KLASIFIKACE VYBRANÝCH TŘÍD POKRYTÍ ÚZEMÍ Z CORINE SYSTÉMU S VYUŽITÍM DRUŽIOVÝCH DAT SENTINEL-2

Lucie Stará[1]

1. ČVUT v Praze, Fakulta stavební, Katedra geomatiky, Thákurova 2077/7, 166 29, Praha, lucie.stara@fsv.cvut.cz

ABSTRAKT

Práce se věnuje klasifikaci problematických tříd zavlažovaná orná půda, pastviny a přírodní travní porosty. Klasifikace byla provedena ve třech evropských lokalitách (Španělsko, Makedonie, Turecko) se zařazením multitemporálních dat. Kromě optických dat Sentinel-2 ke klasifikaci přispěly především kanály NDVI a topografická data. Trénovací data byla vytvořena na podkladu databáze CORINE. Pro klasifikaci byla využita metoda Random Forest a určeny nejdůležitější příznaky. Nejlépe se podařilo klasifikovat třídu zavlažovaná orná půda (precision 98,25 %), dále přírodní traviny (89,30 %) a nakonec pastviny (81,17 %). Pozorována byla záměna mezi přírodními travinami a pastvinami, kterou se zvoleným postupem nepodařilo plně eliminovat.

KLÍČOVÁ SLOVA

pokrytí území, pastviny, přírodní travní porosty, trvale zavlažovaná orná půda, Sentinel-2, CORINE, Random forest

ÚVOD

Obsah příspěvku vychází z mezinárodního projektu Geo-harmonizer, který se zaměřuje na propojení a harmonizaci dostupných geografických dat a jejich poskytování skrz webově orientovaný systém. Rámec zpracování stojí na využití otevřených dat a metodách strojového učení dostupných v open source software. Data s celoevropským rozsahem v něm budou poskytována se zaměřením na různé tematické okruhy. [1], [2] Jedním z připravovaných okruhů je i land cover (jinak také pokrytí území, LC). Jedná se o charakteristiku, která popisuje fyzický pokryv povrchu Země (např. tráva, voda nebo les).

Postup klasifikace LC ve zmíněném projektu byl z obecného hlediska již zpracován [3]. Klasifikace byla provedena pro různou tematickou podrobnost (až 28 tříd) ve třech odlišných evropských lokalitách. Využita byla data družice Sentinel-2 a produkt CORINE Land Cover (CLC) jako referenční data. Práce představila postupy pro klasifikaci se zaměřením na metodu maximum likelihood, zhodnotila její úspěšnost a identifikovala některá omezení.

V návaznosti i tato práce používá družicová data Sentinel-2 a trénovací plochy z produktu CLC. Pozornost se zde soustředí na vybrané třídy, které se v provedených klasifikacích ukázaly jako problematické. Dle nomenklatury CLC se jedná o třídy trvale zavlažovaná orná půda, pastviny a přírodní traviny. Pro práci byly zvoleny tři oblasti, kde se všechny vybrané třídy nachází. Na základě odlišnosti těchto lokalit bude možné posoudit univerzální využití zvoleného postupu. Jde o oblasti v Makedonii, Španělsku a Turecku.

Problém klasifikace vybraných tříd spočívá především v jejich spektrální podobnosti. Jedním z cílů je proto identifikovat vhodné charakteristiky při tvorbě příznakového prostoru. Práce má dále za cíl určit vhodný výběr trénovacích a testovacích ploch. Tento krok obnáší navržení úprav podkladových dat CORINE a jejich rozdělení do trénovacího a testovacího setu. Na základě připravených dat provést

klasifikaci a vyhodnocení přesnosti. A v neposlední řadě jde o zhodnocení dosažených výsledků v závislosti na příznacích, metodě a lokalitách. Zároveň posoudit možnosti a případné limity klasifikace těchto tříd s využitím metody CORINE.

METODIKA

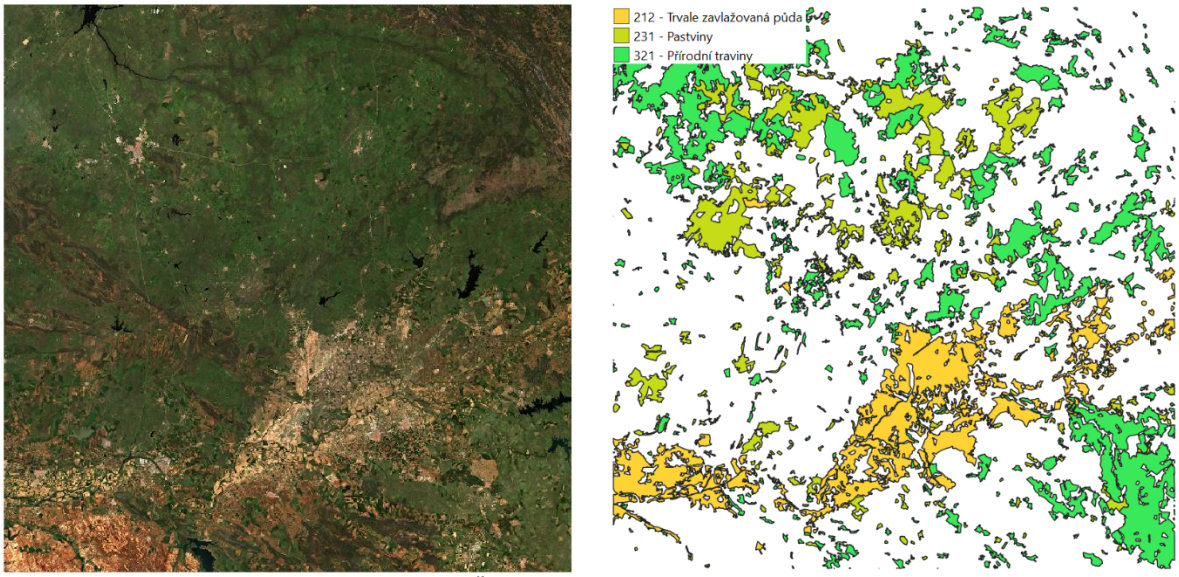
Zájmové území

Pro srovnání klasifikace v různých částech Evropy byly vybrány tři odlišné lokality. Pro výběr bylo podstatné, aby se všechny sledované třídy nacházely uvnitř jedné dlaždice, dále aby rozloha třídy uvnitř dlaždice měla přijatelné zastoupení a v neposlední řadě geografická rozmanitost vybraných dlaždic.

Zatímco pastviny a přírodní traviny se dle CORINE vyskytují téměř po celé Evropě, trvale zavlažovaná orná půda je dominantou Středomoří. Dle vizuálního posouzení jsou na kombinaci těchto tříd jednoznačně nejbohatší Turecko a Španělsko. V rámci zachování rozmanitosti bylo nutné najít třetí lokalitu. V tomto případě byly rozhodující jak geografická poloha, tak rozloha tříd v dané oblasti. Přijatelné zastoupení všech tříd bylo nalezeno v oblasti v Severní Makedonii.

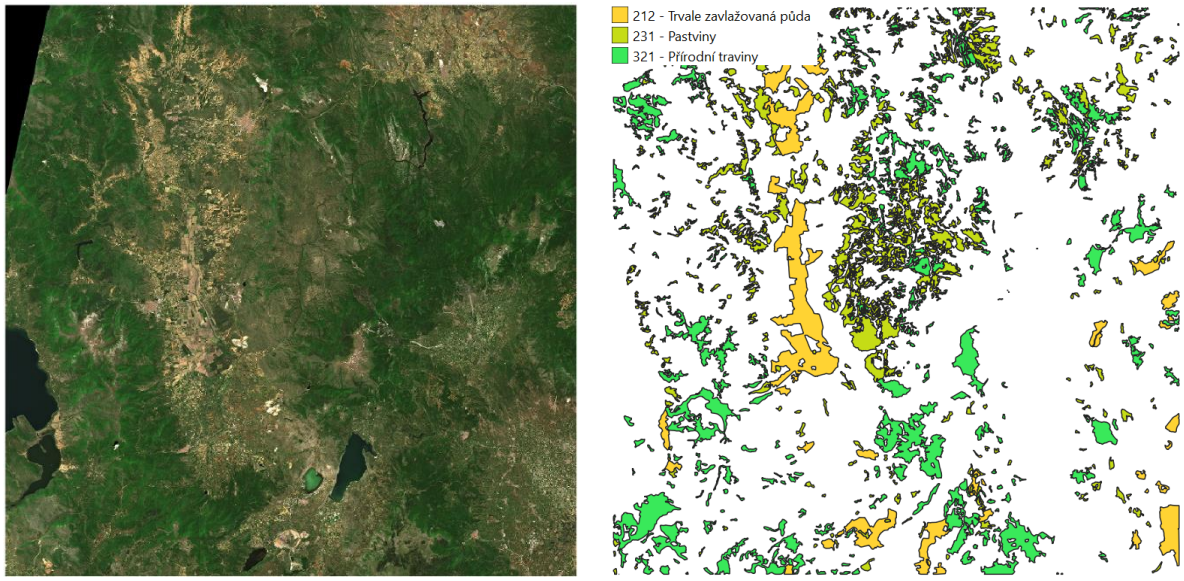
Obr. 1 Zájmová území: poloha scén Sentinel-2 a jejich označení

Scéna ze Španělska zachycuje zemědělsky využívanou oblast s minimem zástavby a vodních ploch. Nadmořská výška se zde pohybuje od 200 do 1 500 m n. m. Sledované třídy jsou rozptýleny po celé oblasti, s výjimkou orné půdy, která se rozkládá spíše v jižní nížinaté části. Jde o suchou oblast se středomořským podnebím, léta jsou velmi suchá, zimy mírné a vlhké [4].



