Döntési fa és véletlen erdő osztályozási módszerekkel készített felszínborítási térképek pontosságának összehasonlító elemzése

Gudmann András[1] – Mucsi László[2]

[1] Doktorandusz, Szegedi Tudományegyetem, Természettudományi és Informatikai Kar, Természeti Földrajzi és Geoinformatikai Tanszék, 6722, Szeged, Egyetem u. 2-6., gudmannandras@gmail.com
[2] Egyetemi docens, Szegedi Tudományegyetem, Természettudományi és Informatikai Kar, Természeti Földrajzi és Geoinformatikai Tanszék, 6722, Szeged, Egyetem u. 2-6., mucsi@geo.u-szeged.hu

Absztrakt: Az osztályozó módszerek közül egyre népszerűbbek az adatbányászati, illetve gépi tanuláson alapuló eljárások. Ebben a tanulmányban két ilyen módszert, a döntési fa és a véletlen erdő osztályozó eljárást hasonlítjuk össze úgy, hogy a két módszerrel elkészített felszínborítási térképek pontosságát vizsgáljuk meg. A vizsgálathoz két 5x5 km nagyságú mintaterületet választottunk ki Csongrád megyén belül. Az osztályozás számára bemenő adatként Landsat 7 műholdfelvételeket, a belőlük számított mutatókat, illetve SRTM magassági adatokat használtunk fel a 2000-es évre vonatkozóan. Referencia adatként a CLC00-ás adatbázist használtuk föl. Az osztályozás eredményeként sikeresen állítottunk elő felszínborítási térképeket mindkét mintaterületre, amely térképek összpontossága 81 és 97% között volt. A tematikus rétegek összevetése alapján a véletlen erdő eljárás hatékonyabban képes ezt a feladatot megoldani.

Bevezetés

A felszínborítás térképezése a műholdas távérzékelés egyik állandó kutatási iránya, mivel ismerete szükséges a környezeti változások tanulmányozásához, a földi erőforrás kutatáshoz és kezeléshez, várostervezéshez, illetve a fenntartható fejlődéshez szükséges környezeti politika kialakításához. Az ingyenesen elérhető közepes felbontású műholdfelvételek növekedésével, új lehetőségek nyíltak a felszínborítás térképezésére. Ezen műholdfelvételek, egyre több új információt szolgáltatnak a Föld felszínéről, köszönhetően egyre javuló tulajdonságainak (térbeli felbontás, spektrális felbontás, időbeli felbontás, radiometrikus felbontás). Mindazonáltal, ahhoz, hogy a nagy mennyiségű adatot elemezni tudjuk és információt tudjunk kinyerni belőlük, gyors és hatékony módszerekre van szükségünk (Abriha et al. 2018). A hagyományos képosztályozási módszerek legtöbbje nagyfokú felhasználói beavatkozást igényel, illetve alkalmazásuk ekkora adathalmazon már nehézkes és nem hoz kielégítő eredményeket. Nagy mennyiségű adat osztályozására megfelelőek lehetnek, a gépi tanulás és az adatbányászat által is egyaránt használt algoritmusok (Döntési fák, véletlen erdők, szupport vektor gépek, neurális hálók). Ebben a tanulmányban két ilyen osztályozó eljárást, a döntési fát és a véletlen erdőt hasonlítjuk össze, az ezen osztályozókkal elkészített térképek alapján. A vizsgálatban két mintaterületre vonatkozóan az osztályozó algoritmusok változó parametrizálásával felszínborítási térképeket állítottunk elő Landsat 7-es műholdfelvételek spektrális sávjai, ezen sávokból számított mutatók, illetve SRTM magassági adatok alapján.

Felhasznált adatok

A kutatáshoz Landsat-7 műholdfelvételeket, az egyes felvételekből számított indexeket, transzformációkat, valamint SRTM magassági adatokat használtunk fel. Az SRTM adatot az Amerikai Egyesült Államok Geológiai Szolgálatának (USGS) adatbázisából, ingyenesen töltöttük le (https://earthexplorer.usgs.gov). A vizsgálatba vont műholdképek a 2000-es évben hat időpontban készültek áprilistól októberig terjedően, így egymástól spetrálisan jelentősen eltérő térbeli adatok álltak rendelkezésre az osztályozáshoz, így növelve az osztályozás pontosságát (Liska et al. 2017) (*1. táblázat*). A felszíni reflektancia értékeket tartalmazó műholdképeket és az ezekből számított NDVI, EVI, SAVI, MSAVI, NDMI értékeket az Earth Resource Observation and Science (EROS) Center Processing Architecture (ESPA) (https://espa.cr.usgs.gov) rendszerén keresztül rendeltük meg és töltöttük le. A műholdképek a WRS katalógus 187-es sorának 028-as oszlopának csempéi, melyeket Universal Transverse Mercator (UTM WGS84 N34) vetületben töltöttünk le.

