

A CORINE felszínborítási térkép automatikus előállításának lehetősége döntésifa-osztályozó segítségével

Gudmann András – Mucsi László – Henits László

DOI: <https://doi.org/10.30921/GK.71.2019.2.2>

Absztrakt: Az egyik legismertebb felszínborítási térkép a CORINE (Coordination of information on the environment) Land Cover (CLC), amely az Európai Környezetvédelmi Ügynökség (EEA) koordinálásával készül el az ügynökség 39 tagországának területére vonatkozóan. Ez az ingyenes adatbázis, részletes nomenklatúrájával (44 osztály), kellő tematikus pontosságával és időbeli felbontásával (1990-től napjainkig) megfelelő alapot biztosít a különböző környezeti folyamatok vizsgálatához. Az adatbázis hátránya, hogy úrfelvételek vizuális interpretációjával készül, ezért előállítása időigényes, és az eredmények nagyban függenek az interpretáló személyek szakmai tudásától. Ezen probléma megoldását, az úrfelvételek osztályozása jelenti. Ezen tanulmányban döntésifa-osztályozást alkalmaztunk, hogy e módszerrel előállítsuk a CLC00-s adatbázist Csongrád megye területére. A vizsgálat alapját az ingyenesen elérhető Landsat 7 ETM+ adatok szolgáltatták. Az osztályozáshoz három különböző időpontban készült úrfelvételt (2000. április 30., 2000. július 3. és 2000. augusztus 20.), valamint a távérzékelte adatokból levezetett Normalizált Differenciált Vegetációs Index (NDVI) és Tasseled Cap (TC) értékeket és Shuttle Radar Topography Mission (SRTM) magassági adatokat használtunk. Az összes adat és az egész adathalmazt tanulóterületként felhasználva 78,6%-os összpontosságot értünk el. Az osztályozó eljárás képes volt lehatárolni nem csak a tisztán felszínborítási, hanem a legtöbb területhasználati osztályt is, így alkalmas CLC felszínborítási és egyes területhasználati osztályainak automatikus előállítására.

Abstract: One of the best known land cover maps is the CORINE Land Cover (CLC), which is co-ordinated by the European Environment Agency (EEA) for the 39 member states of the agency. This freely available data, with its detailed nomenclature (class 44), with a sufficient thematic accuracy and temporal resolution (from 1990 to the present) provides an adequate basis for examining various environmental processes. The disadvantage of the database is that it is made with visual interpretation of remote sensed images, so its production is time-consuming and the results are highly dependent on the professional knowledge of the interpreters. The solution to this problem is image classification. In this study, decision tree classification was used to produce the CLC00 database for Csongrád County, in Hungary. The study was based on the free available Landsat 7 ETM+ data. Three different Landsat image (30 April 2000, 3 July 2000 and 20 August 2000) as well as the Normalized Difference Vegetation Index (NDVI) and Tasseled Cap (TC) derived from the Landsat images and Shuttle Radar Topography Mission (SRTM) elevation data were used. Using all data and the entire data set as a training set, we achieved an overall accuracy of 78,6%. The classification process was able to delimit not only the pure land cover, but also most of the land use classes, so it is suitable for automatic production of the land use and some land use classes of the CLC map.

Kulcsszavak: döntési fa, felszínborítás, gépi tanulás, területhasználat

Keywords: decision tree, land cover, machine learning, land use,

Bevezetés

A Föld felszínének vizsgálata és a felszínborítás térképezése a távérzékelés alapvető feladata. Az egyik legismertebb felszínborítási térkép a CORINE (Coordination of information on the environment) Land Cover (CLC), amely az Európai Környezetvédelmi Ügynökség (EEA) koordinálásával készül el az EEA 39 tagországának területére. A háromszintes, részletes nomenklatúra (44 osztály a 3. szinten), a kellő tematikus pontosság és az időbeli felbontás (1990-től napjainkig) megfelelő alapot biztosít a különböző felszínborítottsággal kapcsolatos folyamatok vizsgálatához. Az adatbázis