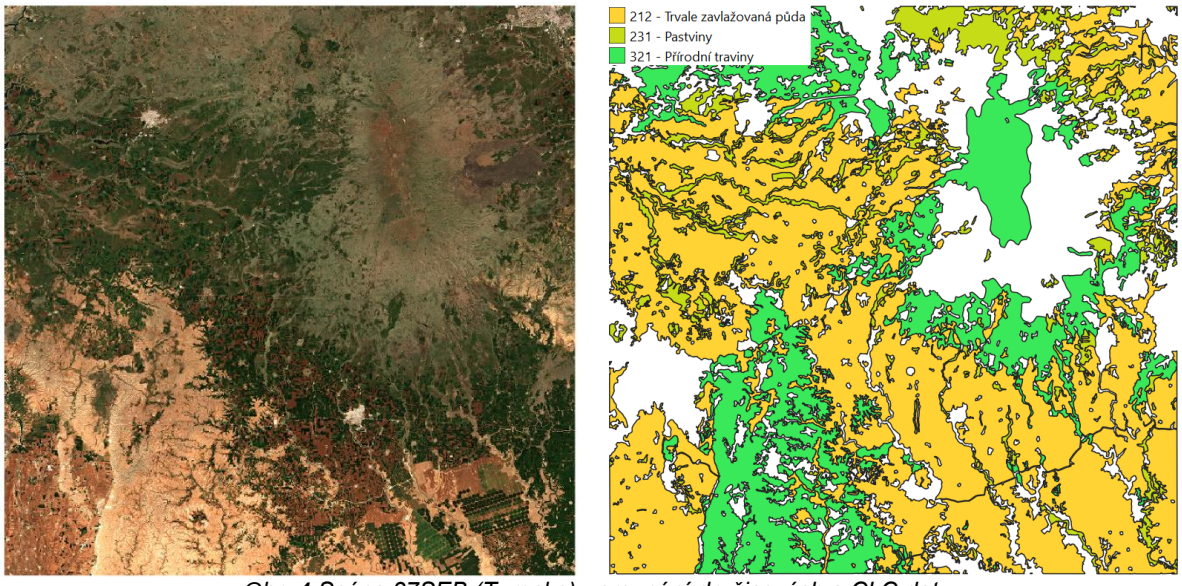
Obr. 2 Scéna 29SQD (Španělsko) - srovnání družicových a CLC dat

V Severní Makedonii byla vybrána scéna zobrazující jižní část země na hranici s Řeckem. Z vybraných oblastí jsou zde nejvýraznější přechody mezi nadmořskou výškou, která se rozpíná od 0 do 2 500 m n. m. Významná část oblasti se nachází ve výškách nad 1 500 m n. m., kde se nachází i nejpočetnější LC této oblasti - lesy a polopřírodní oblasti. Druhou nejpočetnější třídou jsou zemědělské oblasti. Klima je zde kontinentální, s mírnějšími letními teplotami a studenými zimami [4].



Obr. 3 Scéna 34TEL (Makedonie) - srovnání družicových a CLC dat

Zachycená oblast v Turecku se nachází na jihovýchodě státu. Nadmořská výška od jihu (400 m n. m.) stoupá až ke 2 000 m n. m. Sledované třídy zde zabírají největší část oblasti. Opět dominuje zemědělství, ve vyšších nadmořských výškách se nacházejí přírodní traviny. Pastviny se zde vyskytují především v severní části a v okolí řek. Obdobně jako ve Španělsku i tato oblast je velmi suchá [4].



Obr. 4 Scéna 37SEB (Turecko) - srovnání družicových a CLC dat

Použitá data

CORINE

CORINE Land Cover (CLC) představuje seznam 44 tříd, do kterých je LC v Evropě klasifikován. CLC je produkt vytvořený v rámci programu Copernicus. Vzniká jako jeden z produktů monitorování LU/LC a zapojeny jsou všechny státy EEA39 [5]. Obsah databáze CLC je vytvářen na úrovni jednotlivých zemí. Nejmenší mapovací jednotka (MMU) je 25 ha u plošných objektů a 100 m u liniových objektů. V důsledku toho mohou být objekty, jejichž rozloha je menší než stanovený limit, generalizované a zahrnuté do jiné třídy. Je proto nutné mít na paměti, že přesnost tohoto modelu je ≥ 85 % [5].

Třídy, které jsou sledovány v této práci, pochází z nejpodrobnější - třetí - úrovně systému CLC. Dle CLC směrnice [6] jsou definovány takto:

- 2.1.2 Trvale zavlažovaná orná půda: „Obdělávané a zemědělsky využívané parcely pro ornou půdu, které jsou trvale nebo periodicky zavlažované s použitím trvalé infrastruktury (zavlažovací kanály, drenážní síť a přídavné zavlažovací zařízení). Většinu těchto plodin nelze pěstovat bez umělé dodávky vody. Nezahrnuje sporadicky zavlažovanou půdu.“
- 2.3.1 Pastviny, louky a ostatní trvalé travní porosty se zemědělským využitím: „Oblasti jsou soustavně využívány (minimálně po dobu 5 let) pro produkci krmiva. To zahrnuje přírodní nebo oseté byliny, nekultivované nebo lehce kultivované louky a spásané nebo mechanicky sklízené louky. (...) Pastviny lze popsat jako extenzivně nebo intenzivně obdělávané trvalé traviny, kde se nachází prvky zemědělské infrastruktury, jako například ploty, přístřešky, napajedla, a probíhají zde tyto procesy: spásání, zavlažování, osev a hnojení. Typickým znakem je pravidelný tvar pozemku a/nebo vyšlapané cestičky od zvířat.“
- 3.2.1 Přírodní travní porosty: „Travní porosty s žádným nebo mírným zásahem člověka. Traviny s nízkou produktivitou. Často se nachází v drsném, nerovném terénu, ve strmých svazích; nezřídka zahrnují i skalnaté oblasti nebo shluky jiné vegetace. (...) Typickou charakteristikou této třídy je velká rozloha, nepravidelný tvar, zpravidla se nacházejí ve větší vzdálenosti od lidských obydlí.“

Sentinel-2

Použití družicových dat pro mapování má oproti pozemním metodám mnoho výhod, např. častá aktualizace, v mnoha případech téměř okamžitá dostupnost dat a především schopnost zachytit velké území ve velmi krátkém čase. Výběr družice pro konkrétní úlohu se může odvíjet od dostupnosti dat pro řešené období nebo požadovaného rozlišení (prostorové, spektrální, časové). Pro účely této práce byla vybrána data Sentinel-2, která jsou bezplatně poskytována skrz Copernicus Open Access Hub [7].

Sentinel-2 je mise programu Copernicus, která je technicky zajišťována Evropskou kosmickou agenturou (ESA). Tvoří ji 2 družice (Sentinel-2A a Sentinel-2B) vypuštěné v roce 2015, resp. 2017. Obíhají po stejné slunečně synchronní dráze a pohybují se v průměrné výšce 786 km. Časové rozlišení jedné družice je 10 dní, při zvážení vzájemného odfázování družic o 180 stupňů je to pouze 5 dní [8], [9]. Spektrální data jsou na palubě každé družice snímána pomocí senzoru, tzv. MSI (MultiSpectral Instrument). Ten snímá informace ve 12 pásmech elektromagnetického záření od viditelného až po střední infračervené (SWIR). Pásma mají různou plošnou rozlišovací schopnost v rozsahu 10 až 60 m. Šířka záběru senzoru je 290 km. MSI používá tzv. push-broom skener, což umožňuje zaznamenat informace v jeden moment po celé šířce záběru [8].

Data Sentinel-2 jsou zpracována v několika úrovních (L0, L1A, L1B, L1C a L2A). V tomto případě byla použita data L2A, která obsahují radiometrické i atmosférické korekce a jsou ortorektifikována. V práci bylo použito 9 optických pásem Sentinel-2 s rozlišením 20 m (B2, B3, B4, B5, B6, B7, B8a, B11 a B12).

Zvýraznění obrazu

Zvýraznění obrazu přispívá ke zlepšení vizuální interpretace dat, a tedy i potenciálnímu zvýšení úspěšnosti klasifikace. V tomto případě byly použity kombinace pásem pro výpočet indexu NDVI a metody hlavních komponent. Z metod lokálního zvýraznění obrazu byly použity Haralickovy funkce pro vytvoření texturálních měr.

