EÖTVÖS LORÁND TUDOMÁNYEGYETEM
INFORMATIKAI KAR

Geológiai előfordulások modellezése LiDAR adatok alapján

DIPLOMAMUNKA
TÉRKÉPÉSZ MESTERSZAK

Készítette:
Gurály Attila

Témavezető:
Dr. Albert Gáspár
egyetemi docens
ELTE Térképtudományi és Geoinformatikai Tanszék

Budapest, 2022

Tartalomjegyzék

Bevezetés

Elegendőek-e a morfometriai változók ahhoz, hogy – korábbi geológiai térképek alapján kijelölt mintaterületek segítségével – egy gépi tanuláson alapuló osztályozást végrehajtva pontos modellt kapjunk egy növényzettel borított terület geológiai formációiról? Milyen pontossággal tudunk modellezni kizárólag a morfometriai változókra hagyatkozva? Jelen dolgozatban elsősorban ezekre a kérdésekre keresem a válaszokat. A megfogalmazott kérdések azon alapulnak, hogy összefüggést feltételezünk a geológiai formációk és a földfelszín morfometriai paraméterei között.

A különféle földtudományi alkalmazásokon kívül gazdasági okokból (az ércek, az ásványok, az energiahordozók vagy a vízkészletek kutatása), valamint a környezeti veszélyek megfigyelése, az azok elleni védekezés szempontjából is fontos a geológiai adatbázisok naprakészen tartása és modernizálása (Bachri I., Hakdaoui M., Raji M., Benbouziane A., 2020). A geológiai térképezés azonban igencsak időigényes feladat, köszönhetően a nélkülözhetetlen terepi munkának. Ezen a helyzeten javíthat némiképp a távérzékelési módszerek alkalmazása, mellyel a terepen töltött idő minimalizálható.

Egy friss kutatás eredményeiből kiindulva egy osztályozás-alapú geológiai térkép hasznos segítség lehet a terepi felméréseket megelőzően (Albert G., Ammar S., 2021). Különféle nagyfelbontású műholdfelvételeket felhasználva már számos gépi tanulás alapú (a mesterséges intelligencia egyik területe) osztályozás készült. Műholdképeket azonban csak a növényzettől és a mesterséges objektumoktól mentes területeken célszerű felhasználni, ahol az alapkőzet a felszínen látható. Korábbi kutatások eredményei alapján, amelyekben különböző gépi tanulási módszereket hasonlítottak össze geológiai osztályozás céljára, a random forest osztályozás jónak (He J., Harris J. R., Sawada M., Behnia P., 2015), illetve a legjobbnak (Cracknell M. J., Reading A. M., 2014) bizonyult az összevetett módszerek közül. Egyes tanulmányokban egy nem felügyelt osztályozó algoritmust (SOM – self-organizing maps) alkalmaztak a műholdfelvételek spektrális sávjain alapuló litológiai (Bedini, 2009), illetve geofizikai jellemzők alapján (Carneiro C. D. C., Fraser S. J., Crósta A. P., Silva A. M., Barros C. E. D. M., 2012). A random forest osztályozással születtek korábban jó eredmények spektrális, illetve morfometriai változókat felhasználva marokkói területeken (Bachri I., Hakdaoui M., Raji M., Benbouziane A., 2020); valamint geofizikai tényezők vizsgálatával tasmániai területeken (Radford D. D. G., Cracknell M. J., Roach M. J., Cumming G. V., 2018). Egy ciprusi, növényzettel fedett területet vizsgáló tanulmányban spektrális jellemzők és LiDAR-ból

levezetett morfometriai változók felhasználásával használták a SOM-módszert (Grebby S., Naden J., Cunningham D., Tansey K., 2011). Tunéziai területeken pedig spektrális jellemzőkön alapuló litológiai, illetve globális domborzatmodell-alapú morfometriai változókkal alkalmazták a random forest osztályozást (Albert G., Ammar S., 2021). Ugyanezen tanulmány, illetve egy korábbi kutatás eredményei igazolták, hogy egyes morfometriai változók kellően pontosak lehetnek a helyes predikcióhoz, és abban fontos szerepet játszhatnak (Bachri I., Hakdaoui M., Raji M., Benbouziane A., 2020).

Jelen kutatás során a modellezéshez a random forest algoritmust alkalmaztam. Az algoritmus az osztályozáskor az előre meghatározott morfometriai változók fontosságát figyelembe véve sorolja be a felszín pontjait az előre definiált geológiai osztályokba (formációkba), tehát felügyelt osztályozási módszert alkalmaztam. A munkám alapjául szolgáló referencia-adatbázist a vizsgált területek nagy méretarányú geológiai térképei jelentették, a morfometriai változók kiszámításához pedig LiDAR (lézerszkennelt) adatokból generált terepmodelleket (DTM) használtam fel. A felhasznált adatokat mindkét adattípus esetében a USGS (Amerikai Egyesült Államok Földtani Szolgálata) ingyenesen hozzáférhető adatbázisai szolgáltatták. Dolgozatomban három Amerikai Egyesült Államok-beli, különböző geológiai eredetű területet vizsgáltam (vulkanikus, üledékes, illetve metamorf), melyek Washington, Nyugat-Virginia, illetve Virginia államokban találhatók. Munkám fő célja a referencia-térképhez viszonyítva a lehető legpontosabb geológiai osztályozás végrehajtása mindhárom említett mintaterületről. Eredményként várhatóan olyan osztályozott térképeket kapok, amelyekből rekonstruálhatók a referenciaként használt formációk poligonjai. A kutatás elsősorban módszertani, így a bemutatott módszerek bármilyen területen alkalmazhatók, ahol adottak az ehhez hasonló földrajzi, morfológiai és geológiai feltételek. Tudomásunk szerint jelen dolgozat paramétereivel egyező tanulmány korábban nem készült.

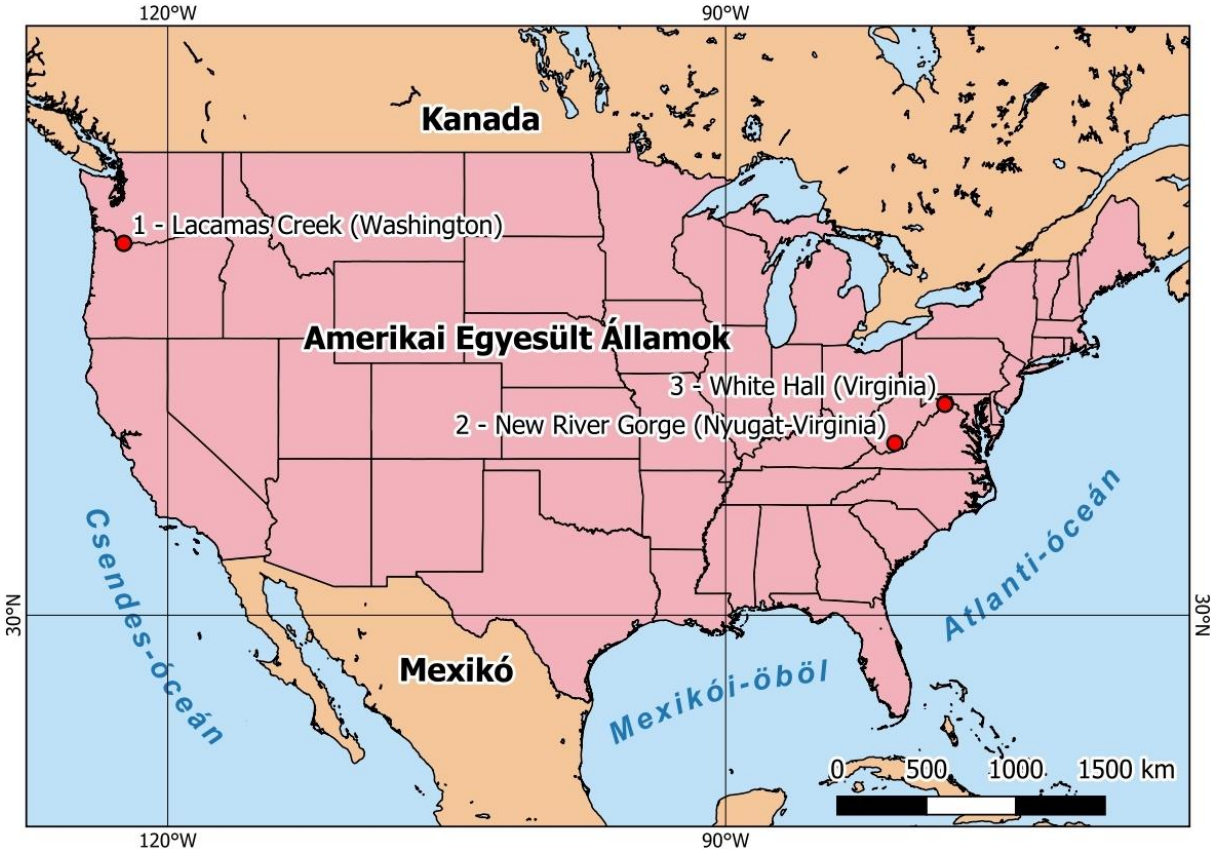
A vizsgált területek

Ebben a fejezetben a mintaterületek kiválasztásának szempontjai, illetve a három vizsgált terület kerül röviden bemutatásra.

A területválasztás egyik fő szempontja az volt, hogy átfogó képet kapjunk a modellezési módszer működéséről eltérő geológiai környezetekben. Emellett igyekeztem elkerülni a sűrűbben lakott területeket, törekedtem az emberi beavatkozásoktól mentes területek kiválasztására. Alapvető feltétel volt, hogy a területekről szabadon hozzáférhető térinformatikai adatbázis, illetve LiDAR pontfelhő is rendelkezésre álljon. Törekedtem a lehetőleg minél

nagyobb méretarányú geológiai térképek kiválasztására. A felhasznált geológiai adatbázisok és a LiDAR adatok a „Felhasznált adatok” című fejezetben kerülnek bővebben kifejtésre.

Ezen feltételeknek megfelelve esett a választás egy Kordillerák-beli, túlnyomórészt vulkanikus eredetű kőzetekből álló területre Washington államban; egy jellemzően üledékes kőzeteket felvonultató területre az Appalache-hegységből Nyugat-Virginia államban; illetve egy metamorf eredetű területre Virginia és Nyugat-Virginia államok határán (szintén az Appalache-hegységből). A későbbiekben az egyszerűség és könnyebb követhetőség kedvéért az alábbi alcímekben szereplő számokkal fogom azonosítani a területeket (pl. „1-es terület” vagy „1-es számú terület”). A vizsgált területek elhelyezkedése az 1. ábrán látható.



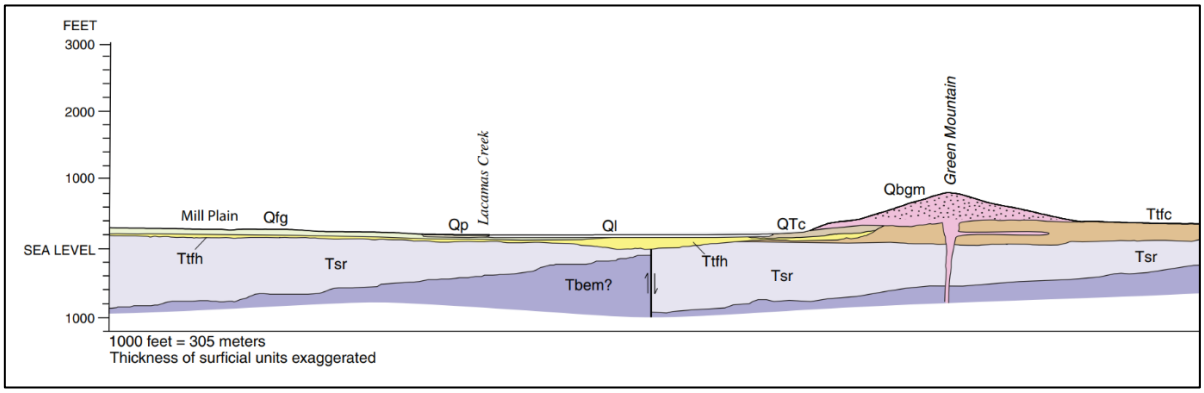
1. ábra: A vizsgált területek elhelyezkedése.

1. Lacamas Creek (Washington)

A Lacamas Creek-ről (Lacamas-patak) elnevezett 1-es számú terület az Amerikai Egyesült Államok északnyugati sarkában lévő Washington állam délnyugati határán, az Oregon állambeli Portland-től 25 km-re északkeletre található. A vizsgált terület dombjain eredő patakok vizét összegyűjtve nyugat, majd dél és délkelet felé folyó Lacamas-patak a területet délen elhagyva a Columbia folyóba torkollik. A vizsgált terület a Portland-medence keleti

szélén található, amely a Cascade-hegység és az oregoni Parti-hegység között helyezkedik el. A késői eocén kortól kezdve a Cascade-hegység epizodikusan aktív vulkáni területként van számon tartva (Evarts, 2006). A vulkáni aktivitás oka az alábukó Csendes-óceáni-lemez és az Észak-Amerikai-lemez találkozásánál létrejött Cascadia szubdukciós zóna. Számos korábbi vulkánkitörés és lávaömlés alakította a terület mai tagolt képét, melynek legmagasabb pontja 625 m-es, a legalacsonyabb pontja pedig 55 m-es tengerszint feletti magasságban található. A térség keleti, hegyvidéki része északkelet felé emelkedik, melynek nyugati, laposabb oldalán folyóvízi üledék halmozódott fel. A szelvény területének alapkőzete a teljes területen bazalt, illetve bazaltos andezit, melyek az oligocén korban kerültek a felszínre lávafolyások formájában. Ezek a vulkanikus lejtők lankásnak mondhatóak, ugyanis nem meredekebbek, mint 5°. A középső pleisztocénben bazalt és bazaltos andezit került a felszínre a terület déli részén lévő kisebb vulkánok kitörésekor. A késő pleisztocénben bekövetkező áradások következményeként pedig kavicsos üledékek rakódtak le a terület délnyugati részén. Az üledékes kőzetek anyaga nagyrészt homokkő, illetve konglomerátum. A terület rétegszerkezetébe a 2. ábra nyújt bepillantást. A terület érintetlen részei nagyrészt sűrű növényzettel borítottak, így csak elvétve bukkan a felszínre az alapkőzet. Erre csupán meredek sziklaszirteken, földcsuszamlások helyén, vagy útbevágásokban találunk példát. A modellezett terület kiterjedése 110,535 km^2.

Jelen alfejezet végén az 1. táblázatban ismertetem a területen előforduló formációkat, melyben azok nevei angol nyelven, eredeti formájukban szerepelnek. A formáció a kőzetrétegtan alapegysége, amely olyan kőzetrétegekből áll, melyek fizikai paramétereiket tekintve hasonlóak, például kőzettani vagy fácies tulajdonságaikban (Boggs, 1987).



2. ábra: Részlet az 1-es számú terület geológiai térképének egyik metszetéből (forrás: Geologic Map of the Lacamas Creek Quadrangle, Clark County, Washington). A jelek magyarázatát lásd az 1. táblázatban.

Kód	Név	Kőzetek	Terület (km^2)
Qa	Alluvium	iszap, homok, kavics	6,870
Qbbh	Basaltic andesite of Brunner Hill	bazaltos andezit	0,732
Qbgm	Basaltic andesite of Green Mountain	bazaltos andezit	1,419
Qbmc	Basalt of Matney Creek	bazalt	0,682
Qfg	Gravel facies	kavics	9,689
Qfs	Sand and silt facies	iszap, homok, agyag	5,936
Ql	Lake deposits	agyag, sár, szerves üledékek	5,585
Qls	Landslide deposits	földcsuszamlásos törmelék	8,579
QTc	Conglomerate	konglomerátum	25,053
Tbem	Basaltic andesite of Elkhorn Mountain	bazaltos andezit, bazalt, breccsa	40,682
Tsr	Sandy River mudstone	homokkő, aleurolit, agyagkő	1,271
Tftc	Troutdale Formation, Conglomerate member	konglomerátum	26,492
Tfth	Troutdale Formation, Hyaoclastic sandstone member	homokkő, konglomerátum, bazaltos törmelék	0,845

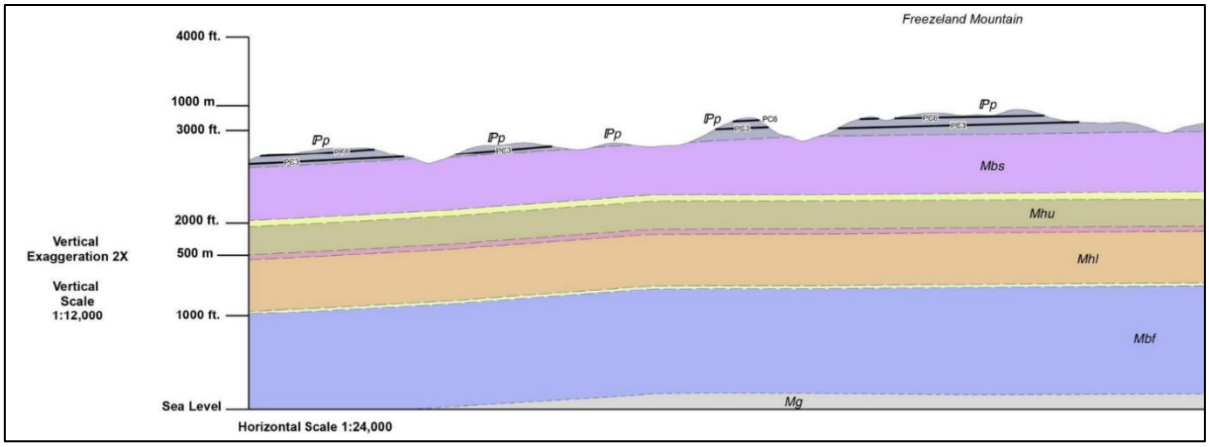
1. táblázat: Az 1-es számú területen előforduló formációk.