az úrfelvételek számítógéppel segített vizuális interpretációjával készül, ezért előállítása időigényes, és az eredmények nagyban függenek az interpretáló személyek szakmai tudásától (Mari-Mattányi 2002). Jelen tanulmány fő céljai a következők voltak: a CORINE nomenklatúra szerinti tematikus térkép automatikus előállításának lehetőségének vizsgálata Csongrád megye területére, döntésifa-osztályozás használatával; illetve annak elemzése, hogy milyen mértékben befolyásolják a bemenő adatok (a tanítóterület nagyságának, illetve a döntési fa egyes paramétereinek változtatása) a végeredményt.

Az adatbányászat, azaz a nyers adat értelmezésének, megjelenítésének és

az adatban rejlő minták és információk kinyerésének egyik leghatékonyabb módja a döntési fa típusú osztályozás, (Bányász 2010). 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 osztályozás szempontjából fontos adatok a döntési fa gyökeréhez közel, a másodlagos adatok a gyökértől

távolabb helyezkednek el. A döntési fa előnye, hogy a végeredményt a felhasználó könnyen értelmezheti, valamint a módszer, különböző változótípusú (nominális, szám, szöveg) információkat is tud együttesen kezelni. A módszer képes a hibás adatok kiszűrésére és a hiányzó adatok ellenében is működni, valamint a felhasználó nagy mennyiségű adatot képes kis munkacél-ráfordítással kezelni (Bhargava et al. 2013).

Mintaterület

Csongrád megye Magyarország dél-délkeleti részén fekszik, 4 262 km²-es területének nagy része síkság, jellemzően kicsik a magasságkülönbségek és a tengerszint feletti magasság sehol se éri el a 200 métert. Két legnagyobb városa, melyekben a népesség több mint fele él, Szeged és Hódmezővásárhely. A megyén áthaladó legjelentősebb vízfolyások a Tisza, a Körös és a Maros, legnagyobb tava a Szeged közelében található Fehér-tó. A térségben főleg agrártermelés folyik, köszönhetően annak, hogy a terület mezőgazdasági művelésre különösen alkalmas, területének 84%-a termőföld, amelynek nagy

része jó minőségű csernozjom, öntés-és réti talajok. Ebből adódik, hogy a termőterületek aránya kimagasló, a megye 75%-a, illetve, hogy a leggyakoribb területhasználati forma a szántó, a megye 60%-a (Csőszné-Tölcsér 2015). Ezen kívül a térségben nagy arányban található erdőket, gyepeket, illetve gyümölcsösöket.