Normovaný rozdílový vegetační index (normalized difference vegetation index, NDVI) je někdy zvaný také „index zelenosti“. NDVI je založen na rozdílné odrazivosti vegetace v červeném (R) a blízkém infračerveném (NIR) pásmu [10] a je určen vztahem

$$NDVI = \frac{NIR - R}{NIR + R}.$$

Výsledek nabývá hodnot od -1 do 1 a indikuje množství vegetace v rámci jednoho pixelu. I pro snadnou interpretaci je jedním z nejčastěji používaných indexů. [10]

Použití multispektrálních dat s sebou nese riziko korelace mezi jednotlivými pásmy. Ta může být eliminována použitím metody hlavních komponent (Principal Component Analysis, PCA). Jedná se o transformaci, která zavádí nově orientované osy a nový počátek. Nová hlavní osa vede ve směru, kde je rozptyl hodnot ze všech použitých pásem největší. Nový počátek je určen průměrem a druhá osa jím vede kolmo na osu první. Vzájemná kolmost os zajišťuje, že korelace mezi daty je odstraněna [11]. Kanály PCA byly vytvořeny na základě 9 zmíněných pásem Sentinel-2. Do klasifikace byly použity první 3 komponenty (PCA1, PCA2, PCA3), ve kterých byly spektrální rozdílnosti nejpatrnější.

Texturální míry patří k metodám lokálního zvýraznění obrazu [11]. Častou metodou určení textury jsou tzv. Haralickovy funkce [12]. Matice (gray level co-occurence matrix, GLCM) určuje, kolikrát se ve sledované oblasti vyskytuje dvojice pixelů o dané hodnotě a v dané vzdálenosti. Na základě této matice jsou vypočteny konkrétní metriky. Textury mohou být využity v případech, kdy jsou spektrální rozdíly sledovaných tříd malé [13]. Na základě studované literatury [14], [15] byly vytvořeny míry, jejichž použití se opakuje: korelace, homogenita, entropie a druhý úhlový moment (angular second moment, ASM). Jako podklad pro vytvoření textur bylo použito pásmo B2.

EU-DEM

Výsledky a přesnost klasifikace lze podpořit přidáním dalších informací, které dodají širší kontext, např. v podobě topografické informace. V tomto případě byl použit digitální model povrchu, který představuje model Země i s objekty, které se na ní nacházejí, vč. vegetace, zástavby ad. [16]. Produkt EU-DEM je digitální model povrchu z roku 2016 a je dostupný pod hlavičkou programu Copernicus [17]. Jedná se o hybridní model vytvořený na základě vážených průměrů dřívějších výškových modelu misí SRTM a ASTER GDEM. Použita byla verze EU-DEM v 1.1. Model je poskytován ve formátu GeoTIFF po dlaždicích 1000 x 1000 km s rozlišením 25 m a směrodatnou odchylkou výškových dat 7 m. [18] Na základě topografického rastru byl dále vytvořen rastr sklonitosti, který vyjadřuje sklon terénu v daném místě (pixelu) ve stupních.

Multitemporální data

Klasifikace na základě jedné scény může být velkou výzvou [19], obzvlášť pokud je obtížné některé typy LC odlišit [20]. V takových případech je vhodné použití multitemporálních dat. Jejich výhoda tkví v možnosti klasifikovat daný LC na základě odlišností v různých obdobích [20]. Vegetační aktivita se v závislosti na klimatických podmínkách a přístupu k zemědělství může lišit napříč lokalitami, především částí roku, kdy probíhá. V Turecku a Španělsku byla aktivita zjevná již v březnu. Od května přirozená vegetace postupně usychala v závislosti na stoupající teplotě. V Makedonii svou roli sehrála i nadmořská výška, jelikož zde byla část oblasti do konce dubna pokryta sněhem. Výběr dat omezovala i často přítomná oblačnost. Družicová data pro Španělsko a Turecko byla vybrána z roku 2018, jelikož k tomuto roku byla vydána použitá verze CLC. V Makedonii byly scény z roku 2018 velmi oblačné, proto byla použita data z roku 2019.

Tab. 1 Multitemporální data - vybrané scény

Španělsko (2018)	Makedonie (2019)	Turecko (2018)
28.3.	8.6.	19.3.
17.4.	3.7.	23.4.
17.5.	7.8.	23.5.
16.6.	16.9.	7.6.
16.7.	16.10.	12.7.

Pro vstup do klasifikace bylo použito množství příznaků shrnutých v Tab. 2. Jedná se o 9 optických pásem (OPT) a vegetační indexy jednotlivých scén (NDVI). Dále byla použita 3 pásma hlavních komponent (PCA), která v některých kombinacích nahradila optická data. Zařazena byla i texturální (TEX) a topografická data (TOPO). Do klasifikace byly tyto skupiny zaváděny v různých kombinacích.

Tab. 2 Použité příznaky

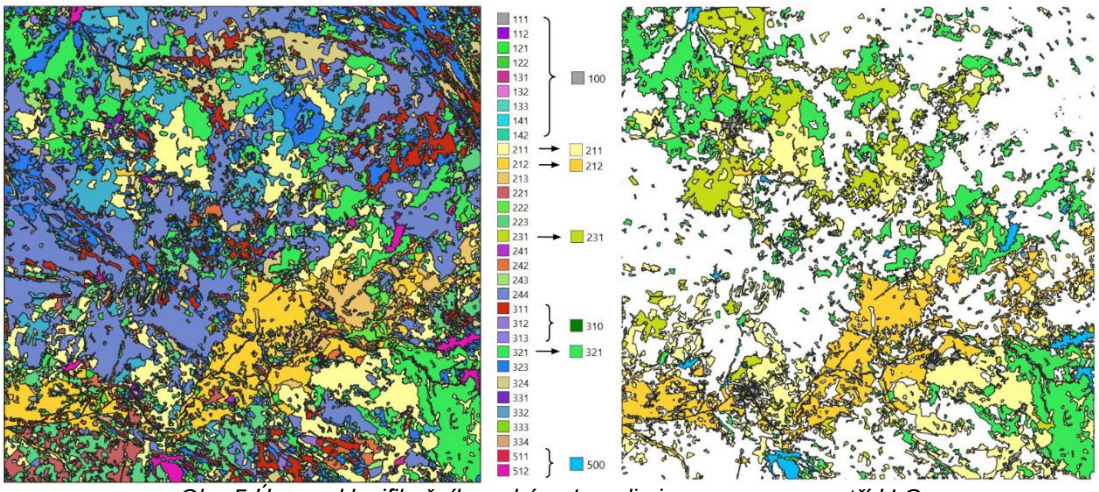
kategorie	jednotlivé příznaky	celkový počet příznaků (5 scén)
optická (OPT)	B2, B3, B4, B5, B6, B7, B8a, B11, B12	45
NDVI	NDVI	5
PCA	PCA1, PCA2, PCA3	15
texturální (TEX)	ASM, korelace, homogenita, entropie	20
topografická (TOPO)	DEM, sklon	2

Klasifikace

Klasifikace je automatický proces, během kterého jsou vstupní data převedena do tematické mapy [21]. Samotná klasifikace je součástí komplexního procesu. Co se týče řízené klasifikace, prvními kroky jsou předzpracování a výběr trénovacích ploch. Následuje výběr vhodných příznaků, samotná klasifikace a vyhodnocení výsledků. Výsledky klasifikace nezávisí pouze na použité metodě. Do klasifikace vstupuje celá rada faktorů, které se ve výsledcích projeví. S ohledem na dosažení uspokojivých výsledků lze celý proces opakovat se zaměřením na jednotlivé činitele, kterými jsou trénovací plochy, vstupní data a použitá metoda a její parametry.

Předzpracování a výběr trénovacích ploch

Před výběrem trénovacích ploch bylo nutné vybrat třídy, pro které bude klasifikace probíhat. V každé oblasti se vyskytuje přes 20 tříd ze třetí úrovně CLC nomenklatury. Hlavním cílem úpravy bylo vybrat třídy tak, aby počet ostatních tříd výrazně nepřevažoval nad počtem sledovaných tříd. Některé třídy byly na základě vzájemné podobnosti LC agregovány, jiné byly vyřazeny. Úpravu klasifikačního schématu demonstruje Obr. 5. Do třídy s označením 100 byly zahrnuty všechny třídy třetí úrovně CLC, které reprezentují zástavbu. Obdobně vznikly i třídy 310 a 500, které reprezentují lesní porost a vodní plochy. Jedná se o typy LC, které jsou v oblasti zachovány, ale pouze ve zobecněné podobě. Nové klasifikační schéma (Tab. 3) pochopitelně obsahuje sledované třídy (212, 231, 321). Pro srovnání byla zahrnuta i třída 211 - nezavlažovaná orná půda. Ostatní třídy byly ponechány stranou.



Obr. 5 Úprava klasifikačního schématu - eliminace a agregace tříd LC

Tab. 3 Klasifikační schéma - všechny lokality

Španělsko	Makedonie	Turecko	popis
100	100	100	zástavba
211	211	211	nezavlažovaná orná půda
212	212	212	trvale zavlažovaná orná půda
231	231	231	pastviny
310	310	-	lesy
321	321	321	přírodní travní porost
500	500	500	vodní plochy

Trénovací plochy (jinak také referenční data) nesou informaci o rozdělení sledované oblasti do jednotlivých tříd. Apriorní informace byla v tomto případě obsažena v datech CORINE. Hranice mezi jednotlivými typy LC jsou v tomto produktu generalizované a nemusí věrně vystihovat realitu. V důsledku toho je možné, že mezi pixely na této hranici dojde k záměně. Na polygony byl aplikován vnitřní buffer (40 m, odpovídá 2 pixely při rozlišení 20 m), aby mezi jednotlivými LC vznikla mezera a zamezilo se tím případné nesprávné klasifikaci. Případná oblačnost byla ze scény odstraněna pomocí vrstvy SCL (Scene classification), která je součástí dat Sentinel-2. Na závěr byla z trénovacích ploch vytvořena bodová vrstva. Rozestup bodů byl zvolen 500 m. Tato generalizovaná bodová vrstva byla použita pro další zpracování při klasifikaci.

