint az szokott felmerülni problémaként, hogy mivel véletlen mintavételen alapul, ezért nem ismételhető meg kétszer úgy, hogy a végeredmény is ugyanaz legyen. Ez a legtöbb szoftver esetében igaz is, de R és Python környezetben, ahol a random mintavétel paraméterezhető, már megoldott.
3. A Support Vector Machine elnevezése magyarul inkább támasztó vektorok módszereként elterjedtebb.

Az eredmények újszerűsége és tartalmi értékelése

A dolgozat egésze egy életmű bemutatása. Nem rövid, a Szerzőnek volt mondanivalója és azt sem lehet mondani, hogy azért hosszú, mert felesleges részek lennének benne. Természetesen lehetett volna rövidíteni, de a Szerző szándéka jól láthatóan az volt, hogy a talajtan teljes általa művelt szegmense, a digitális talajtérképezés a lehető legrészletesebben kerülhessen bemutatásra. Mivel ez az az utolsó olyan tudományos megmérettetés, ahol ezt nemcsak, hogy megteheti, hanem meg is kell tennie, ezért – bár így meglehetősen sokfelé ágazó lett a téma – nem hibáztathatjuk érte.

Az eredmények a Szerző irodalmi munkásságának hivatkozott publikációi alapján a Szerző sajátjai.

A Digitális Kreybig Talajinformációs rendszer reambulációja és információtartalma
A fejezet a Kreybig-féle talajtérképezés bemutatásával indul, melyben a digitális változat kialakításának az alapfeltételeit ismerjük meg részleteiben. Erre a fejezetre kifejezetten igaz, amit a fenti bekezdésben írtam, vagyis a leírás nem egy vagy néhány publikációra épül, hanem szintetizáló jellegű, akár úgy is tekinthető, mint az életmű egyfajta szintetizációja, mivel közel 20 éves munka eredménye a Kreybig térképek vektorizációjától egészen a többszöri reambulációkig.

A 3D talajtextúra adatbázis előállításához használt indikátor krigelés (IK) egy kifejezetten előremutató módszer, ami nemzetközi viszonylatban is újdonság. A természeti hátrányokkal érintett területek térképzése kapcsán, ahol szintén IK-t alkalmazott, viszont felmerül a kérdés, hogy a bináris térképek milyen input adatból jöttek létre. Írja ugyan a Szerző, hogy ezt a részt itt csak felvillantja és később tárgyalja, ugyanakkor helyes lett volna az első előfordulásnál egy kicsit több információt szolgáltatni. A validálás eredményét nem közli, a táblázatból viszont én kiszámoltam és a közel 86%-os eredmény biztatónak tűnik.

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.
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 lehatárolás funkcionális talajtérkép vonatkozásai

A különböző (akár egyszerre több) talajtulajdonságok kategória szintű lehatárolása történt meg kvantitatív alapokon. A téma önmagában nem új, ilyen térképek már korábban is készültek, viszont nem kvantatív és digitális környezetben. A kapott eredmények indikátor és regressziós krigelésen alapulnak.

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.
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)?
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.
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 Szerző termőhelyi minőséget meghatározó pontozásos rendszerrel kapcsolatos kritikai észrevételeivel egyetértek, a legelső lépés, a pontérték meghatározása a legnagyobb objektivitás mellett is tartalmaz szubjektivitást.

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?

A hazai térbeli talajinformációk előállításának és szolgáltatásának megújítása

A korábbiakban leírt alapokon itt ismerkedhetünk meg azzal, ahogy a talajinformáció fogalma kiterjed a korábbi talajszelvény/poligon és a hozzájuk tartozó információ (pl. magyarázó füzet, táblázat, kód) a változó felhasználói igények szerinti, lehető legjobb és az adott feladathoz igazodó felbontású talajtérképekig, ahol ráadásul a mindig jelen lévő bizonytalanság is hozzárendelésre kerül a térképekhez.

Egyetértek azzal a kezdőgondolattal, hogy erre nézve nincs és nem is lehet univerzálisan alkalmazható módszer, bár sokan szeretnének ilyet kidolgozni, vagy egyszerűen csak valakitől egy ilyen módszertant átvenni a szakirodalomból. Nincs két egyforma feladat és nincs két egyforma terület, ahonnan minden input adat ugyanúgy (pl. mintaszám, mintasűrűség, mért változók, mérési körülmények azonossága, mérési pontosság) állna rendelkezésre egy modellhez. Ha ezt felismerjük, akkor nyilvánvalóvá válik, hogy rugalmas és adaptív környezetre és egyedi megoldásokra van szükség ahhoz, hogy a szolgáltatás valóban a felhasználói igényeket elégítse ki. Az ebben a fejezetben ismertetett módszer (halmaz) ennek maximálisan eleget tesz.

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

Kisebb megjegyzés a CarpatClimmel kapcsolatban, hogy nem 60, csak 50 év hosszú idősort tartalmaz (110. oldal).