Az Shuttle Radar Topograhy Mission (SRTM) a Nemzeti Térinformatikai Hírszerző Ügynökség (National Geospatial-Intelligence Agency) és a NASA közös projektje, amelynek célja magassági adatok gyűjtése a Föld legnagyobb részéről egy űrsiklóra telepített radarberendezés segítségével. A küldetést az Endeavour űrsikló hajtotta végre 2000. február 11. és 22. között. A küldetés során felmérték az északi szélesség 60° és déli szélesség 56° közötti területeket, ezzel lefedve a Föld szárazföldi területeinek 80%-át. Az adatok a feldolgozás után 3 fokmásodperces felbontással (körülbelül 90 méter) állnak rendelkezésre a felvételezett területekre. Az osztályozáshoz referencia adatként a Coordination of Information on the Environment (CORINE) Land Cover (CLC) felszínborítási térkép 2000-es évre vonatkozó adatát használtuk fel (CLC00). A CLC programot 1985-ben indította el az Európai Közösség azzal a céllal, hogy adatokat állítson elő a Közösség tagországainak felszínborításáról,

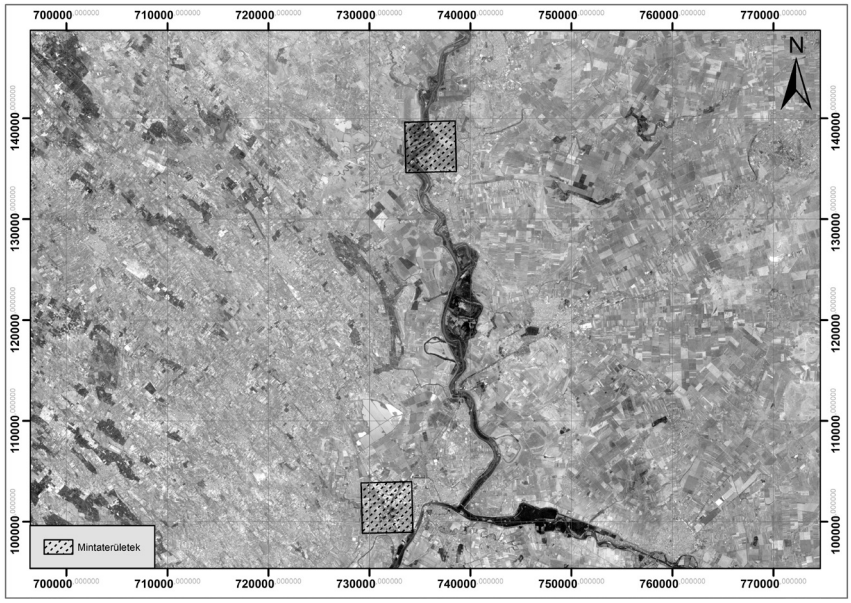
1. táblázat A Landsat 7-es műholdképek és a belőlük származtatott mutatók

Felvételezés időpontja	Felhasznált adatok
2000.04.30	Felszíni reflektancia(6) + termális sáv(1) + NDVI(1) + EVI(1) + SAVI(1) + MSAVI(1) + NDMI(1) + NDBI(1) + URBANI(1) + Tasseled Cap(6)
2000.05.16	Felszíni reflektancia(6) + termális sáv(1) + NDVI(1) + EVI(1) + SAVI(1) + MSAVI(1) + NDMI(1) + NDBI(1) + URBANI(1) + Tasseled Cap(6)
2000.07.03	Felszíni reflektancia(6) + termális sáv(1) + NDVI(1) + EVI(1) + SAVI(1) + MSAVI(1) + NDMI(1) + NDBI(1) + URBANI(1) + Tasseled Cap(6)
2000.08.04	Felszíni reflektancia(6) + termális sáv(1) + NDVI(1) + EVI(1) + SAVI(1) + MSAVI(1) + NDMI(1) + NDBI(1) + URBANI(1) + Tasseled Cap(6)
2000.08.20	Felszíni reflektancia(6) + termális sáv(1) + NDVI(1) + EVI(1) + SAVI(1) + MSAVI(1) + NDMI(1) + NDBI(1) + URBANI(1) + Tasseled Cap(6)
2000.10.23	Felszíni reflektancia(6) + termális sáv(1) + NDVI(1) + EVI(1) + SAVI(1) + MSAVI(1) + NDMI(1) + NDBI(1) + URBANI(1) + Tasseled Cap(6)

ezzel segítve az egységes környezeti politika kialakítását (Heymann 1993). A CLC adatbázis ehhez a célhoz mérten, 1:100 000-es méretarányban készült el 25 hektáros minimális térképezési egységgel (MMU), vonalas elemeknél legalább 100 méteres vastagsággal. A CLC 5 fő csoportban (mesterséges felszínek, mezőgazdasági területek, erdők és természetközeli területek, vizenyős területek és vízfelületek) 44 osztályt tartalmaz. (Mari – Mattányi 2002). A részletes nomenklatúra, a nagy időbeli felbontás, a nagy lefedettség és a közepes térbeli felbontás miatt ez az adatbázis megfelelő alapot biztosít a különböző környezeti folyamatok vizsgálatához.