2. New River Gorge (Nyugat-Virginia)

A 2-es számú terület az ország keleti oldalának középső harmadában, az Appalache-hegység nyugat-virginiai részén helyezkedik el. A vizsgált terület a New River nevű folyó völgyét és annak környezetét foglalja magában, Hinton város környezetében. A terület Nyugat-Virginia fővárosától, Charleston-tól 100 km-re délkeletre helyezkedik el, az Allegheny-fennsík délkeleti részén. Az Allegheny-fennsík egy vízfolyások völgyeivel szabdalt vidék. Ezek a völgyek általában nem túl mélyek, bár viszonylag szélesek és lapos talpúak. A környék hegyhátai és tetői szintén laposak, a dolgozatban vizsgált terület legmagasabb pontja 1041 m. A New River 400 m körüli magasságban folyik át a területen, melynek legalacsonyabb pontja 397 m. A folyó észak-északnyugat felé haladva folyik át a területen, kiközben széles kanyarokat tesz. A terület mai képét elsősorban a folyóvízi erózió alakította ki.

Az Appalache-hegység az egyik legősibb földtörténeti képződmény. Idős korára utal az is, hogy kristályos és üledékes kőzetek egyaránt előfordulnak a területen. Előbbiek a hegység nyugati-délnyugati, míg utóbbiak keleti-északkeleti oldalára jellemzőek. Az Allegheny-fennsíkot elsősorban karbon kori üledékes kőzetek jellemzik. A hegyvidék ezen részét gyakran emlegetik „Új-Appalache”-ként is fiatalabb korára utalva. A területen a kőzetrétegek szinte tökéletesen vízszintes településben figyelhetők meg, az adott helyeken jellemző felszínközeli kőzeteket tehát a vízfolyások bevágódásának mértéke határozza meg (lásd 3. ábra). A terület

legjellemzőbb kőzetei a homokkő, az iszapkő, a mészkő, és a különféle palák (Peck R. L., Matchen D. L., Hunt P. J., 2013). A vizsgált terület kiterjedése 163,815 km^2. A területen előforduló formációkat a 2. táblázat foglalja össze.



3. ábra: Részlet a 2-es számú terület geológiai térképének egyik metszetéből (forrás: Bedrock Geology of the New Rivre Gorge National River). A jelek magyarázatát lásd a 2. táblázatban.

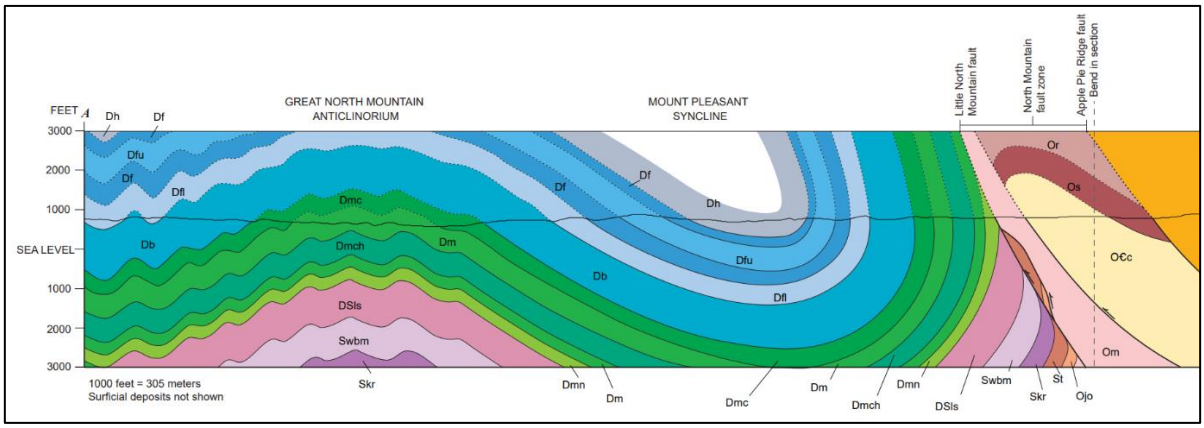
Kód	Név	Kőzetek	Terület (km2)
Mbf	Bluefield Formation, undifferentiated	meszes agyagpala, meszes iszapkő, aleurolit	2,260
Mbs	Bluestone Formation, undifferentiated	iszapkő, agyagpala, aleurolit, homokkő	48,235
Mhl	Hinton Formation, Lower Hinton member	iszapkő, homokkő, mészkő	43,868
Mhlsg	Hinton Formation, Little Stone Gap Member, Avis Limestone	mészkő, meszes agyagpala, homokkő	3,511
Mhsg	Hinton Formation, Stony Gap Sandstone Member	arenit	2,495
Mhu	Hinton Formation, Upper Hinton member	iszapkő, homokkő, mészkő	33,695
Mpn	Princeton Formation, undifferentiated	homokkő, konglomerátum, agyagpala, iszapkő	13,058
PNnrp	New River Formation, Pineville Sandstone	arenit	0,890
PNp	Pocahontas Formation, undifferentiated	homokkő, aleurolit, agyagpala, iszapkő, szén, sziderit	15,799

2. táblázat: A 2-es számú területen előforduló formációk.

3. White Hall (Virginia, Nyugat-Virginia)

A 3-as számú terület a 2-es területtől nem messze, 300 km-re északkeletre található. A vizsgált terület Virginia államban, a nyugat-virginiai határ közvetlen szomszédságában található, bő 100 km-re északnyugatra a fővárostól, Washingtontól. A terület az Appalache-

hegység Valley and Ridge (völgyek és gerincek) régió hullámzó és gerincekkel szabdalt vidékének keleti szélén helyezkedik el, délkeleti része pedig a Shenandoah-völgy lapos térszínébe nyúlik bele, amely már a Great Valley (Nagy-völgy) elnevezésű táj része. A terület legmagasabb pontja az északon található 417 m magas Little Mountain, a legalacsonyabb pont pedig 159 m-es magasságban található. Említésre méltó továbbá a terület keleti oldalán a North Mountain északról benyúló közel 400 m magasságú gerince. E két kiemelkedés között található a Back-patak (Back Creek) völgye, amely kisvízfolyások mellékvölgyeivel sűrűn szabdalt. A hegység Valley and Ridge régióját „Öreg-Appalache" néven is szokás emlegetni. A hegységképződés első szakaszának időszaka a késő paleozoikumra tehető, ekkor ütközött ugyanis az Észak-amerikai-lemez az Eurázsiai-lemezzel és az Afrikai-lemezzel. A legidősebbnek számító gneisz mellett gránit és különféle palák jöttek létre; a mészkő márvánnyá, a homokkő pedig kvarcittá alakult. A területen arra is látunk példát, hogy a vetődések és gyűrődések következményeként az idősebb kőzettömegek a fiatalabb rétegekre torlódtak áttolt redőként. Erről árulkodik a területen található több északkelet-délnyugat csapásirányú vetősík. A terület felszínközeli kőzetei főként a tengeri és folyóvízi üledékek közül kerülnek ki. A vidék hegygerinceit nagyrészt az ellenállóbb homokkő és kvarcit alkotja. A vizsgált terület mészkővel borított térszínein karsztos jelenségek is megfigyelhetők, mint például a dolinák és a víznyelők. A mai térszín kialakításában az erózió is fontos szerepet játszott (Doctor D. H., Orndorff R. C., Parker R. A., Weary D. J., Repetsky J. E., 2010). A modellezett terület kiterjedése 130,426 km^2. A terület rétegszerkezetét a 4. ábra jellemzi, az előforduló formációk pedig a 3. táblázatban olvashatók.



4. ábra: Részlet a 3-as számú terület geológiai térképének egyik metszetéből (forrás: Geologic Map of the White Hall Quadrangle). A jelek magyarázatát lásd a 3. táblázatban.

Kód	Név	Kőzetek	Terület (km2)
Ce	Elbrook Formation	mészkő, dolomit, agyagpala	11,834
Db	Brallier Formation	agyagpala, aleurolit, homokkő	18,224
Df	Foreknobs Formation	homokkő, aleurolit, agyagpala	9,329
Dfl	Foreknobs Formation	homokkő, aleurolit, agyagpala	7,831
Dfu	Foreknobs Formation	homokkő, aleurolit, agyagpala	8,879
Dh	Hampshire Formation	homokkő, iszapkő, aleurolit	18,088
Dm	Mahantango Formation	agyagpala, aleurolit	17,894
Dmc	Mahantango Formation	homokkő, aleurolit	11,518
Dmch	Mahantango Formation	aleurolit, agyagpala	8,796
Dmn	Marcellus Shale and Needmore Shale, undivided	agyagpala, mészkő	0,72
DSls	Oriskany Sandstone, Helderberg Group, Tonoloway Limestone, undivided	mészkő, homokkő, agyagpala	1,604
Oc	Chambersburg Formation	mészkő, agyagpala	0,194
OCc	Conococheague Limestone	mészkő, dolomit, homokkő	12,928
Ojo	Juniata Formation and Oswego Sandstone, undivided	homokkő, konglomerátum, agyagpala	0,993
Om	Martinsburg Formation	agyagpala, homokkő, aleurolit	9,276
On	New Market Limestone	mészkő	0,074
Op	Pinesburg Station Dolomite	dolomit	0,318
Or	Rockdale Run Formation	mészkő, dolomit	2,124
Os	Stonehenge Limestone	mészkő	3,727
Qal	Alluvium	agyag, iszap, homok	10,117
Qt	Terrace Deposits	homok, iszap, agyag	0,06
Skr	Keefer Sandstone and Rose Hill Formation, undivided	homokkő, kvarcit, agyagpala	1,782
St	Tuscarora Quartzite	kvarcit, konglomerátum	1,261
Swbm	Wills Creek, Bloomsburg, and McKenzie Formations, undivided	agyagpala, homokkő, aleurolit, mészkő	1,294

3. táblázat: A 3-as számú területen előforduló formációk.

Felhasznált adatok

Ebben a fejezetben ismertetésre kerülnek a domborzatmodellek és a későbbi morfometriai változók alapjául szolgáló LiDAR adatok, valamint a tanulóterületek létrehozásához és az osztályozás referenciáiként szolgáló geológiai térképek és a hozzájuk kapcsolódó térinformatikai adatbázisok. Jelen dolgozat kizárólag nyílt hozzáférésű, ingyenesen letölthető adatokon alapszik. Az Amerikai Egyesült Államok az élen jár az ilyen típusú tudományos adatszolgáltatásban, így az adatok letöltése nem ütközött akadályokba.

LiDAR adatok

A LiDAR adatokat a USGS (United States Geological Survey – az Amerikai Egyesült Államok Földtani Intézete) 3DEP LidarExplorer nevű online felületéről töltöttem le (USGS LidarExplorer, 2022). A 3DEP (3-Dimension Elevation Program, azaz Háromdimenziós Magasság Program) projekt célja kiszolgálni az egyre növekvő igényeket a nagy felbontású magassági adatok iránt (USGS 3D Elevation Program, 2022). A cél az ország minden államára kiterjedő adatszolgáltatás mind a LiDAR pontfelhők, mind a domborzatmodellek tekintetében. A weboldalon egy interaktív térképes felület fogad minket, ahol kiválaszthatjuk, hogy milyen típusú adatot szeretnénk letölteni, majd meg kell adnunk a kívánt területet. A böngészés megkönnyítése érdekében bepipálhatjuk a „Show where Lidar is available" felirat melletti négyzetet; ennek eredményeként láthatóak lesznek a térképen a letölthető LiDAR adatok minőség szerint kategorizálva. A USGS három minőségi kategóriát különböztet meg: QL 1 (a térképen sötétzöld, ez a legnagyobb térbeli felbontású, körülbelül 0,5 és 1 m közötti), QL 2 (világoszöld, 1 m körüli felbontással), illetve QL 3 (limezöld, ez pedig a kisebb felbontást jelenti, ami körülbelül 1–3 m-t jelent). Az adatbázisban megtalálható minden projekt teljes egészében légi lézerszkenneléses (ALS – Airborne Laser Scanning) módszerrel készült. A kívánt terület kiválasztása után linkeken keresztül juthatunk el a metaadatok, valamint a LiDAR adatok letöltéséhez. A fájlok a lézerszkenneléses adatok tömörített formátumában, .laz kiterjesztésben érhetőek el. A .laz fájlok letöltése előtt azonban a „csempék" elhelyezkedését tartalmazó .shp fájlt töltöttem le, amelyből meghatározható volt, hogy mely fájlokra van szükségem. A felhasznált LiDAR projektek legfontosabb adatait az alább ismertetem. Ezek az információk a letöltés előtt olvashatók a LidarExplorer felületén.

1-es terület:

- projekt neve: WA Western South 2016
- minőség: QL 1
- adatgyűjtés ideje: 2016. 03. 17. – 2016. 05. 28.
- összesen letöltve: 93 db .laz fájl, méret: 15,2 GB

2-es terület:

- projekt neve: WV FEMA R3 East 2016
- minőség: QL 2
- adatgyűjtés ideje: 2016. 11. 22. – 2016. 12. 28.
- összesen letöltve: 176 db .laz fájl, méret: 11,4 GB

3-as terület:

- projekt neve: VA NShenandoah 1 2020
- minőség: QL 2
- adatgyűjtés ideje: 2020. 11. 29. – 2021. 01. 12.
- összesen letöltve: 78 db .laz fájl, méret: 5,7 GB

Látható tehát, hogy a lézerszkennelt adatok viszonylag naprakésznek mondhatóak. Mindhárom terület .laz fájljai kellően részletesnek bizonyultak, az esetek túlnyomó többségében szubméteres, legrosszabb esetekben is 1 m körüli térbeli felbontással.

Geológiai térképek és térinformatikai adatbázisok

A vizsgált területek geológiai térképeit és GIS adatbázisait szintén egy, a USGS által üzemeltetett felületről töltöttem le. A forrás a National Geologic Map Database (NGMDB), azaz a Nemzeti Geológiaitérkép-adatbázis nevű interaktív weboldal volt, melynek a „MapView” menüpontját használtam a térképek kiválasztásához. Az oldal a működését tekintve hasonló a LidarExplorerhez; egy térképes felületen kell elnavigálnunk a számunkra érdekes területekhez, miközben láthatjuk az adott területekről elérhető geológiai térképeket. A felületen különféle szűrők használatával szűkíthetjük a találatokat, mint például a szerző, a méretarány vagy a készítés idejének a megadása. A méretarányt a lehető legnagyobbra, 1 : 24 000-esre állítottam be, a cél ugyanis a minél részletesebb adatbázis letöltése volt. A választott térkép megnyitása után az „External resources” mező alatti linkek vezetnek a megfelelő webhelyre, ahonnan letölthetők a térképek .pdf vagy raszteres formátumban, a GIS adatbázisok, valamint a metaadatok és a kiegészítő dokumentumok egyaránt. Az alábbiakban ismertetem a felhasznált térképek fontosabb paramétereit.

1-es terület: Geologic map of the Lacamas Creek quadrangle, Clark County, Washington (A Lacamas-patak-szelvény geológiai térképe, Clark megye, Washington állam)

- szerző: Russell C. Evarts
- kiadó: USGS
- sorozat és azonosító: Scientific Investigations Map SIM-2924 (Tudományos Vizsgálatok térképsorozat)
- kiadás éve: 2006
- méretarány: 1 : 24 000

2-es terület: Bedrock Geology of the New River Gorge National River, Map Sheet 4: Hinton and Talcott 7.5' Quadrangles, West Virginia (A New River-szurdok Természetvédelmi Terület fedetlen földtani térképe, 4-es számú térképlap: Hinton és Talcott 7,5’-es szelvények, Nyugat-Virginia)

- szerző: Robert L. Peck, David L. Matchen, Paula J. Hunt, Sarah E. Gooding
- kiadó: West Virginia Geological and Economic Survey (Nyugat-Virginiai Geoloógiai és Gazdasági Intézet)
- sorozat és azonosító: Geologic Quadrangle Maps of West Virginia OF 1301 (Nyugat-Virginia geológiai térképszelvényei)
- kiadás éve: 2013
- méretarány: 1 : 24 000

3-as terület: Geologic map of the White Hall quadrangle, Frederick County, Virginia, and Berkeley County, West Virginia (A White Hall-szelvény geológiai térképe, Frederick megye, Virginia és Berkeley megye, Nyugat-Virginia)

- szerző: Daniel H. Doctor, Randall C. Orndorff, Ronald A. Parker, David J. Weary, John E. Repetski
- kiadó: USGS
- sorozat és azonosító: Open-File Report OF-2010-1265
- kiadás éve: 2010
- méretarány: 1 : 24 000

A későbbi modellezés során a fent felsorolt térképekkel tartalmilag megegyező térinformatikai adatbázisokat használtam fel, melyek a fent felsorolt geológiai térképek digitalizált változatai. Többféle tartalmú vektoros állomány volt elérhető mindhárom területről (például vetősíkok, kémiai mérési pontok rétegei, fúráspontok), nekem azonban csak a kőzet- és formációtípusok poligonjait tartalmazó alábbi .shp fájlokra volt szükségem.