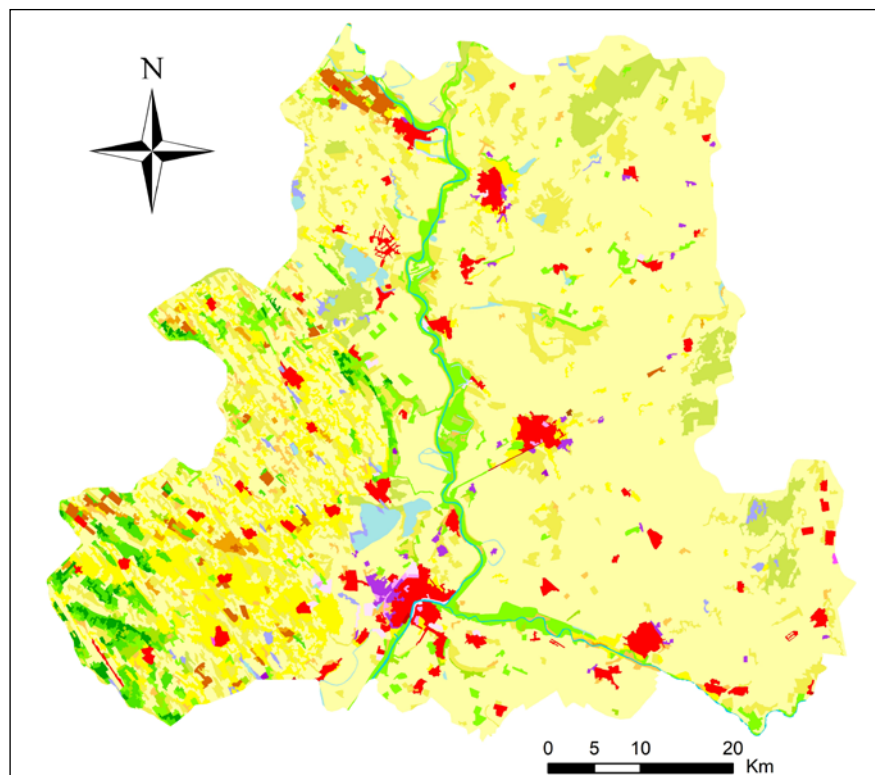
Adatok

A vizsgálat alapját a Landsat 7-es műhold ETM+ szenzorjának felszíni reflektanciaértékké transzformált adatai szolgáltatták a 2000-es év, augusztus 20-i időpontjára vonatkozóan, amely a mintaterület vizuális interpretációjának alapjául is szolgált a CORINE-térképezés során. A döntésifa-módszer hatékonyságának (térképpontosságának) növelése érdekében további kettő, 2000. április 30-án és július 3-án készült Landsat 7-es műholdfelvételt, illetve SRTM magassági adatokat vettünk be az elemzésbe. Ezen távérzékelte adatokon kívül további információkat tartalmazó rétegeket vettünk be az elemzésbe, melyeket gyakran használnak a tájértékelési és a felszínborítással kapcsolatos elemzések során (Hussein et al. 2017, Szabó et al. 2016, Szilassi

et al. 2017.). Ezek szerint a műholdképek adataiból levezetett Normalizált Differenciált Vegetációs Index (NDVI), illetve a Tasseled Cap transzformáció 3 adata (brightness, greenness, wetness) azaz rétegek kerültek az adatbázisba. Referenciaadatként a CLC00 vektoros adatainak 30×30 méteres rasterre alakított tematikus rétege szolgált.

Módszerek

Az adatok feldolgozásához ERDAS Imagine¹ és WEKA² szoftvereket, a programok közötti adattípusok transzformációjához pedig python scripteket használtunk. A műholdfelvételek reflektanciaértékei az előkészítés során tisztítva lettek egy ERDAS-ban készített modell segítségével, mely kiszűrte minden olyan értéket, amely nem valós adat (10 000 fölötti értékek). További lépésként a műholdképekből a mintaterületet kivágtuk Csongrád megye határát tartalmazó poligon segítségével. Az Landsat 7 ETM+ képekből számított NDVI és a Tasseled Cap transzformáció első 3 sávja, valamint a műholdképekhez hasonlóan a mintaterületre kivágott és a műholdfelvételek rácshálójára igazított SRTM magassági értékek egy állományba kerültek egysítésre a CLC00 adataival együtt. Az így létrejött adathalmaz (4 805 195 pixel 34 attribútummal) szolgált a döntésifa-osztályozás alapjául. A 34 dimenziós adathalmazt az ERDAS-ból tagolás nélküli ASCII-ban egyszerű szöveges (.asc) formátumba kiexportáltuk, majd egy python script segítségével a WEKA-program számára olvasható, vesszővel tagolt szövegfájl (.csv) formátumra átalakítottuk. A WEKA-programba beolvasott adatokat véletlenszerűvé alakítottuk a programba beépített szűrővel. Ezen adatokból többféle tanulólánc került kijelölésre, melyekben változtattuk a felhasznált adatok (pixelek) mennyiségét a teljes adathalmazhoz képest, valamint a felhasznált adattípusokat. Az osztályozáshoz a programban elérhető J48-as döntési fát használtunk. Ez az algoritmus az ID3-as algoritmus kiterjesztése, és a lehető legkisebb modellt hozza létre.