Mintaterületek

A vizsgálathoz két 5x5 kilométeres területet választottunk ki Csongrád megyén belül: (a) egy mesterséges felszínekkel fedettet, Szeged nyugati városrészén, és (b) egy agrárterületet, ahol erdő is található, Szentestől délnyugatra a Tisza keleti oldalán (*1. ábra*). Az osztályozó eljárások bemutatásához és megfelelő összehasonlításához további kritérium volt, hogy a mintaterületeken minél több felszínborítási osztály legyen, de egyik aránya se haladja meg az 50%-ot. A szegedi mintaterületen 11 kategória van jelen, ezek közül legnagyobb arányban a 211-es, „Nem öntözött szántóföldek" (~27%), a 121-es, „Ipari vagy kereskedelmi területek" (~26%), illetve a 112-es osztály, „Nem összefüggő településszerkezet" (~24%). Az agrárterületen 7 osztály van jelen, a legnagyobb kiterjedésű kategória itt is a 211-es, „Nem öntözött szántóföldek" (~43%), a második és a harmadik legnagyobb a 311-es, „Lomblevelű erdők" (~28%), illetve a 231-es osztály, „Rét, legelő" (~19%).



1.ábra A mintaterületek elhelyezkedése Csongrád megyében

Módszerek

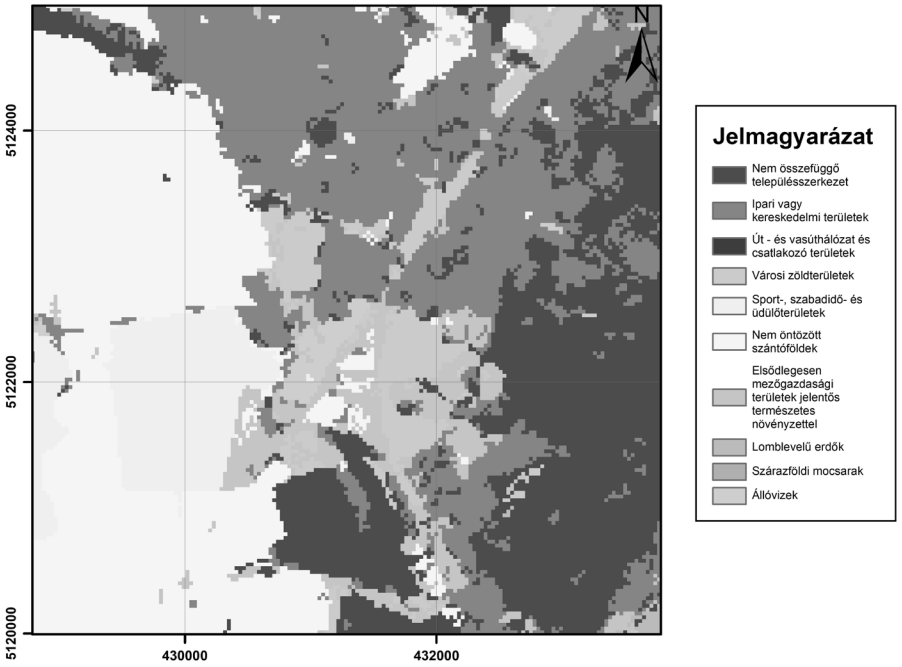
A tanulmányban a két mintaterületre állítottunk elő felszínborítási térképet a 2000-es évre vonatkozóan két osztályozó eljárást felhasználva. Az osztályozáshoz az adott évben hat különböző időpontban készült Landsat 7-es műholdfelvétel spektrális sávjait (7 sáv), a belőlük kiszámított vegetációs, nedvesség és beépítettségi indexeket (NDVI, EVI, SAVI, MSAVI, NDMI, NDBI, Urban index) (Szabó et al. 2016), Tasseled Cap transzformáció értékeit (6 sáv), és az SRTM magassági adatait használtuk föl. A műholdképekből eltávolítottunk minden hibás adatot (a maximális felszíni reflektanciaértéknél nagyobb, 10000 feletti pixelértékeket). A tisztított adatokat, a mutatókat és az SRTM adatokat a CLC00-ás referencia adatokkal együtt összefűztük egy többsávos képpé. Az így létrejött adathalmazt használtuk fel bemenő adatként az osztályozáshoz (122 attribútum, 28561 pixel/rekord). Két osztályozó eljárást alkalmaztunk, a döntési fát és a véletlen erdőt. A döntési fa egy hierarchikus osztályozási módszer, amely egy fára hasonlít (gyökér, ágak, csomópontok, levelek). Az algoritmus az adatokat úgy osztályozza, hogy azokat rekurzív módon egyre kisebb és homogénebb részekre bontja szét. A részekre bontás folyamata addig történik, amíg az összes pixel egy olyan osztályba nem kerül, amely teljesen elkülönül a többi osztálytól vagy az előre