- 1-es terület: LacamasCrkGeologyPolys.shp (dátum és vetület: Clarke 1866, UTM zone 10N)
- 2-es terület: NERI_bedrock_geologic_units.shp (dátum és vetület: NAD83, UTM zone 17N; EPSG-kód: 26917)

- 3-as terület: BedrockMapUnitPolys.shp, SurficialMapPolys.shp (dátum és vetület: NAD83, UTM zone 17N; EPSG-kód: 26917)

Alkalmazott módszerek

Ebben a fejezetben elsőként bemutatásra kerülnek az adatok, illetve a modellezés előkészítésének lépései. Szó lesz a domborzatmodellekről, majd az ezekből később létrehozott morfometriai változókról, a tanulóterületek meghatározásáról, illetve a térinformatikai vektoros adatbázisokon végzett műveletekről. A fejezet második felében a random forest osztályozás általános ismertetésére kerül sor, majd az általam lefuttatott modellek, azok részeredményei, valamint a legjobbra becsült modellek pontosságvizsgálatának módszere kerül kifejtésre.

Domborzatmodellek előállítása

A domborzatmodellek (DEM – Digital Elevation Model) létrehozásához a Global Mapper 22-es verziójú szoftverét használtam. A szoftver jól kezeli a nagyobb méretű .laz fájlokat is, így ideálisnak bizonyult a LiDAR adatokból történő domborzatmodellek előállításához. A fájlok betöltésekor a „lidar load options" ablakban a „select LiDAR point classifications to import" lehetőségnél kiválasztottam a „2 - ground", azaz talajpontok opciót, ugyanis a létrehozandó domborzatmodellnek csak a talaj pontjait tartalmazó modellnek, tehát terepmodellnek (DTM – Digital Terrain Model) kell lennie a kutatás geológiai témájából eredően. Ebből fakadóan a növényzet és az épületek pontjait semmiképp sem vehetjük figyelembe. A pontfelhő számomra hasznos részének, tehát a talajpontoknak az adatsűrűsége, felbontása ezzel a lépéssel nem változott. A .laz fájlok betöltése után a vetületi rendszer megváltoztatása volt a következő lépés. Az egységesség érdekében az új vetület minden területen az UTM (Universal Transverse Mercator) megfelelő zónája lesz a WGS84 ellipszoidon értelmezve. Ezt a lépést a tools → configure → projection menüben tudjuk megtenni. Fontos megjegyezni, hogy ezzel csak a megjelenítés vetületét állítottuk át, a fájlokét nem! Az ezután exportált fájlok azonban már az újonnan megadott vetületetben lesznek értelmezve.

A terepmodellek létrehozásához az analysis → create elevation grid from 3D vector/LiDAR data eszközt használtam. A fájlok kiválasztása után megjelenő „grid creation options" ablak „grid options" füle alatt a „grid method" lehetőségnél a „binning (minimum value - DTM)" opciót választottam, majd a „grid spacing" beállításnál a „manually specify the grid spacing to use" lehetőség kiválasztása után horizontális felbontásként 1 m-t adtam meg. A könnyebb

kezelhetőség érdekében a létrehozandó modell kiterjedését (extents) egész méterben határoztam meg. Ehhez a „grid bounds” fülnél a „current projection (UTM - meters)” opció kiválasztása után megadhatjuk a kívánt határokat minden irányban. A programban létrejött ideiglenes állományt ezután még exportálnunk kell, melyet a rétegen történő jobb kattintás után a layer → export paranccsal tudunk megtenni. A kívánt formátum (jelen esetben geotiff) kiválasztása után a „geotiff export options” ablakban a „geotiff options” fül alatt bepipálandó az „always generate square pixels”, hogy minden esetben négyzet alakú képpontokat kapjunk, valamint az „interpolate to fill small gaps in data” annak érdekében, hogy az esetlegesen előforduló kisebb adathiányos területeket egy algoritmus interpoláció útján kitöltse. Előbbire azért volt szükség, mert az általam használt UTM térképi koordinátarendszer metrikus alapú, tehát az általam megadott határoknak és felbontásnak köszönhetően így minden képpont (méterben) meghatározható lesz egész számú koordinátákkal.

GIS-rendszer létrehozása és térinformatikai műveletek

A formációkat tartalmazó adatbázison végzett műveletekhez, illetve az egyéb térinformatikai munkafázisokhoz minden területen létrehoztam egy-egy QGIS-projektfájlt. A QGIS egy nyílt forráskódú térinformatikai szoftver, melynek a 3.16-os verzióját használtam. A projektfájlok, valamint minden felhasznált vagy létrehozott vektoros, illetve raszteres állomány vetületeit az alábbiakban határoztam meg. Amennyiben szükséges volt, úgy a vector → data management tools → reproject layer eljárást alkalmaztam a rétegek cél-koordinátarendszerbe történő átvetítéséhez.

- 1-es terület: WGS 84 / UTM zone 10N, EPSG: 32610
- 2-es terület: WGS 84 / UTM zone 17N, EPSG: 32617
- 3-as terület: WGS 84 / UTM zone 17N, EPSG: 32617

A geológiai adatbázisok és a LiDAR adatok, tehát az abból létrehozott domborzatmodellek az 1-es és a 3-as területen nem fedik egymást teljes egészében, ugyanis ezeken a területeken a LiDAR felmérések nem terjedtek ki a teljes vizsgált területre. Ennek összeegyeztetése érdekében létrehoztam egy vágási poligont a DTM határvonalát követve, majd ezzel a poligonnal megvágtam a formációk poligonjait. A 2-es terület geológiai rétegei pedig a teljes térképművet (a többi szomszédos szelvény) tartalmazták, így azt is meg kellett vágni a vizsgált térképszelvény területével.

Az 1-es területen megtalálható „Qls” (landslide deposits – holocén és pleisztocén kori tömegmozgásos törmelék) és „Qls?” kódú (bizonytalan meghatározású „Qls”) kategóriák összevonásra kerültek.

A 3-as terület térképén megtalálható harmad- és negyedidőszaki „Qal” (alluvium – folyóvízi hordalék), illetve „Qt” (teraszos üledékek) kódokkal illetett poligonok – felszíni képződmények lévén – egy másik rétegen (SurficialMapPolys) kaptak helyet. Ezek a poligonok átfednek a „BedrockMapUnitPolys” réteg poligonjaival. A két réteg egyesítéséhez a „symmetrical difference” (szimmetrikus különbség) eljárást alkalmaztam. Ugyanerről a vizsgált területről létezik egy további említésre méltó poligon-réteg, amely végül nem került feldolgozásra. A felszíni breccsa előfordulásait tartalmazó „BrecciaPolys” réteg poligonjai ugyanis túl kicsik ahhoz, hogy a későbbi osztályozáskor egy különálló geológiai kategóriát határozzanak meg, emellett pedig fedettek, a felszínen nem láthatóak (a poligonok legtöbbje beépített vagy megművelt területen található), tehát semmilyen morfológiai nyoma nincsen a jelenlétüknek.

Az egy adott formáció-kategórián belüli poligonokat mindhárom területen összevontam, így többrészű poligonok jöttek létre. Ezután minden formáció-poligont egyedi azonosítóval láttam el: a mező kalkulátorral egy új, integer típusú, „ID” nevű mezőt hoztam létre, melynek értékeit a „$ID” kifejezéssel töltöttem fel. Továbbá töröltem minden, a jelen kutatás szempontjából felesleges információt tartalmazó mezőt mindhárom terület adatbázisából.

Az 1-es és a 2-es számú terület számottevő kiterjedésű vízfelületeket tartalmaz, így a későbbi osztályozás pontossága érdekében ezeket kivágtam a formációk poligonjaiból. Ugyan geológiai értelemben véve ezek alatt is jelen van az alapkőzet, viszont ezt a LiDAR-ból generált morfometriai változók segítségével nem tudjuk lemodellezni.

Annak ellenére, hogy a DTM-ek létrehozásánál bepipáltam a „fill small gaps” opciót, a raszteren lévő nagyobb rések továbbra is megmaradtak. A terepmodell külső határain belül nem megengedhetőek a „no data” (ez bináris fájlokban legtöbbször egy negatív, tízezres nagyságrendű szám) értékű képpontok, ezért szükségszerű volt ezek kitöltése. Ebben a raster → analysis → fill nodata (azaz a „nincs adat” képpontok kitöltése) eljárás volt a segítségemre. Az algoritmus IDW interpolációval (inverse distance weighting, ami a távolsággal fordítottan arányos súlyozást jelent) számolja ki a keresett pixelek értékeit. Itt a „maximum distance (in pixels) to search out for values to interpolate” lehetőségnél paraméterként meg kell adni, hogy mekkora keresési sugárral dolgozzon az algoritmus. Itt a 150 pixeles (azaz 150 méteres) sugár megadása elegendő volt. Mivel az eljárás nem csak a belső határokon belül interpolált, hanem

a külső határokon túl is extrapolált, így egy nem kívánt „pufferzóna” is létrejött a DTM körül. Ez a korábban említett vágási poligonnal levágásra került. A módszer buktatója lehet, hogy vízfelületek esetében nem valós, ferde felszíneket eredményezhet. Ez a mi esetünkben nem okozott problémát, ugyanis a vízfelületeket már korábban kivágtuk a poligonokból.

Tanulóterületek létrehozása

A random forest osztályozás előkészítésének első lépéseként tanulóterületeket kellett létrehozni. Tanulóterületeknek azokat a – teljes vizsgált területhez képest – kisebb területeket nevezzük, amelyeken a különböző prediktor változók (esetünkben a morfometriai változók) adott pontokon felvett értékeinek figyelembe vételével betanítjuk az osztályozó algoritmust. A betanulás eredményeként később minden osztály (formációtípus) minden változójáról kialakul egy jellemző értéktartomány. Amennyiben az osztályok láthatóan elkülönülnek bizonyos morfometriai paraméterek szerint, úgy a tanulóterületeket manuálisan is kijelölhetjük a rá jellemző területek alapján, azaz magunk határozhatjuk meg azok helyét és méretét. Fontos, hogy a tanulóterületek olyan területek legyenek, ahol a változók értékei a lehető legjobban jellemzik az adott osztályt.

A tanulóterületek számát minden területen 10 db-ban határoztam meg formációtípusonként. Ezek elhelyezéséhez elsőként a QGIS vector → research tools → random points in polygons algoritmusát alkalmaztam. Az eljárás a kiválasztott réteg minden poligonján belül elhelyezi a megadott számú pontot, random módon meghatározott helyekre. A korábban említett poligon-összevonásnak köszönhetően minden formáció egy poligonként szerepel az adatbázisban, tehát a random pontok is ennek megfelelően lettek elhelyezve formációnként egyenlő számban. Ezek a pontok öröklik annak a poligonnak minden attribútumát, amelyben el lettek helyezve. A mi esetünkben ezt azért fontos megjegyezni, mert az „ID” mezőnek köszönhetően a pontok és a poligonok összekapcsolhatók egymással.

Ezzel létrehoztam minden vizsgált területen a tanulóterületek alapjául szolgáló „training_points” réteget. A következő lépés az elhelyezett pontok manuális felülvizsgálata volt. A lakott területek és az egyéb emberi beavatkozások nyomainak jelenléte miatt egyes pontokat át kellett helyezni. Ilyenek például az utakra vagy épületekre lerakott pontok (lásd 5. ábra). Továbbá a formációk határaihoz túl közel eső pontokat is új helyre kellett helyezni, ellenkező esetben a későbbi tanulóterület átlógna a szomszédos poligonba. Ez mindenképpen elkerülendő, ugyanis pontatlanságokhoz vezetne.

5. ábra: Nem megfelelő helyekre generált random pontok.

A tanulóterületek alakja és mérete egységes körökben lett meghatározva, ezek poligonjainak létrehozásához a meglévő pontok köré generáltam pufferzónát a vector → geoprocessing → buffer eljárással. A később végrehajtott random forest osztályozás paraméterezési lehetőségeinek bővítése érdekében három eltérő méretű tanulóterületet hoztam létre 2,5, 5, illetve 10 m-es sugárral. Ezek alapterületei 19,63 m^2, 78,54 m^2, illetve 314,16 m^2 lettek.

A tanulóterületek poligon-rétegei tehát az átmérőikre utalva a „training_areas_5m", a „training_areas_10m", illetve a „training_areas_20m" nevet kapták.

Morfometriai változók létrehozása

A morfometria szó jelentése alakmérés, és lényegében jelentheti bármilyen tárgy felszínének vagy alakjának a megmérését (Kertész Á., Karátson D., 1997). Jelen dolgozatban azonban a földfelszín alakját vizsgáljuk, tehát helyesen mondva geomorfometriáról beszélünk, azonban ez az előtag gyakran lemarad a szakirodalomban. Egy definíció szerint a geomorfometria a kvantitatív földfelszín-elemzés tudománya (Pike R. J., Evans I. S., Hengl T., 2009). A morfometria szó alatt a dolgozatban geomorfometriát értek.

A morfometriai változók pedig olyan, helyről helyre (esetünkben pixelről pixelre) változó számértékek, amelyek egy domborzatmodellből kiszámíthatók egy adott pontnak (legtöbbször grid-képpontnak) a szűkebb vagy bővebb környezetéhez való viszonya alapján. Ilyen levezetett változó például a lejtőszög, a kitettség vagy a görbület (a lejtőszög megváltozása), de ebben a dolgozatban a tengerszint feletti magasság is ezekkel egyenértékűnek tekintendő.

A morfográfia (geomorfográfia) pedig a földfelszín kvalitatív osztályozásával foglalkozó tudományág, tehát minőségi kategóriákat határoz meg (Pavlopoulos K., Evelpidou N., Vassilopoulos A., 2009). Ez azért említendő, mert szigorú értelemben véve a morfometriai

változók mellett néhány morfográfiai változóval is dolgoztam, mint például a geomorfonok vagy a „fuzzy landform element classification” felszínosztályozó eljárása.

A morfometriai változók kiszámítását a SAGA GIS (System for Automated Geoscientific Analyses, azaz automatizált földtudományi elemzőrendszer) 8.1.1-es verziószámú nyílt forráskódú szoftverben végeztem el. Ahogyan erre a neve is utal, a szoftver fő erőssége az automatizált elemzések végrehajtása.

A SAGA „tools” (elemzőeszközök) főmenüjének „terrain analysis” (földfelszínelemzés) eszköztárát használtam a változók létrehozásához. Ezen belül további menüpontok találhatóak, melyek közül az alábbi eszköztárak algoritmusait használtam. A változók meghatározásakor törekedtem a már bevált, ismert metódusok alkalmazására, illetve az adott területek morfometriai paramétereit minél több „szemszögből” vizsgáló eljárások kiválasztására. A morfometriai változókat – a nem mindig egyértelmű magyar elnevezéseik miatt – a továbbiakban legtöbbször az angol nevükön fognak említésre kerülni. Maguk a felhasznált morfometriai változók az alább félkövér betűtípussal szerepelnek, melyek részletes bemutatása nem a dolgozat része.