1. ábra. A CLC00 felszínborítási térkép Csongrád megye területére

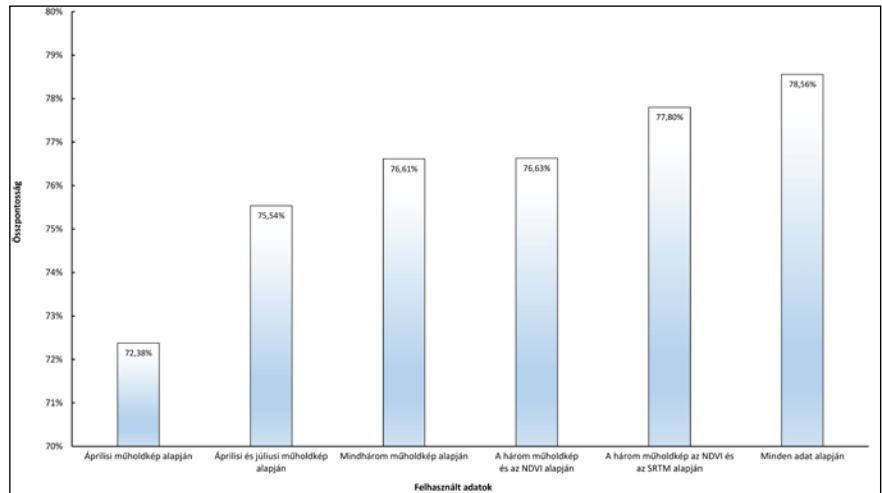
¹ <https://www.hexagongeospatial.com/products/power-portfolio/erdas-imagine>
² <https://www.cs.waikato.ac.nz/ml/weka/>

A modell építéséhez három alaplépést ismételt az algoritmus: 1. ellenőrzi, hogy az ágon lévő összes eset egy osztályba tartozik-e, ha igen akkor az egy végződés (levél) lesz és megkapja az osztály nevét, ha nem, akkor az adatok szétválasztása tovább folyik. 2. minden attribútumra az információ és információnyereség kiszámítása; 3. a számítások alapján a legjobb attribútum kiválasztása a felosztáshoz és a felosztás elvégzése. A felépítéséhez utómetszést alkalmaztunk, ami a modell felépítése után eltávolít minden olyan végződést (összevonja magasabb szintre), ami nem növeli a fa összpontosságát. A modell felépítéséhez pedig feltételként megadtuk, hogy minimum száz rekordként (pixelenként) alakítson ki végzódéseket a modellépítő algoritmus (Bhargava et. al. 2013). A különböző tanulódatok mindegyikére döntésifa-modell készült, és ezek által a teljes adathalmazt osztályoztuk. Az osztályozás pontosságát az összpontosság (*overall*), a felhasználói (*user's*) és a készítői (*producer's*) pontosság vizsgálatával határoztuk meg. A pontossági adatokat árnyalja, hogy az osztályozási eljárás pixelalapú, így sokkal kisebb a minimális térképezési egysége (900 négyzetméter), mint a CLC-térkép legkisebb foltja, ami legalább 25 hektáros.