meghatározott feltételek nem teljesülnek (Jiang et al. 2010). Az általunk használt döntési fa a J48-as, ami az ID3-mas algoritmus kiterjesztése, és a lehető legkisebb modellt hozza létre (Bhargava et. al. 2013). A felépítéséhez utómetszést alkalmaztunk, ami a modell felépítése után eltávolít minden olyan végződést (összevonja magasabb szintre), ami nem növeli a fa összpontosságát. A modellalkotó algoritmus eljáráshoz két paramétert kell megadni, a (1.) konfidencia faktort (C): ami az utómetszésnél, mint határérték jelenik meg (a megadott konfidencia értéknél kisebb végződések lesznek „lemetszve”), illetve (2.) minimális objektum nagyságot (M): ennél az értéknél kisebb végződést nem alakíthat ki az algoritmus. A véletlen erdő módszer valójában döntési fák halmaza. Több döntési fa által leadott előrejelzéseket kombinálnak egy többségi szavazási séma segítségével. Mindegyik fa rögzített valószínűségi eloszlásból számított véletlen vektorok egy független halmazának értékei alapján jön létre. A véletlenszerűséget tovább fokozhatjuk ’zsákolás’ alkalmazásával. A zsákolás során az eredeti tanulóhalmazból véletlenszerűen választott mintákat viszünk a modellépítő eljárásba (Breiman 1996). Az erdő általánosított hibája függ az erdőben lévő egyes fák erejétől és a köztük lévő korrelációtól (Breiman 2001). A modellépítő eljárás során a véletlenszerűség növelésével csökkenthetjük a fák közötti korrelációt, ezáltal az általánosított hibát is (Tan et al. 2011). A véletlen erdő előnye a döntési fához képest, hogy robusztusabb a zajra és a túltanításra, képes kiegyensúlyozatlan és hiányos adathalmazokat is kezelni, miközben osztályozási pontossága nem romlik (Pal 2005). A modellépítő eljárás számára több paramétert kell megadni: a „zsák” nagyságát (P), azaz, hogy a teljes adathalmazhoz képest mekkora területű adathalmazt válasszon ki véletlenszerűen egy fa létrehozásához; az erdő nagyságát (N), hogy mennyi fát hozzon létre az algoritmus, illetve, hogy mennyi attribútumot (M) használjon fel az eljárás az egyes

fák létrehozásakor. A P nagyságú véletlenszerűen kiválasztott adatokon, M számú attribútum segítségével N db létrehozott döntési fa mindegyike lead egy szavazatot, és végeredménynek a leggyakoribbat fogjuk kapni (Breiman 2001). Ezen tanulmányban a zsákolásnak 33-at (a teljes adathalmaz 33%-a), 301 fát és az attribútumok számának (log2(összes_attribútum_száma) + 1) egyenlet alapján 7-et adtunk meg.