- Hydrology:
    - **SAGA Wetness Index** – nedvességi index
- Lighting, Visibility:
    - **Geomorphons** – geomorfonok szerinti felszínosztályozás
- Morphometry:
    - **Convergence Index** – a kitettségen alapuló összefolyási index
    - **Downslope Distance Gradient** – lejtőhossz-gradiens
    - **Fuzzy Landform Element Classification** – tagolt felszínforma-osztályozás
    - **Morphometric Protection Index** – morfometriai védettségi index
    - Relative Heights and Slope Positions – relatív magasságok és lejtőpozíciók
        - **Slope Height** – lejtőmagasság
        - **Mid-Slope Position** – lejtőn elfoglalt pozíció
    - Slope, Aspect, Curvature – lejtőszög, kitettség, görbület
        - **Slope** – lejtőszög
        - **Aspect** – kitettség
        - **General Curvature** – általános görbület, a vízszintes és a lejtőirányú görbület kombinációja

        - **Profile Curvature** – lejtőirányú görbület
        - **Plan Curvature** – vízszintes görbület
        - **Tangential Curvature** – érintőirányú görbület
        - **Longitudinal Curvature** – hosszanti görbület, a lejtőirányú görbület számításához hasonló algoritmus
        - **Cross-Sectional Curvature** – keresztirányú görbület, a vízszintes görbület számításához hasonló algoritmus
        - **Minimal Curvature** – a lejtőn mért legkisebb görbület egy keresési sugáron belül
        - **Maximal Curvature** – a lejtőn mért legnagyobb görbület egy keresési sugáron belül
        - **Total Curvature** – teljes görbület
        - **Flow Line Curvature** – folyásirányú görbület, a lejtőirányú görbület számításához hasonló algoritmus
    - **Terrain Ruggedness Index (TRI)** – felszíntagoltsági mutató, a teljes magasságkülönbség (relief) egy keresési sugáron belül
    - **Terrain Surface Convexity** – a földfelszín konvex felületeinek aránya
    - **Terrain Surface Texture** – a földfelszín textúrája, azaz mélyedések és kiemelkedések gyakorisága egy keresési sugáron belül
    - **Topographic Position Index (TPI)** – topográfiai pozíció index, amely egy keresési sugáron belül vizsgálja egy pont szomszédsági viszonyait
        - ebből a változóból két eltérő változatot is felhasználtam egyidejűleg (lásd lentebb)
    - Upslope and Downslope Curvature – lejtő- és emelkedőirányú görbületek
        - **Local Curvature** – helyi görbület, a szomszédos cellákba mutató gradiens vektorok összege
        - **Upslope Curvature** – emelkedőirányú görbület, a helyi görbület gradiensének átlaga a cella emelkedőirányú környezetéből
        - **Local Upslope Curvature** – helyi emelkedőirányú görbület, az emelkedőirányú görbületnek csak a szomszédos cellákkal számoló változata
        - **Downslope Curvature** – a lejtőirányú görbülethez hasonló, a helyi görbület gradiensének átlaga a cella lejtőirányú környezetéről

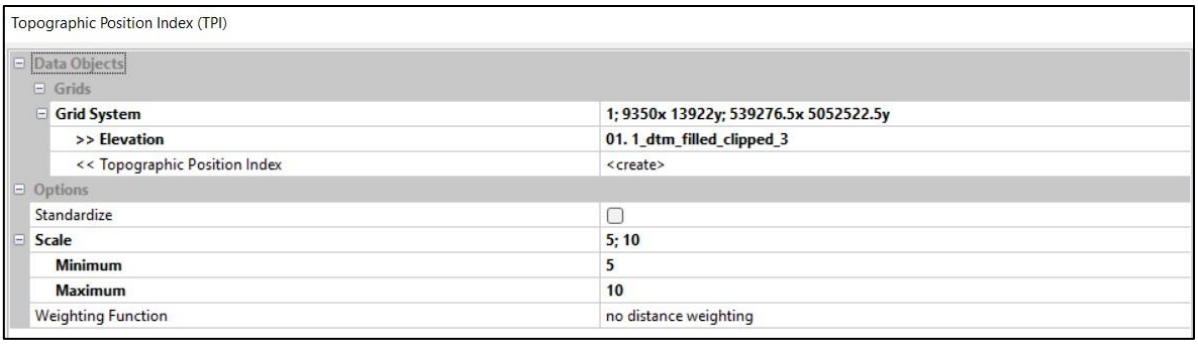
        - **Local Downslope Curvature** – helyi lejtőirányú görbület, a lejtőirányú görbületnek csak a szomszédos cellákkal számoló változata
    - **Vector Ruggedness Measure (VRM)** – a lejtőszögből és a kitettségből számolt felszíntagoltsági mutató

A 32. változó, a magasság pedig a korábban létrehozott DTM formájában került felhasználásra, ami továbbá a többi változó alapját is jelentette.

A fentebb felsorolt eljárások mindegyike egy grid típusú domborzatmodellt kér bemeneti fájlként, ezen felül pedig különféle paraméterek beállításával módosíthatjuk a kapott eredményt. A legtöbb algoritmust az alapértelmezett beállításokkal futtattam, néhány esetben azonban ettől eltértem. Ezek közül az alább olvashatóak a jelentősebb változtatások.

A „Terrain Ruggedness Index”, illetve a „Vector Ruggedness Measure” esetében a keresési sugarat (search radius) a tanulóterületek méretéhez igazítottam (itt 3, 5, és 10 m-es sugár felelt meg az 5, 10 és 20 m átmérőjű tanulóterületeknek), tehát ezekből a változókból három eltérő verziót hoztam létre területenként. Legelső esetben 3 m-t adtam meg az optimális 2,5 m helyett, ugyanis a SAGA ezen paramétere csak egész számot fogadott el bemenetként (ez az érték a pixelszámra vonatkozik). A geomorfonok (Geomorphons) létrehozásakor a sugárirányú határértéket (radial limit) 200 m-ben határoztam meg, míg a „Morphometric Protection Index” keresési sugara 10 m-ben lett megadva. A „Topographic Position Index” 5 és 10 m közötti, illetve 90 és 100 m közötti tartományban (scale: minimum, scale: maximum) is futtatásra került, tehát a sugárirányú távolságok között értelmezett gyűrű magasságértékeivel számolt az algoritmus.

A morfometriai változókat kiszámító algoritmusok futtatásának eredményeként raszteres fájlok jöttek létre, melyeket egyenként exportáltam a SAGA Grid (.sgrd) fájlként. A létrejött gridek térbeli vízszintes felbontása a bemeneti domborzatmodellek felbontásával megegyező, tehát 1 m-es. A cél a random forest osztályozás egyik modellparamétere okán 32 morfometriai változó létrehozása volt, amely teljesült.



6. ábra: A morfometriai változók létrehozásának párbeszédablaka a SAGA-ban. A képen példaként a "Topographic Position Index" látható.

A random forest osztályozás bemutatása (random forest classification - RFC)

A random forest vagy „véletlen erdő” algoritmus (random forest classification – a továbbiakban: RFC) egy 2001-ben kidolgozott, gépi tanulási módszereken alapuló osztályozó algoritmus, amely az egyedülálló döntési fák helyett azok egész együttesét építi fel, azaz egy erdőt hoz létre a döntéshozatal során (Breiman, 2001).

A döntési fáknak megvan az az előnye, hogy könnyű őket felépíteni és felhasználni. Azonban a tévesztés, az esetleges pontatlanság lehetősége miatt ezek alkalmazása nem a legelőnyösebb választás a prediktív tanulás céljára (Hastie T., Tibshirani R., Friedman J., 2009). Ugyan kitűnően működnek azoknak az adatoknak az osztályozásánál, amelyekkel létrehozták őket, új minták besorolásánál azonban rugalmatlannak bizonyul az alkalmazásuk. Az RFC a döntési fák egyszerűségét kombinálja a rugalmassággal, amely számottevő javulást eredményez a pontosságban. Az eljárás a felügyelt osztályozások közé tartozik, tehát az osztályok, amelyekbe be kívánjuk sorolni az új, ismeretlen osztályú adatpontot, általunk előre meghatározottak. Az RFC egyaránt használható kvantitatív (számszerű), illetve kvalitatív (minőségi) adatok osztályozására. Előbbi esetében regresszióról (regression), míg utóbbi esetében klasszifikációról (classification) beszélünk. Az RFC hatékony eljárásnak bizonyul a földtudományokban, különösképpen a kvalitatív geológiai osztályozásban (Cracknell M. J., Reading A. M., 2014).

Az algoritmus betanulásának első lépésében a tanulóadatokból, esetünkben a tanulóterületek képpontjaiból történő véletlenszerű mintavételre kerül sor, amely történhet visszatétellel (sampling with replacement) vagy visszatétel nélkül. Előbbi esetében ezt a lépést a szakirodalom a „bootstrapping” kifejezéssel illeti. Az ezt követő lépésben az algoritmus felépíti a döntési fákat a kiválasztott mintaadatokon alapulva, azonban a fák elágazásainál (node) a

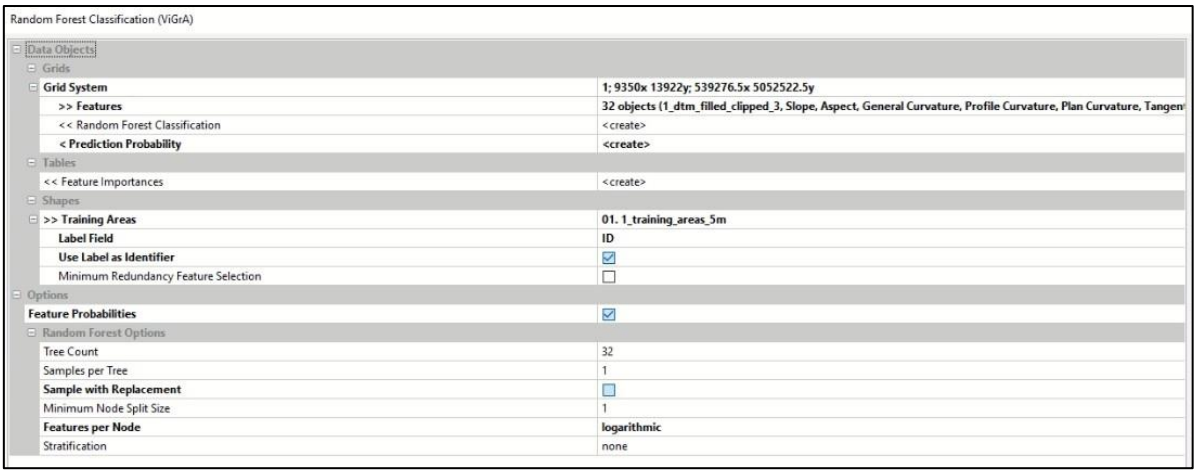
változóknak csak egy véletlenszerűen kiválasztott részhalmazát felhasználva. Új minta osztályozása esetében lefuttatjuk az új adatot a fákon. A döntéshozatalnál a fák közül a legtöbb szavazatot kapó osztály lesz a győztes. A „bootstrapping” és az eredmények aggregációját követő döntéshozatal együttes elnevezése a szakirodalomban a „bagging”.

Az eljárás eredményeként a fák változatos variációi jönnek létre, melyek között a korreláció az esetek többségében igen alacsony. Ez a sokféleség teszi az RFC-t az egyszerű döntési fáknál sokkal hatékonyabb eljárássá. A random forest az önálló döntési fákkal ellentétben nem érzékeny a betanítási adatokban történő változtatásokra. Amennyiben egy döntési fában kevésbé fontos változók szerepelnek, akkor hibás döntés születhet, de a változatosságnak köszönhetően lesznek ugyanilyen eredetű hibák „ellentétes” irányban is, ezáltal a téves döntések a végső döntéshozatalban kiegyenlítik egymást. Felmerülhet a kérdés, hogy az algoritmus miért a random forest nevet kapta? Ennek egyszerű oka a tanulóadatok, illetve a felhasznált változók véletlenszerű (random) kiválasztásában rejlik.

A modellezés végrehajtása

A random forest osztályozásokat szintén a SAGA-ban hajtottam végre. Az ehhez szükséges eszközt az alábbi úton érjük el a program feldolgozó eszköztárában: imagery → ViGrA → random forest classification (ViGrA). A ViGrA a „Vision with Generic Algorithms” angol kifejezés rövidítése, amely egy SAGA által is használt újkeletű, gépi látáson alapuló képfeldolgozó- és elemző algoritmus-könyvtár. Ennek fő erőssége az algoritmusok és az adatstruktúrák testreszabhatósága (VIGRA Homepage, 2021).

Az eszközt megnyitva a párbeszédablakban számos beállítás fogad minket (lásd 7. ábra). A paramétereket három kategóriára bontva mutatom be; elsőként a bemeneti feltételeket, majd a kimeneti eredményeket említem, végül pedig a modellezési opciókat magyarázom. Az futtatott modellek konkrét paramétereit a későbbiekben ismertetem. Egyes beállítások ismertetéséhez a SAGA online dokumentációja volt a segítségemre (SAGA-GIS Tool Library Documentation (v8.1.1), 2021).



7. ábra: A random forest osztályozás párbeszédablaka a SAGA-ban.

1. Input (bemeneti) beállítások és paraméterek:

- **Grid system**: Elsőként ki kell választanunk az általunk vizsgált terjedelmet az adott vetületi rendszer koordinátáiban.
- **Features**: Itt kell megadnunk a morfometriai változók grid-jeit.
- **Training areas**: Ennél a lehetőségnél pedig ki kell választani a tanulóterületeket tartalmazó vektoros fájlt.
    - **Label field**: A tanulóterületeknek az adott osztályokat azonosító mezője, amely csak egy egész szám lehet.
    - **Use label as identifier**: Amennyiben az előbb kiválasztott mező az osztály egyértelmű azonosítója, akkor azt itt megadhatjuk.

Fontos megjegyezni, hogy csak a programba előzetesen importált fájlok adhatók meg az input opcióknál.

2. Output (kimeneti) eredmények:

- **Random forest classification** (automatikusan létrejön): Maga az osztályozás rétege grid formátumban. Értékei egész számok, melyek megfeleltethetők a formációknak.
- **Prediction probability** (opcionális): Az osztályozás helyességének valószínűsége grid formátumban. A grid 0 és 1 közötti értékeket tartalmaz, a nagyobb értékek nagyobb valószínűséget jelentenek.
- **Feature importances** (automatikusan létrejön): Egy táblázat formájában látjuk, hogy mely változók játszottak fontos szerepet az osztályozásban, és melyek voltak kevésbé erősek.

- **Feature probabilities** (opcionális): Az osztályok valószínűségei a teljes területen, melyből annyi grid jön létre, ahány formációtípus (osztály) van az adott területen.

A kimeneti grid-eket célszerű a „create” lehetőségre állítani a futtatás előtt, ellenkező esetben (második vagy későbbi futtatás esetén) egy korábban létrehozott grid felülírása történik. Az RFC modellek minden lefuttatása után létrejöttek az előbb említett grid rétegek, illetve táblázatok. A létrejött grid-eket .sgrd formátumban, míg a táblázatokat .csv-ként exportáltam minden modellnél.

3. RFC paraméterek:

- **Minimum redundancy feature selection**: Ennek kiválasztása esetén egy algoritmus (minimum Redundancy Maximum Relevance – mRMR) megvizsgálja a megadott változókat a fontosságuk szerint, és csak a legfontosabbnak ítélteket veszi figyelembe az osztályozásnál. Ezeknek a számát mi adhatjuk meg.
- **Tree count (ntree)**: A generálni kívánt fák darabszámát jelenti.
- **Samples per tree**: Itt egy 1-nél kisebb törtszámmal megadhatjuk, hogy a tanulóterületek hanyadrészét vegye figyelembe az algoritmus a betanuláskor. Ezzel várhatóan gyorsíthatjuk annak lefutását. 1 megadása esetén a teljes tanulóterület-állományból vesz mintát az algoritmus.
- **Sample with replacement**: Megadható, hogy visszatétellel vagy anélkül történjen a mintavétel a betanulás során.
- **Minimum node split size**: Az elemek minimális száma, amely egy elágazáshoz szükséges.
- **Features per node (mtry)**: A fának egy adott elágazásánál vizsgát változók száma. A választható opciók: „logarithmic” (logaritmikus), „square root” (négyzetgyök), „all” (összes). Ez az érték az összes változó számának 10-es alapú logaritmusában vagy négyzetgyökében értendő, ezenkívül az összes változó felhasználását is megadhatjuk.
- **Stratification**: Itt megadható, hogy figyelembe legyenek-e véve a mintában található osztályok előfordulási arányai.

Az alábbiakban az állandó paramétereket sorolom fel, amelyek beállításain nem változtattam a modellezés során.

- Label field: ID
- Use label as identifier: igen

- Minimum redundancy feature selection: nem
- Feature probabilities: nem
- Samples per tree: 1
- Sample with replacement: nem
- Minimum node split size: 1
- Stratification: nincs

A különböző modellvariációk futtatásához az itt felsorolt paramétereket viszont rendre változtattam.

- Features (morfometriai változók)
- Training areas (tanulóterületek)
- Tree count (fák száma)
- Features per node (változók elágazásonként)

Az RFC modellezés első köre

Az 1-es és a 3-as terület osztályozásánál első körben mind a 32 létrehozott morfometriai változót felhasználtam. A 2-es területen annak nagyobb kiterjedése miatt a másik két területen elért becsült pontosságok függvényében határoztam meg a futtatandó modelleket, de első körben szintén mind a 32 változót felhasználva. Azonban később a 2-es területen a SAGA sajnos memóriaproblémákba ütközött (itt a morfometriai változók 32 db 1,2 GB-os rasztert jelentettek). Ebből fakadóan a teljes vizsgált terület modellezésére nem volt lehetőség, ezért annak levágása mellett döntöttem. A területcsökkentéskor úgy jártam el, hogy minden formációtípus továbbra is elegendő területtel legyen jelen a modellezéshez. Az új terület kiterjedése az eredetinek körülbelül 50%-át teszi ki. A 250 fát generáló modell 1-es és 3-as területen mutatott, a többihez képest nem kimagasló eredményeit, illetve az algoritmus hosszú futási idejét figyelembe véve nem került futtatásra a 2-es számú területen. Ellenben a fák csökkentésével 32 (ez az alapértelmezett érték a SAGA-ban) fát generáló modelleket is futtattam ezen a területen. A létrehozott modelleket a lentebb található táblázat összegzi az osztályozási valószínűségek értékeivel.