V průběhu klasifikace byly v trénovacích datech zjištěny některé nedostatky, které zde budou popsány. Na základě vizuálního posouzení v QGIS byly zřetelné různé typy záměn (vegetace v zástavbě, vodní plochy v zástavbě, vegetace ve vodní plochách, zástavba ve vegetaci). V jednotlivých třídách CLC se mohou objevit jiné typy LC, které do nich správně nepatří, ale jejichž velikost je menší než stanovená MMU při produkci CLC (např. silnice užší než 100 metrů nebo plochy menší než 25 ha). Právě tyto případy byly nejčastější příčinou výše zmíněných záměn. Zjištěné nedostatky byly v referenčních datech opraveny na základě hodnoty NDVI. Z daného typu LC byly vyřazeny všechny body, jejichž hodnoty NDVI do daného intervalu nespadaly. Mezní hodnoty NDVI (Tab. 4 Orientační rozmezí hodnot NDVI pro různé typy povrchuTab. 4) byly zvoleny kompromisem mezi studovanou literaturou [22], [23] a pozorovanými hodnotami v dané oblasti.

Tab. 4 Orientační rozmezí hodnot NDVI pro různé typy povrchu

hodnota NDVI	typ povrchu
< 0	voda
0 - 0,2	holá půda, zástavba
0,2 - 0,5	řídká až středně hustá vegetace
$> 0,5$	velmi hustá vegetace

Další provedenou úpravou trénovacích dat bylo vyřazení odlehlých měření. Hodnoty těchto bodů nebyly pro sledovaný jev reprezentativní, a tak byly odstraněny. Pro stanovení odlehlých měření byly použity následující vztahy:

$$odlehlé\ měření < 1QR\ - 1{,}5 * IQR$$

$$odlehlé\ měření > 3QR\ + 1{,}5 * IQR$$

kde 1QR je první kvartil, 3QR je třetí kvartil a jejich rozdíl tvoří mezikvartilové rozpětí (IQR). Tento proces byl aplikován pro každý příznak, ve kterém byly postupně upraveny body jednotlivých tříd.

V posledním kroku se úprava referenčních dat týkala počtu vstupních bodů. Ten dosud nebyl nijak regulován a každá třída byla v klasifikaci zastoupena jiným počtem vzorků. V tomto kroku byl z každé třídy vybrán stejný počet bodů, tak aby byl jejich počet vyrovnán.

Random Forest a výběr příznaků

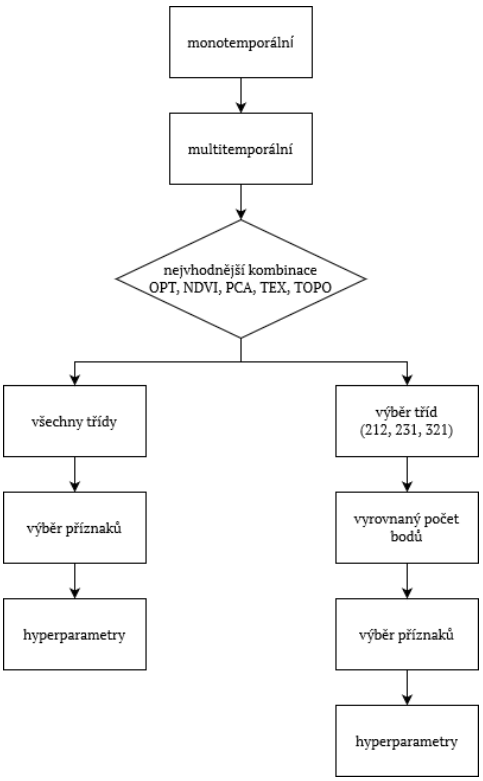
Náhodný les (dále Random Forest, RF) je metoda strojového učení používaná pro vysokou přesnost, rychlost a nízké riziko přetrénování. RF se skládá z určitého počtu rozhodovacích stromů (obr. 7.6). Ze všech příznaků, které jsou k dispozici na vstupu, používá každý strom pouze náhodný a nezávislý výběr (subset) těchto veličin. Klasifikace následně probíhá nezávisle v každém stromu. Výsledná hodnota je modus, tedy nejčastěji zvolená hodnota napříč všemi stromy. Každý strom rozhoduje nezávisle, čímž se snižuje korelace mezi výsledky. [24], [25]

V DPZ byl použit v mnoha úlohách a byl úspěšně aplikován i na klasifikace LC [25], [26]. RF poskytuje velmi dobré výsledky pro klasifikace, které používají data z různých zdrojů, tedy nejen optická, ale i jejich deriváty, výšková data ad. [14], [25]. Umí si poradit i s problémy, které obsahují velké množství (tisíce) příznaků [24]. Navíc, v množství příznaků, které vstupují do klasifikace, dokáže tento algoritmus samostatně vyhodnotit jejich důležitost. Na základě významnosti (feature importance, FI) lze eliminovat méně přínosné příznaky nebo jim přiřadit odpovídající váhu a výsledek klasifikace dále zlepšit [25].

Ladění **hyperparametrů** bývá jednou z finálních úprav klasifikace metodou RF. Jedná se o proces, při kterém se určují takové hodnoty parametrů, které zvýší přesnost. Běžně jsou laděny tyto parametry:

- n_estimators - počet rozhodovacích stromů (vyšší počet může zvýšit časovou náročnost),
- max_depth - maximální výška stromu; s vetší výškou se zvyšuje přesnost, ale od určité hodnoty může nastat přetrénování,
- min_samples_split - minimální počet bodů v uzlu (node) než se rozroste o další uzly,
- min_samples_leaf - minimální počet bodů, které musí být v uzlu po rozdělení (split). [27]

Klasifikace byla nejprve prováděna pro určené klasifikační schéma (Tab. 3). V druhé fázi klasifikace byly ostatní třídy ponechány stranou a do klasifikace byly zařazeny pouze sledované třídy (212, 231, 321, označení varianty – výběr). Cílem výběru bylo zjistit výsledky klasifikace, pokud budou odstraněny zbylé potenciálně rušivé třídy, a získat představu o odlišitelnosti pouze mezi vybranými typy LC. V tomto případě byl pro další klasifikaci počet bodů na vstupu upraven na stejný počet pro všechny tři třídy (označení - vyrov.). Pro obě varianty (všechny třídy i výběr tříd) byly určeny důležité příznaky z dané kombinace a hyperparametry. Popsaný postup shrnuje Obr. 6.



Obr. 6 Pracovní postup klasifikace

Vyhodnocení

Při hodnocení klasifikátoru se používají tzv. testovací data, která se liší od trénovacích. Jedná se o porovnání výsledků klasifikace se skutečností. Při tomto posouzení je nutné mít na paměti, že „kvalita jakéhokoliv odhadu přesnosti je pouze tak dobrá, jak je dobrá informace o skutečném stavu“ [11].

K vyhodnocení modelu lze použít metodu křížové validace. Tato metoda umožňuje zhodnotit, jak dobře bude model pracovat na nezávislém datasetu (testovacím vzorku), a odhalit případné přetrénování. Použita byla metoda tzv. k-fold validace, při které jsou trénovací data rozdělena na k podmnožin (folds), jedna z nich je ponechána stranou a na zbylých podmnožinách proběhne trénink klasifikátoru. Výsledek je zhodnocen pomocí nepoužité části dat. Tento proces se opakuje podle stanoveného počtu k. [28]

Jeden ze způsobů vyhodnocení výsledků je tzv. chybová matice. S její pomocí lze zhodnotit výsledky jak jednotlivých tříd, tak celkové klasifikace. Na jedné straně jsou umístěna referenční data, tedy skutečné hodnoty pixelu, proti nim stojí výsledky klasifikace, tedy pixely zařazené pomocí klasifikátoru [21]. V příslušném směru lze zhodnotit záměnu u referenčních a klasifikovaných bodů.

Informaci o výsledku klasifikace jednotlivých tříd podávají metriky precision a recall. Precision (spolehlivost) je poměr správně klasifikovaných pixelů dané třídy ku všem pixelům, které byly do této třídy klasifikovány. Značí, nakolik klasifikované pixely odpovídají skutečnému stavu. Recall, jinak senzitivita, je poměr správně klasifikovaných pixelů dané třídy a celkového počtu pixelů, které do této třídy patří ve skutečnosti. Říká, nakolik jsou pixely z dané třídy rozpoznané při klasifikaci [21], [29]. Tyto metriky se chovají jako spojené nádoby a v úlohách DPZ je žádoucí mezi nimi dosáhnout vyrovnaných hodnot. Soulad těchto metrik shrnuje ukazatel F1. Jedná se o vážený harmonický průměr precision a recall. Udává, nakolik je model přesný, stejně jako nakolik je robustní. Nabývá hodnot 0 až 1, čím je hodnota vyšší, tím je model lepší [29].