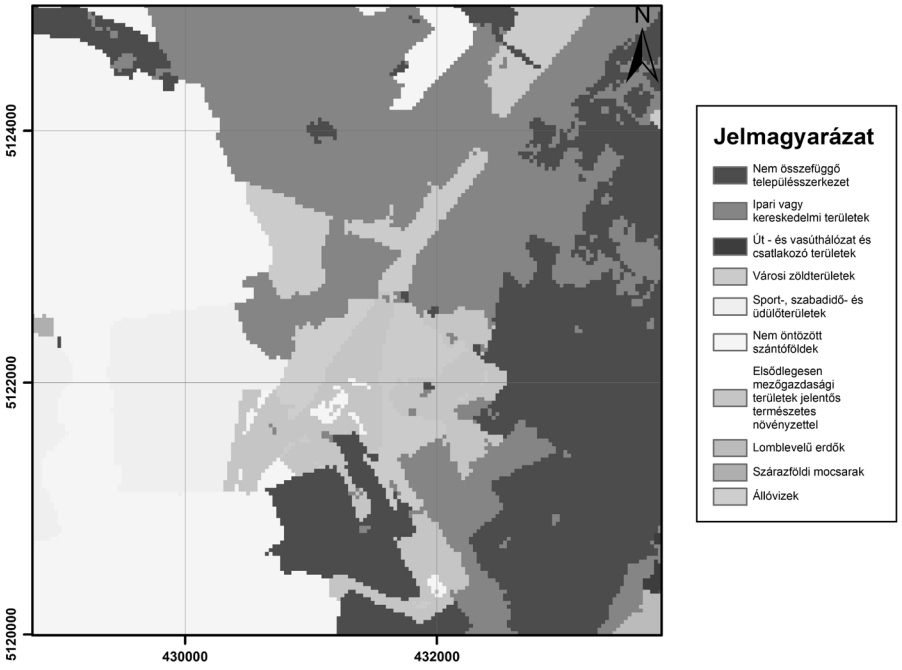
Eredmények

Az osztályozás eredményeként 2–2 tematikus réteg jött létre mindkét mintaterületre. Illetve létrehoztunk mindegyik osztályozott képhez egy térképet, amely az adott pixelre adott érték valószínűségét/megbízhatóságát adja meg. Utófeldolgozási műveletekkel eltávolítottuk a 3 pixelnél kisebb foltokat.

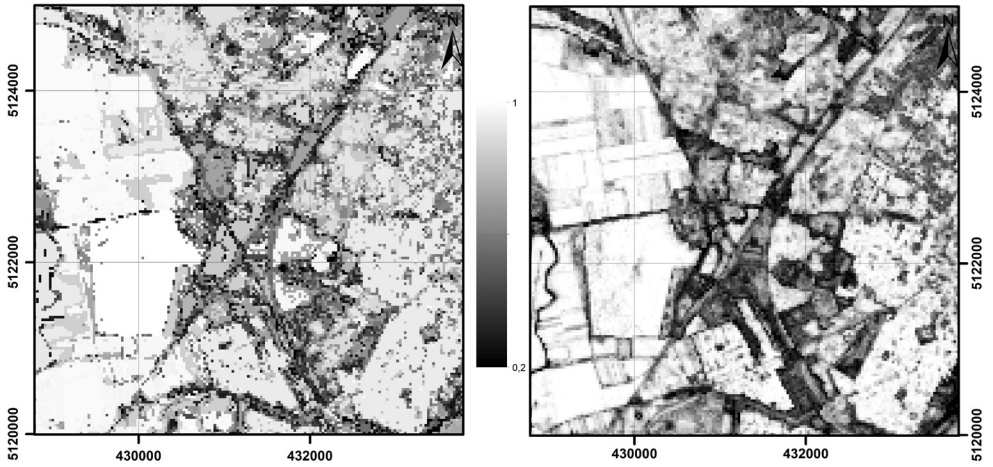
Az osztályozott képek pontosságbecslését a CLC00-ás adatbázis alapján végeztük el, meghatározva az egyes képek összpontosságát, felhasználói és előállítói pontosságát. A városi mintaterületre készített tematikus rétegek összpontossága mindkét osztályozó esetében kisebb, mint az agrár mintaterületre készítetteké. A döntési fa esetében a szegedi minterületre készített térkép (*2. ábra*) összpontossága 81,47%, míg az agrár területre készített képé 89,75% (*5. ábra*). A véletlen erdő osztályozó esetében az összpontosságok között kisebb az eltérés, 95,13% (Szeged – *3. ábra*) és 97,05% (Szentes – *6. ábra*). A mintaterületek közötti pontosság különbséget a városi területen található több osztály, illetve a mesterséges felszín osztályok komplexitása okozhatja. A szegedi területen található 11 osztály közül a döntési fa osztályozó eljárás esetében 3 nem osztályozódott (122 – Út- és vasúthálózat, csatlakozó területek, 231– Rét, legelők, 411 – Szárazföldi mocsarak), azaz nem jelent meg az eredmény térképen. Továbbá a mesterséges felszínekbe tartozó kategóriák mint a 112-es, 121-es, 141-es, 142-es előállítói és felhasználói pontossága megfelelő, 80% körül mozognak. Ez alól kivétel a 141-es kategória (Városi zöldterületek), amely nagymértékű átosztályozást mutat a 112-es, 121-es, 243-as osztályokkal. Ezen osztály előállítói pontossága 71,6%, míg felhasználói 58,3% csupán. Az említett osztályok között (112, 121, 141, 243) általános az átosztályozás, tehát ezek a legnehezebben elkülöníthető (egymáshoz spektrálisan leginkább hasonlító) osztályok. A legmagasabb pontossági értékeket a 142 – Sport-, szabadidő-és üdülő területek (előállítói: 94,7%, felhasználói: 91,6%) illetve a 211 – Nem-öntözött szántóföldek (előállítói: 90%, felhasználói: 94%). A véltlen erdő eljárással készített tematikus rétegen mind a 11 osztály jelen van, mindegyik osztály előállítói pontossága 90% fölötti. A felhasználói pontosságot megvizsgálva, csupán 4 osztály nem éri el a 90% fölötti pontosságot. Ezek a kategóriák a 122 – Út- és vasúthálózat és csatlakozó területek (33%), 231 – Rét / legelő (40%), 243 – Elsődlegesen mezőgazdasági területek jelentős természetes növényzettel (84%), 411 – Szárazföldi mocsarak (87,8%), ebből három olyan, amelyik nem lett osztályozva a döntési fa esetén. Ha az osztályozók által készített tematikus rétegek valószínűségi térképére tekintünk (*4. ábra*), láthatjuk, hogy a valószínűség alapján kisebb-nagyobb foltok azonosíthatók, nem osztály szinten értelmezhetőek az adatok. Ez az osztályozó eljárás működéséből adódik.