A modellezés első körében az alábbi paramétereket alkalmaztam.

- változók: 32 db (külön 3, 5 és 10 m keresési sugarú változókkal, melyek a tanulóterületek méreteivel összhangban lettek alkalmazva)
- tanulóterületek: 5, 10, 20 m

- fák száma: 32, 50, 100, 250 db
- változók száma elágazásonként: logaritmus alapú: $\log_{10} 32 = 1{,}505 \rightarrow 2$ változó

Ez utóbbi paraméter miatt törekedtünk a 32 morfometriai változó létrehozására. Az algoritmus futási ideje legnagyobb mértékben a fák számától függött, kisebb mértékben pedig a tanulóterületek mérete is befolyásolta azt. 250 fánál közel 3, míg 500 fánál 8 óra volt az osztályozás lefutásának ideje. A várhatóan legpontosabb modellek kiválasztásához az eredményként létrejött „prediction probability" (azaz előrejelzési valószínűség, a továbbiakban: PP) raszter „arithmetic mean", azaz számtani középértékeit vetettem össze. A PP-értékek a cellánként kiszámított „feature probabilities"-értékek átlagaiként lettek számítva. Alább röviden összegzem a területekre jellemző trendeket, illetve a PP-értékekre hagyatkozva a legerősebb modelleket. A „rad3" a 3 méteres keresési sugárral létrehozott morfometriai változókra utal, melyek az 5 méter átmérőjű tanulóterületek modelljeiben szerepelnek. Ennek megfelelően a „rad5" változók a 10 méteres, a „rad10" változók pedig a 20 méteres tanulóterületekkel szerepelnek együtt. A „t..." a fák számára utal. Bár az első körben minden modellt logaritmikus változószámmal futtattam, de a későbbi tévesztések elkerülése érdekében a feltüntetett neveikben a „log" is szerepel.

Az **1-es terület** legjobbra becsült modellje: **rad3 t100 log**, azaz 5 m átmérőjű tanulóterületekkel és 100 fával dolgozó modell (PP arithmetic mean: **0.04011**)

A kapott PP-értékek láthatóan és egyértelműen elkülönülnek a tanulóterületek mérete szerint. A legjobban az 5 m-es (azaz a legkisebb) tanulóterületek szerepeltek. A 10 m-es modellek ennél valamivel rosszabb, míg a legrosszabb valószínűséggel a 20 m-es modellek zártak. A fák száma alapján számottevő különbségek nem mutatkoztak.

A **2-es terület** legjobbja: **rad5 t100 log** (PP arithmetic mean: **0.06879**)

Itt már kevésbé egyértelmű a helyzet, határozott trend nem állapítható meg, de a legerősebbnek szintén az 5 m-es tanulóterületeket tartalmazó modellek bizonyultak. A 10 és 20 m-es tanulóterületek modelljei hasonló valószínűségeket mutatnak, de előbbiek PP-értékei között nagyobb a szórás. A legerősebb modell 10 m-es tanulóterületekkel lett futtatva.

A **3-as terület** legjobbja: **rad5 t250 log** (PP arithmetic mean: **0.03648**)

Ezen a területen szintén az 5 m-es tanulóterületek bizonyultak a legjobbnak, bár a legjobb valószínűséget egy századdal egy 10 m-es tanulóterületű modell érte el. Az 1-es területhez hasonlóan a 20 m-es tanulóterületű modellek itt is a leggyengébb eredményeket produkálták.

Az alábbi táblázatban lévő eredményeket szemlélve kijelenthető, hogy az előrejelzési valószínűséget tekintve a legtöbb esetben az 5 m átmérőjű tanulóterületekkel futtatott modellek (rad3...) bizonyultak a legjobbnak.

	1-es terület			2-es terület			3-as terület		
tanulóterület - fák száma - változók elágazásonként	**PP max.**	**PP átlag**	**PP szórás**	**PP max.**	**PP átlag**	**PP szórás**	**PP max.**	**PP átlag**	**PP szórás**
rad3 - 32 - log				0.46875	**0.06844**	0.08667			
rad3 - 50 - log	0.43999	**0.03970**	0.05731	0.47999	**0.06463**	0.08839	0.47999	**0.03625**	0.05270
rad3 - 100 - log	0.43999	**0.04011**	0.05666	0.47999	**0.06871**	0.08619	0.47999	**0.03640**	0.05030
rad3 - 250 - log	0.44400	**0.03958**	0.05509				0.48399	**0.03647**	0.04888
rad5 - 32 - log				0.46875	**0.06384**	0.09159			
rad5 - 50 - log	0.46000	**0.03544**	0.05922	0.47999	**0.06216**	0.08798	0.47999	**0.03584**	0.05383
rad5 - 100 - log	0.46000	**0.03668**	0.05818	0.49000	**0.06879**	0.08983	0.47999	**0.03544**	0.05212
rad5 - 250 - log	0.44400	**0.03693**	0.05665				0.47999	**0.03648**	0.05149
rad10 - 32 - log				0.46875	**0.06314**	0.09266			
rad10 - 50 - log	0.47999	**0.03283**	0.05829	0.47999	**0.06420**	0.09111	0.47999	**0.03510**	0.05579
rad10 - 100 - log	0.49000	**0.03369**	0.05756	0.49000	**0.06483**	0.09068	0.49000	**0.03458**	0.05387
rad10 - 250 - log	0.48399	**0.03417**	0.05817				0.48800	**0.03511**	0.05320

4. táblázat: Az első körben futtatott random forest modellek paraméterei. Narancssárgával az adott forduló legjobbra becsült modellje látható, zölddel pedig az abszolút legjobb modell az adott területről.

A változók erősségeinek vizsgálatában a „**mean decrease gini**” (MDG) értékeket vizsgáltam, amelyek alapján láthatjuk, hogy melyik változó milyen mértékben járult hozzá a predikcióhoz. Ezek az értékek a létrejött „Feature importances” táblázatban találhatók. A gini index egy olyan arányszám, melynek segítségével egy adathalmaz diverzitásának, egyenlőtlenségének fokát mérhetjük. Minél nagyobb annak az esélye, hogy két eltérő osztályú mintát választunk ki, annál nagyobb a „gini impurity” érték, tehát annál diverzebb adathalmazzal állunk szemben. A fák betanítása közben pedig az algoritmus kiszámolja, hogy az egyes változók milyen mértékben csökkentik ezt a diverzitást. Minél nagyobb mértékben képes egy változó ennek a csökkentésére, annál erősebbnek nevezhetjük. Egy változó „mean decrease gini” értéke tehát az összes fára számított „gini decrease” értékének átlaga.

A legerősebb változó mindhárom területen a magasság volt. Az 1-es és a 2-es területek esetében több, mint 2-szer, a 3-as területen pedig több, mint másfélszer bizonyult erősebbnek, mint az azt követő legerősebb változó. Rendre erősnek mutatkozott továbbá mindhárom mintaterületen a „Terrain Surface Convexity”, a „Topographic Position Index” nagyobb keresési sugarú verziója, illetve a „Slope Height”.

A különböző görbületi (curvature) értékek mindegyike a gyenge változók közé sorolható, értékeik minden modellben nagyon alacsonyak voltak. Mindhárom területen gyengének nevezhető továbbá a „Convergence Index”, illetve a „Fuzzy Landform Element Classification”.

Az RFC modellezés második köre

A modellezés második körében a magasság és a gyenge változók elhagyásával már csak az erősnek ítélt változókat vettem figyelembe. A magasság a korábban említett erősségéből kifolyólag negatívan is befolyásolhatta a kapott eredményt, így tehát a modellezés további lépéseiből kihagytam. Ezt abból a megfontolásból tettem, miszerint a legjobb változók nem feltétlenül egyeznek a legerősebb változókkal (Albert G., Ammar S., 2021). A 2-es terület vízszintes települését figyelembe véve a magasság nélküli modellektől gyengébb eredmények várhatóak.

Hasonlóképp jártam el tehát a gyengének ítélt változókkal is, az összes „curvature” (görbületi) változó el lett hagyva. Ezzel az eredmények várható javulásán túl nem utolsósorban a futási időn is javulást érhető el. Az alábbiakban felsorolom a második körös modellezéshez redukált számú változókat, valamint azok MDG értékeit az első kör adott területeken legerősebb modelljeiből. Ezek az értékek csak egy adott modellen belül vethetők össze, ugyanis azok paramétereitől függenek.

Meghagyott változók az 1-es területen (összesen 13): terrain surface convexity (254), topographic position index 90-100 (239), slope height (229), vector terrain ruggedness (141), mid-slope position (124), terrain ruggedness index (122), saga wetness index (74), topographic position index 5-10 (73), downslope distance gradient (72), aspect (69), morpometric protection index (45), slope (39), terrain surface texture (39)

A továbbiakban figyelmen kívül hagyott magasság MDG értéke ebben a modellben 513 volt (azaz 2,02-szor erősebb az azt követő legerősebb változónál).

Meghagyott változók az 2-es területen (összesen 14): topographic position index 90-100 (732), slope height (551), terrain ruggedness index (441), terrain surface convexity (440), vector terrain ruggedness (321), aspect (295), mid-slope position (259), terrain surface texture (175),

downslope distance gradient (170), morphometric protection index (151), slope (135), geomorphons (128), saga wetness index (126), topographic position index 5-10 (105)

Itt a magasság MDG értéke 1900 volt (ami 2,6-szor erősebb az azt követőnél). Ez az arány ennek a területnek a modelljeiben volt a legmagasabb, ami szintén a magasság döntő szerepére utal.

Meghagyott változók az 3-as területen (összesen 14): topographic position index 90-100 (2279), slope height (1795), terrain surface convexity (1469), mid-slope position (1246), terrain ruggedness index (1242), aspect (1132), vector terrain ruggedness (861), saga wetness index (606), downslope distance gradient (511), topographic position index 5-10 (443), geomorphons (397), morphometric protection index (310), slope (227), terrain surface texture (194)

Ezen a területen pedig 3952-es MDG értékkel szerepelt a magasság (ami 1,73-szor erősebb az azt követő változónál).

A három területen egy kivétellel ugyanazok a változók maradtak meg. A geomorfonok az erősségéből kifolyólag a 2-es és a 3-as területen meg lett hagyva, azonban az 1-es területen annak gyengesége miatt elhagytam. A generálandó fák számát a második kör modelljeiben 100-ban maximáltam. A második körben az alábbi paraméterekkel modelleztem.

- változók: 13, illetve 14 db területtől függően (külön 3, 5 és 10 m keresési sugarú változókkal, melyek a tanulóterületekkel összhangban lettek alkalmazva)
- tanulóterületek: 5, 10, 20 m
- fák száma: 32, 50, 100 db
- változók száma elágazásonként:
    - négyzetgyök alapú
        - 13 változó esetében: $\sqrt{13} = 3{,}606 \rightarrow 4$ változó
        - 14 változó esetében: $\sqrt{14} = 3{,}742 \rightarrow 4$ változó
    - logaritmus alapú
        - 13 változó esetében: $\log_{10} 13 = 1{,}114 \rightarrow 1$ változó
        - 14 változó esetében: $\log_{10} 14 = 1{,}146 \rightarrow 1$ változó

A PP átlagértékeit nézve mindhárom területről elmondható, hogy a négyzetgyök-alapú változószámmal futtatott modellek esetében nagyon gyenge eredmények születtek. Az 1-es és a 3-as terület esetében az 5 m-es tanulóterületek szerepeltek a legjobb valószínűséggel. A 2-es területen a 10 és a 20 m-es tanulóterületek produkálták a legjobb értékeket.

A logaritmus-alapú változószámok modelljei esetében az előzőekhez viszonyítva átlagosan több, mint 3-szor jobb várható valószínűségeket kaptunk. Az 1-es terület esetében az első körös modellekhez hasonlóan az 5 m-es tanulóterületek érték el a legnagyobb valószínűségeket, majd a 10, ezután pedig a 20 m-es modellek követeznek. A terület legerősebb modellje: **rad3 t50 log** (PP arithmetic mean: **0.05079**).

A 2-es területen ezzel ellentétben a 10 és a 20 m-es előrejelzések bizonyultak erősnek. A terület legerősebb modellje: **rad10 t100 log** (PP arithmetic mean: **0.07059**).

A 3-as területen kiegyenlített becsült pontossági eredményeket láthatunk. A terület legerősebb modellje: **rad3 t100 log** (PP arithmetic mean: **0.03351**).

Az alábbi táblázatban a második kör modelljeinek becsült valószínűségei olvashatók.

	1-es terület			2-es terület			3-as terület		
tanulóterület - fák száma - változók elágazásonként	**PP max.**	**PP átlag**	**PP szórás**	**PP max.**	**PP átlag**	**PP szórás**	**PP max.**	**PP átlag**	**PP szórás**
rad3 - 32 - sqrt	0.46875	**0.01733**	0.07136	0.46875	**0.01859**	0.07399	0.46875	**0.01270**	0.06210
rad3 - 50 - sqrt	0.47999	**0.01866**	0.07595	0.47999	**0.01764**	0.07263	0.47999	**0.01368**	0.06488
rad3 - 100 - sqrt	0.49000	**0.01817**	0.07505	0.49000	**0.01926**	0.07710	0.49000	**0.01303**	0.06280
rad5 - 32 - sqrt	0.46875	**0.00856**	0.05244	0.46875	**0.02213**	0.08632	0.46875	**0.00803**	0.04938
rad5 - 50 - sqrt	0.47999	**0.00876**	0.05214	0.47999	**0.02464**	0.09256	0.47999	**0.00850**	0.05027
rad5 - 100 - sqrt	0.49000	**0.00951**	0.05568	0.49000	**0.02062**	0.08613	0.49000	**0.00890**	0.05181
rad10 - 32 - sqrt	0.46875	**0.00924**	0.05317	0.46875	**0.01589**	0.07039	0.46875	**0.00652**	0.04416
rad10 - 50 - sqrt	0.47999	**0.00977**	0.05536	0.47999	**0.02546**	0.09551	0.47999	**0.00647**	0.04476
rad10 - 100 - sqrt	0.49000	**0.00991**	0.05562	0.49000	**0.02067**	0.08463	0.49000	**0.00692**	0.04626
rad3 - 32 - log	0.46875	**0.04769**	0.07344	0.46875	**0.06110**	0.08250	0.46875	**0.03207**	0.05283
rad3 - 50 - log	0.47999	**0.05079**	0.07342	0.47999	**0.06340**	0.08416	0.47999	**0.03157**	0.05036
rad3 - 100 - log	0.46999	**0.04918**	0.07207	0.49000	**0.06377**	0.08227	0.47999	**0.03351**	0.05016
rad5 - 32 - log	0.46875	**0.04259**	0.07038	0.46875	**0.06955**	0.08986	0.46875	**0.03066**	0.05372
rad5 - 50 - log	0.47999	**0.04432**	0.07060	0.47999	**0.07041**	0.08952	0.47999	**0.03137**	0.05315
rad5 - 100 - log	0.49000	**0.04459**	0.07034	0.49000	**0.06981**	0.08716	0.49000	**0.03212**	0.05193
rad10 - 32 - log	0.46875	**0.03807**	0.06910	0.46875	**0.06863**	0.09050	0.46875	**0.02993**	0.05508
rad10 - 50 - log	0.47999	**0.04003**	0.06955	0.47999	**0.06941**	0.08895	0.47999	**0.03171**	0.05548
rad10 - 100 - log	0.49000	**0.03899**	0.06745	0.49000	**0.07059**	0.08927	0.49000	**0.03193**	0.05419

5. táblázat: A második körben futtatott random forest modellek paraméterei. Narancssárgával az adott forduló legjobbra becsült modellje látható, zölddel pedig az abszolút legjobb modell az adott területről.

Összességében, mindkét forduló eredményeit figyelembe véve az **1-es terület** legjobbra becsült modellje a második kör „**rad3 t50 log**” modellje, a **2-es területé** a második kör „**rad10 t100 log**” modellje, a **3-as területé** pedig az első kör „**rad5 t250 log**” modellje. A modellezés mindkét körét összegezve elmondható, hogy az 1-es és a 2-es területen sikerült javulást

elérnünk a PP-értékekkel a második körben, míg a 3-as területen a várható legjobb modell az első körben lefuttatott modellek közül került ki.

A pontosságvizsgálat előtt a legjobb modellek RFC-osztályozott grid-jeit megvágtam a vektoros geológiai poligonok kiterjedésével, ugyanis az eredeti grid kiterjedése téglalap alakú volt.