VÝSLEDKY

Proces klasifikace proběhl opakovaně za účelem optimalizace výsledků. Pro srovnání byla klasifikace provedena nejprve bez úprav podkladových dat a s užitím monotemporálních dat (data ze začátku vegetační aktivity). Následně byla upravena trénovací data a do klasifikace byla zařazena data ze všech

5 scén. Byly zahrnuty příznaky v různých kombinacích (Tab. 2), postupně byl upraven i počet typů LC a počet bodů pro jednotlivé třídy. Výsledky jsou prezentovány po jednotlivých lokalitách.

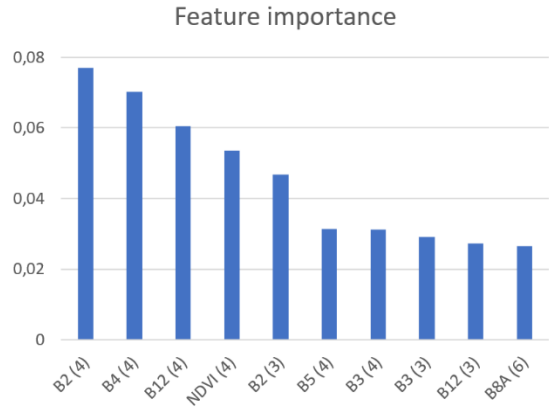
Turecko

Výsledky klasifikace s nejvyššími hodnotami F1 znázorňuje Tab. 5. V této klasifikaci byla použita kombinace příznaků OPT + NDVI + TOPO s vyrovnaným počtem bodů pro všechny tři typy LC a s určením hyperparametrů. Levá část tabulky je chybová matice, která zobrazuje počet bodů a jak byly body dané třídy klasifikovány. Pravá část výsledky shrnuje pomocí ukazatelů recall, precision a F1. Je patrné, že úplnou záměnu mezi pastvinami a přírodními travinami se eliminovat nepodařilo. Nicméně, v porovnání s průběžnými výsledky lze vyrovnání vstupního počtu bodů hodnotit jako velmi úspěšný krok.

Tab. 5 Turecko: výsledky klasifikace - vybrané třídy (OPT + NDVI + TOPO)

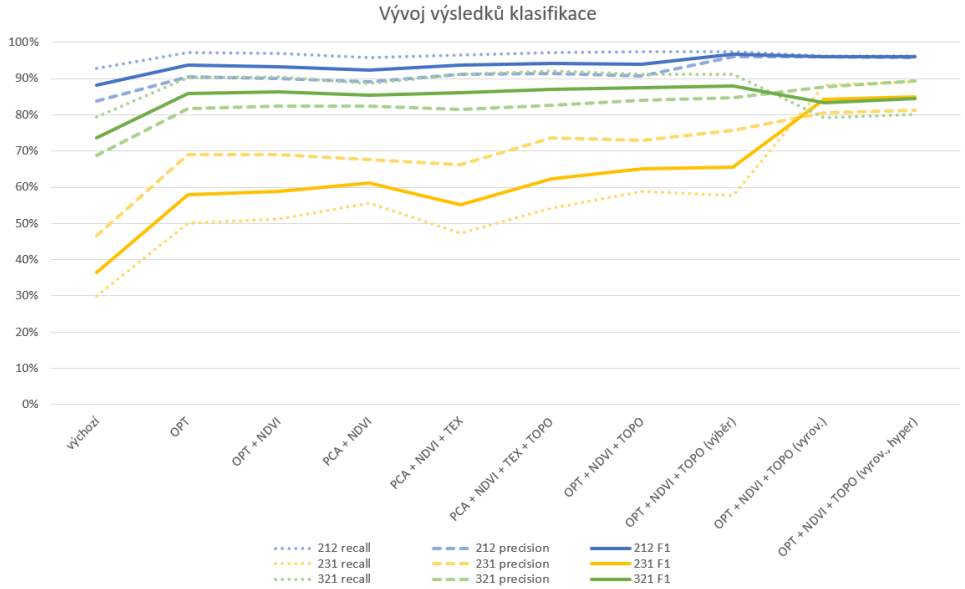
		klasifikace						
		212	231	321	celkem	recall [%]	precision [%]	F1 [%]
reference	212	702	20	7	729	96,30	95,77	96,03
	231	16	651	63	730	89,18	81,17	84,99
	321	15	131	584	730	80,00	89,30	84,39

Na sledované typy LC měly největší vliv spektrální příznaky, jmenovitě B2 (490 nm), B4 (665 nm) a B12 (2190 nm). Významně přispívají příznaky z dubna a března (Obr. 7).



Obr. 7 Turecko: srovnání důležitých příznaků (vybrané třídy OPT + NDVI + TOPO)

Vývoj výsledků sledovaných typů LC zachycuje Obr. 8. Na vodorovné ose jsou vyneseny použité kombinace počínaje první klasifikací s monotemporálními daty (výchozí), přes použití různých kombinací multitemporálních dat (OPT, OPT + NDVI atd.), konče výsledky tří klasifikací, které byly provedeny pouze pro vybrané třídy. Na svislé ose jsou vyneseny hodnoty v procentech. Ústřední zobrazenou veličinou je F1, kolem které vytyčují pás hodnoty recall a precision.



Obr. 8 Turecko: vývoj výsledků klasifikace - vybrané třídy

Výsledky pro zavlažovanou ornou pudu i přírodní traviny v celém průběhu vykazovaly velmi dobré výsledky přes 80 %. Výsledky pro pastviny výrazně znázorňují význam jednotlivých kroků pro klasifikaci této třídy. Nejvýraznější nárůst lze pozorovat při přidání multitemporálních dat (OPT) a u vyrovnání počtu bodů (vyrov.). V momentě, kdy se počet bodů vyrovnal, výsledky této třídy se zlepšily o 20 %. V porovnání s výchozí pozicí se výsledky této třídy se zlepšily téměř o 50 %.

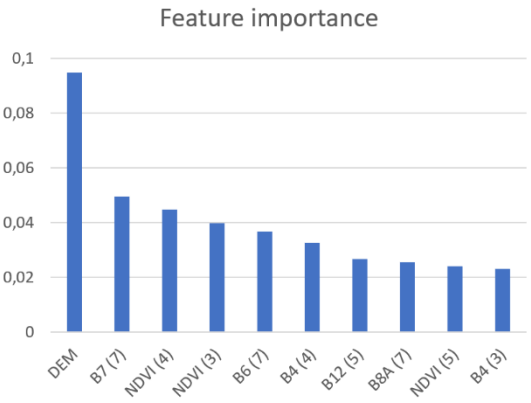
Španělsko

Výsledky klasifikace vybraných tříd s vyrovnaným počtem bodů a hyperparametry zobrazuje Tab. 6. Nejlepší výsledky stejně jako v předchozí lokalitě vykázala zavlažovaná orná půda (precision 99 %, recall 98,25 %). Výrazné zlepšení výsledku bylo pozorováno u pastvin, kde hodnota F1 narostla o téměř 30 %. Ve výsledcích travních porostů došlo především k nárůstu hodnoty precision o více než 20 %.

Tab. 6 Španělsko: výsledky klasifikace - vybrané třídy (OPT + NDVI + TEX + TOPO)

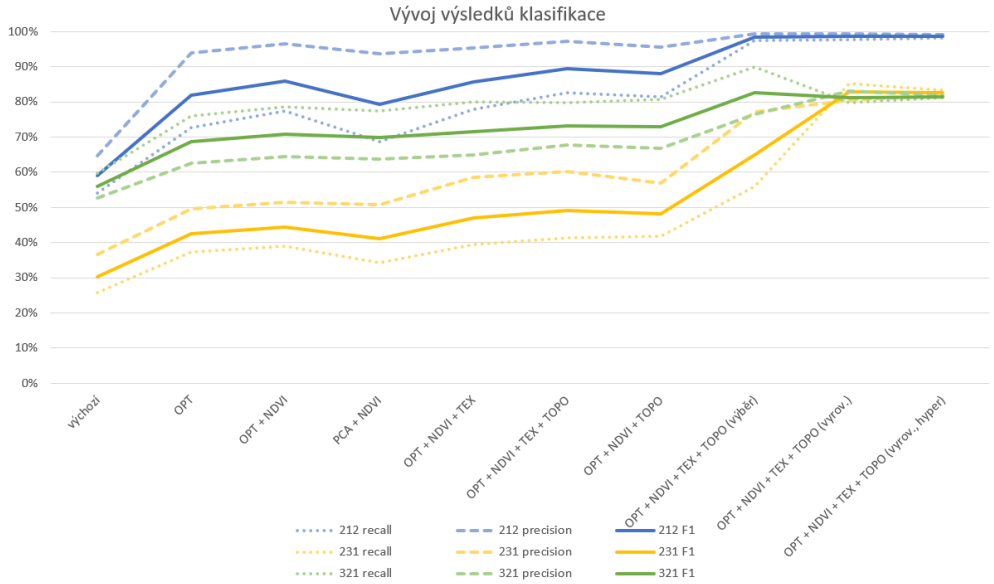
		klasifikace						
		212	231	321	celkem	recall [%]	precision [%]	F1 [%]
reference	212	392	1	6	399	98,25	99,24	98,74
	231	0	333	66	399	83,46	82,02	82,73
	321	3	72	324	399	81,20	81,82	81,51

Při klasifikaci vybraných tříd jednoznačně vynikl vliv DEM (Obr. 9). Následující příznaky byly různorodější, opakovaly se pouze NDVI (březen, duben, květen) a B4 (březen, duben). Ani v jedné variantě se mezi 10 nejvýznamnějšími neobjevil žádný texturální příznak.