2.ábra A döntési fa osztályozás eredményeként létrejött tematikus réteg Szeged mintaterületre



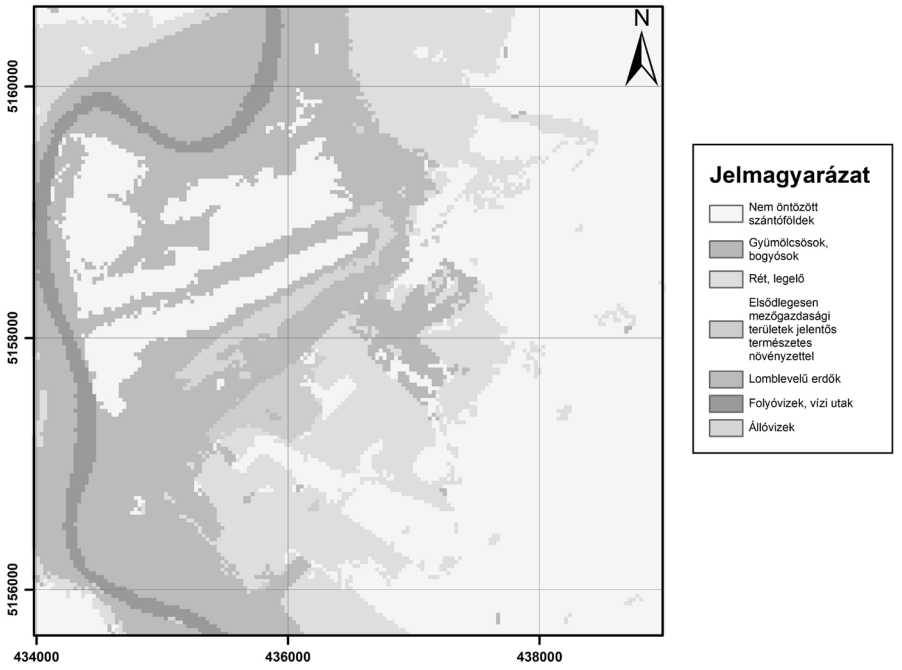
3.ábra A véletlen erdő osztályozás eredményeként létrejött tematikus réteg Szeged mintaterületre



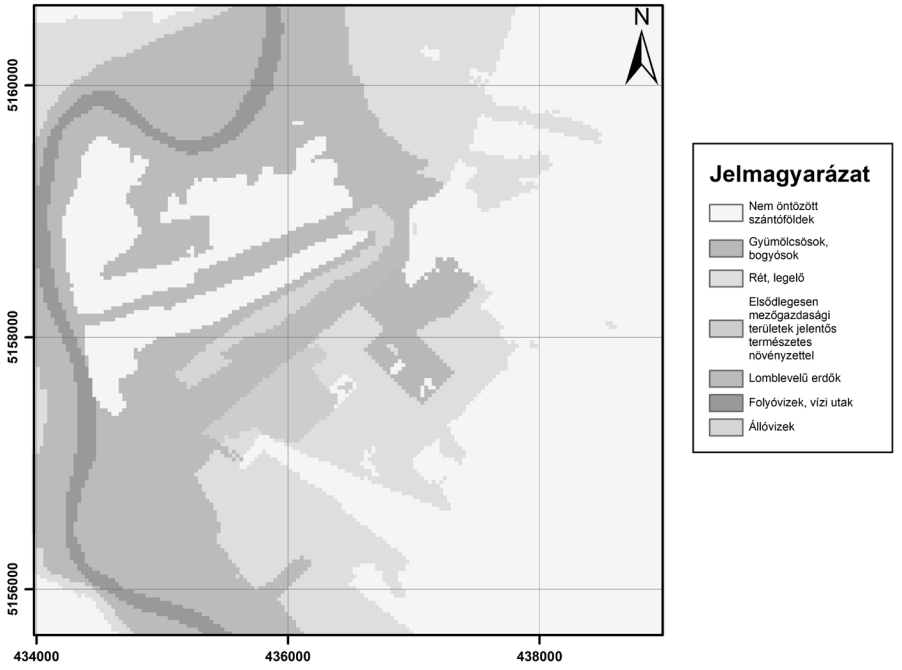
4.ábra A döntési fa (bal) és véletlen erdő (jobb) osztályozók által készített tematikus képek valószínűségi térképe Szeged mintaterületre