Pontosságvizsgálat

A kiértékelés legegyszerűbb módja a találati arány (accuracy - pontosság) kiszámítása. Ehhez a SAGA eszköztárának imagery → classification → confusion matrix (polygons / grid) eszközét használtam (lásd 8. ábra). Az eljárás az általam megadott, az adott területen legjobbnak bizonyuló RFC raszter osztályait (predikált osztályok) vetette össze a formációtípusok poligonjaival (elvárt osztályok). Bemenetként tehát egy gridet és egy vektoros poligon réteget kell kiválasztanunk. Előbbi esetében meg kell adnunk, hogy az egyes osztályokra az adott képpont-értékek utalnak, vagy egyéb keresőtáblából (look-up table) értelmezze az értékeket. Utóbbi esetében pedig az osztályok azonosítására szolgáló mezőt kell kiválasztanunk (mely esetünkben az „ID”).

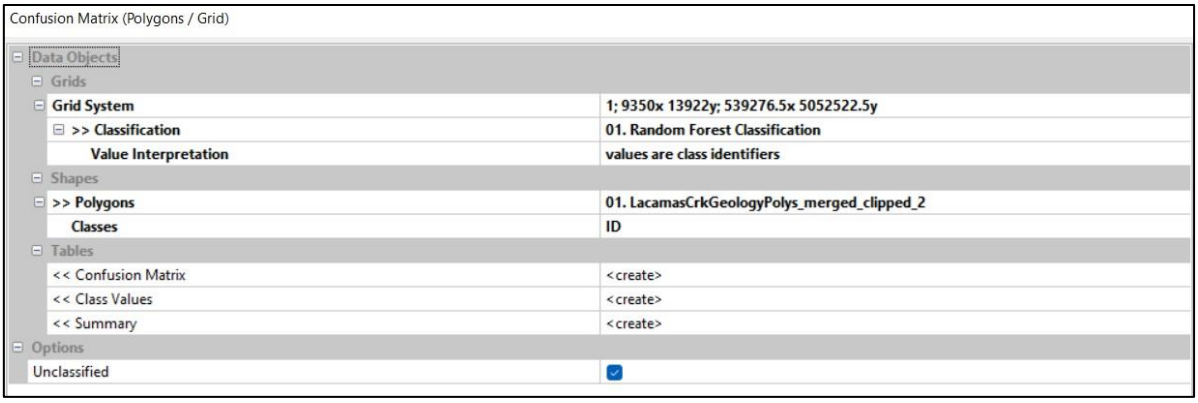
Az algoritmus statisztikát készít az elvárt osztályokba eső helyesen predikált képpontok arányairól (**producer accuracy**), illetve a predikált osztályok képpontjainak a nekik megfelelő elvárt poligonokba eső arányairól (**user accuracy**). A producer accuracy tehát a referenciaadatok szemszögéből, a user accuracy pedig az osztályozás eredményeképp létrejött osztályok szemszögéből vizsgálja a pontosságot. Mindkét mutató egy 0 és 1 közötti arányszám, amely minél nagyobb, annál pontosabb osztályozásról beszélünk. Az algoritmus eredményként egy úgynevezett tévesztési mátrixot (**confusion matrix**) hoz létre, valamint kiszámítja az összesített pontossági mutatót (**overall accuracy**). A tévesztési mátrix futtatása után az alábbi kimeneti állományok jönnek létre.

- class values (osztályértékek): összegzés az osztályok pontosságairól
- confusion matrix (tévesztési mátrix): statisztikai összegző táblázat az összes elvárt és predikált osztály viszonyáról
- summary (összegzés): rövid összefoglaló a kappa-értékkel és az összesített pontosság (overall accuracy) értékével

A kappa egy olyan – véletlenszerű mintavételen alapuló – arányszám, amely megmutatja, hogy a mintaként kiválasztott cellák között milyen arányban vannak a helyesen osztályozott

(úgynevezett „true positive”) pixelek. Az összesített pontosság pedig a helyesen osztályozott adatok (amely a tévesztési mátrix főátlójának összegeként is felfogható) és az összes adat aránya (Accuracy Metrics, 2021).

A fentebb felsorolt állományok mindegyike táblázatos adat, melyeket .csv formátumban exportáltam. A tévesztési mátrix futtatásával kapott eredmények bemutatását a következő fejezetben teszem meg.



8. ábra: A tévesztési mátrix futtatásának paraméterei a SAGA-ban.

Eredmények

Ebben a fejezetben sor kerül az eredmények szöveges kiértékelésére, valamint a pontossági értékek táblázatos bemutatására. A kiértékelésnél csak a fentebbi „Modellezés végrehajtása” című fejezet végén feltüntetett, a területenként legpontosabbra becsült modelleken futtattam le a tévesztési mátrixot. Ez alól kivételt jelentett a 2-es számú terület, ahol a legpontosabbra becsült (2. fordulós – 14 változós) modell (rad10 t100 log) ugyanilyen elnevezésű 1. fordulós (32 változós) verzióját is kiértékeltem. Erre azért került sor, mert a vizsgált területet meghatározó vízszintes rétegtelepülést várhatóan nagy pontossággal osztályozza a magasság változót felhasználó modell, valamint ezzel lehetőségünk nyílik a két modell konkrét összehasonlítására. A fejezet második felében az osztályozás hibáinak lehetséges okaira is kitérek, majd bemutatom a kapott eredményeket térképeken.

A területek legjobb modelljeinek kiértékelése

1-es terület: Lacamas Creek (Washington)

Az 1-es számú terület legjobbra becsült modelljének összesített pontossága 0,187 (azaz **18,7%**). A Kappa értéke 0,102 (10,2%).

Class	Code	SumRef	AccProd	SumClassified	AccUser
0	Qfs	1789659	0.383171	5587133	0.122736
1	Qa	4915185	0.088337	3917763	0.110827
2	Qbbh	731694	0.33327	7958755	0.030639
3	Qbgm	1419462	0.26651	7348022	0.051483
4	Qbmc	682196	0.124782	8524455	0.009986
5	Qfg	689560	0.161575	7498047	0.014859
6	Ql	456023	0.279567	10700578	0.011914
7	QTc	23623966	0.24003	11143116	0.508875
8	Tbem	39988058	0.204236	12368951	0.660283
9	Tsr	1270450	0.137129	5564411	0.031309
10	Tftc	26490010	0.087478	8665993	0.267402
11	Tfth	311737	0.147211	5579545	0.008225
12	Qls	8155366	0.273617	12171702	0.18333
-1	unclassified	0		3494895	0

6. táblázat: Az 1-es számú terület formációinak pontossági mutatói (class: a formáció "ID"-ja, code: a formáció kódja, SumRef: referencia-képpontok száma, AccProd: referencia-poligonok pontossága, SumClassified: eredmény-képpontok száma, AccUser: eredmény-osztályok pontossága).

A legjobb **producer accuracy**-vel bíró formáció a homokos és iszapos fácies (ID-ja: **0** - kódja: **Qfs**), melynek értéke 0,383. Ezt követte a Brunner Hill bazaltos andezit formációja (**2 - Qbbh**) 0,333-as értékkel. Megemlítendő még a tavi üledékes formáció (**6 - Ql**) 0,28-os; a földcsuszamlásos törmelékek osztályának (**12 - Qls**) 0,273-es, illetve a Green Mountain bazaltos andezit formációjának (**3 - Qbgm**) 0,267-es referencia-pontossága, mely a területen viszonylag jónak mondható. Megállapítható tehát, hogy a legnagyobb pontosságú osztályok között két vízi vagy vízközeli üledékes, illetve két kiömléses vulkanikus formáció található. Ezek lapos térszínei, illetve hegykúpjai könnyebben osztályozhatónak bizonyultak a többi formációhoz képest. Alacsony producer accuracy értékeket produkált a Troutdale formáció konglomerátum tagja (**10 - Tftc**) 0,087-es, és az alluvium (**1 - Qa**) 0,088-es értékkel, tehát ezek referencia poligonjai nagyrészt hibásan lettek osztályozva.

Ami pedig a **user accuracy**-t illeti, az osztályok közül magasan kiemelkedik az Elkhorn Mountain bazaltos andezit formációjának (**8 - Tbem**) 0,66-os, illetve a konglomerátum (**7 - QTc**) 0,509-es pontossága. Ez a nekik megfelelő referencia-poligonok nagy kiterjedésével is összefüggésbe hozható. Sajnos több osztály mutat nagyon gyenge, akár 0,01 alatti értékeket, ami a predikció bizonytalanságára utal. Ezek közé tartozik a Troutdale formáció homokkő tagja (**11 - Tfth**) és a Matney Creek bazaltja (**4 - Qbmc**).

Ezen a vizsgált területen nem tudunk megállapítani olyan osztályt, amelynek mindkét értékét jónak lehetne mondani. A formációk két pontossági értéke között nem látható összefüggés.

A táblázatban az „unclassified” értékek a besorolás nélküli cellákat jelentik, ilyen esetekben a random forest algoritmus nem tudott elfogadható döntést hozni. Ez egyik területen sem fordult elő számottevő mértékben. A táblázat „AccProd” és „AccUser” oszlopainak értékei pirosból sárgába, majd zöldbe átmenő árnyalatai jelentik az egyre nagyobb pontosságértékeket. A „SumRef” és a „SumClassified” oszlopok értékei a referencia-pixelek, valamint az osztályozott pixelek számait jelentik az adott kategóriákban.

2-es terület: New River Gorge (Nyugat-Virginia) – 2. fordulós, abszolút legjobb PP-értékű modell 14 változóval

A 2-es számú terület legjobbra becsült modelljének összesített pontossága 0,1687 (**16,87%**). A Kappa értéke 0,0465 (4,65%).

Class	Code	SumRef	AccProd	SumClassified	AccUser
0	**Mbf**	2259778	0.419894	10933501	0.086785
1	**Mbs**	48235931	0.121823	17551768	0.334796
2	**Mhl**	43867429	0.124707	19246601	0.284236
3	**Mhlsg**	3511757	0.28386	30628933	0.032546
4	**Mhsg**	2494804	0.25445	13475714	0.047107
5	**Mhu**	33695386	0.259755	33545345	0.260917
6	**Mpn**	13058670	0.125966	14403600	0.114204
7	**PNnrp**	890500	0.365241	7014008	0.046371
8	**PNp**	15800494	0.188921	16270840	0.18346
-1	**unclassified**	0		744439	0

7. táblázat: Az 2-es számú terület formációinak pontossági mutatói (14 változóval).

A **producer accuracy** értékeket tekintve a Bluefield formáció differenciálatlan tagjának (**0 - Mbf**) 0,42-os, illetve a New River formáció pineville-i homokkő tagjának (**7 - PNnrp**) 0,365-es referencia-pontossága bizonyult a legjobbnak. Előbbi a folyóvölgyhöz legközelebbi üledékréteg, míg utóbbi egy hegykúpok magasságában jelen lévő formáció. Gyengének mondható a Bluestone formáció differenciálatlan tagjának (**1 - Mbs**) 0,122-es, a Hinton formáció alsó tagjának (**2 - Mhl**) 0,125-es, illetve a Princeton formáció differenciálatlan tagjának (**6 - Mpn**) 0,126-es pontossága.

User accuracy tekintetében kiemelhető a Bluestone formáció differenciálatlan tagjának (**1 - Mbs**), a Hinton formáció alsó tagjának (**2 - Mhl**) valamint a Hinton formáció felső tagjának (**5 - Mhu**) viszonylagosan jó eredménye 0,334-es, 0,284-es és 0,261-es értékekkel. Sajnos néhány osztály pontossága itt is jócskán elmarad a többitől. Ilyen a Hinton formáció mészkő tagja (**3 - Mhlsg**), a New River formáció pineville-i homokkő tagja (**7 - PNnrp**), vagy a Hinton formáció homokkő tagja (**4 - Mhsg**), melyek értékei 0,03 és 0,05 közöttiek.

Az említett példákból, illetve a fenti táblázatból is látható, hogy ezen a területen a producer és a user accuracy között sok esetben fordított arányosság áll fent egy adott formációnál. A területre jellemző vízszintes rétegzettséget nem tudta visszaadni ez a modell.

2-es terület: New River Gorge (Nyugat-Virginia) – az előbbi modell paramétereinek 1. fordulós megfelelője 32 változóval

A magasságot is figyelembe vevő, a többi paraméter tekintetében viszont a fentivel egyező modell összesített pontossága 0,5112 (**51,12%**). A Kappa értéke 0,4198 (41,98%).

Class	Code	SumRef	AccProd	SumClassified	AccUser
0	Mbf	2259778	0.628665	3770174	0.376811
1	Mbs	48235931	0.579195	35768957	0.781069
2	Mhl	43867429	0.397127	18526492	0.940327
3	Mhlsg	3511757	0.516421	20559729	0.088209
4	Mhsg	2494804	0.614603	10072863	0.152222
5	Mhu	33695386	0.516031	29908618	0.581366
6	Mpn	13058670	0.692079	29706625	0.30423
7	PNnrp	890500	0.603314	3311471	0.162239
8	PNp	15800494	0.420541	11445381	0.580562
-1	unclassified	0		744439	0

8. táblázat: A 2-es számú terület formációinak pontossági mutatói (32 változóval).

Láthatjuk, hogy az előzőeknél lényegesen jobb eredményeket kaptunk mind a producer accuracy, mind a user accuracy tekintetében.

Előbbi esetében kiemelendő a Princeton formáció differenciálatlan tagjának (**6 - Mpn**) 0,692-es pontossága, de jónak mondható még a Bluestone formáció differenciálatlan tagja (**0 - Mbf**), a Hinton formáció homokkő tagja (**4 - Mhsg**), illetve a New River formáció pineville-i homokkő tagja (**7 - PNnrp**), mind 0,6 fölötti pontosságértékkel. Csupán egyetlen formáció pontossága 0,4 alatti, mégpedig a Hinton formáció alsó tagja (**2 - Mhl**).

User accuracy tekintetében az előbb említett Hinton formáció alsó tagja (**2 - Mhl**) magasan a legpontosabb a maga 0,64-os értékével. A leggyengébb user accuracy értéket a Hinton formáció mészkő tagja (**3 -Mhlsg**) produkálta 0,088-es értékkel.

A terület előzőleg bemutatott modelljéhez hasonlóan itt is megfigyelhető a fordított arányosság az adott formációk két pontosságértékét illetőleg. A két modell között pedig hasonlóság figyelhető meg a gyenge és az erős osztályok tekintetében, ami az user accuracy számait figyelve szembetűnőbb. Megfigyelhető továbbá, hogy a producer accuracy értékei között kisebb a szórás, mint a user accuracy esetében.

3-as terület: White Hall (Virginia, Nyugat-Virginia)

A 3-as számú terület legjobbra becsült modelljének összesített pontossága 0,2151 (**21,51%**), így a három modellezett terület közül ez lett a legpontosabb. A Kappa értéke 0,1668 (16,68%).

Class	Code	SumRef	AccProd	SumClassified	AccUser
0	Ce	11622588	0.180588	7310743	0.287098
1	DSls	579744	0.589457	2861822	0.119411
2	Swbm	612660	0.72798	1023553	0.435741
3	Db	16353901	0.261054	13484860	0.316596
4	Df	8814726	0.13996	8855607	0.139313
5	Dfl	7366400	0.147431	6660346	0.16306
6	Dfu	8633245	0.150833	5687817	0.228942
7	Dh	16954159	0.092526	5861283	0.267637
8	Dm	10998773	0.210893	9180516	0.252661
9	Dmc	8524924	0.217277	10022615	0.184809
10	Dmch	4398519	0.211921	6404807	0.145537
11	Dmn	708260	0.285222	7043102	0.028682
12	Oc	116861	0.513311	3136859	0.019123
13	Occ	12681905	0.150755	6145931	0.311077
14	Ojo	565428	0.838238	3168810	0.149571
15	Om	5248712	0.23334	6084231	0.201296
16	On	70062	0.497859	2846721	0.012253
17	Op	77692	0.6876	2630611	0.020307
18	Or	2033188	0.178215	3269034	0.110841
19	Os	3726659	0.262419	2581862	0.378776
20	Skr	1157295	0.368617	1725098	0.24729
21	St	766896	0.642846	3349530	0.147184
22	Qal	8351869	0.520343	5987436	0.725826
23	Qt	60457	0.635162	2904818	0.013219
-1	unclassified	0		2196911	0

9. táblázat: A 3-as számú terület formációinak pontossági mutatói.

A legjobb **producer accuracy**-t mutató formációk ezen a területen a Juanita formáció és Oswego homokkő osztatlan kategóriája (**14 - Ojo**) 0,838-as értékkel, illetve a Wills Creek, Bloomsburg, és McKenzie formációk osztatlan kategóriája (**2 - Swbm**) 0,728-es pontossággal. Ezek az értékek a többi formációhoz képest nem kiemelkedőek, ugyanis vannak további nagyon jó eredmények, amelyek nem sokkal maradnak el az említettektől. Ilyen például a Pinesburg Station-i dolomit (**17 - Op**), a Tuscarora kvarcit formáció (**21 - St**), a teraszos üledékek (**23 - Qt**), az Oriskany homokkő, Helderberg-csoport és Tonoloway-i mészkő osztatlan kategóriája (**1 - DSls**), az alluvium (**22 - Qal**), illetve a Chambersburg formáció (**12 - Oc**), melyek mind 0,5 feletti pontossággal bírnak. Ezzel ellentétben gyengének bizonyult a Hampshire formáció

(**7 - Dh**) 0,093-as értéke, a Foreknobs formáció tagjainak (**4 - Df**, **5 - Dfl**, **6 - Dfu**) 0,14–0,15 körüli értékei, valamint a Conococheague mészkő (**13 - Occ**) 0,151-es pontossága.