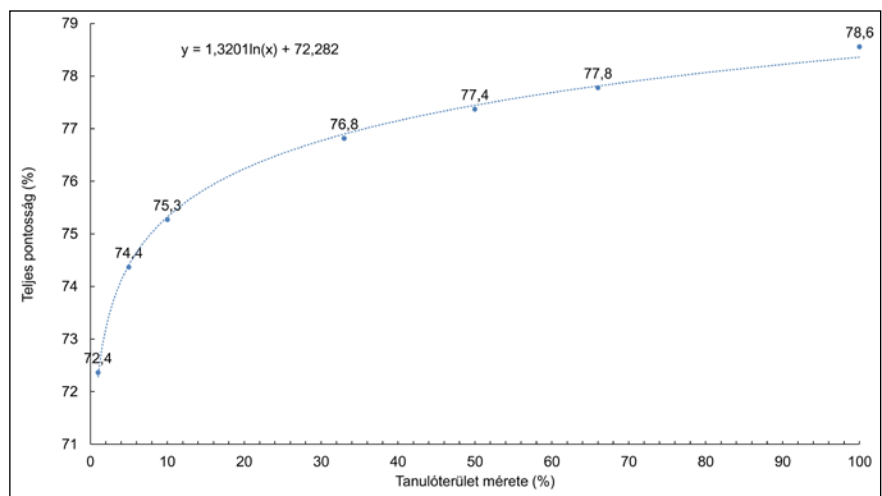
Eredmények

A futtatásokból kiderült, hogy a különböző időpontban készített összes műholdfelvétel, az SRTM magassági adatok és a Tasseled Cap transzformáció adatainak együttes felhasználása esetén érhető el a legnagyobb összpontosság (78,56%) (2. ábra).

Megállapítható, hogy csupán az NDVI-index felhasználása nem eredményezett pontosságjavulást (>0,02%), mely abból adódhat, hogy a -1, 1 közötti NDVI-értékek egydimenziós histogramjában nem különülnek el élesen a CLC-osztályok. A legnagyobb pontosságjavulást a júliusi műholdkép felhasználása okozta. Feltehetően az időbeli eltérés miatt kialakuló reflektanciakülönbségek információtartalma eredményezte a pontosság javulását. Teszteltük a tanulótérlet növelésének hatását az osztályozás



2. ábra. A teljes pontosság (Overall Accuracy) változása a bevitt adatok függvényében