Obr. 9 Španělsko: srovnání důležitých příznaků (vybrané třídy OPT + NDVI + TEX + TOPO)

Vývoj výsledků zachycuje Obr. 10. Výrazný nárůst je patrný především při zařazení multitemporálních dat, dále při klasifikaci pouze vybraných tříd (výběr) a při vyrovnání vstupního počtu bodů. Pás kolem F1 měl u všech tříd rozpětí vetší než 10 %, až v posledních klasifikacích se zúžil na hodnotu kolem 1 %. Po určení hyperparametrů se rozdíl mezi hodnotami recall a precision zmenšil a vyrovnal.



Obr. 10 Španělsko: vývoj výsledků klasifikace - vybrané třídy

V průběhu optimalizace se hodnoty recall a precision výrazně zvýšily a vyrovnaly pro všechny typy LC, což je známkou dobře vytrénovaného modelu. Co se týče záměny mezi pastvinami a přírodními travinami, pohybovala se v tomto případe mezi 16 a 18 %. Záměnu se podařilo snížit především u pastvin. To muže být jak v důsledku vyrovnání počtu bodů, tak odebráním třídy nezavlažovaná orná půda, za kterou byly body obou tříd často zaměňovány.

Pro zavlažovanou pudu a pastviny se objevil znatelný pokles při zařazení příznaků PCA. Jejich použití se pro tuto lokalitu neosvědčilo. Co se týče zavlažované orné pudy, po zařazení multitemporálních dat se hodnoty pohybovaly nad 80 %, což je dobrý výsledek, nicméně v závěru hodnoty vyrostly až k 99 %. Při porovnání počátečních a koncových hodnot třídy pastviny, zlepšily se výsledky o více než 50 %, což je zdaleka největší nárůst. Oproti ostatním třídám se u této třídy projevilo výrazné zlepšení i po vyrovnání vstupního počtu bodů. To může být v důsledku toho, že tato třída byla v klasifikaci zastoupena nejmenším počtem bodů.

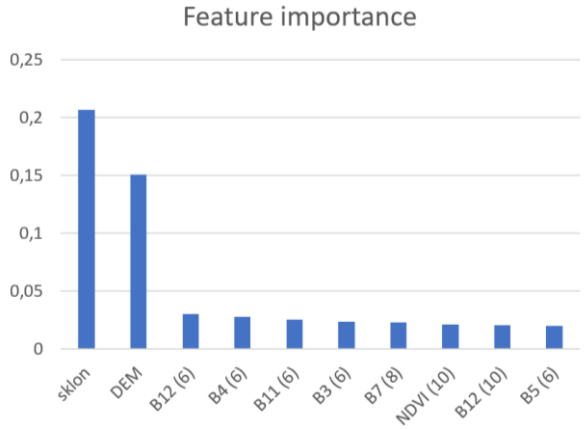
Severní Makedonie

Tab. 7 ukazuje výsledek klasifikace vybraných tříd pro vyrovnaný počet vstupních bodů a s použitím hyperparametrů. Hodnoty precision a recall se vyrovnaly pouze u zavlažované orné pudy. Hodnota F1 narostla téměř o 30 % a tradičně se výrazně přiblížila 100 %. Hodnoty zbylých tříd zůstaly pod hranicí 80 %. Ani v tomto případě záměna mezi třídami pastviny a přírodní traviny nebyla zcela odstraněna.

Tab. 7 Makedonie: výsledky klasifikace - vybrané třídy (OPT + NDVI + TOPO)

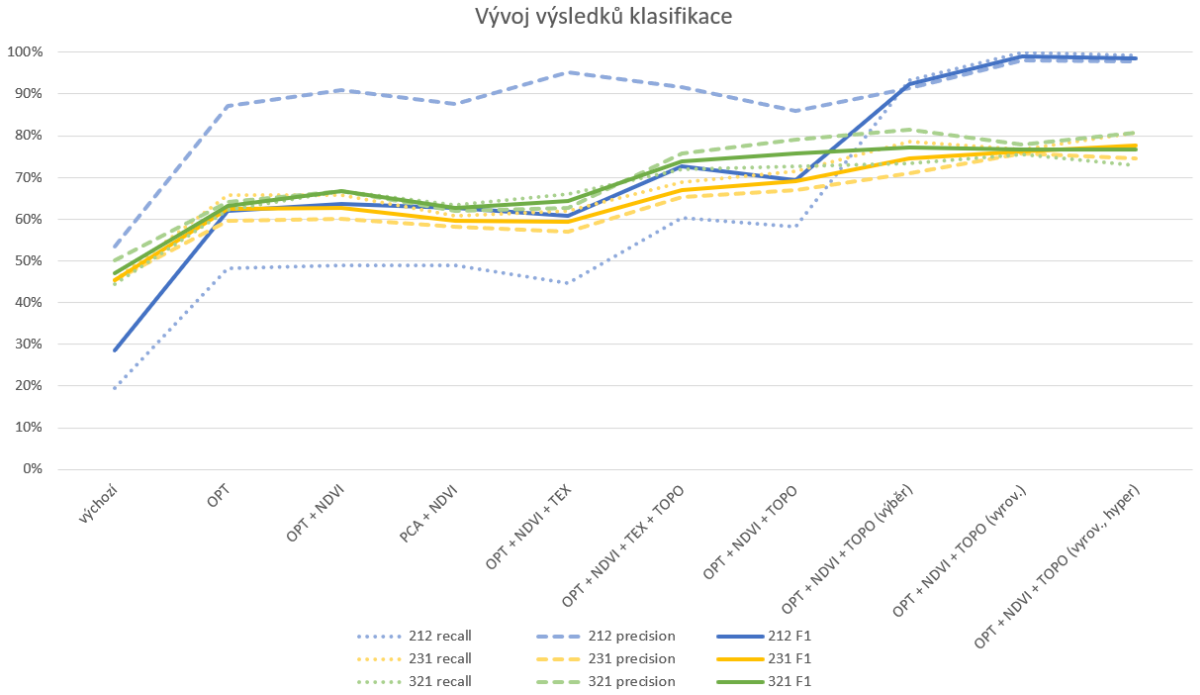
		klasifikace						
		212	231	321	celkem	recall [%]	precision [%]	F1 [%]
reference	212	264	2	0	266	99,25	97,78	98,51
	231	5	215	46	266	80,83	74,65	77,62
	321	1	71	194	266	72,93	80,83	76,68

Co se týče významných příznaků (Obr. 11) největší vliv byl zaznamenán pro sklon. Pro klasifikaci vybraných tříd významně převažují sklon a DEM, topografický aspekt měl v této oblasti největší vliv. Na dalších pozicích se nejčastěji objevily různé příznaky z června.

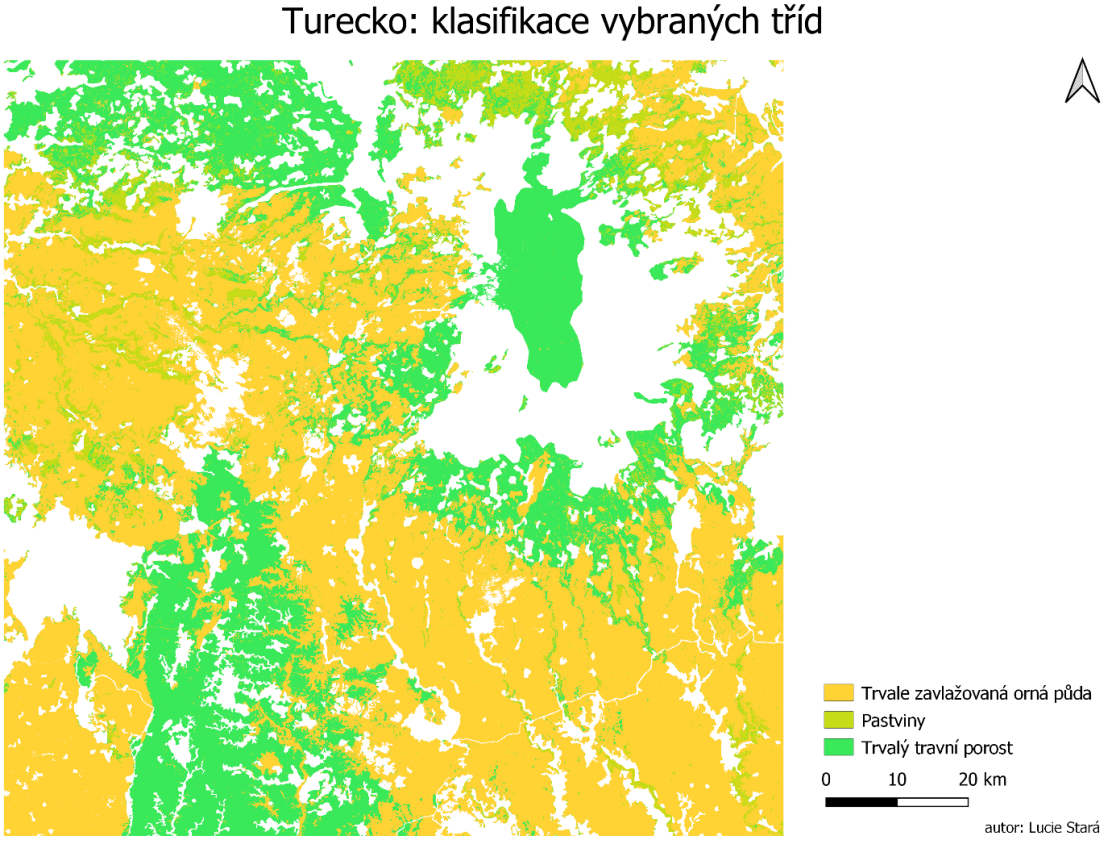


Obr. 11 Makedonie: srovnání důležitých příznaků (vybrané třídy OPT + NDVI + TOPO)