A **user accuracy** értékeit nézve az alluvium (**22 - Qal**) mondható magasan a legjobbnak a maga 0,726-es pontosságával. Az alluviummal borított lapos térszínek tehát – figyelembe véve a viszonylag jónak mondható producer accuracy értékeket is – nagy pontossággal lettek meghatározva. Kiemelhető még a Wills Creek, Bloomsburg, és McKenzie osztatlan kategóriájának (**2 - Swbm**) 0,436-es pontossága, amely emellett a második legmagasabb producer accuracy értéket produkálta. Néhány formáció kifejezetten alacsony user accuracy értéket ért el, mint például a New Market-i mészkő (**16 - On**), a teraszos üledékek (**23 - Qt**), a Chambersburg formáció (**12 - Oc**), a Pinesburg Station-i dolomit (**17 - Op**), vagy a Marcellus és Needmore agyagpala formációk (**11 - Dmn**) 0,01 és 0,03 közötti értékekkel.

A 3-as számú területen nem fedezhető fel összefüggés a két pontossági mutató között az egyes osztályoknál.

Az eredmények összegzése

Az alábbi táblázat röviden összefoglalja a kiértékelt modellek Kappa és összesített pontosság-értékeit.

	1-es ter.	2-es ter. (14 vált.)	2-es ter. (32 vált.)	3-as ter.
Kappa	0.101996	0.046523	0.419844	0.166792
Összesített pontosság	0.187041	0.168698	0.51115	0.215108

A három terület eredményeit összegezve elmondható, hogy a 2-es számú terület 32 változós modellje érte el a legjobb eredményeket 51,12%-os összesített pontossággal. Ebben a modellben a legtöbb formáció predikciója jónak vagy nagyon jónak mondható. Meg kell jegyezni azonban, hogy ez a magasság döntő szerepének köszönhető, ugyanis a területet határozottan kirajzolódó vízszintes rétegtelepülés jellemzi. A 2-es számú terület 14 változós modellje ezzel ellentétben gyenge összesített eredményeket mutat.

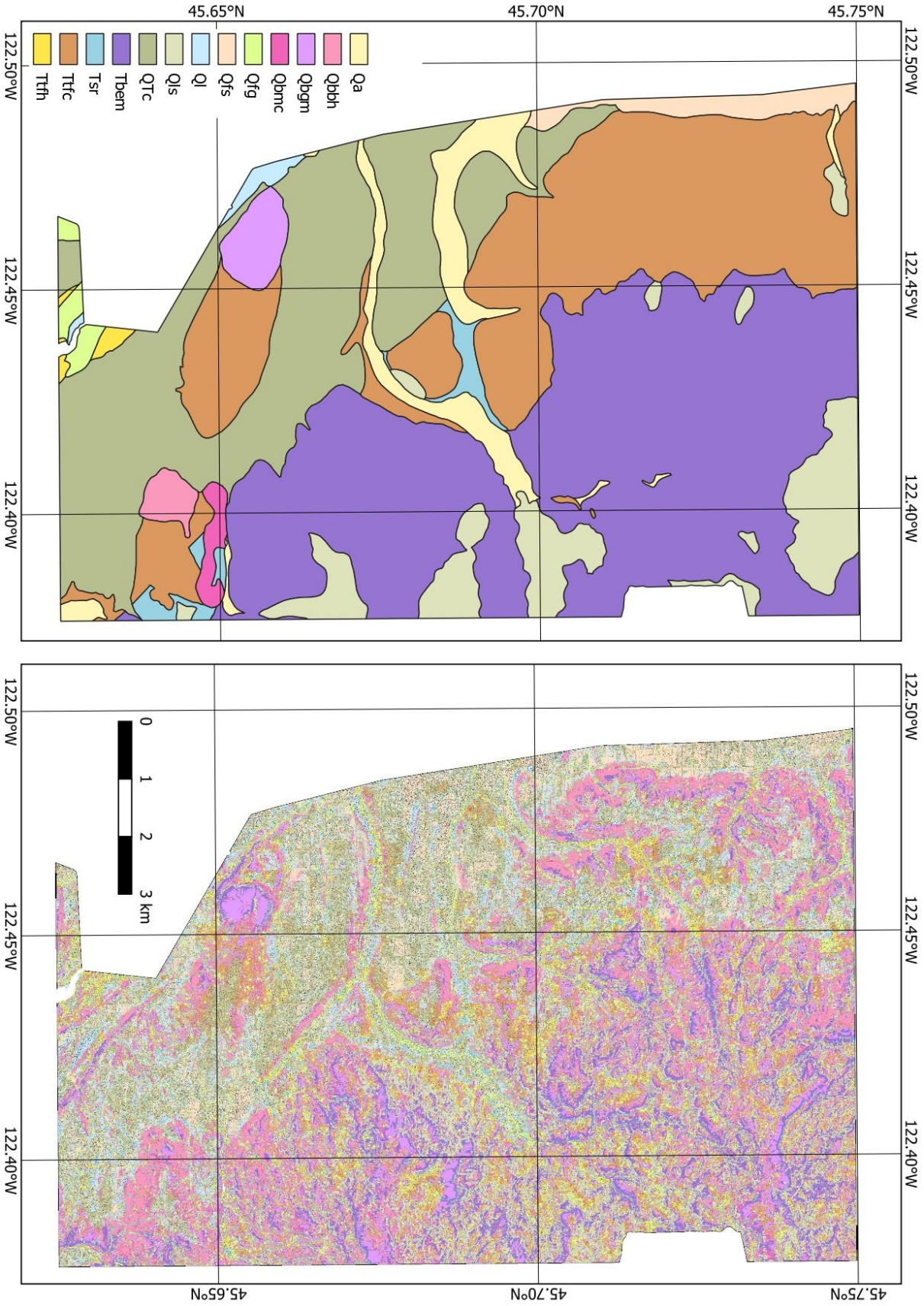
A 3-as számú terület osztályozása moderáltan pontosnak, közepesnek nevezhető. A előbb említett 2-es számú terület erősebb modelljét leszámítva azonban ezt nevezhetjük a legpontosabb modellnek. Az összesített pontosság korántsem tekinthető kiemelkedőnek, ennek aránya 21,51%. A területen viszont vannak pontosan meghatározott formációk, melyek javítják a modell összesített értékelését.

Az 1-es számú terület lényegében semmilyen szempontból nem produkált jónak értékelhető eredményeket a pontossági mutatók alapján (összesített pontosság: 18,7%).

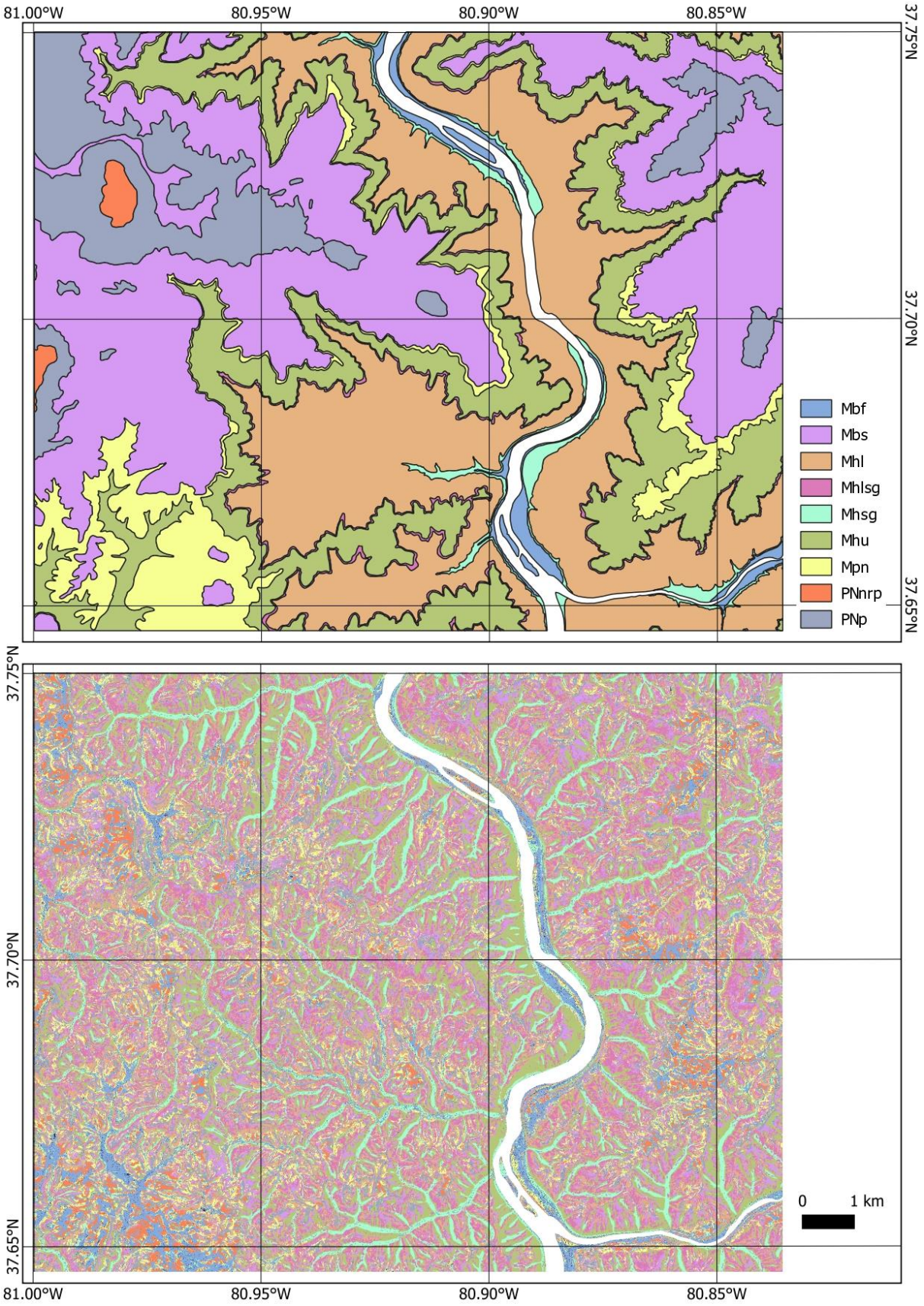
Összefoglalva elmondható, hogy azoknál a formációknál értünk el nagy pontosságot, ahol azok egyedi és egyértelműen meghatározható morfometriai paraméterekkel rendelkeztek.

Az eredmények bemutatása térképeken

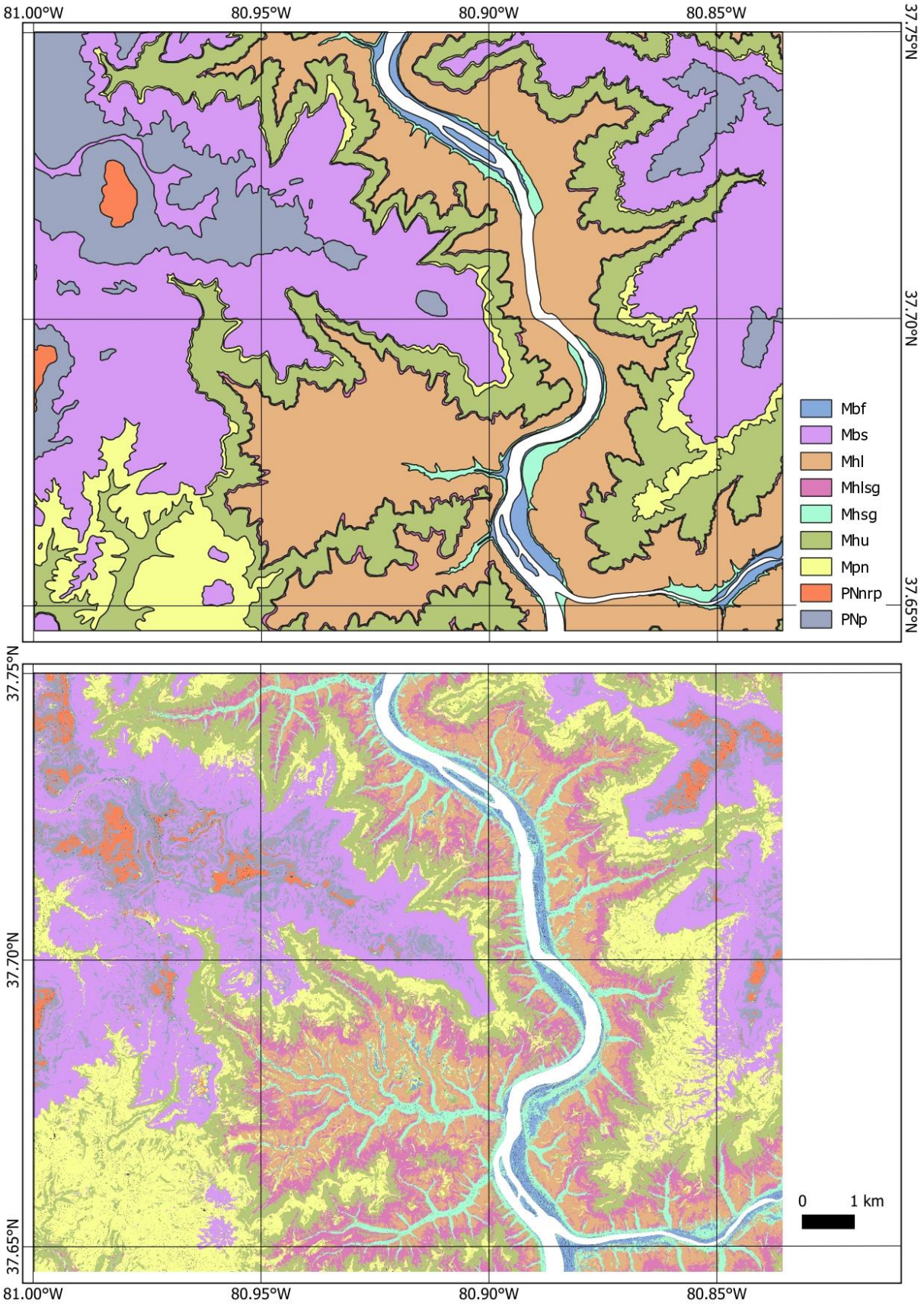
A következő oldalakon az eredmények térképes bemutatására kerül sor, ahol összehasonlíthatjuk a referencia-térképeket az általam létrehozott osztályozott térképekkel.



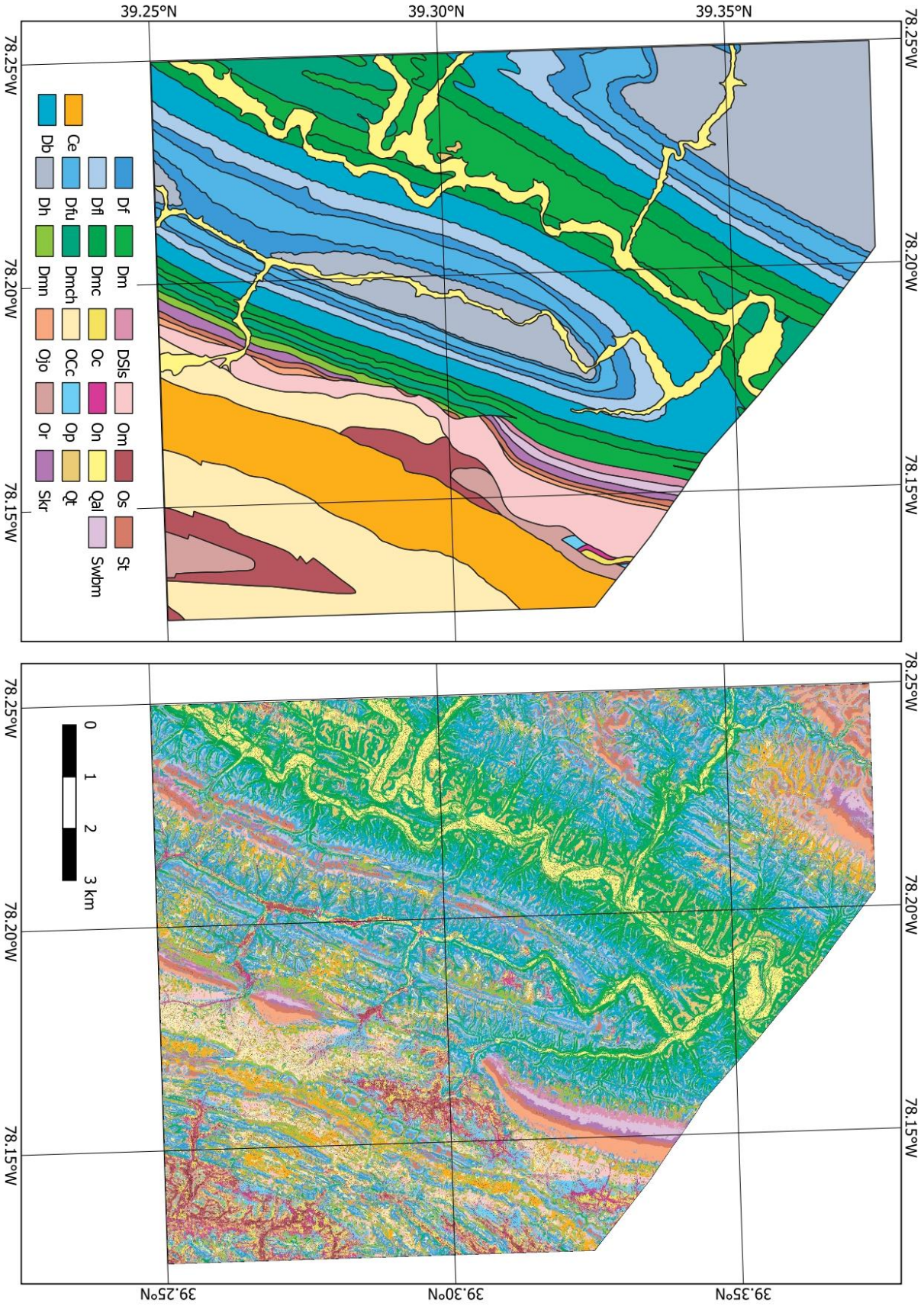
9. ábra: Az 1-es számú terület referencia-térképe (felül), illetve osztályozott térképe (alul).