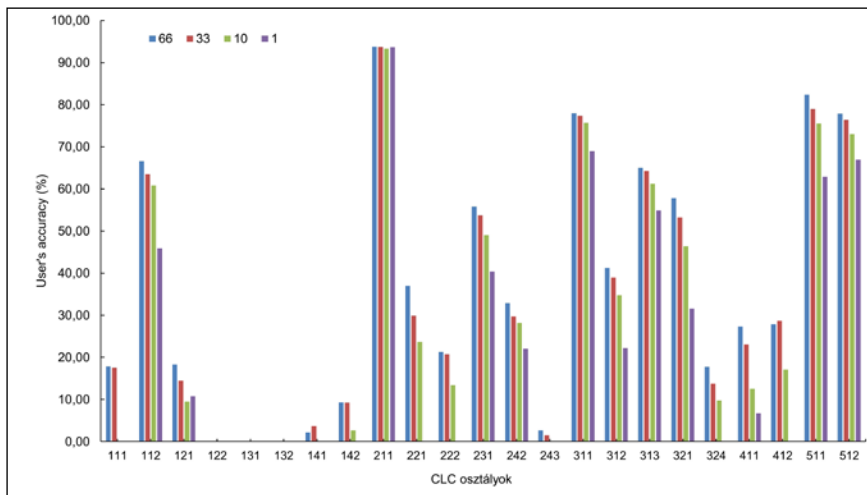


3. ábra. A teljes pontosság változása a tanulótérlet méretének függvényében

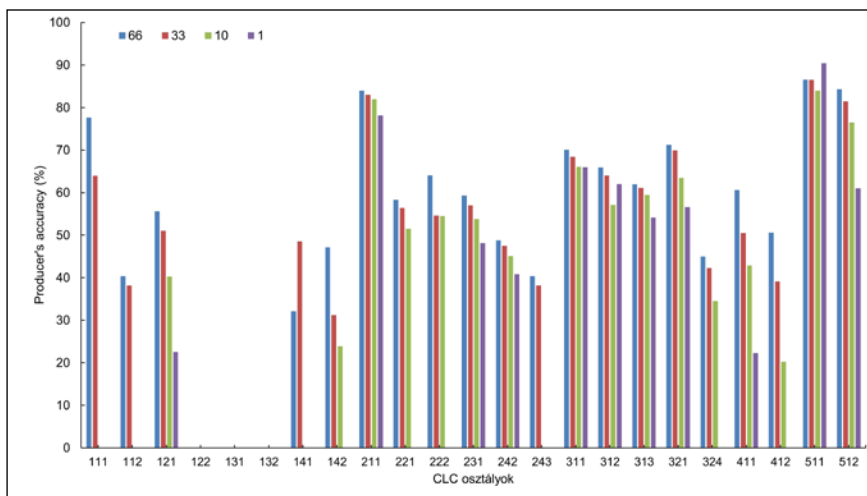
pontosságára és megállapítható, hogy a növelés mértékével együtt nő a létrehozott döntési fa mérete és végzódéseinek száma is. Mivel a modellépítéshez több adat áll rendelkezésre, így több végzódést tud kialakítani a modellépítő eljárás.

A 3. ábrán jól látszik, hogy a pontosság logaritmikusan változik a tanulótérlet méretének változásával. Vagyis, viszonylag kis tanulótérlet-méretnél elérhető a 75%-os pontosság, és újabb tanulótérlet bevonása már nem eredményez jelentős pontosságjavulást (2%-nál nem nagyobb egyik esetben sem a javulás). Így a legnagyobb növekedés annál a döntési fánál mutatkozik, amelyben a modellépítő eljárás az összes pixel 1%-át használja fel a tanuláshoz. Az osztályozások felhasználói és készítői pontosságát megfigyelve (4., 5. ábra) több megállapítás is tehető. Egyrészt, a ténylegesen felszínborítási osztályoknak tekinthető

CLC-kategóriák (2.3.1. – rét/legelő, 3.1.1. – lomblevelű erdők, 4.1.1. – szárazföldi mocsarak, 5.1.1. – folyóvizek, vízi utak, 5.1.2. – állóvizek) és a mintaterületen belül nagy kiterjedésű osztályok (1.1.2. – nem összefüggő településszerkezet, 2.1.1. – nem öntözött szántóföldek) jobban elkülöníthetők a többi osztálytól, így pontosságuk nagyobb, illetve ezek az osztályok a tanulótérlet csökkentésre is robusztusabban reagálnak, kisebb mértékben romlik az osztályozási pontosságuk. Másrészt, a kis területtel (az egész mintaterülethez képest) rendelkező osztályok a véletlenszerűsített adathalmazban (ami később tanulódattáként szolgált) kisebb valószínűséggel jelentek meg, és ha megjelentek, akkor se volt elegendő attribútum róluk, hogy osztályozva legyenek. Ez a jelenség erősödött a tanulótérlet-nagyság csökkentésével, így ezek az osztályok egyre nagyobb mértékben nem



4. ábra. A tanulóterület csökkentéssel létrehozott döntési fák osztályainak felhasználói pontossága



5. ábra. A tanulóterület csökkentéssel létrehozott döntési fák osztályainak készítői pontossága

osztályozhatók. A területhasználati jellegű osztályok (pl.: 1.2.1. – ipari vagy kereskedelmi területek, 1.4.2. – sport-, szabadidő- és üdülőterületek, 2.2.1. – szőlők, 2.4.2. – komplex művelési szerkezet) elkülönítése a spektrális térben komplikált, sokszor nem lehetséges, ebből adódóan a döntésfá-osztályozás összpontossága ezeknél az osztályoknál 50% körül mozog.