Általánosan elmondható, hogy a véletlen erdő osztályozás által adott becslések sokkal megbízhatóbbnak mondhatók, nagyobb a valószínűsége, hogy helyesen sorolta be a pixeleket az adott kategóriába. Továbbá jól kivehető, hogy a legkisebb valószínűségi értékek a foltok határain vannak, ahol spektrálisan vegyes pixelek találhatók. A döntési fa esetén kiugró értéket kapott a városi repülőtér egybefüggő nagy foltja (a kép nyugati oldalán), illetve a kép Észak-keleti sarkában, a Lencsés-horgásztó. Ezen kívül magas értéket kaptak a terület nyugati részén fekvő szántók, illetve a Vadaspark erdős részei, amelyek a kép közepén találhatók. A legkisebb valószínűségi értékek a több osztály foltjai közötti részeken láthatók. A véletlen erdő valószínűségi képén az általános valószínűség magasabb, illetve élesebbek a foltok közötti határok is, még egyes parcellák is kivehetők a kép Észak-nyugati részén. Ez a foltok szélein látható alacsony értékek és a foltok belsejében lévő nagy értékek közötti kontrasztnak köszönhető.

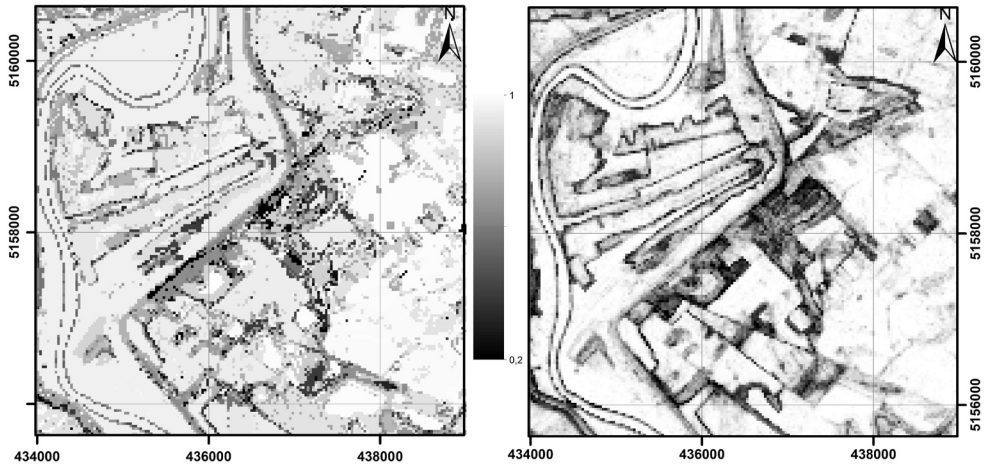
A szentesi mintaterületen 7 osztály volt található, többségük mezőgazdasági kategória. A döntési fa osztályozó esetében mind a 7 kategóriát sikerült lehatárolni. Ezen 7 kategória közül 3 kiválóan osztályozódott, a 211-es, 311-es, 511-es, mindegyik osztály előállítói és felhasználói pontossága 90% fölötti. A mezőgazdasági kategóriák (211, 222, 231, 243) és a lomblevelű erdők (311) között nagymértékű átosztályozás jelentkezett, illetve a lomblevelű erdők (311) és a vizek között. Emiatt az átosztályozás miatt a többi kategória pontosság változóan alakul, de egyik kategória pontossága sem kisebb mint 50%. A véletlen erdő osztályozás eredménytérképén minden osztály jól osztályozódott, előállítói pontosságuk 95% fölötti. A felhasználói pontosság csupán 3 esetben nem 95% fölötti, a 222 – Gyümölcsösök, bogyósok (90,7%), 243 – Elsődlegesen mezőgazdasági területek jelentős természetes növényzettel (82,4%), 512 – Állóvizek (83,3%). Ennek oka lehet, hogy a 243-mas nem jól definiált, nem egyértelmű felszínborítási osztály és ezért az osztályozó algoritmus nem tud megfelelő