10. ábra: A 2-es számú terület referencia-térképe (felül), illetve 14 változón alapuló osztályozott térképe (alul).



11. ábra: A 2-es számú terület referencia-térképe (felül), illetve 32 változón alapuló osztályozott térképe (alul).



12. ábra: A 3-as számú terület referencia-térképe (felül), illetve osztályozott térképe (alul).

A hibák lehetséges okai, értékelés

Egyes területek gyengébb eredményei mögött egész egyszerűen azok efféle osztályozásra való alkalmatlansága áll. Például a változatos domborzati formákkal rendelkező, nagyobb kiterjedésű formációk kimutatása nehéz, ezek ugyanis morfometriailag túlságosan sokszínűek ahhoz, hogy helyesen osztályozhatók legyenek. Ilyen esetekben nem áll fent egyértelmű kapcsolat a vizsgált morfometriai változók és a formáció között.

Ezenkívül azonban bizonyosan vannak olyan tényezők vagy paraméterek, amik rontottak a pontosságon, vagy amikkel tovább javítható lenne a kapott pontosság. A következőkben tehát röviden áttekintem a hibás osztályozás lehetséges okait.

Mesterséges felszínek befolyása

Egyes területrészeken nehezítette a pontos osztályozást a beépített területek jelenléte, habár ezt előzetesen igyekeztem elkerülni. Ez a tényező nehezen iktatható ki, ugyanis ilyen területeken egy helyesen osztályozott LiDAR pontfelhőből kiindulva is valótlan felszínek jönnek létre a domborzatmodell létrehozásakor, amikor talajpontok hiányában nagyobb adathiányos „lyukaknál” interpolálunk.

Továbbá a mesterséges eredetű domborzati, mikrodomborzati viszonyok, mint például a díszkertek és szántóföldek sík területei, valamint a kőfejtők és útbevágások meredek, hirtelen változó peremei egyaránt nehezítették a helyes besorolást. Az eredmények térképes elemzéséből látszik, hogy esetenként a növényzet is befolyásoló tényező volt.

Túl kicsi vagy rosszul megválasztott tanulóterületek esete

Előfordulhat, hogy az uniform módon kijelölt tanulóterületek túl kicsinek bizonyultak, vagy nem a legmegfelelőbb helyekre kerültek ahhoz, hogy visszaadják az adott formáció legjellemzőbb morfometriai paramétereit. Egy tanulmány továbbá alátámasztja, hogy a tanulóterületek nagyobb térbeli elszórtsága növeli a pontosságot (Cracknell M. J., Reading A. M., 2014).

Több formációtípus hasonló morfometriai paraméterekkel

Előállhat olyan helyzet is, amikor egyes morfometriai változók együttes jellemzői több formációtípusban is fellépnek. Ilyen esetekben „egy a többhöz” típusú kapcsolatokról beszélünk; ekkor a predikció gyengének mondható.

A felhasznált adatok hibái

A geológiai térképen, ebből következően pedig a felhasznált térinformatikai adatbázisban is előfordulhatnak pontatlanságok, melyek téves osztályozáshoz vezethetnek. Ez leginkább a kisebb méretarányú geológiai térképeknél merülhet föl, köszönhetően a generalizált megjelenítésnek. Jelen kutatásban azonban nagy méretarányú térképekkel dolgoztam, így ez a jelenség kevésbé játszhatott szerepet. Szélsőséges esetben előállhat az a helyzet is, hogy a modellezéssel létrehozott osztályok a valósak, melyek viszont a referencia-térképtől való eltérés miatt hibákként mutatkoznak meg. Ezeket az eltéréseket felhasználóként nem tudjuk kiszűrni. A lézerszkenneléses adatnyerésből is fakadhatnak bizonyos kalibrációs vagy egyéb eltérések, továbbá hibás lehet a LiDAR pontfelhő osztályozása is.

Fogalmi generalizálás

Ezalatt a formációk és a kőzettípusok közötti különbségekre kell gondolni, ugyanis egyes formációk úgy lettek meghatározva, hogy többféle kőzettípust is tartalmaznak. Ugyan a geológiai formáció definíciójában meg van határozva, hogy azt egymáshoz fizikailag hasonló kőzetek alkotják, mégis előfordulhat, hogy az egy formációban lévő különböző kőzettípusok eltérő morfometriai jellemzőkkel rendelkeznek.

„Rossz" területválasztás

Előfordulhat, hogy az általam futtatott modellek hasonló paraméterekkel jobb eredményeket értek volna el más területeken. Ezt nem lehet feltétlenül előre tudni, de bizonyosan létezik olyan hasonló geológiájú terület, ahol például kevesebb a beépített, vagy egyéb módon hasznosított terület.

A pontosságról

Az osztályozás és a pontosságvizsgálat módszertanából adódik, hogy kisebb területeket vizsgálva nagy számban jönnek létre valótlan, hibás helyzetek. Itt elsősorban arra gondolok, hogy az algoritmus nem veszi figyelembe annak a fontosságát, hogy a térben egymáshoz közel lévő pontok nagyobb valószínűséggel tartoznak ugyanabba az osztályba, mint az egymástól távolabb esők. Ez a gyakorlatban megfogalmazva azt jelenti, hogy minden 1-2 (kevés) cellából álló predikált osztály geológiailag valótlan helyzetet mutat, ugyanis ekkora kiterjedésű, és a dolgozatban vizsgált eljárással kimutatható formációk nem léteznek. Ez a helyzet abból adódik, hogy a felhasznált adatok belső struktúráját (például a kőzetrétegek sorrendjét) az algoritmus nem veszi figyelembe (Albert G., Ammar S., 2021). Ilyen esetekben sokszor helyesnek

mondható az osztályozás, és az említett képpontok csak „zajnak” tekintendők. Az alkalmazott pontosságvizsgálati módszer azonban ezt nem veszi, és nem is veheti figyelembe, tehát ezek cellák a statisztikát rontó tényezőkként lesznek jelen. Az eredeti geológiai térkép pedig esetenként a létrejött osztályozott térkép generalizált verziójaként is felfogható, ekkor a számszerű adatnál nagyobb pontosság is fennállhat. A korábban említett zaj-jelenséget különféle raszteres szűrési eljárásokkal (például medián szűrővel) némiképp ki lehetne iktatni, de az efféle manipulációkkal óvatosan kell bánni, ugyanis az összesített pontosságot egyaránt ronthatják vagy javíthatják, mivel nem valós helyzeteket is produkálhatnak.

Összefoglalás

Munkám során három eltérő geológiai eredetű területen végeztem el a geológiai formációk modellezését a random forest osztályozás segítségével. Célom annak a kérdésnek a kutatása volt, miszerint elegendő-e a morfometriai változók felhasználása a helyes osztályozáshoz.

Első lépésben felkutattam az ingyenesen hozzáférhető geológiai adatbázisokat és LiDAR (lézerszkennelt) adatokat az Amerikai Egyesült Államok területéről. Minden felhasznált adathoz a USGS (az Amerikai Egyesült Államok Földtani Szolgálata) online forrásaiból fértem hozzá. Ismertettem a vizsgált területeket földrajzi és geológiai szempontból.

Létrehoztam a LiDAR adatokból a területek terepmodelljeit (DTM), majd ezeket felhasználva legeneráltam a különféle morfometriai változókat. A tanulóterületek meghatározása után végrehajtottam a gépi tanuláson alapuló, random forest algoritmust felhasználó osztályozást. A módszertanon túl igyekeztem röviden bemutatni a random forest algoritmus működését.

Az algoritmus futtatása során több paraméter (mint például a fák száma, a változók száma vagy a tanulóterületek mérete) beállításával igyekeztem megkeresni a vizsgált területek legpontosabb modelljeit. Az kapott legjobb modellek eredményeinek kiértékeléséhez létrehoztam azok tévesztési mátrixait, melyekből meghatározhatók a pontossági értékek.

Végül pedig bemutattam a kapott eredményeket mindhárom vizsgált területről. Statisztikai elemzést végeztem a legjobb modellek pontosságértékeit illetően, valamint ismertettem az adott területeken a legjobb, illetve a legrosszabb eredményeket elérő formációkat. Kitértem az osztályozás hibáinak lehetséges okaira is. A kiértékelt modelleket térképek formájában is bemutattam.

Az 1-es számú, vulkanikus eredetű területet illetően arra a következtetésre jutottam, hogy kizárólag a morfometriai változókra hagyatkozva az osztályozás itt nem elvégezhető. Az elért összesített pontosság (overall accuracy) értéke csupán 18,7% lett. A terület legpontosabb formációja is csak 38%-os pontosságot ért el.

A 2-es számú, üledékes eredetű területen az eredmények kifejezetten jók lettek a 32 változós modell esetében 51,12%-os összesített pontossággal, legpontosabb formáció 69,2%-ot ért el. Ugyanezen terület 14 változós modellje ellenben jóval gyengébb eredményeket produkált 16,87%-os összesített pontossággal. Utóbbi modell legpontosabb formációja 42%-ot ért el. Az előbbi változat eredményeiből kiindulva kijelenthetnénk, hogy a vizsgált terület formációi előrejelezhetőek a morfometriai változókból, azonban ez a magasság meghatározó szerepe miatt – megfigyelve a két modell közötti különbségeket – nem teljesen igaz.

A 3-as számú, metamorf (átalakult) jellegű terület esetében jónak mondható eredmények is születtek egyes formációk esetében. A terület formációi túlnyomórészt modellezhetőek morfometriai változókkal. A számszerű pontosságok ezt nem támasztják alá, azonban a létrejött osztályok térbeli eloszlásait tekintve bizakodóak lehetünk. Az itt kapott összesített pontosság értéke azonban nem lett kiemelkedő, 21,51%-os. A legpontosabb formációk 83,8%-os, illetve 72,8%-os értékekkel produkáltak.

Az általam ismertetett módszer alkalmazható arid, csupasz területeken is, mellyel növelhető a pontosság. Ez kiegészíthető a LiDAR adatok intenzitásértékeivel is, mint változó. Továbbá a LiDAR adatokból levezetett morfometriai változókkal a globális vagy a helyi domborzatmodellekhez képest nagyobb felbontással modellezhetünk.

Irodalomjegyzék

(2021. 05 04). Forrás: VIGRA Homepage: http://ukoethe.github.io/vigra/

(2021. 05 04). Forrás: SAGA-GIS Tool Library Documentation (v8.1.1): https://saga-gis.sourceforge.io/saga_tool_doc/8.1.1/index.html

(2021. 05 04). Forrás: Accuracy Metrics: http://gsp.humboldt.edu/olm_2019/courses/GSP_216_Online/lesson6-2/metrics.html

(2022. 05 04). Forrás: USGS LidarExplorer: https://prd-tnm.s3.amazonaws.com/LidarExplorer/index.html#/

(2022. 05 04). Forrás: USGS 3D Elevation Program: https://www.usgs.gov/3d-elevation-program

Albert G., Ammar S. (2021). Application of random forest classification and remotely sensed data in geological mapping on the Jebel Meloussi area (Tunisia). *Arabian Journal of Geosciences*, 2240.

Bachri I., Hakdaoui M., Raji M., Benbouziane A. (2020). Geological mapping using random forests applied to remote sensing data: a demonstration study from Msaidira-Souk Al Had, Sidi Infi inlier (Western Anti-Atlas, Morocco). *IEEE International conference of Moroccan Geomatics.*

Bedini, E. (2009). Mapping lithology of the Sarfartoq carbonite complex, southern West Greenland, using HyMap imaging spectrometer data. *Remote Sensing of Environment*, 1208-1219.

Boggs, S. (1987). *Principles of sedimentology and stratigraphy (1st ed.).* Merrill Pub. Co.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

Carneiro C. D. C., Fraser S. J., Crósta A. P., Silva A. M., Barros C. E. D. M. (2012). Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon. *Geophysics*, 17-24.

Cracknell M. J., Reading A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, 22-33.

Doctor D. H., Orndorff R. C., Parker R. A., Weary D. J., Repetsky J. E. (2010). *Geologic Map of the White Hall Quadrangle, Frederick County, Virginia and Berkeley County, West Virginia.* U.S. Geological Survey.

Evarts, R. C. (2006). *Geologic Map of the Lacamas Creek Quadrangle, Clark County, Washington.* U.S. Geological Survey.

Grebby S., Naden J., Cunningham D., Tansey K. (2011). Integrating airborne multispectral imagery and airborne LiDAR data for enhanced lithological mapping in vegetated terrain. *Remote Sensing of Environment*, 214-226.

Hastie T., Tibshirani R., Friedman J. (2009). *The Elements of Statistical Learning.*

He J., Harris J. R., Sawada M., Behnia P. (2015). A comparison of classification algorithms using Landsat-7 and Landsat-8 data for mapping lithology in Canada's Arctic. *International Journal of Remote Sensing*, 2252-2276.

Kertész Á., Karátson D. (1997). Morfometria. In D. Karátson, *Magyarország földje.* Budapest: KERTEK 2000.

Pavlopoulos K., Evelpidou N., Vassilopoulos A. (2009). *Mapping Geomorphological Environments.*

Peck R. L., Matchen D. L., Hunt P. J. (2013). *Bedrock Geology of the New River Gorge National River, Map Sheet 4: Hinton and Talcott 7.5' Quadrangles, West Virginia.* West Virginia Geological and Economic Survey.

Pike R. J., Evans I. S., Hengl T. (2009). Geomorphometry: A Brief Guide. *Developments in Soil Science,* 3-30.

Radford D. D. G., Cracknell M. J., Roach M. J., Cumming G. V. (2018). Geological mapping in Western Tasmania using radar and random forests. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 3075-3087.

Köszönetnyilvánítás

Köszönettel tartozom témavezetőmnek és tanáromnak, Dr. Albert Gáspárnak a diplomamunka témájának meghatározásában nyújtott segítségéért, valamint a munkám során nyújtott folyamatos útmutatásáért és segítő tanácsaiért.

Továbbá a USGS (Amerikai Egyesült Államok Földtani Intézete) ingyenesen letölthető adatai nélkül sem jöhetett volna létre a dolgozat jelen formájában.

DIPLOMAMUNKA LEADÁSI
és
EREDETISÉG NYILATKOZAT

Alulírott **Gurály Attila**, Neptun-kód: **DAQK1S**

az Eötvös Loránd Tudományegyetem Informatikai Karának Térképtudományi és Geoinformatikai Intézetén

„**Geológiai előfordulások modellezése LiDAR adatok alapján**” című diplomamunkámat a mai napon leadtam.

Témavezetőm neve: **Dr. Albert Gáspár**

CD-t / DVD-t mellékelek *(aláhúzandó)*: igen nem

Büntetőjogi és fegyelmi felelősségem tudatában nyilatkozom, hogy jelen szakdolgozatom/diplomamunkám saját, önálló szellemi termékem; az abban hivatkozott szakirodalom felhasználása a szerzői jogok általános szabályainak megfelelően történt.
Tudomásul veszem, hogy szakdolgozat/diplomamunka esetén plágiumnak számít:

- szószerinti idézet közlése idézőjel és hivatkozás megjelölése nélkül;
- tartalmi idézet hivatkozás megjelölése nélkül;
- más publikált gondolatainak saját gondolatként való feltüntetése.

A témavezető által benyújtásra elfogadott szakdolgozat PDF formátumban való elektronikus publikálásához a tanszéki honlapon

HOZZÁJÁRULOK NEM JÁRULOK HOZZÁ

Budapest, 2022. május 15.

hallgató aláírása