Ezen osztályoknál azonban a tanulóterület csökkentésének hatására jelentősen romlik mind a felhasználói, mind a készítői pontosság. A legjobb pontossággal rendelkező modell által osztályozott kép a 6. ábrán látható; ez a modell a teljes mintaterületet használta fel tanulóterületként, és minimum 100 rekordként (pixelenként) alakított ki végződéseket. A térképen lévő osztályok a térben a CLC-adatbázissal megegyezően helyezkednek el, durva osztályozási hiba nincs. A folyók, tavak és még

a holtágak is jól kirajzolódnak, könnyen azonosíthatók. A települések jól osztályozódtak, egységes halmazoknak látszanak, az azonosítás ebben az esetben sem probléma, ennek oka az SRTM magassági adatok által, a környező felszínektől való jó elkülöníthetőség (a modellben a legtöbb városi osztályhoz köthető döntési szabály elején áll SRTM-adat). A folyók mentén jól kivehetők az ártéri erdők, és a Duna–Tisza közén lévő erdőségek is könnyen felismerhetők. A szántóföldekben különböző osztályok foltjai találhatók, főleg a Duna–Tisza térségét vizsgálva, ez a pixelalapú osztályozás és az eredeti térkép minimális térképezési egysége közötti eltérésből adódik. Az egyes osztályok a földrajzi térben a megfelelő helyen jelennek meg, a legtöbb osztályozási hiba az osztályok határain lévő pixelek téves kategorizálásából ered. Ez több okból is származhat, egyrészt

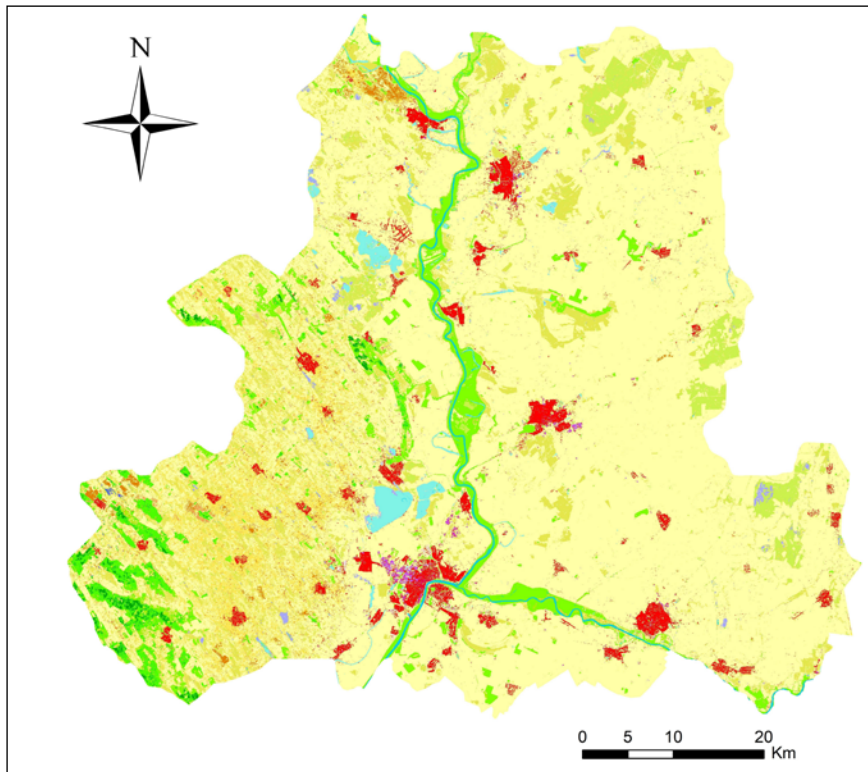
ezen pixelek lehetnek spektrálisan vegyesek, másrészt a pixelalapú osztályozás részletesebb eredményt adhat, mint a CLC 25 hektáros minimális térképezési egysége, amely így hibaként jelenik meg. A létrehozott térképen a CLC-osztályok közül a felszínborítási vagy ahhoz közel álló osztályok jól osztályozódtak, a CORINE-adatbázissal nagyban egyező térképet adtak vissza. A CLC-osztályok területhasználati kategóriái közül a legtöbb közepesen osztályozódtak, de több közülük egyáltalán nem osztályozódtak, azaz meg sem jelentek a tematikus rétegen.