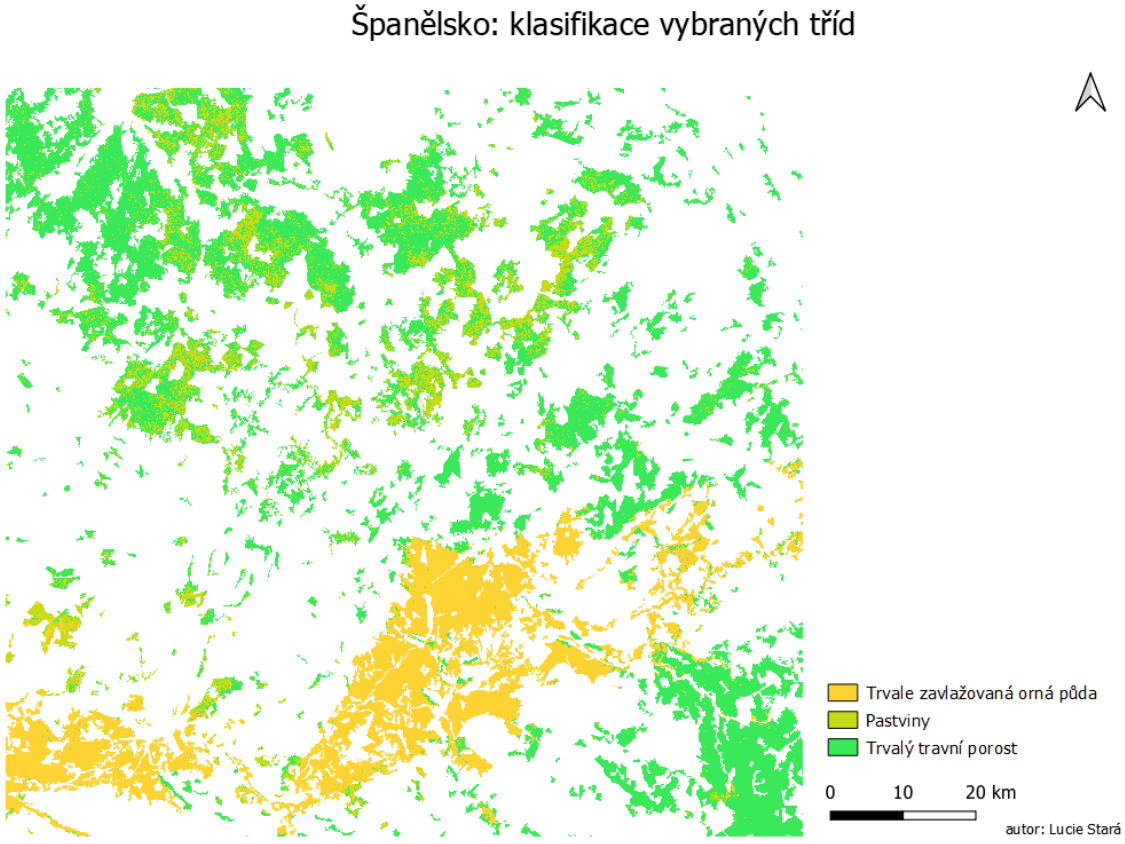
Vývoj výsledků (Obr. 12) byl v této oblasti pozvolnější, a to především pro pastviny a přírodní travní porosty. Výsledky klasifikace těchto tříd byly velmi podobné. To mohlo souviset i s faktem, že počet bodů těchto tříd byl oproti jiným lokalitám přibližně stejný. Odfiltrování ostatních tříd se významně projevilo u třídy zavlažovaná orná půda, která v přechozích výsledcích vykazovala významnou záměnu s nezavlažovanou ornou půdou. Tento jev se výrazně projevil i na hodnotách precision a recall.



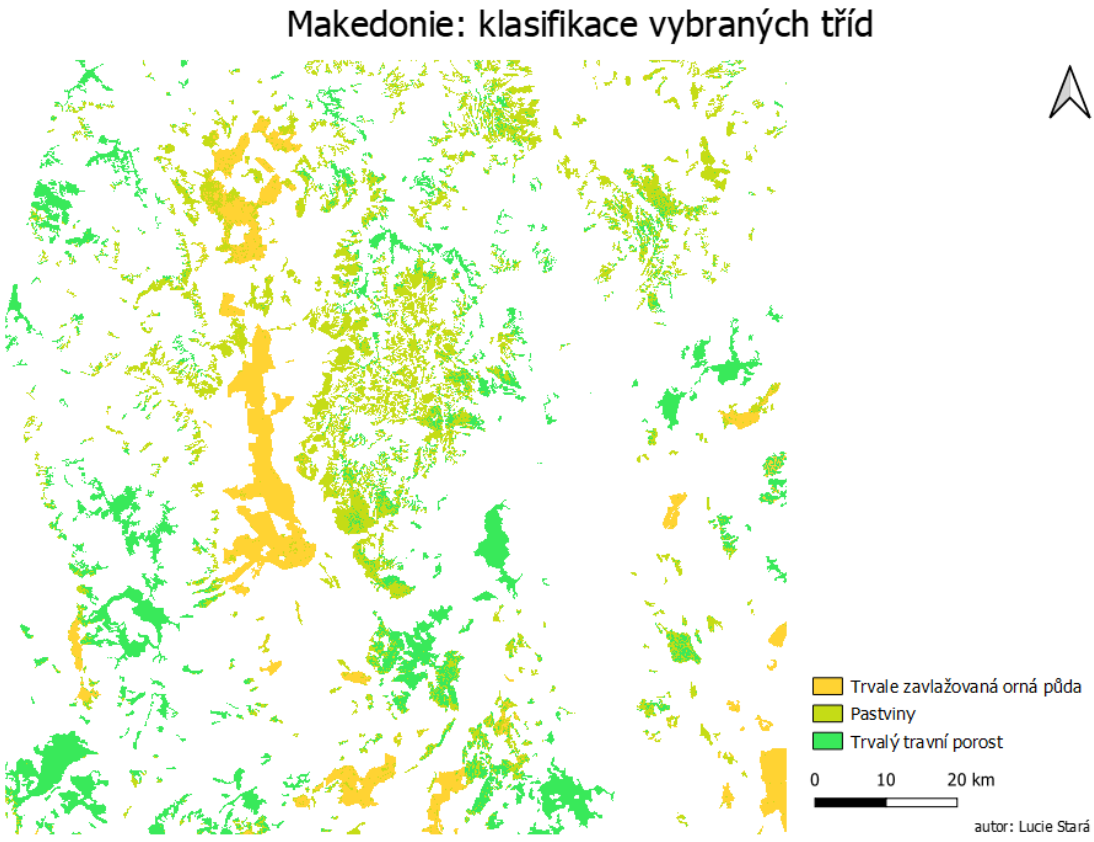
Obr. 12 Makedonie: vývoj výsledků klasifikace - vybrané třídy



Obr. 13 Turecko: výstup klasifikace (OPT + NDVI + TOPO)



Obr. 14 Španělsko: výstup klasifikace (OPT + NDVI + TEX + TOPO)



Obr. 15 Makedonie výstup klasifikace (OPT + NDVI + TOPO)

ZÁVĚR

Obdobný postup klasifikace byl aplikován ve třech lokalitách s rozdílným podnebím i topografií. Odlišnost se projevila i ve výsledcích, ačkoli některé jevy lze pozorovat shodně. Dosažené výsledky jednotlivých tříd budou zhodnoceny ve dvou rovinách. První jsou výsledky sledovaných tříd v kontextu ostatních tříd z klasifikačního schématu. V tomto případě se počet vstupních bodu odvíjí od velikosti původních trénovacích ploch CORINE a je pro každou třídu jiný. Druhá rovina je klasifikace pouze třech sledovaných tříd, ve které byl počet vzorku pro každou třídu vyrovnaný.