5.ábra A döntési fa osztályozás eredményeként létrejött tematikus réteg Szentes mintaterületre



6.ábra A véletlen erdő osztályozás eredményeként létrejött tematikus réteg Szentes mintaterületre



7.ábra A döntési fa (bal) és véletlen erdő (jobb) osztályozók által készített tematikus képek valószínűségi térképe Szentes mintaterültre

szabályrendszert kialakítani rá, így más osztályokkal keveredik. A *7. ábrán* láthatjuk a két tematikus réteg valószínűség képét. Ugyanazon szabályszerűségek mondhatók el róluk, mint a szegedi mintaterület esetében, azaz a jó osztályozás valószínűsége a foltok szélein kisebb, mint a foltok belsejében, a vegyes pixelek miatt. Továbbá a több osztály találkozásánál lévő pixelekre adott kategória helyességének esélye a legalacsonyabb. Azonban ezen a mintaterületen a két osztályozó által adott becsléseknek a valószínűsége hasonló mértékű, kisebb a különbség a döntési fa és a véletlen erdő között. Mindazonáltal a véletlen erdő által adott becslések egy adott folton belül egységesek, így a kép kompaktabbnak nézz ki, nincsenek elszórt pixelek benne.

Összefoglalás

Az eredmények alapján megállapítható, hogy a döntési fa és a véletlen erdő osztályozó eljárások alkalmazásával sikeresen lehet felszínborítási térképet előállítani mind agrár, mind városi környezetre vonatkozóan. A két mintaterületre 2-2 tematikus kép készült el ezzel a két osztályozó eljárással, amely mintaterületeken 11, illetve 7 felszínborítási osztály található. Az osztályozás eredményeképp létrejött tematikus rétegek összevetésével megvizsgáltuk a két osztályozó eljárás hatékonyságát. Mind a döntési fa, mind a véletlen erdő eljárás sikerrel képes volt a területeken található kategóriák többségét osztályozni. A döntési fa osztályozó által adott osztályozott képekben azonban nagyobb a bizonytalanság, mint a véletlen erdő eljárással készített képek esetében. További előnyt jelent, hogy a véletlen erdő eljárás robusztusabb a túltanítással szemben, azonban a modell készítési ideje jóval több, mint egy döntési fáé. Ez alapján a véletlen erdő, amely több döntési fa felhasználásával készül, hatékonyabban képes ezt a feladatot megoldani, illetve az eljárás által adott becslések valószínűségének helyessége nagyobb.

Irodalomjegyzék

Abriha, D. – Kovács, Z. – Ninsawat, S. – Bertalan, L. – Balázs, B. – Szabó, Sz. (2018): Identification of roofing materials with Discriminant Function Analysis and Random Forest classifiers on pan-sharpened WorldView-2 imagery – a comparison, Hungarian Geographical Bulletin, 67, pp. 375–392.

Bhargava, N. – Sharma, G. – Bhargava, R. – Mathuria, M. (2013): Decision Tree Analysis on J48 Algorithm for Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, 3(6), pp. 1114–1119.

Breiman, L. (1996): Bagging predictors. Machine Learning, 26, pp. 123–140.

Breiman, L. (2001): Random forests. Machine Learning, 45, pp. 5–32.

Heymann, Y. (1993): CORINE land cover Technical guide – EUR 12585, Office for Official Publications of the European Communities, Luxembourg, pp. 14–15.

Liska, Cs. M. – Mucsi, L. – Henits, L. (2017): Hosszú-távú felszínborítás-változások vizsgálata Csongrád megyében idősoros adatok felhasználásával, Random Forest módszerrel, Földrajzi Közlemények CXLI. (L.), pp. 71–83.

Mari, L. – Mattányi, Zs. (2002): Egységes európai felszínborítási adatbázis a CORINE Land Cover program. – Földrajzi Közlemények CXXVI. (L.), pp. 31–38.

Pal, M. (2005): Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26(1), pp. 217–222.

Jiang, L. – Wang, W. – Yang, X. – Xie, N. – Cheng, Y. (2010): Classification Methods of Remote Sensing Image Based on Decision Tree Technologies, Computer and Computing Technologies in Agriculture IV. Li, D., Liu Y., Chen, Y. (Ed.), pp. 353–358.

Szabó, Sz. – Gácsi, Z. – Balázs, B. (2016): Specific features of NDVI, NDWI and MNDWI as reflected in land cover categories, Landscape & Environment, 10(3–4), pp.194–202.

Tan, P. N. – Steinbach, M. – Kumar, V. (2011): Bevezetés az adatbányászatba.