Következtetések

A döntési fa osztályozása a vizsgálatba vont úrfelvételek és egyéb független adatrétegek helyes kiválasztása esetén alkalmas volt a CLC-adatbázis felszínborítási osztályait nagy pontossággal lehatárolni. A területhasználati jellegű osztályoknak viszont csak kisebb részét sikerült megfelelően osztályozni, ezért további, elsősorban a mintázatot, a térbeli struktúrákat jobban leíró adatok bevonására van szükség a későbbi kutatások során. Megállapítható továbbá, hogy a döntési fa osztályozási módszer hatékonyan képes nagy mennyiségű többdimenziós térbeli információt feldolgozni, elemezni, és ilyen nagy összetettségű feladatot megoldani, mint az általánosságban használt automatikus vagy irányított osztályozási módszerek. Elmondható továbbá, hogy a különböző, egymástól független adatok bevonása az osztályozásba növeli az összpontosságot, illetve hogy a tanulók területének növelése logaritmikusan növeli a pontosságot, de ezzel arányosan növeli a modellépítés időtartamát, és a modell nagyságát, így a tanulóterület nagyságát egy bizonyos határ fölé nem érdemes növelni.

Köszönetnyilvánítás

A kutatás az NKFIH 124648K azonosítójú „A felszínborítás dinamikájának idősoros vizsgálata közepes- és nagyfelbontású úrfelvételek segítségével” c. OTKA kutatási projekt és az Emberi Erőforrások Minisztériuma 20391-3/3018/FEKUSTRAT támogatásával valósult meg.



6. ábra. A döntésifa-osztályozás eredményeképp létrejövő tematikus térkép

Irodalomjegyzék

- Bányász M. 2010. *Klasszifikáció az adatbányászatban*, Eötvös Lóránd Tudományegyetem, Budapest
- 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.
- Csószné, S. I. - Tölcsér M. (2015) *Csongrád megye számokban.*, KSH
- Hussein, S. O. - Kovács, F. - Tobak, Z. 2017. *Spatiotemporal Assessment of Vegetation Indices and Land Cover for Erbil City and Its Surrounding Using Modis Imageries*, Journal of Environmental Geography Vol. 10 (1-2.) pp. 31-39., DOI: <https://doi.org/10.1515/jengeo-2017-0004>
- 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.) 71-83.
- Mari, L. - Mattányi, Zs. 2002. *Egységes európai felszínborítási adatbázis a CORINE Land Cover program.*, Földrajzi Közlemények CXXVI. (L.) pp. 31-38.
- Mucsi, L. - Liska, Cs. M. - Henits, L. - Tobak, Z. - Csendes, B. - Nagy, L. 2017. *The evaluation and application of an urban land cover map with image data fusion and laboratory measurements*, Hungarian Geographical Bulletin 66. pp. 145-156., DOI: <https://doi.org/10.15201/hungeobull.66.2.4>
- 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., DOI: https://doi.org/10.1007/978-3-642-18333-1_41

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., DOI: <https://doi.org/10.21120/LE/10/3-4/13>

Szilassi, P. - Bata, T. - Szabó, Sz. - Czúcz, B. - Molnár, Zs. - Mezősi, G. 2017. *The link between landscape pattern and vegetation naturalness on a regional scale*, Ecological Indicators 81:252-259, DOI: <https://doi.org/10.1016/j.ecolind.2017.06.003>



Gudmann András
doktorandusz

Szegedi Tudományegyetem, Természeti Földrajzi és Geoinformatikai Tanszék
gudmannandras@gmail.com



Dr. Mucsi László
egyetemi docens

Szegedi Tudományegyetem, Természeti Földrajzi és Geoinformatikai Tanszék
mucsi@geo.u-szeged.hu



Dr. Henits László
egyetemi adjunktus

GPS TUNER SYSTEM Kft.
henits@geo.u-szeged.hu