Co se týče referenčních dat, množství podkladových dat je jeden z aspektů, které se při tomto druhu analýzy ovlivnit nedá. Na základe dosažených výsledků však lze soudit, že výsledky jsou závislé i na počtu bodů, přesněji na poměru mezi počtem bodů jednotlivých tříd. V první části analýzy počet vstupních bodů do klasifikace nebyl nijak regulován. Odvíjel se od početnosti zastoupení jednotlivých tříd v trénovacích datech, a byl pro každou třídu různý. Počet bodů sledovaných tříd byl v závěru klasifikace vyrovnán. Tento krok nejen že přinesl významné zlepšení výsledků jednotlivých tříd, navíc se v závěru ve všech lokalitách shodně projevila jednoznačná klasifikovatelnost zavlažované orné půdy od pastvin a přírodních travin.

Pro řešení se osvědčilo použití multitemporálních dat. Uplatněna byla kombinace optických dat, kanálů NDVI a topografických informací. V případě Španělska přispěla ke klasifikaci navíc i texturální data, ačkoli v ostatních lokalitách se jejich využití neprojevilo zlepšením výsledků. Referenční data byla použita z celoevropské databáze CORINE. V klasifikaci se projevil generalizační limit spojený s MMU, kterou CORINE používá. Na tento problém byly aplikovány odpovídající úpravy.

Klasifikace byla provedena s použitím metody Random Forest. Na základě této metody byly určeny i významné příznaky. Význam jednotlivých příznaků se projevil odlišně v závislosti na lokalitě. Mezi nejvýznamnější byl ve Španělsku a Makedonii zařazen DEM, v Makedonii k tomu ještě sklon terénu. Ze spektrálních příznaků byla ve všech oblastech nejčastější optická pásma 4 (665 nm) a 12 (2190 nm). Mezi deseti nejvýznamnějšími příznaky byly zaznamenáni také zástupci kanálu NDVI. Spíše neaplikovatelné byly příznaky texturální a kanály PCA.

Navržený postup vykázal nejlepší výsledky klasifikace vybraných tříd při vyrovnaném počtu vstupních bodů. Provedení klasifikace v různých lokalitách umožnilo odhalit společné rysy sledovaných tříd. Nejlépe klasifikovatelná se za daných podmínek ukázala trvale zavlažovaná orná půda, která dosáhla nejlepších hodnot ve Španělsku (precision 99,24 %, recall 98,25 %). Pastviny a trvalé travní porosty dosáhly sice nižších hodnot, ale hodnoty F1 těchto tříd se v závěru pohybovaly na přibližně stejných hodnotách. Nejlepších výsledků bylo dosaženo v Turecku - pastviny (precision 81,17 %, recall 89,18 %), přírodní travní porosty (precision 89,30 %, recall 80,00 %).

Ve všech lokalitách se projevila záměna mezi pastvinami a přírodními travinami a přes všechny aplikované kroky tato záměna přetrvala. V závislosti na lokalitě se pohybovala v rozpětí 8 až 27 %. Ačkoli se sledovanou záměnu mezi třídami nepodařilo zcela eliminovat, bylo pro sledované třídy dosaženo uspokojivých výsledků.

Na základě dosažených výsledku lze říci, že použití CORINE jako podkladových dat s adekvátními úpravami se osvědčilo. Pro tyto polygony byl navržen způsob výběru trénovacích a testovacích dat (bodů) s využitím přístupu strojového učení. Pomocí metody Random Forest byla provedena klasifikace i výběr vhodných příznaků, které přispěly ke klasifikovatelnosti sledovaných tříd.

PODĚKOVÁNÍ

Ráda bych poděkovala své vedoucí, paní prof. Ing. Leně Halounové, CSc., za vstřícnost a cenné připomínky, které mi v během zpracování poskytovala. Za odborné konzultace a obohacující diskuze nad řešeným problémem dekuji panu Ing. Lukáši Brodskému, PhD. V neposlední řade děkuji rodině a přátelům za podporu a trpělivost, které mi projevovali po celou dobu studia.

REFERENCE

[1] „Geo-harmonizer: EU-wide automated mapping system for harmonization of Open Data based on FOSS4G and Machine Learning,“ [Online]. Available: https://opendatascience.eu/geoharmonizer-project/. [Přístup získán 2021-11-05].

[2] „Iniciační fond fakulty stavební,“ [Online]. Available: https://web. fsv.cvut.cz/aktuality/490/. [Přístup získán 2021-11-05].

[3] T. Bouček, „Testování zpusobu klasifikace pokrytí území vybraných evropských oblastí,“ 2020.

[4] J. A. Arnfield, „Köppen climate classification,“ [Online]. Available: https://www.britannica.com/science/Koppen-climate-classification.. [Přístup získán 2021-11-05].

[5] „CORINE Land Cover — Copernicus Land Monitoring Service,“ [Online]. Available: https://land.copernicus.eu/pan-european/corineland- cover. [Přístup získán 2021-11-05].

[6] *Updated CLC illustrated nomenclature guidelines.*

[7] „Copernicus Open Access Hub,“ [Online]. Available: https://scihub.copernicus. eu/. [Přístup získán 2021-11-05].

[8] „Sentinel-2 User Handbook,“ 2015. [Online]. Available: https://sentinels.copernicus. eu/documents/247904/685211/Sentinel-2_User_Handbook. [Přístup získán 2021-11-05].

[9] „Launch of Sentinel-2B satellite,“ [Online]. Available: https://land.copernicus. eu/user-corner/events/launch-of-sentinel-2b-satellite-for-copernicus. [Přístup získán 2021-11-05].

[10] N. Pettorelli, „The Normalized Difference Vegetation Index,“ 2013.

[11] L. Halounová, Dálkový průzkum Země, Praha: Vydavatelství ČVUT, 2005.

[12] R. M. Haralick, K. Shanmugam a I. Dinstein, „Textural Features for Image Classification,“ *IEEE Transactions on Systems, Man, and Cybernetics,* Sv. %1 z %2SMC-3, č. issue 6, pp. 610-621, 1973.

[13] „GRASS GIS 7.9.dev Reference Manual,“ [Online]. Available: https://grass.osgeo.org/grass79/manuals/index.html. [Přístup získán 2021-11-05].

[14] Y. Jin, X. Liu, Y. Chen a X. Liang, „Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: a case study of central Shandong,“ *International Journal of Remote Sensing,* sv. vol. 39, č. issue 23, pp. 8703-8723, 2018.

[15] P. Kupidura, „The Comparison of Different Methods of Texture Analysis for Their Efficacy for Land Use Classification in Satellite Imagery,“ *Remote Sensing,* sv. vol. 11, č. issue 10, 2019.

[16] „Terminologický slovník zeměměřictví a katastru nemovitostí,“ [Online]. Available: https://www.vugtk.cz/slovnik/index.php. [Přístup získán 2021-11-05].

[17] „EU-DEM v1.1 Download,“ [Online]. Available: https://land.copernicus. eu/imagery-in-situ/eu-dem/eu-dem-v1.1?tab=download. [Přístup získán 2021-11-05].

[18] „EU-DEM,“ [Online]. Available: https://land.copernicus.eu/ imagery-in-situ/eu-dem. [Přístup získán 2021-11-05].

[19] W. S. McInnes, B. Smith a G. J. McDermid, „Discriminating Native and Nonnative Grasses in the Dry Mixedgrass Prairie With MODIS NDVI Time Series," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* sv. vol. 8, č. issue 4, pp. 1395-1403, 2015.

[20] S. K. Langley, H. M. Cheshire a K. S. Humes, „A comparison of single date and multitemporal satellite image classifications in a semi-arid grassland," *Journal of Arid Environments,* sv. vol. 49, č. issue 2, pp. 401-411, 2001.

[21] S. Khorram, S. A. Nelson, F. H. Koch a C. F. van der Wiele, „Remote Sensing".

[22] A. Rutkay, K. Kaan a G. Önder, „Determining The Forest Fire Risk with Sentinel 2 Images," 2020. [Online]. Available: https://dergipark.org. tr/en/pub/turkgeo. [Přístup získán 2021-11-05].

[23] H. Hashim, Z. Abd Latif a N. A. Adnan, „URBAN VEGETATION CLASSIFICATION WITH NDVI THRESHOLD VALUE METHOD WITH VERY HIGH RESOLUTION (VHR) PLEIADES IMAGERY," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* Sv. %1 z %2XLII-4/W16, pp. 237-240, 2019.

[24] L. Breiman, „Random Forests," *Machine Learning,* sv. vol. 45, č. issue 1, pp. 5-32, 2001.

[25] P. O. Gislason, J. A. Benediktsson a J. R. Sveinsson, „Random Forests for land cover classification," *Pattern Recognition Letters,* sv. vol. 27, č. issue 4, pp. 294-300, 2006.

[26] O. Tokar, O. Vovk, L. Kolyasa, S. Havryliuk a M. Korol, „Using the Random Forest Classification for Land Cover Interpretation of Landsat Images in the Prykarpattya Region of Ukraine," *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT),* pp. 241-244, 2018.

[27] „Hyperparameters of Random Forest Classifier," [Online]. Available: https: //www.geeksforgeeks.org/hyperparameters-of-random-forest-classifier/. [Přístup získán 2021-11-05].

[28] „User Guide," [Online]. Available: https://scikit-learn.org/stable/user_ guide.html. [Přístup získán 2021-11-05].

[29] „Metrics to Evaluate your Machine Learning Algorithm," [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machinelearning- algorithm-f10ba6e38234. [Přístup získán 2021-11-05].