A mintavétel tervezéstől a célspecifikus talajtérképekig

A két lépésben végrehajtott mintavétel a tervezéstől a kivitelezésig és a térképek előállításáig egy jól átgondolt – és megrendelői célokat kiszolgáló munka volt. Ez utóbbi azért is fontos,

mert az eddigiekben részben tudományos, részben olyan felhasználói igényeket kielégítő megközelítéseket mutatott be a Szerző, melyek alapvetően egy általános cél elérést tűzték ki célul, mint pl. talajgenetikai, termőréteg vastagság, mechanikai összetétel térkép, itt viszont volt egy konkrét megkeresés a Tokaj-Hegyaljai borvidék részletes talajtulajdonságainak a feltárására. Vagyis, itt érvényesült az elmélet találkozása a gyakorlattal. A kétlépcsős megközelítés ennek a célnak volt alárendelve és a végeredmény eléréshez tartották magukat ahhoz a térbeli mintavételi tervhez, ami 850 talajminta begyűjtését és feldolgozását irányozta elő. A térképezésből sokat ugyan nem láthatunk a jogi kérdések miatt, ami viszont bemutatásra került, az meggyőző.

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?

Digitális környezeti térképezés a talajtérképezés koncepciója alapján

Az utolsó fejezetben a Szerző a belvíz és radontérképezés példáján keresztül mutatja be a környezeti tényezők/folyamatok térképezésének a lehetőségeit. Az eredmény is itt is meggyőző és nem a Szerző itt sem takargatja a modellek hibáit, bizonytalanságukat. Ez az a gondolat, amivel a magam részéről a legmesszebb menőkig egyet tudok érteni: nem a legmagasabb R^2 elérése a cél, hanem a rendelkezésre álló adatok és módszerek/algoritmusok kombinációjában egy precíz elemzés eredményének a bemutatása. Itt előfordult, hogy az R^2 csak 0,35 volt, ami nem nevezhető soknak, de kiindulási alapnak jó, amin lehet javítani. A 6.3. ábra egy nagyon komoly munka eredménye, de lemaradt a jelmagyarázata (bár ki lehet találni, hogy melyik a gyakoribb).

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.
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?

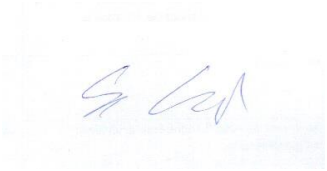
A tézisfüzet kissé hosszú lett, de tény, hogy jó áttekintést ad magáról a feldolgozott témáról. Az itt ismertetett új tudományos eredmények 8 pontban kerültek bemutatásra. Tartalmilag egyik tézis ellen sincs kifogásom, részemről elfogadom mindet; azonban a 6. és 8. tézisek

megfogalmazása nem tézisszerű. A megfogalmazás ezeken a helyeken óvatos, ami dicsérendő, nem mondja csak sajátjának azokat az eredményeket, amiket a mögötte lévő csapatban valósított meg. Ugyanakkor a dolgozatban nem 8 tézisnyi eredmény van, ennyi önmagában a DKTIR talajtérképezés reambulációjában benne lehet.

Összefoglalóan a dolgozat egy részletes és szakmailag igen sokrétű anyag. A módszerek napjaink legmodernebbjeinek számítanak, amihez nagy szaktudásra, felkészültségre és lényeglátásra van szükség. A Szerző sok éves munkája bontakozik a rövidnek nem nevezhető értekezésből, aminek még az utolsó oldalai is újdonsággal szolgálnak. Az alaposság nem kérdéses, mint ahogyan az sem, hogy egy kiforrott kutató dolgozata készült el, szintetizáló jellegű megközelítéssel. A megjegyzéseim és kérdéseim nem kérdőjelezik meg a munka hitelességét, vagy tartalmát, leginkább arról van szó, hogy a hosszú dolgozat ellenére a sok felvetett téma nem fejthető ki ebben a terjedelemben és a Szerzőnek valahol spórolnia kellett a hellyel. A meglátásom az a munkássága tükrében, hogy a Szerző láthatóan sokat fejlődött a talajfolt poligonok vektorizálása óta, lépést tartott a módszerek és a technika fejlődésével és be tudta építeni ezeket az aktuális munkáiba. Az univerzális krigelés, regressziós krigelés, indikátor krigelés, hagyományos statisztikai eljárások, gépi tanulás (ehhez kapcsolódóan döntési fák, neurális hálózatok, stb.) alkalmazása vitte el nemcsak őt, hanem az MTA ATK TAKI-t is egy olyan tudásszintre, melyet nemzetközi viszonylatban is jegyeznek – ezt bizonyítja, hogy az értekezés témájában 42 publikációja jelent meg, melyek közül csak a D1 minősítésű cikke száma 5 db.

Mindezeket figyelembe véve megállapítható, hogy Pásztor László értekezése eddigi tudományos munkásságával együtt messzemenően kielégíti az MTA doktori cím követelményeit. Ezek alapján a benyújtott dolgozat nyilvános vitára bocsátását javaslom, majd sikeres védés esetén a cím odaítélését támogatom.

Debrecen, 2019. 07. 09.



Dr. Szabó Szilárd
az MTA doktora